Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development of an Open-Source RDKit-Based Pipeline for Early-Stage Drug Discovery and Virtual Screening of Large Chemical Libraries

Saeed Tayeb
Umm al-Qura University
Abdulrahman Rustom
Umm al-Qura University
Abdulelah Alfattani
Umm al-Qura University
Ahmed Alqurashy
Umm al-Qura University
Abdulrahman Almaghrabi
Umm al-Qura University
Khaled Alharbi
Umm al-Qura University
Hasan Altowairqi
Umm al-Qura University
Mostafa Marhoomi
Umm al-Qura University
Mohamed Gomaa
Imam Abdulrahman Bin Faisal University
Abdulaziz Al Khzem
ahalkhzem@iau.edu.sa

Imam Abdulrahman Bin Faisal University

Research Article

Keywords: Virtual screening, Drug-likeness, Cheminformatics, RDKit, Scaffold diversity, Medicinal chemistry filters, Chemical library curation, PAINS, Open-source

Posted Date: May 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9460567/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Drug discovery from large chemical databases is hindered by issues of structural redundancy and computational costs of virtual screening. Existing filtration tools either rely on restrictive commercial licenses or lack efficient multi-layered filtering capabilities.

Objective

To develop and validate an open-source, RDKit-based pipeline that integrates standardized molecular curation, drug-likeness filtering, and scaffold diversity optimization for large-scale pharmaceutical virtual screening workflows.

Methods

The pipeline was implemented in Python using RDKit, Pandas, and JupyterLab. It applies sequential standardization, PAINS and BRENK structural filters, and Lipinski, Ghose, Veber, Egan, and Muegge drug-likeness rules. Efficiency was benchmarked against ChEMBL, SwissADME, and Schrödinger using a dataset of 1,980 approved small-molecule drugs. Computational speed was evaluated across four dataset sizes (1k to 1M compounds) derived from merged PubChem and COCONUT natural product databases (n = 1,562,874). Scaffold diversity was assessed using the Murcko framework decomposition and cumulative scaffold frequency plots (CSFP) before and after optimization.

Results

The RDKit pipeline achieved > 97.8% agreement with ChEMBL across all physicochemical descriptors, including molecular weight (99.7%), HBA (99.4%), and LogP (98.5%). Processing throughput remained stable at 816–866 molecules/second across all tested scales, outperforming Schrödinger by up to 5.3-fold. Scaffold diversity improved markedly after optimization, with the scaffold-to-molecule ratio increasing from 0.480 to 0.753 at 100k compounds and the maximum scaffold frequency reduced from 6,351 to 42.

Conclusion

The proposed pipeline delivers efficient physicochemical profiling and effective scaffold redundancy reduction compared to established platforms. Its fully open-source design, multi-format compatibility, and scalable architecture make it a practical and reproducible tool for large-scale virtual screening.

1. Introduction

Drug discovery remains one of the major challenges in biomedical research. Conventionally, the process starts with identifying targets for specific diseases, followed by the development of thousands of synthesized compounds, multiple protein supplies, and established methods for laboratory bioactivity testing [1]. Development of a novel drug has a typical cost range of 1 to 2 billion USD and would take 10 to 17 years [2]. To accelerate drug discovery, computational methods have been widely used in the past thirty years. Virtual screening (VS) is one of widely utilized method, defined by the International Union of Pure and Applied Chemistry (IUPAC) as “computational methods that classify molecules in a database according to their ability to present biological properties against a given molecular target” [1, 2]. Computational methods can screen databases containing billions of compounds, at low cost and prioritize those to be tested before conducting biological testing, synthesized, or purchased [1, 3]. This approach has been playing a major role recently in drug discovery and development [4].

The chemical space is believed to consist of up to 10^{60} unique small compounds. To investigate and classify this chemical space, large chemical databases have been created, which have become a crucial resource for drug discovery. Public databases such as PubChem [5] and ChEMBL [6] have become essential drug discovery resources [7]. Similarly, natural product databases such as COCONUT [8] now contain over 700,000 compounds [9]. The continued generation of data has helped in the systematic collection of precise chemical and biological information. Over the years, the public databases have increased. Since researchers rely on the availability of such data for computations and experiments, the curation of the data becomes important [10, 11].

However, Evaluation of these large databases is challenging because of the significance of computing power needed. In addition, databases will continue expanding and reach trillions of compounds in the near future, which will be difficult to screen even with the fastest algorithms [12]. In addition, the challenge of data integrity remains unsolved, since many molecular databases still contain errors in the structure of the compounds. These observed errors can be summarized in a few categories, such as structure integrity, presence of undesired salts or mixtures, and duplicates [4]. Moreover, computational chemists still have not reached an agreement on a golden standard in molecular database preparation. This results in a variety of existing software, each having specific parameters fine-tuned by the user. Furthermore, most protocols use software with restrictive licenses, either commercial or academic. To validate or reproduce published results, restrictive licenses that are not always freely accessible are a limited factor for the scientific community [13]. Therefore, there is an urgent need for more efficient virtual screening approaches able to evaluate these vast chemical libraries [12].

Addressing these challenges requires integration of chemoinformatics, which is defined by Gasteiger and Engel as "the application of informatics methods to solve chemical problems" [14]. Considering chemoinformatics as an interdisciplinary field that integrates chemistry with computer science and data analysis to process chemical data at scale while ensuring quality and reproducibility [14, 15]. Using this approach has rapidly become a cornerstone of modern chemical research [14]. Initially in the

pharmaceutical industry, the field focused on quantitative structure activity relationships, molecular docking, and virtual screening, but has eventually developed to include data-driven methods that enable storage, retrieval, and analysis of large databases [14].

Among available chemoinformatics tools, RDKit has emerged as a widely used Python open-source cheminformatics toolkit that can be used in command-line use or as a web application to provide core functionalities, including molecular representation, molecular descriptor calculation, fingerprint generation, and substructure searching, making it suitable for constructing flexible and reproducible drug discovery workflows [16, 17]. Chemical knowledge obtained from scientific literature is the cornerstone of drug discovery, particularly with the aid of cheminformatics techniques to facilitate the systematic application of medicinal chemistry concepts [10, 11]. Implementing the RDKit toolset allows us to measure data quality before virtual screening. It will apply structural and physicochemical filters to compounds to determine which compounds pass through to be enrolled in in the virtual screening pipeline. Drug-likeness filters were established by major pharmaceutical companies aiming to improve the quality of their proprietary chemical collections [17]. These include Lipinski (Pfizer) [18] which is the pioneer rule-of-five, Ghose (Amgen) [19], Veber (GSK) [20], Egan (Pharmacia) [21], and Muegge (Bayer) [22]. Any violation of any mentioned rule appears in the output. Apart from the physicochemical evaluation, a structure-based inspection is used to improve the quality of the data and to identify problematic compounds. Pan Assay Interference Compounds (PAINS) are compounds that contain substructures showing potent response in assays regardless of the protein target. Such a compound, leading to false positive biological output [23]. In addition, the BRENK framework consists of a list of 105 fragments that are known to be potentially toxic, chemically reactive, unstable, and display poor pharmacokinetic properties [24]. These structural and drug-likeness filters are used to cleanse chemical libraries by removing compounds that are unstable, reactive, toxic, and interfere with biological assays [17]. On the other hand, one of many parameters that may be used to assess large databases is scaffold diversity, which assesses chemical diversity based on the scaffolds. One of the frameworks used for scaffolds is the Murcko framework, proposed by Bemis and Murcko. The Murcko framework of a structure consists of all the ring systems and all the linkers that connect to the ring systems. The frequency of Murcko frameworks or scaffolds has been used to define the structural diversity of chemical databases [25, 26].

To address these challenges, this study introduces an RDKit-based, fully open-source pipeline that integrates multi-layered data curation, drug-likeness filtering, and scaffold-aware redundancy control into a single, scalable workflow. Especially, by integrating standardized data cleaning, drug-like filtration, medicinal chemistry filters such as PAINS and Brenk, and scaffold-aware redundancy control into a single open-source workflow. Finally, this study aims to accelerate drug discovery research and enable large-scale studies by translating the principles of cheminformatics into an efficient computational workflow.

