Journal of Molecular Graphics and Modelling 120 (2023) 108414

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/jmgm

Chemo-structural diversity of anti-obesity compound database

Prabhakar Lavanya, Dicky John Davis G *

Department of Bioinformatics, Faculty of Engineering and Technology, Sri Ramachandra Institute of Higher Education and Research (DU), Chennai, India

ARTICLE INFO

Keywords:
Anti-obesity database
Natural compounds
Fingerprint diversity
Chemical descriptors
Scaffold diversity and molecular properties

ABSTRACT

Nature plays a major role in the development of new drugs which helps in preventing and treating human diseases. Anti-obesity compound database (AOCD) contains comprehensive information on all published small molecules from natural sources with anti-obesity potential targeting pancreatic lipase (PL), appetite suppressant (AS) and adipogenesis (AD). Presently the database contains 349 compounds isolated from 307 plants, 26 marine and 16 microbial sources. Users can query the AOCD database (https://aocd.swmd.co.in/) in several ways. The database was divided into three datasets (PL, AS and AD) to perform chemoinformatic analysis using Platform for Unified Molecular Analysis (PUMA), which were analyzed based on molecular descriptors, scaffold diversity and structural fingerprint diversity. Chemoinformatics study inferred the PL dataset has the highest diversity of compounds based on the Euclidean distance on molecular properties, scaffold diversity and pairwise similarity on fingerprint diversity. This study would hasten the process of anti-obesity drug discovery.

1. Introduction

The prevalence of obesity is increasing with the substantial lifestyle changes over the past three decades. Due to this remarkable increase in lifestyle changes, the risk factors are triggered which results in affecting people's health condition [1]. These changes lead to significant illnesses such as hypertension, type-2 diabetes, cardiovascular diseases and hyperlipidaemia [2]. The effectiveness of synthetic drugs in targeting obesity was studied, which was found to possess consequent side effects [3]. To overcome these effects, there is a need to explore natural resources to treat obesity.

Nature plays a vital role as a rich, active pharmaceutical that aids in preventing and treating diseases. Natural-product drugs have emerged from an extensive variety of natural sources such as plants, marine organisms, microbes and animals [4]. Recent studies have also reported that 50% of the present drugs have originated from natural sources [5]. The inherent structural diversity of natural products is significantly large when compared to that of synthetic compounds which have given rise to modern drug discovery [6]. Natural products are proven to be an important source of lead structures that can be used as a template for the development and design of new drugs [7]. Natural products and their structural analogs can be a promising pool that is directly developed or used as starting points for optimization into novel drugs [8]. Previous studies have reported that novel marine metabolites from the algal source are found to exhibit anti-cancer and anti-obesity properties [9, 10]. Natural products serve as the most traditional source for the development of new drugs [11]. Thus, the information on natural products becomes easy to access when it gets standardized and assembled into a database. A publicly available database that provides extensive information about natural products is much needed to facilitate the importance of sharing it with relevant communities. This knowledge sharing from the database can minimize the timeline in drug discovery and helps in understanding the mechanism of compounds [12].

Significant progress has happened over the last few decades which had given rise to the curation of databases for various ailments. The accessibility of curated databases of all possible traditional medicinal plants and their chemical and biological functions generally aids in drug development. A repository which possesses collective information on the phytochemistry of plants, their chemical structure and ethnopharmacological behavior has already been reported in preceding databases such as KNApSAcK, TCMID, CVDHD, Phytochemica, Nutrichem, OCDD, DiaNAT- DB and COCONUT. KNApSAcK database describes the plant species and their metabolites and also relates it to their geographical zones [13]. The Traditional Chinese Medicine Integrative Database (TCMID) portrays the relationship between the different herbs and their diseases and facilitates the understanding of underlying mechanisms at a molecular level [14]. Cardiovascular disease herbal database for drug discovery and network pharmacology (CVDHD) depicts herbs of medicinal use, natural products and proteins that target

* Corresponding author.
E-mail address: dicky@sriramachandra.edu.in (D.J. Davis G).

https://doi.org/10.1016/j.jmgm.2023.108414
Received 18 August 2022; Received in revised form 28 December 2022; Accepted 12 January 2023
Available online 19 January 2023

