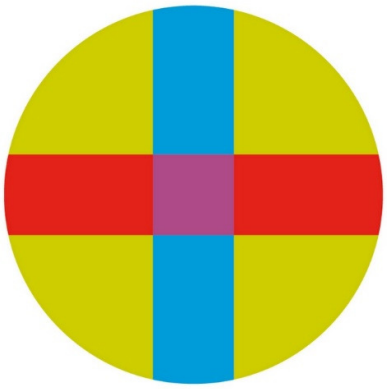
UNIVERSITY CEU - SAN PABLO

POLYTECHNIC SCHOOL

BIOMEDICAL ENGINEERING DEGREE



BACHELOR THESIS

Design and Development of a Compound-Centred Computational Pipeline for Mechanisms-of-Action Exploration in Phenotypic Drug Discovery

Author: Paula Blanco González
Supervisors: Carlos Óscar Sorzano Sánchez and Javier Tejedor Noguerales

June 2026

UNIVERSIDAD SAN PABLO-CEU
ESCUELA POLITÉCNICA SUPERIOR
División de Ingeniería

Datos del alumno

NOMBRE:

Datos del Trabajo

TÍTULO DEL PROYECTO:

Tribunal calificador

PRESIDENTE:	FDO.:

SECRETARIO:	FDO.:

VOCAL:	FDO.:

Reunido este tribunal el _____/_____/_____, acuerda otorgar al Trabajo Fin de Grado presentado por Don_____la calificación de _____.

ACKNOWLEDGMENTS

ABSTRACT

Phenotypic drug discovery identifies compounds through their observable biological effects, but the molecular mechanisms responsible for those effects are often unknown or partially understood. The current growing availability of chemical and functional information in public biomedical databases creates new opportunities for computational workflow capable of connecting phenotypically active compounds with potential targets, pathways and biological functions.

This project presents the design and implementation of a compound-centred computational workflow that starts from user-provided SMILES codes and automates the retrieval, integration and interpretation of compound, target, pathway and Gene ontology data in a single web application. The application accesses external web services programmatically to identify compounds through PubChem, retrieves compound-protein interactions and pathway-associated proteins, standardize target identifiers using NCBI Gene and UniProt, and obtain Gene Ontology annotations.

A validation case study using statins is performed to evaluate whether the retrieved results are consistent with known pharmacological mechanisms.

Overall, the application provides an interactive exploratory environment in which users can inspect recurrent targets, pathways and ontology terms across groups of compounds sharing a known phenotype, as well as export a final consolidated report.

RESUMEN

El descubrimiento fenotípico de fármacos identifica compuestos a partir de efectos biológicos empíricamente, pero los mecanismos responsables de dichos efectos suelen ser desconocidos o parcialmente conocidos. La creciente disponibilidad de información química, molecular y funcional en bases de datos biomédicas públicas genera nuevas oportunidades para el desarrollo de flujos de trabajo computacionalmente, capaces de conectar compuestos fenotípicamente activos con posibles dianas, rutas y funciones biológicas.

Este Proyecto presenta el diseño y la implementación de un flujo de trabajo computacional centrado en el compuesto que, a partir de códigos SMILES proporcionados por el usuario, automatiza la recuperación, integración e interpretación de datos sobre el compuesto, dianas, rutas metabólicas y anotaciones de Gene Ontology en una sola aplicación web. La aplicación accede computacionalmente a servicios web externos para identificar compuestos como PubChem, recupera interacciones compuesto-proteína y proteínas asociadas a rutas, estandariza identificadores de dianas con NCBI Gene y UniProt, y obtiene anotaciones de Gene Ontology.

Un caso de estudio ha sido realizado usando estatinas para evaluar si los resultados obtenidos son coherentes con la información farmacológica ya disponible y sabida.

En conclusión, la aplicación proporciona un entorno interactivo en el que los usuarios pueden inspeccionar dianas, rutas y términos ontológicos recurrentes en compuestos que comparten un fenotipo conocido y, exportar un informe final consolidado.

INDEX

FIGURE INDEX

TABLE INDEX

1 INTRODUCTION

1.1 Bioinformatics and Biomedical Data Growth

Within the last decade, the field of biomedical science has dramatically changed by the rapid technological advancements and the growth of large-scale digital data, making bioinformatics an essential pillar for managing and analysing large amounts of complex biological information [1]. With an increase in available high-throughput technologies and the proliferation of open-access databases and repository systems, large amounts of biomedical big data have accumulated, including genomics data, proteomics data, and metabolomics data among others. [1], [2]. These developments have played a pivotal role in revolutionizing biomedical research, offering unprecedented insights into molecular mechanisms and disease processes, and enhancing traditional drug discovery methodologies [1]. As a result, the use of bioinformatics, understood as the application of computational methods to the analysis and interpretation of biological data, has become indispensable in modern biomedical research community, serving as a backbone for the management, integration and computational exploration of massive volumes of data [1]. Nonetheless, the effective exploitation of these resources remains challenging, since biomedical data are not only large in scale, but also complex, unstructured, and heterogeneous [1], [2]. The data are also often dispersed among multiple databases, each with its own format and access mechanisms [3].

Thus, the importance of bioinformatics extends beyond being useful for data analysis. It is fundamental for the integration of heterogeneous resources into coherent, interpretable and unified biological workflows that can support innovative biomedical research, including drug discovery [1], [3]. By combining advanced computational tools with large volumes of data, enables researchers to uncover patterns or associations among the data that were previously difficult to detect.

1.2 Computational Drug Discovery

As previously discussed, bioinformatic applications serve as tools for researchers to enhance biomedical research. Drug discovery, for instance, has increasingly become an important application area of bioinformatics, since the

identification of potential therapeutic drugs depends on the integration and characterization of chemical, biological and computational information. In general terms, drug discovery is a sophisticated and lengthy process that consists of the identification of molecules that interact with chemical targets in the body to treat diseases [4]. Traditionally, this procedure has been put into practice through "trial and error", which can be tedious and time-consuming [5]. Novel computational approaches, such as homology modelling or protein structure simulation, together with the accumulation and expansion of high-throughput data and public resources, have paved the way for more systematic and accelerated processes. Drug Discovery has notably benefited from these technological advances, giving rise to what is broadly understood as computational or computer-aided drug discovery. Essentially, these strategies facilitate the identification, characterization and prioritization of candidate compounds by combining computational analysis and with chemical and biological knowledge.

Within this broad computational framework, the conventional thinking in modern drug discovery has been strongly shaped by two major and complementary perspectives: target-based drug discovery (TDD), which begins with a predefined molecular target; and phenotypic drug discovery (PDD), which starts from an observed phenotypic or biological effect [6], [7].

Even though both approaches have the same objective (i.e., discovery potential therapeutic compounds for disease treatment), they differ substantially in their starting point logic. Target-based discovery methods have dominated modern drug discovery throughout the past decade and continue to be one of the most used strategies. On the other hand, phenotypic approaches have slowly re-emerged as an important complementary support, especially in complex or rare diseases where biological mechanistic effects cannot be reduced to a single predefined target.

1.2.1 Target-Based Drug Discovery

Target-based drug discovery has been predominantly used in the modern drug discovery field, particularly following the expansion of structural biology, genomics, and molecular biology, which has shifted scientific attention towards the identification of

specific disease-associated molecular targets. Target-based screening relies on a starting molecular target, such as enzymes, ion channels, or other biomolecule considered relevant to the pathology under study. Once this target is identified, compounds are searched for, designed or optimized to modulate its activity and ultimately produce a therapeutic effect [7].

As previously mentioned, thanks to the growth of biomedical data, the on-going development of widely used computational approaches such as structure-based drug design methods, including molecular docking, ligand-based drug design tools (QSAR), and sequence-based approaches, in which sequence information is analysed and compared, continue [8]. Together, these methodologies have contributed to an improvement in identifying candidate compounds in a more systematic and precise manner compared to purely empirical predictions.

The purpose of TDD is therefore the identification of compounds capable of acting on a selected target in a both effective and efficient manner. This strategy has successfully been recognized as the most established frameworks in contemporary drug discovery and has led to major advancement in pharmaceutical research. However, its underlying assumption of being able to generate therapeutic benefit by modulating one single well-defined target has been questioned throughout recent years [6]. The conventional "one disease, one target, one drug" schema, may oversimplify disease biology, which is often driven by interconnected signalling pathways rather than isolated molecular events [6]. Furthermore, approximately 10% of drug candidates are successfully approved during clinical evaluation, which emphasizes the current limitation of TDD, especially in diseases of high heterogeneity and unmet therapeutic need [6].

Consequently, target-based approaches are often most effective when used in combination with complementary computational strategies that provide clear and meaningful biological context, especially when investigating complex diseases [8]. From this perspective, computational approaches capable of retrieving pathway-level effects and chemical-target interactions, have gained relevance as forms of early-stage analysis that support and enrich the biological interpretation of compound activity [6], [9]

1.2.2 Phenotypic Drug Discovery

Since 2011, PDD has experienced major relevance as a potential complementary approach to target-based discovery following the observation that best-seller drugs were discovered empirically without a previous drug target hypothesis [9]. Whereas TDD process begins with a predefined molecular target, PDD consists of the collection of lead compounds based on an observable phenotypic effect in a biological system, without prior knowledge of the molecular target responsible for that effect [6]. This strategy focuses heavily on the modulation of a disease phenotype rather than on the direct interaction of a pre-selected target. This approach, together with the now strengthened contemporary tools and computational methods, allows researchers to have a novel perspective of earlier empirical drug discovery as well as a valuable bridge between therapeutic biology and unknown mechanisms, signalling pathways and potential drug targets [9].

The renaissance of phenotypic discovery approaches is explained by the successful discoveries in a majority in first in class drugs, by observing empirically its phenotypic effect on models. [9]. PDD follows a biology-first logic, therefore it is remarkably attractive in diseases where the complex biological mechanisms are not yet fully understood and thus, cannot be simply reduced to a single molecular target. Due to this claim, phenotypic strategies provide deeper understanding of pathway-level responses and uncover underlying mechanisms of action (MoAs). Moreover, many effective therapies do not rely exclusively on one single drug, but rather on combinations of drugs that interact with multiple targets or pathways [6], [9]. This point is especially relevant for complex diseases and therapies that struggle with secondary effects. On this basis, opting for phenotypic strategies, therapeutic benefit may rise from the coordinated modulation of diverse compounds. Additionally, it might potentially reveal biologically meaningful insight on compound effects in realistic systems, including those associated with multi-target drugs or drug combinations.

Nonetheless, the identification of underlying molecular interactions and mechanisms is only the first step to determine which targets may induce that response. This issue is often described as target deconvolution or mechanistic interpretation. [6], [9], [10]. Identifying the target of a compound does not automatically explain its full

biological behaviour, and that meaningful mechanistic understanding requires the integration of additional evidence from broader biological profiling strategies. This is where the focus is shifted towards bioinformatics and data integration [6]. When compounds are already known to induce a significant phenotypic response, the challenge is not only to observe their activity, but to interpret the biological evidence associated. In order to address this limitation, biomedical data analysis can provide an important complementary strategy for organizing, integrating and retrieving relevant information and thereby, aiding the mechanistic interpretation of phenotypically relevant compounds. This perspective highly motivates the development of the workflow proposed in the present project.

1.3 Problem Statement

Despite the increasing technological advancements brought by computational drug discovery strategies, the biological interpretation of compound activity remains a major challenge, especially in contexts where compounds are already known to induce a significant phenotypic effect in biological systems, but their underlying mechanisms are not fully yet understood. This limitation is especially present within PDD, where the focus is a quantifiable biological effect rather than a predefined molecular target. As discussed in section 1.2.2, the interpretation of phenotypic responses often requires the biological understanding of MoAs of a drug across multiple levels of biological complexity. Viewed in this way, drugs can be explored from their known interactions, pathway associations, and MoAs, using phenotypic, transcriptomic and proteomic technologies capable of integrating with evolving chemical and biological databases and new computational methods [6]. Therefore, there is a need for integrated computational workflows capable of connecting complex disease association with drug mechanisms of action and transforming dispersed public biomedical data into unified intelligible analyses.

Previous bioinformatics research has demonstrated the relevance of using data-mining techniques to large-scale public biomedical databases. For instance, Pan et al. described a pathway-analysis strategy to identify and retrieve new potential applications of already known drugs from databases (i.e., PubChem, BioSystems and BioAssay), followed by a pathway enrichment to characterize biologically relevant mechanisms and

potential clinical uses. Additionally, the authors noted that earlier studies had mainly focused their investigation on compound-target associations or broader drug-gene-pathway relationships but using very narrow datasets and therefore, did not get relevant outcome [11]. Thus, their work both highlighted the challenge and the importance of integrating large-scale biomedical data: while useful information is available, it is dispersed throughout heterogeneous resources and requires computational organization before it can provide significant biological insight or generate biologically meaningful hypotheses regarding the mechanisms underlying that response.

For this reason, compound-centred workflows demonstrate to be a valuable complementary strategy for phenotypic approaches, as they help organize, integrate and transform large volumes of distributed biomedical information into a unified and interpretable biological framework.

1.4 Objectives

This project aims to design and develop an automated compound-centred computational workflow that starts from a user-provided SMILES codes and supports the biological and mechanistic interpretation of compounds with known phenotypic effects.

Essentially, the purpose of the application is to facilitate the early-stage biological interpretation of chemical compounds by creating a comprehensive software tool capable of integrating, transforming and handling large volumes of biomedical data from open-access resources. Rather than relying on target-based prediction alone, this approach addresses the key limitation previously mentioned in the Problem Statement section: underlying relevant information about compound-associated targets, pathways, and biological functions is often dispersed across heterogeneous databases and is therefore difficult to integrate into a coherent analytical framework. The presented workflow adopts a complementary compound-centred strategy in which the SMILES input serves as the starting point for the exploration of compound-target and pathway interactions, pathway-associated proteins, and functional annotations, with the aim of proposing biological insight potentially associated to the observed phenotypic response.

To achieve this general goal, the project pursues the following specific objectives that can be expressed as functional and non-functional requirements, as described in Table 1.

Type	*Requirements*	*Description*
F	SMILES input handling	The application must accept one or more SMILES code as the starting point of the analysis from the user.
F	SMILES validation and canonicalization	The application must validate, reject invalid entries and standardize valid compounds into canonical representations.
F	Compound identification and metadata retrieval	The application must identify the corresponding chemical compounds in PubChem from their SMILES code and retrieve relevant compound metadata (CID, name, molecular formula and molecular weight).
F	Compound-target interaction information retrieval	The application must retrieve compound-associated chemical-target interaction information form PubChem.
F	Pathway interaction information retrieval	The application must retrieve compound-associated pathway interaction information from PubChem.
F	Pathway-associated protein retrieval	The application must identify and retrieve the proteins involved in the pathways associated with each compound.
F	GeneID translation	The workflow must retrieve gene-related information from NCBI Entrez.
F	UniProt mapping	The workflow must map GeneID identifiers to UniProtKB accession codes to standardize protein-level information.
F	GO annotation retrieval	The workflow must retrieve Gene Ontology annotations associated with the identified proteins through QuickGO.
F	Functional enrichment	The application must group GO terms by aspect and summarize functionality across proteins and compounds.
F	Final data integration	The application must integrate compound, interaction, pathway, protein, and GO information into a unified analytical summary.
F	Comparative analysis support	The application must support the identification of recurrent targets, pathways, or biological functions across multiple analysed compounds.
F	Interactive visualization	The application must display intermediate and final outputs through an interactive web-based interface
F	Report export	The application must generate a downloadable report containing the final integrated analysis.

NF	Usability	The application should provide an intuitive interface suitable for exploratory analysis by non-expert users.
NF	Reproducibility	The workflow should execute the same analytical steps consistently for the same input data in different devices.
NF	Interoperability	The application should integrate heterogeneous public biomedical resources despite differences in formats, identifiers, and access mechanisms.
NF	Robustness	The application should handle invalid input, missing records, and external request failures without interrupting the complete analysis.
NF	In-memory processing	Intermediate results should be processed in memory and not in disk to avoid unnecessary file generation.
NF	Traceability	The application should preserve the correspondence between compounds, retrieved targets, pathways, proteins and annotations throughout the analysis.

Table 1: Functional (F) and Non-Functional (NF) Requirements of the project.

By consolidating these processes into a single workflow, this project aims to provide a structure, automated and accessible approach for the interpretation of PDD. It intends to contribute to the organization and biological interpretation of large-scale public biomedical data, facilitating exploratory analysis and supporting hypotheses generation regarding the molecular targets and pathways potentially involved in the phenotypic effects.