2. Methods

2.1 Implementation of RDKit-based pipeline

The pipeline has been developed using RDKit [16], open cheminformatics tool [27], Pandas library for data handling [28, 29], and management using Jupyter lab [30] as an environment. All commands were performed using Python version 3.10.15. The pipeline was built as a multistep process that can run and stream databases, molecules, and apply each phase or step on it [31]. The pipeline can handle multiple formats from a dataset, including SMILES (.smi), a command-separated file (.csv) that has a SMILE column, and a structure data file (.sdf). compressed SDF (.sdf.gz) (Fig. 1). The pipeline provides a sequential screen that can be modified based on the dataset size to ensure scalability and a smooth screen.

Screening starts by standardization for the dataset, removing invalid molecules, multicomponent structures such as salts and mixtures, metals, and non-organic molecules. Ensuring and focusing the dataset on small organic molecules [13, 32]. The standardization process aims to reduce the errors during virtual screening related to molecular representation.

To further eliminate problematic molecules, medicinal chemistry filters are applied, including PAINS (A, B, and C) and BRENK. Any molecule that triggers an alert will result in exclusion from the dataset. Resulting in enhancing the quality of the dataset by removing molecules with potential toxicity or assay interference.

Further evaluation involves assessing physicochemical properties by calculating molecular descriptors in accordance with drug-likeness rules [27, 29, 33]. The Lipinski Rule of Five and Veber rules are used. Compounds that meet the criteria are included for further assessment, which involves additional rules by Ghose, Egan, and Muegge. This step aims to rank compounds, and no compounds are excluded if any deviation from the criteria happens.

2.2 The process of evaluation for the RDKit-based pipeline

2.2.1 Accuracy and Correctness Evaluation

This step aims to evaluate the consistency of the results for the proposed RDKit-based pipeline screening tool with SwissADME [17], one of the well-established and widely used tools. The evaluation will be based on the physicochemical properties calculation that is included in the drug likeness rules criteria.

The ChEMPL (version 36) database, one of the well-established databases [6], was chosen to extract the dataset. The dataset for approved drugs was extracted with (n = 4.005) molecules, then we restricted it to small-molecule drugs (n = 3.280). Furthermore, specific criteria were applied to ensure comparable runs between RDKit pipeline and SwissADME. Invalid SMILES were removed, along with multicomponent and non-organic structures. Additionally, SMILES exceeding 200 characters were removed, as SwissADME cannot process molecules beyond this length. After applying these criteria, the final dataset

used for this step contained (n = 1,980) molecules. To quantify descriptor agreement between RDKit and the reference platforms, an “integer-digit” comparison metric was used. For each descriptor, values from RDKit and each comparator (ChEMBL, SwissADME, Schrödinger) were rounded to the nearest integer and classified as matching if their integer values were identical, and non-matching otherwise. Agreement was reported as the percentage of compounds with matching integer values for each descriptor and summarized as heatmaps and bar plots (Fig. 3).

2.2.2 Processing Speed Analysis

To evaluate the screening efficiency of the RDKit pipeline, natural product data were obtained from two publicly available databases: PubChem Natural Products (n = 853,148) [5] and COCONUT database (n = 715,822) [8]. After merging and removing redundant SMILES representations, the final curated dataset comprised n = 1,562,874 distinct natural product structures. Then the dataset was across four dataset sizes (n = 1,000, n = 10,000, n = 100,000, and n = 1,000,000) compounds to enable systematic performance assessment. Each subset underwent five independent processing iterations under standardized conditions on a MacBook equipped with an Apple M4 processor and 16 GB RAM. All metrics including runtime per second and molecule per second were recorded using Python.

2.2.3 Structural Diversity Analysis Using Murcko Scaffolds

To assess redundancy control, Murcko scaffolds were extracted from all molecules before and after the implementation of the developed RDKit-based pipeline. Screening was performed across four dataset sizes (n = 1,000, n = 10,000, n = 100,000, and n = 1,000,000) to evaluate pipeline stability and scalability. Each scaffold was represented as a canonical SMILES string. For each dataset, the total number of molecules (M), unique scaffolds (N), and singleton scaffolds (Ns) were quantified. Scaffold ratio (N/M) and singleton scaffold ratio (Ns/N) were calculated to measure library diversity. Cumulative scaffold frequency plots (CSFP) were generated to visualize molecular distribution across scaffolds using Matplotlib [34]. To examine structural diversity, Morgan fingerprints were computed for each molecule across (n = 1,000, n = 10,000, n = 100,000, n = 1000,000) dataset and reduced to two dimensions using. t-SNE. These 2D projections were plotted to assess chemical space preservation after filtering using Scikit-learn [35].

2.3 Statistical Analysis

The comparisons were performed using one-way ANOVA for normally distributed variables and the Friedman test for nonparametric repeated-measures data, as appropriate. In this work, statistical significance was defined at $p < 0.05$. All statistical analyses were performed using [GraphPad Prism version 10.0], and all tests were two-tailed unless otherwise specified.

3. Results

3.1 Physicochemical Descriptors Analysis

This study utilized several filtration criteria to compare the efficiency of the four molecular platforms (ChEMBL, Swiss, Schrödinger, and our developed code “RdKit”). One of the filtration criteria implemented in the developed workflow was the application of the Lipinski (rule of five), which is commonly used to evaluate drug-likeness based on key physicochemical properties, including molecular weight, lipophilicity (LogP), hydrogen bond donors (HBD), and hydrogen bond acceptors (HBA) [12]. To assess whether the implemented filtration process produced comparable physicochemical profiles, the filtered dataset was statistically compared with compounds from the other three platforms.

The molecular weight was observed to show no statistically significant differences among ChEMBL, RDKit, Swiss, and Schrödinger datasets (all comparisons, ns). However, the mean molecular weight values remained comparable across all platforms, indicating that the filtration workflow did not introduce systematic bias in molecular size selection. A statistically significant difference was detected on rotatable bonds. The Swiss dataset showed a higher mean number of rotatable bonds than both ChEMBL and RDKit ($p < 0.001$) and differed significantly from Schrödinger ($p < 0.01$). No significant difference was observed between ChEMBL and our code, RDKit (Fig. 2).

The Hydrogen Bond Acceptors (HBA) of Schrödinger-derived compounds showed a significantly lower mean HBA count compared with the other datasets (**$p < 0.001$). However, no significant differences were detected among ChEMBL, RDKit, and Swiss (ns). This indicates a stricter constraint on hydrogen bond acceptors within the Schrödinger filtering output. Similarly, Schrödinger displayed significantly reduced Hydrogen Bond Donors HBD values relative to the other three groups (**$p < 0.001$), while ChEMBL, RdKit, and Swiss did not significantly differ from each other (ns). This trend aligns with a more conservative hydrogen bonding profile in the Schrödinger-filtered compounds. Regarding the analysis of Topological Polar Surface Area (TPSA), the data revealed significant variations across platforms. Specifically, the Swiss and Schrödinger datasets demonstrated higher TPSA values than both ChEMBL and RDKit ($p < 0.001$), while no significant difference was also found between ChEMBL and RDKit.

This observation suggests that the implementation of our code RDKit workflow maintains physicochemical distributions that are closer to those observed in the reference database (ChEMBL). A reasonable explanation for this similarity is that the filtration rules implemented in the RDKit code were designed to closely follow commonly accepted drug-likeness constraints (e.g., Lipinski criteria), which are also broadly reflected in curated medicinal chemistry databases such as ChEMBL.

A more detailed comparison of descriptor agreement is presented in Fig. 3A, which reports the percentage of compounds with matching descriptor values relative to RDKit using an integer-digit comparison. The highest level of agreement was observed between RDKit and the ChEMBL dataset across all evaluated descriptors. Specifically, the matching percentage reached 99.7% for molecular weight (MW), 97.8% for rotatable bonds, and 97.8% for TPSA. Similarly high agreement was observed for hydrogen bonding descriptors, with 99.4% agreement for hydrogen bond acceptors (HBA) and 98.5% for hydrogen bond donors (HBD). Lipophilicity (LogP) also showed a high level of consistency at 98.5%.