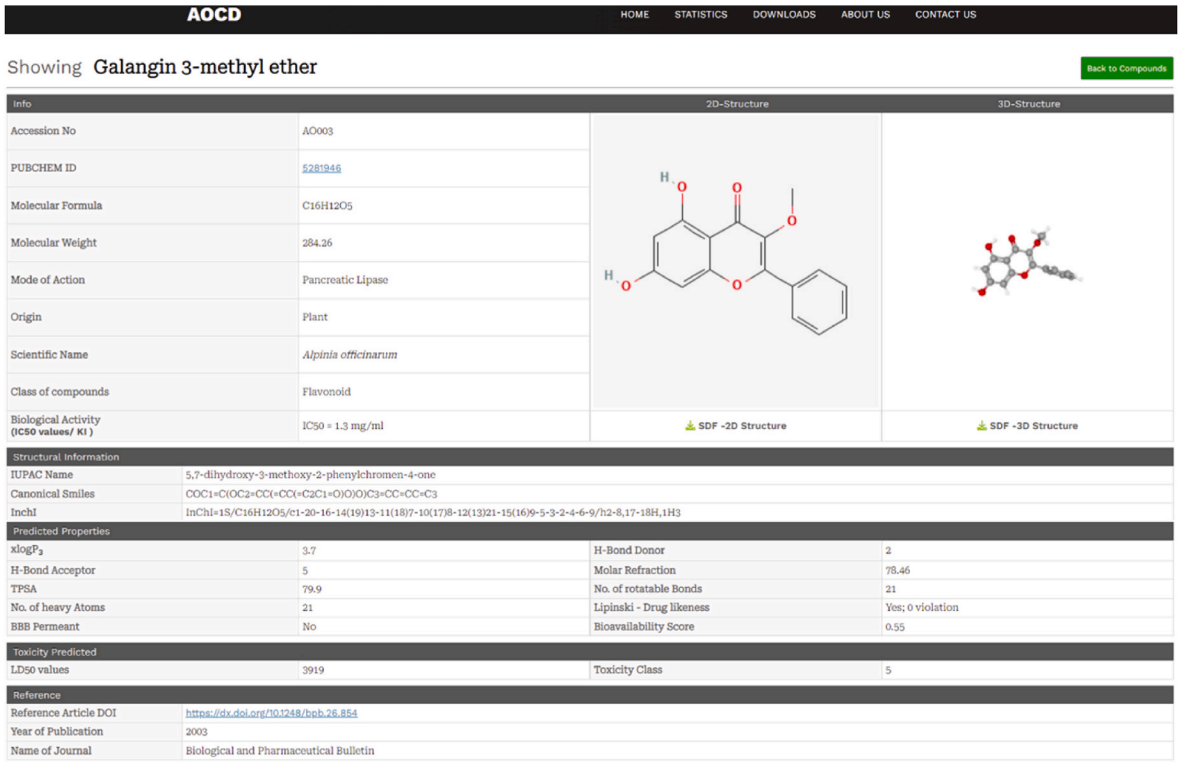

Fig. 1. The search result page of AOCD Database.

cardiovascular disease [15]. The tendency of producing therapeutic molecules extracted from medicinal plants are compiled in one single platform named Phytochemica [16]. Obesity and Co-morbid disease database (OCDD) involve the relationship of genes that target both obesity and its co-morbid diseases. The interaction among gene networks, functional illustration of common genes and key driver analyses have made it a more valuable and effective database [17]. DiaNAT DB is a comprehensive collection of unique anti-diabetic, natural product from medicinal plants [18]. COlleCtion of Open Natural ProdUcTs (COCONUT) is an open-source project for natural products storage, search and analysis [19]. In addition, IMPPAT (Indian Medicinal Plants, Phytochemistry and Therapeutics) database helps in providing a unified platform for carrying out all possible chemoinformatic approaches to accelerate drug discovery using natural sources [20]. However, for treating specific underlying condition such as obesity, an initial effort of the Anti-Obesity Compound Database (AOCD) was created, comprising natural anti-obesity compounds. This database stands out in delivering extensive information regarding the natural compounds that can be used for conducting in-silico approaches with the illustrated molecular properties, toxicity properties, and 3D structures of compounds. It also provides additional information on IC_{50}/KI values to know about the biological activity of the plant, microbial or marine source. The compouunds in AOCD database consists of different class of compounds such as flavonoids, terpenes, hydrolase, prenol lipids, polyketide, tannin, alkaloids, saponins, glycosides, phenolic compounds, fatty acids, steriod derivatives, cinnamic acids, napthoquinone derivates, etc. The database covers a datasets of anti-obesity compounds from three different mechanisms of action against obesity namely Pancreatic Lipase, Appetite suppressant and Adipogenesis.