To illustrate the functionality of the proposed application, a validation case study based on statins was included in the present work. Statins constitute a well-characterized drug class known for their therapeutic effects on cholesterol-related diseases [12]. Because they share well-established biological effects, they provide a suitable framework for evaluating whether the workflow can retrieve meaningful recurrent targets, pathway-associated proteins and functional annotations that are biologically consistent with their known pharmacological context.

2 MATERIAL AND METHODS

The following sections describe the materials and the methodological workflow applied in the development of the proposed application. It is divided into two main parts: first, the software environment and the data sources employed for the application are presented; and second, the workflow methodology applied is also described, where each step is outlined and explained from compound identification through SMILES code to the integration of target, pathway and functional annotation data into a unified final report. To make the understanding clearer, the following Figures show different colours to differentiate the tools applied and used throughout the process. PubChem's services are represented in purple, and the PUG-REST service is represented in light blue; Entrez and NCBI services are represented in light green; UniProtKB services and databases are represented in orange; Python tools and libraries are represented in turquoise; and DataFrames are represented in yellow.

2.1 Materials

2.1.1 Software environment

The presented application was developed in Python as the main programming language used to implement the interactive web-based workflow. Python was selected due to its flexibility, modularity, and extensive collection of scientific and data-processing libraries, which made it particularly suitable for the proper design of this application.

On the same note, the user interface was developed using Streamlit, an open-source Python framework that enabled the construction of accessible and interactive web environments. Because of its simplicity and Python compatibility, it was an attractive choice for running the workflow and visualizing the results.

For data handling and processing, the pandas library was used. This open-source data analysis tool allowed all intermediate and final outputs to be processed in memory as DataFrames, thereby facilitating the integration of heterogeneous information retrieved throughout the pipeline from external sources. Moreover, the communication with

external public resources and databases was performed through HTTP requests, while specific queries were handled using the Biopython package.

Finally, Artificial Intelligence was employed as a collaborative tool, specifically ChatGPT. Throughout the project, it was used to generate simple Python functions and for debugging and optimizing parts of the code in the final development. It is important to highlight that the overall design, logical structure, execution order, conceptual basis of the pipeline and the rationale behind each stage, were developed independently and not produced by AI. Regarding the writing of the thesis, ChatGPT was used occasionally to help reformulate certain sections of the methodology and introduction in a more technical and academically appropriate manner.

2.1.2 Public Biomedical Resources

The pipeline was designed to both retrieve and integrate heterogeneous biological data from various external biomedical resources. PubChem, a National's Center for Biotechnology Information public repository and knowledgebase, contains and provides chemical and biological data (https://pubchem.ncbi.nlm.nih.gov). The information is divided in three internal databases: Substance, Compound and BioAssay [13]. This PubChem's database is organized as shown in Figure 1.

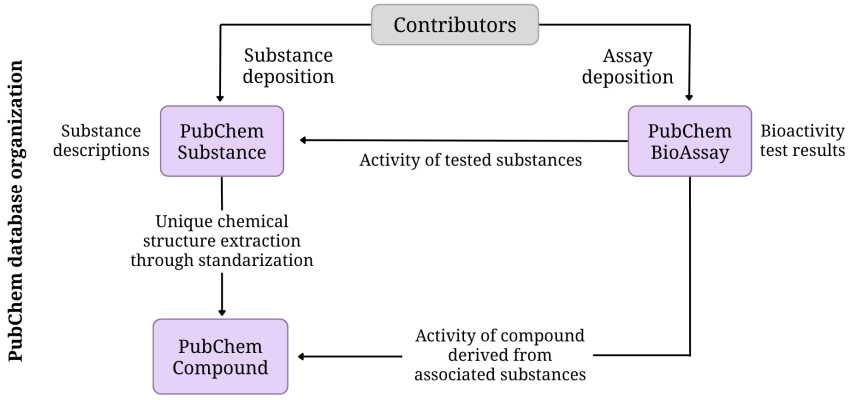


Figure 1. PubChem database organization into three internal databases and their relationships.

It was the primary source for collecting valuable information. PubChem's information can be reached through multiple programmatic access routes (see Figure 2); the collection of these interfaces is called Power User Gateway (PUG). PUG is a suite of

APIs for the NCBI PubChem resource that provides specialised programmatic access to PubChem's information and functions via a single common gateway interface called *pug.cgi*. This is a central gateway to multiple PubChem services and interprets user requests, initiates the proper action and returns results. PUG offers HTTP, REST, View and SOAP interfaces, and integrates with the Entrez utilities (E-utilities) to facilitate access to other useful NCBI databases that will be mentioned further ahead. (https://www.ncbi.nlm.nih.gov/home/develop/api/)

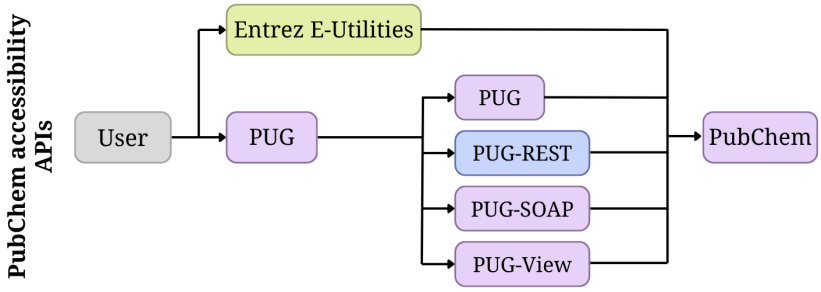


Figure 2. PubChem different accessibility routes.

PubChem's PUG-REST service (Representational State Transfer (REST)-style interface) was used as the main tool for compound identification from user-provided SMILES codes, as well as for compound metadata, chemical-target interactions, and pathway-associated interactions. PUG-REST service enables a simplified access route to PubChem, creating convenient ways to facilitate the retrieval of relevant compound information. [14], [15]. PUG-REST requests are encoded in a three-part HTTP URL: input, describing the identifiers of interest (SID for Substance, CID for Compound or AID for BioAssay databases); operation, specifying what operation to perform with the identifiers; and output, defining the output format (see Figures 3 and 4) [14].

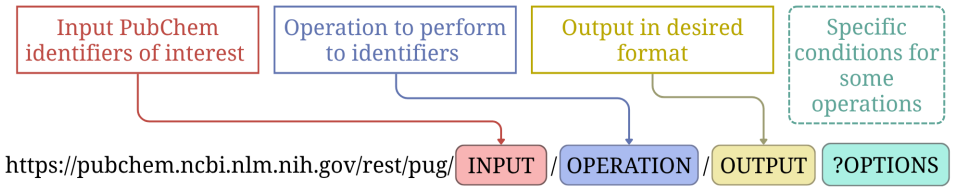


Figure 3. PubChem's PUG-REST URL request syntax.

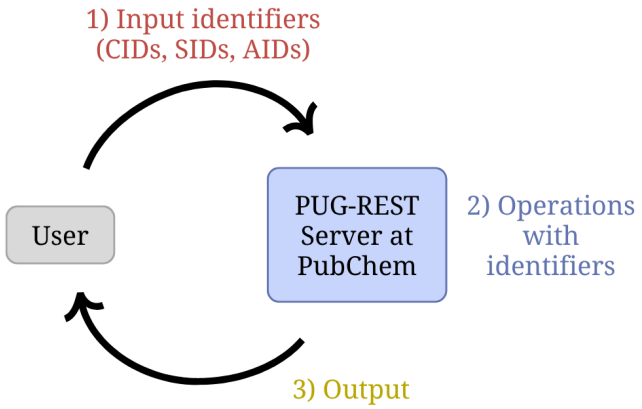


Figure 4. User and PUG-REST request internal workflow.

This service is mainly intended to handle short and synchronous requests (less than 30 seconds), making it a very useful tool for compound metadata retrieval. Another alternative and easy way to access PubChem database in Python via the PUG-REST API is through PubChemPy. This package handles the complexity of the PubChem PUG-REST API, thereby facilitating the syntax and comprehension of chemical informatics workflows. It allows retrieval of PubChem data, for instance, searching in PubChem Substance and Compound databases by name, SMILES and retrieve compound metadata (chemical structures, properties, 3D conformer data, substructure, etc). Furthermore, it allows the conversion between SMILES, PubChem CID and compound name (https://docs.pubchempy.org/en/)

In addition to this, a cheminformatics toolkit called RDKit was included for SMILES code verification (https://www.rdkit.org). RDKit is an open-source toolkit useful for programmatic operations with chemical data [16].

Additionally, Biopython, an open source of tools and libraries for biological computation available in Python, was used to access to NCBI services through Entrez [17]. Entrez is a publicly accessible database system designed to integrate bibliographic data and molecular biology databases. (https://www.ncbi.nlm.nih.gov/Web/Search/entrezfs.html). It integrates an intuitive user interface for quick data retrieval. To access this service and query it programmatically, E-utilities were used as the official API between Entrez and Biopython. They were employed to retrieve each target's gene-related information from the NCBI Gene dataset.

The Entrez Programming Utilities (E-utilities) are composed of eight server-side programs that enable programmatic access into the Entrez data retrieval system, which provides users access to NCBI's databases such as PubChem, GenBank or GEO. This service aids with the retrieval or search of data using a fixed URL syntax that translates into the necessary values of interest as described in Figure 5 [18]. Because of this, Biopython's Bio.Entrez module was used for programmatic access to Entrez. This module implements the E-utilities and checks if proper URL syntax is used when requesting data to NCBI page, as well as ensuring that proper guidelines for responsible data access are being followed.

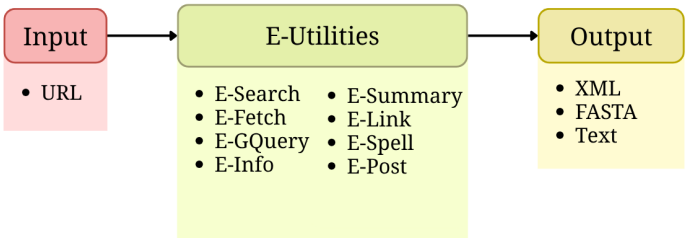


Figure 5. Entrez eight E-utilities services and request workflow.

Moreover, UniProt (https://www.uniprot.org), the world's leading freely accessible resource of protein sequence and functional information, was used to programmatically standardize the collected biological targets through its UniProt ID Mapping service [19]. This tool allowed the correspondence between GeneID entries and UniProt accession codes to be established as a form of target standardization.

Finally, QuickGO was accessed to retrieve Gene Ontology annotations associated with the identified proteins. The Gene Ontology database (https://www.geneontology.org) is the world's largest source of information on the function of genes. It provides vast amounts of detailed ontologies of terms describing molecular functions that enrich the biomedical computational research and molecular biology computational analyses. QuickGO is a highly valuable tool in providing biological insights for large proteomic datasets. The QuickGO browser allows easy access of the Gene Ontology (GO) annotations in the GOA (Gene Ontology Annotation) database (https://www.ebi.ac.uk/GOA), which include more than 45 million high-quality Gene Ontology annotations to millions of proteins in the UniProt Knowledgebase

(UniProtKB) [20] The QuickGO web-browser allows users to search protein-associated GO annotations (molecular function, biological process and cellular component) from their UniProtKB accession number. The relationships between these databases is graphically represented in Figure 6. Moreover, the interface provides a REST-style query interface compatible with programmatically HTTP requests, which is highly convenient for this project's Python workflow [20].

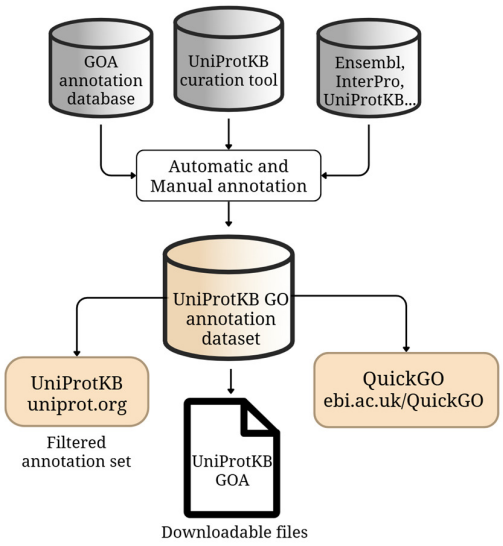


Figure 6. Schematic representation of the relationship between GOA, UniProtKB, external annotation sources and QuickGO

In this project, APIs and web services constituted a fundamental component of the workflow. They were essential for accessing public biomedical databases programmatically and collecting data into the workflow. An Application Programming Interface (API) is a programmatic mechanism that allows software applications to communicate with external services and retrieve information automatically [21]. All the information handled in the application was retrieved programmatically from the previously mentioned resources through publicly available APIs (PUG-REST, E-Utilities, QuickGO) and web services (PubChem, NCBI, UniProtKB, Gene Ontology Annotations), allowing the application to automatically retrieve and integrate distributed biomedical information into a unified analytical framework.

2.2 Workflow Methodology

The methodological design of this project follows a computational pipeline that intends to automatically retrieve biological insight from a collection of compounds that are known to have a phenotypic effect on biological systems, but whose underlying mechanisms are not yet fully understood. Starting from user-provided SMILES codes, the workflow collects compound-associated information from public biomedical resources and integrates it into a unified analysis. The code used for this process is uploaded into the GitHub repository as https://github.com/paulablancog/Drug-Discovery-App.git and deployed as an application in Streamlit Community Cloud service as https://drugdiscoveryapp-tfg.streamlit.app.

The pipeline of the web platform is described in Figure 7, involving the following four layers:

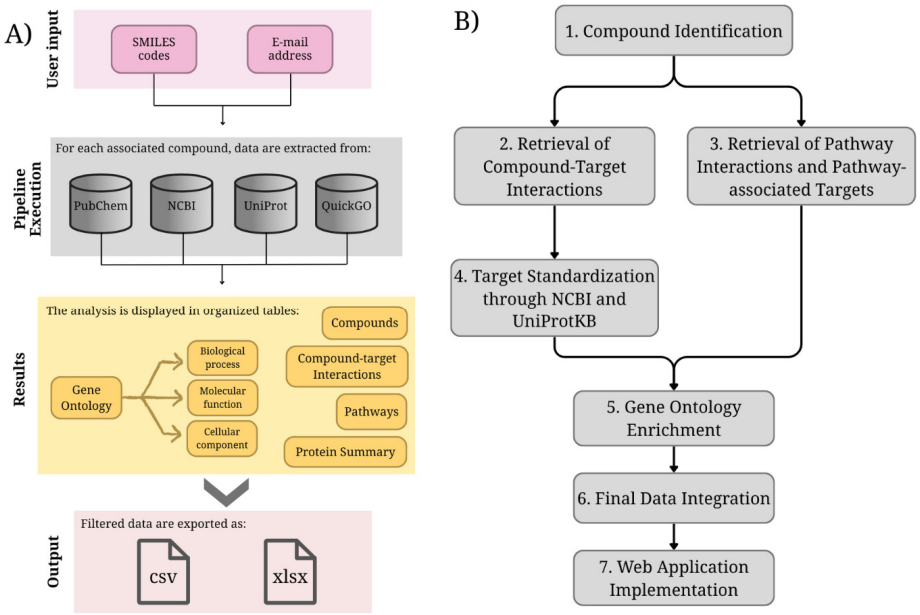


Figure 7. Application summaries. (A) Application data path and tools. (B) Application pipeline stages.

Throughout the workflow, all retrieved responses, such as JSON, CSV, and TXT data are temporarily stored in in-memory variables rather than on disk storage. This facilitates reproducible execution across different computing environments. Furthermore, the application includes input-validation and code-injection mitigation to improve security.

Sensitive user information including the email address introduced, is handled only to send and retrieve requests, thereby contributing to privacy protection and secure use of the application.

2.2.1 Compound Identification from SMILES codes

The first stage of the workflow corresponds to the identification of chemical compounds from the input SMILES codes. Text string encoded molecular structures are a supported chemical line notations that become useful when constructing programmatic database queries in PubChem and other chemical databases. SMILES offer a human-friendly character syntax representation of molecule structures that are compatible for compound analyses [22], [23]. Although they facilitate the processing of molecules, SMILES strings are complex and can be represented in alternative ways while being chemically equivalent for the same compound. Therefore, an initial validation and standardization step was applied before querying external databases.

