Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5dd00387c

Back to the future of lead optimization: benchmarking compound prioritization strategies

Pablo Mas,[ab] Bruno Filoche-Rommé,[b] Marc Bianciotto *[b] and Rodolphe Vuilleumier *[a]

Drug discovery requires traversing vast chemical spaces to identify compounds exhibiting favorable potency, selectivity and absorption–distribution–metabolism–excretion–toxicity (ADMET) profiles. During this process, the synthetic and assay throughput is generally markedly lower than the ideation of new propositions by the project team members, so that the prioritization of new syntheses is indispensable. We herein introduce a framework for simulating the outcome of multi-objective prioritization strategies during lead optimization. Based on the Design–Make–Test–Analyze (DMTA) paradigm, historical discovery programs are replayed round by round using user-defined compound selection strategies. We develop qualitative and quantitative tools to assess their performance in retrieving the best compounds and exploring the project's chemical space. We demonstrate our pipeline using four industrial datasets, each containing chemical structures, assay values and time stamps. Multiple selection strategies are implemented, including approaches inspired by active learning (AL), multi-criteria decision analysis (MCDA), and medicinal chemistry heuristics, that display distinct behavior in the selection of compounds. Retrospective analysis provides a rigorous, low-cost test bed for investigating selection strategies in lead optimization and could help reduce the cost, duration and risk of lead optimization projects.

Received 29th August 2025
Accepted 22nd April 2026

DOI: 10.1039/d5dd00387c

rsc.li/digitaldiscovery

1 Introduction

Drug discovery relies on efficiently exploring and exploiting vast chemical spaces to identify active molecules, which are then optimized for key properties such as potency, selectivity, safety, and pharmacokinetics. Identifying a novel drug candidate for clinical development is both resource-intensive and time-consuming, encompassing everything from target identification to preclinical studies. In recent years, machine learning (ML) has emerged as a promising tool to accelerate early-stage drug discovery by enhancing virtual screening,[1] improving structure–activity relationship (SAR) modeling,[2–4] and generating novel and optimized compounds.[5–8] However, the limited availability of experimental data at this stage often constrains the full potential of ML-driven approaches.

Active learning (AL), a subfield of ML, uses an acquisition function to selectively query unlabeled samples from a pool for training in order to improve the model's performance with minimal labeling effort by focusing on the most informative examples. It has demonstrated significant utility in data-scarce settings across various domains.[9–11] Consequently, AL has gained traction in early-stage drug discovery, particularly for hit identification. Most applications focus on leveraging AL to enhance Quantitative Structure Activity Relationship (QSAR) and Quantitative Structure Property Relationship (QSPR) modeling with minimal data, thereby improving the efficiency of virtual screening and high-throughput screening campaigns.[12–18]

The strong interest in applying AL to hit identification stems from its critical role in the drug discovery pipeline, as well as the availability of large screening datasets[19–21] and commercial compound libraries. In these cases, docking software often serves as a low-cost oracle, aiding in the benchmarking of AL-based selection strategies. However, extending AL methodologies to later stages of drug discovery such as hit-to-lead and lead optimization poses significant challenges.

Unlike hit identification, where the molecular pool remains static, subsequent stages involve iterative hypothesis generation within the Design–Make–Test–Analyze (DMTA) cycle. This process causes shifts in the chemical distribution over time, requiring predictive models to adapt accordingly. Additionally, lead optimization involves multi-objective optimization, where potency, selectivity, and ADMET properties must all be taken into account, further complicating the application of AL.

To rigorously benchmark AL strategies in lead optimization, one would ideally conduct real DMTA cycles. In such cycles, the design of molecules, whether by medicinal chemists or AI,

[a]Chimie Physique et Chimie du Vivant, École Normale Supérieure, PSL Université, Sorbonne Université, CNRS, 75005 Paris, France. E-mail: rodolphe.vuilleumier@ens.psl.eu

[b]Integrated Drug Discovery, R&D, Sanofi, 94400 Vitry-sur-Seine, France. E-mail: marc.bianciotto@sanofi.com

would be coupled with an acquisition function to guide which compounds to synthesize and test. Although automated synthesis platforms have emerged,[22–24] this approach remains costly and difficult to implement, especially when attempting to benchmark multiple strategies in parallel.

An alternative is to conduct fully virtual DMTA scenarios. Such an approach has been tested but only with one binary objective, with molecules being labeled as active or inactive based on a simple oracle using various ratios between the number of carbons, oxygens and nitrogens.[25] Beyond the fact that the realism of this oracle is limited, this way of evaluation does not take into account the multi-objective nature of lead optimization.

In this work, we investigated another approach that is the retrospective analysis of legacy drug discovery projects through virtual scenarios in which an acquisition function guides compound selection (in this paper, the terms acquisition function and selection strategy will be used interchangeably). Unfortunately, we are not aware of publicly available lead optimization datasets that document the temporal evolution of a project, making this kind of retrospective study difficult to conduct outside of industry.

In response to these challenges, we developed a multi-objective molecular prioritization toolkit that incorporates selection strategies inspired by AL algorithms, multi-criteria decision analysis (MCDA) methods, and classical medicinal chemistry approaches. We also devised a methodology for running retrospective studies on legacy drug discovery projects, along with analytical tools to assess the exploration and exploitation performance of these strategies. By combining the toolkit with this retrospective methodology, we can run virtual "what if" scenarios to evaluate how various selection strategies might have performed historically. When conducting these simulations, we only have access to experimental data for the compounds actually tested during the project (not for those that were designed but never tested, as that would require an oracle). Thus, each simulation involves selecting a fraction of the real project's molecules and comparing the outcomes with the project's actual results. Acquisition functions are evaluated for their ability to match the real project outcomes (both in terms of exploitation and exploration of the chemical space) while selecting less compounds.

The aim of this paper is to present the prioritization toolkit, retrospective methodology and analytical tools, based on four partial datasets from previous projects.[26–28] It is not intended to show which selection strategy is best but to present tools to benchmark and compare them. Additional insights and conclusions gleaned from larger and more consistent datasets will be presented in future work.

In summary, this work makes the following contributions:

- We created a molecular prioritization toolkit, drawing on selection strategies from active learning, multi-criteria decision analysis, and classical medicinal chemistry approaches.
- We developed a methodology for conducting virtual scenarios on legacy drug discovery projects.
- We introduced an analysis framework to benchmark strategies for both exploiting and exploring the chemical space.
- We utilized datasets comprising previously disclosed molecular structures[15,26–33] to illustrate our simulation procedures and analytical methodologies. In this work, we enrich these datasets by including experimental timestamps for each compound and extend one of them with data for two additional experimental endpoints beyond the primary potency assay.

2 Materials

2.1 Data

To illustrate the capabilities of our prioritization toolkit and to demonstrate the retrospective simulation methodology and analyses of exploration *versus* exploitation strategies, we used four datasets that are derived from legacy Sanofi projects but that do not contain either all compounds or all readouts considered in these projects. For each dataset, each chemical structure is associated with one or more assay values and corresponding timestamps.

The first dataset is derived from a project targeting Factor Xa (FXa),[26–30] 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 six years, comprises 1015 compounds synthesized and evaluated for potency against FXa, as well as for solubility and LogD. It serves as the primary demonstration dataset for showcasing the analytical tools developed in the Results section.

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 (PPARδ), a nuclear receptor crucial for metabolic and inflammatory regulation, implicated in metabolic disorders such as diabetes and cardiovascular diseases. It covers a period of nearly six years, comprising 498 compounds.

Finally, the fourth dataset comes from a project targeting renin,[31–33] an enzyme essential for blood pressure and electrolyte balance regulation, primarily associated with hypertension and related cardiovascular disorders. This dataset spans over two years, with 388 compounds being evaluated.

2.2 Blueprint

During the lead optimization phase, the project team develops a blueprint, also referred to as a drug candidate target profile, that outlines the desired characteristics of an optimal drug candidate across a list of endpoints. For each endpoint—such as potency, selectivity, or ADMET properties—the blueprint specifies several key parameters (Table 1):

- 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).
- Lowest threshold value (LTV): for the higher is better trend defines the minimum value for a compound to be acceptable.

Table 1 Blueprint tables outlining the parameter space for the FXa, renin, PPARδ 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 that bound the acceptable property window. The weight column indicates the relative importance of each endpoint

Dataset	Endpoint	Trend	LTV	HTV	Weight
FXa	pKi	H	8.5	9	3
FXa	Solubility (pH 7.40) (μM)	H	50	100	1
FXa	LogD (pH 7.40)	V	1.5	3.5	1
Renin	pIC_{50}	H	8	9	1
PPARδ	pEC_{50}	H	8	9	1
MMP-8	pIC_{50}	H	8.5	9	1

For the lower is better trend 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 the lower is better trend 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,[34] by defining a priority order among different endpoints. A weight of 0 can be assigned to endpoints that are monitored but not actively optimized.

