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.

MDSLabChemBridge: Multi-Engine Molecular Descriptor Generation and ML-Ready Feature Engineering

Yogesh Kumar

Kumar.yogesh601@gmail.com

Centre of Biomedical Research

Research Article

Keywords: Cheminformatics, Molecular descriptors, Chemical fingerprints, Machine learning, Drug discovery, RDKit, CDK, Mordred

Posted Date: March 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9251519/v1

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

Additional Declarations: No competing interests reported.

Title:

MDSLabChemBridge: Multi-Engine Molecular Descriptor Generation and ML-Ready Feature Engineering

Abstract

Background: Molecular descriptors and fingerprints are the primary means of encoding chemical structures into numerical features that capture physicochemical, topological, geometric, and electronic properties of molecules, which is essential for a wide range of cheminformatics applications, including Quantitative Structural Activity Relationship (QSAR/SAR) analysis. They are widely used as machine-readable representations for statistical modeling and artificial intelligence workflows. Over the past few decades, several cheminformatics toolkits have been developed to calculate molecular descriptors and fingerprints, including RDKit, PaDEL-Descriptor, Mordred, and the Chemistry Development Kit (CDK). While these tools provide extensive descriptor libraries and computational capabilities, and are implemented in different programming environments such as Python and Java. As a result, integrating them into a single workflow often requires complex multi-language pipelines and additional scripting. Furthermore, most of these tools rely primarily on command-line interfaces and lack unified graphical environments, posing challenges for researchers without extensive programming expertise.

Results: A freely available R package, MDSLabChemBridge, was developed with an integrated platform that bridges multiple descriptor engines within the R environment. The package enables seamless descriptor calculation from RDKit, Mordred, PaDEL, and CDK. In addition to existing descriptor engines, MDSLabChemBridge introduces a custom descriptor calculator, MDSLab Custom, that computes additional structural and functional group-based descriptors to complement traditional descriptor libraries. The package provides a unified interface and an interactive Shiny-based graphical user interface to simplify descriptor generation and cheminformatics analysis.

Conclusion: MDSLabChemBridge provides a Shiny-based user interface with multiple data output options. The tool is designed to generate machine learning-ready descriptor matrices, enabling seamless integration with statistical modeling and AI workflows. The package is

available at https://github.com/yogesh601/MDSLabChemBridge and can be installed directly in R.

Keywords: Cheminformatics, Molecular descriptors, Chemical fingerprints, Machine learning, Drug discovery, RDKit, CDK, Mordred

Background

In modern drug discovery, the transformation of a physical chemical entity into a digital data point is achieved by calculating molecular descriptors and fingerprints, which constitute the foundational language of cheminformatics. Molecular descriptors provide quantitative information about a molecule's properties, ranging from simple one-dimensional constitutional data such as molecular weight to complex three-dimensional geometric and electronic characteristics that describe spatial arrangement and surface potential [1, 2]. Molecular fingerprints complement these descriptors by encoding structural information as bit strings or vectors, indicating the presence or absence of specific substructures and fragments. This numerical representation enables the conversion of unstructured chemical information into a machine-readable format suitable for high-dimensional statistical analysis and artificial intelligence applications [3–6]. Both molecular descriptors and fingerprints play a crucial role in QSAR/SAR analysis, drug Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADME/T) prediction, and toxicity prediction modeling [7, 8]. Molecular descriptors and fingerprints also play an important role in current scientific research, including the identification of biomolecular targets and the analysis of protein-ligand interaction networks. Several other studies have shown that descriptors and fingerprints play a role in predicting drug-target interactions and understanding drug action mechanisms [9–12]. Over the past two decades, several cheminformatics toolkits have been developed to calculate molecular descriptors and fingerprints. Among the most widely used are RDKit, PaDEL-Descriptor, Mordred, and Chemistry Development Kit (CDK) [13–16], BlueDesc, Mold2, ChemAxon Jchem, ADMEWORKS Model-Builder, Chemopy, AvlaDesc, etc [17–21]. Several other commercial software licenses were also available, such as SYBYL- iX, Molecular Operating Environment (MOE), Dragon, BIOVIA Discovery Studio, and Schrödinger (Maestro/MacroModel) [22–25]. Each of these tools provides a large number of descriptors and has been widely adopted in cheminformatics and computational chemistry research. However, despite their popularity and computational capabilities, these tools present several

practical challenges that limit their usability in integrated cheminformatics workflows [26]. RDKit is one of the most widely used open-source cheminformatics libraries and provides an extensive set of descriptors and molecular fingerprints [27]. It offers robust cheminformatics functionality, including molecular parsing, substructure searching, fingerprint generation, and descriptor calculation. However, RDKit is primarily implemented in Python and C++, which may pose challenges for researchers who primarily work within the R ecosystem for statistical analysis and machine learning [13]. While it is a comprehensive tool, it does not cover all available descriptors present in other specialized engines. Users often need to combine RDKit with other tools to obtain more comprehensive descriptors. PaDEL-Descriptor is another widely used software package that can calculate more than a thousand molecular descriptors and fingerprints [14]. It is implemented in Java and typically distributed as a standalone software or command-line tool. Although PaDEL provides a large descriptor library and supports multiple fingerprint types, its command-line interface and dependency on Java environments can make it difficult to integrate seamlessly into automated data analysis workflows. Furthermore, PaDEL lacks native support for interactive visualization or integration with modern machine learning pipelines without additional scripting and data preprocessing steps.

