WEBLEM 1

Importance of 3D structures and method of generation from 1D and 2D representation for Drug designing and Chemoinformatics studies

The intensity and selectivity of therapies' interactions with proteins and other bioactive molecules are significantly influenced by the 3-dimensional (3D) structure of those molecules. Previous research has taken into account the strain introduced and tolerated by protein-induced conformational changes.

Ligand-protein interactions control how well new therapies are recognised by their intended targets. In order to create potent and targeted medications, structure-activity relationships (SAR) are often developed to examine noncovalent interactions, such as hydrogen and halogen bonding, salt bridges, and pi-pi stacking. The shape of inhibitors in the solvated vs. bound states can also influence the energetic favorability of inhibitor binding, in addition to many other protein-ligand interactions that can be adjusted. To produce effective medications, the variety of druggable protein targets needs structural and conformational flexibility in ligands. It has been investigated how much molecular strain and associated energy expenditures medications can withstand when they attach to proteins, and it was discovered that molecules binding to proteins might easily use 5–9 kcal/mol of strain energy. Although significant rearrangements are feasible, more diverse structural shapes have been linked to a spectrum of biological activities and positive therapeutic consequences. Modeling and docking of three-dimensional (3D) structures into protein active sites are now possible thanks to the accessibility and abundance of structures in the Protein Data Bank (PDB). In order to improve lead identification and therapeutic success, libraries with greater topological variety have been prepared for drug-discovery campaigns as a result of these observations and methods.

Method of Generation from 1D and 2D representatinog for drug designing and chemoinformatics studies

Mapping Chemical Graphs into 1D descriptors:

Chemical compound information is significantly reduced when mapping to a 0D representation, and 0D descriptors are typically too rough to encode molecules. As an alternative, 1D descriptors like:

- Systematic nomenclature
- Line Notations

Appeared to be enough for the efficient coding of molecules and for database searching.

The idea of a systematic nomenclature scheme (chemical name) by naming rules is human-oriented, hence it has been given top importance to make chemical names that are simple to pronounce and names that human chemists can easily read. But systematic nomenclature, like the IUPAC system, frequently results in names that are overly complex, reflecting the intricacy of chemical structures. A computer-oriented system does not always necessitate a nomenclature that is user-friendly. A preferred name programme (PNP) is currently being considered to address the difficulty of limiting the name generated for a particular structure to a single value, for instance. In 1986, Beilstein created the first system that permitted the entry of chemical structures in the form of their chemical names.

Another significant method for translating a 2D molecular graph into a linear list of letters and numbers is called a line notation. The first systems for this mapping were created even before computer applications, and their success peaked in the 1960s and 1970s when they gave scientists a quick way to code even big molecules.

Due of the small amount of space needed for data storage in computers at the time, this was particularly appropriate. Among crucial line notations throughout history:

- Wiswesser (WLN)
- ROSDAL
- SYBYL (SLN)
- SMILES

SMILES, the later of the two, obviously rules today. This is brought on by the system's intuitive design and its extremely basic semantic rules. We provide some sample cases in Fig. 3. Every chemical entry on Wikipedia is supported by SMILES codes, and practically every molecular editor can produce SMILES. In the other direction, SMILES can take the place of a molecular input into a number of databases, such as Reaxys and/or molecular editors. If we consider that a SMILES input can currently conduct a Google search without any intermediaries, there is a strong odour of success for this system.

A line notation accomplishes a mathematical reduction of data from a 2D into a 1D representation. The key to this success is dimensional reduction, which keeps 2D information intact. Data manipulation is simplified when the dimension is lower. The fact that this dimensionality reduction is completely reversible is a noteworthy observation. In order to recreate the original molecule, we can both obtain a SIMLES for every molecule.

Mapping Chemical Graphs into 2D Descriptors:

For the needs of human scientists, molecular graphs were created to code molecules; yet, they first seemed unsuitable for providing computers with chemical data. The approaches created for mapping molecular graphs into computer-friendly data include matrix notations and linear codes. There are many different matrix systems (2D descriptors) that have been developed, and the numeric representation depends on the coding scheme for atoms and bonds. The following matrices are examples of popular notations:

- Adjacency
- Atom connectivity
- Distance (topological or geometrical)
- Bond
- Bond electrons
- Incidence

Although a matrix system works well, it is not practical for storing huge molecules or enormous datasets since the number of entries required to represent a molecule squares with the growth in the number of atoms in the molecule.

An alternative approach is the connection table (CT), which enables us to display molecular graphs as lists of the atoms and bonds within a molecule that grow linearly as the number of atoms increases. There are numerous CT variants available, and Gasteiger (2003) is recommended for a thorough introduction.

The most common method for depicting molecular structures today is a CT notation. In Fig. 6, we demonstrate a Molfile that codes the three-dimensional structure of salicylic acid while also providing the atoms' three-dimensional coordinates. Standard notations like Molfiles and SDFiles have been established and are widely used for information communication. In essence, a globally standardised computer-readable form has been created from the global language of chemical graphs. (These representations can also be read by a human because they are coded in ASCII.) Bobach et al. provide an intriguing example of coding vitamin C, a chemical molecule containing all tautomers, via a CT (2012).

The molecular topology is defined by a 2D molecular representation (atomic constitution and connectivity). However, in its purest form, this does not encode stereochemistry, also known as atomic configuration—the spatial arrangement of atoms around an atomic centre. Chemical graphs can be enhanced with solid and dashed

wedged bonds to indicate above-the-plane or below-the-plane pointing in order to encode this. The configuration is a streamlined 3D molecular illustration that shows the stereoisomer's atomic arrangement but not its precise 3D atomic coordinates.

The development of stereochemical codes in chemistry aims to greatly simplify the viewing of 3D molecules in 2D for the human chemist. Information about stereochemistry lies between topology (2D) and topography (3D).

There are really just two representations for the actual 3D structure. The first step is to measure the real property of compounds (substances) using 3D atomic coordinates, such as by crystallography. As a result, databases contain such data (see section "Databases and Database Searching"). Second, there are 3D structural predictions made using molecular modelling. For the second scenario, read the section under "Representation of 3D Structures and Their Mapping into Lower Dimensionality" for an explanation of how this process maps a 2D (CT) or 1D (SMILES) molecular representation into 3D. Technically, there are numerous data formats in which 3D structures can be noted:

- SDfile (Structure Data) *.sdf
- Molfile (Connection Table Format) *.mol
- CIF (Crystallographic Information Fil) *.cif
- PDB (Protiens) *.pdf

These data files can represent both descriptor and property data and can e converted into each other.

PUBCHEM DATABASE:

Introduction and history:

PubChem is a database that the general public can access to learn more about chemical compounds and their biological effects. Since its debut in 2004 as a part of the US National Institutes of Health's (NIH) Molecular Libraries Roadmap Initiatives, PubChem has quickly developed into a crucial source of chemical information that supports scientific communities in a variety of fields, including cheminformatics, chemical biology, medicinal chemistry, and drug discovery.

One of the greatest collections of freely accessible chemical information may be found on PubChem. As of September 2015, it included more than 157 million descriptions of chemical substances contributed by depositors, 60 million descriptions of distinctive chemical structures, and 1 million descriptions of biological assays, encompassing roughly 10,000 different protein target sequences. These enormous amounts of data are organised by PubChem into three interconnected databases called Substance, Compound, and BioAssay. Information that was provided by depositors is kept in the Substance database. The Compound database stores unique chemical structures that are taken from the Substance database. Descriptions of biological tests on chemical compounds are kept in the BioAssay database. SID (SubstanceID), CID (CompoundID), and AID (AssayID), respectively, are the main identifiers for the Substance, Compound, and BioAssay databases.

Working:

Web interfaces for textual search:

The three main PubChem databases as well as other significant NCBI databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Genome, Taxonomy, BioSystems, Gene Expression Omnibus (GEO), and many others, use the search and retrieval engine Entrez. By starting a search from the PubChem home page, which also offers launch points to other PubChem services, tools, assistance materials, and more, one can search the PubChem databases through Entrez. The search can also be started from the NCBI main page. Entrez searches all accessible Entrez databases by default if a specific database is not

selected in the search menu, and it reports the number of records in each database that are returned for this "global query."

