DATE: 05/09/22

WEBLEM 1

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

IMPORTANCE OF 3D STRUCTURES:

The 3-dimensional (3D) structure of therapeutics and other bioactive molecules is an important factor in determining the strength and selectivity of their protein–ligand interactions. Previous efforts have considered the strain introduced and tolerated through conformational changes induced upon protein binding.

The recognition of novel therapeutics by their targets is dictated by ligand–protein interactions. Noncovalent interactions, including hydrogen and halogen bonding, salt bridges, and pi–pi stacking, are typically explored through the development of structure–activity relationships (SAR) to produce potent and selective drugs. In addition to many protein–ligand interactions that can be optimized, the conformation of inhibitors in the solvated vs bound states can also contribute to the energetic favorability of inhibitor binding. The diversity of druggable protein targets necessitates structural and conformational variability in ligands to generate effective pharmaceuticals. The amount of molecular strain and associated energy costs tolerated by drugs upon protein binding has been explored, where it was found that molecules binding to proteins could readily incur 5–9 kcal/mol of strain energy. While extensive rearrangements are possible, increased conformational shape diversity has been related to broad biological activity and successful clinical outcomes. The accessibility and large number of structures in the Protein Data Bank (PDB) has enabled the modeling and docking of 3-dimensional (3D) structures into protein active sites. These observations and tools have motivated the preparation of libraries with greater topological diversity for drug-discovery campaigns with the aim of improved lead identification and therapeutic success.

METHOD OF GENERATION FROM 1D AND 2D REPRESENTATION FOR DRUG DESIGNING AND CHEMOINFORMATICS STUDIES:

Mapping Chemical Graphs into 1D Descriptors:

Mapping to a 0D representation substantially decreases information on chemical compounds, and 0D descriptors are usually too coarse to code molecules. Instead, 1D descriptors such as:

• Systematic nomenclature

• Line notation

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

The concept of a systematic nomenclature scheme (chemical name) by naming rules is human-oriented, which means that easy chemical nameability and name readability by human chemists have been given high priority. However, the systematic nomenclature such as the IUPAC system quite often leads to rather complicated names reflecting the complexity of chemical structures. A human-friendly nomenclature does not necessarily meet the requirements of a computer-oriented system. For example, restricting the name generated for a single structure to a unique value is still a challenge and is to be solved by a preferred name program (PNP) currently under consideration. The first system that allowed the input of chemical structures in the form of their chemical names was developed by Beilstein in 1986. This was operated internally at the Beilstein Institute, and the structure input was restricted to the Beilstein notation sub-rules. Currently, the systematic nomenclature is a fully reversible system with software, allowing the mapping of a molecule to a name and a name to a molecule structure.

A line notation is another important system that maps a 2D molecular graph into a linear sequence of letters and numbers. The first systems for this mapping were developed even before computer applications, and a peak of success was in the 1960s and 1970s when this provided a rapid system for coding even large molecules.

In a computer system, this was especially suitable due to the compactness of data storage requirements, an important problem at that time. Among historically important line notations:

• Wiswesser (WLN)

• ROSDAL

• SYBYL (SLN)

• SMILES

the latter one, SMILES, evidently predominates today. This is caused by its very simple semantic rules and the intuitive nature of the system. In Fig. 3, we show some illustrative examples. SMILES codes support any chemical entry in Wikipedia, and almost each molecular editor can generate SMILES. Vice versa, SMILES can replace a molecular input into a variety of databases, e.g., Reaxys and/or molecular editors. A smell of success for this system is in the air if we realize that a Google search can be performed today directly by a SMILES input.

Mathematically, a line notation performs a reduction of information from a 2D into a 1D representation. Dimensionality reduction, while preserving 2D information, explains the success here. Lower dimension means simplification in data manipulation. An important remark is that this dimensionality reduction is fully reversible. So we can both obtain a SIMLES for any molecule and remap this back to reproduce the original molecule.

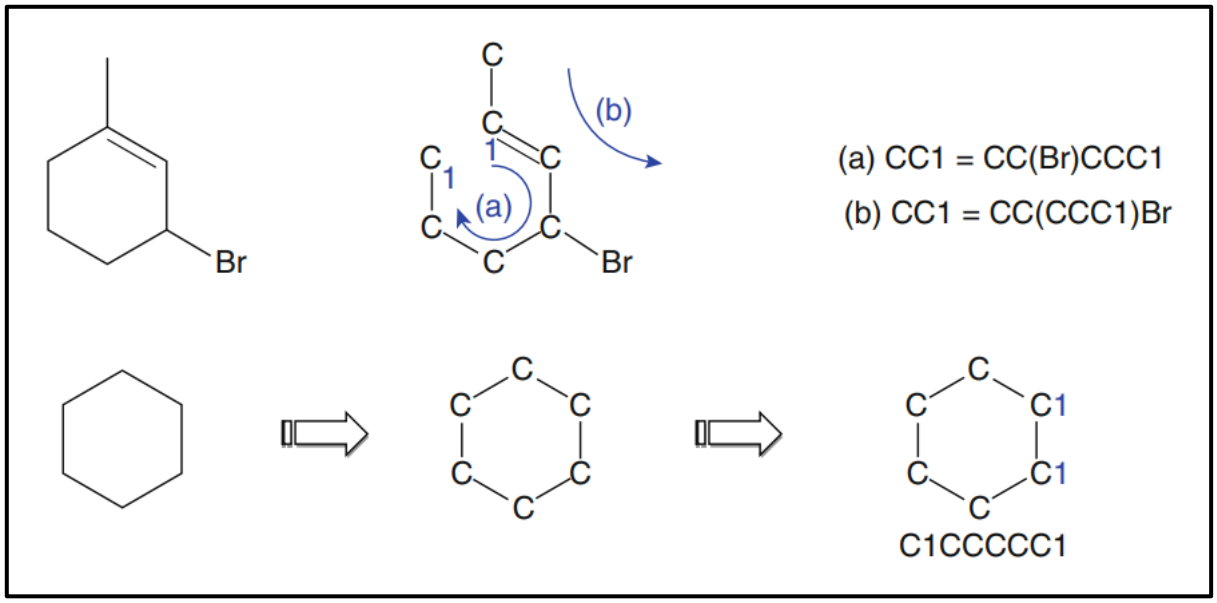


FIG 1. Example for 1D representation

Mapping Chemical Graphs into 2D Descriptors:

Molecular graphs were developed for coding molecules for the needs of human chemists; however, they appeared impractical for feeding computers with molecular data. Matrix notations or linear codes are the methods developed for mapping molecular graphs into the computer-friendly data. A variety of matrix systems (2D descriptors) were designed, where the numeric representation depends upon the scheme used for coding atoms and bonding. Popular notations include the following matrices:

• Adjacency

• Atom connectivity

• Distance (topological or geometrical)

• Bond

• Bond electrons

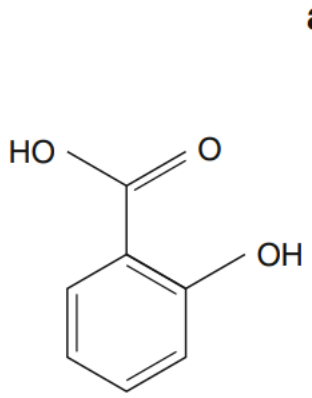
• Incidence

Although a matrix system performs well, the number of entries needed to describe a molecule squares with the increase in the number of atoms in the molecule, making it impractical for storing large molecules or large datasets.

A connection table (CT) is an alternative method, allowing us to present molecular graphs in a form of lists of atoms and bonds in a molecule which increase linearly with the number of atoms. A number of CT versions are available, and the reader is referred to Gasteiger (2003) for a detailed introduction.

A CT notation has now become the predominant form of representing molecular structures. Standard notations such as Molfiles and SDFiles have been developed that are globally used for information exchange, e.g., in Fig. 6, we illustrate a Molfile coding the 3D structure of salicylic acid by also giving the 3D coordinates of the atoms. In effect, the international language of chemical graphs has been translated into an internationally standardized computer-readable form. (As these representations are coded in ASCII, they can also be read by a human.) An interesting example of coding vitamin C, a chemical compound with all tautomers, by a CT can be found in Bobach et al. (2012).

Today, graphical molecule editors have been developed as tools for the direct creation, modification, or inputting of molecular graph 2D information. Vice versa tools are available for obtaining molecular graphs as outputs from computerstored CTs.



a

n	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	1	0	0	0	0
2	1	0	1	0	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0	0
4	0	0	1	0	1	0	0	0	0	0
5	0	0	0	1	0	1	0	0	0	1
6	1	0	0	0	1	0	1	0	0	0
7	0	0	0	0	0	1	0	1	1	0
8	0	0	0	0	0	0	1	0	0	0
9	0	0	0	0	0	0	1	0	0	0
10	0	0	0	0	1	0	0	0	0	0

FIG 2. Example of 2D representation

Mapping Chemical Graphs into 3D Descriptors:

A 2D molecular representation defines the molecular topology (atomic constitution and connectivity). However, this in its cleanest form does not code stereochemistry, i.e., the spatial arrangement of the atoms around an atomic center called atomic configuration. This can be coded within chemical graphs by enriching them with solid and dashed wedged bonds, indicating above-the-plane or below-the-plane pointing of the bonds. The configuration is a simplified 3D molecular representation indicating a version of the atomic arrangement (stereoisomer) but not the real 3D atomic coordinates.

Stereochemical codes in chemistry are developed to achieve a significant simplification in the visualization of 3D molecules in 2D for the human chemist. Stereochemical information is somewhere in between topology (2D) and topography (3D).