In contrast, the agreement between RDKit and other computational platforms was noticeably lower for several descriptors. When compared with Schrödinger, matching percentages decreased substantially for certain properties, including HBA (27.6%), LogP (56.4%), and rotatable bonds (70.2%), while higher agreement remained for MW (99.2%) and TPSA (75.3%). A similar pattern was observed with SwissADME, where lower agreement was seen for LogP (32.4%), HBA (49.8%), and rotatable bonds (56.4%), while descriptors such as MW (99.8%), HBD (98.9%), and TPSA (73.7%) showed relatively higher consistency. These variations may reflect methodological differences in descriptor calculation, such as alternative algorithms for estimating lipophilicity, differences in hydrogen bond assignment rules, or variations in the treatment of molecular flexibility across cheminformatics platforms.

The distribution of matched and non-matched compounds is further illustrated in Fig. 3B, which provides the absolute number of compounds for each descriptor. Consistent with the percentage analysis, the ChEMBL comparison showed nearly complete agreement with RDKit across all descriptors, with almost all compounds classified as matched and only a very small number falling into the non-matched category. In contrast, both SwissADME and Schrödinger showed noticeably more non-matched compounds, particularly for descriptors such as LogP, HBA, and rotatable bonds. For example, SwissADME showed a substantial number of non-matching compounds for LogP and HBA, while Schrödinger also displayed increased non-matching counts for HBA and LogP, despite maintaining relatively high agreement for MW.

Taken together, the results presented in Figures B and C indicate that the RDKit-based workflow produces descriptor values that are highly consistent with those observed in the ChEMBL reference dataset. The lower agreement observed with SwissADME and Schrödinger appears to be descriptor-specific and likely arises from methodological differences in how these platforms calculate certain physicochemical properties rather than from inconsistencies in the compound dataset itself.

3.2 Processing Speed Analysis

To evaluate the runtime computational performance of the proposed RDKit pipeline, a large dataset of natural products was constructed by merging two publicly available databases: PubChem Natural Products and COCONUT. After removing duplicate SMILES representations, the final curated dataset contained 1,562,874 unique molecules. From this master dataset, four subsets were generated (1k, 10k, 100k, and 1M compounds) to enable a systematic comparison of processing speed across different dataset scales.

The comparison in runtime between RDKit and Schrödinger across datasets ranges from 1k to 1M compounds. RDKit consistently demonstrated lower execution times than Schrödinger across all tested scales. For the smallest dataset (1k compounds), RDKit completed processing in 1.19 seconds, whereas Schrödinger required approximately 3 seconds, corresponding to roughly a 2.5-fold speed improvement. As the dataset size increased, the performance difference became more pronounced. At 10k compounds, RDKit required 11.92 seconds compared with 63 seconds for Schrödinger, representing approximately a 5.3-fold speed advantage. For 100k compounds, RDKit processed the dataset in 122.53

seconds, while Schrödinger required 311 seconds, indicating a 2.5-fold faster runtime. At the largest scale tested (1 million compounds), RDKit completed the task in 1156.62 seconds, compared with 1807 seconds for Schrödinger, corresponding to an approximately 1.6-fold improvement in runtime. These results indicate that RDKit provides consistently faster processing across different dataset scales, particularly for medium-sized datasets (Fig. 4A).

Additionally, the processing throughput, expressed as molecules processed per second, was measured across the same dataset sizes. RDKit demonstrated relatively stable throughput across all tested scales, ranging from 816 to 866 molecules per second, indicating strong scalability and consistent computational performance. Specifically, RDKit processed approximately 837.8 molecules/s for 1k compounds, 839.1 molecules/s for 10k, 816.1 molecules/s for 100k, and 865.6 molecules/s for 1M compounds. In contrast, Schrödinger exhibited lower and more variable throughput, with values ranging from 158.8 to 333.2 molecules per second. For example, throughput dropped to 158.8 molecules/s at the 10k dataset size before increasing again at larger scales. This variability suggests that the Schrödinger pipeline may involve additional processing steps or overhead that affect performance consistency (Fig. 4B).

The overall performance trend was observed across the runtime and throughput analyses. The results collectively indicate that the RDKit-based workflow maintains higher processing efficiency and more stable throughput compared with the Schrödinger platform across increasing dataset sizes. The relatively constant throughput observed for RDKit suggests that the workflow scales efficiently with increasing dataset size, making it suitable for large-scale compound screening and cheminformatics workflows. In contrast, the greater runtime and variability in throughput observed for Schrödinger indicate comparatively higher computational overhead under the tested conditions.

The number of compounds passing the filtration step differed between the two platforms. Schrödinger retained a higher number of compounds after filtration compared with the RDKit pipeline across all dataset sizes. For example, in the 1k dataset, Schrödinger retained 684 compounds, whereas the RDKit workflow retained 229 compounds. A similar trend was observed for larger datasets, where Schrödinger retained 6,627 compounds from the 10k dataset, 66,529 from the 100k dataset, and 665,062 from the 1M dataset, while the RDKit pipeline retained 2,057, 17,057, and 124,087 compounds, respectively. This difference can be attributed to the more comprehensive filtration strategy implemented in the RDKit workflow. In the Schrödinger run, the filtration primarily relied on Lipinski's rule of five, which focuses on basic drug-likeness properties such as molecular weight, LogP, hydrogen bond donors, and hydrogen bond acceptors. In contrast, the RDKit pipeline applied multiple additional molecular descriptors and structural filters such as Veber, Ghose, Egan and Muegge (Fig. 1) which may result in a greater number of compounds being excluded during the screening process compared with the Schrödinger setup used in this study (Fig. 4C).

3.3 Scaffold Diversity Analysis

To evaluate the impact of the optimization procedure on the chemical diversity of generated compound libraries, we performed a scaffold-based analysis using Murcko scaffold decomposition across four library sizes (1k, 10k, 100k, and 1M molecules). Scaffold diversity was assessed both visually through Cumulative Scaffold Frequency Plots (CSFPs) and quantitatively through key scaffold statistics, comparing distributions before and after optimization (Table 1, Fig. 5).

The CSFP curves consistently demonstrated that optimization improved scaffold diversity across all library sizes. In all four panels, the "Before" curves (orange) exhibited pronounced convexity, indicating that a small fraction of scaffolds accounted for a disproportionately large share of molecules a hallmark of redundant, scaffold-biased libraries. Following optimization, the "After" curves (green) shifted markedly toward the diagonal reference line, which represents an ideal uniform scaffold distribution, indicating a more balanced representation of chemical scaffolds. This trend was observed at every scale tested, and became increasingly pronounced at larger library sizes (100k and 1M), suggesting that the optimization procedure scales effectively and remains beneficial even as the number of molecules grows substantially.

Quantitative Scaffold Statistics. The scaffold-level statistics corroborated the visual findings (Table 1). The ratio of unique scaffolds to total molecules (N/M) increased substantially after optimization at every library size: from 0.837 to 0.954 at 1k, from 0.689 to 0.864 at 10k, and from 0.480 to 0.753 at 100k, reflecting a marked reduction in scaffold redundancy. Similarly, the proportion of singleton scaffolds (Ns/N) those represented by only a single molecule increased after optimization across all conditions, further confirming the shift toward greater scaffold novelty and reduced over representation of common scaffolds.

Most strikingly, the maximum number of molecules sharing a single scaffold decreased dramatically after optimization. At the 1k scale, this value dropped from 65 (before) to just 3 (after), and at the 10k scale from 628 to 10, indicating that the optimization procedure effectively dismantled dominant scaffold clusters. At the 100k scale, the maximum was reduced from 6,351 to 42, demonstrating that even highly recurrent scaffold families were substantially suppressed.

Taken together, these results demonstrate that the optimization procedure reliably enhances scaffold diversity across library sizes spanning three orders of magnitude, producing compound collections with more uniform scaffold distributions, higher proportions of singleton scaffolds, and substantially reduced scaffold redundancy all desirable properties for broad chemical space exploration in drug discovery applications.

The distribution of chemical space was visualized using Morgan fingerprints projected into two dimensions for four dataset sizes (1k, 10k, 100k, and 1M compounds), comparing the compound sets before and after the application of the workflow (Fig. 6). At the smallest scale (1k), both datasets appear sparsely distributed with partial overlap, indicating that some regions of chemical space are shared while others remain unique to each set. As the dataset size increases to 10k and 100k compounds, the density of points increases and the overlap between the two distributions becomes more pronounced,

suggesting broader chemical space coverage. At larger scales (100k and 1M), the “After” dataset occupies a wider region of the chemical space, indicating improved representation of structural diversity. Overall, these results demonstrate that the workflow preserves broad chemical space coverage while enabling exploration of diverse molecular regions as the library size increases.