Firstly, each submitted SMILES string is validated using RDKit, a tool that allows checking and canonizing each string [16]. Empty entries, and duplicates and invalid SMILES codes are ignored. For valid SMILES, its canonical representation is generated, preserving stereochemical information. This canonicalization ensures that all SMILES codes introduced are registered in a standardized manner for their further analysis.

After validation and standardization, each SMILES code is queried against PubChem for compound identification. The identification is carried out by programmatic queries through PubChem's PUG-REST service, using the SMILES code string as the query input. The PUG-REST service allows users to retrieve PubChem's structured data through the submission of specific queries as HTTP URLs [15]. To optimize identification specificity, a two-step filter is employed beforehand. Compound Identification Numbers (CIDs) are retrieved through PubChem's PUG-REST and saved temporarily for validation. Firstly, the exact stereochemical identity is searched; if no compound is still identified, a second search based on exact connectivity is performed. This approach prioritizes the most chemically precise match while permitting

identification when stereochemical information is incomplete or not available in the database. This distinction is important because two SMILES strings may share the same connectivity while differing in explicit stereochemical annotation. For instance, Cc1ccc(C2OC(CO)C(O)C(O)C20)cc1Cc1ccc(-c2ccc(F)cc2)s1 and CC1=C(C=C(C=C1)[C@H]2[C@@H]([C@H]([C@@H]([C@H](O2)CO)O)O)O)CC3=CC=C(S3)C4=CC=C(C=C4)F are related representations of the same compound (canagliflozin), but only the second one specifies stereochemistry explicitly. Because the workflow preserves stereochemistry during canonicalization, these two inputs produce two independent canonical SMILES and are therefore treated as distinct unique compounds. Once compounds are finally identified, their candidate compound identifiers are returned. The retrieved entries are ranked according to the availability of informative metadata, and the highest-ranked candidate is selected as the representative and final compound.

Compound name and metadata are retrieved from PubChem, including CID, preferred compound name, molecular formula and molecular weight using PubChemPy, a simple Python wrapper around the PubChem PUG-REST API that allows conversion between SMILES, compound name and CID.

Finally, the resulting compounds and metadata are organized into a structured table containing the generated canonical SMILES code, the identified compound name, its PubChem CID, molecular formula, molecular weight, and an identification status. Invalid or non-identified compounds are also displayed in the output table, ensuring traceability of all user inputs throughout the workflow (see Figure 8).

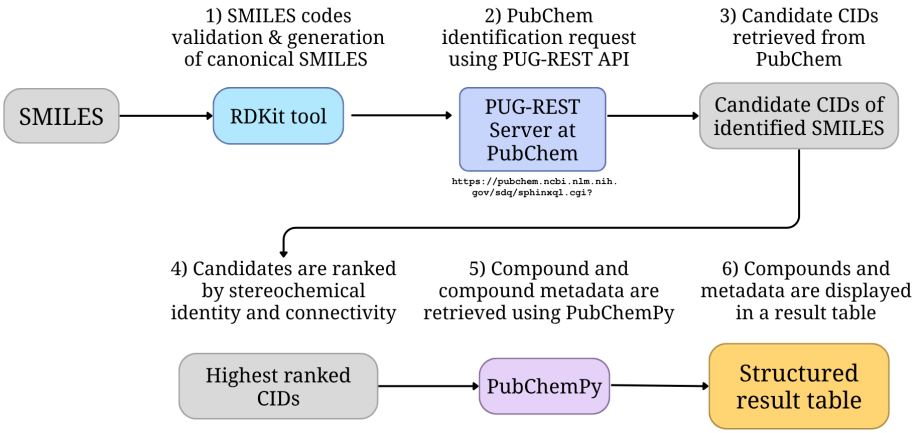


Figure 8. SMILES identification workflow and all programmatical tools used.

2.2.2 Retrieval of Compound-Target Interactions

After the successful completion of compound identification and identifier retrieval from the user-provided SMILES code, the next stage is the retrieval of the compound's chemical interactions. The process begins with an initial request to the PubChem PUG-View index for the compound record in a JSON output format. This first call is necessary to examine programmatically the structural organization of the compound page in the web, including the headings and nested subsections available for the compound, and determine whether the compound contains an "Interactions and Pathways" section.

Hence, the JSON extracted is examined by recursively flattening the section headings and nested subsections. The pipeline searches for a section whose heading matches "Interactions and Pathways". If the section is not found, the compound is considered not to contain interaction or pathway information accessible through this route, and the interaction retrieval process is omitted. When the "Interactions and Pathways" section is detected, its internal PUG-View data JSON is retrieved through a second PubChem request to differentiate two main subsections: "Chemical-Target Interactions" and "Pathway Interactions", as shown in Figure 9. This second, more specific JSON contains the actual structured content and the valuable data of interest.

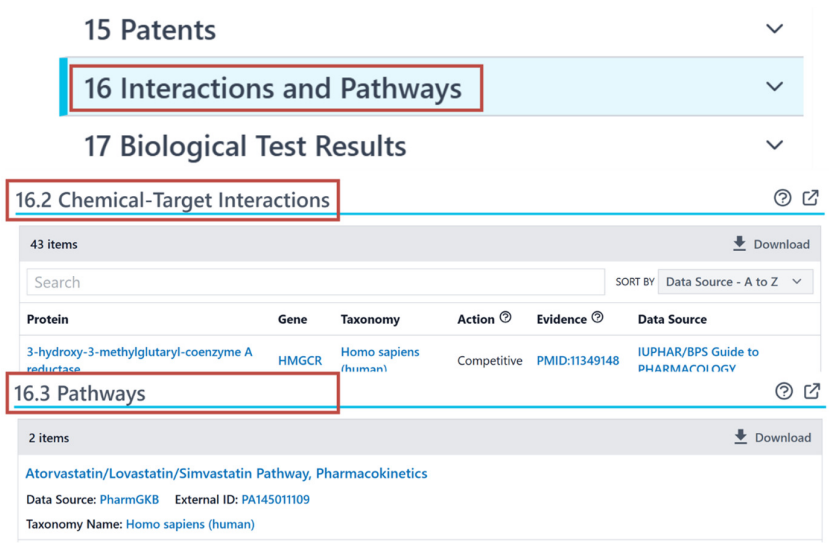


Figure 9: PubChem's index section "Interactions and Pathways" and subsection "Chemical-Target Interactions" used for data retrieval.

Once the section JSON has been retrieved, its internal structure is parsed to identify the relevant subsections. For each subsection, the workflow extracts the associated external table provided by PubChem, when available, through the *ExternalTableName* field in the JSON record. This step is essential because these external tables contain the structured interaction and pathway data used in the subsequent stages of the workflow. However, this information is not directly embedded in the PUG-View response itself. Instead, PubChem exposes the underlying tabular data through an associated external table. At this point, the workflow explicitly distinguishes between the two subsections found: “Chemical-Target Interactions” and “Pathway Interactions”. For compound-target interaction retrieval, only the external tables associated with “Chemical-Target Interactions” are considered in this first stage.

The external table is queried through PubChem’s SDQ service. The PubChem Structured Data Query (SDQ) is an internal service used by PubChem web pages which allows the retrieval of tabular biological information linked to a compound [22].

The PubChem SDQ agent is conceptually similar to the PUG-REST since a structured HTTP request is submitted to PubChem and machine-readable data are returned, typically in JSON format. No official documentation nor fixed syntax is present for the construction of the URLs. Nonetheless, in practice, basic SDQ syntax can be

deconstructed by inspecting the network traffic generated by the PubChem web interface when downloading the relevant external tables. In the present workflow, this exploration revealed that the table associated with the Chemical-Target section is retrieved through the `sdq/sphinxql.cgi` endpoint. Accordingly, the pipeline constructs SDQ queries against this endpoint to recover the tabular interaction data needed. [22].

Therefore, the pipeline submits a SDQ query containing the collection or table name, the needed fields, the sorting order (based on gene ascending order), the starting row, the maximum number of rows to be returned, and the filtering condition (only rows associated with the analysed compound are meant to be returned). Moreover, the corresponding SDQ query is requested with automatic pagination to retrieve the complete set of available rows in the external table. The workflow first collects the initial block of rows and then verifies if additional pages are available by checking the total row count of the table. In this way, the pipeline ensures the recovery of all interaction associated with each compound. The table rows are retrieved in a JSON format for an easier handling of information.

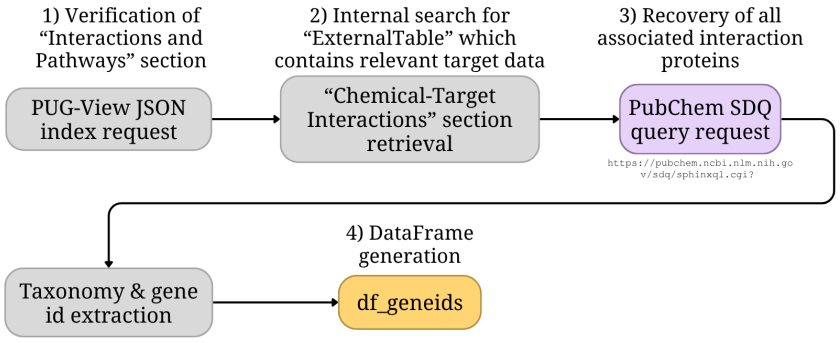


Figure 10: Workflow of Chemical Interactions retrieval into df_geneids DataFrame.

The final rows constitute the initial compound-target interaction dataset. To summarize and preserve an organized structure, the workflow extracts a key target identifier that will be used for downstream analysis: the GeneID. For every interaction row, the pipeline examines each target individually, and whenever a valid *geneid* and *taxid* field is present, it is collected and linked to the corresponding compound name (see Figure 10). The resulting table is a simplified target table containing the compound and its associated GeneID value as shown in Table 2:

compound	*geneid*	*taxid*	*taxname*
atorvastatin	196	9606	*Homo sapiens*

Table 2: df_geneids DataFrame.

2.2.3 Retrieval of Pathways and Pathway-associated Targets

In parallel with compound-target interaction data retrieval, the pipeline follows identical steps to extract pathway-related information associated with each identified compound. This stage expands the analysis from compound-level interactions to a broader pathway-centred biological interpretation, allowing the recovery of, not only pathways in which the compound is known to be involved, but also of the proteins involved in those pathways.

2.2.3.1 Pathways Retrieval

The process begins once a valid PubChem CID has already been obtained from the user-provided SMILES code and the compound has been successfully identified. Using this CID, the workflow queries the PubChem PUG-View index for the compound record in a JSON format to programmatically inspect the general structure of the compound page and verify whether an "Interactions and Pathways" section is present. If the namely section is absent, the compound is considered not to contain interactions or pathway information accessible through this route. The retrieval of pathway interactions stage is completely omitted, and no relevant data are retrieved. On the other hand, if the pipeline verifies the existence of a section whose heading matches "Interactions and Pathways", its internal PUG-View data JSON is retrieved to differentiate two main subsections: "Chemical-Target Interactions" and "Pathway Interactions".

For each subsection contained inside "Interactions and Pathways", the workflow examines the associated *ExternalTableName* whenever present. By doing this, the corresponding PubChem external tables linked to each subsection are recovered, as well as the underlying tabular resources associated with the visual information displayed in the PubChem compound record. For the pathway component, the corresponding external table is retrieved through a dedicated PubChem SDQ query against PubChem's pathway

collection. The workflow submits a structured query to the PubChem SDQ service to retrieve the complete set of pathway records associated with the compound's CID. The query filters the pathway collection by the CID and results are requested in JSON format. In addition to this, because PubChem responses tend to be paginated, the workflow implements an automatic pagination mechanism. Basically, it first retrieves the initial result block of rows and iteratively requests additional pages until the total count of rows corresponds to the number of rows retrieved. The output of this stage is a complete set of pathway rows associated with the compound. This marks the starting point for the further pathway-level analysis.

After the collection of all pathway rows, both *pathwayid* and *pwacc* JSON fields that serve as pathway identifiers are examined. The field *pathwayid* corresponds to the internal PubChem identifier used to query pathway-specific resources, whereas *pwacc* is the pathway accession code which allows further processing for labelling and grouping proteins involved in that specific pathway. Only rows containing both identifiers were maintained for downstream processing (see Table 3). Figure 11 graphically summarizes this workflow.

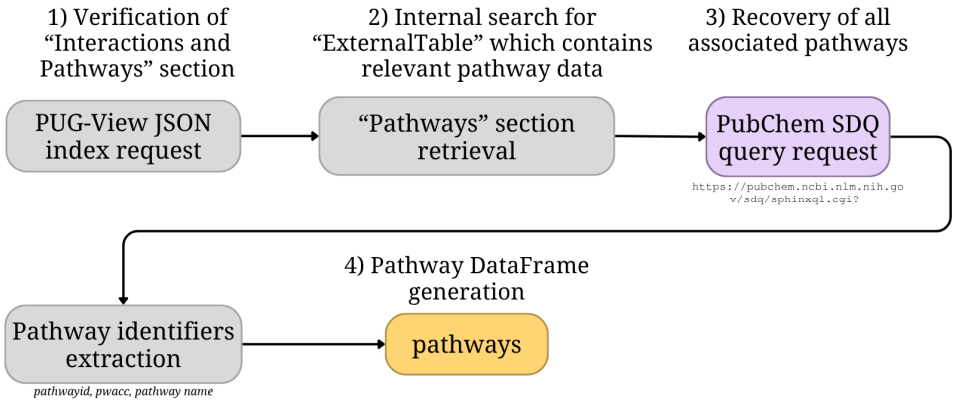


Figure 11:Workflow of Pathways retrieval into pathways DataFrame

cid	*pathwayid*	*pwacc*	*name*
atorvastatin	1184773	PathBank:SMP0000131	Atosvastatin Action Pathway

Table 3: pathways DataFrame

2.2.3.2 Pathway-Associated Proteins Retrieval

Once all pathway rows are collected and filtered, the pipeline initiates a new process and moves from the compound-level pathway list to the collection of pathway-associated proteins. From each individual pathway row, a new PubChem SDQ query is built to request a JSON table containing all possible proteins associated with that pathway. This request is configured to return a JSON output and to allow sufficient large row limit so that all pathway-associated proteins are collected in a single submission whenever possible. If this request is successful, the pipeline extracts each associated protein record individually.

Within each target record, the workflow retrieves valuable data of interest: *acc* and *protname*. The field *acc* corresponds to the UniProt accession identifier associated with the protein, while *protname* is the protein name. These two fields are then merged into a table in addition to their associated pathway accession code (*pwacc*), and the name and CID of the initial compound to ensure a structured protein-level record.

Therefore, the resulting standard row for each pathway-associated protein contains the following variables: UniProt accession, protein name, pathway accession identifier, compound name and CID. During this process, a taxonomic filtering is executed. Proteins whose taxonomic identifier or organism name corresponds to the user-input taxonomy are retained as pathway-associated proteins. However, those that do not match the previously user-selected taxonomic fields are removed. If the pathway ends up without any proteins associated, it is removed from the Pathway table (see Table 4).

Finally, the workflow iterates this process throughout all pathways associated with the compound, and the resulting tables were concatenated into a single DataFrame. After concatenation, data are filtered and duplicated entries are removed using the combination of UniProt accession, protein name, pathway, compound and CID as the uniqueness criterion all together (see Figure 12). This ensures that repeated PubChem records do not bias or disturb the biological evidence of targets. The resulting DataFrame consists of all pathway-derived proteins of the analysed compound as shown in Table 5.

