



SORBONNE UNIVERSITÉ

ED 388 : Chimie Physique et Chimie Analytique de Paris Centre
Laboratoire CPCV (ENS - PSL, CNRS, Sorbonne Université) / Pôle de théorie

Rational Exploration/Exploitation Strategies for Drug Discovery in Low-Data Regimes

Par PABLO MAS

Thèse de doctorat

Dirigée par RODOLPHE VUILLEUMIER et MARC BIANCIOTTO

Présentée et soutenue publiquement le 29/09/2025

Devant un jury composé de :

BENOIT DEPREZ, PROF.	Université de Lille	Président du jury
NATHALIE LAGARDE, MCF-HDR	CNAM	Rapporteuse
OLIVIER SPERANDIO, CHEF DE GROUPE	Institut Pasteur	Rapporteur
ANNE-CLAUDE CAMPROUX, PROF.	Université Paris Cité	Examinatrice
FRANCESCA GRISONI, PROF.	Eindhoven University of Technology	Examinatrice
RODOLPHE VUILLEUMIER, PROF.	Sorbonne Université	Co-directeur de thèse
MARC BIANCIOTTO, DOCTEUR	Sanofi	Co-directeur de thèse

Acknowledgements

This thesis would not have been possible without the support, guidance, and kindness of many people.

First, my deepest gratitude to my supervisors. Prof. Rodolphe Vuilleumier was there from the very beginning, as my teacher when I arrived in France from Morocco in 2016, through to the writing of this manuscript. Your ability to move between quantum chemistry and AI for drug discovery while heading the ENS chemistry department is matched only by your kindness. Dr. Marc Bianciotto, your experience, rigor, mentorship, and countless hours of discussion shaped this work at every stage.

I also thank my committee, Prof. Benoît Deprez, Dr. Nathalie Lagarde, Dr. Olivier Sperandio, Prof. Anne-Claude Camproux, and Prof. Francesca Grisoni, for reviewing this manuscript and for the rich exchanges during the defense.

To all my colleagues: thank you for making these years scientifically rich and humanly warm.

At Sanofi: Dr. Bruno Filoche-Rommé, a wellspring of knowledge and, though never an official supervisor, a true mentor through whom I learned so much about the pharmaceutical world. Dr. Hervé Minoux, for our talks by the coffee machine, about the industry as much as the overnight snowfall. Dr. Jean-Philippe Rameau, for consistently trusting and championing the team's PhD students. Dr. Nicolas Muzet, for the cheerfulness, the humor, and the many films I will finally get around to watching. And to Dr. Kwame Amaning, Dr. David Papin, Dr. Corinne Terrier, Dr. Anna Gogolou, Dr. Bruno Cornet, Dr. Laurent Schio, Dr. Yann Fauricher, Dr. Francesco-Paolo Panei, Charlotte Chapellier, and Matteo Marengo, thank you for taking part, each in your own way.

At ENS: Dr. Maxime Langevin, for taking me under your wing early on. Dr. Théodore Bellon, for the coffee breaks and the board games. Juliette Fenogli, for an unforgettable Japan trip. Olaia Anton, for brightening the office. Dr. Miha Srdinšek, for the Lebanese veggie lunches. And thanks to the whole theoretical chemistry group: Dr. Damien Laage, Dr. Guillaume Stirnemann, Dr. Nicolas Cheron, Dr. Marie Labeye, Dr. Simone Pezzotti, Ari Seitsonen, Paul Guenon, Subashini Kennedy, Dr. Axel Descamps, Dr. Sélène Forget, Dr. Zakarya Benayad, and Dr. Axel Gomez.

To my friends, who made these three years what they were.

The original PPNL colloc, Niels Sageaux, Dr. Louis Barthe, and Pepe Bonet Faus, for the early days. The newer PNL colloc, Dr. Louis Barthe (again) and Dr. Nicolas Ricci, for getting me through an ACL rupture and the writing of this manuscript.

To those who became far more than colleagues: Dr. Pierre Llompart, for conversations as comforting as a glass of Chablis, and Dr. Jaime de la Fuente Diez, for the laughs and for teaching me how to get free KitKats from the vending machine.

To my friends from Descartes, ESPCI and the Collège des Ingénieurs, thank you for walking this road with me.

To my family, who supported me without always understanding what I did.

Lolie, the best big sister I could hope for; Mathias, Simon, and Jules, the brothers I never had; my aunt and uncle, with whom I began this French adventure; and my mother and father, always there for me, and the ones who bought me Science & Vie Junior as a child, perhaps the very seed of this manuscript.

To Ambre, for supporting me through thick and thin, in Paris and London.

Thank you to everyone, named or unnamed but no less appreciated, who took part in this journey.

Contents

Introduction	1
1 Drug discovery and <i>in silico</i> methods	5
1.1 The drug discovery pipeline	5
1.1.1 Modern drug discovery	6
1.1.2 Research	9
1.1.3 Clinical development	16
1.2 <i>In silico</i> tools in drug discovery	19
1.2.1 Molecular representation	20
1.2.2 Molecular modeling	24
1.2.3 Artificial intelligence	28
1.3 Focus on lead optimization	39
1.3.1 Optimization objectives	40
1.3.2 DMTA cycle	44
1.3.3 Challenges	48
2 Benchmarking molecular descriptors for low-data regime modeling	51
2.1 Challenges in QSAR/QSPR modeling	52
2.1.1 Data quality and quantity	52
2.1.2 Real-world proof models	52
2.1.3 Navigating the models jungle	54
2.2 Material and methods	54
2.2.1 Molecular descriptors	55
2.2.2 Predictive algorithms	57
2.2.3 Final set of benchmarked descriptors	59
2.2.4 Datasets	59
2.2.5 Benchmarking methodology	62
2.3 Results	65
2.3.1 A brief overview	65
2.3.2 Benchmarking single descriptors	71
2.3.3 Benchmarking combinations of descriptors	74
2.3.4 Time-based Splitting vs. Cross-validation	79

2.4	Discussion	85
2.4.1	Benefits of combining descriptors	85
2.4.2	Deep learning is not all we need	85
2.4.3	Differentiating ADMET and potency	85
2.4.4	The need for real-world proof benchmarks	85
2.5	Conclusion	86
3	Molecular prioritization toolkit and retrospective simulations	87
3.1	Enhancing lead optimization	88
3.1.1	Challenges in lead optimization	88
3.1.2	From DMTA to D(S)MTA	88
3.1.3	Active learning in drug discovery	89
3.1.4	Multi-objective prioritization toolkit	90
3.1.5	Challenge of benchmarking selection strategies	90
3.1.6	Retrospective simulation of legacy projects	91
3.2	Retrospective simulation methodology	92
3.2.1	Data	92
3.2.2	Setup of a virtual scenario	95
3.2.3	Running a virtual scenario	96
3.2.4	Defining what is a good molecule	98
3.3	Acquisition functions	100
3.3.1	Model-agnostic acquisition functions	102
3.3.2	Predictive models	106
3.3.3	Model-dependent acquisition functions	106
3.3.4	Oracle acquisition function	112
3.4	Retrospective analysis tools	112
3.4.1	Visualizing simulations	113
3.4.2	Measuring exploitation	114
3.4.3	Measuring exploration	116
3.5	Practical demonstration on the FXa dataset	118
3.5.1	Simulations	118
3.5.2	Understanding the chemical space with TMAPs	121
3.5.3	Visualizing simulated scenarios	121
3.5.4	Exploration	123
3.5.5	Exploitation	125
3.5.6	The exploitation/exploration trade-off	126
3.5.7	Impact of the predictive model	128
3.6	Conclusion	128
3.6.1	A realistic retrospective analysis framework	128
3.6.2	Methodological limitations	131
3.6.3	The need for more realistic datasets	132

4	Insights from retrospective analysis of legacy projects	133
4.1	Real industrial datasets	133
4.1.1	Exploratory data analysis	134
4.1.2	Desirable molecules	137
4.1.3	Simulation parameters	139
4.2	Visualizing the chemical space	142
4.2.1	Early stage of lead optimization	142
4.2.2	Navigating between chemical series	143
4.2.3	Exploiting, exploring, or both	143
4.3	Insights from individual analysis	146
4.3.1	Exploration	146
4.3.2	Exploitation	150
4.3.3	Exploitation/exploration trade-off	150
4.4	Aggregating information across datasets	152
4.4.1	Exploitation	152
4.4.2	Exploration	155
4.4.3	Exploitation/exploration trade-off	157
4.5	Discussion and perspectives	161
4.5.1	Exploring to better exploit	161
4.5.2	Rationalizing lead optimization	161
4.5.3	Perspectives	161
5	Conclusions and perspectives	165
5.1	Predictive modeling in a shifting low-data regime	165
5.1.1	No fancy models but more data	165
5.1.2	Temporality is key	166
5.2	Rationalizing lead optimization	166
5.2.1	Molecular prioritization toolkit	166
5.2.2	Retrospective simulation methodology	167
5.2.3	Translating virtual insights to practice	168
5.3	Challenges	168
5.3.1	Data management and data cleaning	168
5.3.2	Democratizing <i>in silico</i> tools	169
5.3.3	Convincing with explainable AI	169
5.4	The role of industry	170
5.5	Conclusion	171
	Appendix	173
	List of figures	191
	List of tables	193

List of abbreviations	195
Bibliography	197

List of conferences

Event	Type	Date	Country	Title
GGMM 2023	Poster	May 15–17 2023	France	Evaluation of active learning strategies for improving QSAR modeling: A regression-based benchmarking study
6th AI in Chemistry Symposium	Poster	September 4-5 2023	UK	Evaluation of active learning strategies for improving QSAR modeling: A regression-based benchmarking study
R&D Sanofi	Day Poster	October 19 2023	France	Exploring and exploiting the chemical space with active machine learning
MLSS 2024	Poster	March 4-15 2024	Japan	Exploring and exploiting the chemical space with active learning
CS3 2024	Poster	June 24-28 2024	France	Enhanced Molecule Selection via Active Learning in Lead Optimization for Drug Discovery
R&D Sanofi	Day Poster	October 31 2024	France	Thoughtful Risk Taking in Drug Design: Active Learning for Lead Optimization
GGMM 2025	Oral	June 10-12 2025	France	Efficient compound selection strategies in lead optimization: insights from retrospective analysis

List of publications

1. **Mas, P.**; Filoche-Romme, B.; Bianciotto, M.; Vuilleumier, R. **Back to the Future of Lead Optimization: Benchmarking Compound Prioritization Strategies.** Digital Discovery 2026. (accepted)
2. Llompart, P.; Amaning, K.; Bianciotto, M.; Filoche-Romme, B.; Foricher, Y.; **Mas, P.**; Papin, D.; Rameau, J.-P.; Schio, L.; Marcou, G Varnek, A.; Moussaid, M.; Gkeka, P. **Harnessing Medicinal Chemical Intuition from Collective Intelligence.** Journal of Medicinal Chemistry 2024. <https://doi.org/10.21203/rs.3.rs-4365958/v1..>
3. Bailey, M.; Moayedpour, S.; Li, R.; Corrochano-Navarro, A.; Kötter, A.; Kogler-Anele, L.; Riahi, S.; Grebner, C.; Hessler, G.; Matter, H.; Bianciotto, M.; **Mas, P.**; Bar-Joseph, Z.; Jager, S. **Deep Batch Active Learning for Drug Discovery.** eLife 2024. <https://doi.org/10.7554/elife.89679.2>.
4. Moayedpour, S.; Corrochano-Navarro, A.; Sahneh, F.; Noroozizadeh, S.; Koetter, A.; Vymetal, J.; Kogler-Anele, L.; **Mas, P.**; Jangjou, Y.; Li, S.; Bailey, M.; Bianciotto, M.; Matter, H.; Grebner, C.; Hessler, G.; Bar-Joseph, Z.; Jager, S. **Many-Shot In-Context Learning for Molecular Inverse Design.** arXiv 2024. <https://doi.org/10.48550/arxiv.2407.19089>.

Introduction

Context and motivation

Drug discovery is a complex, costly, and time-consuming process that aims to identify and develop new therapeutic molecules. It encompasses several distinct phases, starting from target identification and progressing through hit discovery, lead optimization, and finally clinical candidate selection. Among these phases, lead optimization stands out as particularly critical, as it involves the iterative refinement of promising molecules to enhance their efficacy, selectivity, safety, and pharmacokinetic profiles ultimately yielding a viable drug candidate.

The lead optimization phase is inherently challenging due to its multi-parametric nature. Medicinal chemists are required to simultaneously optimize multiple, often conflicting objectives. This optimization problem is now generally tackled through iterative Design-Make-Test-Analyze (DMTA) cycles. In each cycle, new chemical compounds are designed based on previous results, synthesized, experimentally tested, and analyzed to inform the next design cycle. However, these cycles frequently operate in a low-data regime, characterized by limited experimental data points due to constraints in time, resources, and cost associated with chemical synthesis and biological testing.

The advent of generative artificial intelligence tools has significantly modified the landscape of lead optimization. These tools facilitate the generation of numerous novel chemical entities, complementing and extending the chemical space traditionally explored by medicinal chemists. While this technology presents tremendous opportunities, it also introduces an unprecedented challenge: an overload of molecules to choose from. This rapid expansion in the number of designed compounds necessitates the development of rational prioritization strategies to systematically and effectively select the most promising ones for synthesis and testing.

Studying the lead optimization phase in a rigorous, data-driven manner is particularly challenging outside of the industry. The main reason is the scarcity of publicly available datasets reflecting the iterative nature and multi-parametric decision-making processes intrinsic to lead optimization. Proprietary constraints, confidentiality issues, and competitive considerations often restrict access to such data, limiting academic and public research efforts. This thesis being part of a CIFRE collab-

oration with Sanofi gives the unique opportunity to work with usually inaccessible data.

Objectives and principal contributions

This thesis begins by addressing the challenge of predictive modeling in low-data regimes under realistic conditions. As preliminary work, we conducted an extensive benchmarking study evaluating over 400 molecular descriptors combined with traditional machine learning methods and state-of-the-art deep learning models. This evaluation was performed using 121 diverse datasets extracted from legacy drug discovery projects from Sanofi, representing a wide spectrum of predictive tasks. To closely mimic the real-world conditions of the lead optimization phase, characterized by shifting chemical spaces over time, we primarily employed a time-based data splitting strategy. Additionally, we compared this approach against conventional cross-validation splits to highlight significant differences in predictive performance assessment methodologies.

Building upon the insights gained from the benchmarking study, we next developed a practical toolkit tailored for multi-objective compound selection during lead optimization. The toolkit integrates various compound selection strategies inspired by active learning, multiple-criteria decision analysis (MCDA), and the practical behaviors exhibited by medicinal chemists. These strategies collectively facilitate systematic exploration and exploitation of chemical space, enhancing the efficiency and effectiveness of the compound optimization process.

Finally, recognizing the high costs associated with prospective evaluations, we introduced a robust framework designed for retrospective performance assessment of the toolkit’s selection strategies. Leveraging historical data from real-world drug discovery projects, we developed a simulation methodology to run retrospective simulations where compound selection is guided by rational selection strategies. To rigorously analyze these simulations, we developed analytical tools capable of qualitatively and quantitatively comparing the outcomes of different virtual scenarios. This work includes the release of open-source code and publicly available datasets enriched with critical temporal information and supplementary experimental assay results, enabling the broader research community to further explore and validate these methodologies.

Thesis outline

This manuscript is structured around four primary chapters, aligned with the objectives previously mentioned:

Chapter 1 provides an in-depth overview of the drug discovery pipeline, highlighting the key phases and emphasizing the critical role of lead optimization. It also reviews relevant contemporary

in silico tools and methodologies employed in drug discovery, contextualizing the contributions presented in subsequent chapters.

Chapter 2 details the comprehensive benchmarking study of molecular descriptors, comparing classical machine learning and deep learning methodologies in low-data scenarios taking into account a distribution shift and examining the performance of an extensive set of molecular descriptors.

Chapter 3 presents the methodological framework underlying the molecular prioritization toolkit, describing its various prioritization strategies. It also outlines the retrospective simulation approach and demonstrates the analytical tools developed for evaluating prioritization strategies, showcasing their application on publicly available datasets.

Chapter 4 discusses initial findings derived from applying the toolkit and retrospective simulation methodology to twelve complete datasets sourced from legacy Sanofi projects. It highlights preliminary insights and underscores the practical implications of the prioritization strategies within realistic lead optimization scenarios.

Chapter 1

Drug discovery and *in silico* methods

Discovering a new drug is not a sprint but a marathon, it demands endurance through a long and arduous journey. In practice, bringing a single medicine from lab bench to patient bedside typically takes well over a decade and can cost on the order of several billion dollars. This journey, known as the drug discovery pipeline, winds through a series of stages: scientists first identify a potential biological target and find initial "hit" compounds, then refine those hits into promising lead molecules, test these leads extensively in preclinical studies, and finally conduct clinical trials before any regulatory approval. At every step along this path, many hopeful candidates are left behind – some prove ineffective, others reveal toxic effects or fall short on crucial properties – so that only a tiny handful of initial compounds ultimately make it to market as approved medicines. In recent years, however, researchers have gained powerful new tools to help tilt the odds in their favor: advanced computational methods now enable *in silico* experiments that complement traditional lab work. These tools can rapidly search through enormous chemical libraries, model how molecules bind to biological targets, and predict key pharmacological properties, helping to prioritize the most promising leads and flag problematic compounds early in development. In this chapter, we will first walk through the overall drug discovery pipeline, then explore how *in silico* techniques are being used within this process to make it more efficient, and finally zoom in on the crucial lead optimization stage, where a promising molecule is carefully optimized through iterative refinement into a viable drug candidate ready for further development. This PhD work focuses on this specific phase of the drug discovery process, by developing methodological and analytical tools to improve its efficiency.

1.1 The drug discovery pipeline

Drug discovery and development is a long, complex, and costly endeavor to bring a new therapeutic from lab to market. On average, it takes on the order of a decade or more (often 10–15 years) and costs around \$2.5 billion (or more) to develop a single new medicine [1]. The process involves

multiple stages, from early research (identifying a biological target and finding "hits" that affect it) through lead optimization and preclinical testing, and finally clinical trials in humans and regulatory approval [2]. At each stage, many candidate compounds are filtered out due to issues with efficacy, safety, or other properties [3]. Indeed, only a tiny fraction of initial candidate molecules will ultimately succeed as an approved drug [4, 5]. The high attrition rate reflects the stringent requirements a compound must meet in terms of potency, selectivity, pharmacokinetics, safety, and more. This section provides an overview of the modern drug discovery pipeline, with an emphasis on small-molecule drugs. We also note historical context and recent advancements, such as computational methods and AI-driven approaches, that have shaped current practices.

1.1.1 Modern drug discovery

Modern drug discovery has evolved from a trial-and-error art into a more systematic science over the past century [6]. Historically, many drugs were derived from natural sources or discovered serendipitously based on observed medical effects, often without understanding their molecular targets [7]. For example, compounds like penicillin (an antibiotic from mold) and aspirin (from willow bark) were used for their therapeutic properties long before their precise mechanisms were known [8, 9]. In the latter half of the 20th century, advances in chemistry and biology enabled more deliberate approaches. The paradigm shifted to high-throughput screening and rational design at the beginning of the 21st century: researchers could screen large libraries of synthetic compounds (often 10^3 – 10^6 molecules) to find those that elicit a desired biological effect [10]. In parallel, the rise of molecular biology and genomics expanded the repertoire of drug targets [11], specific biomolecules in the body that a drug can interact with to produce a therapeutic effect. Overall, modern drug discovery couples deep biological insights (to select the right targets and evaluate compounds on the most relevant assays) with powerful chemical and computational tools (to design or find the right drugs) [12].

1.1.1.1 Drugs

In drug discovery, the term "drug" refers to the active pharmaceutical ingredient that produces a beneficial effect. Drugs can be broadly divided into two categories that are small-molecules and biologics. Small-molecule drugs are low molecular weight organic compounds – typically on the order of a few hundred daltons – that can readily enter cells and are usually administered orally (as pills) or by other convenient routes [13]. These molecules have relatively simple, well-defined structures that chemists can synthesize and modify. Over 90% of marketed drugs are small-molecules, reflecting their pharmacological versatility and the historical focus of drug discovery on this class [14, 15]. Small-molecules can often be designed to selectively bind to a target (such as an enzyme's

active site or a receptor's ligand pocket) and modulate its function [2]. They are chemically stable and do not typically require special handling (unlike some fragile biologics). Importantly, because small-molecules are chemically synthesized, once a small-molecule drug is optimized, it can be manufactured at scale with high reproducibility. This also facilitates the production of generic drugs after patent expiry, contributing to their widespread use [16].

A small-molecule drug modulates the activity of its target, generally a protein. When targeting molecular receptors, small-molecule drugs typically function as agonists or antagonists. An agonist binds to a receptor and activates it, triggering a biological response similar to the receptor's natural ligand. Agonists can be classified as full agonists, eliciting the receptor's maximal response, partial agonists, inducing a submaximal response, or inverse agonists, which produce an opposite biological effect. Conversely, antagonists bind to the receptor without activating it, thus preventing or reducing the effect of natural ligands or agonists. Antagonists can be competitive, binding directly to the agonist's binding site, or non-competitive, binding to an alternative site and indirectly inhibiting receptor activation.

For enzyme targets, different terminology is employed. Small-molecules interacting with enzymes to reduce or block catalytic activity are termed inhibitors, preventing substrate conversion into product. Enzyme inhibitors include competitive inhibitors, which compete directly with substrates for the active site; non-competitive inhibitors, binding to an allosteric site and indirectly reducing enzyme activity; and uncompetitive inhibitors, binding exclusively to the enzyme-substrate complex, thereby halting the reaction progression. Although enzyme activators, which enhance enzymatic activity, do exist, they are relatively uncommon.

Biologics, on the other hand, are large, complex molecules (or even living cells) produced by biological systems [17]. These include therapeutic proteins such as monoclonal antibodies, peptides, vaccines, and nucleic acid-based therapies (like siRNA or mRNA vaccines) [18]. Biologics have risen in prominence since the late 20th century, offering treatments with extremely high specificity, for example, an antibody drug can bind to a unique antigen on cancer cells [19]. However, biologics often must be injected or infused (since proteins would be digested if taken orally) [20], and they typically require more complex development and manufacturing processes than small-molecules [21]. Biologic drugs also pose challenges like immunogenicity (potential to trigger immune reactions) [22]. Despite these differences, both small-molecules and biologics follow a similar high-level pipeline: discovery, optimization, testing, and approval [5].

Figure 1.1 illustrates the annual number of FDA-approved drugs since 1997, categorized by their respective classes, alongside the proportion represented by small-molecules. Although the significance of biologics has steadily increased over recent years, small-molecules continue to comprise the majority of approved drugs each year. The research presented in this manuscript will specifically

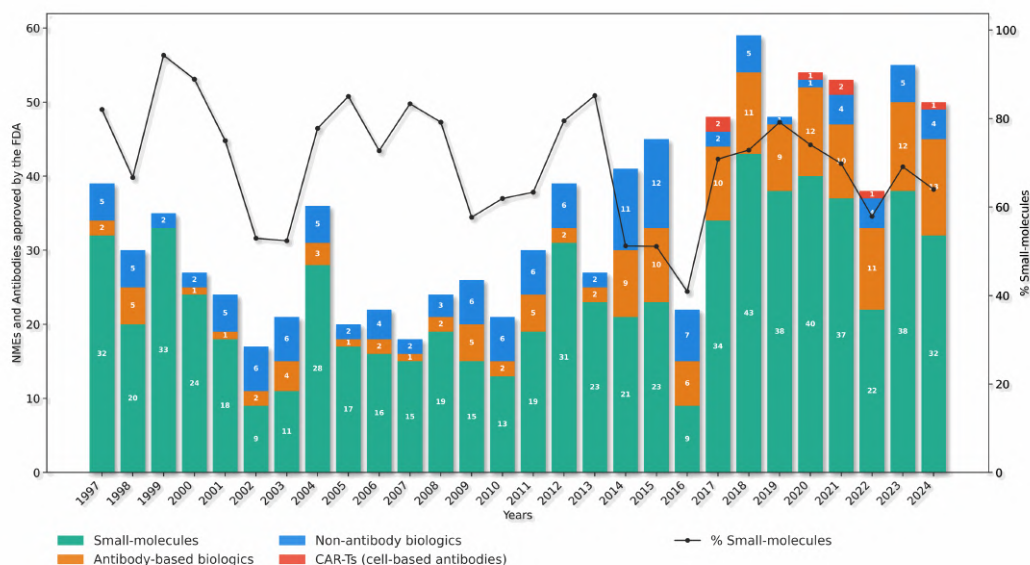


Figure 1.1: Number of FDA approved drugs per year, with a distinction between small-molecules, antibody biologics, non-antibody biologics and CAR-Ts. The proportion of small-molecules approved per year is also displayed. Data from the FDA website.

address small-molecule drug discovery, but could also be applied to biologics.

1.1.1.2 Targets

Every drug achieves its therapeutic effect by interacting with one or more biological targets. Proteins represent the overwhelming majority of druggable entities, accounting for more than 90% of approved therapeutic targets [15]. Common classes of protein targets include enzymes, G-protein coupled receptors (GPCRs) and other cell-surface receptors, ion channels, and transporters [15, 23]. Drugs act on these targets through various mechanisms: enzymes can be inhibited or modulated; receptors can be blocked by antagonists or activated by agonists; ion channels can be opened or blocked; and transporters can be inhibited or facilitated. Targets of marketed drugs are found across diverse protein families; for instance, numerous antihypertensive medications target GPCRs or ion channels, while many anticancer agents target protein kinases involved in cell proliferation. Figure 1.2 depicts the distribution of human protein drug targets across major protein families and highlights the significant proportion of small-molecule drugs targeting GPCRs, ion channels, and nuclear receptors.

While proteins dominate, other biomolecules can also be targeted by drugs. For example, some drugs bind to nucleic acids (DNA/RNA) or alter gene expression. Therapies targeting DNA/RNA include certain anticancer agents that intercalate DNA, as well as the newer class of oligonucleotide drugs that bind RNA [24, 25]. These remain a minority of drugs, but their importance is growing

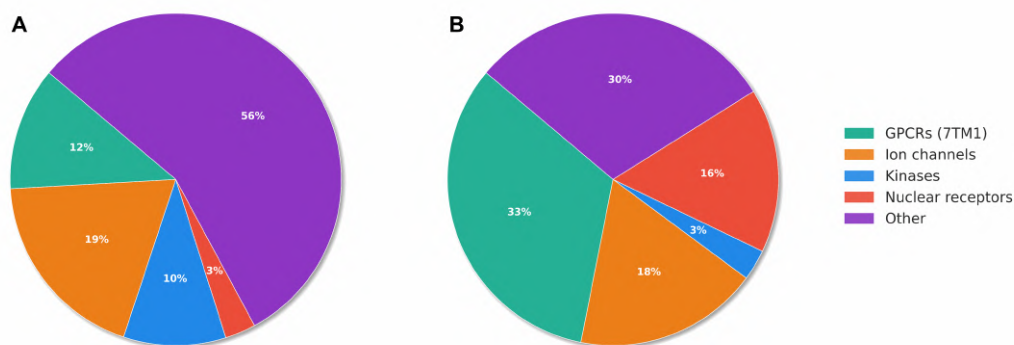


Figure 1.2: Distribution of human protein drug targets by major families (A) and distribution of small-molecule drugs targeting these major families (B), as reported up to 2017. Figure reproduced from Santos et al. [15].

(e.g., antisense therapies, RNA interference, etc.). Regardless of molecular type, a good therapeutic target is one that plays a causal role in disease pathology in human and can be modulated safely. Target selection is typically disease-driven: researchers choose a target believed to be critical in a disease process, with the hypothesis that a drug affecting that target will alter the disease course. For example, in Alzheimer's disease a long-standing target has been beta amyloid (a protein that forms plaques in the brain); drugs have been developed to prevent plaque formation or to clear existing plaques [26]. In infectious diseases, targets are often essential proteins of a pathogen (like an enzyme vital for a bacterium's survival). In cancer, many drugs target proteins that control cell division or survival (to selectively kill or stop them from duplicating). Target selection and validation are crucial, choosing the wrong target (one that is not actually central to the disease or cannot be modulated without causing toxicity) is a common reason for failure later in development [27].

1.1.2 Research

The "research" phase of the pipeline encompasses the early discovery and optimization of a potential drug. This phase can be further broken down into several key stages: target identification and validation, hit identification (finding initial active compounds, or "hits"), hit-to-lead transition (refining hits into promising lead compounds), lead optimization (iteratively improving those leads' properties), and preclinical studies (testing the optimized candidate in biological models before human trials) [2, 5]. Each of these stages are described in Figure 1.3 and in the following sections. Notably, these activities are often iterative and overlapping. For example, insights from lead optimization might loop back to identify new hits, or additional targets might emerge. Throughout the research phase, an interdisciplinary team (medicinal chemists, biologists, pharmacologists, computational modelers, etc.) works in cycles of design, synthesis, and testing to identify a promising drug candidate.

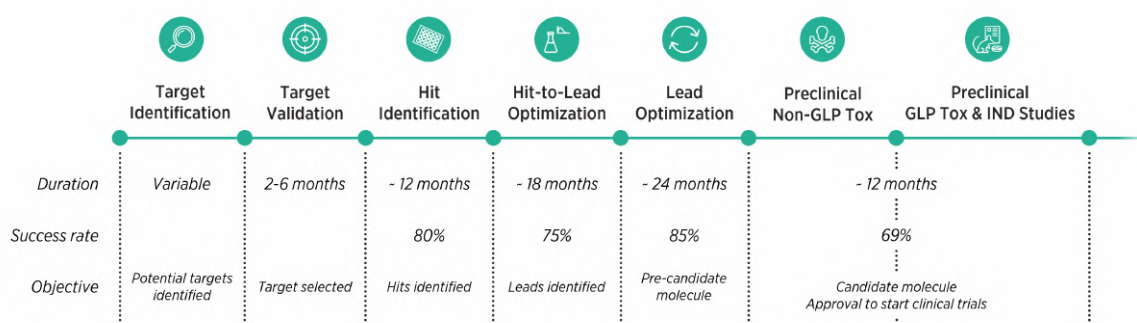


Figure 1.3: Drug discovery pipeline illustrating key stages from target identification to preclinical studies. Success rates and durations are estimated based on Paul et al. [5] and may vary depending on the specific project context.

1.1.2.1 Target identification and validation

Target identification is the process of pinpointing a biological molecule (often a protein, as previously stated) that could serve as a leverage point for treating a disease [2]. This often arises from fundamental research. For instance, scientists might discover that a certain protein is mutated or overactive in a disease, or that knocking out a gene in an animal model has a beneficial effect [2, 28]. Once a prospective target is identified, target validation is the critical step of proving that modulating this target actually impacts the disease in the desired way. Validation can involve multiple approaches: genetic techniques (e.g. deleting or silencing the gene in cell or animal models to see if disease traits are reversed) or use of tool compounds or antibodies (to pharmacologically inhibit or activate the target in models). Only if target modulation shows efficacy in preclinical models, and preferably a lack of unacceptable side effects, is the target considered "validated" enough to embark on a drug discovery program.

This stage is time-consuming and intellectually challenging, because biology is complex and disease mechanisms are not always fully understood [5]. A significant portion of drug failures can be traced to inadequate target validation. The drug may effectively hit the target, but if the target itself was not truly driving the disease, the drug won't help patients [27]. In fact, analyses of development failures have found that lack of efficacy (often due to the wrong target or pathway being addressed) is one of the largest cause of drug attrition, especially during development [27]. Therefore, modern drug discovery puts great emphasis on choosing the right target up front. New technologies have bolstered target identification and validation in recent years: for example, genome-wide association studies (GWAS) can implicate genes in disease risk [29]; CRISPR-based screens can identify which genes are essential for certain disease processes [30]; and advanced omics (proteomics, transcriptomics) can reveal dysregulated pathways [31]. Chemical biology also plays a role: researchers use small "probe" molecules to test in cells or animals whether a hypothesized target, when inhibited,

yields a therapeutic-like effect [32, 33]. If the target passes these tests (showing a clear disease-modifying role with tolerable on-target effects in models), the stage is set to start searching for actual drug molecules against that target.

1.1.2.2 Hit identification

Once a validated target is in hand, the next challenge is to find molecules that can interact with the target in the desired way. For instance, this might involve inhibiting an enzyme's activity or blocking a receptor. These initial active molecules are called hits. There are several strategies for hit identification:

- **High-throughput screening (HTS):** A classical approach is to perform an assay (a biochemical or cell-based test) on a large library of diverse compounds to see which ones produce the desired effect, such as inhibiting the target's activity above a threshold [10]. Modern robotic HTS can test hundreds of thousands of compounds relatively quickly [34]. This approach is essentially empirical: it casts a wide net to find any "hit" that shows activity. Hit rates in a random HTS are typically quite low, often around 1% or less of the compounds screened show significant activity [35]. However, given a large library, it is still possible to find a number of hits to start with. The diversity of the screening library is important; it may include synthetic compounds collected over years, natural product extracts, or combinatorial chemistry libraries. Historically, pharmaceutical companies amassed collections of up to a few million compounds, though even a million-compound library covers only an infinitesimal fraction of the estimated 10^{60} possible drug-like molecules [36].
- **Virtual screening:** Instead of (or in addition to) physically testing compounds, researchers commonly use computational methods to triage compounds before experimental screening. If the 3D structure of the target (usually a protein) is known from X-ray crystallography or cryo-EM, one can do structure-based virtual screening: computer programs "dock" each compound into the binding site of the target and score how well it might fit and bind. Docking can rapidly evaluate millions of virtual compounds, filtering out those that are geometrically or chemically incompatible with the target [37, 38]. This enriches the pool of compounds that are actually tested in the lab, focusing on those most likely to bind. Structure-based virtual screening has become a standard step when a high-resolution target structure is available. On the other hand, if a detailed binding site is not known or the target is less defined, ligand-based virtual screening can be used: here, the idea is to use known active molecules (perhaps from literature or earlier screens) as models, and then search databases for similar compounds that might also be active. For example, given one hit with a certain substructure, a computational similarity search might find analogs containing that substructure or

with similar physico-chemical properties. Ligand-based methods can employ 2D fingerprints or 3D shape/pharmacophore matching to pick molecules that resemble known actives in key features [39]. Virtual screening (both structure-based and ligand-based) greatly expands the pool of candidates by evaluating compounds *in silico* beyond what may be immediately available; some vendors even offer "make-on-demand" compound libraries [40] numbering in the billions that can be virtually screened, after which top-scoring molecules are synthesized only if predicted to be active. Virtual screening can be enhanced using active learning approaches, where the screening is iterative, each iteration being optimized based on the previous' outcomes [41, 42].

- **Phenotypic screening:** In some cases, drug discovery begins without a predefined target, by screening compounds in cell or organism models for a desired phenotype (e.g., cell death, reversal of disease markers). This was the classical mode of drug discovery in earlier decades and has seen a renaissance for diseases where the exact targets aren't clear [6, 43, 44]. Phenotypic screening yields active compounds first, and later efforts (using techniques like affinity pull-down or CRISPR resistance screens) identify what the compound's target is. While our structured pipeline assumes target-first (the dominant paradigm now), it is worth noting that many first-in-class drugs have come from phenotypic approaches [45]. Phenotypic hit identification still requires going through the subsequent optimization stages, but the path starts from an observed effect rather than a chosen target.

Regardless of the initial screening method (such as high-throughput screening, virtual screening, or phenotypic assays), identified hits must undergo a validation process. Typically, a small number of hits (ranging from a few to a few dozen) are selected based on their biological activity and novelty. However, these compounds often present several liabilities. They may exhibit weak potency, poor selectivity, instability, or other undesirable characteristics, making them suitable as starting points rather than as viable drug candidates. Chemists carry out hit confirmation and triage by re-testing compounds, verifying their chemical structures, and discarding false positives. Some false positives arise from assay interference rather than genuine target engagement, as is the case with PAINS (Pan Assay Interference Compounds) ??.

1.1.2.3 Hit-to-lead

After identifying confirmed hits, the subsequent phase involves refining these into lead compounds. A lead is typically defined as a molecule or series of related molecules demonstrating improved drug-like characteristics [2], thus warranting extensive optimization. The transition from a hit to a lead involves exploring the chemical space around the original hit to check that its liabilities can be

modulated and enhance its essential properties. This process usually proceeds iteratively through cycles of design and experimental evaluation [46].

Central to the hit-to-lead transition are structure-activity and structure-property relationship (SAR/SPR) analysis. During this step, chemists synthesize a series of analogs, which are compounds structurally similar to the hit, and evaluate their biological activity to understand how modifications influence potency and other desirable properties [47]. The primary objective of SAR analysis is to identify structural elements vital for maintaining or enhancing activity and to determine which regions of the molecule can be altered without loss of efficacy or selectivity. Genuine hit series typically exhibit discernible SAR/SPR patterns, whereby specific molecular changes consistently affect activity or other properties positively or negatively.

However, researchers must remain vigilant about assay artifacts at this stage [48, 49]. For example, a uniformly active analog series ("flat SAR") often indicates non-specific mechanisms such as compound aggregation or general assay interference, signaling potential false positives. Such artifacts are typically identified and eliminated through more rigorous experimental validation, including counter-screening against structurally or functionally related targets to confirm selectivity.

During hit-to-lead, teams often explore multiple chemical series or chemotypes (different chemical scaffolds) in parallel if they have more than one promising hit series [46]. They will compare these series on various criteria to decide which to advance. Key factors in selecting a lead series include:

- **Potency:** How strong is the effect on the target?
- **Selectivity:** Does it preferentially affect the target versus other related proteins? Early selectivity assays can flag pan-assay interference compounds or broadly toxic compounds.
- **Synthetic accessibility:** Can chemists readily make analogs and scale up this series? A chemotype accessible by straightforward chemistry is preferable.
- **Intellectual property (IP):** Is the chemotype novel and patentable? If the hit is identical or very similar to a competitor's compound or a known drug, it may not be viable due to IP restrictions.
- **Preliminary ADME/Tox:** Simple tests for absorption, distribution, metabolism, and excretion (ADME) properties, as well as toxicity assays (e.g., cytotoxicity in cell lines), can be performed on hits. For example, if a series is extraordinarily insoluble or is rapidly metabolized in liver microsomes, it might be less attractive. If a series has toxicophores (structural features associated with toxicities), medicinal chemists might favor different series.

- **Hypothesized optimization potential:** Does a series have "room" for further optimization? If the initial hit is already as bulky or polar as can be tolerated, it may be hard to improve its properties without losing activity. A lead that's moderately active but tunable is often better than one that's very potent but limited for improvements.

By the end of hit-to-lead, the goal is to have one or a few lead series that show acceptable potency on target and initial signs of drug-like behavior. The lead should be one for which there is a clear path to optimize remaining issues (such as metabolic stability or selectivity). At this point, the project enters the more intensive lead optimization phase.

1.1.2.4 Lead optimization

Lead optimization is the stage where lead compounds are turned into candidates suitable for clinical trials [2]. Here the emphasis shifts to multi-objective optimization – the compound must not only be potent at the target, but also have a whole profile of properties acceptable for a drug (pharmacokinetics, safety, formulation, etc.) [50, 51, 52]. This is a challenging balancing endeavor because improving one property can often worsen another [50]. For example, increasing potency by making a molecule more lipophilic might reduce its solubility or increase toxicity [53, 54]. Thus, lead optimization is truly a multi-objective optimization problem where the team should exploit and explore the chemical space around the initial leads in order to have at least one drug candidate suitable for preclinical development [50, 55].

The research presented in this manuscript specifically aims to enhance the efficiency of this critical phase through computational methods. A detailed section addressing the mechanics, methodologies, and principal challenges of lead optimization will be presented in a subsequent section.

1.1.2.5 Preclinical development

Preclinical development refers to the testing of a drug candidate in non-human systems to gather efficacy, toxicity, and pharmacokinetic data. This phase is necessary to satisfy regulatory requirements and to make a rational go/no-go decision on whether the compound is likely to be safe and effective in humans. Preclinical studies can be divided into *in vitro* studies (experiments in test tubes, cell cultures or biochemical assays) and *in vivo* studies (experiments in living animals) [2].

During lead optimization, many *in vitro* assays would already have been done; but now they are expanded and GLP (Good Laboratory Practice) compliant studies begin for regulatory submission. For example, *in vitro* genotoxicity tests (like the Ames test for DNA mutations [56]) are performed to flag any carcinogenic risk. The candidate's interactions with drug-metabolizing enzymes are

characterized to predict drug-drug interaction potential [57, 58]. If the drug is intended to be oral, researchers examine its stability in simulated gastric fluid, its intestinal absorption in cell monolayers, etc... [59]

The *in vivo* preclinical studies usually involve testing the compound in at least two animal species (typically, one rodent and one non-rodent) to evaluate safety and pharmacokinetics in whole organisms [60]. Initial pharmacology studies should confirm that the compound engages the target and produces the expected pharmacological effect in an animal model of the disease [2, 61]. For instance, an antihypertensive drug might be tested in a rat model of high blood pressure to see if it lowers their blood pressure. Alongside efficacy models, toxicology studies are conducted to detect any harmful effects on organ systems. These include acute toxicity (giving escalating single doses to animals to determine what dose causes adverse effects), sub-chronic and chronic toxicity (repeated dosing for weeks or months to observe longer-term effects), safety pharmacology (specific studies focusing on vital systems like cardiac, respiratory, CNS side effects), and reproductive/developmental toxicity studies if needed.

A major focus is on pharmacokinetics (PK) and drug metabolism in animals: researchers measure how the drug is absorbed, how long it circulates (half-life), how it is metabolized (which enzymes break it down into which metabolites, and whether any of them are toxic or still active on target), and how it is excreted. Often, these properties differ across species, but understanding them in animals helps predict human dosing and any risks. For example, if a drug is rapidly metabolized in mice but not in dogs, one has to consider how humans might metabolize it (*in silico* and *in vitro* models using human liver microsomes or hepatocytes aid in this extrapolation). The pharmacodynamics (PD), which describes the relationship between drug concentrations and their effects, is also characterized. This informs what blood concentration is needed for efficacy and how that compares to concentrations that cause toxicity.

A key outcome of preclinical efficacy and toxicology studies is to establish the compound's therapeutic index (the ratio of doses that cause toxicity to doses that cause efficacy). A wide therapeutic index is desirable: it means there is a comfortable safety margin between effective doses and toxic doses. Many compounds get terminated at the preclinical stage if they show toxicity (for instance, causing cancer in animal models, or organ damage) or if they have impractically poor PK (e.g., they never reach sufficient levels in blood/tissue without toxic co-solvents, or they have to be dosed too frequently).

Ethical considerations are paramount in animal testing. Researchers follow the 3Rs (Replacement, Reduction, Refinement) to minimize animal use [62, 63]. Nevertheless, at present, regulatory agencies require certain animal studies before human trials can begin, to safeguard volunteers and patients safety. Companies are continually looking for improved *in vitro* and *in silico* models (such

as organ-on-a-chip systems, human cell spheroids, and advanced computational models [64, 65]) that might predict human responses and reduce the reliance on animal tests. However, until such methods are proven equally predictive, animal studies remain a crucial gatekeeper. It's noteworthy that even extensive preclinical testing cannot guarantee human safety, but it can at least eliminate candidates with obvious hazards [66, 67].

By the end of the preclinical development phase, usually one lead compound (backed up by a close analog as a contingency if possible) is selected as the clinical candidate. This compound will have demonstrated a promising combination of efficacy (in disease models) and safety (in toxicity studies) in animals, and the data are compiled into an Investigational New Drug (IND) application for regulatory authorities. In the IND (or equivalent, like a CTA in Europe), the sponsor summarizes all the preclinical data and the manufacturing information, and proposes a plan for human clinical trials. If regulators are satisfied that the candidate looks reasonably safe to test in humans, they grant approval to proceed to the development phase.

1.1.3 Clinical development

The development phase encompasses the testing of the drug in humans (clinical trials) and the steps required to obtain regulatory approval for marketing. Whereas the research phase is largely exploratory and iterative, the development phase is highly regulated and follows a structured path: Phase I, II, and III clinical trials, leading up to a New Drug Application (NDA) or Biologics License Application (BLA) for approval [68]. Development is extremely costly and time-consuming – often constituting the majority of time and cost in the pipeline – because human trials must be conducted with rigorous controls and sufficient statistical power [1]. Many drug candidates that looked good in preclinical stages will fail in clinical trials due to lack of efficacy or unforeseen safety issues. In fact, the overall probability of success for a drug candidate from the start of Phase I trials to approval is only on the order of 10% [69, 70, 5]. Below we outline the key aspects of clinical trials and the approval process.

1.1.3.1 Clinical trials

Clinical trials are typically conducted in three main phases (I, II, III), each with a distinct purpose.

Phase I – Safety and dosing In Phase I the new drug is given to humans for the first time. The primary goal is to assess safety, tolerability, and pharmacokinetics in people, rather than efficacy (though for some severe diseases like cancer, even Phase I may look for some efficacy signals). Phase I trials are usually done in a small number of healthy volunteers or people with the disease (on the

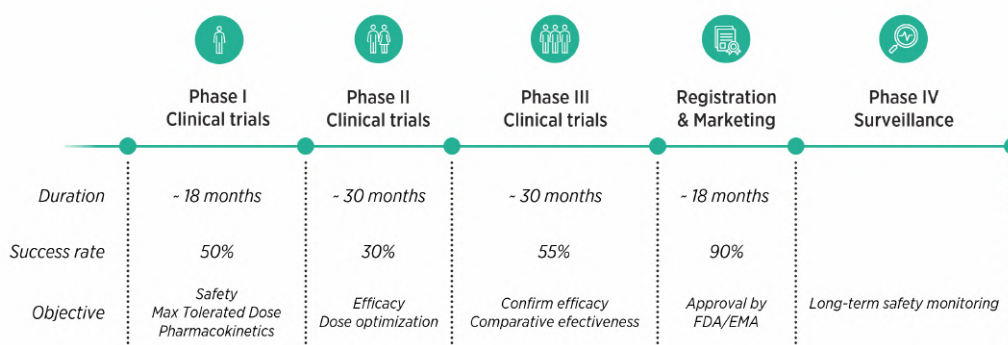


Figure 1.4: Drug development pipeline illustrating key stages from Phase I trials to Approval and Phase IV. Success rates and durations are estimated based on Paul et al. [5] and may vary depending on the specific project context.

order of 20–100 subjects). In a typical Phase I, escalating doses of the drug are given to sequential groups of volunteers under careful monitoring [71]. Investigators look for any acute adverse effects and determine the maximum tolerated dose (MTD). They also measure how the drug is absorbed and eliminated in humans (taking blood samples to quantify drug levels over time, thus assessing parameters like half-life, C_{max} , bioavailability, etc., and comparing with the animal PK to see how predictive the models were). If the drug is oral, Phase I will test different formulations if needed; if intravenous, they will examine infusion-related effects. By the end of Phase I, the hope is to have an initial safety profile and to know the dosing range that can be used for further studies (for instance, Phase I might tell us that doses up to 100 mg/kg are generally safe and achieve blood levels expected to be efficacious based on animal data). Common side effects are recorded and any dose-limiting toxicities are noted. Many drugs fail in Phase I if they show unacceptably toxic or unpredictable behavior in humans. For example, organ toxicity or effects on heart rhythm could terminate a program early [69, 70].

Phase II – Efficacy and dose refinement Phase II trials are the first trials only with patients that have the target disease, and their main goal is to explore efficacy and continue safety evaluations [68]. Phase II is sometimes split into Phase IIa (pilot efficacy or proof-of-concept) and IIb (dose-ranging). Typically, a Phase II will enroll on the order of 100 to a few hundred patients who have the condition of interest. These trials may be controlled (often against placebo, and sometimes against an active comparator). In Phase II, researchers try to answer these questions: Does the drug show signs of benefit in patients (e.g., lower blood pressure, tumor shrinkage, improved symptoms)? And what dosing regimen maximizes benefit while minimizing risk? Different dose levels are often tested (hence "dose-ranging" study) to find an optimal dose to carry forward. Efficacy in Phase II is assessed with appropriate endpoints. For some diseases, it might be a biomarker or intermediate endpoint [72], since full outcomes can take too long or require Phase III scale. Safety continues to be

monitored; with more people and longer exposure than Phase I, less common side effects emerge. It's not uncommon for a drug to show efficacy but also some toxicity in Phase II, requiring adjustments (for instance, managing side effects or narrowing which patient subgroup benefits most). Roughly speaking, only a minority of drugs that enter Phase II successfully pass it. Many fail because they lack sufficient efficacy in patients, even if they hit the target as designed [69, 70]. In fact, insufficient efficacy is a top reason for discontinuation, underscoring that translating preclinical promise to human benefit is difficult. By the end of Phase II, the sponsor aims to define the effective dose and have a sense of how to use the drug (e.g., in which population, with which concomitant therapies). If results are positive (a clear signal that the drug works at a tolerable dose), the program moves to the decisive Phase III.

Phase III – Confirmatory trials Phase III trials are large studies intended to definitively establish the drug's efficacy and safety profile for regulatory approval [68]. These trials often enroll hundreds to thousands of patients, across multiple hospitals and multiple countries, and are randomized and controlled (often placebo-controlled or against the current standard of care). The design of Phase III trials (e.g., endpoints, duration, sample size) is typically discussed with regulatory agencies beforehand to ensure that if the trial is successful, it will be considered sufficient evidence for approval [73]. Phase III trials measure clinically meaningful outcomes, such as survival rate, disease exacerbation rate, functional improvement, and others, depending on the disease type. They also collect extensive safety data from a larger population to detect rarer side effects that smaller trials couldn't reveal. Phase III trials are complex and expensive. Many drug candidates that looked promising in Phase II fail at this stage because they do not outperform placebo or existing treatments, or because previously unseen safety issues emerge when the study population is broadened [69, 70]. However, if Phase III results are robustly positive, showing that the drug provides a statistically significant and clinically relevant benefit with an acceptable safety profile, then the drug is a strong candidate for approval. By the end of Phase III, if the evidence indicates that the drug's benefits outweigh its risks for the intended use, the company will proceed to file for regulatory approval.

1.1.3.2 Approval and Phase IV

Gaining regulatory approval is the final hurdle before a drug can be marketed and prescribed to patients. In the United States, the sponsor must submit a New Drug Application (NDA) for small-molecules (or a Biologics License Application (BLA) for biologic drugs) to the Food and Drug Administration (FDA). In Europe, a Marketing Authorization Application (MAA) is submitted to the European Medicines Agency (EMA), and other regions have analogous processes. Regulators then conduct a thorough review during which expert evaluators assess the efficacy data (Does the drug actually work? How robust are the statistics?), the safety data (What are the risks? Are certain

populations at higher risk? Is risk management needed?), and the quality data (Is the drug consistently manufacturable at high purity?). If approval is granted, the drug can then be marketed for the specific indication(s) that were tested and approved.

After approval, the post-marketing (Phase IV) phase begins. The drug is prescribed to the general patient population, and companies as well as regulators continue to monitor its performance. Adverse events are reported and logged in databases; sometimes very rare side effects only become apparent when millions of people take the drug. If serious issues are found, regulators can update the drug's labeling with new warnings, restrict its use, or in extreme cases withdraw approval. On the positive side, new therapeutic uses for the drug might be discovered post-approval (leading to off-label use or new clinical trials for additional indications).

1.2 *In silico* tools in drug discovery

As the traditional drug discovery pipeline is lengthy and prone to high attrition, one strategy to improve its efficiency and success rates has been to integrate computer-aided (*in silico*) approaches into the process. These computational tools can, for example, rapidly screen large libraries of compounds and predict drug properties, helping to prioritize promising candidates and eliminate less viable ones early in development. Accordingly, the next section examines the role of *in silico* tools in drug discovery, exploring how such methods complement and enhance the conventional experimental pipeline.

Computer-aided or *in silico* drug discovery refers to the use of computational methods to design and evaluate potential drug candidates. These approaches have become integral to modern drug development, complementing traditional *in vitro* and *in vivo* experiments. In recent years, there has been a shift toward embracing computational technologies in both academia and industry [74, 12], driven by an explosion of biological and chemical data, increased computing power, and the availability of virtual libraries containing billions of compounds [75]. By leveraging fast algorithms for virtual screening and predictive modeling, *in silico* tools can substantially cut down the cost and time of early-stage drug discovery [76, 75, 77]. This section provides an overview of some key *in silico* methodologies in drug discovery. It covers how molecules are represented and featurized for computation, the main paradigms of molecular modeling (structure-based and ligand-based) including the concept of model applicability domain, and the integration of artificial intelligence techniques. Each subsection highlights fundamental concepts and methods, with references to notable algorithms, tools, and literature.

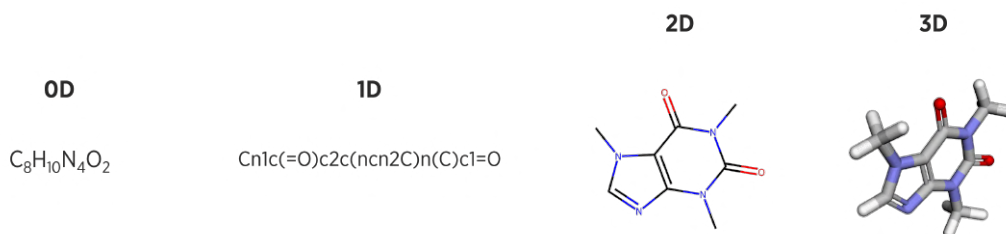


Figure 1.5: Representations of the caffeine molecule at various dimensionalities: molecular formula, SMILES, two-dimensional, and three-dimensional structures.

1.2.1 Molecular representation

A fundamental step in computational drug discovery is how to represent small-molecules in a form suitable for computer processing. Unlike the intuitive 2D chemical drawings or 3D models chemists use, computational representations encode molecules as data structures or descriptors that algorithms can manipulate. Several levels of information can be used to represent a molecule, ranging from simple formula-based encodings to complex three-dimensional models. Choosing an appropriate representation is crucial because it determines what chemical information is preserved and how effectively an algorithm can learn or make decisions from the data.

1.2.1.1 Levels of information

Molecular representations can be categorized according to their dimensionality, which reflects the level of structural detail they encode. Generally, higher-dimensional representations are more informative but less concise and computationally more demanding. Selecting the optimal level of information involves balancing task-specific requirements with available computational resources. Figure 1.5 demonstrates these different levels of representation using caffeine as an illustrative example.

Zero-dimensional (0D) Zero-dimensional molecular representations encode basic compositional information independently of molecular structure. An example of this encoding is the chemical formula, which indicates only the type and count of atoms present without detailing connectivity or spatial arrangement. While concise and easy to handle computationally, these representations lack structural context essential for predicting molecular behavior.

One-dimensional (1D) One-dimensional molecular representations encode sequential connectivity information through textual line notations such as SMILES or InChI. These encodings explicitly specify how atoms are connected within a molecule, providing more structural detail than

zero-dimensional encodings. Although compact and straightforward to store, one-dimensional representations still omit precise spatial configurations.

Two-dimensional (2D) Two-dimensional representations capture the topology of a molecule, describing its atoms and their bonds without reference to explicit spatial coordinates. In this representation, the molecule is depicted as a graph, with atoms represented by nodes and bonds by edges. Although visualized in two-dimensional space, these representations can encode three-dimensional information, such as stereochemistry, through specialized notations. Due to their balance of clarity and informational completeness, 2D representations are the most widely utilized in drug discovery and constitute the primary molecular representation employed in this PhD research.

Three-dimensional (3D) Three-dimensional molecular representations encode the spatial geometry of a molecule, encompassing atomic positions in Cartesian coordinates along with bond angles and torsion angles. This detailed spatial information is critical for accurately describing stereochemistry, conformational flexibility, and potential non-covalent interactions, including hydrogen bonding and shape complementarity. Typically, 3D coordinates are acquired either through experimental techniques or computationally generated from 2D structures using conformer generation algorithms.

Four-dimensional (4D) Four-dimensional molecular representations extend beyond spatial geometry by incorporating conformational ensembles or capturing time-dependent behavior. For instance, a 4D representation might include multiple energetically favorable conformations or track molecular trajectories derived from molecular dynamics simulations. Such representations are crucial for modeling conformational flexibility, a key determinant for accurately predicting ligand binding and induced-fit mechanisms. Within the classical mechanics framework, 4D representations provide the highest level of structural detail available for molecular modeling.

1.2.1.2 Featurization

Once a molecular representation has been selected, it typically requires featurization: conversion into a quantitative form, such as numerical vectors or computational graph structures, suitable for algorithmic processing. Featurization strategies vary widely, ranging from the calculation of pre-defined molecular descriptors to the direct use of molecular graphs. In the following paragraphs, we outline common approaches to featurization: directly leveraging the molecular graph structure, computing molecular descriptors, and generating molecular fingerprints.

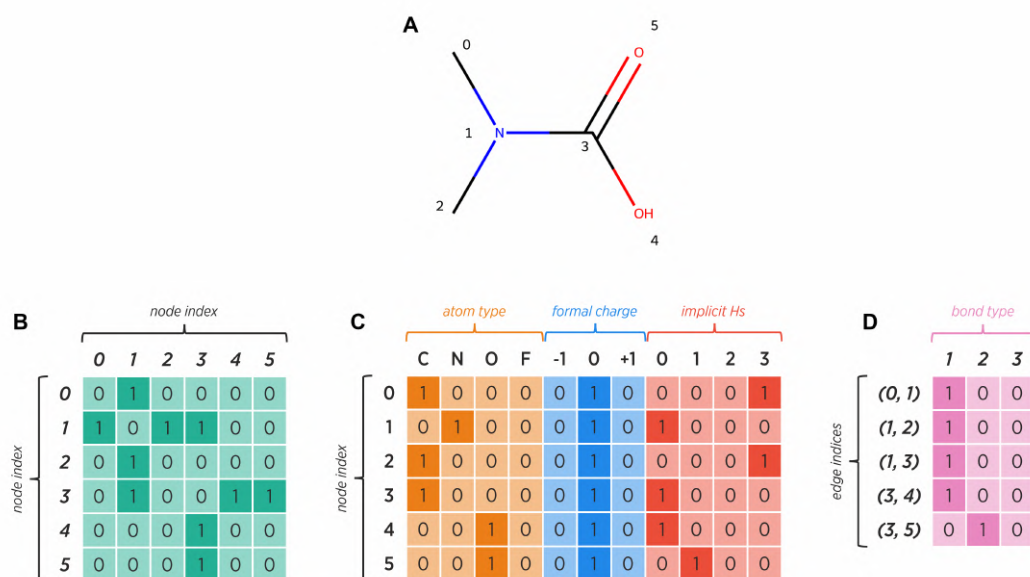


Figure 1.6: Graph representation of dimethylcarbamic acid (A), its adjacency matrix (B), node feature matrix (C), and edge feature matrix (D).

Molecular graph Representing a molecule as a graph is a natural choice in cheminformatics, as it preserves the full 2D connectivity information [78]. In a molecular graph, nodes represent atoms and edges represent bonds, as depicted on Figure 1.6. Graph representations abstract away specific coordinates and focus on which atoms are connected to which. These connections are represented under the form of an adjacency matrix (Figure 1.6B). A graph-based featurization can be as simple as this, but additional attributes can be assigned to nodes (e.g. atom type, formal charge, implicit number of hydrogens) or to edges (e.g. simple, double or triple bond) as described in Figure 1.6C and Figure 1.6D. The advantage of graph-based featurization is that it retains structural detail (all bonds and their relationships) and does not assume a fixed feature vector length a priori.

Molecular descriptors Molecular descriptors can be defined as "the final result of a logic and mathematical procedure that transforms chemical information encoded within a symbolic representation of a molecule into a useful number or result" [78]. Molecular descriptors are numerical features calculated from a molecule's representation, intended to encode various chemical properties or patterns. Descriptors can be as simple as a single number (molecular weight, atom count) or as complex as a vector of dozens or hundreds of calculated values (topological indices, physicochemical properties, etc.). Thousands of descriptors have been defined in the literature. They are often classified by the type of information used in their calculation: for example, 0D descriptors include counts of atoms, bonds, or functional groups; 1D descriptors may include fragment counts or properties like logP (octanol-water partition coefficient) which summarize whole-molecule characteristics; 2D descriptors are typically graph invariants or topological indices computed from the 2D

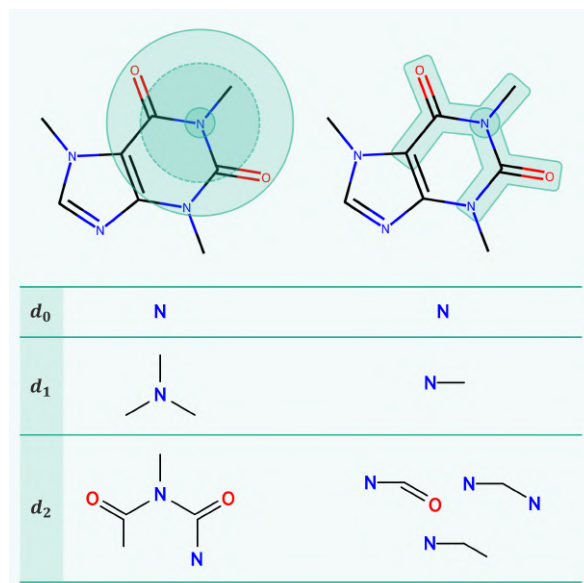


Figure 1.7: Fragments generated at difference distances starting from a nitrogen atom of the caffeine molecule using circular and linear fragmentation.

structure (e.g. molecular fingerprints – see below); 3D descriptors capture geometrical or shape information (e.g. molecular volume, surface area, 3D pharmacophore features); 4D descriptors extend 3D by considering multiple conformations or even interactions with a target, such as descriptors derived from sampling a molecule's conformational ensemble.

Molecular fingerprints Molecular fingerprints are descriptors that encode the presence or absence of structural features in a molecule, typically in the form of a binary vector or sometimes as count-based vector. Fingerprints are widely used for rapid similarity searching [79] and for featureizing molecules in ligand-based modeling [80]. There are various types of fingerprints, but most fall into two categories: structural key fingerprints and topological fingerprints.

Structural key fingerprints use a predefined list of substructural patterns (keys) and set a bit to 1 if the molecule contains that pattern. A common structural key fingerprint is the MACCS keys [81] which check for fragments like "benzene ring" or "phosphate group" etc. If the fragment is present, the corresponding bit is 1, else 0. This results in a fixed-length bit string where each position corresponds to a specific chemical feature.

On the other hand, topological fingerprints do not rely on a fixed list of patterns; instead, they algorithmically generate features based on the molecule's graph topology. Linear and circular fingerprints are the two major classes of topological fingerprints (Figure 1.7).

Linear (or path-based) fingerprints represent molecules by encoding linear sequences of atoms and bonds, commonly referred to as "paths." These paths are typically limited in length (e.g., 1 to 7

atoms) and are enumerated throughout the molecular graph. Each unique path is translated into a hashed identifier and used to set corresponding bits in a fixed-length binary vector. This approach captures detailed information about molecular connectivity and is especially useful for detecting functional groups or recurring motifs. However, path-based fingerprints can be sensitive to small topological changes and may struggle to represent local atomic environments compactly if they are not part of a linear sequence.

Circular fingerprints, such as the widely used Extended Connectivity Fingerprints (ECFPs) [82], focus on atom-centered substructures defined by a given radius. Starting from each atom, the algorithm iteratively collects and hashes information about neighboring atoms and bonds up to a certain distance. Each resulting substructure (or "atom environment") is encoded into the fingerprint vector. These fingerprints are particularly effective at capturing local chemical features and are more robust to small molecular changes than path-based ones.

1.2.2 Molecular modeling

Molecular modeling comprises computational techniques designed to simulate or predict molecular behavior or properties [83]. Typically, molecular modeling methods are classified into two categories: structure-based approaches, which require the three-dimensional structure of the biological target, and ligand-based approaches, which depend exclusively on information derived from known ligands. Although both methodologies aim to identify or optimize compounds with desired biological activities, they leverage different sources of information. A crucial aspect to consider in all modeling methods is their applicability domain, defined as the region of chemical space within which a model's predictions are deemed reliable. In the following sections, we first outline structure-based and ligand-based approaches and subsequently discuss the concepts of applicability domain and activity cliffs.

1.2.2.1 Structure-based modeling

Structure-based drug design (SBDD) utilizes the 3D structures of biological targets, typically protein binding sites, to facilitate the discovery of new ligands. High-resolution structural information obtained experimentally through methods such as X-ray crystallography or cryo-electron microscopy (cryo-EM), or computationally via predictive approaches like AlphaFold[84], enables a detailed, atomic-level comprehension of ligand-target interactions. Prominent techniques within structure-based drug design include molecular docking, molecular dynamics simulations, free energy perturbation calculations or pharmacophore modeling. A concise overview of these methodologies is provided in Table 1.1. However, this PhD research primarily focuses on ligand-based approaches, which are detailed in the subsequent sections.

Method	Description
Molecular Docking	Predicts ligand binding modes and affinities by computationally fitting molecules into the binding site of a target protein.
Molecular Dynamics	Simulates atomic movements over time to explore protein-ligand flexibility and stability in a dynamic environment.
Free Energy Perturbation	Calculates binding free energy differences between similar ligands by rigorously sampling their interactions using statistical thermodynamics.
Pharmacophore Modeling	Identifies and represents spatial arrangements of chemical features essential for biological activity to guide ligand design.

Table 1.1: Overview of common structure-based drug design (SBDD) methods used to identify and optimize ligand interactions with biological targets.

1.2.2.2 Ligand-based modeling

Ligand-based drug design (LBDD) depends primarily on information derived from known active ligands rather than structural data about the target. This approach is particularly valuable when the target structure is unavailable or insufficiently resolved. Additionally, ligand-based methodologies effectively predict various molecular properties beyond biological activity, such as pharmacokinetic endpoints. Central to ligand-based modeling is the similarity principle, which posits that chemically similar molecules typically exhibit comparable biological activities. Frequently employed ligand-based methods include Quantitative Structure-Activity Relationships (QSAR) modeling, Quantitative Structure-Property Relationships (QSPR) modeling, similarity searching, and *de novo* ligand design.

QSAR/QSPR modeling QSAR and QSPR modeling constitute fundamental techniques within ligand-based drug design [85]. These methods employ statistical and machine learning algorithms to establish predictive relationships between molecular structures, represented through descriptors, and their corresponding biological activities or physicochemical properties. The QSAR concept originated in the 1960s from Hansch’s foundational work [86, 87], which linked substituent hydrophobicity and electronic parameters to biological activity. Contemporary QSAR methodologies now leverage extensive descriptor sets and advanced, non-linear machine learning algorithms (further discussed in the next section on artificial intelligence). QSAR models serve various purposes, including virtual screening of extensive compound libraries to identify potentially active molecules, or guiding the lead optimization phase by modeling analogues around a core scaffold to prioritize structural modifications that enhance potency or optimize drug-like characteristics [88, 89, 83, 90]. In this thesis, QSAR and QSPR terms are utilized interchangeably.

Similarity search In similarity searches, a known active molecule or a set of active molecules serve as a reference against which a database of molecules is compared using a chosen similarity metric [91, 92]. These metrics typically rely on molecular fingerprints or other descriptors. In cheminformatics, a commonly employed metric is the Tanimoto similarity, also known as the Jaccard index in other domains. For two binary molecular fingerprints, A and B, the Tanimoto similarity $\delta_{A,B}$ is defined in Eq. 1.1:

$$\delta_{A,B} = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1.1)$$

Similarly, the Tanimoto distance can be defined as $1 - \delta_{A,B}$. Molecules exceeding a predefined Tanimoto similarity threshold are presumed to share similar biological activities with the reference compounds. Consequently, similarity searches are widely used in drug discovery to identify novel hits structurally analogous to known actives.

De novo design Ligand-based *de novo* design involves the creation of entirely new molecular structures guided by predictive QSAR models [12, 93]. These models function as evaluation metrics, directing the search towards molecular structures anticipated to exhibit optimal biological activities and favorable pharmacokinetic properties. Computational strategies often employed include genetic algorithms, reinforcement learning, and generative neural networks (e.g., Variational Autoencoders or Generative Adversarial Networks). These methodologies systematically explore chemical space, optimizing molecular properties iteratively and efficiently generating novel molecules that potentially address specific drug design challenges [94, 95, 96, 97, 98, 99, 100].

1.2.2.3 Applicability domain

No predictive model in drug discovery is universally reliable for all of the chemical space [101]. The applicability domain (AD) of a model defines the region of chemical space for which the model is considered valid. In simpler terms, it answers the question: for which new compounds can we trust this model’s prediction? The AD is typically determined by the characteristics of the training set compounds used to build the model [102, 103, 104, 105]. If a new molecule is very dissimilar to all the molecules on which the model was trained, the prediction is an extrapolation that might be unreliable.

A simple approach for defining an AD is descriptor range: ensure that for each descriptor, the new compound’s value lies within the range (or a tolerance around it) seen in the training set [103]. A more comprehensive approach is distance-based: for example, calculate the distance (Euclidean, Mahalanobis, etc.) from the new compound’s descriptor vector to the centroid of the training set,

or to the nearest training compounds, and set a threshold beyond which predictions are flagged as unreliable. Related to this, similarity-based AD assessment uses metrics like Tanimoto similarity (for fingerprint-based models): if the maximum similarity of a new compound to any training compound is below a certain cutoff, the compound is outside the AD [104]. More advanced techniques for defining AD include probabilistic and model-based methods such as Isolation Forest, One-Class Support Vector Machine or k-nearest neighbors [106].

1.2.2.4 Activity cliffs

Activity cliffs refer to pairs or sets of structurally similar molecules that exhibit substantial differences in biological activity [107, 108]. They illustrate how minor molecular modifications can profoundly alter activity, thus challenging the fundamental similarity principle underlying ligand-based modeling. Identifying and analyzing activity cliffs is crucial for QSAR model development [109, 110], as it reveals the sensitivity of biological responses to subtle chemical variations, guiding medicinal chemistry efforts towards more effective scaffold optimization.

Several methods have been developed to systematically detect and quantify activity cliffs. The simplest and most commonly used is the similarity-based pairwise approach, which defines a cliff as a pair of molecules displaying high structural similarity (typically a Tanimoto coefficient > 0.9) but exhibiting large differences in activity (usually > 2 log unit) [109, 110]. A more refined measure is the Structure–Activity Landscape Index (SALI) [111], which quantifies the steepness of activity changes relative to structural similarity (Eq. 1.2). High SALI values indicate pronounced activity cliffs and help in identifying discontinuities within the activity landscape.

$$\text{SALI}(m_1, m_2) = \frac{|A_{m_1} - A_{m_2}|}{1 - \delta_{m_1, m_2}} \quad (1.2)$$

where m_1 and m_2 are two molecules, A_{m_1} and A_{m_2} represent their respective biological activities, and δ_{m_1, m_2} is the Tanimoto similarity between the two molecules.

Another frequently used method for analyzing activity cliffs involves Matched Molecular Pairs (MMPs) [112]. This technique considers pairs of molecules differing by a specific, chemically interpretable transformation, as illustrated in Figure 1.8. When such minor structural changes lead to significant differences in biological activity, the pair is classified as an activity cliff [113].

Activity cliffs thus provide critical insights into structural determinants of bioactivity, allowing researchers to pinpoint influential molecular features and refine predictive models and optimization strategies.