Mordred is a descriptor calculation tool that generates a large number of molecular descriptors, including both two- and three-dimensional descriptors. It can compute more than 1800 descriptors and is commonly used in machine learning studies requiring extensive molecular feature sets. Despite its powerful descriptor generation capabilities, Mordred is implemented in Python and requires additional setup and configuration of Python environments and dependencies [16]. For researchers primarily working in R-based statistical environments, integrating Mordred outputs into analysis pipelines can involve additional data processing and environment management. The Chemistry Development Kit (CDK) is an open-source Java library that provides a wide range of cheminformatics functionalities, including molecular descriptor calculations, atom typing, aromaticity detection, and chemical structure manipulation. In contrast, CDK provides valuable descriptor sets and structural analysis tools, which is implemented in Java. It is often accessed through interfaces such as rcdk in R [15]. This indirect access can introduce complexity in installation, configuration, and compatibility with other descriptor engines. Although each of these tools provides valuable capabilities for descriptor generation, they operate largely as independent systems implemented in different programming languages, including Python, Java, and C++. As a result, researchers often need

to construct complex multilingual pipelines to integrate these tools into a single workflow. Managing multiple programming environments, dependencies, and software configurations can become a significant barrier, particularly for researchers without extensive programming experience.

Another practical limitation is the lack of a unified graphical interface that allows users to easily compute descriptors from multiple engines within a single environment. Most existing descriptor tools rely on command-line interfaces or scripting environments, which may be difficult for researchers who prefer interactive workflows or lack programming expertise [3, 4, 6]. While some graphical interfaces exist for individual tools, there is currently no widely available platform that integrates multiple descriptor engines into a single interactive framework. Furthermore, many cheminformatics workflows require the generation of large descriptor matrices suitable for machine learning and artificial intelligence applications. Preparing such datasets often requires multiple preprocessing steps, including molecular format conversion, descriptor calculation using different engines, data merging, and feature cleaning. These tasks can be time-consuming and error-prone when performed manually across multiple software platforms.

To overcome these limitations, we developed MDSLabChemBridge, an integrated R package that connects multiple descriptor engines within a unified computational framework. The platform combines RDKit, Mordred, PaDEL, and CDK in the R environment using the reticulate interface and the rcdk package. In addition to existing descriptor engines, MDSLabChemBridge introduces a custom descriptor calculator, MDSLab Custom, that calculates additional structural and functional group-based descriptors designed to complement traditional descriptor libraries. The package also offers a Shiny-based graphical user interface that allows users to compute descriptors, explore chemical data interactively, and generate machine-learning-ready feature matrices without complex multi-language workflows. We believe that MDSLabChemBridge will be useful to researchers across different fields of chemistry, computational chemistry, bioinformatics, and computational biology, who can retrieve ML-ready data from the tool and directly use it for AI/ML model development.

Figure 1: Graphical Abstract of the tool, showing the workflow.

Implementation

Software architecture

The bioinformatics and computational biology communities widely rely on both R and Python for data analysis, modeling, and tool development. However, many cheminformatics tools are distributed across multiple programming languages, making it challenging to integrate them into a single, streamlined workflow. During the development of MDSLabChemBridge, multiple programming environments were required to support different descriptor calculation engines and functionalities.

To address this complexity, MDSLabChemBridge was designed as a unified platform that seamlessly bridges multiple programming languages, including R, Python, Java, and C++, within a single R package. This integration enables users to perform complex descriptor calculations and cheminformatics analyses through a single R-based interface, eliminating the need to manually manage separate environments.

The package follows a hybrid architecture in which R serves as the central framework for data processing, workflow management, and visualization. Integration with Python is achieved using the reticulate package, which provides a robust interface for executing Python code within the R environment. This allows the incorporation of widely used Python-based libraries such as *pandas, NumPy, RDKit, PaDEL (via padelpy),* and *Mordred*. In addition, descriptor calculations from the Chemistry Development Kit (CDK) are accessed through the rcdk package in R [28–30].

Beyond existing descriptor engines, MDSLabChemBridge also includes a custom descriptor module, referred to as the MDSLab Custom calculator, which extends the descriptor space by capturing additional structural and functional properties of molecules.

The graphical user interface is developed using Shiny, with enhanced user interface components from the libraries *ShinyWidgets, bslib,* and *shinycssloaders* to improve usability and responsiveness [31–34]. Chemical structure parsing and descriptor calculations via the Chemistry Development Kit are performed using the rcdk package, while the fingerprint package supports molecular fingerprint operations.

Data manipulation and preprocessing are handled using the *tidyverse suite* of packages, ensuring efficient data handling and transformation. Interactive data tables are implemented using *DT*, and dynamic visualization of chemical space and descriptor distributions is achieved through *plotly*. Dimensionality reduction for chemical space exploration is supported using the *uwot* package for UMAP, while machine learning workflows and model development are facilitated through the *caret* package [35, 36].

Together, these libraries enable MDSLabChemBridge to provide an integrated, scalable, and user-friendly platform for descriptor generation, visualization, and machine learning-ready data preparation. The Shiny-based graphical user interface enables users to perform descriptor calculations, explore chemical data, and visualize results interactively without requiring extensive programming expertise.

Figure 2: Represents the Homepage of the tool when launch, showing the instruction guide for user to get started.

Molecular input processing

MDSLabChemBridge supports multiple commonly used chemical file formats to ensure flexibility and compatibility with diverse cheminformatics workflows. The platform accepts molecular inputs in SDF, SMILES (.smi), and plain text (.txt) formats that contain SMILES strings, enabling users to analyze both structure-based and text-based chemical datasets[37].

Upon input, molecular structures are automatically parsed and standardized using the underlying cheminformatics libraries. For SMILES-based inputs, the system converts textual representations into internal molecular objects suitable for descriptor calculation. For structure-based formats such as SDF, atomic coordinates and bonding information are preserved, enabling both two-dimensional and three-dimensional descriptor calculations.

The input processing pipeline includes validation and preprocessing steps to ensure data quality and compatibility across different descriptor engines. These steps include format checking, removing invalid or unsupported molecules, and normalizing molecular representations. Where required, molecules are prepared for downstream analysis by hydrogenation, aromaticity detection, and structure sanitization.

This flexible input handling enables MDSLabChemBridge to seamlessly process datasets from a wide range of sources, from small compound libraries to large-scale chemical databases, thereby supporting a broad range of cheminformatics and machine learning applications.

Figure 3: Figure represents the MDSLabChemBridge technical report page.