Basically, two possible representations for the real 3D structure are available. First are 3D atomic coordinates measured as the actual property of compounds (substances), e.g., through crystallography. Accordingly, such data are available in databases (compare section "Databases and Database Searching"). Second, there are predictions of 3D structures based on molecular modeling. In the second case, this operation can be explained as a mapping of a 2D (CT) or a 1D (SMILES) molecular representation into 3D (see section "Representation of 3D Structures and Their Mapping into Lower Dimensionality"). Technically, 3D structures can be noted in several data formats:

• SDfile (Structure Data) *.sdf

• Molfile (Connection Table Format) *.mol

• CIF (Crystallographic Information File) *.cif

• PDB (proteins) *.pdb

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

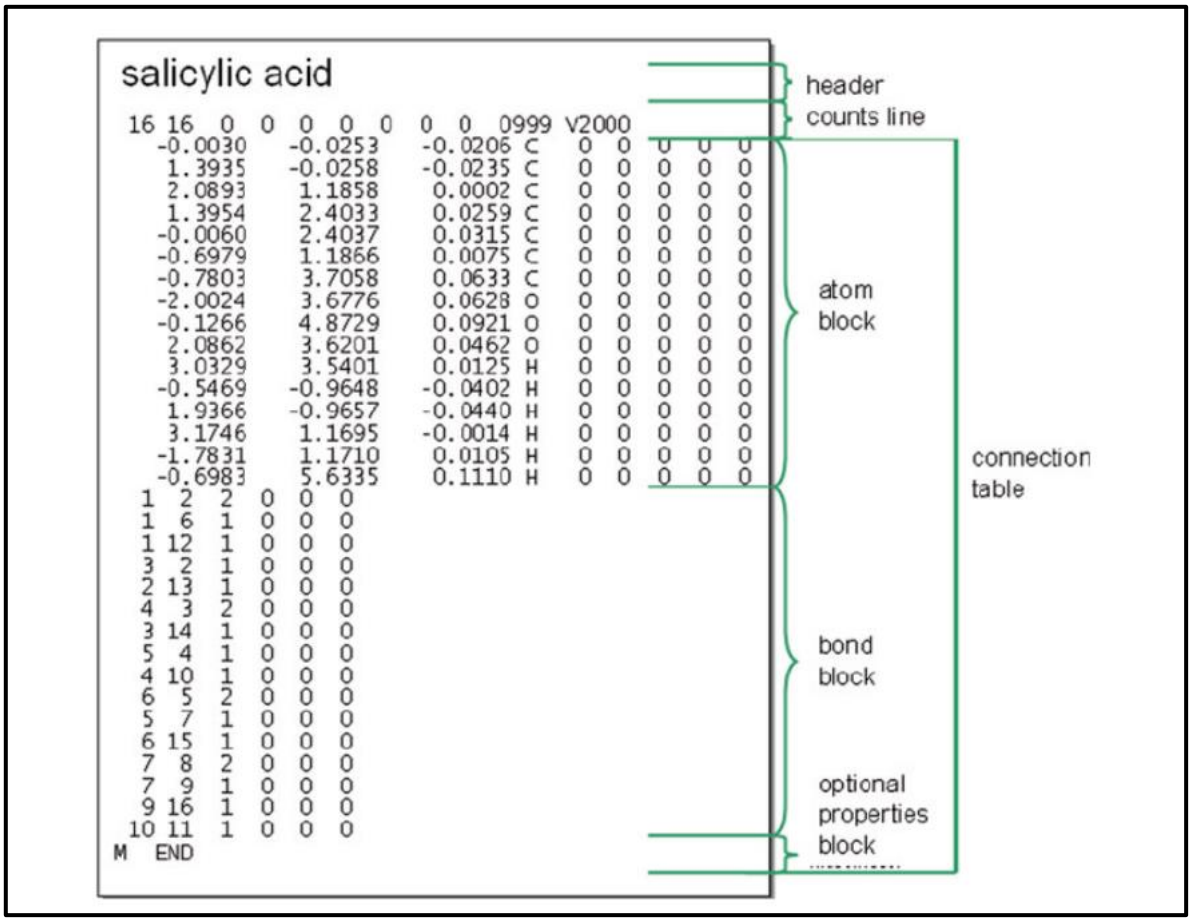


FIG 3. Example of 3D reprentation

PUBCHEM DATABASE:

Introduction and history:

PubChem is a public repository for information on chemical substances and their biological activities. Since launched in 2004 as a component of the Molecular Libraries Roadmap Initiatives of the US National Institutes of Health (NIH), PubChem has rapidly grown to a key chemical information resource that serves scientific communities in many areas such as cheminformatics, chemical biology, medicinal chemistry and drug discovery.

PubChem contains one of the largest corpus of publicly available chemical information. As of September 2015, it has more than 157 million depositor-provided chemical substance descriptions, 60 million unique

chemical structures and 1 million biological assay descriptions, covering about 10 thousand unique protein target sequences. PubChem organizes this vast amount of data into three inter-linked databases: Substance, Compound and BioAssay. The Substance database stores depositor-contributed information. Unique chemical structures are extracted from the Substance database and stored in the Compound database. The BioAssay database stores descriptions of biological assays on chemical substances. The primary identifiers for the Substance, Compound and BioAssay databases are SID (SubstanceID), CID (CompoundID) and AID (AssayID), respectively.

Working:

Web interfaces for textual search:

Entrez is the search and retrieval system used for PubChem's three primary databases and other major NCBI databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Genome, Taxonomy, BioSystems, Gene Expression Omnibus (GEO) and many others. One can search the PubChem databases through Entrez by initiating a search from the PubChem home page, which also provides launch points to various PubChem services, tools, help documents and more. Alternatively, one can begin the search from the NCBI home page. By default, if a specific database is not selected in the search menu, Entrez searches all Entrez databases available and lists the number of records in each database that are returned for this 'global query'. Simply by selecting one of the three PubChem database from the global query result page, one can see the query result specific to that database.

If an Entrez search returns multiple records, they are displayed in a document summary (DocSum) report. For each record in the DocSum page, some data-specific information is provided with a link to the Summary page for that record. The DocSum page contains controls to change the display type, to sort the results by various means, or to export the page to a file or printer. In addition, the icons and links on the right column of the DocSum page allow users to perform further analysis on the query result, to download the corresponding records, to refine or modify the search, to obtain associated records in other databases and so on.

If a search against the Compound database returns a single record, the Compound Summary page for that record is displayed. The Compound Summary page provides a comprehensive view that recaps all information known about a particular chemical, collected from different data sources. If a search against the Substance database returns a single substance, the Substance Summary page for that substance is displayed, which shows information provided by the data contributor for that record.

Non-textual search using the chemical structure search tool:

Because Entrez is primarily a text-based search system, it cannot be used for searching that involves data types specific to PubChem, such as chemical structures. The Chemical Structure Search tool enables one to query and subset the Compound database using various chemical structure search types, including identity search, substructure/superstructure search, molecular formula search and 2-D and 3-D similarity searches.

The Chemical Structure Search tool supports a variety of query formats, including SMILES, SMARTS, InChI, CID, molecular formula and SDF. One can also manually draw a query chemical structure using the PubChem Chemical Structure Sketcher. This JavaScript-based structure editor is platform-independent and compatible with major web browsers, and does not require the user to download or install special software. In addition, it contains import and export features such as support for chemical structure files.

The Chemical Structure Search tool allows users to narrow a search to the result from a previous Entrez or chemical structure search or to the set of CIDs uploaded in a file. Optional filters may be applied to limit the search result, based on various properties, such as molecular weight, heavy atom count, presence or absence of stereochemistry, depositor name or category and so on. A query can be exported to an XML file, which allows one to import the query from the XML file and to repeat the search without filling out the search form again. This XML file can also serve as an example for constructing queries for the PUG interface.

Applications:

The massive volume of chemical structure and bioactivity data in PubChem and its online services has been used globally in various fields including chemical biology, medicinal chemistry and informatics research. PubChem supports drug discovery in many aspects such as lead identification and optimization, compound–target profiling, polypharmacology studies and unknown chemical identity elucidation. PubChem has also become a valuable resource for developing secondary databases, informatics tools and web services. The growing PubChem resource with its public availability offers support and great opportunities for the interrogation of pharmacological mechanisms and the genetic basis of diseases, which are vital for drug innovation and repurposing.

Chemoinformatics is a relatively new principle of chemistry and is based upon the processing of data concerning chemical and molecular structures through the use of computational analysis. The analysis of these data allows the relationship between chemical structure, chemical properties, and molecular activity to be studied. It is an in-silico technique, which means it is a form of scientific study which is performed virtually on a computer via software and simulations.

The normal process of drug discovery entails selecting a disease to target, then searching for potential compounds and molecules which can be used to reduce the severity of the disease in some way. This is done through many stages of screening, which normally compare the effectiveness of these potential molecules to stop a biochemical mechanism. Chemoinformatics can drastically enhance this process, as one of the principal applications of chemoinformatics in research is the discovery and development of drugs. There are many techniques available in order to achieve this, and the use of software to calculate and visualize structures is crucial.

In order to reduce costs and speed up drug discovery when screening for new potential compounds that could be developed into drugs, virtual screening can be used to filter out certain compounds early on that aren't compatible without the need for physical screening. This method uses computer software to build virtual screens and simulations which can check for potential molecules that have the potential to be developed into drugs with much higher efficiency than conventional methods. This is where 3D structures of compounds as well as their representation in 1D and 2D formats is important in drug discovery and chemoinformatics to find new potential compounds to be used as drugs.

REFERENCES:

1. 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
2. Polanski, J., & Gasteiger, J. (2017). Computer Representation of Chemical Compounds. *Handbook Of Computational Chemistry*, 1997-2039. doi: 10.1007/978-3-319-27282-5_50
3. 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
4. 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
5. Jack Davis, B. (2022). How is Chemoinformatics Used in Drug Discovery?. Retrieved 5 September 2022, 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

DATE: 05/09/22

WEBLEM 1a

Retrieve 2D/3D structures using Pubchem Database

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

AIM:

To retrieve 2D/3D structure for Quercetin (Pubchem id-5280343) using Pubchem Database and apply filters as:

(i) To refine search using Substructure, Exact and Similar structure approach.
(ii) To screen structures based on chemical properties approach.