The simulation results presented in the following sections were made with this fixed blueprint. However, the blueprint can be dynamic and may evolve throughout the course of a project. Weights assigned to different endpoints can shift as priorities change over time, and new properties may be added as additional data become available or new challenges arise. The framework enables changing the blueprint during a simulation if needed.

3 Methods

3.1 Retrospective simulation methodology

3.1.1 Simulation setup. Prior to launching a simulation, the following parameters must be specified: the starting date, ending date, timestep, batch size, type of predictive model and acquisition function.

• Starting date (t_0): the simulation can begin when the conditions for building a QSAR model on the primary endpoint are met (see below): at least 100 compounds with experimental data for the primary endpoint have been synthesized and tested, with at least one of them having an activity beyond the threshold on that endpoint.

• Ending date (t_f): the simulation terminates on the same calendar date as the corresponding real-world project, allowing a direct, time-aligned comparison between virtual and real scenarios.

• Timestep (Δt): Δt represents the standard duration of a DMTA cycle, defined as the interval between consecutive project meetings for compound prioritization. Typical values for Δt range from 1 to 3 months.

• Memory effect: to simulate the natural de-prioritization of untested hypotheses over time, we introduced a memory parameter. This mechanism controls the retention of candidates in the pool. It can be configured for infinite retention (where compounds remain available indefinitely) or assigned a specific expiration threshold (*e.g.*, removing unselected compounds after 6 consecutive iterations).

• Batch size: the batch size can be expressed as either an integer or a fraction. If an integer is provided, exactly that many compounds are selected at each iteration, provided a sufficient pool of virtual candidates exists. If a fraction is supplied, a proportionate subsample of number of compounds synthesized and tested in the real project during each Δt is selected.

• QSAR/QSPR models: at each iteration, Quantitative Structure–Activity Relationship (QSAR) or Quantitative Structure–Property Relationship (QSPR) models are trained for every endpoint that satisfies two prerequisites: (i) a minimum of 100 data points, ensuring baseline model reliability and (ii) at least one compound within the endpoint's optimal range. Feature extraction is performed using Extended Connectivity FingerPrints[35] (ECFP, 2048 bits, radius 2, with chirality and count) fed into the algorithms shown in Table 2. We deliberately utilize default hyperparameters to maintain a neutral baseline and reduce computational overhead, and the objective of this work is to explain the methodology and framework workflow. However, in a real-world application, model-specific fine-tuning would be required to optimize predictive performance.

• Acquisition function: the acquisition function governs compound selection at every iteration and is the principal variable under investigation in the simulations. It may favor exploitation, exploration, or a balance of both. In this study, each simulation employs a single, fixed acquisition function; neither mixtures of functions nor dynamic switching strategies are considered in this work. Further details about available strategies are provided in a following section and in the SI.

3.1.2 Simulation. Once the setup parameters are defined, the simulation can begin. All compounds (and associated data) 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 the training set meets the requirements defined in the previous section: at least 100 documented compounds, including one

Table 2 QSAR/QSPR models implemented in the framework and their associated uncertainty quantification methods

Model	Uncertainty estimation methods
Random forest	Standard deviation of trees or bootstrap
LightGBM	Quantile regression or bootstrap
XGBoost	Quantile regression or bootstrap
Gaussian process	Posterior variance

within the optimal range. Fig. 1 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 $t_0 + \Delta t$ (Fig. 1A). 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 (Fig. 1B), 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 (Fig. 1C). Note that a molecule can remain in the pool indefinitely if it is never selected by the acquisition function.

3.1.3 Desirable molecules. 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,[36] 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 assay values are normalized between 0.05 and 1 to define a utility value, thanks to a utility function d inspired by Cummins and Bell.[37] Utility functions, illustrated in Fig. 2, 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 (k) 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 (eqn (1), Fig. 2A), maximized (eqn (2), Fig. 2B), or kept within a designated interval (eqn (3), Fig. 2C).

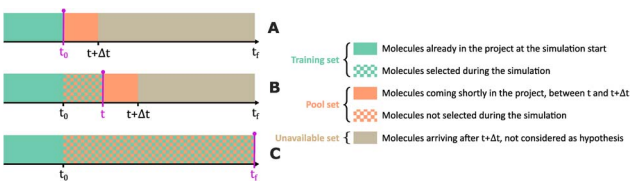

Fig. 1 Evolution of different molecular sets over the course of a simulation. (A) Initial configuration (iteration 1) consisting entirely of newly introduced molecules. (B) Intermediate stage showing a mixture of newly introduced and previously unselected molecules. (C) Final state comprising exclusively of previously unselected molecules.

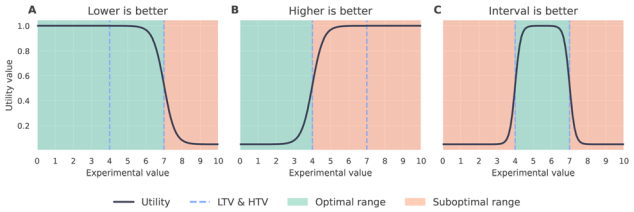

Fig. 2 Utility functions for the three optimization directions. Plots illustrate (A) lower is better, (B) interval is better, and (C) higher is better optimization trends. All functions are plotted with a lower threshold of 4 and an upper threshold of 7. Molecules are defined as being in the optimal region when the utility score is ≥0.525, corresponding to the midpoint of the sigmoidal curve.

$$d_i^{\text{lower-is-better}}(x) = \frac{1-k}{1+e^{\frac{x-\text{LTV}_i}{\eta_{\text{HTV}_i-\text{LTV}_i}}}} + k \quad (1)$$

$$d_i^{\text{higher-is-better}}(x) = \frac{1-k}{1+e^{\frac{x-\text{HTV}_i}{\eta_{\text{LTV}_i-\text{HTV}_i}}}} + k \quad (2)$$

$$d_i^{\text{interval}}(x) = \begin{cases} d_i^{\text{lower-is-better}}(x), & x < \dfrac{\text{LTV}_i + \text{HTV}_i}{2} \\ d_i^{\text{higher-is-better}}(x), & x \geq \dfrac{\text{LTV}_i + \text{HTV}_i}{2} \end{cases} \quad (3)$$

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 (eqn (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.

$$D_{\text{score}}(x) = \left(\prod_{i=1}^{n} d_i(x)\right)^{\frac{1}{\sum_i^n \omega_i}} \quad (4)$$

With each molecule in the project assigned a desirability score, various "top categories" can be defined. One such category, "Fixed ($t = 0.5$)" includes those with a desirability score exceeding a fixed threshold of 0.5. While this threshold-based approach is useful for monitoring progress in live projects, it presents challenges when applied to legacy data. A strict threshold may exclude all molecules, whereas a lenient one could classify nearly every molecule as "good," rendering the classification ineffective.

To address these limitations, we introduce additional top categories based on score distribution. For instance, the Top DScore (X%) category include compounds within the highest X% of desirability scores, offering a more nuanced approach to the analysis. However, this method always selects a fixed proportion of molecules, irrespective of their absolute scores. To enhance classification flexibility, we also define categories based on deviations from the mean: Top DScore ($n\sigma$) includes molecules with scores at or above $\mu + n\sigma$, where μ is the mean desirability score, σ is the standard deviation and n is a number

greater than 0. This approach provides a more adaptive and context-sensitive way to identify promising molecules.

3.2 Acquisition functions

The retrospective pipeline incorporates a comprehensive set of acquisition functions for regression tasks, drawing on methodologies from active learning, multi-criteria decision analysis, and medicinal chemistry strategies. These functions can be classified along several dimensions. All the strategies implemented are summarized in Table 3, and detailed explanations for each can be found in the Supplementary Information.

First, acquisition functions can rely on predictive (QSAR/QSPR) models. For example, some strategies use predicted values or prediction uncertainty to select molecules, such as Desirability, Desired Spread or Uncertainty. Other strategies, such as Coverage or K-means do not require predicted values and only rely on structural information. As discussed previously, Table 2 lists machine learning models and the associated uncertainty estimation methods that are available to choose from in the framework for building predictive models.

Second, acquisition functions can be distinguished by their primary objective. Some are designed to retrieve promising molecules, hence prioritizing exploitation, and others tend to foster exploration by selecting compounds all around the chemical space. Trade-off strategies also exist, balancing more or less equally exploitation and exploration.

