O-ST 009

การประชุมเสนอผลงานวิจัยระดับชาติ มหาวิทยาลัยสุโขทัยธรรมาธิราช ครั้งที่ 14

The 14th STOU National Research Conference

Development of a Data Processing Framework to Construct a Bioactivity-focused Dataset from ChEMBL Database for Drug Discovery

Pitchapong Chouchanakij[1] Natapol Pornputtapong[2]

Abstract

Despite Thailand's rich biodiversity and potential for drug discovery, it has made limited progress in fully utilizing its natural resources for pharmaceutical development. To address this gap, this study focused on curating and refining bioactivity data from the ChEMBL database, aiming to create a dataset that supports drug discovery efforts. The ChEMBL database contains over 20 million bioactivity records from various laboratories worldwide; however, the high volume and complexity of the data present significant challenges. We developed an efficient data processing framework that systematically filtered, cleaned, and curated the dataset, ultimately producing approximately 1.9 million bioactivity records related to small molecules and their interactions with human single protein targets. The study resulted in two distinct datasets: the Original Bioactivity Dataset, which preserves experimentally derived data, and the Processed Bioactivity Dataset, which uses pChEMBL values to infer the bioactivity of compounds lacking explicit classification. This curated dataset serves as a foundational resource for future drug discovery research, particularly for integrating factors related to natural products. The resulting datasets provide flexibility to researchers, enabling both immediate and long-term contributions to pharmaceutical research.

Keywords: Drug discovery, Bioactivity, Data processing framework, Data curation, ChEMBL database

[1] Student of Pharmaceutical Sciences and Technology Program, Faculty of Pharmaceutical Sciences, Chulalongkorn University 6370005333@student.chula.ac.th

[2] Assistant Professor in Department of Biochemistry and Microbiology, Faculty of Pharmaceutical Sciences, Chulalongkorn University natapol.p@pharm.chula.ac.th

Introduction

One of the fundamental processes in advancing healthcare is drug discovery, leading to the development of new medications that can treat diseases and improve quality of life worldwide. In Thailand, the importance of drug discovery is significant because of specific health challenges and the need for accessible and effective treatments. However, the National Drug Policies in Thailand require further support and participation from various sectors such as research institutes, universities, and stakeholders, along with additional resources for effective implementation (Jitruknatee et al., 2020).

Natural products have historically played a critical role in drug discovery (Atanasov et al., 2021), Meanwhile, Thailand possesses immense potential for natural product discovery, featuring rich biodiversity and a long-standing tradition of herbal medicine (Kittakoop, 2022). A nation's diverse ecosystems are home to an extensive variety of plant and marine species, many of which remain unexplored for their therapeutic potential.

Drug discovery is driven by an understanding of bioactivity, which simply refers to how drug molecules interact with biological targets, often human proteins (Kim et al., 2022). These interactions are critical for therapeutic interventions, as they determine whether a compound can modulate the function of a protein to treat diseases. Identifying compounds with desired bioactivity is a key step in the development of effective drugs.

Despite the rich biodiversity in Thailand and its great potential for natural product-based drug discovery, the country has struggled to fully utilize these resources for pharmaceutical development. Only a small number of herbal medicines have been officially approved by the Thai Food and Drug Administration (FDA), highlighting the underutilization of the country's enormous natural wealth in drug discovery (Kwankhao et al., 2020). This deficit emphasizes the need for more efficient approaches to harness natural products for the development of new therapeutic agents.

In recent years, technological advancements have provided researchers with access to vast biological data resources, which can accelerate drug discovery. The ChEMBL database, maintained by the European Bioinformatics Institute, is one of the most comprehensive bioactivity databases available. It contains over 20 million records of biological activities, capturing interactions between small molecules and their biological targets, such as proteins, enzymes, and receptors (Zdrazil et al., 2023). ChEMBL’s data is manually curated from experimental results published by laboratories worldwide, ensuring reliability and consistency across a wide range of assays and biological contexts. The database includes crucial information such as molecular identifiers, assay types, bioactivity measurements, and organism taxonomy, making it a valuable tool for drug discovery. In theory, this abundance of information can significantly aid drug discovery (Zhu et al., 2022).

However, the vast volume and complexity of the data pose significant challenges. The biological data records from ChEMBL may not always align with the specific needs of targeted drug discovery projects. The size and diversity of the dataset often overwhelm researchers, making it difficult to extract the most relevant information for advancing drug development. As a result, several research groups have developed specialized frameworks designed to specific tasks. For example, a study on SARS coronavirus retrieved bioactivity data on 3C-like protease inhibitors from ChEMBL, which is a key drug target in the coronavirus genome (Ishola et al., 2021). Another study curated drug safety data focusing on toxicity and warnings to aid in safety-related drug discovery efforts (Hunter et al., 2021). Additionally, a gene ontology-based tool was developed using ChEMBL to help researchers navigate protein-ligand target spaces, linking biological processes with drug information, which is valuable for drug-disease pathway studies (Mutowo et al., 2016).