INTRODUCTION:

Quercetin is a natural flavonoid found abundantly in vegetables and fruits. There is growing evidence suggesting that quercetin has therapeutic potential for the prevention and treatment of different diseases, including cardiovascular disease, cancer, and neurodegenerative disease. Mechanistically, quercetin has been shown to exert antioxidant, anti-inflammatory, and anticancer activities in a number of cellular and animal models, as well as in humans through modulating the signaling pathways and gene expression involved in these processes.

PubChem is a public repository for information on chemical substances and their biological activities. Since launched in 2004 as a component of the Molecular Libraries Roadmap Initiatives of the US National Institutes of Health (NIH), PubChem has rapidly grown to a key chemical information resource that serves scientific communities in many areas such as cheminformatics, chemical biology, medicinal chemistry and drug discovery.

PubChem contains one of the largest corpus of publicly available chemical information. As of September 2015, it has more than 157 million depositor-provided chemical substance descriptions, 60 million unique chemical structures and 1 million biological assay descriptions, covering about 10 thousand unique protein target sequences. PubChem organizes this vast amount of data into three inter-linked databases: Substance, Compound and BioAssay. The Substance database stores depositor-contributed information. Unique chemical structures are extracted from the Substance database and stored in the Compound database. The BioAssay database stores descriptions of biological assays on chemical substances. The primary identifiers for the Substance, Compound and BioAssay databases are SID (SubstanceID), CID (CompoundID) and AID (AssayID), respectively.

Filtering for data is important because redundant or impartial pieces of data can confuse or disorient a user. Filtering data can also make results more efficient. In some other cases, data filters work to prevent wider access to sensitive information.

METHODOLOGY:

- Open homepage for Pubchem database. (URL: https://pubchem.ncbi.nlm.nih.gov/)
- Search for compound “Quercetin”.
- Open result for best match.
- Refine results obtained for Quercetin using Substructure, Exact and Similar structure approach.
- Screen results for quercetin using chemical properties approach.
- Observe and interpret the results.

OBSERVATIONS:

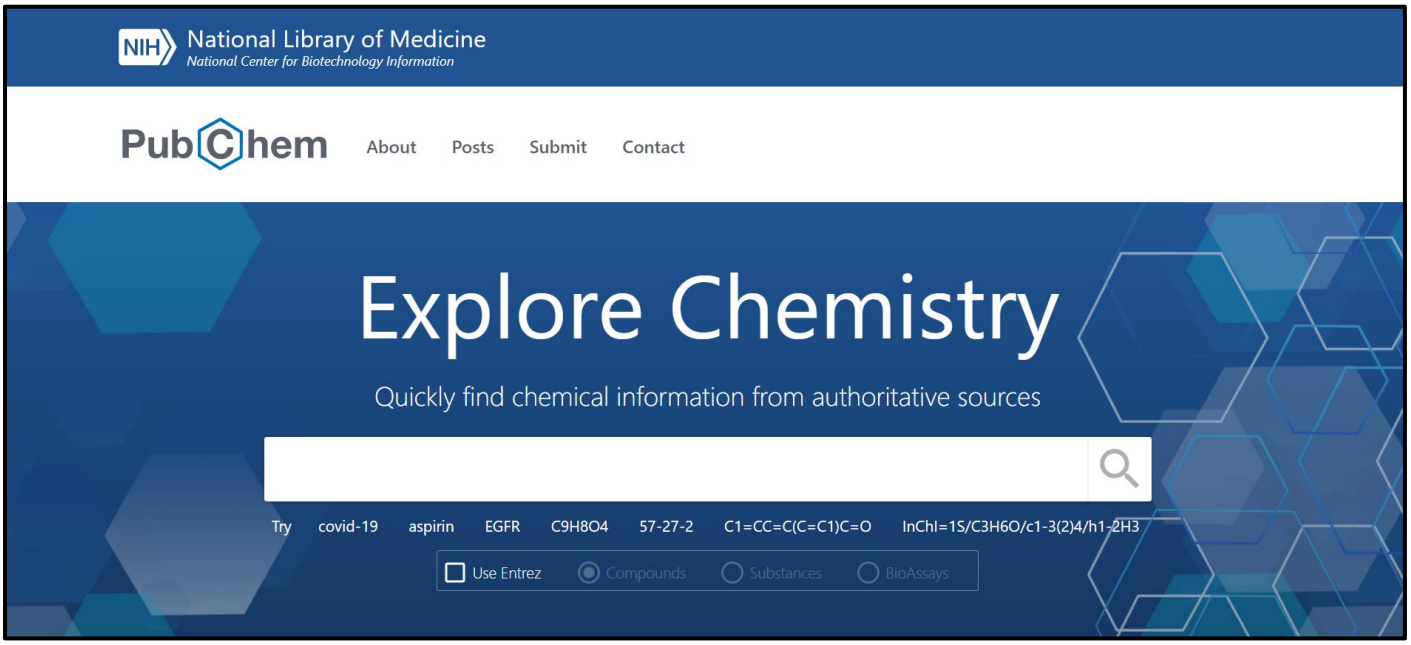


FIG 1. Homepage of Pubchem Database

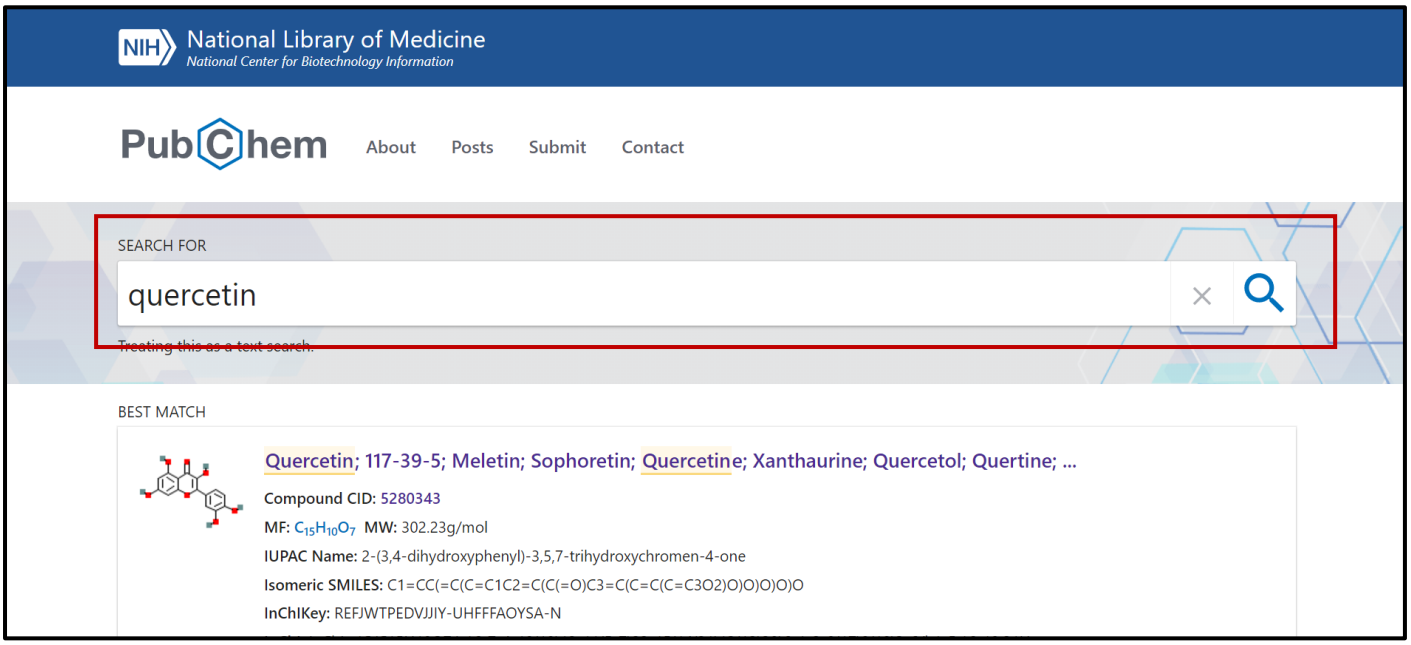


FIG 2. Hit page for compound “Quercetin”

