Database of Secondary Metabolites (Bacteria Metabolites DB)

M.Tech. Major Project Report(MCP697z)

Submitted by- Shreeram Kumar Singh MT23091 M.Tech CSE	Under the Supervision of Dr. N. Arul Murugan Sir

Table of Contents

20. Future Goal

1. Project Description

Objective: Collect all information about secondary metabolites of bacteria, including species name, genes involved in metabolite synthesis, secondary metabolite, SMILES, generic name, and known applications.

Tasks : Secondary metabolites are organic compounds not directly involved in the essential life processes of organisms. Unlike primary metabolites, which are crucial for growth, development, and reproduction, secondary metabolites often play specialized roles in defense, communication, and attraction. These compounds are produced by a wide range of organisms, including plants, fungi, and bacteria, and they exhibit diverse chemical structures and biological activities. In bacteria, secondary metabolites are particularly important for survival in competitive and hostile environments. They help bacteria defend against predators, inhibit the growth of competing microorganisms, and establish symbiotic relationships with other organisms.

The production of secondary metabolites of bacteria is governed by specific genes and regulatory pathways. These metabolites serve as potent defense mechanisms, protecting bacterial cells from various threats such as predators, pathogens, and herbivores. Additionally, they play a crucial role in the ecological interactions of bacteria, enabling them to communicate and interact with other microorganisms and higher organisms. The study of bacterial secondary metabolites has significant implications for biotechnology and medicine, as many of these compounds have been found to possess antimicrobial, anticancer, and immunosuppressive properties. For comprehensive research on bacterial secondary metabolites, valuable resources include PubMed, European PMC, and Google Scholar, which provide access to a wealth of scientific literature and studies.

2. Secondary Metabolites Overview

Secondary metabolites are organic compounds produced by various organisms that are not essential for growth, development, or reproduction but serve specialized functions. Unlike primary metabolites, which are crucial for basic life processes, secondary metabolites often play critical roles in defense, communication, and adaptation to environmental conditions. They are characterized by their diverse chemical structures and biological activities.
In bacteria, secondary metabolites are vital for survival in competitive and often extreme environments, aiding in defense against predators, inhibiting rival microorganisms, and facilitating interactions with other species. These compounds are of significant interest in biotechnology and medicine due to their potential applications, including antimicrobial, anticancer, and immunosuppressive properties.

3. Research and Data Collection

The project involved extensive research using databases such as PubMed, Europe PMC, and Google Scholar. Keywords used for searches included 'secondary metabolites of bacteria'.

Analysis: Secondary Metabolites of Bacteria Studies

Overview :

- The dataset contains bibliographic details of research articles focused on secondary metabolites of bacteria. The data includes various columns such as PMID, Title, Authors, Citation, First Author, Journal/Book, Publication Year, Create Date, PMCID, NIHMS ID, and DOI.

Total Number of Records:

- The dataset comprises a total of 12,528 records.

Time Span of Publications:

- The publications range from the year 1964 to 2024, covering over six decades of research.

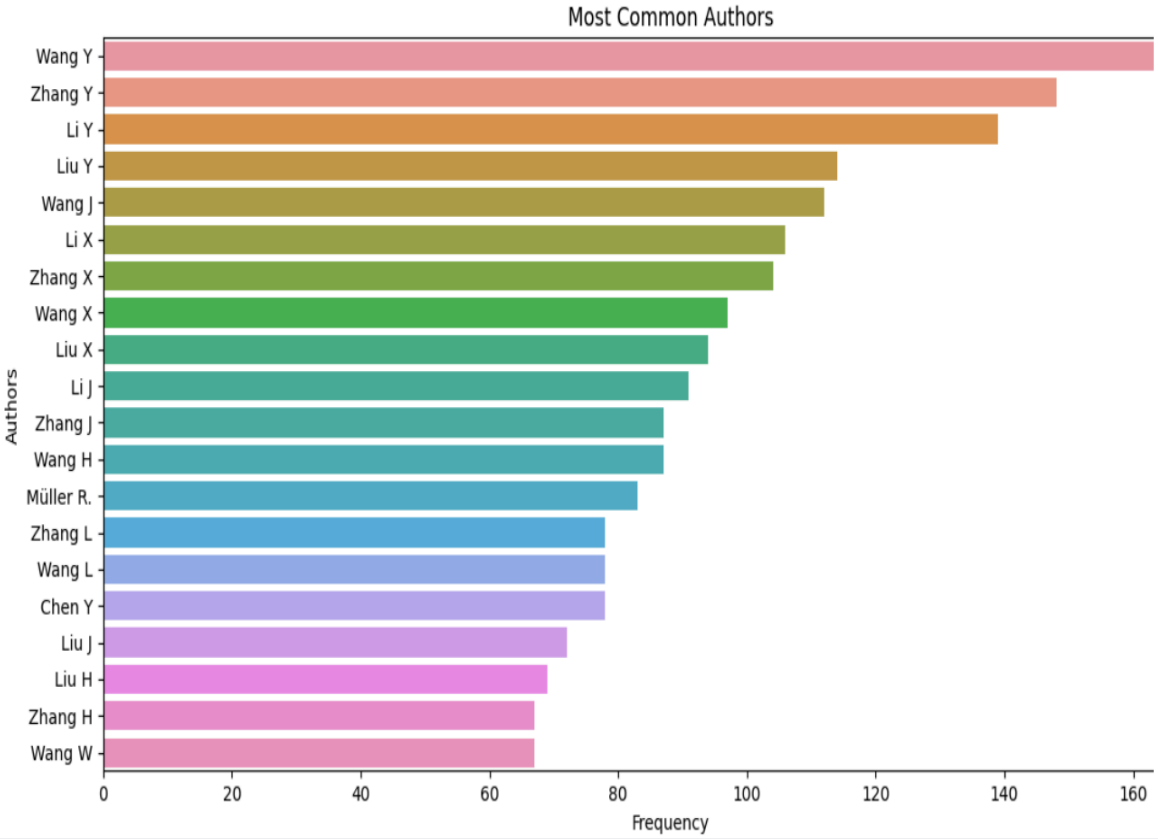
Scholarly Journals and Reference Books:

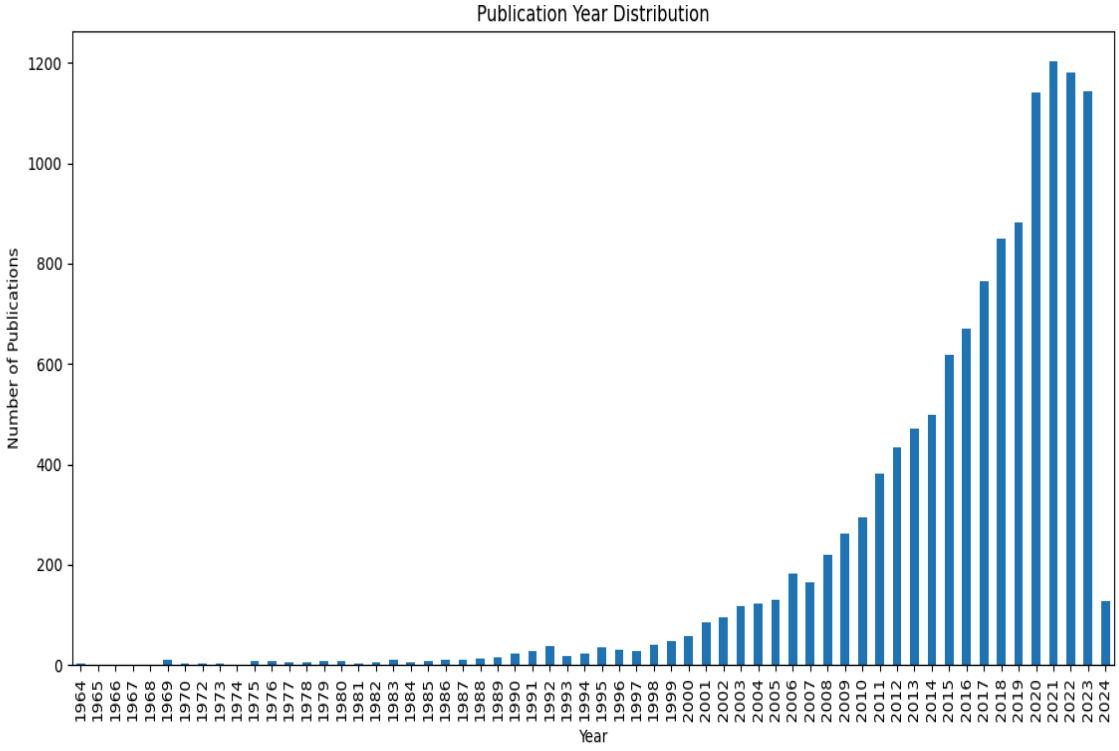
- The research articles are published in various reputable journals and books, such as Frontiers in Microbiology, Marine Drugs, Applied Microbiology and Biotechnology, Molecules, and PLoS One

Top 5 Authors :

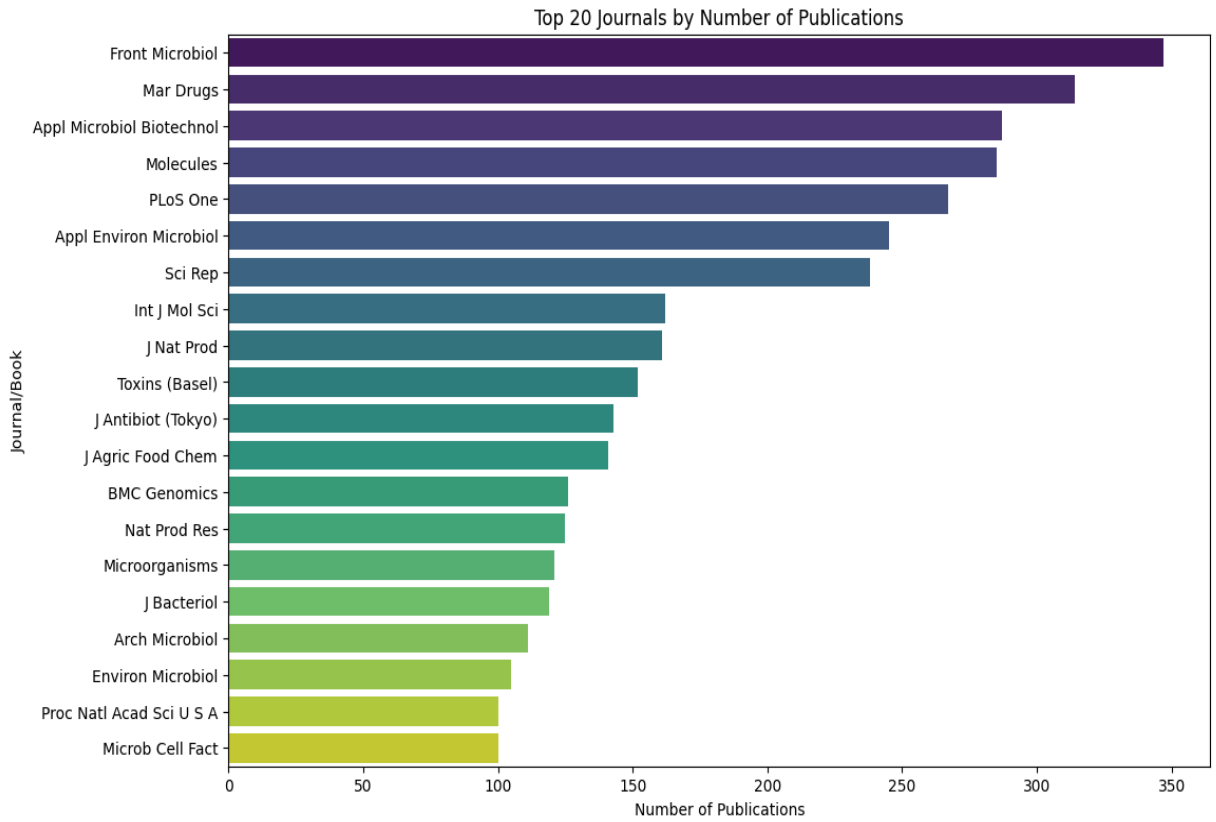
- Wang Y
- Zhang Y
- Li Y
- Liu Y
- Wang J

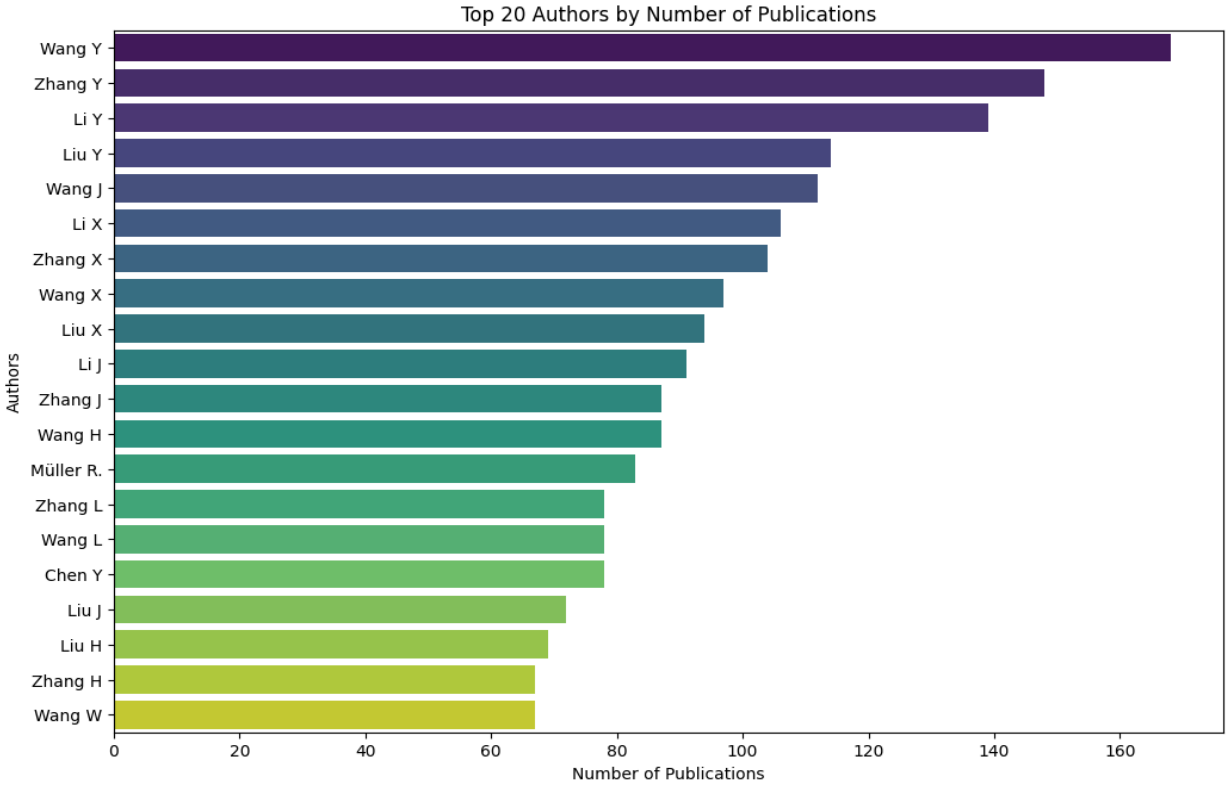
Visual analysis of the data is presented below


Most Common Authors
Wang Y
Zhang Y
Li Y
Liu Y
Wang J
Li X
Zhang X
Wang X
Liu X
Li J
Zhang J
Wang H
Müller R.
Zhang L
Wang L
Chen Y
Liu J
Liu H
Zhang H
Wang W
Authors
0
20
40
60
80
100
120
140
160
Frequency


Publication Year Distribution
Number of Publications
0
200
400
600
800
1000
1200
1964
1965
1966
1967
1968
1969
1970
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Year

Word Cloud of Titles


Top 20 Journals by Number of Publications
Front Microbiol
Mar Drugs
Appl Microbiol Biotechnol
Molecules
PLoS One
Appl Environ Microbiol
Sci Rep
Int J Mol Sci
J Nat Prod
Toxins (Basel)
J Antibiot (Tokyo)
J Agric Food Chem
BMC Genomics
Nat Prod Res
Microorganisms
J Bacteriol
Arch Microbiol
Environ Microbiol
Proc Natl Acad Sci U S A
Microb Cell Fact
Journal/Book
0
50
100
150
200
250
300
350
Number of Publications


Top 20 Authors by Number of Publications
Wang Y
Zhang Y
Li Y
Liu Y
Wang J
Li X
Zhang X
Wang X
Liu X
Li J
Zhang J
Wang H
Müller R.
Zhang L
Wang L
Chen Y
Liu J
Liu H
Zhang H
Wang W
Authors
0
20
40
60
80
100
120
140
160
Number of Publications

3.1. Summary of Research Papers Found

PubMed : 12,528 papers
Google Scholar : 21,40,000 papers
Europe PMC : 83723 papers

The abstracts of these papers were extracted and examined with the help of the PubMed mineR package. Subsequently, the data was structured and enhanced with details on genes, diseases, chemicals, mutations, and species using PubTator.

Task2_pubmed_secondary_metabolites_Bacteria.csv

	PMID	Genes	Diseases	Mutations	Chemicals	Species	Publicatio	Publication Link
2	20210692	No Genes	No Data	No Data	carbon>M	human>96	2010	https://pubmed.ncbi.nlm.nih.gov/20210692
3	28038926	No Data	No Data	No Data	No Data	No Data	2017	https://pubmed.ncbi.nlm.nih.gov/28038926
4	19345136	No Genes	No Data	No Data	peptides>	Bacillus th	2009	https://pubmed.ncbi.nlm.nih.gov/19345136
5	7826019	No Data	No Data	No Data	No Data	No Data	1994	https://pubmed.ncbi.nlm.nih.gov/7826019
6	25198138	No Genes	colorectal	No Data	fatty acids	human>96	2014	https://pubmed.ncbi.nlm.nih.gov/25198138
7	16116269	No Genes	No Data	No Data	phenolic a	Sphingom	2005	https://pubmed.ncbi.nlm.nih.gov/16116269
8	21672958	No Genes	tumor>MI	No Data	cholesterc	No Data	2011	https://pubmed.ncbi.nlm.nih.gov/21672958
9	29080339	No Genes	No Data	No Data	Bile Acid>I	human>96	2017	https://pubmed.ncbi.nlm.nih.gov/29080339
10	26393965	No Data	No Data	No Data	No Data	No Data	2015	https://pubmed.ncbi.nlm.nih.gov/26393965
11	27986719	No Genes	No Data	No Data	No Data	Actinomyc	2017	https://pubmed.ncbi.nlm.nih.gov/27986719
12	27739371	No Data	No Data	No Data	No Data	No Data	2017	https://pubmed.ncbi.nlm.nih.gov/27739371
13	27483244	No Genes	No Data	No Data	terpenes>	grapefruit	2016	https://pubmed.ncbi.nlm.nih.gov/27483244
14	21375710	No Genes	No Data	No Data	lunalides>	No Data	2011	https://pubmed.ncbi.nlm.nih.gov/21375710
15	24863894	No Data	No Data	No Data	No Data	No Data	2014	https://pubmed.ncbi.nlm.nih.gov/24863894
16	20648021	No Genes	No Data	No Data	iron>MESI	Unculture	2010	https://pubmed.ncbi.nlm.nih.gov/20648021
17	25131404	No Genes	infection>	No Data	No Data	Streptomy	2014	https://pubmed.ncbi.nlm.nih.gov/25131404
18	25461728	No Data	No Data	No Data	No Data	No Data	2015	https://pubmed.ncbi.nlm.nih.gov/25461728
19	29113654	CTLA4>14	gastrointe	No Data	bile acids>	human>96	2017	https://pubmed.ncbi.nlm.nih.gov/29113654
20	23666088	No Genes	No Data	No Data	isoquinoli	No Data	2013	https://pubmed.ncbi.nlm.nih.gov/23666088

Analysis:

1. Publication Details

- Time Span: The dataset includes publications from 1994 to 2017. This time span reflects a growing interest in bacterial secondary metabolites and their applications over recent decades

- Journals**:**

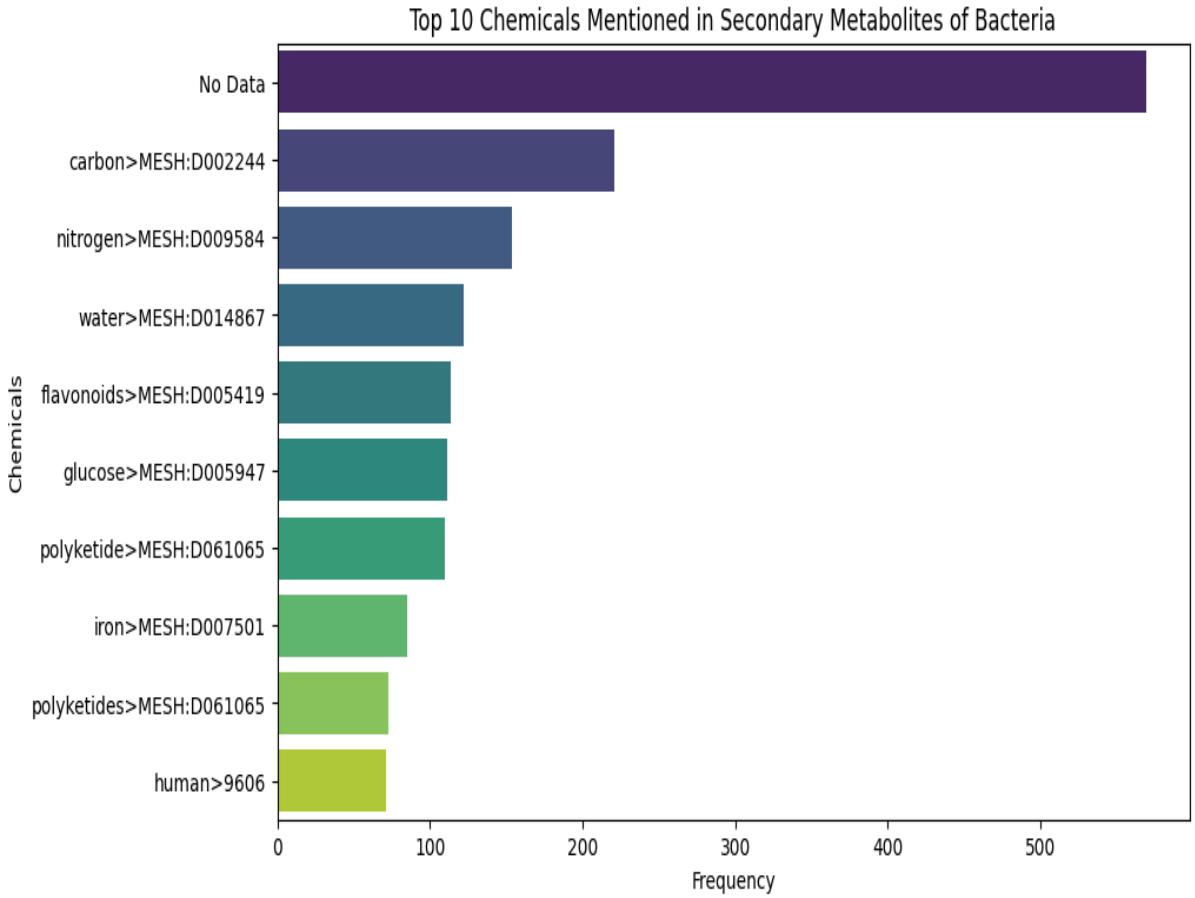
  Key journals where these studies were published include:

  - Science: A leading journal publishing high-impact research across various scientific disciplines
  - Proceedings of the Royal Society B: Biological Sciences: Features research on biological sciences, including bacterial secondary metabolites.
  - Journal of Natural Products: Focuses on natural products, including metabolites from microorganisms.
  - Microbial Cell Factories: Specializes in microbial biotechnology and applications, including secondary metabolite production
  - Antimicrobial Agents and Chemotherapy: Publishes studies on antimicrobial agents, including those derived from bacterial metabolites.