Finally, acquisition functions can be categorized by their operational approach: independent strategies, in which each compound is chosen independently, *versus* batch strategies, in which the choice of one compound is influenced by the others selected. Notably, some batch selection strategies can be effectively approximated through sequential selection.[38]

Table 3 Summary of acquisition functions categorized by objective, selection type and reliance on predictive modeling. A detailed description of each strategy is available in the Supplementary Information. Functions newly introduced in this work are marked with an asterisk (*)

Acquisition function	Objective	Selection	Model
Coverage score[39]	Exploration	Batch	No
Desirability[37]	Exploitation	Independent	Yes
Desirability with AD[40]	Exploitation	Independent	Yes
Desired coverage *	Trade-off	Batch	Yes
Desired spread *	Trade-off	Batch	Yes
Dissimilarity-to-known	Exploration	Independent	No
Dissimilarity-to-known-good	Exploration	Independent	No
GRA[41]	Exploration	Independent	Yes
Greediverse[42]	Trade-off	Batch	Yes
K-means[43]	Exploration	Batch	No
K-medoids[44]	Exploration	Batch	No
Random	Exploration	Independent	No
Similarity-to-known	Exploration	Independent	No
Similarity-to-known-good	Exploration	Independent	No
Spread[45]	Exploration	Batch	No
TOPSIS[46]	Exploitation	Independent	Yes
Uncertainty *	Exploration	Independent	Yes
U/D harmonic *	Trade-off	Independent	Yes
U/D ratio *	Trade-off	Independent	Yes

3.3 Retrospective analysis tools

3.3.1 Visualizing simulations. The primary objective of the lead optimization phase is to efficiently explore the chemical space to identify molecules that align with a predefined blueprint and are suitable candidates for progression to preclinical development. In retrospective analyses, these molecules typically demonstrate high desirability and fall within previously defined top-tier categories, each characterized by varying degrees of stringency. The evaluation of the exploitative capacity of an acquisition function involves determining whether it preferentially selects molecules from these top-tier categories during simulation cycles. Furthermore, monitoring how the chemical space is explored provides valuable insights into the strategy's behavior.

Visualization of the chemical space across iterative cycles facilitates qualitative assessment of different acquisition functions. Commonly used visualization techniques include Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP),[47] and Tree MAP (TMAP).[48] In this study, TMAP was selected due to its tree-like structure, which effectively preserves both global and local similarity. TMAP produces a two-dimensional (2D) graph where each node corresponds to a molecule. Extending this methodology, we developed the Kernel Density Estimate TMAP (KDE TMAP). KDE TMAP applies a Gaussian kernel to the 2D embeddings, transforming the discrete chemical space into a continuous representation that highlights areas with varying molecular densities, making interpretation quicker and simpler.

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

$$kde(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \quad (5)$$

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

$$K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right). \quad (6)$$

To illustrate how a virtual scenario populates the chemical space, we first compute the TMAP for all molecules in the project (Fig. 3A) and subsequently transform it into a KDE TMAP (Fig. 3B). We then derive the KDE TMAP envelope (Fig. 3C), serving as a reference chemical space progressively populated across simulation iterations.

Fig. 4 shows the projection of different sets of molecules onto the project's envelope across iterations. The figure offers an intuitive way to gauge how effectively the simulation balances exploration of new chemical regions and exploitation of high-value areas. After the final iteration, the plot provides a holistic view of the sampled chemical space. Zones populated

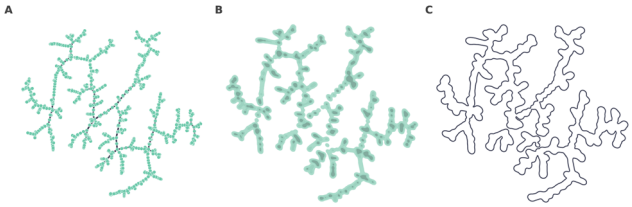

Fig. 3 Visualization of the full FXa chemical space using TMAP. (A) Discrete TMAP embedding showing individual data points. (B) The same embedding colored by a kernel density estimate (KDE) to highlight regions of high molecular density. (C) The resulting density envelope, delineating the boundaries of the occupied chemical space.

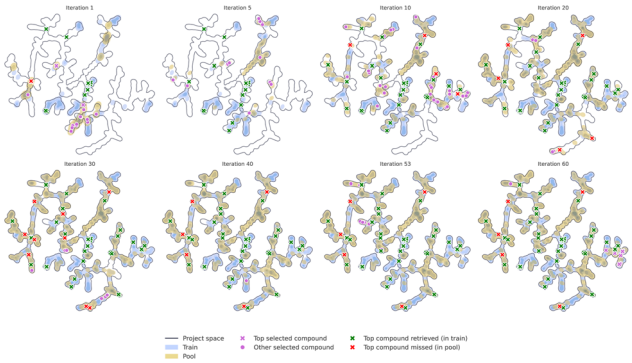

Fig. 4 Evolution of the FXa chemical space under random acquisition (iterations 1–60). Snapshots of the TMAP embedding display the progression from iterations 1 to 60. As molecules are selected, they transition from the pool set (yellow) to the training set (blue). In each iteration, newly acquired molecules are highlighted in purple, with crosses indicating the top 5% DScore. To visualize recall efficiency, green crosses track top-tier compounds already found, while red crosses mark those still missing from the training set.

by compounds in the final training set correspond to regions that are explored by the acquisition function, whereas areas still containing only molecules from the pool set mark the chemical space that remains unvisited. The arrangement of the plotted markers thus also reveals the exploitative capability of the strategy, highlighting how well it selected the most promising compounds.

3.3.2 Exploitation. While visualizing how a strategy explores chemical space can be insightful, comparing multiple strategies at a larger scale necessitates a quantitative approach. To address this requirement, we propose a simple metric called the Proportion of Top Molecules Retrieved (PTMR). The PTMR quantifies the proportion of molecules selected by a strategy that belong to a specified top-tier category. Formally, let S represent the set of molecules selected up to a given iteration and T denote the set of molecules in a particular top-tier category, such as the Top DScore 5%. The PTMR is defined as follows:

$$\mathrm{PTMR}(S,T) = \frac{|T \cap S|}{|T|} \tag{7}$$

This monotonic metric can be computed at each iteration of a simulation to gauge how effectively an acquisition function selects the most desirable molecules. A PTMR of 1 indicates that a strategy has retrieved all top molecules, whereas a value of 0 indicates that it has not retrieved any.

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

3.3.3.1 Internal diversity. Internal diversity is a widely adopted metric commonly employed in chemical library design and for evaluating chemically diverse outputs from AI-driven molecular generative models.[49–51] Defined as the average Tanimoto distance (δ) among all pairs of compounds within a set S, it is formally expressed as:

$$\text{Internal Diversity}(S) = \frac{2}{n(n-1)} \sum_{i,j \in S \ i \neq j} \delta_{i,j} \tag{8}$$

Despite its simplicity and popularity, Internal Diversity has notable limitations.[52] High internal diversity may result from structurally disparate molecules that do not necessarily exhibit drug-like properties or align with specific biological targets. Additionally, it does not explicitly measure the coverage of chemical space or biological relevance, highlighting the necessity for complementary metrics.

3.3.3.2 #Circles. Recent metrics leveraging sphere-exclusion clustering algorithms,[53] such as *SEDiv*[54] and #Circles,[52] have emerged for quantifying chemical diversity. The #Circles metric specifically counts the maximum number of mutually exclusive clusters (circles) with diameter d that fit within a set S, formally defined as:

$$\#\text{Circles}(S,d) = \max_{C \subseteq S} \ \max_{i \neq j \in C} |C| \quad s.t. \quad \delta_{i,j} > d \tag{9}$$

3.3.3.3 Neighborhood Coverage. Metrics like Internal Diversity and #Circles are reference-free methods for analyzing the chemical diversity within a single set S. It is also useful to use coverage metrics for quantifying the extent to which the set S covers the chemical space defined by a reference set R. In our simulations, R represents the full dataset, with S being the subset selected during 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). This metric employs an indicator function I:

$$I(i,\tau) = \begin{cases} 1, & \text{if} \quad \min_{j \in S} \delta_{j,i} \leq \tau \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Then, Neighborhood Coverage is computed as:

$$\mathrm{NC}(S,R,\tau) = \frac{1}{|R|} \sum_{i \in R} I(i,\tau) \tag{11}$$

The proposed metric utilizes a distance threshold of 0.3 (corresponding to a Tanimoto similarity of 0.7) based on 2048 bit ECFP4 binary fingerprints. This value aligns with established benchmarks for structural similarity[55,56] and has been shown to provide qualitatively consistent results for medicinal chemists.[57] Nevertheless, threshold selection remains inherently subjective, as its appropriateness depends on the specific fingerprint architecture and the requirements of the domain expert.[58]

3.3.3.4 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.

$$\mathrm{NCAUC}(S, R, \tau) = \int_0^1 \left[\frac{1}{|R|} \sum_{i \in R} I(i, \tau) \right] \tag{12}$$

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

4 Results

We illustrate the full retrospective pipeline from simulating virtual scenarios to evaluating outcomes on the FXa dataset. All corresponding figures for the other three datasets are provided in the Supplementary Information.