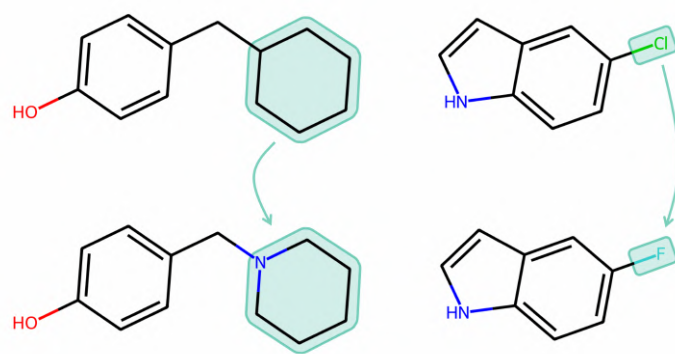


Figure 1.8: Illustration of Matched Molecular Pairs (MMPs), with minor chemical transformation highlighted.

1.2.3 Artificial intelligence

Artificial intelligence (AI) techniques have significantly transformed the landscape of *in silico* drug discovery over the past decade. These data-driven methods are capable of learning meaningful patterns from existing datasets, including known compound activities, physicochemical properties, and molecular structures, subsequently enabling predictions or informed decisions for novel compounds. Within the realm of drug discovery, AI methodologies have been effectively utilized for tasks such as predicting biological activity or designing new molecules. This section provides an overview of fundamental AI concepts relevant to ligand-based drug discovery.

1.2.3.1 Supervised learning

Supervised learning is a paradigm where models are trained on datasets consisting of input data paired with known outputs (labels). More formally, given a set of data points $X = \{x^1, \dots, x^n\}$ and corresponding outcomes $Y = \{y^1, \dots, y^n\}$, a supervised learning model is a mathematical function f that aims to accurately predict the outputs Y from inputs X :

$$Y = f(X) \quad (1.3)$$

In the context of drug discovery, input data typically comprise molecular features, such as descriptors, fingerprints, or graph-based representations, while outputs represent experimentally measured properties or activities. These outputs may include continuous values (e.g., potency or solubility) or categorical labels (e.g., active/inactive). The primary objective of supervised learning in this setting is to establish a robust mapping from molecular features to outcomes, thereby enabling reliable predictions for new, previously unseen compounds.

Regression vs Classification In supervised learning, the type of dependent variable determines whether a task is framed as regression or classification. Regression involves predicting a continuous variable $y \in \mathbb{R}$, whereas classification assigns each data point to one of K discrete classes $y \in \{1, \dots, K\}$. Within drug discovery, many outcomes, such as potency (e.g., IC_{50}/EC_{50}) or physicochemical properties like solubility, are inherently continuous, making their prediction a regression task. Nevertheless, classification is also widely employed, particularly during the hit identification phase, where compounds are commonly labeled as either active or inactive.

Machine learning Throughout this thesis, the term "machine learning" specifically refers to traditional, non-deep learning algorithms. Examples include decision trees, ensemble methods (e.g., random forests and gradient boosting), support vector machines, and k-nearest neighbors. Many of these methods gained prominence during the 1990s and early 2000s in QSAR modeling and remain extensively employed due to their robustness, interpretability, and ease of implementation [114, 85].

Decision Tree A decision tree [115] is a hierarchical, flowchart-like model that predicts outcomes by recursively partitioning data according to feature values, as illustrated in Figure 1.9. Regression trees output numerical values at their leaf nodes, whereas classification trees assign categorical class labels. Trees are constructed by selecting splits that best separate the data, guided by criteria such as information gain for classification or variance reduction for regression. Within drug discovery, decision trees represent some of the earliest nonlinear QSAR models [114], appreciated particularly for their intuitive interpretability. Users can explicitly trace decision paths to identify which descriptors influenced predictions, sometimes even deriving interpretable rules (e.g., "if molecular weight $< n$ and ring count $> p$, then active"). However, single decision trees are susceptible to high variance, often resulting in overfitting. To address this limitation, ensemble methods were developed [116].

Random Forest A random forest (RF) [117] is an ensemble method consisting of multiple decision trees, each trained independently on randomly sampled subsets of data and features at each split, as illustrated in Figure 1.10. Introduced in 2001, this approach aggregates predictions from numerous decorrelated trees, enhancing robustness and prediction accuracy. Random forests have proven particularly effective in QSAR and molecular property prediction tasks due to their capability to handle extensive descriptor sets, resistance to overfitting (given sufficient trees), and ability to estimate feature importance [114, 118]. Although RFs offer reduced interpretability compared to single trees, insights can still be extracted through feature importance metrics [119]. Thus, random forests represent a balanced and practical initial modeling choice in cheminformatics.

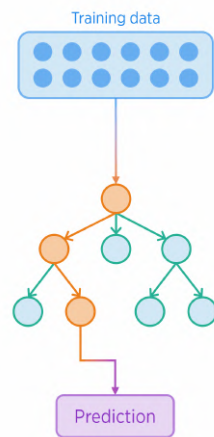


Figure 1.9: Illustration of a decision tree, highlighting its hierarchical splitting by recursively partitioning data.

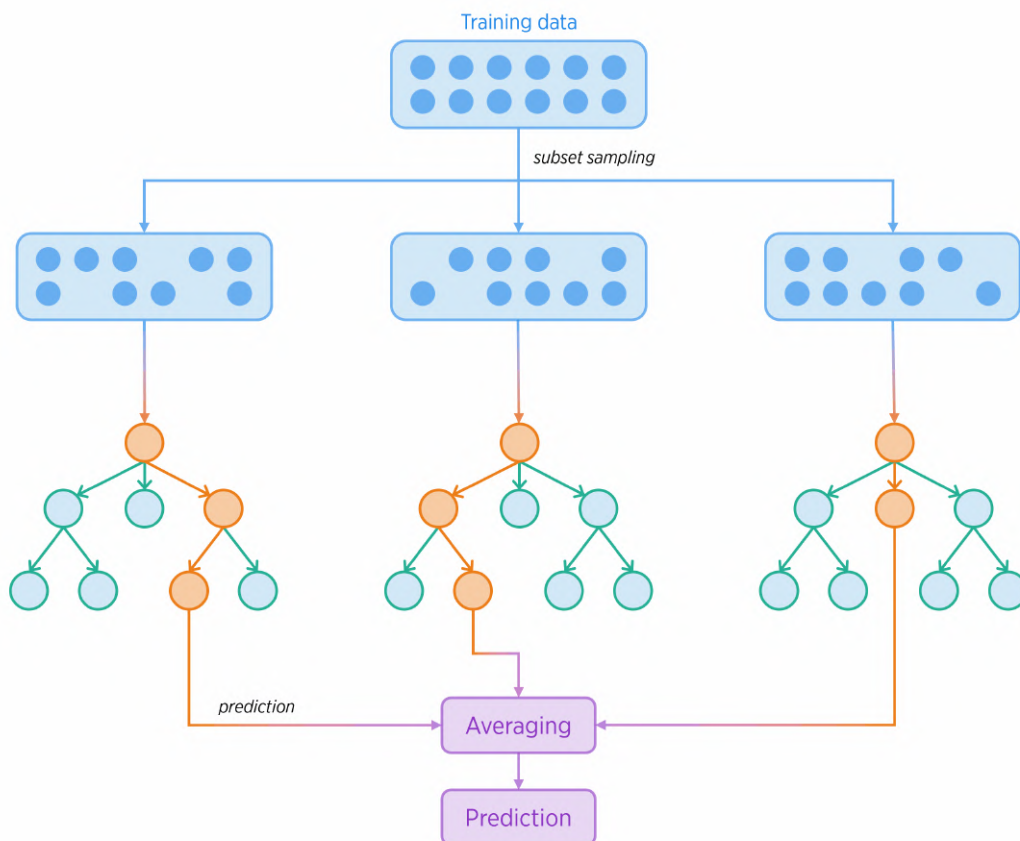


Figure 1.10: Illustration of a random forest, emphasizing the ensemble structure where predictions from multiple independently trained decision trees are aggregated.

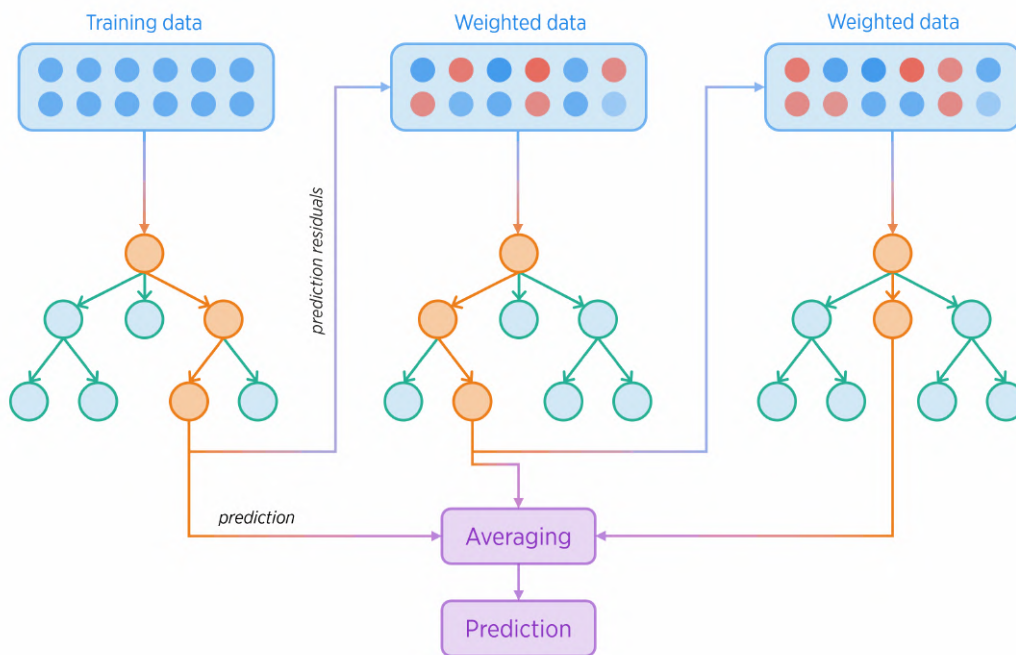


Figure 1.11: Illustration of gradient boosted trees illustrating how each tree sequentially corrects the residual errors of its predecessors.

Gradient boosted trees Gradient boosting [120] is another highly effective ensemble approach that sequentially constructs decision trees, with each subsequent tree specifically trained to correct the errors made by its predecessors, as illustrated in Figure 1.11. Unlike random forests, which generate decision trees independently, gradient boosting incrementally directs attention toward samples misclassified or poorly predicted by previous iterations. Typically consisting of numerous shallow trees, boosted models progressively capture intricate nonlinear relationships and interactions between descriptors. These methods frequently outperform other machine learning algorithms in competitions involving tabular data and demonstrate exceptional efficacy for QSAR modeling tasks [118].

Gaussian Process Gaussian Processes (GPs) [121] are probabilistic machine learning models defined by distributions over functions rather than explicit parametric forms. Formally, a GP sets a prior over functions, which, when combined with observed data via Bayesian inference, yields a posterior predictive distribution. The kernel (covariance function) is a key component, encoding assumptions about the modeled function's smoothness, periodicity, or other properties. GPs are valuable in QSAR modeling [122] due to their inherent robustness and their ability to provide uncertainty estimates alongside predictions, an essential capability in drug discovery for prioritizing molecules for experimental validation. However, GPs' computational complexity scales cubically with the number of data points, which restricts their application to relatively small datasets.

It’s worth noting that other ML methods, such as Support Vector Machines (SVM) [123, 124] and k-Nearest Neighbors (kNN) [124, 125], have also been widely used in cheminformatics. SVMs with appropriate kernels (e.g., Tanimoto kernel for fingerprints) were state-of-the-art in many mid-2000s virtual screening exercises [126]. However, in recent years, tree ensembles and deep learning have often surpassed these in performance.

Deep learning Deep learning refers to neural network-based models with multiple layers of non-linear processing units (neurons). These models can automatically learn complex feature representations from data, given sufficient data and computational resources [127]. In drug discovery, deep learning has gained prominence for its success in tasks like property prediction [128, 129] or *de novo* molecular design [128, 130]. Here we will discuss two classes of deep learning models: Multilayer Perceptron (MLP) [131] and Graph Neural Network (GNN), especially Message Passing Neural Network (MPNN) [132] as a specific example of GNN.

Multilayer Perceptron A Multilayer Perceptron [131] is a classic deep neural network architecture consisting of an input layer, one or more hidden layers of neurons, and an output layer. Each neuron in a layer computes a weighted sum of inputs from the previous layer, applies a nonlinear activation function, and passes the result to the next layer. MLPs are fully connected networks (each neuron connects to all neurons in the next layer) and can approximate complex functions given enough hidden units and layers. In cheminformatics, if we use an MLP, typically the input would be a vector of molecular features, such as a fingerprint or a set of descriptors, and the output could be a predicted activity or property value. Essentially, a MLP is a nonlinear regression or classification model that can capture interactions between features that linear models or shallow models might miss.

Neural networks were applied to QSAR problems as early as the 1990s [133, 134, 135], but they often suffered from limited data and computational limitations. Now with larger datasets and modern training techniques (better activation functions, regularization like dropout, powerful GPUs etc.), MLPs are more viable. MLPs also serve as components in larger frameworks (e.g., a generative model might use an MLP as part of its property prediction module [96]).

Additionally, MLPs and similar deep learning architectures support multi-task learning [136, 137], enabling simultaneous prediction of multiple molecular properties or biological activities. This multi-output capability can potentially enhance predictive accuracy, computational efficiency, and generalization by leveraging shared feature representations across related tasks. Figure 1.12 illustrates single-output and multi-output neural networks.

In summary, while MLPs remain foundational in deep learning for cheminformatics, their per-

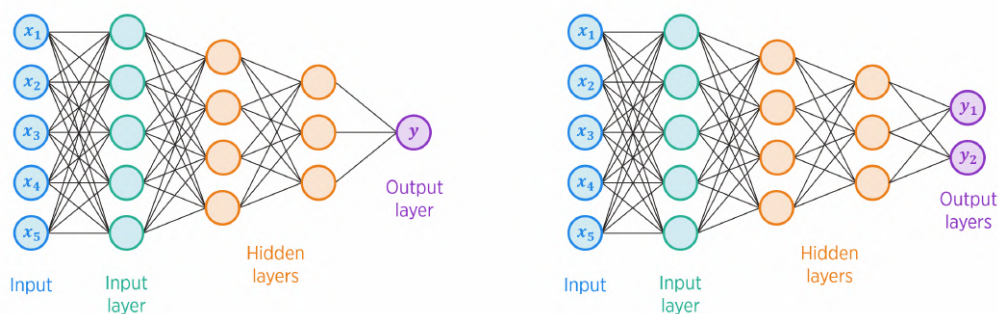


Figure 1.12: Illustration of single-output (left) and multi-output (right) neural networks.

formance is often surpassed by more specialized neural network architectures, such as graph neural networks [138, 132], which exploit molecular structure, or recurrent neural networks (RNN), which capitalize on molecular sequential representations such as SMILES.

Graph Neural Network Graph Neural Networks [139] are a category of deep learning models designed to operate on graph-structured data, which makes them particularly well-suited for molecular data (where, as discussed, a molecule is naturally represented as a graph of atoms and bonds). In contrast to feeding an array of pre-computed features into an MLP, a GNN takes the molecular graph as input directly, with initial features for each atom (and sometimes bonds) such as atom type, degree, formal charge, etc. The GNN then propagates information along the graph to compute an embedding (vector representation) for each atom and ultimately for the whole molecule.

A common framework for GNNs in chemistry is the Message Passing Neural Network (MPNN) paradigm [132]. In an MPNN, each round of message passing consists of each atom (node) receiving "messages" from its neighbors (based on the neighbor atoms' current state features and the bond features connecting them). These messages are aggregated (summed/averaged) and used to update the atom's state via a neural network function. This process is iterated for several steps (T steps), allowing information to travel farther across the graph – after T steps, each atom's state contains information about atoms in its T -bond radius neighborhood. Finally, a readout function (usually a MLP) combines the atom states to predict the property of the whole molecule. Many GNN variants have been developed (Graph Convolutional Networks, Graph Attention Networks, etc.), but most can be seen as instances of the message-passing concept.

1.2.3.2 Unsupervised learning

Unsupervised learning is a category of machine learning wherein models are trained on datasets consisting solely of input data without associated outputs (labels). More formally, given a set of

unlabeled data points $X = \{x^1, \dots, x^n\}$, an unsupervised learning model seeks to uncover inherent patterns, groupings, or structures within these inputs. Unlike supervised learning, the goal is not prediction but rather extracting meaningful representations or insights from the data, typically through methods such as clustering or dimensionality reduction:

$$Z = f(X) \tag{1.4}$$

Here, Z represents the discovered structures or latent representations derived from the input data X .

Clustering Clustering algorithms aim to group a set of objects (molecules, in our case) such that objects in the same group (cluster) are more similar to each other than to those in other groups. This is done without any examples of correct groupings (hence unsupervised). In drug discovery, clustering is used for diversity selection (picking a diverse subset of compounds) [140, 141, 142, 143], defining chemotype families, organizing screening results, or analyzing structure-activity relationships by grouping similar structures and looking at their activities. Here we present three commonly used clustering algorithms used in cheminformatics:

- **k-Means clustering:** k-means [144] is a very common algorithm that partitions the data into a predefined number of k clusters. It operates on a feature space (so one needs a numeric representation of molecules, e.g., fingerprints or descriptor vectors). The algorithm works by randomly initializing k cluster centroids, assigning each data point to the nearest centroid, then recomputing centroids as the mean of the assigned points, and iterating this process. The objective is to minimize the within-cluster variance (squared distance of points to their cluster center). K-means is relatively efficient and works well when clusters are roughly spherical in the feature space. In chemoinformatics, one might apply k-means to a set of compounds described by descriptors to identify groups of similar compounds. For example, one could cluster a chemical library into 100 clusters and then pick one representative from each cluster to ensure diversity in a screening subset.
- **k-Medoids clustering:** k-medoids [145] is similar in spirit to k-means but instead of using the mean point as the cluster center, it chooses an actual data point (a medoid) to represent each cluster. The algorithm greedily finds a set of k medoids that minimize the sum of distances of points to the nearest medoid. Because it uses actual points, k-medoids is more robust to outliers and can use any distance metric (not just Euclidean, which k-means implicitly uses for means). In chemoinformatics, one might use k-medoids with a Tanimoto distance matrix of fingerprints, which is a common approach for clustering compounds based on structural sim-

ilarity. Each cluster is then defined by a medoid compound – which is one of the compounds in the cluster most centrally located – which can be useful for selecting representatives (the medoids themselves can serve as a diverse subset). K-medoids is more computationally expensive than k-means, especially for large datasets (as it often operates on a full distance matrix), but for moderate datasets it is quite practical. It’s also applicable when an average of compounds doesn’t make sense (you can’t really “average” chemical structures in feature space meaningfully if using binary features or similarities, so medoids are more appropriate).

- **Butina clustering:** Butina clustering [146] is a hierarchical clustering algorithm widely used in cheminformatics for grouping molecules based on structural similarity. Unlike k-means or k-medoids, Butina clustering does not require predefining the number of clusters (k). Instead, it operates by first computing a pairwise distance (or similarity) matrix, often using Tanimoto similarity derived from binary molecular fingerprints. Molecular pairs that exceed a user-defined similarity threshold are considered neighbors. The algorithm then identifies the molecule with the largest number of neighbors (most connected), which becomes the center of the first cluster. All its neighbors within the similarity threshold are grouped into that cluster. The process repeats, excluding already clustered molecules, progressively forming clusters around the remaining most connected molecules until no molecules remain above the similarity cutoff. This makes Butina clustering particularly useful for selecting representative subsets of structurally diverse compounds, as it naturally forms clusters based on the intrinsic similarity distribution rather than enforcing a fixed number of clusters.

In summary, clustering helps split up chemical space into meaningful regions. This can reveal, for instance, that a screening hit list actually contains, say, 3 distinct chemotypes, which might become separate starting points for chemistry efforts. Or, in the design of a screening library, clustering can ensure coverage by selecting compounds from each cluster (diversity picking).

Dimensionality reduction High-dimensional data (like hundreds or thousands of molecular descriptors) can be difficult to analyze or visualize. Dimensionality reduction techniques aim to project such data into a lower-dimensional space (often 2D or 3D for visualization, or to a few dimensions for simplifying models) while preserving as much meaningful structure in the data as possible. Several techniques are popular:

- **Principal Component Analysis (PCA):** PCA [147] is a linear dimensionality reduction technique that finds new orthogonal axes (principal components, PCs) that successively explain the largest variance in the data. Essentially, PCA diagonalizes the covariance matrix of the features. The first principal component is the direction in feature space along which the data

vary the most, the second PC is the orthogonal direction of next largest variance, and so on. By projecting data onto the first few PCs, one can capture the majority of variance in fewer dimensions. PCA is often used as an exploratory tool: e.g., to visualize a chemical library in a 2D PC plot and see if certain clusters or trends emerge (maybe known actives cluster separate from inactives, etc.). It's also used for pre-processing – reducing collinearity and dimensionality before applying a supervised machine learning model. The limitation of PCA is that it's linear – complex nonlinear structure in data might not be captured by a linear subspace.

- **t-Distributed Stochastic Neighbor Embedding (t-SNE):** t-SNE [148] is a nonlinear technique specifically aimed at visualizing high-dimensional data in 2 or 3 dimensions. It does not preserve large-scale structure or relative distances per se; instead, it preserves local neighborhood affinities. t-SNE works by converting distances between points in high-dimensional space into conditional probabilities (similar points have high probability of being picked as neighbors) and does the same in the low-dimensional space, then it tries to minimize the mismatch (Kullback-Leibler divergence) between these two distributions. The result is that points that were close in original space tend to cluster together in the t-SNE plot, and points far apart remain far, but the actual distances and axes in the embedding are not directly interpretable beyond that. t-SNE has been very useful in visualizing chemical data – for example, one could take a set of molecules represented by high-dimensional fingerprints and use t-SNE to plot them in 2D, often revealing clusters that correspond to different scaffold families or property clusters. One must be cautious: t-SNE can sometimes create visual artifacts (e.g., it might always produce some clusters even if data is fairly uniform, due to how it emphasizes local structure) and it has a perplexity parameter that needs tuning to the data size.
- **Uniform Manifold Approximation and Projection (UMAP):** UMAP [149] is a newer nonlinear dimensionality reduction technique that, like t-SNE, aims to preserve local structure but also retains more of the global structure of the data. It is rooted in manifold learning and topology – it constructs a graphical representation of the high-dimensional data (fuzzy simplicial complex) and optimizes a low-dimensional embedding to have a similar structure. In practice, UMAP often produces plots similar to t-SNE, but it tends to better maintain some large-scale relationships and is usually faster and can handle larger datasets with ease. For example, UMAP might be used to embed millions of compound fingerprints into 2D for visualization or for pre-clustering. A noted advantage is that UMAP can be used as a general-purpose dimensionality reducer for downstream modeling (since it can preserve more structure, one could reduce dimensions and then use those for clustering or even for training models). According to its authors, UMAP often preserves more of the global geometry compared to t-SNE, meaning distances in the low-D space can be more meaningful globally (though still not as strictly as PCA). For instance, if you used UMAP on a set of compounds, the distance between two

points in the UMAP 2D plot has some correlation with their distance in original space (more than in t-SNE, where distances in the plot are not linearly meaningful). In chemoinformatics, UMAP has quickly become popular for visualizing high-dimensional chemistry data.

- **Tree MAP (TMAP):** TMAP [150] is a dimensionality reduction and visualization technique particularly well-suited for analyzing large-scale, high-dimensional datasets prevalent in cheminformatics. Unlike linear methods such as PCA or nonlinear embedding methods like t-SNE and UMAP, TMAP employs a two-stages approach using locality-sensitive hashing (LSH) combined with a minimum spanning tree (MST) to maintain both local and global data relationships. Initially, molecules are encoded as high-dimensional binary fingerprints, after which LSH efficiently approximates their similarities by clustering similar fingerprints together. Subsequently, an MST connects these clusters, forming a graph that effectively captures local neighborhoods and global data topology simultaneously. Due to its explicit graph-based representation of molecular relationships, TMAP naturally enables interactive exploration, allowing users to intuitively navigate chemical space by examining connected nodes or local neighborhoods. Nevertheless, careful interpretation is essential when assessing distances within TMAP visualizations: while local proximities accurately reflect molecular similarity, global distances between distant clusters may not strictly represent precise quantitative measures.

Figure 1.13 illustrates the four visualization methods discussed above. In this thesis, the primary visualization approach employed is TMAP due to its unique graph-based structure, which effectively captures both similarity and connectivity between compounds. This capability is especially valuable in the context of lead optimization, where chemical series commonly arise from a limited number of core scaffolds and differ by incremental structural modifications. The TMAP representation thus aids in clarifying the structural logic underlying compound relationships.

1.2.3.3 Active learning

Active learning (AL) is a machine learning approach in which the algorithm iteratively selects the most informative data points to label, thereby enhancing model performance with minimal labeling effort. The criterion used to guide this selection process is commonly referred to as an acquisition function; within this thesis, the terms "acquisition function" and "selection strategy" will be employed interchangeably.

In drug discovery, the active learning framework naturally applies to scenarios where obtaining biological data (labels) is costly or time-intensive, such as experimental compound screening in wet laboratories. Here, an active learning strategy can effectively guide the model to propose specific

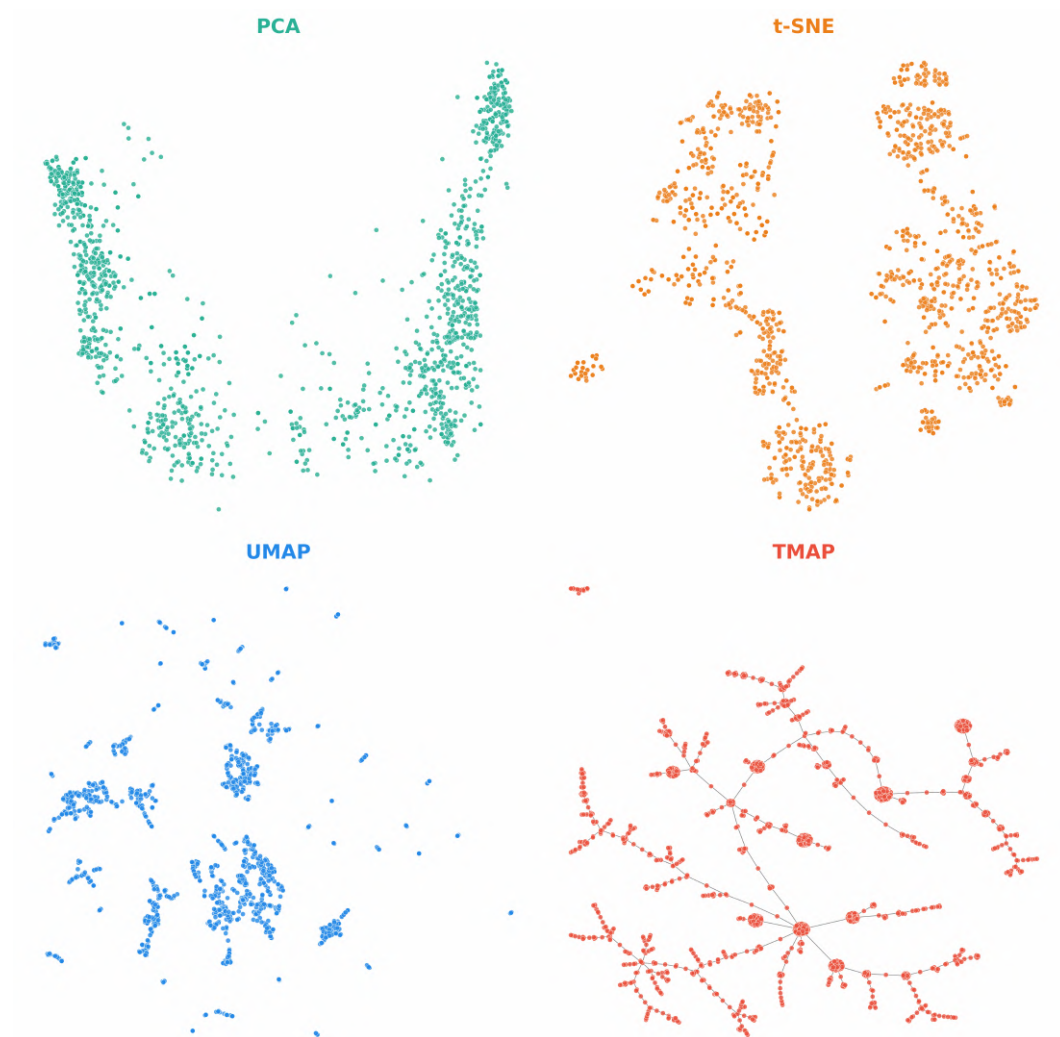


Figure 1.13: PCA, t-SNE, UMAP and TMAP representations of the same dataset, based on ECFP4 fingerprint descriptors.

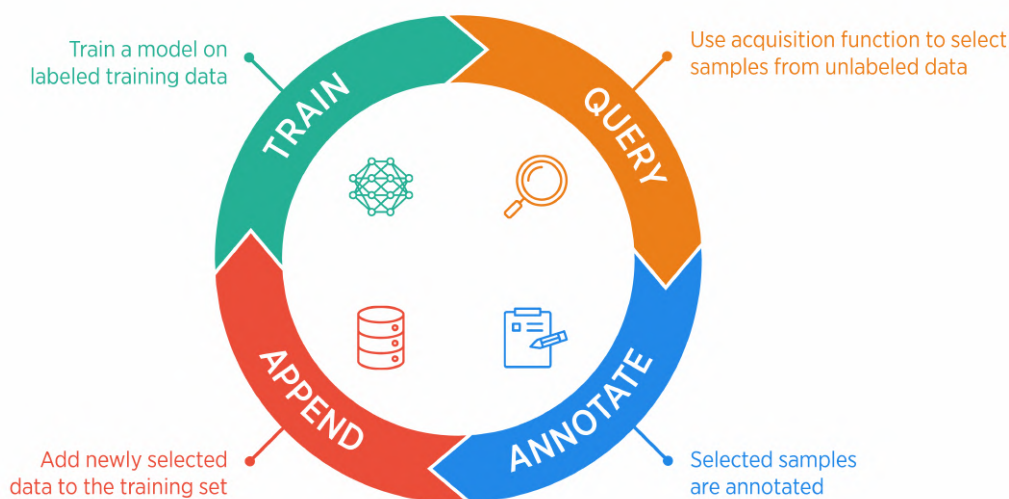


Figure 1.14: Overview of the active learning cycle.

compounds for testing, specifically those most likely to improve model performance or identify active molecules [151, 152, 153, 41, 42]. Subsequently, these experimentally validated compounds are integrated into the training set for future iterations.

A typical active learning workflow involves several iterative steps, as shown in Figure 1.14: (1) Training a model using currently available labeled data, which initially may consist of a small number of experimentally tested compounds; (2) Employing the model to prioritize candidate compounds from a pool of unlabeled molecules according to an acquisition function, which often targets maximizing uncertainty or structural diversity; (3) Experimentally testing these prioritized compounds to obtain new labels; (4) Augmenting the training dataset with the newly acquired experimental results; and repeating this cycle. The primary objective of active learning is to develop a model whose predictive performance matches that of a model trained on a substantially larger random dataset, while significantly reducing the number of required experiments by strategically focusing on the most informative compounds.

1.3 Focus on lead optimization

The preceding section emphasized the increasing role of *in silico* approaches in contemporary drug discovery, highlighting their ability to accelerate early-stage research and facilitate informed decision-making throughout the pipeline. Currently, most computational methods are predominantly utilized in the initial stages, particularly around hit identification [154, 12, 41, 42], due to the availability of extensive public datasets. However, lead optimization remains underrepresented in public datasets, as detailed records of lead optimization processes are rarely available outside industrial settings.

Conducting this thesis research at Sanofi provided a unique opportunity to delve into this specific stage. Therefore, the following section will concentrate on the lead optimization phase, examining its pivotal role as a major bottleneck in drug development and exploring how strategic approaches, including computational methodologies, can help overcome associated challenges.

As previously mentioned, lead optimization constitutes a multi-objective optimization challenge, in which initial lead compounds must be refined to simultaneously enhance potency, selectivity, AD-MET properties, and minimize toxicity. This optimization is performed iteratively through Design-Make-Test-Analyze (DMTA) cycles [155, 156], a process that lies at the core of this thesis and is described in detail in section 1.3.2.

The lead optimization stage is particularly critical due to its initial reliance on limited chemical knowledge, typically involving only a small set of compounds from a restricted number of chemical series. Compound synthesis and property assessment at this stage are resource-intensive and expensive. Consequently, the project team must strategically navigate chemical space by adopting efficient compound prioritization strategies during DMTA cycles, balancing the exploitation of promising chemical leads with the exploration of novel chemotypes.

1.3.1 Optimization objectives

In the following sections we describe the main categories of endpoints to be optimized within the lead optimization phase.

1.3.1.1 Modulating target's activity

Modulating the biological activity of the target is a fundamental objective in drug discovery. Central to this process are the pharmacological concepts of efficacy and potency, which define how drugs interact with their biological targets to elicit therapeutic responses. Efficacy refers to the maximal biological effect achievable when a drug binds to its target, reflecting its intrinsic activity. Highly efficacious drugs can fully activate their targets, thereby inducing maximal therapeutic responses, whereas drugs with lower efficacy yield only partial activation. Potency, meanwhile, denotes the concentration or dose necessary for a drug to produce a specific biological response. A highly potent drug achieves its desired effect at lower concentrations compared to one with lower potency.

Potency is typically the primary endpoint and is quantified using metrics such as IC_{50} and EC_{50} . For antagonists or inhibitors, the IC_{50} (half-maximal inhibitory concentration) represents the drug concentration required to inhibit a particular biological process or target activity by 50%. Conversely, for agonists, the EC_{50} (half-maximal effective concentration) is the concentration at which the drug achieves 50% of its maximal effect. Figure 1.15 illustrates a dose-response for inhibition

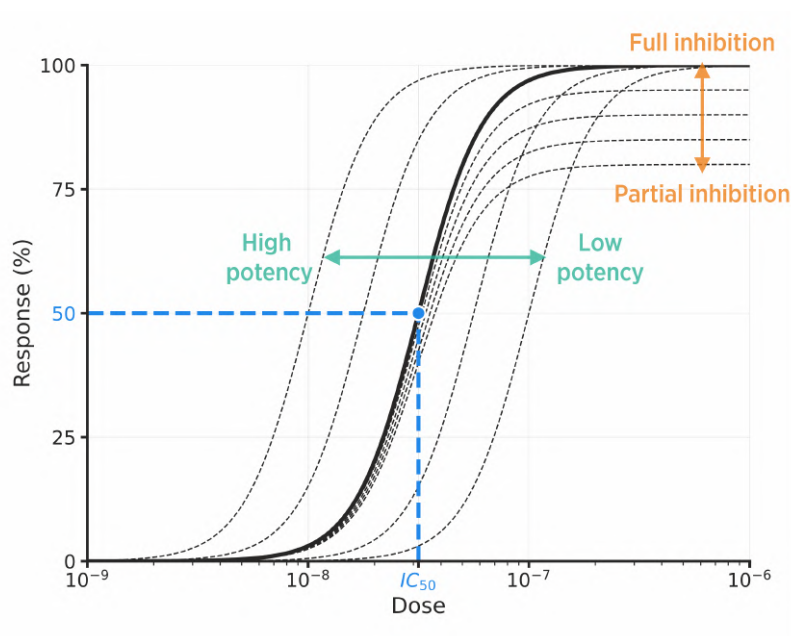


Figure 1.15: Example of a dose-response relationship curve for an inhibitor. The response (in %) is based on a relative maximum response. The IC_{50} is the concentration (dose) at which 50% of the response is obtained.

with the directions for low/high efficacy and potency. For ease of interpretation and comparison, these values are often converted into their negative logarithmic forms, pIC_{50} and pEC_{50} as described in Equation 1.5. Figure 1.15 illustrates a dose-response curve with the directions for high/low efficacy and high/low potency. Figure 1.15 illustrates a dose-response curve with the directions for high/low efficacy and high/low potency.

$$pIC_{50} = -\log_{10}(IC_{50}) \quad (1.5)$$

1.3.1.2 Selectivity

Beyond potency, a successful drug must act selectively on its intended target without significantly affecting other biological targets. Optimizing selectivity is crucial because off-target interactions are a common source of adverse effects and late-stage failures [27, 3]. A drug-like molecule in a biological system encounters a myriad of potential binding partners – other enzymes, receptors, ion channels, etc. – and unintended binding can lead to toxicity or side effects. Therefore, narrowing a compound's activity spectrum to preferentially modulate the disease-relevant target (while avoiding other proteins) improves its safety profile and therapeutic index. Indeed, appropriate tuning of binding selectivity is considered a primary objective during lead optimization. In practice, this means the team must optimize the difference between the compound's potency on the desired target

and its potency on off-targets.

Achieving selectivity, however, is often challenging. Many drug targets belong to protein families that have closely related isoforms or homologous binding sites (for example, the ATP-binding pockets of kinases). As a result, a molecule designed to fit one active site may also fit another with only subtle differences. In the case of protein kinases, for instance, the structural similarity across the kinase family makes it notoriously difficult to develop inhibitors that are highly selective for a single kinase [157, 158]. It is common for kinase inhibitors and many other classes of drugs to exhibit some degree of polypharmacology, interacting with multiple targets at therapeutic concentrations [158, 159, 160]. While such promiscuity can sometimes be exploited for polypharmacological therapies, it more often manifests as off-target toxicity or side effects [161]. Another challenge is that early in a project, not all off-target liabilities are known; a compound might appear selective in a limited panel, only to later show activity on an unanticipated target when broader profiling is done. Furthermore, certain molecular properties can predispose a compound to low selectivity: for example, very lipophilic or large molecules can engage many proteins non-specifically (sticking to hydrophobic pockets in various macromolecules) [162, 163], leading to "promiscuous" behavior. Thus, the lead optimization team must proactively address selectivity by characterizing off-target activities and refining the molecule accordingly.

1.3.1.3 ADME-Tox properties

Another critical endpoint in lead optimization is the set of Absorption, Distribution, Metabolism, Excretion, and Toxicology (ADME-Tox or ADMET) properties. No matter how potent and selective a compound is at its target, it will not become a successful drug unless it can be delivered to the site of action at sufficient concentrations for an adequate duration without causing harm to the organism. Thus, optimizing ADMET properties is fundamental to transforming a potent lead into a viable drug candidate. The scientific rationale is clear: a drug must be absorbable (if oral administration is desired, it should cross the gut barrier), adequately distributed (reach the tissue or organ where the target resides), metabolically stable (not cleared or inactivated too rapidly), and excretable at a rate that maintains the needed exposure but avoids accumulation. In parallel, the molecule should not trigger toxicity, whether through off-target pharmacology or via reactive metabolites and other deleterious pathways. A large fraction of drug candidates in the past failed in preclinical or clinical development due to suboptimal pharmacokinetics or unexpected toxicities. In fact, before the widespread implementation of ADMET optimization strategies, poor pharmacokinetics was responsible for an estimated 40% of drug failures in clinical trials [4]; intensive early ADMET screening has since reduced that attrition to around 10% [4]. This shift underlines why modern lead optimization devotes substantial effort to ADMET: by identifying and improving liabilities early (during lead

Endpoint	Target value	Units	Description
Permeability (Caco-2)	$> 10^{-6}$	$cm.s^{-1}$	Ability to cross intestinal epithelial cells modeled by Caco-2 cell line.
Solubility (aqueous)	> 50	μM	Ability of a compound to dissolve in water, crucial for bioavailability.
LogP (Lipophilicity)	$1 - 3 (\leq 5)$	<i>unitless</i>	Partition coefficient between n-octanol and water, quantifying how much the molecule prefers the non-polar (octanol) phase over the aqueous one.
LogD (Lipophilicity)	$1 - 4 (\leq 4)$	<i>unitless</i>	Distribution coefficient between n-octanol and water at a specified pH, so it accounts for both neutral and ionized forms present at that pH.
Oral bioavailability	$> 30\%$	%	Fraction of administered dose reaching systemic circulation.
Plasma Protein Binding (PPB)	< 90	%bound	Extent to which a compound binds to blood proteins, impacting free drug concentration.
Volume of distribution (Vd)	$0.1 - 3$	$L.kg^{-1}$	Hypothetical volume the drug would occupy if evenly distributed.
CYP450 Inhibition (IC_{50})	> 10	μM	Concentration at which 50% of cytochrome P450 enzyme activity is inhibited, critical to avoid drug-drug interactions.
Metabolic Stability (liver microsomes)	> 30	<i>min</i>	Time a drug remains stable in liver microsomes before metabolism.
Clearance (CL)	$10 - 30$	$mL.min^{-1}.kg^{-1}$	Rate at which drug is eliminated from body or plasma.
Elimination half-life ($t_{1/2}$)	$12 - 48$	<i>hours</i>	Time it takes for the concentration (or total amount) of a drug in the body to fall by 50% once the absorption/distribution phases are over.
hERG inhibition (IC_{50})	> 30	μM	Indicates cardiac risk; inhibition of hERG potassium channel linked to QT prolongation.
Genotoxicity (Ames Test)	<i>negative</i>	<i>qualitative</i>	Indicates potential to cause genetic mutations, tested in bacterial assays.

Table 1.2: Some key ADMET endpoints with typical target values and associated biological relevance for small-molecule drug candidates. The target value is indicative, and may change depending on the project.

optimization), the risk of late-stage failure is significantly lowered.

Optimizing ADMET is challenging because it involves a complex interplay of many molecular properties, and improving one aspect can adversely affect another. A classic example is the compound’s lipophilicity: increasing lipophilicity often boosts membrane permeability and target potency (since many binding pockets are hydrophobic), but overly lipophilic molecules tend to have poor solubility [53]. Thus, while some lipophilicity is needed for activity, there is an optimal range beyond which drug-like properties deteriorate. Similar trade-offs occur with molecular size and polarity: adding polar groups may improve solubility and reduce off-target binding, but if done excessively it can reduce permeability and oral absorption [164]. Identifying such issues early requires thorough profiling and is not always straightforward. Balancing all these factors simultaneously is a high-dimensional optimization problem. Often, a change that benefits one parameter harms another, so the chemist must iteratively test analogs to find a sweet spot where the compound’s overall profile is acceptable [52]. Table 1.2 summarizes some common ADMET endpoints, their typical target values, and respective biological significance for successful progression in drug discovery.

1.3.1.4 Blueprint

All of the above optimization efforts are ultimately directed toward one goal: producing a drug candidate that meets a predefined target profile for clinical success. At the outset of a drug discovery project, teams typically establish a blueprint that is also commonly called a Drug Candidate Target Profile (DCTP) or Target Product Profile (TPP). It is a document which enumerates the desired characteristics of the eventual drug. This profile includes essential attributes such as the intended patient population and therapeutic indication, the required properties and their range (i.e. potency, selectivity, ADMET), the route of administration and dosing frequency, and any other properties critical for the drug's use (stability, formulation, cost of goods, etc.). The scientific rationale for defining such a profile is to ensure the lead optimization is guided by the end-goal – a compound that not only works *in vitro* or in animal models, but also has the practical characteristics to be a medicine that is better than existing therapies. For example, if the blueprint calls for an oral drug taken once daily, the lead optimization must deliver a molecule with sufficient oral bioavailability and a plasma half-life long enough to support once-daily dosing. If the target patient population is likely to be elderly or pediatric, that might influence the acceptable safety and dosing parameters. By articulating these objectives early, the team works with a clear vision of what "success" looks like for the candidate.

The challenge in the lead optimization phase is that all the various endpoints, such as potency, selectivity, and ADMET, must be balanced together to fit the target profile. Often, improving one parameter can make it harder to meet another, and there may be multiple possible lead series or chemotypes in play, each with its own strengths and weaknesses relative to the profile. The lead optimization process thus becomes one of multi-parameter trade-offs and decision-making: the team must decide how to weight different criteria and when a compound is "good enough" on all fronts to merit advancement. It is rarely possible to maximize everything; instead, the goal is to fulfill the minimum criteria of the profile with an optimal overall balance. For instance, a slightly lower potency might be acceptable if it dramatically improves bioavailability and safety, yielding a better overall therapeutic index, as long as the potency still meets the minimal requirement for efficacy in humans. The blueprint also serves as a benchmark for go/no-Go decisions: if a lead series cannot achieve a certain critical parameter (for example, it consistently fails to improve metabolic stability to the required level), the team may decide to halt work on that series in favor of alternative approaches, rather than continuing to iterate ineffectively.

1.3.2 DMTA cycle

Small-molecule drug discovery is to be performed as an iterative Design-Make-Test-Analyze (DMTA) cycle [155, 156]. In each iteration, chemists design new molecular structures, make (synthesize)

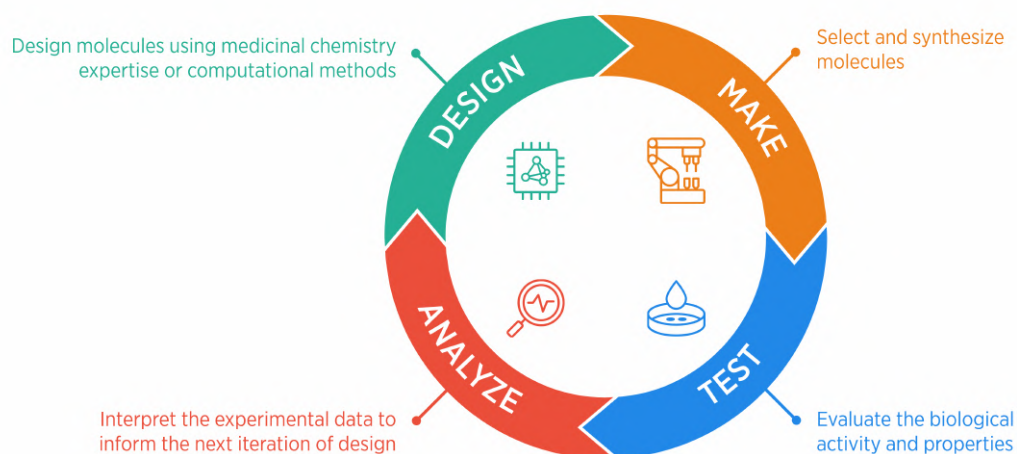


Figure 1.16: Illustration of the Design-Make-Test-Analyze cycle where compounds are iteratively optimized.

those compounds, test them in biological assays, and analyze the results to inform the next round of design. Multiple DMTA cycles are needed to optimize a "hit" molecule into a viable drug candidate, refining properties like potency, selectivity, and pharmacokinetics with each loop. Below, we discuss the general principles and current practices at each stage of the DMTA cycle, followed by an examination of key challenges in modern small-molecule drug discovery.

1.3.2.1 Design

Design is the creative front-end of the DMTA cycle, where new compound ideas are conceived. Medicinal chemists leverage both experimental data and computational tools to propose molecules with a high chance of desired activity. Thanks to the chemist's imagination and recent generative AI models [94, 95, 96, 97, 98, 99, 100], it is now easier than ever to design thousands of compounds at each iteration of the DMTA cycle. However, the synthesis capabilities are limited, typically around 20 - 50 compounds per month on a project. Hence the need to prioritize compounds efficiently. To do so, medicinal chemists mainly rely on their knowledge and intuition, also considering the ease of synthesis. They (sometimes) listen to proposition from computational chemists that uses molecular simulations or predictive QSAR and QSPR models to predict the properties of compounds in order to propose the most promising ones.

1.3.2.2 Make

The "Make" stage refers to the synthesis of designed compounds, i.e. turning ideas on paper (or *in silico*) into real molecules that can be tested. Modern medicinal chemistry employs a vast repertoire of synthetic organic reactions to assemble complex drug-like molecules [165]. A key principle is

to design efficient synthetic routes that can produce target compounds in a few steps, using readily available starting materials or building blocks. Synthetic feasibility often influences the design strategy itself, since there is little value in a brilliant drug design if it cannot be made. Thus, medicinal chemists consider synthetic accessibility from the outset and may prefer designs that leverage known chemistry or commercial scaffolds. Despite these advances, synthesizability remains a significant challenge in the design-make cycle. It is often observed that some molecules proposed by computational methods (or even by imaginative chemists) are extremely difficult to make in practice. There may be no known reaction to form a particular bond, or the required sequence might be too lengthy or low-yielding. In AI-driven *de novo* design especially, many generated structures fall outside the realm of accessible chemistry [166]. To address this, chemists employ retrosynthetic analysis tools [167, 168, 169] and heuristic metrics [170, 171, 172] to evaluate synthetic feasibility during design. For example, computational retrosynthesis algorithms can attempt to find a synthetic route to a proposed molecule; if even the AI struggles to find a route, the molecule might be set aside as "too hard to make." Likewise, synthetic accessibility scores (based on fragment rarity or known building blocks) can flag designs that are likely problematic. Increasingly, generative design algorithms themselves are being modified to "generate what you can make" by incorporating synthetic rules or even running a retrosynthesis step as a filter [173]. Overall, the Make stage is a balance between synthetic creativity and practicality.

1.3.2.3 Test

Once new compounds are synthesized, the Test stage evaluates their biological and physicochemical properties based on a screening tree. A screening tree is a structured, hierarchical framework to efficiently evaluate and prioritize chemical compounds based on progressively stringent selection criteria. Early stages typically involve straightforward and cost-effective computational or *in vitro* assays, whereas later stages encompass more demanding and resource-intensive evaluations, including detailed ADMET profiling and *in vivo* studies. At each step, compounds failing to meet predetermined thresholds are eliminated, thus concentrating resources exclusively on the most promising candidates. An illustrative example of a screening tree is shown in Figure 1.17; however, it should be emphasized that, despite general principles, screening trees are highly tailored to individual project needs.

1.3.2.4 Analyze

In the Analyze stage, the data collected from testing are examined in detail to derive insights and drive decision-making [156]. A primary focus is building the SAR: understanding how different chemical modifications impact biological activity and other properties [47]. Medicinal chemists will

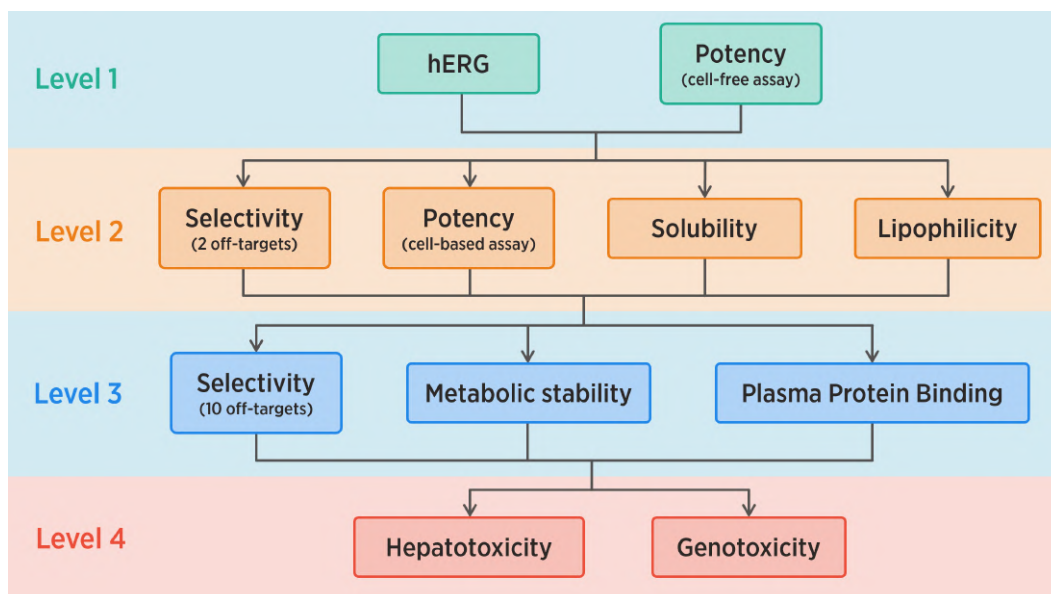


Figure 1.17: Example of a lead optimization screening tree with 4 stages.

look at all the assay results for a series of compounds and try to identify patterns – for example, "adding a fluorine at this position increased potency threefold" or "bulky substitutions on the ring reduce activity." By interpreting these SAR trends, the team can hypothesize the structural features responsible for desired activity (the pharmacophore) as well as features influencing off-target effects or ADMET profile.

The Analyze phase closes the loop of the DMTA cycle by providing feedback to the design stage. The insights gained, such as which functional groups are tolerated and what the preliminary structure–property trade-offs are, directly inform the next round of molecule design. Often, computational models are updated during analysis. If sufficient data has been generated, a QSAR/QSPR model can be retrained with the new data, or molecular docking poses might be revisited in light of SAR [174] ("the model didn't predict that compound X binds well, yet it's potent; perhaps the binding mode is different from what was originally assumed").

Human intuition and expert knowledge remain key at this stage [175] with medicinal chemists drawing on their experience to propose why a certain chemical change had the observed effect, integrating the data with chemical theory and prior literature. The outcome of analysis is essentially the strategy for the next DMTA iteration: which compound(s) to prioritize, what modifications to try next, and sometimes the decision that a series should be abandoned if the data reveals insurmountable issues [156]. In summary, the Analyze stage ensures that the huge amount of data generated in modern drug discovery is translated into knowledge and actionable decisions, completing the DMTA cycle and propelling the project forward.

1.3.3 Challenges

Despite tremendous technological advances, small-molecule drug discovery still faces several broad challenges. Modern automation, computational modeling, and high-throughput methods have improved parts of the pipeline, yet fundamental difficulties remain [12]. Researchers must contend with the astronomical size of chemical space (estimated to exceed 10^{60} drug-like molecules [36]), which makes it impossible to exhaustively search and demands efficient strategies for choosing what to make or test. They also face an ever-growing flood of experimental and computational data; the sheer volume of data can be overwhelming [74]. On top of these scientific complexities, practical constraints of time and resources continue to be severe. This prolonged timeline (sometimes described by Eroom’s Law [176], the observation that drug discovery has become slower and less efficient over time) means that any inefficiency in the process compounds the overall delay and cost. In the following subsections, we discuss some key challenges.

1.3.3.1 The exploitation/exploration trade-off

One core challenge in small-molecule drug discovery is balancing exploitation and exploration during iterative optimization [177, 178]. The vastness of chemical space means there are countless directions one could explore, but only a tiny fraction of compounds can ever be made and tested. Exploitation refers to focusing on areas of chemical space known to be promising – for example, refining a lead series or analogues that already showed good activity. Exploration, in contrast, means venturing into novel chemotypes or untested regions of chemical space to discover new opportunities. Both approaches are necessary, yet they conflict: investing too heavily in exploitation can lead to local optima or blind spots, whereas overly broad exploration may waste resources on many low-probability ideas.

In practice, medicinal chemists must decide at each DMTA cycle how to allocate their limited "design slots" between exploiting known SAR trends versus trying out new structural motifs. This trade-off is complicated by uncertainty. A compound that scores highest on current models might seem the best choice, but if all top-scoring candidates share the same scaffold, they may also share the same hidden liability or failure risk. Indeed, selecting only the highest-scoring, most similar molecules can be risky. If an unmodeled issue affects that chemotype, all those compounds could fail together. On the other hand, introducing diversity (exploration) by testing compounds with novel structures can mitigate correlated risks and increase the chance that at least one molecule succeeds. This intuition has been formalized by viewing a batch of candidate molecules like an investment portfolio [179]: concentrating on a single promising "asset" maximizes expected return but also maximizes risk, whereas diversifying into chemically different candidates reduces the chance of a total failure. The optimal strategy often lies in a middle ground, striking a balance between ex-

exploiting known good features and exploring enough diversity to hedge against errors in predictions.

1.3.3.2 Time and resource constraints

Another major challenge in small-molecule drug discovery is the limitation of time and resources. Even with efficient workflows, each DMTA cycle requires significant time. Designing compounds, synthesizing them (the "Make" step, which is often a bottleneck), testing in biological assays, and analyzing results can collectively take weeks to months [5]. Over the many iterations needed to optimize a drug candidate, these cycle times accumulate. It is not uncommon for lead optimization to involve dozens of cycles and many hundreds (or thousands) of compounds, which in turn can span several years of work.

Closely tied to time are resource constraints. A discovery team usually operates under limits on the number of chemists, amount of available reagents, assay throughput, and funding. Practically, this means only a certain number of compounds can be synthesized and tested in each cycle. Even with emerging automation that can parallelize compound production, resources like material and researcher time remain precious. Thus, choosing which compounds to prioritize for synthesis is critical. A suboptimal choice wastes not only that particular synthesis and assay effort, but also the calendar time until the next cycle when a better candidate could have been tried. Furthermore, late discovery stages often face hard deadlines (e.g., go/no-go decision points, patent cliffs, or competitive pressures), so inefficient use of cycles can mean missed opportunities or project termination due to time run-out.

1.3.3.3 Data overload for humans

Medicinal chemists' expertise lies at the core of the Design-Make-Test-Analyze (DMTA) process. However, as projects advance and data accumulates, it becomes increasingly challenging for individuals to retain and accurately recall the breadth of information generated. Although extensive data ideally leads to improved decision-making, the sheer volume and complexity of this information can paradoxically impede effective judgments. Teams frequently experience information overload, causing critical insights, such as subtle structure-toxicity correlations or notable physicochemical outliers, to become obscured or overlooked amid numerous measured parameters. Furthermore, within large pharmaceutical organizations, decades of data are typically stored in fragmented repositories, complicating rapid retrieval and interpretation of relevant information. Without efficient tools for information management, project scientists often depend on incomplete datasets or personal recollections, resulting in suboptimal decision-making.

An additional complexity arises from the inherently multi-objective nature of drug design de-

cisions. Selecting the most promising compound never hinges on a single criterion; instead, it necessitates balancing multiple properties as previously described. Human cognition struggles to intuitively handle such multi-dimensional trade-offs. Historically, heuristic guidelines like Lipinski’s Rule of Five provided basic filters for managing this complexity by setting general boundaries for acceptable drug properties. While these guidelines have been helpful, they represent coarse approximations and often fail to capture project-specific nuances, potentially excluding promising candidates prematurely. The contemporary, data-rich context demands more sophisticated, data-driven decision-support tools capable of addressing such nuanced complexities.

1.3.3.4 Data shortage for machines

Although the volume of data generated during drug discovery projects may overwhelm human decision-makers, it represents a low-data regime from a machine learning perspective. Typical hit-to-lead and lead optimization campaigns seldom produce more than 10^2 - 10^3 molecules per chemical series, and individual in-vitro assays generally yield even fewer experimentally validated data points per measured property. In contrast, fields such as computer vision or natural language processing routinely employ datasets ranging from tens of thousands to billions of samples to train state-of-the-art models. For instance, ImageNet, a foundational computer vision dataset introduced in 2012, comprises over one million labeled images [180]. Another notable example is GPT-2, the last open-source language model released by OpenAI, trained on approximately 8 million web pages in 2019 [181]. Even within computational drug discovery context, successful models like AlphaFold[84] (prediction of protein structures) were trained on extensive databases containing hundreds of thousands of experimentally determined protein structures, highlighting that substantial data availability often underpins significant advances in the field [84].

1.3.3.5 Benchmarking lead optimization strategies

While significant progress has been made in computational methods within drug discovery, most *in silico* approaches predominantly target hit identification [182]. This focus arises due to readily available public screening datasets and the comparatively straightforward objective of optimizing potency alone. Conversely, lead optimization constitutes a fundamentally different challenge, characterized by multi-objective optimization alongside dynamically evolving chemical spaces.

A major barrier to benchmarking computational approaches for lead optimization is the lack of publicly available multi-objective datasets that accurately reflect industrial conditions [183, 184]. Collecting comprehensive datasets representative of the multi-objective and dynamic nature of real drug discovery projects remains particularly challenging outside industrial settings, significantly limiting academic research.

Chapter 2

Benchmarking molecular descriptors for low-data regime modeling

QSAR and QSAR/QSPR models are commonly employed in cheminformatics to predict biological activities and physicochemical properties from molecular structures [185, 186]. The fundamental principle of QSAR/QSPR modeling is that changes in molecular structures directly influence biological activity and physicochemical characteristics [187]. These models are particularly valuable in drug discovery and materials science, as they accelerate research and reduce experimental costs [188]. Despite their widespread adoption, QSAR/QSPR models still encounter substantial challenges, particularly regarding accurate performance assessment [189, 183] and robust predictive accuracy when applied to real-world pharmaceutical datasets [190, 183].

As previously discussed, molecular structures can be encoded in various ways to enable processing by computers, and numerous algorithms leverage these representations for predictive tasks. Most modern methods rely on machine learning or deep learning algorithms. Although conventional data science wisdom suggests that larger datasets yield better models, the lead optimization phase of drug discovery typically involves limited data. This situation underscores the importance of molecular descriptors and predictive models suited for low-data contexts.

In this chapter, we benchmark several widely adopted 2D molecular descriptors in combination with classical machine learning algorithms. Their predictive performances are compared against a state-of-the-art deep learning model. Evaluations are performed using datasets from real-world drug discovery projects, explicitly considering temporal distribution shifts. This approach provides practical insights into model performance under realistic conditions.

2.1 Challenges in QSAR/QSPR modeling

This section examines key challenges facing QSAR/QSPR modeling in modern drug discovery.

2.1.1 Data quality and quantity

The foundation of any QSAR/QSPR model is the underlying data, and data quality represents perhaps the most fundamental challenge to overcome. Experimental errors in datasets can severely compromise model performance and reliability.

2.1.1.1 Experimental errors and inconsistencies

Experimental errors frequently appear in QSAR/QSPR datasets, potentially misleading model development. Errors and uncertainties in activity measurements can create misleading activity cliffs, obscuring true structure-activity relationships [191]. Public datasets often present notable problems, including incorrect chemical structure representations and conflicting data points for identical compounds [192].

2.1.1.2 Data scarcity

Limited dataset size poses another significant challenge. Medicinal chemistry typically generates small compound series, hindering the development of statistically robust predictive models [193]. Small datasets are especially vulnerable to experimental errors, complicating modeling efforts further [192]. The issue intensifies when modeling complex biological endpoints involving multiple mechanisms.

2.1.2 Real-world proof models

2.1.2.1 The public and industrial data gap

A critical but often overlooked aspect of QSAR/QSPR modeling is the substantial difference between public and industrial datasets. These differences significantly impact model performance when transitioning from academic research to industrial applications. Table 2.1 summarizes some of the main differences between public and private datasets.

Public datasets, such as those provided by ChEMBL or PubChem [194, 195, 196], offer several advantages for QSAR/QSPR modeling. These datasets typically encompass chemically diverse compounds across a wide range of biological activities, supporting the development of predictive

Dataset type	Typical size	Chemical diversity	Curation level	Availability
Industrial	Small/Medium	Low-Medium	High	Low
Public	Medium/Large	High	Variable	High

Table 2.1: Key differences between public and industrial datasets in term of size, chemical diversity, curation level and availability.

models with broad applicability domains. Moreover, their open access promotes reproducibility, benchmarking, and cross-method comparisons in model development[196, 197]. Nonetheless, public datasets often present notable limitations. Common issues include experimental noise, inconsistent assay conditions, and incomplete or missing metadata [198]. Such deficiencies require extensive curation to resolve chemical and contextual inconsistencies prior to model training [189].

Industrial datasets, in contrast, typically originate from targeted research programs within pharmaceutical companies. These proprietary datasets usually comprise compounds belonging to specific chemical series designed explicitly for particular biological targets, leading to more structurally homogeneous distributions compared to public datasets. Companies also generally adhere to standardized internal protocols, leading to more consistent and reproducible assay conditions compared to those publicly available.

Empirical studies highlight substantial performance gaps when applying models trained on public data to industrial datasets [199]. This indicates that strong performance on public benchmarks may not reliably translate to industrial settings, a pattern also observed in internal evaluations with proprietary datasets.

2.1.2.2 Temporal distribution shift

In lead optimization campaigns, the chemical space explored progressively diverges from that of the initial leads, causing the distribution of features encountered by models during production to differ significantly from the training set. This distribution shift can substantially decrease the predictive performance and degrade the uncertainty calibration of QSAR/QSPR models, thus seriously impairing decision-making. QSAR/QSPR models are fundamentally interpolative; they yield reliable predictions only within the applicability domain established by their training compounds. When applied to novel chemotypes outside this domain, their predictions may become biased and inaccurate [190]. Retrospective temporal and scaffold-split studies, which better mimic prospective deployment, consistently show predictive performance losses compared with random splits, confirming the magnitude of this problem [200, 201]. During lead optimization, where each prediction guides costly synthesis and testing cycles, unmitigated distribution shift therefore translates directly into wasted laboratory effort and slower project timelines.

2.1.3 Navigating the models jungle

QSAR/QSPR models combine molecular descriptors with predictive algorithms to estimate how structural features influence biological activities and physicochemical properties. Since model performance depends on both components, careful selection of descriptors and algorithms is crucial.

2.1.3.1 Choosing the right predictive algorithm

Machine learning methods dominate contemporary QSAR/QSPR studies. Early investigations used k-nearest neighbors (k-NN) [202, 203] and support-vector machines (SVM) [204]. Tree-based ensembles, particularly random forests [205, 206, 202] and gradient boosting [207], later became preferred choices. Although neural networks were explored in the 1990s, they gained widespread acceptance only recently [206, 208, 202]. State-of-the-art deep learning architectures now range from multilayer perceptrons to models such as graph neural networks [209, 210, 211, 212, 213, 214]. Conventional machine learning models typically require fewer parameters than neural networks. Thus, carefully considering the choice between traditional algorithms and more data-intensive deep learning models is essential prior to developing QSAR/QSPR models.

2.1.3.2 Choosing the right featurization

Descriptors span simple physicochemical properties to richer encodings (e.g. extended-connectivity fingerprints or graph-based embeddings) that capture complementary facets of molecular structure relevant to biological function [186, 82, 215]. Crucially, each representation pairs best with specific algorithms: vector formats such as fingerprints suit methods like SVMs or Random Forest, whereas graphs are processed most effectively by GNNs.

2.2 Material and methods

This study benchmarks models combining widely used 2D molecular descriptors with various machine learning and deep learning algorithms, comparing their performance across 121 proprietary industrial datasets derived from actual lead optimization campaigns. These datasets span a wide range of sizes, from a few dozen compounds up to more than 8000. To better replicate prospective deployment scenarios typical in drug discovery, we employ a temporally stratified (time-based) validation scheme, which separates training and test sets based on the assay dates of the compounds. This way of separating data is both the gold-standard [216, 217] but is also related to the rest of the thesis, presented in chapters 3 and 4, where temporality plays an essential role.

2.2.1 Molecular descriptors

The aim of this work is to explore which featurization technique is the best in low-data QSAR/QSPR modeling, comparing vectorial and graph descriptors.

2.2.1.1 Molecular fingerprints

Two-dimensional (2D) molecular fingerprints are computationally inexpensive, capture key molecular features, and consistently underpin highly predictive models [218, 219]. We evaluated eight fingerprint featurization, several of which yield multiple variants owing to alternative parameterizations.

Extended-connectivity fingerprints Extended-Connectivity FingerPrints (ECFPs) [82] are circular, atom-centered molecular fingerprints generated by iteratively hashing the environment around each atom out to a specified radius (e.g., ECFP4 uses a radius of two bonds and ECFP6 a radius of three), with each unique local substructure mapped to a bit position (or index) in a vector; the resulting vectors can be of various fixed lengths (commonly 1 024 or 2 048 bits), making the fingerprint size tunable to balance memory footprint and collision rate. During generation, you can choose to encode stereochemical information, known as "chiral" ECFPs, by incorporating atom chirality and bond stereo descriptors into the hash keys, or omit it for achiral versions when stereochemistry is unimportant. Finally, the output can be stored as a sparse binary presence/absence vector, or as a count-based/hashed integer vector that records how many times each environment occurs.

SMILES extended-connectivity fingerprints SMILES-ECFP (SECFP) [220] is a circular fingerprint that keeps the size-tunable, fixed-length bit-vector format of classical ECFP but generates each feature differently: instead of iteratively hashing atom-based invariants, it "shingles" the molecule by writing out the canonical SMILES of every circular substructure around each atom, then hashes those fragment strings and folds the resulting integers into the bit array by a simple modulo-n operation.

Avalon fingerprints Avalon fingerprints, implemented in the open-source Avalon toolkit and exposed in RDKit [221], are path-and-substructure-based molecular fingerprints that enumerate a rich set of fragments, including linear atom/bond paths up to a configurable length, ring fragments, atom pairs/triplets, heteroatom environments, and valence/charge patterns. Each fragment description is then hashed into positions of a fixed-length bit vector with tunable size.

Topological torsion fingerprints Topological torsion fingerprints [222] (also called atom-sequence or Torsion fingerprints) describe a molecule by enumerating every unique sequence of four contiguous, non-branching atoms along its topological graph, capturing both atom types and bond orders. This encodes the "torsion" (bond–bond dihedral) environments that underlie conformational preferences. Each four-atom string is converted to an integer key, commonly by hashing, and mapped into a fixed-length vector, producing a size-tunable representation. This approach emphasizes local sequential chemistry rather than broader rings or paths.

AtomPair fingerprints Atom Pair fingerprints [223] encode a molecule by enumerating every unordered pair of atoms, labeling each atom with simple descriptors (e.g., element, ring status, formal charge) and tagging the pair with their shortest topological separation in bonds; the resulting "atom-type–distance–atom-type" strings are then hashed into indices of a fixed-length vector with tunable size.

Mordred Mordred [224] is more accurately described as a set of descriptors rather than a structural fingerprint. They are generated by the open-source Mordred Python package, which computes an extensive catalogue of predefined 2-D and 3-D molecular descriptors (over 1 800 by default) that span simple constitutional counts, topological and connectivity indices, charge and electronic terms, radial distribution functions, geometrical moments, and more. For each molecule, RDKit is used to perceive atom and bond attributes (and 3-D coordinates if supplied); each descriptor class then applies its specific formula to yield a named scalar value, so the final "fingerprint" is a dense, fixed-ordering numeric vector rather than a hashed bitset. Users can tailor its dimensionality by selecting or omitting descriptor classes (e.g., excluding 3-D descriptors when coordinates are unavailable, what is done in this work).

MACCS keys MACCS (Molecular ACCess System) keys are a curated, fixed-length set of structural yes/no questions. Originally, the system included 960 proprietary patterns, but the most commonly used version is the public 166-bit subset [81], where each bit corresponds to a predefined SMARTS-like substructure such as " ≥ 1 aromatic ring," "alkyl halide fragment," or "terminal alkyne". To generate the fingerprint, a cheminformatics toolkit scans the molecule against every pattern; if the motif is found the associated bit is switched from 0 to 1, producing an easily interpretable binary vector whose positions and meanings are standardised across datasets. Because the patterns are hand-selected to cover diverse functional groups and topologies, MACCS keys provide a compact and human-readable fingerprint. Note that full public definitions are not available, and packages differ. Here, we use the definition refined by Greg Landrum that can be found in RDKit [221].

RDKit2D Descriptors RDKit 2D Descriptors are the collection of 200 scalar properties that RDKit can calculate directly from a molecule’s 2D graph, covering simple constitutional counts (atoms, hetero-atoms, rotatable bonds), fragment-based physicochemical terms (cLogP, topological polar surface area), ring and heteroatom environments, charge and valence statistics. When a molecule is loaded, RDKit perceives atom types, bond orders, aromaticity, and ring membership, then each descriptor class applies its specific formula (often summing pre-tabulated atom contributions or performing path-based graph traversals) to yield a numeric value; the results are returned in a fixed, ordered vector whose elements retain their explicit chemical meaning. Because these descriptors are deterministic, interpretable, and unit-consistent, they serve as a convenient and human-readable fingerprint.

2.2.1.2 Molecular graph : Chemprop

As seen in the first chapter, a molecule can be featurized as a graph. Graph featurization is the preprocessing step that converts a molecule into the data structures a graph neural network can ingest: each atom becomes a node with a feature vector and chemical bonds becomes an edge with a feature vector. Here, we used the default graph featurization method from the Chemprop model [209]. Chemprop is an open-source toolkit that uses message-passing neural networks (MPNN) to predict chemical and materials properties directly from molecular graphs.