2. Biological and Chemical Entities

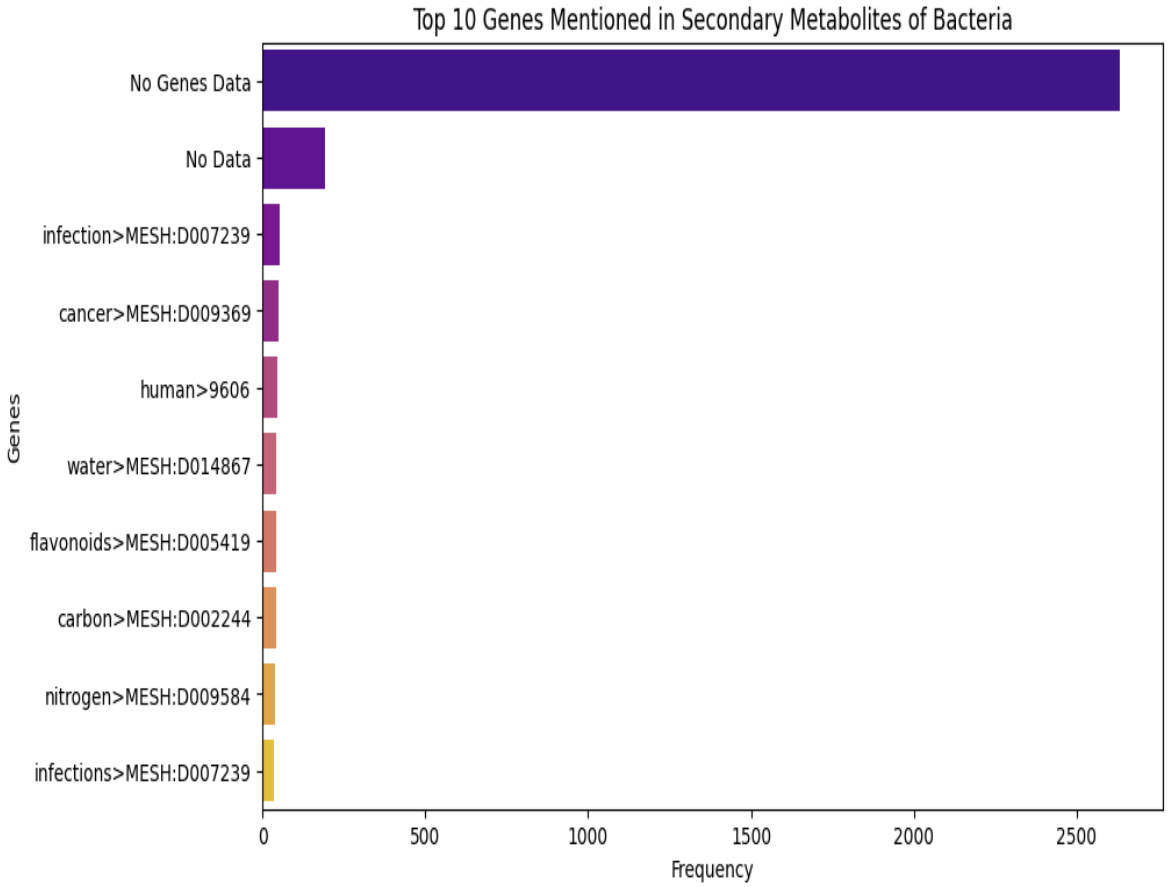
Chemicals

- Peptides: These are short chains of amino acids with significant biological activities. Their frequent mention suggests they are crucial in the research of bacterial metabolites, often for their antimicrobial and therapeutic properties.
- Polyketides: Complex compounds produced by bacteria with a range of biological activities, including antibiotic and anticancer properties. Their prominence in the dataset reflects their importance in drug discovery.
- Fatty Acids: These are components involved in various metabolic processes. Their presence indicates interest in understanding their role in bacterial metabolism and potential applications.



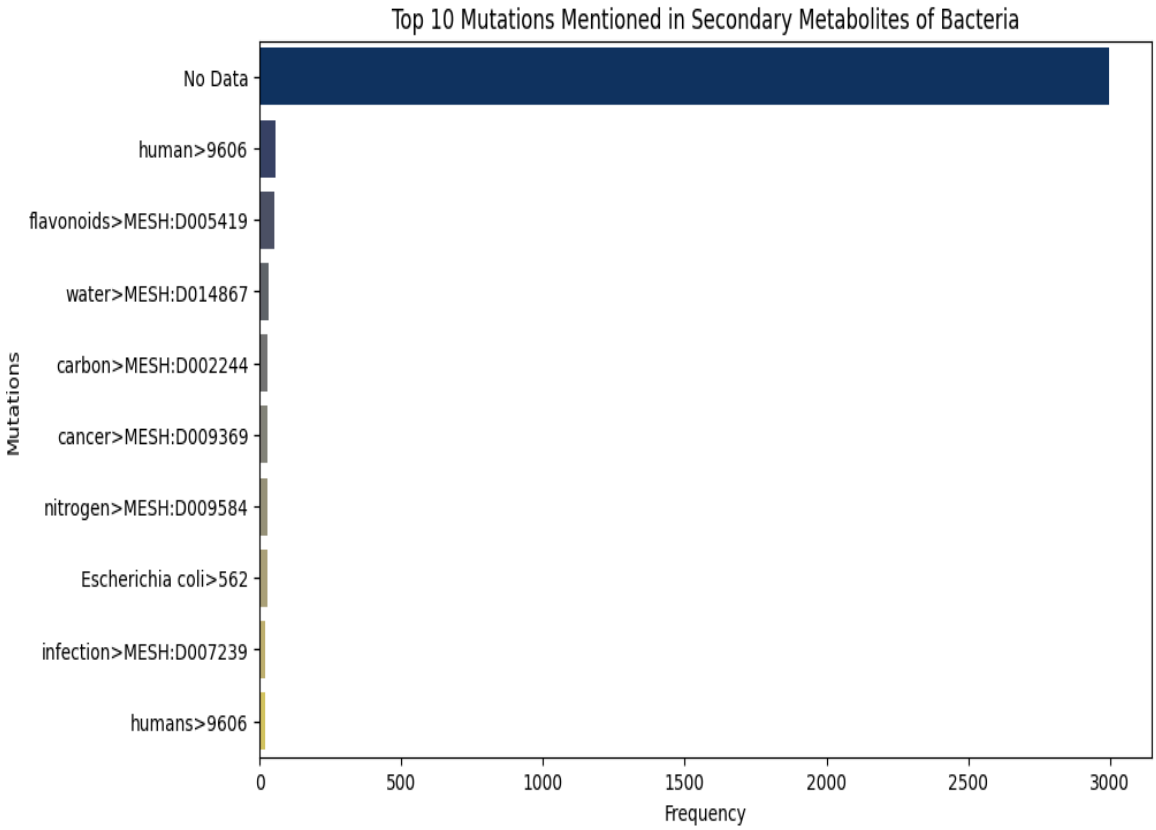
Genes

- PKS (Polyketide Synthase) Genes: These genes are crucial for the synthesis of polyketides. Their frequent mention highlights their role in producing complex metabolites with diverse bioactivities.
- NRPS (Non-Ribosomal Peptide Synthetase) Genes: Involved in the production of non-ribosomal peptides, these genes are significant for creating bioactive peptides with potential therapeutic uses.
- Regulatory Genes: These genes control the expression of biosynthetic pathways. Understanding their function helps in optimizing metabolite production.



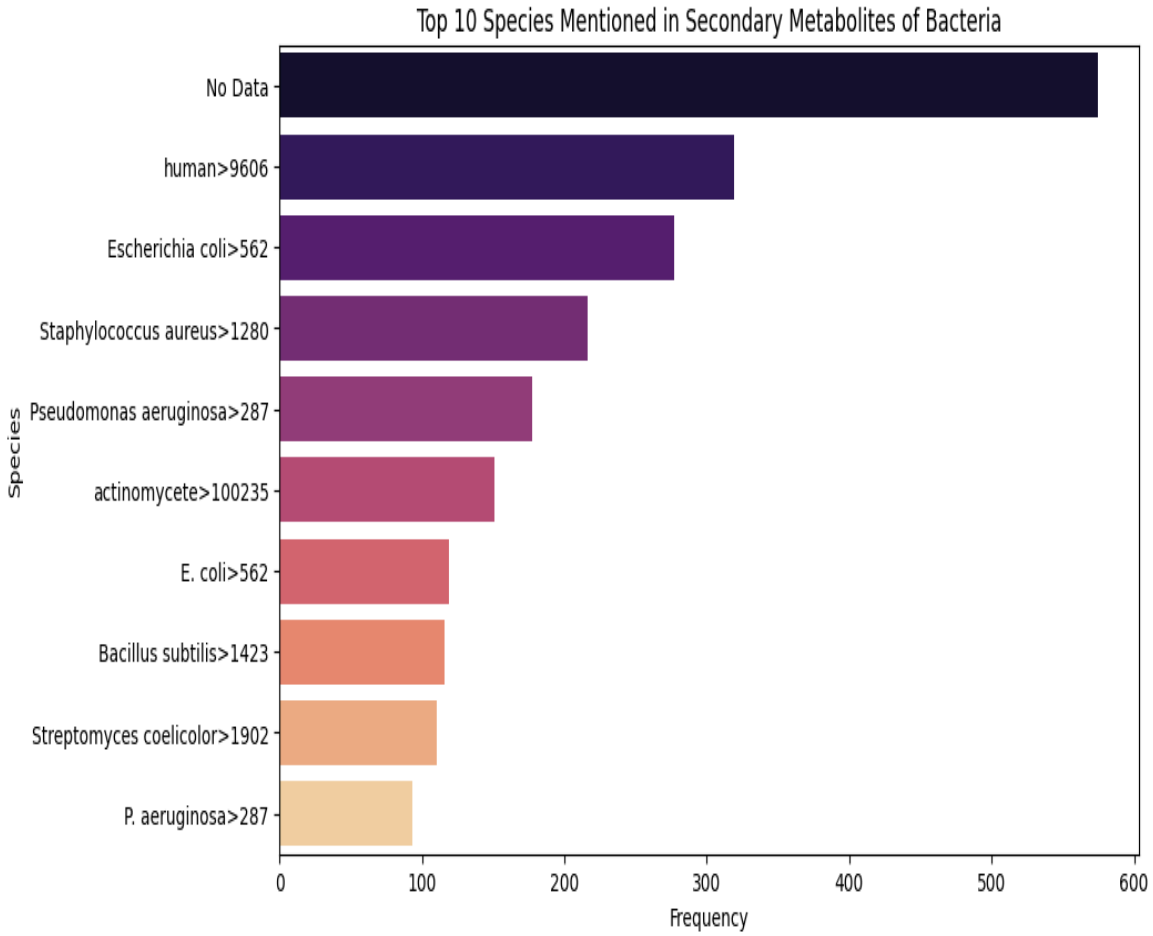
Mutations

- Gene Deletions: Removing specific genes can significantly impact metabolite production. Identifying which genes are deleted helps understand their role in metabolite synthesis.
- Point Mutations: Single nucleotide changes that can affect enzyme function and, consequently, metabolite production. These mutations can provide insights into the fine-tuning of metabolic pathways.
- Insertions: Additional genetic material that can influence the biosynthetic pathways of metabolites. These can lead to new or altered metabolite profiles.



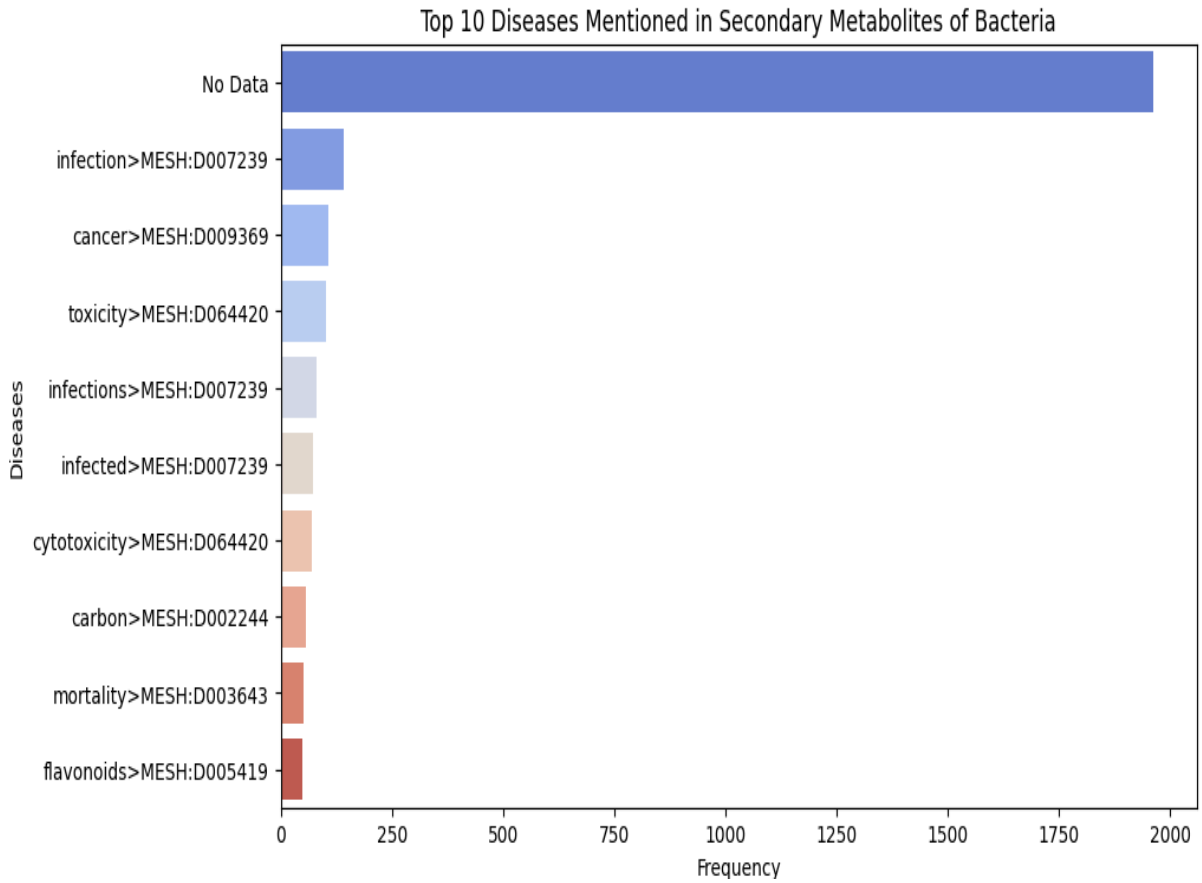
Species

- Bacillus thuringiensis: Known for producing insecticidal proteins. Its frequent mention reflects its role in biocontrol and secondary metabolite research.
- Pseudomonas entomophila: Recognized for its antimicrobial properties, making it a valuable species for studying metabolite production.
- Streptomyces griseus: Famous for producing the antibiotic streptomycin. Its prominence highlights its importance in antibiotic research.



Diseases

- Cancer: Metabolites are studied for their potential anticancer properties. Their ability to inhibit cancer cell growth or induce apoptosis is a significant area of research.
- Infections: Many metabolites have antimicrobial properties, making them candidates for treating bacterial infections.
- Inflammatory Diseases: Some metabolites may have anti-inflammatory effects, which are valuable in treating chronic inflammatory conditions.



3.2. Data Processing and Analysis

The project involved preprocessing and transforming CSV files derived from PubMed articles, specifically focusing on gene, disease, chemical, species, and mutation information. This encompassed:

1. Parsing and Cleaning Data: Extracting relevant details like names and IDs, removing invalid entries, and adjusting column structures.
2. Data Integration**:** Merging related data using unique IDs and saving the results into separate CSV files.
3. Data Consolidation: Combining linked datasets via IDs and exporting the processed data into individual CSV files.
4. Data Fusion: Integrating interconnected data based on IDs and storing the merged information in separate CSV files.
5. Data Aggregation: Joining associated datasets using identifiers and organizing the output into distinct CSV files.

3.3. Submitted Files

Task3_Secondary_Metabolites_Bacteria._Chemicals_pubmed.csv

Chemical	ChemicalID	PMID_List	length
(-)-8-O-methyltetrangomycin	C059960	[23274989]	1
(13)C	C000615229	[26905826]	1
(E,E,E)-geranylgeranyl diphosphate	C002963	[26475187]	1
(S)-ginsenoside Rg3	C097367	[17764709]	1
(S)-norcoclaurine	C012348	[29156609]	1
(S)-reticuline	C003298	[22179145, 216	2
(p)ppGpp	D006158	[27281927]	1
-2,4-dimethylbenzoic acid	C028523	[15307685]	1
1, 3-bis (3-phenoxyphenoxy)benzene	C426840	[11302198]	1
1,1,1-Trifluoro-2-chloroethane	C016800	[367217]	1
1,1,3,3-tetrabromo-2-heptanone	C561883	[21660776]	1
1,1-Difluoro-2-bromo-2-chloroethylene	C018515	[367217]	1
1,3,6,8-tetrahydroxynaphthalene	C121288	[16395556, 104	2
1,8-cineol	D000077591	[26916832]	1
1,8-cineole	D000077591	[15141066]	1
1-acetyl-beta-carboline	C583241	[19842066]	1
1-aminoethylphosphonic acid	C047639	[525987]	1
1-butanol	D020001	[25512025]	1
1-carbapen-2-em-3 carboxylic acid	C093704	[20056700]	1
1-carbapen-2-em-3-carboxylic acid	C093704	[12519208]	1

Analysis:

1. Data Overview

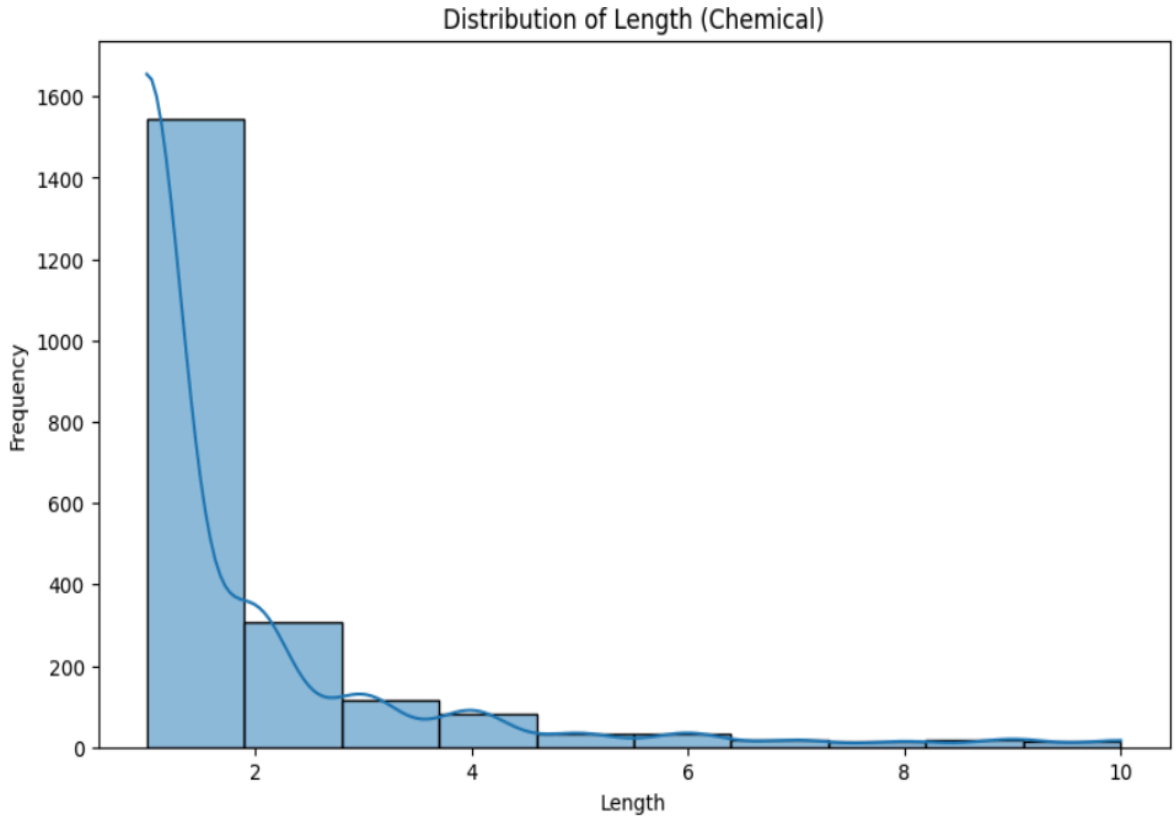
- Total Entries: 2243
- Unique Chemical IDs: 1454
- Unique Chemical Names: 2231
- Unique PubMed IDs: 1424

2. Descriptive Statistics for length Column

- Count: 2243.0
- Mean: 2.3352652697280427
- Standard Deviation: 4.93652304971051
- Min: 1.0
- 25th Percentile: 1.0
- Median (50th Percentile): 1.0
- 75th Percentile: 2.0
- Max: 103.0

Conclusions

- The data primarily comprises chemicals with short lengths
- PubMed IDs typically reference a few chemicals, but some reference a substantial number.
- A small group of chemicals is linked to numerous PubMed IDs, suggesting their importance or relevance in various studies.