Descriptor calculation workflow

MDSLabChemBridge implements a unified, automated workflow for large-scale molecular descriptor and fingerprint generation by integrating multiple cheminformatics engines into a single computational framework. The workflow is designed to streamline descriptor calculation while ensuring compatibility across different tools and data formats.

The process begins with the ingestion of molecular input files, followed by optional preprocessing and standardization steps, including hydrogen addition, aromaticity perception, charge assignment, and structure normalization. Once the molecules are prepared, they are passed to the selected descriptor engines for feature generation.

Descriptor calculation is performed in parallel across multiple integrated engines, including RDKit, Mordred, PaDEL, CDK, and the MDSLab Custom descriptor module. Each engine independently computes its own set of descriptors, capturing diverse molecular properties, including physicochemical attributes, topological indices, electronic features, functional group composition, and surface property distributions.

The outputs from each descriptor engine are then systematically aggregated into a single unified descriptor matrix. During this step, the platform ensures consistent formatting, alignment of molecular identifiers, and removal of redundant or incompatible features. Basic data cleaning procedures, such as handling missing values and filtering non-informative descriptors, are also applied to improve dataset quality.

The final output is a high-dimensional, machine-learning-ready feature matrix that can be used directly for downstream applications, including QSAR modeling, clustering, classification, and regression analysis. The workflow is optimized to handle both small and large molecular datasets efficiently, enabling scalable descriptor generation for diverse cheminformatics applications.

Additionally, the results can be visualized interactively within the Shiny interface or exported in standard formats for further analysis. This integrated workflow eliminates the need for manual data merging and cross-platform scripting, significantly simplifying descriptor generation and enabling reproducible cheminformatics pipelines.

Custom descriptors

In addition to standard descriptor engines, MDSLabChemBridge implements a proprietary descriptor module, MDSLab Custom, designed to generate complementary molecular features that capture structural complexity, functional group composition, and property-based surface characteristics. These descriptors are computed using a combination of rule-based algorithms, SMARTS pattern matching, and property-driven feature engineering approaches.

Bridging of multiple cheminformatics tools

MDSLabChemBridge was designed to integrate multiple cheminformatics descriptor engines implemented across different programming environments into a single unified framework. The platform achieves this by combining R, Python, Java, and C++-based tools within a cohesive architecture, enabling seamless communication between heterogeneous software components.

Cross-language integration framework: The R environment serves as the central orchestration layer for workflow execution, data handling, and visualization. Integration with Python-based descriptor engines, including RDKit, Mordred, and PaDEL (via padelpy), is achieved using the reticulate package. This interface allows the direct execution of Python code within R and facilitates the exchange of data structures, such as data frames and molecular representations, between the two environments.

Java-based descriptor calculations from the Chemistry Development Kit (CDK) are accessed through the rcdk package, which provides a high-level R interface to CDK functionalities. This enables descriptor computation and molecular manipulation without requiring direct interaction with Java code.

C++-based operations underlying libraries such as RDKit are accessed indirectly through Python bindings, ensuring efficient execution of computationally intensive tasks while maintaining usability within the R framework.

Workflow orchestration: MDSLabChemBridge coordinates descriptor calculations by invoking each engine sequentially or in parallel, depending on the workflow configuration.

Molecular inputs are first standardized and converted into compatible formats, after which they are passed to individual descriptor engines. The outputs from each engine are returned to the R environment, where they are aggregated into a unified descriptor matrix.

Data exchange between different programming environments is handled through intermediate representations, ensuring consistency in molecular identifiers and descriptor alignment. The platform also includes mechanisms for error handling and validation to ensure robustness across different toolchains.

Unified execution and user interface

Despite the complexity of integrating multiple programming languages and descriptor engines, MDSLabChemBridge abstracts these operations behind a single R-based interface. Users can execute the entire workflow using a single command or through the Shiny-based graphical user interface, without requiring direct interaction with Python or Java environments.

This bridging strategy eliminates the need for manual pipeline construction and simplifies the use of multiple descriptor engines, enabling reproducible and scalable cheminformatics workflows within a unified computational environment.

Results

Graphical User Interface

MDSLabChemBridge provides a highly interactive, guided GUI. When opened with the launch command in R, it displays a pop-up page with quick-start instructions. These instructions were helpful for new users. Once clicked, the program window appears. Users see several options in the left panel of the GUI window. We have provided a Descriptor Glossary for MDSLab custom descriptors. Next is the source option, which allows users to upload a compound file in different file formats, such as SDF, SMILES, smi, and txt, with the default SDF file format. The next step will be a flexible molecular preparation module that enables users to standardize input structures before descriptor calculations. Proper molecular preparation is essential to ensure consistency and accuracy across different descriptor engines. The platform included the following preprocessing options: addition of explicit hydrogen to enable accurate representation of molecular valence and to improve descriptor calculations that depend on

hydrogen atoms. Aromaticity perception standardizes the identification of aromatic systems across molecules and the assignment of Gasteiger charges, allowing the estimation of partial atomic charges required for electronic and physicochemical descriptors. Kekulization of structure, converting aromatic representations into alternating single and double bond forms for compatibility with specific descriptor calculations [38]. The second option is the selection of engines, which has been bridged in this package. Among the available options, the user can select one or more engine types. They can click all the engines by selecting All. Once the engines are selected, the user will choose the descriptor type option among 1D, 2D, and 3D. Another option for selecting fingerprints was provided in the package, allowing one to enable or disable selection, or select the types of fingerprints they want. After all the selections, the user needs to click the Start calculation button and wait for the calculations to complete. The package will show a progress bar in the bottom-right corner, indicating which engine type is running at the moment.

MDSLab Descriptor design and implementation

The MDSLab Custom descriptors are computed directly within the R framework, with supporting functions leveraging cheminformatics operations from integrated libraries such as RDKit and CDK. Molecular structures are first converted into standardized internal representations, after which descriptor calculations are performed using predefined rules and mathematical formulations. Two descriptors were developed to quantify molecular complexity and potential biochemical behavior. The MDS_Lability descriptor is a rule-based metric derived from the presence of specific structural motifs, particularly phenyl and methoxy groups. These features are associated with molecular flexibility and potential metabolic susceptibility. The descriptor is computed by identifying occurrences of these motifs using substructure matching and assigning a weighted score.