4.1 Simulations

4.1.1 Desirable molecules. To evaluate the exploitative capabilities of different selection strategies, we categorize molecules based on their experimental desirability scores. These scores were computed for all compounds in the FXa dataset, and multiple top-tier categories were defined with varying levels of selectivity, as previously described. The distribution of compounds across these categories is presented in Fig. 5. In the analyses that follow, we primarily focus on the Top DScore 5% category for statistical robustness and the Top DScore 1% category to capture the highest-performing compounds.

4.1.2 Simulation setup. The FXa simulation commences on October 1, 2002, adhering to the criteria outlined previously. At initialization, 154 compounds had already been evaluated for FXa pKi. The simulation spans the historical project timeline, concluding on September 19, 2007. During this period, the real project synthesized and tested approximately 14 compounds per month.

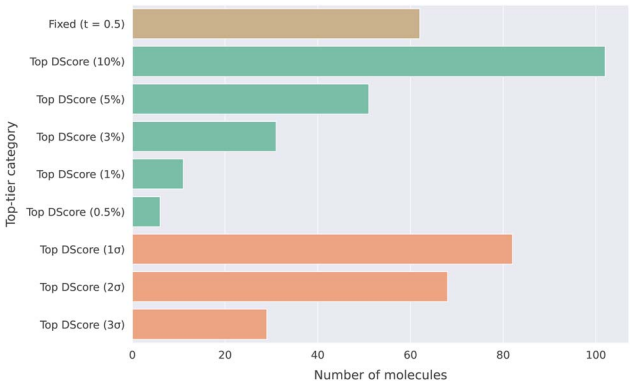

Fig. 5 Number of molecules classified as "top-tier" across different threshold definitions (FXa dataset). The definitions are grouped by methodology: brown indicates fixed desirability thresholds, orange indicates percentile-based thresholds, and green indicates standard deviation-based thresholds.

To reflect the operational cycle of a typical drug discovery project, simulations were conducted with a one-month time-step. The memory effect was disabled by default, ensuring that unselected compounds remained within the candidate pool for subsequent iterations. For comparison, results for simulations incorporating a six-month memory decay are provided in the Supplementary Information.

The batch size is set to a ratio of 0.5 to demonstrate the potential for reducing experimental effort; this means that at each iteration, the acquisition function selects 50% of the number of new molecules synthesized in the real project (Fig. 6A). Over 60 total iterations, the simulation selects 58% of the total molecules evaluated in the historical project (15% *via* initialization and 43% *via* the acquisition function; Fig. 6B).

4.1.3 Acquisition functions. For demonstration purposes, we selected ten representative acquisition functions from the

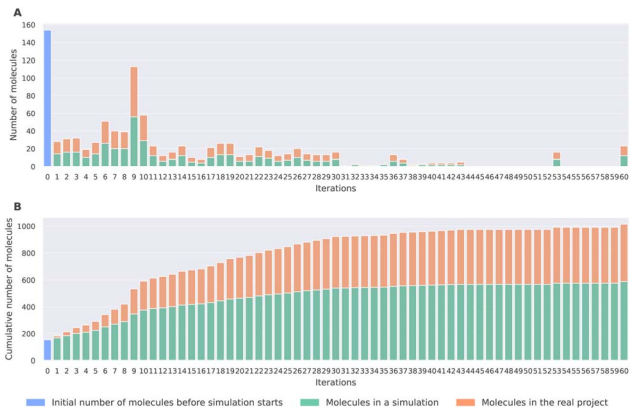

Fig. 6 Comparison of compound selection between the real project and a simulation (FXa dataset). (A) The number of compounds selected per iteration. (B) Cumulative number of compounds selected over time. The simulation is initialized with 154 previously synthesized compounds. By the end of the run (batch size: 0.5%), 58% of all the molecules are selected. The real project averages approximately 14 molecules per iteration, compared to 7 molecules per iteration in the simulations.

broader set available (see Table 3 and SI). These strategies were chosen to cover a diverse spectrum of exploratory and exploitative behaviors. To ensure statistical reliability, multiple simulation replicates were performed for each strategy: 100 independent runs for the Random baseline and 10 for all other methods, with each replicate initialized using a unique random seed. This approach allows for the computation of 95% confidence intervals across all result plots. The acquisition functions evaluated are:

•Random: randomly selects compounds from the pool set. Often treated as a dummy baseline in active learning studies, a purely random selection is informative in our retrospective context. Because each simulated batch contains fewer molecules than the corresponding experimental batch, random sampling provides a realistic estimate of project performance at comparable batch size. To obtain reliable statistics, Random simulations are repeated 100 times. All the following acquisition functions are repeated 10 times.

• Desirability: a purely exploitative strategy in which, at every iteration, the algorithm selects the molecules with the highest predicted desirability scores.

• Desirability (with applicability domain): this strategy extends the Desirability approach by restricting selection to molecules within the predictive models' applicability domain (AD). This reflects standard practice, as predictions are generally considered reliable only within the validated chemical space. We employ a common AD definition based on a similarity criterion:[57,59] a molecule from the pool set is included only if its similarity to at least one training set compound exceeds 0.7. This threshold ensures consistency with the value defined for Neighborhood Coverage (Section 3.3.3.3). We encourage people that will use this framework to implement other applicability domain definitions.

• Similarity-to-known-good: emulates a popular strategy in medicinal chemistry where compounds are selected based on their structural similarity to the best molecules from the training set. Here, the best molecules are defined as those whose experimental desirability score is above 0.5.

• Greediverse: a strategy originally developed in-house for *de novo* drug design[42] which optimizes batch desirability, while imposing a penalty, scaled by the parameter λ, on each pair of molecules whose similarity surpasses a predefined similarity threshold in order to introduce some diversity in the selection.

• Uncertainty: a widely used active learning strategy which selects the molecules with the most uncertain predictions, adapted here to the multi-objective case by aggregating the predicted uncertainties of each individual endpoint. As in the Desirability approach, all scores are first normalized—here with MinMax scaling—and subsequently combined by taking their geometric mean to yield a single uncertainty score. Molecules with the highest resulting score are then prioritized for selection.

• Coverage: a model-agnostic strategy leveraging Bayesian statistics and information entropy to select informative subsets.[39]

• U/D harmonic: a trade-off approach that prioritizes molecules with the highest harmonic mean between desirability and uncertainty scores.

• Desired spread: a hybrid exploration–exploitation strategy inspired by the demerit spread design method.[60] It integrates predicted desirability scores with a diversity criterion, where desirability acts as a scaling factor for the compound's distance to its nearest neighbor among the training set or previously selected molecules.

•Oracle: this baseline selects the compounds with the highest experimental desirability scores, providing an upper bound for pure exploitation.

If a tie occurs where multiple compounds share identical acquisition scores but exceed the remaining batch capacity, selection is prioritized by the best predicted value of the highest-weighted endpoint defined in the blueprint.

4.2 Understanding the chemical space with TMAPs

The TMAP visualization can be a powerful tool to understand how the chemical space of a drug discovery project is structured. Fig. 7A 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.

Fig. 7B overlays the main chemical series, revealing distinct yet partially overlapping domains. Fig. 7C displays some of the best structures 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.

4.3 Visualizing simulated scenarios

Before quantitatively evaluating the exploration–exploitation performance of the acquisition functions, we first conduct a visual inspection of how each function traverses the project's chemical space. Fig. 8 shows the final state reached with every acquisition function after 60 iterations on the FXa dataset, projecting both the corresponding training and pool sets together with the top-tier compounds recovered or not during the simulation. The extent of overlap between training and pool sets serves as a proxy for exploratory capability, whereas the number of retrieved top-tier molecules gauges the exploitative efficiency.

Exploration-focused strategies such as Coverage (Fig. 8B) and Uncertainty (Fig. 8C) effectively sample compounds across the entire chemical space, leaving few areas unrepresented. However, these strategies often underperform in terms of exploitation, recovering only a small number of top-tier compounds.