Atom featurization Chemprop’s default atom featurization generates a vector that encodes information about the atom’s atomic number, degree, formal charge, chirality, number of hydrogens, hybridization, aromaticity and mass.

Bond featurization Chemprop’s default bond featurization generates a vector that encodes information about the bond type (single, double, triple or aromatic), its conjugation, its presence (or not) in a ring system and its stereochemistry.

2.2.2 Predictive algorithms

The molecular descriptors described above were subsequently evaluated using multiple predictive algorithms. To ensure the robustness of the molecular fingerprints and mitigate bias associated with the strengths or limitations of any single method, we employed a diverse set of widely used models. Specifically, we benchmarked three machine learning algorithms frequently applied in QSAR/QSPR modeling and broader data science applications: Random Forest [205], XGBoost [225], and LightGBM [226]. Each model was trained using its default hyperparameters. In parallel, we assessed

molecular graph representations using Chemprop’s default implementation of the message-passing neural network (MPNN), trained for 30 epochs.

2.2.2.1 Random Forest

Random Forest [227] is an ensemble learning algorithm that builds a "forest" of decision trees, each trained on a different bootstrap sample (random resampling with replacement) of the training data; at every split in every tree, it further injects randomness by selecting only a random subset of features to consider, which decorrelates the trees. During inference, the outputs are aggregated by averaging in the case of regression or by majority vote in classification. This allows the diverse, individually high-variance trees to combine into a low-variance, accurate predictor.

2.2.2.2 XGBoost

XGBoost (eXtreme Gradient Boosting) [225] is an optimized, regularized implementation of gradient-boosted decision trees that builds an ensemble sequentially: each new tree is fit to the residual errors (negative gradients) of the current model, then added with a shrinkage (learning-rate) weight so the ensemble gradually corrects its mistakes. Key innovations compared to regular gradient boosting include second-order gradient information for precise split scoring, L1/L2 regularization on leaf weights to curb over-fitting, column and row subsampling to decorrelate trees and sophisticated handling of sparse "missing" values.

2.2.2.3 LightGBM

LightGBM (Light Gradient Boosting Machine) [226] is a fast, memory-efficient implementation of gradient-boosted decision trees that grows its trees leaf-wise rather than level-wise: at each boosting iteration it picks the leaf with the largest loss reduction and splits it, producing unbalanced, deeper trees that capture complex patterns with fewer nodes. To keep training quick and RAM-friendly, LightGBM bins continuous features into histograms on the fly, reuses histogram differences to evaluate splits cheaply, and supports exclusive feature bundling and gradient-based one-side sampling to down-sample easy examples without hurting accuracy. The framework then adds each new tree’s predictions, scaled by a learning rate, to the ensemble, iteratively minimizing a differentiable loss.

2.2.2.4 Chemprop

Chemprop is an open-source cheminformatics toolkit designed for predicting chemical properties directly from molecular graphs using graph neural networks, specifically MPNN, described in the

first chapter . We used Chemprop’s default MPNN implementation consisting of a message passing across bonds with a mean aggregation on nodes and a feed forward network for prediction.

2.2.3 Final set of benchmarked descriptors

Each molecular fingerprint was assessed both individually and in concatenated pairs, enabling complementary information to be exploited [228, 229, 230]. Pairs combining two fingerprints from the same family but only differ by parameterization are excluded. A summary of the various fingerprint families and variations is described in Figure . Hereafter, combination size denotes the number of fingerprints concatenated (1 or 2), whereas dimension refers to the length of the resulting feature vector. In total, 483 descriptors were tested :

- 35 unique fingerprints (combination size = 1) to be used with Random Forest or XGBoost or LightGBM.
- 447 paired fingerprints (combination size = 2) to be used with Random Forest or XGBoost or LightGBM.
- 1 graph embedding to be used with Chemprop model.

2.2.4 Datasets

2.2.4.1 Industrial datasets

A curated collection of 121 datasets originating from legacy lead optimization projects was assembled for benchmarking purposes. Of these, 38 quantify potency (i.e., they predict activity on a biological target), while the remaining 83 are ADMET endpoints. All datasets contain chemical structures associated to a numerical value. They are all designed for regression (continuous target) and not for classification. For instance, the potency sets report pIC_{50} values instead of providing a binary active/inactive label.

Dataset sizes span from 23 to 8 730 molecules, with a median of approximately 1 000 compounds. Most datasets contain fewer than 3 000 molecules (Fig.2.1). The few very large datasets all concern ADMET endpoints (Fig.2.1B,C); none of the potency datasets reach comparable sizes (Fig.2.1B,D). This disparity is expected: bioactivity against a specific therapeutic target is usually confined to a single project, whereas many ADMET assays are shared across projects, enabling the aggregation of substantially larger ADMET datasets on some very common endpoints.

Descriptor	Size (bits)	Chirality	Counts
ECFP4	1024		
ECFP4	1024	✓	
ECFP4	1024		✓
ECFP4	1024	✓	✓
ECFP6	1024		
ECFP6	1024	✓	
ECFP6	1024		✓
ECFP6	1024	✓	✓
ECFP4	2048		
ECFP4	2048	✓	
ECFP4	2048		✓
ECFP4	2048	✓	✓
ECFP6	2048		
ECFP6	2048	✓	
ECFP6	2048		✓
ECFP6	2048	✓	✓
SECFP4	1024		
SECFP6	1024		
SECFP4	2048		
SECFP6	2048		
AtomPair	1024		
AtomPair	1024	✓	
AtomPair	2048		
AtomPair	2048	✓	
Topological Torsion	1024		
Topological Torsion	2048		
Avalon	512		
Avalon	512		✓
Avalon	1024		
Avalon	1024		✓
Avalon	2048		
Avalon	2048		✓
RDKit2D Descriptors	200		
Mordred	1613		
MACCS	166		

Table 2.2: Summary of the 8 vector-based descriptors with each of their variants.

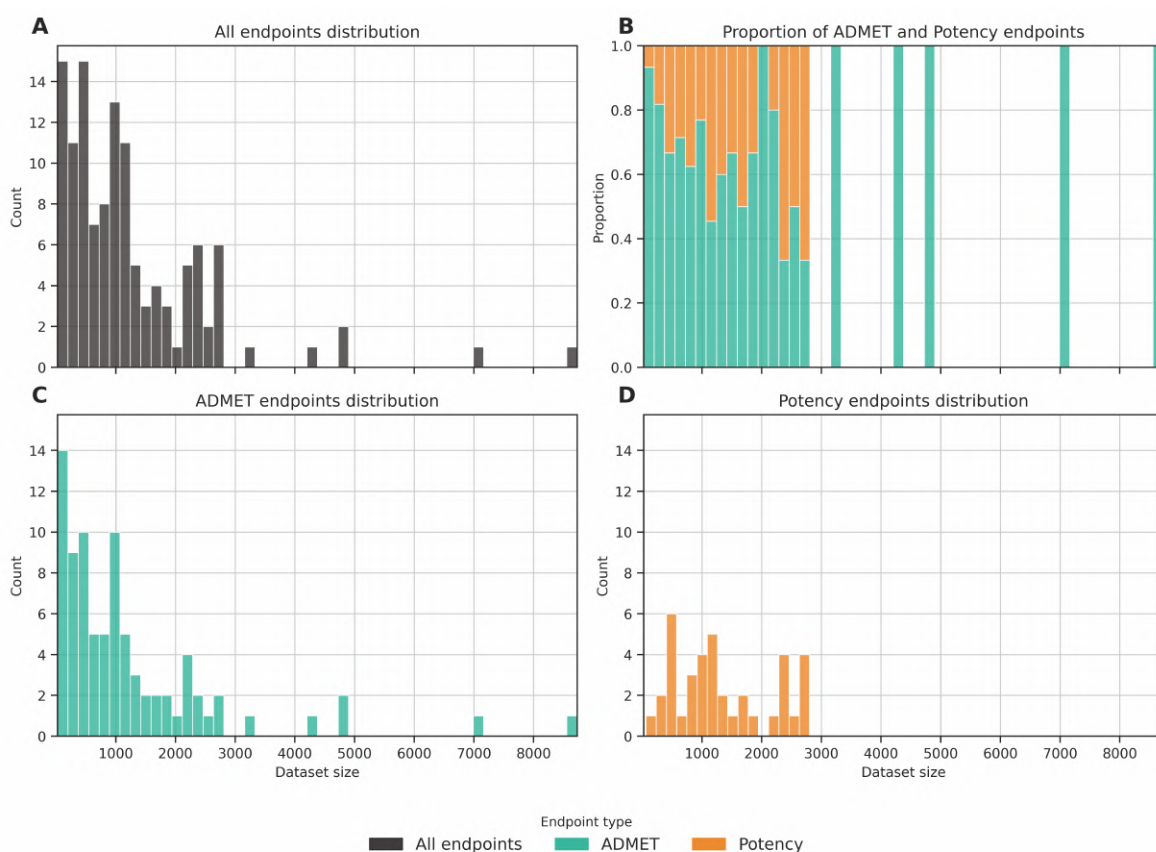


Figure 2.1: Size distribution of the 121 benchmarking datasets. Panel A shows the overall distribution of molecule counts per dataset. Panel B separates potency and ADMET sets, highlighting that all potency datasets lie below 3 000 molecules, whereas several ADMET datasets exceed this range. Panels C and D show the distinct distributions of respectively ADMET and potency datasets.

2.2.4.2 Data curation

Data curation, although being a laborious task, is a crucial step to develop efficient models as previously described. Here, we followed a curation pipeline inspired by Tropsha [189].

- Removal of solvents and salts from SMILES.
- Aggregation of data based on internal ID of molecules and not only on SMILES.
- Correction of chirality based on internal ID of molecules and not only on SMILES.
- Standardization of SMILES.
- Treatment of tautomeric forms (harmonization of SMILES).
- Removal of duplicates.
- Removal of abnormal values.
- Manual inspection.

2.2.5 Benchmarking methodology

To find out which descriptor performs best for model building, we ranked them based on their predictive performance on the 121 datasets. To assess this performance, multiple metrics exist in a regression context.

2.2.5.1 Metrics

Metrics are quantitative measures used to evaluate the performance of a model. For this work we used the three most common regression metrics that are the coefficient of determination (R^2), the Mean Squared Root Error (RMSE) and the Mean Absolute Error (MAE).

Coefficient of determination The coefficient of determination R^2 is a statistical metric used to evaluate how well a regression model explains the variability of the target variable. It quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables. An R^2 value of 1 indicates perfect prediction, while an of 0 means the model does no better than simply predicting the mean of the target. The R^2 can also be negative, meaning that the model performs worse than a horizontal line at the mean value, it adds error compared to simply predicting the average. So while R^2 is bounded above by 1, it can be negative and unbounded below if the model fits the data very poorly. In practice, a R^2 value is considered to be acceptable around

0.6-0.7. This metric is convenient because it enables to quickly compare the performance of different models, however one need to be cautious using it. Indeed, a small dataset can make R^2 unstable because of the high variance, leading to misleading high or low values.

The R^2 score is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2.1)$$

where y_i are the true values, $\hat{y}_i$ are the predicted values, $\bar{y}$ is the mean of the true values, and n is the number of data points.

Mean Squared Root Error The Root Mean Squared Error (RMSE) is a widely used metric to assess the performance of regression models. It measures the square root of the average squared differences between the predicted and actual values. RMSE penalizes larger errors more than MAE due to the squaring operation, making it sensitive to outliers. A lower RMSE indicates a better fit of the model to the data.

The RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2.2)$$

While the RMSE provides a measure of error, it can be difficult to compare across datasets with different target value scales. The Normalized RMSE (nRMSE) addresses this by scaling the RMSE with respect to reference quantities. This normalization yields a dimensionless value, making it possible to benchmark model performance across heterogeneous datasets. Many normalization techniques exist, in our case we used a MinMax normalization to rescale, for a given dataset, the metric between 0 and 1. A nRMSE of 0 indicates that it is the best performing model while a value of 1 indicates that it is the worst.

The nRMSE is defined as:

$$\text{nRMSE} = \frac{\text{RMSE} - y_{\min}}{y_{\max} - y_{\min}} \quad (2.3)$$

Mean Absolute Error The Mean Absolute Error (MAE) is a common metric used to evaluate the accuracy of regression models. It measures the average magnitude of the errors between predicted and actual values, without considering their direction. In other words, it gives an idea of how far predictions are from the true values on average. Unlike other metrics like RMSE, MAE is less sensitive to large errors, making it a robust measure in the presence of outliers.

The MAE is defined as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2.4)$$



Figure 2.2: Illustration of the three data-partitioning strategies to assess predictive performance of QSAR/QSPR models. Panel A illustrate the random split: molecules are assigned to training (blue) and test (orange) subsets at random. Panel B illustrate the cross-validation: the data are divided into k equally sized folds (example shown for $k = 5$); each fold is used once as a test set while the remaining $k - 1$ folds form the training set, reducing variance in the performance estimate. Panel C illustrate the time-based split: compounds are ordered by assay date and the most recent portion is held out for testing, emulating prospective deployment.

Similarly to the nRMSE, the nMAE can be defined as:

$$\text{nMAE} = \frac{\text{MAE} - y_{\min}}{y_{\max} - y_{\min}} \quad (2.5)$$

2.2.5.2 Assessing models the right way

Proper evaluation of predictive models is essential to ensure their generalizability and robustness [216, 217]. This is particularly important in lead optimization, where data can be highly heterogeneous and temporally biased due to evolving design priorities. Different strategies exist to split the data into training and test sets, each with its own benefits and limitations. Figure 2.2 describes the three ways of splitting data.

Random train/test split The most straightforward approach to evaluate a model is the random train/test split. In this method, the dataset is randomly divided into training and test subsets, typically using a fixed proportion (e.g., 80% for training and 20% for testing). While simple to implement, this approach assumes that the data are independently and identically distributed (i.i.d.), which may not hold in real-world drug discovery scenarios. In particular, molecules from the same chemical series may be split across training and test sets, leading to overly optimistic performance estimates due to information leakage. This split is represented on the Figure 2.2A.

Cross-validation Cross-validation improves upon the random split by repeating the train/test evaluation multiple times. In k -fold cross-validation, the dataset is partitioned into k subsets of approximately equal size. Each fold is used once as the test set while the remaining $k - 1$ folds serve as the training set. This method reduces the variance of the performance estimate and provides a more reliable assessment, especially for small datasets. However, standard cross-validation still assumes that data points are exchangeable, which is not ideal in settings where temporal or structural biases exist. This split is represented on the Figure 2.2B with a 5 splits cross-validation.

Time-based train/test split To more accurately reflect prospective deployment, particularly in drug discovery projects, a time-based train/test split is preferred. Here, compounds are divided based on their assay dates: the training set contains older compounds, while the test set includes newer ones. This mimics real-world scenarios where models are trained on past data and used to predict future outcomes. Although more challenging, this validation scheme provides a realistic estimate of model performance under prospective use and is better suited for decision-making in dynamic research environments. This split is represented on the Figure 2.2C.

For the 121 datasets extracted from legacy drug discovery projects, we have access to the date of the experimental assays. Based on this, we will only use time-based split for the benchmark.

2.3 Results

2.3.1 A brief overview

2.3.1.1 Datasets predictability

Before directly comparing descriptors, it is useful to first assess the inherent predictability of different properties. Figure 2.3 shows the distribution of R^2 scores obtained for all 482 descriptors (plus Chemprop) across 121 datasets. Throughout this analysis, we refer collectively to Random Forest, LightGBM, and XGBoost models as "ML models," reporting their average performance.

Modeling ADMET-related properties (Figure 2.3A,B) is notably challenging, as indicated by the many R^2 values below zero, highlighting poor predictability. However, Chemprop often emerges among the top-performing models for these tasks.

In contrast, potency prediction (Figure 2.3C) generally yields higher R^2 scores, predominantly above zero, with some reaching around 0.7. Interestingly, Chemprop performs relatively poorly in this category, frequently ranking among the lowest scores.

However, relying solely on R^2 as the primary metric has limitations, especially for small datasets

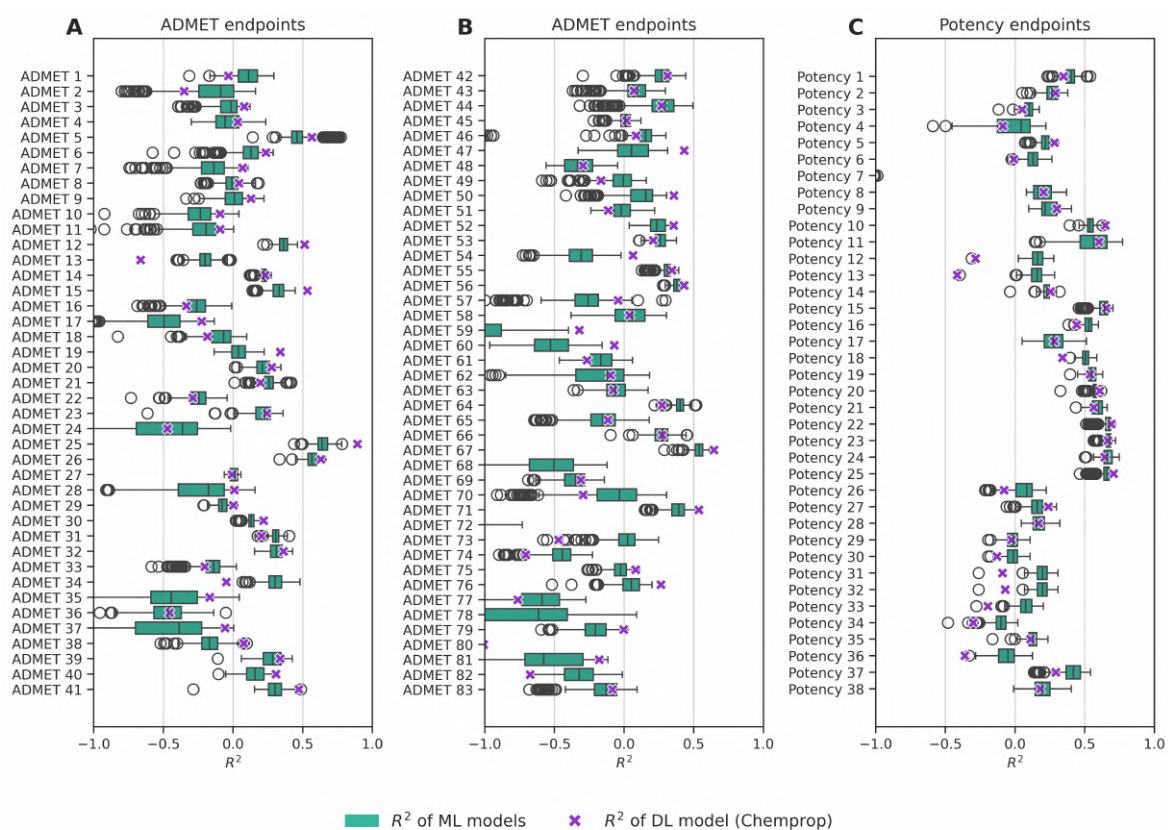


Figure 2.3: Distribution of R^2 scores for ADMET datasets (A, B) and potency datasets (C). Boxplots represent machine learning models trained on vector-based descriptors, while crosses represent Chemprop.

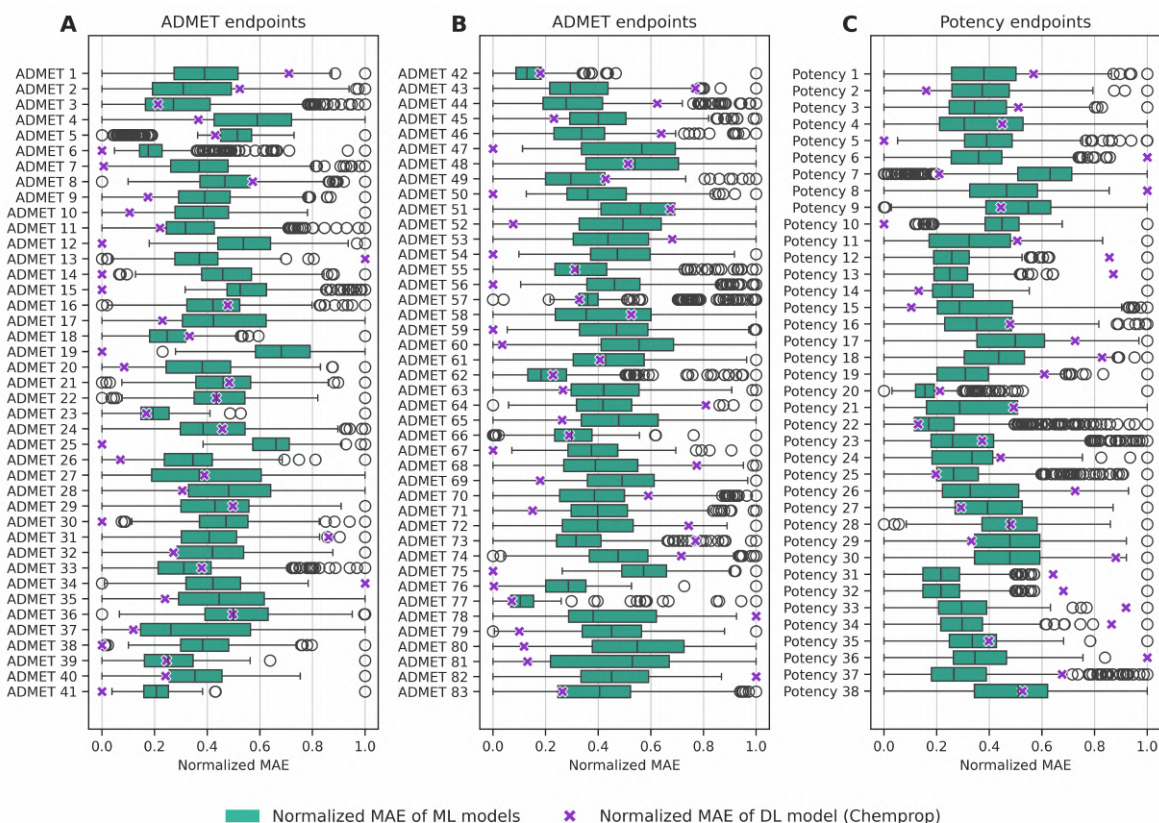


Figure 2.4: Distribution of nMAE scores for ADMET datasets (A, B) and potency datasets (C). Box-plots represent machine learning models trained on vector-based descriptors, while crosses represent Chemprop.

where a single data point can significantly impact the score. Therefore, metrics such as Mean Absolute Error (MAE) or its normalized variant (nMAE) are preferable for evaluating and comparing model performance. MAE is particularly advantageous over RMSE due to its robustness against outliers.

Similar to Figure 2.3, Figure 2.4 presents the distribution of nMAE scores for all 482 descriptors (plus Chemprop) across the 121 datasets. A corresponding detailed figure distinguishing individual models is provided in Appendix A.

Beyond robustness, this figure offers enhanced readability. Consistent with previous results, Chemprop frequently ranks among the best models for ADMET endpoints. However, for potency datasets, Chemprop achieves an nMAE below the median of the ML models in only 11 out of 38 datasets, and falls below the first quartile in just 6 out of 38 cases.

2.3.1.2 Influence of Dataset Size

As previously mentioned, the datasets analyzed in this study differ considerably in size. Figure 2.5 shows how dataset size impacts nMAE scores. The columns represent the analyzed models: column A corresponds to Chemprop, column B displays the average nMAE across all machine learning (ML) models, and column C represents the best-performing ML model for each dataset. The rows separate dataset types, with ADMET datasets in the first row and potency datasets in the second.

For ADMET endpoints, Chemprop (A1) generally achieves lower nMAE values compared to the average of all ML models (B1), especially as dataset size increases. However, Chemprop is often outperformed by the best ML model for each endpoint (C1), particularly with smaller datasets.

These performance differences are more pronounced for potency datasets. Chemprop (A2) consistently exhibits higher nMAE values than the average of all ML models (B2). When compared only to the best-performing ML models for each dataset (C2), the gap widens further, underscoring Chemprop's relatively weaker performance in this context.

2.3.1.3 Algorithms training time

Training time is a critical factor in machine learning, as it directly influences resource efficiency and the speed of experimentation. In this study, models were trained using their default hyperparameters, given the large number of descriptors being benchmarked. However, when models are fine-tuned using methods such as grid search or Bayesian optimization, the number of training iterations can increase significantly. As a result, training speed becomes a key criterion when selecting predictive algorithms.

Figure 2.6 presents the training times of the different models across all datasets and descriptors. Chemprop (Figure 2.6D), trained with the default 30 epochs, is by far the slowest model, particularly on smaller datasets. Among the traditional machine learning models, Random Forest (Figure 2.6A) is the most time-consuming, especially when scaling to large datasets (>1000 compounds). In contrast, XGBoost (Figure 2.6B) and LightGBM (Figure 2.6C) demonstrate high efficiency, with LightGBM showing the fastest training times overall.

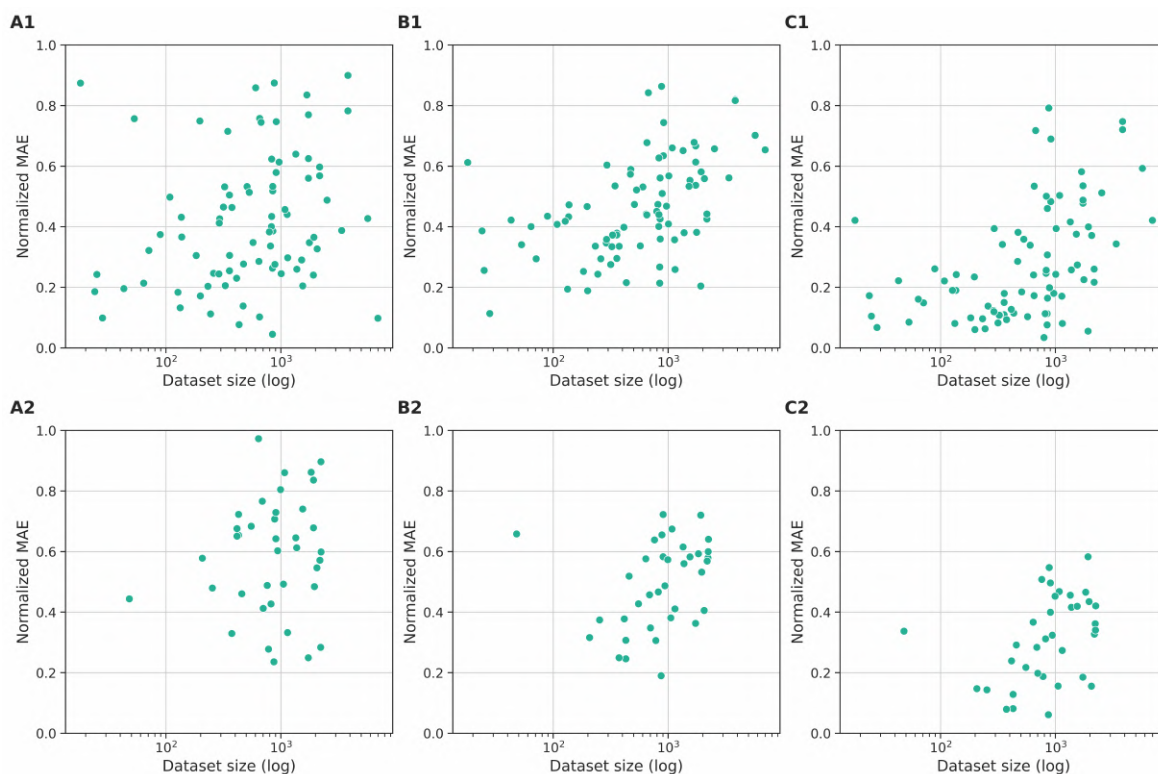


Figure 2.5: Impact of dataset size on normalized mean absolute error (nMAE). Columns represent different modeling approaches: (A) Chemprop model, (B) average nMAE across all machine learning (ML) models, and (C) the best-performing ML model for each dataset. Rows differentiate between dataset types, with the first row showing ADMET datasets and the second potency datasets. The results highlight Chemprop’s relative strength on larger ADMET datasets, but reveal its limitations compared to the best ML models, particularly on smaller and potency datasets.

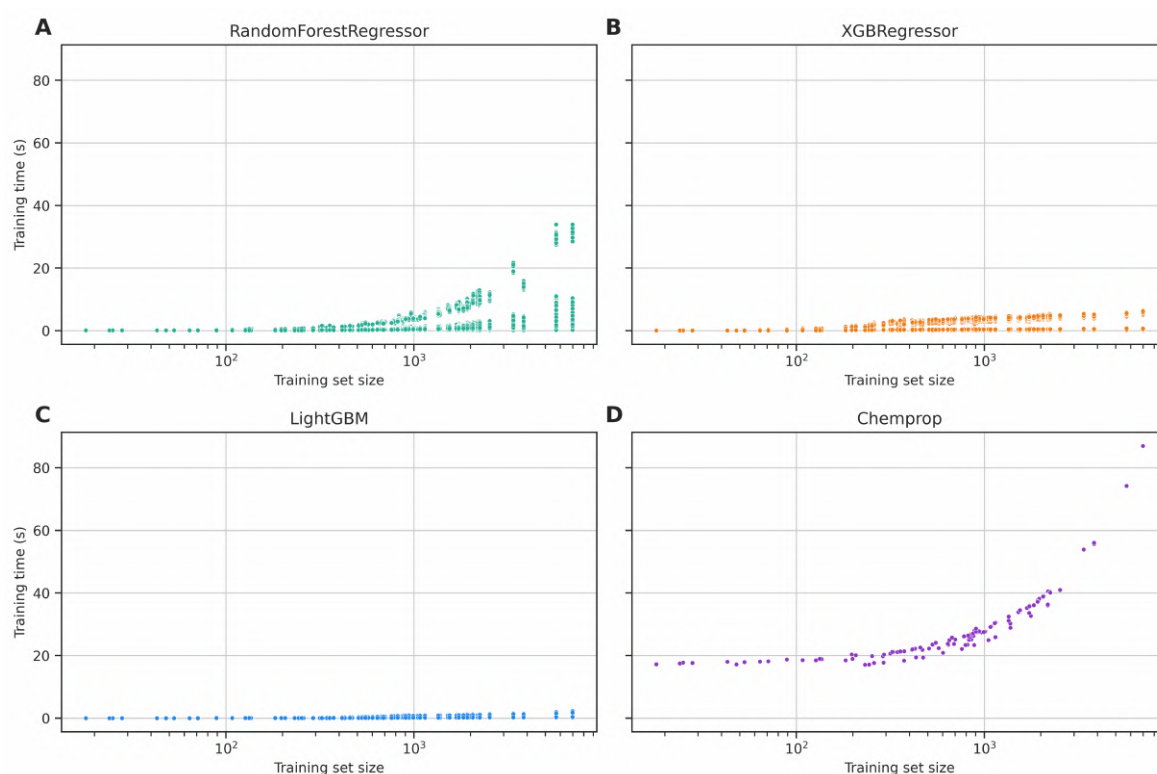


Figure 2.6: Training time as a function of training set size for all models. Panels A to C represent conventional machine learning models: for each of the 121 datasets, the models are retrained with 482 different descriptor sets, so multiple points appear at the same x value, giving 58322 points in total (121×482). Panel D is Chemprop: because Chemprop uses a single molecular representation, only one point is plotted per dataset, for 121 points overall.

2.3.2 Benchmarking single descriptors

First, let’s focus on the 35 single descriptors and compare them to Chemprop.

2.3.2.1 Ranking

To compare descriptor performance, we ranked them based on their average score across the 121 datasets. Figure 2.7 shows the distribution of scores for each descriptor using three evaluation metrics: the R^2 score (Figure 2.7A), the normalized root mean squared error (nRMSE; Figure 2.7B), and the normalized mean absolute error (nMAE; Figure 2.7C). Descriptors are ordered by their mean performance across all datasets.

MACCS, Topological Torsion, and AtomPair fingerprints consistently rank among the lowest-performing descriptors across all three metrics. In contrast, ECFP- and Avalon-based descriptors rank highly, alongside Chemprop. Notably, the score distributions for Chemprop exhibit a wider interquartile range in both the nRMSE and nMAE rankings, suggesting that its performance may be less robust and more sensitive to dataset variability.

It is also worth noting that the score distributions appear more uniform when using the R^2 metric, likely reflecting the instability of this metric on small datasets. For the following results, we will be using the nMAE as reference as the MAE is a more robust metric to outliers compared to the RMSE.

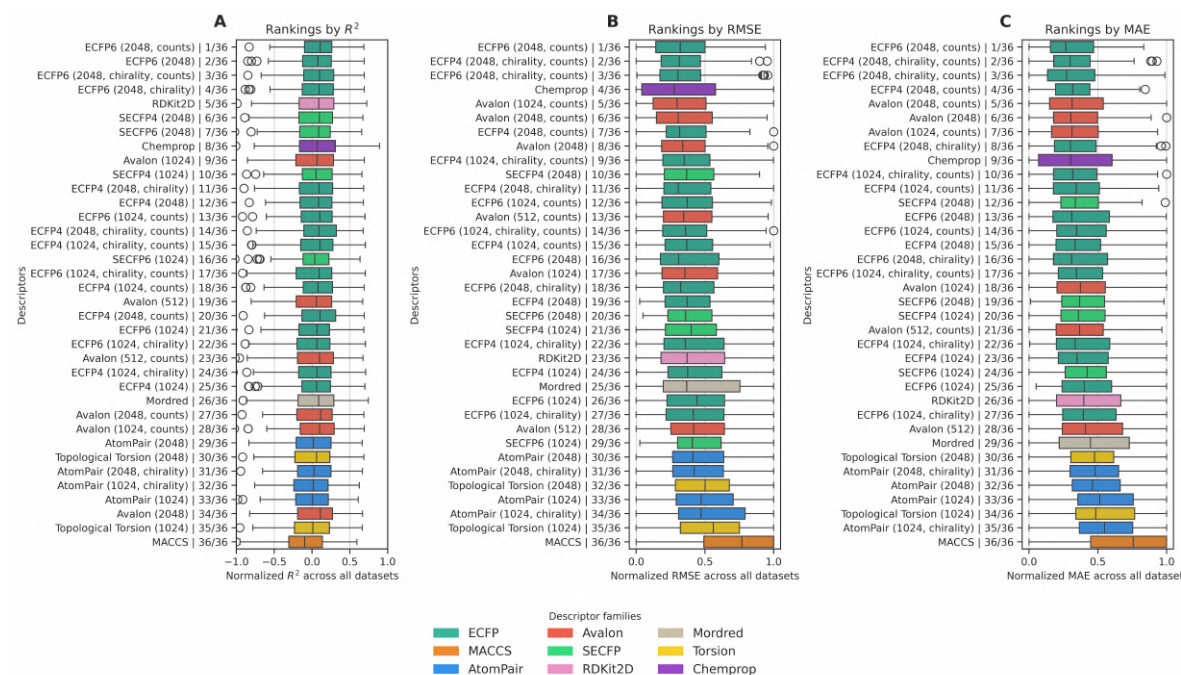


Figure 2.7: Rankings of the single descriptors and Chemprop based on the average on 121 datasets for R^2 (A), nRMSE (B) and nMAE (C) metric distributions. Single descriptor variants are colored based on their descriptor family.

2.3.2.2 Differentiating ADMET and potency

As previously noted in the analysis of datasets predictability, ADMET properties and potency are not modeled with equal ease. Figure 2.8 presents descriptor rankings based on the nMAE. Figure 2.8A shows the overall rankings across all endpoints (as in Figure 2.7C), while Figure 2.8 B and C display rankings specific to ADMET and potency endpoints, respectively.

Chemprop performs well when considering all endpoints together, ranking 9th out of 36 descriptors. However, disaggregating the results by endpoint type reveals stark differences. For ADMET properties, Chemprop achieves the lowest average nMAE, despite a wide interquartile range, making it the top-performing descriptor. In contrast, for potency prediction, Chemprop ranks among the worst performers.

Among the vector-based descriptors used with classical machine learning models, the ranking is relatively stable across tasks. ECFP variants consistently occupy the top positions, particularly those incorporating count information and/or chirality, highlighting their robustness across endpoint types.

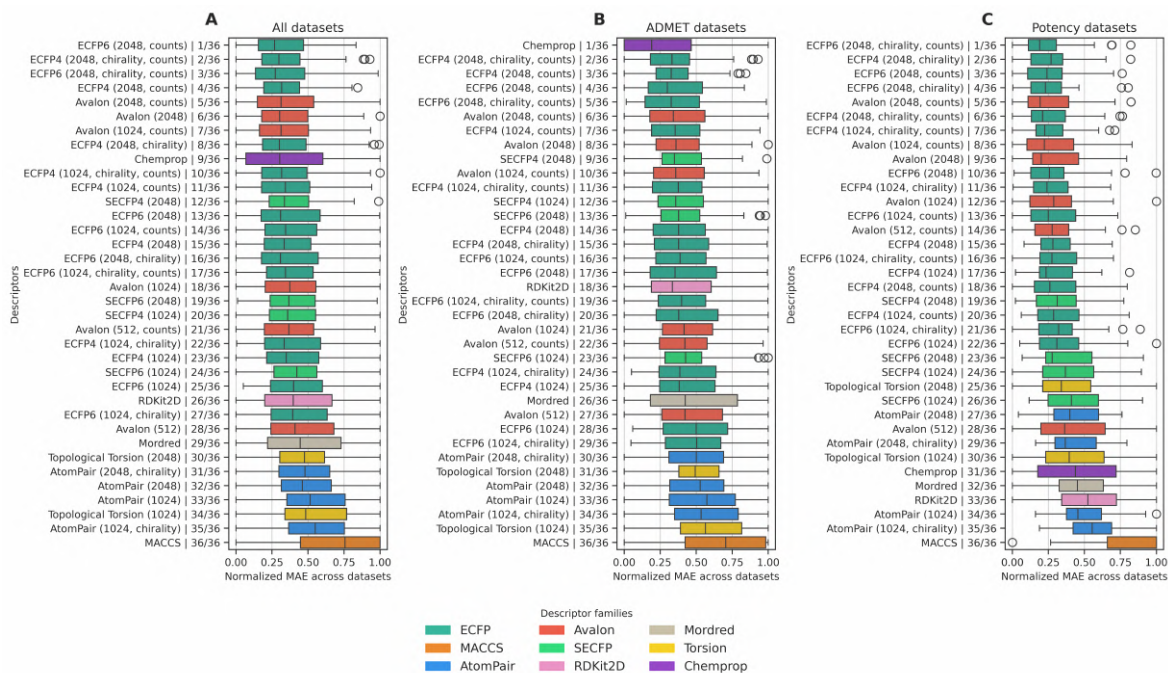


Figure 2.8: Rankings of the single descriptors and Chemprop based on their average nMAE across all the 121 datasets (A), the 83 ADMET datasets (B) and the 38 potency datasets (C). Single descriptor variants are colored based on their descriptor family.

2.3.2.3 Statistical analysis

Although we ranked the descriptors based on their mean nMAE across datasets, the overlap in their score distributions necessitates statistical testing to assess whether the observed differences are significant. To this end, we applied the Friedman test, a non-parametric statistical method designed to detect performance differences among multiple methods evaluated on multiple datasets using rankings. This test is particularly well-suited for repeated measures scenarios such as QSAR/QSPR modeling, where each dataset is evaluated under several modeling conditions.

The Friedman test evaluates whether the average ranks of the descriptors differ more than would be expected by chance, without assuming normality or homoscedasticity. The null hypothesis states that all descriptors perform equally well, meaning that differences in rankings arise purely from random variation.

We applied the Friedman test separately to ADMET and potency endpoints. In both cases, the test returned a p-value inferior to 10^{-20} , allowing us to confidently reject the null hypothesis. This indicates that the performance differences among at least two descriptors is statistically significant. However, the Friedman test does not identify which specific descriptor pairs differ. To explore this, we conducted a Nemenyi post-hoc analysis, which compares all pairwise differences in average ranks. If the difference between two descriptors exceeds a critical threshold, known as the critical

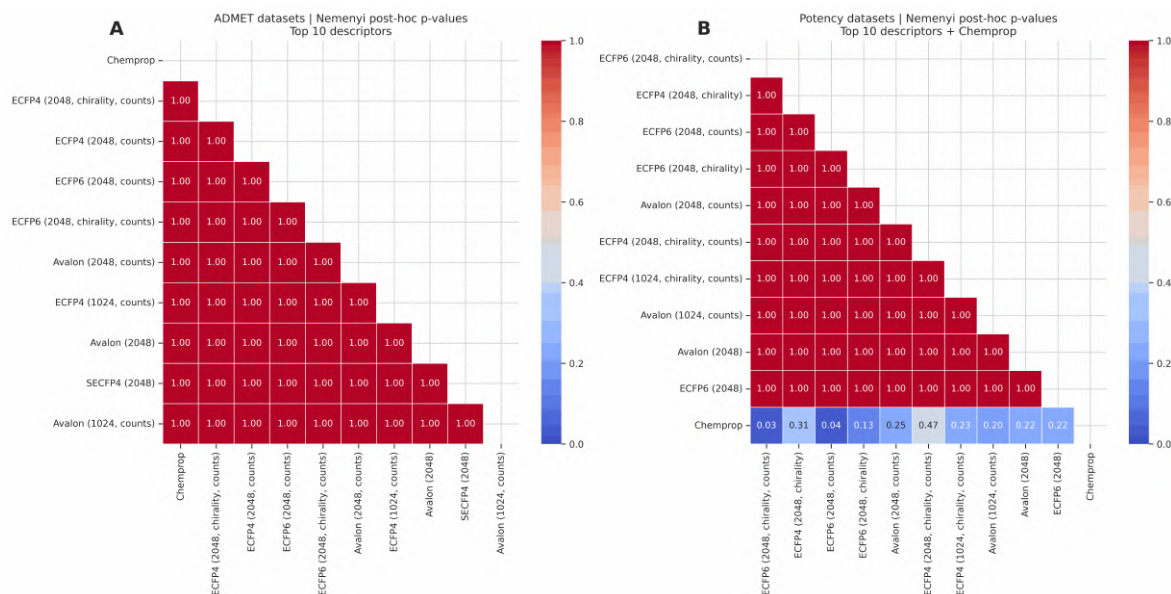


Figure 2.9: Pairwise p-values for the ten highest-ranked single descriptors, plus Chemprop when it falls outside this group, for the ADMET datasets (A) and potency datasets (B). A p-value of 1 indicates that the rank orders of the two descriptors do not differ significantly, whereas values approaching 0 signify a statistically significant difference in their rankings.

difference, their performance is considered significantly different.

Figure 2.9 presents the pairwise p-values from the Nemenyi analysis for ADMET (Figure 2.9A) and potency (Figure 2.9B) endpoints. For clarity, we show only the ten highest-ranked descriptors, and include Chemprop if it falls outside the top ten.

For ADMET endpoints, although Chemprop ranks first, its pairwise p-values with the next nine descriptors are all equal to 1, indicating that it is not statistically better than any of them. In contrast, for potency endpoints, Chemprop yields smaller p-values when compared with the top ten descriptors. In particular, its p-values against ECFP6 (2048, chirality, counts) and ECFP6 (2048, counts) are 0.03 and 0.04, respectively, both below the 0.05 significance threshold. This result indicates that Chemprop performs significantly worse than these two descriptors for potency modeling across the 38 potency datasets.

2.3.3 Benchmarking combinations of descriptors

Among the 482 vector-based descriptors evaluated, only 35 are single descriptors. The remaining 447 are composite descriptors obtained by concatenating pairs of single descriptors from different families. Several studies have suggested that combining fingerprints can be beneficial, as it enables the integration of complementary structural information [228, 229, 230]. In this section, we extend the analysis to all 482 vector-based descriptors used with machine learning models, alongside the

Chemprop model based on graph embedding.

2.3.3.1 Ranking

As in the previous section, descriptors were ranked according to their average nMAE across all datasets. Figure 2.10 presents this ranking. For clarity, only the top 30 descriptors are shown, with Chemprop included if it does not appear among them.

This figure shows that the top-ranked descriptors are all combinations involving ECFP variants, often concatenated with Avalon, and occasionally with RDKit2D descriptors or MACCS keys. Notably, none of the top descriptors are single-feature vectors, reinforcing the idea that combining descriptors can enhance the predictive power by enriching the feature space. While the top 30 descriptors perform similarly in terms of their average nMAE, their distributions overlap considerably, suggesting that further statistical analysis is required to assess the significance of their differences.

Chemprop ranks 223rd out of 483, a substantial drop compared to its earlier position when only single descriptors were considered. Its interquartile range remains wide, once again indicating a lack of robustness relative to the top performing vector-based descriptors, which exhibit narrower performance distributions across datasets.

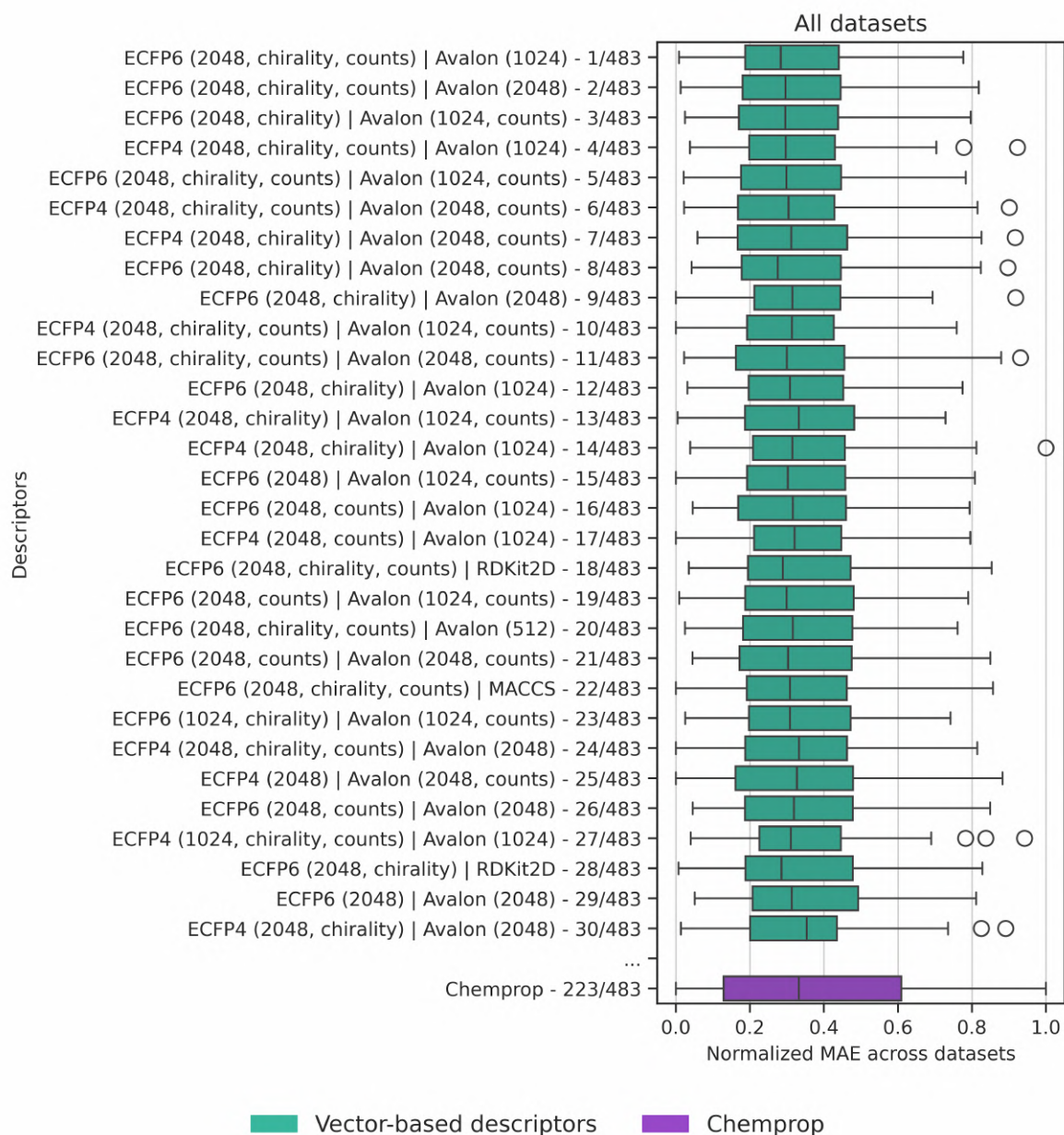


Figure 2.10: Rankings of all the descriptors combinations based on their average nMAE across all the 121 datasets. Vector-based descriptors are in green while Chemprop is in purple.

2.3.3.2 Differentiating ADMET and potency

Differentiating between ADMET and potency datasets reveals markedly distinct descriptor rankings. Figure 2.11A shows the rankings based on ADMET datasets, while Figure 2.11B presents the rankings for potency datasets. As before, only the top 30 descriptors are shown for clarity.

In the ADMET-based ranking (Figure 2.11A), Chemprop again appears as the top-performing model. However, this result should be interpreted with caution due to its wide performance distribution, both in terms of interquartile range and maximum nMAE, highlighting variability and limited robustness across datasets. Among the top-ranked vector-based descriptors, most are combinations involving ECFP variants, consistent with observations from the full dataset. Notably, these combinations now more frequently involve RDKit2D descriptors rather than Avalon fingerprints, indicating a shift in what contributes most to predictive performance in ADMET tasks.

In contrast, the potency-based ranking (Figure 2.11B) is dominated by combinations of ECFP variants and Avalon fingerprints across all top 30 descriptors. The addition of RDKit2D descriptors appears less beneficial in this context, which aligns with previous results showing that RDKit2D alone performs poorly on potency datasets (Figure 2.8C). Chemprop performs substantially worse for potency modeling, ranking 476th out of 483, nearly last, marking a sharp contrast to its performance on ADMET endpoints.

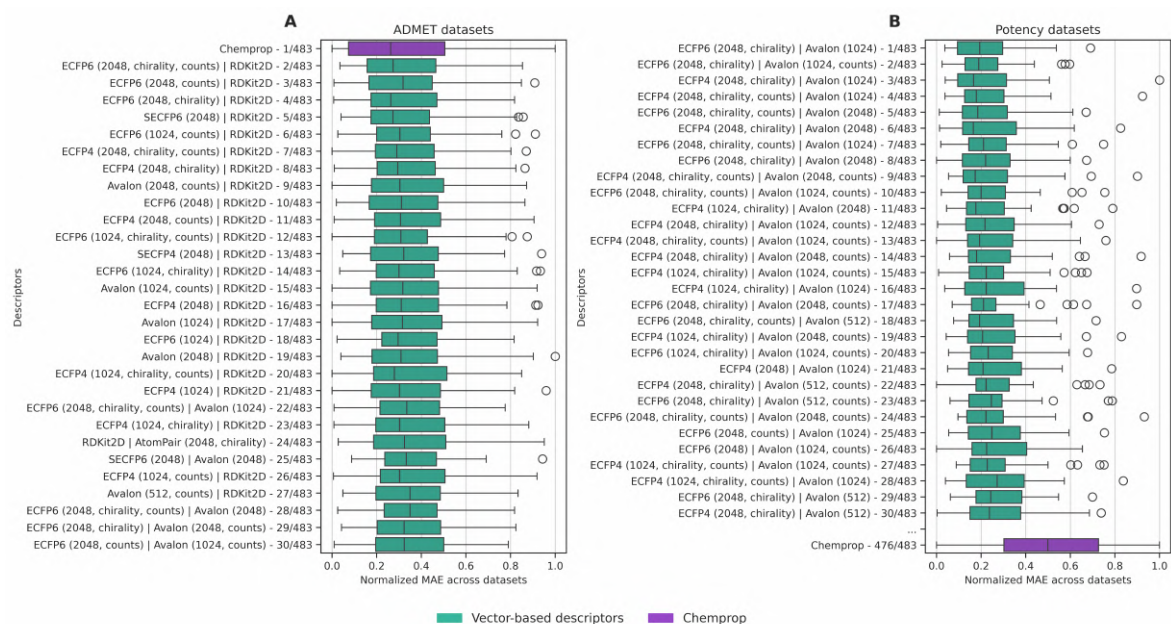


Figure 2.11: Rankings of all the descriptors based on their average nMAE across ADMET datasets (A) and potency datasets (B). Vector-based descriptors are in green while Chemprop is in purple.

2.3.3.3 Statistical analysis

To assess whether the observed rank differences among the 482 descriptors combinations (plus Chemprop) are statistically significant, we applied the same non-parametric Friedman test previously used for single descriptors. This test was conducted separately for the ADMET-based and potency-based rankings. In both cases, the Friedman test returned p-values inferior to 10^{-20} , leading to a clear rejection of the null hypothesis. This means that at least two descriptors combinations are statistically different.

We then conducted a Nemenyi post-hoc analysis, which compares all pairs of descriptors and identifies those with rank differences exceeding the critical difference (CD) as significantly different.

Figure 2.12 presents the pairwise p-values from the Nemenyi test for ADMET (Figure 2.12A) and potency (Figure 2.12B) endpoints. As with the analysis of single descriptors, we show only the ten highest-ranked descriptors and include Chemprop if it falls outside this group.

For ADMET-based rankings, the results mirror those obtained for single descriptors: although Chemprop ranks first, its performance is not statistically different from that of the next highest-ranked descriptors. In contrast, for potency-based rankings, Chemprop is significantly outperformed by the top descriptors. The pairwise p-values are close to 0 ($\ll 0.05$), confirming that models combining machine learning with the top-ranked descriptors significantly outperform Chemprop across the 38 potency datasets.

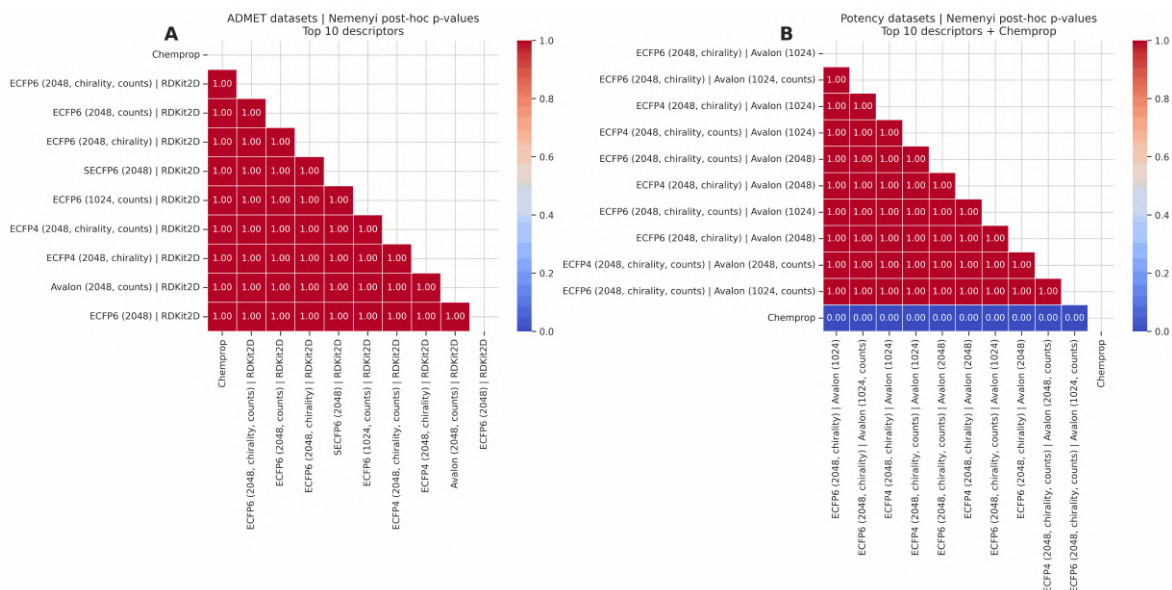


Figure 2.12: Pairwise p-values for the ten highest-ranked descriptors, plus Chemprop when it falls outside this group, for the ADMET datasets (A) and potency datasets (B). A p-value of 1 indicates that the rank orders of the two descriptors do not differ significantly, whereas values approaching 0 signify a statistically significant difference in their rankings.

2.3.4 Time-based Splitting vs. Cross-validation

In previous analyses, we employed an 80/20 time-based split, where the training set included the earliest 80% of compounds within a project (ordered by assay date), and the test set contained the most recent 20%. This strategy closely mirrors how predictive models are typically deployed in real-world drug discovery applications [216, 217]. However, implementing a time-based split is challenging outside industrial contexts, due to limited availability of datasets with timestamped data.

This raises an important question: Is a time-based split necessary, or can standard cross-validation on temporally agnostic datasets yield comparable insights? To explore this, we compared the predictive performance across all datasets using both splitting methods.

For fingerprint-based models, we conducted a total of 174 966 experiments (482 descriptors $\times$ 121 datasets $\times$ 3 predictive algorithms), each evaluated using both cross-validation and time-based splits. Figure 2.13 shows the distribution of normalized Mean Absolute Error (nMAE) scores. Cross-validation generally yielded lower and narrower nMAE distributions than time-based splits, suggesting that temporal shifts in chemical properties significantly reduce predictive accuracy when using time-based splits. Similar trends were observed for R^2 scores (Figure 2.14). Quantitatively, among the 174 966 experiments, 93.7% had higher R^2 scores under cross-validation compared to time-based splitting.

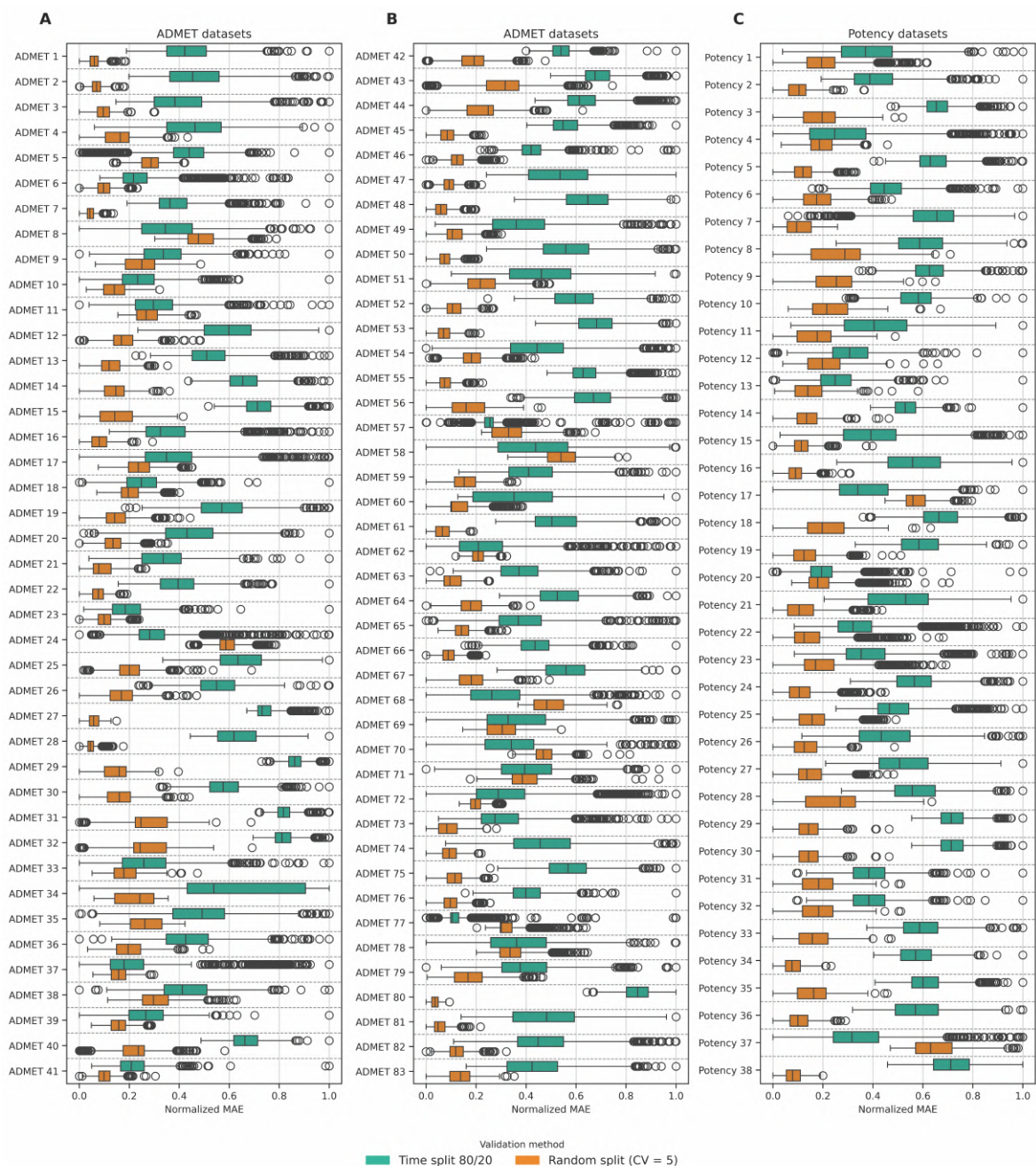


Figure 2.13: Distribution of nMAE values across all vector-based descriptors for ADMET datasets (A and B) and potency datasets (C). Evaluation using a temporal split are in green and using a cross-validation in orange.

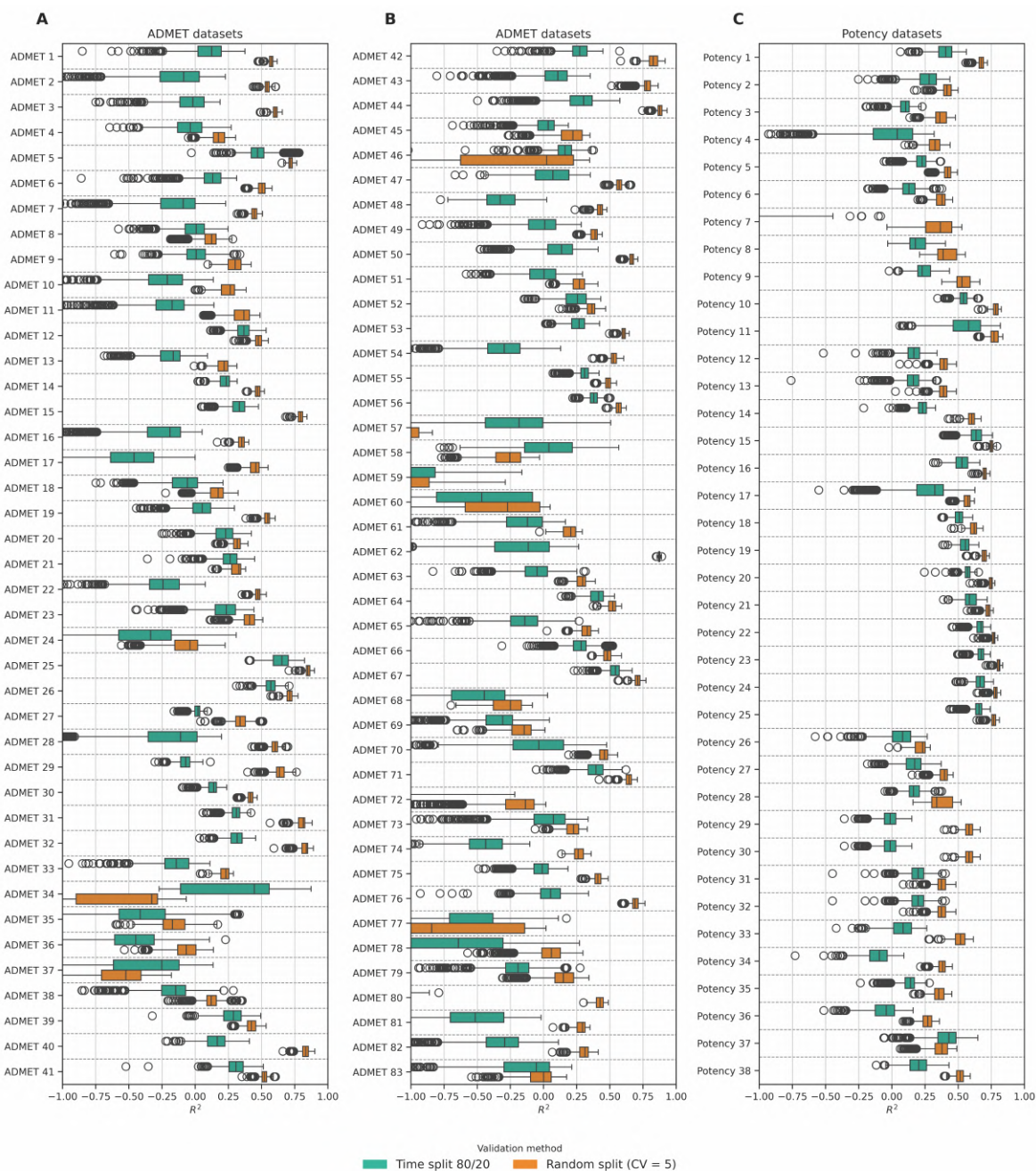


Figure 2.14: Distribution of R^2 values across all vector-based descriptors for ADMET datasets (A and B) and potency datasets (C). Evaluation using a temporal split are in green and using a cross-validation in orange.

We performed a similar comparison using the Chemprop model. Figures 2.15 and 2.16 illustrate the distribution of nMAE and R^2 scores, respectively, across all 121 datasets. The results align with the fingerprint-based models: Chemprop also consistently performed better under cross-validation than time-based splits. Specifically, Chemprop showed superior performance under cross-validation 92.6% of the time (112 out of 121 datasets), reinforcing that temporal shifts pose significant challenges for prospective model generalization.

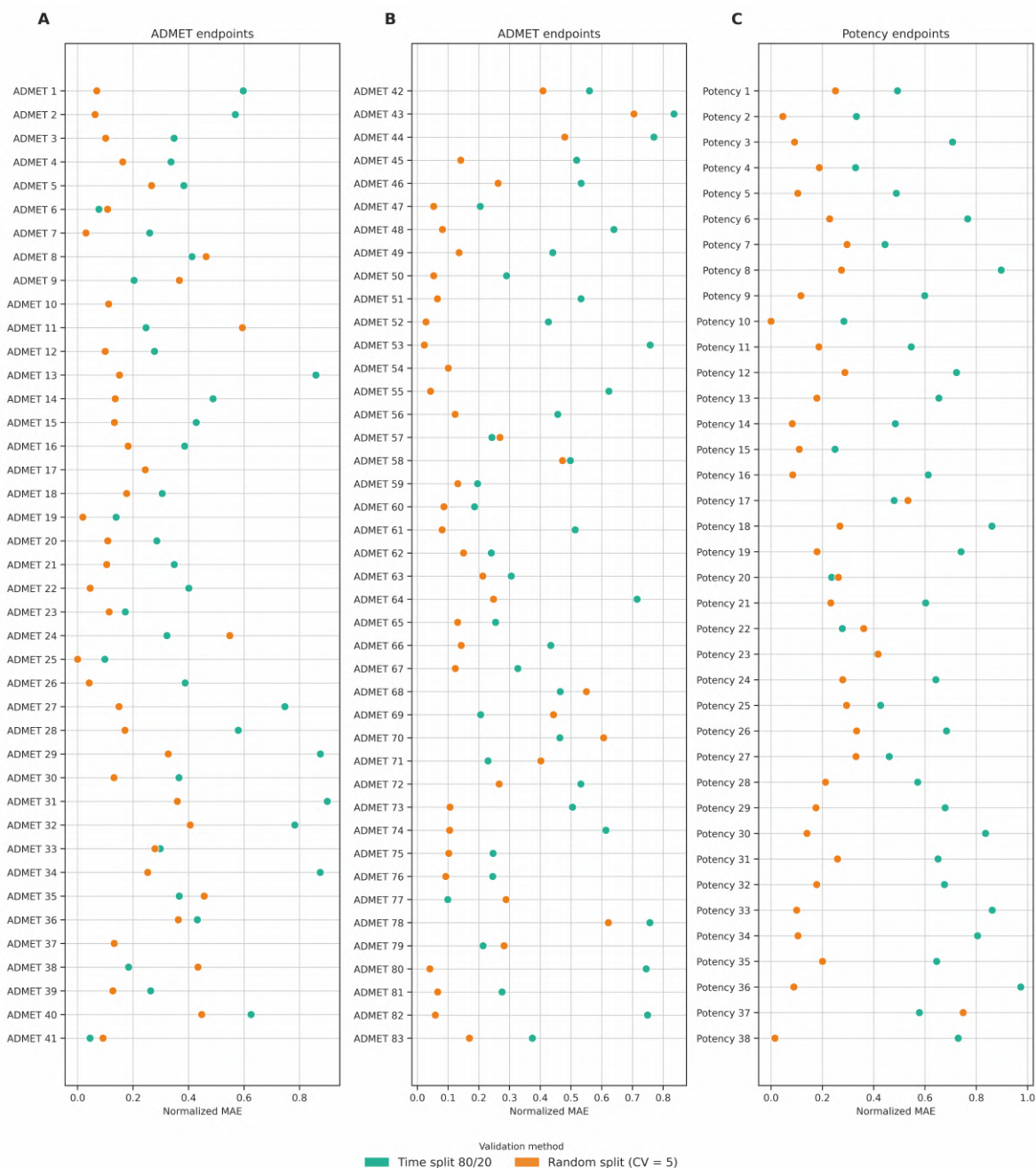


Figure 2.15: Distribution of nMAE values using Chemprop for ADMET datasets (A and B) and potency datasets (C). Evaluation using a temporal split are in green and using a cross-validation in orange.

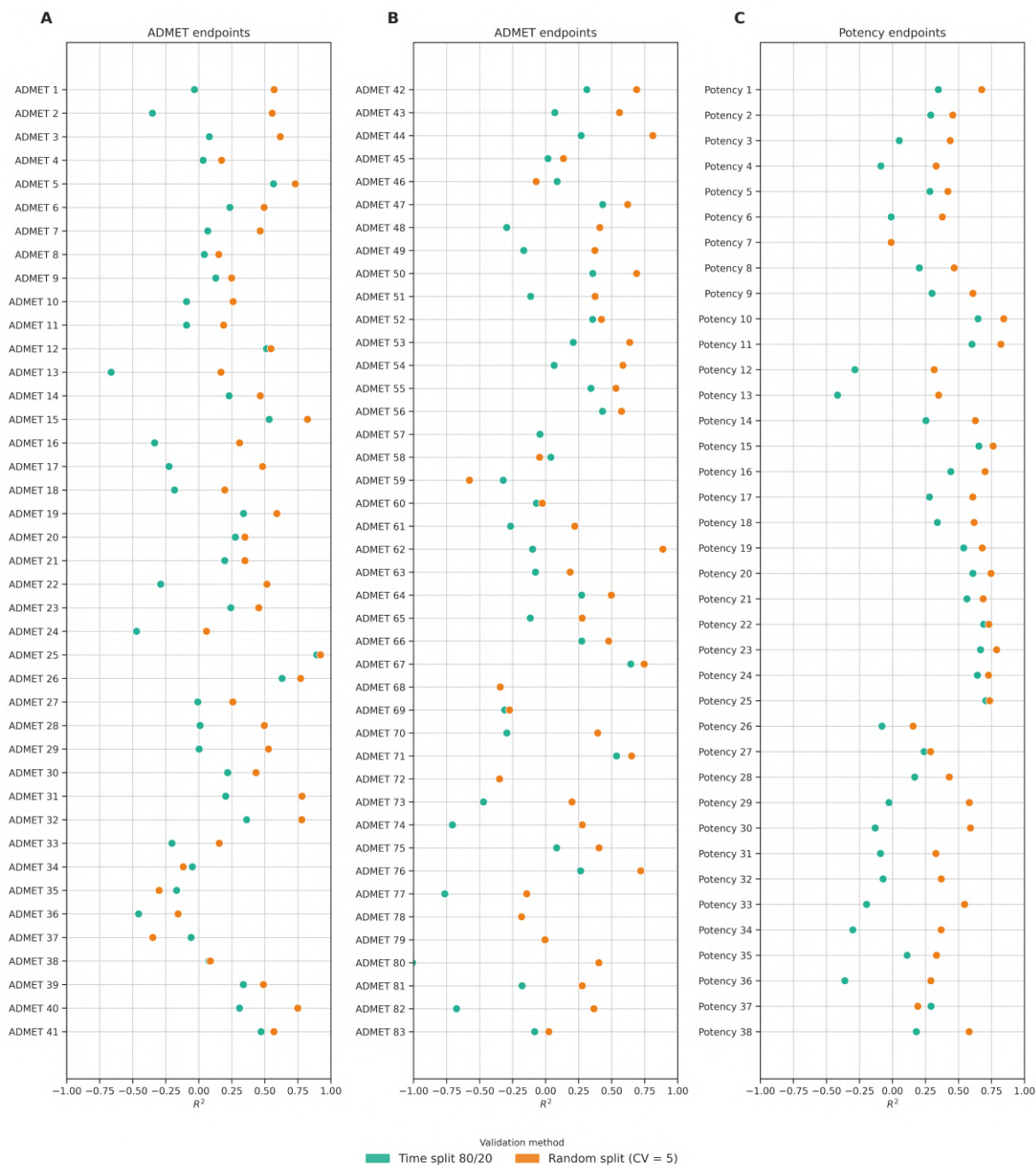


Figure 2.16: Distribution of R^2 values using Chemprop for ADMET datasets (A and B) and potency datasets (C). Evaluation using a temporal split are in green and using a cross-validation in orange.