Recent research emphasizes the growing potential of plant-based natural products and their analogues in modern therapeutic development, underscoring the importance of natural product research (Najmi et al., 2022). As a smaller research country, Thailand faces barriers in conducting large-scale drug discovery due to limited resources. In response to these challenges, this study aimed to develop an efficient data processing framework to curate and refine the massive bioactivity data from ChEMBL. Access to global databases like ChEMBL is crucial for bridging this gap, as it provides a wealth of diverse, high-quality bioactivity data that would be impossible to generate locally. For demonstration, chemical compounds derived from natural sources from different geographic regions may exhibit similar bioactivities due to shared biological mechanisms. Having access to a worldwide dataset allows researchers to compare and validate their findings with global trends, leading to more robust conclusions. This comprehensive perspective is crucial for identifying broader patterns in bioactivity, which can accelerate the development of new drugs from natural products in Thailand and elsewhere. This framework will allow researchers to utilize bioactivity data more effectively, improving the overall efficiency of the drug discovery process. The resulting dataset is intended to serve as a foundational step toward developing a larger, integrated database that incorporates additional factors critical for a comprehensive understanding of drug discovery for future studies.

Purpose

1. To develop a data processing framework that curates, cleans, and refines extensive bioactivity data from the ChEMBL database.
2. To create a manageable and relevant bioactivity-focused dataset that streamlines drug discovery projects.

Research Methodology

The field of drug discovery increasingly relies on vast datasets, often sourced from multiple studies and laboratories around the world. The ability to process ChEMBL database effectively is essential for transforming raw, unstructured information into actionable insights that can accelerate the identification of therapeutic compounds. In this study, the data processing framework was developed specifically to address the complexities inherent in bioactivity data. The methodology developed in this project was designed to systematically process and curate bioactivity data from the ChEMBL database for drug discovery. This process involves several critical steps, as illustrated in Figure 1.

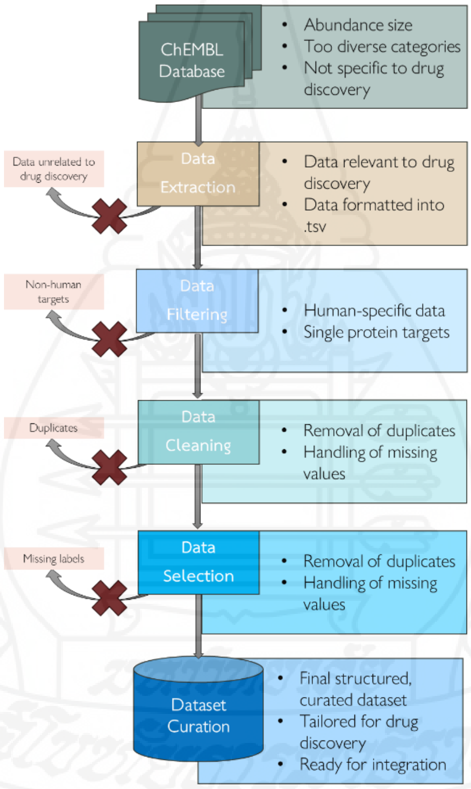

Figure 1 Workflow for Data Processing Framework

To accomplish this, several key steps were employed in the data processing pipeline: data extraction, filtering, cleaning, selection, and curation. Each of these steps was designed to address specific challenges associated with large, heterogeneous datasets, ensuring that the resulting dataset was both robust and optimized for research use.

1. Overview of the ChEMBL Database

The ChEMBL database is widely recognized in cheminformatics, bioinformatics, and drug discovery. It has been maintained by the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) since 2009. ChEMBL contains a diverse range of information, including molecular properties, bioassay results, pharmacokinetics, and target details.

All the data in ChEMBL has been manually curated by the ChEMBL curator team, who gather bioactivity data from a wide variety of sources, including experimental results and journal articles published by laboratories worldwide. These data were carefully processed and validated according to strict curation criteria, ensuring reliability and consistency across the entire database.

The version used in this study was “ChEMBLdb, Current Release: 34, Last Update: March 2024.”

2. Data Extraction

While ChEMBL database is full of valuable data, not all of them is directly relevant to the specific focus of drug discovery projects. Many categories, such as pharmacokinetics or non-human bioassays, fall outside the scope of our project. The first step of this research was to extract only bioactivity data records from the ChEMBL database.