Conversely, exploitation-oriented strategies display an inverse pattern. The Desirability strategy (Fig. 8E), which targets molecules with the highest predicted desirability scores, successfully retrieves many top-performing compounds but neglects regions

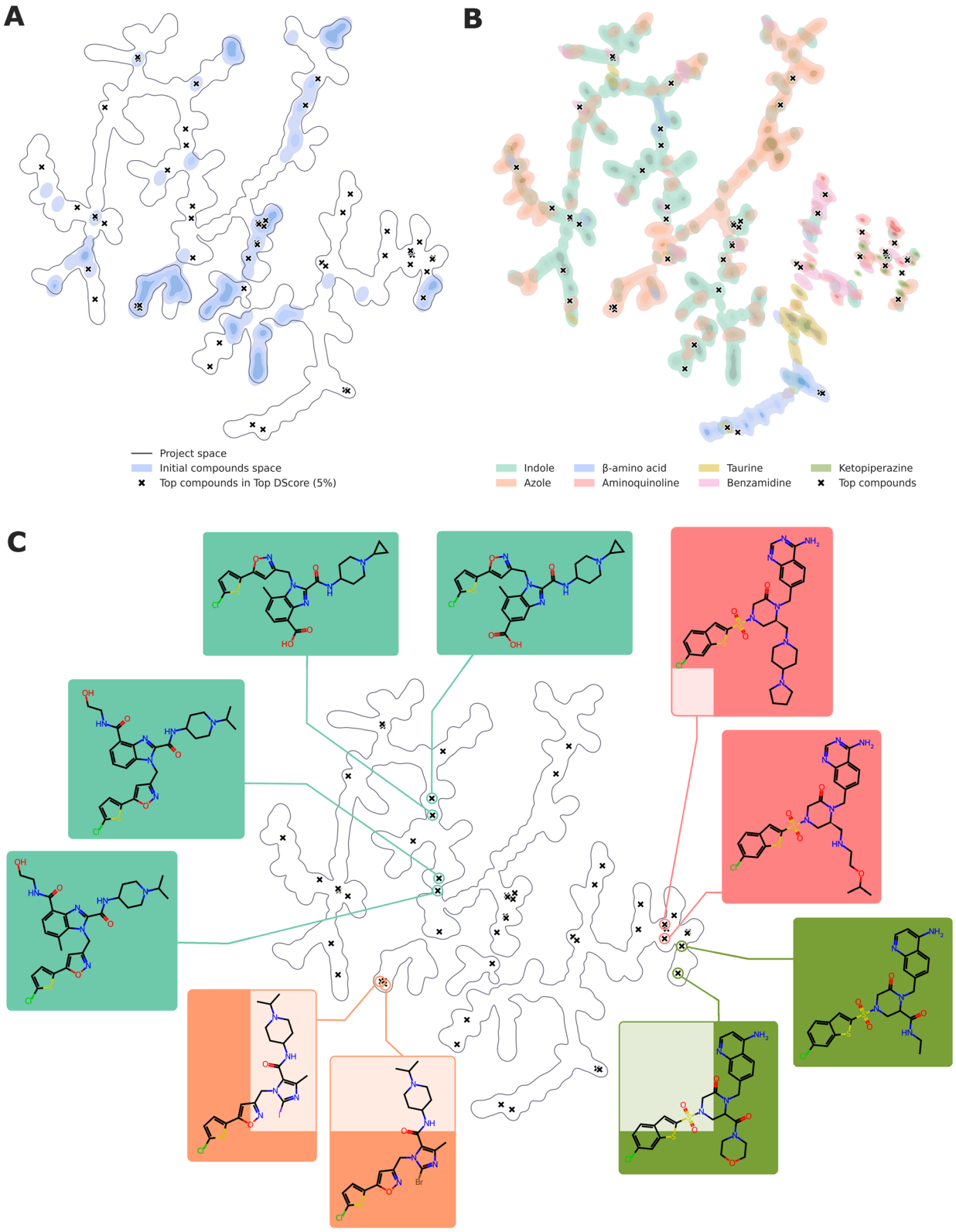

Fig. 7 Overview of the FXa dataset chemical space. (A) Global view of the chemical space at the start of the simulation. (B) Distributions of the major chemical series. (C) Representative chemical structures of top-tier compounds (top 5% DScore) from various series. The box colors in (C) correspond to the chemical series colors used in (B).

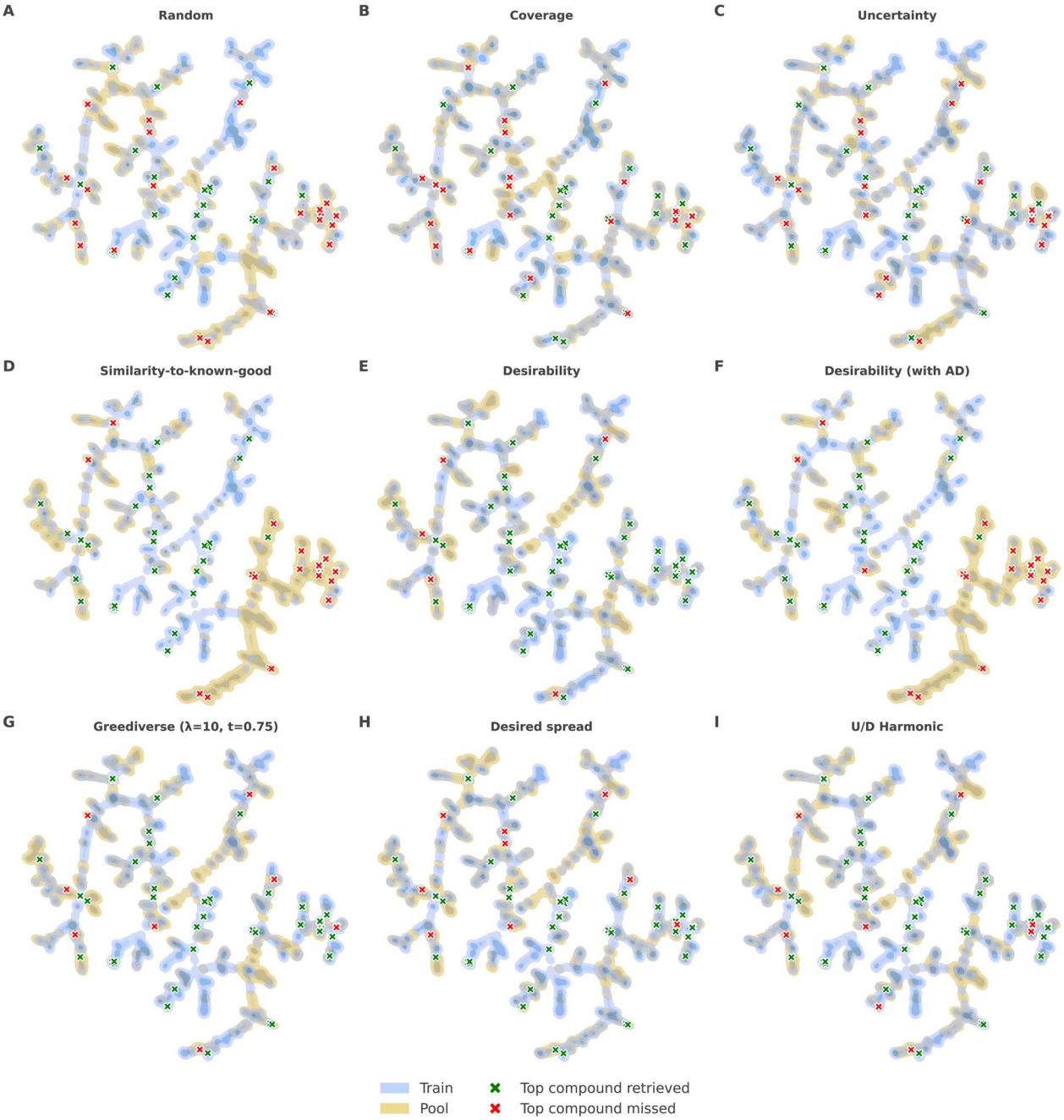

Fig. 8 TMAP visualizations of the final chemical space across nine acquisition functions (FXa dataset, (A) Random, (B) Coverage, (C) Uncertainty, (D) Similarity-to-known-good, (E) Desirability, (F) Desirability with AD, (G) Greediverse, (H) Desired Spread, (I) U/D Harmonic). Regions: blue areas represent the sampled subspace (selected molecules), while yellow areas indicate the remaining unselected pool. Markers: crosses identify top-tier compounds (top 5% DScore). Green crosses denote top-tier molecules that were successfully retrieved, whereas red crosses denote those that were missed.

of the chemical space. When the applicability domain constraint is applied (Fig. 8F), exploration is even more limited, resulting in missed opportunities to capture promising compounds from series such as aminoquinoline or ketopiperazine. Likewise, the Similarity-to-known-good strategy (Fig. 8D) shows excellent exploitation within the azole and indole series, successfully identifying all top-ranked molecules in these series, but it fails to explore beyond them, overlooking entire chemical series.