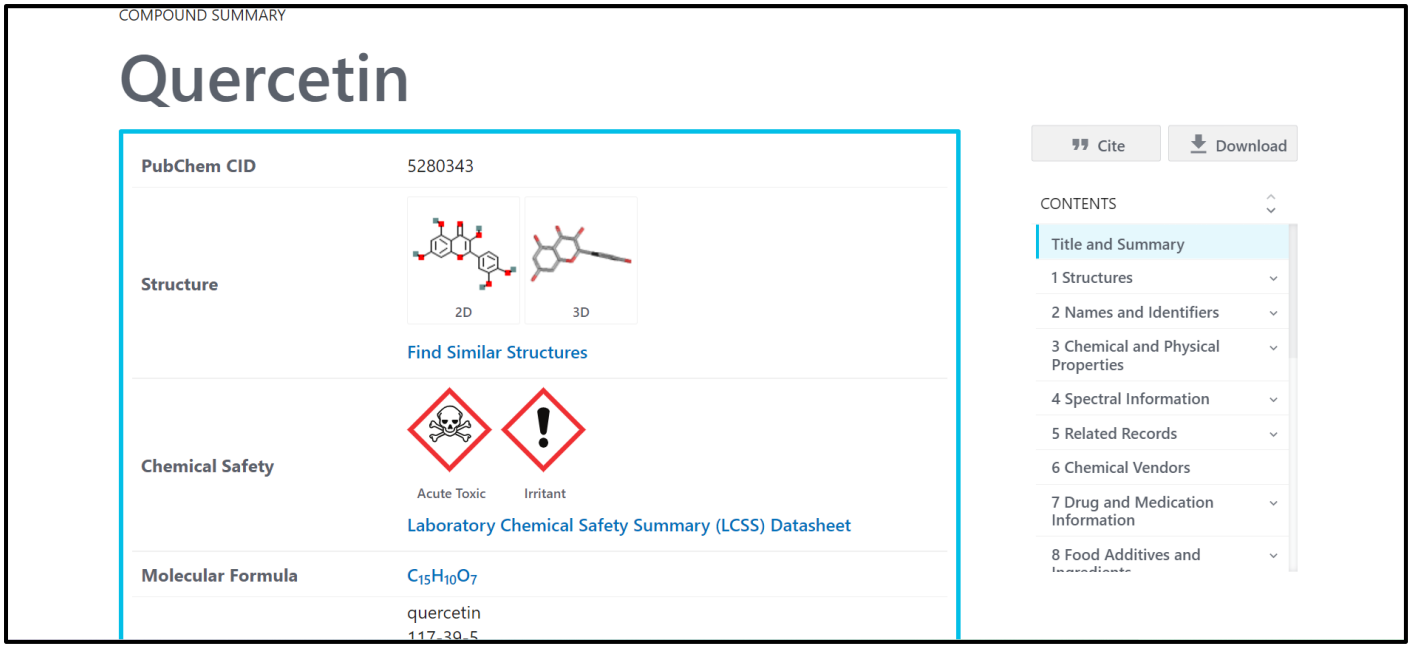


FIG 3. Result page with summary for Quercetin

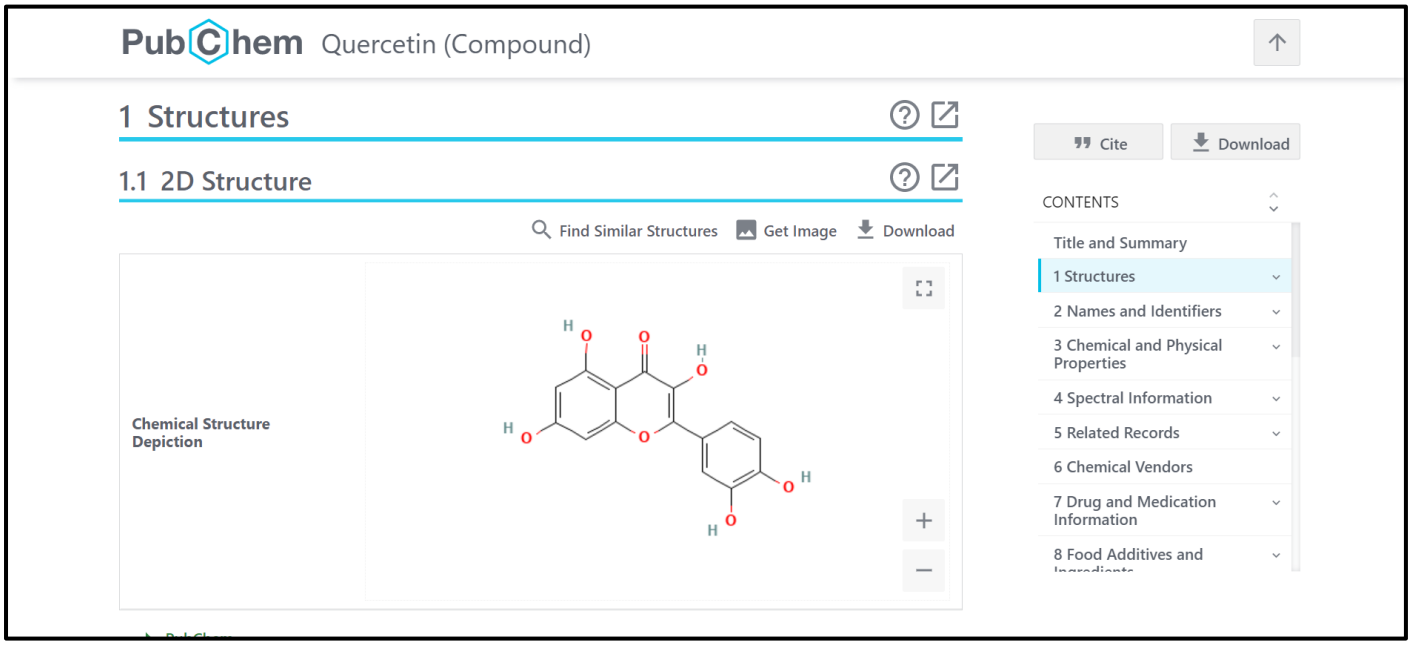


FIG 3.1. Structure information for Quercetin

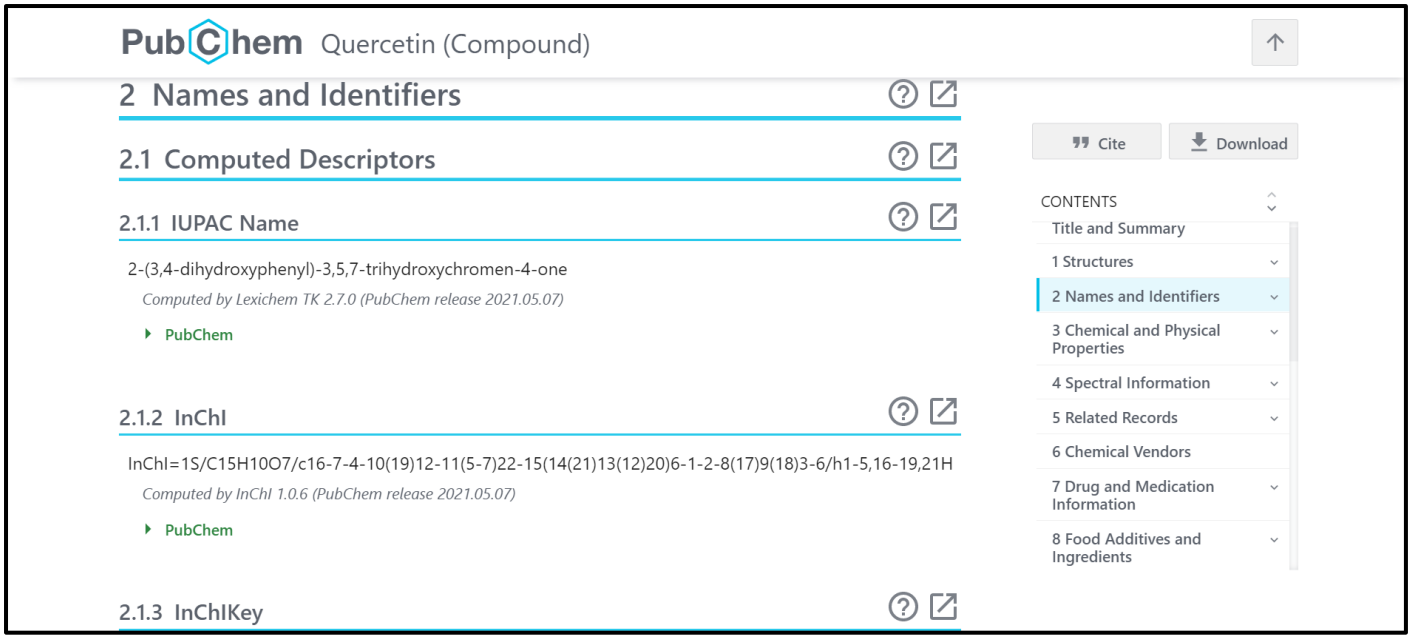


FIG 3.2. Names and Identifiers for Quercetin

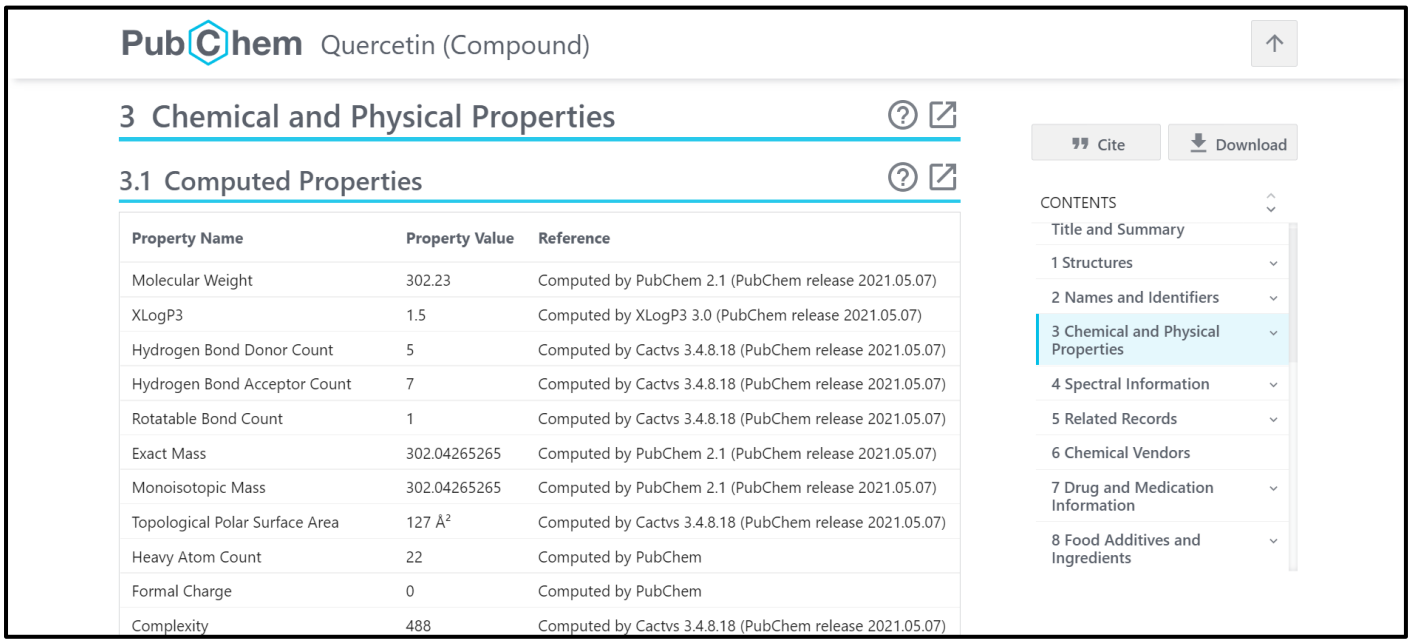


FIG 3.3. Chemical and physical properties of Quercetin