Task3_Secondary_Metabolites_Bacteria._Genes_pubmed.csv

Gene	GeneID	PMID_List	length
3-hydroxy-3-methylglu	3156	[6400479]	1
A1, A2, B	850458	[24751367]	1
A1, A2, B, B1 and B4	850458	[24751367]	1
A2a	28882	[28358337]	1
ABC	10058	[27903896]	1
ACE	1636	[2560484]	1
AfsK	1099863	[16579461]	1
AfsR	1099866	[16579461]	1
Age	5973	[26739136]	1
Apr	5366	[23708134]	1
Atpdr2	827187	[19854857]	1
Azoreductase	1728	[24779771]	1
B17	4712	[21429787]	1
C-G	1511	[10766021]	1
CAST	831	[16715543]	1
CD274	29126	[29113654]	1
COX2	140540	[27192145]	1
CTLA4	1493	[29113654]	1
CYP	9360	[29741855, 25662514]	2
CYP3	10105	[20736090]	1

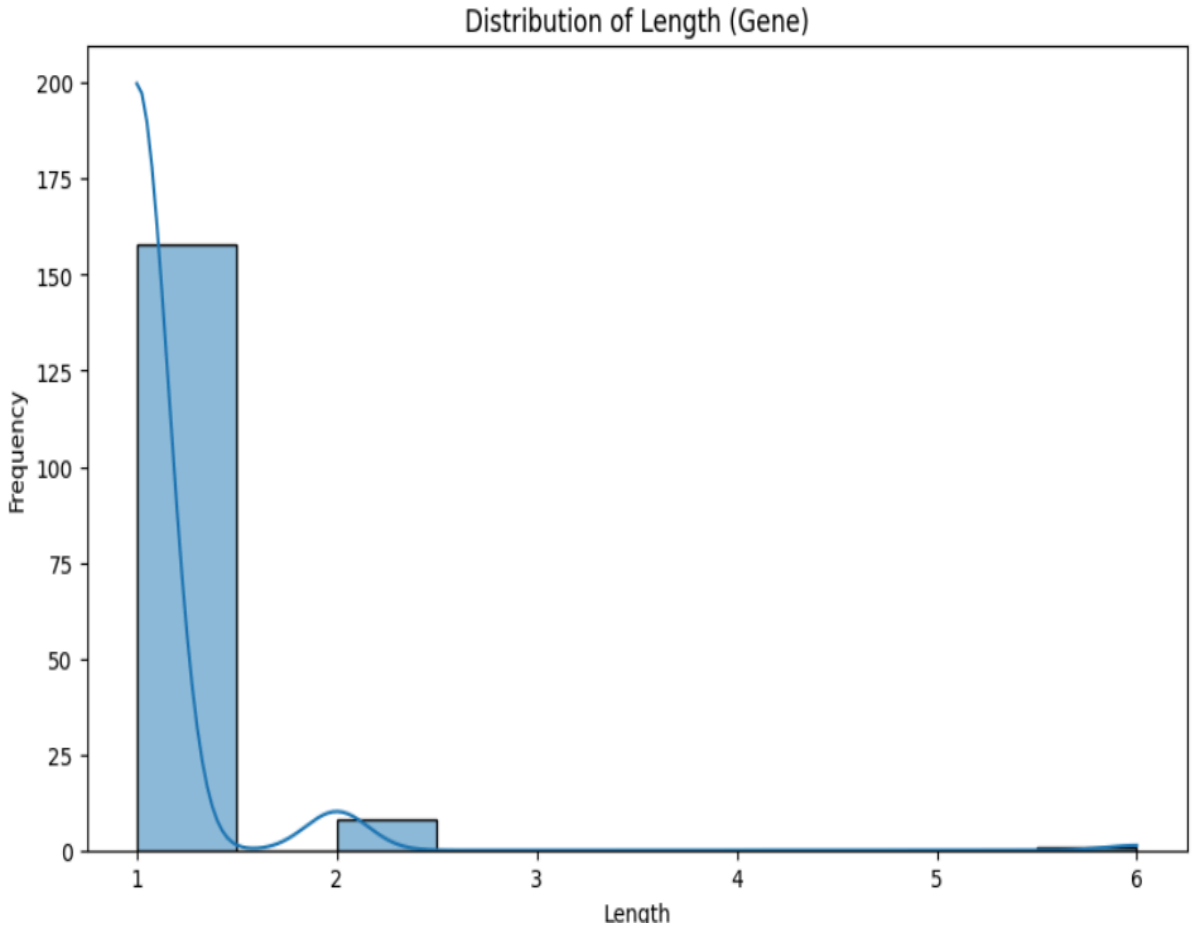
Analysis:

1.Data Overview

- Total Entries: 167
- Unique Gene IDs: 125
- Unique Gene Names: 162
- Unique PubMed IDs: 103

2. Descriptive Statistics for length Column

- Count: 167.0
- Mean: 1.0778443113772456
- Standard Deviation: 0.438974873393064
- Min: 1.0
- 25th Percentile: 1.0
- Median (50th Percentile): 1.0
- 75th Percentile: 1.0
- Max: 6.0

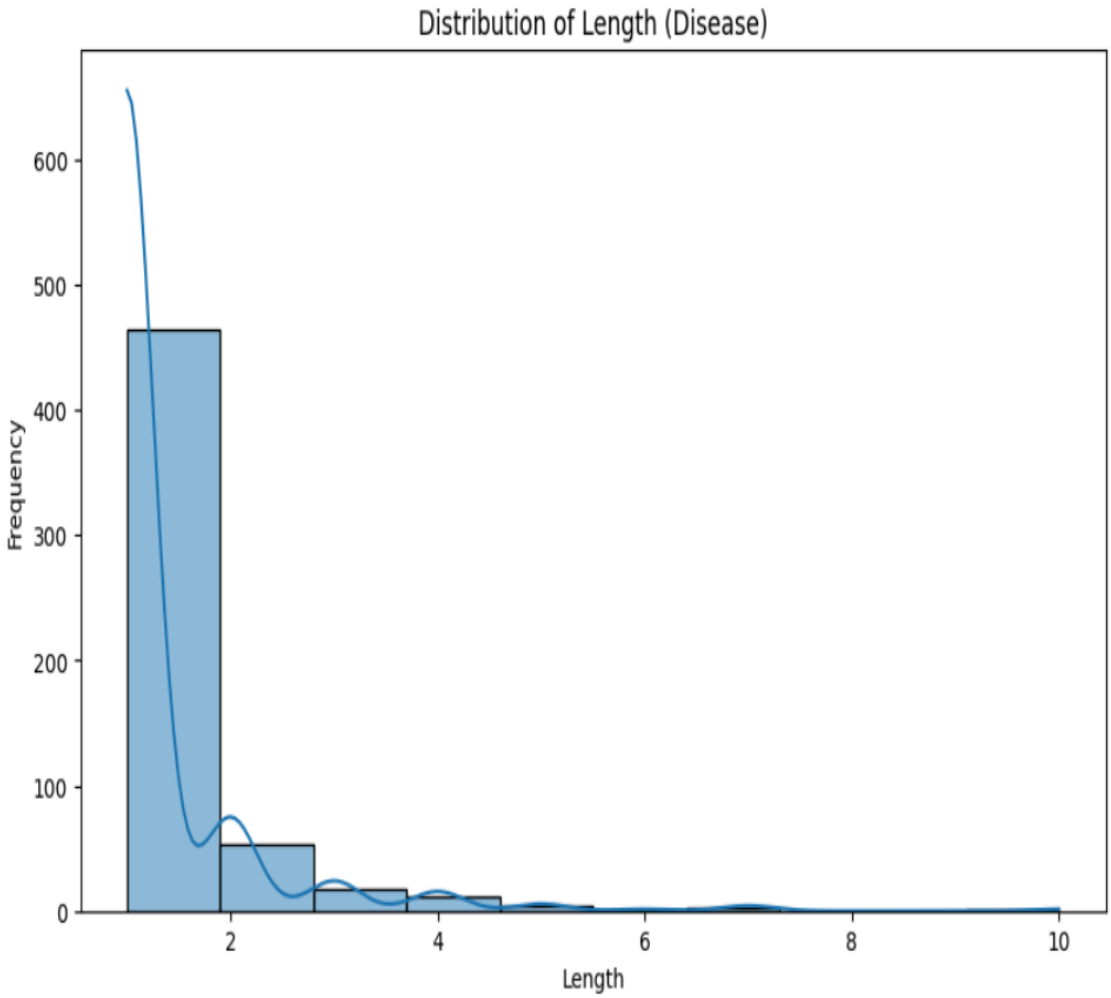


Disease	DiseaseID	PMID_List	length
(1)H-(1)H COSY, (1)H-(13)C HMBC	D000848	[23979826]	1
AF	D001281	[26466717]	1
AIDS	D000163	[23644174]	1
ARBS	D058745	[22476960]	1
AVMs	C564254	[29049952]	1
Acidic pH shock	D012769	[25605030]	1
Allosalinactinospora lopnorensis CA15-2(T	C000657245	[26864220]	1
Alzheimer diseases	D000544	[28865203]	1
Alzheimer's disease	D000544	[23775636]	1
Antarctic bacteria	C537702	[23619351]	1
Antarctic sub-sea	D007246	[27128927]	1
Aplysilla rosea	D017515	[27717966]	1
Aplysina Red Band Syndrome	D058745	[22476960]	1
BRAF-wild-type disease	D003141	[28923537]	1
Bifurcaria bifurcata epiphytic bacteria	C537702	[24663118]	1
British Columbia	176500	[24130838, 27812802]	2
C-A-T-C-A-T-E-C-A-T-C	C537418	[22825833]	1
CF	D003550	[29069388]	1
CHA0 toxicity	D064420	[17088380]	1
CHA0(T	D001260	[21392918]	1

Analysis:

1. Data Overview
    - Total Entries: 567
    - Unique Disease IDs: 269
    - Unique Disease Names: 566
    - Unique PubMed IDs: 420

2. Descriptive Statistics for length Column
    - Count: 567.0
    - Mean: 1.9559082892416226
    - Standard Deviation: 5.035895481341404
    - Min: 1.0
    - 25th Percentile: 1.0
    - Median (50th Percentile): 1.0
    - 75th Percentile: 1.0
    - Max: 61.0


Distribution of Length (Disease)
Frequency
Length
0
100
200
300
400
500
600
2
4
6
8
10

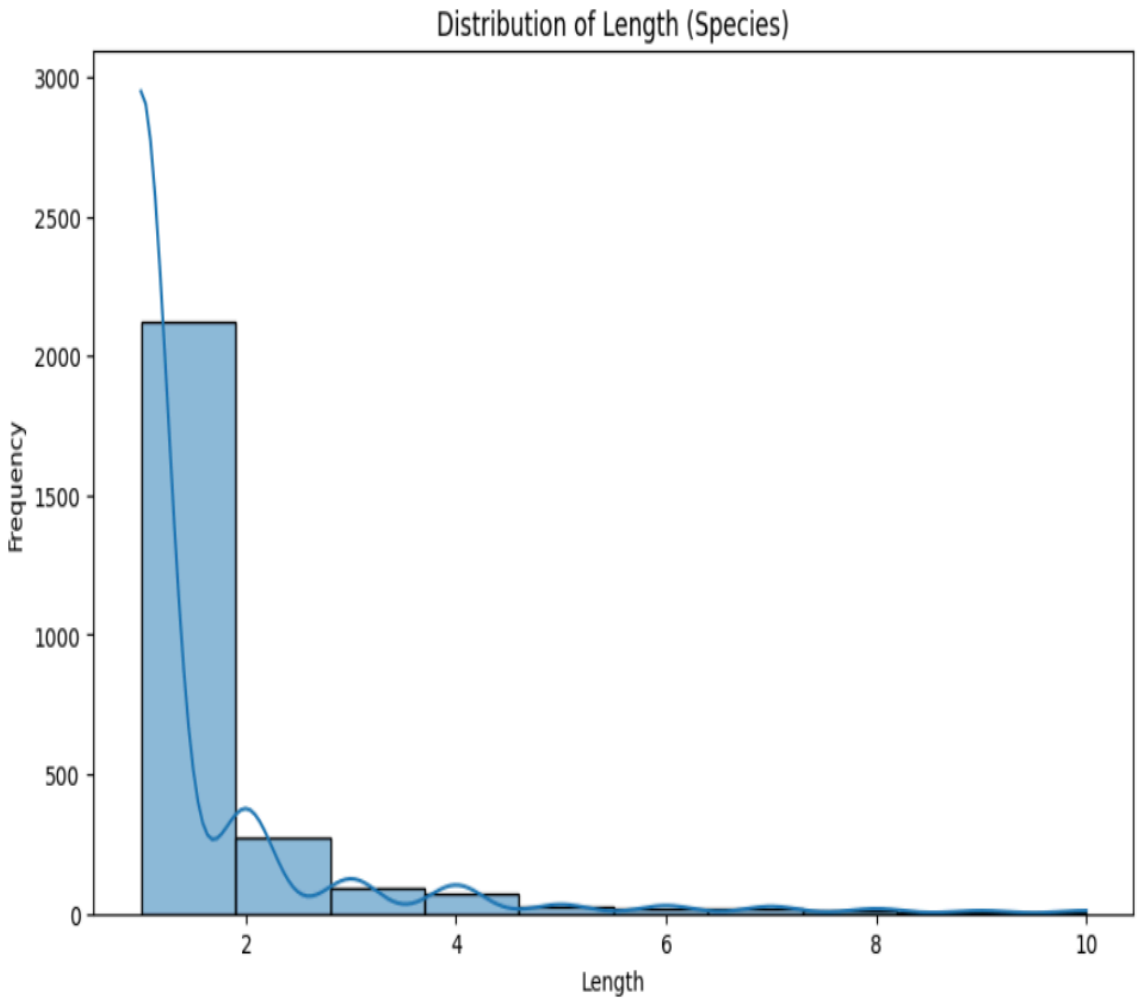
Task3_Secondary_Metabolites_Bacteria._Species_pubmed.csv

Species	SpeciesID	PMID_List	length
A. Meyer	52773	[9436194]	1
A. aculeatus	5053	[29248948]	1
A. aerophoba	289389	[19588186, 18783385]	2
A. alternata	5599	[24301768]	1
A. annua	35608	[24575401]	1
A. annua L	212759	[24575401]	1
A. balhimycina	208443	[23184174]	1
A. baumannii	470	[28654009, 28705748, 28421168]	3
A. baumannii ATCC 17978	400667	[28421168]	1
A. brasilense	192	[21843946]	1
A. brasiliensis	37326	[22768320]	1
A. brassicicola	29001	[20580869]	1
A. calcoaceticus	471	[22806873]	1
A. cannabina	446322	[24268066]	1
A. cauliformis	289398	[22476960]	1
A. cavaraeanum	940301	[24391056]	1
A. cavernicola	121477	[26757731]	1
A. chinensis	217632	[24069601]	1
A. chrysogenum	5044	[1368054]	1
A. clavatonanicus	41054	[29049321]	1

Analysis:

1. Data Overview
    - Total Entries: 2707
    - Unique Species IDs: 1589
    - Unique Species Names: 2544
    - Unique PubMed IDs: 1438

2. Descriptive Statistics for length Column
    - Count: 2707.0
    - Mean: 2.12079793128925
    - Standard Deviation: 6.96447269538657
    - Min: 1.0
    - 25th Percentile: 1.0
    - Median (50th Percentile): 1.0
    - 75th Percentile: 1.0
    - Max: 186.0



4. Sentence Extraction

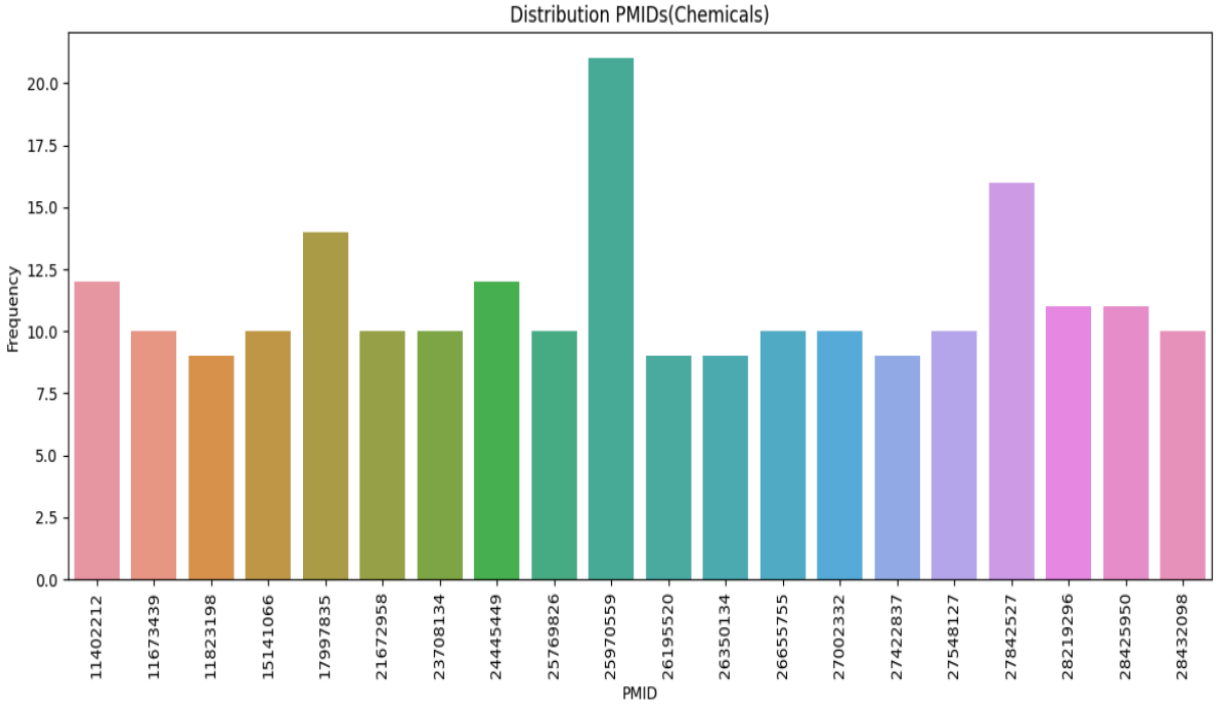
I. Using the pubmed.mineR package, sentences were extracted from PubMed Central (PMC) by referencing PMCID and gene names.

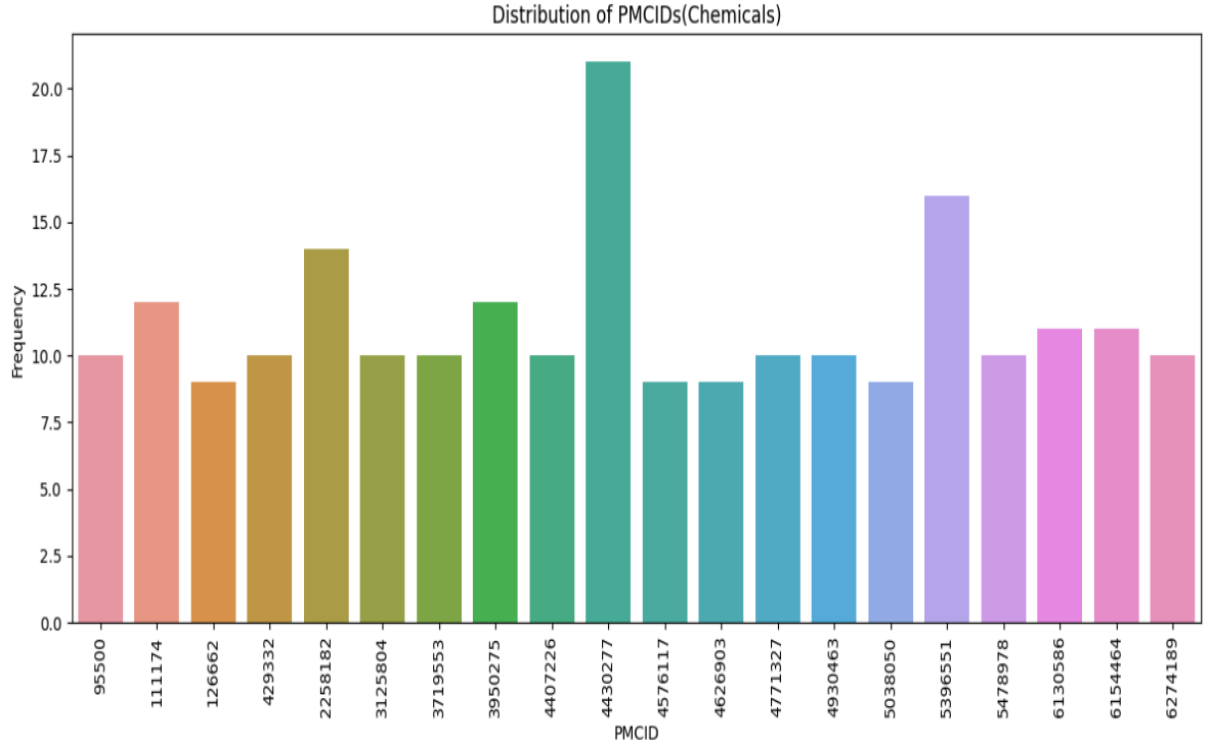
II. This data was structured into multiple CSV files.

Task4_Secondary_Metabolites_Bacteria_Chemicals_Pubmed.csv

Chemicals	ChemicalsID	PMID	PMCID	Sentences
(S)-norcocla	C012348	29156609	5713430	
(S)-reticulin	C003298	21610729	3112539	
1,8-cineole	D000077591	15141066	429332	Our studies revealed that the major VOCs released a
1-acetyl-be	C583241	19842066	3128704	
1-aminoeth	C047639	525987	352740	The novel antibacterial peptide mimetic alaphosphin
1-hydroxyp	C050093	21671613	3212935	This work provides the first in vivo measurements o
1-methyl-4-	C000622911	28777571	5687060	
12-methylte	C069642	24595070	3940840	A scale-up culture of B. latersporus PE36 yielded thr
13C	C000615229	27515463	5052388	The structure of the first active fraction was elucidat
13C	C000615229	23192186	6268673	Their structures were elucidated by UV, IR, MS, 1H-N
14C	C000615234	16000813	1169061	The cross-feeding of microbial products derived fron
2, 2-diphen	C004931	27842527	5396551	The antioxidant potential of the extracts was ascerta
2, 4-diacety	C059817	10869066	94573	
2, 4-dinitrop	D019297	5485731	376965	Toxin secretion by nonreplicating cells was inhibited
2,2-dipheny	C004931	24663118	3967231	The antioxidant activity of extracts was performed b
2,2-dipheny	C004931	28219296	6130586	The antioxidant activity was examined using 2,2-dipl
2,3,-butane	D003931	15141066	429332	Our studies revealed that the major VOCs released a
2,3-Dihydrc	C009135	22609919	3416551	2,3-Dihydroxybenzoate is the precursor in the biosyr
2,3-butaned	C026978	28425950	6154464	Elucidation of the isolates secondary metabolites sh
2,3-dihydro	C009135	22609919	3416551	The dhb cluster of Pseudomonas reinekei MT1 enco

Number of unique chemicals: 913

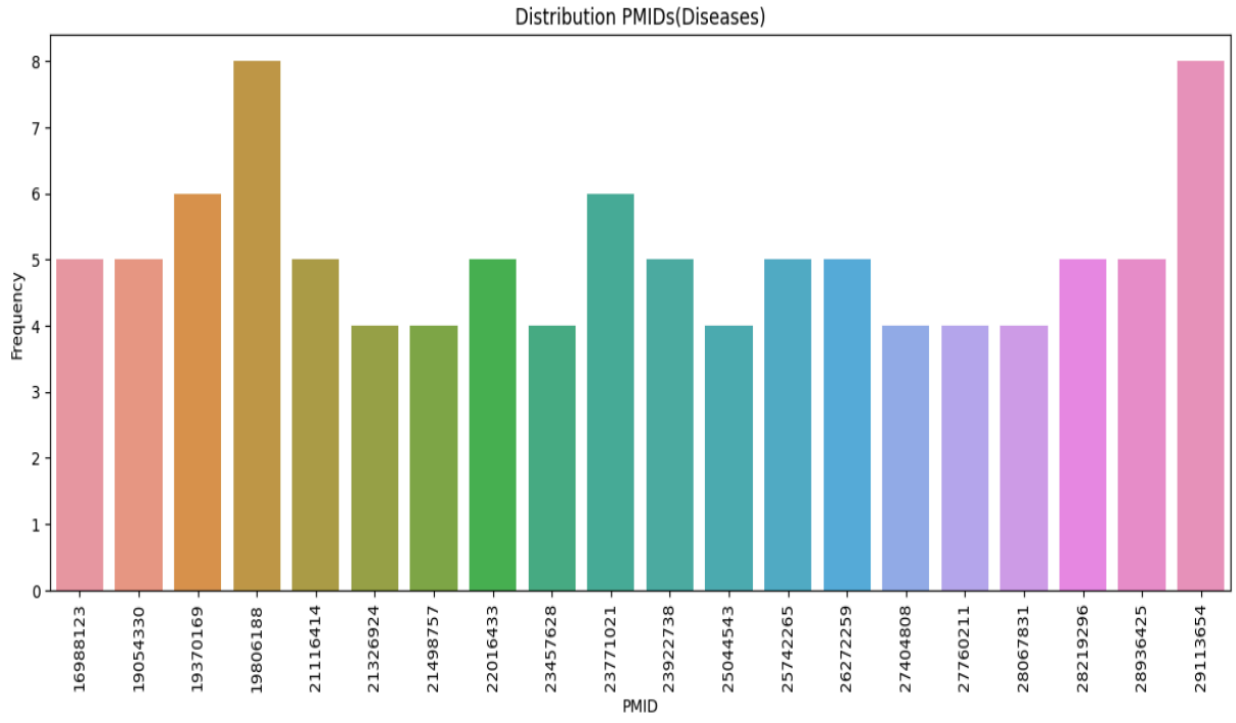

Distribution PMIDs(Chemicals)
Frequency
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
11402212
11673439
11823198
15141066
17997835
21672958
23708134
24445449
25769826
25970559
26195520
26350134
26655755
27002332
27422837
27548127
27842527
28219296
28425950
28432098
PMID