Strategies where the exploration–exploitation balance is built-in offer a more favorable trade-off. Greediverse (Fig. 8G) moderately improves exploration, particularly in the β-amino acid and benzamidine series. Likewise, the Desired Spread (Fig. 8H) preserves the strong exploitative performance of Desirability while significantly enhancing the exploratory coverage, approaching that of exploration-focused methods.

4.4 Exploration

Although visualizing the chemical space provides a qualitative glimpse of each acquisition function, quantitative comparisons are crucial for objectively identifying the most effective strategies. We track four complementary indicators: Neighborhood Coverage and its area-under-the-curve variant (both monotonic and reference-based); Internal Diversity (non-monotonic and reference-free); and #Circles (monotonic, yet reference-free).

Fig. 9 displays the evolution of these metrics over iterations and creates a coherent image. In early iterations, most acquisition functions cluster together, but Desirability (with AD) and Similarity-to-known-good quickly fall behind and remain the weakest throughout. Exploration-oriented Uncertainty and Coverage consistently top the curves but the trade-off Desired Spread strategy often matches or surpasses them, suggesting that judicious mixing of exploration and exploitation can pay off for exploration. Neighborhood Coverage and its AUC variant offer a very fine decomposition between the different scenarios. The non-monotonic nature of Internal Diversity enables highlighting phases when the chemical space is being focused with consistent drops such as the one between iterations 2 and 7. For #Circles, plateau phases obscure fine differences during simulations, but Desired Spread again secures the highest final score, followed closely by Uncertainty and U/D Harmonic.

4.5 Exploitation

Although exploration remains essential in drug discovery, the lead optimization phase ultimately aims to exploit chemical space efficiently to surface the most promising candidates. We therefore evaluated exploitation performance using the Percentage of Top Molecules Retrieved (PTMR) within the top

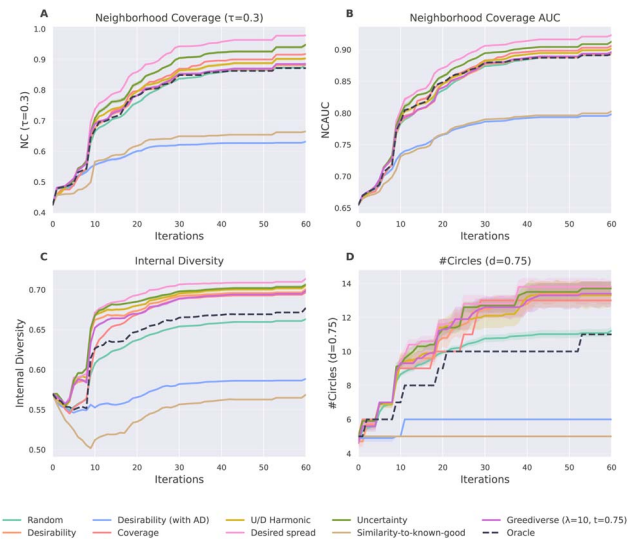

Fig. 9 Exploration metrics over iterations for selected acquisition functions (FXa dataset). (A) Evolution of Neighborhood Coverage ($\tau = 0.3$). (B) Neighborhood Coverage AUC. (C) Internal Diversity. (D) Number of circles (#Circles). Shaded regions indicate the 95% confidence interval. These intervals are calculated from 100 replicates for the Random strategy and 10 replicates for all other strategies.

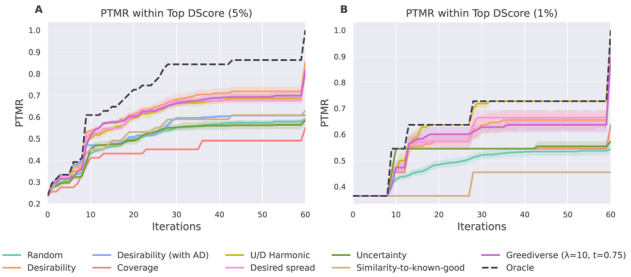

Fig. 10 Exploitation performance of various acquisition functions (FXa dataset). The plots track the retrieval of top-tier compounds defined by two stringency levels: (A) the top 5% DScore and (B) the top 1% DScore categories. Shaded regions indicate the 95% confidence interval. These intervals are calculated from 100 replicates for the Random strategy and 10 replicates for all other strategies.

5% and top 1% of DScores, corresponding to 51 and 11 compounds, respectively, out of the project's 1015 molecules. Each strategy is assessed on its ability to recover these top compounds while selecting only a subset of the full dataset. Two baseline reference scenarios provide context: Random, which approximates the real project selection performance at comparable batch size, and Oracle, which represents the ideal exploitative outcome.

Fig. 10 shows the exploitation performance of the different acquisition functions by monitoring the PTMR over iterations. All acquisition functions but Coverage perform similarly or outperform the Random baseline for both the top 5% and top 1% categories. Purely exploratory acquisition functions—Coverage and Uncertainty—have a PTMR of ≈ 0.55, mirroring Random. Desirability (with AD) and Similarity-to-known-good have similar performance. In contrast, exploitation-oriented methods excel. In the top 5% category (Fig. 10A), Desirability, Greediverse and Desired Spread stay close to the Oracle throughout optimization and finish with a PTMR ≈ 0.85. The distinction becomes sharper in the top 1% category (Fig. 10B). Desirability and Greediverse retrieve 90% of the top molecules and Desired Spread is almost indistinguishable from Oracle, ultimately retrieving all the top molecules. By contrast, Coverage and Uncertainty remain similar to random, with a final PTMR ≈ 0.5.

5 Discussion

5.1 Balancing and understanding exploration/exploitation

Effective lead optimization requires a strategic equilibrium between exploitation, iteratively refining a promising chemotype to enhance its profile, and exploration, which probes alternative scaffolds to uncover latent opportunities. The lead optimization decision-making process integrates both objectives to maximize SAR learning, generate novel hypotheses, mitigate late-stage failures and, ultimately, deliver differentiated, developable drugs. We propose a set of tools to benchmark compound selection strategies by running simulations on legacy projects and evaluating them qualitatively and quantitatively.

The tree-based TMAP visualization helps understand structural relationships among project compounds, thanks to its tree-based nature. The toolkit augments this view with KDE

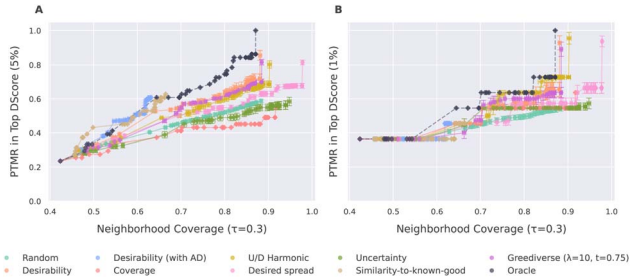

Fig. 11 Trajectories of exploitation *vs.* exploration (FXa dataset). Performance is mapped by plotting exploitation, monitored *via* PTMR within the top 5% DScore (A) and 1% DScore (B) against exploration, monitored *via* Neighborhood Coverage at $\tau = 0.3$. Error bars indicate the 95% confidence interval, calculated from 100 replicates for the Random baseline and 10 replicates for all other acquisition strategies.

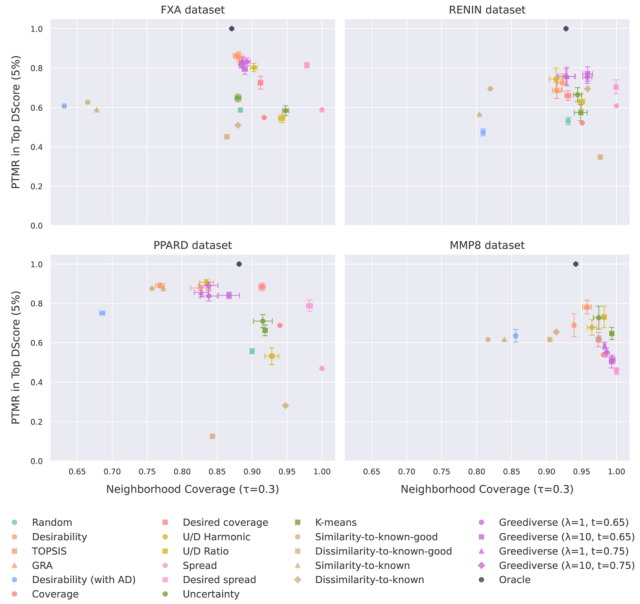

Fig. 12 Final overview of exploitation *vs.* exploration performance across four datasets. The panels display results for (A) FXa, (B) renin, (C) PPARδ, and (D) MMP-8. All strategies available in the framework are compared based on their ability to retrieve top-tier compounds (top 5% DScore). Error bars indicate the 95% confidence interval, calculated from 100 replicates for the Random baseline and 10 replicates for all other acquisition strategies.