2.4 Discussion

2.4.1 Benefits of combining descriptors

Concatenating molecular fingerprints significantly improves predictive performance, surpassing results from individual descriptors alone. This improvement occurs because descriptor concatenation expands and enriches the feature space, allowing machine learning algorithms to identify additive structure–activity relationship (SAR) patterns. Further research should investigate incorporating additional descriptors and combinations beyond descriptor pairs to enhance these benefits.

2.4.2 Deep learning is not all we need

While graph neural networks like Chemprop are often viewed as state-of-the-art for QSAR/QSPR modeling [209, 231, 232, 233], our comparative analysis against classical machine learning models reveals nuanced insights. Chemprop performs competitively when single descriptors are used but underperforms compared to classical methods using descriptor combinations. The data-intensive nature of graph neural networks likely contributes to this gap, especially noticeable in smaller potency datasets compared to larger ADMET datasets. Chemprop also exhibited greater variability across datasets and required longer training times, highlighting practical limitations relative to classical methods.

2.4.3 Differentiating ADMET and potency

When considering all datasets collectively, combinations of ECFP variants and Avalon fingerprints generally led to the best model performance. However, endpoint-specific analyses revealed notable differences. For potency datasets, combinations of ECFP and Avalon remained highly effective, while ADMET datasets performed better using combinations involving RDKit2D descriptors. This difference likely arises because ADMET properties depend more heavily on bulk physicochemical characteristics captured by RDKit2D, whereas potency is influenced primarily by structural complementarity between ligand and target.

2.4.4 The need for real-world proof benchmarks

Many benchmarks rely on noisy datasets that do not fully reflect the complexity and practical constraints of drug discovery. Our findings emphasize the necessity of temporal validation: more than 90% of the experiments showed that cross-validation splits inflated performance estimates compared to time-based splits. This underscores the importance for the pharmaceutical industry to share ro-

bust, temporally annotated datasets, as advocated by both academia and industry [183].

2.5 Conclusion

In conclusion, state-of-the-art deep learning models do not consistently outperform classical machine learning approaches based on molecular fingerprints, especially when working with industrial datasets in low-data regimes. This trend becomes even clearer when multiple fingerprints are combined to create richer molecular descriptors, effectively capturing complementary aspects of structure-activity and structure-property relationships. Thus, for future QSAR/QSPR modeling, combining ECFP-based fingerprints with classical algorithms like Random Forest or LightGBM provides a robust baseline that will be used for the rest of this work.

Chapter 3

Molecular prioritization toolkit and retrospective simulations

This chapter introduces a framework for evaluating molecular prioritization strategies aimed at accelerating the lead-optimization phase of drug discovery. Lead optimization is intrinsically challenging: it requires the simultaneous optimization of multiple, often-conflicting objectives, must operate under shifting data distributions, and typically begins with low information.

To meet these challenges we developed a comprehensive decision-support toolkit that unifies selection strategies derived from active learning (AL), multiple-criteria decision analysis (MCDA), and classical medicinal-chemistry heuristics. We benchmark these strategies with a retrospective simulation protocol that replays entire historical projects under controlled virtual scenarios. This chapter details both the methodological foundations and the practical implementation of the toolkit.

The chapter is structured as follows: initially, we discuss the theoretical underpinnings and practical constraints of molecular prioritization within lead optimization. Subsequently, we detail the retrospective simulation methodology, encompassing virtual scenario configuration, predictive modeling techniques, and selection strategies. Lastly, we introduce analytical tools crafted to quantitatively and qualitatively assess how effectively each strategy balances exploration and exploitation of the chemical space, exemplifying the approach through practical demonstrations involving a dataset from a project targeting Factor Xa.

3.1 Enhancing lead optimization

3.1.1 Challenges in lead optimization

Lead optimization inherently involves multi-objective decision-making, as previously outlined in Chapter 1. A viable drug candidate must exhibit not only sufficient on-target potency but also acceptable selectivity, pharmacokinetics, and safety profiles. These criteria often conflict, necessitating careful management of complex trade-offs.

Adding complexity is the dynamic nature of drug discovery workflows. As projects advance, the synthesis and evaluation of new compounds continually alter underlying data distributions. Initially, SAR data may derive from narrow chemical series; however, the introduction of new scaffolds or substituents expands chemical space and reduces the predictive accuracy of earlier models. Essentially, the "rules" governing biological activity and potential liabilities evolve as teams gain deeper insights into the biological target and associated chemistry. This temporal distribution shift, discussed in Chapter 2, significantly degrades model performance, with historically trained models exhibiting notably lower accuracy when predicting new compounds compared to cross-validation on randomly shuffled datasets. Studies have further indicated that temporal shifts substantially impact uncertainty estimations [234].

An additional challenge arises from the prevalent low-data environment common to early-stage discovery. New projects typically start with limited active compound data and sparse experimental outcomes. Consequently, lead optimization teams operate with minimal SAR data, particularly when exploring novel targets or unprecedented mechanisms. Thus, teams must carefully balance exploiting known chemical spaces to optimize promising scaffolds against exploring novel chemical spaces to discover potentially superior compounds.

3.1.2 From DMTA to D(S)MTA

Historically, molecular design relied primarily on the intuition and creativity of medicinal chemists, complemented by predictive modeling tools and computational methods. Today, chemists are accompanied by AI models capable of both predicting molecular properties and generating novel compounds [75]. Consequently, each design cycle now yields significantly more hypotheses from dual sources: the chemists and generative AI, as illustrated in Figure 3.1. Molecules proposed by generative models often deviate from traditional chemist-generated structures, presenting assessment challenges due to their unconventional nature [95, 96, 97, 98, 99]. While some generative models facilitate exploration around known scaffolds [94, 100], their primary strength lies in introducing entirely novel chemical entities, substantially broadening chemical space exploration.

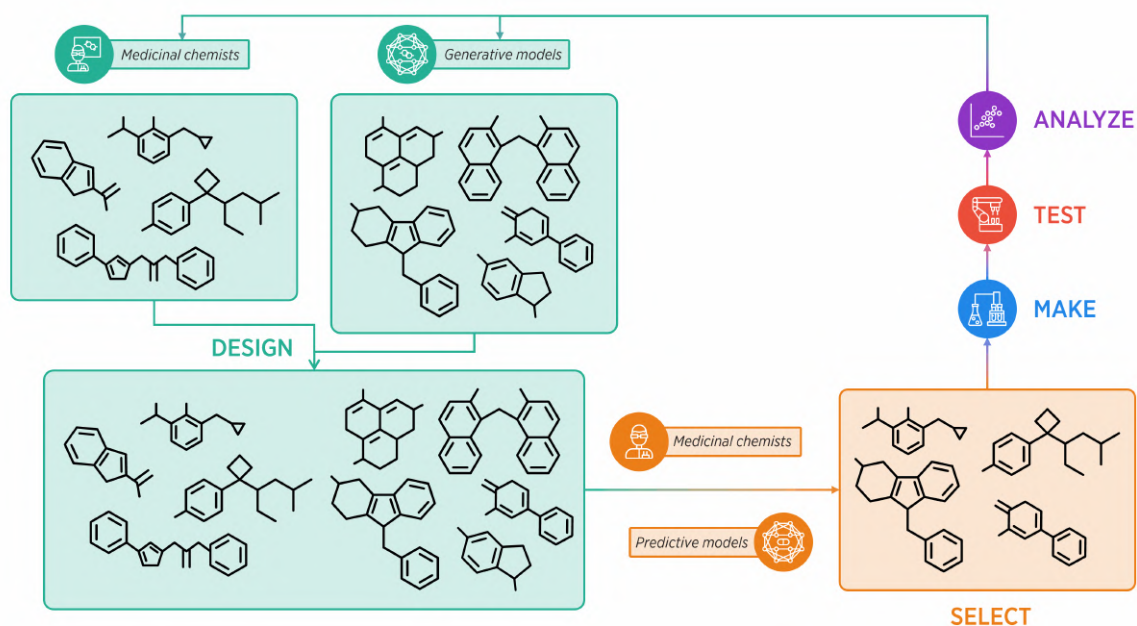


Figure 3.1: Overview of the D(S)MTA cycle, highlighting compound design by medicinal chemists and generative models.

This new design paradigm is both a blessing and a curse. Historically, selecting molecules to synthesize aligned intuitively with the chemists' own design rationale, potentially backed by predictive models. Now, however, the selection step has grown substantially more intricate, requiring chemists to evaluate a large array of compounds, including those generated by AI that may be counterintuitive to their expertise [235]. Consequently, one could say that selection has evolved into a dedicated stage of the D(S)MTA cycle [236], demanding careful consideration and integration of hypotheses from both human intuition and generative AI.

3.1.3 Active learning in drug discovery

Drug discovery, characterized by a vast chemical space combined with costly and time-intensive experimental validations, represents an ideal context for applying AL methodologies. AL has already demonstrated significant utility across various research stages, particularly in virtual screening for hit identification [41, 42, 237, 238, 239]. Typically, existing AL applications focus on efficiently scanning extensive compound libraries to discover initial hits or to construct QSAR models aimed at predicting a single property based on limited assay data [153, 240]. These approaches generally emphasize a single objective: maximizing the number of identified active compounds or enhancing predictive model accuracy while minimizing data requirements.

Although AL implementations have predominantly targeted early discovery phases concerned with hit identification and single-objective optimization, this approach also holds considerable promise

for addressing the inherently multi-objective and iterative nature of lead optimization. Lead optimization frequently operates in scenarios characterized by low-data availability, where each iterative cycle yields only a small number of newly synthesized analogs accompanied by corresponding assay results, thereby greatly increasing the value of each compound tested. AL is particularly suited to such low-data conditions due to its capacity to identify and select the most informative compounds for subsequent experimental evaluation.

Consequently, AL principles could be strategically employed to guide compound selection within DMTA cycles, significantly enhancing the efficiency and efficacy of knowledge acquisition regarding SAR and the intricate trade-offs among desired properties. Conceptually, the iterative cycle inherent to AL, comprising modeling, querying, experimentation, and updating, closely aligns with the DMTA cycle of drug discovery, which involves the iterative processes of designing, synthesizing, testing, and analyzing compounds. Both frameworks utilize recent experimental outcomes to iteratively refine and inform strategic decisions for subsequent cycles, underscoring the direct adaptability and substantial potential of active learning methodologies for lead optimization contexts.

3.1.4 Multi-objective prioritization toolkit

To effectively support medicinal chemists in managing the extensive set of hypotheses generated during the molecular design phase, it is crucial to employ rational selection strategies. A substantial part of this thesis was devoted to developing a comprehensive decision-support toolkit specifically designed to aid in the prioritization of candidate molecules within a multi-objective context.

This toolkit incorporates a broad range of acquisition functions (selection strategies), combining both established methods and novel approaches developed as part of this research. It draws extensively from active learning concepts by implementing strategies that prioritize compounds according to uncertainty estimates. Additionally, methodologies from multiple-criteria decision analysis have been adapted to effectively address and reconcile conflicting molecular properties, ensuring that selected compounds align closely with the complex, multifaceted objectives encountered in drug discovery projects. Finally, the toolkit integrates strategies inspired by medicinal chemists' typical behavioral patterns and their intuitive decision-making processes.

3.1.5 Challenge of benchmarking selection strategies

The vast majority of research on the development of selection strategies focuses primarily on virtual screening, specifically targeting the hit identification phase [42, 41]. In this setting, the compound pool from which molecules are selected remains fixed and typically comprises very large screening

datasets, such as LIT-PCBA [241]. Furthermore, these studies generally optimize only a single endpoint, which is biological activity against the target. However, the lead optimization phase differs significantly from virtual screening. Here, the compound pool is dynamic, evolving continuously as the project advances, with each optimization cycle informing subsequent cycles. Additionally, lead optimization inherently involves a multi-objective optimization problem; beyond potency, it requires simultaneous consideration of multiple ADMET properties.

Rigorously benchmarking selection strategies for lead optimization ideally involves conducting actual DMTA cycles within authentic drug discovery projects. In this ideal scenario, the compound design stage explicitly integrates an acquisition function to guide decisions regarding synthesis and testing. However, executing real-world DMTA cycles is prohibitively expensive, as it effectively requires carrying out a complete drug discovery campaign. Although recent advancements have introduced automated synthesis platforms [168, 242, 243, 244], these technologies remain costly and insufficiently mature to emulate complete drug discovery pipelines. Furthermore, benchmarking multiple selection strategies through this approach would necessitate running separate, parallel projects for each strategy, making such comparisons inconceivable.

To circumvent these practical constraints, one alternative is simulating fully virtual D(S)MTA cycles. Previous research from AstraZeneca [236] proposed fully virtual cycles where the design is made using REINVENT [245, 99] and the selection done using various strategies that are compared. However here again it was a single-objective optimization, where the compound were labeled as active or inactive. This labeling was not made on a biological value (real or predicted) but on a binary score based on ratios of carbon, nitrogen and oxygen atoms in the molecule. This choice is debatable in terms of its realism compared to real potency and ADMET endpoints. Recent work by Iktos [246] proposed molecular assay simulators to mimic biological endpoints but here again only for single or bi-objective experiments, and the outputs of such oracles showed limited realism when comparing the distributions of values to internal datasets from Sanofi.

Consequently, benchmarking selection strategies under realistic conditions, characterized by a multi-objective context involving authentic drug discovery endpoints, shifts in compound distributions, and the constraint of avoiding costly laboratory experiments, remains a challenge.

3.1.6 Retrospective simulation of legacy projects

To overcome the limitations inherent to both purely virtual and fully experimental benchmarking methodologies, we have explored a hybrid approach that is based on retrospective simulations utilizing historical drug discovery projects. This retrospective approach leverages real-world datasets, accurately replicating each project’s original timeline and workflow within a virtual environment.

During these simulations, an acquisition function iteratively selects subsets of compounds from the historical dataset. Importantly, experimental data are exclusively available for compounds tested in the actual projects; no information exists for compounds designed but never experimentally validated. Each simulation thus involves selecting subsets of the real project’s compounds and comparing the resulting outcomes with actual historical results. The central questions addressed by this method are: Did the chosen acquisition function efficiently exploit the chemical space by identifying and selecting the most promising compounds? Did it adequately explore the project’s chemical space?

To systematically address these questions, we developed the previously described molecular selection toolkit, specifically designed for multi-objective optimization and integrating various selection strategies. Additionally, we established a methodology for conducting retrospective scenario analyses and created analytical tools to qualitatively and quantitatively evaluate the performance and characteristics of these selection strategies during simulations. This framework enables a deeper understanding of how effectively each strategy balances exploration and exploitation within the chemical space of a given project.

3.2 Retrospective simulation methodology

This section details the structured approach we used to construct the retrospective simulations and the datasets used to illustrate them.

3.2.1 Data

3.2.1.1 Datasets

To demonstrate how the prioritization toolkit works, we employed four datasets derived from legacy Sanofi projects [247, 248, 249, 250, 251, 252, 253, 254], targeting Factor Xa (FXa), MMP-8, PPAR δ and Renin. Intellectual property constraints prevent these datasets from including all compounds and experimental readouts from the original projects. The chemical structures and corresponding potency data for the four targets were already publicly available. In this work, we further enriched these datasets by adding experimental timestamps for each compound and extending the FXa dataset with two additional endpoints, subsequently releasing these enhanced versions publicly. Figure 4.1 shows the distributions of molecules overtime for the four datasets.

The first dataset is derived from a project targeting Factor X, the activated form of coagulation Factor X, which plays a pivotal role in blood clot formation within the coagulation cascade. This dataset, spanning just over 6 years, comprises 1015 compounds synthesized and evaluated for po-

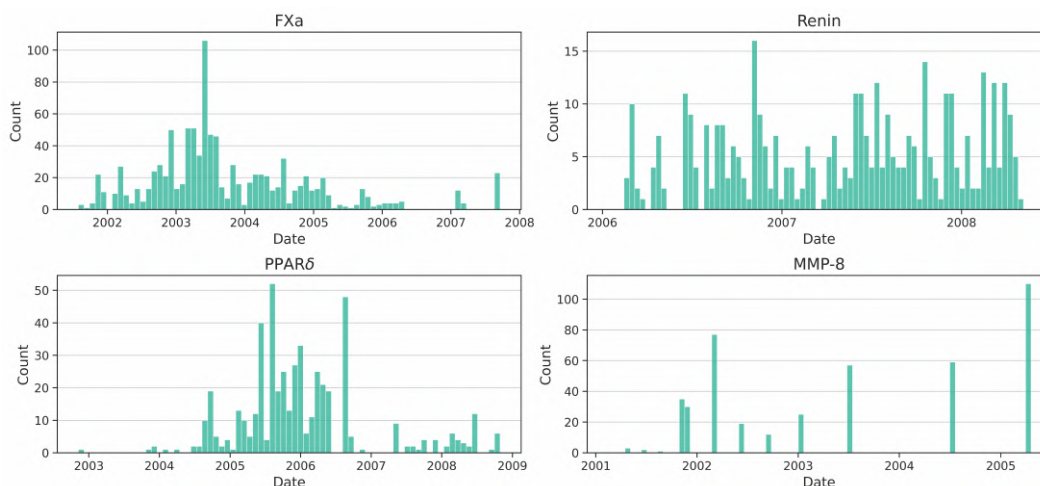


Figure 3.2: Temporal distribution of new molecules in FXa, MMP-8, PPAR δ and Renin datasets.

tency against FXa, as well as for solubility and LogD. It serves as the primary demonstration dataset in the coming sections to showcase the analytical tools.

The second dataset originates from a project focused on Matrix Metalloproteinase-8 (MMP-8), an enzyme involved in degrading extracellular matrix components, implicated in inflammatory disorders, cancer metastasis, and periodontal disease. This dataset spans four years and includes 430 compounds.

The third dataset pertains to a project targeting Peroxisome Proliferator Activated Receptor-delta (PPAR δ), a nuclear receptor crucial for metabolic inflammatory regulation, implicated in metabolic disorders such as diabetes cardiovascular diseases. It covers a period of nearly six years, comprising 498 compounds.

Finally, the fourth dataset comes from a project targeting Renin, an enzyme essential for blood pressure and electrolyte regulation, primarily associated with hypertension and related cardiovascular disorders. This dataset spans over two years, with 388 compounds evaluated.

3.2.1.2 Blueprints

As previously explained, the project team should define a blueprint that outlines the desired characteristics of an optimal drug candidate across a list of endpoints. For each endpoint, the blueprint specifies several key parameters.

- **Optimization Trend:** Specifies whether the property should be maximized (higher is better, denoted as H), minimized (lower is better, denoted as L), or maintained within a specified range (interval trend, denoted as V).

Dataset	Endpoint	Trend	LTV	HTV	Weight
FXa	pKi	H	8.5	9	3
FXa	Solubility (pH 7.40) (μ M)	H	50	100	1
FXa	LogD (pH 7.40)	V	1.5	3.5	1
Renin	pIC ₅₀	H	8	9	1
PPAR δ	pEC ₅₀	H	8	9	1
MMP-8	pIC ₅₀	H	8.5	9	1

Table 3.1: Blueprint tables outlining the parameter space for the FXa, Renin, PPAR δ and MMP-8 datasets. For each endpoint, the required optimization direction is given—H (higher-is-better), L (lower-is-better) or V (value must fall within the stated range)—together with the lower (LTV) and upper (HTV) threshold values. The Weight column indicates the relative importance of each endpoint.

- **Lowest Threshold Value (LTV):** For the higher is better trend, defines the minimum value for a compound to be acceptable. For lower is better, defines the minimum value below which compounds are not considered "more acceptable" (i.e. utility of 1).
- **Highest Threshold Value (HTV):** For the higher is better trend, defines the maximum value after which compounds are not considered "more acceptable" (i.e. utility of 1). For lower is better, defines the minimum value below which compounds are considered acceptable.
- **Weight:** A numerical value indicating the relative importance of the property in comparison to others. Weights can be assigned arbitrarily by the project team or medicinal chemists, or determined using MCDA methods, such as the Analytical Hierarchy Process [255] by defining a priority order among the different endpoints. A weight of 0 can be assigned to endpoints that are monitored but not actively optimized.

Table 3.1 presents the blueprints we used for simulating the four datasets. It should be noted that these blueprints differ from those used in the original projects, both in terms of assigned values and the number of considered endpoints. Additionally, although the simulation results presented here utilized static blueprints, these blueprints can, in practice, evolve throughout the duration of a project. For instance, weights assigned to different endpoints may shift as project priorities change over time, and new properties could be integrated as additional data become available or as new challenges emerge. The framework is designed to accommodate changes to the blueprint during simulation, if required.

The Figure 3.3 illustrates the values distribution of the various endpoints on the four datasets, highlighting unacceptable, acceptable and optimal ranges based on the blueprints parameters.

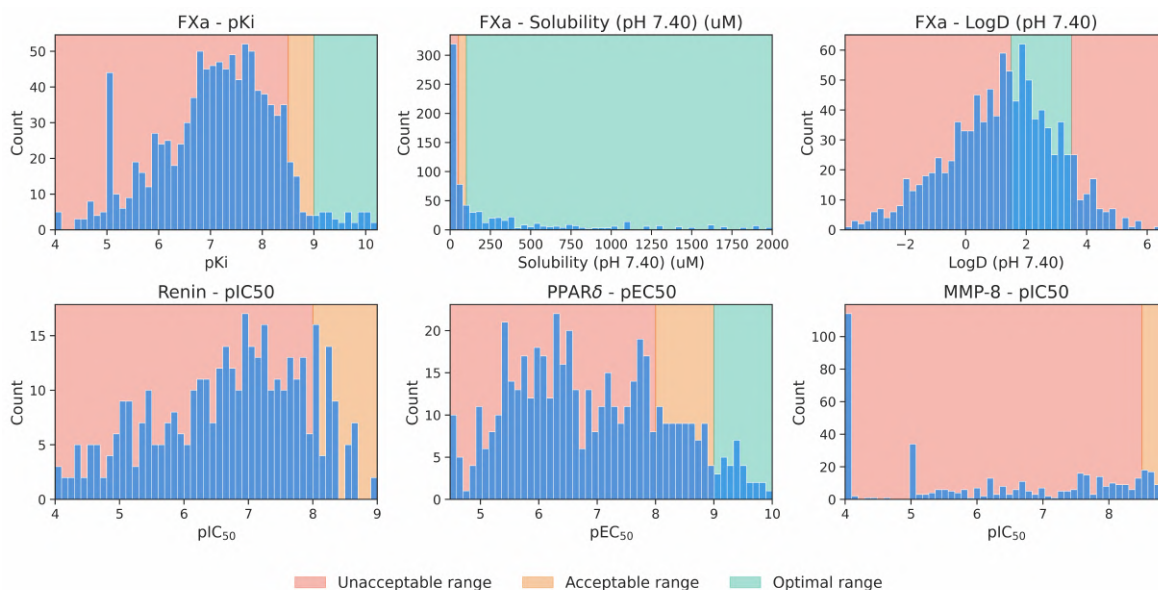


Figure 3.3: Distribution of endpoints assays values for FXa, MMP-8, PPAR δ and Renin datasets.

3.2.2 Setup of a virtual scenario

After establishing the blueprint and prior to initiating the simulation, several parameters must be specified to configure the virtual scenario accurately.

Starting date (t_0) A simulation can begin when two conditions are met:

1. A minimum of 100 compounds are documented for the primary endpoint, ensuring sufficient data for training the initial QSAR model.
2. At least one compound for the primary endpoint lies within the acceptable or optimal range. This criterion ensures that the initial QSAR model recognizes what constitutes a "good" molecule, as QSAR models typically struggle to extrapolate predictions outside their training distributions.

Ending date (t_f) A simulation concludes on the same date as its corresponding real-world project, facilitating a direct and temporally aligned comparison between actual and virtual scenarios.

Timestep (Δt) The timestep Δt corresponds to the typical duration of a DMTA cycle, defined as the interval between successive project meetings dedicated to compound prioritization. During the simulation, each iteration advances the scenario by one timestep. Typical values for Δt range from 1 to 3 months.

Dataset	t_0	t_f	Δt	Batch size	Model
FXa	2002-10-01	2007-09-19	1 month	0.5	Random Forest (std)
Renin	2006-12-01	2008-05-06	1 month	0.5	Random Forest (std)
PPAR δ	2005-07-01	2008-10-29	1 month	0.5	Random Forest (std)
MMP-8	2002-03-01	2005-04-20	5 months	0.5	Random Forest (std)

Table 3.2: Simulation parameters for virtual scenarios on the FXa, Renin, PPAR δ , and MMP-8 datasets. The starting date (t_0), ending date (t_f), timestep duration (Δt), batch size, and the predictive model utilized for predictive modeling are specified for each dataset.

Batch size The batch size refers to the number of compounds selected at each iteration and can be defined either as an absolute integer or as a fractional value. If an integer is specified, exactly that number of compounds is selected at each iteration, provided there is a sufficient number of candidates available in the virtual pool. Alternatively, if a fractional value is given, the number of selected compounds at each iteration is proportional to the number of new compounds synthesized and tested during the corresponding time interval (Δt) in the real project. In this fractional scenario, the number of compounds selected is based on newly synthesized and tested molecules, but the actual selection is conducted on the pool set that includes new molecules but also previously unselected ones.

Predictive model QSAR or QSPR models are retrained at each iteration for endpoints meeting the same two prerequisites required for the starting date: a minimum of 100 data points and at least one compound within the acceptable or optimal range. Further details regarding the available machine learning models and their respective uncertainty estimation methodologies are presented in a subsequent section.

Acquisition function The acquisition function determines compound selection at each iteration and represents the core parameter investigated within the simulations. Acquisition functions may emphasize exploitation, exploration, or balance both. Comprehensive details on these strategies are discussed in a later section. Table 3.2 summarizes the parameters used for the simulations presented in this chapter.

3.2.3 Running a virtual scenario

Once the setup parameters are defined, the simulation can begin. All compounds (and associated readouts) obtained before t_0 are designated as "already selected", forming the initial training set. Using this dataset, the first QSAR model is trained on the main endpoint. At any point in time, additional QSAR/QSPR models are trained on new endpoints if their training set meets the requirements

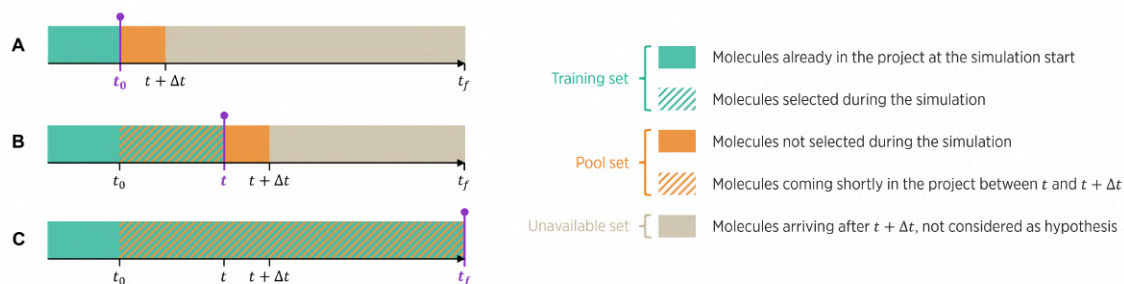


Figure 3.4: Simulation snapshots illustrating the three key stages: (A) the initial configuration at iteration 1, where the pool contains only newly introduced molecules; (B) an intermediate stage midway through the run, with a mix of newly introduced and previously unselected molecules; and (C) the final state at the end of the simulation, in which the pool consists solely of previously unselected molecules.

defined in the previous section about data quantity and distribution of values. Figure 3.4 represents three different stages of a simulation from start to finish.

The initial pool of molecules available for selection consists of those appearing between t_0 and $t_0 + \Delta t$ (Figure 3.4A). Molecules that appear after $t_0 + \Delta t$ are not considered part of the current hypothesis space and are therefore unavailable for selection. From this pool, a batch of molecules is selected based on the specified acquisition function, after which the simulation proceeds to the next iteration.

For subsequent iterations, previously selected compounds are added to the training set and removed from the pool. The training set is also updated with any newly experimental data that appeared for its molecules. The pool is then updated to include compounds appearing between t and $t + \Delta t$, as well as any remaining unselected compounds (Figure 3.4B), leaving apart the molecules appearing after $t + \Delta t$. The updated pool set is evaluated using the acquisition function and a new set of compounds is selected. The training set is then updated and the simulation continues to the following iteration.

Upon reaching the final date, every molecule will have been made available for selection in at least one iteration (Figure 3.4C). Note that a molecule can remain in the pool indefinitely if it is never selected by the acquisition function.

For each simulation, the batch size is set to 0.5, meaning that at each iteration, the number of selected compounds is equal to 50% of the compounds synthesized and tested during the corresponding timestep (Δt) in the real project, but the selection itself is conducted from the entire pool of available molecules, including previously unselected compounds. At the end of a simulation, only a subset of the total compounds from the real-world project is selected. Table 3.3 summarizes various information for each dataset: the total number of compounds, the number of simulation iterations,

Dataset	Dataset size	Iterations	Selected at t_0	Selected at t_f	% dataset sampled
FXa	1015	60	154	588	58%
Renin	388	18	128	258	66%
PPAR δ	498	40	131	318	64%
MMP-8	430	8	148	289	67%

Table 3.3: The table details, for each dataset, the total number of compounds in the original project, the number of iterations of the simulations, the count of compounds included at simulation start (t_0), the cumulative number of compounds selected by simulation completion (t_f), and the proportion of the entire dataset selected within the simulation.

the number of compounds present at the simulation’s start date (t_0), the number selected at the end date (t_f), and the proportion of the entire dataset sampled during a simulation.

3.2.4 Defining what is a good molecule

The effectiveness of a strategy should be assessed based on its ability to prioritize compounds with favorable activity, selectivity, and ADMET properties. For this purpose, we employ the concept of a desirability score [256], which assigns a unique value to each molecule by integrating its experimental data across documented endpoints. A high desirability score indicates that a compound meets the blueprint’s criteria and vice versa. The computation of the desirability score is a multi-step process.

First, for each documented endpoint of a molecule, experimental assays values are normalized between 0.05 and 1 to define a utility value, thanks to a utility function d inspired by Cummins and Bell [257]. Utility functions, illustrated in Figure 3.5, are smooth sigmoidal functions related to the trend of optimization and parametrized by the lowest threshold value (LTV) and the highest threshold value (HTV), along with shift (δ) and steepness (η) parameters, which default to 0.05 and 10, respectively. When passed into a utility function, an experimental assay value x from endpoint i is normalized according to whether it should be minimized (Eq. 3.1, Figure 3.5A), maximized (Eq. 3.2, Figure 3.5B), or kept within a designated interval (Eq. 3.3, Figure 3.5C).

$$d_i^{\text{lower-is-better}}(x) = \frac{1 - \delta}{1 + e^{\eta \frac{x - LTV_i}{HTV_i - LTV_i}}} + \delta \quad (3.1)$$

$$d_i^{\text{higher-is-better}}(x) = \frac{1 - \delta}{1 + e^{\eta \frac{x - HTV_i}{LTV_i - HTV_i}}} + \delta \quad (3.2)$$

$$d_i^{\text{interval}}(x) = \begin{cases} d_i^{\text{lower-is-better}}(x), & x < \frac{LTV_i + HTV_i}{2} \\ d_i^{\text{higher-is-better}}(x), & x \geq \frac{LTV_i + HTV_i}{2} \end{cases} \quad (3.3)$$

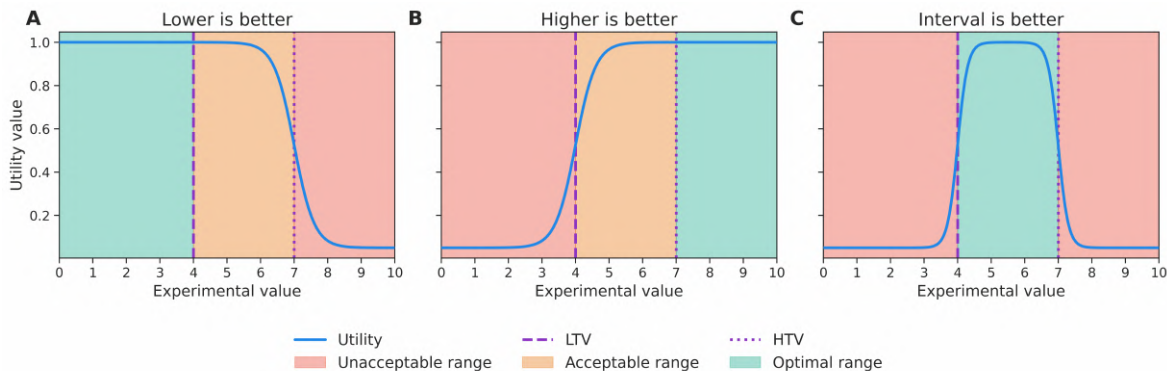


Figure 3.5: Utility-function for the three optimization directions—(A) *lower-is-better*, (B) *interval*, and (C) *higher-is-better*—all plotted with a lower threshold value of 4 and a higher threshold value of 7. Molecules belong to the acceptable/optimal regions when the utility score is ≥ 0.525 , the midpoint of the sigmoidal curve.

Next, after computing the utility for each documented endpoint, the values are aggregated into a final desirability score (D_{score}) using a weighted geometric mean (Eq. 3.4), with weights ω_i defined in the blueprint. Previously setting the shift to 0.05 instead of 0 ensures that a poor property does not reduce the final score to zero.

$$D_{score}(x) = \left(\prod_{i=1}^n d_i(x)^{w_i} \right)^{\frac{1}{\sum_{i=1}^n \omega_i}} \quad (3.4)$$

Now that we’ve assigned a score to each of the molecules, the most obvious option for defining what a good molecule is is to choose those with the highest scores. Although tempting, this approach needs to be questioned slightly. Indeed, as mentioned above, the desirability score is calculated by taking into account the experimental values for the different endpoints available. If an endpoint is not available for a molecule, it is not taken into account in the calculation of the desirability score. Let’s imagine a project where 25 endpoints are optimized, and a molecule is documented on only one of these endpoints. Assuming that this single value is within the optimal range of the endpoint according to the blueprint, this would mean that the molecule has a desirability score of 1. However, is it reasonable to say that this molecule is "good" when we have no idea of all its experimental values on all the other endpoints?

To address this concern, we introduce the intuitive notion of documentation. The documentation of a molecule corresponds to the proportion the project’s endpoints that are documented for this molecule. Taking the previous example with 25 endpoints, a molecule with all 25 endpoints documented would have a documentation of 1, a molecule with 15 documented endpoints a documentation of 0.6 etc... Using the desirability score and the documentation, we define multiple top categories of molecules that are summarized in Table 3.4. These categories can depend on fixed

Top category	Description
Fixed ($t = 0.5$)	Molecules with desirability scores above 0.5 and documentation above 0.65.
Top Dscore ($X\%$)	Molecules within the top $X\%$ of desirability scores with $X \in [10, 5, 3, 1, 0.5]$, and having documentation above 0.65.
Top Dscore ($X\sigma$)	Molecules exceeding the mean desirability score by at least X standard deviation with $X \in [1, 2, 3]$, and with documentation above 0.65.
Top DocDScore ($X\%$)	Molecules that are within the top $X\%$ of DocDScores with $X \in [10, 5, 3, 1, 0.5]$.

Table 3.4: Summary of the different top categories of molecules with their respective definition, based on desirability and documentation.

thresholds for both desirability and documentation (Fixed) or distribution-based threshold for desirability and fixed for documentation (Top DScore). The documentation threshold for a molecule to be "documented enough" is arbitrarily fixed to 0.65. However, a third class of top category, relies on a hybrid score between the desirability score and documentation called DocDScore (Eq. 3.5). This hybrid category can be less restrictive on compounds that have very high desirability but medium documentation.

$$DocD_{score}(x) = D_{score}(x) \times Documentation(x) \quad (3.5)$$

The different top categories are illustrated in Figure 3.6, using an undisclosed dataset comprising 12 endpoints.

Considering documentation is particularly valuable for datasets that include multiple endpoints with uneven levels of documentation, as will be discussed in the subsequent chapter. However, in this chapter, we focus exclusively on smaller datasets that are fully documented across all endpoints.

3.3 Acquisition functions

The retrospective analysis pipeline integrates a comprehensive set of acquisition functions, leveraging methodologies from active learning, multi-criteria decision analysis, and medicinal chemistry practices. These acquisition functions can be systematically categorized according to several criteria, as summarized in Table 3.5.

First, the acquisition functions are differentiated based on their reliance on predictive modeling (QSAR/QSPR). Certain strategies, including Desirability, Desired Spread, and Uncertainty, explicitly utilize predicted molecular properties or prediction uncertainties to guide molecule selection.

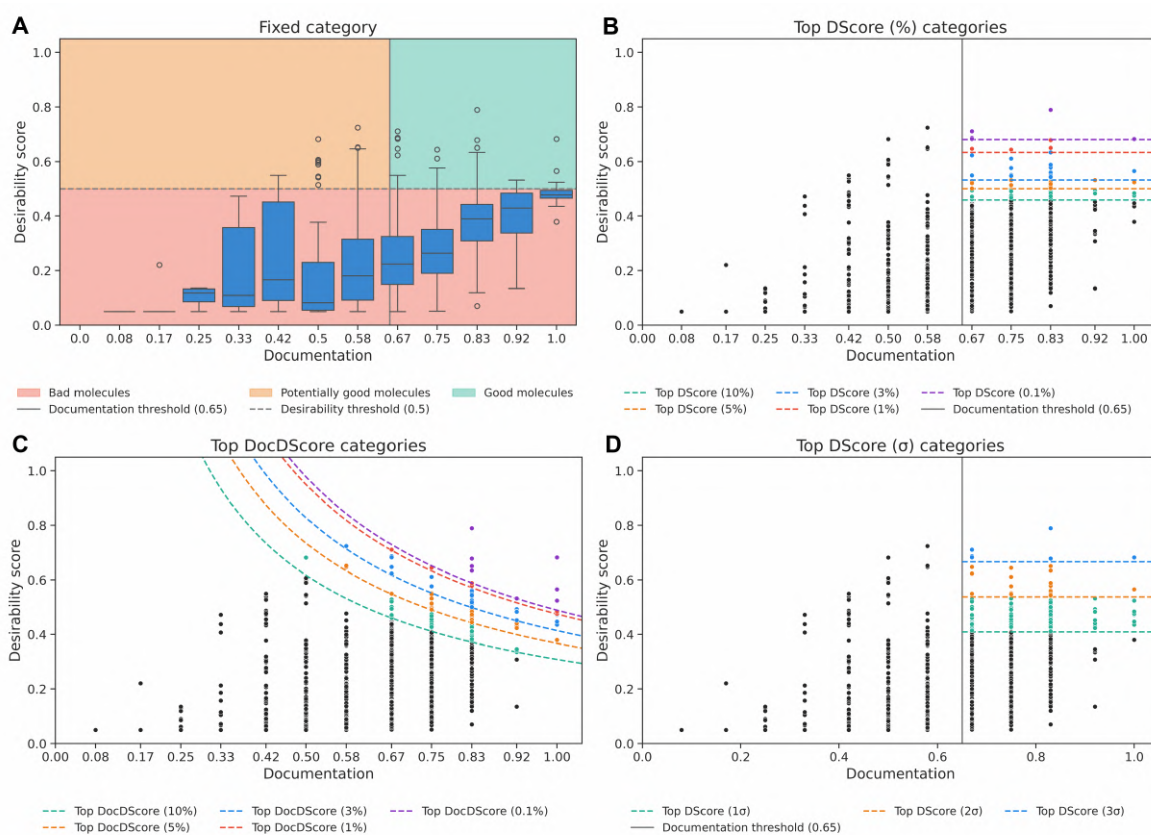


Figure 3.6: Plots illustrating desirability versus documentation for an undisclosed dataset with 12 endpoints. Panel **A** represents the Fixed category, panel **B** shows the Top Dscore (%) categories, panel **C** displays the Top DocDScore categories, and panel **D** depicts the Top Dscore (σ) categories.

Conversely, methods like Coverage and K-means rely solely on structural descriptors, thus circumventing the need for predictive models.

Second, these functions can be distinguished by their primary operational objectives: exploitation or exploration. Exploitation-focused strategies prioritize the identification of promising molecular candidates based on existing data, while exploration-focused approaches aim to systematically sample compounds throughout the chemical space to uncover novel structural regions. Additionally, there exist trade-off strategies that aim to balance exploitation and exploration objectives to varying degrees.

Finally, the acquisition functions can be classified according to their selection methodology. Independent strategies select compounds individually based solely on their inherent characteristics, whereas batch strategies select compounds collectively, considering interdependencies between candidates. Importantly, certain batch strategies can be effectively approximated through sequential selection processes, facilitating computational efficiency [258].

3.3.1 Model-agnostic acquisition functions

Random The random acquisition function selects molecules from the pool set randomly, disregarding uncertainty, diversity or any predicted endpoints values. While typically viewed as a simplistic or "dummy" strategy in active learning, random sampling offers practical advantages for retrospective analyses. Because our pool consists exclusively of molecules previously synthesized and tested in actual projects, and given that our simulated batch sizes are smaller than those employed originally, random sampling repeatedly selects distinct subsets from the historical pool. Consequently, when executed numerous times, this approach approximates the original selection process, adjusted to align with the current simulation's parameters. In our study, the Random strategy thus serves as an essential baseline, approximating the real project selection at comparable batch size.

K-Means Clustering K-means clustering [144] promotes diversity by grouping compounds based on their molecular feature representations. In our method, building on the findings from Chapter 2, the feature space combines Extended Connectivity Fingerprints (ECFP, 2048 bits, radius of 2, incorporating chirality and count) and MACCS keys, computed via the scikit-fingerprints library [264].

K-Medoids Clustering K-medoids clustering [145] similarly promotes diversity by clustering compounds within molecular feature space but selects each cluster's medoid rather than its centroid. Although computationally more demanding than k-means, this method is more robust against outliers by using actual data points rather than mean positions, thus reducing sensitivity to noise and

Acquisition function	Objective	Selection	Model	Reference	Description
Coverage score	Exploration	Batch	No	[259]	Select batch of compounds that has the highest coverage score, based on bayesian statistics and information entropy.
Desirability	Exploitation	Independent	Yes	[257]	Select compounds with the highest predicted desirabilities.
Desirability with AD	Exploitation	Independent	Yes	[260]	Select compounds with the highest predicted desirability within a distance-based applicability domain.
Desired Coverage	Trade-off	Batch	Yes	This work	Select compounds with the highest composite score combining desirability score and coverage score.
Desired Spread	Trade-off	Batch	Yes	This work	Select compounds with the highest composite score combining desirability score and spread score.
Dissimilarity-to-known	Exploration	Independent	No	-	Select most dissimilar compounds to the ones previously selected.
Dissimilarity-to-known-good	Exploration	Independent	No	-	Select most dissimilar compounds to the good ones previously selected.
GRA	Exploration	Independent	Yes	Adapted from [261]	Select compounds that are the closest predicted values to an ideal solution based on relational coefficients between alternatives. Modified from its original methodology to use a utility function-based normalization.
Greedydiverse	Trade-off	Batch	Yes	[177]	Select compounds that optimize the predicted desirability scores within a batch, with a penalty if the molecules from the batch are too similar.
K-means	Exploration	Batch	No	[144]	Select the compound that is the closest to each cluster's centroid.
K-medoids	Exploration	Batch	No	[145]	Select the compound that is the medoid for each cluster.
Random	Exploration	Independent	No	-	Select random compounds. Proxy to the real project selection at comparable batch size.
Similarity-to-known	Exploration	Independent	No	-	Select most similar compounds to the ones previously selected.
Similarity-to-known-good	Exploration	Independent	No	-	Select most similar compounds to the good ones previously selected.
Spread	Exploration	Batch	No	[262]	Select diverse compounds by iteratively selecting the ones that maximize the minimum distance to their closest neighbor among previously selected compounds.
TOPSIS	Exploitation	Independent	Yes	Adapted from [263]	Select compounds that are the closest predicted values to an ideal solution based on euclidean distance.
Uncertainty	Exploration	Independent	Yes	This work	Select compounds that have the highest predicted uncertainty score aggregating uncertainty across the different endpoints.
U/D Harmonic	Trade-off	Independent	Yes	This work	Select compounds that have the highest harmonic score, combining predicted uncertainty and desirability scores.
U/D Ratio	Trade-off	Independent	Yes	This work	Select compounds that have the highest ratio score, combining predicted uncertainty and desirability scores.

Table 3.5: Summary of acquisition functions categorized by objective, selection type and reliance on predictive modeling.

extreme values. The default distance metric is Euclidean, but other pairwise distance metrics may also be applied.

Coverage Score The coverage score [259] is a subset-based selection technique leveraging Bayesian statistics and information entropy to balance representativeness and diversity.

First, we define the compound sets involved in calculating the coverage score:

- S_{pool} : Set from which compounds are selected.
- $S_{selected}$: Set of compounds selected in the current iteration, $S_{selected} \subseteq S_{pool}$.
- S_{train} : Training set containing compounds selected in previous iterations.
- S_{full} : Full set defined as $S_{pool} \cup S_{train}$.
- $S_{sampled}$: Sampled set defined as $S_{selected} \cup S_{train}$.

Each compound is characterized by a molecular fingerprint (concatenation of ECFP4 and MACCS keys) comprising k features. The coverage score for each fingerprint feature is determined based on its frequency within the sampled and full sets. Bayesian statistics are applied to estimate the probability of a fingerprint feature f_k occurring in the sampled set. This probability, smoothed using Laplacian smoothing, is defined as:

$$P_{smooth}(\text{sampled} \mid f_k) = \frac{F_{f_k, \text{sampled}} + \alpha}{F_{f_k, \text{full}} + \frac{\alpha}{P(\text{sampled})}} \quad (3.6)$$

where α is an additive smoothing parameter set to one, and $F_{f_k, \text{full}}$ and $F_{f_k, \text{sampled}}$ denote the frequencies of fingerprint feature f_k in the full and sampled sets, respectively.

The term $P(\text{sampled})$ represents the probability of a compound belonging to the sampled set and is calculated as:

$$P(\text{sampled}) = \frac{|S_{sampled}|}{|S_{full}|} = \frac{N_{sampled}}{N_{full}} \quad (3.7)$$

where N_{full} and $N_{sampled}$ are the cardinalities of the full and sampled sets, respectively.

The base coverage score for a fingerprint feature, $C_{base}(f_k)$, is defined as the negative log-likelihood ratio:

$$C_{base}(f_k) = -\ln \left(\frac{P_{smooth}(\text{sampled} \mid f_k)}{P_{sampled}} \right) \quad (3.8)$$

To assess the informativeness of each feature within the sampled set, a Shannon entropy term

$H(f_k \mid \text{sampled})$ is incorporated:

$$H(f_k \mid \text{sampled}) = -\frac{p_1 \ln(p_1) + p_2 \ln(p_2)}{\ln(2)} \quad (3.9)$$

where $p_1 = P(f_k \mid \text{sampled}) = \frac{F_{f_k, \text{sampled}}}{N_{\text{sampled}}}$ and $p_2 = 1 - p_1$. The division by $\ln(2)$ normalizes entropy to a $[0, 1]$ scale, with entropy explicitly set to 0 when $F_{f_k, \text{sampled}} = N_{\text{sampled}}$.

The final coverage score for each fingerprint feature, $C_{\text{final}}(f_k)$, is computed as:

$$C_{\text{final}}(x_i) = \begin{cases} C_{\text{base}}(f_k)(2 - H(f_k \mid \text{sampled})) & \text{if } C_{\text{base}}(f_k) < 0 \\ & \cap \frac{F_{f_k, \text{sampled}}}{N_{\text{sampled}}} > \frac{1}{2} \\ C_{\text{base}}(f_k)H(f_k \mid \text{sampled}) & \text{otherwise} \end{cases} \quad (3.10)$$

The compound-level coverage score, $C_{\text{comp}}(x)$, is calculated as the sum of coverage scores of the features present in the compound’s fingerprint:

$$C_{\text{comp}}(x) = \sum_{f_k} C_{\text{final}}(f_k) \quad (3.11)$$

Finally, the overall coverage score of the selected set S_{selected} is obtained by summing the coverage scores of its constituent compounds:

$$C_{\text{score}}(S_{\text{selected}}) = \sum_{x \in S} C_{\text{comp}}(x) \quad (3.12)$$

The set of molecules selected is the one achieving the highest coverage score. The original paper utilizes the NSGA-II genetic algorithm [265] to identify this set. However, internal evaluations demonstrated that a greedy optimization strategy consistently yields comparable or better performance while offering superior computational efficiency and reproducibility. Consequently, we employ a greedy optimization approach in our implementation.

Spread Spread sampling [262] is a sequential, greedy, space-filling approach designed to maximize diversity in the selected subset. At each iteration, the method selects the molecule that maximizes the minimum distance from its nearest neighbors within both the training set and previously selected molecules (noted S_{score}). Distances are calculated using the Tanimoto similarity metric with binary ECFP fingerprints (2048 bits, radius of 2). By emphasizing diversity, spread sampling ensures a representative coverage of the molecular feature space.

Similar-to-known The similar-to-known acquisition function prioritizes pool-set compounds that are highly similar to molecules already included in the training set, determined via Tanimoto similarity. Specifically, it selects batches exhibiting the highest average similarity to existing training compounds. However, this method does not consider diversity within the selected batch.

Dissimilar-to-known Conversely, the dissimilar-to-known acquisition function emphasizes selecting compounds least similar to the training set, again employing Tanimoto similarity. It selects batches with the lowest average similarity to known compounds, also neglecting internal batch diversity.

Similar-to-good-known The similar-to-good-known acquisition function specifically selects pool-set compounds closely resembling "good" compounds in the training set, where "good" is defined by an experimental desirability score greater than 0.5. Similarity is assessed using Tanimoto similarity. If the training set lacks "good" compounds, this strategy defaults to the similar-to-known approach previously described.

Dissimilar-to-good-known Similarly, the dissimilar-to-good-known acquisition function targets pool-set compounds least similar to the "good" compounds from the training set, with the same desirability threshold (>0.5). Tanimoto similarity again measures similarity. If the training set does not include any "good" compounds, the strategy defaults to the dissimilar-to-known approach described earlier.

3.3.2 Predictive models

As previously discussed, several strategies within our framework rely on predictive modeling. The toolkit integrates seven distinct machine learning models, each employing unique methodologies for generating predictions and estimating their associated uncertainties. These models are summarized in Table 3.6. All models utilize molecular fingerprints as inputs, specifically the concatenation of ECFP fingerprints (2048 bits, radius of 2, incorporating chirality and counts) and MACCS keys by default.

3.3.3 Model-dependent acquisition functions

Desirability The desirability acquisition function relies on the QSAR/QSPR models trained at each iteration. Each molecule in the pool set receives predicted endpoint values, from which a predicted desirability score D_{score} is computed. Normalization and aggregation of predicted values

Predictive Model	Description
Random Forest (std)	Prediction: Mean prediction from all decision trees. Uncertainty: Standard deviation across all decision trees.
Random Forest (bootstrap)	Prediction: Mean prediction from multiple bootstrapped models. Uncertainty: Standard deviation across these bootstrapped models. By default, 50 bootstrapped models are used.
LightGBM (bootstrap)	Prediction: Mean prediction from bootstrapped LightGBM models. Uncertainty: Standard deviation across these bootstrapped models. By default, 50 bootstrapped models are used.
LightGBM (quantile)	Prediction: Median (50th percentile) predicted via quantile regression. Uncertainty: Range defined by the 2.5th and 97.5th percentile predictions from quantile regression.
XGBoost (bootstrap)	Prediction: Mean prediction from bootstrapped XGBoost models. Uncertainty: Standard deviation across these bootstrapped models. By default, 50 bootstrapped models are used.
XGBoost (quantile)	Prediction: Median (50th percentile) predicted via quantile regression. Uncertainty: Range defined by the 2.5th and 97.5th percentile predictions from quantile regression.
Gaussian Process	Prediction: Posterior mean of the Gaussian Process model. Uncertainty: Posterior standard deviation provided by the Gaussian Process model.

Table 3.6: Overview of predictive models implemented in the framework, detailing their respective approaches to prediction and uncertainty estimation.

mirror the methodology applied to experimental desirability scoring (3.2.4), using the utility functions and a weighted geometric mean. Molecules with the highest predicted desirability scores are subsequently selected.

Desirability (with applicability domain) This variant extends the basic desirability approach by introducing an applicability domain (AD) filter, defined by molecular similarity. Molecules are considered within the AD if their maximum similarity to any training set molecule exceeds a specified threshold (default: 0.7). Prioritization occurs first among molecules within the AD based on predicted desirability scores. If fewer molecules fall within the domain than required for batch selection, remaining slots are filled by out-of-domain molecules according to their predicted desirability scores.

Uncertainty The uncertainty acquisition function is widely employed in active learning to select samples with the highest predictive uncertainty. In multi-objective optimization, however, combining uncertainties across several endpoints can be challenging. To address this, we compute an uncertainty score U_{score} in a manner analogous to the desirability score. First, predicted uncertainties for all modeled endpoints i are obtained. These uncertainties are then normalized using MinMax scaling with a lower threshold of 0.05 (Eq. 3.13). Finally, the normalized uncertainties are combined with a weighted geometric mean (Eq. 3.14) to yield a final uncertainty score for each compound.

$$u_i(x) = \max\left(\frac{x_i - \min(X_i)}{\max(X_i) - \min(X_i)}, 0.05\right) \quad (3.13)$$

where x_i is the predicted uncertainty on endpoint i for a compound x , X_i is the set of all predicted uncertainties on endpoint i for all molecules.

$$U_{score}(x) = \left(\prod_{i=1}^n u_i(x)^{w_i}\right)^{\frac{1}{\sum_{i=1}^n w_i}} \quad (3.14)$$

where n is the number of endpoints and w_i the weight associated to endpoint i .

U/D Harmonic The U/D Harmonic acquisition function combines the predicted desirability and uncertainty scores into a harmonic score (H_{score}) using the harmonic mean (Eq. 3.15). This approach balances exploitation and exploration by highlighting compounds that are both promising and uncertain.

$$H_{score}(x) = \frac{2 \times U_{score}(x) \times D_{score}(x)}{U_{score}(x) + D_{score}(x)} \quad (3.15)$$

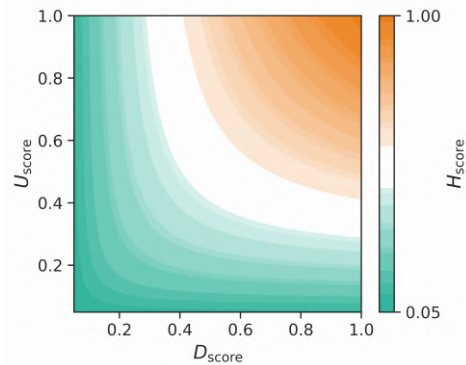


Figure 3.7: U/D Harmonic score distribution based on U_{score} and D_{score} .

where $U_{score}(x)$ and $D_{score}(x)$ are respectively the predicted uncertainty and predicted desirability scores of a molecule x .

U/D Ratio The U/D Ratio acquisition function combines predicted desirability and prediction uncertainty scores into a ratio score (R_{score}) using an adjusted ratio (Eq. 3.16). Since the desirability lowest possible value is 0.05, the adjustment constant ϵ is typically set to 0.95 to have a maximum R_{score} value of 1. The ratio score balances exploitation and exploration by prioritizing uncertain molecules that are more or less desirable, depending on the ϵ value. Higher values of epsilon will promote highly desirable molecules.

$$R_{score}(x) = \frac{U_{score}(x)}{\epsilon + D_{score}(x)} \quad (3.16)$$

where $U_{score}(x)$ and $D_{score}(x)$ are respectively the predicted uncertainty and predicted desirability scores of a molecule x .

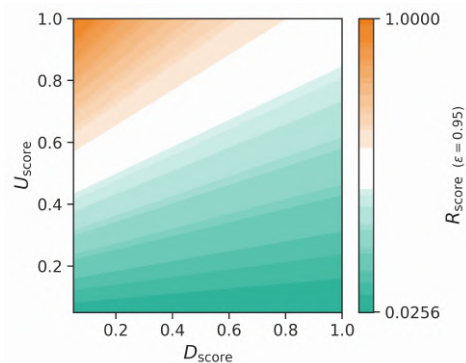


Figure 3.8: U/D Ratio scores distribution based on U_{score} and D_{score} for $\epsilon = 0.95$.

Desired Coverage The desired coverage acquisition proposes a trade-off between exploration and exploitation by combining the predicted desirability score and the coverage score to create a desired coverage score (DC_{score}). Using a greedy approach, it optimizes a composite score for a subset of molecules that is its coverage score multiplied by its mean predicted desirability score (Eq. 3.17). This approach favors batches with high desirability and high coverage.

$$DC_{score}(S) = \sum_{x \in S} C_{comp}(x) \times \frac{1}{|S|} \sum_{x \in S} D_{score}(x) = C_{score}(S) \times D_{score}(S) \quad (3.17)$$

where S is the set of selected compounds, $C_{comp}(x)$ and $D_{score}(x)$ are respectively the compound-level coverage score and desirability score of compound $x \in S$. $C_{score}(S)$ is the coverage score of set S and $D_{score}(S)$ is the average predicted desirability score of compounds in S .

Desired Spread The desired spread acquisition function proposes a trade-off between exploration and exploitation by combining the predicted desirability score and the spread strategy to create a desired spread score (Eq. 3.18). It is inspired by the demerit spread design strategy developed by Higgs et al. [266] using the desirability score as a scaling factor to the distance between the pool set and its nearest neighbors from the train set or previously selected compounds.

$$DS_{score}(x) = D_{score}(x) \times S_{score}(x) \quad (3.18)$$

where $D_{score}(x)$ and $S_{score}(x)$ are respectively the desirability score and spread score of compound x .

Greediverse The greediverse acquisition function is based on the work of Langevin et al. [177]. It optimizes the desirability scores within a batch, while yielding a penalty if the similarity of two molecules from the batch is above a defined threshold. The function is optimized using a greedy iterative approach as mentioned in the original paper.

TOPSIS The Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) [267] is a widely used method in multiple-criteria decision analysis (MCDA). It ranks alternatives across multiple, often conflicting criteria by identifying solutions that are closest to an ideal option while remaining farthest from the least desirable one.

In its standard formulation, TOPSIS assumes that all criteria either increase or decrease monotonically, which justifies the common use of L2 normalization. However, in drug discovery, certain endpoints require maximization, others minimization, and some must fall within a specific range.

This variability renders L2 normalization inappropriate. To address this limitation, we propose a modified TOPSIS approach that replaces L2 normalization with the utility functions previously described. These functions rescale endpoint values to a range between 0.05 and 1, ensuring a more suitable representation of diverse optimization objectives.

The first step involves normalizing the predicted values for each endpoint i across all molecules j within the pool set using utility functions described in Equations 3.1, 3.2 and 3.3. A normalized decision matrix $m_{j,i}$ is constructed with these normalized values.

This matrix is then transformed into a weighted version, denoted as $m_{j,i}^*$:

$$m_{j,i}^* = m_{j,i} \times \frac{\omega_i}{\sum_{i=1}^n \omega_i} \quad (3.19)$$

where ω_i is the weight associated with endpoint i and n is the total number of endpoints.

Next, the positive ideal solution p_i^{ideal} and negative ideal solution n_i^{ideal} are determined for each endpoint by identifying the maximum and minimum values in the weighted matrix, respectively: :

$$p_i^{ideal} = \max_j (m_{j,i}^*), \quad n_i^{ideal} = \min_j (m_{j,i}^*) \quad (3.20)$$

The Euclidean distance of each molecule to these ideal solutions are then computed as follows:

$$dist_j^+ = \sqrt{\sum_{i=1}^n (m_{j,i}^* - p_i^{ideal})^2} \quad dist_j^- = \sqrt{\sum_{i=1}^n (m_{j,i}^* - n_i^{ideal})^2} \quad (3.21)$$

where n is the number of endpoints.

Finally, the relative closeness of a molecule j to the ideal solution is calculated:

$$c_j = \frac{dist_j^-}{dist_j^+ + dist_j^-} \quad (3.22)$$

For selection, compounds having the highest relative closeness are prioritized.

GRA Grey Relational Analysis (GRA) [261] is a prominent Multi-Criteria Decision Analysis (MCDA) method developed within the framework of Grey System Theory. It is particularly effective in handling uncertain, incomplete, or imprecise information, hence the term "grey". Unlike conventional methods that rely on fully defined datasets, GRA can accommodate varying data quality, making it especially useful in real-world decision-making scenarios. Similar to the TOPSIS algorithm, GRA

evaluates alternatives based on their similarity to an ideal solution.

Originally, GRA employs Min-Max normalization for attributes that should be maximized and Max-Min normalization for those that should be minimized. However, like TOPSIS, it does not inherently account for criteria that should be maintained within a specific range. To address this limitation, we define a normalized decision matrix $m_{j,i}$ similarly to the TOPSIS method, using utility functions.

Next, the deviation sequence $\Delta_{j,i}$ is computed:

$$\Delta_{j,i} = 1 - m_{j,i} \quad (3.23)$$

Then, the grey relational coefficient $\varphi_{j,i}$ is calculated for each data point using an ϵ parameter, typically set to 0.5:

$$\varphi_{j,i} = \frac{\epsilon}{\Delta_{j,i} + \epsilon} \quad (3.24)$$

where $\Delta_{j,i}$ is the deviation sequence of molecule j on endpoint i .

Finally, the grey relational coefficients for different endpoints are aggregated into a grey relational grade using a weighted arithmetic mean:

$$GRG_j = \frac{1}{n} \sum_{i=1}^n \varphi_{j,i} \cdot \omega_i \quad (3.25)$$

where n is the number of endpoints, ω_i the weight associated to endpoint i and $\varphi_{j,i}$ the grey relational coefficient for molecule j on endpoint i .

Compounds with the highest grey relational grades are prioritized for selection.

3.3.4 Oracle acquisition function

The oracle acquisition function operates as an omniscient strategy, having access to the complete set of experimental data at each stage of the selection process. Consequently, it identifies and selects the compounds with the highest experimental desirability scores from the pool set at every iteration. This strategy leads to optimal exploitation to which other strategies can be compared.

3.4 Retrospective analysis tools

To evaluate the performance of different acquisition functions, we need to consider how well they balance exploitation and exploration of the chemical space within a given project. For this purpose, we developed several qualitative and quantitative analytical tools, described in detail in the following

subsections. These tools are integrated into a user-friendly application to facilitate their practical use.

3.4.1 Visualizing simulations

Visualizing the chemical space through iterative cycles enables qualitative assessment of various acquisition strategies. Techniques such as PCA [147], t-SNE [148], UMAP [149], and TMAP [150], introduced in the first chapter, are commonly employed for such purposes. In this work, we chose TMAP due to its ability to effectively preserve both global and local structural similarity through its tree-like graphical representation.

Extending the capabilities of TMAP, we developed Kernel Density Estimate TMAP (KDE TMAP). This approach applies a Gaussian kernel to two-dimensional embeddings, converting the discrete chemical space into a continuous density representation. This transformation accentuates areas with varying molecular densities, thus enhancing interpretability.

Given a set of data points $x_1, x_2, \dots, x_n$, the kernel density estimate at a point x is defined as:

$$kde(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) \quad (3.26)$$

where n is the number of data points, h is the bandwidth (smoothing parameter), and $K(\cdot)$ is typically a smooth, symmetric kernel function. For this work, we set the bandwidth to 0.1 to achieve a fine-grained density estimation that closely aligns with the underlying structure of the 2D graph, utilizing the Gaussian kernel defined by:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right). \quad (3.27)$$

To illustrate how virtual simulations populate the chemical space, we first compute the TMAP for all molecules in the project (Figure 3.9A), subsequently transforming it into a KDE TMAP (Figure 3.9B). The resulting KDE TMAP envelope (Figure 3.9C) serves as a reference map of the chemical space, progressively populated throughout simulation iterations.

Figure 3.10 demonstrates how different molecule sets are projected onto the KDE TMAP envelope across simulation iterations. This visualization provides intuitive insights into how effectively each acquisition function manages the exploration of novel chemical regions and exploitation of areas with high-value compounds. At the end of the simulation, the visualization depicts the chemical space sampled by the strategy. Regions populated by the final training set illustrate areas explored by the strategy, whereas regions containing only pool set molecules represent chemical spaces that

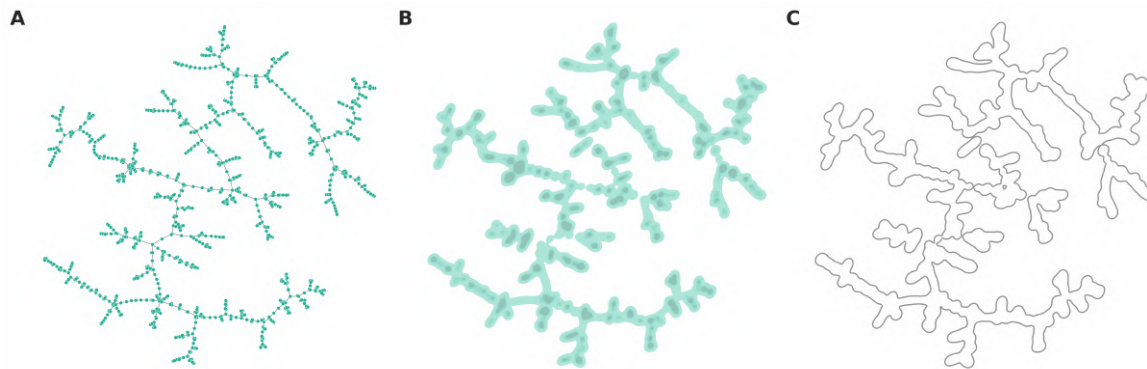


Figure 3.9: Visualization of the FXa chemical space using three complementary TMAP representations: (A) discrete TMAP embedding, (B) embedding colored according to kernel-density estimation (KDE) of point density, and (C) resulting density envelope defining the occupied chemical space.

remain unvisited. The distribution of markers additionally highlights the acquisition function’s ability to select the best molecules.

3.4.2 Measuring exploitation

While visual assessments of how acquisition functions navigate chemical space offer valuable insights, systematically comparing multiple strategies at scale needs quantitative measures. To address this need, we introduce the Proportion of Top Molecules Retrieved (PTMR), a straightforward metric designed to quantify the effectiveness of a strategy in retrieving molecules belonging to a specified top-tier category.

Formally, let S denote the set of molecules selected up to a given iteration, and T represent the set of molecules classified within a defined top-tier category. The PTMR metric is then calculated as follows:

$$\text{PTMR}(S, T) = \frac{|T \cap S|}{|T|} \quad (3.28)$$

This metric, inherently monotonic with regards to the number of iterations, can be evaluated iteratively throughout simulations to quantify how effectively an acquisition function retrieves high-value molecules. A PTMR value of 1 signifies that the strategy has successfully identified all molecules from the defined top-tier category, whereas a PTMR value of 0 indicates that no top molecules have been retrieved.

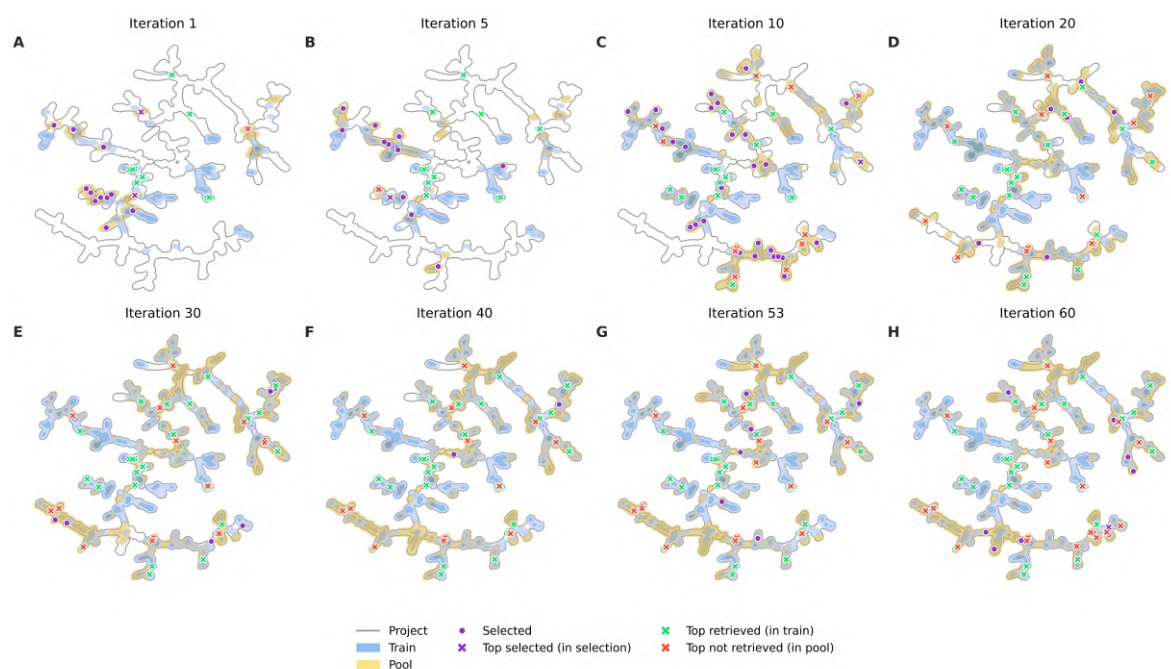


Figure 3.10: Visualization of the FXa chemical space using KDE TMAP across iterations in a representative simulation with the Random acquisition function. Snapshots are displayed at multiple iterations, from initial (iteration 1) to final (iteration 60). Compounds selected in the current iteration are depicted in purple, with crosses representing top-tier compounds (Top DScore 5%) and dots representing all others. Green crosses indicate previously selected top-tier compounds, while red crosses mark top-tier compounds not selected ("missed"). Over iterations, selected molecules form the training set (blue), whereas unselected molecules constitute the pool set (yellow).