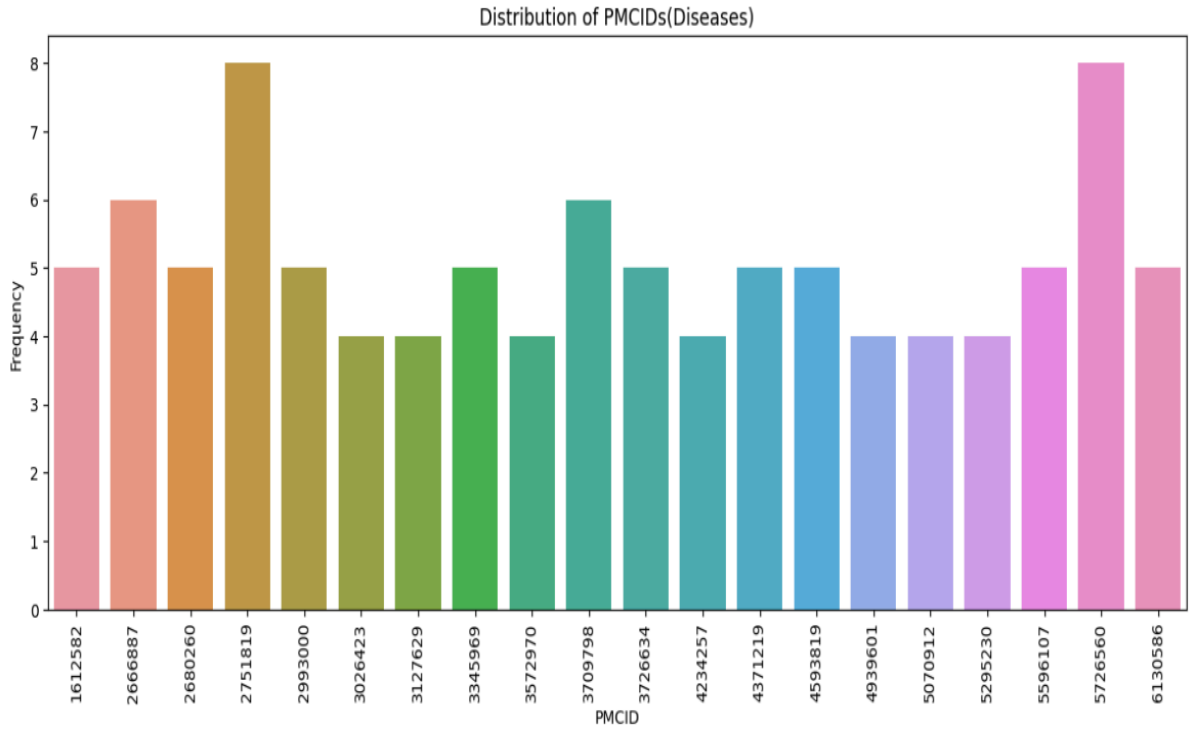

Distribution of PMCIDs(Chemicals)
Frequency
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
95500
111174
126662
429332
2258182
3125804
3719553
3950275
4407226
4430277
4576117
4626903
4771327
4930463
5038050
5396551
5478978
6130586
6154464
6274189
PMCID

Task4_Secondary_Metabolites_Bacteria_Diseases_Pubmed.csv

Diseases	DiseasesID	PMID	PMCID	Sentences
Allosalinactir	C000657245	26864220	4749953	In genus level, strain Allosalinactinospora lopnorensis CA15-2T is the first new actinomycete isolated from the Lo
Antarctic sub	D007246	27128927	4882557	Antarctic sub-sea sediments were collected from the Ross Sea, and used to isolate 25 microorganisms, which wer
BRAF-wild-ty	D003141	28923537	5602478	Whereas many melanoma patients exhibit profound response to ICT, there are fewer options for patients failing
Bifurcaria bif	C537702	24663118	3967231	Bifurcaria bifurcata epiphytic bacteria were revealed to be excellent sources of natural antioxidant and antimicro
British Colum	176500	24130838	3794959	
CHA0 toxicity	D064420	17088380	1636139	By analyzing the responses of protists to bacterial supernatants obtained from different isogenic mutants whose
California ha	C537062	23229438	4517938	These results indicate that marine sediments from the Gulf of California harbor diverse Actinobacterial taxa with
Cancer	D009369	18955814	2963942	Cancer cells are more sensitive to the proapoptotic effects of proteasome inhibition than normal cells.
Clonostachys	D017515	28438205	5404306	
Colon Tumor	D003110	19370169	2666887	Over 2,000 actinomycetes were isolated and of these approximately 20%, 5%, and 10% inhibited the growth of Hu
Crohn's disea	D003424	22016433	3345969	In addition, specific metabolic profiles can function as a diagnostic tool for the identification of several gastrointe
Cutaneous cc	D017577	4911442	376755	Cutaneous cocci are known to be ureolytic but few diphtheroids had urease activity.
DW	D015431	28219296	6130586	
Dysbiosis	D064806	28936425	5596107	
Entomopath	C537702	24599183	3944712	Entomopathogenic bacteria Xenorhabdus spp. produce secondary metabolites with potential antimicrobial activit
Frankia genu	D056304	21498757	3127629	This work supports the value of bioinformatic investigation in natural products biosynthesis using genomic inforn
Frankia genu	D056304	20190089	2849203	While these mechanisms are well studied in the rhizobia-legume symbiosis, little is known about the role of plant
Glucose Soya	D044882	25406714	4243295	Glucose Soyabean meal broth was found to be the suitable medium for antibiotic production at 28Â°C for seven d
HIV	D015658	21116414	2993000	Some of these bioactive secondary metabolites of microbial origin with strong antibacterial and antifungal activiti
HMQC, (1)H-	D000848	22611356	3347017	

Number of unique Diseases: 230

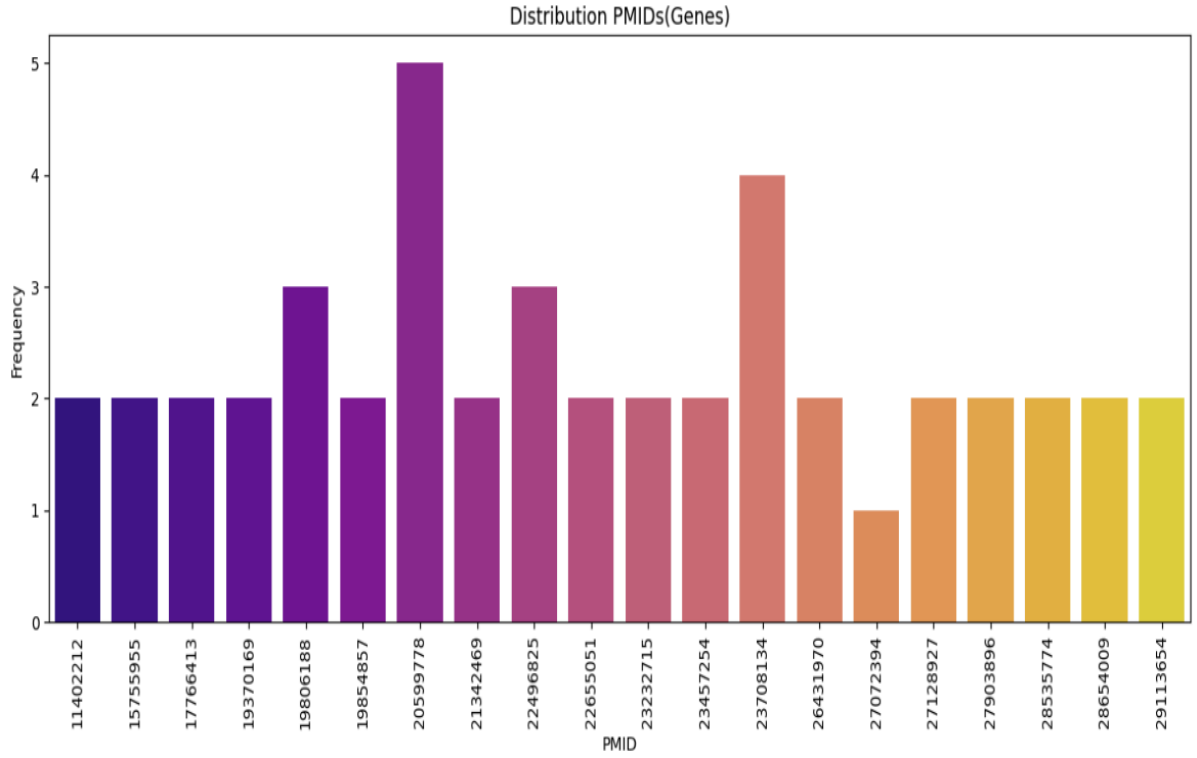


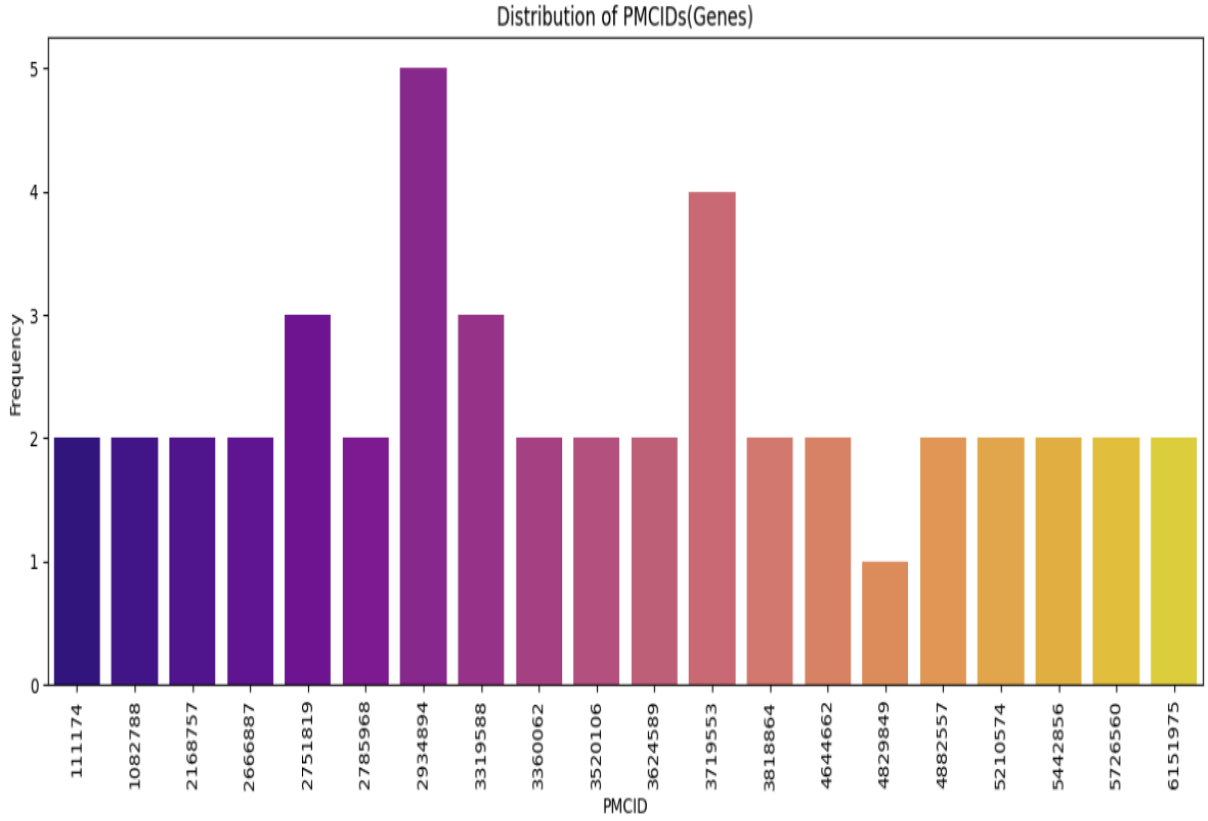


Task4_Secondary_Metabolites_Bacteria_Genes_Pubmed.csv

Genes	GenesID	PMID	PMCID	Sentences
A2a	28882	28358337	6154602	Two novel benzamido nonacyclic dilactones, namely neoantimycins A (1)
ABC	10058	27903896	5210574	In order to advance the exploration of microbial secondary metabolism,
Apr	5366	23708134	3719553	Another important category of the GacA regulon was secretion systems, i
Atpdr2	827187	19854857	2785968	After two generations, only the Arabidopsis abcg30 (Atpdr2) mutant had
B17	4712	21429787	3947797	Today, this burgeoning class of natural products encompasses a structura
CD274	29126	29113654	5726560	Emerging data demonstrate that intestinal bacteria can modulate the effi
CTLA4	1493	29113654	5726560	Emerging data demonstrate that intestinal bacteria can modulate the effi
Crp	20468888	23232715	3520106	Cyclic AMP receptor protein (Crp) is a transcription regulator controlling d
ER	2099	20599778	2934894	Estrogen Receptor (ER) dimerization is required for target gene transcript
ERalpha/a	2099	20599778	2934894	
ERalpha/b	2099	20599778	2934894	
ERbeta/be	2099	20599778	2934894	
Esi	5266	28535774	5442856	We
Estrogen F	2099	20599778	2934894	Estrogen Receptor (ER) dimerization is required for target gene transcript
Gac	2744	23708134	3719553	The antibiotic biosynthesis of M18 is coordinately controlled by multiple
GcvT	275	28784988	5547118	Further characterization of several other berberine-utilizing bacteria and
Has	3036	23708134	3719553	Another important category of the GacA regulon was secretion systems, i
IF1	93974	15755955	1082788	Subsequent sections describe the structure, function, and interactions of
IF2	9669	15755955	1082788	Subsequent sections describe the structure, function, and interactions of
MCT1	6566	26272259	4593819	Monocarboxylate transporter 1 (MCT1) plays a major role in colonic lumi

Number of unique Genes: 69

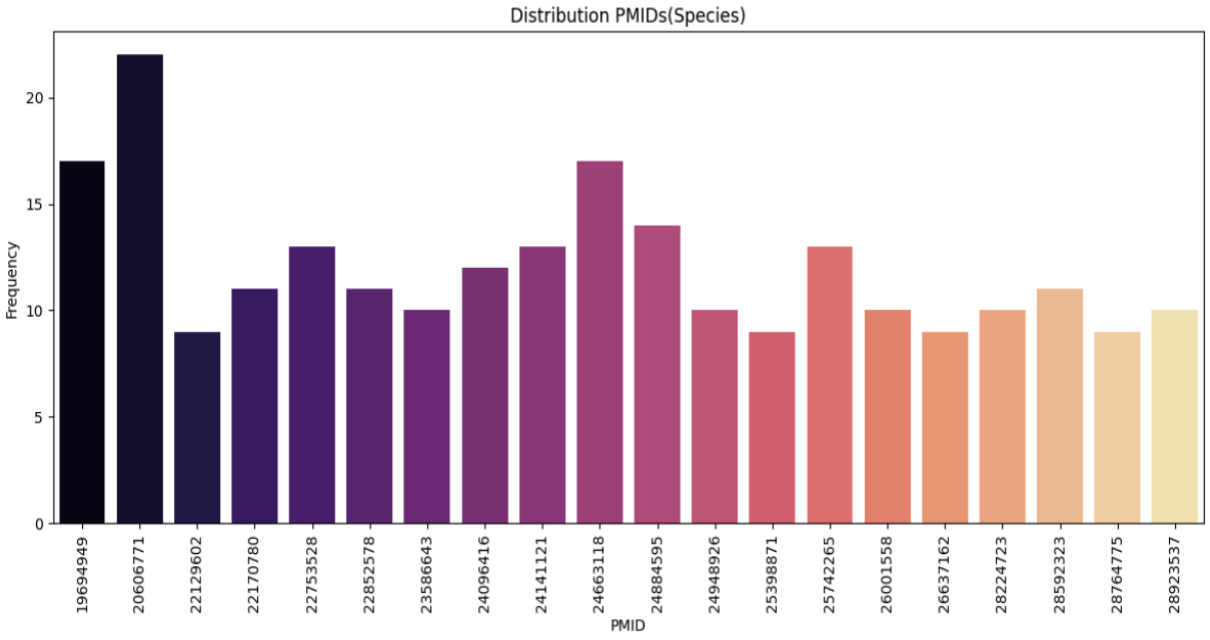


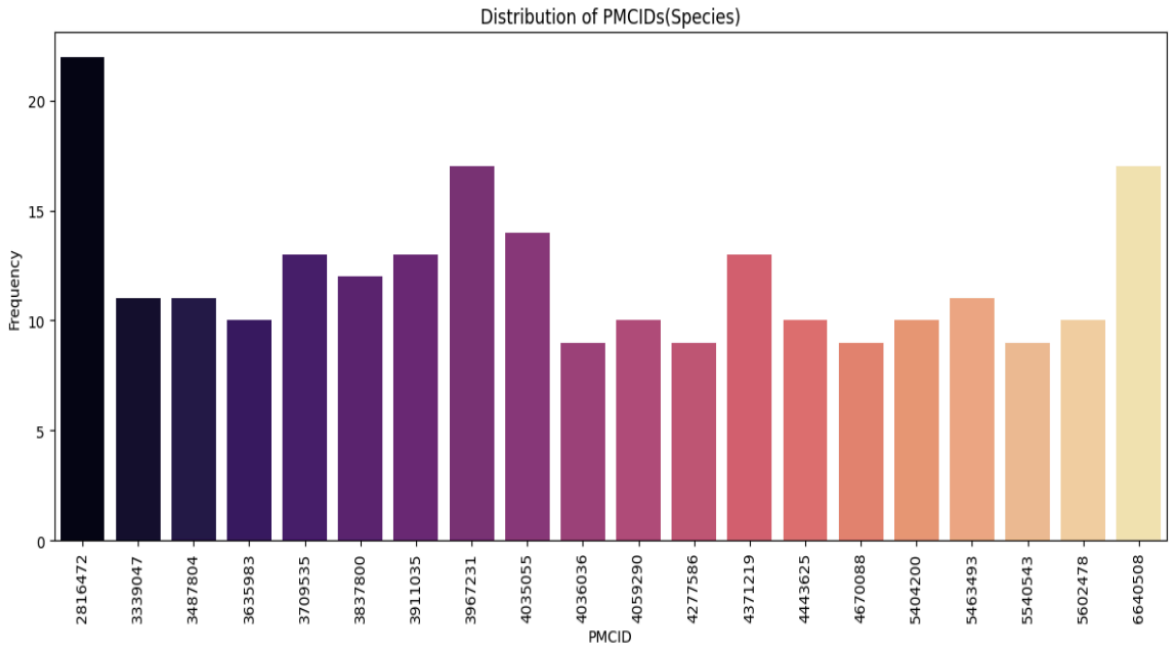


Task4_Secondary_Metabolites_Bacteria_Species_Pubmed.csv

Species	SpeciesID	PMID	PMCID	Sentences
A. annua	35608	24575401	3915762	In
A. annua L	212759	24575401	3915762	In this study, the leaves of three in vitro A. annua L. clones were, extracted and two
A. bauman	470	28654009	6151975	The minimal inhibitory concentrations (MIC) of compounds 1 and 2 against S. aureu
A. bauman	470	28421168	5376624	Therefore, other bacterial phenotypes were analyzed to determine the role of this g
A. bauman	400667	28421168	5376624	Therefore, other bacterial phenotypes were analyzed to determine the role of this g
A. brasilien	37326	22768320	3388064	In the
A. chinensi	217632	24069601	3771249	In this study the gut bacterial community of larvae and adults of A. chinensis, collect
A. clavaton	41054	29049321	5648158	We
A. echinatic	103372	21245311	3033269	Valinomycins and actinomycins were also directly identified in samples from the was
A. echinatic	103372	28717220	5514151	These
A. flavus	5059	24948950	4059316	In overall, 120 different extracts were prepared, their HPLC profiles were obtained a
A. fumigatu	746128	25532893	4286458	(A) Phenazine biosynthesis pathway of P. aeruginosa (solid arrows) and bioconversio
A. hymenia	340345	28770445	5676828	Digital DNA: DNA hybridization and ANI values between strains ADI 127-17T and GBA
A. japonicu	208439	25114137	4187939	In contrast to most Amycolatopsis strains, A. japonicum is genetically tractable with
A. kerguele	254788	27515463	5052388	The strain A. kerguelensis VL-RK_09 exhibited a broad spectrum of in vitro antimicro
A. nidulans	162425	25532893	4286458	Gliotoxin can signal A. nidulans development under mildly reducing conditions throu
A. niger	5061	21245311	3033269	Valinomycins and actinomycins were also directly identified in samples from the was
A. niger	5061	20606771	2816472	Also, these two species of Cola demonstrated activities on C. albicans and A. niger at
A. ochraceu	40380	23569796	3614191	To isolate and characterize the bioactive secondary metabolites from Aspergillus och
A. odoratis	158562	28219296	6130586	

Number of unique Species: 1031





5. Database Compilation

A comprehensive compilation of databases related to secondary metabolites from bacteria was created, summarizing available databases, publication years, and descriptions. This serves as a valuable resource for scientists exploring bacterial secondary metabolites' chemical diversity, biosynthetic pathways, and biotechnological applications.

I've assembled a CSV file summarizing two databases focusing on secondary metabolites derived from various bacterial species. Each entry provides links to the respective database and publication, along with the publication year and a concise description. This compilation stems from keyword searches focusing on "Secondary Metabolites from Bacteria."

Understanding bacterial secondary metabolites is crucial for various fields, from drug discovery to synthetic biology. Such research sheds light on the biosynthesis and potential applications of these compounds, including their roles in ecological interactions and potential therapeutic uses. These databases serve as valuable resources for scientists exploring the chemical diversity, biosynthetic pathways, and biotechnological applications of bacterial secondary metabolites.

1. StreptomeDB (2013)
StreptomeDB is a repository of compounds obtained from Streptomyces spp. The data within it was gathered through a combination of text mining and manual creation of numerous abstracts and full papers, utilizing both an in-house platform and two external databases.
StreptomeDB

2. DoBISCUIT (2013)
The latest iteration of DoBISCUIT is dedicated to bacterial-derived secondary metabolites, with a particular emphasis on those originating from actinomycetes. The primary dataset of DoBISCUIT is derived from entries in the INSDC, each detailing a biosynthetic cluster associated with a known bacterial secondary metabolite.
DoBISCUIT

Submitted File

Available Databases:

Task5_Available_databases.csv

Name	Descriptio	Publicatio	Publicationlink
Streptome	Streptome	2013	http://www.pharmbioinf.uni-freiburg.de/streptomedb
DoBISCUI	atest itera	2013	https://www.nite.go.jp/en/index.html

6. Protein and Gene Analysis

1. Dataset Overview
    - The dataset contains 123,135 entries with the following columns:
    - Entry: Unique identifier for each protein.
    - Entry Name: Descriptive name for the entry.
    - Protein names: Names of the proteins.
    - Gene Names: Names of the genes associated with the proteins.