Multiple records from an Entrez search are shown in a document summary (DocSum) report. There is some data-specific information and a link to the record's summary page for each record in the DocSum page. Controls are available on the DocSum page to alter the display type, sort the outcomes in different ways, or export the page to a file or printer. Additionally, the icons and buttons in the right column of the DocSum page give users the option to download the relevant records, alter or refine their search, get related records from other databases, perform additional analysis on the query result, and more.

Non-textual search using the chemical structure search tool:

Entrez cannot be used for searching that requires data types peculiar to PubChem, including chemical structures, since it is largely a text-based search system. A variety of chemical structure search methods, such as identity search, substructure/superstructure search, molecular formula search, and 2-D and 3-D similarity searches, can be used to query and subset the Compound database using the Chemical Structure Search tool. A number of query formats, including SMILES, SMARTS, InChI, CID, molecular formula, and SDF, are supported by the Chemical Structure Search tool. The PubChem Chemical Structure Sketcher can also be used to manually construct a query chemical structure. The user does not need to download or install any other software in order to use this JavaScript-based structure editor, which is cross-platform compatible and works with all major web browsers.

Users of the Chemical Structure Search tool have the option of limiting their search to a set of CIDs submitted in a file, the results of a prior Entrez or chemical structure search, or both. The search result can be restricted using optional filters depending on a variety of factors, including molecular weight, the number of heavy atoms, whether stereochemistry is present or not, the name or category of the depositor, and others. One can export a query to an XML file, import the query from the XML file, and repeat the search without having to fill out the search form once more. This XML file can also be used as a model for creating PUG interface questions.

Applications:

Numerous domains, including chemical biology, medicinal chemistry, and informatics research have utilised the vast amount of chemical structure and bioactivity data in PubChem and its online services. Drug discovery is supported by PubChem in various ways, including lead identification and optimization, compound-target profiling, polypharmacology research, and the clarification of unidentified chemical identities. Additionally useful for creating auxiliary databases, informatics tools, and online services, PubChem has grown into a platform. With its rising public accessibility, the PubChem resource provides assistance and fantastic opportunities for the investigation of pharmacological mechanisms and the genetic underpinnings of disorders, both of which are essential for drug development and repurposing.

A relatively recent concept in chemistry, chemoinformatics is based on the processing of information on chemical and molecular structures using computational analysis. These data can be analysed to examine the connection between molecular activity, chemical characteristics, and chemical structure. It is an in-silico technique, which is a type of scientific investigation carried out virtually on a computer using software and simulations. Choosing an illness to target is the first step in the typical drug discovery process. Next, prospective chemicals and molecules that could be used to lessen the disease's severity in some way are looked for.

Virtual screening can be used to filter out specific compounds early on that aren't compatible without the requirement for physical screening in order to lower expenses and accelerate drug development while screening for new prospective compounds that could be developed into medications. This strategy, which is

significantly more effective than traditional approaches, builds virtual screens and simulations using computer software to search for prospective compounds that could be turned into medications. In order to locate new prospective compounds to be utilised as medications, chemoinformatics and drug discovery depend on the 3D structures of compounds as well as their representation in 1D and 2D forms.

REFERENCES:

1. Kim, S., Thiessen, P., Bolton, E., Chen, J., Fu, G., & Gindulyte, A. et al. (2015). PubChem Substance and Compound databases. *Nucleic Acids Research*, *44*(D1), D1202-D1213. doi: 10.1093/nar/gkv951
2. Cheng, T., Pan, Y., Hao, M., Wang, Y., & Bryant, S. (2014). PubChem applications in drug discovery: a bibliometric analysis. *Drug Discovery Today*, *19*(11), 1751-1756. doi: 10.1016/j.drudis.2014.08.008
3. Prosser, K., Stokes, R., & Cohen, S. (2020). Evaluation of 3-Dimensionality in Approved and Experimental Drug Space. *ACS Medicinal Chemistry Letters*, *11*(6), 1292-1298. doi: 10.1021/acsmedchemlett.0c00121
4. Polanski, J., & Gasteiger, J. (2017). Computer Representation of Chemical Compounds. *Handbook OfComputational Chemistry*, 1997-2039. doi: 10.1007/978-3-319-27282-5_50
5. Jack Davis, B. (2022). How is Chemoinformatics Used in Drug Discovery?. Retrieved 5 September2022, from https://www.azolifesciences.com/article/How-is-Chemoinformatics-Used-in-Drug-Discovery.aspx#:~:text=Chemoinformatics%20can%20drastically%20enhance%20this%20process%2C%20as%20one,calculate%20and%20visualize%20structures%20is%20crucial.%20Virtual%20screening

WEBLEM 1a

Retrieve 2D/3D structures using Pubchem Database

(URL: https://pubchem.ncbi.nlm.nih.gov/)

AIM:

To Retrieve 2D/3D structure for Penicillin (Pubchem id - 22502) using PubChem Database and apply filters as:

1. To refine search using Substructure, Exact and Similar structure approach.
2. To screen structure based on chemical properties approach.

INTRODUCTION:

Penicillin is the procaine salt form of penicillin G, a broad-spectrum, beta-lactam, naturally occurring penicillin antibiotic with antibacterial activity. Penicillin G binds to and inactivates the penicillin binding proteins (PBPs) located inside the bacterial cell wall. Inactivation of PBPs interferes with the cross-linkage of peptidoglycan chains necessary for bacterial cell wall strength and rigidity. This interrupts bacterial cell wall synthesis and results in the weakening of the bacterial cell wall, eventually causing cell lysis.

PubChem is a database open to the public that contains details on chemical compounds and their biological effects. PubChem was first introduced in 2004 as a part of the Molecular Libraries Roadmap Initiatives of the US National Institutes of Health (NIH), and since then has grown quickly to become a major source of chemical information that supports scientific communities in a variety of fields, including cheminformatics, chemical biology, medicinal chemistry, and drug discovery.

One of the greatest collections of freely accessible chemical information may be found on PubChem. As of September 2015, it included more than 157 million descriptions of chemical substances contributed by depositors, 60 million descriptions of distinctive chemical structures, and 1 million descriptions of biological assays, encompassing roughly 10,000 different protein target sequences. These enormous amounts of data are organised by PubChem into three interconnected databases called Substance, Compound, and BioAssay. Information that was provided by depositors is kept in the Substance database. The Compound database stores unique chemical structures that are taken from the Substance database. Descriptions of biological tests on chemical compounds are kept in the BioAssay database. SID (SubstanceID), CID (CompoundID), and AID (AssayID), respectively, are the main identifiers for the Substance, Compound, and BioAssay databases.

Data filtering is crucial because redundant or unreliable pieces of information can make users confused or disoriented. Results can also be more effective by filtering the data. Data filters can also be used in other situations to limit access to private information.

METHODOLOGY:

- Open homepage for PubChem daabase
- Search for compound “Penicillin”
- Open result for best match
- Refine results obtained for Penicillin using Substructure, Exact and similar structure approach
- Screen results for penicillin using chemical properties approach
- Observe and interpret the results

OBSERVATIONS:

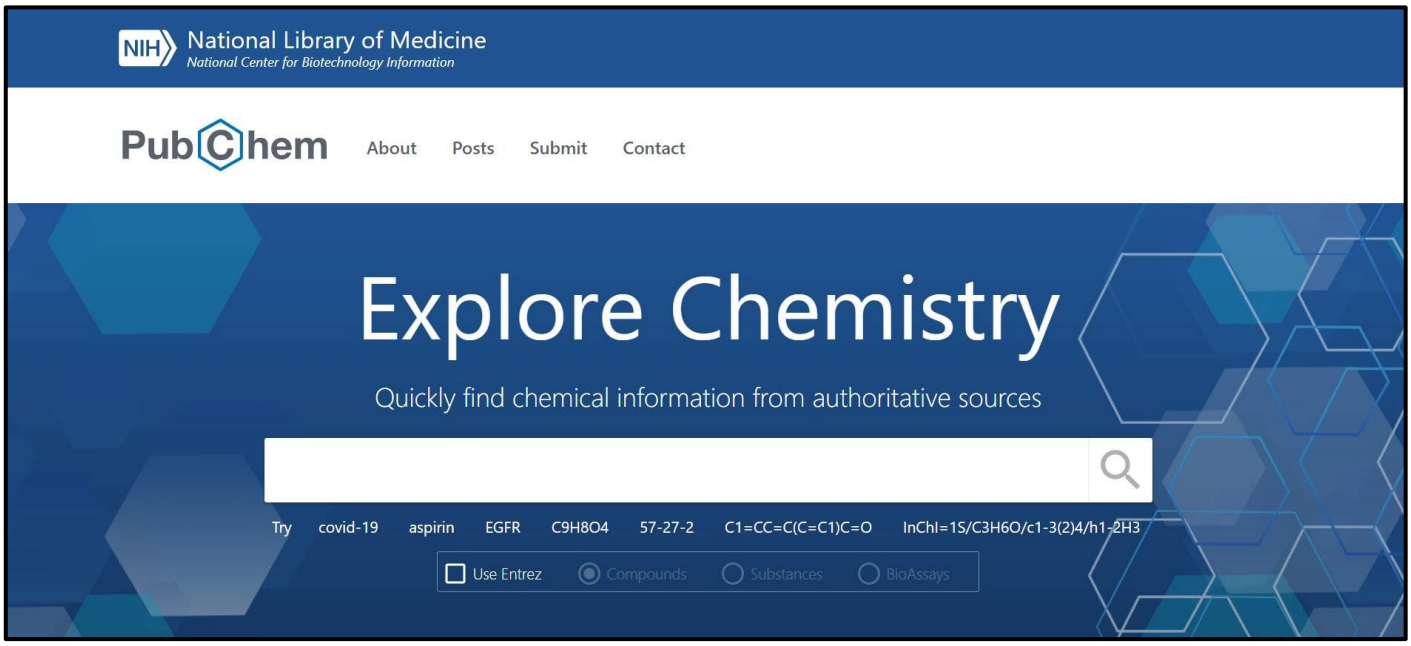

FIG 1. Homepage of Pubchem Database

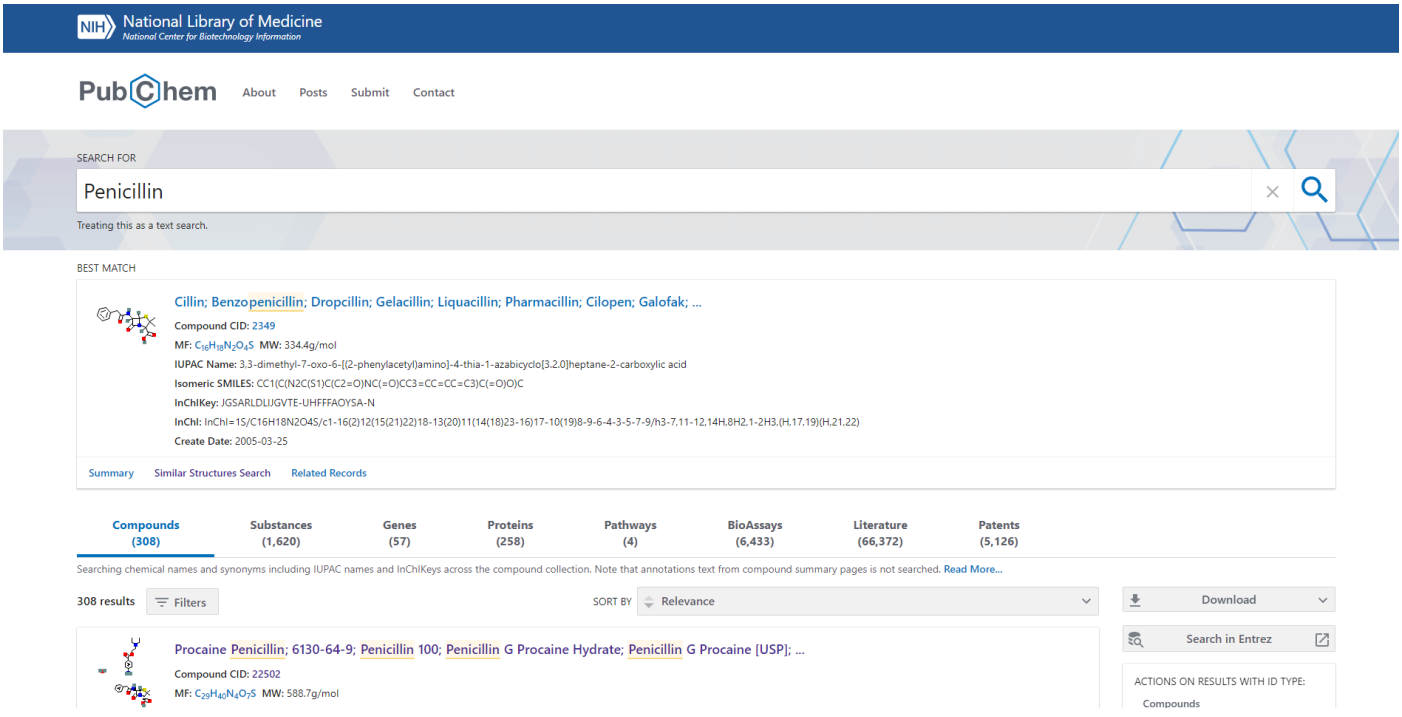

Fig 2. Hit page for compound "Penicillin"