ChEMBL contains over 20 million records of biological activities, also referred to as bioactivity records. The bioactivity records contain a range of information, including molecular identifiers, assay descriptions, bioactivity measurements, and taxonomic details. These bioactivity records capture interactions between small molecules and their specific biological targets, such as proteins, enzymes, and receptors. This massive scale required a computational approach to handle and process the data efficiently.

For this study, we chose to use tab-separated values (.tsv) format for the extracted data. The .tsv format offers several advantages over other formats. The structure of .tsv files ensures that data fields were separated by tabs rather than commas, which reduces the risk of conflicts when data values contain commas, such as molecule names or assay descriptions. Additionally, .tsv files provide better readability for datasets containing textual or numerical information with complex structures, making it easier to parse and manipulate data during further stages of processing.

To accomplish this, Python a programming language widely used in scientific research, was utilized. The following packages were employed to automate data extraction and ensure precision:

Pandas: Essential for data manipulation, Pandas allows for efficient data cleaning, filtering, and restructuring of large datasets into a manageable format.

NumPy: Used for performing numerical calculations on bioactivity metrics, including array-based operations that facilitate mathematical processing.

Python scripts were written to automatically retrieve the necessary records from ChEMBL, ensuring accuracy and reproducibility throughout the extraction process.

The specific fields of data extracted for this study were detailed in the Results section.

3. Data Filtering

Specific filtering criteria were applied to narrow the focus to data relevant to drug discovery. The ChEMBL dataset contains millions of records representing bioactivity data across a wide range of biological systems and organisms. Given the vast size and complexity of this dataset, a critical step in this methodology was to filter the data to focus on records most relevant to drug discovery, specifically targeting human proteins. This focus was crucial because most modern drug discovery efforts involve small molecules designed to modulate the activity of single proteins, which are often critical components of disease pathways (Gao & Skolnick, 2013; Tabana et al., 2023).

To refine the dataset for drug discovery purposes, we applied specific filtering criteria designed to reduce the data to only the most pertinent entries. The key filtering steps involved:

Protein Targets: Since this research focuses on drug discovery for human, we filtered the dataset to retain only bioactivity records where the Target Type was a single protein. Single protein targets are often of primary interest in drug discovery as they represent specific biological molecules, such as enzymes or receptors, that drugs can interact with to produce a therapeutic effect.

Organism Filtering: To focus on human biology, the dataset was further filtered to retain only bioactivity records where the Target Organism was *Homo sapiens.* Human proteins are the most relevant targets for pharmaceutical research, and filtering out non-human targets ensures that the dataset was aligned with the goal of identifying therapeutically relevant interactions for human diseases.

This filtering process was automated using Python programming to guarantee precision and reproducibility.

4. Data Cleaning

After the filtering process, the dataset was cleaned to ensure consistency and data quality. Duplicate records were identified and only one record was left to avoid redundancy. Entries with missing critical fields, such as bioactivity comments, were marked and separated into the missing value group to maintain the reliability of the dataset without losing information. In addition, bioactivity values were standardized to ensure that all units were consistent across the dataset, enabling comparability of the data.

The entire cleaning process was conducted via Python on .tsv files, a format that offers several advantages for handling large datasets. Pandas, a well-known Python library was utilized to handle the data. Pandas is well-optimized for manipulating tabular data, making it easy to load, process, and save cleaned datasets efficiently. Key cleaning steps included:

Duplicate Removal: Duplicate records were identified and removed to avoid redundancy. Only one unique record was retained for each bioactivity entry to maintain data integrity.

Handling Missing Values: Critical fields, such as bioactivity comments and molecular identifiers, were assessed for missing data. For example, missing values in the 'Comment' field, which notes the activity status “active/inactive” of bioactivity records, were flagged for further handling during the Data Selection step.

5. Data Selection

This step was crucial for ensuring that the dataset was both relevant and applicable to drug discovery projects, particularly those aimed at identifying promising compounds for therapeutic development. In this step, we choose and justify the data based on the ideal criteria focused on “Active” or “Inactive” of bioactivities for drug discovery tasks.

Since the bioactivity measurements were reported in various units across different studies. To ensure consistency, these measurements were standardized using pChEMBL values.

$$\text{pChEMBL} = -\log_{10}(\text{Molar concentration})$$

Compounds with a pChEMBL value ≥ 5 were classified as active.