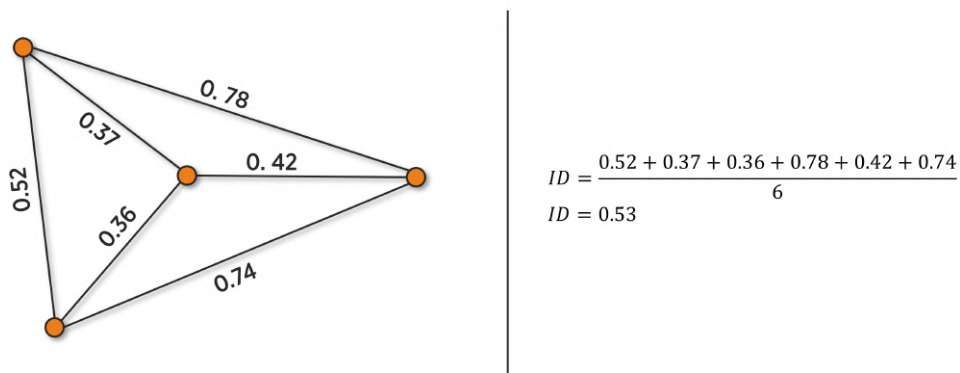


Figure 3.11: Illustration of the internal diversity metric. Left: Tanimoto similarity of each pair of molecules, represented as dots. Right: expression of the Internal Diversity of the set.

3.4.3 Measuring exploration

Similarly to exploitation, quantifying exploration is crucial for comparing a large number of strategies. Many exploration metrics related to coverage, diversity, or novelty of a molecular set have been proposed in the literature. Our analysis framework integrates several of them, which are detailed in the following sections.

3.4.3.1 Internal Diversity

Internal Diversity is a widely adopted metric commonly employed in chemical library design and for evaluating the chemical diversity of outputs from AI-driven molecular generative models [268, 269, 270]. Defined as the average Tanimoto distance among all pairs of compounds within a set S , it is illustrated on Figure 3.11 and formally expressed as follows:

$$\text{InternalDiversity}(S) = \frac{2}{n(n-1)} \sum_{i,j \in S, i \neq j} \delta_{i,j} \quad (3.29)$$

where n is the number of compounds within set S and $\delta_{i,j}$ the Tanimoto similarity between two molecules $i \in S$ and $j \in S$.

3.4.3.2 Neighborhood Coverage

Internal Diversity is a reference-free method for analyzing the chemical diversity within a single set S . It is also useful to use coverage metrics for quantifying the extent to which a set S covers the chemical space defined by a reference set R . In our simulations, R represent the full dataset, with

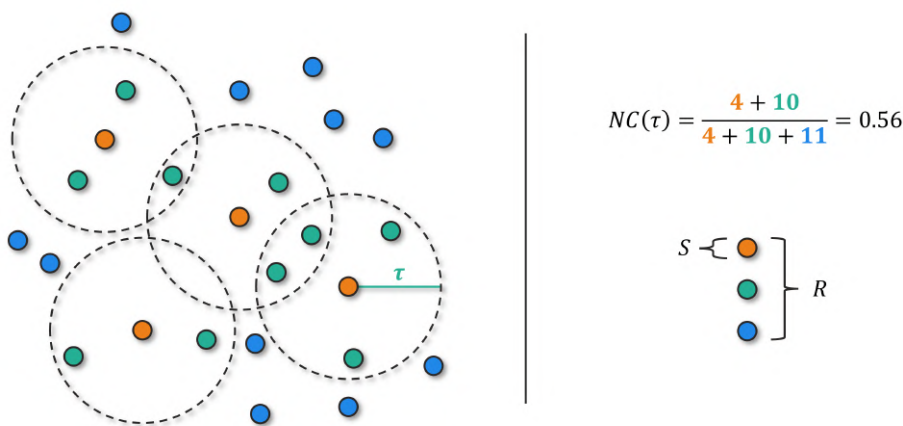


Figure 3.12: Illustration of the Neighborhood Coverage metric for an undefined threshold τ . All molecules are part of the reference set R . The metric is computed based on the neighborhood (dashed circles) of molecules from S (orange). Molecules that lie within the specified Tanimoto-distance threshold τ from molecules in S are counted and divided by the number of molecules in R .

S being the subset selected during at a given iteration of a simulation.

We define the Neighborhood Coverage (NC), a distance-based metric measuring the fraction of molecules in R that have at least one neighbor in S within a specified Tanimoto-distance threshold τ , as illustrated in Figure 3.12. This metric uses an indicator function I defined as:

$$I(i, \tau) = \begin{cases} 1, & \text{if } \min_{j \in S} \delta_{j,i} \leq \tau \\ 0, & \text{otherwise} \end{cases} \quad (3.30)$$

Neighborhood Coverage is then computed as:

$$NC(S, R, \tau) = \frac{1}{|R|} \sum_{i \in R} I(i, \tau) \quad (3.31)$$

This metric effectively captures chemical space coverage but requires careful distance threshold selection, typically around 0.3 (corresponding to a similarity of 0.7) using the ECFP4 binary fingerprint with 2048 bits. However, selecting the threshold remains subjective and depends on the choice of fingerprint and the expert’s judgment [271].

3.4.3.3 Neighborhood Coverage AUC

A limitation of the Neighborhood Coverage metric is its sensitivity to the choice of the distance threshold, τ . To overcome this dependency, we use the Neighborhood Coverage Area Under the

Curve (NCAUC). This metric integrates the Neighborhood Coverage values across the entire range of possible thresholds (typically 0 to 1). By calculating the area under the curve generated by plotting Neighborhood Coverage against τ , NCAUC provides a single, more robust measure of coverage. This eliminates the need to select a specific threshold, offering a convenient, parameter-free evaluation to the price of a lower dynamic range.

$$\text{NCAUC}(S, R, \tau) = \int_0^1 \left[\frac{1}{|R|} \sum_{i \in R} I(i, \tau) \right] d\tau \quad (3.32)$$

In our implementation, the Neighborhood Coverage is computed at discrete thresholds (e.g., $\tau = 0, 0.05, 0.10, \dots, 1.0$), followed by numerical integration to estimate the area under the curve.

3.4.3.4 Other metrics

Although the four previously described metrics are the main ones that we consider in this chapter, other have been integrated to the framework. Table 3.7 summarizes all the exploration metrics available in the framework, specifying whether they are monotonic or not and whether they depend on a reference set or not.

3.5 Practical demonstration on the FXa dataset

We illustrate the retrospective analysis pipeline, from the simulation of virtual scenarios to the evaluation of outcomes, using the FXa dataset. This dataset was selected because it is the largest among the four, both in terms of the number of molecules and the number of endpoints. Similar figures for the other datasets can be found in Appendix B.

3.5.1 Simulations

3.5.1.1 Desirable molecules of the FXa dataset

To evaluate the exploitative capabilities of the different selection strategies, we categorized molecules according to their experimental desirability scores. These scores were calculated for all compounds in the FXa dataset, and several top-tier categories were defined, as previously described. The number of compounds in these categories is illustrated in Figure 3.13. In subsequent analyses, we primarily focus on the "Top DScore 5%" category for statistical robustness, while additionally examining the "Top DScore 3%" and "Top DScore 1%" category to specifically highlight the highest-performing compounds. Since all molecules have complete documentation across all endpoints, the "Top DScore"

Exploration metric	Monotonic	Reference-based	Description
Internal Diversity	No	No	Average pairwise Tanimoto distance among all molecules in a set.
Neighborhood Coverage	Yes	Yes	Measures the extent to which a subset covers the chemical space of a larger reference set. It calculates the fraction of molecules in the reference set that have at least one nearby neighbor in the subset, based on a distance threshold.
Neighborhood Coverage AUC	Yes	Yes	Integrates Neighborhood Coverage across all possible distance thresholds to produce a robust, threshold-independent measure of chemical space coverage, mitigating sensitivity to a single arbitrary threshold choice.
Bemis-Murcko Coverage	Yes	Yes	Proportion of unique Bemis-Murcko scaffolds present in a reference set that are also represented within a subset. It explicitly highlights structural diversity and scaffold representativeness.
#Circles	Yes	No	Maximum number of mutually exclusive clusters (circles) of molecules within a specified diameter in a chemical set, providing a clustering-based measure of diversity.
Functional Group Coverage	Yes	Yes	Proportion of unique functional groups present in a reference set that are also represented within a subset. It emphasizes chemical diversity based on specific molecular substructures or functional moieties, reflecting chemically relevant structural diversity.
Ring Systems Coverage	Yes	Yes	Proportion of unique ring systems present in a reference set that are also represented within a subset. It specifically captures diversity in cyclic structural frameworks, highlighting coverage of chemically significant ring motifs.
Diameter	Yes	No	Maximum pairwise distance (e.g., Tanimoto distance) between any two molecules within a set.
Bottleneck	Yes	No	Minimum pairwise distance (e.g., Tanimoto distance) between any two molecules within a set.

Table 3.7: Summary of exploration metrics available in the framework, including monotonicity, dependency on a reference set, and a brief description of each metric.

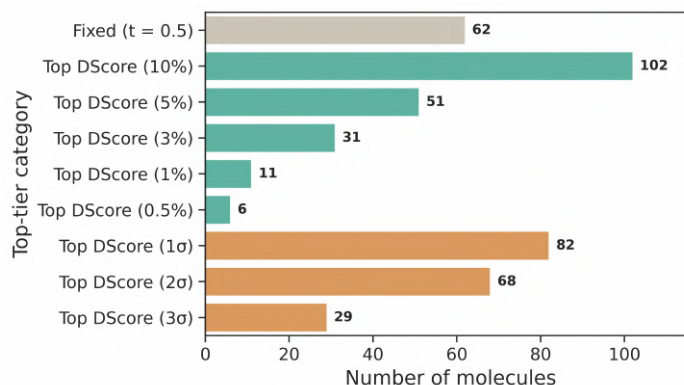


Figure 3.13: Number of molecules falling into each top-tier category for the FXa dataset out of the 1015 molecules. Categories are color-coded: brown indicates a fixed desirability threshold; orange indicates percentile-based thresholds; and green indicates standard deviation-based thresholds.

and "Top DocDScore" categories are identical.

3.5.1.2 Simulation setup

The FXa simulation begins on October 1, 2002, following the previously outlined criteria. At the start of the simulation, 154 compounds had already been evaluated for FXa potency. The simulation continues until September 19, 2007, matching the timeline of the actual project (Table 3.2). During this period, approximately 14 compounds were synthesized and tested each month. The simulation uses a timestep of one month and a batch-size factor of 0.5. At each iteration, the acquisition function selects from the pool set a number of new molecules equal to half of those synthesized in the actual project during the corresponding timestep (Figure 3.14A). Overall, a simulation consists of 60 iterations, during which 58% of the molecules from the original project are selected (15% initially plus an additional 43% via the acquisition function; Figure 3.14B).

3.5.1.3 Acquisition functions

Although a broader set of acquisition functions is available (see Table 3.5), our analysis focuses on ten representative strategies selected to cover a spectrum of exploratory and exploitative behaviors. Specifically, we include the Random and Oracle strategies as baseline references for comparative evaluation. The exploitative strategies considered are Desirability, Desirability with applicability domain, and Similar-to-good-known. Conversely, Uncertainty and Coverage exemplify purely explorative approaches. Lastly, U/D Harmonic, Greediverse, and Desired Spread represent trade-off strategies that balance exploration and exploitation.

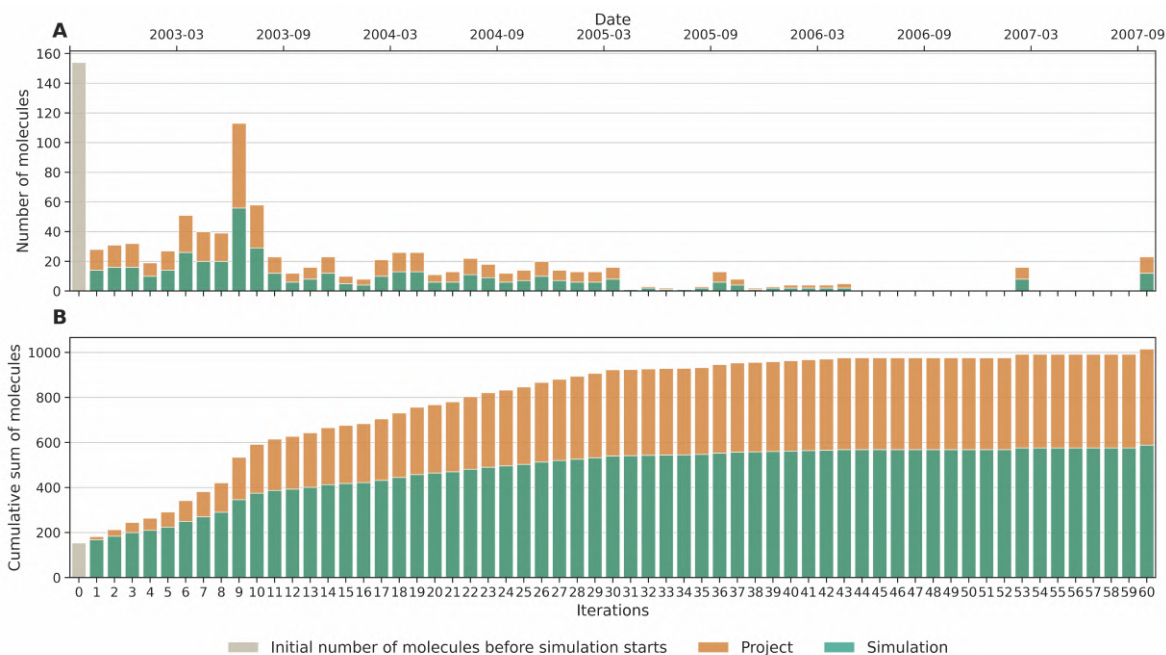


Figure 3.14: Comparison of compound selection between the real project and the simulated scenario for the FXa dataset: (A) the number of compounds selected per iteration, and (B) the cumulative number of selected compounds.

3.5.2 Understanding the chemical space with TMAPs

Figure 3.15A depicts the chemical space at simulations' start (t_0), illustrating that the regions initially covered are small compared to the full chemical landscape. While a few compounds from the Top DScore 5% category are located in these areas, most of them lie in regions that are unexplored at the start of the simulation.

Figure 3.15B shows the projection of the main chemical series on the KDE TMAP, revealing distinct yet partially overlapping domains. Figure 3.15C displays the structure of some of the best compounds and their location in the tree-graph layout, revealing patterns of structural similarity and diversity among them, and providing insight into how features transition across the chemical space within a series or between two similar series. For instance, the close spatial proximity between the Azole and Indole series is explained by their high scaffold similarity, a relationship mirrored between the Aminoquinoline and Ketopiperazine series.

3.5.3 Visualizing simulated scenarios

Prior to quantitatively assess the different acquisition functions, we conducted a qualitative visual inspection of the subset of the chemical space of the FXa project (Figure 3.16 124) selected by each of them at the end of the simulation.

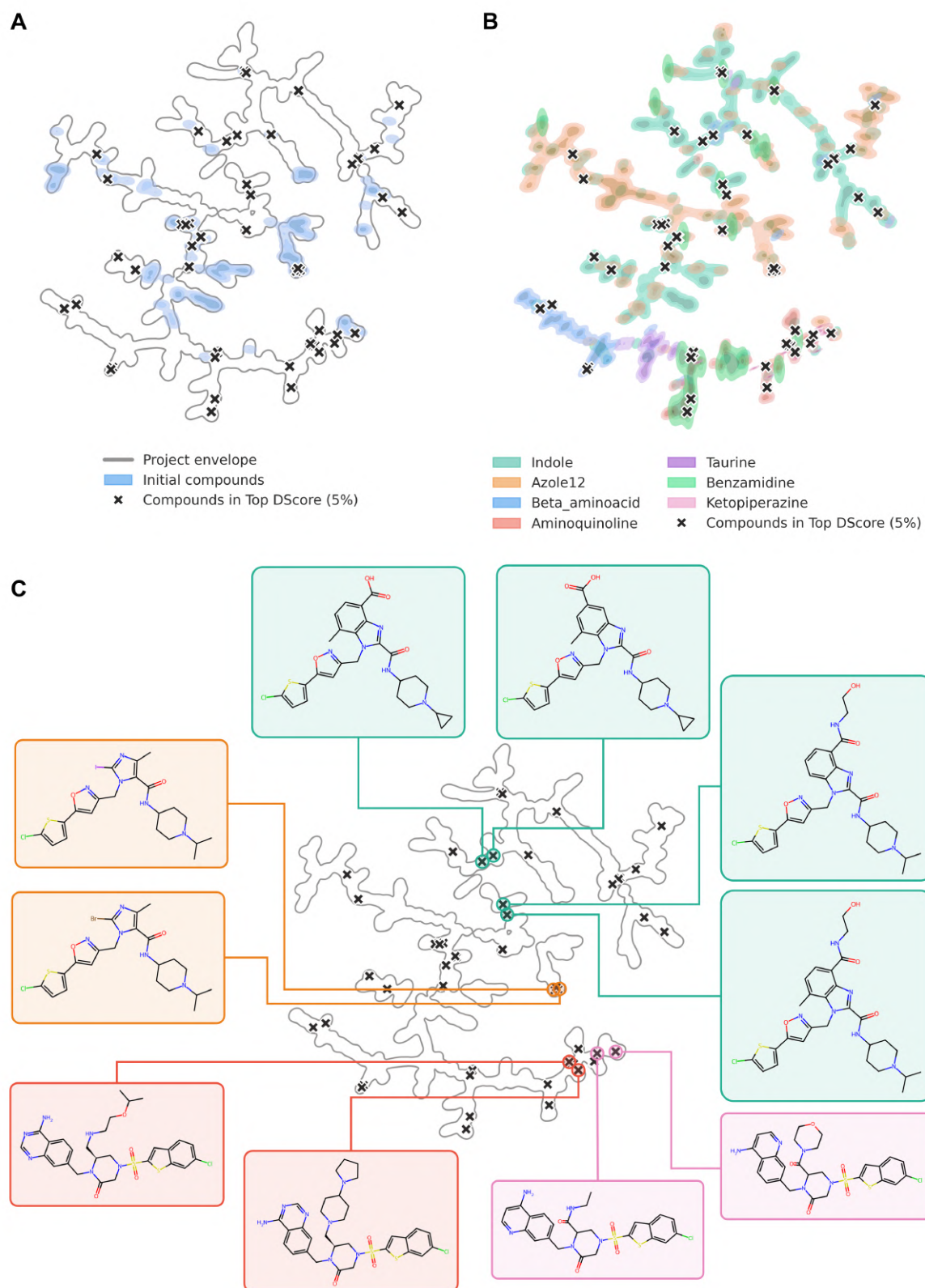


Figure 3.15: Overview of the chemical space at the start of a simulation (A), distributions of major chemical series (B) and examples of chemical structures of compounds belonging to Top DScore 5% from various chemical series (C). Color code of boxes for C are the same as the color code of B.

Strategies explicitly designed for exploration, such as Coverage (Figure 3.16B) and Uncertainty (Figure 3.16C), broadly sample compounds across the chemical space, leaving very few regions unexplored. However, these strategies often exhibit weaker exploitative capabilities, recovering only a limited subset of top-tier molecules.

Conversely, strategies emphasizing exploitation present an opposite pattern. For instance, the Desirability strategy (Figure 3.16E), aimed at selecting compounds with the highest predicted desirability scores, recovers most of the top molecules but neglects portions of the chemical space. Imposing an applicability domain constraint (Figure 3.16F) further restricts exploration, missing top compounds in the Aminoquinoline, Ketopiperazine, and Taurine series. Similarly, the Similarity-to-known-good strategy (Figure 3.16D) displays excellent exploitation within Azole and Indole series, capturing almost all top-tier molecules within these clusters but neglecting broader exploratory coverage beyond these areas.

Acquisition functions incorporating a built-in balance between exploration and exploitation perform as expected. The Greediverse strategy (Figure 3.16G) moderately enhances exploration, especially within the Beta Aminoacid and Benzamidine series. Additionally, the Desired Spread strategy (Figure 3.16H) successfully retains the strong exploitative performance of Desirability while substantially improving exploratory coverage, approaching the breadth seen in purely exploration-focused strategies.

3.5.4 Exploration

We quantified the acquisition functions’ exploration capabilities through three complementary metrics: Neighborhood Coverage and its area-under-the-curve (AUC) variant, both monotonic and reference-dependent, and Internal Diversity, which is non-monotonic and reference-independent.

Figure 3.17 shows the progression of these metrics over iterations, yielding coherent comparative insights. Initially, most acquisition functions exhibit similar behaviors, but the Desirability strategy (particularly with applicability domain constraints) and Similarity-to-known-good lag behind early, consistently underperforming throughout the simulation. Strategies designed explicitly for exploration, such as Coverage, consistently achieve high metric scores. Notably, the trade-off Desired Spread strategy surpasses purely exploratory strategies, indicating that integrating exploitation and exploration thoughtfully can enhance overall exploration effectiveness. Neighborhood Coverage and its AUC variant provide nuanced insights into each strategy’s progression. The non-monotonic behavior of Internal Diversity additionally highlights specific periods of chemical space concentration (decreases) or expansion (increases).

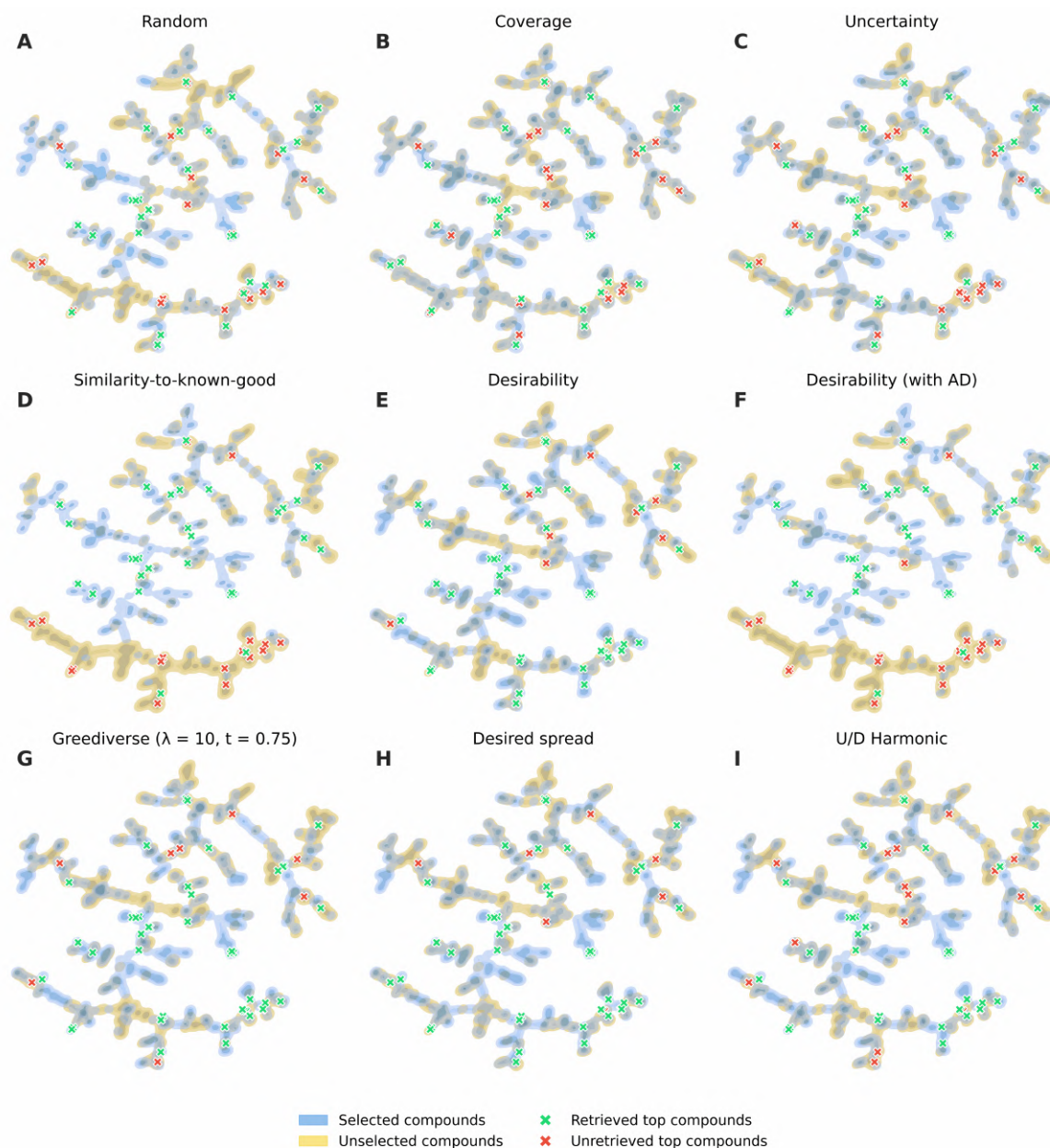


Figure 3.16: TMAP visualizations illustrating the final chemical space after 60 iterations for nine acquisition functions on the FXa dataset. Blue regions indicate sampled subspaces (selected molecules), while yellow denotes unsampled compounds. Crosses highlight compounds within the Top DScore 5%, where green marks successful retrieval and red indicates missed compounds.

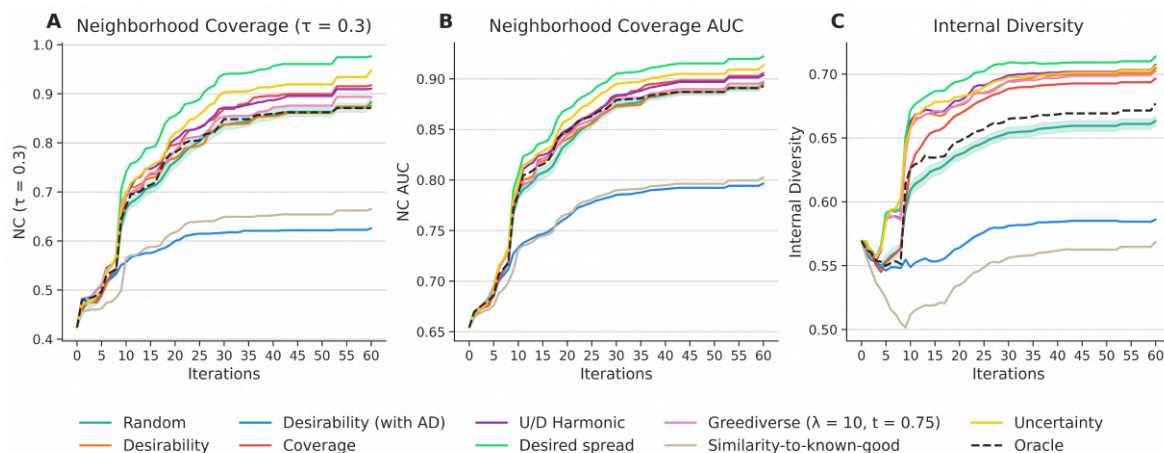


Figure 3.17: Evolution of exploration metrics over iterations for selected acquisition functions applied to the FXa dataset: (A) Neighborhood Coverage (with threshold $\tau = 0.3$), (B) Neighborhood Coverage AUC, and (C) Internal Diversity. The shaded area around the Random acquisition function curve represents the standard deviation across 100 replicates.

3.5.5 Exploitation

While exploration remains integral to drug discovery, the lead optimization phase primarily aims to efficiently exploit chemical space to identify the most promising candidates. To evaluate exploitation performance, we utilized the Percentage of Top Molecules Retrieved (PTMR) metric, focusing specifically on molecules within the Top DScores 5%, 3% and 1% categories. These categories correspond to 51, 31, and 11 compounds, respectively, out of the total dataset of 1015 molecules. Each selection strategy’s effectiveness was assessed by its capability to recover these top-ranked compounds while sampling only a subset of the complete dataset. For comparison, we included two baseline scenarios: the Random strategy, approximating actual project performance at a similar batch size, and the Oracle strategy, representing an optimal exploitative outcome.

Figure 3.18 illustrates the exploitation performance of the acquisition functions considered, tracking their PTMR across the simulation. For all defined categories of top-ranked molecules, exploratory strategies such as Coverage and Uncertainty showed performance similar to the Random baseline, recovering approximately 60% of the top compounds. In contrast, explicitly exploitative strategies, namely Desirability and Greediverse, significantly surpassed the Random baseline by retrieving between 80% and 100% of the top molecules, depending on the category considered.

Desirability (with AD) and Similarity-to-known-good strategies demonstrated performance equivalent to the Random baseline. On the other hand, trade-off strategies, including U/D Harmonic and particularly Desired Spread, delivered some of the best results, surpassing purely exploitative approaches.

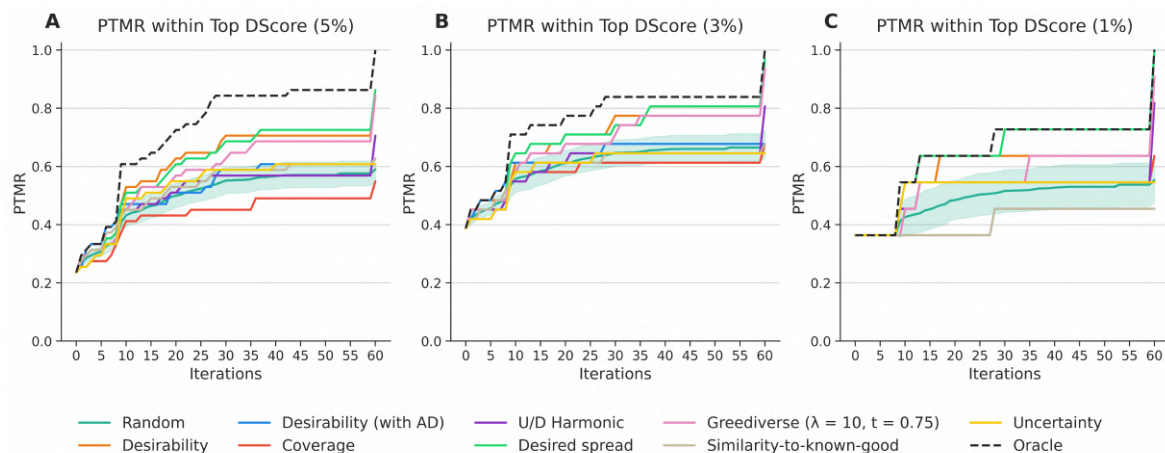


Figure 3.18: Exploitation performance of various acquisition functions on the FXa dataset monitoring the good molecules within Top DScore 5% (A), Top DScore 3% (B) and Top DScore 1% (C) categories. For the Random acquisition function, the envelope around the curve is the standard deviation across the 100 replicates.

A noteworthy observation is that the narrower the top-ranking category, the more closely the best performing strategies approached the Oracle baseline, achieving higher final PTMR scores. Importantly, these high exploitation performances were accomplished despite selecting only 58% of the total available molecules.

3.5.6 The exploitation/exploration trade-off

Efficient lead optimization demands a careful balance between exploitation and exploration of the chemical space. In previous sections, we introduced metrics specifically designed to quantitatively assess and compare strategies with respect to these two objectives. To visualize and compare the exploitation-exploration trade-off performed by the different strategies, we plot these metrics against each other.

In Figure 3.19, each of the three previously described exploration metrics (Neighborhood Coverage, Neighborhood Coverage AUC, and Internal Diversity) is plotted against the PTMR exploitation metric across three distinct top-tier categories (Top DScore 5%, 3%, and 1%). These visualizations summarize how strategies balance exploitation with exploration during the simulations. Notably, strategies excelling in exploitation generally exhibit exploration comparable to the Random baseline, with the exception of the Desired Spread strategy, which achieves markedly superior exploration. These quantitative findings confirm the qualitative observations from the TMAP visualizations shown in Figure 3.16.

To facilitate comparisons among a broader set of strategies, we display on Figure 3.20 the exploitation-exploration performance at the end of the simulations. The Random baseline provides a critical

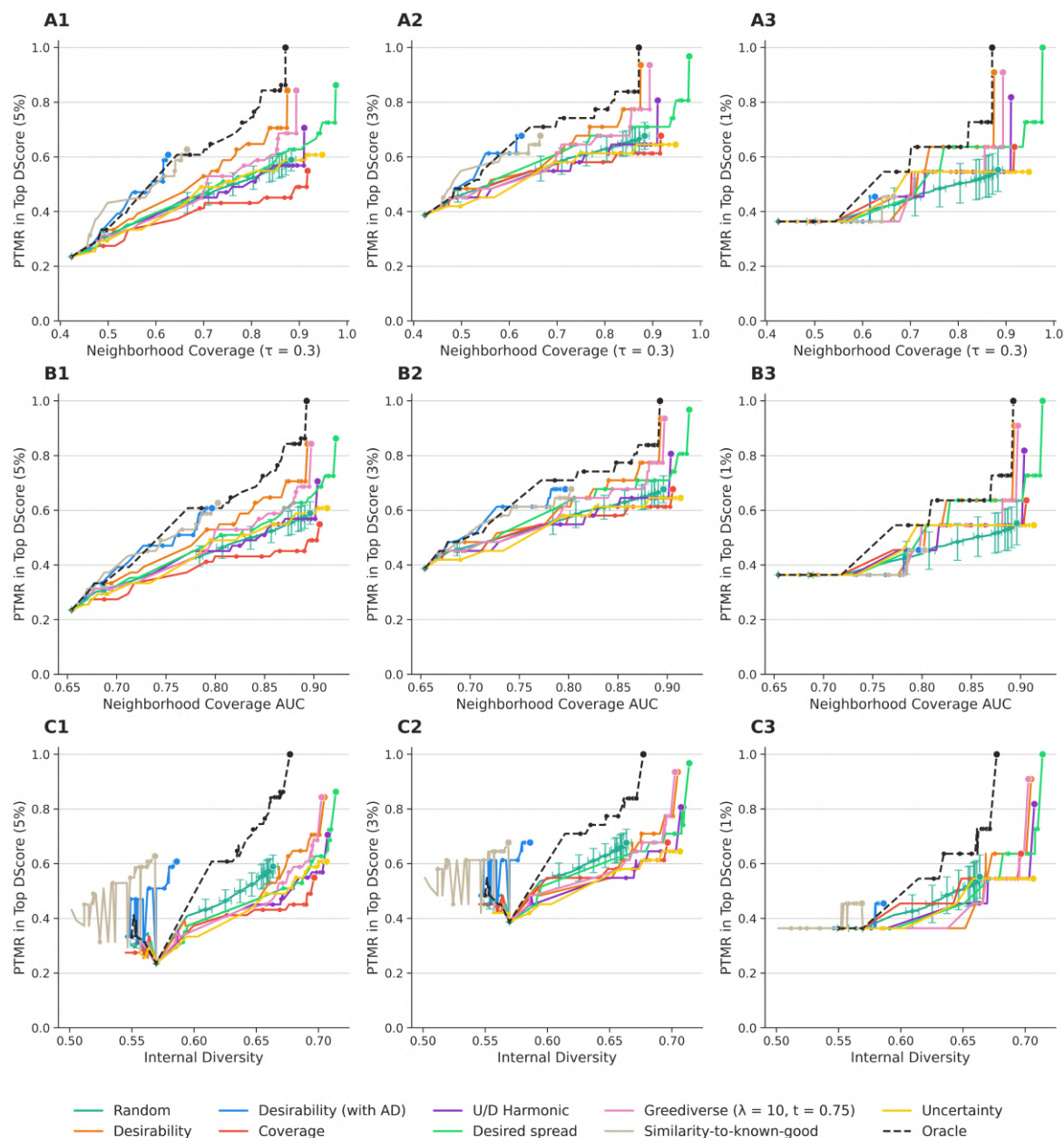


Figure 3.19: Exploitation-exploration trajectories considering PTMR within Top DScore 5%, 3%, and 1% (columns 1, 2, and 3, respectively), plotted against exploration metrics: Neighborhood Coverage ($\tau = 0.3$), Neighborhood Coverage AUC, and Internal Diversity (rows A, B, and C, respectively).

reference point, delineating three distinct performance regions:

- Strategies exhibiting performance that is inferior to Random in both exploitation and exploration.
- Strategies achieving superior exploitation but inferior exploration than Random
- Strategies better than Random in both exploitation and exploration.

Identifying strategies that fall into this last category is a primary objective of our analysis. These strategies effectively optimize exploitation and exploration, maximizing both outcomes while sampling a fraction of the whole dataset.

3.5.7 Impact of the predictive model

Some acquisition strategies rely on predictive modeling for their selection process. So far, was used the default predictive model set in our framework, a Random Forest regressor with uncertainty estimation derived from the standard deviation across its trees. To assess the influence of the predictive model on the performance of the selection strategies, we performed simulations using different predictive models that are available in the framework.

Figure 3.21 (page 130) illustrates the final exploitation-exploration trade-offs achieved by these strategies using different predictive models. In this figure, marker colors correspond to different selection strategies, while marker shapes denote the predictive models employed during simulations. Additionally, the Random and Oracle baselines are included for reference.

The plots reveal that strategies consistently occupy similar regions of the exploitation-exploration space regardless of the predictive model chosen. Although minor variability is observed within individual strategies, no specific predictive model significantly outperforms the others.

3.6 Conclusion

3.6.1 A realistic retrospective analysis framework

The work presented in this chapter aims to advance methods that enhance the efficiency of the real-world drug discovery process, especially during the lead optimization phase. We have developed a comprehensive toolkit incorporating a diverse set of selection strategies inspired by active learning, multiple-criteria decision analysis, and standard practices of medicinal chemists. Alongside

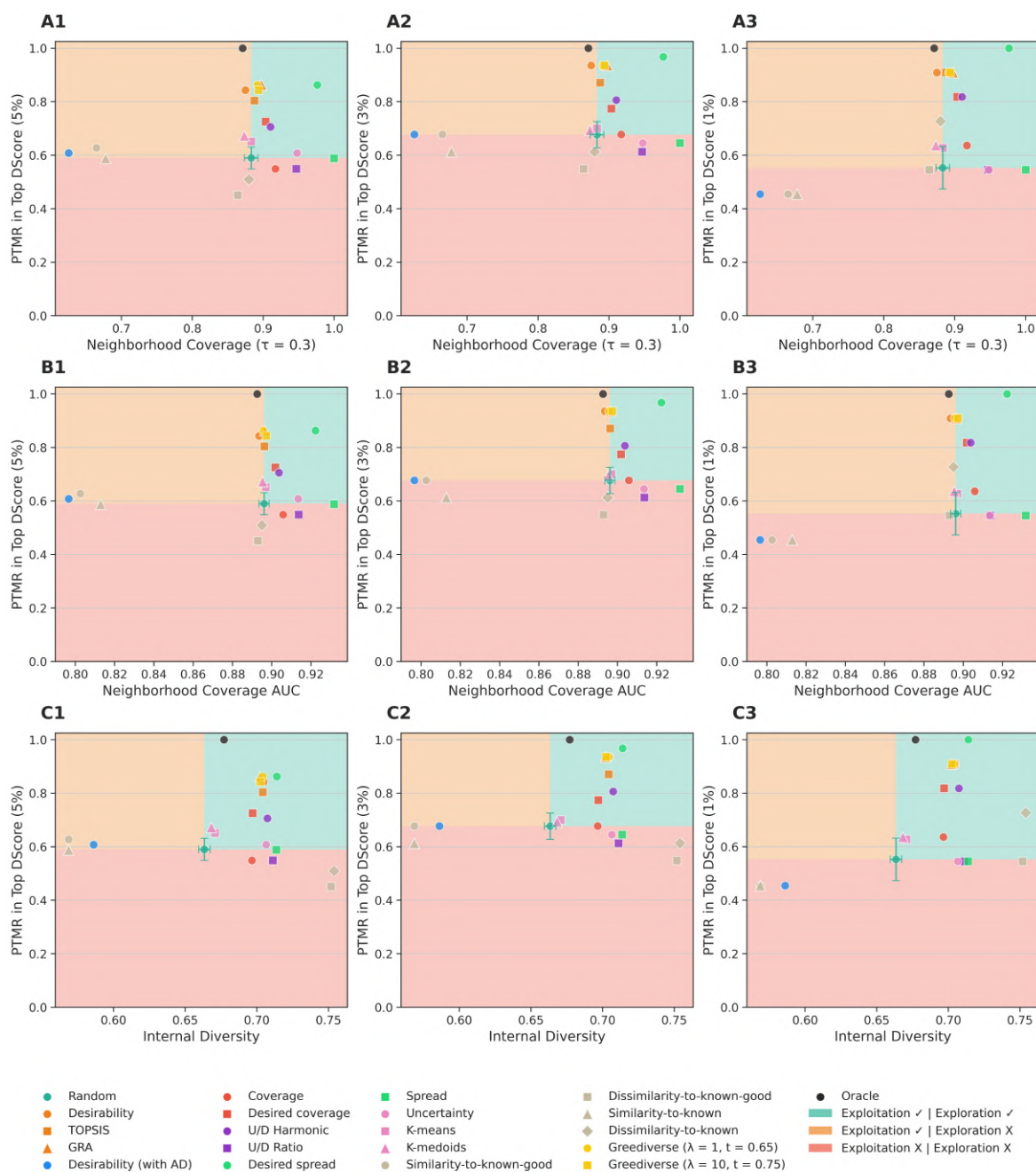


Figure 3.20: Final-state exploitation-exploration performance at iteration 60, considering PTMR within Top DScore 5%, 3%, and 1% (columns 1, 2, and 3, respectively) plotted against exploration metrics: Neighborhood Coverage ($\tau = 0.3$), Neighborhood Coverage AUC, and Internal Diversity (rows A, B, and C, respectively).

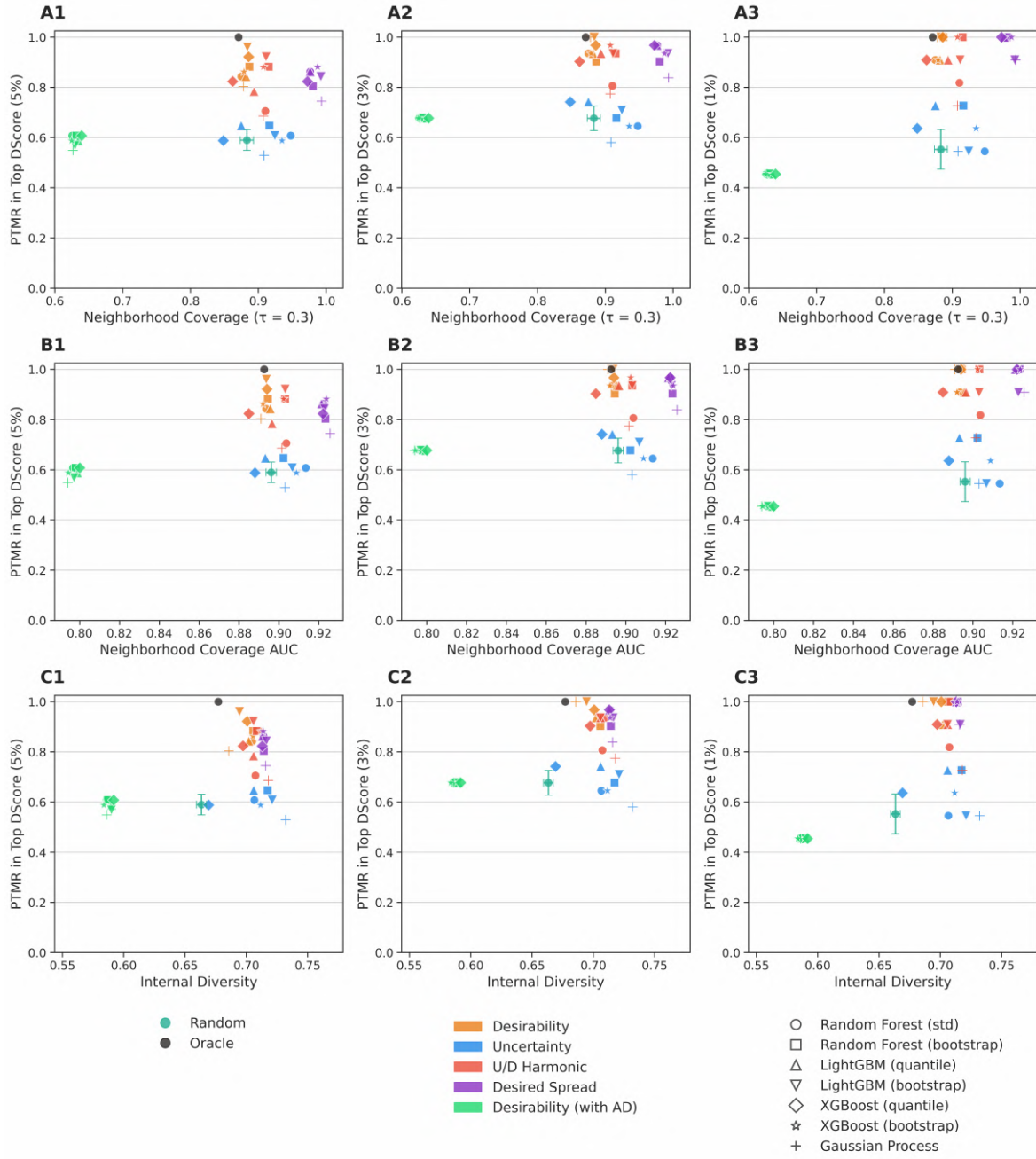


Figure 3.21: Impact of the predictive model: final-state exploitation-exploration performance, considering PTMR within Top DScore 5%, 3%, and 1% (rows A, B, and C, respectively) plotted against exploration metrics: Neighborhood Coverage ($\tau = 0.3$), Neighborhood Coverage AUC, and Internal Diversity (columns 1, 2, and 3, respectively). The color of a marker represent the strategy and its shape represent the type of predictive model. Random and Oracle baseline are also displayed.

this toolkit, we have established a methodology for benchmarking these strategies through retrospective simulations using legacy lead optimization datasets. Unlike previous studies [236, 246, 41], our approach explicitly incorporates real experimental data, accounting for temporal dynamics and simultaneously optimizing multiple endpoints, thus providing a more realistic assessment.

The virtual simulations developed in this work offer realistic yet cost-effective platforms to systematically evaluate different strategies without requiring laboratory experimentation. To facilitate the analysis of these simulations, we introduced a combination of qualitative and quantitative analytical tools designed to reveal how various strategies manage the exploitation-exploration trade-off within chemical space. Understanding this balance is crucial for identifying strategies that can be relevant to use in practice. Consequently, our framework acts as a proof-of-concept to highlight strategies with strong potential for adoption in real-world drug discovery projects. For instance, in the FXa dataset, strategies such as the Desired Spread demonstrated superior performance, achieving greater exploitation efficiency compared to the greedy Desirability strategy, while also surpassing the Random baseline in terms of exploration.

3.6.2 Methodological limitations

We recognize several methodological limitations inherent in our approach, mainly arising from the complexities of simulating realistic lead optimization scenarios.

First, retrospective simulations assume that all historically available compounds at any given time point t can be selected immediately. This simplification disregards sequential dependencies present in real-world scenarios, where synthesizing and testing molecule "A" may inspire the subsequent design of molecule "B," which then again could inspire the design of molecule "C". Retrospective simulations might select molecule "A" and then directly molecule "C," bypassing molecule "B". This behavior can inflate exploitation metrics by enabling the selection of good compounds without having to select the ones that logically led to them. Consequently, strategies focusing mainly on exploitation such as Desirability may demonstrate reduced effectiveness in practice. However, hybrid strategies incorporating broader exploration of chemical space might partially mitigate this bias.

Second, the retrospective approach only relies on molecules that were designed and synthesized during projects, as it is a requirement to access their experimental data. Hence, it does not account for compounds that were designed by the chemists during the project but not selected, therefore limiting the pool of hypothesis compared to what the chemists faced during the real project.

Lastly, the datasets used in this chapter represent only a fraction of the original lead optimization projects they are taken from, limited in terms of compounds and measured endpoints with a modified blueprint. Therefore, caution is necessary when interpreting results derived from using the Random

strategy as a baseline to approximate the historical performance of project teams since the original optimization was made on more molecules considering more endpoints.

3.6.3 The need for more realistic datasets

In the context of this work, we provided datasets for the academic community. Although the molecular structures and potency endpoints included were previously available publicly, we have published additional data, notably supplementary endpoints and, critically, the introduction of timestamps for each compound. These timestamps document the chronological progression of projects, addressing the significant yet often overlooked temporal aspect inherent in drug discovery. As emphasized in Chapter 2, temporality profoundly affects predictive modeling due to the chemical shifts occurring over time in real-world projects. To our knowledge, this represents the first publicly available industry-derived dataset explicitly capturing the temporal dynamics of lead optimization. We hope that these contributions will benefit the broader research community.

Nevertheless, these datasets remain partial and do not represent complete drug discovery projects. Given that the primary objective of this chapter was to introduce the toolkit and methodology, this limitation does not undermine our immediate goals but does restrict the breadth of conclusions derived from simulations on these datasets. The subsequent chapter addresses this by utilizing complete datasets from legacy projects, encompassing thousands of molecules optimized on dozens of endpoints.

Chapter 4

Insights from retrospective analysis of legacy projects

In the previous chapter, the aim was to introduce the prioritization toolkit along with the analytical tools available to interpret the behavior of simulated scenarios. To achieve this, we used a partial dataset derived from a project targeting FXa, which had the benefit of including publicly available molecular structures. However, this dataset was limited both in terms of the number of compounds (as it did not encompass all molecules from the project) and in the scope of its blueprint, covering only three endpoints.

In this chapter, we present preliminary results from simulations conducted on twelve comprehensive industrial datasets. These datasets fully encompass all molecules from their respective projects and feature complete blueprints. They vary considerably in size, containing between approximately 1400 and 5500 compounds, and are documented across 10 to 35 endpoints. Our objective here is to highlight initial insights derived from these real datasets, laying the groundwork for further analyses planned on additional datasets.

4.1 Real industrial datasets

For intellectual property reasons, the names of datasets, endpoints and chemical series are anonymized, and no molecular structure can be shown. The datasets will be named "dataset X", with X being an integer between 1 and 12.

4.1.1 Exploratory data analysis

4.1.1.1 Uniqueness of each project

Each lead optimization project is inherently unique due to variations in scientific context, team composition, and resource allocation. Each target protein poses distinct challenges, demanding specific strategies and the optimal balance between potency, pharmacokinetics, and toxicity differs. Furthermore, the uniqueness of each project is accentuated by the distinct composition and dynamics of its interdisciplinary team, characterized by diverse expertise (and biases) or communication styles. Finally, management priorities directly influence the resources allocated to each project, highly impacting the synthesis efforts or timeline of a project.

The 12 datasets originate from projects conducted since the early 2010s. Figure 4.1 illustrates the temporal dynamics of these projects by showing the distribution of new molecules appearing over time. Project timelines vary significantly: some, such as dataset 1, began slowly from mid-2008 to early 2011 before experiencing a marked increase in activity. Others, including datasets 3, 9, and 12, started intensively and exhibited a gradual decline in the appearance of new compounds. Certain datasets, notably dataset 5 and dataset 6, display irregular patterns with clearly distinguishable phases separated by extended pauses, reflecting projects that were halted and subsequently restarted based on renewed scientific or strategic interest. Furthermore, a commonly observed phenomenon across these projects is the occurrence of distinct peaks, indicating periods when a substantial number of new molecules are synthesized. Such peaks often result from initiating large-scale screening campaigns, though they may also emerge from mergers or the integration of external datasets. A typical example occurs when a major pharmaceutical company acquires another entity, subsequently merging its existing projects and leading to a sudden influx of data.

In addition to their differing temporal dynamics, the datasets also exhibit substantial variation in size. Table 4.1 summarizes the number of unique compounds across the 12 projects, which range from 1432 compounds in dataset 12 to 5505 compounds in dataset 8. Furthermore, the number of endpoints per project varies considerably, spanning from 10 to 35. It is important to take note that, in this analysis, only endpoints containing at least 200 documented molecules throughout the duration of the respective projects were considered to ensure that QSAR/QSPR model would have enough data to be trained on.

4.1.1.2 Chemical heterogeneity and homogeneity

Projects may differ significantly from one another, but how variable are the molecules within a single project? Figure 4.2 presents the distribution of pairwise similarities among all molecules within

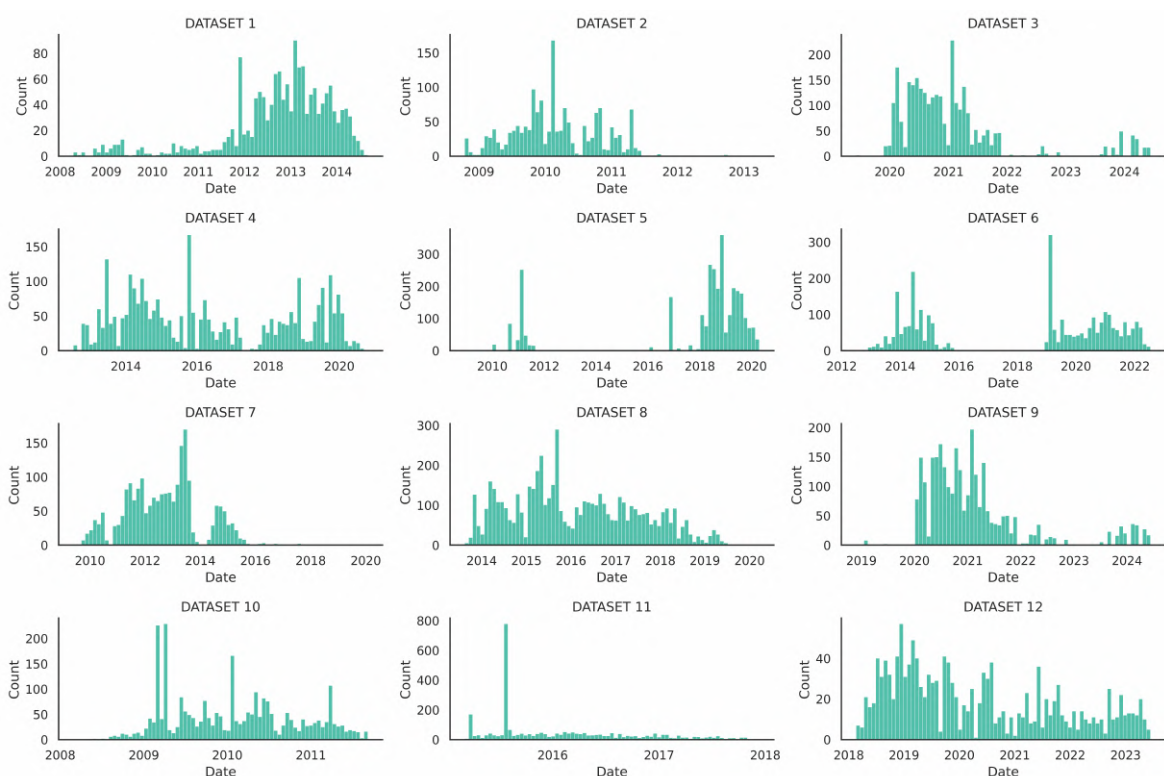


Figure 4.1: Temporal distribution of new molecules appearing over time across 12 lead optimization projects.

Dataset	Size	Endpoints
DATASET 1	1560	12
DATASET 2	1562	18
DATASET 3	2859	30
DATASET 4	3087	33
DATASET 5	2960	10
DATASET 6	2913	20
DATASET 7	2068	26
DATASET 8	5505	30
DATASET 9	2777	35
DATASET 10	2733	12
DATASET 11	2785	18
DATASET 12	1432	17

Table 4.1: Summary of the 12 lead optimization project datasets analyzed, showing substantial variation in dataset sizes (number of unique compounds, ranging from 1432 to 5505) and endpoint counts (ranging from 10 to 35).

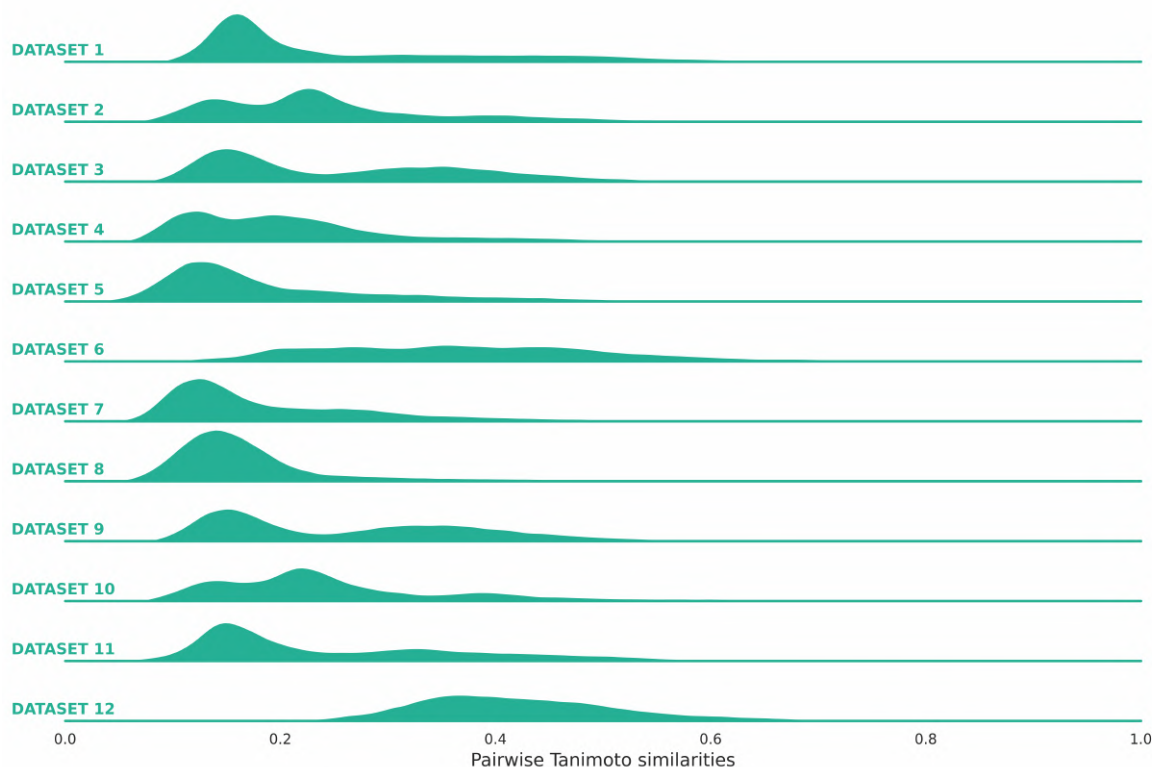


Figure 4.2: Distribution of pairwise molecular similarities within each of the 12 projects. Similarity values are calculated using the Tanimoto coefficient on ECFP4 binary fingerprints with 2048 bits. The distributions are generally skewed towards lower similarity, indicating significant chemical diversity across entire projects.

each project. Interestingly, these distributions are skewed towards lower similarity values, which is somewhat counterintuitive. As discussed in Chapter 1, projects typically consist of molecules belonging to a few chemical series, implying that they should share one of the common scaffolds and thus exhibit higher similarity. Rather than considering overall project distributions alone, it is more informative to analyze pairwise similarities within each chemical series, as shown in Figure 4.3. Here, both the overall similarity distribution and the distributions broken down by chemical series are presented. These results demonstrate that chemical series generally exhibit high internal similarity, aligning with previous explanations. While similarity within a chemical series is high, chemical series themselves can differ substantially from one another, accounting for the low overall similarity observed in Figure 4.2.

4.1.1.3 Data scarcity

To further investigate the characteristics of the individual projects, we analyze the documentation levels of various endpoint, a critical metric in defining a project's "good" molecules, as discussed previously in section 3.2.4 (page 98). As a reminder, a molecule's documentation is defined as the

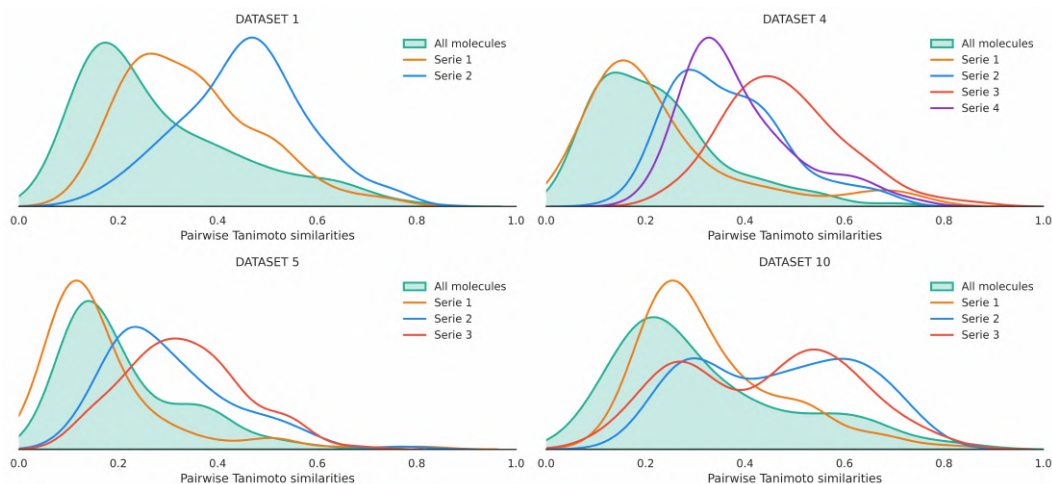


Figure 4.3: Pairwise molecular similarity distributions, both overall (filled) and separated by chemical series (lines). Similarities within individual chemical series tend to be higher, illustrating that internal diversity within chemical series is low, while diversity across different chemical series is high.

proportion of endpoints for which experimental data have been collected. Figure 4.4 presents the distribution of documentation levels across projects as histograms, revealing substantial variability among the distributions. Such variability is likely influenced by the specific design of each project's screening tree (Figure 1.17, page 47). Generally, since attrition tends to increase with tree depth, screening trees with numerous endpoints tested in their initial stages are expected to show documentation distributions skewed toward higher values, and vice versa. Nevertheless, given each project's unique context, other very specific factors may also significantly influence the shape of the documentation distributions.

4.1.2 Desirable molecules

The aim of the lead optimization phase is to design compounds meeting all the blueprint criteria. To characterize these compounds, we rely on the desirability score as defined in section 3.2.4 (page 98). Compounds with high desirability scores exhibit properties within optimal ranges across their documented endpoints. However, high desirability alone is insufficient if a compound has limited documentation, as this leads to significant uncertainty regarding its true quality. Consequently, we employ compound definitions that combine both desirability and documentation.

A straightforward approach applicable during active projects is to use fixed thresholds for documentation and desirability. As discussed in the previous chapter, we used threshold values of 0.65 for documentation and 0.5 for desirability. Figure 4.5 shows the distribution of documentation and desirability across the 12 datasets. However, in our retrospective analyses, this fixed-threshold definition shows limits. Some datasets, such as datasets 4 and 9, yield almost no compounds categorized

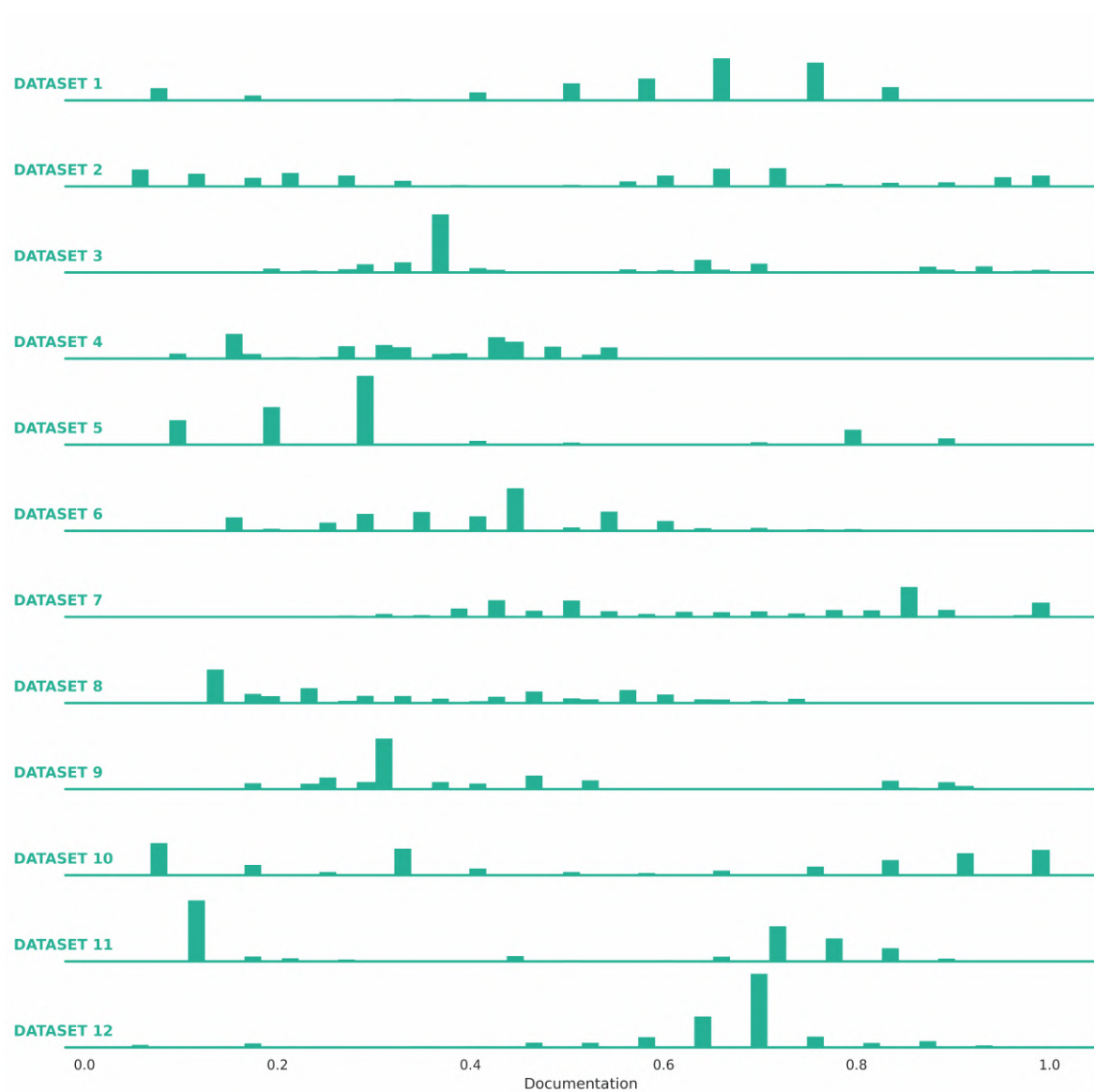


Figure 4.4: Distribution of documentation levels across the different projects. Documentation refers to the proportion of experimentally evaluated endpoints per molecule.

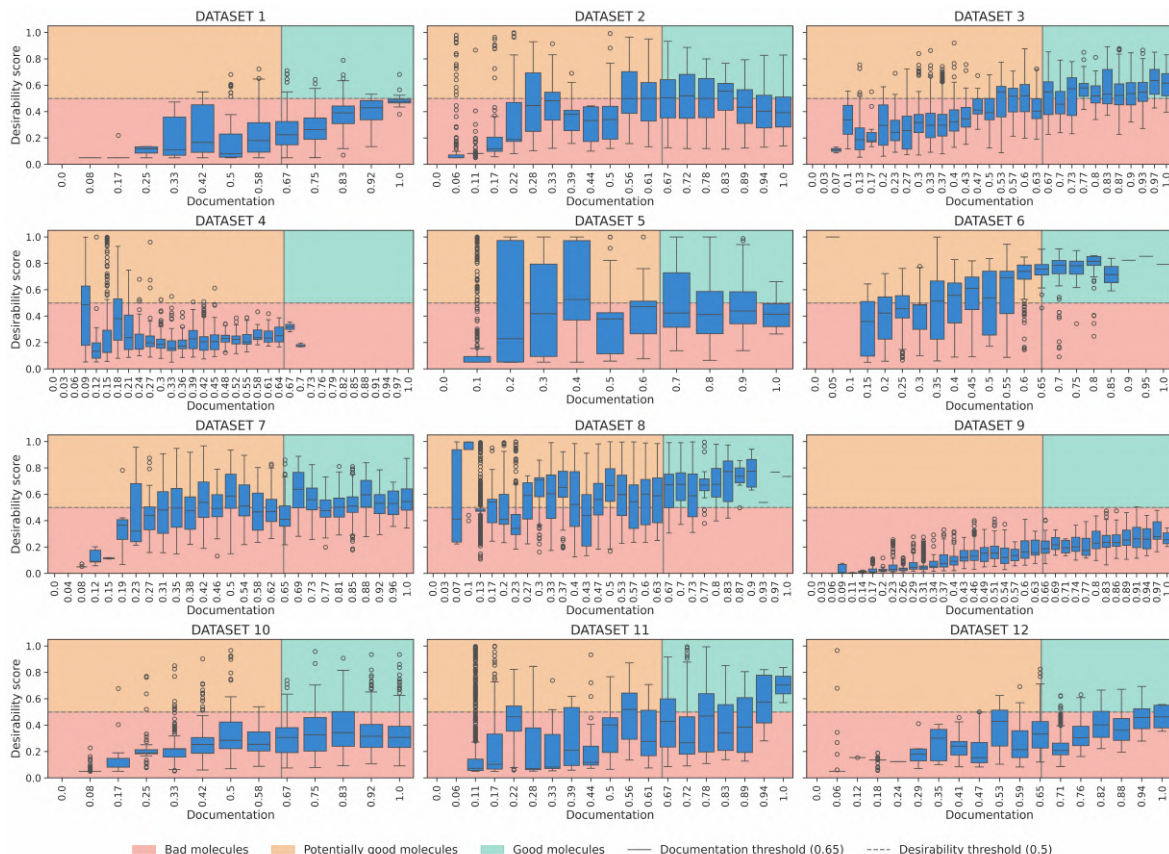


Figure 4.5: Distribution of the documentation and desirability for compounds from the 12 datasets and categorization using fixed thresholds (desirability ≥ 0.5 and documentation ≥ 0.65).

as good using this definition, while others, such as datasets 2, 3, 7, 8, and 11, categorize a large fraction of compounds as good, potentially indicating that the definition is overly permissive.

Therefore, we propose an alternative approach that combines a dataset-specific desirability threshold derived from the distribution of desirability scores with a fixed documentation threshold of 0.65. This approach yields the "Top DScore (X%)" categories, depicted in Figure 4.6. Although this method provides a more realistic assessment, some datasets, notably dataset 4, still contain few good compounds due to very low documentation levels. Figure 4.7 summarizes the number of compounds classified within the top categories, clearly demonstrating the improved consistency achieved by the distribution-based definition compared to the fixed-threshold approach.

4.1.3 Simulation parameters

This section describes the parameters used to conduct simulations on the 12 datasets. The starting date t_0 for each simulation is determined using the criteria defined in Chapter 3: at least 100 molecules documented on the main endpoint, with at least one molecule falling within the acceptable or optimal range as specified by the blueprint. At t_0 , the initial number of compounds varies



Figure 4.6: Distribution of the documentation and desirability for compounds from the 12 datasets and categorization using dataset-specific desirability thresholds derived from each dataset's desirability score distribution ("Top DScore (X%)"), combined with a fixed documentation threshold (≥ 0.65).