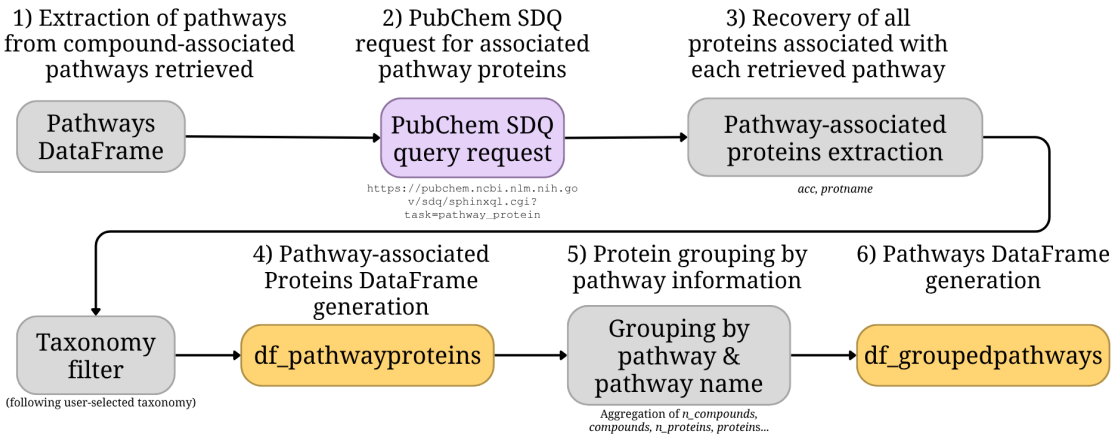


Figure 12: Workflow of Pathway-associated protein retrieval into df_pathwayproteins and df_groupedpathways DataFrames.

pathway	*pathway_name*	*n_proteins*	*n_compounds*	*compounds*	*uniprot_accessions*	*taxid*	*taxname*
PathBank:SMP0000131	Atosvastatin Action Pathway	21	1	atosvastin	[protein codes]	9606	*Homo sapiens*

Table 4: df_groupedpathways DataFrame

uniprot_accession	*protein_name*	*pathway*	*pathway_name*	*compound*	*taxid*	*taxname*
P53602	Diphosphomevalonate decarboxylase	PathBank:SMP0000131	Atorvastatin Action Pathway	atorvastatin	9606	*Homo sapiens*

Table 5: df_pathwayproteins DataFrame

2.2.4 Target Standardization Through NCBI Gene database and UniProtKB ID Mapping

From this stage, targets have been retrieved and stored in two different simplified tables: compound-target interaction table (see Table 2), including the compound name and the associated GeneID values extracted from PubChem, and pathway-associated target table (see Table 4), which contain protein rows and their UniProt accession, protein name, pathway, compound and associated CID. To gain significant and complete biological insight from these data, both tables need to be combined in an organized and

structured way. For this reason, the next stage of the workflow consists of a target standardization process, in which GeneID identifiers from compound-target interaction rows are translated into interpretable gene information and mapped to standardized UniProtKB accession codes.

The standardization process is divided into two complementary operations. Firstly, the pipeline queries the NCBI Gene database through Entrez to retrieve gene symbols and descriptive annotations associated with each GeneID. Secondly, the workflow maps those same GeneIDs to UniProtKB accessions using the UniProt ID Mapping service. After completing these two steps, interaction-derived target GeneIDs are successfully converted into biologically interpretable data, as well as standardized into a protein representation suitable for downstream integration with pathway-derived proteins.

2.2.4.1 Retrieval of gene-related information from the NCBI Gene database through Entrez

The purpose of this first retrieval stage is to obtain significant information about the compound-interaction proteins for their further analysis and interpretation. To do so, the workflow handles the retrieval from NCBI Gene database using E-utilities provided through Biopython package in Python. But before accessing the NCBI services, all repeated target GeneIDs rows from interaction-derived target table are removed. The ones that remain, are reduced to a unique list.

To access NCBI databases, users are required to provide an email address to control requests and access. The introduced email is assigned to the Entrez configuration before the gene-related information request is submitted. GeneIDs list is now processed in batches of 200. The pipeline submits an Entrez summary request against the NCBI Gene database for each batch of GeneIDs. As it is an Entrez call, the service returns an XML response containing the summary information for all genes in the batch. The XML response is parsed using Biopython's Entrez reader, extracting three main fields for each returned gene record:*geneid*, *gene symbol*, and *description* fields. These values are inserted and reorganized into a structured DataFrame (see Table 6) in which each row

includes the gene's GeneID, gene symbol, a descriptive annotation and the associated compound's name, as described in Figure 13.

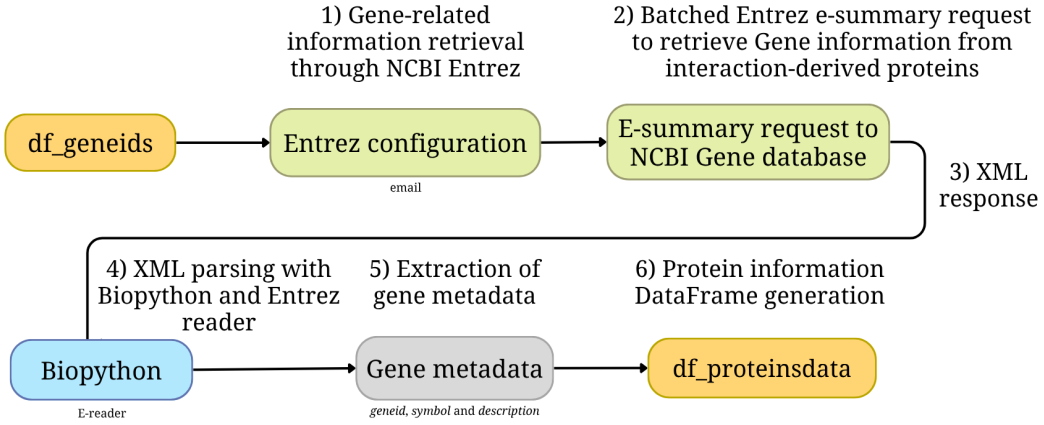


Figure 13: Workflow for Protein information extraction into df_proteinsdata DataFrame.

The result of this first operation is mainly to obtain valuable information from the targets retrieved from "Chemical-Target Interactions" data. Adding human-readable symbol and description for each interaction-derived target enriches the previous GeneID representation with a more interpretable biological context.

compound	*geneid*	*symbol*	*description*
atorvastatin	196	AHR	aryl hydrocarbon receptor

Table 6: df_proteinsdata DataFrame

2.2.4.2 Mapping GeneID identifiers to UniProtKB accession codes

On the other hand, while translating GeneIDs through NCBI Gene database improves the interpretability of the data retrieved, the workflow must compute an additional normalization step to integrate interaction-derived proteins and pathway-derived proteins into a unified framework. To achieve this, the workflow performs a second standardization process in which unique GeneIDs are mapped to UniProtKB accession codes using UniProt ID Mapping tool. The UniProtKB accession code is then used as the common key for protein identification throughout the workflow (Figure 14).

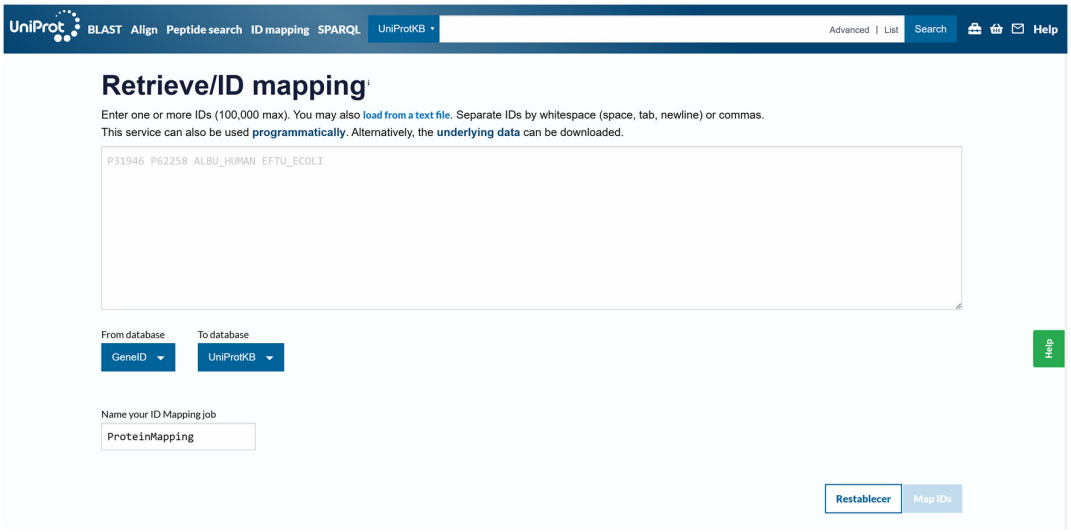


Figure 14. UniProt's Mapping tool in UniProt webpage.

The UniProt ID Mapping tool aids with mapping between imported databases of interest. In this workflow, the mapping is carried out programmatically through the UniProt REST service, like PubChem's PUG-REST performance. The tool requires a source and a target identifier system. Therefore, the workflow first queries the UniProt configuration endpoint to identify the internal database definitions corresponding to GeneID and UniProtKB. Once these are resolved, the full GeneID is submitted to the `idmapping/run` endpoint. The requests made are computed as an asynchronous job. This is due to UniProt's mapping process on the server side. Results are not available immediately after submission, so UniProt provides a job identifier to detect whether the job has been completed or not. The pipeline implements a polling procedure to control the status of the mapping job. At regular intervals, the workflow queries the UniProt `idmapping/status` endpoint until one of the following conditions is reached: successful completion of the job, job failure, or timeout.

If the job is completed successfully, the mapping results are downloaded from the UniProt `idmapping/results` endpoint in JSON format. Additionally, since the results set may be distributed across multiple pages, the workflow examines the HTTP Link header of each response to ensure a complete recovery of all mapping results, as it is the place where the page number is programmatically displayed. Each returned record contains the original GeneID and the corresponding UniProt accession (see Figure 15).

To preserve uniqueness, the workflow ensures that each GeneID identifier corresponds to a single representative UniProt accession. This is done by the selection of accessions that preferentially started with the letter "P", since these typically correspond to standard UniProtKB entries. The final output of this stage is a mapping table containing for each GeneID its representative UniProt accession as well as the list of all mapped accessions retrieved as displayed in Table 7.

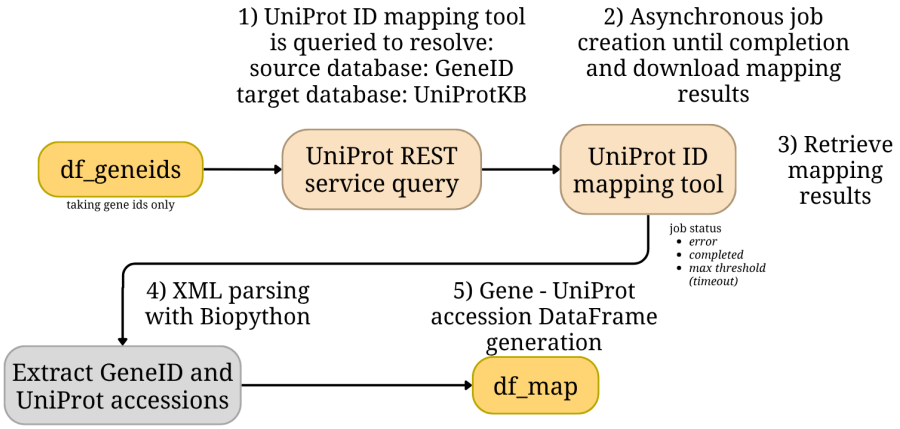


Figure 15: Workflow of UniProtID mapping into df_map DataFrame.

geneid	*Uniprot_accession*	*Uniprot_accessions*
196	P35869	[all uniprot codes associated]

Table 7: df_map DataFrame

2.2.4.3 Integration of gene translation and protein mapping

Once both stages of standardization are completed, information is integrated into a single table. The previously generated Entrez translation table (see Table 6) is merged into the mapping table (see Table 7), containing its associated compound name, GeneID, gene symbol, gene description, and its representative UniProt accession.

In this integrated representation, each interaction-derived target retains its original GeneID identity while also being associated with a symbol, a description, and a unique UniProtKB accession code. This workflow is graphically represented in Figure 16 and displayed as a table inside the application in Figure 17. This multistep normalization process ensures that heterogeneous biological evidence retrieved from distinct different

public resources can be accessed through multiple ways and can be effectively compared and merged consistently in the subsequent integration steps.

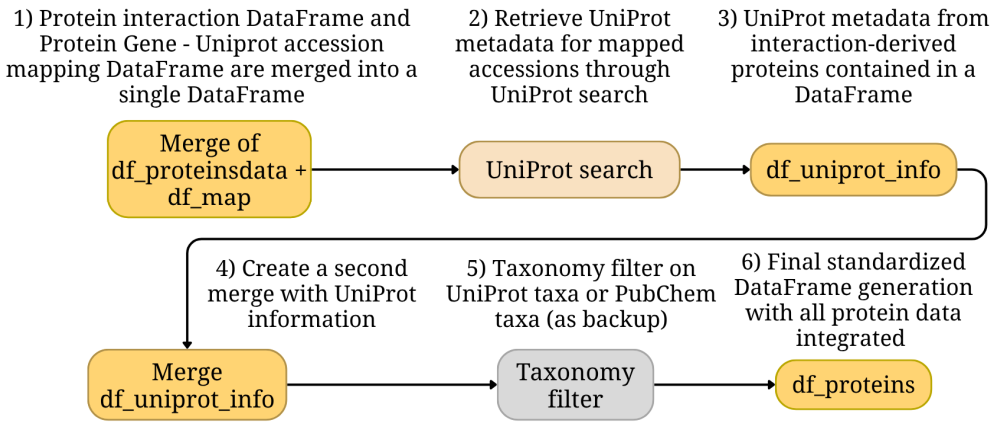


Figure 16: Workflow of protein standardization and filtering into df_proteins DataFrame.

Compound-Protein Interactions

Show 10 entries Search:

compound	geneid	symbol	protein_name	uniprot_accession	description	taxid	taxname
Search compound	Search geneid	Search symbol	Search protein_name	Search uniprot_acces	Search description	Search taxid	Search taxname
atorvastatin	196	AHR	Aryl hydrocarbon receptor	P35869	aryl hydrocarbon receptor	9606	Homo sapiens
atorvastatin	213	ALB	Albumin	P02768	albumin	9606	Homo sapiens
atorvastatin	1244	ABCC2	ATP-binding cassette sub-family C member 2	Q92887	ATP binding cassette subfamily C member 2	9606	Homo sapiens
atorvastatin	1551	CYP3A7	Cytochrome P450 3A7	P24462	cytochrome P450 family 3 subfamily A member 7	9606	Homo sapiens
atorvastatin	1555	CYP2B6	Cytochrome P450 2B6	P20813	cytochrome P450 family 2 subfamily B member 6	9606	Homo sapiens
atorvastatin	1557	CYP2C19	Cytochrome P450 2C19	P33261	cytochrome P450 family 2 subfamily C member 19	9606	Homo sapiens
atorvastatin	1558	CYP2C8	Cytochrome P450 2C8	P10632	cytochrome P450 family 2 subfamily C member 8	9606	Homo sapiens
atorvastatin	1559	CYP2C9	Cytochrome P450 2C9	P11712	cytochrome P450 family 2 subfamily C member 9	9606	Homo sapiens
atorvastatin	1565	CYP2D6	Cytochrome P450 2D6	P10635	cytochrome P450 family 2 subfamily D member 6 (gene/pseudogene)	9606	Homo sapiens
atorvastatin	1576	CYP3A4	Cytochrome P450 3A4	P08684	cytochrome P450 family 3 subfamily A member 4	9606	Homo sapiens

Showing 1 to 10 of 670 entries Previous 1 2 3 4 5 ... 67 Next

Figure 17. Compound-Protein Interactions table in Streamlit application. The table refers to the DataFrame called df_interactions in the application data flow

2.2.5 Gene Ontology Target Enrichment

Once interaction-derived targets have been standardized to UniProtKB accession codes and integrated into the pathway-associated protein table, the workflow starts the functional enrichment process with Gene Ontology. Functional characterization with GO inserts an additional biological interpretation layer to the targets retrieved throughout the whole pipeline. This enrichment process is implemented using QuickGO services

provided by EMBL-EBI. QuickGO browser allows easy access to Gene Ontology annotations in the GOA database.

2.2.5.1 Gene Ontology term enrichment with QuickGO

The annotation process begins with the protein summary produced in the previous stage, a fully standardized summary of interaction- and pathway- based evidence protein table. From now on, the workflow retrieves protein-level information using each target's UniProt accession code. Thereby, each unique accession code is queried to the QuickGO annotation search service as a REST-style HTTP request. The QuickGO annotation service requires the identifier of the biological entity (*geneProductId*), which is set to the corresponding Uniprot accession in the form `UniProtKB:<accession>` by the pipeline. The retrieval is configured to include the three principal Gene Ontology aspects: biological process, molecular function, and cellular component. These aspects are automatically defaulted by the workflow directly to the service. The pipeline, just like previously mentioned in older stages, implemented a pagination mechanism. Since the number of annotations may exceed the maximum size of a single response page, each request is iteratively submitted until all pages associated with the accession were retrieved. As a result, from the returned records, the GO identifier and the GO aspect are stored together with the corresponding UniProt accession in a unified table. This table constitutes the first level of Gene Ontology information in the workflow, as it captures the direct relationship between proteins and GO identifiers.