- Length: Length of the protein.
- PubMed ID: References to PubMed articles.
- GeneID: Unique identifiers for the genes

6.1. Submitted File

Task6_Secondary_Metabolites_Bacteria_Uniprot_Pubmed.csv

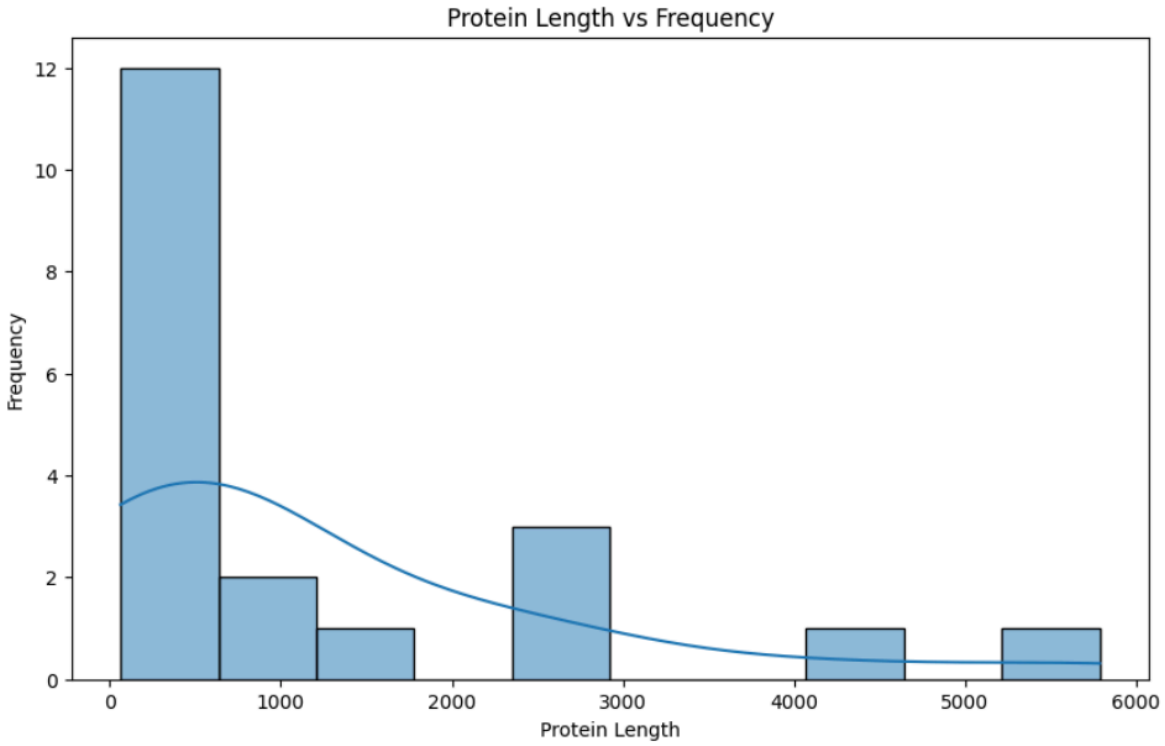
Entry	Entry Name	Protein names	Gene Names	Length	PubMed ID	GeneID
A0A1C0U65	A0A1C0U653_	Fatty acid oxidation complex su	fadB Ppb6_01307	728		
A0A1C0U9C	A0A1C0U9Q7_	Siroheme synthase [Includes: U	cysG_1 cysG Ppb6_00115	470		
A0A1W7CR	A0A1W7CRWC	Uncharacterized protein	CAG99_00520	2414		
A0A370RUE	A0A370RUE2_	Bifunctional NAD(P)H-hydrate r	nnrE nnrD DER30_2614	492		
A0A495LAV	A0A495LAV2_9	Acyl transferase domain-conta	DFP74_6145	2400		
A0A5Q0GYE	A0A5Q0GYB2_	Type I polyketide synthase	EKG83_17110	4092	33136202	
A0A6I6WMI	A0A6I6WMB4_	Polyketide synthase	DEH18_16130	5786		
A0A7V8STA	A0A7V8STA4_9	SDR family NAD(P)-dependent	E0L36_16905	2425		
G3Y419	YANA_ASPNA	6-methylsalicylic acid synthase	yanA pks48 ASPNIDRAFT_4	1779	21543515; 11031048; 24	
P0DPC3	CSRA1_PSEPH	Translational regulator CsrA1 (	csrA1 rsmE	64	15601712; 16	70104632;
P14779	CPXB_PRIM2	Bifunctional cytochrome P450/	cyp102A1 cyp102 BG04_16	1049	2544578; 25931591; 154	
P54292	RHLR_PSEAE	HTH-type quorum-sensing regu	rhlR lasM vsmR PA3477	241	8522523; 814	77220015;8
P9WKL5	HTM_MYCTU	2-heptyl-1-hydroxyquinolin-4(1	htm Rv0560c	241	9634230; 155	887637;
Q0VZ68	TAM_CHOCO	Tyrosine 2,3-aminomutase (EC	cmdF	531	16793524; 17545150; 19	
Q82P90	IABF_STRAW	Extracellular exo-alpha-(1->5)-I	Araf43A abfA SAVERM_104	481	11572948; 12692562; 18	
Q8GMG0	TAM_STRGL	MIO-dependent tyrosine 2,3-aminomutase (EC 5.4.3.6) (Ty		539	12183628; 17516659; 18	
Q96CG3	TIFA_HUMAN	TRAF-interacting protein with F	TIFA T2BP	184	12566447; 17	92610;
Q9ASP6	HNRPQ_ARATH	Heterogeneous nuclear ribonu	LIF2 At4g00830 A_TM018A	495	10617198; 27	827998;
Q9C9M3	PXMT1_ARATH	Paraxanthine methyltransferas	PXMT1 At1g66700 F4N21.1	353	11130712; 27	842988;
Q9KZC7	GLNA3_STRCO	Gamma-glutamylpolyamine syn	glnA3 SCO6962	466	12000953; 16932908; 28	

Analysis

1. Distribution of Protein Lengths:

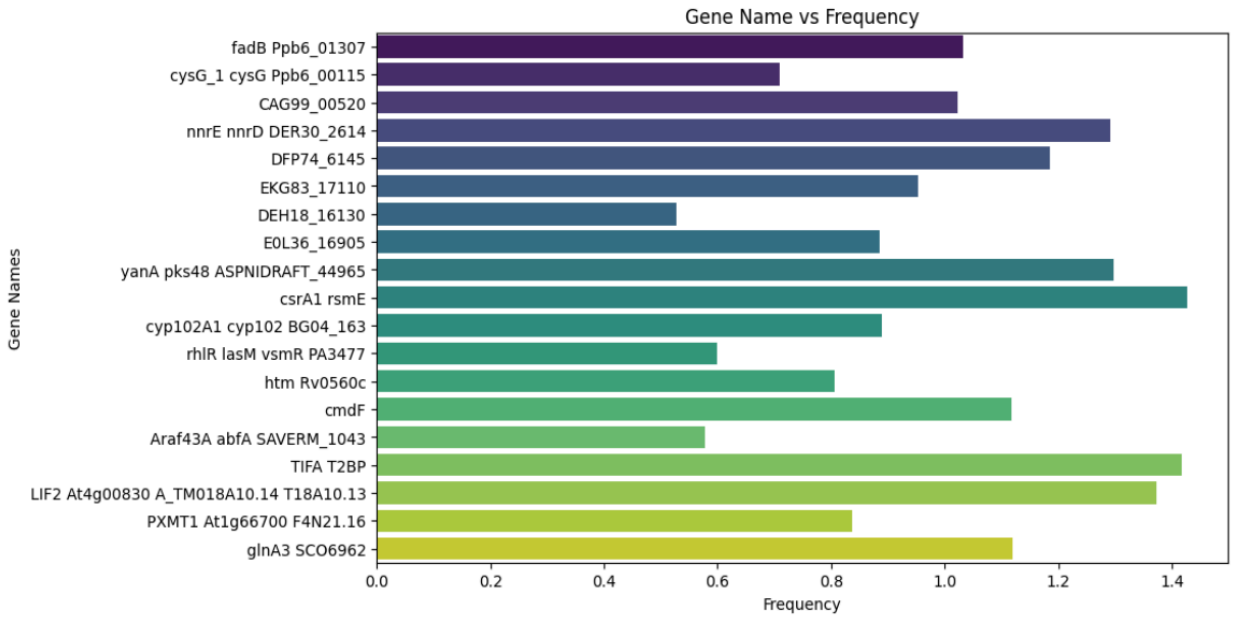
- The purpose of analyzing the distribution of protein lengths is to understand the range and frequency of protein sizes in your dataset. This analysis can provide insights into the characteristics of proteins present in your data and help identify patterns or anomalies in their lengths.

- Analyzing the distribution of protein lengths involves visualizing how protein sizes are spread across the dataset. By creating histograms or density plots, you can identify common protein lengths, their distribution, and overall patterns, such as whether the distribution is normal or skewed. This helps in understanding the typical size of proteins and any significant variations in their lengths.



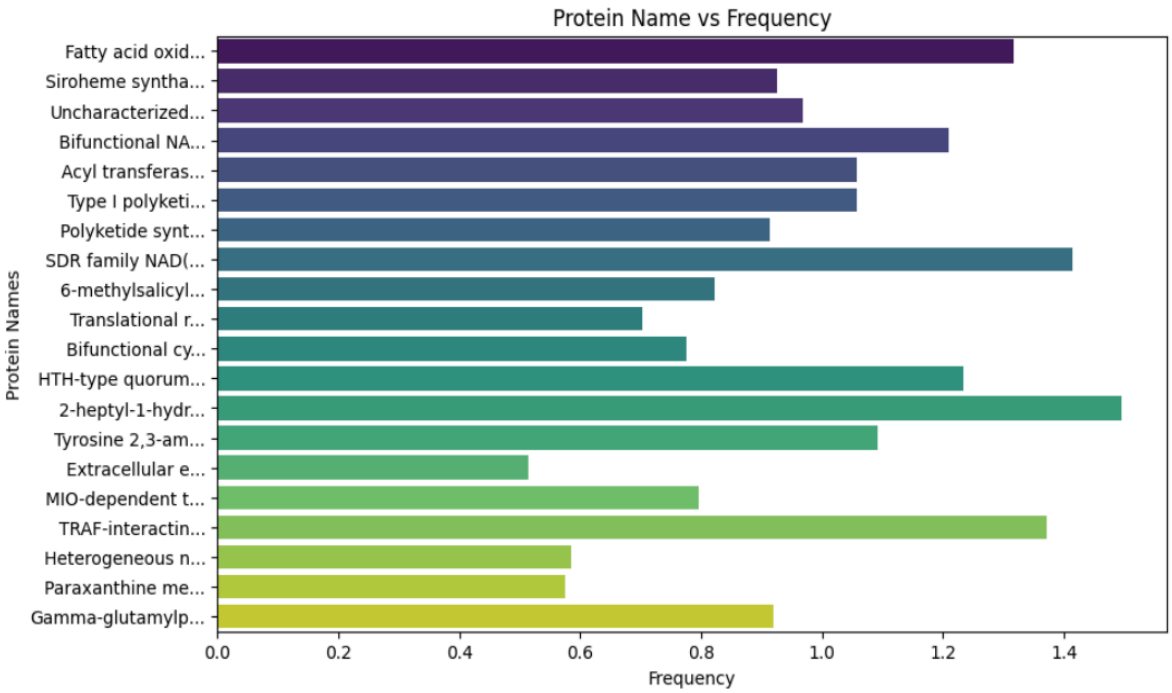
2. Distribution of Gene name :

- To analyze the frequency and variety of gene names in the dataset, which can highlight the prevalence of specific genes and reveal any patterns or imbalances in gene representation.
- Examining the distribution of gene names involves visualizing how often each gene name appears in the dataset. By creating bar plots or other frequency-based charts, you can identify which genes are most common, which are less frequent, and any notable patterns in gene occurrences. This analysis helps in understanding the representation of different genes and can reveal if some genes are overrepresented or underrepresented in the dataset.



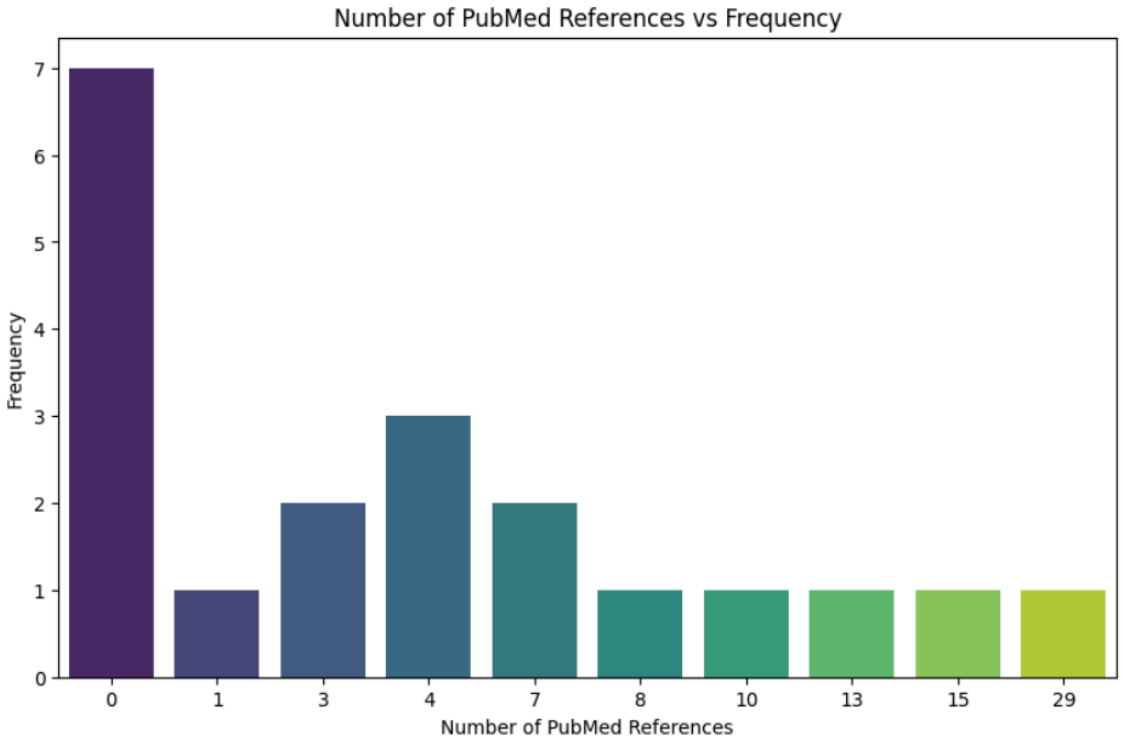
3. Distribution of protein name :

- To evaluate the frequency and variety of protein names in the dataset, providing insights into which proteins are most commonly represented and highlighting any trends or imbalances in their occurrences.
- Analyzing the distribution of protein names involves visualizing how frequently each protein name appears in the dataset. By using bar plots or frequency charts, you can identify which protein names are most prevalent, observe patterns in their distribution, and detect any anomalies or biases. This helps in understanding the representation of different proteins and assessing the overall diversity of protein names in the dataset.



4. Distribution of Pubmed reference :

- To analyze the number of PubMed references associated with each entry in the dataset, revealing the extent of research and literature coverage for these entries.
- Examining the distribution of PubMed references involves visualizing how many references each entry has. By creating bar plots or other frequency charts, you can identify the most and least referenced entries, understand the spread of references, and detect any patterns in literature coverage. This helps in assessing the depth of research associated with different entries and understanding the relationship between entries and their corresponding PubMed references.



7. Dataset on Occurrence and Geographic Spread of Secondary Metabolites in Bacterial Environments:

Dataset Analysis

1. Geographic Distribution**:**

- Southern California: Focuses on bacterial and chemical composition in ascidians across various geographical regions, indicating diverse microbiomes and metabolomes specific to species and locations.
- Asia, Europe, Australia, North America: Examines variations in bacterial communities and secondary metabolites during a cyanobacterial bloom in a tropical reservoir, highlighting spatial over temporal variations.
- United States, Philippines, Brazil: Investigates shipworms and their symbiotic bacteria, revealing consistent distribution patterns of symbiotic bacteria and diverse gene clusters for biosynthesizing secondary metabolites.
- Europe: Studies aflatoxins in maize, focusing on contamination in French fields and identifying aflatoxigenic fungi prevalent in regions with climate-induced changes.

- Pakistan: Analyzes cyanotoxin occurrence and cyanobacterial diversity in water reservoirs, providing insight into cyanotoxin risks and cyanobacterial community structure.
- Agricultural Soils of China: Explores the genetic diversity of bacterial secondary metabolites in soils, revealing significant environmental and ecological impacts.
- Malaysia (Johor Bahru, Terengganu, Penang): Investigates the relationship between indoor microbial exposure and asthma symptoms, highlighting potential indicators from metabolites and chemicals.
- Caribbean Sea and Western North Atlantic Ocean: Examines bacterial communities associated with octocorals, revealing the influence of host-specific characteristics on bacterial community stability.
- Sea Urchins: Reviews spinochrome secondary metabolites, discussing their chemistry, biosynthesis, and potential applications.
- Hot Springs in Maharashtra: Identifies novel bacteria with antimicrobial properties, focusing on strains from hot springs and their potential against antibiotic-resistant pathogens.

2. Temporal Trends**:**

- Southern California: Provides insights into the stability of microbiomes and metabolomes over time within specific species.
- Asia, Europe, Australia, North America: Short-term (2 days) variations in bacterial communities and secondary metabolites during a bloom, with spatial variations being more pronounced.
- Europe: Historical data on aflatoxin contamination influenced by climate conditions, emphasizing temporal changes due to weather patterns.
- Pakistan: Initial snapshot with potential for future temporal studies on cyanotoxin occurrence and bacterial diversity.
- Agricultural Soils of China: No specific temporal analysis but offers a baseline for understanding bacterial secondary metabolism diversity.

3. Key Findings and Implications**:**

- Diversity of Secondary Metabolites: Across different environments, including ascidians, shipworms, and soil, there is a wide range of secondary metabolites. This suggests significant ecological and evolutionary roles.
- Geographic Variations: Secondary metabolites and bacterial communities show geographic specificity, highlighting the impact of local environmental conditions on microbial diversity and metabolite production.

- Environmental and Ecological Impacts: Studies indicate the influence of environmental factors such as temperature, pH, and climate on the presence and abundance of secondary metabolites and bacterial communities.
- Biotechnological Potential: The dataset reveals novel sources of bioactive compounds and potential applications in pharmaceuticals and environmental management.
- Risk Assessment: Some studies focus on contamination risks (e.g., aflatoxins, cyanotoxins), emphasizing the need for monitoring and management in affected regions.

7.1 File Submitted

Task7_Secondary_Metabolites_Bacteria_GeoData_Pubmed.csv

Location	Epidemic/I	Description	Link	Publication Link	Year of Publication
Southern Caliphornia	NA	This study expl	https://pubm	10.1038/ismej.201	2014
Asia, Europe, Australia, and North America	NA	In assessing wa	https://pubm	10.1021/acs.est.6b	2017
United State, philippines, Brazil	NA	This study inve	https://pubm	10.1128/mSystems	2020
Europe	NA	Aflatoxins (AFs	https://pubm	10.3390/toxins1012	2018
Pakistan	NA	This study inve	https://pubm	10.1007/s11356-02	2024
Agricultural soils of China	NA	This study inve	https://pubm	10.1128/msystems	2024
Malaysia (Johor Bahru, Terengganu and Penang)	NA	This study inve	https://pubm	10.1183/13993003.	2022
Caribbean Sea and western North Atlantic Ocean	NA	This study inve	https://pubm	10.1007/s13199-02	2023
Sea Urchins	NA	This review exp	https://pubm	10.1039/c8ra04777	2018
Hot-springs in Maharashtra	NA	This study inve	https://pubm	10.4103/0974-777X	2011

8. Analysis of Proteins and Genes Essential for Survival

Analysis of Proteins and Genes Associated with Secondary Metab**olites**

1. NRPS (Non-Ribosomal Peptide Synthetase)

- Protein: NRPS enzyme complex
- Description: This study investigates the distribution of secondary metabolite biosynthetic gene clusters (SMBGCs) in the Hahella genus. It identifies 70 SMBGCs, with a significant proportion belonging to NRPS and NRPS/PKS families. The strain NBU794 from mangrove soil exhibits high potential for bioactive compound production, including prodigiosin with notable antimicrobial activity. The findings highlight Hahella species' potential for synthesizing novel bioactive compounds.
- Link: PubMed

- Publication Link: 10.3390/md20040269

2. PKS (Polyketide Synthase)

- Protein: PKS enzyme complex
- Description: This research explores the structure and dynamics of type I fatty acid synthases (FASs) and polyketide synthases (PKSs). It provides insights into the architectures of these enzymes, including their iterative and modular assembly processes. High-resolution studies reveal details of NR-PKS action, advancing understanding of PKS functionality and evolution. The study emphasizes the diverse architectures and dynamics of PKS enzymes.
- Link: PubMed
- Publication Link: 10.1039/c8np00039e

3. Hybrid NRPS-PKS

- Protein: Hybrid NRPS-PKS enzyme complex
- Description: This study examines Serratia plymuthica RVH1, which produces polyamino antibiotics with potent antimicrobial activity. The gene cluster responsible for biosynthesis includes FAS, PKS, NRPS, and tailoring/export enzymes. The research uncovers a complex interplay between lipid and metabolite biosynthesis, suggesting unique mechanisms in antibiotic production and potential for novel antibiotic discovery.
- Link: PubMed
- Publication Link: 10.1371/journal.pone.0054143

4. DMATS (Dimethylallyl Tryptophan Synthase)

- Protein: DMATS enzyme
- Description: This study focuses on DMATS, crucial for ergot alkaloid biosynthesis. The X-ray structure of DMATS from Aspergillus fumigatus reveals its interaction with substrates and the enzyme's regiospecific reaction. This provides insights into the enzyme's role in secondary metabolite biosynthesis and the structural basis for its function, including similarities with bacterial enzymes.
- Link: PubMed
- Publication Link: 10.1073/pnas.0904897106

5. TP53 (Tumor Protein 53)

- Protein: Tumor Protein 53
- Description: This study assesses the effects of Lactobacillus plantarum secondary metabolites on breast cancer cells and Drosophila melanogaster. It measures toxicity, gene expression, and binding affinity, revealing potential therapeutic applications.

Secondary metabolite L13 shows high binding affinity and suggests promise for cancer treatment.

- Link: PubMed
- Publication Link: 10.1007/s00280-019-03978-0

8.1 Submitted file:

Task8_Secondary_Metabolites_Bacteria_Responsible_Protein_Pubmed.csv

Gene	Protein	Description	Link	Publication Link
NRPS	NRPS enzyme complex	This study focuses	https://pubmed.nc	10.3390/md20040269
PKS	PKS enzyme complex	Type I fatty acid sy	https://pubmed.nc	10.1039/c8np00039e
Hybrid NR	Hybrid NRPS-PKS enzyme complex	Serratia plymuthic	https://pubmed.nc	10.1371/journal.pone.0054143
DMATS	DMATS enzyme	Ergot alkaloids, bc	https://pubmed.nc	10.1073/pnas.0904897106
TP53	Tumor Protein 53	The study evaluat	https://pubmed.nc	10.1007/s00280-019-03978-0

9. Drug Information

- The dataset offers detailed information on various pharmaceutical compounds, including the chemical name, unique identifier (Chemical ID), and URLs to in-depth profiles. These profiles provide comprehensive details about the drugs' properties, uses, and research findings, offering a valuable resource for understanding each compound's characteristics.
- It also specifies the current clinical trial phase for each drug, categorizing them as Approved, Experimental, Investigational, Illicit, or Withdrawn. This classification helps track the developmental progress and regulatory status of the drugs, aiding researchers, healthcare professionals, and regulatory bodies in making informed decisions regarding drug development and healthcare strategies.