The MDS_Bio_Complexity descriptor is defined as:

$$MDS_Bio_Complexity = TPSA \times Atom\ Count/100$$

where TPSA represents the topological polar surface area and Atom Count corresponds to the total number of atoms in the molecule. This descriptor captures the density of polar functionality relative to molecular size.

Fragment-based descriptors: The MDS_Frag descriptor family is calculated using SMARTS-based pattern matching to quantify the occurrence of predefined functional groups. A curated set of twelve functional group patterns, including hydroxyl, phenyl, amine, carbonyl, carboxyl, ester, nitro, and halogen groups, is used. Each fragment is encoded as a count-based descriptor that represents its frequency within the molecular structure.

Property-encoded surface area descriptors: The MDS_PVSA family consists of 50 features that represent property-encoded surface area distributions. These descriptors partition molecular surface areas into discrete bins based on physicochemical properties and encode them as vectorized features. The approach captures spatial and property-based heterogeneity within molecular structures and provides high-dimensional representations suitable for machine learning applications.

High-density core descriptors: The MDS_Core descriptor set comprises 275 high-density features derived using a CDK-based computational bridge. These descriptors combine topological indices, electronic descriptors, and hybrid structural features. The integration of these descriptors enhances the overall feature space by capturing deeper structural and physicochemical relationships within molecules.

Table 1: Overview of the MDSLab Custom descriptor categories implemented in MDSLabChemBridge. The descriptor set includes complexity metrics, fragment-based features, surface area representations, and high-density structural descriptors designed to complement standard cheminformatics descriptor libraries.

Output and integration: All MDSLab custom descriptors, along with descriptors generated by external engines, are computed in a structured, standardized format within the platform. These features are seamlessly integrated into a unified descriptor matrix, ensuring compatibility with downstream statistical analyses and machine learning workflows. To evaluate the package's performance, a dataset comprising 1,000 biologically active compounds against the Zika virus was retrieved from the ChEMBL database in SDF format and processed via the Shiny-based graphical user interface. After selecting the appropriate computational options, the platform generated 5,936 features. These included 529 descriptors from the MDSLab custom module, 454 from CDK, 228 from RDKit, 1,826 from Mordred, 1,875 from

PaDEL, and 1,069 fingerprint features. Classification of the computed features yielded 336 1D descriptors, 3,759 2D descriptors, and 772 3D descriptors, demonstrating the platform's ability to generate a comprehensive, diverse descriptor space suitable for advanced cheminformatics and machine learning applications. The package includes multiple informative files for download, such as a technical report and a detailed log file. These files provide extensive information about the data, calculations, results, and direction for the users to interpret the results. The technical report provides information on feature engineering, including the total number of descriptors calculated in our case (5936), the total number of fingerprints (1069), the number of ML-ready, optimized features (1620), the feature retention rate (27.3%), and the correlation redundancy cutoff (0.9). These ML-ready features were selected from 5936 initial candidates. Again, it also provides information on the calculation engine's integrity, a detailed glossary of descriptors for MDSLab standards, external engine references, an ML-ready, optimized feature methodology for users, expanded physicochemical profiles, drug likeness & quality audits, and guidelines for ML & QSAR data integration.

Machine learning-ready descriptor matrix generation

The package enables the automated generation of machine-learning-ready descriptor matrices by integrating outputs from multiple cheminformatics engines into a unified, standardized feature table. The ML-ready data was generated into two steps, first program remove all the descriptors which contains the NA/NaN/blank values within the calculation table, second step was the removal of correlated descriptors among each other, user can set the rabnge of its requirement to remove the highly correlated descriptors from the table, the minimum range was set to 70% and maximum was 99%. This sliding bar dynamically changes the cutoff. This functionality eliminates the need for manual data preprocessing and significantly streamlines dataset preparation for downstream modeling applications.

Following descriptor calculation, outputs from RDKit, Mordred, PaDEL, CDK, and the MDSLab Custom descriptor module are aggregated into a single matrix, with rows corresponding to individual molecules and columns representing descriptor features. The platform ensures consistent alignment of molecular identifiers across all descriptor sets, thereby preserving data integrity during integration.

To enhance the usability of machine learning workflows, the generated descriptor matrix is subjected to automated preprocessing. These include removing invalid or non-computable

descriptors, handling missing values, and filtering out constant or low-variance features. Such preprocessing ensures that the resulting dataset is suitable for direct use in statistical analysis and predictive modeling.

The integrated descriptor matrix is inherently high-dimensional, capturing a wide range of molecular properties, including physicochemical attributes, topological features, functional group composition, and surface property distributions. This comprehensive feature representation enhances the potential to capture structure–activity relationships in cheminformatics studies.

The final output can be readily exported to standard formats such as CSV and directly used in machine learning frameworks in R or Python. The generated datasets are compatible with a wide range of modeling approaches, including classification, regression, clustering, and feature selection methods.

By automating descriptor integration and preprocessing, MDSLabChemBridge provides a robust, scalable solution for generating high-quality, machine-learning–ready datasets, thereby facilitating the efficient development of predictive models in drug discovery and cheminformatics research.

Visualization and chemical space analysis (PCA and UMAP)

MDSLabChemBridge provides integrated tools for visualizing and exploring chemical space using dimensionality reduction techniques. High-dimensional descriptor matrices generated by the platform are transformed into lower-dimensional representations to enable intuitive analysis of molecular similarity, diversity, and clustering patterns.

The platform supports Principal Component Analysis (PCA), a linear dimensionality reduction method that projects high-dimensional descriptor data into a lower-dimensional space while preserving as much variance as possible. PCA enables users to identify dominant trends in the dataset, assess molecular diversity, and detect outliers. It is particularly useful for obtaining an overview of chemical space and understanding the contribution of major descriptor components.