However, GO identifiers alone are not sufficient for biological interpretation, so the workflow implements a second retrieval step focusing on GO term metadata. Firstly, all unique GO identifiers present in the previous annotation table are gathered and reduced to a unique set and queried in chunks to the QuickGO ontology service. From the returned ontology records, the workflow extracts the GO identifier and its term name. These names are then merged back into the main annotation table using the GO identifier as the linking key (see Figure 18). In this way, the initial protein-GO mapping is transformed into a more informative structure containing, for each UniProt accession, the associated GO identifier, GO term name, and GO aspect as displayed in Table 8.

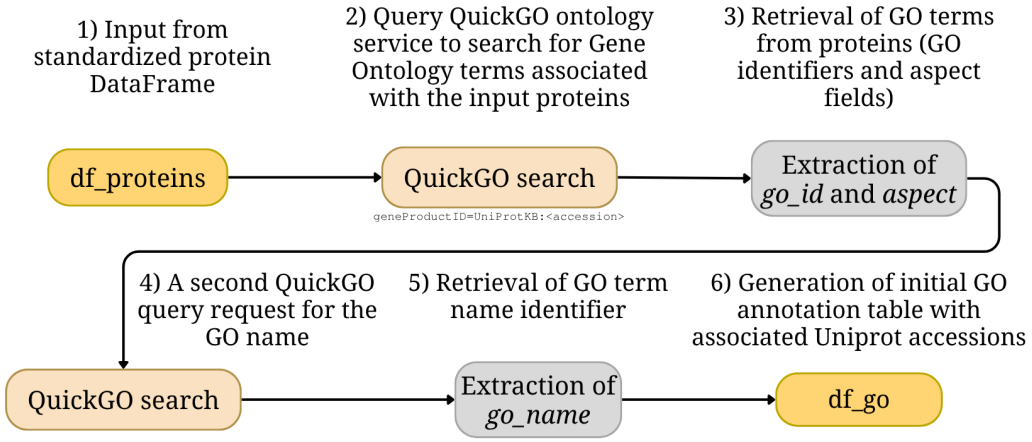


Figure 18: Workflow of GO term retrieval and annotation into df_go.

uniprot_accession	*go_id*	*go_name*	*aspect*
P53602	GO:0005737	cytoplasm	Cellular component

Table 8: df_go DataFrame.

2.2.5.2 Addition of protein and compound information

For the subsequent biological interpretation of this GO information, it must be connected to the broader compound-centred interpretation framework. For this reason, the GO annotation table (see Table 8) is enriched with protein and compound information.

Firstly, a mapping between each UniProt accession code and its corresponding protein symbols is extracted from the protein summary and merged into the GO table. Each UniProt accession-symbol match is associated to their GO annotation, which improves readability.

Secondly, the workflow incorporates compound association information by mapping each UniProt accession and the compound or compounds linked to that protein and merging them into the GO table. Through this operation, each GO annotation is also connected to the compound from which that protein was originally identified. After these enrichments included into the GO table, it is reduced to the following fields displayed in Table 9.

uniprot_accession	*go_id*	*go_name*	*symbol*	*aspect*	*compounds*
P53602	GO:0005737	cytoplasm	MVD	Cellular component	Rosuvastatin; atorvastatin; fluvastatin

Table 9: df_go DataFrame updated with protein information.

A deduplication process is additionally applied to prevent repeated annotations and artificially inflating the later summaries.

2.2.5.3 Differentiation into Gene Ontology aspects

Once the complete annotation table has been standardized, the workflow separates the records according to the three Gene Ontology aspects: Biological Process, Molecular Function and Cellular Component, into three independent tables as shown in Figure 19.

Figure 19. GO Term grouping by aspect: Biological process, Molecular function and Cellular component.

Each of these aspect-specific tables maintain the previous fields from the GO annotation table. This separation allows the application to present the GO results in a more structured and biologically interpretable way, since each aspect represents a different biological dimension of protein function. At this stage, the workflow generates four complementary GO-level outputs: a global annotation table containing all Gene Ontology terms, and three additional tables corresponding to Biological Process, Molecular Function, and Cellular Component.

go_name	*go_id*	*n_compounds*	*compounds*	*n_proteins*	*proteins*

cytoplasm	GO:0005737	7	[…]	236	[…]

Table 10: df_go final update DataFrame.

To further improve interpretability, each aspect-specific table is then grouped by the pair (*go_name, go_id*). Through this aggregation, all proteins and compounds associated with the same GO term are summarized into a single row, together with the number of proteins and compounds linked to that term (see Table 10). This grouping highlights which GO terms are shared by multiple proteins and compounds and hence, allows the identification of recurrent functional categories potentially linked to the shared phenotypic effects of the analysed compounds.

2.2.6 Final Data Integration

Finally, after the retrieval of interaction-derived proteins, pathway-associated proteins, and Gene Ontology annotations, the workflow integrates all evidence layers into a unified protein-centred framework in which each standardized protein is represented in a single table with its interaction support, pathways associations, compound links and functional annotation.

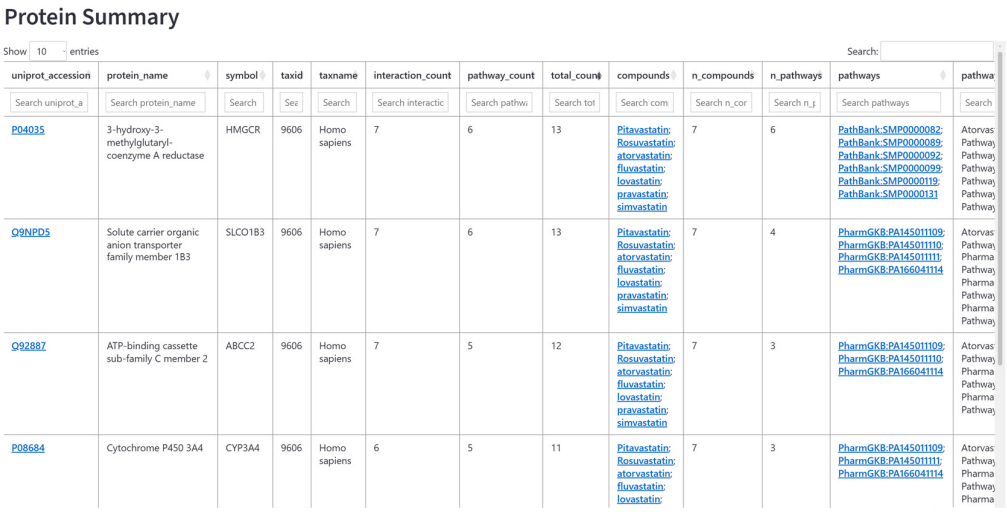
This integration process starts with the two first generated tables: the interaction summary table and the pathway-derived protein table.

Firstly, the pipeline generates an interaction-based protein summary (*df_interactions*), in which for every protein, it calculates the number of distinct associated compounds to observe how many times the protein is associated to any of the compounds submitted as a direct chemical interaction. In addition to this, Gene symbols and GeneID identifiers are also grouped into the generated DataFrame.

Secondly, the pipeline generates a pathway-based protein summary (*df_pathways*), in which for every protein, it calculates the number of distinct associated pathways and compounds to observe how many time the protein is present in our list of compounds from pathway interactions. Additionally, because some proteins lack gene

symbols, the corresponding Uniprot accessions are queried again through UniProt to obtain their mapped symbol whenever possible.

Even though both summary tables are generated independently and contain different fields, the UniProt accession serves as the common identifier for both, since it had already been established as the standardized protein-level reference during the target normalization stage. As a result, the workflow can merge both tables using the UniProt accession field. This merging strategy ensures that all proteins collected throughout the pipeline either supported exclusively by interaction evidence, exclusively by pathway evidence, or by both evidence layers, remain represented in the final integrated table.



Protein Summary

Show 10 entries Search:

uniprot_accession	protein_name	symbol	taxid	taxname	interaction_count	pathway_count	total_count	compounds	n_compounds	n_pathways	pathways	pathwa
Search uniprot_a	Search protein_name	Search	Sea	Search	Search interactic	Search pathw	Search tot	Search com	Search n_cor	Search n_p	Search pathways	Search
P04035	3-hydroxy-3-methylglutaryl-coenzyme A reductase	HMGCR	9606	Homo sapiens	7	6	13	Pitavastatin; Rosuvastatin; atorvastatin; fluvastatin; lovastatin; pravastatin; simvastatin	7	6	PathBank:SMP0000082; PathBank:SMP0000089; PathBank:SMP0000092; PathBank:SMP0000099; PathBank:SMP0000119; PathBank:SMP0000131	Atorvas Pathwa Pathwa Pathwa Pathwa Pathwa Pathwa Pathwa
Q9NPD5	Solute carrier organic anion transporter family member 1B3	SLCO1B3	9606	Homo sapiens	7	6	13	Pitavastatin; Rosuvastatin; atorvastatin; fluvastatin; lovastatin; pravastatin; simvastatin	7	4	PharmGKB:PA145011109; PharmGKB:PA145011110; PharmGKB:PA145011111; PharmGKB:PA166041114	Atorvas Pathwa Pharma Pathwa Pharma Pathwa Pharma Pathwa
Q92887	ATP-binding cassette sub-family C member 2	ABCC2	9606	Homo sapiens	7	5	12	Pitavastatin; Rosuvastatin; atorvastatin; fluvastatin; lovastatin; pravastatin; simvastatin	7	3	PharmGKB:PA145011109; PharmGKB:PA145011110; PharmGKB:PA166041114	Atorvas Pathwa Pharma Pathwa Pharma Pathwa
P08684	Cytochrome P450 3A4	CYP3A4	9606	Homo sapiens	6	5	11	Pitavastatin; Rosuvastatin; atorvastatin; fluvastatin; lovastatin;	7	3	PharmGKB:PA145011109; PharmGKB:PA145011111; PharmGKB:PA166041114	Atorvas Pathwa Pharma Pathwa Pharma

Figure 20: Final Streamlit protein summary.

After the merging is done, missing values are handled appropriately: numeric counts are replaced by zero and textural fields are replaced by empty strings. In addition to this, the workflow computes the total evidence count for each protein as the sum of the interaction-based and pathway-based counts. Moreover, a categorical source tag is implemented indicating whether the protein is interaction-derived, pathway-derived or both. This classification facilitates the interpretation of the relative origin of the protein. Finally, as both independent interactions and pathway tables initially stored compounds in separate fields, both sources are merged into a single compounds field. This unified table is displayed in Figure 20.

Once the integrated protein summary is established, the workflow incorporates the Gene Ontology annotation layer previously generated. For each associated protein, using their UniProt accession code, the workflow calculates the total number of distinct GO terms and merges this value into the protein summary table, together with the concatenated list of all GO identifiers and all GO term names. Additionally, for each aspect and UniProt accession, the corresponding GO identifiers and GO names are aggregated into dedicated fields.

The resulting GO-enriched final summary therefore combines, in a single protein-centred table, all major evidence layers of the workflow: Compound-target interactions; Pathway associated interactions; Standardized protein-level integration; and Gene Ontology functional annotation. By linking each standardized protein to its associated compounds, pathways, evidence counts, and GO terms, this final integrated table represents the most complete biological output of the application. The workflow transforms distributed public medical information into a coherent analytical structure suitable for exploratory interpretation and hypothesis generation for computational drug discovery.

2.2.7 Web Application Implementation

To make the workflow accessible, the whole analysis process is implemented as an interactive application using Streamlit. This environment, provides a graphical wrapper around the computational pipeline and acts as the execution tool through which the user submits the input data, launches the analysis, inspects intermediate outputs, and accesses the final integrated results. The application is designed to accept one or more user-provided SMILES codes as the main and initial input, together with the user's email address required internally for NCBI Entrez queries as seen in Figure 21. Additionally, the user can select which protein taxonomies to include in the analysis (see Figure 22).

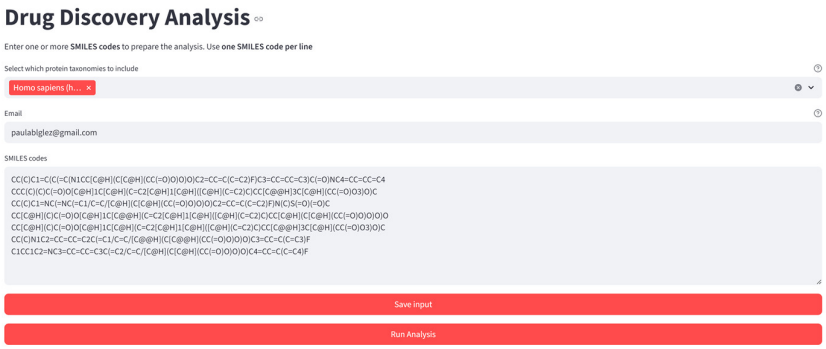


Figure 21: Streamlit's initial stage for analysis where SMILES codes and email are introduced.

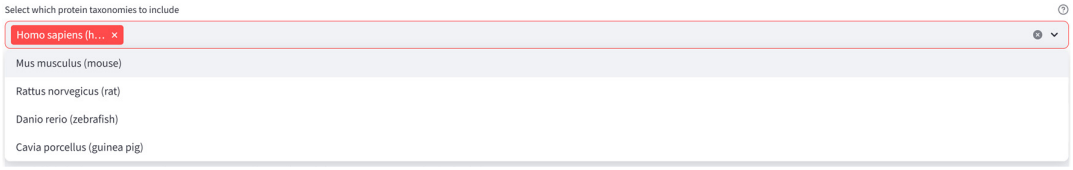


Figure 22: Streamlit's taxonomy selection.

Once the user clicks "Save Input" and the SMILES are successfully introduced and validated by the pipeline as observed in Figure 23 (as explained in the compound identification stage in subsection 2.2.1), they are displayed for the user. After the "Run Analysis" button is pressed, the application runs the pipeline progressively (Figure 24).

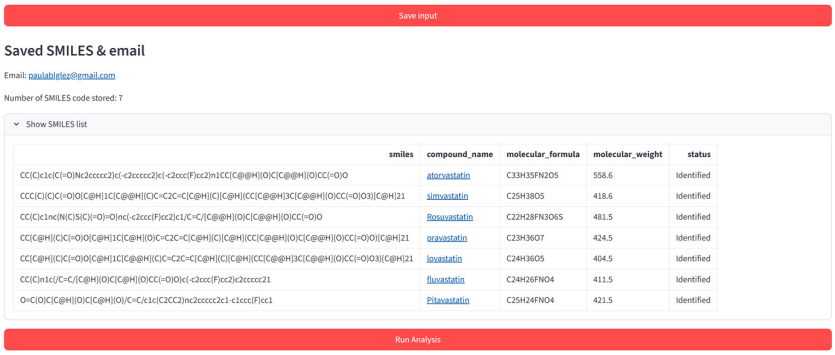


Figure 23: Streamlit's saved SMILES and email verification tables before analysis.

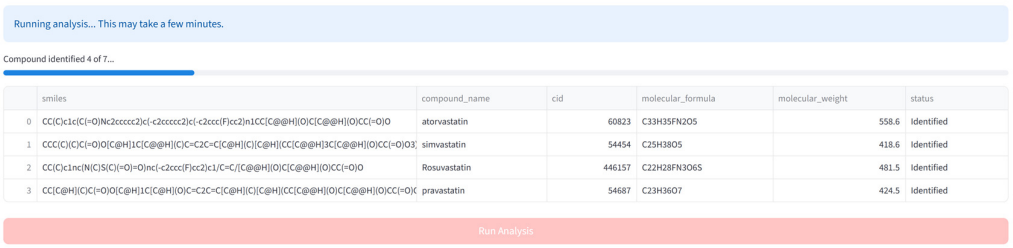


Figure 24: Streamlit's table progressive display during analysis.

As each stage of the application is completed, the user can observe the corresponding partial output in the interface. For instance, after compound identification stage is completed, the application displays the table of recognized compounds together with their PubChem metadata and identification status, when completing the target-interaction retrieval step, the compound-protein interaction summary is shown, etc. This progressive visualization allows the user to follow correctly the analysis and confirm that the pipeline is not stuck but rather taking time to complete each stage of the analysis.

The workflow stores the retrieved and generated information primarily in memory using pandas and DataFrames. This design maintains the exploratory nature of the application, as it supports the interactive inspection of the outputs without requiring to generate/download unnecessary files, or having internal data management to worry about.