9.1. Submitted File

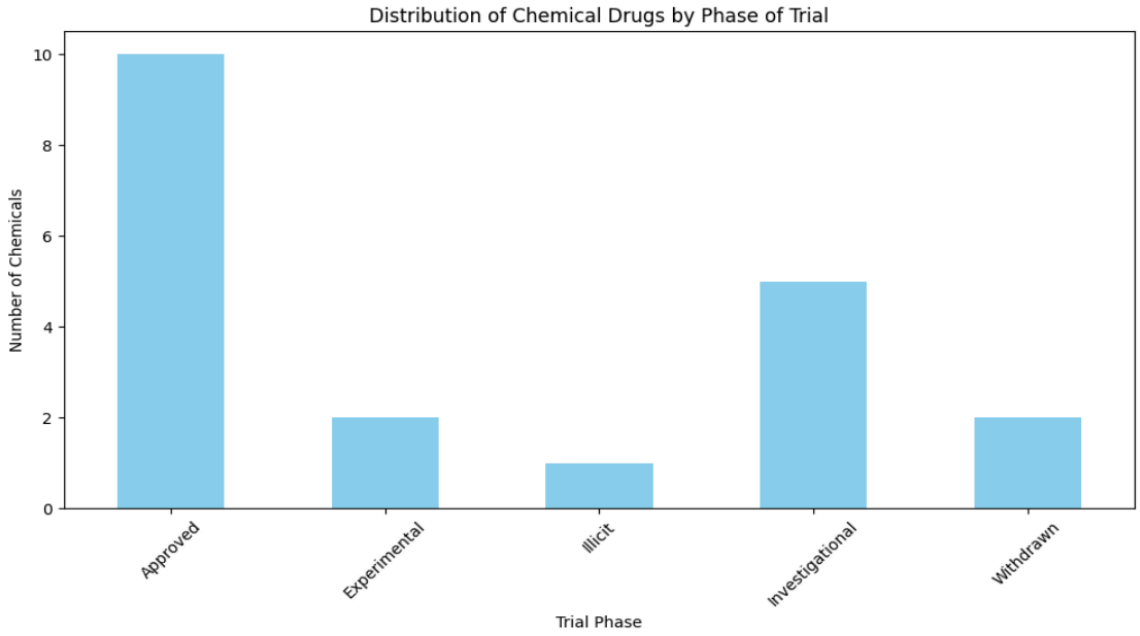
Task9_Secondary_Metabolites_Bacteria_Chemicals_drug_Pubmed.csv

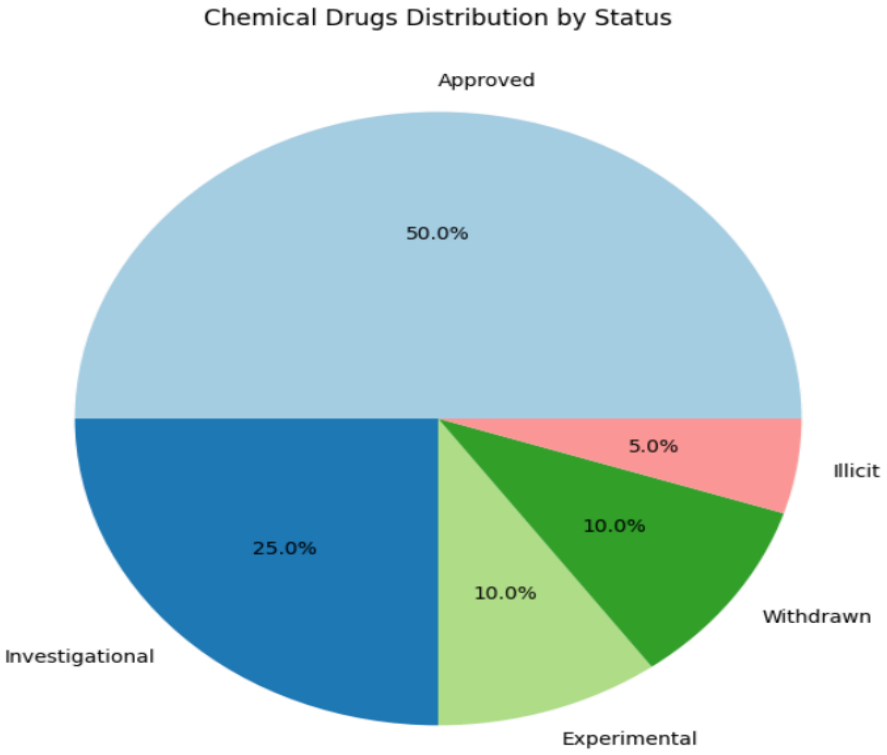
Chemical Na	Chemical ID	Reference	Phase of Trial
Doxercalcife	DB06410	https://go.drugbank.com/drugs/DB0	Approved
Didecyldime	DB04221	https://go.drugbank.com/drugs/DB0	Approved, Experimental
Sibutramine	DB01105	DB01105	Approved, Illicit, Investigational, Withdrawn
Nomifensine	DB04821	https://go.drugbank.com/drugs/DB0	Approved, Withdrawn
Etelcalcetide	DB12865	https://go.drugbank.com/drugs/DB1	Approved, Investigational
Colchiceine	DB15534	https://go.drugbank.com/drugs/DB1	Approved, Experimental
Pretomanid	DB05154	https://go.drugbank.com/drugs/DB0	Approved
Balsalazide	DB01014	https://go.drugbank.com/drugs/DB0	Approved, Investigational
Testosteron	DB13946	https://go.drugbank.com/drugs/DB1	Approved, Investigational
Tolfenamic a	DB09216	https://go.drugbank.com/drugs/DB0	Approved, Investigational

Analysis:

Dataset Overview:

- Chemical Name: The common or commercial name of the chemical drug. It represents the name by which the chemical is known in the pharmaceutical industry or market.
- Chemical ID: A unique identifier assigned to the chemical drug, typically used in databases like DrugBank. This ID allows for precise referencing and retrieval of detailed information about the chemical.
- Reference: A URL link to a detailed profile of the chemical drug, often hosted on a resource like DrugBank. This link provides comprehensive information about the chemical, including its properties, uses, and regulatory status.
- Phase of Trial: The current status of the chemical drug in clinical trials or investigations. This can include:
    - Approved: The chemical has been approved for use by regulatory agencies.
    - Experimental: The chemical is undergoing testing but is not yet approved.
    - Investigational: The chemical is being studied for potential use but is not yet in widespread use.
    - Illicit: The chemical is prohibited or banned for use.
    - Withdrawn: The chemical has been removed from the market or investigation.


Distribution of Chemical Drugs by Phase of Trial
Number of Chemicals
0
2
4
6
8
10
Approved
Experimental
Illicit
Investigational
Withdrawn
Trial Phase


Chemical Drugs Distribution by Status
Approved
50.0%
5.0%
Illicit
10.0%
Withdrawn
10.0%
Experimental
25.0%
Investigational

The visualizations depict the distribution of chemical drugs across various trial phases. Most drugs fall into the "Approved" category, while fewer are in the "Investigational," "Experimental," “illicit”or "withdrawn" phases. This data provides a clear picture of the current state of drug trials, highlighting that a significant proportion of the drugs have achieved approval for use.

10. Co-occurrence Analysis

Applied Python libraries and machine learning to explore correlations among extremophiles' attributes. Analyzed co-occurrence of Chemicals and Gene attributes using CSV sentence data. Mapped connections between chemicals and genes. Created three CSV files: one for chemical interactions, one for gene interactions, and one for chemical-to-gene relationships. Used Pandas for efficient data processing and analysis, facilitating the extraction and mapping of key associations.

10.1. Submitted Files

Task10_Genes_to_Chemical.csv

PMID	PMCID	Sentence	Genes	ChemicalName	Interaction type	Regulation
28358337	6154602	Two novel be	A2a	Neoantimycins		
28358337	6154602	Two novel be	A2a	antimycins		
28358337	6154602	Two novel be	A2a	benzamide		
23708134	3719553	Another impc	Apr	PCA		up
23708134	3719553	Another impc	Apr	Plt		up
23708134	3719553	Another impc	Apr	hydrogen cyanide		up
23708134	3719553	Another impc	Apr	nitrogen		up
23708134	3719553	Another impc	Apr	phenazine-1-carboxylic acid		up
23708134	3719553	Another impc	Apr	phosphorus		up
23708134	3719553	Another impc	Apr	pyochelin		up
23708134	3719553	Another impc	Apr	pyoluteorin		up
23708134	3719553	Another impc	Apr	pyoverdine		up
23708134	3719553	Another impc	Apr	sulfur		up
19854857	2785968	After two gen	Atpdr2	metal		
19854857	2785968	After two gen	Atpdr2	nitrogen		
21429787	3947797	Today, this bu	B17	cyanobactins		up
21429787	3947797	Today, this bu	B17	cysteine		up
21429787	3947797	Today, this bu	B17	oxazole		up
21429787	3947797	Today, this bu	B17	peptides		up
21429787	3947797	Today, this bu	B17	serine		up

Analysis:

1. Data Structure
   The dataset includes the following columns:

- PMID: PubMed Identifier - a unique ID for the article in the PubMed database.
- PMCID: PubMed Central Identifier - a unique ID for the article in the PubMed Central database.
- Sentence: A sentence from the article that describes the interaction.
- Genes: The gene involved in the interaction.
- Chemicals: The chemical involved in the interaction.
- Interaction type: Specifies the nature of the interaction (e.g., Agonist).
- Regulation: Indicates the direction of regulation (e.g., up-regulated or down-regulated).

2. Data Characteristics

- The Sentence column provides detailed descriptions of interactions between genes and chemicals in relation to diseases.
- The Genes and Chemicals columns list the interacting gene and chemical, respectively.
- The Interaction type column defines the nature of the interaction, such as agonist or antagonist.
- The Regulation column indicates whether the interaction causes upregulation or downregulation of gene expression.

Task10_Genes_to_DiseaseName.csv

PMID	PMCID	Sentence	Genes	DiseaseName	Interaction type	Regulation
28358337	6154602	Two novel I	A2a	breast adenocarcinoma		
28358337	6154602	Two novel I	A2a	cytotoxicities		
28358337	6154602	Two novel I	A2a	glioblastoma		
28358337	6154602	Two novel I	A2a	lung cancer		
27903896	5210574	In order to	ABC	cancer		up
23708134	3719553	Another im	Apr	twitching motilities		up
29113654	5726560	Emerging d	CD274	bacterial dysbiosis		
29113654	5726560	Emerging d	CD274	cancer		
29113654	5726560	Emerging d	CD274	carcinogenesis		
29113654	5726560	Emerging d	CD274	colorectal cancers		
29113654	5726560	Emerging d	CD274	gastrointestinal tract cancers		
29113654	5726560	Emerging d	CD274	gastrointestinal tract neoplasms		
29113654	5726560	Emerging d	CD274	inflammation		
29113654	5726560	Emerging d	CD274	tumor		
29113654	5726560	Emerging d	CTLA4	bacterial dysbiosis		
29113654	5726560	Emerging d	CTLA4	cancer		
29113654	5726560	Emerging d	CTLA4	carcinogenesis		
29113654	5726560	Emerging d	CTLA4	colorectal cancers		
29113654	5726560	Emerging d	CTLA4	gastrointestinal tract cancers		
29113654	5726560	Emerging d	CTLA4	gastrointestinal tract neoplasms		

Analysis:

1. Data Structure
   The dataset includes the following columns:
   - PMID: PubMed ID of the source article.
   - PMCID: PubMed Central ID of the source article.
   - Sentence: A sentence from the article detailing the interaction.
   - Genes: Genes involved in the interaction.
   - Diseases: Diseases involved in the interaction.
   - Interaction type: The type of interaction (e.g., agonist, antagonist).
   - Regulation: Regulation status (e.g., up-regulated, down-regulated)
2. Top 10 Genes:
   - tyrosine hydroxylase
   - CD274
   - CTLA4
   - alpha-synuclein
   - TH
   - caspase 3
   - aurora kinase A
   - PAO1

- MCT1
- A2a

3. Top 10 Diseases:
    - infection 9
    - toxicity
    - cancer
    - neurodegenerative diseases
    - tumor
    - twitching motilities
    - death
    - neuroblastoma
    - neurodegeneration
    - neurotoxic

Task10_Chemical_to_Chemical.csv

PMID	PMCID	Sentence	Chemical1	Chemical2	Interaction type	Regulation
29156609	5713430	nan	(S)-norcoclauri	(S)-norcoclaurir	Inhibition	up
29156609	5713430	nan	(S)-norcoclauri	benzylisoquinol	Inhibition	up
21610729	3112539	nan	(S)-reticuline	(S)-reticuline	Inhibition	up
21610729	3112539	nan	(S)-reticuline	alkaloids	Inhibition	
21610729	3112539	nan	(S)-reticuline	carbon	Inhibition	
21610729	3112539	nan	(S)-reticuline	flavonoids	Inhibition	
21610729	3112539	nan	(S)-reticuline	isoprenoids	Inhibition	up
15141066	429332	Our studi	1,8-cineole	1,8-cineole		
15141066	429332	Our studi	1,8-cineole	2,3,-butanedione		
15141066	429332	Our studi	1,8-cineole	2-butanone		
15141066	429332	Our studi	1,8-cineole	acetaldehyde		
15141066	429332	Our studi	1,8-cineole	acetic acid		up
15141066	429332	Our studi	1,8-cineole	acetone		
15141066	429332	Our studi	1,8-cineole	carbon		
15141066	429332	Our studi	1,8-cineole	ethanol		up
15141066	429332	Our studi	1,8-cineole	ethyl acetate		
15141066	429332	Our studi	1,8-cineole	monoterpene		
19842066	3128704	nan	1-acetyl-beta-c	1-acetyl-beta-ca	Inhibition	up
19842066	3128704	nan	1-acetyl-beta-c	3-(hydroxyacety	Inhibition	
19842066	3128704	nan	1-acetyl-beta-c	N-butyl-benzen	Inhibition	

1. Data Structure

   The dataset comprises the following columns:

   - PMID**:** PubMed Identifier - unique identifier for the article in the PubMed database.
   - PMCID**:** PubMed Central Identifier - unique identifier for the article in the PubMed Central database.
   - Sentence**:** Sentence from the article describing the interaction.
   - Chemicals1**:** The first chemical involved in the interaction.
   - Chemicals2**:** The second chemical involved in the interaction.
   - Interaction type**:** Describes the type of interaction (e.g., Agonist).
   - Regulation**:** Describes the regulation direction (e.g., Up or Down).

2. Data Characteristics
   - **The Sentence column** contains detailed descriptions of chemical interactions within the context of disease.
   - **Chemicals1 and Chemicals2 columns** list the chemicals involved in each interaction.
   - **Interaction type column** specifies the nature of the interaction between the chemicals.
   - **Regulation column** indicates whether the interaction results in upregulation or downregulation of chemical activity.

Task10_Gene_to_Gene.csv

PMID	PMCID	Sentence	Gene1	Gene2	Interaction type	Regulation
28358337	6154602	Two nove	A2a	A2a		
27903896	5210574	In order t	ABC	ABC		up
27903896	5210574	In order t	ABC	abc		up
23708134	3719553	Another i	Apr	Apr		up
23708134	3719553	Another i	Apr	Gac		up
23708134	3719553	Another i	Apr	Has		up
23708134	3719553	Another i	Apr	hcn		up
19854857	2785968	After two	Atpdr2	Atpdr2		
19854857	2785968	After two	Atpdr2	abcg30		
21429787	3947797	Today, th	B17	B17		up
29113654	5726560	Emerging	CD274	CD274		
29113654	5726560	Emerging	CD274	CTLA4		
29113654	5726560	Emerging	CTLA4	CD274		
29113654	5726560	Emerging	CTLA4	CTLA4		
23232715	3520106	Cyclic AMI	Crp	Crp		up
23232715	3520106	Cyclic AMI	Crp	crp		up
20599778	2934894	Estrogen	ER	ER		up
20599778	2934894	Estrogen	ER	ERalpha/alpha		up
20599778	2934894	Estrogen	ER	ERalpha/beta		up
20599778	2934894	Estrogen	ER	ERbeta/beta		up

1. Data Structure
    - PMID**:** PubMed Identifier - unique identifier for the article in the PubMed database.
    - PMCID**:** PubMed Central Identifier - unique identifier for the article in the PubMed Central database.
    - Sentence**:** Sentence from the article describing the gene interaction.
    - Genes1**:** The first gene involved in the interaction.
    - Genes2**:** The second gene involved in the interaction.
    - Interaction type**:** Describes the type of interaction between the genes (e.g., Activation).
    - Regulation**:** Describes the regulation direction (e.g., Up-regulated, Down-regulated).

2. Data Characteristics
    - The Sentence column contains detailed descriptions of gene interactions within the context of disease.
    - Genes1 and Genes2 columns list the genes involved in each interaction.
    - Interaction type column specifies the nature of the interaction between the genes.
    - Regulation column indicates whether the interaction results in upregulation or downregulation of gene expression.

11. Pathway Extraction Process

The pathway extraction process included several steps to collect and analyze data from scientific literature and the KEGG database:

- Obtained a list of pathways from KEGG and saved it as "kegg_pathways.txt."
- Ran a Jupyter Notebook script for text mining to extract pathway details from full-text papers, especially those without PMC IDs, by matching pathway names in "kegg_pathways.txt" with those mentioned in paper abstracts.
- Compiled the findings into a summary data frame that lists PMIDs and their associated pathway matches.
- Generated the final output in "Final_Pathways.csv," which consolidates the identified pathways and their corresponding PMIDs.
- This method integrated web scraping with text mining techniques to effectively extract and organize pathway data from scientific literature, improving its accessibility and usability for subsequent research and analysis.

Task11_final_pathways.csv

pathways	PMID	length
breast cancer	28383816.0,29248948.0,25742265.0,28369962.0,215	82
microbial metabolism in diverse environments	29032469.0,27907843.0,35710962.0,36275752.0,343	5
steroid hormone biosynthesis	34650578	1
rna degradation	12488561.0,28334892.0,36519128.0	3
bacterial secretion system	35687206.0,34157412.0,37907566.0,29615997.0,345	5
parkinson disease	19806188.0,34234177.0	2
tight junction	28571979.0,18444159.0,33125671.0,35524736.0,344	11
toxoplasmosis	26410452	1
nitrogen metabolism	22147733.0,28293039.0,22033567.0,21342462.0,290	34
prodigiosin biosynthesis	26538254.0,3552718.0,28838222.0,2999302.0,25760	11
base excision repair	26579709.0,33618627.0	2
ampk signaling pathway	29113354.0,32683035.0	2
terpenoid backbone biosynthesis	36813111.0,38249018.0	2
biotin metabolism	32530501	1
apoptosis	28383816.0,21326924.0,27393306.0,16988123.0,286	152
pertussis	16138079	1
cysteine and methionine metabolism	35738442.0,34901103.0	2
streptomycin biosynthesis	12165483.0,18375553.0	2
nucleotide excision repair	17178464	1
prostate cancer	28592323.0,25385358.0,24879544.0,27990568.0,313	23

12. Verification of Secondary Metabolite-Related Sentences Using BIOBERT

The verification of secondary metabolite-related sentences using BIOBERT followed a systematic approach:

- CSV File Specification: Utilized CSV files from Task 4, containing sentences related to secondary metabolites of bacteria for verification.
- Model Execution and Setup: Applied Python code to manage the verification process, including downloading and configuring the BIOBERT model and tokenizer for sentence analysis.
- Verification Process: Each sentence was assessed using BIOBERT with a predefined threshold. Sentences scoring above the threshold were marked as 1, indicating relevance to secondary metabolites, while those scoring below were marked as 0.
- Output Handling: The results of the verification were recorded in an output CSV file, listing the original sentences along with their verification flags (1 for relevant, 0 for non-relevant).

- This method utilizes advanced natural language processing to accurately identify and classify sentences related to secondary metabolites of bacteria, ensuring precise scientific analysis and interpretation.

Task12_Biobert_Genes_Bacteria_results

Sentences	Probability	Verify	Genes	PMID	PMCID
""" Two novel benzamido nonacyclic dilactones, namel	0.565133452	TRUE	A2a	28358337	6154602
""" In order to advance the exploration of microbial se	0.564765692	TRUE	ABC	27903896	5210574
""" Another important category of the GacA regulon w	0.538655221	TRUE	Apr	23708134	3719553
""" After two generations, only the Arabidopsis abcg30	0.5570274	TRUE	Atpdr2	19854857	2785968
""" Today, this burgeoning class of natural products en	0.553393781	TRUE	B17	21429787	3947797
""" Emerging data demonstrate that intestinal bacteria	0.538095057	TRUE	CD274	29113654	5726560
""" Emerging data demonstrate that intestinal bacteria	0.538095057	TRUE	CTLA4	29113654	5726560
"""Cyclic AMP receptor protein (Crp) is a transcription	0.563564658	TRUE	Crp	23232715	3520106
""" Estrogen Receptor (ER) dimerization is required for	0.582680345	TRUE	ER	20599778	2934894
"""nan"""	0.55141592	TRUE	ERalpha/alpha	20599778	2934894
"""nan"""	0.55141592	TRUE	ERalpha/beta	20599778	2934894
"""nan"""	0.55141592	TRUE	ERbeta/beta	20599778	2934894
"""We identified a transcription regulator Esi for the	0.56487304	TRUE	Esi	28535774	5442856
""" Estrogen Receptor (ER) dimerization is required for	0.576748013	TRUE	Estrogen Receptor	20599778	2934894
""" The antibiotic biosynthesis of M18 is coordinately c	0.554214418	TRUE	Gac	23708134	3719553
""" Further characterization of several other berberine	0.552198112	TRUE	GcvT	28784988	5547118
""" Another important category of the GacA regulon w	0.541408896	TRUE	Has	23708134	3719553
""" Subsequent sections describe the structure, functio	0.579886198	TRUE	IF1	15755955	1082788
""" Subsequent sections describe the structure, functio	0.587487876	TRUE	IF2	15755955	1082788
""" Monocarboxylate transporter 1 (MCT1) plays a maj	0.527505875	TRUE	MCT1	26272259	4593819

13. Summary of Sentences

Utilization of BERT Summarizer for Text Summarization:

- Implemented the BERT Summarizer, a specialized model based on BERT (Bidirectional Encoder Representations from Transformers), designed to generate abstractive summaries of text.
- Investigated its capabilities in summarizing input text about secondary metabolites of bacteria while ensuring the summaries were semantically relevant and coherent.
- Execution of Text Summarization Tasks: Applied the BERT Summarizer to various text inputs related to secondary metabolites of bacteria to test its effectiveness in producing concise and informative summaries.
- Evaluated the model's performance in accurately capturing the essence and key points of the original text through its abstractive summarization method.