Each subplot represents the chemical space for a specific number of compounds, with Component 1 and Component 2 serving as the reduced dimensions. The primary objective of such visualizations is to understand the diversity, coverage, and structural relationships within a chemical library, and how these properties change under different conditions or processes (represented by 'Before' and 'After')

4 Discussion

The development of open source and efficient cheminformatics pipelines is essential to address the rising costs and timelines of drug discovery. Standardizing chemical structures remains a primary challenge, as uncurated datasets often contain inconsistencies that hinder reproducible research [1, 36]. Integrating RDKit within distributed computing environments, such as Apache Hadoop and Spark, allows for the processing of millions of compounds with significantly improved speed compared to traditional single-node setups [3]. This approach is particularly critical given the variability in descriptor implementations across different software packages [2]. Although the present implementation was evaluated on a single consumer-grade laptop, our results demonstrate that the RDKit-based workflow already achieves stable throughput on libraries up to 1 million molecules, and its modular design makes it amenable to future deployment on distributed architectures. Furthermore, the availability of comprehensive, standalone packages like ChemSuite enables the calculation of diverse 1D, 2D, and 3D descriptors and fingerprints within a unified interface, facilitating the development of machine learning models for predicting biological and toxicological properties [4, 33].

In this study, three libraries have been used in the developed pipeline which offer scaffold diversity. The collections of smaller libraries built around distinct scaffolds offer significantly higher shape diversity and a better chance of addressing a broad range of biological targets [5, 36]. This is further supported by comparative analyses of screening libraries and databases, which use tools like Murcko frameworks scaffolds to quantify diversity through unique scaffold counts [37]

By utilizing optimized open-source pipelines to curate and filter these diverse libraries, researchers can move beyond redundant chemical space and prioritize compounds with the highest potential for biological activity and novelty.

One of the primary requirements of any virtual screening workflow is the efficient computation of physicochemical descriptors and the application of drug-likeness rules [38], such as the Lipinski Rule of Five [18], the Ghose filter [19], Veber criteria [20], and the Egan and Muegge rules [21, 22]. Our results indicate that the RDKit pipeline produces descriptor values that are highly concordant with those of the ChEMBL reference dataset across all evaluated properties. In contrast, notable discrepancies were

observed between RDKit and both SwissADME and Schrödinger for specific descriptors, particularly LogP and HBA.

For SwissADME, the agreement for LogP was 32.4%, and for HBA it was 49.8%. For Schrödinger, matching was 56.4% for LogP and 27.6% for HBA. These discrepancies are not unexpected, as different platforms employ distinct algorithms for lipophilicity estimation and hydrogen bond assignment [38]. For example, SwissADME uses the WLOGP and iLOGP methods among others, while RDKit applies the Wildman-Crippen approach. Such methodological divergence has been previously documented in the literature and reflects a broader lack of standardization in cheminformatics descriptor calculation [13, 15, 19, 38]. In addition, the differences observed between RDKit and Schrödinger with respect to HBA and HBD are consistent with the more conservative hydrogen bonding criteria imposed by Schrödinger's filtration settings, which appear to apply different rules than those used in the Lipinski framework employed by our pipeline.

From a practical standpoint, the close agreement between RDKit and ChEMBL one of the most widely used and rigorously curated medicinal chemistry databases validates our implementation of the drug-likeness filters. It also suggests that the pipeline reflects the physicochemical profiles of established drugs, which is a critical attribute for any tool intended to support early-stage drug discovery. The variability observed with SwissADME and Schrödinger for certain descriptors does not necessarily indicate errors in any platform, but rather reflects the inherent sensitivity of certain properties to the underlying calculation method. These differences highlight the importance of using consistent tools throughout a virtual screening campaign and underscore the value of employing openly documented, reproducible methods such as those provided by RDKit [16].

The processing speed analysis demonstrated that the RDKit pipeline consistently outperforms the Schrödinger platform across all tested dataset scales [39]. Additionally, RDKit maintained stable throughput across all scales, whereas Schrödinger exhibited greater variability (Fig. 4). This throughput stability is a key indicator of scalability and suggests that the RDKit pipeline will continue to perform efficiently as databases grow toward the trillion-compound scale anticipated in the near future [12].

These performance characteristics are particularly significant given the path of chemical library expansion. Public databases such as PubChem and COCONUT are growing rapidly [39], and emerging ultra-large virtual screening approaches routinely handle libraries of billions of compounds [13, 40]. The ability to process one million molecules at a stable rate of ~ 866 molecules per second on a consumer-grade laptop without requiring specialized hardware is a meaningful practical advantage. This stands in contrast to many high-throughput docking solutions that require GPU clusters or institutional computing infrastructure [4, 12].

Furthermore, it is important to consider the difference in retention rates observed between the two platforms. Schrödinger retained substantially more compounds across all dataset sizes compared to RDKit. This disparity reflects the broader and more multi-layered filtration strategy added to RDKit

pipeline, which incorporates not only Lipinski's Rule of Five but also PAINS [23] and BRENK [24] structural alerts, alongside additional physicochemical descriptors.

The development of the RDKit-based pipeline addresses a critical bottleneck in modern drug discovery: the need for efficient, open-source, and reproducible workflows for curating vast chemical libraries. The high level of agreement observed between the calculated descriptors and ChEMPL datasets supports the reliability of the developed pipeline. This suggests that the proposed workflow can serve as a reliable alternative to other computational platforms, which may exhibit greater variability in descriptors calculations [1].

The results regarding scaffold diversity analysis highlight a central theme in chemoinformatics: library size is not an inherent proxy for structural diversity [41]. The optimization procedure implemented in this workflow successfully managed dominant scaffold clusters, moving the library toward a more uniform exploration of chemical space. This shift is quantitatively reflected in the increase of the unique scaffold-to-molecule (N/M) ratio and the proportion of singleton scaffolds.

These findings align with the principle that smaller, multi-scaffold collections can produce higher shape diversity and more consistent biological hit rates than numerically larger, single-scaffold libraries [4, 41]. By reducing scaffold redundancy, the workflow effectively addresses the "early enrichment" problem, where a few common chemotypes disproportionately dominate a collection, a phenomenon frequently observed even in large public databases [5]. Furthermore, the integration of comprehensive descriptor calculations within a unified interface, as seen in similar standalone packages, facilitates the subsequent development of predictive models [6].

Integrating natural products into such optimized workflows is particularly promising for expanding structural variety [42]. These compounds often occupy unique regions of the chemical space that are structurally distant from approved drugs while maintaining favorable drug-like properties [7]. The ability of this RDKit-based pipeline to curate these diverse sets efficiently while ensuring structural complexity provides a valuable resource for identifying hits against challenging biological targets.

The pipeline workflow in the future could be extended to incorporate machine learning-based pre-filtering to further accelerate the screening of ultra-large libraries [39]. The filtration design of the pipeline also enables future integration of machine learning approaches to support compound prioritization and improve virtual screening workflows. Additionally, incorporating protein-specific pharmacophore or target chemical family filtering could improve hit rates for focused virtual screening workflows.

5. Conclusions

In conclusion, this work establishes that the proposed RDKit-based pipeline is an efficient and freely accessible tool for molecule curation and large-scale virtual screening. By combining descriptor calculations, scalable throughput, effective compound filtration, and an entirely open-source architecture, the pipeline addresses a critical unmet need in the cheminformatics community namely, the availability

of reproducible, license-free workflows capable of handling the rapidly expanding chemical libraries used in modern drug discovery. By translating established medicinal chemistry principles into an efficient, freely accessible workflow, this study contributes a valuable tool for accelerating early-stage drug discovery in both academic and resource-limited applications and provides a strong foundation for future integration with three-dimensional conformer generation, machine learning-based pre-filtering, and target-based pharmacophore screening strategies.