Once the pipeline finished, the application displays all the generated tables in a structured manner within the interface. These outputs include the Compound results table, Compound-Target Interaction table, Pathway-associated Protein summary, Grouped Pathway table, Aspect-specific GO tables, Grouped GO summaries, and finally the Final Protein summary table integrating all executed outputs. As observed in Figures 20 and 23, the application also supports interactive navigation through direct links associated with compounds, gene identifiers, UniProt accession codes, pathways and GO terms. Moreover, the application enables users to export the analysis results into a downloadable Excel report to preserve the integrated analysis outside the web interface for further inspection or processing. Therefore, the web application serves both as an execution environment and as an exploratory analysis platform for users to observe and inspect the systematic transformation of chemical input into an integrated biologically interpretable output with valuable insight for future drug research.

3 RESULTS

To validate the proposed workflow, a validation case study based on statins was performed. Statins constitute a well-characterized class of lipid-lowering drugs widely used in the treatment of hypercholesterolemia, a condition defined by elevated circulating cholesterol levels and strongly associated with increased cardiovascular risk. Their shared phenotypic effect results in the reduction of circulating cholesterol levels, particularly low-density lipoprotein cholesterol (LCL-C) [12], [24]. Because their pharmacological context and principal mechanisms of action are well characterized, this makes them a suitable compound group for evaluating whether the application successfully retrieves recurrent targets, pathways, and functional annotations consistent with known pharmacological mechanisms.

Statins act primarily through the inhibition of 3-hydroxy-3-methylglutaryl-coenzyme A reductase (HMG-CoA reductase) by blocking the active site of the enzyme. This enzyme is the rate-limiting enzyme of the mevalonate pathway and central regulator of endogenous cholesterol synthesis. By reducing hepatic or “de novo” cholesterol biosynthesis, statins promote compensatory upregulation of LDL receptor-mediated clearance, and therefore, decrease plasma cholesterol (LDL-C) levels. Besides this canonical mechanism, they are involved in other pleiotropic responses independent of their cholesterol-lowering capacity such as antioxidant and anti-inflammatory responses, anti-fibrotic effects, cardiovascular protective effects, or enhancement of bone formation. This is why it makes them drugs worth evaluating [25], [26], [27], [28].

In this context, the present application is not intended to rediscover statin pharmacology de novo but rather determine whether a compound-centred integration of public biomedical data workflow can recover biologically meaningful recurrent evidence consistent with the known statin response.

To start off, the analysis ran without errors, and the Excel report was ready before going deep into results. Regarding the overview of the results at the beginning of the page (see Figure 25)

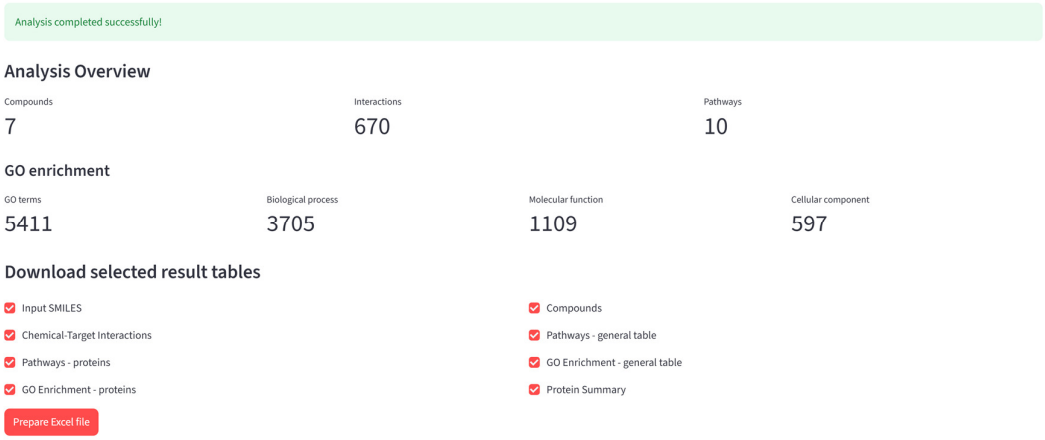
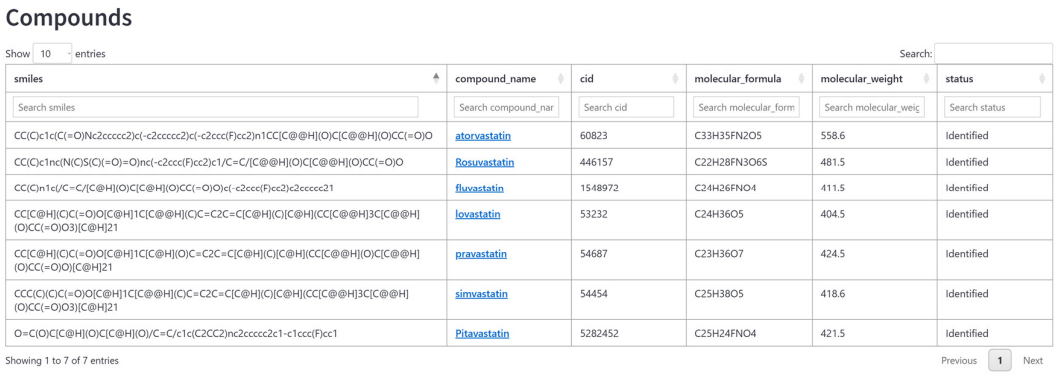


Figure 25: Streamlit's Analysis overview analysis information

It displayed in Figure 26 that the seven statins' SMILES codes introduced were successfully recognized as valid chemical entities. The introduced statins' SMILES codes correspond to the following:

- CC(C)C1=C(C(=C(N1CC[C@H](C[C@H](CC(=O)O)O)O)C2=CC=C(C=C2)F)C3=CC=CC=C3)C(=O)NC4=CC=CC=C4 (Atorvastatin)
- CCC(C)(C)C(=O)O[C@H]1C[C@H](C=C2[C@H]1[C@H]([C@H](C=C2)C)CC[C@@H]3C[C@H](CC(=O)O3)O)C (Simvastatin)
- CC(C)C1=NC(=NC(=C1/C=C/[C@H](C[C@H](CC(=O)O)O)O)C2=CC=C(C=C2)F)N(C)S(=O)(=O)C (Rosuvastatin)
- CC[C@H](C)C(=O)O[C@H]1C[C@@H](C=C2[C@H]1[C@H]([C@H](C=C2)C)CC[C@H](C[C@H](CC(=O)O)O)O)O (Pravastatin)
- CC[C@H](C)C(=O)O[C@H]1C[C@H](C=C2[C@H]1[C@H]([C@H](C=C2)C)CC[C@@H]3C[C@H](CC(=O)O3)O)C (Lovastatin)
- CC(C)N1C2=CC=CC=C2C(=C1/C=C/[C@@H](C[C@@H](CC(=O)O)O)O)C3=CC=C(C=C3)F (Fluvastatin)
- C1CC1C2=NC3=CC=CC=C3C(=C2/C=C/[C@H](C[C@H](CC(=O)O)O)O)C4=CC=C(C=C4)F (Pitavastatin)

This confirmed that the compound identification step worked properly and the downstream analysis was performed on the intended validation set.



Compounds

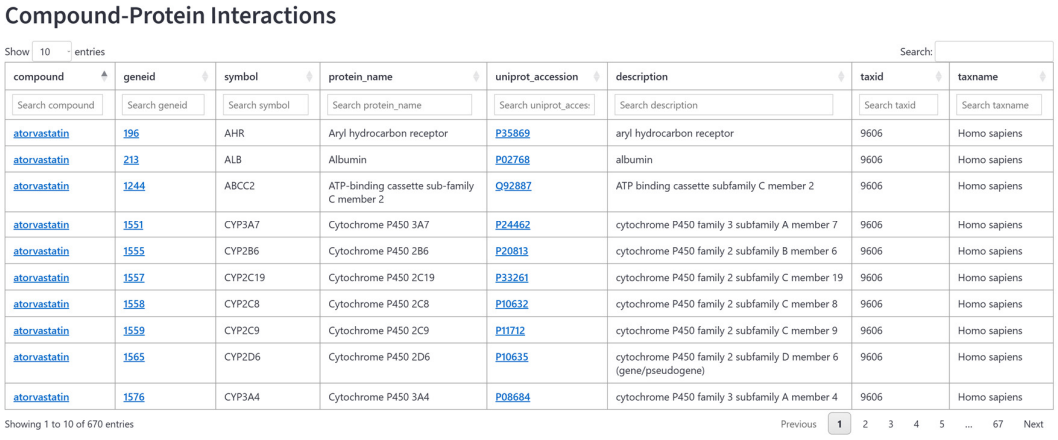
Show 10 entries Search:

smiles	compound_name	cid	molecular_formula	molecular_weight	status
Search smiles	Search compound_nar	Search cid	Search molecular_form	Search molecular_weig	Search status
CC(C)c1c(C(=O)Nc2ccccc2)c(-c2ccccc2)c(-c2ccc(F)cc2)n1CC[C@@H](O)C[C@@H](O)CC(=O)O	atorvastatin	60823	C33H35FN2O5	558.6	Identified
CC(C)c1nc(N(C)S(C)(=O)=O)nc(-c2ccc(F)cc2)c1/C=C/[C@@H](O)C[C@@H](O)CC(=O)O	Rosuvastatin	446157	C22H28FN3O6S	481.5	Identified
CC(C)n1c(/C=C/[C@H](O)C[C@H](O)CC(=O)O)c(-c2ccc(F)cc2)c2ccccc21	fluvastatin	1548972	C24H26FNO4	411.5	Identified
CC[C@H](C)C(=O)O[C@H]1C[C@@H](C)C=C2C=C[C@H](C)[C@H](CC[C@@H]3C[C@@H](O)CC(=O)O3)[C@H]21	lovastatin	53232	C24H36O5	404.5	Identified
CC[C@H](C)C(=O)O[C@H]1C[C@H](O)C=C2C=C[C@H](C)[C@H](CC[C@@H](O)C[C@@H](O)CC(=O)O)[C@H]21	pravastatin	54687	C23H36O7	424.5	Identified
CCC(C)(C)C(=O)O[C@H]1C[C@@H](C)C=C2C=C[C@H](C)[C@H](CC[C@@H]3C[C@@H](O)CC(=O)O3)[C@H]21	simvastatin	54454	C25H38O5	418.6	Identified
O=C(O)C[C@H](O)C[C@H](O)/C=C/c1c(C2CC2)nc2ccccc2c1-c1ccc(F)cc1	Pitavastatin	5282452	C25H24FNO4	421.5	Identified

Showing 1 to 7 of 7 entries Previous 1 Next

Figure 26: Streamlit's Compound table.

After compound recognition, the application retrieved compound-associated proteins (see Figure 27). A total of 670 target interaction records were obtained (corresponding to 469 unique UniProt proteins after removal of duplicated protein entries, for instance, HMGCR was retrieved for all seven compounds but represents a single unique protein) as observed in Figure 25. Taking a deeper look into these targets, the number of retrieved interactions differed substantially between statins: Simvastatin and lovastatin showed the highest number of retrieved associations, whereas pravastatin and pitavastatin showed fewer associations. This uneven distribution does not affect the interpretation of the results as being "better" or "more complex" compounds, instead it reflects the difference in database coverage, literature curation, and the number of computationally annotated associations available for each compound in PubChem.



Compound-Protein Interactions

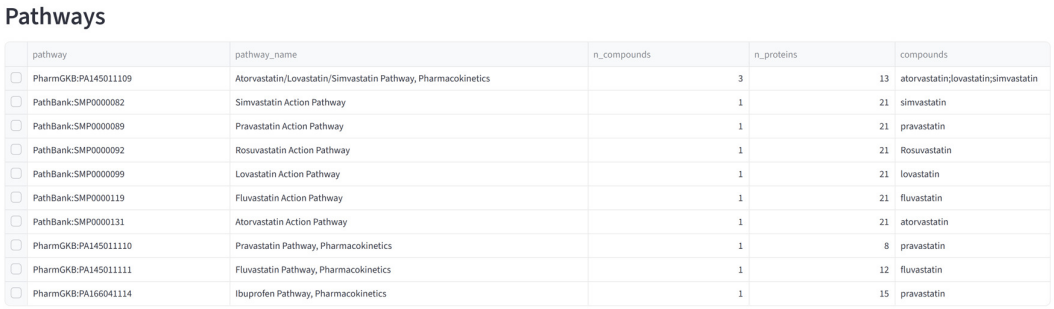
Show 10 entries Search:

compound	geneid	symbol	protein_name	uniprot_accession	description	taxid	taxname
Search compound	Search geneid	Search symbol	Search protein_name	Search uniprot_acces	Search description	Search taxid	Search taxname
atorvastatin	196	AHR	Aryl hydrocarbon receptor	P35869	aryl hydrocarbon receptor	9606	Homo sapiens
atorvastatin	213	ALB	Albumin	P02768	albumin	9606	Homo sapiens
atorvastatin	1244	ABCC2	ATP-binding cassette sub-family C member 2	Q92887	ATP binding cassette subfamily C member 2	9606	Homo sapiens
atorvastatin	1551	CYP3A7	Cytochrome P450 3A7	P24462	cytochrome P450 family 3 subfamily A member 7	9606	Homo sapiens
atorvastatin	1555	CYP2B6	Cytochrome P450 2B6	P20813	cytochrome P450 family 2 subfamily B member 6	9606	Homo sapiens
atorvastatin	1557	CYP2C19	Cytochrome P450 2C19	P33261	cytochrome P450 family 2 subfamily C member 19	9606	Homo sapiens
atorvastatin	1558	CYP2C8	Cytochrome P450 2C8	P10632	cytochrome P450 family 2 subfamily C member 8	9606	Homo sapiens
atorvastatin	1559	CYP2C9	Cytochrome P450 2C9	P11712	cytochrome P450 family 2 subfamily C member 9	9606	Homo sapiens
atorvastatin	1565	CYP2D6	Cytochrome P450 2D6	P10635	cytochrome P450 family 2 subfamily D member 6 (gene/pseudogene)	9606	Homo sapiens
atorvastatin	1576	CYP3A4	Cytochrome P450 3A4	P08684	cytochrome P450 family 3 subfamily A member 4	9606	Homo sapiens

Showing 1 to 10 of 670 entries Previous 1 2 3 4 5 ... 67 Next

Figure 27: Streamlit's Compound-Protein Interactions table.

Regarding the Pathway tables, the results displayed in Figure 28 show 10 retrieved pathway entries in total.

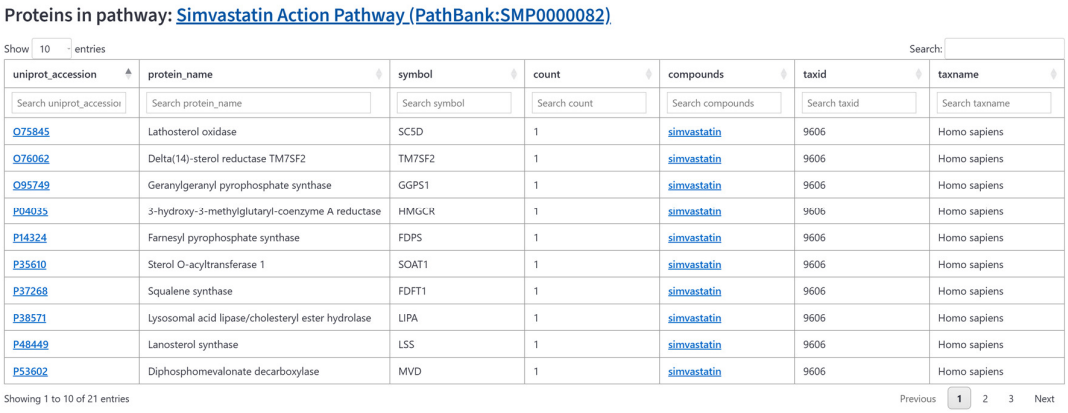


Pathways

pathway	pathway_name	n_compounds	n_proteins	compounds
PharmGKB:PA145011109	Atorvastatin/Lovastatin/Simvastatin Pathway, Pharmacokinetics	3	13	atorvastatin;lovastatin;simvastatin
PathBank:SMP0000082	Simvastatin Action Pathway	1	21	simvastatin
PathBank:SMP0000089	Pravastatin Action Pathway	1	21	pravastatin
PathBank:SMP0000092	Rosuvastatin Action Pathway	1	21	Rosuvastatin
PathBank:SMP0000099	Lovastatin Action Pathway	1	21	lovastatin
PathBank:SMP0000119	Fluvastatin Action Pathway	1	21	fluvastatin
PathBank:SMP0000131	Atorvastatin Action Pathway	1	21	atorvastatin
PharmGKB:PA145011110	Pravastatin Pathway, Pharmacokinetics	1	8	pravastatin
PharmGKB:PA145011111	Fluvastatin Pathway, Pharmacokinetics	1	12	fluvastatin
PharmGKB:PA166041114	Ibuprofen Pathway, Pharmacokinetics	1	15	pravastatin

Figure 28: Streamlit's Pathways Interactions table.