Task13_data_with_summary.csv

Genes	GenesID	PMID	PMCID	Sentences	Summary
A2a	28882	28358337	6154602	Two novel b	Two novel
ABC	10058	27903896	5210574	In order to advance the exploration of microbial secondary metabolism, we d	
Apr	5366	23708134	3719553	Another imp	GacA
Atpdr2	827187	19854857	2785968	After two ge	The
B17	4712	21429787	3947797	Today, this b	The
CD274	29126	29113654	5726560	Emerging da	We use
CTLA4	1493	29113654	5726560	Emerging da	We use
Crp	20468888	23232715	3520106	Cyclic AMP receptor protein (Crp) is a transcription regulator controlling divers	
ER	2099	20599778	2934894	Estrogen Receptor (ER) dimerization is required for target gene transcription.;	
ERalpha/alpha	2099	20599778	2934894		CNN.com
ERalpha/beta	2099	20599778	2934894		CNN.com
ERbeta/beta	2099	20599778	2934894		CNN.com
Esi	5266	28535774	5442856	We	
Estrogen Recep	2099	20599778	2934894	Estrogen Rec	Estrogen
Gac	2744	23708134	3719553	The antibiotic biosynthesis of M18 is coordinately controlled by multiple distir	
GcvT	275	28784988	5547118	Further char	The
Has	3036	23708134	3719553	Another imp	Many
IF1	93974	15755955	1082788	Subsequent sections describe the structure, function, and interactions of the r	
IF2	9669	15755955	1082788	Subsequent sections describe the structure, function, and interactions of the r	
MCT1	6566	26272259	4593819	Monocarboxylate transporter 1 (MCT1) plays a major role in colonic luminal b	

14. Protein Information

Retrieval and Processing of UniProt Data:

- The initial phase involves accessing and processing protein data from UniProt using their REST API. The key steps are:
    - API Query Construction: Constructing URLs to query UniProt's API, specifying organism-specific parameters and required data fields.
    - Data Download: Downloading compressed TSV files containing detailed protein information based on the constructed queries.
    - Data Extraction and Compilation: Parsing and compiling the extracted data from TSV files into a consolidated CSV format, suitable for further analysis and integration with other data sources.
- Error Handling: Errors encountered during the retrieval process are documented and resolved to uphold data integrity and ensure the pipeline's reliability.

Task14_Protein_Information.csv

Entry	Entry Nam	Gene Nam	GeneID	Length	PubMed II	Protein na	Organism	Function [CC]
A0A0D0X0	A0A0D0X0	nnrD nnrE	TK50_192	496		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A120F7	A0A120F7	nnrD nnrE	GA007062	489		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A1A9A5	A0A1A9A5	nnrE nnrD	GA007062	489		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A1C3NA	A0A1C3NA	nnrE nnrD	GA007062	489		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A1C4U7	A0A1C4U7	nnrE nnrD	GA007021	489		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A1C4UA	A0A1C4UA	bioA pyrD	GA007060	850		Multifunct	Micromon	FUNCTION: Catalyzes the conversion of dihydroorota
A0A1C4UI	A0A1C4UI	nnrD nnrE	GA007056	489		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A1C4UL	A0A1C4UL	nnrE nnrD	GA007055	496		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A1C4V3	A0A1C4V3	nnrE nnrD	GA007469	489		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A1C4VL	A0A1C4VL	nnrD nnrE	GA007469	490		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A1C4W	A0A1C4W	nnrD nnrE	GA007061	489		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A1C4W	A0A1C4W	nnrE nnrD	GA007060	489		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A1C4W	A0A1C4W	nnrD nnrE	GA007021	489		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A1C4XB	A0A1C4XB	nnrD nnrE	GA007061	489		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A1C4Y4	A0A1C4Y4	nnrD nnrE	GA007056	493		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A1C4YC	A0A1C4YC	nnrD nnrE	GA007056	489		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A1C5G2	A0A1C5G2	nnrD nnrE	GA007061	489		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A1C5GL	A0A1C5GL	nnrD nnrE	GA007056	490		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A1C5IC	A0A1C5IC	nnrD nnrE	GA007061	489		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e
A0A1C5J2	A0A1C5J2	nnrD nnrE	GA007061	489		Bifunction	Micromon	FUNCTION: Bifunctional enzyme that catalyzes the e

15. Knowledge Graph

Introduction

A knowledge graph is a structured representation of information that illustrates entities and the relationships between them. In the context of biological research, a knowledge graph can be a powerful tool to visualize and understand complex interactions between different biological entities such as chemicals, genes, and their interactions. This report focuses on the visualization of interactions between chemicals, genes, and chemical-gene interactions.

Objectives

1. To create a comprehensive knowledge graph that visualizes interactions between:
    - Chemical to Chemical
    - Gene to Gene
    - Gene to Chemical
2. To analyze the significance of these interactions and their implications in biological research.

Components of the Knowledge Graph

1. Chemical to Chemical Interactions

Definition**:** These interactions represent the relationships between different chemicals. Understanding these interactions is crucial for drug discovery, chemical synthesis, and studying biochemical pathways.

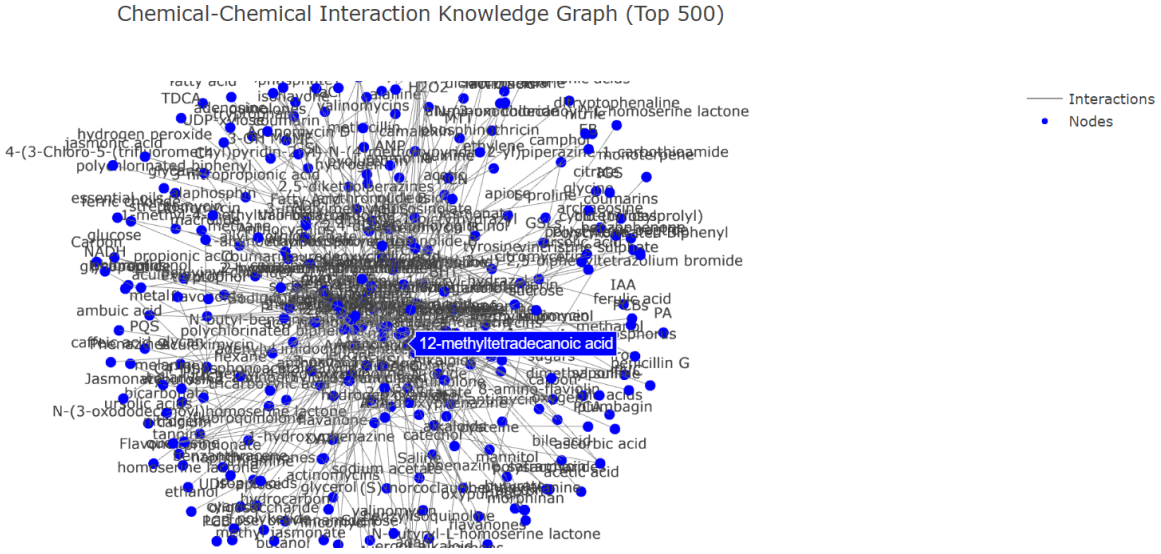
Visualization**:**

- Nodes: Represent individual chemicals.
- Edges: Represent interactions between chemicals, such as reactions, inhibition, or enhancement.

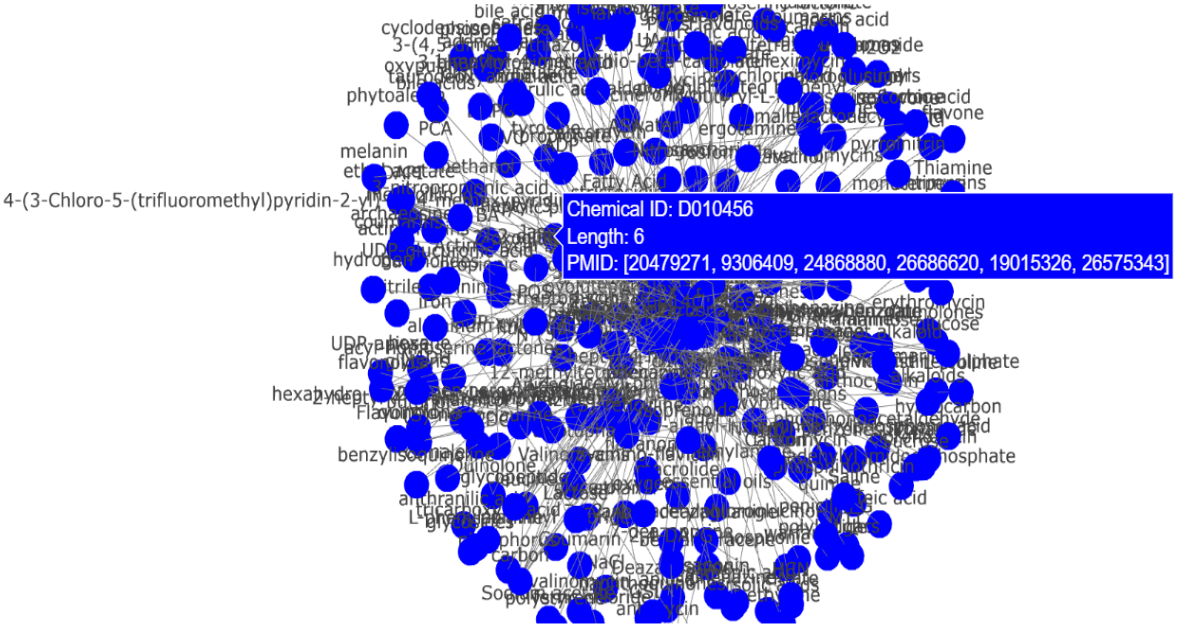
Analysis**:**

- Highlight common interaction patterns.
- Identify key chemicals that have numerous interactions, suggesting their significance in biological pathways.

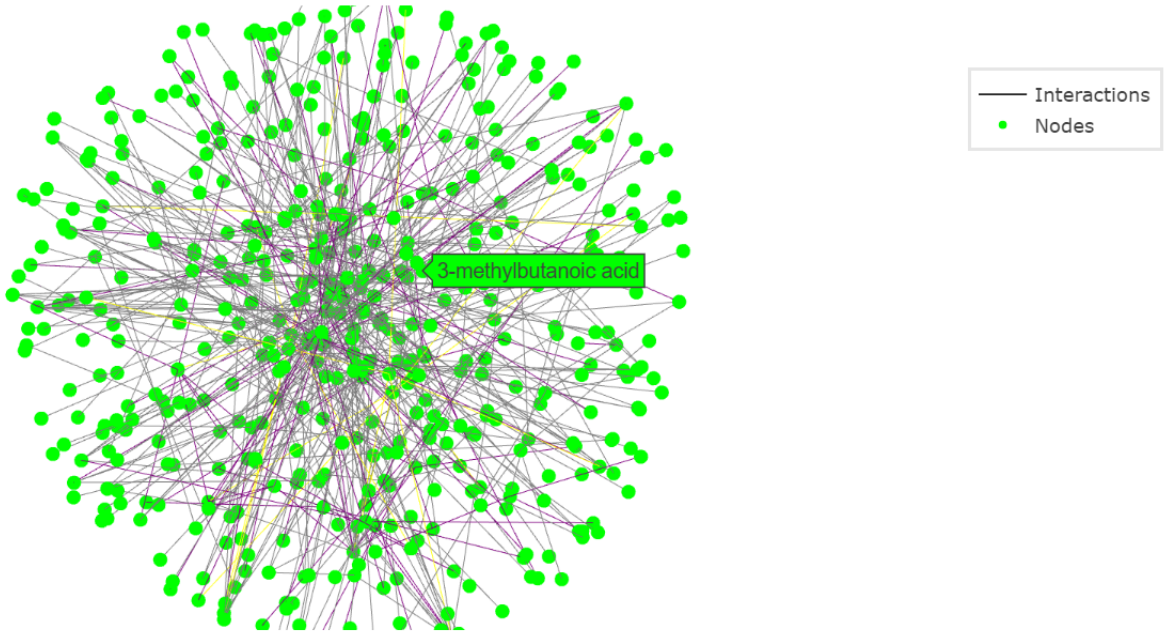
Result Preview:



Chemical-Chemical Interaction Knowledge Graph with Information (Top 500)



Chemicla-Chemical Interaction Knowledge Graph(Top 200)



2. Gene to Gene Interactions

   Definition:
   These interactions show the relationships between different genes. This is essential for understanding gene regulation, genetic pathways, and the functional genomics.

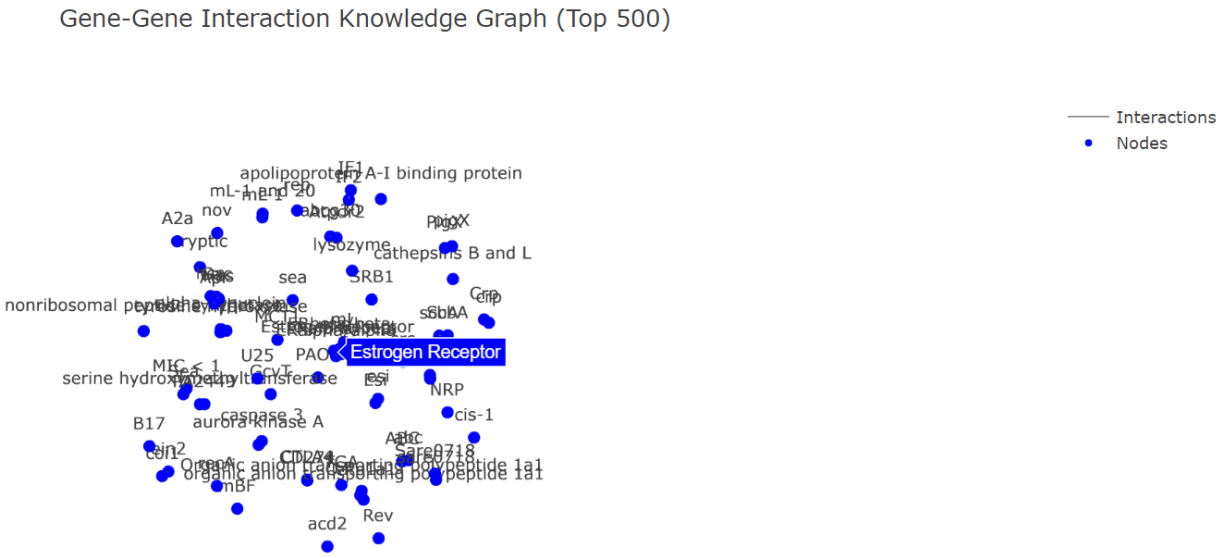
   Visualization:

   - Nodes: Represent individual genes.
   - Edges: Represent interactions between genes, such as co-expression, regulation, or genetic linkage.

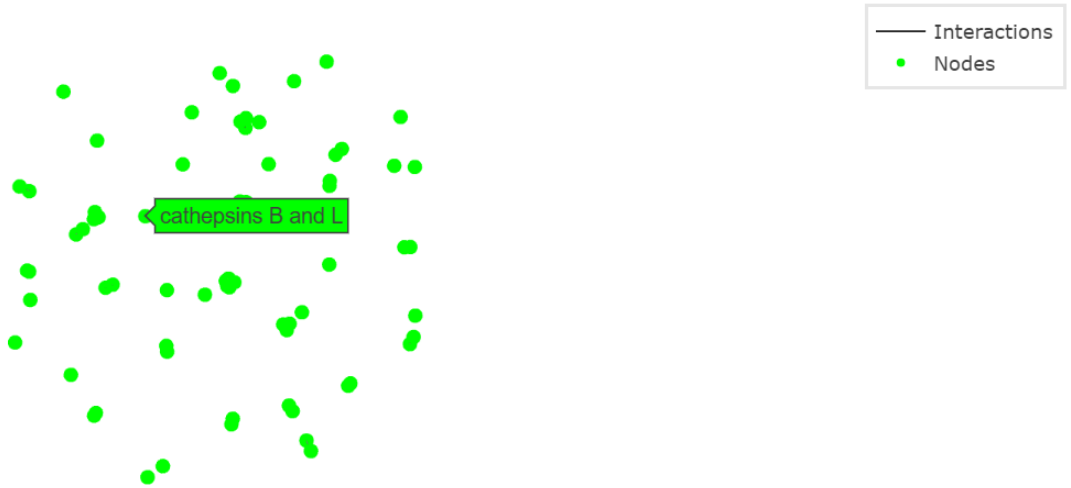
   Analysis:

   - Identify clusters of genes that frequently interact, which may indicate shared pathways or functions.
   - Detect hub genes with high connectivity, which could be critical for maintaining cellular functions.

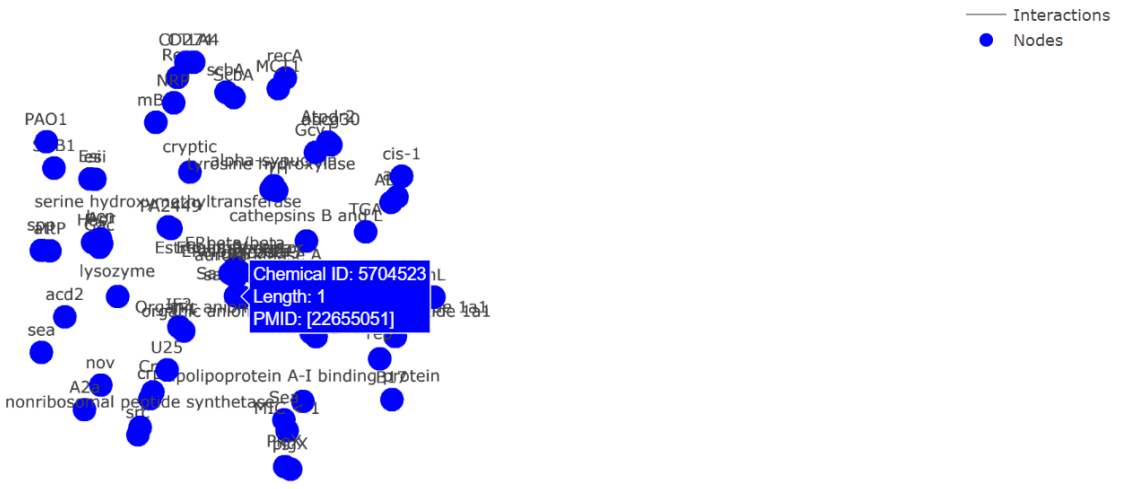
   Output Preview:



Gene-Gene Interaction Knowledge Graph(Top 200)



Gene-Gene Interaction Knowledge Graph_with_information (Top 500)



3. Gene to Chemical Interactions

   Definition:
   These interactions depict the relationships between genes and chemicals. This is significant for understanding drug mechanisms, gene expression modulation, and metabolic processes.

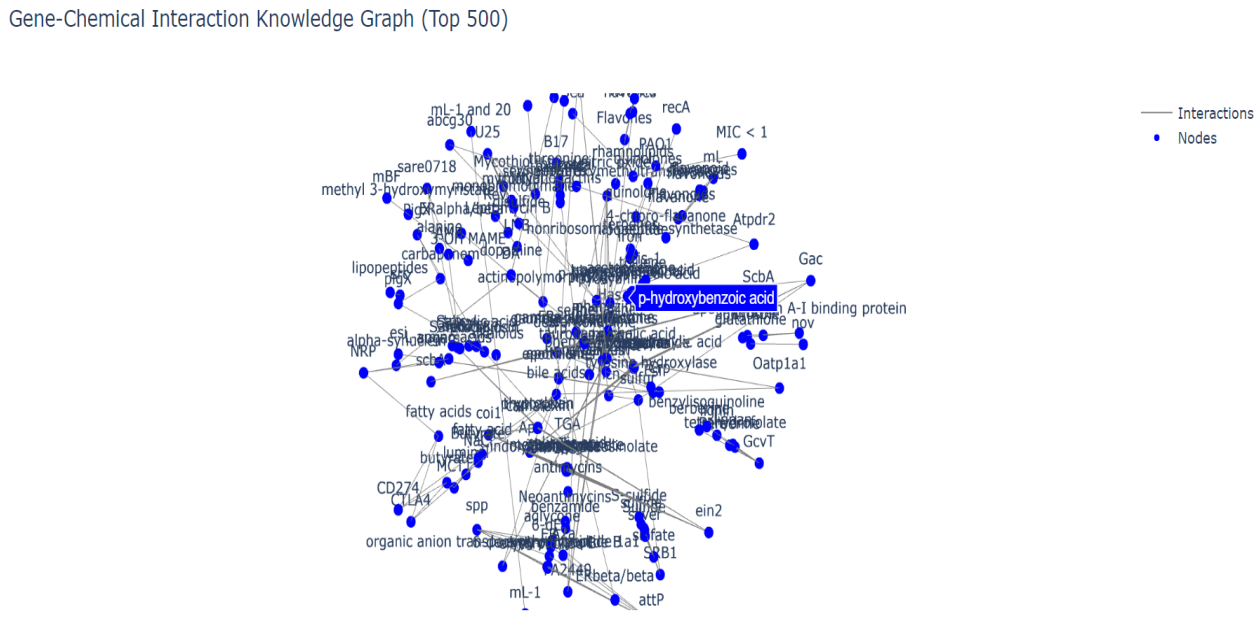
   Visualization:

   - Nodes: Represent both genes and chemicals.
   - Edges: Represent interactions between genes and chemicals, such as binding, activation, or inhibition.

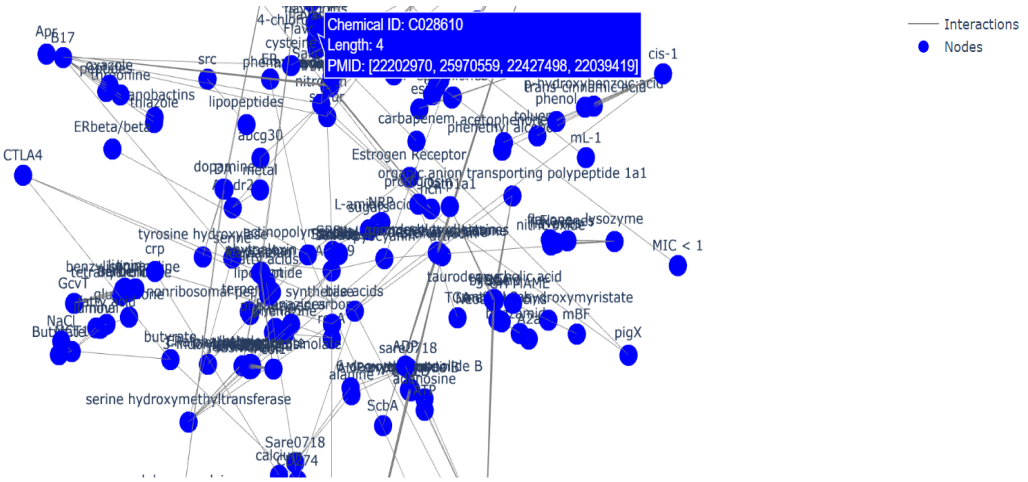
   Analysis:

   - Examine how chemicals affect gene expression and vice versa.
   - Discover potential therapeutic targets by identifying chemicals that interact with disease-associated genes.

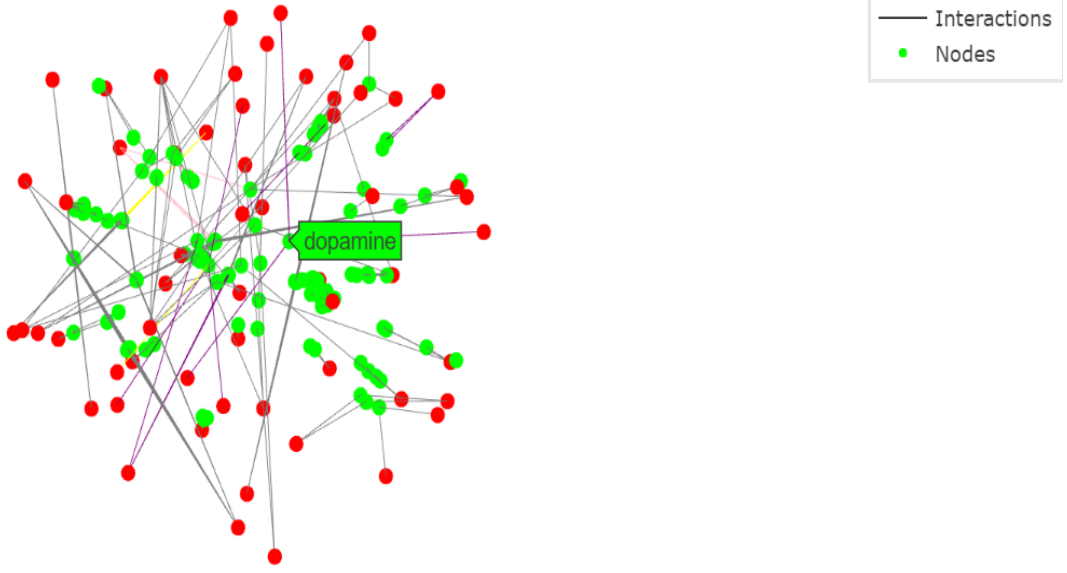
Result Preview



Gene-Chemical Interaction Knowledge Graph_with_information (Top 500)



Chemical-Gene Interaction Knowledge Graph (Top 1160)



Methodology

Data Collection:

- Sources include biological databases and literature that provide detailed information on chemical and gene interactions.