Declarations

Author Contributions: Conceptualization, S.M.T., A.H.A.K., ; Methodology, S.M.T., A.A.A., A.H.A.K., M.S.G., A.K.R, A.E.A., A.I.A., K.T.A., H.Y.A., M.J.M.; Software, S.M.T., A.A.A., A.K.R, A.E.A., A.I.A., K.T.A., H.Y.A., M.J.M.; Formal Analysis, S.M.T., A.A.A., A.H.A.K., M.S.G., A.K.R, A.E.A., A.I.A., K.T.A., H.Y.A., M.J.M ;Investigation, S.M.T., A.A.A., A.H.A.K., M.S.G.; Resources, S.M.T., A.A.A., A.H.A.K., M.S.G., A.K.R, A.E.A., A.I.A., K.T.A., H.Y.A., M.J.M.; Data Curation, S.M.T., A.A.A., A.H.A.K., M.S.G., A.K.R, A.E.A., A.I.A., K.T.A., H.Y.A., M.J.M; Writing—Original Draft Preparation, S.M.T., A.A.A., A.H.A.K., M.S.G.; Writing—Review and Editing, A.K.R, A.E.A., A.I.A., K.T.A., H.Y.A., M.J.M.; Visualization, A.K.R, A.E.A., A.I.A., K.T.A., H.Y.A., M.J.M.; Supervision, S.M.T., A.A.A., Project Administration, S.M.T., A.H.A.K.; Funding Acquisition, S.M.T., A.A.A., A.H.A.K., M.S.G., A.K.R, A.E.A., A.I.A., K.T.A., H.Y.A., M.J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The developed RDKit-based pipeline can be accessed in https://github.com/abdulrhmanrustom-prog/RDKit-screening-pipeline-for-large-databases.git

Acknowledgments: The authors would like to thank Umm Al-Qura University for providing the academic environment that supported this research. We also acknowledge the developers of open-source cheminformatics tools, particularly RDKit, whose contributions made this work possible.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kimber TB, Chen Y, Volkamer A (2021) Deep learning in virtual screening: recent applications and developments. Int J Mol Sci 22:4435
2. Oliveira TAd, Silva MPd, Maia EHB, Silva AMd, Taranto AG (2023) Virtual Screening Algorithms in Drug Discovery: A Review Focused on Machine and Deep Learning Methods. Drugs Drug Candidates 2:311–334

3. Giordano D, Biancaniello C, Argenio MA, Facchiano A (2022) Drug Design by Pharmacophore and Virtual Screening Approach. Pharmaceuticals 15:646
4. Zhou G, Rusnac D-V, Park H, Canzani D, Nguyen HM, Stewart L, Bush MF, Nguyen PT, Wulff H, Yarov-Yarovoy V et al (2024) An artificial intelligence accelerated virtual screening platform for drug discovery. Nat Commun 15:7761. 10.1038/s41467-024-52061-7
5. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B et al (2024) PubChem 2025 update. Nucleic Acids Res 53:D1516–D1525. 10.1093/nar/gkae1059
6. Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, de Veij M, Ioannidis H, Lopez DM, Mosquera JF et al (2023) The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res 52:D1180–D1192. 10.1093/nar/gkad1004
7. Kuan J, Radaeva M, Avenido A, Cherkasov A, Gentile F (2023) Keeping pace with the explosive growth of chemical libraries with structure-based virtual screening. WIREs Comput Mol Sci 13:e1678. https://doi.org/10.1002/wcms.1678
8. Chandrasekhar V, Rajan K, Kanakam, Sri Ram S, Sharma N, Weißenborn V, Schaub J, Steinbeck C (2024) COCONUT 2.0: a comprehensive overhaul and curation of the collection of open natural products database. Nucleic Acids Res 53:D634–D643. 10.1093/nar/gkae1063
9. Zeng T, Li J, Wu R (2024) Natural product databases for drug discovery: Features and applications. *Pharmaceutical Science Advances 2*, 100050. https://doi.org/10.1016/j.pscia.2024.100050
10. Zhu H (2020) Big Data and Artificial Intelligence Modeling for Drug Discovery. Annu Rev Pharmacol Toxicol 60:573–589. https://doi.org/10.1146/annurev-pharmtox-010919-023324
11. Marbán-González A, Ramírez-Cid V, Cristóbal-Ramírez A, Medina-Franco JL (2025) Exploiting PubChem and other public databases for virtual screening in 2025: what are the latest trends? Expert Opin Drug Discov 20:1387–1403. 10.1080/17460441.2025.2558161
12. Luttens A, Cabeza de Vaca I, Sparring L, Brea J, Martínez AL, Kahlous NA, Radchenko DS, Moroz YS, Loza MI, Norinder U et al (2025) Rapid traversal of vast chemical space using machine learning-guided docking screens. Nat Comput Sci 5:301–312. 10.1038/s43588-025-00777-x
13. Gally J-M, Bourg S, Do Q-T, Aci-Sèche S, Bonnet P, VSPrep: (2017) A General KNIME Workflow for the Preparation of Molecules for Virtual Screening. Mol Inf 36:1700023. https://doi.org/10.1002/minf.201700023
14. Chetry AB, Ohto K (2025) From molecules to data: the emerging impact of chemoinformatics in chemistry. J Cheminform 17:121. 10.1186/s13321-025-00978-6
15. Lo Y-C, Rensi SE, Torng W, Altman RB (2018) Machine learning in chemoinformatics and drug discovery. Drug Discovery Today 23:1538–1546. https://doi.org/10.1016/j.drudis.2018.05.010
16. Landrum G, Tosco P, Kelly B, Rodriguez R, Cosgrove D, Vianello R, Gedeck P, Jones G, Kawashima E, Schneider (2015), N. RDKit
17. Daina A, Michielin O, Zoete V (2017) SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep 7:42717.

10.1038/srep42717

18. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 3–25.1. *Advanced Drug Delivery Reviews* 2001, *46*, 3–26. https://doi.org/10.1016/S0169-409X(00)00129-0
19. Ghose AK, Viswanadhan VN, Wendoloski JJ (1999) A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. J Comb Chem 1:55–68. 10.1021/cc9800071
20. Veber DF, Johnson SR, Cheng H-Y, Smith BR, Ward KW, Kopple KD (2002) Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. J Med Chem 45:2615–2623. 10.1021/jm020017n
21. Egan WJ, Merz KM, Baldwin JJ (2000) Prediction of Drug Absorption Using Multivariate Statistics. J Med Chem 43:3867–3877. 10.1021/jm000292e
22. Muegge I, Heald SL, Brittelli D (2001) Simple Selection Criteria for Drug-like Chemical Matter. J Med Chem 44:1841–1846. 10.1021/jm015507e
23. Baell JB, Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J Med Chem 53:2719–2740
24. Brenk R, Schipani A, James D, Krasowski A, Gilbert IH, Frearson J, Wyatt PG (2008) Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. ChemMedChem: Chem Enabling Drug Discovery 3:435–444
25. Bemis GW, Murcko MA (1996) The Properties of Known Drugs. 1. Molecular Frameworks. J Med Chem 39:2887–2893. 10.1021/jm9602928
26. Egieyeh S, Syce J, Christoffels A, Malan SF (2016) Exploration of Scaffolds from Natural Products with Antiplasmodial Activities, Currently Registered Antimalarial Drugs and Public Malarial Screen Data. Molecules 21:104
27. Aires-de-Sousa JGUIDEMOL (2024) A Python graphical user interface for molecular descriptors based on RDKit. Mol Inf 43:e202300190. https://doi.org/10.1002/minf.202300190
28. McKinney W (2010) Data structures for statistical computing in Python. *scipy 445*, 51–56
29. Kunnakkattu IR, Choudhary P, Pravda L, Nadzirin N, Smart OS, Yuan Q, Anyango S, Nair S, Varadi M, Velankar S (2023) PDBe CCDUtils: an RDKit-based toolkit for handling and analysing small molecules in the Protein Data Bank. J Cheminform 15:117
30. Wijayaningrum VN, Lestari VA Jupyter lab platform-based interactive learning. In Proceedings of 2022 International Conference on Electrical and Information Technology (IEIT); pp. 295–301
31. Ramírez-Márquez CD, Medina-Franco JL (2025) KNIME workflows for chemoinformatic characterization of chemical databases. Mol Inf 44:e202400337