Compounds with a pChEMBL value < 5 were classified as inactive.

pChEMBL value is a transformed bioactivity metric that allows for standardized comparisons across various assays created by the ChEMBL database. It is derived from negative logarithms of bioactivity measurements like IC50, EC50, and Ki, enabling a uniform scale similar to the pH scale. This standardization helps compare compound potencies and activities across diverse experimental setups. For example, a pChEMBL value of 6 corresponds to a bioactivity concentration of 1 μM, whereas a value of 9 corresponds to 1 nM. Using pChEMBL ensures consistency across the dataset, allowing for more precise and meaningful.

The choice of pChEMBL values as a criterion for activity was justified by several studies, which demonstrated that over 90% of pChEMBL values accurately reflect bioactivity outcomes (Bender et al., 2007; Kawai et al., 2021; Lenselink et al., 2017; Nidhi et al., 2006). These studies support the use of pChEMBL values as reliable indicators for bioactivity, particularly in cases where explicit comments or labels regarding activity were missing or unclear.

By assigning a pChEMBL threshold, we aimed to reduce the bias introduced by inconsistent labeling across different records. This method ensures a consistent and fair approach to labeling bioactivity status by classifying compounds based on standardized bioactivity metrics, rather than relying on potentially subjective observations found in text-based comments.

The selection process was automated using Python scripts to ensure precision and reproducibility. The newly labeled dataset will be discussed in the Results section, detailing how this selection contributes to the overall drug discovery pipeline.

6. Dataset Curation

Once the data selection process was complete, the final dataset was curated to ensure that it was well-structured and ready for further analysis. The curation step involved organizing the filtered and cleaned bioactivity data into a standardized tabular format, optimizing it for ease of use in downstream drug discovery research. By curating the dataset in this manner, it was made compatible with a variety of analytical tools, including machine learning and cheminformatics platforms.

Python was utilized to conduct the data curation process. The selected bioactivity records were stored in a tab-separated values (.tsv) format, which maintains a clear structure for easy access and readability. This format was chosen for its versatility in handling both textual and numerical data, and it provides a lightweight, platform-independent way to store large datasets.

In addition to its immediate use in drug discovery, this curated dataset has been designed to allow for future integration into more advanced systems, such as graph databases, which can represent complex relationships between compounds and biological targets. The tabular format ensures that researchers can easily import and manipulate the dataset as new tools and technologies emerge.

Results

1. Data Extraction

During the data extraction process, specific fields were chosen from the ChEMBL database to focus on the information most relevant to the drug discovery tasks. The extracted information was listed in Table 1. These fields provide essential details about molecules, assays, and bioactivities, ensuring that the dataset was tailored for use in drug discovery efforts.

The initial extraction process from the ChEMBLdb (Current Release: 34 Last Update: March 2024) resulted in 20,772,701 bioactivity data records. The initial dataset includes critical information, such as molecule identifiers, assay types, biological target details, bioactivity measurements, and organism taxonomy from various types of bioassays, many of which may not be directly related to drug discovery.

As shown in Table 1, each field represents essential information needed to analyze bioactivity data. For example, the Molecule ChEMBL ID (CHEMBL113) uniquely identifies the compound caffeine, and the Compound Key field provides the InChI-based identifier for its chemical structure. The Assay ChEMBL ID (CHEMBL1803433) references the specific assay used to measure the bioactivity, while the Assay

Description provides context about the experiment, such as the displacement of a ligand from a receptor using human cells in a cell-based assay.

Other critical fields include the Target ChEMBL ID (CHEMBL251), identifying the biological target (Adenosine A2a receptor), and the pChEMBL Value (5.3), which offers a standardized metric for comparing compound potency across different bioassays. Finally, the Comment field indicates that the bioactivity for this specific interaction was originally marked as "active."

Table 1 Extracted Fields Selected for Drug Discovery and Their Example of Result

Fields	Description	Importance	Example Record*
Molecule ChEMBL ID	Unique identifier for the molecule in the ChEMBL database	References the specific compound throughout the database	CHEMBL113
Compound Key	InChI-based identifier for chemical structure	Ensures accurate identification of the chemical structure	InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H, 1-3H3
Molecule Name	Name of the molecule	Provides a common reference for known molecules or drugs in research	CAFFEINE
Assay ChEMBL ID	Unique identifier for the assay	References the specific bioassay throughout the database	CHEMBL1803433
Assay Description	Details about the assay used	Offers context about the experimental conditions	Displacement of [3H]ZM241385 from stabilized human adenosine receptor A2a expressed in HEK293 cells followed by receptor capturing on Biocore chips by SPR method
Assay Type	Type of bioassay conducted	Indicates the type of the experiment	Binding
Target ChEMBL ID	Unique identifier for the biological target	References the biological target used in the experiment	CHEMBL251