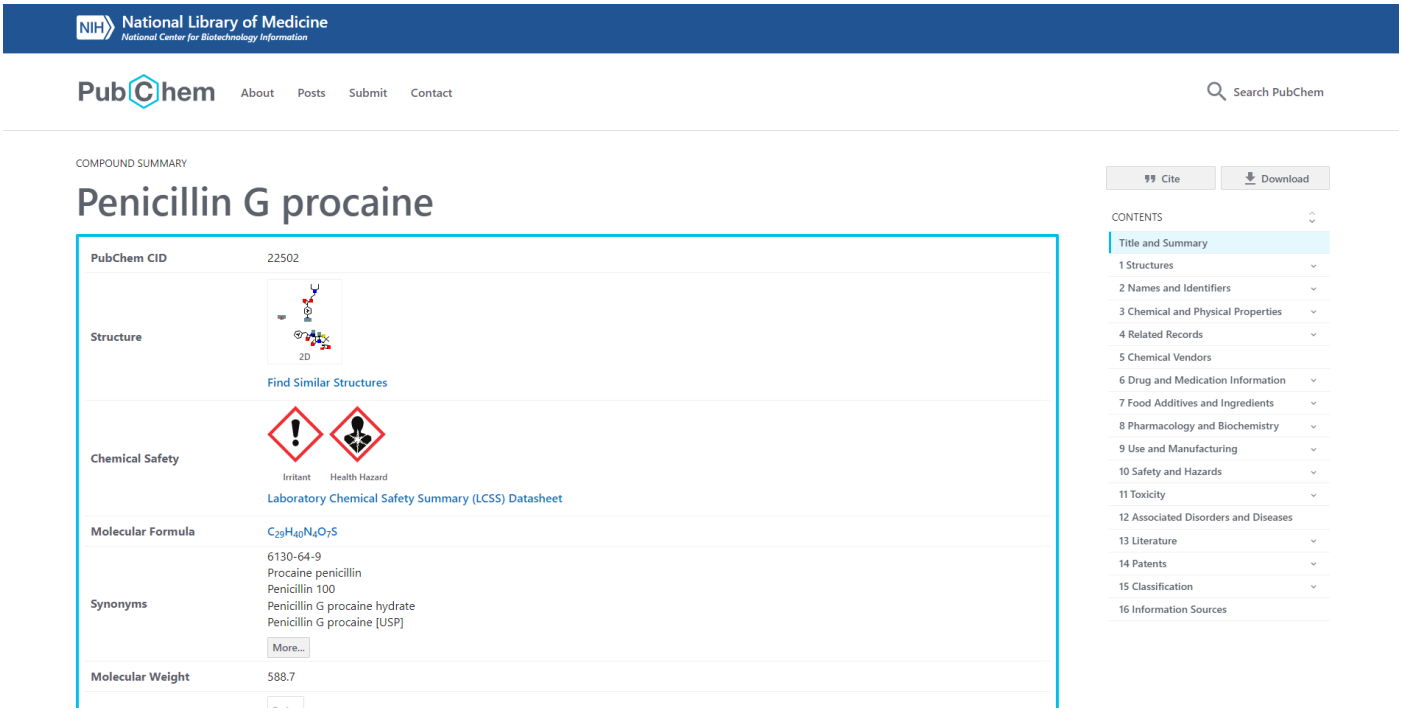

Fig 3. Result page with summary for Penicillin

Fig 3.1 Structure information for Penicillin

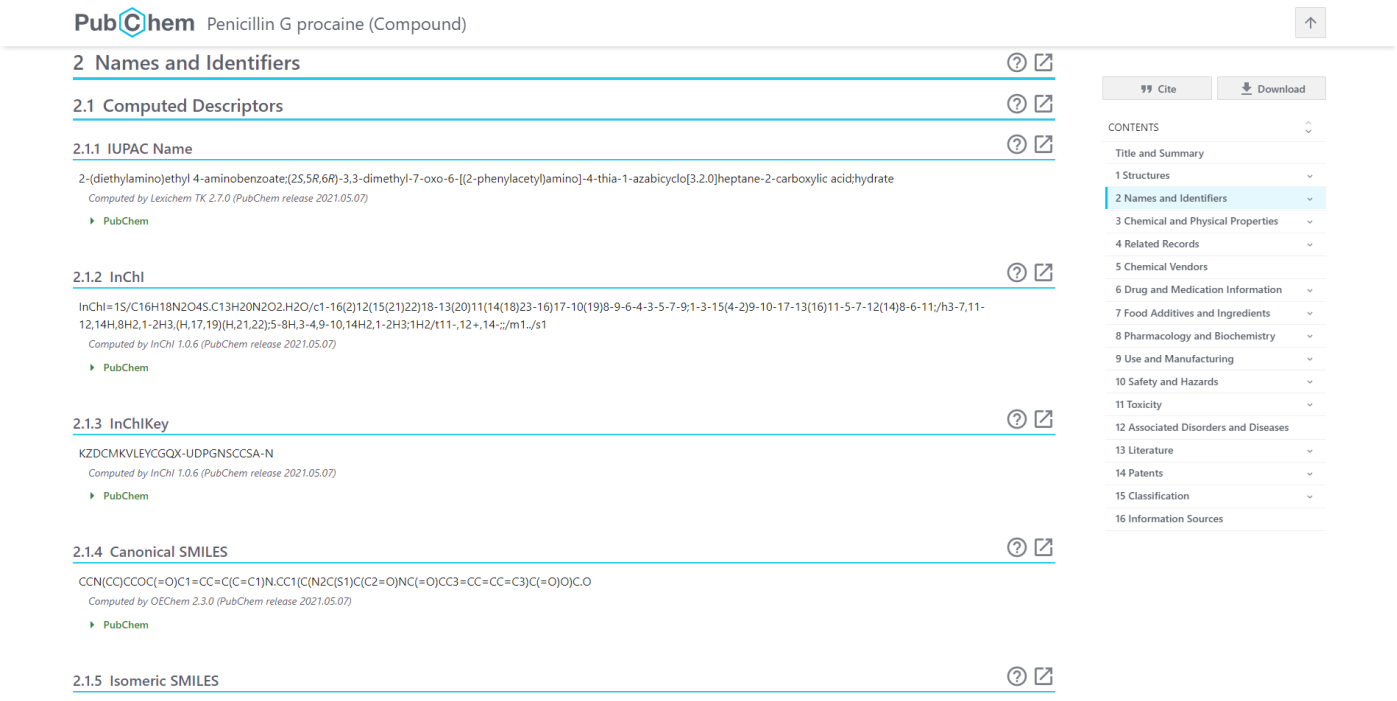

Fig 3.2. Names and Identifiers for Penicillin

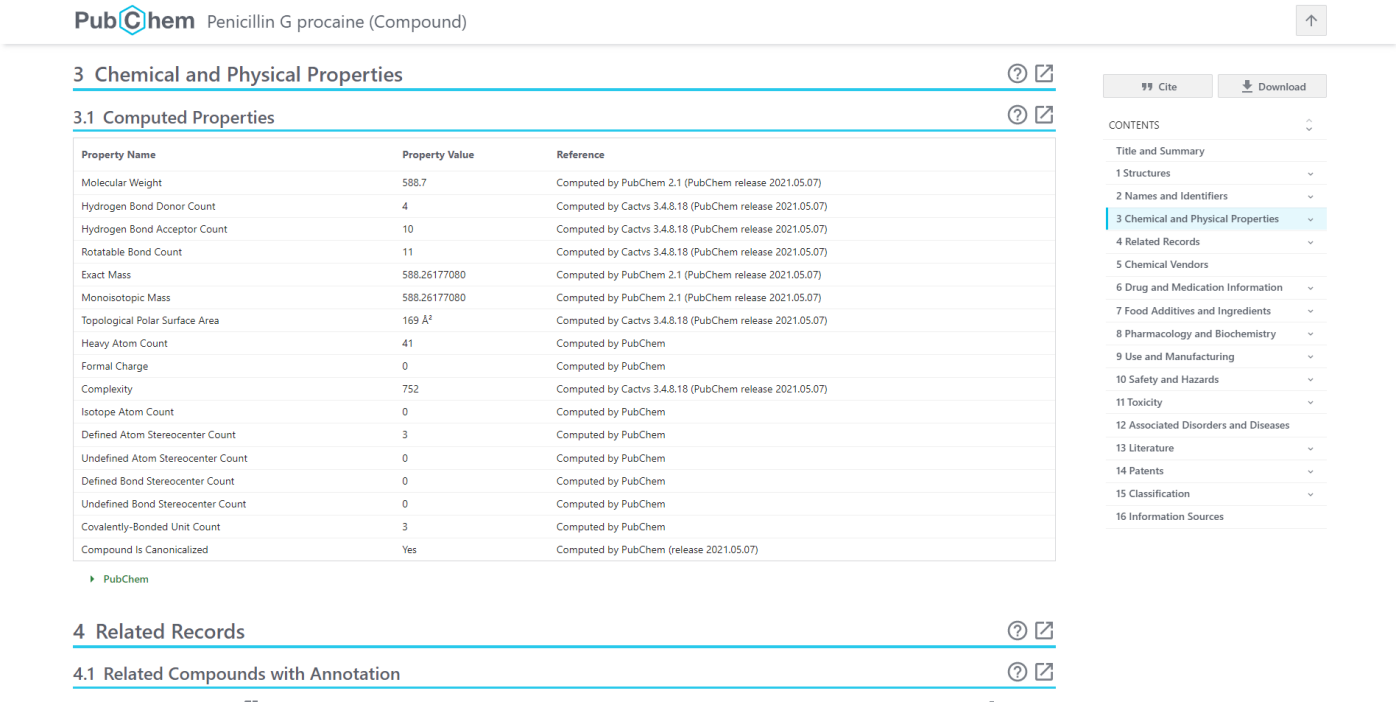

Fig 3.3. Chemical and Physical properties of Penicillin

4 Related Records

4.1 Related Compounds with Annotation

9 items View More Rows & Details

Download

SORT BY Compound CID

Structure	Compound CID	Name	Molecular Formula	Molecular Weight, g/mol
	5903	Procaine penicillin G	$C_{29}H_{38}N_4O_5S$	570.7
	22502	Penicillin G procaine	$C_{29}H_{40}N_4O_7S$	588.7
	23570	2-(Diethylamino)ethyl 4-amino-2-chlorobenzoate;3,3-dimethyl-7-oxo-6-[(2-phenylacetyl)amino]-4-thia-1-azabicyclo[3.2.0]heptane-2-carboxylic acid	$C_{29}H_{37}ClN_4O_6S$	605.1
	517319	2-(Diethylamino)ethyl 4-aminobenzoate;3,3-dimethyl-7-oxo-6-[(2-phenylacetyl)amino]-4-thia-1-azabicyclo[3.2.0]heptane-2-carboxylic acid	$C_{29}H_{38}N_4O_5S$	570.7
	3081515	Jenicillin A	$C_{45}H_{56}N_6O_{10}S_2$	905.1

1 2 Next >

Cite Download

CONTENTS

Title and Summary
1 Structures
2 Names and Identifiers
3 Chemical and Physical Properties
4 Related Records
5 Chemical Vendors
6 Drug and Medication Information
7 Food Additives and Ingredients
8 Pharmacology and Biochemistry
9 Use and Manufacturing
10 Safety and Hazards
11 Toxicity
12 Associated Disorders and Diseases
13 Literature
14 Patents
15 Classification
16 Information Sources