In addition to PCA, MDSLabChemBridge incorporates Uniform Manifold Approximation and Projection (UMAP) for nonlinear dimensionality reduction [35]. UMAP captures complex relationships in high-dimensional descriptor data and often improves the separation of

molecular clusters compared to linear methods. This makes it especially suitable for identifying structural groupings, scaffold diversity, and hidden patterns within chemical libraries.

Both PCA and UMAP visualizations are implemented within the Shiny-based graphical interface and rendered using interactive plotting tools. Users can dynamically explore chemical space by zooming, filtering, and inspecting individual data points, facilitating deeper insight into molecular relationships. PCA coordinates can be used as principal components for principal component regression. This is highly beneficial for QSAR models where descriptors are naturally intercorrelated. Among the calculated 1000 compounds, 300 fell into cluster 1, 571 into cluster 2, 7 into cluster 3, and 122 into cluster 4 (**Figure PCA**). UMAP excels at preserving local structural similarities. By overlaying biological activity on UMAP plots, researchers can identify 'Activity Cliffs' structural neighbors with vastly different bioactivities, which are critical for SAR optimization. In our example, the UMAP figure shows that clusters 1, 3, and 4 largely overlap, which is justified because all input compounds were partially active against the Zika virus, so they must share similar SAR properties.

These visualization capabilities enable researchers to perform exploratory data analysis, identify clusters of structurally similar compounds, and evaluate dataset diversity, thereby supporting downstream applications such as compound selection, library design, and machine learning model development.

Performance and scalability

MDSLabChemBridge is designed to generate high-quality, machine learning–ready descriptor datasets by integrating multiple descriptor engines and systematically organizing their outputs into a unified feature space. Rather than focusing solely on computational speed, the platform emphasizes the generation of comprehensive, structured, and analysis-ready data suitable for downstream cheminformatics and artificial intelligence applications.

The system produces a high-dimensional descriptor matrix that includes a diverse range of molecular features derived from multiple sources. These include:

1D descriptors (e.g., atom counts, molecular weight, elemental composition)

2D descriptors (e.g., topological indices, connectivity features, functional group counts)

3D descriptors (e.g., geometry-dependent and surface-based features, where applicable)

Molecular fingerprints (e.g., structural and substructure-based binary representations)

Custom descriptors (MDSLab Custom) capturing complexity, fragment composition, and surface property distributions

The integration of these descriptor types results in a rich and multi-scale representation of chemical structures, enabling improved capture of structure-property and structure–activity relationships. To ensure data quality and usability, MDSLabChemBridge incorporates automated preprocessing and feature curation steps. These include removing redundant descriptors, filtering for constant or low-variance features, and handling missing or non-computable values [39]. This process reduces noise in the dataset and enhances the robustness of downstream analyses. The resulting descriptor matrix is inherently machine learning-ready, requiring minimal additional preprocessing before model development. The standardized format allows direct integration with statistical and machine learning frameworks for tasks such as classification, regression, clustering, and feature selection.

Furthermore, the availability of a unified descriptor dataset combining outputs from multiple engines reduces the need for manual data merging and feature engineering. This not only improves reproducibility but also enables researchers to focus on model development and interpretation rather than data preparation. Overall, MDSLabChemBridge provides a scalable framework for generating clean, comprehensive, and ML-ready chemical feature datasets, supporting advanced cheminformatics analyses and accelerating data-driven drug discovery workflows.

Conclusion

In this work, we present MDSLabChemBridge, a unified and extensible cheminformatics platform that simplifies and standardizes molecular descriptor and fingerprint generation across multiple computational environments. By bridging widely used descriptor engines, including RDKit, Mordred, PaDEL, and CDK, within a single R-based framework, the platform addresses a key challenge in cheminformatics: the fragmentation of tools across different programming languages and workflows.

A central contribution of MDSLabChemBridge is its ability to generate comprehensive, machine learning–ready descriptor matrices by integrating diverse molecular features, including 1D, 2D, and 3D descriptors, molecular fingerprints, and a novel set of MDSLab Custom descriptors. These custom descriptors further enrich the feature space by capturing

molecular complexity, functional group composition, and surface property distributions, thereby complementing existing descriptor libraries.

The platform emphasizes data quality and usability, incorporating automated preprocessing steps such as removing redundant and low-informative descriptors, handling missing values, and standardizing feature matrices. This ensures that the generated datasets are immediately suitable for downstream applications, including QSAR modeling, clustering, classification, and regression analysis, without requiring extensive manual preprocessing. In addition to its computational capabilities, MDSLabChemBridge provides an accessible, user-friendly interface via a Shiny-based graphical user interface, enabling both programmers and non-expert users to perform descriptor calculations, visualize chemical space using PCA and UMAP, and explore molecular datasets interactively. This significantly lowers the barrier to entry for complex cheminformatics analyses.

By integrating multiple descriptor engines into a single workflow and providing ML-ready outputs, MDSLabChemBridge enhances reproducibility, reduces workflow complexity, and facilitates efficient data-driven research in cheminformatics and drug discovery. The platform is scalable, extensible, and adaptable to a wide range of applications, from small-scale exploratory studies to large-scale virtual screening and machine learning pipelines.

Future developments of MDSLabChemBridge may include integrating deep learning–based molecular representations, automated feature selection pipelines, and cloud-enabled deployment to support high-throughput cheminformatics workflows further.

Overall, MDSLabChemBridge provides a robust and comprehensive solution for molecular descriptor generation and analysis, advancing computational methodologies in cheminformatics and enabling more efficient development of predictive models in drug discovery.

List of abbreviations

QSAR: Quantitative Structural Activity Relationship

SAR: Structural Activity Relationship

ADME/T: Absorption Distribution Metabolism Excretion/ Toxicity

ML: Machine Learning

AI: Artificial Intelligence

Declarations

Availability of data and materials

Project MDSLabChembridge is available at the GitHub link (https://github.com/yogesh601/MDSLabChemBridge). It can be retrieved directly from GitHub and installed in R/RStudio. The project required several packages and libraries to be installed, which could be easily installed in R.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

YK has designed and implemented the package, written the original draft, and reviewed the manuscript.