TMAP plots that qualitatively track the exploration–exploitation balance across iterations (see *e.g.*, Fig. 8). TMAP visualization enable a quick and intuitive assessment of different strategy behaviors.

For exploration, the four proposed metrics are complementary. Neighborhood Coverage is a monotonic, reference-based measure well-suited to post-hoc benchmarking against the complete project set; with an appropriately chosen distance threshold, it finely discriminates among strategies (Fig. 9A). Its threshold-free variant, Neighborhood Coverage AUC, shares these advantages but tends to smooth inter-strategy differences (Fig. 9B). Internal Diversity and #Circles are reference-free, making them applicable to live projects in which the final chemical space is unknown. A drop in Internal Diversity signals a concentration on a limited region of chemical space (Fig. 9C), whereas #Circles provides an interpretable count of newly explored regions (Fig. 9D). Collectively, these metrics yield consistent, interpretable rankings of the tested strategies.

In Fig. 11, we plot an exploration metric (Neighborhood Coverage) against an exploitation metric (PTMR within the Top 5% DScore) to reveal the trade-off between the two objectives. On the FXa dataset, the Desired Spread strategy exhibits both strong exploration and strong exploitation performance.

Plotting only the final points for each strategy further streamlines comparisons. Fig. 12 illustrates this endpoint trade-off for all implemented strategies across the FXa, renin, PPARδ, and MMP-8 datasets. The first three datasets exhibit similar patterns, with each strategy occupying approximately the same region of chemical space relative to the Random baseline strategy. Conversely, the fourth dataset (MMP-8) demonstrates notably different outcomes, likely due to its distinct and sparser experimental timeline (see the MMP-8 section in Supplementary Information). Based on their positions relative to the Random baseline, the strategies we evaluated can be categorized as follows:

- Strategies demonstrating worse exploitation than random.
- Strategies with improved exploitation—which is the primary objective in lead optimization—but reduced exploration.
- Strategies showing superior performance in both exploitation and exploration; these represent the optimal ones.

5.2 Methodological considerations

We acknowledge that this work operates under several methodological limitations regarding the complexity of simulating lead optimization.

First, the datasets utilized represent only a subset of real-world projects, particularly regarding compound diversity, readouts, and design blueprints. Consequently, these should be viewed primarily as demonstration datasets intended to illustrate our simulation methodology and analytical framework. The results presented here are not intended to establish universal rules or general principles of lead optimization. Larger-scale studies involving dozens of comprehensive datasets are currently underway.

Second, our simulations relied on a single combination of ECFP + Random Forest model using standard deviation for uncertainty estimation. While model selection can influence results, our sensitivity analyses (detailed in the SI) demonstrate that performance differences across various predictive architectures are minimal for a given acquisition strategy.

Third, the primary simulations presented in this manuscript do not explicitly model the "memory effect," which is the practical tendency for untested hypotheses to be de-prioritized or discarded over time. This simplification appears to have a negligible impact on the study's conclusions. As detailed in Section 3.1, 4.8, 5.8 and 6.8 of the Supplementary Information, introducing a six-iteration memory constraint yields results comparable to those obtained in memory-free simulations.

Fourth, retrospective simulations assume that all compounds historically available at time t are immediately

selectable, overlooking synthetic lineage. In practice, chemical synthesis is a dependency chain: molecule A may be a required intermediate for B, which in turn is necessary for C. If a simulation selects A and C while bypassing B, it artificially inflates exploitation metrics. Correcting this is difficult because the historical logs needed to reconstruct these dependencies are often buried in physical laboratory notebooks, making them nearly inaccessible for older projects.

5.3 Future directions

While the primary goal of this study was to demonstrate the pipeline's capabilities using simple datasets, its full potential lies in application to larger and more complex optimization campaigns. Future work will therefore extend this framework to a broader and more realistic collection of datasets.

All simulations in this study employed a fixed acquisition strategy throughout the iterations. In contrast, real-world projects often begin with broad exploration to identify promising regions, then gradually shift toward focused exploitation as structure–activity insights accumulate. Because our framework supports dynamic switching or blending of strategies, future investigations will explore adaptive acquisition policies that evolve over time, better mirroring the iterative reasoning used by project teams.

Finally, the blueprint used in this work only consider endpoint-related criteria. However, additional constraints could be incorporated such as synthetic accessibility,[61–63] estimated synthesis cost, or computable molecular properties like molecular weight, topological polar surface area (TPSA), or cLogP.

6 Conclusion

The aim of this work is to contribute to the development of new methods for analyzing real-world drug discovery projects. It presents a toolkit for compound prioritization that integrates a set of selection strategies, paired with a methodology for retrospective simulations on historical projects. The toolkit streamlines the development and evaluation of rational decision-making policies for compound prioritization. These simulations provide a realistic, laboratory-free test bed. Accompanying analytical tools deliver qualitative and quantitative insights into the behavior of the acquisition functions, clarifying how each strategy balances the exploitation and the exploration of chemical space. This balance is critical in lead optimization to identify promising candidates within the current chemical series while preserving the opportunity to discover better compounds in new series. Rigorous analysis of the exploration–exploitation trade-off is mandatory to highlight strategies worth testing in real settings. This methodological work paves the way for future analyses on a larger scale, with application to a larger number of more realistic datasets, with the ultimate objective to enable drug-hunting teams to allocate more effectively synthesis and testing resources.

Author contributions

P. M. developed the selection framework/analysis tools and drafted the manuscript. B. F.-R. discussed the methodology and provided data. M. B. and R. V directed the research, discussed the methodology and results and reviewed the manuscript.

Conflicts of interest

P. M., B. F.-R. and M. B. are or have been Sanofi employees and may hold shares and/or stock options in the company. R. V. has nothing to disclose.

Data availability

The code and data for this work are available at the GitHub repository: https://github.com/Sanofi-Public/RetroDMTA and on Zenodo (DOI: https://doi.org/10.5281/zenodo.18758006).

Supplementary information: datasets information, simulations setup, detailed description of the acquisition functions, additional figures for the FXa dataset, and the full set of corresponding figures for the other datasets. See DOI: https://doi.org/10.1039/d5dd00387c.

Acknowledgements

Dr Hans Matter is gratefully acknowledged for sharing with us the datasets used in this work. The French National Association of Research and Technology (ANRT) is acknowledged for supporting P. M. (grant 2022/0928).

References

1 T. A. d. Oliveira, M. P. d. Silva, E. H. B. Maia, A. M. d. Silva and A. G. Taranto, *Drugs and Drug Candidates*, 2023, **2**, 311–334.
2 A. Tropsha, O. Isayev, A. Varnek, G. Schneider and A. Cherkasov, *Nat. Rev. Drug Discovery*, 2024, **23**, 141–155.
3 T. A. Soares, A. Nunes-Alves, A. Mazzolari, F. Ruggiu, G.-W. Wei and K. Merz, *J. Chem. Inf. Model.*, 2022, **62**, 5317–5320.
4 J. Mao, J. Akhtar, X. Zhang, L. Sun, S. Guan, X. Li, G. Chen, J. Liu, H.-N. Jeon, M. S. Kim, K. T. No and G. Wang, *iScience*, 2021, **24**, 103052.
5 V. D. Mouchlis, A. Afantitis, A. Serra, M. Fratello, A. G. Papadiamantis, V. Aidinis, I. Lynch, D. Greco and G. Melagraki, *Int. J. Mol. Sci.*, 2021, **22**, 1676.
6 M. Wang, Z. Wang, H. Sun, J. Wang, C. Shen, G. Weng, X. Chai, H. Li, D. Cao and T. Hou, *Curr. Opin. Struct. Biol.*, 2022, **72**, 135–144.
7 D. D. Martinelli, *Computers in Biology and Medicine*, 2022, **145**, 105403.
8 J. Meyers, B. Fabian and N. Brown, *Drug Discovery Today*, 2021, **26**, 2707–2715.
9 A. Tharwat and W. Schenck, *Mathematics*, 2023, **11**, 820.
10 P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen and X. Wang, *ACM Computing Surveys*, 2022, **54**, 1–40.

