Can Invalid Bioactives Undermine Natural Product-Based Drug Discovery?

Jonathan Bisson,[1] James B. McAlpine,[1,2] J. Brent Friesen,[1,2,3] Shao-Nong Chen,[1,2] James Graham,[1] and Guido F. Pauli,[1,2]*

[1]Department of Medicinal Chemistry and Pharmacognosy and [2]Institute for Tuberculosis Research, College of Pharmacy, University of Illinois at Chicago, 833 S. Wood Street, Chicago, IL 60612, United States

[3]Physical Sciences Department, Rosary College of Arts and Sciences, Dominican University, River Forest, Illinois 60305, United States

■ SUPPORTING INFORMATION

Contents

CRITERIA FOR OCCURRING COMPOUNDS S3
DESCRIPTIVE STATISTICS OF COMPOUNDS BY PLANT AND DISTRIBUTION OF OCCURRENCES OF COMPOUNDS .. S4
MOST STUDIED PLANTS IN NAPRALERT S5
MOST STUDIED FAMILIES IN NAPRALERT S6
DESCRIPTIVE STATISTICS OF THE DISTRIBUTION OF THE THREE SETS S7
ANALYSIS OF PUBCHEM CONFIRMATORY ASSAYS DATA S8
LIST OF ATTACHED FILES S9

Criteria for occurring compounds

These are the different NAPRALERT codes used for identification, separation, isolation or review of compound presence.

Code	Name
AK	LIPID ANALYSIS REPORTED FOR
AM	IDENTIFICATION IN BY TLC AFTER ACID HYDROLYSIS
A8	ISOLATION OF COMPOUND WITH UNKNOWN STRUCTURE
BT	ISOLATION FROM DURING BIOSYNTHESIS STUDY
BX	IDENTIFICATION IN BY GC/MS AFTER ACID HYDROLYSIS
B1	ISOLATION OF A COMPOUND WITH NOVEL STRUCTURE
DC	IDENTIFICATION IN BY DROPLET CCD
EL	ELECTROPHORESISa
ET	COMPOUND ISOLATED FROM FOLLOWING ENZYMATIC HYD...
FM	FLUORESCENT MICROSCOPIC IDENTIFICATION OF
GF	IDENTIFICATION IN BY CMR SPECTROSCOPY
HK	ISOLATION OF NEW COMPOUND FOLLOWING SAPONIFICA...
HL	ISOLATION OF NEW COMPOUND FOLLOWING ACID HYDRO...
HM	ISOLATION OF NEW COMPOUND FOLLOWING ENZYMATIC ...
HN	IDENTIFICATION IN BY RIA
PA	ISOLATION OF PHYTOALEXIN FROM
PC	PAPER CHROMATOGRAPHIC PATTERN REPORTED FOR
PH	IDENTIFICATION IN BY PC AFTER ACID HYDROLYSIS
P9	ISOLATION OF A PHYTOALEXIN FROM
RC	PRESENT IN (REVIEW)
T1	IDENTIFICATION IN BY TLC
T2	IDENTIFICATION IN BY GLC
T3	IDENTIFICATION IN BY GC/MS
T4	IDENTIFICATION IN BY PAPER CHROMATOGRAPHY
T5	IDENTIFICATION IN BY UV SHIFTS AFTER ADDITION ...
T6	IDENTIFICATION IN BY GLC AFTER SAPONIFICATION
T7	IDENTIFICATION IN AFTER SAPONIFICATION
T8	SPECTROPHOTOMETRIC QUANTIFICATION OF IN
T9	IDENTIFICATION IN BY HPLC
UA	ISOLATION FROM AFTER SAPONIFICATION
UB	IDENTIFICATION IN BY COLOR REACTIONS
UC	IDENTIFICATION IN BY GLC AFTER ACID HYDROLYSIS
11	ISOLATION AND IDENTIFICATION OF
13	SEPARATION OF (EXCEPT OILS, VENOM, PROTEINS, F...
14	IDENTIFICATION OF
16	ISOLATION AND IDENTIFICATION AFTER CHEMICAL TR...
35	ISOLATION AND IDENTIFICATION OF AN ARTIFACT
TA	IDENTIFICATION IN BY HPLC/MS
TB	IDENTIFICATION IN BY LC/MS
TC	IDENTIFICATION IN BY CONTINUOUS FLOW FAB
LM	IDENTIFICATION IN BY LICHEN MASS SPECTROMETRY
AR	IDENTIFICATION IN AFTER ACID HYDROLYSIS
SN	ISOLATION OF A NEW COMPOUND USING SUPERCRITICA...
SK	ISOLATION OF A KNOWN COMPOUND USING SUPERCRITI...
CB	ISOLATION BY CCC
CC	IDENTIFIED IN BY CAPILLARY CHROMATOGRAPHY

Descriptive statistics of compounds by plant and distribution of occurrences of compounds

	Compounds	Occurrences
count	43578	189740
*Mean**	11.39095	3.67271
*Standard deviation**	23.5447	37.87223
min	1	1
0.25	2	1
0.5	5	1
0.75	12	2
max	795	7640

* Standard deviation and mean are not really meaningful as the distribution is long-tailed.