32. Bento AP, Hersey A, Félix E, Landrum G, Gaulton A, Atkinson F, Bellis LJ, De Veij M, Leach AR (2020) An open source chemical structure curation pipeline using RDKit. J Cheminform 12:51
33. Tangadpalliwar SR, Vishwakarma S, Nimbalkar R, Garg P (2019) ChemSuite: A package for chemoinformatics calculations and machine learning. Chem Biol Drug Des 93:960–964
34. Hunter JD, Matplotlib (2007) A 2D Graphics Environment. Comput Sci Eng 9:90–95. 10.1109/MCSE.2007.55
35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830
36. Shang J, Sun H, Liu H, Chen F, Tian S, Pan P, Li D, Kong D, Hou T (2017) Comparative analyses of structural features and scaffold diversity for purchasable compound libraries. J Cheminform 9:25
37. Yongye AB, Waddell J, Medina-Franco JL (2012) Molecular scaffold analysis of natural products databases in the public domain. Chem Biol Drug Des 80:717–724
38. Guha R, Willighagen E (2012) A survey of quantitative descriptions of molecular structure. Curr Top Med Chem 12:1946–1956
39. Lovrić M, Molero JM, Kern R (2019) PySpark and RDKit: moving towards big data in cheminformatics. Mol Inf 38:1800082
40. Fang ZH, Sim BYC, Gunasinghe KKJ, Shabbir S, Ginjom IRH, San HS, Lau BT, Wezen XC (2026) Hit identification in ultra large virtual screening: an integrative review and future challenges. Drug Discovery Today 104616
41. Sauer WH, Schwarz MK (2003) Size doesn't matter: Scaffold diversity, shape diversity and biological activity of combinatorial libraries. Chimia 57:276–276
42. Vivek-Ananth R, Sahoo AK, Baskaran SP, Samal A (2023) Scaffold and structural diversity of the secondary metabolite space of medicinal fungi. ACS omega 8:3102–3113

Tables

Table 1 is not available with this version.

Figures

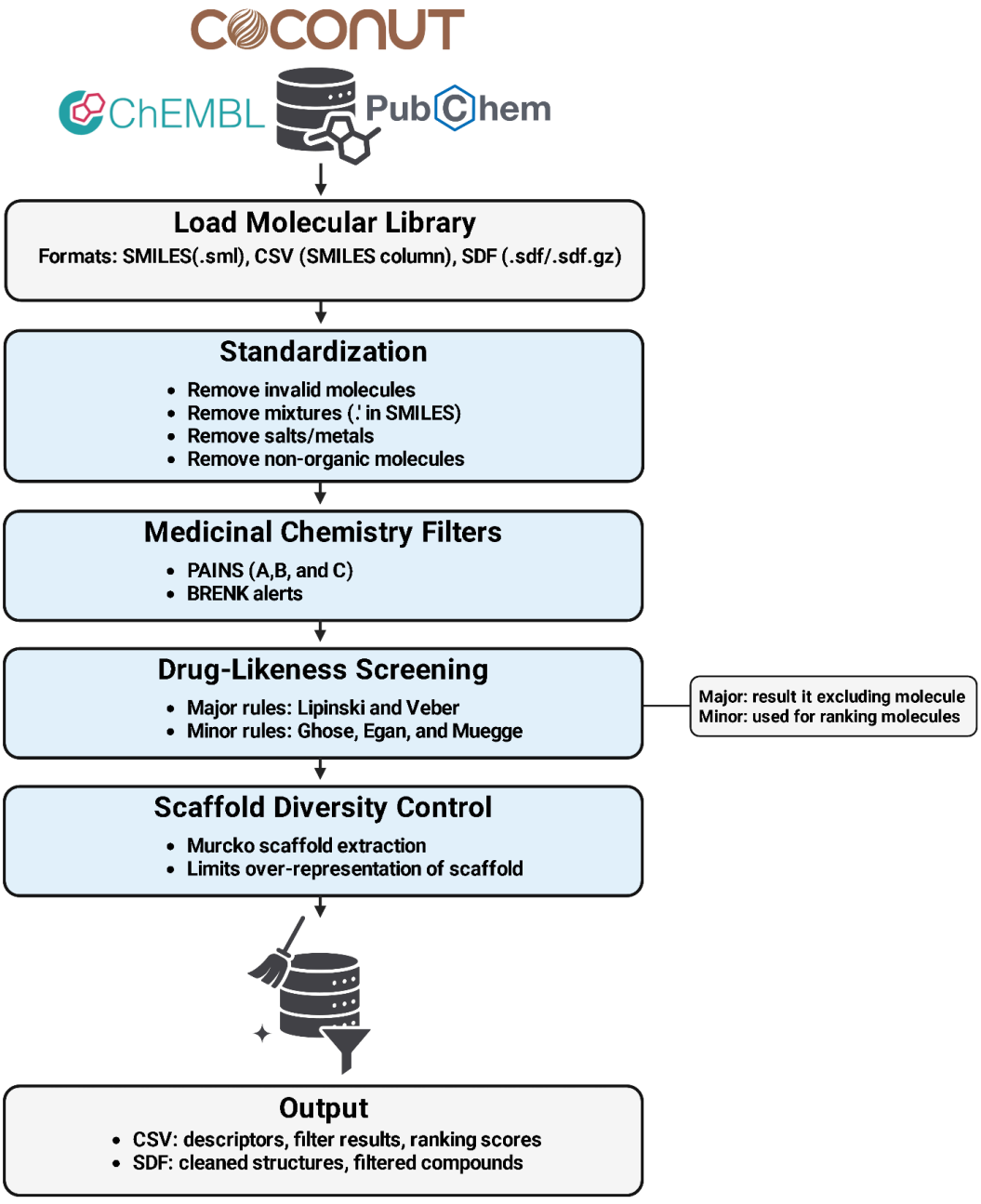


Figure 1

RDKit-Based Pipeline for early-stage Drug Discovery and Virtual Screening of Large Chemical Libraries. First, the process began by reading molecular libraries in different formats. Second, the data were processed by standardizing the chemical structures, resulting in a collection of chemically valid small organic molecules. Third, applying medicinal chemistry filters, the remaining molecules are evaluated according to drug-likeness criteria. Structural diversity was maintained by extracting Murcko

scaffolds and limiting the representation of highly populated chemotypes in the final dataset. Finally, the dataset was exported as both a CSV SDF containing the cleaned and filtered molecular structures, ready for subsequent virtual screening.