Acknowledgements

The author(s) gratefully acknowledge the Centre of Biomedical Research (CBMR) for providing the necessary infrastructure facilities and technical support that made this work possible.

Authors' information

Department of Data Sciences, Centre of BioMedical Research, SGPGIMS Campus. Lucknow. India-226014

Web links and URLs:

GitHub link https://github.com/yogesh601/MDSLabChemBridge

References:

1. Consonni V, Todeschini R (2010) Molecular Descriptors. Challenges Adv Comput Chem Phys 8:29–102. https://doi.org/10.1007/978-1-4020-9783-6_3

2. Dunn W (2001) Handbook of Molecular Descriptors. Methods and Principles in Medicinal Chemistry Series. Volume 11 By Roberto Todeschini and Viviana Consonni (Universita degli. https://doi.org/10.1021/ja015204

3. Zhang K, Yang X, Wang Y, et al (2025) Artificial intelligence in drug development. Nat Med 2025 311 31:45–59. https://doi.org/10.1038/s41591-024-03434-4

4. Zhou Y, Wang F, Tang J, et al (2020) Artificial intelligence in COVID-19 drug repurposing. Lancet Digit Heal 2:e667–e676. https://doi.org/10.1016/S2589-7500(20)30192-8

5. Burki T (2020) A new paradigm for drug development. Lancet Digit Heal 2:e226–e227. https://doi.org/10.1016/S2589-7500(20)30088-1

6. De Rycker M, Baragaña B, Duce SL, Gilbert IH (2018) Challenges and recent progress in drug discovery for tropical diseases. Nat 2018 5597715 559:498–506. https://doi.org/10.1038/s41586-018-0327-4

7. Peter SC, Dhanjal JK, Malik V, et al (2018) Quantitative structure-activity relationship (QSAR): Modeling approaches to biological applications. Encycl Bioinforma Comput Biol ABC Bioinforma 1–3:661–676. https://doi.org/10.1016/B978-0-12-809633-8.20197-0

8. Paul Gleeson M, Hersey A, Hannongbua S (2011) In-Silico ADME Models: A General Assessment of their Utility in Drug Discovery Applications. Curr Top Med Chem 11:358–381. https://doi.org/10.2174/156802611794480927

9. Cereto-Massagué A, Ojeda MJ, Valls C, et al (2015) Molecular fingerprint similarity search in virtual screening. Methods 71:58–63. https://doi.org/10.1016/j.ymeth.2014.08.005

10. Polgar T, M. Keseru G (2011) Integration of Virtual and High Throughput Screening in Lead Discovery Settings. Comb Chem High Throughput Screen 14:889–897. https://doi.org/10.2174/138620711797537148

11. Schwaller P, Vaucher AC, Laplaza R, et al (2022) Machine intelligence for chemical reaction space. Wiley Interdiscip Rev Comput Mol Sci 12:e1604. https://doi.org/10.1002/wcms.1604

12. Clark RD, Roe DC (2010) Ligand- and structure-based virtual screening. Handb

Chemoinformatics Algorithms 145–171. https://doi.org/10.1201/9781420082999

13. RDKit: Open-source cheminformatics (2023) The RDKit Documentation. In: https://www.rdkit.org. https://www.rdkit.org/docs/index.html. Accessed 24 Mar 2026

14. Yap CW (2011) PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474. https://doi.org/10.1002/jcc.21707

15. Willighagen EL, Mayfield JW, Alvarsson J, et al (2017) The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. J Cheminformatics 2017 91 9:33-. https://doi.org/10.1186/s13321-017-0220-4

16. Moriwaki H, Tian YS, Kawashita N, Takagi T (2018) Mordred: a molecular descriptor calculator. J Cheminformatics 2018 101 10:4-. https://doi.org/10.1186/s13321-018-0258-y

17. GitHub - OlivierBeq/BlueDesc: BlueDesc Molecular Descriptor Calculator · GitHub. https://github.com/OlivierBeq/BlueDesc. Accessed 24 Mar 2026

18. Hong H, Xie Q, Ge W, et al (2008) Mold2, Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. J Chem Inf Model 48:1337–1344. https://doi.org/10.1021/ci800038f

19. https://independent.academia.edu/MauriAndrea (2020) alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. Methods Pharmacol Toxicol 801–820. https://doi.org/10.1007/978-1-0716-0150-1_32

20. ChemAxon (2011) Instant JChem 5.11.3

21. Cao DS, Xu QS, Hu QN, Liang YZ (2013) ChemoPy: freely available python package for computational biology and chemoinformatics. Bioinformatics 29:1092–1094. https://doi.org/10.1093/bioinformatics/btt105

22. (2021) Free Download: BIOVIA Discovery Studio Visualizer - Dassault Systèmes. https://discover.3ds.com/discovery-studio-visualizer-download. Accessed 24 Mar 2026

23. SYBYL-X 2.1 | Scientific Computing World. https://www.scientific-computing.com/press-releases/sybyl-x-21. Accessed 24 Mar 2026

24. MOPAC. https://openmopac.net/. Accessed 24 Mar 2026

25. Molecular Operating Environment (2013) Molecular Operating Environment (MOE) | MOEsaic | PSILO. In: 08; Chem. Comput. Gr. Inc., 1010 Sherbooke St. West, Suite #910,.Montreal, QC, Canada, H3A 2R7, 2019. https://www.chemcomp.com/en/Products.htm. Accessed 24 Mar 2026

26. Dong J, Cao DS, Miao HY, et al (2015) ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. J Cheminformatics 2015 71 7:60-. https://doi.org/10.1186/s13321-015-0109-z

27. Desaphy J, Raimbaud E, Ducrot P, Rognan D (2013) Encoding Protein–Ligand Interaction Patterns in Fingerprints and Graphs. J Chem Inf Model 53:623–637. https://doi.org/10.1021/ci300566n