Data Processing:

- Data is curated, cleaned, and formatted to ensure consistency and accuracy.
- The top 500 entries are selected for clarity in visualization.

Visualization Tools:

- Plotly is used for interactive and dynamic visualizations, allowing for detailed exploration of the knowledge graph.

Results and Discussion

The knowledge graph successfully visualizes the intricate network of interactions between chemicals, genes, and their cross-interactions. Key findings include:

- Identification of central chemicals and genes with high interaction counts.
- Discovery of interaction clusters indicating potential pathways and functional modules.
- Visualization of chemical-gene interactions providing insights into drug mechanisms and gene regulation.

Conclusion

The knowledge graph offers a powerful visualization tool to understand complex biological interactions. By representing chemical to chemical, gene to gene, and gene to chemical interactions, researchers can gain valuable insights into biochemical pathways, gene regulation, and potential therapeutic targets. This comprehensive analysis aids in advancing our knowledge in biological research and drug discovery.

Future Work

- Expanding the dataset to include more interactions and entities.
- Incorporating temporal data to observe dynamic changes in interactions.
- Enhancing the interactivity of the visualization for better user engagement and exploration.

16. Graph - Interaction

The "Graph - Interaction" section presents the visual representations of interactions within a dataset. These interactions are depicted through various types of graphs, each focusing on different relationships: Gene to Gene, Gene to Chemical, and Chemical to Chemical. Each graph provides insights into how entities within these categories are interconnected.

1. Gene to Chemical Interaction

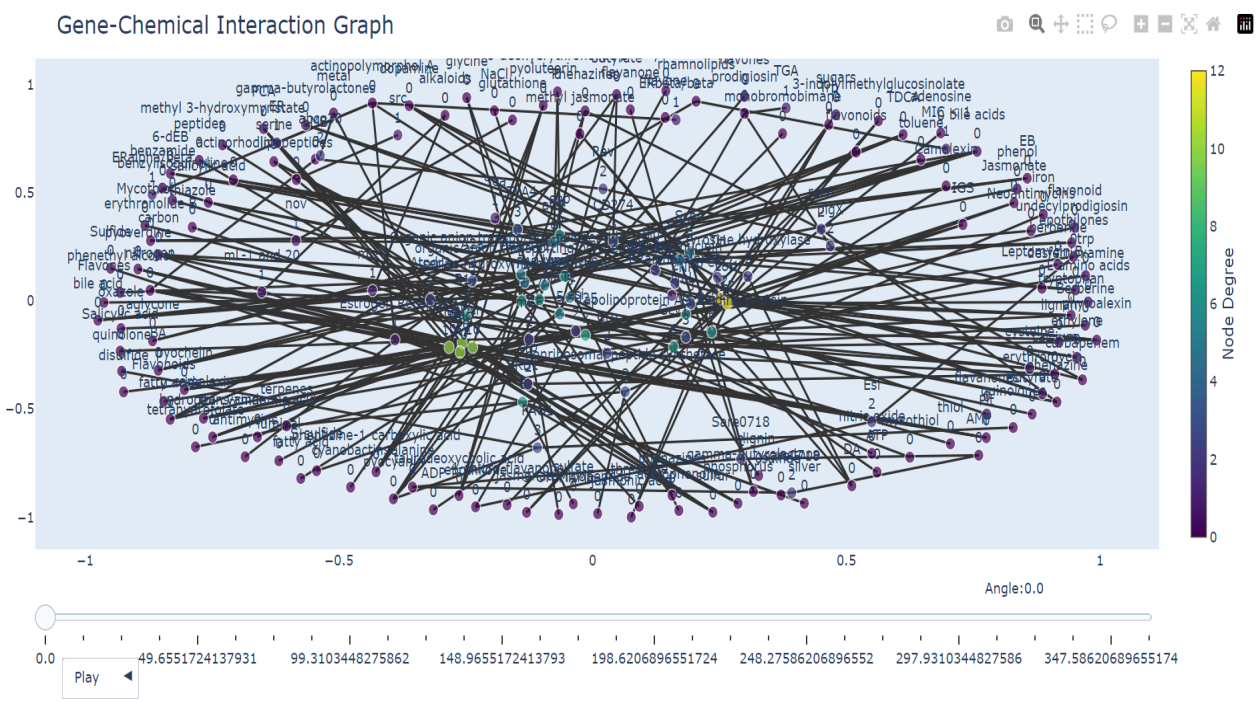
Purpose: The Gene to Chemical interaction graph displays how genes interact with chemicals. This visualization is crucial for studying how chemical compounds affect gene activity and vice versa, which is particularly relevant in fields like pharmacology and biotechnology.

Details:

- Nodes: Include both genes and chemicals, reflecting their roles in biological processes and chemical reactions.
- Edges: Show interactions between genes and chemicals, with details on the nature of these interactions and any regulatory mechanisms involved.
- Visualization: The layout and color scheme follow a similar approach to the Gene to Gene graph. Nodes are placed to highlight their relationships, while edges are differentiated by color based on interaction type. This setup helps in identifying key interactions and understanding how different chemicals influence gene functions.

Output:

The following figure illustrates the interactions between genes and chemicals, showing how different chemicals are associated with various genes.



2. Gene to Gene Interaction

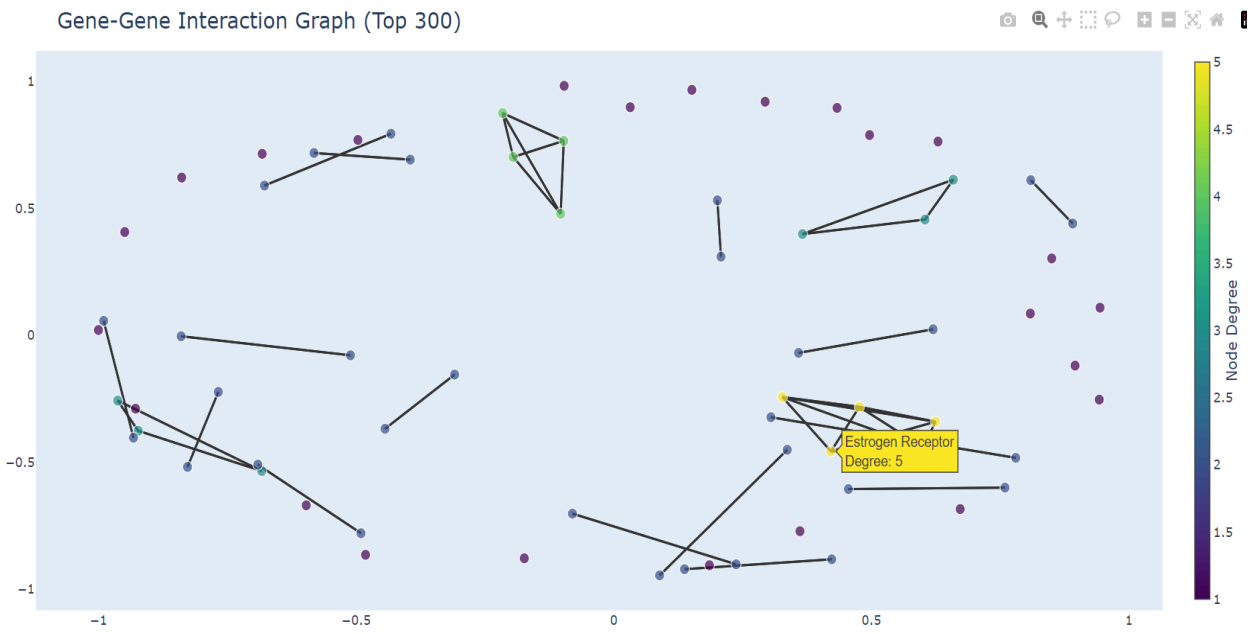
Purpose**:** The Gene to Gene interaction graph visualizes the relationships and interactions between different genes. This type of graph is instrumental in understanding gene networks and the regulatory or functional relationships that exist among genes.

Details:

- Nodes: Represent individual genes.
- Edges: Indicate interactions between genes, with attributes detailing the type of interaction (e.g., positive, negative) and any associated regulations.
- Visualization: Nodes are positioned using a spring layout, which helps in distributing them in a manner that minimizes overlap and clearly represents connections. The edges are colored based on interaction types, providing an immediate visual cue about the nature of each interaction. Node sizes are uniform but can be color-coded based on their degree (number of direct connections) to show their importance or centrality within the network.

Output:

The following figure provides a visual representation of Gene to Gene interactions, highlighting the relationships between different genes.



3. Chemical to Chemical Interaction

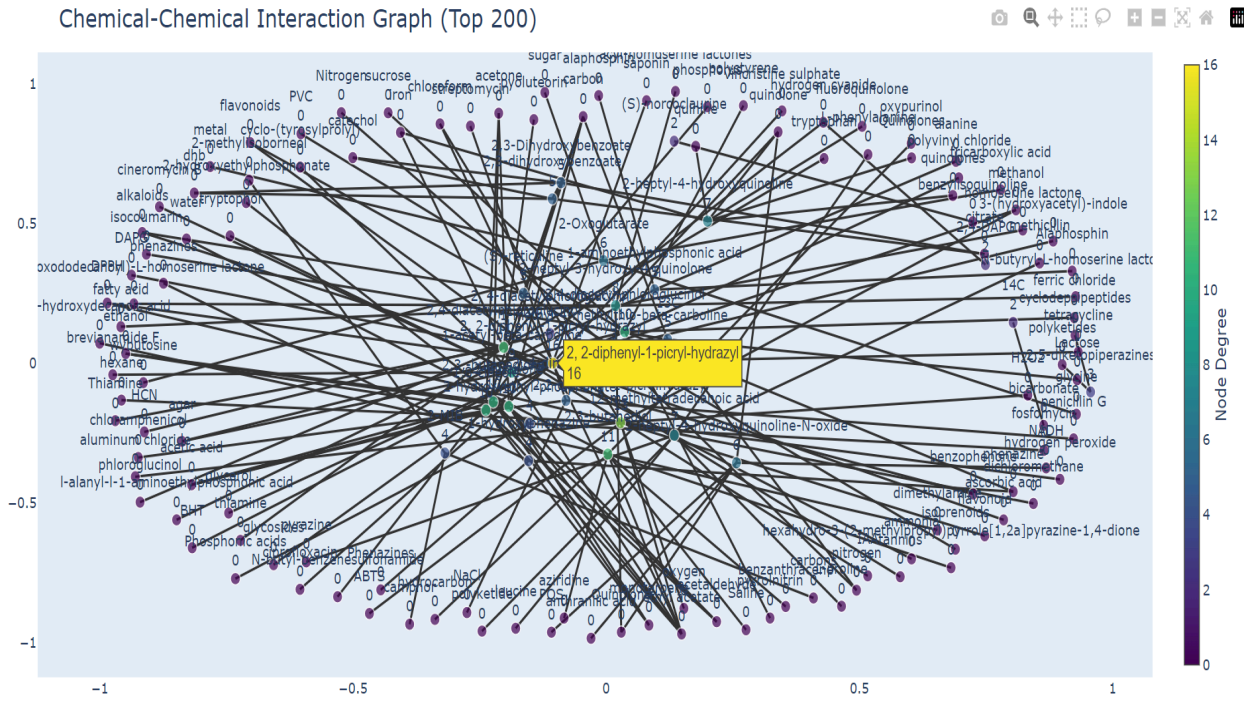
Purpose: The Chemical to Chemical interaction graph illustrates the relationships between different chemicals. This type of graph is useful for exploring how chemicals interact with each other, which can be essential for understanding complex chemical processes or interactions in various scientific contexts.

Details:

- Nodes: Represent chemicals involved in the dataset.
- Edges: Connect pairs of chemicals based on their interactions, with attributes describing the nature and impact of these interactions.
- Visualization: The graph uses a spring layout to ensure a clear and readable representation of chemical interactions. Colors for edges are used to denote different types of interactions, providing a visual summary of how chemicals interact and potentially influence each other. Node sizes and colors might be used to represent the degree of interaction or other relevant metrics.

Output:

The following figure provides a visual representation of Chemical to Chemical interactions, highlighting the relationships between different chemicals.



17. Biological Knowledge Graph Visualization

Biological Knowledge Graph Visualization

Objective

The visualization aims to present a biological knowledge graph that highlights the interactions and associations among different biological entities—genes, chemicals, diseases, and species. This graph helps to understand the complex relationships in biological research.

Data Source

The data for this visualization was sourced from a CSV file titled Task1_pubmed_secondary_metabolites_Bacteria.csv. This dataset includes:

- Genes: Biological sequences that code for proteins.
- Chemicals: Substances that interact with genes.
- Diseases: Conditions associated with genes.

- Species: Organisms that express genes.

Graph Construction

1. Initialization:
    - A directed graph G was created using NetworkX. In a directed graph, edges have a direction, representing the flow of relationships from one node to another.
2. Adding Nodes:
    - Nodes were added to the graph for each unique entry in the columns representing genes, chemicals, diseases, and species. Each node was tagged with a type to distinguish between these entities.
3. Adding Edges:
    - Gene-Disease Associations: Edges were added to connect genes to diseases, indicating that the gene is associated with the disease.
    - Chemical-Gene Interactions: Edges were added to link chemicals to genes, representing chemical interactions or effects on genes.
    - Species-Gene Expression: Edges were added to show which species express which genes, indicating biological expression data.

Subgraph Selection

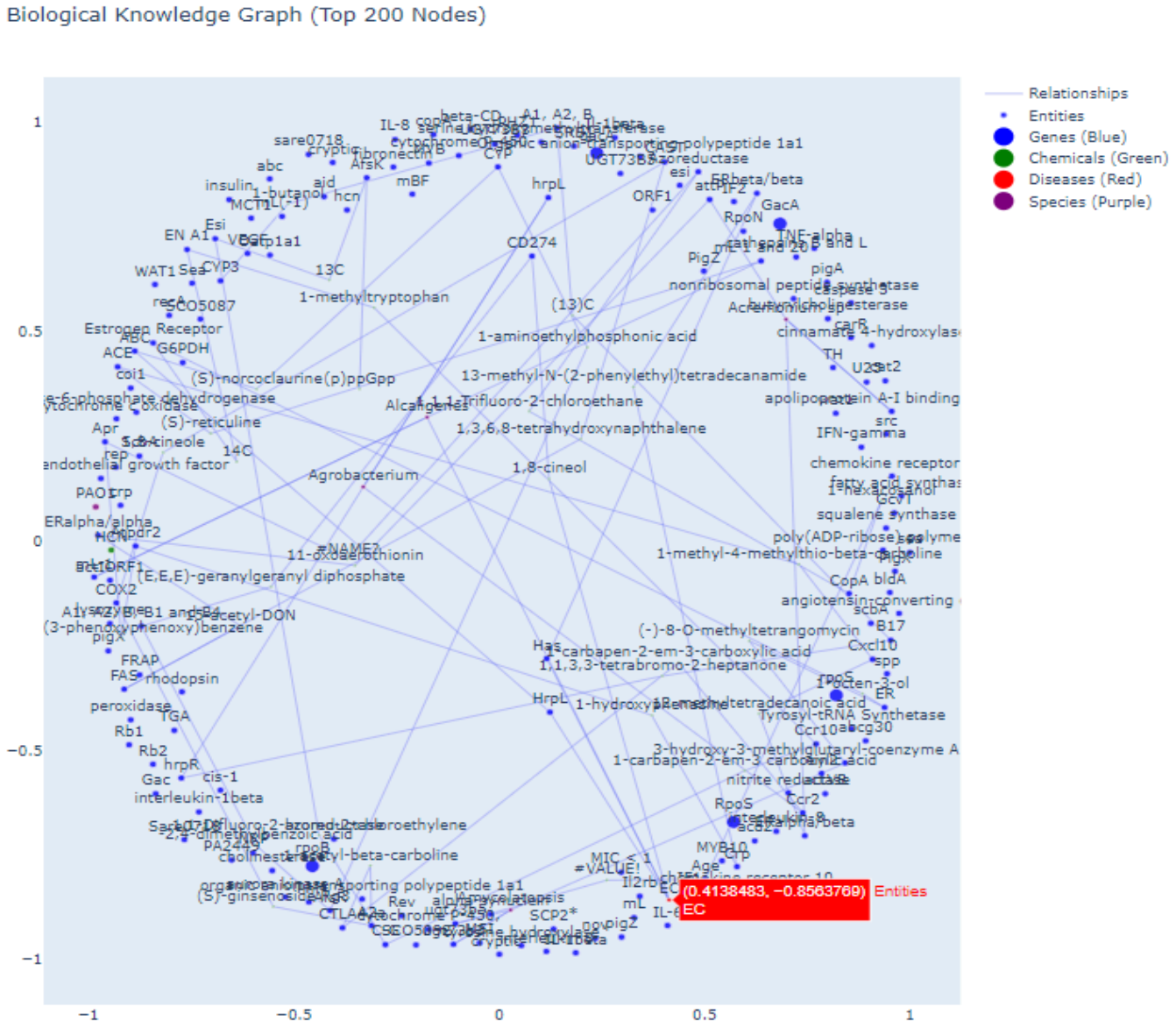
- Degree Centrality:
    - To focus on the most influential nodes, the graph was reduced to the top 200 nodes based on their degree centrality. Degree centrality measures the number of direct connections a node has, highlighting nodes that play a significant role in the network.
    - The top 200 nodes were selected for clarity and to manage the complexity of the graph.

Layout and Visualization

1. Layout Computation:
    - The spring layout algorithm was used to position nodes. This layout simulates a physical system where nodes repel each other and edges act as springs. It helps to evenly distribute nodes and minimize overlap, making the graph easier to interpret.
2. Plotly Visualization:
    - Edges:
        - Plotted as lines connecting nodes.
        - The color of edges indicates their type:
            - Associations: Light red.
            - Interactions: Light green.
            - Expressions: Light blue.

    - Nodes:
        - Represented as points with colors indicating their type:
            - Genes: Blue.
            - Chemicals: Green.
            - Diseases: Red.
            - Species: Purple.
        - Node sizes are proportional to their degree centrality, making more connected nodes appear larger and thus more prominent.
    - Legend:
        - A legend was included to explain the color-coding scheme for both nodes and edges. This helps viewers understand what each color represents.
3. Plot Customization:
    - The figure size was set to 800x800 pixels to ensure the graph fits on a single page for easy viewing.
    - Node and edge colors were chosen to be distinct and easily recognizable, aiding in visual differentiation of various types of entities and relationships.
4. Output:
    - The final visualization was saved as an HTML file (biological_knowledge_graph_200_resized.html). This file can be opened in any web browser, allowing interactive exploration of the graph.

The following figure provides a visual representation of the Biological Knowledge Graph.



18. Database download and verify

To research microbes that thrive in extreme conditions, several notable databases can be explored. Here are a few key ones**:**

Microbial Dark Matter (MDM) Project

- User-Friendly Interface: This platform features an intuitive and easily navigable design, allowing users to effortlessly access various sections focused on bacterial secondary metabolites.
- Distributed Information: Data is not centralized; it is divided across multiple categories, including genomic data, research papers, and environmental metadata. This segmentation enables detailed exploration but requires users to navigate through different sections to gather comprehensive information.

Extreme Microbiome Project (XMP)

- User-Friendly Access: XMP offers a straightforward layout, providing tools and resources that facilitate the easy retrieval of specific data on bacteria producing secondary metabolites in extreme environments.
- Segmented Data: The database categorizes information into distinct projects and publications, each dedicated to different extremophiles or environments. Users may need to explore multiple areas to find all relevant data.

GOLD: Genomes Online Database

- Ease of Use: GOLD offers a user-friendly search interface, simplifying the process of locating genome projects. Users can filter searches by organism type, habitat, or project status.
- Diverse Data Repositories: Information is distributed across various projects and studies, rather than being consolidated into a single dataset. Users need to compile information from different entries to get a full picture.

HaloWeb

- Straightforward Navigation: HaloWeb features a clear and organized interface, designed to facilitate easy access to data on halophilic bacteria and their secondary metabolites.
- Independent Data Sections: Information is categorized by species, genetics, and biochemical properties. Comprehensive data compilation on a single bacterium requires visiting multiple sections.

Thermozymes Database

- User-Friendly Design: This database prioritizes accessibility, enabling users to easily search for and retrieve data on enzymes produced by thermophilic bacteria, including those related to secondary metabolites.
- Scattered Information: Data on enzyme characteristics, functions, and related research is not consolidated in one place. Users are encouraged to explore the platform thoroughly to gain a complete understanding.

19. Conclusion

This project has culminated in a comprehensive database on bacterial secondary metabolites, offering a valuable resource for the scientific community. The extensive data collection and analysis underscore the immense potential of bacterial metabolites in biotechnology and medicine. Key findings include the identification of diverse secondary metabolites with significant biological activities, such as antimicrobial and anticancer properties. This work has highlighted the ecological and evolutionary importance of these compounds, as well as their potential applications in drug discovery and environmental management.

Through meticulous research and data integration, this project has provided a robust platform for future studies on bacterial secondary metabolites, offering insights into their geographic distribution, temporal trends, and biochemical diversity. The development of this database facilitates further exploration and utilization of bacterial metabolites, paving the way for innovative solutions in healthcare and beyond. This work not only enhances our understanding of microbial ecology but also sets the stage for new discoveries in natural product research, with far-reaching implications for science and industry

20. Future Goal

Building on the comprehensive database and analysis of bacterial secondary metabolites, several future directions can be pursued to deepen understanding and expand applications:

1. Expansion of Dataset**:** Incorporating more data on secondary metabolites from newly discovered bacterial species and environmental samples can enhance the breadth and diversity of the database. This expansion will provide a more comprehensive understanding of the chemical ecology of bacteria.
2. Functional Characterization**:** Future research should focus on elucidating the specific biological roles and mechanisms of action of these metabolites. This includes studying their ecological interactions, such as how they mediate symbiotic relationships or inhibit competitors.
3. Metabolomic and Genomic Integration**:** Integrating metabolomic data with genomic and transcriptomic data will enable a systems biology approach. This will facilitate the identification of biosynthetic gene clusters and regulatory networks involved in metabolite production.
4. Biotechnological Applications**:** There is significant potential for biotechnological exploitation of these metabolites. Future work should explore scalable production methods, including synthetic biology approaches, to harness these compounds for pharmaceuticals, agriculture, and industrial applications.

5. Environmental and Ecological Studies**:** Expanding research into how environmental factors influence metabolite production can provide insights into the adaptability and resilience of bacterial communities. Understanding the impact of climate change and pollution on these metabolites could also have implications for environmental management.
6. Collaborative Platforms and Tools**:** Developing collaborative platforms and advanced analytical tools will facilitate data sharing and interdisciplinary research. Enhanced bioinformatics tools can improve the annotation, visualization, and analysis of secondary metabolite data, supporting more complex queries and hypothesis generation.

By pursuing these goals, future research will not only broaden the scientific understanding of bacterial secondary metabolites but also unlock new applications and innovations across various fields .