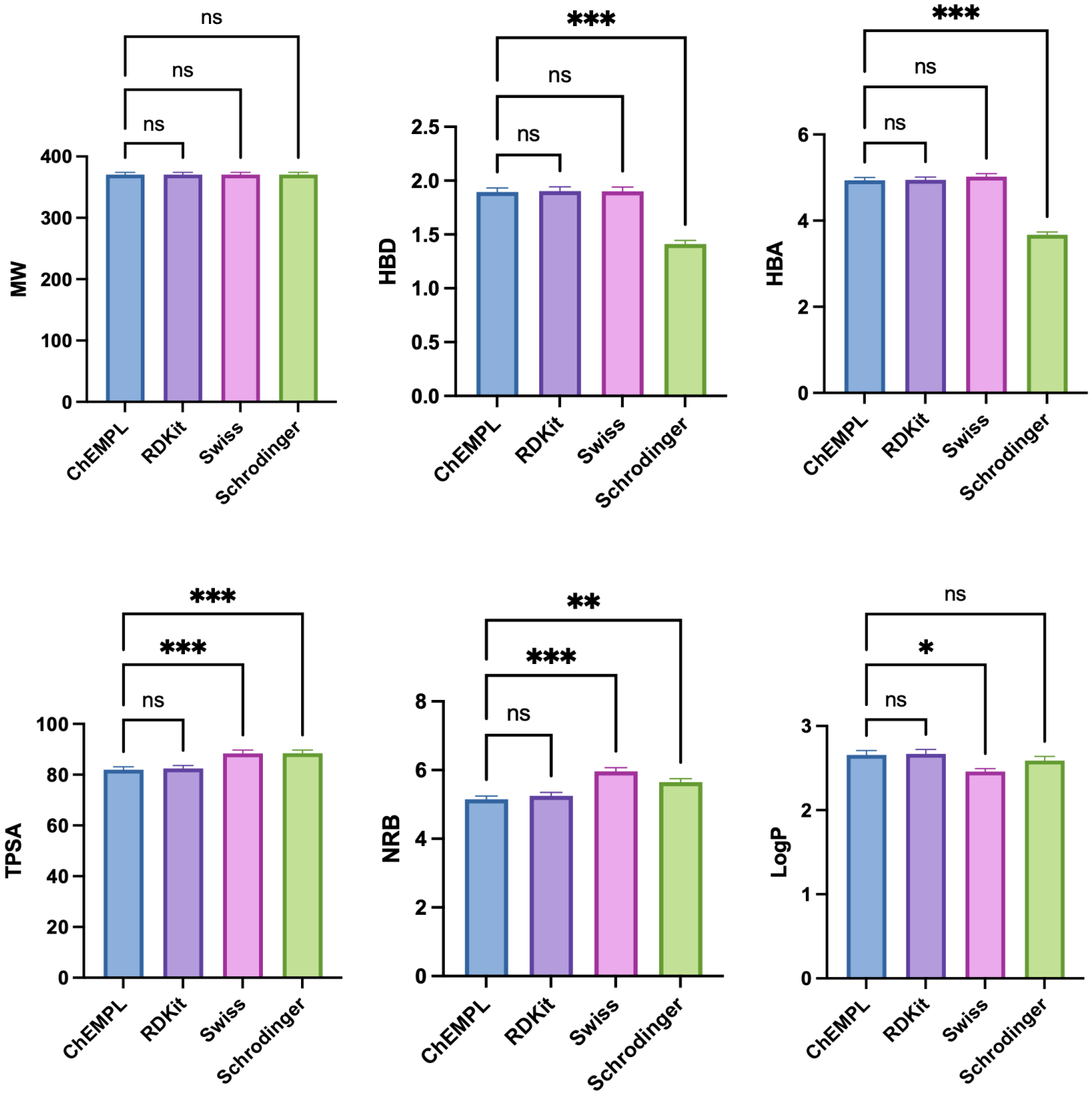


Figure 2

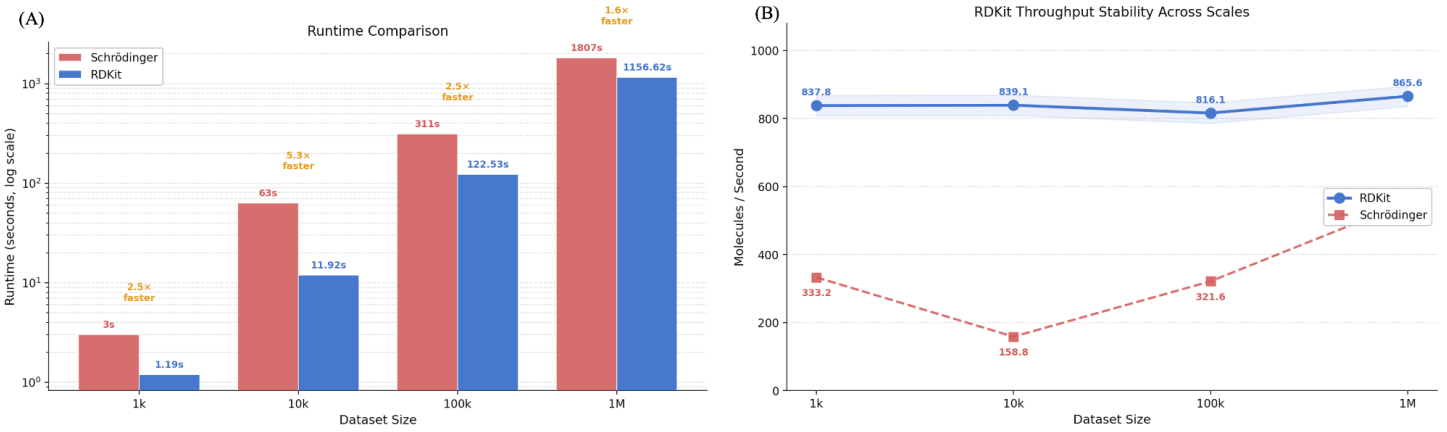
RDKit-Based Pipeline for early-stage Drug Discovery and Virtual Screening of Large Chemical Libraries. First, the process began by reading molecular libraries in different formats. Second, the data were processed by standardizing the chemical structures, resulting in a collection of chemically valid

small organic molecules. Third, applying medicinal chemistry filters, the remaining molecules are evaluated according to drug-likeness criteria. Structural diversity was maintained by extracting Murcko scaffolds and limiting the representation of highly populated chemotypes in the final dataset. Finally, the dataset was exported as both a CSV SDF containing the cleaned and filtered molecular structures, ready for subsequent virtual screening.



Figure 3

(A) Agreement analysis of physicochemical descriptors calculated using RDKit compared with values obtained from ChEMBL, Schrödinger, and SwissADME. Agreement was assessed using an integer-digit comparison for six descriptors: molecular weight (MW), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), lipophilicity (LogP), topological polar surface area (TPSA), and rotatable bonds. Figure B presents the percentage of compounds with matching descriptor values relative to RDKit, visualized as a heatmap where darker green indicates higher agreement. **(B)** The analysis shows the total number of compounds with matching and non-matching descriptor values for each evaluated physicochemical property. Overall, the ChEMBL dataset demonstrates the highest agreement with RDKit across all descriptors, while SwissADME and Schrödinger show lower agreement for some properties, particularly LogP, HBA, and rotatable bonds. These differences likely reflect variations in descriptor calculation methods among the computational platforms.



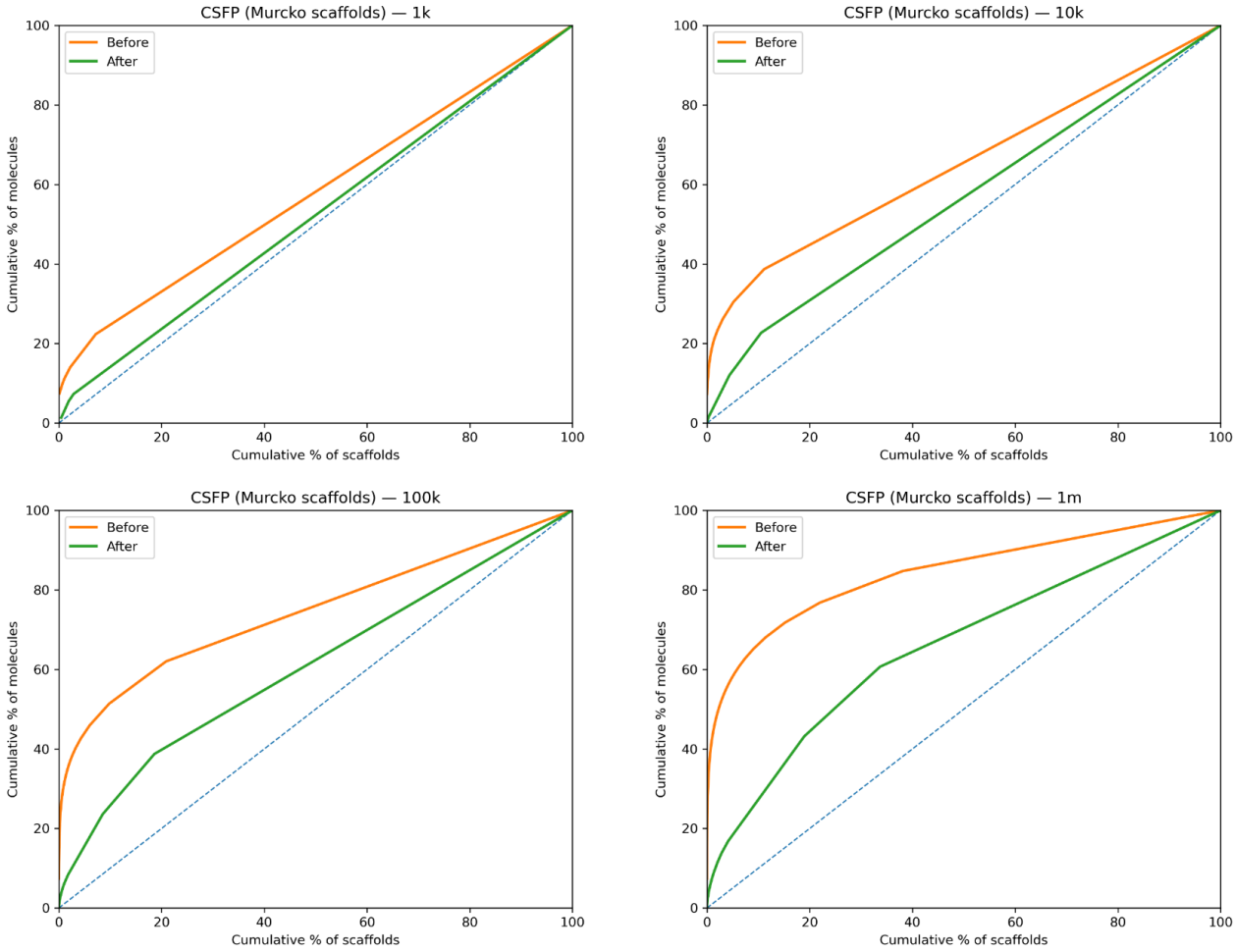
(C)