These included six statin-specific action pathways, which contain proteins involved in cholesterol and sterol biosynthesis such as: HMGCR, MVK, PMVK, MVD, FDPS, GGPS1, LIPA, LSS, etc., as observed in Figure 29.



Proteins in pathway: Simvastatin Action Pathway (PathBank:SMP0000082)

Show 10 entries Search:

uniprot_accession	protein_name	symbol	count	compounds	taxid	taxname
O75845	Lathosterol oxidase	SC5D	1	simvastatin	9606	Homo sapiens
O76062	Delta(14)-sterol reductase TM7SF2	TM7SF2	1	simvastatin	9606	Homo sapiens
O95749	Geranylgeranyl pyrophosphate synthase	GGPS1	1	simvastatin	9606	Homo sapiens
P04035	3-hydroxy-3-methylglutaryl-coenzyme A reductase	HMGCR	1	simvastatin	9606	Homo sapiens
P14324	Farnesyl pyrophosphate synthase	FDPS	1	simvastatin	9606	Homo sapiens
P35610	Sterol O-acyltransferase 1	SOAT1	1	simvastatin	9606	Homo sapiens
P37268	Squalene synthase	FDFT1	1	simvastatin	9606	Homo sapiens
P38571	Lysosomal acid lipase/cholesteryl ester hydrolase	LIPA	1	simvastatin	9606	Homo sapiens
P48449	Lanosterol synthase	LSS	1	simvastatin	9606	Homo sapiens
P53602	Diphosphomevalonate decarboxylase	MVD	1	simvastatin	9606	Homo sapiens

Showing 1 to 10 of 21 entries Previous 1 2 3 Next

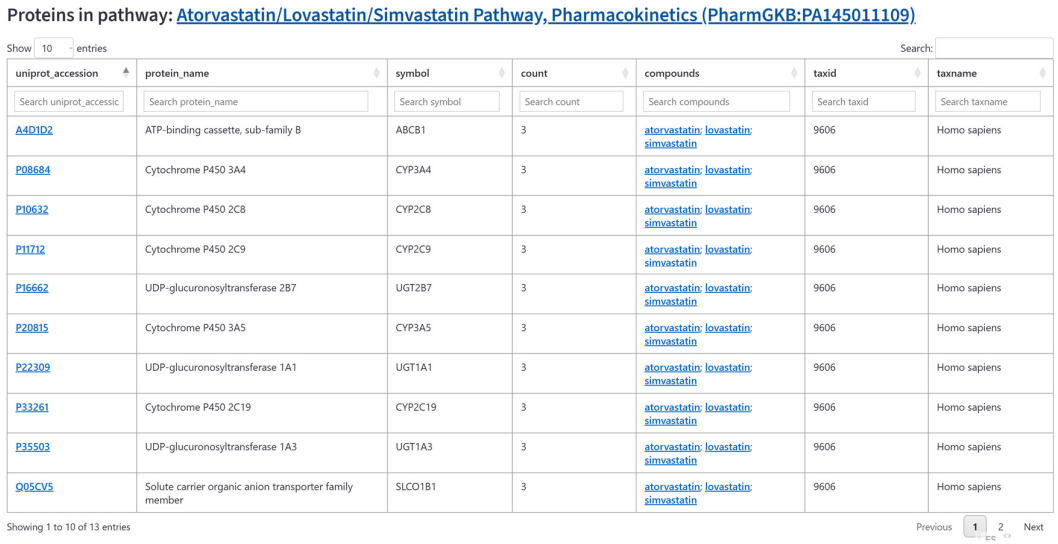
Figure 29: Streamlit's protein table associated with Simvastatin Action Pathway.

This is biologically coherent because statins act in the mevalonate pathway by blocking HMGCR, thus reducing mevalonate production. Mevalonate is not only used to produce cholesterol, but it is also needed to produce isoprenoid intermediates, such as FPP and GGPP. These protein intermediates serve as an important role for posttranslational modifications (called prenylation) of a variety of proteins, including Ras and Rho family GTPases. Members of the Ras and Rho GTPase family are major substrates for post-translational modification by prenylation to move from the cytoplasm

to other membrane compartments. Therefore, statins inhibit both Ras and Rho prenylation, leading to the accumulation of inactive Ras and Rho in the cytoplasm. This helps to understand why statins have a pleiotropic effect. In addition to this, mainly due to the inhibition of the production of prenylated proteins in the cholesterol biosynthesis pathway, statins also influence the cardiovascular system by reducing disease progression. These effects include improved endothelial function (increased vasodilation), anti-inflammatory effects (for atherosclerotic plaque rupture by reducing pro-inflammatory cytokine concentrations), and atherosclerotic plaque stabilization in vessels.

By looking at this from the Pathway section, we could expect GO terms related to not only cholesterol metabolism, but also to gene expression regulation, response to stimulus, inflammation-related pathways or cell proliferation.

Furthermore, additional pharmacokinetic pathways were retrieved as seen in Figure 28. Pharmacokinetics is the branch of pharmacology that studies the interactions between drugs and organisms, so the proteins expected in these pathways include transports, regulators, and enzymes involved in absorption, distribution, metabolism, and excretion (ADME-proteins). The proteins observed include CYP3A4, CYP2C9, CYP2C8, UGT2B7 and CYP3A5. This is biologically meaningful because these proteins are involved in statin metabolism and conjugation processes. Atorvastatin, simvastatin and lovastatin are strongly associated with CPY3A4 and CYP3A5 metabolism, whereas fluvastatin, pitavastatin and rosuvastatin are mainly associated with CYP2C9 and CYP2C8 enzymes. [28], [29]. Other relevant enzymes observed in Figure 30 include ABCB1 and SLCO1B1, which belong to two major transporter families: the solute carrier (SLC) family and the ATP-binding cassette (ABC) family. Both protein families are fundamental for statin uptake into cells and for transport between intracellular and extracellular compartments. Proteins associated with the Atorvastatin/Lovastatin and Simvastatin pathway are observed in Figure 30 and correspond to the protein families mentioned.



Proteins in pathway: Atorvastatin/Lovastatin/Simvastatin Pathway, Pharmacokinetics (PharmGKB:PA145011109)

Show 10 entries Search:

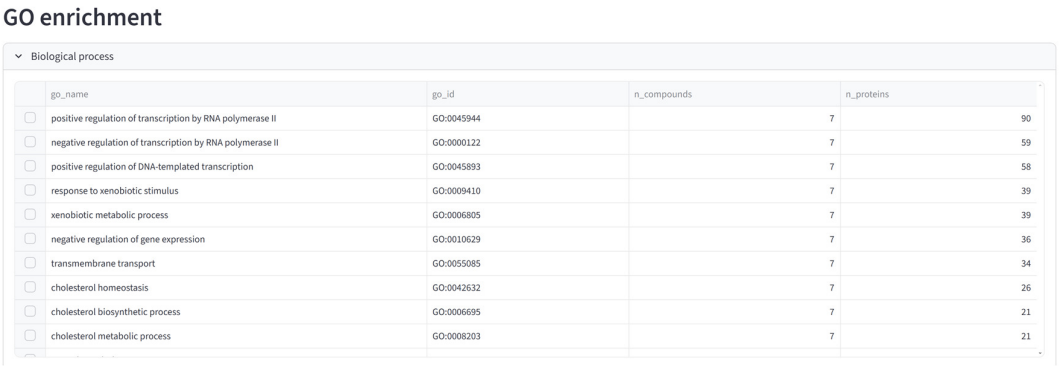
uniprot_accession	protein_name	symbol	count	compounds	taxid	taxname
A4D1D2	ATP-binding cassette, sub-family B	ABCB1	3	atorvastatin; lovastatin; simvastatin	9606	Homo sapiens
P08684	Cytochrome P450 3A4	CYP3A4	3	atorvastatin; lovastatin; simvastatin	9606	Homo sapiens
P10632	Cytochrome P450 2C8	CYP2C8	3	atorvastatin; lovastatin; simvastatin	9606	Homo sapiens
P11712	Cytochrome P450 2C9	CYP2C9	3	atorvastatin; lovastatin; simvastatin	9606	Homo sapiens
P16662	UDP-glucuronosyltransferase 2B7	UGT2B7	3	atorvastatin; lovastatin; simvastatin	9606	Homo sapiens
P20815	Cytochrome P450 3A5	CYP3A5	3	atorvastatin; lovastatin; simvastatin	9606	Homo sapiens
P22309	UDP-glucuronosyltransferase 1A1	UGT1A1	3	atorvastatin; lovastatin; simvastatin	9606	Homo sapiens
P33261	Cytochrome P450 2C19	CYP2C19	3	atorvastatin; lovastatin; simvastatin	9606	Homo sapiens
P35503	UDP-glucuronosyltransferase 1A3	UGT1A3	3	atorvastatin; lovastatin; simvastatin	9606	Homo sapiens
Q05CV5	Solute carrier organic anion transporter family member	SLCO1B1	3	atorvastatin; lovastatin; simvastatin	9606	Homo sapiens

Showing 1 to 10 of 13 entries Previous 1 2 Next

Figure 30: Streamlit's proteins table containing the proteins involved in the Atorvastatin/Lovastatin/Simvastatin Pathway.

However, the presence of the ibuprofen pharmacokinetic pathway (see Figure 28) should be interpreted cautiously. It was probably retrieved due to shared general ADME-related proteins such as CYP enzymes, UGT enzymes and transporters.

Observing the GO terms retrieved, and starting off by the Biological Process aspect terms, the presence of cholesterol-related terms indicated that the analysis was well done.



GO enrichment

Biological process

go_name	go_id	n_compounds	n_proteins
positive regulation of transcription by RNA polymerase II	GO:0045944	7	90
negative regulation of transcription by RNA polymerase II	GO:0000122	7	59
positive regulation of DNA-templated transcription	GO:0045893	7	58
response to xenobiotic stimulus	GO:0009410	7	39
xenobiotic metabolic process	GO:0006805	7	39
negative regulation of gene expression	GO:0010629	7	36
transmembrane transport	GO:0055085	7	34
cholesterol homeostasis	GO:0042632	7	26
cholesterol biosynthetic process	GO:0006695	7	21
cholesterol metabolic process	GO:0008203	7	21

Figure 31: Streamlit's Biological process Gene Ontology table

Terms such as cholesterol biosynthesis or metabolic process are expected and highly relevant, as observed in Figure 31. The xenobiotic-related terms are also appropriated, since statins are administered drugs and thus the recurrence of xenobiotic metabolism, xenobiotic stimulus response and transmembrane transport is expected. Nonetheless, the

top GO terms included both positive and negative transcription by RNA polymerase II, which seemed contradictory. However, these GO terms are broad annotations. Many proteins in the retrieved network are transcription factors, nuclear receptors, inflammatory regulators, etc., for instance: SREBF2, STAT1, NFKB1, PPARA, TP53, etc. When comparing both negative and positive transcription regulation GO terms, 44 proteins overlapped, which means that these terms represent multifunctional regulatory proteins and refer to the more general pleiotropic effects of statins, rather than opposite direct effects of statins.

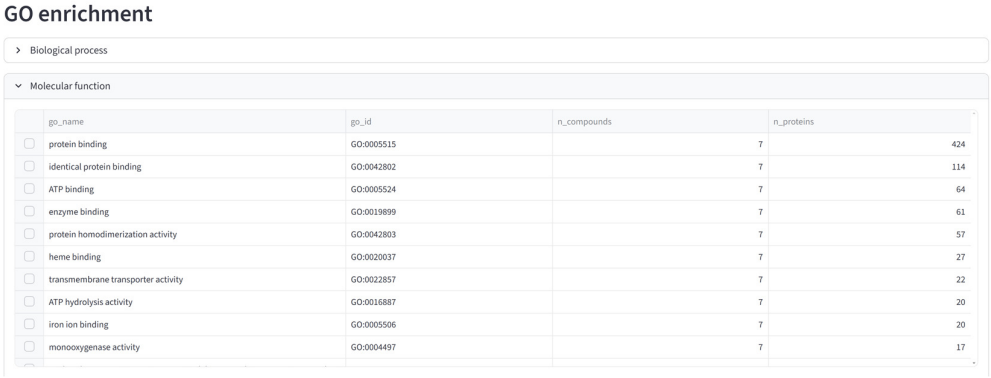
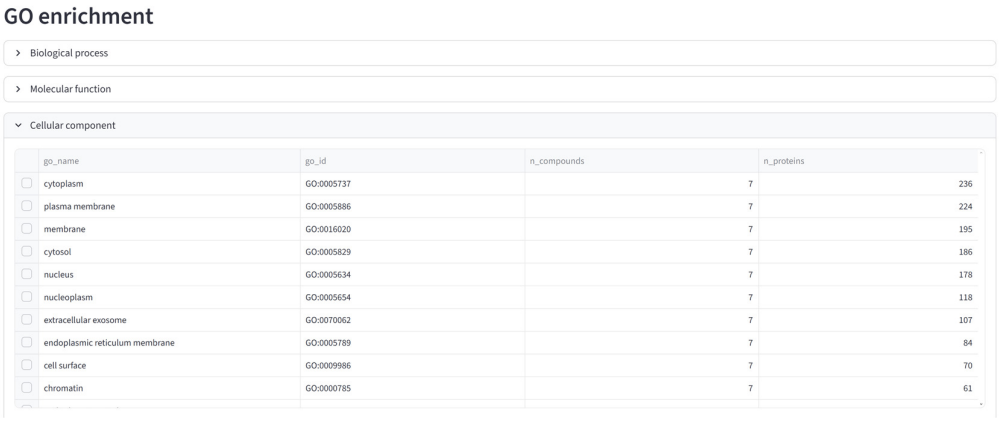


Figure 32: Streamlit's Molecular function Gene Ontology table

Following onto GO Molecular function terms (Figure 32), broad categories such as protein binding, ATP binding and enzyme binding are found as more general functions. Additionally, terms such as transmembrane transporter activity (consistent with SLCO and ABC transporters), xenobiotic transmembrane transporter activity (relevant to drug transport), oxidoreductase activity (consistent with CYPs and sterol-metabolism enzymes), or ATP hydrolysis activity (consistent with ABC transporters), are more specific and related to statin biology and drug handling. Terms CYP-related are biologically meaningful, as cytochrome P450 enzymes are liver enzymes responsible for metabolizing many of the statins introduced.



go_name	go_id	n_compounds	n_proteins
cytoplasm	GO:0005737	7	236
plasma membrane	GO:0005886	7	224
membrane	GO:0016020	7	195
cytosol	GO:0005829	7	186
nucleus	GO:0005634	7	178
nucleoplasm	GO:0005654	7	118
extracellular exosome	GO:0070062	7	107
endoplasmic reticulum membrane	GO:0005789	7	84
cell surface	GO:0009986	7	70
chromatin	GO:0000785	7	61

Figure 33: Streamlit's Cellular component Gene Ontology table.

Thirdly, GO cellular component terms (see Figure 33) were dominated by broad localizations such as cytoplasm, plasma membrane, membrane, cytosol, etc. These terms also make sense biologically, as the presence of transporters such as SLCO1B1, SLCO1B3, SLCO2B1, ABCC2, ABCB1 and ABCG2 is explained. It is important to highlight the term endoplasmic reticulum membrane, because HMGCR and several cholesterol biosynthesis enzymes are associated with the ER membrane. Cellular components related to the plasma membrane are likewise relevant, as they are consistent with the known cardiovascular protective effects of statins, including plaque stabilization, inflammation reduction, and improved endothelial function.

To sum up, the Protein Summary table integrates all interaction, pathway and GO information.