Most studied plants in NAPRALERT

Family name	Genus name	Species name	Distinct occurring compounds
Solanaceae	*Nicotiana*	*tabacum*	795
Araliaceae	*Panax*	*ginseng*	791
Cannabaceae	*Cannabis*	*sativa*	677
Theaceae	*Camellia*	*sinensis*	630
Vitaceae	*Vitis*	*vinifera*	614
Poaceae	*Zea*	*mays*	540
Malvaceae	*Gossypium*	*hirsutum*	514
Ginkgoaceae	*Ginkgo*	*biloba*	497
Rubiaceae	*Coffea*	*arabica*	494
Zingiberaceae	*Zingiber*	*officinale*	446
Fabaceae	*Glycine*	*max*	443
Celastraceae	*Tripterygium*	*wilfordii*	432
Liliaceae	*Allium*	*cepa*	428
Anacardiaceae	*Mangifera*	*indica*	426
Cannabaceae	*Humulus*	*lupulus*	412
Apiaceae	*Daucus*	*carota*	405
Meliaceae	*Azadirachta*	*indica*	400
Fabaceae	*Glycyrrhiza*	*glabra*	391
Solanaceae	*Capsicum*	*annuum*	389
Rutaceae	*Ruta*	*graveolens*	378

Most studied families in NAPRALERT

Family name	Distinct occurring compounds
Asteraceae	29420
Fabaceae	13424
Lamiaceae	8639
Rutaceae	6334
Apiaceae	5769
Liliaceae	5160
Euphorbiaceae	4775
Apocynaceae	4490
Solanaceae	4076
Hyphomycetes	4076
Rubiaceae	3540
Ranunculaceae	3520
Streptomycetaceae	3316
Annonaceae	3279
Rosaceae	3112
Lauraceae	2627
Araliaceae	2615
Meliaceae	2609
Poaceae	2569
Clusiaceae	2484

Descriptive statistics of the distribution of the three sets

	Occurrences	Activities	Distinct activities
count	190712	190712	190712
*Mean**	2.15	0.64	0.42
*Standard deviation **	7.10	1.50	0.93
min	1	0	0
0.25	1	0	0
0.5	1	0	0
0.75	2	1	1
max	715	9	9

* Standard deviation and mean are not really meaningful as the distribution is long-tailed.

Analysis of PubChem confirmatory assays data

	CID	Active	Total	Active / Total (%)
Kaempferol	5280863	45	179	25.1
Luteolin	5280445	77	215	35.8
Quercetin	5280343	132	252	52.4
Rutin	5280805	15	105	14.3
Resveratrol	445154	84	352	23.9
Genistein	5280961	60	322	18.6
Capsaicin	1548943	16	248	6.5
Taxol	36314	43	233	18.5
Cineol,1-8:	2758	3	229	1.3
Caffeic Acid	689043	35	220	15.9
Apigenin	5280443	63	207	30.4
Curcumin	969516	37	206	18.0
Myricetin	5281672	82	203	40.4
Catechin, Epi (-):	72276	18	194	9.3
Eugenol	3314	5	182	2.7
Gossypol	3503	71	172	41.3
Ursolic Acid	64945	21	156	13.5
Tetrandrine,(+):	73078	9	146	6.2
Linalool	6549	0	144	0
Glycyrrhizin	14982	6	123	4.9
Catechin,(+):	9064	10	117	8.5
EGCG	65064	41	116	35.3
Gallic Acid	370	38	110	34.5
Quercitrin, Iso-	5280804	14	81	17.3
Pinene, Alpha-	6654	0	80	0
Berberine	2353	4	72	5.6
Myrcene	31253	1	59	1.7
Limonene	22311	3	45	6.7
Sitosterol, Beta-	222284	1	18	5.6
Cymene, Para-	7463	0	14	0
Pinene, Beta-	14896	0	14	0
Ginsenoside Rb-1	9898279	0	2	0
Daucosterol	5742590	1	2	50
Stigmasterol	5280794	0	1	0
Lupeol	259846	1	1	100
Ginsenoside Rg-1	441923	0	0	0
Oleic Acid	445639	17	192	8.8
Linoleic Acid	5280450	19	101	18.8
Palmitic Acid	985	11	54	20.4

List of attached files

All files have been renamed as .txt as no open-format was allowed by the ACS policy.

compounds_chembl_pubchem.csv.txt	List of compounds, SMILES, ChemBL, and Pubchem codes
table_compounds.csv.txt	Table 3 of the article as csv file
data_exploration/gather_chembl_data.py.txt	Gather ChEMBL information from the NAPRALERT compound name
data_exploration/data_exploration.ipynb.txt data_exploration/data_exploration.py.txt data_exploration/data_exploration.html.txt	The Notebook used for the data exploratory part and the generation of most of the figures in the article.
powerlaw/fitting.py.txt powerlaw/fitting.ipynb.txt powerlaw/fitting.html.txt	Powerlaw fitting code Python file (if you are not using IPyNB or Jupyter), IPython/Jupyter Notebook and HTML output of the notebook
LICENSE.txt, gpl.txt	License files (GNU GPL v3)