Dataset	Molecules	Pass filtration (Schrodinger)	Pass filtration (RDKit)	Runtime (s) (Schrodinger)	Runtime (s) (RDKit)	Mol/sec (Schrodinger	Mol/sec (RDKit)
1k	1,000	684	229	3	1.19	333.2	837.8
10k	10,000	6627	2,057	63	11.92	158.8	839.1
100k	100,000	66529	17,057	311	122.53	321.6	816.1
1M	1,000,000	665062	124,087	1807	1156.62	553.4	865.6

Figure 4

Runtime comparison between the RDKit pipeline and the Schrödinger platform across increasing dataset sizes (1k, 10k, 100k, and 1M compounds). (A) Runtime comparison showing the total processing time required by each platform for screening datasets of different scales. RDKit consistently demonstrates shorter execution times compared with Schrödinger. (B) Throughput performance is expressed as molecules processed per second, illustrating stable performance and scalability of the RDKit workflow across increasing dataset sizes. (C) Overall, the results highlight the higher computational efficiency and scalability of the RDKit-based pipeline for large-scale compound screening.

(A)



(B)

Dataset	Molecules (M)		With Scaffold		Unique Scaffolds (N)		Singleton Scaffolds (Ns)		N/M		Ns/N		Max Mol / Scaffold	
	B.	A.	B.	A.	B.	A.	B.	A.	B.	A.	B.	A.	B.	A.
1k	1k	229	863	219	722	209	670	203	0.8366	0.9543	0.928	0.9713	65	3
10k	10k	2057	8545	1964	5894	1697	5239	1518	0.6898	0.8641	0.8889	0.8945	628	10
100k	100k	17057	85951	16124	41263	12136	32643	9878	0.4801	0.7527	0.7911	0.8139	6351	42
1M	1M	124087	859102	114442	211864	67878	131162	45009	0.2466	0.5931	0.6191	0.6631	63727	143

Figure 5

Scaffold diversity of generated compound libraries before and after optimization, assessed by Cumulative Scaffold Frequency Plots (CSFP) using Murcko scaffolds. Each panel displays the cumulative percentage of molecules (y-axis) as a function of the cumulative percentage of unique Murcko scaffolds (x-axis), plotted in decreasing order of scaffold frequency, for library sizes of 1k, 10k, 100k, and 1M molecules. The dashed diagonal line represents a perfectly uniform scaffold distribution.

Curves closer to the diagonal indicate greater scaffold diversity. Orange curves ("Before") and green curves ("After") represent the scaffold frequency distributions prior to and following optimization, respectively. The shift of the "After" curves toward the diagonal across all library sizes demonstrates that the optimization procedure substantially improves scaffold diversity, reducing the dominance of frequently recurring scaffolds and yielding a more uniform exploration of chemical space.

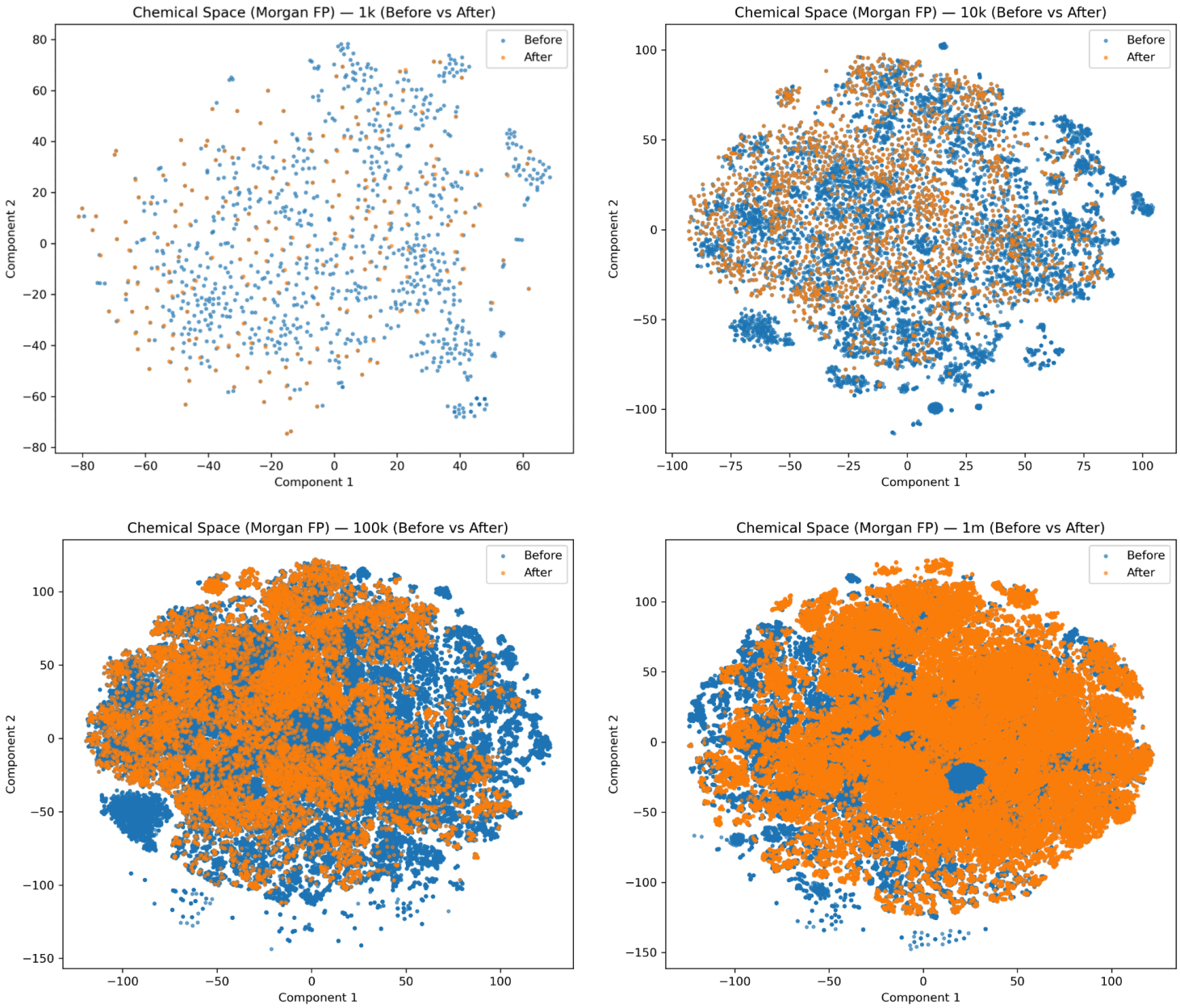


Figure 6

Visualization of chemical space of the four throughput scales which compares the compound libraries before and after application of the RDKit filtering workflow. The figure was generated using Morgan fingerprints (FP) and a dimensionality reduction technique (t-SNE). These plots illustrate the distribution of chemical compounds in a two-dimensional space across different dataset sizes: 1k, 10k, 100K and 1M compounds.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- floatimage1.png