Fields	Description	Importance	Example Record*
Target Name	Name of the biological target	Helps in identifying the protein or enzyme of interest	Adenosine A2a receptor
Target Type	The type of biological target	Focuses on the specific kind of target, such as proteins, that are common drug targets	SINGLE PROTEIN
Target Organism	The organism from which the biological target is derived	References the target organism used in the experiment	*Homo sapiens*
Standard Type	The type of bioactivity measurement	Key for comparing the potency and efficacy of the compounds tested	Ki
pChEMBL Value	Standardized bioactivity value by ChEMBL database	Provides a unified measure of bioactivity, enabling comparison across assays	5.3
Comment	Notes on activity	Offers insights into the compound's bioactivity status	Active

**The example shown in the table represents only one bioactivity data record.*

2. Data Filtering

From Table 2, the results show that the dataset was filtered to retain only records where the Target Type was set to a Single Protein, resulting in 2,676,265 (12.88% remaining) records. Next, the dataset was further reduced by filtering for the Target Organism = *Homo sapiens*, which narrowed the focus to human-related bioactivity, resulting in 1,981,391 (9.54% remaining) records. After applying these two critical filtering steps, the dataset was significantly reduced in size. This ensures that the resulting dataset was directly applicable to the development of therapeutics targeting human diseases because small molecules interacting with human proteins are crucial for identifying potential drug candidates.

Table 2 Data Filtering Result

Step	Criteria	Description	Records	Percentage (%)
1	Initial Data	Total records from data extraction	20,772,701	100.00
2	Target Type = Single Protein	Filtered to retain entries with single protein targets (all organisms)	2,676,265	12.88
3	Target Organism = *Homo sapiens*	Further filtered to retain human targets (*Homo sapiens*)	1,981,391	9.54

3. Data Cleaning

After filtering the data, the dataset was checked for duplicate entries. It initially contained 1,981,391 records, and 815 duplicates were identified and removed, resulting in a final dataset of 1,980,576 records. Duplicates accounted for only 0.04% of the total data (Table 3), which was an acceptable figure given the size of the dataset and does not significantly affect the integrity or diversity of the dataset.

Table 3 Duplicates Result

Step	Data	Description	Records	Percentage (%)
1	Filtered Data	Total records from data filtering	1,981,391	100.00
2	Duplicates	Duplicated data found	815	0.04
3	After Removing Duplicates Data	Total records after removing duplicates	1,980,576	99.96

When duplicates were cleaned, the dataset was checked for missing values of critical information extracted from the dataset. The results were presented in Table 4.

Table 4 Missing Values Result

Fields	# Missing Values	Percentage (%)
Molecule ChEMBL ID	0	0.00
Compound Key	10	0.0005
Molecule Name	1,840,447	92.92
Assay ChEMBL ID	0	0.00
Assay Description	0	0.00
Assay Type	0	0.00
Target ChEMBL ID	0	0.00
Target Name	0	0.00
Target Type	0	0.00
Target Organism	0	0.00
Standard Type	0	0.00
pChEMBL Value	0	0.00
Comment	1,019,778	51.49

Percentages were calculated based on the total number of records after removing duplicates (1,980,576).

The most significant missing data categories were molecular names and comments. However, missing Molecule Name values were acceptable, as many chemical structures lack common names and can still be identified using their InChI via the Compound Key. The 10 missing Compound Key entries were flagged for further analysis.

Additionally, the missing values in the Comment field were noted, but these entries were addressed during the data selection process, in which the pChEMBL values were used to justify their inclusion or exclusion from the dataset.

4. Data Selection and Dataset Curation

In drug discovery, accurate identification and classification of compounds as active or inactive are crucial for determining their potential as therapeutic agents. This study focused on curating a bioactivity-focused dataset that provides clear insights into the activity status of each compound. The "Comment" field was one of the primary pieces of information that allows us to differentiate between compounds based on their bioactivity.

However, the "Comment" field can fall into one of the four categories, as shown in Table 5. The data records were curated to “Bioactivity Dataset Type 1: Original Data.” This dataset contains compounds with their bioactivity labels as originally provided, encompassing records that were not modified during data processing. Researchers seeking bioactivity data with minimal interventions can utilize this dataset.

Table 5 “Bioactivity Dataset Type 1: Original Data” with Bioactivity Comment Results

Bioactivity Comment	# Data Records	Percentage (%)
Active	204,815	10.34
Inactive	86,481	4.37
Not both	669,502	33.80
Missing Values	1,019,778	51.49
Total Data	**1,980,576**	**100.00**

Active compounds accounted for 204,815 records (10.34%), where the bioactivity of the compounds was clearly labeled as "Active," signifying a significant interaction with a biological target. Inactive compounds comprised 86,481 records (4.37%) and were classified as having no significant activity. A significant portion of the dataset fell into the "Not Both" category, with 669,502 records (33.80%). These records were neither explicitly labeled as "Active" nor "Inactive." This category includes various uncertain labels, such as entries containing only numerical values, incomplete information, or results labeled as "inconclusive."