28. pandas - Python Data Analysis Library. https://pandas.pydata.org/. Accessed 25 Mar 2026

29. NumPy. https://numpy.org/. Accessed 25 Mar 2026

30. GitHub - Hamada-Noreldeen/PaDELPy: A Python wrapper for PaDEL-Descriptor software · GitHub. https://github.com/Hamada-Noreldeen/PaDELPy. Accessed 25 Mar 2026

31. GitHub - dreamRs/shinyWidgets: shinyWidgets : Extend widgets available in shiny. https://github.com/dreamRs/shinyWidgets. Accessed 24 Mar 2026

32. Tedyverse (2023) Tidyverse packages

33. Sponsor @lucidrains on GitHub Sponsors · GitHub. https://github.com/sponsors/daattali. Accessed 24 Mar 2026

34. {GitHub} - rstudio/reticulate: R Interface to Python

35. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction — umap 0.5.8 documentation. https://umap-learn.readthedocs.io/en/latest/. Accessed 25 Mar 2026

36. Plotly (2024) Plotly Python Graphing Library. In: Plotly Open Source Graph. Libr. Python. https://plotly.com/python/. Accessed 25 Mar 2026

37. Weininger D (2002) SMILES, a chemical language and information system. 1.

Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31–36. https://doi.org/10.1021/ci00057a005

38. Roy K, Mitra I (2012) Electrotopological State Atom (E-State) Index in Drug Design, QSAR, Property Prediction and Toxicity Assessment. Curr Comput Aided-Drug Des 8:135–158. https://doi.org/10.2174/157340912800492366

39. Ibrahim H Molecular Descriptors and its Types and Importance. 7:106

Figures, tables and additional files

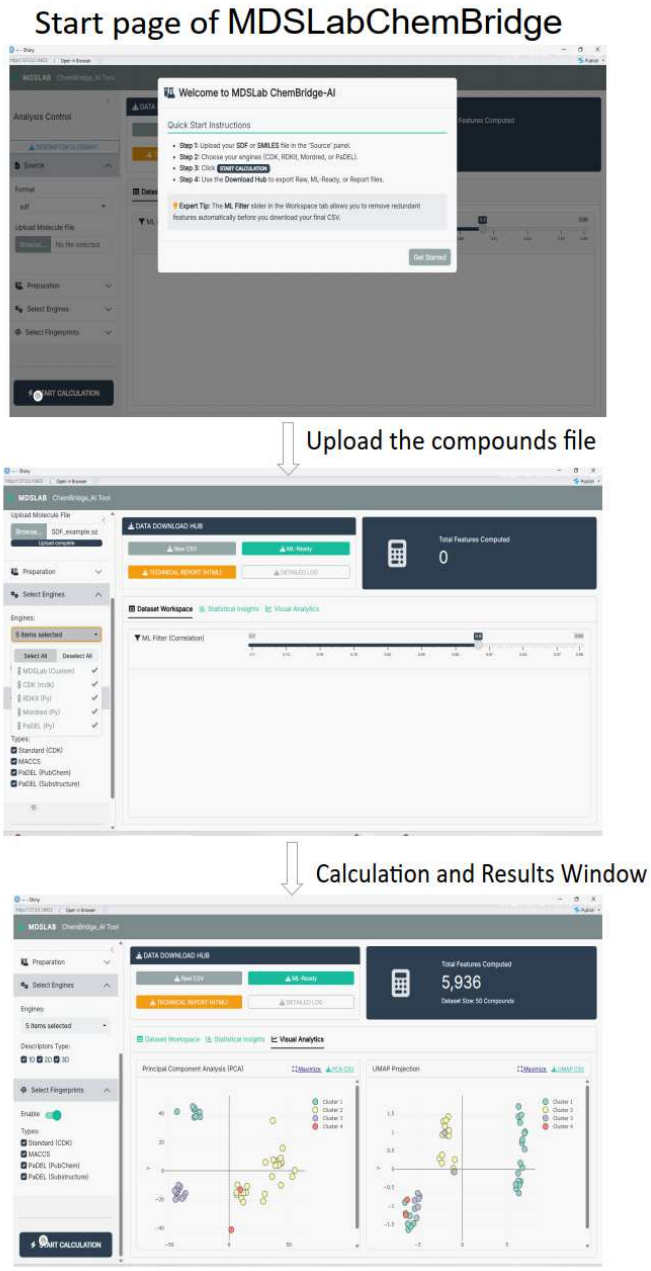


Figure1: Graphical Abstract of the tool, showing the workflow.