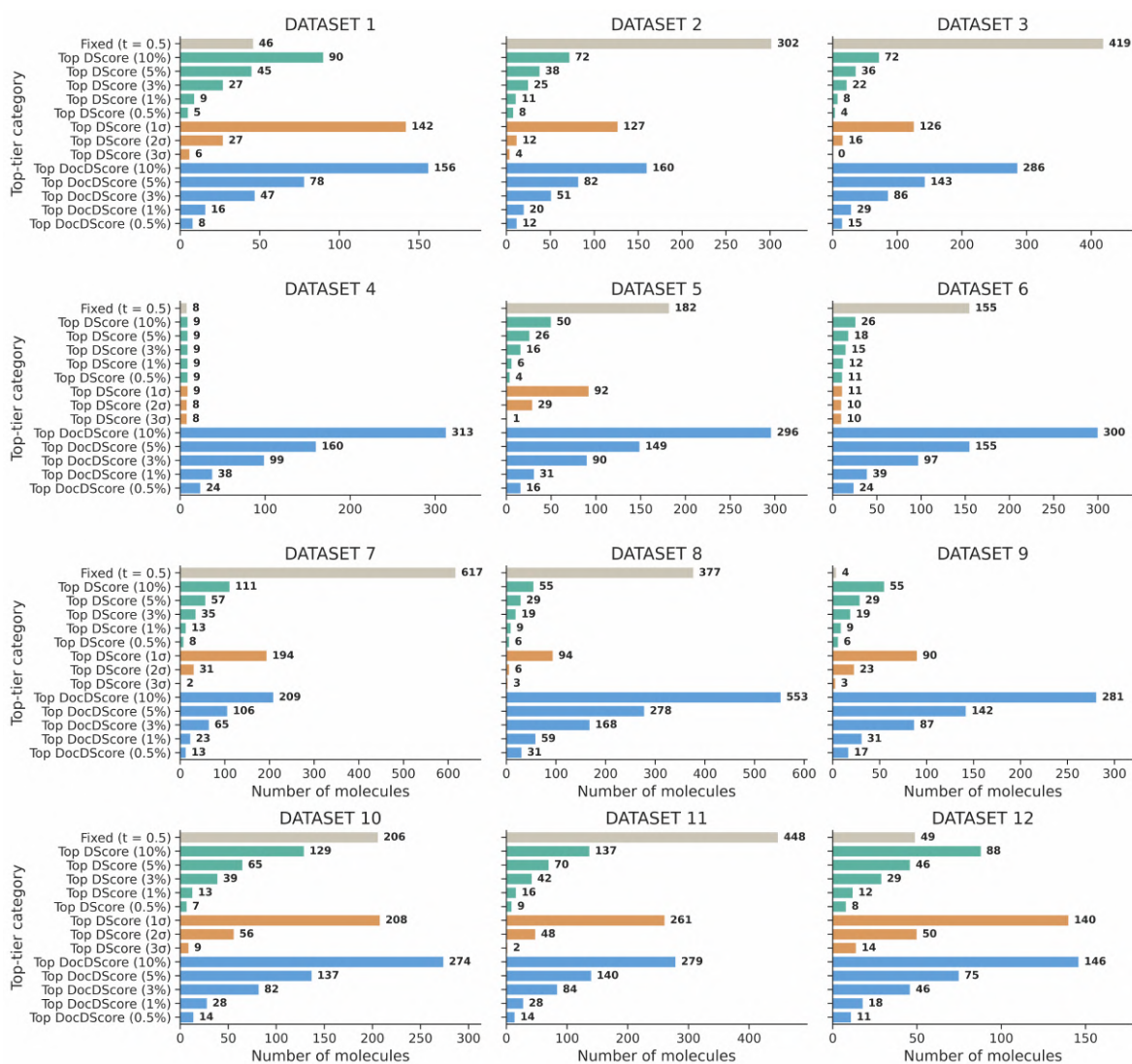


Figure 4.7: Summary of the number of compounds in each top category across the 12 datasets, highlighting the variability of fully fixed threshold-based definition compared to distribution-based.

Dataset	t_0	t_f	Δt	Iterations
DATASET 1	2012-07-01	2014-09-03	1	27
DATASET 2	2009-05-01	2013-03-26	1	47
DATASET 3	2020-06-01	2024-06-11	1	49
DATASET 4	2013-07-01	2021-01-01	2	44
DATASET 5	2018-01-01	2020-05-01	1	28
DATASET 6	2013-12-01	2022-07-18	2	52
DATASET 7	2010-08-01	2020-01-28	2	57
DATASET 8	2013-11-01	2020-03-20	2	39
DATASET 9	2020-03-01	2024-06-11	1	52
DATASET 10	2009-02-01	2011-09-07	1	32
DATASET 11	2015-04-01	2017-12-11	1	33
DATASET 12	2019-03-01	2023-06-15	1	52

Table 4.2: Simulation parameters for virtual scenarios across the 12 datasets. For each dataset, the starting date (t_0), ending date (t_f), and timestep duration (Δt) are specified. The default batch size is 0.5 for all datasets, and Random Forest models are used for predictions.

significantly across datasets, ranging from 120 in dataset 8 to 682 in dataset 5. If a dataset has a large number of compounds when the simulation starts, it means that none of the previous compound was in the acceptable or optimal range for the main endpoint. If we look back at the compound distribution over time on Figure 4.1 (page 135), we can see that those hundreds of compounds are the ones in the first project phase between 2010 and 2012, that was then resumed in 2016. The final date t_f aligns with the conclusion of each actual project. The timestep Δt is typically set to one month; however, for certain datasets covering an extended time span, a timestep of two months is used to reduce computational load. The number of iterations per simulation is calculated based on these parameters. Additionally, the default batch size is set to 0.5, and Random Forest models are used as predictors.

4.2 Visualizing the chemical space

Similarly to the previous chapter, we visualize the subset of the chemical space occupied at the beginning of the simulations.

4.2.1 Early stage of lead optimization

At the start of a project, the chemical space is generally limited, typically comprising one or a few chemical series that have not yet been extensively explored. Figure 4.8 shows the chemical space

at the beginning of simulations for each dataset compared to the overall project chemical space, visualized as KDE TMAPs. The results are consistent across all datasets, with the initial chemical space occupying only a small portion of the total project envelope. Furthermore, the top compounds, defined here as those within the "Top DScore 5%" category, are typically located outside this initial space. This indicates that the lead optimization process is still in its early stages. In the coming sections of this chapter, we will analyze how the different selection strategies exploit and explore beyond these initial chemical spaces.

4.2.2 Navigating between chemical series

Although chemical structures cannot be disclosed, TMAP visualizations effectively reveal structural relationships among project compounds. Figure 4.9 illustrates the distribution of the main chemical series across their respective chemical spaces for datasets previously described in Figure 4.3 (page 137). Typically, each chemical series occupies distinct regions within the chemical space, although some overlapping occurs, reflecting structural similarities between specific series.

4.2.3 Exploiting, exploring, or both

After completing the simulations, the final outcomes can be visualized by projecting both selected and non-selected compounds, highlighting those identified as top compounds. Figures 4.10 (page 147) and 4.11 (page 148) illustrate these outcomes for datasets 1 and 11, confirming patterns previously observed with the FXa dataset in Chapter 3.

The first row of each figure shows results from simulations employing exploitation-focused strategies: Similarity-to-known-good, Desirability, and Desirability combined with an applicability domain. In these TMAPs, the majority of top compounds (indicated by green crosses) are successfully identified. However, considerable portions of the chemical space remain unexplored (yellow areas), leading to isolated top compounds being overlooked. This limitation is more pronounced when the applicability domain is integrated with Desirability.

The second row depicts exploration-focused strategies: Uncertainty, Spread, and Coverage acquisition functions. These strategies display behaviors opposite to those in the first row. Although they thoroughly explore the chemical space (marked predominantly in blue), they fail to sufficiently exploit areas containing top compounds (red crosses), missing a substantial portion of them. Individually, these exploration-only approaches do not effectively meet the primary goal of lead optimization, which is identifying molecules that meet the project's blueprint. Nevertheless, when combined with exploitation strategies, they can leverage the strengths of both exploration and exploitation.

Finally, the third row illustrates three of these trade-off strategies that balance exploration and

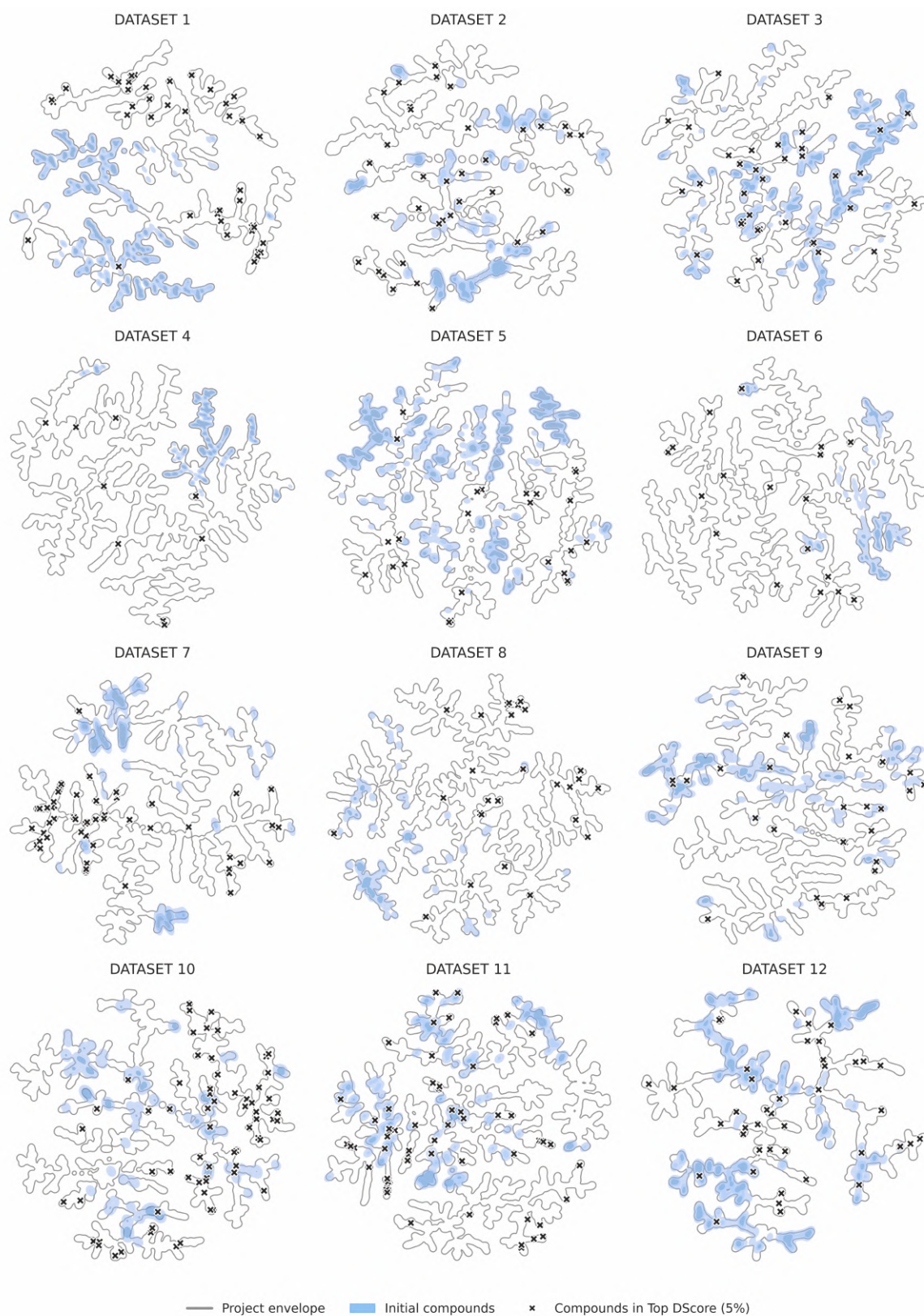


Figure 4.8: Overview of the initial chemical spaces (in blue) projected into the whole projects' envelopes (grey). The top compounds, defined as the ones belonging to the Top DScore % category, are defined by black crosses with a majority of them outside of the initial spaces.

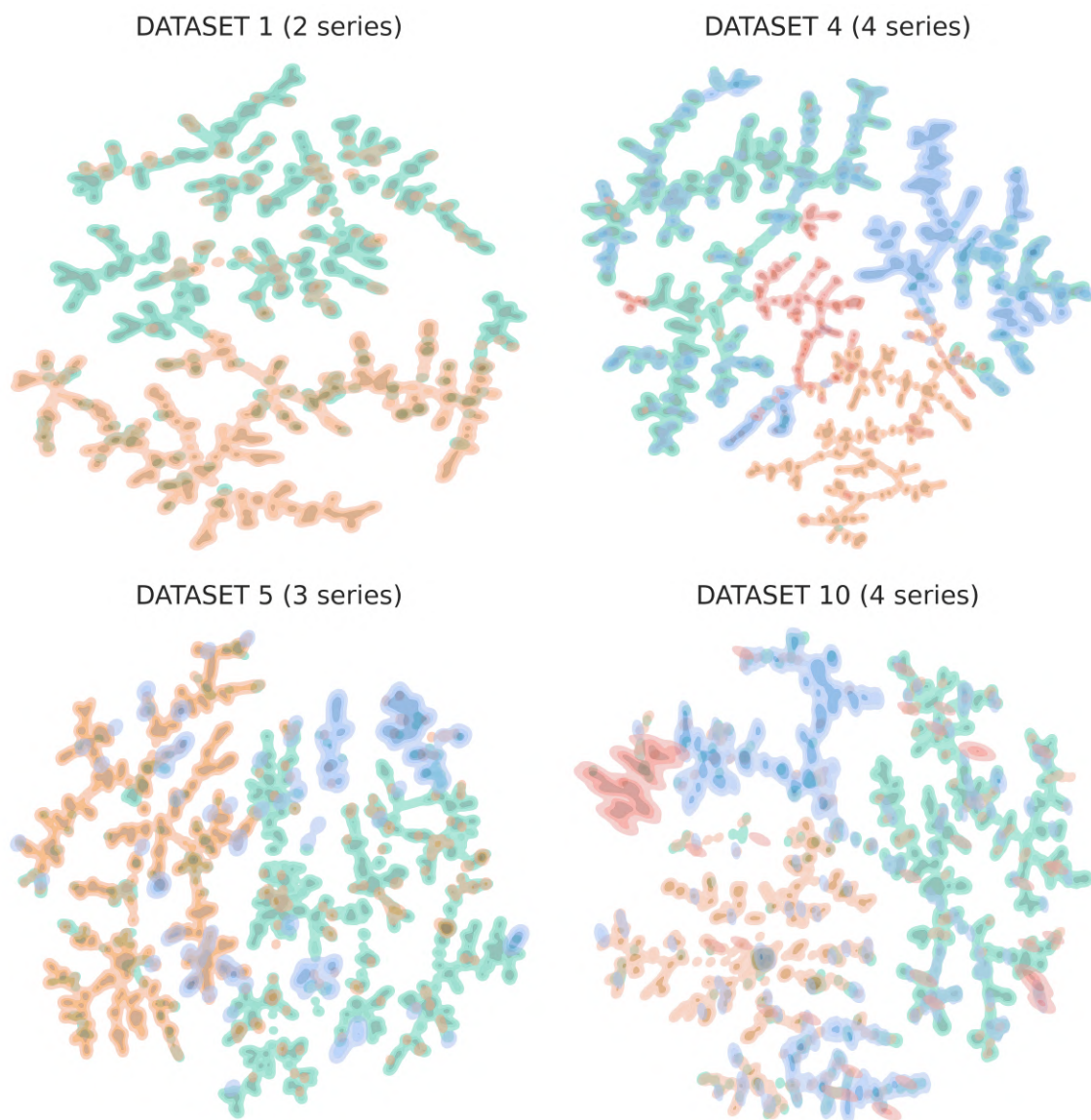


Figure 4.9: Overview of the distribution of the main chemical series for the datasets 1, 4, 5 and 10. Each color is associated to one chemical series within a project.

exploitation: Greediverse (parameterized with $\lambda = 10$ and $t = 0.75$), Desired Spread, and Desired Coverage. The figures show that these strategies efficiently explore a large part of the chemical space while successfully identifying most top compounds.

4.3 Insights from individual analysis

In this section, we will analyze results on all datasets using some of the quantitative metrics introduced in Chapter 3 (sections 3.4.3 page 116 and 3.4.2 page 114).

4.3.1 Exploration

We begin our analysis by examining the Neighborhood Coverage (NC) exploration metric, using a fixed distance threshold of 0.3. Supplementary analyses employing the parameter-free variant Neighborhood Coverage Area Under the Curve (NCAUC) and Internal Diversity are presented in Appendix C.

Figure 4.12 illustrates the evolution of NC across iterations of 14 acquisition functions across all datasets. The results consistently reveal similar patterns and rankings among strategies across all datasets. Notably, methods explicitly designed to prioritize exploration, such as Spread, Coverage, and Uncertainty, consistently achieve the highest NC scores. Hybrid strategies including Desired Spread, Desired Coverage, and occasionally U/D Harmonic, although not solely exploration-oriented, achieve comparable levels of exploratory performance.

Random selection, which mimics realistic project conditions by uniformly sampling chemical space at similar batch sizes, frequently emerges among the top-performing strategies. This outcome aligns with expectations, as uniform sampling naturally covers diverse regions of the chemical space.

In contrast, the Desirability-based strategy consistently underperforms relative to random selection. This reduced performance occurs because this strategy predominantly identifies promising compounds within regions already represented in its training set, highlighting its limited capacity for extrapolation.

Strategies combining Desirability with applicability domain and heuristics based on similarity to previously selected molecules form the lowest-performing group. These methods exhibit significantly reduced exploratory capabilities, remaining predominantly confined to the initial chemical space.

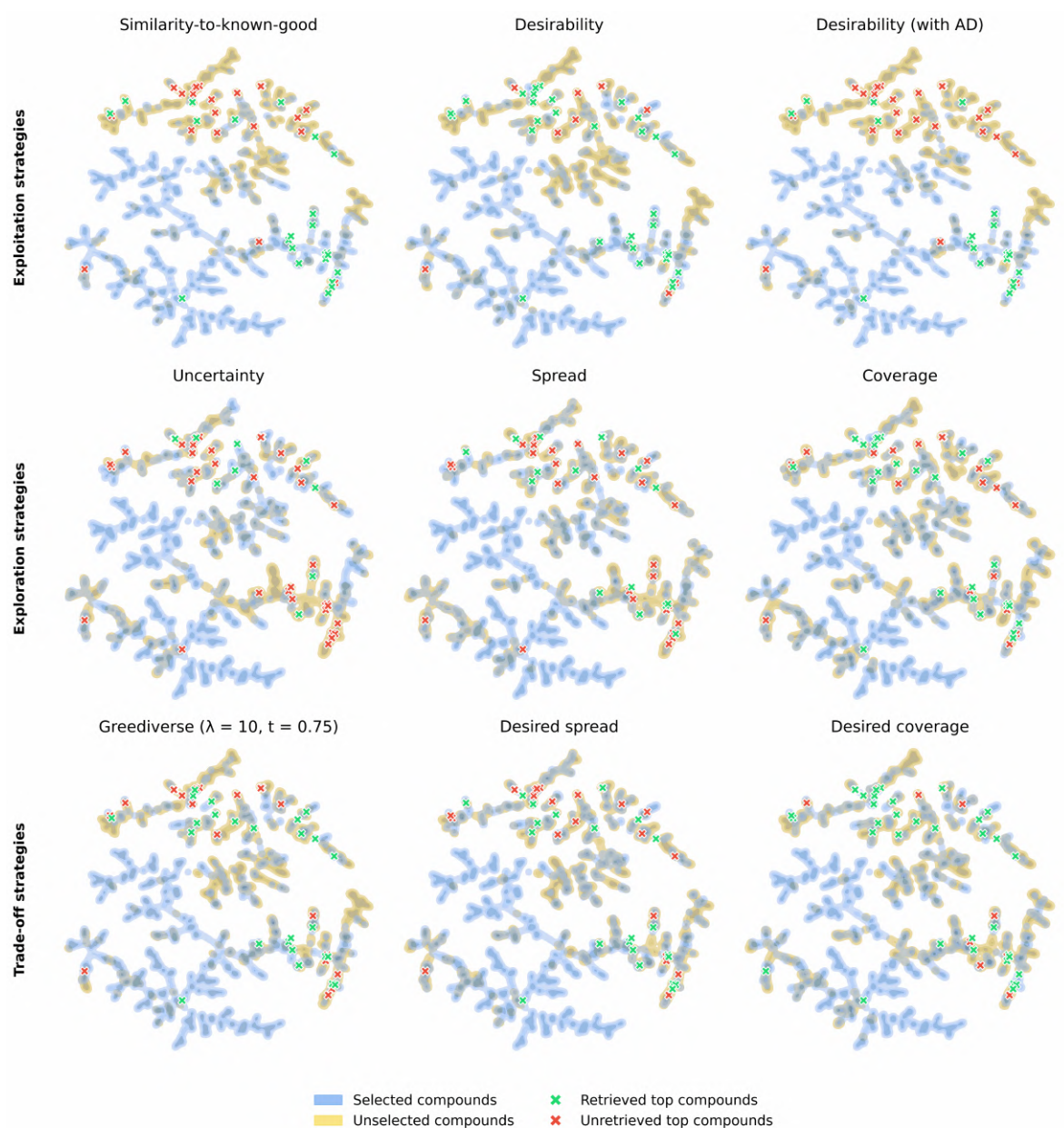


Figure 4.10: TMAPs of the final-state of simulations performed on dataset 1, sorted by the type of selection strategy used. Top row: strategies designed for exploitation. Middle row: strategies designed for exploration. Bottom row: trade-off strategies.

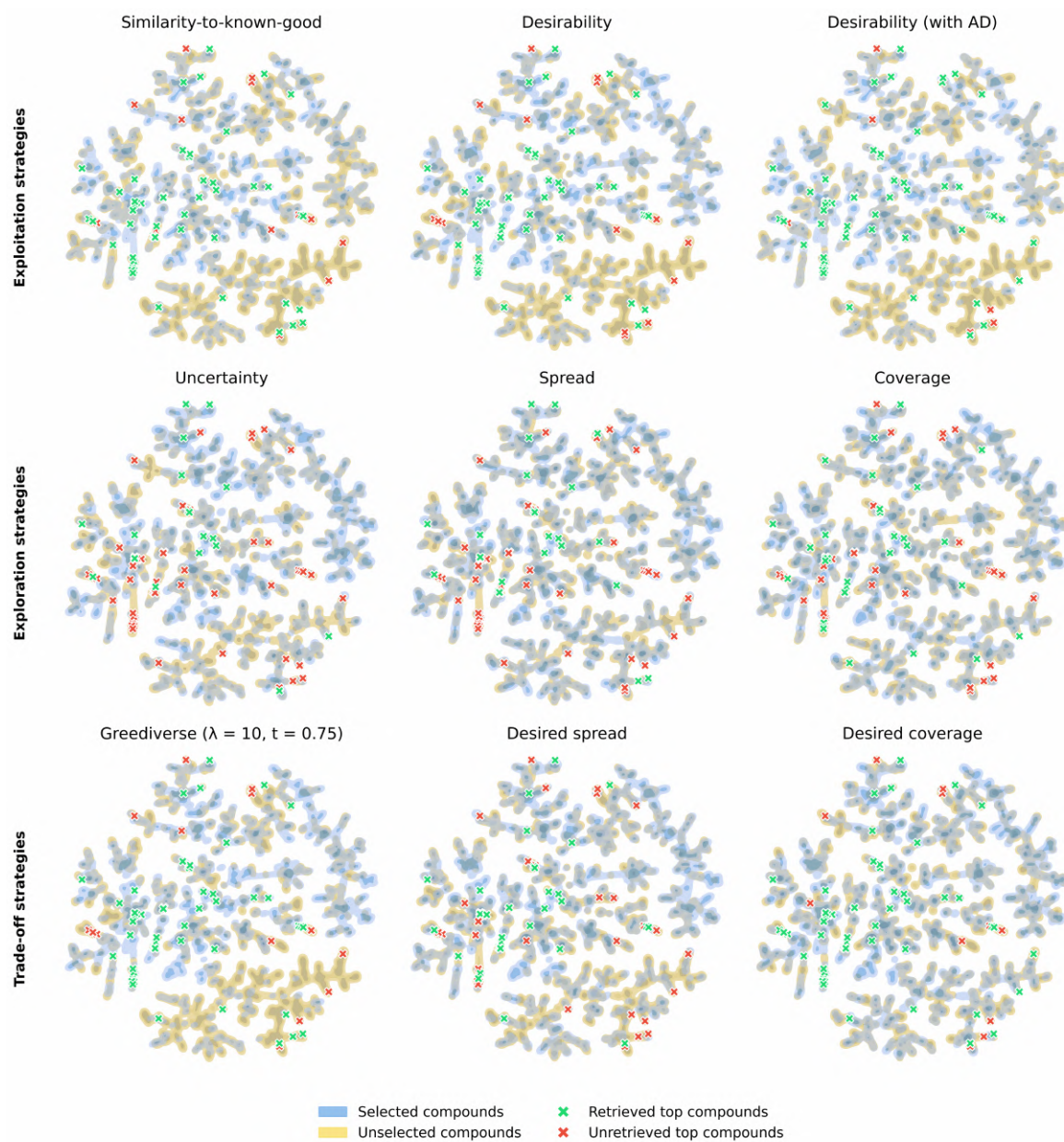


Figure 4.11: TMAPs of the final-state of simulations performed on dataset 11, sorted by the type of selection strategy used. Top row: strategies designed for exploitation. Middle row: strategies designed for exploration. Bottom row: trade-off strategies.

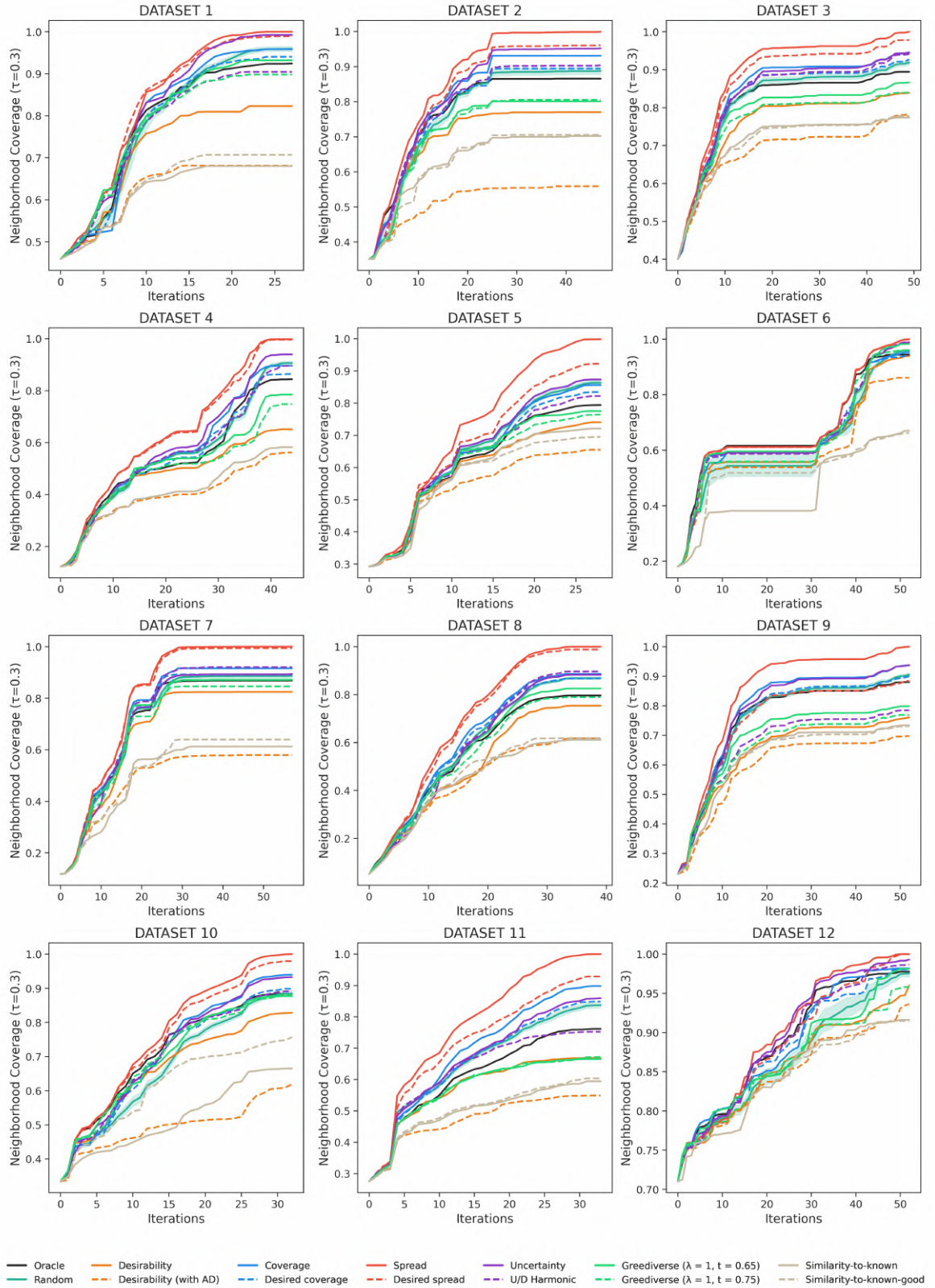


Figure 4.12: Neighborhood Coverage (τ of 0.3) for various selection strategies over iterations for each dataset.

4.3.2 Exploitation

Similar to our exploration analysis, we evaluate the effectiveness of strategies in identifying good compounds over time. This is illustrated in Figure 5.17 using the PTMR metric, described in section 3.4.2 (page 114). As outlined previously in Figure 4.7 (page 141), multiple definitions of good molecules exist. Here, we consider good molecules as those falling within the Top DScore 5% category. This criterion strikes a balance by being restrictive enough to identify high-quality compounds yet inclusive enough to ensure statistical robustness. Corresponding figures for the more restrictive Top DScore 3% and Top DScore 1% categories are presented in Appendix C.

Monitoring the retrieval of compounds in the Top DScore 5% category indicates that certain strategies consistently achieve strong performance, successfully identifying between 70% and 100% of these top compounds. This is especially noteworthy considering these strategies typically sample only about half of the available project compounds, due to a batch size of 0.5. High-performing strategies include Desirability as well as hybrid methods such as Greediverse, Desired Spread, and Desired Coverage. These quantitative findings align with prior qualitative assessments obtained from TMAP visualizations on Figures 4.10 and 4.11. Desirability (with AD) and Similarity-to-known-good metrics also generally exhibit strong exploitation performance, though their results are more variable. Notably, these strategies occasionally underperform significantly, as observed in datasets 2, 6, and 10. This variability could stem from the substantial differences between top compounds and the initial chemical space, limiting the effectiveness of these strategies in exploring beyond initial regions.

4.3.3 Exploitation/exploration trade-off

Having previously examined exploitation and exploration separately, we now jointly analyze these aspects to identify strategies that offer the best trade-off between these two objectives. Although the primary goal of lead optimization is to identify high-quality molecules, maintaining exploratory capacity remains critical. Exploration ensures that potentially superior compounds are not overlooked and provides viable alternatives should the selected clinical candidate fail, as previously discussed.

Figures 4.14 illustrate the trade-offs between the PTMR Top DScore 5% metric and Neighborhood Coverage ($\tau = 0.3$) at the end of simulations for all strategies. This figure makes it easier to confirm previous findings, such as:

- Purely exploratory methods, including Spread, Coverage, Uncertainty, as well as clustering-based methods such as k-means and k-medoids, exhibit poor exploitation performance.
- Desirability serves as a strong baseline for exploitation but performs poorly in terms of ex-

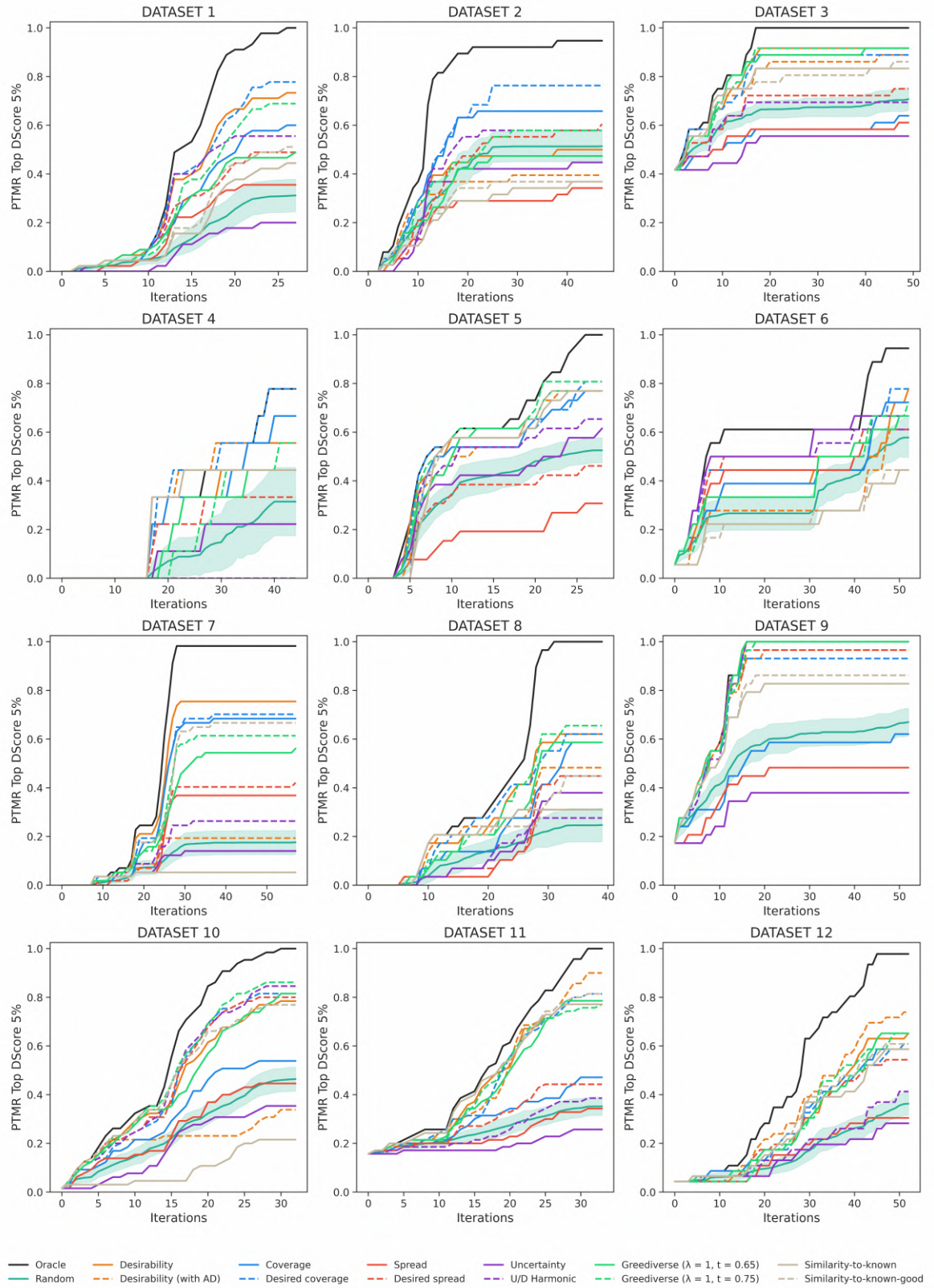


Figure 4.13: PTMR Top DScore 5% for various selection strategies over iterations for each dataset.

ploration. However, as stated in the limitations in Chapter 3, the exploitation can be overestimated for strategies that do not explore well.

- Using an applicability domain with the Desirability strategy consistently reduces exploratory performance and may also negatively impact exploitation, a trend similarly observed with heuristic approaches such as Similarity-to-known-good.
- Desired Coverage emerges as a robust strategy, often matching or surpassing Desirability in exploitation performance while also equaling or exceeding Random selection in exploration capabilities.

4.4 Aggregating information across datasets

Thus far, each dataset has been analyzed individually, enabling a detailed understanding of their specific dynamics and characteristics. We now aggregate these findings across all datasets to derive broader conclusions. Specifically, we examine both the intrinsic performance of each strategy, as measured by various exploration and exploitation metrics, and their relative performance compared to the Random baseline, which normalizes the results across datasets.

Currently, this pipeline has been applied to 12 datasets. Further analysis is planned with additional datasets to support more robust conclusions and enable formal statistical testing, similar to the tests described in Chapter 2 for benchmarking molecular descriptors. Although 12 datasets do not yet provide sufficient data points for rigorous statistical evaluation, preliminary insights can still be drawn.

4.4.1 Exploitation

Figure 4.15 illustrates the distribution of PTMR values for the Top DScore 5%, Top DScore 3%, and Top DScore 1% categories across all datasets, with strategies ranked according to their average PTMR performance on the Top DScore 5% category. The Random baseline is indicated in purple for comparison. Notably, the highest-performing strategies consistently achieve an average PTMR of approximately 0.8, irrespective of the restrictiveness of the top category considered.

Desirability, previously identified as a strong exploitation-focused strategy in individual dataset analyses, maintains its effectiveness by ranking among the top strategies. However, several strategies, particularly Desired Coverage, surpass Desirability, validating earlier observations. Additionally, Desired Coverage demonstrates notable robustness, exhibiting the narrowest interquartile range among all strategies analyzed. This limited variability is similarly evident in the purely exploratory Coverage strategy, despite its comparatively weaker exploitation performance.

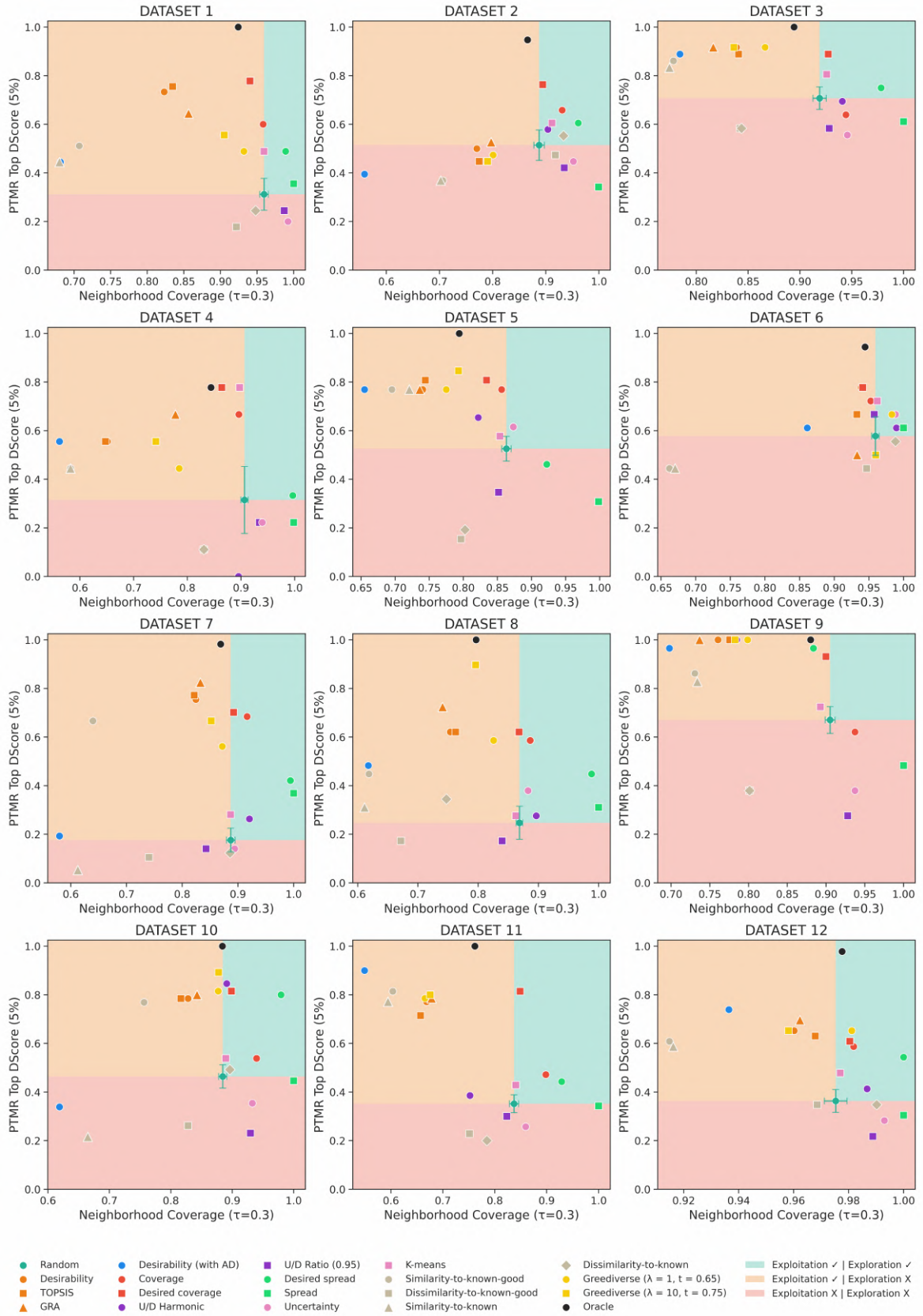


Figure 4.14: Final overview of PTMR Top DScore 5% vs Neighborhood Coverage ($\tau = 0.3$) for all datasets and all selections strategies.

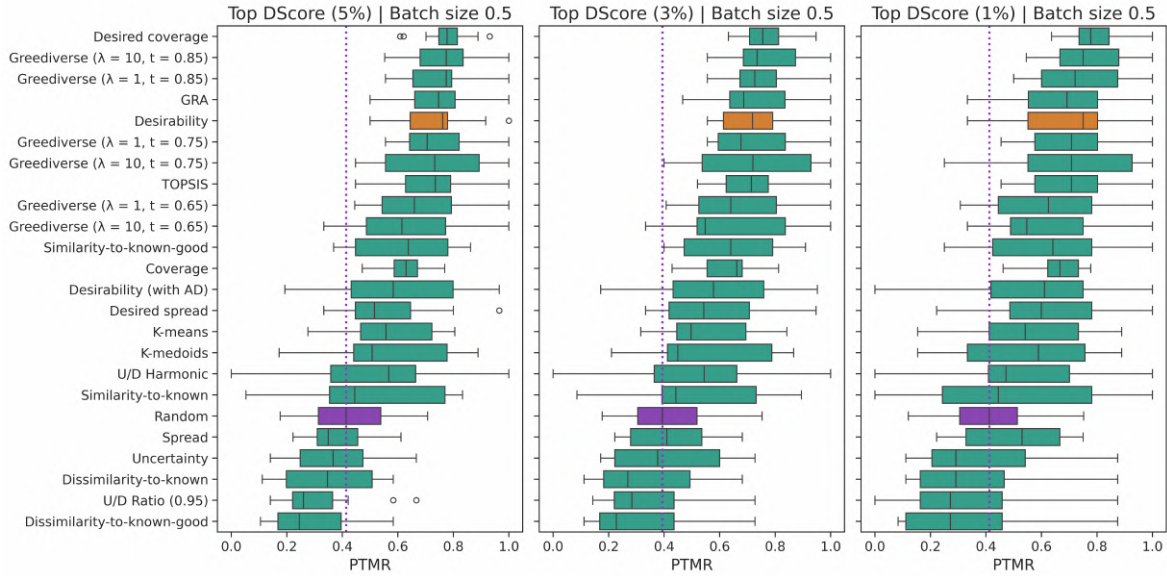


Figure 4.15: Box-plots of PTMR for the Top-DScore 5%, 3% and 1% subsets across all datasets. Strategies are ordered by their mean PTMR on the Top DScore 5%; the Random baseline is shown in purple and Desirability in orange.

MCDA methods, specifically GRA and TOPSIS, yield performances similar to Desirability. Greediverse strategies also achieve comparable outcomes, with a clear trend observed: higher similarity thresholds (t) correlate with improved average performance and reduced variability.

In contrast, Desirability combined with an applicability domain, along with Similarity-to-known-good strategies, exhibit highly inconsistent results. While capable of excellent performance in some cases, these strategies frequently produce poor outcomes, leading to low average performance and large variability.

Finally, strategies primarily designed for exploration consistently display poor exploitation capabilities and high variability.

Figure 4.16 does not use the PTMR metric but the ratio, for each strategy, between its PTMR and the PTMR of the Random baseline, in order to harmonize results taking into account each dataset uniqueness. This representation is very practical because it allows you to quickly see how much better a strategy could have been than Random, which, as a reminder, is a proxy for the performance of the real project.

Table 4.3 summarizes the Ratio to Random (using the PTMR Top DScore 5%) for each strategy, presenting the minimum, average, maximum, and standard deviation across datasets. The results clearly indicate that Desired Coverage is among the top-performing strategies, exhibiting the highest minimum and average ratios, as well as one of the lowest standard deviations among leading strategies. While the average ratio is essential for evaluating overall performance, the minimum

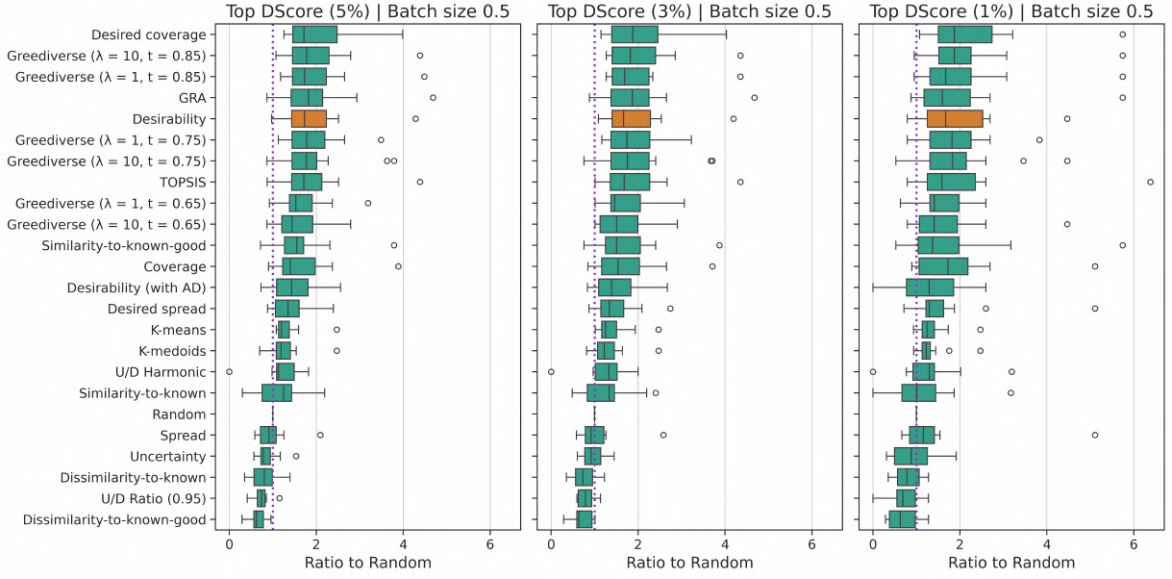


Figure 4.16: Box-plots of each strategy’s PTMR divided by the Random PTMR for the Top DScore 5%, 3% and 1%, highlighting the relative improvement achievable across datasets. The purple dashed line is the ratio of 1 corresponding to Random.

ratio is equally important as it reflects the robustness of the strategy in a "worst-case scenario". For Desired Coverage, this minimum value is 1.25, indicating consistent outperformance compared to the Random baseline across all datasets. Conversely, Desirability, despite a strong average ratio of 1.93, close to Desired Coverage’s 2.01, exhibits a minimum ratio of 0.97. This indicates that Desirability underperformed the Random baseline on at least one dataset. Therefore, an ideal strategy should maximize the average ratio across datasets while maintaining a high minimum performance. It is also important to highlight the performance of Greediverse strategies, which closely aligns with that of Desired Coverage in exploitative capability.

4.4.2 Exploration

Figure 4.17 shows the distribution across datasets of the Neighborhood Coverage ($\tau = 0.3$), Neighborhood Coverage AUC and Internal Diversity metrics for all acquisition functions. This figure highlights the strong baseline that Random is for exploration, conversely to exploitation. Here, very few strategies outperform it. Indeed, all the strategies designed for exploitation such as Desirability, GRA or TOPSIS are outperformed on all three metrics. Greediverse strategies, although introducing a trade-off with exploration, also underperform the baseline on average. Fully exploratory strategies such as Coverage, Spread and Uncertainty slightly outperform Random while also having a narrower distribution, indicating that they are more robust exploration strategies. The same observation can be made for Desired Spread, especially when looking at NC and NCAUC. For Desired Coverage, it matches the performance of Random, similarly to clustering methods that are k-means and k-

Strategy	PTMR Top DScore 5% Ratio to Random			
	min	average	max	std
Desired coverage	1.25	2.01	3.98	0.78
Greedydiverse ($\lambda = 10$, $t = 0.85$)	1.07	2.00	4.38	0.89
GRA	0.86	1.98	4.68	1.02
Greedydiverse ($\lambda = 1$, $t = 0.85$)	1.17	1.98	4.48	0.90
Desirability	0.97	1.93	4.28	0.86
Greedydiverse ($\lambda = 10$, $t = 0.75$)	0.86	1.92	3.78	0.92
TOPSIS	0.87	1.90	4.38	0.91
Greedydiverse ($\lambda = 1$, $t = 0.75$)	1.12	1.88	3.48	0.67
Greedydiverse ($\lambda = 1$, $t = 0.65$)	0.92	1.72	3.18	0.61
Coverage	0.90	1.68	3.88	0.82
Similarity-to-known-good	0.71	1.64	3.78	0.80
Greedydiverse ($\lambda = 10$, $t = 0.65$)	0.86	1.60	2.79	0.61
Desirability (with AD)	0.72	1.46	2.55	0.54
Desired spread	0.87	1.41	2.39	0.42
K-means	1.08	1.34	2.47	0.39
K-medoids	0.69	1.29	2.47	0.43
U/D Harmonic	0.00	1.19	1.82	0.47
Similarity-to-known	0.29	1.16	2.19	0.52
Spread	0.58	0.98	2.09	0.40
Uncertainty	0.56	0.87	1.53	0.27
Dissimilarity-to-known	0.35	0.80	1.39	0.31
U/D Ratio (0.95)	0.41	0.73	1.15	0.19
Dissimilarity-to-known-good	0.29	0.64	0.95	0.20

Table 4.3: Minimum, mean, maximum and standard deviation of the PTMR Top DScore 5% Ratio to Random across datasets.

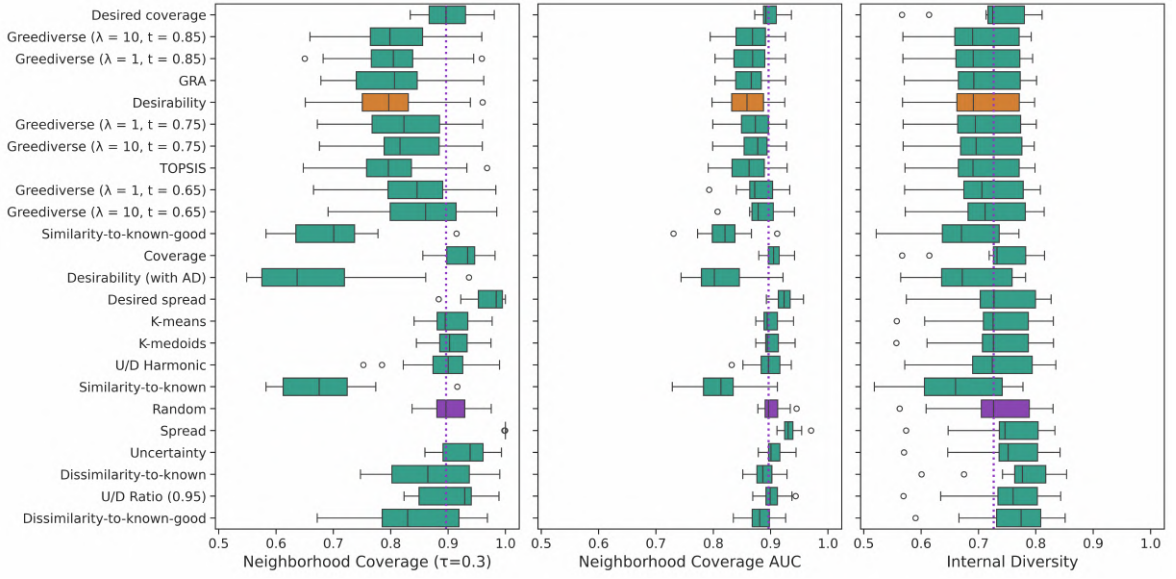


Figure 4.17: Box-plots of Neighborhood Coverage ($\tau = 0.3$), Neighborhood Coverage AUC and Internal Diversity across all datasets. Strategies are ordered by using the same order as the one in the previous section for exploitation, using the average PTMR on the Top DScore 5%.

medoids.

When looking at the Ratio to Random based on the Neighborhood Coverage ($\tau=0.3$) metric on Figure 4.18, the same observations can be made. Although Neighborhood Coverage ($\tau=0.3$) is the metric that displayed the biggest gaps to Random, the maximum average ratio is 1.2 for the Spread strategy, as described in Table 4.4, that is relatively small and shows the strong baseline that Random is for exploration. The last strategy in this table, based on the average, is the Desirability with applicability domain, highlighting the impact that an applicability domain, as it is commonly defined in industry, can have on exploration.

4.4.3 Exploitation/exploration trade-off

Finally, to quickly assess the trade-off between exploitation and exploration, one can plot the respective metrics against each other, averaging performance across all datasets. Figure 4.19 display the ratios to Random, considering all three exploration metrics (rows) and the three top-tier exploitation categories (Top DScore 5%, 3%, and 1%, columns). These figures synthesize previously discussed findings, confirming consistent performance patterns across the exploration and exploitation dimensions. Specifically, they illustrate how strategies that excel in exploitation, such as Desirability, GRA, TOPSIS, and Greediverse, typically underperform in exploration. Conversely, strategies optimized for exploration—Spread, Coverage, Uncertainty, U/D Ratio, and k-means—generally exhibit weaker exploitation capabilities. Desired Coverage emerges as the standout strategy, demonstrating strong exploitation performance (across all top categories) while maintaining exploration effective-

Strategy	NC ($\tau=0.3$) Ratio to Random			
	min	average	max	std
Spread	1.02	1.10	1.19	0.05
Desired spread	0.97	1.07	1.13	0.04
Uncertainty	1.00	1.03	1.07	0.01
Coverage	0.98	1.02	1.07	0.02
U/D Ratio (0.95)	0.95	1.00	1.05	0.03
K-medoids	0.98	1.00	1.01	0.00
K-means	0.98	1.00	1.02	0.01
Random	1.00	1.00	1.00	0.00
Desired coverage	0.95	0.99	1.01	0.01
U/D Harmonic	0.86	0.98	1.03	0.05
Dissimilarity-to-known	0.86	0.96	1.05	0.06
Oracle	0.90	0.96	1.00	0.03
Greedyverse ($\lambda = 10, t = 0.65$)	0.82	0.94	1.02	0.06
Greedyverse ($\lambda = 1, t = 0.65$)	0.79	0.93	1.02	0.06
Dissimilarity-to-known-good	0.77	0.92	1.03	0.07
Greedyverse ($\lambda = 10, t = 0.75$)	0.80	0.91	1.00	0.06
Greedyverse ($\lambda = 1, t = 0.75$)	0.80	0.91	1.00	0.06
GRA	0.81	0.89	0.98	0.05
Greedyverse ($\lambda = 10, t = 0.85$)	0.76	0.89	0.98	0.06
Greedyverse ($\lambda = 1, t = 0.85$)	0.71	0.88	0.98	0.07
TOPSIS	0.71	0.88	0.99	0.07
Desirability	0.71	0.87	0.98	0.07
Similarity-to-known-good	0.64	0.77	0.93	0.08
Similarity-to-known	0.64	0.76	0.93	0.08
Desirability (with AD)	0.61	0.74	0.96	0.10

Table 4.4: Minimum, mean, maximum and standard deviation of the Neighborhood Coverage ($\tau=0.3$) Ratio to Random across datasets.

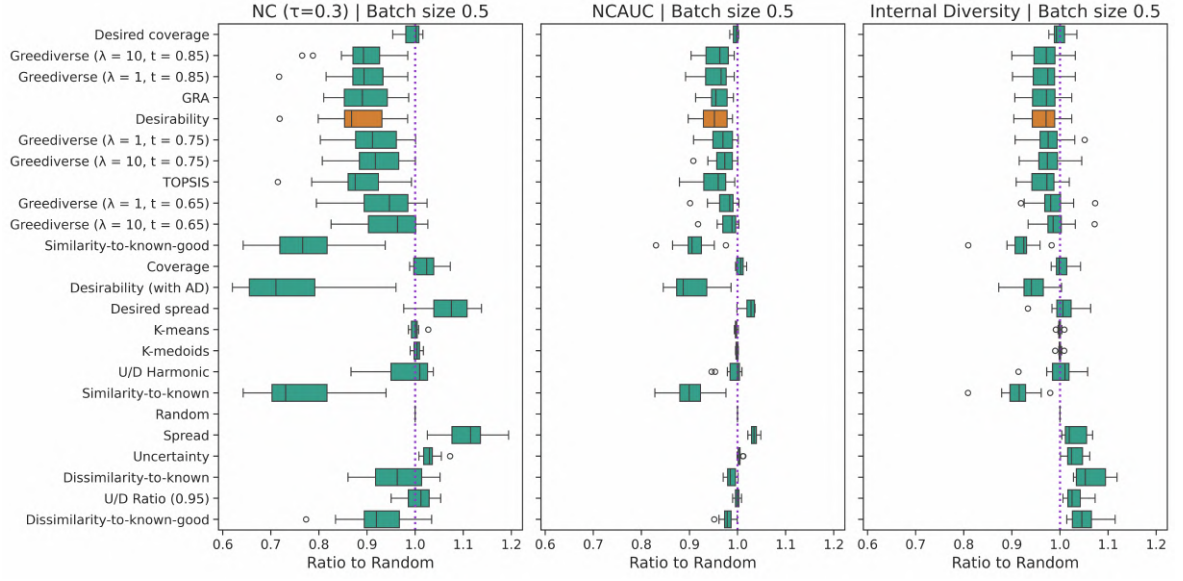


Figure 4.18: Box-plots of Neighborhood Coverage ($\tau = 0.3$), Neighborhood Coverage AUC and Internal Diversity Ratio to Random across all datasets. Strategies are ordered by using the same order as the one in the previous section for exploitation, using the average PTMR on the Top DScore 5%.

ness comparable to the Random baseline (across all exploration metrics).

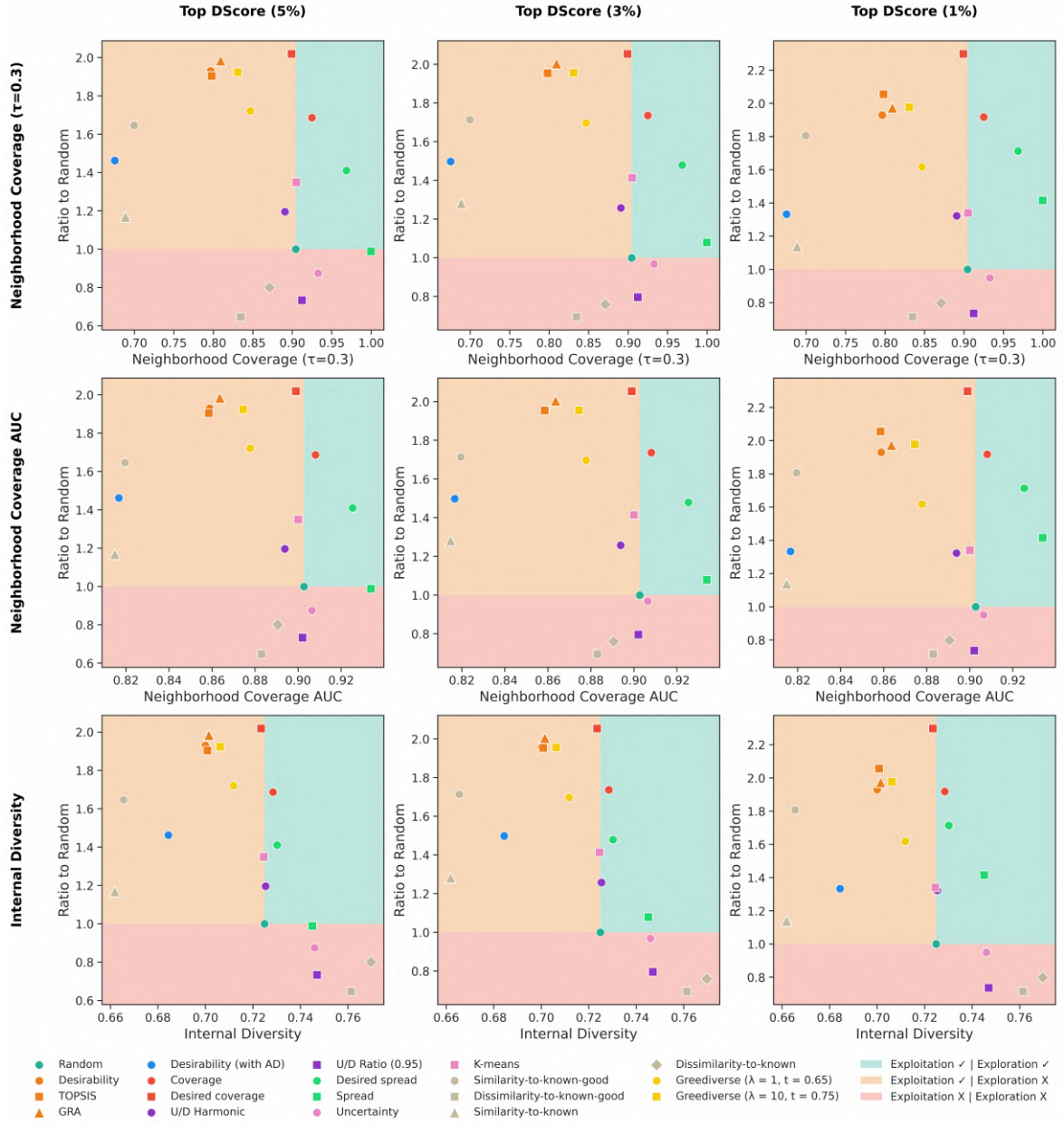


Figure 4.19: Exploitation/exploration trade-off (Ratio to Random) averaged over all datasets for various top categories (columns) and exploration metrics (rows) at the end of simulations.

4.5 Discussion and perspectives

4.5.1 Exploring to better exploit

Exploration can directly influence exploitation. In the early stages of a lead optimization campaign, the chemical space is relatively unexplored, as illustrated in Figure 4.8 (page 144). At this point, immediate exploitation may not produce optimal outcomes, and research teams typically emphasize broader exploration. This approach is exemplified by the Random strategy, which provides a strong baseline for exploration.

Several findings presented above reinforce this idea. First, inadequate exploration can negatively impact exploitation performance, as demonstrated by the relatively poor exploitation results of Desirability with AD compared to standard Desirability, highlighting that constrained exploration limits exploitative effectiveness. Similarly, robust exploration can significantly enhance exploitation outcomes. For instance, the Desired Coverage strategy achieves exploration performance similar to Random and still demonstrates superior exploitation results compared to purely exploitative strategies such as Desirability, TOPSIS, and GRA, even though these strategies might benefit from the overestimation artifact discussed previously section 3.6.2 (page 131). These observations support that thorough exploration can benefit exploitation.

4.5.2 Rationalizing lead optimization

The preliminary analysis of a dozen datasets supports the hypothesis that rationalizing the DMTA cycle through *in silico* methods can provide significant advantages. Systematic approaches employing computational tools consistently surpass the Random baseline, which approximates real-world project performance. For example, strategies like Desired Coverage always outperform the Random baseline in exploitation while matching the exploration. Moreover, many methods can recover from 70% to 100% of the project’s best compounds while selecting only about half of all the compounds. This indicates that incorporating systematic, computationally driven strategies into lead optimization could significantly reduce both costs and time while maintaining or even enhancing project outcomes.

4.5.3 Perspectives

4.5.3.1 Dynamic strategies

The preliminary evaluation of the toolkit on legacy projects employed selection strategies that remained constant throughout the simulation period. Our results indicate that strategies effectively

balancing exploration and exploitation consistently achieve superior overall performance. However, the balance between exploration and exploitation within each individual strategy remained static. Introducing a dynamic adjustment to this balance as the project progresses could potentially enhance performance outcomes. For instance, combining an exploration-centric approach during the early stages of a project with a progressively increasing emphasis on exploitation as the project advances and the need for exploration diminishes may lead to improved results. Another approach would be to initiate exploration phases upon specific triggers, such as the emergence of a new chemical series.

Extending this concept further, it is plausible that the selection of an appropriate strategy could itself be automated through algorithmic decision-making, framing the selection as a reinforcement learning problem analogous to a multi-armed bandit approach [272]. By leveraging extensive data from simulations, reinforcement learning algorithms could be trained to identify the optimal selection strategy based on the current simulation state. In such a context, the reward function for the algorithm could be explicitly designed to encourage the maximization of both exploration and exploitation, utilizing the metrics previously defined.

4.5.3.2 Varying the batch size

In the presented results, simulations were conducted using a batch size of 0.5. However, investigating the impact of varying the batch size could yield further insights. For instance, decreasing the batch size could help determine the minimal threshold at which comparable performance can still be maintained. Conversely, increasing the batch size could reveal whether it is possible to consistently identify 100% of desirable molecules without the necessity of selecting the entire compound set.

4.5.3.3 Going beyond biological and chemical properties

The blueprints employed in the simulations currently consider only biological endpoints, such as potency and pharmacokinetic properties. However, these could be extended to incorporate additional criteria, such as synthetic accessibility, which can be estimated using various scoring methods [170, 171, 172]. This criterion, however, is not applicable in a retrospective context, since the molecules have already been synthesized. Another relevant criterion is cost, which may be assessed either by direct prediction methods [273] or indirectly by predicting synthetic routes using computational tools like ASKCOS [168] or AiZynthFinder [274], subsequently estimating cost based on the cumulative price of the required starting materials [275].

4.5.3.4 Chemical space exploration monitoring activity cliffs

In the work presented in this manuscript, the defined chemical space could be categorized as "structural-based" (caution, nothing to do with structure-based drug design), as it is characterized primarily by structural similarities and differences among compounds. However, this structural definition could be complemented by alternative definitions, such as "activity-based" or "property-based" chemical spaces, which directly relate to the analysis of structure-activity and structure-property relationships (SAR/SPR). Preliminary investigations on this topic, not detailed in the current manuscript, included the monitoring of activity cliffs during simulations to determine which strategies preferentially select compound pairs exhibiting significant activity (or property) cliffs. Rather than penalizing these cliffs as noise, targeted strategies could deliberately exploit them by prioritizing structurally similar yet contrasting analogues. This approach may enhance understanding of steep SAR and refine local predictive models.

Chapter 5

Conclusions and perspectives

This thesis explored the development and application of computational methodologies aimed at rationalizing and improving the lead optimization phase in drug discovery. Through extensive methodological development, rigorous benchmarking on industrial datasets, and detailed analyses, we gained valuable insights into predictive modeling and compound prioritization strategies essential to this critical stage of drug discovery.

5.1 Predictive modeling in a shifting low-data regime

5.1.1 No fancy models but more data

The first major finding of this thesis, discussed in Chapter 1, is that increasing the complexity of predictive models, such as using deep neural networks, does not automatically solve the challenges posed by limited data in drug discovery. Experiments performed on real-world industrial datasets showed that advanced deep learning approaches did not consistently outperform simpler, traditional machine learning methods like random forests or gradient-boosted trees trained on classical molecular fingerprints. In fact, simpler algorithms frequently achieved performance equal to or better than neural networks, particularly when datasets were small. Combining multiple molecular fingerprints to form richer molecular descriptors, and then applying ensemble methods, resulted in better results for molecular property prediction.

These results highlight that, in sparse-data scenarios, the predictive capability of models depends more on the quality and quantity of available data than on algorithmic complexity. Moreover, simpler methods provide practical benefits such as fewer hyperparameters to adjust, faster training times enabling quicker iterative cycles, and enhanced interpretability, all crucial factors for informed decision-making in drug discovery.

5.1.2 Temporality is key

Beyond model selection, our research highlights the critical importance of temporal dynamics when evaluating and deploying predictive models. At the conclusion of Chapter 1, we compared the time-split validation method to a conventional cross-validation approach. We observed a significant decline in predictive performance when models trained on earlier project data were evaluated against chronologically later compounds. This performance degradation illustrates the dynamic nature of chemical space: as drug discovery projects progress, research teams continually explore novel scaffolds and chemotypes, causing underlying data distributions to evolve over time. Evaluation methods that ignore this temporal shift thus yield overly optimistic assessments of a model’s true predictive capabilities. Despite this, the vast majority of studies proposing or benchmarking QSAR/QSPR methodologies still do not incorporate temporal validation splits, mainly because of the lack of such data outside of industry.

5.2 Rationalizing lead optimization

While previous sections focused on modeling techniques, the main objective of this thesis was to propose rational strategies for lead optimization, the iterative refinement of chemical leads to enhance their properties. Specifically, we aimed to integrate more systematic and data-driven decision-making into a stage of drug discovery traditionally dominated by human intuition. Our contributions include a comprehensive molecular prioritization toolkit designed for compound selection, as well as a novel retrospective simulation methodology to evaluate how various selection strategies would have performed in practice. Collectively, these advances are intended to help project teams efficiently and confidently identify high-quality drug candidates.

5.2.1 Molecular prioritization toolkit

We developed a decision-support toolkit for molecular prioritization that formalizes the compound selection step in lead optimization, presented in Chapter 3. This toolkit integrates a broad range of compound acquisition functions (i.e. selection strategies), combining well-established methods from the literature with several novel approaches introduced in this work. The core idea was to gather a diverse toolset of rational criteria that a team might use to decide which compounds to synthesize or advance at each round, and implement them in a flexible, computational framework. By doing so, we move toward a more transparent and testable compound selection process.

Furthermore, the modular nature of our toolkit provides a solid foundation for continuous improvement and customization. Future enhancements could include integrating additional selec-

tion criteria, such as predictors of synthetic feasibility or cost, to refine prioritization further. Additionally, the toolkit could incorporate adaptive algorithms that dynamically adjust their exploration/exploitation balance based on real-time feedback. For instance, it could increase exploration when initial selections yield overly similar compounds or when a new chemical series is introduced.

5.2.2 Retrospective simulation methodology

To rigorously evaluate different selection strategies, we developed a retrospective simulation framework described in Chapter 3. This framework simulates historical drug discovery projects to answer a critical question: "How would outcomes have changed if we had used a specific prioritization strategy during a past project?" We utilized historical data from multiple lead optimization projects and inserted our selection algorithms into their original timelines. Specifically, the retrospective simulation reveals experimental data to the algorithms exactly as it became available in the real projects. At each iteration, the algorithm selects a batch of compounds either from those newly available or from those previously not chosen. This simulation continues until it reaches the actual project's end date. By comparing simulated outcomes, such as the speed at which high-quality leads are identified, to the original project's outcomes or to those from other strategies, we can quantitatively assess the effectiveness of each prioritization strategy under realistic conditions.

This retrospective approach provides a practical balance between purely virtual benchmarks and expensive, prospective experimental studies. Purely virtual benchmarks, commonly found in existing literature, typically optimize only one endpoint and rely on unrealistic assumptions, neglecting the inherent complexity and variability of real experimental conditions. Conversely, prospective comparative studies would require parallel execution of multiple experimental strategies in a laboratory setting, which is both costly and time-intensive. Our retrospective simulations address these limitations by grounding evaluations in real experimental outcomes. This method achieves realism without incurring additional laboratory costs. By accurately replicating each project's timeline and workflow, including constraints and the sequential availability of SAR data, our method effectively allows for "rewinding" a project and exploring alternate decisions within a controlled virtual environment.