Meanwhile, for the "Not Both" and “Missing Values” categories, additional steps were taken to assess their relevance to drug discovery tasks. To address the gaps in the "Not Both" and “Missing Values” categories, pChEMBL was used as an alternative method to classify the bioactivities of these compounds. Notably, as indicated by the zero missing values in the pChEMBL value column, pChEMBL data was available for all records, ensuring comprehensive coverage of bioactivity information. Since text-based comments such as “inconclusive” or incomplete information, can vary between studies and laboratories, introducing inconsistencies in how bioactivity was recorded.

The results obtained after processing were listed in Table 6. The data records were curated to “Bioactivity Dataset Type 2: Processed Data.” This dataset contained compounds whose bioactivity was determined using the pChEMBL values. A higher pChEMBL value indicates stronger bioactivity, making it easier to rank compounds in terms of potency.

This method ensures consistency across the dataset and reduces the potential for discrepancies arising from differences in experimental setups. This provides researchers with an additional layer of data derived through computational methods, enabling broader insights into drug discovery. By setting a clear threshold mentioned in Research Methodology section, we ensure a uniform classification across all records, reducing the potential for bias that could arise from human interpretation of experimental results.

Table 6 "Bioactivity Dataset Type 2: Processed Data" with Bioactivity Comment Results

Bioactivity Comment	# Data Records	Percentage (%)
Active	1,446,788	73.05
Inactive	533,788	26.95
Total Data	**1,980,576**	**100.00**

Discussion

In drug discovery research, the ability to handle large volumes of data and extract meaningful insights is critical for identifying potential therapeutic agents. This study focused on developing a systematic framework to process bioactivity data from the ChEMBL database, which involved curating, cleaning, and refining the dataset to ensure its relevance and manageability for drug discovery purposes. Our primary objectives were to create a data processing framework capable of handling the complexities of bioactivity data, and to develop a streamlined, bioactivity-focused dataset that can accelerate drug discovery projects.

A key component of this study was the use of pChEMBL values to classify bioactivity when explicit comments on activity (such as "active" or "inactive") were missing or unclear. This step played a critical role in streamlining the dataset for drug discovery tasks, as it allowed us to assign bioactivity statuses in a standardized and objective method. This approach not only reduced bias in the data but also provided a consistent method for evaluating the potential of compounds, making the dataset more reliable and actionable for future research.

One example that highlights the effectiveness of this framework is the case study of "MANGOSTIN," which was ranked among the top 3 compounds with the highest pChEMBL values in our dataset in a Table 7. Originally, MANGOSTIN had missing or unclear activity comments in the raw ChEMBL data. However, after applying our data selection method based on pChEMBL values, we classified the compound as "Active" due to its high pChEMBL values across multiple assays. The refined data showed that MANGOSTIN was involved in inhibition activities against human protein targets, such as 7,8-dihydro-8-oxoguanine triphosphatase, Arachidonate 12-lipoxygenase, and Phosphodiesterase 4D. Without the pChEMBL-driven selection process, the potential therapeutic relevance of MANGOSTIN might have been overlooked, illustrating how our framework streamlines the drug discovery process by ensuring that key bioactivity data is retained and classified correctly. By identifying MANGOSTIN as a compound with significant inhibitory potential, we can now propose further experimental testing to confirm its activity in these assays.

Table 7 Example of MANGOSTIN Data Across Top 3 Highest pChEMBL Values After Data Selection