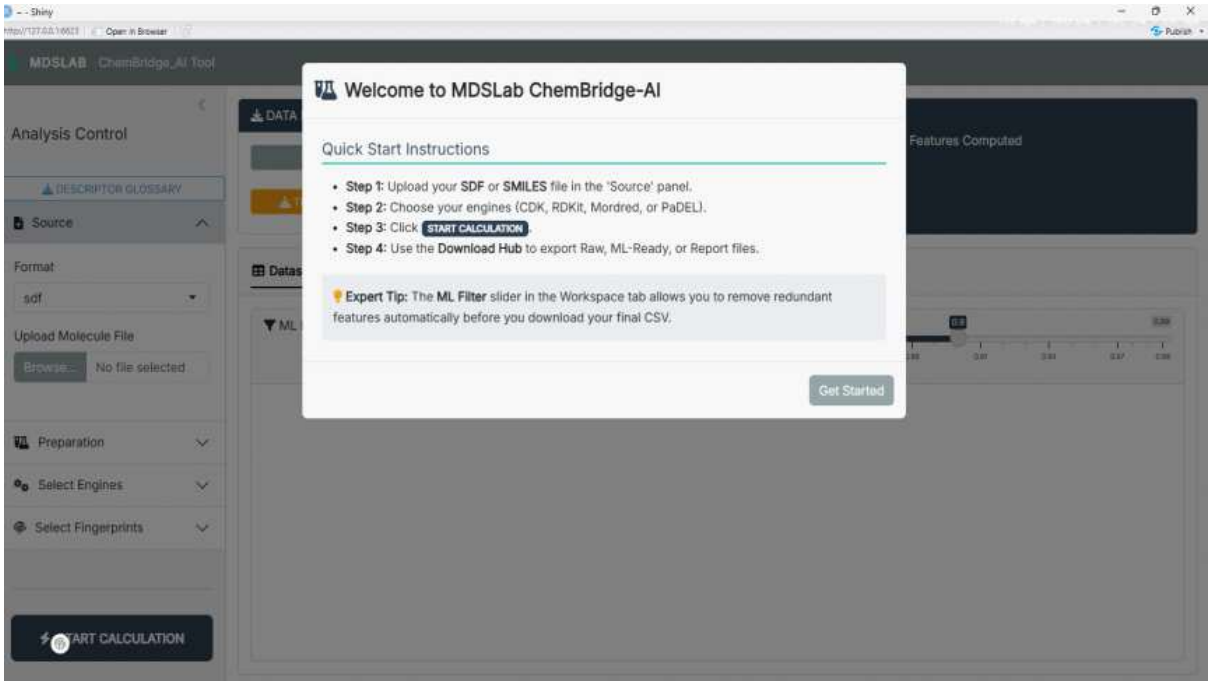


Figure2: Represents the Homepage of the tool when launch, showing the instruction guide for user to get started.

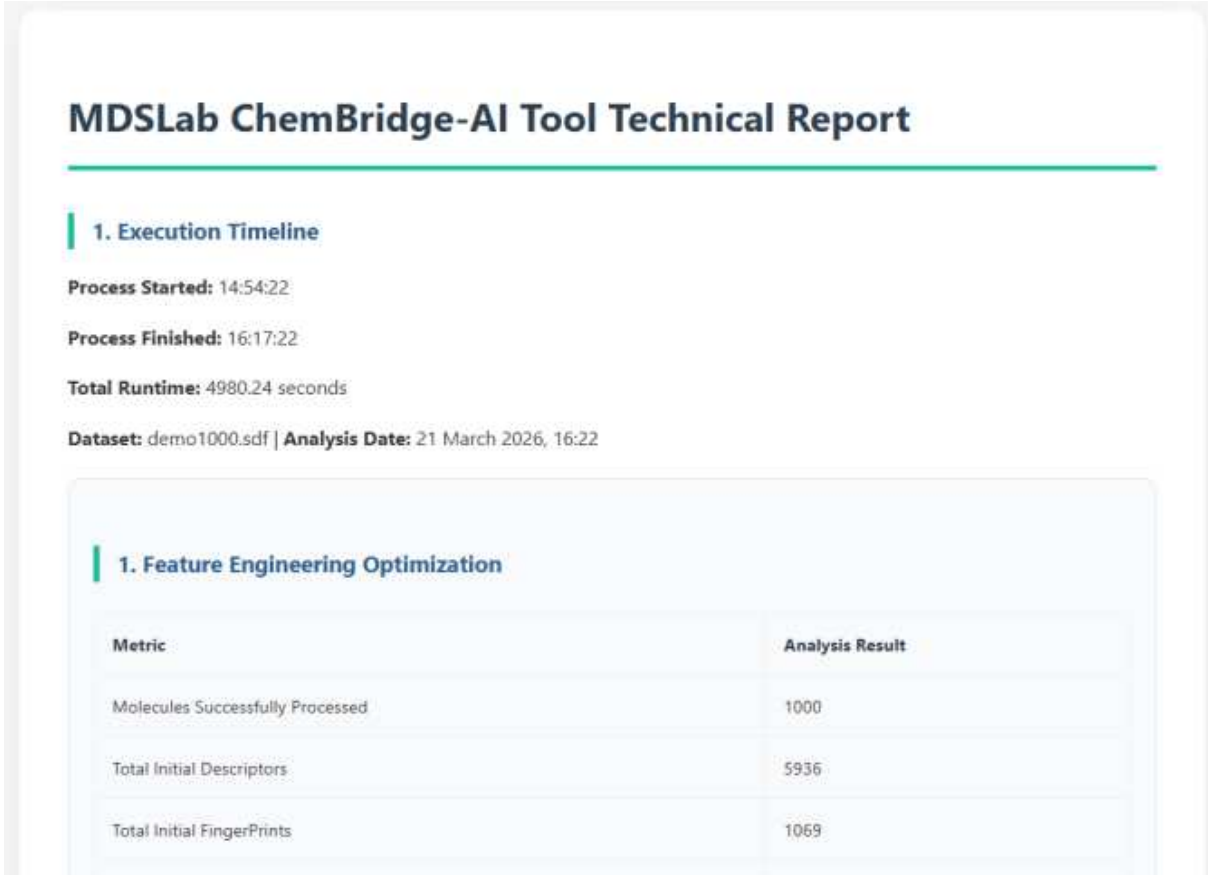


Figure3: Figure represents the MDSLabChemBridge technical report page.

Table 1: Overview of the MDSLab Custom descriptor categories implemented in MDSLabChemBridge. The descriptor set includes complexity metrics, fragment-based features, surface area representations, and high-density structural descriptors designed to complement standard cheminformatics descriptor libraries.

Feature ID	Category	Description
MDS_Lability	Complexity	Descriptor capturing molecular flexibility based on phenyl and methoxy motifs, associated with potential metabolic lability.
MDS_Bio_Complexity	Complexity	Defined as (TPSA × Atom Count) / 100; represents the density of polar functional groups relative to molecular size.
MDS_Frag_[Name]	Fragment-based	Count-based descriptors derived from SMARTS patterns representing 12 predefined functional groups (e.g., hydroxyl, phenyl, amine, carbonyl, carboxyl, ester, nitro, halogen).
MDS_PVSA_[1–50]	Surface area	Vectorized property-encoded surface area descriptors representing molecular surface partitioned into multiple physicochemical bins.
MDS_Core_[1–275]	High-density (CDK)	A set of 275 descriptors derived from CDK, including topological, electronic, and hybrid structural indices.