11 M. Prince, *Journal of Engineering Education*, 2004, **93**, 223–231.
12 D. Reker and G. Schneider, *Drug Discovery Today*, 2015, **20**, 458–465.
13 D. Reker, P. Schneider and G. Schneider, *Chem. Sci.*, 2016, **7**, 3919–3927.
14 D. E. Graff, E. I. Shakhnovich and C. W. Coley, *Chem. Sci.*, 2021, **12**, 7866–7881.
15 M. Bailey, S. Moayedpour, R. Li, A. Corrochano-Navarro, A. Kötter, L. Kogler-Anele, S. Riahi, C. Grebner, G. Hessler, H. Matter, M. Bianciotto, P. Mas, Z. Bar-Joseph and S. Jager, *eLife*, 2023, **12**, RP89679.
16 D. v. Tilborg and F. Grisoni, *Nat. Comput. Sci.*, 2024, **4**, 786–796.
17 F. Gusev, E. Gutkin, F. Gentile, F. Ban, S. B. Koby, F. Li, I. Chau, S. Ackloo, C. H. Arrowsmith, A. Bolotokova, P. Ghiabi, E. Gibson, L. Halabelian, S. Houliston, R. J. Harding, A. Hutchinson, P. Loppnau, S. Perveen, A. Seitova, H. Zeng, M. Schapira, O. Isayev, A. Cherkasov and M. G. Kurnikova, *J. Chem. Inf. Model.*, 2025, **65**, 5706–5717.
18 B. Cree, M. Bieniek, S. Amin, A. Kawamura and D. Cole, *Digital Discovery*, 2025, **4**, 438–450.
19 B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. d. Veij, H. Ioannidis, D. M. Lopez, J. F. Mosquera, M. P. Magarinos, N. Bosc, R. Arcila, T. Kizilören, A. Gaulton, A. P. Bento, M. F. Adasme, P. Monecke, G. A. Landrum and A. R. Leach, *Nucleic Acids Res.*, 2023, **52**, D1180–D1192.
20 T. Liu, Y. Lin, X. Wen, R. N. Jorissen and M. K. Gilson, *Nucleic Acids Res.*, 2007, **35**, D198–D201.
21 *NeurIPS 2024 - Predict New Medicines with BELKA*, Kaggle, https://www.kaggle.com/competitions/leash-belka/overview/neur-ips-2024.
22 A. B. MacConnell, A. K. Price and B. M. Paegel, *ACS Comb. Sci.*, 2017, **19**, 181–192.
23 C. W. Coley, D. A. ThomasIII, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. J. Hart, T. F. Jamison and K. F. Jensen, *Science*, 2019, **365**.
24 G. Schneider, *Nat. Rev. Drug Discovery*, 2018, **17**, 97–113.
25 H. G. Svensson, E. Bjerrum, C. Tyrchan, O. Engkvist and M. H. Chehreghani, *2022 IEEE International Conference on Big Data (Big Data)*, Osaka, Japan, 2022, pp. 5584-5592, DOI: 10.1109/BigData55660.2022.10020357.
26 M. Nazaré, H. Matter, D. W. Will, M. Wagner, M. Urmann, J. Czech, H. Schreuder, A. Bauer, K. Ritter and V. Wehner, *Angew. Chem., Int. Ed.*, 2012, **51**, 905–911.
27 M. Nazaré, D. W. Will, H. Matter, H. Schreuder, K. Ritter, M. Urmann, M. Essrich, A. Bauer, M. Wagner, J. Czech, M. Lorenz, V. Laux and V. Wehner, *J. Med. Chem.*, 2005, **48**, 4511–4525.
28 H. Matter, E. Defossa, U. Heinelt, P.-M. Blohm, D. Schneider, A. Müller, S. Herok, H. Schreuder, A. Liesum, V. Brachvogel, P. Lönze, A. Walser, F. Al-Obeidi and P. Wildgoose, *J. Med. Chem.*, 2002, **45**, 2749–2769.
29 M. Nazaré, M. Essrich, D. W. Will, H. Matter, K. Ritter, M. Urmann, A. Bauer, H. Schreuder, A. Dudda, J. Czech, M. Lorenz, V. Laux and V. Wehner, *Bioorg. Med. Chem. Lett.*, 2004, **14**, 4191–4195.
30 H. Matter, D. W. Will, M. Nazaré, H. Schreuder, V. Laux and V. Wehner, *J. Med. Chem.*, 2005, **48**, 3290–3312.
31 B. Scheiper, H. Matter, H. Steinhagen, U. Stilz, Z. Böcskei, V. Fleury and G. McCort, *Bioorg. Med. Chem. Lett.*, 2010, **20**, 6268–6272.
32 H. Matter, B. Scheiper, H. Steinhagen, Z. Böcskei, V. Fleury and G. McCort, *Bioorg. Med. Chem. Lett.*, 2011, **21**, 5487–5492.
33 B. Scheiper, H. Matter, H. Steinhagen, Z. Böcskei, V. Fleury and G. McCort, *Bioorg. Med. Chem. Lett.*, 2011, **21**, 5480–5486.
34 T. L. Saaty, *J. Math. Psychol.*, 1977, **15**, 234–281.
35 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
36 G. Derringer and R. Suich, *J. Qual. Technol.*, 1980, **12**, 214–219.
37 D. J. Cummins and M. A. Bell, *J. Med. Chem.*, 2016, **59**, 6999–7010.
38 A. Krause and D. Golovin, *Tractability*, Cambridge University Press, 1st edn, 2014, pp. 71–104.
39 D. J. Woodward, A. R. Bradley and W. P. v. Hoorn, *J. Chem. Inf. Model.*, 2022.
40 R. P. Sheridan, B. P. Feuston, V. N. Maiorov and S. K. Kearsley, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1912–1928.
41 D. Ju-Long, *Syst. Control Lett.*, 1982, **1**, 288–294.
42 M. Langevin, M. Bianciotto and R. Vuilleumier, *Digital Discovery*, 2024.
43 J. MacQueen *et al.*, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.
44 L. Kaufman and P. J. Rousseeuw, *Wiley Series in Probability and Statistics*, 2010, pp. 68–125.
45 R. W. Kennard and L. A. Stone, *Technometrics*, 1969, **11**, 137–148.
46 C.-L. Hwang and K. Yoon, *Lecture Notes in Economics and Mathematical Systems*, 1981.
47 L. McInnes, J. Healy and J. Melville, *arXiv*, 2018, preprint, arXiv:1802.03426, DOI: 10.48550/arXiv.1805.11973.
48 D. Probst and J.-L. Reymond, *J. Cheminf.*, 2020, **12**, 12.
49 N. D. Cao and T. Kipf, *arXiv*, 2018, preprint, arXiv:1805.11973, DOI: 10.48550/arXiv.1805.11973.
50 P. Renz, S. Luukkonen and G. Klambauer, *J. Chem. Inf. Model.*, 2024, **64**, 5756–5761.
51 J. Zhang, R. Mercado, O. Engkvist and H. Chen, *J. Chem. Inf. Model.*, 2021, **61**, 2572–2581.
52 Y. Xie, Z. Xu, J. Ma and Q. Mei, *arXiv*, 2021, preprint, arXiv:2112.12542, DOI: 10.48550/arXiv.2112.12542.
53 M. Snarey, N. K. Terrett, P. Willett and D. J. Wilton, *J. Mol. Graphics Modell.*, 1997, **15**, 372–385.
54 M. Thomas, R. T. Smith, N. M. O'Boyle, C. d. Graaf and A. Bender, *J. Cheminf.*, 2021, **13**, 39.
55 Y. C. Martin, J. L. Kofron and L. M. Traphagen, *J. Med. Chem.*, 2002, **45**, 4350–4358.
56 G. Maggiora, M. Vogt, D. Stumpfe and J. Bajorath, *J. Med. Chem.*, 2014, **57**, 3186–3204.

57 M. Langevin, C. Grebner, S. Güssregen, S. Sauer, Y. Li, H. Matter and M. Bianciotto, *ACS Omega*, 2023, **8**, 23148–23167.
58 G. Landrum, Thresholds for "random" in fingerprints the RDKit supports – RDKit blog, https://greglandrum.github.io/rdkit-blog/posts/2021-05-18-fingerprint-thresholds1.html.
59 D. Bajusz, A. Rácz and K. Héberger, *J. Cheminf.*, 2015, **7**, 20.
60 R. E. Higgs, K. G. Bemis, I. A. Watson and J. H. Wikel, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 861–870.
61 P. Ertl and A. Schuffenhauer, *J. Cheminf.*, 2009, **1**, 8.
62 C. W. Coley, L. Rogers, W. H. Green and K. F. Jensen, *J. Chem. Inf. Model.*, 2018, **58**, 252–261.
63 A. Thakkar, V. Chadimová, E. J. Bjerrum, O. Engkvist and J.-L. Reymond, *Chem. Sci.*, 2021, **12**, 3339–3349.