Fields	Example Record A	Example Record B	Example Record C
Molecule ChEMBL ID	CHEMBL323197	CHEMBL323197	CHEMBL323197
Compound Key	InChI=1S/C24H26O6/c1-12(2)6-8-14-16(25)10-19-21(22(14)27)23(28)20-15(9-7-13(3)4)24(29-5)17(26)11-18(20)30-19/h6-7,10-11,25-27H,8-9H2,1-5H3	InChI=1S/C24H26O6/c1-12(2)6-8-14-16(25)10-19-21(22(14)27)23(28)20-15(9-7-13(3)4)24(29-5)17(26)11-18(20)30-19/h6-7,10-11,25-27H,8-9H2,1-5H3	InChI=1S/C24H26O6/c1-12(2)6-8-14-16(25)10-19-21(22(14)27)23(28)20-15(9-7-13(3)4)24(29-5)17(26)11-18(20)30-19/h6-7,10-11,25-27H,8-9H2,1-5H3
Molecule Name	MANGOSTIN	MANGOSTIN	MANGOSTIN
Assay ChEMBL ID	CHEMBL4339319	CHEMBL898687	CHEMBL4404704
Assay Description	Inhibition of recombinant MTH1 (3 to 156 residues) expressed in E. coli	Inhibition of human 12-hLO	Inhibition of human PDE4D2
Assay Type	Binding	Binding	Binding
Target ChEMBL ID	CHEMBL3708265	CHEMBL3687	CHEMBL288
Target Name	7,8-dihydro-8-oxoguanine triphosphatase	Arachidonate 12-lipoxygenase	Phosphodiesterase 4D
Target Type	SINGLE PROTEIN	SINGLE PROTEIN	SINGLE PROTEIN
Target Organism	*Homo sapiens*	*Homo sapiens*	*Homo sapiens*

Fields	Example Record A	Example Record B	Example Record C
Standard Type	IC50	IC50	IC50
pChEMBL Value	6.33	6.24	5.88
Comment	Bioactivity Dataset Type 1: - missing value Bioactivity Dataset Type 2: - Active	Bioactivity Dataset Type 1: - missing value Bioactivity Dataset Type 2: - Active	Bioactivity Dataset Type 1: - missing value Bioactivity Dataset Type 2: - Active

The curated dataset opens opportunities for various advanced drug discovery tasks. In particular, it can be integrated into machine learning models for virtual screening, where the aim is to predict the bioactivity of untested compounds. These computational approaches can significantly reduce the time and cost associated with traditional laboratory testing, accelerating the process of identifying promising drug candidates.

However, this study had several limitations. One potential limitation of the ChEMBL dataset is the variability in bioactivity records, as these were sourced from multiple laboratories worldwide, each with differing experimental protocols and conditions. While ChEMBL manual curation ensures a level of standardization and quality control, these differences in experimental setups may still affect the interpretation of bioactivity results, even after curation.

Analysis of missing values showed that 92.92% of the molecular names were absent. This highlights the limitations of relying on common names in drug discovery, particularly when dealing with small or experimental compounds that may not have widely recognized names. In contrast, the Compound Key (InChI identifier) was almost fully available, emphasizing the importance of standardized chemical identifiers such as InChI. InChI provides a reliable and consistent means of representing chemical structures, enabling more accurate tracking, comparison, and data integration across studies.

Suggestion

This study curated two key datasets from the ChEMBL database: the Original Bioactivity Dataset, preserving the exact experimental data, and the Processed Bioactivity Dataset, which inferred bioactivity using pChEMBL values for compounds without explicit labels. This dual approach offers researchers flexibility, allowing them to choose between validated experimental data and computationally inferred insights based on their specific research needs.

The case study of MANGOSTIN highlights the potential of this curated dataset to support drug discovery, particularly in instances where experimental data is incomplete. By reanalyzing bioactivity data and applying systematic processing, researchers can uncover compounds with promising bioactivity that might otherwise be left unnoticed.

This capability is especially useful for streamlining drug discovery processes, as it enables faster identification of potential drug candidates. Furthermore, the dataset's structure facilitates its integration into predictive models and machine learning tools, offering even greater potential for accelerating drug discovery efforts in the future.

References

Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., Orhan, I. E., Banach, M., Rollinger, J. M., Barreca, D., Weckwerth, W., Bauer, R., Bayer, E. A., Majeed, M., Bishayee, A., Bochkov, V., Bonn, G. K., Braidy, N., Bucar, F., Cifuentes, A., D'Onofrio, G., Bodkin, M., . . . the International Natural Product Sciences, T. (2021). Natural products in drug discovery: advances and opportunities. *Nature Reviews Drug Discovery*, *20*(3), 200-216. https://doi.org/10.1038/s41573-020-00114-z

Bender, A., Young, D. W., Jenkins, J. L., Serrano, M., Mikhailov, D., Clemons, P. A., & Davies, J. W. (2007). Chemogenomic data analysis: prediction of small-molecule targets and the advent of biological fingerprint. *Comb Chem High Throughput Screen*, *10*(8), 719-731. https://doi.org/10.2174/138620707782507313

Gao, M., & Skolnick, J. (2013). A comprehensive survey of small-molecule binding pockets in proteins. *PLoS Comput Biol*, *9*(10), e1003302. https://doi.org/10.1371/journal.pcbi.1003302