Protein Summary

Show 10 entries Search:

uniprot_accession	protein_name	symbol	taxid	taxname	interaction_count	pathway_count	total_count	compounds	n_compounds	n_pathways	pathways	pathwa
Search uniprot_a	Search protein_name	Search	Sea	Search	Search interactic	Search pathwa	Search tot	Search com	Search n_cor	Search n_p	Search pathways	Search
P04035	3-hydroxy-3-methylglutaryl-coenzyme A reductase	HMGCR	9606	Homo sapiens	7	6	13	Pitavastatin; Rosuvastatin; atorvastatin; fluvastatin; lovastatin; pravastatin; simvastatin	7	6	PathBank:SMP0000082; PathBank:SMP0000089; PathBank:SMP0000092; PathBank:SMP0000099; PathBank:SMP0000119; PathBank:SMP0000131	Atorvas Pathwa Pathwa Pathwa Pathwa Pathwa Pathwa
Q9NPD5	Solute carrier organic anion transporter family member 1B3	SLCO1B3	9606	Homo sapiens	7	6	13	Pitavastatin; Rosuvastatin; atorvastatin; fluvastatin; lovastatin; pravastatin; simvastatin	7	4	PharmGKB:PA145011109; PharmGKB:PA145011110; PharmGKB:PA145011111; PharmGKB:PA166041114	Atorvas Pathwa Pharma Pathwa Pharma Pathwa Pharma Pathwa
Q92887	ATP-binding cassette sub-family C member 2	ABCC2	9606	Homo sapiens	7	5	12	Pitavastatin; Rosuvastatin; atorvastatin; fluvastatin; lovastatin; pravastatin; simvastatin	7	3	PharmGKB:PA145011109; PharmGKB:PA145011110; PharmGKB:PA166041114	Atorvas Pathwa Pharma Pathwa Pharma Pathwa
P08684	Cytochrome P450 3A4	CYP3A4	9606	Homo sapiens	6	5	11	Pitavastatin; Rosuvastatin; atorvastatin; fluvastatin; lovastatin; pravastatin; simvastatin	7	3	PharmGKB:PA145011109; PharmGKB:PA145011111; PharmGKB:PA166041114	Atorvas Pathwa Pharma Pathwa Pharma Pathwa

Figure 34: Streamlit's final Protein Summary table.

From Figure 34, HMGCR, SLCO1B3, ABCC2 and CYP3A4 appear as the most recurrent proteins across the analysed statins. This pattern strongly aligns with the biological validity of the workflow, as it captures the core pharmacological and pharmacokinetic context of statins in humans. Overall, the most relevant proteins identified are:

- HMGCR
- SLCO transporters (SLCO1B1, SLCO1B3, SLCO2B1)
- ABC transporters (ABCC2, ABCG2)
- Cytochrome P450 enzymes (CYP2C9, CYP3A4, CYP3A5)

Among these, the most convincing proteins are those supported simultaneously by direct compound-protein interactions pathway level evidence. Together, these proteins represent the most biologically robust findings of the analysis, as they are consistent with the known pharmacological targets and key proteins involved in statin transport and metabolism.

4 DISCUSSION

The aim of this section is to evaluate the performance of the proposed compound-centred workflow for the identification of biologically meaningful insight of compounds with known phenotypic effects. The application is meant to determine whether public medical data could be integrated to recover recurrent proteins, pathway, and functional annotations consistent with known pharmacology.

The statin validation case study confirms that the application can recognize all seven input compounds as valid statins. Therefore, from the very beginning, the analysis can be performed with confidence.

One of the strongest validation outcomes is the recurrent identification of HMGCR across all seven compounds. This target is highly relevant because HMGCR encodes HMG-CoA reductase, which is the widely known pharmacological statin target and the rate-limiting enzyme of the mevalonate pathway. Furthermore, the retrieval of this protein ensures that the workflow is able to collect the central mechanisms responsible for the known phenotype of the compound group. In addition to the main target, the application retrieves several proteins involved in cholesterol biosynthesis and sterol metabolism such as: MVK, PMVK, MVD, IDI1, FDPS, GGPS1, LSS, SQLE and FDFT1. These proteins are involved in the mevalonate pathway, which is biologically coherent because the statins introduced act upstream in the mevalonate pathway. Hence, this is fundamental to understand that HMGCR does not only affect cholesterol formation but also influences other intermediates. By looking at the results, isoprenoid intermediates such as FPP and GGPP which are required for the protein prenylation and are involved in the regulation of small GTPases, provide a mechanistic explanation for why statins are associated with both, cholesterol-associated responses and broader pleiotropic responses. These responses include changes in inflammatory signalling, oxidative stress, endothelial function, cytoskeletal organization, and gene-expression regulation [29], [30].

The pathway results further support the biological validity of the application. The workflow recovered statin action pathways and pharmacokinetic pathways related to individual compounds. Regarding the individual action pathways, they reconstructed the

expected relationship between HMGCR inhibition, mevalonate metabolism and cholesterol biosynthesis. On the other hand, pharmacokinetic proteins retrieved include hepatic uptake transporters, efflux transporters, P450 enzymes and UGT enzymes (SLCO1B1, SLCO1B3, SLCO2B1, ABCC2, CYP3A4, CYP2C9, CYP2C8, UGT1A1, UGT1A3 and UGT2B7). These broader proteins are still associated with statin mechanisms of action, as they are known to be involved in drug absorption, distribution, metabolism and excretion processes. This distinction between both pathway groups retrieval is important because both types of information contribute to complete the biological behaviour of a compound.

The Gene Ontology term results were also quite consistent with statin literature. Firstly, Biological Process section displayed recurrent GO terms such as cholesterol biosynthetic process, sterol metabolic process, xenobiotic metabolic process and transmembrane transport, which matches the expected functional profile of statins and pharmacological mechanisms. However, some GO terms seem contradictory and less specific than the cholesterol and xenobiotic related terms. For instance, positive regulation of transcription by RNA polymerase II and negative regulation of transcription by RNA polymerase II appeared among the recurrent biological processes. These terms should be taken with caution, as they should be interpreted as broader terms. They reflect the presence of multifunctional regulatory proteins. This interpretation is biologically coherent, as the inhibition of the mevalonate pathway can reduce isoprenoid production and thereby affect prenylation-dependent signalling proteins including Ras, Rho and Rac. These signalling changes may indirectly influence transcriptional regulation and cellular adaptation [29]. Secondly, Molecular Function terms mainly reflect activities related to drug handling and metabolism, transporter activity, ATP-binding transporter activity and oxidoreductase activity. These annotations are consistent with the previously retrieved SLCO and ABC transporters, CYP enzymes and cholesterol-biosynthesis proteins. Thirdly, Cellular Component terms are broader, but they also fit the expected biology: membrane and plasma membrane annotations correspond to drug uptake and efflux transporters, while endoplasmic reticulum annotations correspond to cholesterol biosynthesis mechanisms. Nuclear and cytoplasmic terms likely arise due to signalling proteins linked to downstream regulatory responses.

5 CONCLUSIONS

This project presents the design and development of an automated compound-centred computational workflow integrating open-access biomedical information, aimed at supporting the biological and mechanistical interpretation of compounds with known phenotypic effects. Starting from user-provided SMILES codes, the application retrieves, standardizes, integrates and summarizes heterogeneous information distributed across public biomedical resources. The workflow accesses to resources such as PubChem, NCBI Gene, UniProtKB, and QuickGO, from which compound metadata, compound-target interactions, pathway interactions, pathway-associated proteins, and Gene Ontology annotations information is retrieved. By combining these layers of evidence into a single exploratory environment, the workflow addresses an important bioinformatics challenge in modern drug discovery: the difficulty of transforming dispersed data into coherent and biologically interpretable knowledge.

The validation case study based on statins. Since statins constitute a well-characterized drug class with a shared phenotype effect, they are appropriate for evaluating whether the pipeline retrieves recurrent targets, pathways or functional annotations corresponding to known pharmacological data. The obtained results in this work indicate that the workflow can recover meaningful biological insight and integrate them into a unified protein summary, thereby serving as a tool for hypothesis generation and early-stage compound interpretation. However, the retrieved associations should be interpreted with caution as the pipeline depends on the availability, quality and completeness of stored public biomedical databases.

Nonetheless, several limitations are identified that could be upgraded in the future. First, not every retrieved pathway is directly associated to statin effects, for example, the appearance of a non-statin pharmacokinetic pathway (ibuprofen-related pathway) reflects shared ADME proteins rather than a true statin-specific mechanism. Therefore, pathway results may require a filtering process for better specific results. Second, the workflow identifies biological associations but does not prove direct causality. For instance, HMGCR can be considered as a direct statin target, but many transcription factors, inflammatory proteins or signalling mediators should be interpreted

as downstream or associated proteins. The application is most useful as hypothesis-generation and biological interpretation tool or even a complementary tool for other computational drug discovery methods, rather than a standalone method for confirming mechanisms experimentally.

As future improvement, one of the main objectives is to deploy the application web as a functional web on CNB machines. This would allow the project to be further developed and move from a design and validation environment to a real usage setting, as well as enable other researchers to use the workflow in a real institutional environment and aid in computational drug research.

In conclusion, the proposed application is thereby, best understood as a bioinformatics support tool that facilitates biological interpretation and hypothesis generation rather than as a predictive or experimentally validated discovery platform solely. By linking chemical input to targets, pathways and functional annotations into a single web environment, the developed application provides a support framework for the interpretation of phenotypically relevant compounds and for future research in computational drug discovery.

6 REFERENCES

[1] C. Méndez, "Bioinformatics and Big Data: Transforming Biomedical Research," *Journal of Biomedical SYstems & Emerging Technologies*, vol. 11, no. 06, p. 226, 2024, doi: 10.37421/2952-8526.2024.11.226.

[2] X. Yang, K. Huang, D. Yang, W. Zhao, and X. Zhou, "Biomedical Big Data Technologies, Applications, and Challenges for Precision Medicine: A Review," *Global Challenges*, vol. 8, Nov. 2023, doi: 10.1002/gch2.202300163.

[3] J. Ratnam *et al.*, "The application of the Open Pharmacological Concepts Triple Store (Open PHACTS) to support drug discovery research," *PLoS One*, vol. 9, no. 12, Dec. 2014, doi: 10.1371/journal.pone.0115460.

[4] X. Xia, "Bioinformatics and Drug Discovery," *Curr. Top. Med. Chem.*, vol. 17, no. 15, pp. 1709–1726, Apr. 2017, doi: 10.2174/1568026617666161116143440.

[5] N. Wang and Q. He, "Artificial Intelligence and Bioinformatics Applications in Precision Medicine and Future Implications," in *Comprehensive Precision Medicine*, vol. 1, Kenneth S. Ramos, Ed., Elsevier, 2024, ch. 1.02, pp. 9–24. doi: 10.1016/B978-0-12-824010-6.00058-7.

[6] R. E. Hughes, R. J. R. Elliott, J. C. Dawson, and N. O. Carragher, "High-content phenotypic and pathway profiling to advance drug discovery in diseases of unmet need," *Cell Chem. Biol.*, vol. 28, no. 3, pp. 338–355, Mar. 2021, doi: 10.1016/j.chembiol.2021.02.015.

[7] V. Kumar *et al.*, "Target-based drug discovery: Applications of fluorescence techniques in high throughput and fragment-based screening," *Heliyon*, vol. 10, no. 1, Jan. 2024, doi: 10.1016/j.heliyon.2023.e23864.

[8] S. Ou-Yang, J. Lu, X. Kong, Z. Liang, C. Luo, and H. Jiang, "Computational drug discovery," *Acta Pharmacol. Sin.*, vol. 33, no. 9, pp. 1131–1140, Sep. 2012, doi: 10.1038/aps.2012.109.

[9] F. Vincent, A. Nueda, J. Lee, M. Schenone, M. Prunotto, and M. Mercola, "Phenotypic drug discovery: recent successes, lessons learned and new directions," *Nat. Rev. Drug Discov.*, vol. 21, no. 12, pp. 899–914, Dec. 2022, doi: 10.1038/s41573-022-00472-w.

[10] G. C. Terstappen, C. Schlüpen, R. Raggiaschi, and G. Gaviraghi, "Target deconvolution strategies in drug discovery," *Nat. Rev. Drug Discov.*, vol. 6, no. 11, pp. 891–903, Nov. 2007, doi: 10.1038/nrd2410.

[11] Y. Pan, T. Cheng, Y. Wang, and S. H. Bryant, "Pathway analysis for drug repositioning based on public database mining," *J. Chem. Inf. Model.*, vol. 54, no. 2, pp. 407–418, Feb. 2014, doi: 10.1021/ci4005354.

[12] A. Hussain, J. Kaler, and S. D. Ray, "The Benefits Outweigh the Risks of Treating Hypercholesterolemia: The Statin Dilemma.," *Cureus*, vol. 15, no. 1, p. e33648, Jan. 2023, doi: 10.7759/cureus.33648.

[13] Y. Wang *et al.*, "PubChem BioAssay: 2017 update," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D955–D963, Jan. 2017, doi: 10.1093/nar/gkw1118.

[14] S. Kim, P. A. Thiessen, T. Cheng, B. Yu, and E. E. Bolton, "An update on PUG-REST: RESTful interface for programmatic access to PubChem," *Nucleic Acids Res.*, vol. 46, no. W1, pp. W563–W570, Jul. 2018, doi: 10.1093/nar/gky294.

[15] S. Kim, P. A. Thiessen, E. E. Bolton, and S. H. Bryant, "PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W605–W611, Jul. 2015, doi: 10.1093/nar/gkv396.

[16] A. P. Bento *et al.*, "An open source chemical structure curation pipeline using RDKit," *J. Cheminform.*, vol. 12, no. 1, p. 51, Dec. 2020, doi: 10.1186/s13321-020-00456-1.

[17] P. J. A. Cock *et al.*, "Biopython: freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, Jun. 2009, doi: 10.1093/bioinformatics/btp163.

[18] National Center for Biotechnology Information, "Entrez® Programming Utilities Help," in *Entrez® Programming Utilities Help*, Bethesda (MD), 2010. Accessed: Jun. 08, 2026. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK25501/

[19] A. Bateman *et al.*, "UniProt: the Universal Protein Knowledgebase in 2025," *Nucleic Acids Res.*, vol. 53, no. D1, pp. D609–D617, Jan. 2025, doi: 10.1093/nar/gkae1010.

[20] D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'Donovan, and R. Apweiler, "QuickGO: A web-based tool for Gene Ontology searching," *Bioinformatics*, vol. 25, no. 22, pp. 3045–3046, Nov. 2009, doi: 10.1093/bioinformatics/btp536.

[21] F. Nadim Iqbal, "A BRIEF INTRODUCTION TO APPLICATION PROGRAMMING INTERFACE (API)," Red Hat, Inc, Nov. 2023. doi: 10.5281/zenodo.10198423.

[22] V. F. Scalfani, S. C. Ralph, A. Al Alshaikh, and J. E. Bara, "PROGRAMMATIC COMPILATION OF CHEMICAL DATA AND LITERATURE FROM PUBCHEM® USING MATLAB®," *Chem. Eng. Educ.*, vol. 54, no. 4, pp. 230–241, 2020.

[23] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, Feb. 1988, doi: 10.1021/ci00057a005.

[24] R. Chou *et al.*, "Statin Use for the Primary Prevention of Cardiovascular Disease in Adults," *JAMA*, vol. 328, no. 8, p. 754, Aug. 2022, doi: 10.1001/jama.2022.12138.

[25] M. Ahmadi *et al.*, "Pleiotropic effects of statins: A focus on cancer," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1866, no. 12, p. 165968, Dec. 2020, doi: 10.1016/j.bbadis.2020.165968.

[26] Q. Zhou and J. K. Liao, "Pleiotropic Effects of Statins - Basic Research and Clinical Perspectives -," *Circulation Journal*, vol. 74, no. 5, pp. 818–826, 2010, doi: 10.1253/circj.CJ-10-0110.

[27] C. A. German and J. K. Liao, "Understanding the molecular mechanisms of statin pleiotropic effects," *Arch. Toxicol.*, vol. 97, no. 6, pp. 1529–1545, Jun. 2023, doi: 10.1007/s00204-023-03492-6.

[28] Z. Shi and S. Han, "Personalized statin therapy: Targeting metabolic processes to modulate the therapeutic and adverse effects of statins," *Heliyon*, vol. 11, no. 1, p. e41629, Jan. 2025, doi: 10.1016/j.heliyon.2025.e41629.

[29] K. K. Patel, V. S. Sehgal, and K. Kashfi, "Molecular targets of statins and their potential side effects: Not all the glitter is gold," *Eur. J. Pharmacol.*, vol. 922, p. 174906, May 2022, doi: 10.1016/j.ejphar.2022.174906.

[30] J. K. Liao and U. Laufs, "PLEIOTROPIC EFFECTS OF STATINS," *Annu. Rev. Pharmacol. Toxicol.*, vol. 45, no. 1, pp. 89–118, Sep. 2005, doi: 10.1146/annurev.pharmtox.45.120403.095748.