FIG 3.4. Spectral information for Quercetin

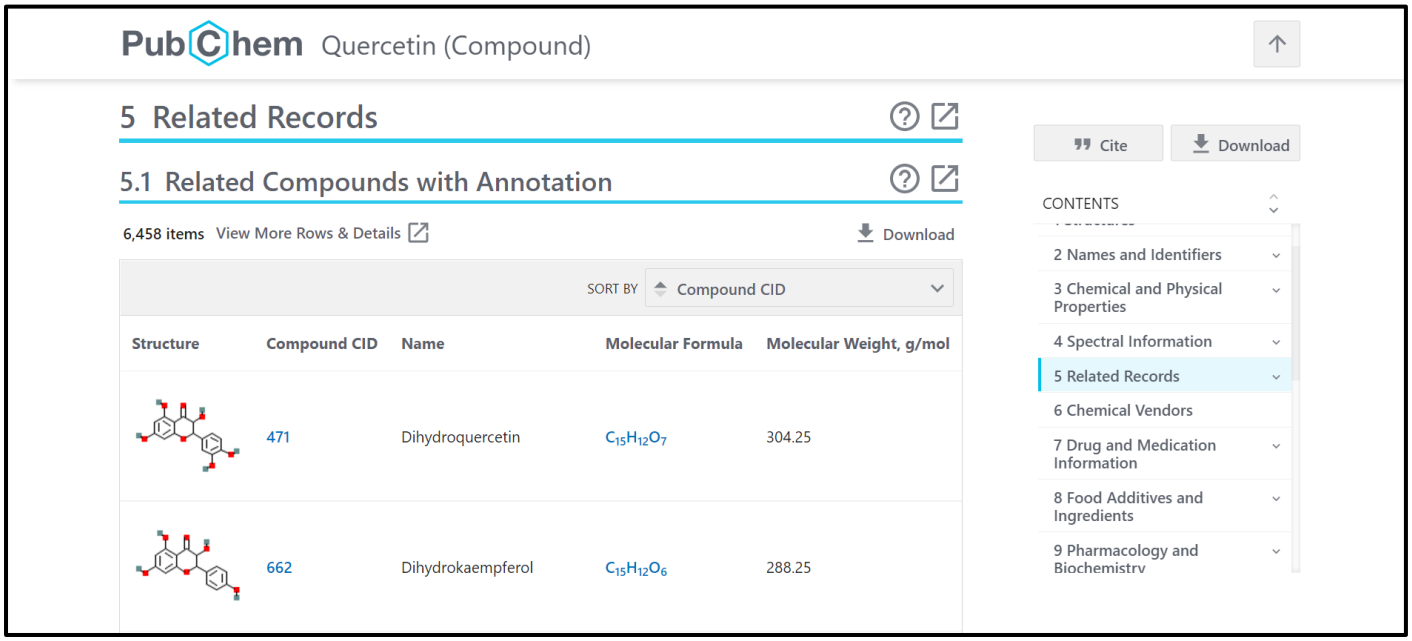


FIG 3.5. Related records information for Quercetin

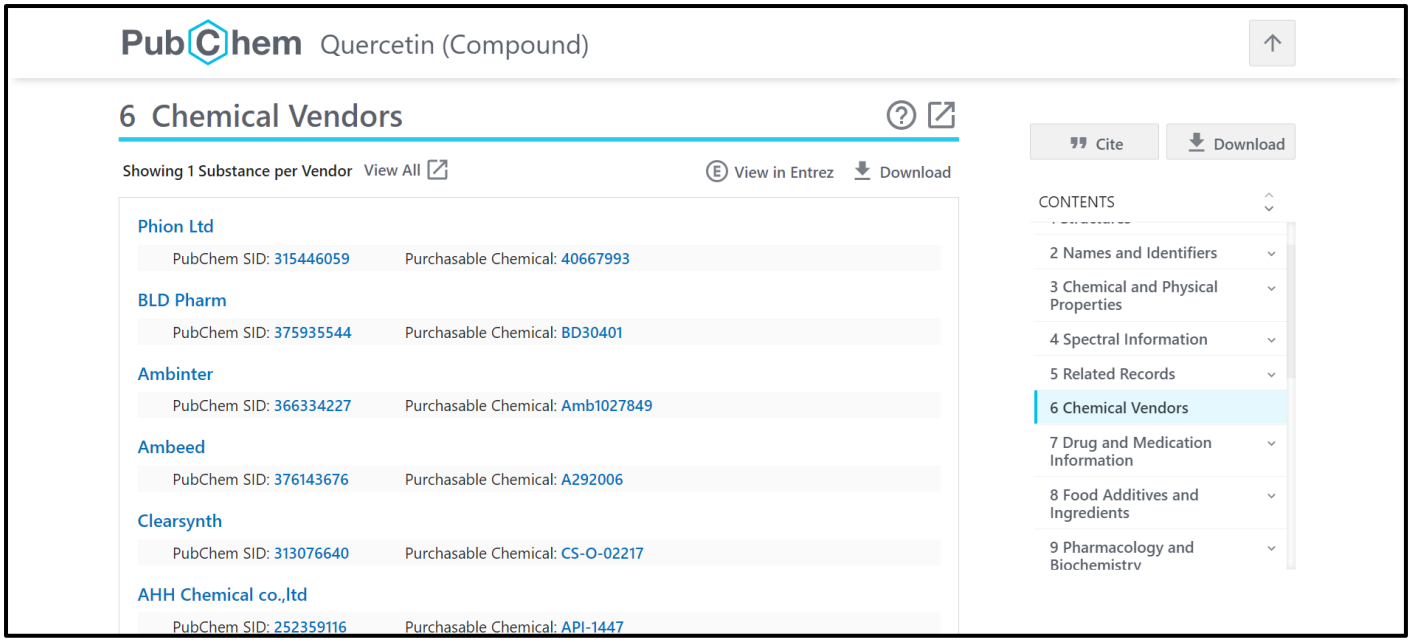


FIG 3.6. Chemical vendors available for Quercetin

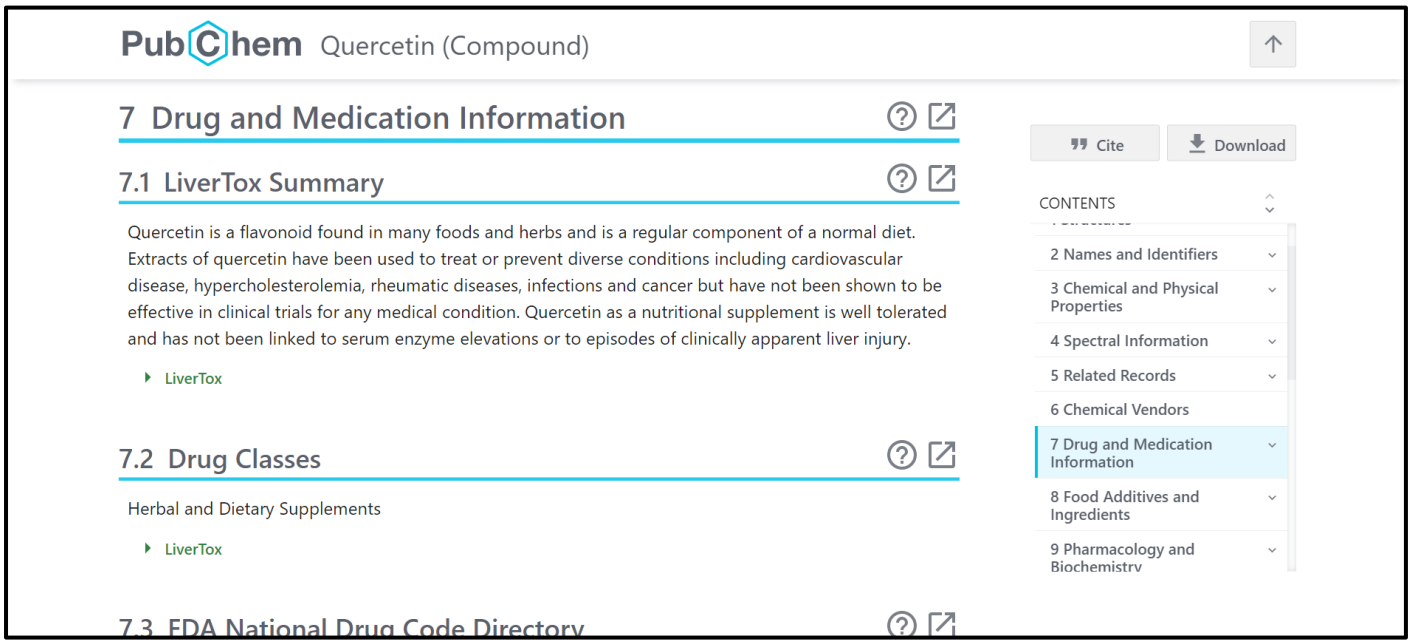


FIG 3.7. Drug and Medication information for Quercetin

Structure search using substructure approach:

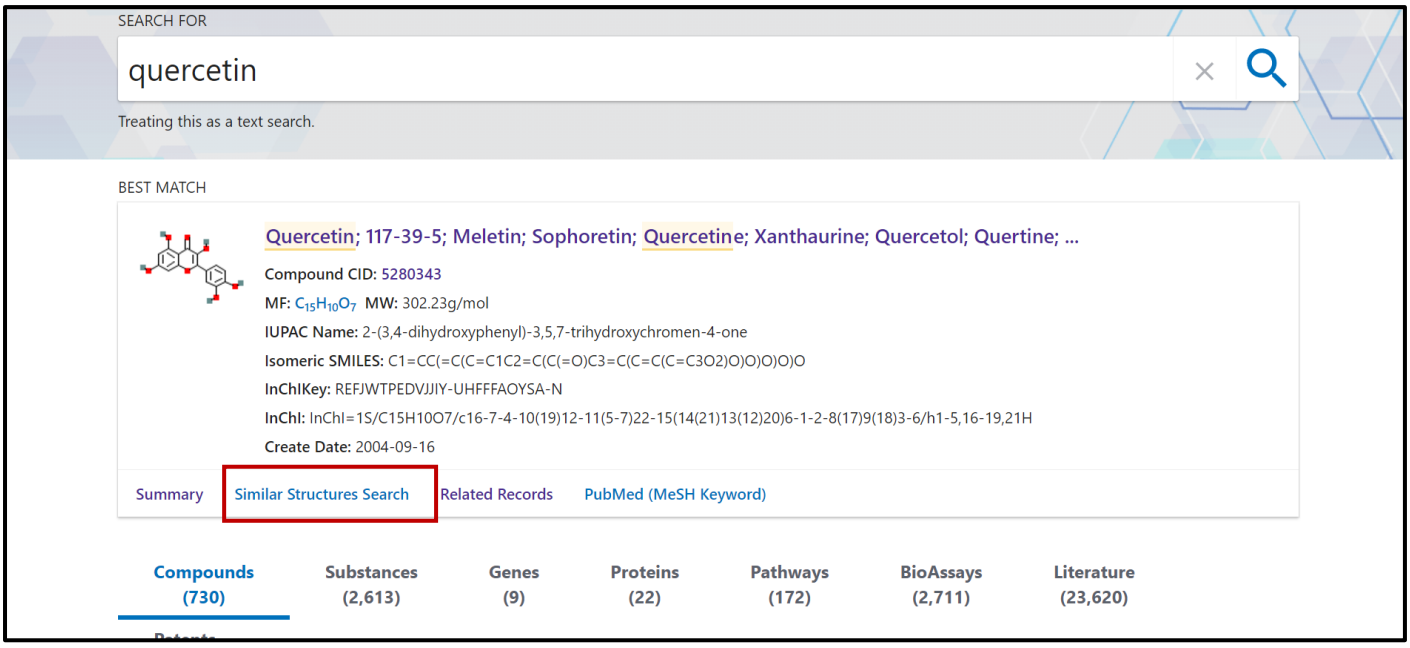


FIG 4. Substructure search for Quercetin

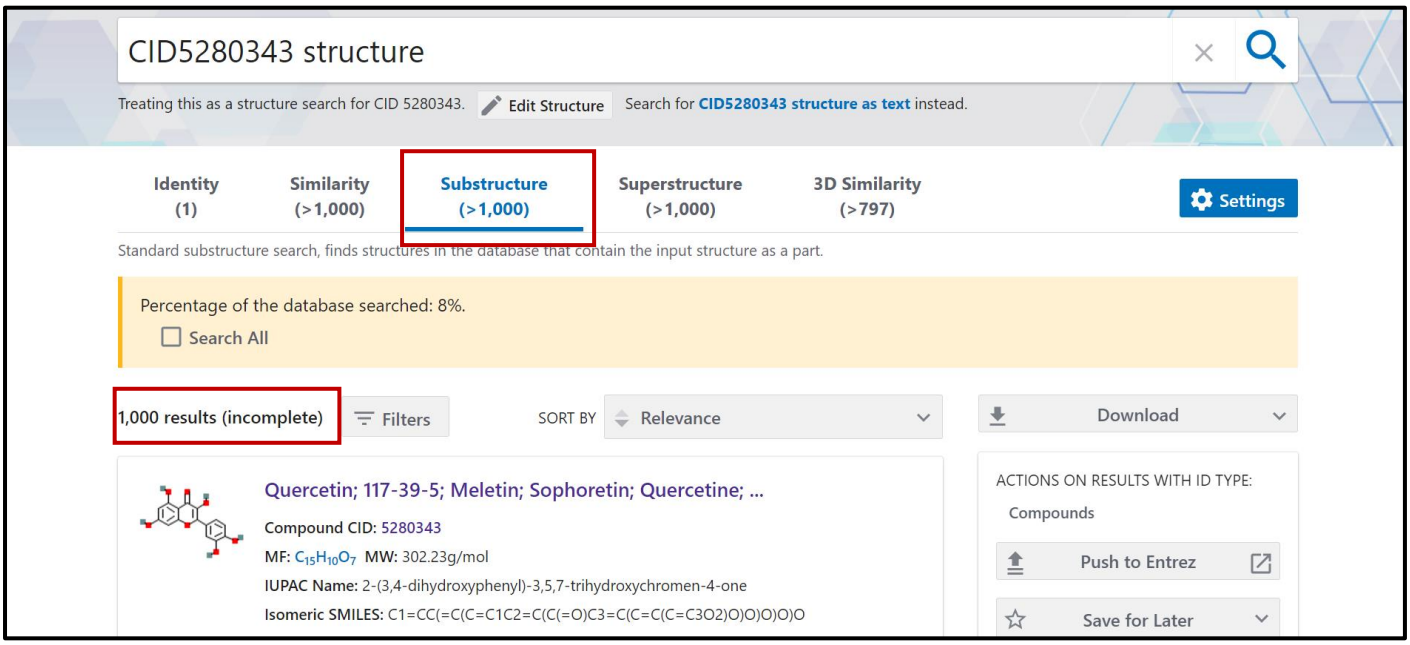


FIG 4.1. Hit page for substructure search for Quercetin

Structure search using exact search approach:

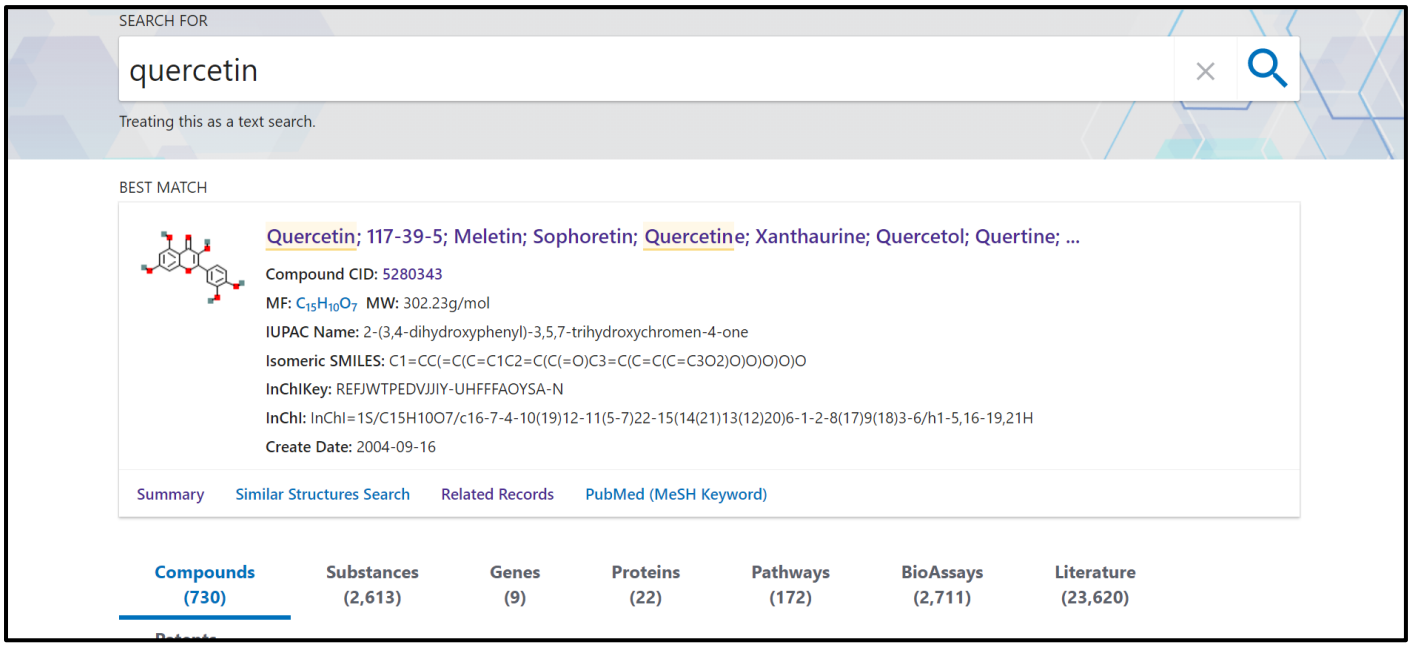


FIG 5. Exact search for Quercetin

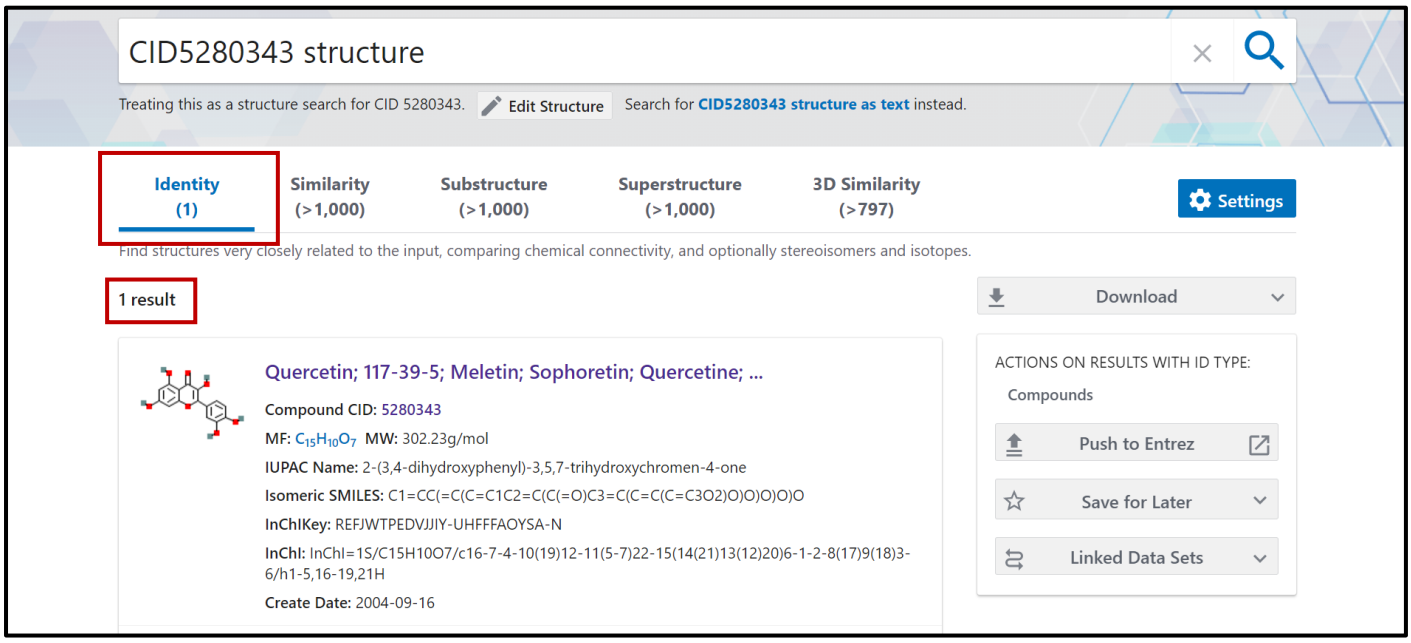


FIG 5.1. Hit page for exact search for Quercetin

Structure search using similarity search approach:

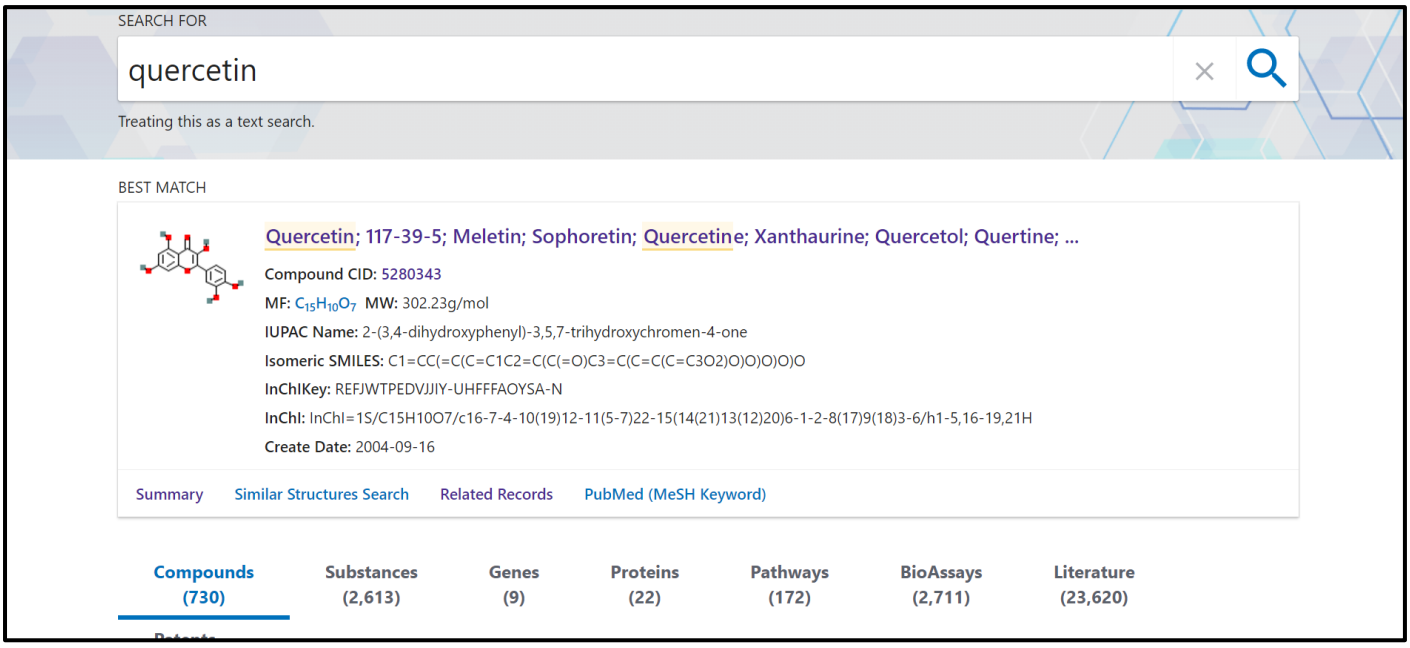


FIG 6. Similarity search for Quercetin

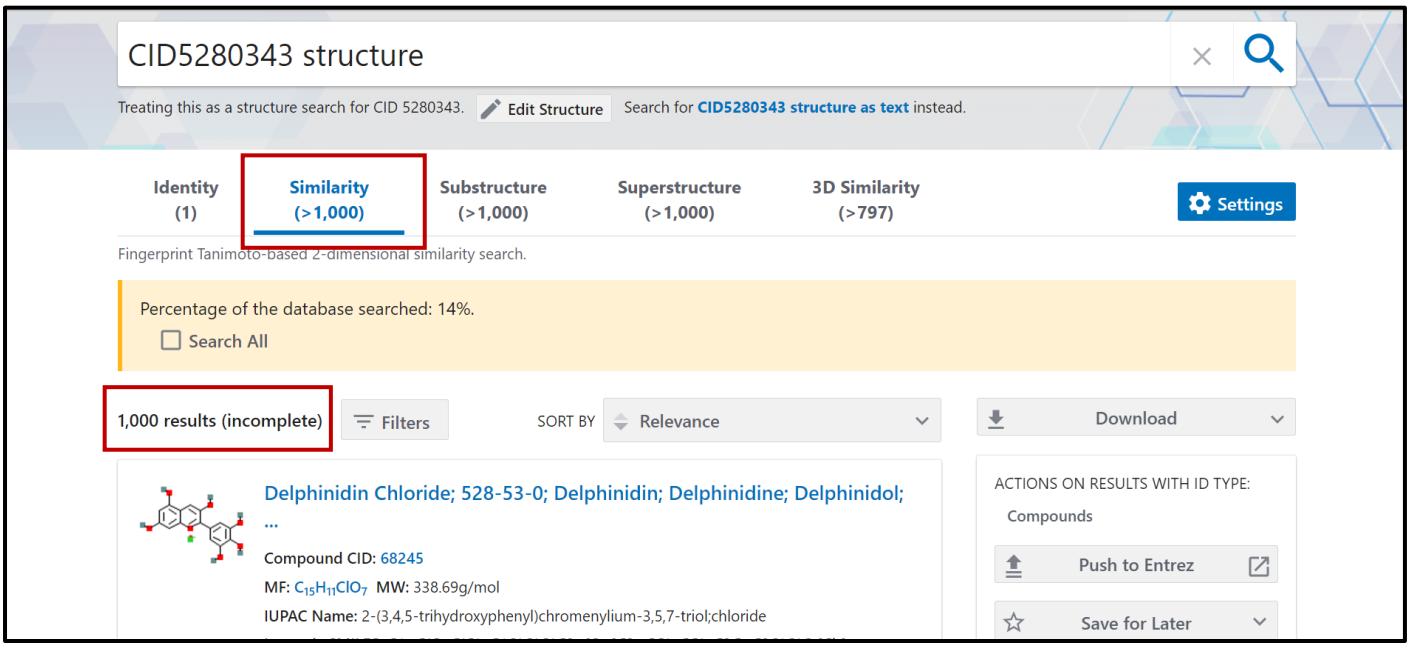


FIG 6.1. Hit page for similarity search for Quercetin

To screen structures based on chemical properties approach:

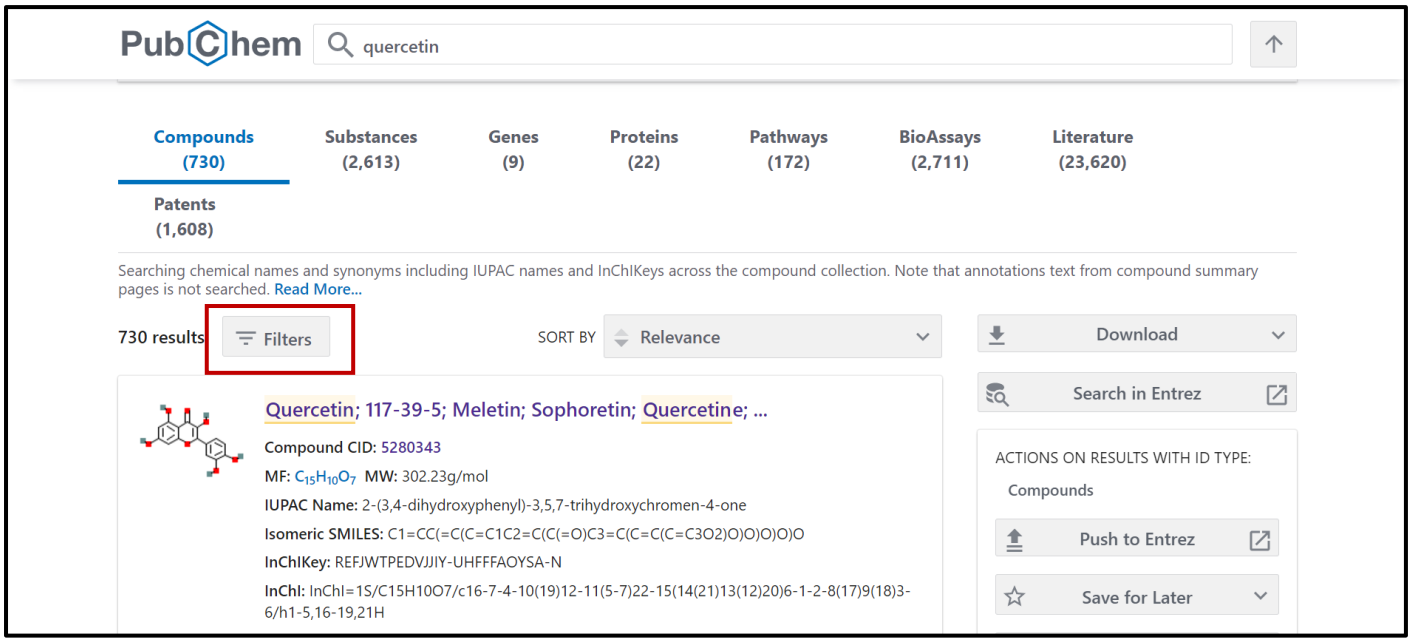


FIG 7. Applying filters to screen structures based on chemical properties

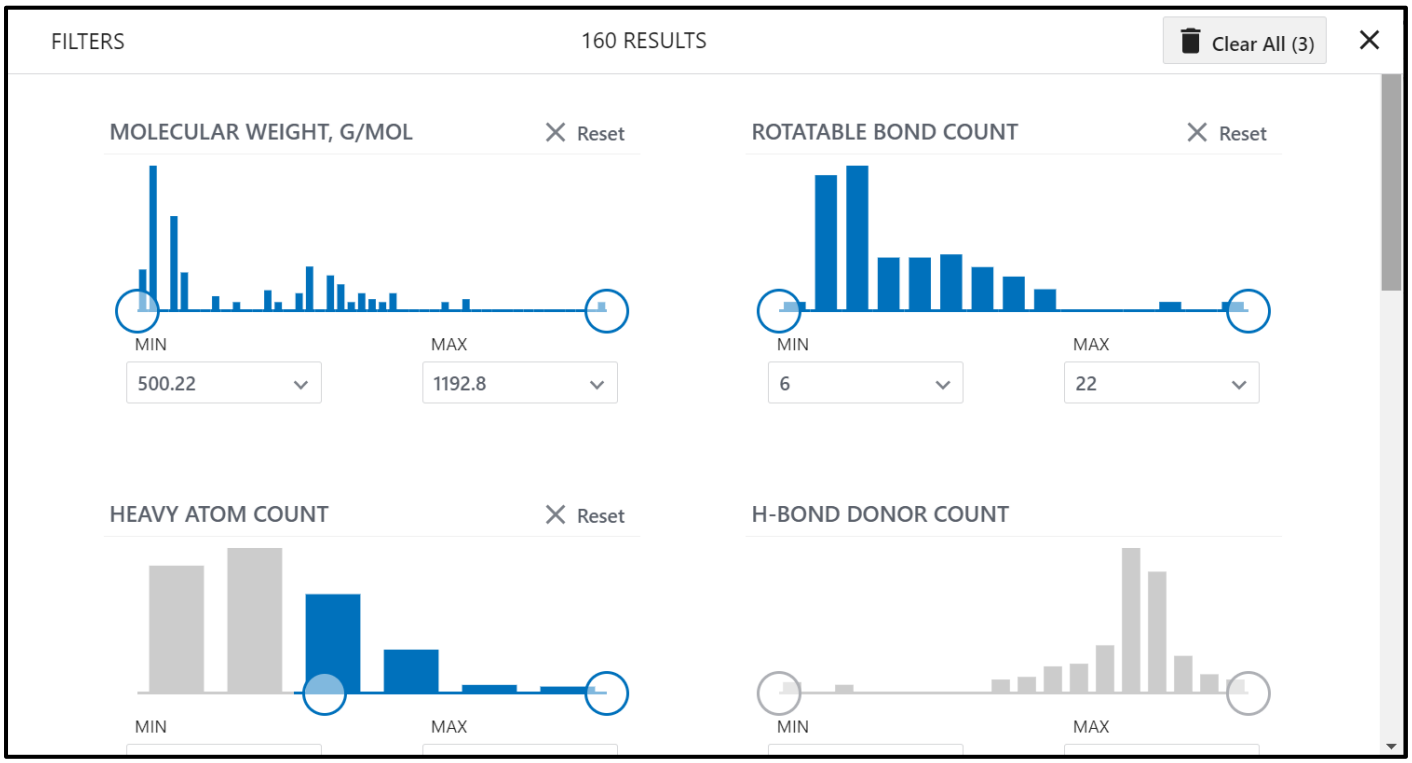


FIG 7.1. Filters applied for refinement based on chemical properties

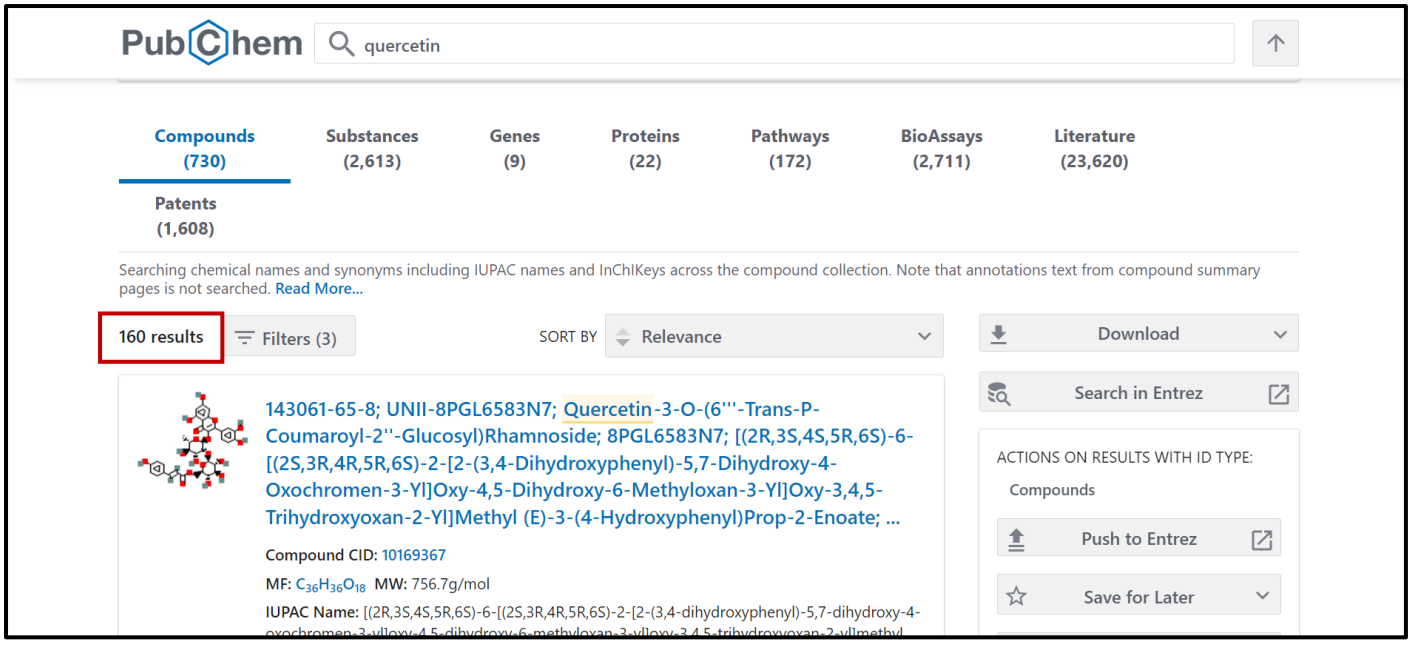


FIG 7.2. Hit page for results obtained after applying filters

RESULTS:

2D/3D structure for Quercetin (Pubchem id-5280343) were retrieved using Pubchem Database. The results showed one best match and 730 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 730 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 160 compounds.

CONCLUSION:

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, from https://www.techopedia.com/definition/26202/data-filtering
2. *Quercetin - an overview | ScienceDirect Topics*. (n.d.). Www.sciencedirect.com. Retrieved 5 September 2022, from https://www.sciencedirect.com/topics/neuroscience/quercetin
3. PubChem. (2022). *The PubChem Project*. Nih.gov; National Library of Medicine. Retrieved 5 September 2022, from https://pubchem.ncbi.nlm.nih.gov/
4. PubChem. (2022). *The PubChem Project*. Nih.gov; National Library of Medicine. Retrieved 5 September 2022, from https://pubchem.ncbi.nlm.nih.gov/#query=quercetin
5. PubChem. (2022). *The PubChem Project*. Nih.gov; National Library of Medicine. Retrieved 5 September 2022, from https://pubchem.ncbi.nlm.nih.gov/compound/5280343