In Chapter 4, by integrating the prioritization toolkit with this retrospective framework, we demonstrated significant improvements in lead optimization efficiency through a systematic, data-driven selection process explicitly balancing exploration and exploitation. The simulations revealed that specific selection strategies consistently outperformed the approximated historical performance of actual projects in selecting optimal molecules. This enhanced selection capability can potentially reduce the total number of compounds and cycles required, thereby accelerating progress toward clinical candidate identification.

5.2.3 Translating virtual insights to practice

Through these simulations, we demonstrated that a systematic, data-driven selection process significantly improved the efficiency of lead optimization. Our results showed that certain strategies consistently outperformed the approximated real project performance. By improving compound selection, these strategies can substantially reduce the number of compounds tested and the number of iterations required, potentially accelerating the discovery of clinical candidates.

A key next step will be to extend our analysis to additional datasets, allowing us to further validate the generalizability of our findings. We recognize, however, that retrospective simulations have inherent limitations, and true validation requires applying these strategies prospectively in active drug discovery projects. The promising results we have presented aim to encourage stakeholders to adopt and test these strategies in practical settings. Demonstrating that these approaches consistently speed up the identification of clinical candidates in practice could significantly advance the lead optimization process.

5.3 Challenges

This work also shed light on several broader challenges that must be addressed to fully realize AI's potential in drug discovery. These challenges are not just technical but also organizational and cultural.

5.3.1 Data management and data cleaning

A fundamental challenge in leveraging machine learning for drug discovery is effectively managing experimental data, a process often burdened by labor-intensive data cleaning requirements. Throughout our research, constructing reliable datasets from databases proved to be a substantial task. Assay data frequently exhibit noise, inconsistent formatting, missing values, and various errors, such as incorrectly recorded units, compound identification mistakes, or outliers resulting from experimental artifacts. These problems are prevalent both in publicly available datasets and proprietary pharmaceutical data. It is well established that poor data quality can undermine even the most sophisticated algorithms, illustrating the classic adage "garbage in, garbage out". Our own experiences underscored this principle: rigorous data curation and preprocessing were essential prerequisites for creating dependable predictive models. For instance, significant effort was devoted to standardizing compound representations, eliminating duplicate records, reconciling conflicting assay outcomes, and normalizing property value ranges. Although such meticulous "data janitorial" efforts were resource-intensive, they resulted in more robust and generalizable predictive models.

Nevertheless, data cleaning remains a notable barrier that can slow down project timelines and introduce errors if not systematically addressed. Ultimately, data quality improvement is a collective responsibility. While data scientists can refine cleaning algorithms and anomaly detection methods, the most impactful improvements originate from enhanced data generation and logging practices. These include employing robust experimental designs, adhering to standardized protocols, and implementing consistent data-sharing procedures.

5.3.2 Democratizing *in silico* tools

A significant challenge is making AI and *in silico* tools accessible to a broader range of drug discovery practitioners, not only specialists. Currently, the complexity of advanced algorithms and pipelines demands substantial expertise in machine learning and programming, creating bottlenecks as only a limited group of specialists, typically computational chemists, can effectively utilize these resources. To genuinely transform drug discovery, AI-driven tools must become user-friendly and broadly accessible. Achieving this involves enhancing user interfaces and tool integration. For instance, acquisition functions developed and tested using the tools presented in this work could be deployed as plugins within existing cheminformatics software, featuring simple drag-and-drop functionalities. Additionally, presenting results in clear visualizations, concise reports, or even natural-language summaries, rather than complicated spreadsheets or obscure model outputs, significantly enhances comprehension among non-specialists. In our own experience, visual aids like chemical space TMAPs have notably facilitated effective communication. Open-source, community-driven development is another key element. Open-source software lowers economic barriers and enables external improvements, such as user-friendly interfaces or broader format compatibility. Coupled with shared benchmark datasets and standardized evaluation methods, open-source initiatives contribute to a supportive ecosystem enabling groups with limited resources to engage, innovate, and advance collective knowledge.

5.3.3 Convincing with explainable AI

A third challenge is building trust in AI recommendations through explainability. In a domain like drug discovery, where decisions can involve large investments and significant risks (e.g. which lead series to advance or which compound to synthesize next), stakeholders understandably require justification for those decisions. An algorithm that simply outputs "Make compound X next" without any rationale is unlikely to convince a seasoned chemist or a project team to change course. Therefore, future work should not only strive to improve the predictive accuracy of computational models but also enhance the transparency of the underlying selection strategies. Developing advanced explainable AI methods tailored specifically for chemical decision-making is crucial. Such approaches

would ideally provide intuitive justifications rooted in chemical reasoning, enabling chemists to clearly understand why certain compounds are prioritized over others. This could involve creating visualization tools that illustrate the rationale behind each recommendation, showing relationships in chemical space, property trade-offs, and uncertainty estimates. Ultimately, these advancements will help foster greater trust and collaboration between AI-driven tools and human experts, ensuring that computational insights translate effectively into actionable strategies in drug discovery.

5.4 The role of industry

A notable call to action was recently articulated by Wognum et al. [183], who advocated for an industry-led, open-science initiative aimed at rigorously evaluating machine learning applications in drug discovery. They identified three essential pillars that resonate closely with the insights presented in this thesis and have substantial potential to advance the field:

- **Establishing shared benchmark datasets** that accurately represent common tasks in drug discovery, incorporating realistic complexities and noise. This initiative would likely involve collaborative curation of new, high-quality datasets that surpass what public organizations could compile alone.
- **Developing standardized guidelines for data curation and methodological evaluation**, covering critical aspects such as dataset splitting, model validation, performance metrics, and statistical comparison techniques. The establishment of these uniform standards would mitigate the prevalent issues of inconsistent or overly optimistic performance assessments, facilitating meaningful comparisons across studies.
- **Creating and distributing open-source software tools** that operationalize these best practices, thereby reducing barriers for researchers and practitioners to adopt state-of-the-art methods and evaluate their effectiveness objectively. This would ideally include reference algorithm implementations, comprehensive data processing pipelines, and accessible benchmarking platforms maintained by the scientific community.

Implementing these recommendations will require concerted effort and extensive collaboration, yet their realization promises a substantial improvement in the field’s ability to accurately assess scientific and methodological advancements. We strongly endorse this vision, as it directly addresses key limitations we encountered in our own research. For instance, the availability of multiple datasets capturing the temporal dynamics of molecular data from actual drug discovery projects would have significantly enhanced the reproducibility of our descriptor benchmarking and retro-

spective simulations. Nonetheless, we hope to actively contribute to this collective endeavor by releasing temporal, multi-objective datasets and providing open-source code.

5.5 Conclusion

In conclusion, the work presented in this thesis, "Rational Exploration/Exploitation Strategies for Drug Discovery in Low-Data Regimes," is another step towards a more data-driven and transparent approach to lead optimization. Our results demonstrated that, even when data is limited, it is possible to improve the selection of compounds by systematically balancing exploration and exploitation of the chemical space. Additionally, we highlighted the current limitations in methods and identified areas requiring improvements, including data quality, tool usability, and method explainability. These insights underscore the necessity for increased openness and collaborative efforts within the field.

We hope the insights and methodologies developed here will not only benefit Sanofi but also serve as foundational tools for the broader community. By openly sharing our findings, releasing datasets, making our code publicly available, and clearly outlining ongoing challenges, we encourage others to expand upon this research. Drug discovery is an inherently complex field, yet by rationally integrating human intuition with artificial intelligence, we can significantly enhance our ability to address modern drug discovery challenges. Ultimately, such collective progress serves the enduring goal of drug discovery: delivering new, effective therapeutics to the patients who need them.

Appendix

This appendix provides supplementary information intended to support and enrich the understanding of the content presented in the main body of the manuscript. The appendix is structured as follows:

- **Appendix A:** Chapter 2 - Benchmarking molecular descriptors for low-data regime modeling
- **Appendix B:** Chapter 3 - Molecular prioritization toolkit and retrospective simulations
- **Appendix C:** Chapter 4 - Insights from retrospective analysis of legacy projects

Appendix A

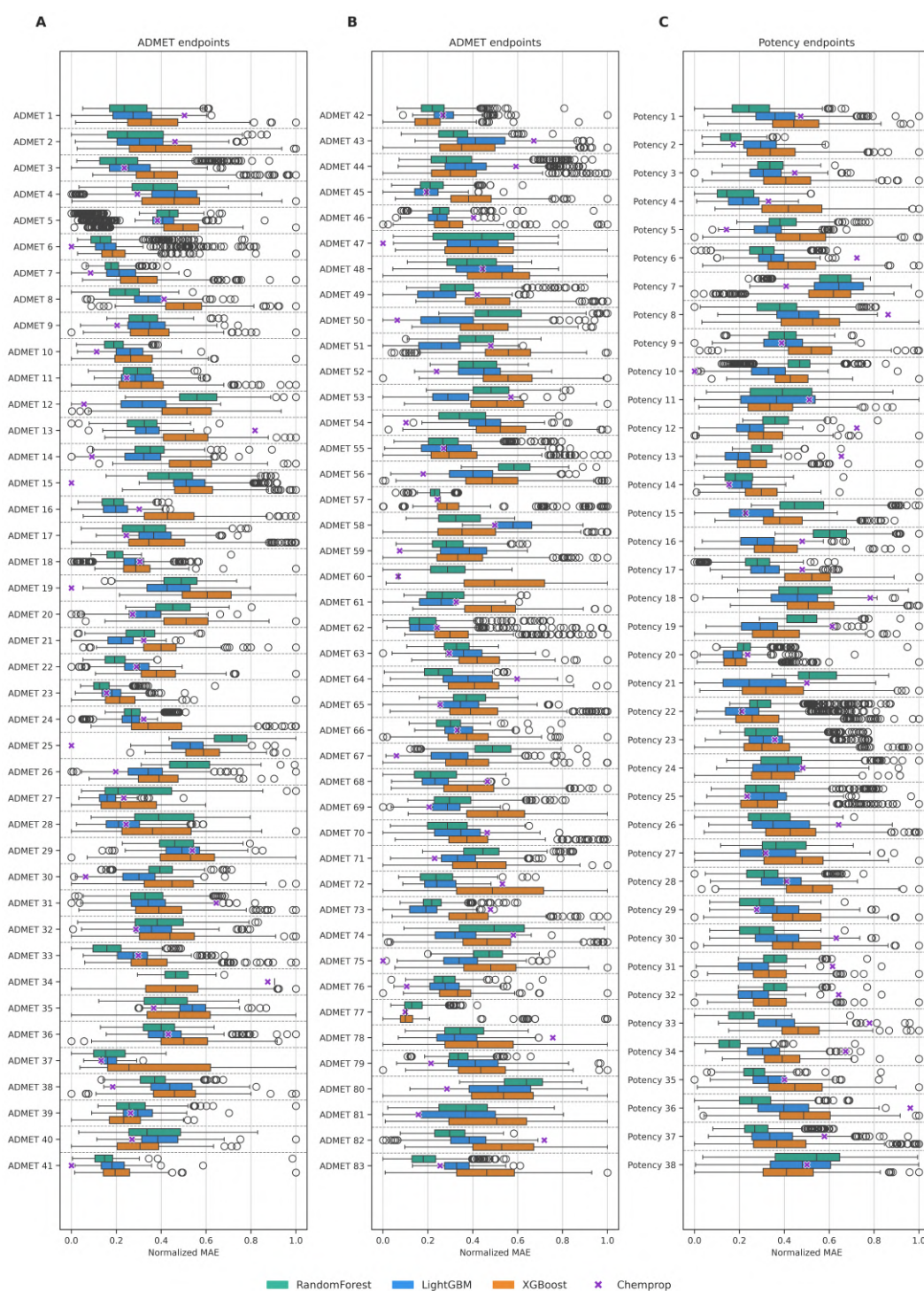


Figure 5.1: Distribution of nMAE scores for ADMET datasets (A, B) and potency datasets (C). Boxplots represent machine learning models trained on vector-based descriptors, while crosses represent Chemprop.

Appendix B

MMP-8

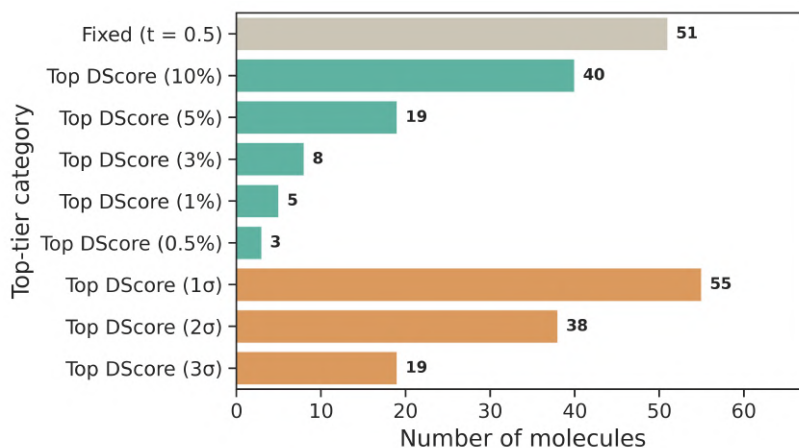


Figure 5.2: Number of molecules falling into each top-tier category for the MMP-8 dataset out of the 430 molecules. Categories are color-coded: brown indicates a fixed desirability threshold; orange indicates percentile-based thresholds; and green indicates standard deviation-based thresholds.

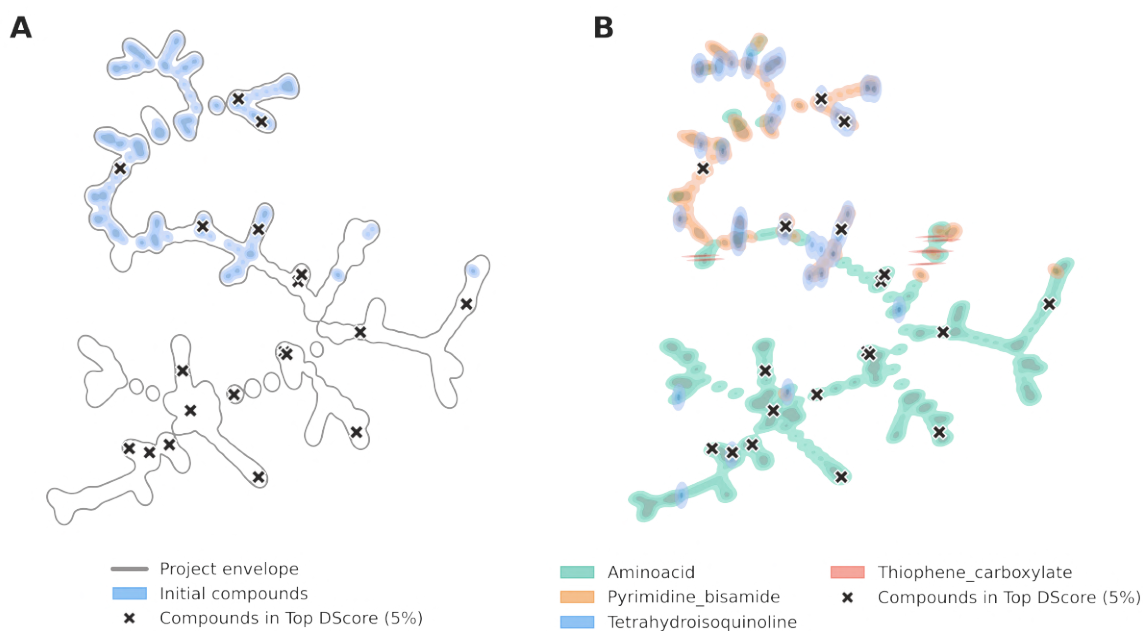


Figure 5.3: Overview of the chemical space of MMP-8 at the start of a simulation (A) and distributions of major chemical series (B).

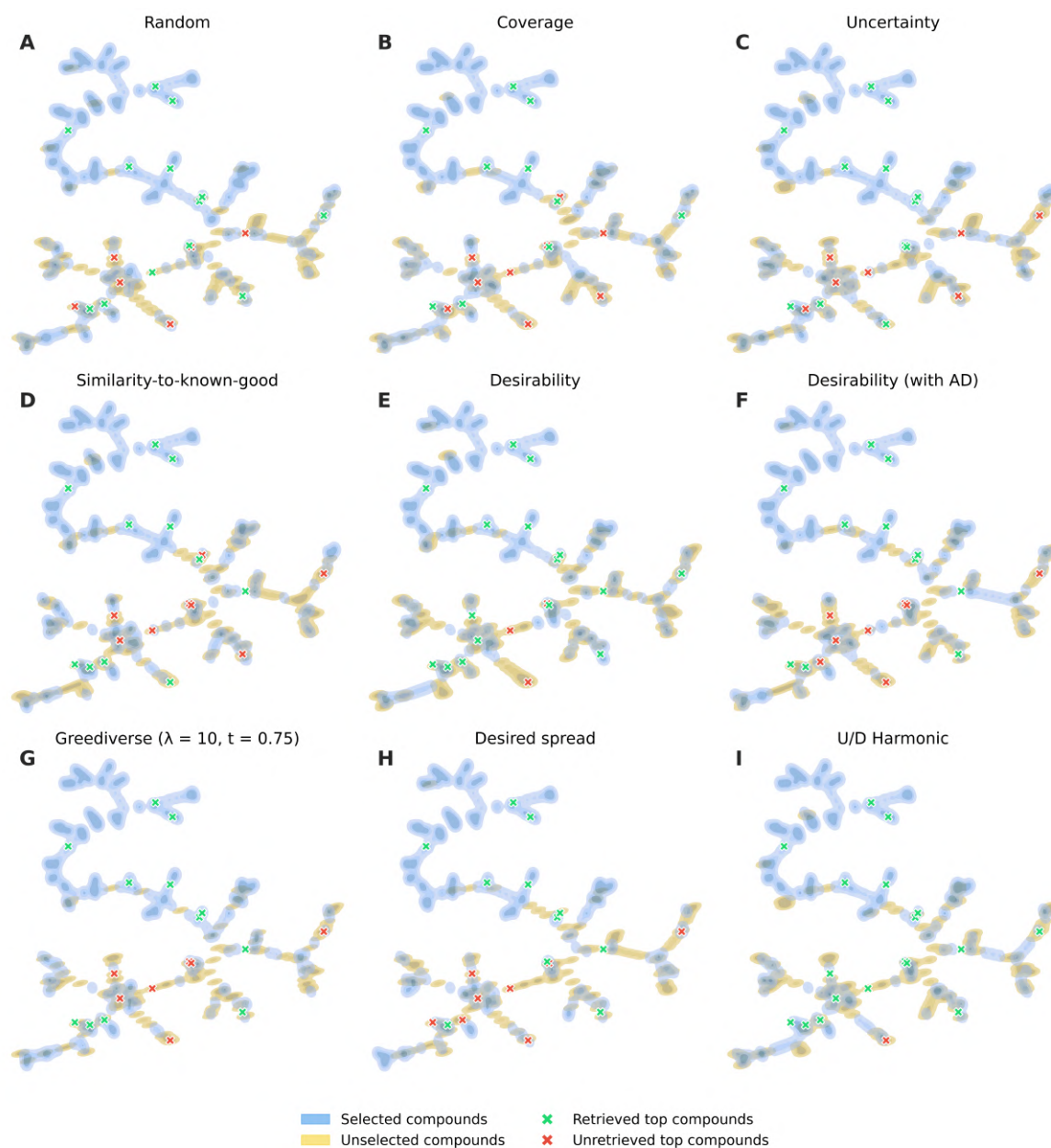


Figure 5.4: TMAP visualizations illustrating the final chemical space for nine acquisition functions on the MMP-8 dataset. Blue regions indicate sampled subspaces (selected molecules), while yellow denotes unselected compounds. Crosses highlight compounds within the Top DScore 5%, where green marks successful retrieval and red indicates missed compounds.

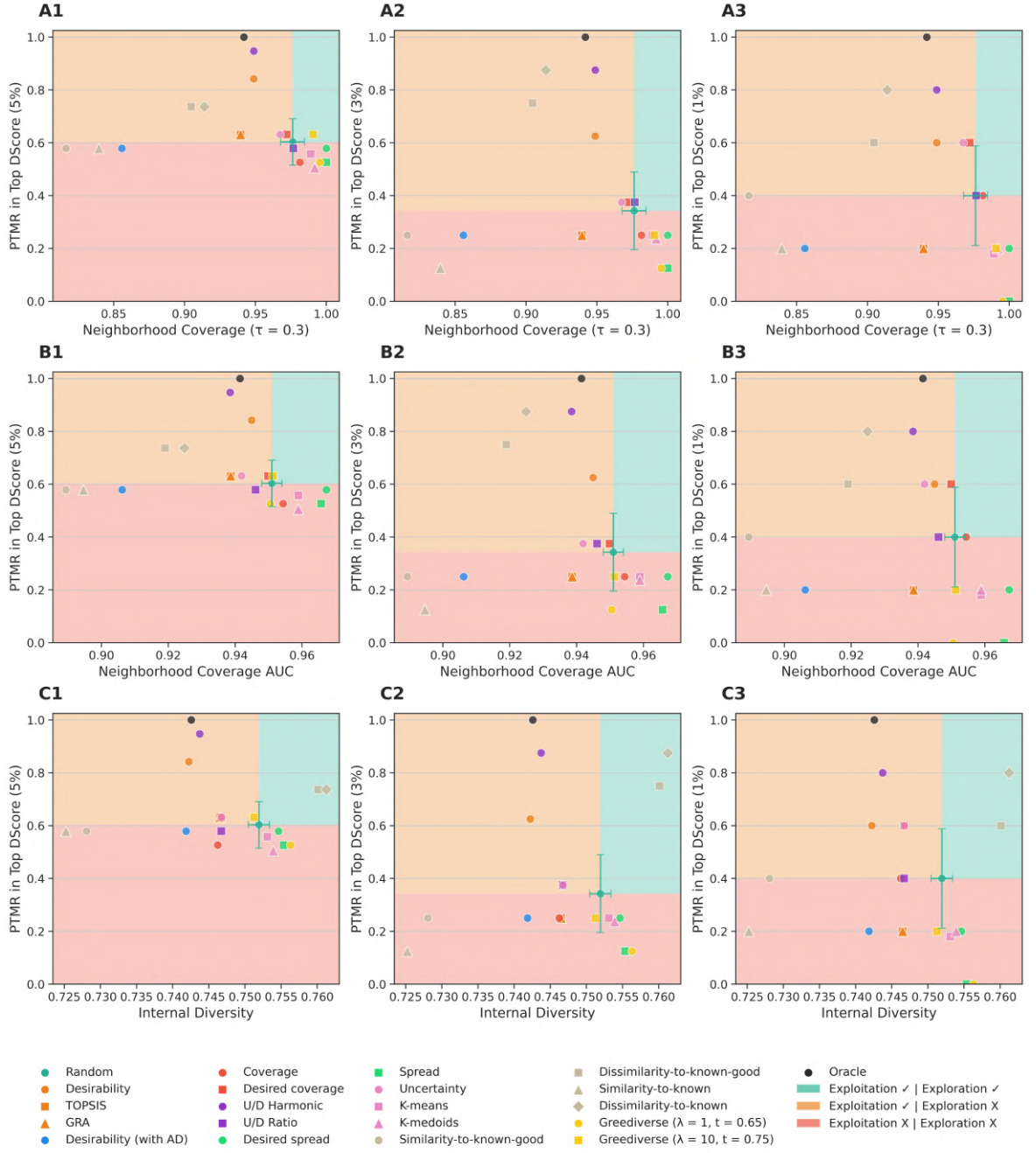


Figure 5.5: Final-state exploitation-exploration performance on the MMP-8 dataset, considering PTMR within Top DScore 5%, 3%, and 1% (columns 1, 2, and 3, respectively) plotted against exploration metrics: Neighborhood Coverage ($\tau = 0.3$), Neighborhood Coverage AUC, and Internal Diversity (rows A, B, and C, respectively).

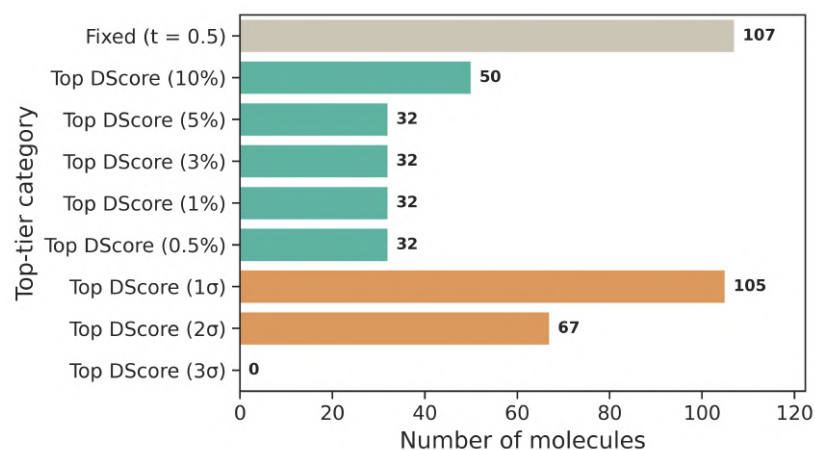
PPAR- δ 

Figure 5.6: Number of molecules falling into each top-tier category for the PPAR- δ dataset out of the 498 molecules. Categories are color-coded: brown indicates a fixed desirability threshold; orange indicates percentile-based thresholds; and green indicates standard deviation-based thresholds.

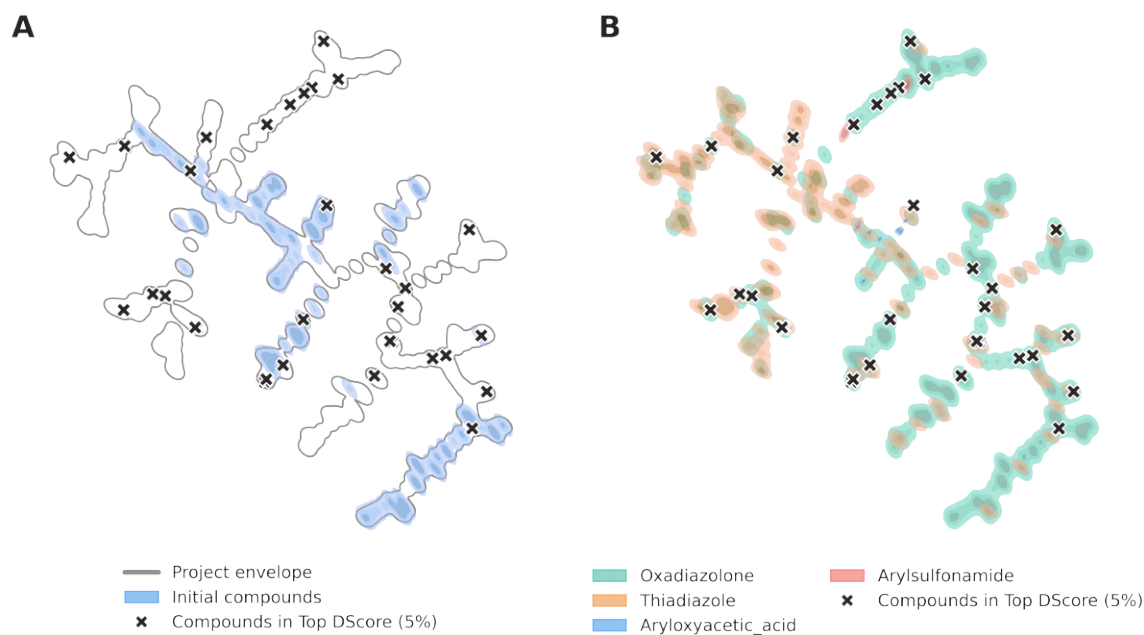


Figure 5.7: Overview of the chemical space of PPAR- δ at the start of a simulation (A) and distributions of major chemical series (B).

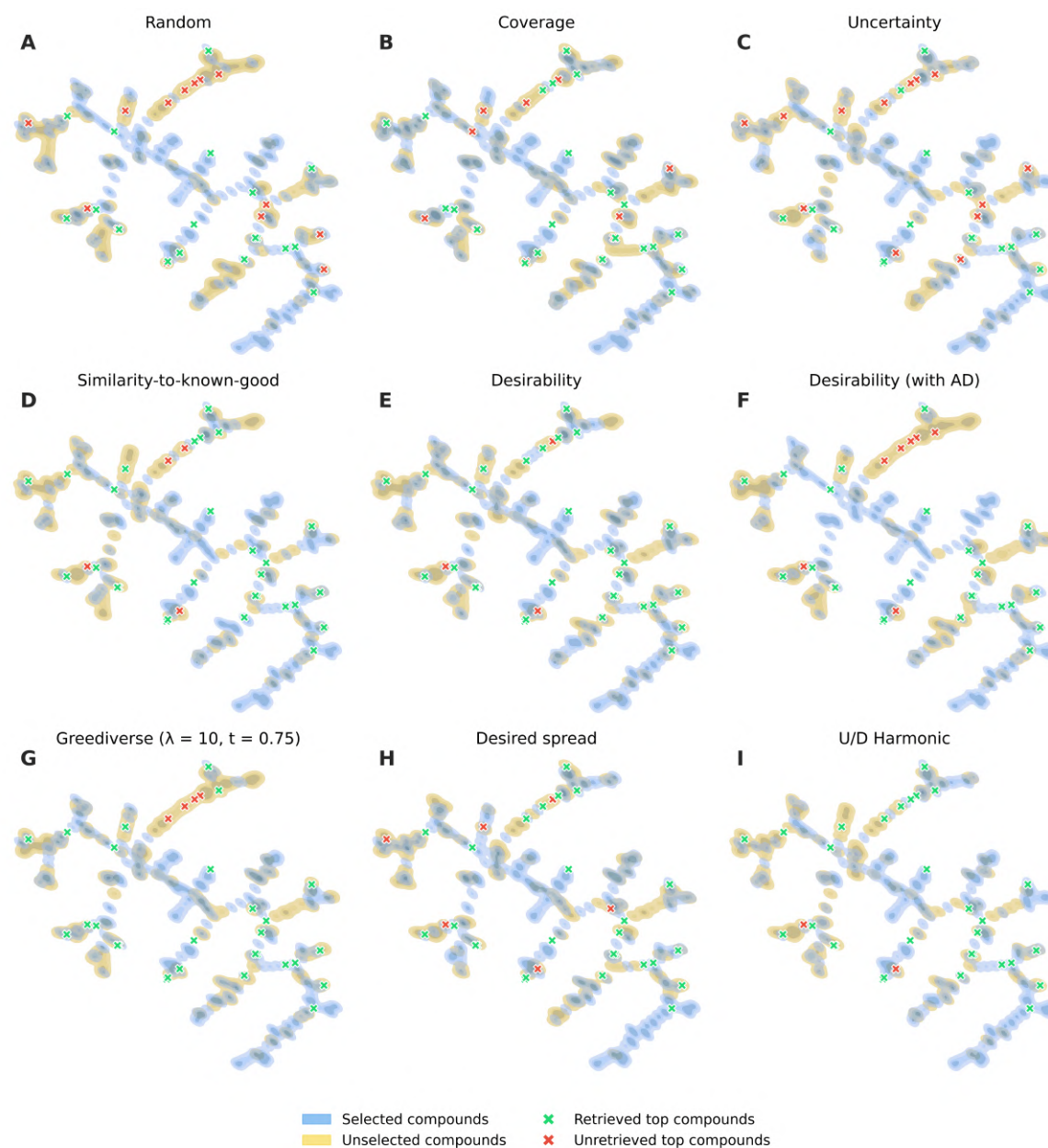


Figure 5.8: TMAP visualizations illustrating the final chemical space for nine acquisition functions on the PPAR- δ dataset. Blue regions indicate sampled subspaces (selected molecules), while yellow denotes unselected compounds. Crosses highlight compounds within the Top DScore 5%, where green marks successful retrieval and red indicates missed compounds.

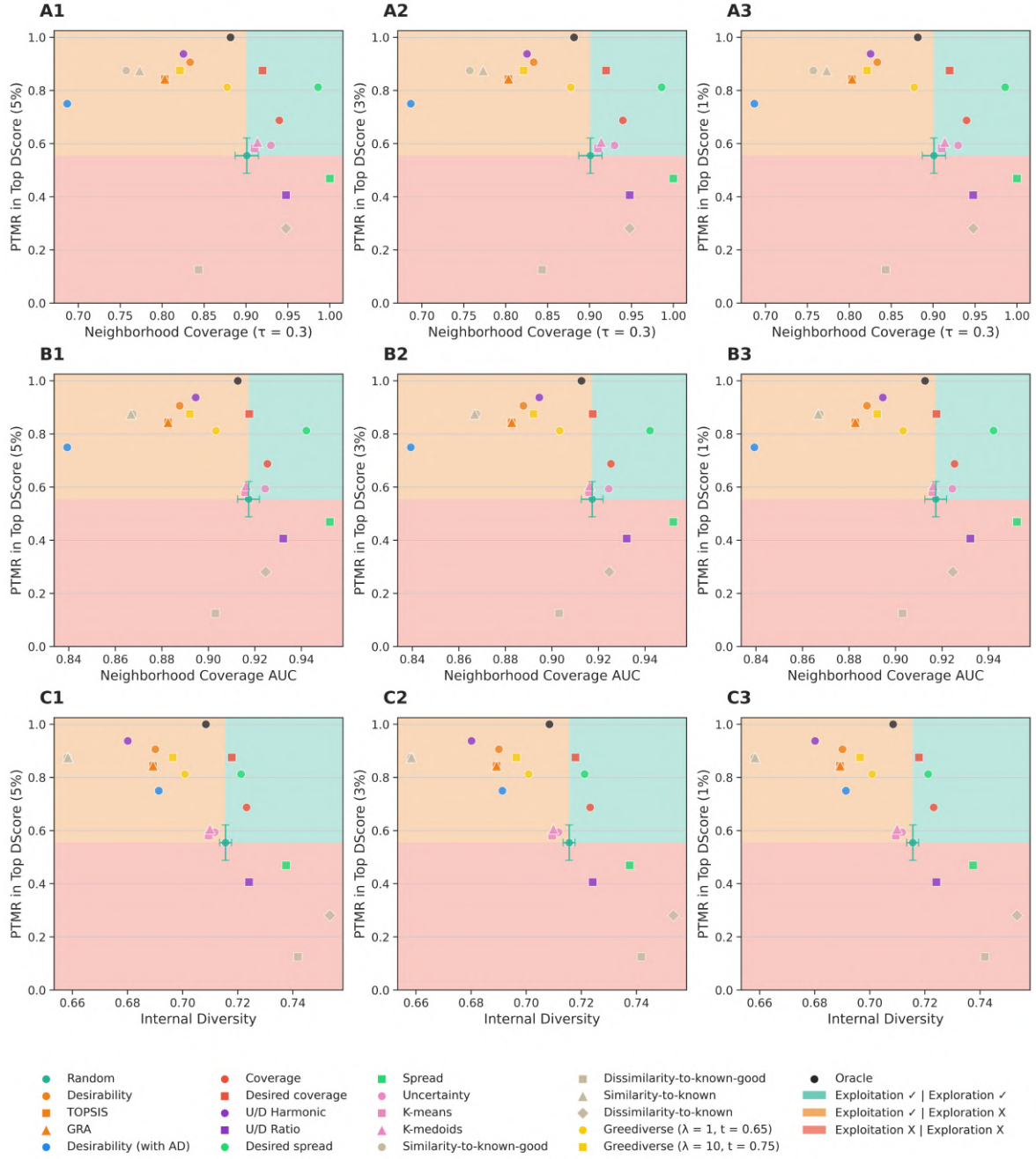


Figure 5.9: Final-state exploitation-exploration performance on the Renin dataset, considering PTMR within Top DScore 5%, 3%, and 1% (columns 1, 2, and 3, respectively) plotted against exploration metrics: Neighborhood Coverage ($\tau = 0.3$), Neighborhood Coverage AUC, and Internal Diversity (rows A, B, and C, respectively).

Renin

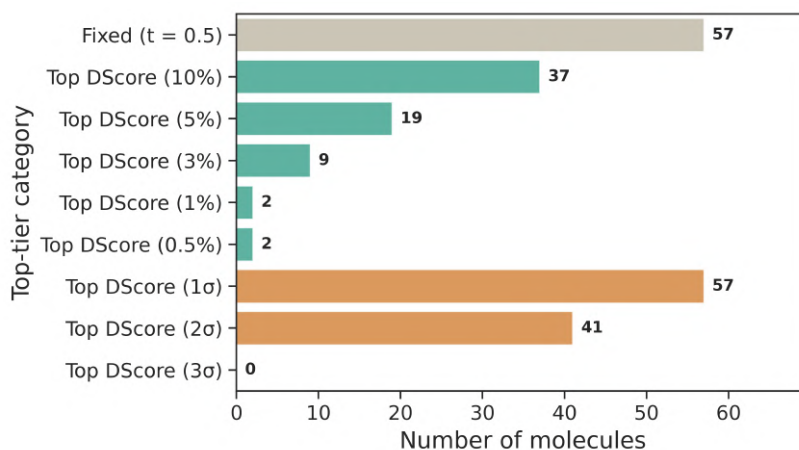


Figure 5.10: Number of molecules falling into each top-tier category for the Renin dataset out of the 388 molecules. Categories are color-coded: brown indicates a fixed desirability threshold; orange indicates percentile-based thresholds; and green indicates standard deviation-based thresholds.

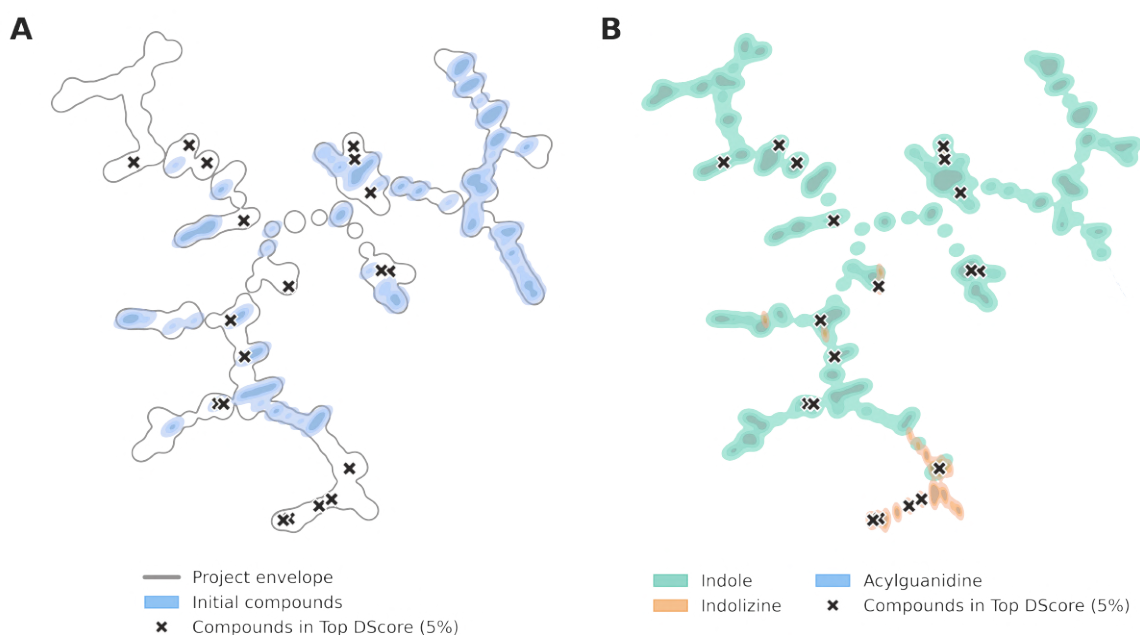


Figure 5.11: Overview of the chemical space of Renin at the start of a simulation (A) and distributions of major chemical series (B).

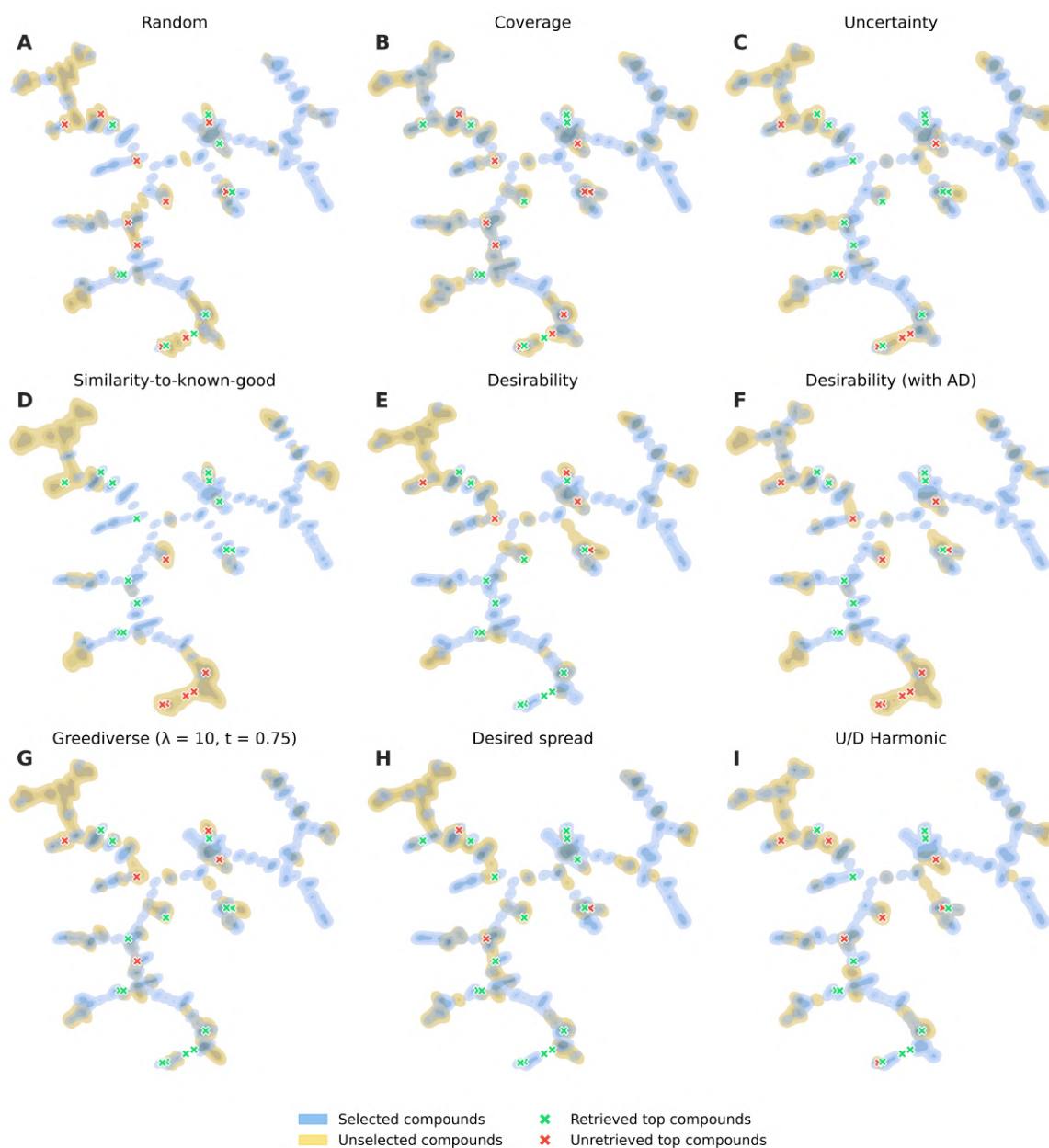


Figure 5.12: TMAP visualizations illustrating the final chemical space for nine acquisition functions on the Renin dataset. Blue regions indicate sampled subspaces (selected molecules), while yellow denotes unselected compounds. Crosses highlight compounds within the Top DScore 5%, where green marks successful retrieval and red indicates missed compounds.

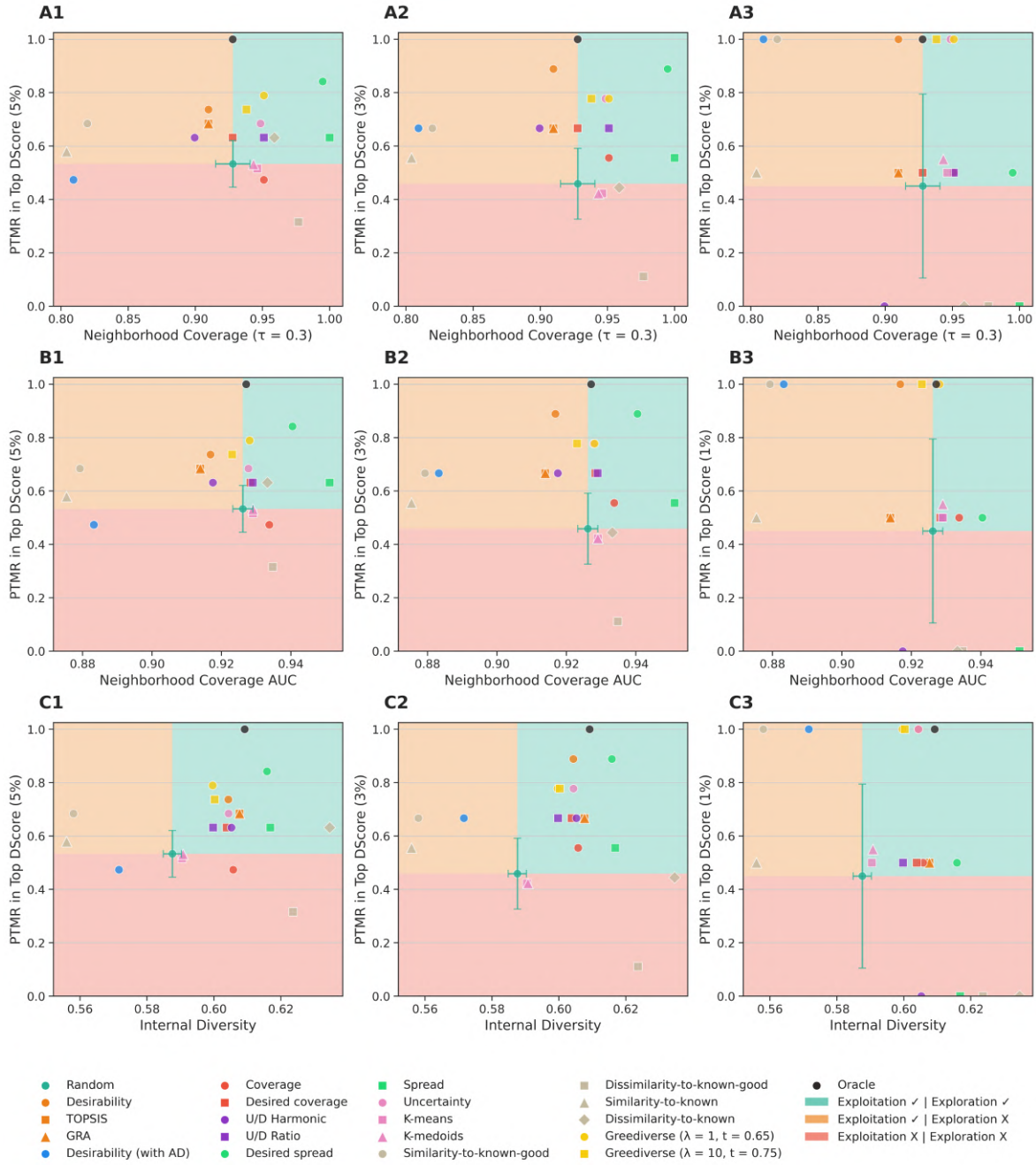


Figure 5.13: Final-state exploitation-exploration performance on the Renin dataset, considering PTMR within Top DScore 5%, 3%, and 1% (columns 1, 2, and 3, respectively) plotted against exploration metrics: Neighborhood Coverage ($\tau = 0.3$), Neighborhood Coverage AUC, and Internal Diversity (rows A, B, and C, respectively).

Appendix C

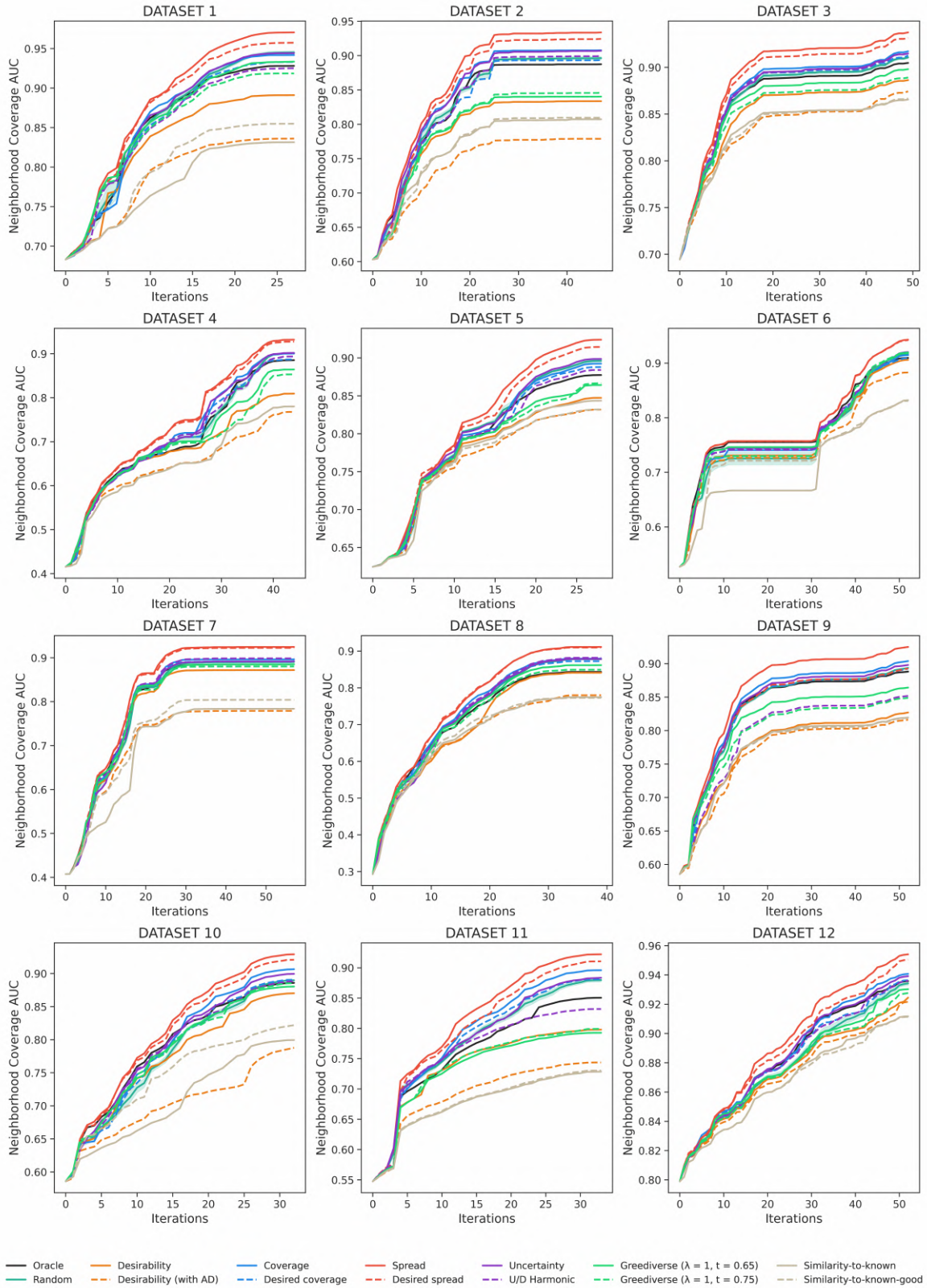


Figure 5.14: Monitoring Neighborhood Coverage AUC for various selection strategies over iterations for each dataset.

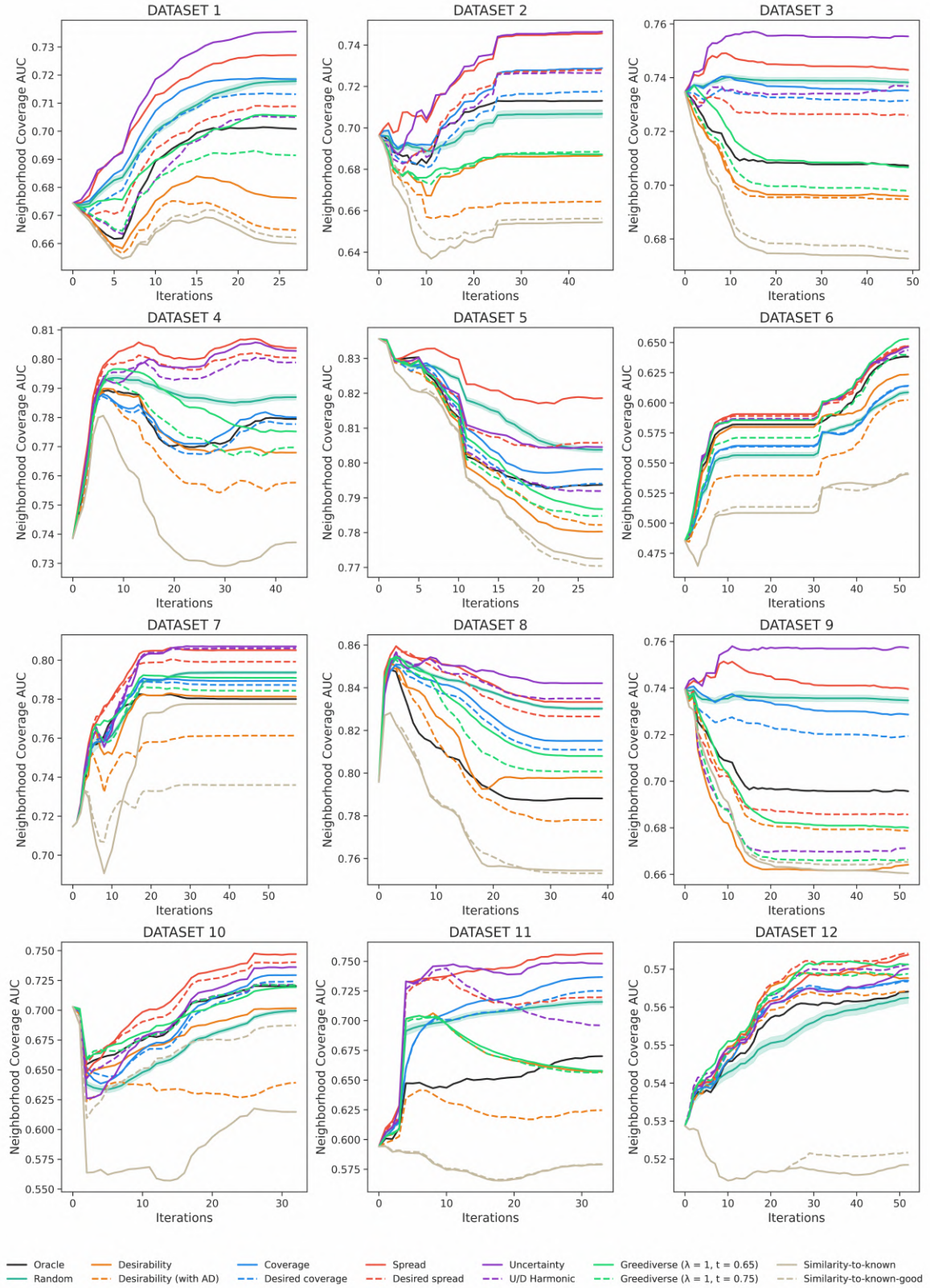


Figure 5.15: Monitoring Internal Diversity for various selection strategies over iterations for each dataset.

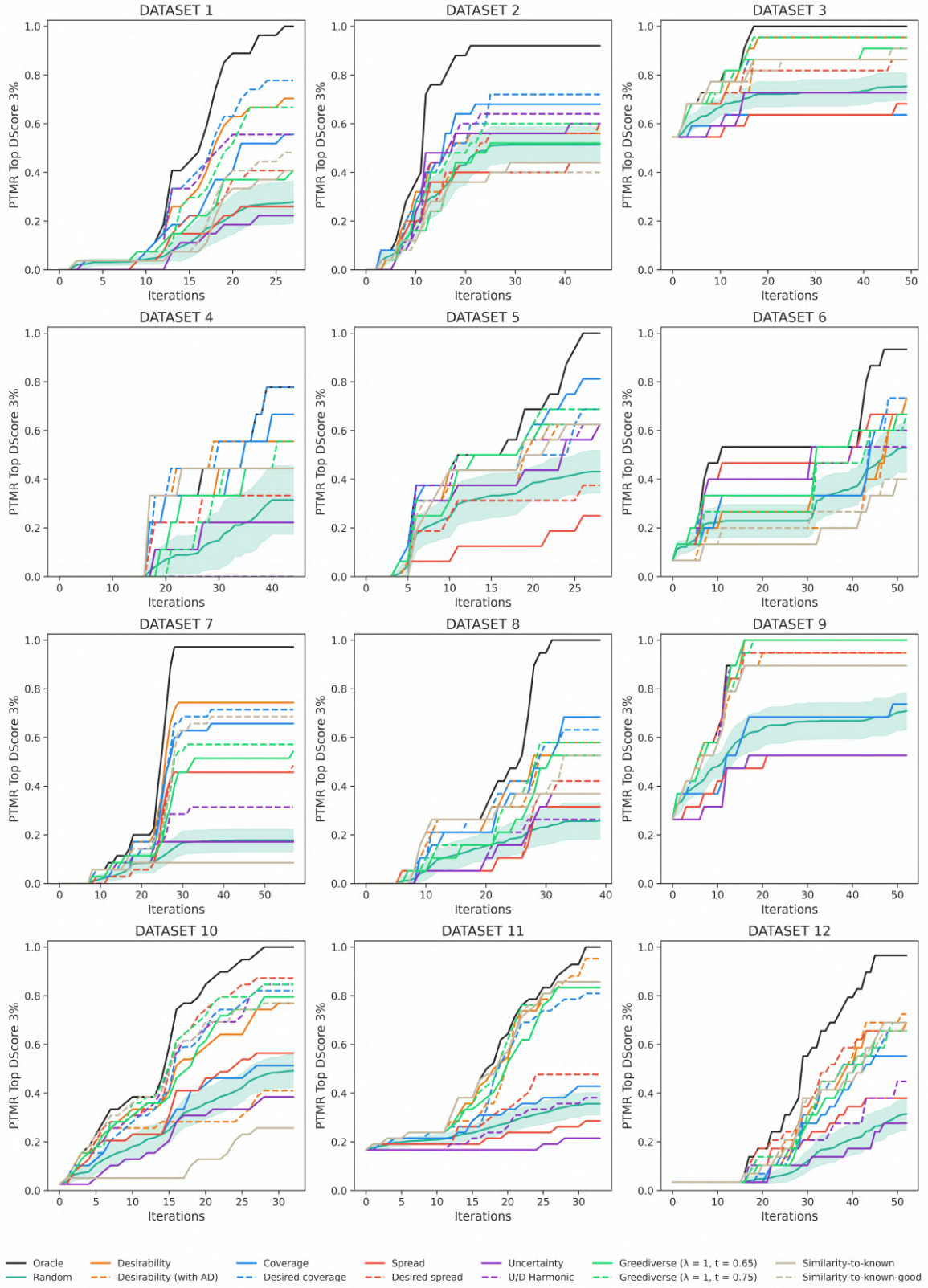


Figure 5.16: PTMR Top DScore 3% for various selection strategies over iterations for each dataset.

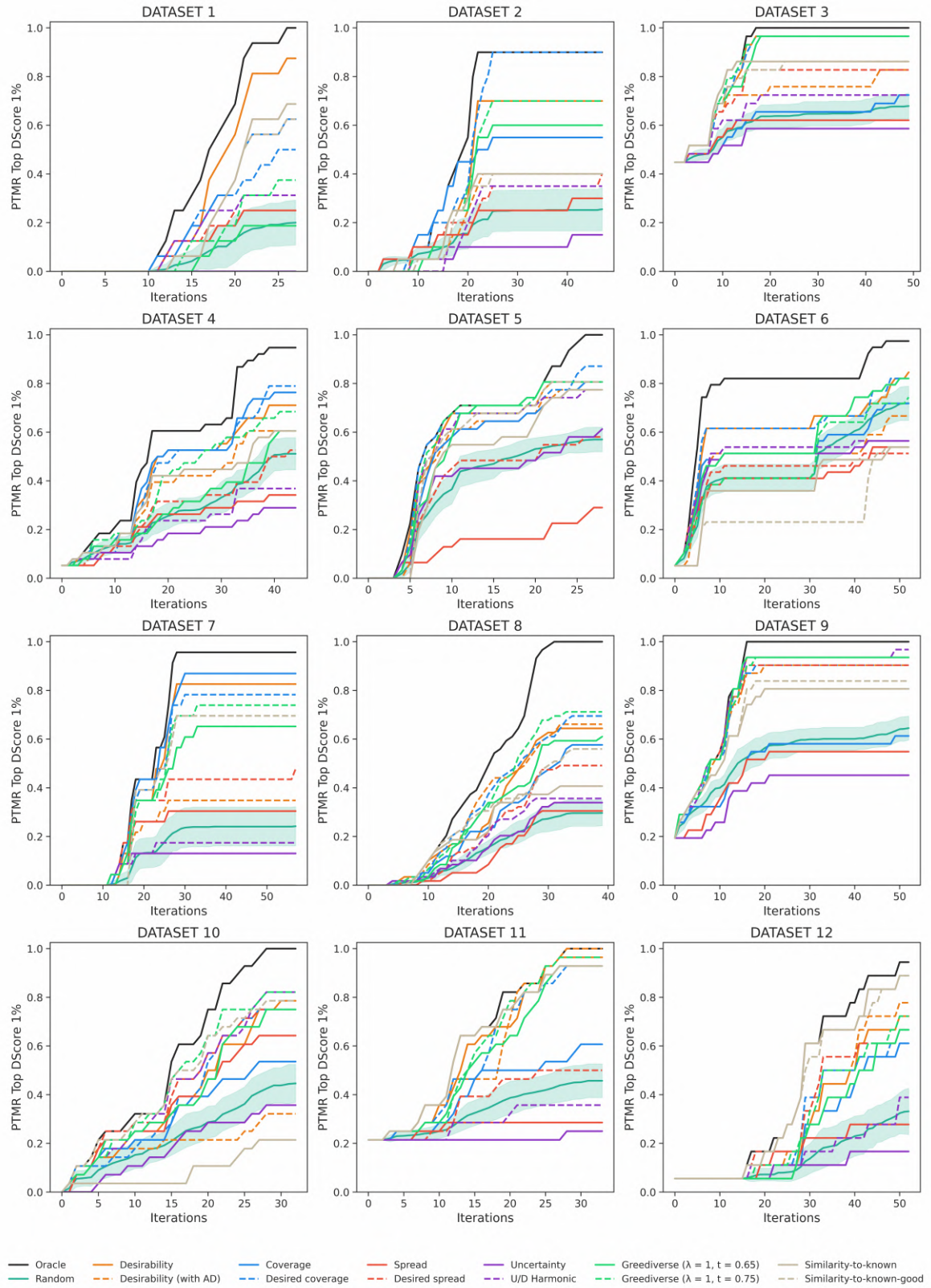


Figure 5.17: PTMR Top DScore 1% for various selection strategies over iterations for each dataset.

List of Figures

1.1	Annual FDA drug approvals by therapeutic modality	8
1.2	Distribution of human protein drug targets and corresponding small-molecule drugs	9
1.3	Research: stages, success rates, and durations	10
1.4	Development: stages, success rates, and durations	17
1.5	From formula to 3D structure with caffeine	20
1.6	Graph-based representation of dimethylcarbamic acid	22
1.7	Circular and linear fingerprints	23
1.8	Illustration of Matched Molecular Pairs	28
1.9	Decision tree model	30
1.10	Random Forest model	30
1.11	Gradient Boosted Trees model	31
1.12	Single and multi-output neural networks	33
1.13	PCA, t-SNE, UMAP and TMAP representations	38
1.14	Active learning cycle	39
1.15	Dose-response curve for inhibitors	41
1.16	Design-Make-Test-Analyze cycle	45
1.17	Example of a lead optimization screening tree	47
2.1	Size distribution of benchmarking datasets	61
2.2	Illustration of data splitting strategies to assess model performance	64
2.3	Distribution of R^2 scores across ADMET and Potency datasets	66
2.4	Distribution of normalized mean absolute error across ADMET and Potency datasets	67
2.5	Impact of dataset size on normalized mean absolute error	69
2.6	Training time versus dataset size	70
2.7	Ranking of single descriptors considering R^2 , nMAE and nRMSE	72

2.8	Ranking of single descriptors and Chemprop considering nMAE for ADMET and potency datasets	73
2.9	Pairwise statistical comparison (p-values) of top-ranked single descriptors and Chemprop for ADMET and potency datasets	74
2.10	Ranking of all descriptors and Chemprop considering nMAE	76
2.11	Ranking of all descriptors and Chemprop considering nMAE for ADMET and potency datasets	78
2.12	Pairwise statistical comparison (p-values) of top-ranked descriptors and Chemprop for ADMET and potency datasets	79
2.13	nMAE distributions under temporal and cross-validation splits for vector-based descriptors	80
2.14	R^2 distributions under temporal and cross-validation splits for vector-based descriptors	81
2.15	nMAE distributions under temporal and cross-validation splits for Chemprop	83
2.16	R^2 distributions under temporal and cross-validation splits for Chemprop	84
3.1	Overview of the D(S)MTA cycle	89
3.2	Temporal distribution of new molecules in FXa, MMP-8, PPAR δ and Renin datasets	93
3.3	Distribution of endpoints assays values for FXa, MMP-8, PPAR δ and Renin datasets	95
3.4	Simulation snapshots of initial, intermediate, and final stages	97
3.5	Utility functions for optimization directions	99
3.6	Desirability vs. documentation	101
3.7	U/D Harmonic scores distribution	109
3.8	U/D Ratio scores distribution	109
3.9	TMAP, KDE TMAP and Envelope	114
3.10	Iterative TMAPs with a random selection on FXa dataset	115
3.11	Illustration of the Internal Diversity metric	116
3.12	Illustration of the Neighborhood Coverage metric	117
3.13	Distribution of FXa molecules by top-tier categories	120
3.14	Real vs. simulated compound selection for FXa dataset	121
3.15	Initial chemical space, series distribution and example structures for the FXa dataset	122
3.16	Final chemical space across acquisition functions	124
3.17	Evolution of exploration metrics for FXa dataset	125
3.18	Evolution of exploitation metrics for FXa dataset	126

3.19	Exploitation/exploration trade-off trajectories for the FXa dataset	127
3.20	Exploitation/exploration trade-off final state	129
3.21	Exploitation/exploration trade-off final state using various predictive models	130
4.1	Temporal distribution of new molecules over time	135
4.2	Distribution of pairwise molecular similarities within each project	136
4.3	Pairwise molecular similarity distributions broken down by chemical series	137
4.4	Distribution of documentation levels across projects	138
4.5	Classification of compounds using fixed thresholds for desirability and documenta- tion across datasets	139
4.6	Classification of compounds using dataset-specific desirability thresholds and fixed documentation thresholds	140
4.7	Comparison of compounds classified in top categories across projects	141
4.8	TMAP overview of the initial chemical spaces	144
4.9	TMAP overview of the chemical series	145
4.10	Final overview on dataset 1	147
4.11	Final overview on dataset 11	148
4.12	Exploration: Neighborhood Coverage (τ of 0.3) over iterations	149
4.13	Exploitation: Top DScore 5% retrieval over iterations	151
4.14	Trade-off: Top DScore 5% vs Neighborhood Coverage ($\tau = 0.3$) final overview . . .	153
4.15	PTMR distribution by strategy for Top DScore 5% / 3% / 1%	154
4.16	PTMR Ratio to Random for Top DScore 5% / 3% / 1%	155
4.17	Distribution of the exploration metrics across datasets for every strategy	157
4.18	Distribution of the Ratio to Random for exploration metrics across datasets for every strategy	159
4.19	Exploitation/exploration trade-off averaged over all datasets (ratio to Random) . . .	160

List of Tables

1.1	Overview of common structure-based drug design (SBDD) methods used to identify and optimize ligand interactions with biological targets.	25
1.2	Some key ADMET endpoints with typical target values and associated biological relevance for small-molecule drug candidates. The target value is indicative, and may change depending on the project.	43
2.1	Public and industrial datasets characteristics	53
2.2	Summary of the descriptors and their variants	60
3.1	Blueprint for FXa, Renin, PPAR δ and MMP-8 datasets	94
3.2	Simulation parameters for virtual scenarios on the FXa, Renin, PPAR δ , and MMP-8 datasets	96
3.3	Summary of compound selection and simulation iterations across datasets	98
3.4	Summary of the different top categories of molecules	100
3.5	Summary of the acquisition functions	103
3.6	Summary of the predictive models available	107
3.7	Summary of exploration metrics	119
4.1	Dataset sizes and endpoint counts across 12 lead optimization projects	135
4.2	Overview of simulation parameters and the number of selected compounds	142
4.3	Statistics of PTMR Top DScore 5% Ratio to Random	156
4.4	Statistics of Neighborhood Coverage ($\tau=0.3$) Ratio to Random	158

List of abbreviations

- **3R**: Replacement, Reduction, and Refinement
- **AD**: Applicability Domain
- **ADME**: Absorption, Distribution, Metabolism, and Excretion
- **ADMET**: Absorption, Distribution, Metabolism, Excretion, and Toxicology
- **AI**: Artificial Intelligence
- **AL**: Active Learning
- **AUC**: Area Under the Curve
- **BLA**: Biologics License Application
- **CAR-T**: Chimeric Antigen Receptor T cell
- **CL**: Clearance
- **CNS**: Central Nervous System
- **CRISPR**: Clustered Regularly Interspaced Short Palindromic Repeats
- **CTA**: Clinical Trial Application
- **CYP450**: Cytochrome P450
- **D(S)MTA**: Design-(Select)-Make-Test-Analyze
- **DCTP**: Drug Candidate Target Profile
- **DMTA**: Design-Make-Test-Analyze
- **DNA**: Deoxyribonucleic Acid
- **EC50**: Half-maximal effective concentration
- **ECFP**: Extended Connectivity Fingerprints
- **EMA**: European Medicines Agency
- **FDA**: Food and Drug Administration
- **FXa**: Factor Xa
- **GLP**: Good Laboratory Practice
- **GNN**: Graph Neural Network
- **GP**: Gaussian Process
- **GPCR**: G protein-coupled receptor
- **GPT**: Generative Pre-trained Transformer
- **GRA**: Grey Relational Analysis
- **GWAS**: Genome-Wide Association Study
- **HTS**: High-Throughput Screening
- **HTV**: Highest Threshold Value
- **IND**: Investigational New Drug
- **IP**: Intellectual Property
- **InChI**: International Chemical Identifier
- **KDE**: Kernel Density Estimate
- **LBDD**: Ligand-Based Drug Design
- **LSH**: Locality-Sensitive Hashing

- **LTV**: Lowest Threshold Value
- **LightGBM**: Light Gradient Boosting Machine
- **MAA**: Marketing Authorization Application
- **MACCS**: Molecular ACCess System
- **MAE**: Mean Absolute Error
- **MCDA**: Multiple-Criteria Decision Analysis
- **ML**: Machine Learning
- **MLP**: Multilayer Perceptron
- **MMP**: Matched Molecular Pair
- **MMP-8**: Matrix Metalloproteinase-8
- **MPNN**: Message Passing Neural Network
- **MST**: Minimum Spanning Tree
- **MTD**: Maximum Tolerated Dose
- **NC**: Neighborhood Coverage
- **NCAUC**: Neighborhood Coverage Area Under the Curve
- **NDA**: New Drug Application
- **NSGA-II**: Non-dominated Sorting Genetic Algorithm II
- **PCA**: Principal Component Analysis
- **PD**: Pharmacodynamics
- **PK**: Pharmacokinetics
- **PPAR δ** : Peroxisome-Proliferator-Activated Receptor- δ
- **PPB**: Plasma Protein Binding
- **PTMR**: Proportion of Top Molecules Retrieved
- **PhD**: Doctor of Philosophy
- **QSAR**: Quantitative Structure-Activity Relationship
- **QSPR**: Quantitative Structure-Property Relationship
- **RF**: Random Forest
- **RMSE**: Root Mean Square Error
- **RNA**: Ribonucleic Acid
- **RNN**: Recurrent Neural Network
- **SALI**: Structure-Activity Landscape Index
- **SAR**: Structure-Activity Relationship
- **SBDD**: Structure-Based Drug Design
- **SECFP**: SMILES Extended Connectivity Fingerprint
- **SMILES**: Simplified Molecular Input Line Entry System
- **SPR**: Structure-Property Relationship
- **SVM**: Support Vector Machine
- **TMAP**: Tree MAP
- **TOPSIS**: Technique for Order of Preference by Similarity to Ideal Solution
- **TPP**: Target Product Profile
- **Tox**: Toxicology
- **UMAP**: Uniform Manifold Approximation and Projection
- **Vd**: Volume of distribution
- **XGBoost**: eXtreme Gradient Boosting
- **hERG**: human Ether-à-go-go-Related Gene
- **kNN**: k-Nearest Neighbors
- **mRNA**: messenger RNA
- **nMAE**: normalized Mean Absolute Error
- **nRMSE**: normalized Root Mean Square Error
- **siRNA**: small interfering RNA
- **t-SNE**: t-distributed Stochastic Neighbor Embedding

Bibliography

- [1] Joseph A. DiMasi, Henry G. Grabowski, and Ronald W. Hansen. "Innovation in the pharmaceutical industry: New estimates of R&D costs". In: *Journal of Health Economics* 47 (2016), pp. 20–33. ISSN: 0167-6296. DOI: 10 . 1016/j . jhealeco . 2016 . 01 . 012.
- [2] JP Hughes et al. "Principles of early drug discovery". In: *British Journal of Pharmacology* 162.6 (2011), pp. 1239–1249. ISSN: 0007-1188. DOI: 10 . 1111/j . 1476-5381 . 2010 . 01127 . x.
- [3] Michael J. Waring et al. "An analysis of the attrition of drug candidates from four major pharmaceutical companies". In: *Nature Reviews Drug Discovery* 14.7 (2015), pp. 475–486. ISSN: 1474-1776. DOI: 10 . 1038/nrd4609.
- [4] Ismail Kola and John Landis. "Can the pharmaceutical industry reduce attrition rates?" In: *Nature Reviews Drug Discovery* 3.8 (2004), pp. 711–716. ISSN: 1474-1776. DOI: 10 . 1038/nrd1470.
- [5] Steven M. Paul et al. "How to improve R&D productivity: the pharmaceutical industry's grand challenge". In: *Nature Reviews Drug Discovery* 9.3 (2010), pp. 203–214. ISSN: 1474-1776. DOI: 10 . 1038/nrd3078.
- [6] David C. Swinney and Jason Anthony. "How were new medicines discovered?" In: *Nature Reviews Drug Discovery* 10.7 (2011), pp. 507–519. ISSN: 1474-1776. DOI: 10 . 1038/nrd3480.
- [7] David J. Newman and Gordon M. Cragg. "Natural Products as Sources of New Drugs from 1981 to 2014". In: *Journal of Natural Products* 79.3 (2016), pp. 629–661. ISSN: 0163-3864. DOI: 10 . 1021/acs . jnatprod . 5b01055.
- [8] Walter Sneader. "The discovery of aspirin: a reappraisal". In: *BMJ* 321.7276 (2000), p. 1591. ISSN: 0959-8138. DOI: 10 . 1136/bmj . 321 . 7276 . 1591.
- [9] Robert Gaynes. "The Discovery of Penicillin—New Insights After More Than 75 Years of Clinical Use - Volume 23, Number 5—May 2017 - Emerging Infectious Diseases journal - CDC". In: *Emerging Infectious Diseases* 23.5 (2017), pp. 849–853. ISSN: 1080-6040. DOI: 10 . 3201/eid2305 . 161556.
- [10] Ricardo Macarron et al. "Impact of high-throughput screening in biomedical research". In: *Nature Reviews Drug Discovery* 10.3 (2011), pp. 188–195. ISSN: 1474-1776. DOI: 10 . 1038/nrd3368.
- [11] Andrew L. Hopkins and Colin R. Groom. "The druggable genome". In: *Nature Reviews Drug Discovery* 1.9 (2002), pp. 727–730. ISSN: 1474-1776. DOI: 10 . 1038/nrd892.
- [12] Gisbert Schneider. "Automating drug discovery". In: *Nature Reviews Drug Discovery* 17.2 (2018), pp. 97–113. ISSN: 1474-1776. DOI: 10 . 1038/nrd . 2017 . 232.
- [13] Christopher A. Lipinski. "Lead- and drug-like compounds: the rule-of-five revolution". In: *Drug Discovery Today: Technologies* 1.4 (2004), pp. 337–341. ISSN: 1740-6749. DOI: 10 . 1016/j . ddttec . 2004 . 11 . 007.
- [14] Favour Danladi Makurvet. "Biologics vs. small molecules: Drug costs and patient access". In: *Medicine in Drug Discovery* 9 (2021), p. 100075. ISSN: 2590-0986. DOI: 10 . 1016/j . medidd . 2020 . 100075.
- [15] Rita Santos et al. "A comprehensive map of molecular drug targets". In: *Nature Reviews Drug Discovery* 16.1 (2017), pp. 19–34. ISSN: 1474-1776. DOI: 10 . 1038/nrd . 2016 . 230.
- [16] Henry Grabowski, Genia Long, and Richard Mortimer. "Recent trends in brand-name and generic drug competition". In: *Journal of Medical Economics* 17.3 (2014), pp. 207–214. ISSN: 1369-6998. DOI: 10 . 3111/13696998 . 2013 . 873723.
- [17] Gary Walsh. "Biopharmaceutical benchmarks 2018". In: *Nature Biotechnology* 36.12 (2018), pp. 1136–1145. ISSN: 1087-0156. DOI: 10 . 1038/nbt . 4305.
- [18] Benjamin Leader, Quentin J. Baca, and David E. Golan. "Protein therapeutics: a summary and pharmacological classification". In: *Nature Reviews Drug Discovery* 7.1 (2008), pp. 21–39. ISSN: 1474-1776. DOI: 10 . 1038/nrd2399.