Hunter, F. M. I., Bento, A. P., Bosc, N., Gaulton, A., Hersey, A., & Leach, A. R. (2021). Drug Safety Data Curation and Modeling in ChEMBL: Boxed Warnings and Withdrawn Drugs. *Chemical Research in Toxicology*, *34*(2), 385-395. https://doi.org/10.1021/acs.chemrestox.0c00296

Ishola, A. A., Adedirin, O., Joshi, T., & Chandra, S. (2021). QSAR modeling and pharmacoinformatics of SARS coronavirus 3C-like protease inhibitors. *Computers in Biology and Medicine*, *134*, 104483. https://doi.org/https://doi.org/10.1016/j.compbiomed.2021.104483

Jitruknatee, A., Martro, J., Tosanguan, K., Doangjai, Y., & Theantawee, W. (2020). National Drug Policies in Thailand: Evolution and Lessons for the Future. *Journal of Health Science of Thailand*, *29*(0), S3-S14. https://thaidj.org/index.php/JHS/article/view/8409

Kawai, K., Tomonou, M., Machida, Y., Karuo, Y., Tarui, A., Sato, K., Ikeda, Y., Kinashi, T., & Omote, M. (2021). Effect of Learning Dataset for Identification of Active Molecules: A Case Study of Integrin αIIbβ3 Inhibitors. *Molecular Informatics*, *40*(6), 2060040. https://doi.org/https://doi.org/10.1002/minf.202060040

Kim, S., Lim, S. W., & Choi, J. (2022). Drug discovery inspired by bioactive small molecules from nature. *Anim Cells Syst (Seoul)*, *26*(6), 254-265. https://doi.org/10.1080/19768354.2022.2157480

Kwankhao, P., Chuthaputti, A., Tantipidok, Y., Pathomwichaiwat, T., Theantawee, W., Buabao, S., Chantraket, R., Puttarak, P., Petrakart, P., Chinsoi, P., Chungsiriporn, D., Bongcheewin, B., & Sermsinsiri, V. (2020). The Current Situation of the Herbal Medicinal Product System in Thailand. *Journal of Health Science of Thailand*, *29*(0), S82-S95. https://thaidj.org/index.php/JHS/article/view/8415

Lenselink, E. B., Ten Dijke, N., Bongers, B., Papadatos, G., van Vlijmen, H. W. T., Kowalczyk, W., AP, I. J., & van Westen, G. J. P. (2017). Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J Cheminform*, *9*(1), 45. https://doi.org/10.1186/s13321-017-0232-0

Mutowo, P., Bento, A. P., Dedman, N., Gaulton, A., Hersey, A., Lomax, J., & Overington, J. P. (2016). A drug target slim: using gene ontology and gene ontology annotations to navigate protein-ligand target space in ChEMBL. *Journal of Biomedical Semantics*, *7*(1), 59. https://doi.org/10.1186/s13326-016-0102-0

Najmi, A., Javed, S. A., Al Bratty, M., & Alhazmi, H. A. (2022). Modern Approaches in the Discovery and Development of Plant-Based Natural Products and Their Analogues as Potential Therapeutic Agents. *Molecules*, *27*(2). https://doi.org/10.3390/molecules27020349

Nidhi, Glick, M., Davies, J. W., & Jenkins, J. L. (2006). Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J Chem Inf Model*, *46*(3), 1124-1133. https://doi.org/10.1021/ci060003g

Qiao, F., Binknowski, T. A., Broughan, I., Chen, W., Natarajan, A., Schiltz, G. E., Scheidt, K. A., Anderson, W. F., & Bergan, R. (2024). Protein Structure Inspired Drug Discovery. *bioRxiv*. https://doi.org/10.1101/2024.05.17.594634

Tabana, Y., Babu, D., Fahlman, R., Siraki, A. G., & Barakat, K. (2023). Target identification of small molecules: an overview of the current applications in drug discovery. *BMC Biotechnology*, *23*(1), 44. https://doi.org/10.1186/s12896-023-00815-4

Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., de Veij, M., Ioannidis, H., Lopez, D. M., Mosquera, Juan F., Magarinos, Maria P., Bosc, N., Arcila, R., Kizilören, T., Gaulton, A., Bento, A P., Adasme, Melissa F., Monecke, P., Landrum, Gregory A., & Leach, Andrew R. (2023). The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, *52*(D1), D1180-D1192. https://doi.org/10.1093/nar/gkad1004

Zhu, Y., Ouyang, Z., Du, H., Wang, M., Wang, J., Sun, H., Kong, L., Xu, Q., Ma, H., & Sun, Y. (2022). New opportunities and challenges of natural products research: When target identification meets single-cell multiomics. *Acta Pharmaceutica Sinica B*, *12*(11), 4011-4039. https://doi.org/https://doi.org/10.1016/j.apsb.2022.08.022