Fig 3.4. Related Records of Penicillin

5 Chemical Vendors

Showing 1 Substance per Vendor View All

View in Entrez Download

3B Scientific (Wuhan) Corp
PubChem SID: 375108673 Purchasable Chemical: 3B3-059508

AA BLOCKS
PubChem SID: 381988336 Purchasable Chemical: AA01CC7G

Selleck Chemicals
PubChem SID: 404639703 Purchasable Chemical: S5515

BenchChem
PubChem SID: 446397691 Purchasable Chemical: B1679273

Smolecule
PubChem SID: 438485372 Purchasable Chemical: S538955

THE BioTek
PubChem SID: 434303065 Purchasable Chemical: bt-278434

Ambinter
PubChem SID: 364182307 Purchasable Chemical: Amb22801185

MedChemexpress MCE
PubChem SID: 405010564 Purchasable Chemical: HY-N7120

LGC Standards
PubChem SID: 340515512 Purchasable Chemical: LGCFOR0302.00

A2B Chem
PubChem SID: 444240261 Purchasable Chemical: AW55208

3WAY PHARM INC
PubChem SID: 438521790 Purchasable Chemical: SWT-00733

CymitQuimica
PubChem SID: 470666208 Purchasable Chemical: CQ 6130-64-9

Cite Download

CONTENTS

Title and Summary
1 Structures
2 Names and Identifiers
3 Chemical and Physical Properties
4 Related Records
5 Chemical Vendors
6 Drug and Medication Information
7 Food Additives and Ingredients
8 Pharmacology and Biochemistry
9 Use and Manufacturing
10 Safety and Hazards
11 Toxicity
12 Associated Disorders and Diseases
13 Literature
14 Patents
15 Classification
16 Information Sources

Fig 3.5. Chemical vendors available for Penicillin

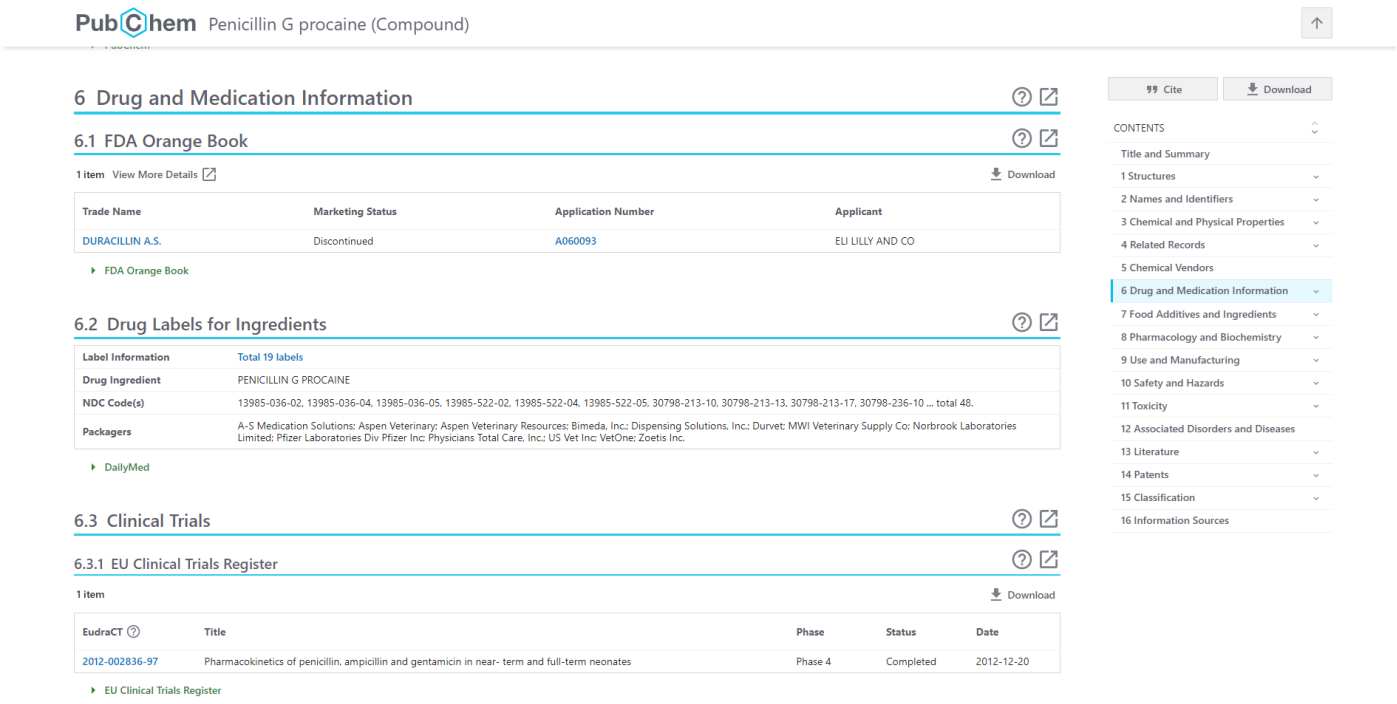

Fig 3.6. Drug and Medication information for Penicillin

Fig 4. Substructure search for Penicillin

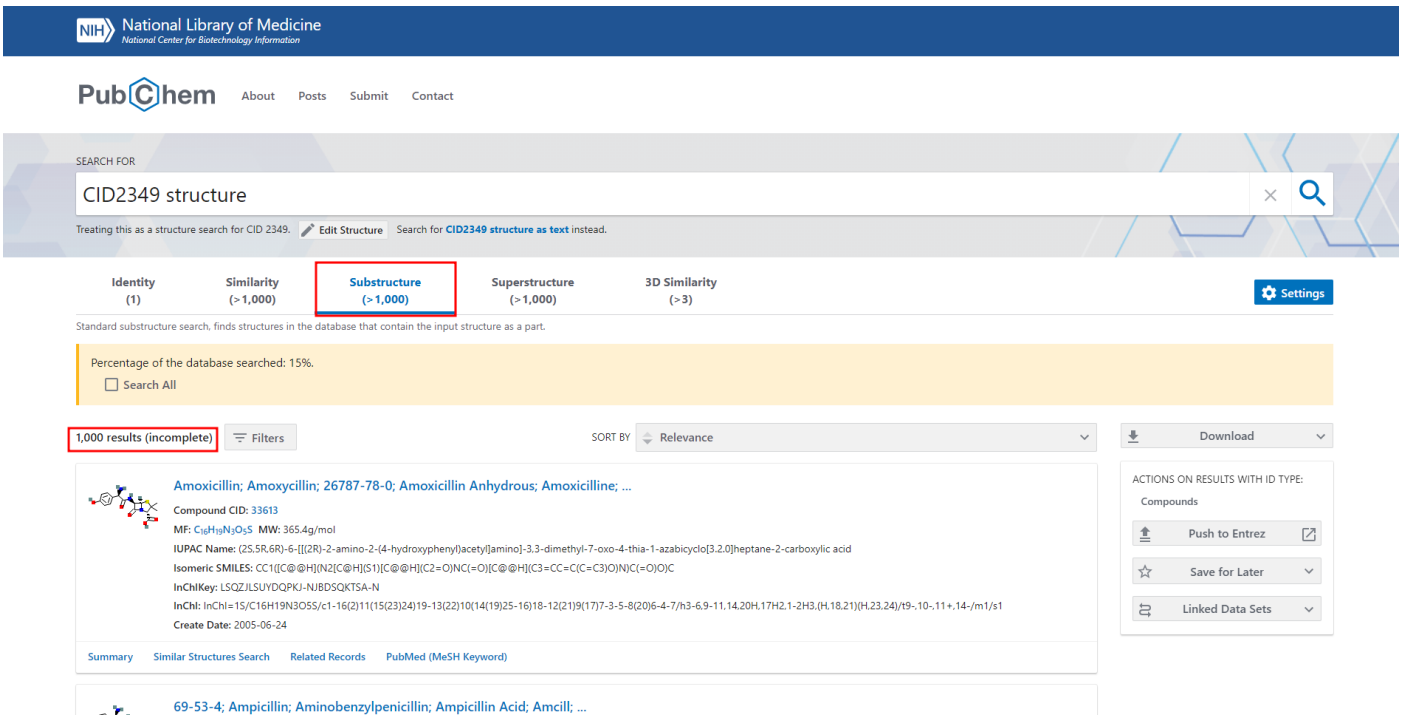

Fig 4.1 Hit page for substructure search for Penicillin

Structure search using exact search approach:

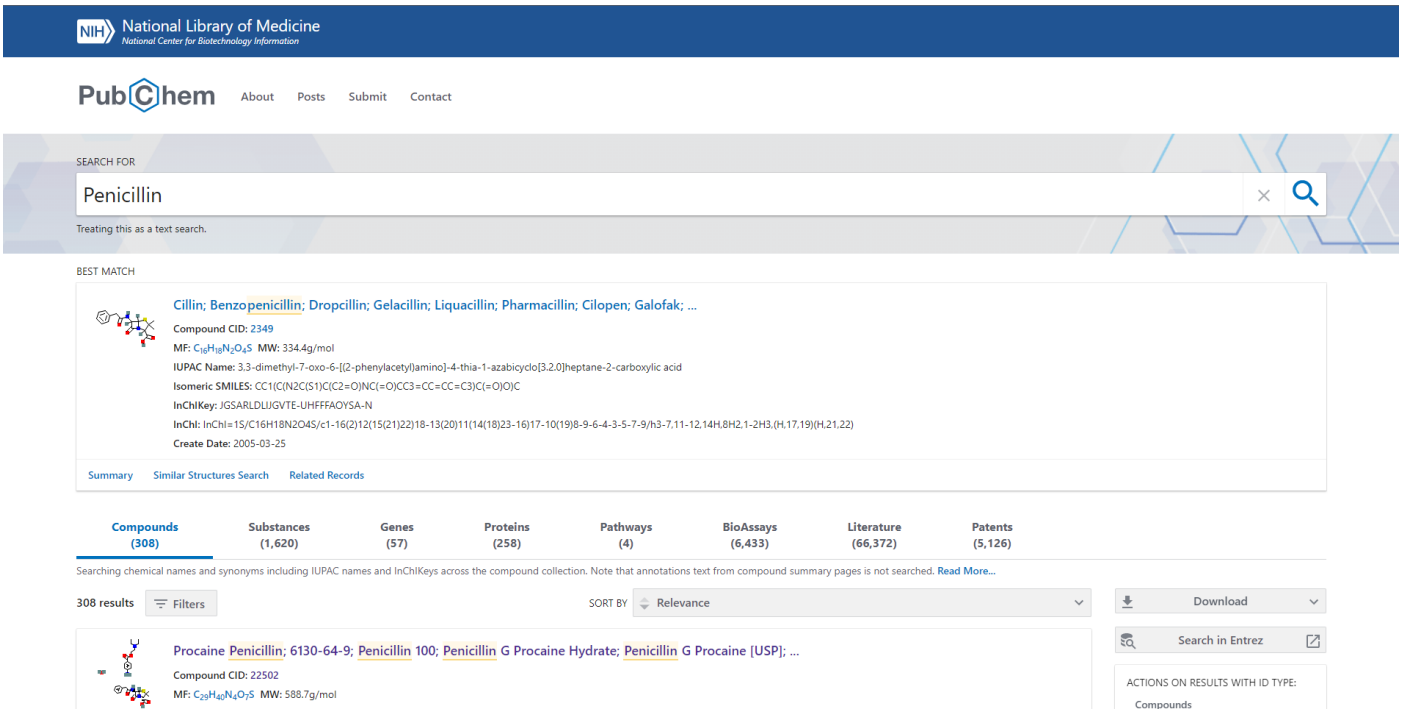

Fig 5. Exact search for Penicillin

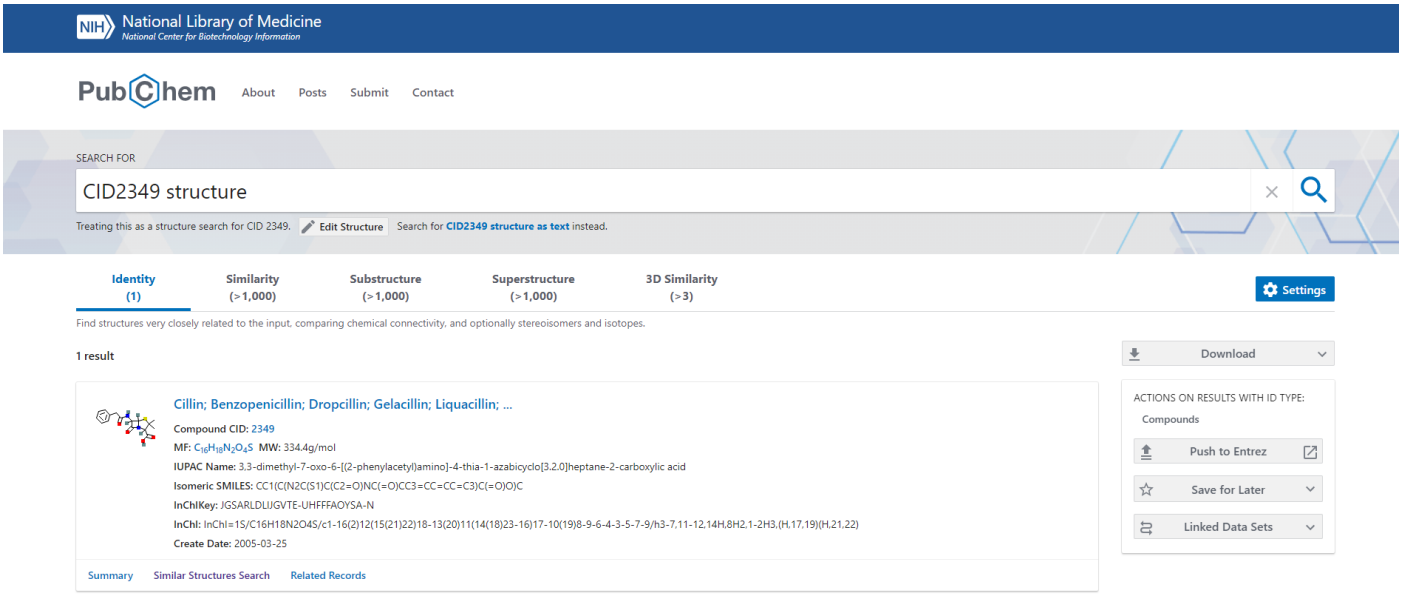

Fig 5.1 Hit page for exact Search for Penicillin

Structure search using similarity search approach:

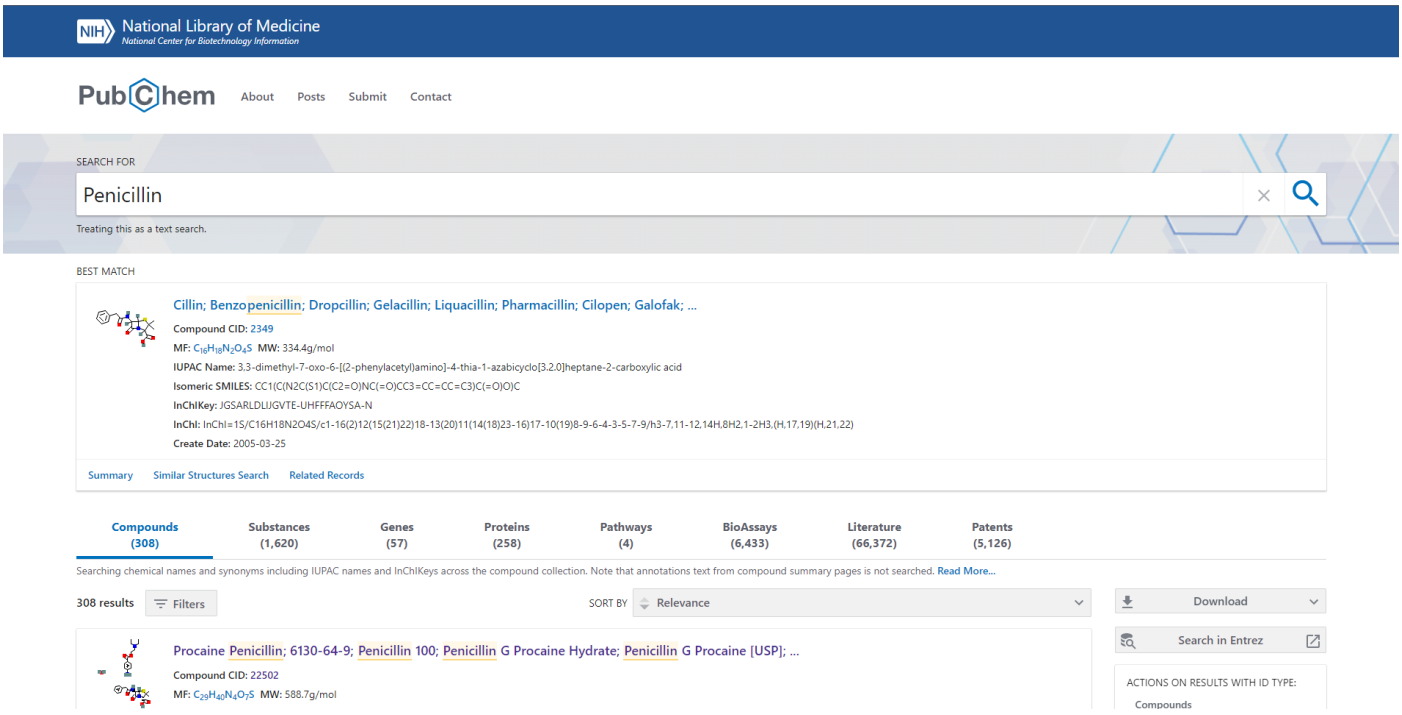

Fig 6. Similarity search for Penicillin

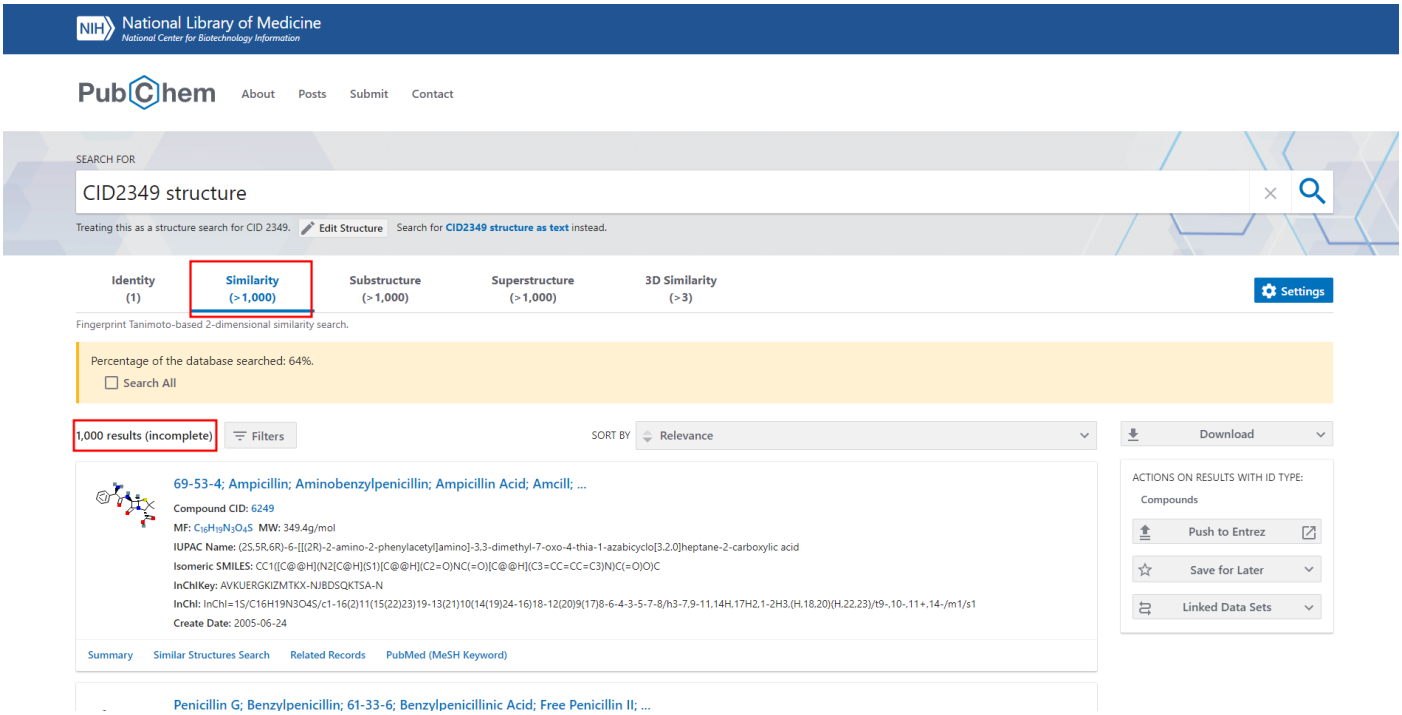

Fig 6.1. Hit page for similarity search for Penicillin

To screen structures based on chemical properties approach:

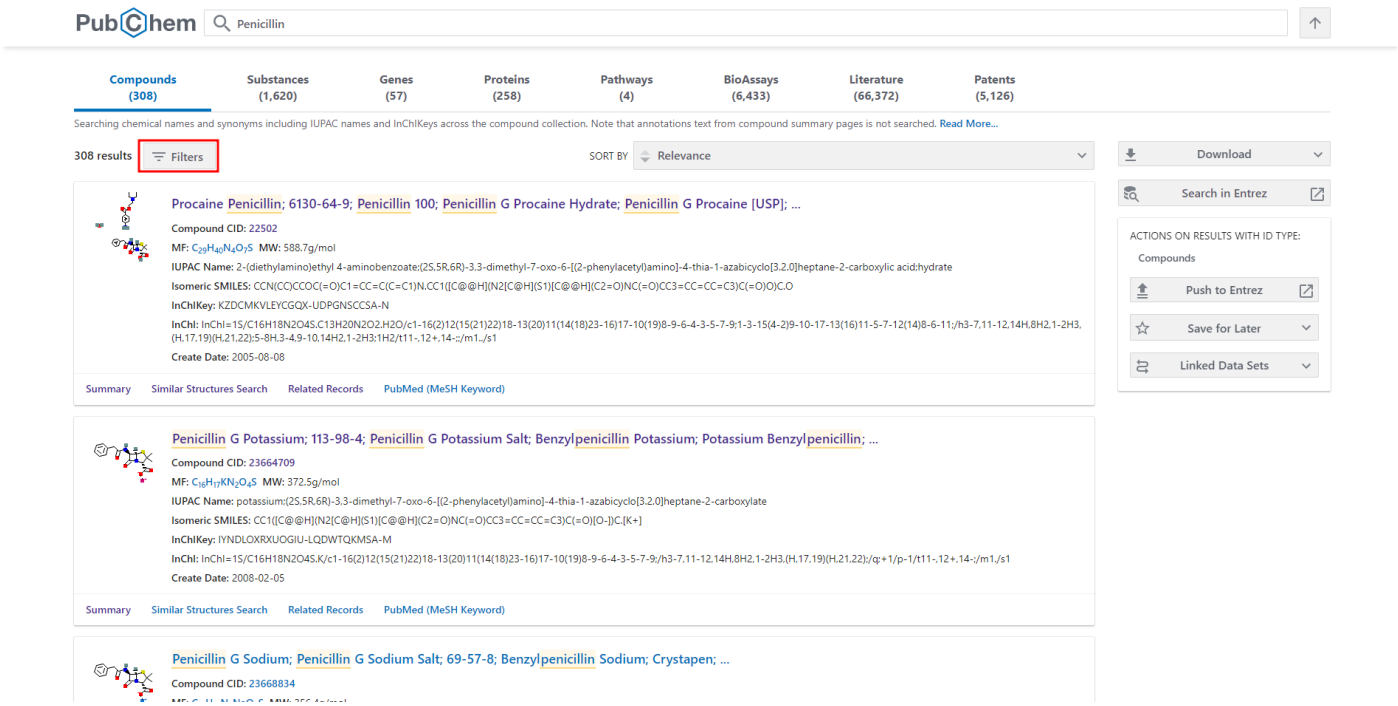

Fig 7. Applying filters to screen structure based on chemical properties

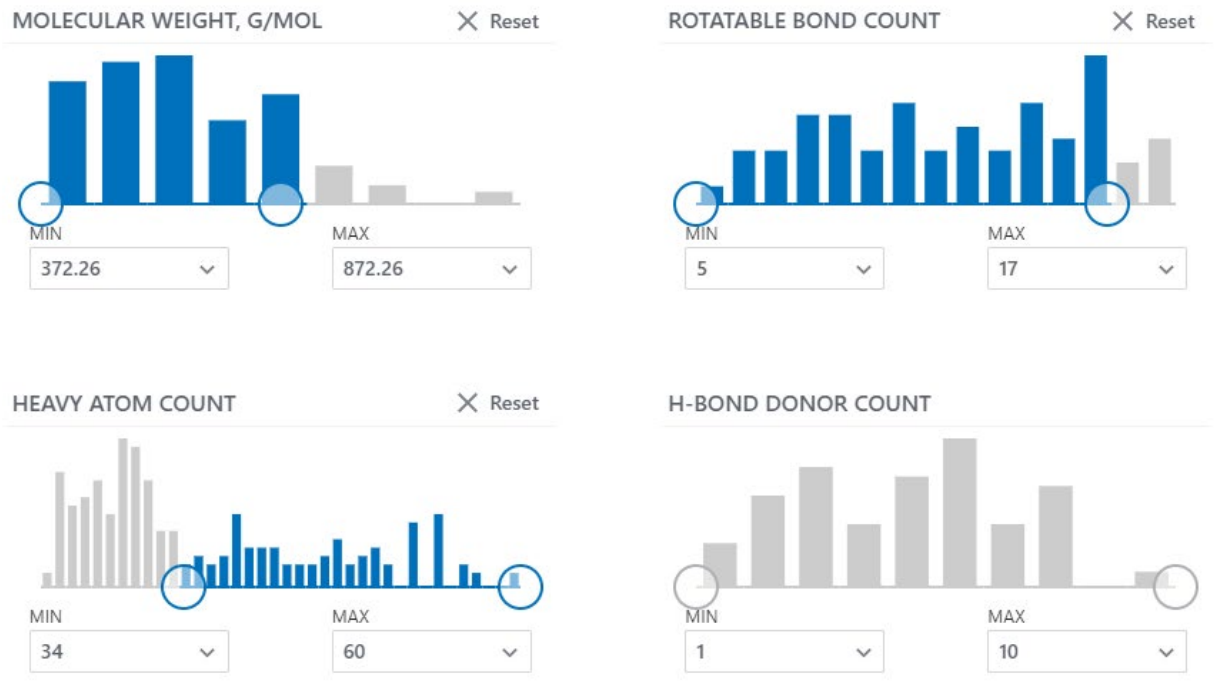

Fig 7.1 Filters applied to refinement based on chemical properties

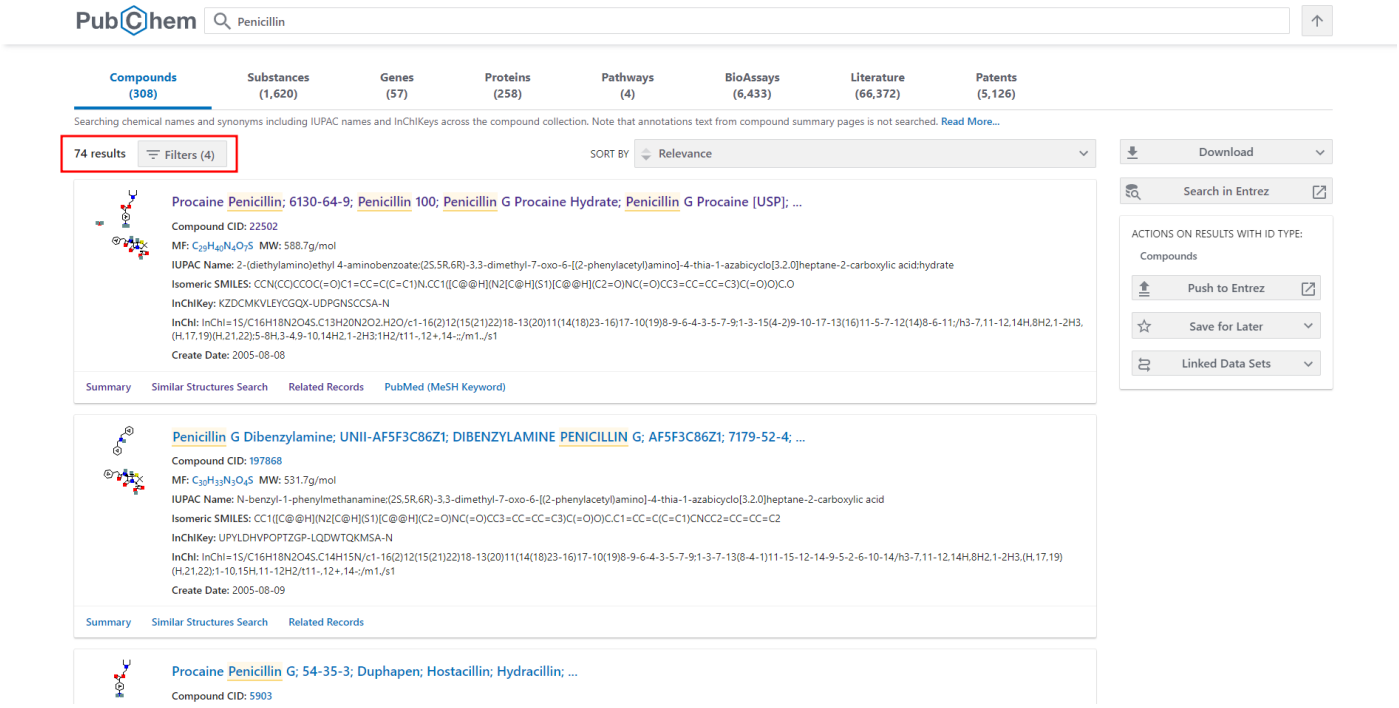

Fig 7.2. Hit page for results obtained after applying filters

Results:

2D/3D structure for Penicillin (Pubchem id - 22502) were retrieved using Pubchem Database. The results showed one best match and 308 similar compounds. The best match was used to refine the search based on substructure and similarity structure approach wherein both results showed <1000 compounds. Search refinement using exact search approach gave one identical compound. The 308 compounds retrieved earlier were filtered using chemical properties approach. Filters were applied for molecular weight, rotatable bond count and heavy atom count and the search was narrowed down to 74 compounds.

Compounds:

PubChem is a public repository for information on chemical substances and their biological activities which can be used to retrieve 2D/3D structures of chemical compounds. The database also provides various search options using different ways of representing molecular structures such as SMILES, InChl key, molecular formula, Pubchem CID or even by drawing the structure. It also provides various options to refine the search to retrieve structures that have substructure of query compounds, structures that are similar or the exact match. Filter can also be applied using chemical properties to narrow down the search results. Thus, Pubchem serves as a useful resource many areas such as cheminformatics, chemical biology, medicinal chemistry and drug discovery where there is requirement for information on chemical compounds.

References:

1. What is Data Filtering? - Definition from Techopedia. (2022). Retrieved 5 September 2022, fromhttps://www.techopedia.com/definition/26202/data-filtering
2. *Penicillin- an overview | ScienceDirect Topics*. (n.d.). Www.sciencedirect.com. Retrieved 5September 2022, from https://www.sciencedirect.com/topics/neuroscience/penicillin
3. PubChem. (2022). *The PubChem Project*. Nih.gov; National Library of Medicine. Retrieved 5September 2022, from https://pubchem.ncbi.nlm.nih.gov/
4. PubChem. (2022). *The PubChem Project*. Nih.gov; National Library of Medicine. Retrieved 5September 2022, from https://pubchem.ncbi.nlm.nih.gov/#query=penicillin
5. PubChem. (2022). *The PubChem Project*. Nih.gov; National Library of Medicine. Retrieved 5September 2022, from https://pubchem.ncbi.nlm.nih.gov/compound/22502