- [19] Hélène Kaplon and Janice M. Reichert. "Antibodies to watch in 2021". In: *mAbs* 13.1 (2021), p. 1860476. issn: 1942-0862. doi: 10.1080/19420862.2020.1860476.
- [20] Mariko Morishita and Nicholas A. Peppas. "Is the oral route possible for peptide and protein drug delivery?" In: *Drug Discovery Today* 11.19-20 (2006), pp. 905–910. issn: 1359-6446. doi: 10.1016/j.drudis.2006.08.005.
- [21] Anurag S Rathore and Helen Winkle. "Quality by design for biopharmaceuticals". In: *Nature Biotechnology* 27.1 (2009), pp. 26–34. issn: 1087-0156. doi: 10.1038/nbt0109-26.
- [22] Amy S. Rosenberg. "Effects of protein aggregates: An immunologic perspective". In: *The AAPS Journal* 8.3 (2006), E501–E507. doi: 10.1208/aapsj080359.
- [23] Krishna Sriram and Paul A. Insel. "G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs?" In: *Molecular Pharmacology* 93.4 (2018), pp. 251–258. issn: 0026-895X. doi: 10.1124/mol.117.111062.
- [24] C. Frank Bennett. "Therapeutic Antisense Oligonucleotides Are Coming of Age". In: *Annual Review of Medicine* 70.1 (2019), pp. 307–321. issn: 0066-4219. doi: 10.1146/annurev-med-041217-010829.
- [25] Laurence H. Hurley. "DNA and its associated processes as targets for cancer therapy". In: *Nature Reviews Cancer* 2.3 (2002), pp. 188–200. issn: 1474-175X. doi: 10.1038/nrc749.
- [26] Jeffrey Cummings et al. "Drug development in Alzheimer's disease: the path to 2025". In: *Alzheimer's Research & Therapy* 8.1 (2016), p. 39. doi: 10.1186/s13195-016-0207-9.
- [27] David Cook et al. "Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework". In: *Nature Reviews Drug Discovery* 13.6 (2014), pp. 419–431. issn: 1474-1776. doi: 10.1038/nrd4309.
- [28] Iain P. Chessell et al. "Disruption of the P2X7 purinoceptor gene abolishes chronic inflammatory and neuropathic pain". In: *Pain* 114.3 (2005), pp. 386–396. issn: 0304-3959. doi: 10.1016/j.pain.2005.01.002.
- [29] Peter M. Visscher et al. "10 Years of GWAS Discovery: Biology, Function, and Translation". In: *The American Journal of Human Genetics* 101.1 (2017), pp. 5–22. issn: 0002-9297. doi: 10.1016/j.ajhg.2017.06.005.
- [30] Ophir Shalem et al. "Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells". In: *Science* 343.6166 (2014), pp. 84–87. issn: 0036-8075. doi: 10.1126/science.1247005.
- [31] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. "Multi-omics approaches to disease". In: *Genome Biology* 18.1 (2017), p. 83. issn: 1474-7596. doi: 10.1186/s13059-017-1215-1.
- [32] Stephen V Frye. "The art of the chemical probe". In: *Nature Chemical Biology* 6.3 (2010), pp. 159–161. issn: 1552-4450. doi: 10.1038/nchembio.296.
- [33] Cheryl H Arrowsmith et al. "The promise and peril of chemical probes". In: *Nature Chemical Biology* 11.8 (2015), pp. 536–541. issn: 1552-4450. doi: 10.1038/nchembio.1867.
- [34] D A Pereira and J A Williams. "Origin and evolution of high throughput screening". In: *British Journal of Pharmacology* 152.1 (2007), pp. 53–61. issn: 0007-1188. doi: 10.1038/sj.bjp.0707373.
- [35] Tong Ying Shun et al. "Identifying Actives from HTS Data Sets: Practical Approaches for the Selection of an Appropriate HTS Data-Processing Method and Quality Control Review". In: *SLAS Discovery* 16.1 (2011), pp. 1–14. issn: 2472-5552. doi: 10.1177/1087057110389039.
- [36] Jean-Louis Reymond. "The Chemical Space Project". In: *Accounts of Chemical Research* 48.3 (2015), pp. 722–730. issn: 0001-4842. doi: 10.1021/ar500432k.
- [37] Douglas B. Kitchen et al. "Docking and scoring in virtual screening for drug discovery: methods and applications". In: *Nature Reviews Drug Discovery* 3.11 (2004), pp. 935–949. issn: 1474-1776. doi: 10.1038/nrd1549.
- [38] John J. Irwin and Brian K. Shoichet. "Docking Screens for Novel Ligands Conferring New Biology". In: *Journal of Medicinal Chemistry* 59.9 (2016), pp. 4103–4120. issn: 0022-2623. doi: 10.1021/acs.jmedchem.5b02008.
- [39] Peter Ripphausen, Britta Nisius, and Jürgen Bajorath. "State-of-the-art in ligand-based virtual screening". In: *Drug Discovery Today* 16.9-10 (2011), pp. 372–376. issn: 1359-6446. doi: 10.1016/j.drudis.2011.02.011.
- [40] W. Patrick Walters. "Virtual Chemical Libraries". In: *Journal of Medicinal Chemistry* 62.3 (2019), pp. 1116–1124. issn: 0022-2623. doi: 10.1021/acs.jmedchem.8b01048.
- [41] Derek van Tilborg and Francesca Grisoni. "Traversing chemical space with active deep learning for low-data drug discovery". In: *Nature Computational Science* 4.10 (2024), pp. 786–796. doi: 10.1038/s43588-024-00697-2.

- [42] David E. Graff, Eugene I. Shakhnovich, and Connor W. Coley. "Accelerating high-throughput virtual screening through molecular pool-based active learning". In: *Chemical Science* 12.22 (2021), pp. 7866–7881. ISSN: 2041-6520. DOI: 10.1039/d0sc06805e. eprint: 2012.07127.
- [43] John G. Moffat et al. "Opportunities and challenges in phenotypic drug discovery: an industry perspective". In: *Nature Reviews Drug Discovery* 16.8 (2017), pp. 531–543. ISSN: 1474-1776. DOI: 10.1038/nrd.2017.111.
- [44] Monica Schenone et al. "Target identification and mechanism of action in chemical biology and drug discovery". In: *Nature Chemical Biology* 9.4 (2013), pp. 232–240. ISSN: 1552-4450. DOI: 10.1038/nchembio.1199.
- [45] D C Swinney. "Phenotypic vs. Target-Based Drug Discovery for First-in-Class Medicines". In: *Clinical Pharmacology & Therapeutics* 93.4 (2013), pp. 299–301. ISSN: 0009-9236. DOI: 10.1038/clpt.2012.236.
- [46] György M. Keserü and Gergely M. Makara. "Hit discovery and hit-to-lead approaches". In: *Drug Discovery Today* 11.15-16 (2006), pp. 741–748. ISSN: 1359-6446. DOI: 10.1016/j.drudis.2006.06.016.
- [47] Rajarshi Guha. "In Silico Models for Drug Discovery". In: *Methods in Molecular Biology* 993 (2013), pp. 81–94. ISSN: 1064-3745. DOI: 10.1007/978-1-62703-342-8_6.
- [48] Natasha Thorne, Douglas S Auld, and James Inglese. "Apparent activity in high-throughput screening: origins of compound-dependent assay interference". In: *Current Opinion in Chemical Biology* 14.3 (2010), pp. 315–324. ISSN: 1367-5931. DOI: 10.1016/j.cbpa.2010.03.020.
- [49] Jonathan Baell and Michael A. Walters. "Chemistry: Chemical con artists foil drug discovery". In: *Nature* 513.7519 (2014), pp. 481–483. ISSN: 0028-0836. DOI: 10.1038/513481a.
- [50] M. Paul Gleeson et al. "Probing the links between in vitro potency, ADMET and physicochemical parameters". In: *Nature Reviews Drug Discovery* 10.3 (2011), pp. 197–208. ISSN: 1474-1776. DOI: 10.1038/nrd3367.
- [51] Paul D. Leeson and Brian Springthorpe. "The influence of drug-like concepts on decision-making in medicinal chemistry". In: *Nature Reviews Drug Discovery* 6.11 (2007), pp. 881–890. ISSN: 1474-1776. DOI: 10.1038/nrd2445.
- [52] Matthew D Segall. "Multi-Parameter Optimization: Identifying High Quality Compounds with a Balance of Properties". In: *Current Pharmaceutical Design* 18.9 (2012), pp. 1292–1310. ISSN: 1381-6128. DOI: 10.2174/138161212799436430.
- [53] John A Arnott and Sonia Lobo Planey. "The influence of lipophilicity in drug discovery and design". In: *Expert Opinion on Drug Discovery* 7.10 (2012), pp. 863–875. ISSN: 1746-0441. DOI: 10.1517/17460441.2012.714363.
- [54] Michael M. Hann and György M. Keserü. "Finding the sweet spot: the role of nature and nurture in medicinal chemistry". In: *Nature Reviews Drug Discovery* 11.5 (2012), pp. 355–365. ISSN: 1474-1776. DOI: 10.1038/nrd3701.
- [55] György M. Keserü and Gergely M. Makara. "The influence of lead discovery strategies on the properties of drug candidates". In: *Nature Reviews Drug Discovery* 8.3 (2009), pp. 203–212. ISSN: 1474-1776. DOI: 10.1038/nrd2796.
- [56] Kristien Mortelmans and Errol Zeiger. "The Ames Salmonella/microsome mutagenicity assay". In: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 455.1-2 (2000), pp. 29–60. ISSN: 0027-5107. DOI: 10.1016/S0027-5107(00)00064-6.
- [57] Urs A. Meyer. "Overview of enzymes of drug metabolism". In: *Journal of Pharmacokinetics and Biopharmaceutics* 24.5 (1996), pp. 449–459. ISSN: 0090-466X. DOI: 10.1007/bf02353473.
- [58] F. Peter Guengerich. "Cytochrome P450s and other enzymes in drug metabolism and toxicity". In: *The AAPS Journal* 8.1 (2006), E101–E111. DOI: 10.1208/aapsj080112.
- [59] Richard B van Breemen and Yongmei Li. "Caco-2 cell permeability assays to measure drug absorption". In: *Expert Opinion on Drug Metabolism & Toxicology* 1.2 (2005), pp. 175–185. ISSN: 1742-5255. DOI: 10.1517/17425255.1.2.175.
- [60] Patricia V Turner et al. "Administration of substances to laboratory animals: routes of administration and factors to consider." In: *Journal of the American Association for Laboratory Animal Science : JAALAS* 50.5 (2011), pp. 600–13.
- [61] Tinneke Denayer, Thomas Stöhr, and Maarten Van Roy. "Animal models in translational medicine: Validation and prediction". In: *New Horizons in Translational Medicine* 2.1 (2014), pp. 5–11. ISSN: 2307-5023. DOI: 10.1016/j.nhtm.2014.08.001.
- [62] W.M.S Russell and R.L Burch. "The Principles of Humane Experimental Technique". In: *Medical Journal of Australia* 1.13 (1960), pp. 500–500. ISSN: 0025-729X. DOI: 10.5694/j.1326-5377.1960.tb73127.x.

- [63] Jerrold Tannenbaum and B Taylor Bennett. "Russell and Burch's 3Rs then and now: the need for clarity in definition and purpose." In: *Journal of the American Association for Laboratory Animal Science : JAALAS* 54.2 (2015), pp. 120–32.
- [64] Lucie A Low and Danilo A Tagle. "Organs-on-chips: Progress, challenges, and future directions". In: *Experimental Biology and Medicine* 242.16 (2017), pp. 1573–1578. issn: 1535-3702. doi: 10.1177/1535370217700523.
- [65] Lorna Ewart et al. "Application of Microphysiological Systems to Enhance Safety Assessment in Drug Discovery". In: *Annual Review of Pharmacology and Toxicology* 58.1 (2017), pp. 1–18. issn: 0362-1642. doi: 10.1146/annurev-pharmtox-010617-052722.
- [66] Harry Olson et al. "Concordance of the Toxicity of Pharmaceuticals in Humans and in Animals". In: *Regulatory Toxicology and Pharmacology* 32.1 (2000), pp. 56–67. issn: 0273-2300. doi: 10.1006/rtph.2000.1399.
- [67] Jarrod Bailey and Michael Balls. "Recent efforts to elucidate the scientific validity of animal-based drug tests by the pharmaceutical industry, pro-testing lobby groups, and animal welfare organisations". In: *BMC Medical Ethics* 20.1 (2019), p. 16. doi: 10.1186/s12910-019-0352-3.
- [68] Lawrence M. Friedman, Curt D. Furberg, and David L. DeMets. "Fundamentals of Clinical Trials". In: (2010), pp. 199–214. doi: 10.1007/978-1-4419-1586-3_11.
- [69] Michael Hay et al. "Clinical development success rates for investigational drugs". In: *Nature Biotechnology* 32.1 (2014), pp. 40–51. issn: 1087-0156. doi: 10.1038/nbt.2786.
- [70] Chi Heem Wong, Kien Wei Siah, and Andrew W Lo. "Estimation of clinical trial success rates and related parameters". In: *Biostatistics* 20.2 (2019), pp. 273–286. issn: 1465-4644. doi: 10.1093/biostatistics/kxx069.
- [71] Christophe Le Tourneau, J. Jack Lee, and Lillian L. Siu. "Dose Escalation Methods in Phase I Cancer Clinical Trials". In: *JNCI: Journal of the National Cancer Institute* 101.10 (2009), pp. 708–720. issn: 0027-8874. doi: 10.1093/jnci/djp079.
- [72] Thomas R. Fleming and John H. Powers. "Biomarkers and surrogate endpoints in clinical trials". In: *Statistics in Medicine* 31.25 (2012), pp. 2973–2984. issn: 0277-6715. doi: 10.1002/sim.5403.
- [73] Charlie McLeod et al. "Choosing primary endpoints for clinical trials of health care interventions". In: *Contemporary Clinical Trials Communications* 16 (2019), p. 100486. issn: 2451-8654. doi: 10.1016/j.conctc.2019.100486.
- [74] Jessica Vamathevan et al. "Applications of machine learning in drug discovery and development". In: *Nature Reviews Drug Discovery* 18.6 (2019), pp. 463–477. issn: 1474-1776. doi: 10.1038/s41573-019-0024-5.
- [75] W. Patrick Walters and Regina Barzilay. "Applications of Deep Learning in Molecule Generation and Molecular Property Prediction". In: *Accounts of Chemical Research* 54.2 (2021), pp. 263–270. issn: 0001-4842. doi: 10.1021/acs.accounts.0c00699.
- [76] Brian K. Shoichet. "Virtual screening of chemical libraries". In: *Nature* 432.7019 (2004), pp. 862–865. issn: 0028-0836. doi: 10.1038/nature03197.
- [77] Alex Zhavoronkov et al. "Deep learning enables rapid identification of potent DDR1 kinase inhibitors". In: *Nature Biotechnology* 37.9 (2019), pp. 1038–1040. issn: 1087-0156. doi: 10.1038/s41587-019-0224-x.
- [78] Dr. Roberto Todeschini Prof. and Dr. Viviana Consonni. "Molecular Descriptors for Chemoinformatics". In: *Methods and Principles in Medicinal Chemistry* (2023). doi: 10.1002/9783527628766.
- [79] Adrià Cereto-Massagué et al. "Molecular fingerprint similarity search in virtual screening". In: *Methods* 71 (2015), pp. 58–63. issn: 1046-2023. doi: 10.1016/j.ymeth.2014.08.005.
- [80] Sereina Riniker and Gregory A Landrum. "Open-source platform to benchmark fingerprints for ligand-based virtual screening". In: *Journal of Cheminformatics* 5.1 (2013), p. 26. issn: 1758-2946. doi: 10.1186/1758-2946-5-26.
- [81] Joseph L. Durant et al. "Reoptimization of MDL Keys for Use in Drug Discovery". In: *Journal of Chemical Information and Computer Sciences* 42.6 (2002), pp. 1273–1280. issn: 0095-2338. doi: 10.1021/ci010132r.
- [82] David Rogers and Mathew Hahn. "Extended-Connectivity Fingerprints". In: *Journal of Chemical Information and Modeling* 50.5 (2010), pp. 742–754. issn: 1549-9596. doi: 10.1021/ci100050t.
- [83] Gregory Sliwoski et al. "Computational Methods in Drug Discovery". In: *Pharmacological Reviews* 66.1 (2014), pp. 334–395. issn: 0031-6997. doi: 10.1124/pr.112.007336.
- [84] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (2021), pp. 583–589. issn: 0028-0836. doi: 10.1038/s41586-021-03819-2.
- [85] Artem Cherkasov et al. "QSAR Modeling: Where Have You Been? Where Are You Going To?" In: *Journal of Medicinal Chemistry* 57.12 (2014), pp. 4977–5010. issn: 0022-2623. doi: 10.1021/jm4004285.

- [86] CORWIN HANSCH et al. "Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients". In: *Nature* 194.4824 (1962), pp. 178–180. ISSN: 0028-0836. doi: 10.1038/194178b0.
- [87] Corwin Hansch and Toshio Fujita. "p- σ - π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure". In: *Journal of the American Chemical Society* 86.8 (1964), pp. 1616–1626. ISSN: 0002-7863. doi: 10.1021/ja01062a035.
- [88] Eugene N. Muratov et al. "QSAR without borders". In: *Chemical Society Reviews* 49.11 (2020), pp. 3525–3564. ISSN: 0306-0012. doi: 10.1039/d0cs00098a.
- [89] Bruno J. Neves et al. "QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery". In: *Frontiers in Pharmacology* 9 (2018), p. 1275. ISSN: 1663-9812. doi: 10.3389/fphar.2018.01275.
- [90] Alexander Golbraikh et al. "Data Set Modelability by QSAR". In: *Journal of Chemical Information and Modeling* 54.1 (2014), pp. 1–4. ISSN: 1549-9596. doi: 10.1021/ci400572x.
- [91] Peter Willett. "Similarity-based virtual screening using 2D fingerprints". In: *Drug Discovery Today* 11.23-24 (2006), pp. 1046–1053. ISSN: 1359-6446. doi: 10.1016/j.drudis.2006.10.005.
- [92] Peter Willett. "The Calculation of Molecular Structural Similarity: Principles and Practice". In: *Molecular Informatics* 33.6-7 (2014), pp. 403–413. ISSN: 1868-1743. doi: 10.1002/minf.201400024.
- [93] Gisbert Schneider and David E. Clark. "Automated De Novo Drug Design: Are We Nearly There Yet?" In: *Angewandte Chemie International Edition* 58.32 (2019), pp. 10792–10803. ISSN: 1433-7851. doi: 10.1002/anie.201814681.
- [94] Maxime Langevin et al. "Scaffold-Constrained Molecular Generation". In: *Journal of Chemical Information and Modeling* 60.12 (2020), pp. 5637–5646. ISSN: 1549-9596. doi: 10.1021/acs.jcim.0c01015. eprint: 2009.07778.
- [95] Nathan Brown et al. "GuacaMol: Benchmarking Models for de Novo Molecular Design". In: *Journal of Chemical Information and Modeling* 59.3 (2019), pp. 1096–1108. ISSN: 1549-9596. doi: 10.1021/acs.jcim.8b00839. eprint: 1811.09621.
- [96] Rafael Gómez-Bombarelli et al. "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules". In: *ACS Central Science* 4.2 (2018), pp. 268–276. ISSN: 2374-7943. doi: 10.1021/acscentsci.7b00572. eprint: 1610.02415.
- [97] Marcus Olivecrona et al. "Molecular de-novo design through deep reinforcement learning". In: *Journal of Cheminformatics* 9.1 (2017), p. 48. ISSN: 1758-2946. doi: 10.1186/s13321-017-0235-x.
- [98] Daniel C. Elton et al. "Deep learning for molecular design—a review of the state of the art". In: *Molecular Systems Design & Engineering* 4.4 (2019), pp. 828–849. doi: 10.1039/c9me00039a. eprint: 1903.04388.
- [99] Hannes H. Loeffler et al. "Reinvent 4: Modern AI-driven generative molecule design". In: *Journal of Cheminformatics* 16.1 (2024), p. 20. ISSN: 1758-2946. doi: 10.1186/s13321-024-00812-5.
- [100] Marwin H. S. Segler et al. "Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks". In: *ACS Central Science* 4.1 (2018), pp. 120–131. ISSN: 2374-7943. doi: 10.1021/acscentsci.7b00512.
- [101] Igor V Tetko, Ola Engkvist, and Hongming Chen. "Does Big Data exist in medicinal chemistry, and if so, how can it be harnessed?" In: *Future Medicinal Chemistry* 8.15 (2016), pp. 1801–1806. ISSN: 1756-8919. doi: 10.4155/fmc-2016-0163.
- [102] Kunal Roy, Supratik Kar, and Pravin Ambure. "On a simple approach for determining applicability domain of QSAR models". In: *Chemometrics and Intelligent Laboratory Systems* 145 (2015), pp. 22–29. ISSN: 0169-7439. doi: 10.1016/j.chemolab.2015.04.013.
- [103] Faizan Sahigara et al. "Comparison of Different Approaches to Define the Applicability Domain of QSAR Models". In: *Molecules* 17.5 (2012), pp. 4791–4810. doi: 10.3390/molecules17054791.
- [104] Joanna Jaworska, Nina Nikolova-Jeliazkova, and Tom Aldenberg. "QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review". In: *Alternatives to Laboratory Animals* 33.5 (2005), pp. 445–459. ISSN: 0261-1929. doi: 10.1177/026119290503300508.
- [105] Horvath Dragos, Marcou Gilles, and Varnek Alexandre. "Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models". In: *Journal of Chemical Information and Modeling* 49.7 (2009), pp. 1762–1776. ISSN: 1549-9596. doi: 10.1021/ci9000579.
- [106] Faizan Sahigara et al. "Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions". In: *Journal of Cheminformatics* 5.1 (2013), p. 27. ISSN: 1758-2946. doi: 10.1186/1758-2946-5-27.

- [107] Dagmar Stumpfe and Jürgen Bajorath. "Exploring Activity Cliffs in Medicinal Chemistry". In: *Journal of Medicinal Chemistry* 55.7 (2012), pp. 2932–2942. issn: 0022-2623. doi: 10.1021/jm201706b.
- [108] Dagmar Stumpfe, Huabin Hu, and Jürgen Bajorath. "Evolving Concept of Activity Cliffs". In: *ACS Omega* 4.11 (2019), pp. 14360–14368. issn: 2470-1343. doi: 10.1021/acsomega.9b02221.
- [109] Gerald M. Maggiora. "On Outliers and Activity Cliffs Why QSAR Often Disappoints". In: *Journal of Chemical Information and Modeling* 46.4 (2006), pp. 1535–1535. issn: 1549-9596. doi: 10.1021/ci060117s.
- [110] Derek van Tilborg, Alisa Alenicheva, and Francesca Grisoni. "Exposing the Limitations of Molecular Machine Learning with Activity Cliffs". In: *Journal of Chemical Information and Modeling* (2022). issn: 1549-9596. doi: 10.1021/acs.jcim.2c01073.
- [111] Rajarshi Guha and John H. Van Drie. "Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs". In: *Journal of Chemical Information and Modeling* 48.3 (2008), pp. 646–658. issn: 1549-9596. doi: 10.1021/ci7004093.
- [112] Alexander G. Dossetter, Edward J. Griffen, and Andrew G. Leach. "Matched Molecular Pair Analysis in drug discovery". In: *Drug Discovery Today* 18.15-16 (2013), pp. 724–731. issn: 1359-6446. doi: 10.1016/j.drudis.2013.03.003.
- [113] Xiaoying Hu et al. "MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs". In: *Journal of Chemical Information and Modeling* 52.5 (2012), pp. 1138–1145. issn: 1549-9596. doi: 10.1021/ci3001138.
- [114] Vladimir Svetnik et al. "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling". In: *Journal of Chemical Information and Computer Sciences* 43.6 (2003), pp. 1947–1958. issn: 0095-2338. doi: 10.1021/ci034160g.
- [115] Leo Breiman et al. "Classification And Regression Trees". In: (2017). doi: 10.1201/9781315139470.
- [116] Thomas G. Dietterich. "Multiple Classifier Systems, First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings". In: *Lecture Notes in Computer Science* (2000), pp. 1–15. issn: 0302-9743. doi: 10.1007/3-540-45014-9_1.
- [117] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. issn: 0885-6125. doi: 10.1023/a:1010933404324.
- [118] Robert P. Sheridan et al. "Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships". In: *Journal of Chemical Information and Modeling* 56.12 (2016), pp. 2353–2360. issn: 1549-9596. doi: 10.1021/acs.jcim.6b00591.
- [119] Kellie J. Archer and Ryan V. Kimes. "Empirical characterization of random forest variable importance measures". In: *Computational Statistics & Data Analysis* 52.4 (2008), pp. 2249–2260. issn: 0167-9473. doi: 10.1016/j.csda.2007.08.015.
- [120] Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine." In: *The Annals of Statistics* 29.5 (2001). issn: 0090-5364. doi: 10.1214/aos/1013203451.
- [121] Carl Edward Rasmussen and Christopher K I Williams. "Gaussian Processes for Machine Learning". In: (2005). doi: 10.7551/mitpress/3206.001.0001.
- [122] Olga Obrezanova et al. "Gaussian Processes: A Method for Automatic QSAR Modeling of ADME Properties". In: *Journal of Chemical Information and Modeling* 47.5 (2007), pp. 1847–1857. issn: 1549-9596. doi: 10.1021/ci7000633.
- [123] Hu Mei et al. "Support vector machine applied in QSAR modelling". In: *Chinese Science Bulletin* 50.20 (2005), pp. 2291–2296. issn: 1001-6538. doi: 10.1007/bf03183737.
- [124] Alexander Tropsha. "Best Practices for QSAR Model Development, Validation, and Exploitation". In: *Molecular Informatics* 29.6-7 (2010), pp. 476–488. issn: 1868-1743. doi: 10.1002/minf.201000061.
- [125] Subhash Ajmani, Kamalakar Jadhav, and Sudhir A. Kulkarni. "Three-Dimensional QSAR Using the k-Nearest Neighbor Method and Its Interpretation". In: *Journal of Chemical Information and Modeling* 46.1 (2006), pp. 24–31. issn: 1549-9596. doi: 10.1021/ci0501286.
- [126] Liva Ralaivola et al. "Graph kernels for chemical informatics". In: *Neural Networks* 18.8 (2005), pp. 1093–1110. issn: 0893-6080. doi: 10.1016/j.neunet.2005.07.009.
- [127] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444. issn: 0028-0836. doi: 10.1038/nature14539.
- [128] Hongming Chen et al. "The rise of deep learning in drug discovery". In: *Drug Discovery Today* 23.6 (2018), pp. 1241–1250. issn: 1359-6446. doi: 10.1016/j.drudis.2018.01.039.

- [129] Alexander Tropsha et al. "Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR". In: *Nature Reviews Drug Discovery* 23.2 (2024), pp. 141–155. issn: 1474-1776. doi: 10.1038/s41573-023-00832-0.
- [130] Mingyang Wang et al. "Deep learning approaches for de novo drug design: An overview". In: *Current Opinion in Structural Biology* 72 (2022), pp. 135–144. issn: 0959-440X. doi: 10.1016/j.sbi.2021.10.001.
- [131] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323.6088 (1986), pp. 533–536. issn: 0028-0836. doi: 10.1038/323533a0.
- [132] Justin Gilmer et al. "Neural Message Passing for Quantum Chemistry". In: *arXiv* (2017). doi: 10.48550/arxiv.1704.01212. eprint: 1704.01212.
- [133] Johann Gasteiger and Jure Zupan. "Neural Networks in Chemistry". In: *Angewandte Chemie International Edition in English* 32.4 (1993), pp. 503–527. issn: 0570-0833. doi: 10.1002/anie.199305031.
- [134] D.J. Livingstone, D.T. Manallack, and I.V. Tetko. "Data modelling with neural networks: Advantages and limitations". In: *Journal of Computer-Aided Molecular Design* 11.2 (1997), pp. 135–142. issn: 0920-654X. doi: 10.1023/a:1008074223811.
- [135] Gisbert Schneider and Paul Wrede. "Artificial neural networks for computer-based molecular design". In: *Progress in Biophysics and Molecular Biology* 70.3 (1998), pp. 175–222. issn: 0079-6107. doi: 10.1016/s0079-6107(98)00026-1.
- [136] George E Dahl, Navdeep Jaitly, and Ruslan Salakhutdinov. "Multi-task Neural Networks for QSAR Predictions". In: *arXiv* (2014). doi: 10.48550/arxiv.1406.1231. eprint: 1406.1231.
- [137] Rodolfo S. Simões et al. "Transfer and Multi-task Learning in QSAR Modeling: Advances and Challenges". In: *Frontiers in Pharmacology* 9 (2018), p. 74. issn: 1663-9812. doi: 10.3389/fphar.2018.00074.
- [138] Zhenqin Wu et al. "MoleculeNet: a benchmark for molecular machine learning". In: *Chemical Science* 9.2 (2017), pp. 513–530. issn: 2041-6520. doi: 10.1039/c7sc02664a.
- [139] Zonghan Wu et al. "A Comprehensive Survey on Graph Neural Networks". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.1 (2021), pp. 4–24. issn: 2162-237X. doi: 10.1109/tnnls.2020.2978386. eprint: 1901.00596.
- [140] Geoff M Downs and John M Barnard. "Clustering methods and their uses in computational chemistry". In: *Reviews in computational chemistry* 18 (2002), pp. 1–40.
- [141] Thibault Varin et al. "Mining for Bioactive Scaffolds with Scaffold Networks: Improved Compound Set Enrichment from Primary Screening Data". In: *Journal of Chemical Information and Modeling* 51.7 (2011), pp. 1528–1538. issn: 1549-9596. doi: 10.1021/ci2000924.
- [142] Peter Willett. "Similarity-based data mining in files of two-dimensional chemical structures using fingerprint measures of molecular resemblance". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.3 (2011), pp. 241–251. issn: 1942-4787. doi: 10.1002/widm.26.
- [143] Mohamed G. Malhat, Hamdy M. Mousa, and Ashraf B. Ei-Sisi. "Clustering of Chemical Data Sets for Drug Discovery". In: *2014 9th International Conference on Informatics and Systems* (2014), DEKM-11-DEKM-18. doi: 10.1109/infos.2014.7036702.
- [144] J. B. MacQueen. "Some Methods for Classification and Analysis of MultiVariate Observations". In: *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by L. M. Le Cam and J. Neyman. Vol. 1. University of California Press, 1967, pp. 281–297.
- [145] Leonard Kaufman and Peter J. Rousseeuw. "Finding Groups in Data". In: *Wiley Series in Probability and Statistics* (2010), pp. 68–125. doi: 10.1002/9780470316801.ch2.
- [146] Darko Butina. "Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets". In: *Journal of Chemical Information and Computer Sciences* 39.4 (1999), pp. 747–750. issn: 0095-2338. doi: 10.1021/ci9803381.
- [147] Karl Pearson F.R.S. "LIII. On lines and planes of closest fit to systems of points in space". In: *Philosophical Magazine Series 6* 2.11 (1901), pp. 559–572. issn: 1941-5982. doi: 10.1080/14786440109462720.
- [148] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.
- [149] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *arXiv* (2018). doi: 10.48550/arxiv.1802.03426. eprint: 1802.03426.

- [150] Daniel Probst and Jean-Louis Reymond. "Visualization of very large high-dimensional data sets as minimum spanning trees". In: *Journal of Cheminformatics* 12.1 (2020), p. 12. issn: 1758-2946. doi: 10.1186/s13321-020-0416-x.
- [151] Daniel Reker and Gisbert Schneider. "Active-learning strategies in computer-assisted drug discovery". In: *Drug Discovery Today* 20.4 (2015), pp. 458–465. issn: 1359-6446. doi: 10.1016/j.drudis.2014.12.004.
- [152] Daniel Reker. "Practical considerations for active machine learning in drug discovery". In: *Drug Discovery Today: Technologies* 32-33 (2020), pp. 73–79. issn: 1740-6749. doi: 10.1016/j.ddtec.2020.06.001.
- [153] Michael Bailey et al. "Deep Batch Active Learning for Drug Discovery". In: *bioRxiv* (2023), p. 2023.07.26.550653. doi: 10.1101/2023.07.26.550653.
- [154] D. Reker, P. Schneider, and G. Schneider. "Multi-objective active machine learning rapidly improves structure–activity models and reveals new protein–protein interaction inhibitors". In: *Chemical Science* 7.6 (2016), pp. 3919–3927. issn: 2041-6520. doi: 10.1039/c5sc04272k.
- [155] Steven S. Wesolowski and Dean G. Brown. "Lead Generation". In: *Methods and Principles in Medicinal Chemistry* (2016), pp. 487–512. doi: 10.1002/9783527677047.ch17.
- [156] Alleyn T. Plowright et al. "Hypothesis driven drug design: improving quality and effectiveness of the design-make-test-analyse cycle". In: *Drug Discovery Today* 17.1-2 (2012), pp. 56–62. issn: 1359-6446. doi: 10.1016/j.drudis.2011.09.012.
- [157] Mindy I Davis et al. "Comprehensive analysis of kinase inhibitor selectivity". In: *Nature Biotechnology* 29.11 (2011), pp. 1046–1051. issn: 1087-0156. doi: 10.1038/nbt.1990.
- [158] Susan Klaeger et al. "The target landscape of clinical kinase drugs". In: *Science* 358.6367 (2017). issn: 0036-8075. doi: 10.1126/science.aan4368.
- [159] James J. Lynch et al. "Potential functional and pathological side effects related to off-target pharmacological activity". In: *Journal of Pharmacological and Toxicological Methods* 87 (2017), pp. 108–126. issn: 1056-8719. doi: 10.1016/j.vascn.2017.02.020.
- [160] Andrew Anighoro, Jürgen Bajorath, and Giulio Rastelli. "Polypharmacology: Challenges and Opportunities in Drug Discovery". In: *Journal of Medicinal Chemistry* 57.19 (2014), pp. 7874–7887. issn: 0022-2623. doi: 10.1021/jm5006463.
- [161] Jens-Uwe Peters. "Polypharmacology – Foe or Friend?" In: *Journal of Medicinal Chemistry* 56.22 (2013), pp. 8955–8971. issn: 0022-2623. doi: 10.1021/jm400856t.
- [162] M. Paul Gleeson. "Generation of a Set of Simple, Interpretable ADMET Rules of Thumb". In: *Journal of Medicinal Chemistry* 51.4 (2008), pp. 817–834. issn: 0022-2623. doi: 10.1021/jm701122q.
- [163] Michael M. Hann. "Molecular obesity, potency and other addictions in drug discovery". In: *MedChemComm* 2.5 (2011), pp. 349–355. issn: 2040-2503. doi: 10.1039/c1md00017a.
- [164] Nicholas A. Meanwell. "Improving Drug Design: An Update on Recent Applications of Efficiency Metrics, Strategies for Replacing Problematic Elements, and Compounds in Nontraditional Drug Space". In: *Chemical Research in Toxicology* 29.4 (2016), pp. 564–616. issn: 0893-228x. doi: 10.1021/acs.chemrestox.6b00043.
- [165] David C. Blakemore et al. "Organic synthesis provides opportunities to transform drug discovery". In: *Nature Chemistry* 10.4 (2018), pp. 383–394. issn: 1755-4330. doi: 10.1038/s41557-018-0021-z.
- [166] Wenhao Gao and Connor W. Coley. "The Synthesizability of Molecules Proposed by Generative Models". In: *Journal of Chemical Information and Modeling* 60.12 (2020), pp. 5714–5723. issn: 1549-9596. doi: 10.1021/acs.jcim.0c00174.
- [167] Connor W. Coley et al. "Computer-Assisted Retrosynthesis Based on Molecular Similarity". In: *ACS Central Science* 3.12 (2017), pp. 1237–1245. issn: 2374-7943. doi: 10.1021/acscentsci.7b00355.
- [168] Connor W. Coley et al. "A robotic platform for flow synthesis of organic compounds informed by AI planning". In: *Science* 365.6453 (2019). issn: 0036-8075. doi: 10.1126/science.aax1566.
- [169] Marwin H. S. Segler, Mike Preuss, and Mark P. Waller. "Planning chemical syntheses with deep neural networks and symbolic AI". In: *Nature* 555.7698 (2018), pp. 604–610. issn: 0028-0836. doi: 10.1038/nature25978. eprint: 1708.04202.
- [170] Peter Ertl and Ansgar Schuffenhauer. "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions". In: *Journal of Cheminformatics* 1.1 (2009), p. 8. doi: 10.1186/1758-2946-1-8.
- [171] Connor W. Coley et al. "SCScore: Synthetic Complexity Learned from a Reaction Corpus". In: *Journal of Chemical Information and Modeling* 58.2 (2018), pp. 252–261. issn: 1549-9596. doi: 10.1021/acs.jcim.7b00622.

- [172] Amol Thakkar et al. "Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning". In: *Chemical Science* 12.9 (2021), pp. 3339–3349. ISSN: 2041-6520. DOI: 10.1039/d0sc05401a.
- [173] Camille Bilodeau et al. "Generative models for molecular discovery: Recent advances and challenges". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* (2022). ISSN: 1759-0876. DOI: 10.1002/wcms.1608.
- [174] Veronika Temml and Zsofia Kutil. "Structure-based molecular modeling in SAR analysis and lead optimization". In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 1431–1444. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2021.02.018.
- [175] Laurent Gomez. "Decision Making in Medicinal Chemistry: The Power of Our Intuition". In: *ACS Medicinal Chemistry Letters* 9.10 (2018), pp. 956–958. ISSN: 1948-5875. DOI: 10.1021/acsmmedchemlett.8b00359.
- [176] Jack W. Scannell et al. "Diagnosing the decline in pharmaceutical R&D efficiency". In: *Nature Reviews Drug Discovery* 11.3 (2012), pp. 191–200. ISSN: 1474-1776. DOI: 10.1038/nrd3681.
- [177] Maxime Langevin, Marc Bianciotto, and Rodolphe Vuilleumier. "Balancing exploration and exploitation in de novo drug design". In: *Digital Discovery* (2024). DOI: 10.1039/d4dd00105b.
- [178] Martin J. Valler et al. "Diversity screening versus focussed screening in drug discovery". In: *Drug Discovery Today* 5.7 (2000), pp. 286–293. ISSN: 1359-6446. DOI: 10.1016/s1359-6446(00)01517-8.
- [179] Iryna Yevseyeva et al. "Application of portfolio optimization to drug discovery". In: *Information Sciences* 475 (2019), pp. 29–43. ISSN: 0020-0255. DOI: 10.1016/j.ins.2018.09.049.
- [180] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 248–255. ISSN: 1063-6919. DOI: 10.1109/cvpr.2009.5206848.
- [181] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: *OpenAI* (2019). Accessed: 2024-11-15. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
- [182] Hyunho Kim et al. "Artificial Intelligence in Drug Discovery: A Comprehensive Review of Data-driven and Machine Learning Approaches". In: *Biotechnology and Bioprocess Engineering* 25.6 (2020), pp. 895–930. ISSN: 1226-8372. DOI: 10.1007/s12257-020-0049-y.
- [183] Cas Wognum et al. "A call for an industry-led initiative to critically assess machine learning for real-world drug discovery". In: *Nature Machine Intelligence* 6.10 (2024), pp. 1120–1121. DOI: 10.1038/s42256-024-00911-w.
- [184] Gregory A. Landrum et al. "SIMPd: an algorithm for generating simulated time splits for validating machine learning approaches". In: *Journal of Cheminformatics* 15.1 (2023), p. 119. ISSN: 1758-2946. DOI: 10.1186/s13321-023-00787-9.
- [185] Kunal Roy, Supratik Kar, and Rudra Narayan Das. "A Primer on QSAR/QSPR Modeling, Fundamental Concepts". In: *SpringerBriefs in Molecular Science* (2015). ISSN: 2191-5407. DOI: 10.1007/978-3-319-17281-1.
- [186] Dr. Roberto Todeschini Prof. and Dr. Viviana Consonni. "Molecular Descriptors for Chemoinformatics". In: *Methods and Principles in Medicinal Chemistry* (2023). DOI: 10.1002/9783527628766.
- [187] "Handbook of Computational Chemistry". In: (2017). DOI: 10.1007/978-3-319-27282-5.
- [188] Artem Cherkasov et al. "QSAR Modeling: Where Have You Been? Where Are You Going To?" In: *Journal of Medicinal Chemistry* 57.12 (2014), pp. 4977–5010. ISSN: 0022-2623. DOI: 10.1021/jm4004285.
- [189] Alexander Tropsha. "Best Practices for QSAR Model Development, Validation, and Exploitation". In: *Molecular Informatics* 29.6-7 (2010), pp. 476–488. ISSN: 1868-1743. DOI: 10.1002/minf.201000061.
- [190] Hannah Rosa Friesacher et al. "Temporal Distribution Shift in Real-World Pharmaceutical Data: Implications for Uncertainty Quantification in QSAR Models". In: *arXiv* (2025). DOI: 10.48550/arxiv.2502.03982. eprint: 2502.03982.
- [191] Robert P. Sheridan et al. "Experimental Error, Kurtosis, Activity Cliffs, and Methodology: What Limits the Predictivity of Quantitative Structure–Activity Relationship Models?" In: *Journal of Chemical Information and Modeling* 60.4 (2020), pp. 1969–1982. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.9b01067.
- [192] Linlin Zhao et al. "Experimental Errors in QSAR Modeling Sets: What We Can Do and What We Cannot Do". In: *ACS Omega* 2.6 (2017), pp. 2805–2812. ISSN: 2470-1343. DOI: 10.1021/acsomega.7b00274.
- [193] Rodolfo S. Simões et al. "Transfer and Multi-task Learning in QSAR Modeling: Advances and Challenges". In: *Frontiers in Pharmacology* 9 (2018), p. 74. ISSN: 1663-9812. DOI: 10.3389/fphar.2018.00074.

- [194] Anna Gaulton et al. "ChEMBL: a large-scale bioactivity database for drug discovery". In: *Nucleic Acids Research* 40.D1 (2012), pp. D1100–D1107. ISSN: 0305-1048. doi: 10.1093/nar/gkr777.
- [195] Sunghwan Kim et al. "PubChem Substance and Compound databases". In: *Nucleic Acids Research* 44.D1 (2016), pp. D1202–D1213. ISSN: 0305-1048. doi: 10.1093/nar/gkv951.
- [196] Kexin Huang et al. "Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development". In: *arXiv* (2021). doi: 10.48550/arxiv.2102.09548. eprint: 2102.09548.
- [197] Zhenqin Wu et al. "MoleculeNet: a benchmark for molecular machine learning". In: *Chemical Science* 9.2 (2017), pp. 513–530. ISSN: 2041-6520. doi: 10.1039/c7sc02664a.
- [198] K. Mansouri et al. "An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling". In: *SAR and QSAR in Environmental Research* 27.11 (2016), pp. 911–937. ISSN: 1062-936X. doi: 10.1080/1062936x.2016.1253611.
- [199] F. Lunghini et al. "Consensus QSAR models estimating acute toxicity to aquatic organisms from different trophic levels: algae, Daphnia and fish". In: *SAR and QSAR in Environmental Research* 31.9 (2020), pp. 655–675. ISSN: 1062-936X. doi: 10.1080/1062936x.2020.1797872.
- [200] Ruifeng Liu, Patric Schyman, and Anders Wallqvist. "Critically Assessing the Predictive Power of QSAR Models for Human Liver Microsomal Stability". In: *Journal of Chemical Information and Modeling* 55.8 (2015), pp. 1566–1575. ISSN: 1549-9596. doi: 10.1021/acs.jcim.5b00255.
- [201] Prudencio Tossou et al. "Real-World Molecular Out-Of-Distribution: Specification and Investigation". In: *Journal of Chemical Information and Modeling* 64.3 (2024), pp. 697–711. ISSN: 1549-9596. doi: 10.1021/acs.jcim.3c01774.
- [202] Thereza A. Soares et al. "The (Re)-Evolution of Quantitative Structure–Activity Relationship (QSAR) Studies Propelled by the Surge of Machine Learning Methods". In: *Journal of Chemical Information and Modeling* 62.22 (2022), pp. 5317–5320. ISSN: 1549-9596. doi: 10.1021/acs.jcim.2c01422.
- [203] Weifan Zheng and Alexander Tropsha. "Novel Variable Selection Quantitative Structure–Property Relationship Approach Based on the k-Nearest-Neighbor Principle". In: *Journal of Chemical Information and Computer Sciences* 40.1 (2000), pp. 185–194. ISSN: 0095-2338. doi: 10.1021/ci980033m.
- [204] H. X. Liu et al. "QSAR Study of Ethyl 2-[(3-Methyl-2,5-dioxo(3-pyrrolinyl)amino]-4-(trifluoromethyl) pyrimidine-5-carboxylate: An Inhibitor of AP-1 and NF-KB Mediated Gene Expression Based on Support Vector Machines". In: *Journal of Chemical Information and Computer Sciences* 43.4 (2003), pp. 1288–1296. ISSN: 0095-2338. doi: 10.1021/ci0340355.
- [205] Vladimir Svetnik et al. "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling". In: *Journal of Chemical Information and Computer Sciences* 43.6 (2003), pp. 1947–1958. ISSN: 0095-2338. doi: 10.1021/ci034160g.
- [206] Pavel Polishchuk. "Interpretation of Quantitative Structure–Activity Relationship Models: Past, Present, and Future". In: *Journal of Chemical Information and Modeling* 57.11 (2017), pp. 2618–2639. ISSN: 1549-9596. doi: 10.1021/acs.jcim.7b00274.
- [207] Vladimir Svetnik et al. "Boosting: An Ensemble Learning Tool for Compound Classification and QSAR Modeling". In: *Journal of Chemical Information and Modeling* 45.3 (2005), pp. 786–799. ISSN: 1549-9596. doi: 10.1021/ci0500379.
- [208] Günter Klambauer, Sepp Hochreiter, and Matthias Rarey. "Machine Learning in Drug Discovery". In: *Journal of Chemical Information and Modeling* 59.3 (2019), pp. 945–946. ISSN: 1549-9596. doi: 10.1021/acs.jcim.9b00136.
- [209] Esther Heid et al. "Chemprop: A Machine Learning Package for Chemical Property Prediction". In: *Journal of Chemical Information and Modeling* 64.1 (2024), pp. 9–17. ISSN: 1549-9596. doi: 10.1021/acs.jcim.3c01250.
- [210] Junshui Ma et al. "Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships". In: *Journal of Chemical Information and Modeling* 55.2 (2015), pp. 263–274. ISSN: 1549-9596. doi: 10.1021/ci500747n.
- [211] Eugene N. Muratov et al. "QSAR without borders". In: *Chemical Society Reviews* 49.11 (2020), pp. 3525–3564. ISSN: 0306-0012. doi: 10.1039/d0cs00098a.
- [212] Alexander Tropsha et al. "Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR". In: *Nature Reviews Drug Discovery* 23.2 (2024), pp. 141–155. ISSN: 1474-1776. doi: 10.1038/s41573-023-00832-0.
- [213] Derek van Tilborg et al. "Deep learning for low-data drug discovery: Hurdles and opportunities". In: *Current Opinion in Structural Biology* 86 (2024), p. 102818. ISSN: 0959-440X. doi: 10.1016/j.sbi.2024.102818.

- [214] Bozheng Dou et al. "Machine Learning Methods for Small Data Challenges in Molecular Science". In: *Chemical Reviews* 123.13 (2023), pp. 8736–8780. issn: 0009-2665. doi: 10.1021/acs.chemrev.3c00189.
- [215] David Duvenaud et al. "Convolutional Networks on Graphs for Learning Molecular Fingerprints". In: *arXiv* (2015). doi: 10.48550/arxiv.1509.09292. eprint: 1509.09292.
- [216] Pat Walters. *Some Thoughts on Splitting Chemical Datasets*. 2024. url: <https://practicalcheminformatics.blogspot.com/2024/11/some-thoughts-on-splitting-chemical.html>.
- [217] Robert P. Sheridan. "Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction." In: *Journal of Chemical Information and Modeling* 53.4 (2013), pp. 783–790. issn: 1549-9596. doi: 10.1021/ci400084k.
- [218] Malkeet Singh Bahia et al. "A comparison between 2D and 3D descriptors in QSAR modeling based on bio-active conformations". In: *Molecular Informatics* 42.4 (2023), e2200186. issn: 1868-1743. doi: 10.1002/minf.202200186.
- [219] Danishuddin and Asad U. Khan. "Descriptors and their selection methods in QSAR analysis: paradigm for drug design". In: *Drug Discovery Today* 21.8 (2016), pp. 1291–1302. issn: 1359-6446. doi: 10.1016/j.drudis.2016.06.013.
- [220] Daniel Probst and Jean-Louis Reymond. "A probabilistic molecular fingerprint for big data settings". In: *Journal of Cheminformatics* 10.1 (2018), p. 66. issn: 1758-2946. doi: 10.1186/s13321-018-0321-8.
- [221] "RDKit: Open-source cheminformatics. <https://www.rdkit.org>". In: (). doi: 10.5281/zenodo.591637.
- [222] Ramaswamy Nilakantan et al. "Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors". In: *Journal of Chemical Information and Computer Sciences* 27.2 (1987), pp. 82–85. issn: 0095-2338. doi: 10.1021/ci00054a008.
- [223] Raymond E Carhart, Dennis H Smith, and R Venkataraghavan. "Atom pairs as molecular features in structure-activity studies: definition and applications". In: *Journal of Chemical Information and Computer Sciences* 25.2 (1985), pp. 64–73. issn: 0095-2338. doi: 10.1021/ci00046a002.
- [224] Hiroto Moriaki et al. "Mordred: a molecular descriptor calculator". In: *Journal of Cheminformatics* 10.1 (2018), p. 4. issn: 1758-2946. doi: 10.1186/s13321-018-0258-y.
- [225] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *arXiv* (2016). doi: 10.48550/arxiv.1603.02754. eprint: 1603.02754.
- [226] Guolin Ke et al. "LightGBM: a highly efficient gradient boosting decision tree". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17*. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 3149–3157. isbn: 9781510860964.
- [227] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. issn: 0885-6125. doi: 10.1023/a:1010933404324.
- [228] Lagnajit Pattanaik and Connor W. Coley. "Molecular Representation: Going Long on Fingerprints". In: *Chem* 6.6 (2020), pp. 1204–1207. issn: 2451-9294. doi: 10.1016/j.chempr.2020.05.002.
- [229] Bimbisai Desai et al. "Rapid Discovery of a Novel Series of Abl Kinase Inhibitors by Application of an Integrated Microfluidic Synthesis and Screening Platform". In: *Journal of Medicinal Chemistry* 56.7 (2013), pp. 3033–3047. issn: 0022-2623. doi: 10.1021/jm400099d.
- [230] Gabriel H.S. Dreiman et al. "Changing the HTS Paradigm: AI-Driven Iterative Screening for Hit Finding". In: *SLAS Discovery* 26.2 (2021), pp. 257–262. issn: 2472-5552. doi: 10.1177/2472555220949495.
- [231] Clemens Isert et al. "Machine Learning for Fast, Quantum Mechanics-Based Approximation of Drug Lipophilicity". In: *ACS Omega* 8.2 (2023), pp. 2046–2056. issn: 2470-1343. doi: 10.1021/acsomega.2c05607.
- [232] Kyle Swanson et al. "ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries". In: *Bioinformatics* 40.7 (2024), btae416. issn: 1367-4803. doi: 10.1093/bioinformatics/btae416.
- [233] Kenneth Goossens et al. "ChemProp multi-task models for predicting ADME properties in the Polaris challenge". In: (2025). doi: 10.26434/chemrxiv-2025-q12vh.
- [234] Emma Svensson et al. "AI in Drug Discovery, First International Workshop, AIDD 2024, Held in Conjunction with ICANN 2024, Lugano, Switzerland, September 19, 2024, Proceedings". In: *Lecture Notes in Computer Science* (2024), pp. 132–148. issn: 0302-9743. doi: 10.1007/978-3-031-72381-0_11.
- [235] W. Patrick Walters and Mark Murcko. "Assessing the impact of generative AI on medicinal chemistry". In: *Nature Biotechnology* 38.2 (2020), pp. 143–145. issn: 1087-0156. doi: 10.1038/s41587-020-0418-2.

- [236] Hampus Gummeson Svensson et al. "Autonomous Drug Design with Multi-armed Bandits". In: *arXiv* (2022). doi: 10.48550/arxiv.2207.01393. eprint: 2207.01393.
- [237] Manfred K. Warmuth et al. "Active Learning with Support Vector Machines in the Drug Discovery Process". In: *Journal of Chemical Information and Computer Sciences* 43.2 (2003), pp. 667–673. ISSN: 0095-2338. doi: 10.1021/ci025620t.
- [238] Yuriy Khalak et al. "Chemical Space Exploration with Active Learning and Alchemical Free Energies". In: *Journal of Chemical Theory and Computation* 18.10 (2022), pp. 6259–6270. ISSN: 1549-9618. doi: 10.1021/acs.jctc.2c00752.
- [239] Filipp Gusev et al. "Active Learning Guided Drug Design Lead Optimization Based on Relative Binding Free Energy Modeling". In: *Journal of Chemical Information and Modeling* 63.2 (2023), pp. 583–594. ISSN: 1549-9596. doi: 10.1021/acs.jcim.2c01052.
- [240] Justin S. Smith et al. "Less is more: Sampling chemical space with active learning". In: *The Journal of Chemical Physics* 148.24 (2018), p. 241733. ISSN: 0021-9606. doi: 10.1063/1.5023802. eprint: 1801.09319.
- [241] Viet-Khoa Tran-Nguyen, Célie Jacquemard, and Didier Rognan. "LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening". In: *Journal of Chemical Information and Modeling* 60.9 (2020), pp. 4263–4273. ISSN: 1549-9596. doi: 10.1021/acs.jcim.0c00155.
- [242] Florian Häse, Loïc M. Roch, and Alán Aspuru-Guzik. "Next-Generation Experimentation with Self-Driving Laboratories". In: *Trends in Chemistry* 1.3 (2019), pp. 282–291. ISSN: 2589-5974. doi: 10.1016/j.trechm.2019.02.007.
- [243] Connor W. Coley, Natalie S. Eyke, and Klavs F. Jensen. "Autonomous Discovery in the Chemical Sciences Part I: Progress". In: *Angewandte Chemie International Edition* 59.51 (2020), pp. 22858–22893. ISSN: 1433-7851. doi: 10.1002/anie.201909987. eprint: 2003.13754.
- [244] Connor W. Coley, Natalie S. Eyke, and Klavs F. Jensen. "Autonomous Discovery in the Chemical Sciences Part II: Outlook". In: *Angewandte Chemie International Edition* 59.52 (2020), pp. 23414–23436. ISSN: 1433-7851. doi: 10.1002/anie.201909989. eprint: 2003.13755.
- [245] Thomas Blaschke et al. "REINVENT 2.0: An AI Tool for De Novo Drug Design". In: *Journal of Chemical Information and Modeling* 60.12 (2020), pp. 5918–5922. ISSN: 1549-9596. doi: 10.1021/acs.jcim.0c00915.
- [246] Philippe Gendreau et al. "Molecular Assays Simulator to Unravel Predictors Hacking in Goal-Directed Molecular Generations". In: *Journal of Chemical Information and Modeling* 63.13 (2023), pp. 3983–3998. ISSN: 1549-9596. doi: 10.1021/acs.jcim.3c00195.
- [247] Hans Matter et al. "Structural Requirements for Factor Xa Inhibition by 3-Oxybenzamides with Neutral P1 Substituents: Combining X-ray Crystallography, 3D-QSAR, and Tailored Scoring Functions". In: *Journal of Medicinal Chemistry* 48.9 (2005), pp. 3290–3312. ISSN: 0022-2623. doi: 10.1021/jm049187l.
- [248] Marc Nazaré et al. "Factor Xa inhibitors based on a 2-carboxyindole scaffold: SAR of neutral P1 substituents". In: *Bioorganic & Medicinal Chemistry Letters* 14.16 (2004), pp. 4191–4195. ISSN: 0960-894X. doi: 10.1016/j.bmcl.2004.06.020.
- [249] Hans Matter et al. "Design and Quantitative Structure-Activity Relationship of 3-Amidinobenzyl-1H-indole-2-carboxamides as Potent, Nonchiral, and Selective Inhibitors of Blood Coagulation Factor Xa". In: *Journal of Medicinal Chemistry* 45.13 (2002), pp. 2749–2769. ISSN: 0022-2623. doi: 10.1021/jm0111346.
- [250] Marc Nazaré et al. "Probing the Subpockets of Factor Xa Reveals Two Binding Modes for Inhibitors Based on a 2-Carboxyindole Scaffold: A Study Combining Structure-Activity Relationship and X-ray Crystallography". In: *Journal of Medicinal Chemistry* 48.14 (2005), pp. 4511–4525. ISSN: 0022-2623. doi: 10.1021/jm0490540.
- [251] Marc Nazaré et al. "Fragment Deconstruction of Small, Potent Factor Xa Inhibitors: Exploring the Superadditivity Energetics of Fragment Linking in Protein–Ligand Complexes". In: *Angewandte Chemie International Edition* 51.4 (2012), pp. 905–911. ISSN: 1433-7851. doi: 10.1002/anie.201107091.
- [252] Bodo Scheiper et al. "Discovery and optimization of a new class of potent and non-chiral indole-3-carboxamide-based renin inhibitors". In: *Bioorganic & Medicinal Chemistry Letters* 20.21 (2010), pp. 6268–6272. ISSN: 0960-894X. doi: 10.1016/j.bmcl.2010.08.092.
- [253] Hans Matter et al. "Structure-based design and optimization of potent renin inhibitors on 5- or 7-azaindole-scaffolds". In: *Bioorganic & Medicinal Chemistry Letters* 21.18 (2011), pp. 5487–5492. ISSN: 0960-894X. doi: 10.1016/j.bmcl.2011.06.112.
- [254] Bodo Scheiper et al. "Structure-based optimization of potent 4- and 6-azaindole-3-carboxamides as renin inhibitors". In: *Bioorganic & Medicinal Chemistry Letters* 21.18 (2011), pp. 5480–5486. ISSN: 0960-894X. doi: 10.1016/j.bmcl.2011.06.114.

- [255] Thomas L. Saaty. "A scaling method for priorities in hierarchical structures". In: *Journal of Mathematical Psychology* 15.3 (1977), pp. 234–281. ISSN: 0022-2496. DOI: 10.1016/0022-2496(77)90033-5.
- [256] George Derringer and Ronald Suich. "Simultaneous Optimization of Several Response Variables". In: *Journal of Quality Technology* 12.4 (1980), pp. 214–219. ISSN: 0022-4065. DOI: 10.1080/00224065.1980.11980968.
- [257] David J. Cummins and Michael A. Bell. "Integrating Everything: The Molecule Selection Toolkit, a System for Compound Prioritization in Drug Discovery". In: *Journal of Medicinal Chemistry* 59.15 (2016), pp. 6999–7010. ISSN: 0022-2623. DOI: 10.1021/acs.jmedchem.5b01338.
- [258] Andreas Krause and Daniel Golovin. *Tractability*. 1st edn. Cambridge University Press, 2014, pp. 71–104.
- [259] Daniel J. Woodward, Anthony R. Bradley, and Willem P. van Hoorn. "Coverage Score: A Model Agnostic Method to Efficiently Explore Chemical Space". In: *Journal of Chemical Information and Modeling* (2022). ISSN: 1549-9596. DOI: 10.1021/acs.jcim.2c00258.
- [260] Robert P. Sheridan et al. "Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR". In: *Journal of Chemical Information and Computer Sciences* 44.6 (2004), pp. 1912–1928. ISSN: 0095-2338. DOI: 10.1021/ci049782w.
- [261] Deng Ju-Long. "Control problems of grey systems". In: *Systems & Control Letters* 1.5 (1982), pp. 288–294. ISSN: 0167-6911. DOI: 10.1016/S0167-6911(82)80025-X.
- [262] R. W. Kennard and L. A. Stone. "Computer Aided Design of Experiments". In: *Technometrics* 11.1 (1969), pp. 137–148. DOI: 10.1080/00401706.1969.10490666.
- [263] Ching-Lai Hwang and Kwangsun Yoon. "Multiple Attribute Decision Making, Methods and Applications A State-of-the-Art Survey". In: *Lecture Notes in Economics and Mathematical Systems* (1981). ISSN: 0075-8442. DOI: 10.1007/978-3-642-48318-9.
- [264] Jakub Adamczyk and Piotr Ludynia. "Scikit-fingerprints: Easy and efficient computation of molecular fingerprints in Python". In: *SoftwareX* 28 (2024), p. 101944. ISSN: 2352-7110. DOI: <https://doi.org/10.1016/j.softx.2024.101944>. URL: <https://www.sciencedirect.com/science/article/pii/S2352711024003145>.
- [265] Kalyanmoy Deb et al. "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II". In: *IEEE Transactions on Evolutionary Computation* 6.2 (2002), p. 182. ISSN: 1089-778X. DOI: 10.1109/4235.996017.
- [266] Richard E. Higgs et al. "Experimental Designs for Selecting Molecules from Large Chemical Databases". In: *Journal of Chemical Information and Computer Sciences* 37.5 (1997), pp. 861–870. ISSN: 0095-2338. DOI: 10.1021/ci9702858.
- [267] Ching-Lai Hwang and Kwangsun Yoon. "Multiple Attribute Decision Making, Methods and Applications A State-of-the-Art Survey". In: *Lecture Notes in Economics and Mathematical Systems* (1981). ISSN: 0075-8442. DOI: 10.1007/978-3-642-48318-9.
- [268] Nicola De Cao and Thomas Kipf. "MolGAN: An implicit generative model for small molecular graphs". In: *arXiv* (2018). DOI: 10.48550/arxiv.1805.11973. eprint: 1805.11973.
- [269] Philipp Renz, Sohvi Luukkonen, and Günter Klambauer. "Diverse Hits in De Novo Molecule Design: Diversity-Based Comparison of Goal-Directed Generators". In: *Journal of Chemical Information and Modeling* 64.15 (2024), pp. 5756–5761. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.4c00519.
- [270] Jie Zhang et al. "Comparative Study of Deep Generative Models on Chemical Space Coverage". In: *Journal of Chemical Information and Modeling* 61.6 (2021), pp. 2572–2581. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.0c01328.
- [271] Greg Landrum. *Thresholds for "random" in fingerprints the RDKit supports – RDKit blog*. URL: <https://greglandrum.github.io/rdkit-blog/posts/2021-05-18-fingerprint-thresholds1.html>.
- [272] Michael N. Katehakis and Arthur F. Veinott. "The Multi-Armed Bandit Problem: Decomposition and Computation". In: *Mathematics of Operations Research* 12.2 (1987), pp. 262–268. ISSN: 0364-765X. DOI: 10.1287/moor.12.2.262.
- [273] Ruben Sanchez-Garcia et al. "CoPriNet: graph neural networks provide accurate and rapid compound price prediction for molecule prioritisation". In: *Digital Discovery* 2.1 (2022), pp. 103–111. DOI: 10.1039/d2dd00071g.
- [274] Samuel Genheden et al. "AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning". In: *Journal of Cheminformatics* 12.1 (2020), p. 70. ISSN: 1758-2946. DOI: 10.1186/s13321-020-00472-1.
- [275] Massina Abderrahmane et al. "Predicting the price of molecules using their predicted synthetic pathways". In: (2024). DOI: 10.26434/chemrxiv-2024-8wcfp.

Stratégies rationnelles d'exploration/exploitation pour la découverte de médicaments en régime de faibles données

Mots clefs: Intelligence artificielle ; Recherche pharmaceutique ; Optimisation de lead ; Apprentissage actif

Cette thèse, réalisée en partenariat avec Sanofi, explore comment l'intelligence artificielle et les méthodes numériques peuvent être intégrées efficacement à la découverte de médicaments dans un contexte industriel. Centrée sur l'optimisation des leads, étape critique souvent peu étudiée en dehors de l'industrie, elle compare des approches traditionnelles de machine learning avec les modèles récents basés sur les réseaux de neurones, évaluant leur pertinence dans des conditions réalistes. Une boîte à outils computationnelle innovante, intégrant apprentissage actif et aide multicritère à la décision, est développée pour faciliter la sélection rationnelle des composés. Validée rétrospectivement sur des données industrielles historiques enrichies et mises à disposition du public, cette approche démontre des avantages stratégiques concrets, tout en soulignant les défis restants à relever pour généraliser l'usage de ces outils dans l'industrie pharmaceutique.

Rational Exploration/Exploitation Strategies for Drug Discovery in Low-Data Regimes

Mots clefs: Artificial intelligence ; Drug discovery ; Lead optimization ; Active learning

This thesis, conducted in partnership with Sanofi, investigates how artificial intelligence and computational methods can be effectively integrated into drug discovery within an industrial context. Focusing on lead optimization, a crucial yet often underestimated phase, it compares traditional machine learning approaches to recent neural network-based methods, evaluating their practical relevance under realistic conditions. An innovative computational toolkit, combining active learning and multi-criteria decision support, is developed to rationally guide compound selection. Retrospectively validated using historical industrial datasets, enhanced and made publicly available, this approach demonstrates concrete strategic benefits, while clearly identifying the challenges that remain to broaden the adoption of these tools across the pharmaceutical industry.