This research deals with the chemoinformatics and characterization of chemical diversity studies of AOCD database compounds. Chemoinformatics is a scientific discipline that employs various computational methods in analysing chemically related data [21]. Applications of chemoinformatics include analysis of physical, chemical and biological properties of molecules, macromolecular interactions, nutraceuticals, polymer chemistry and high throughput screening [22]. Recent report on chemoinformatics interest on natural products pointed of the exploration of natural product database, origination of secondary metabolites databases and their expanded accessibilites. Analysis of chemical properties helps in the understanding of database compounds for screening and lead optimization [23]. Furthermore, chemical diversity was investigated using scaffold diversity, fingerprint diversity and physiochemical properties [24]. Recently, the development of drugs from the chemical properties of natural products have been analyzed [25]. Chemoinformatics characterization aids in designing virtual chemical databases by understanding the molecules characteristics and exploring as to what makes them distinct [26]. It also provides suitable tools for estimating the physicochemical properties of natural products and aids in the selection of macromolecular targets [27]. In addition, the development of chemoinformatics reformed the curation of metabolomics data, specifically with their suitable functional properties of annotated metabolites and their defined roles in metabolic pathways [28].

In this research, AOCD delivers a comprehensive resource on botanicals isolated from plant source(s) and their classification. It also includes compounds derived from marine seaweeds: Microbial sources are found to have a potent source of inhibition targeting obesity. This database provides extensive information on physicochemical descriptors, pharmacokinetic properties, drug-like nature, computational toxicity prediction and lipophilic properties of compounds to support drug discovery. In addition to this, chemoinformatic-based diversity analysis using PUMA (Platform for Unified Molecular Analysis) is performed to calculate properties of molecular descriptors, understand the diversity of scaffolds and to determine fingerprint diversity.

2. Methodology

2.1. Database structure

The records in the database are produced by text mining of published

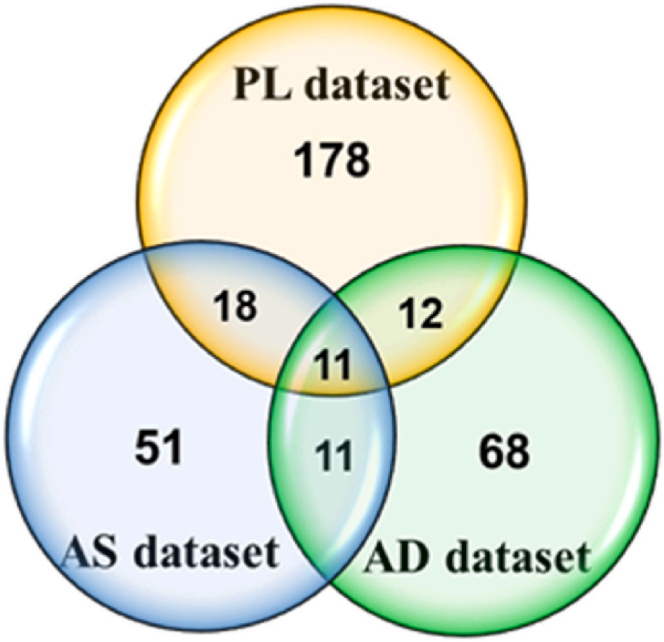

Fig. 2. Distribution of anti-obesity compounds based on three different modes of action. Total number of compounds in AOCD database is 349, 219 compounds targeting Pancreatic Lipase (PL), 91 compounds targeting Appetite Suppressant (AS) and 102 compounds targeting Adipogenesis (AD).

literature. The design of this database is made using PHP 5.3 software (https://www.php.net/releases/5_3_0.php). At present, our AOCD database is an "in-house" database comprising 349 compounds out of which 307 compounds were retrieved from plant sources, 26 compounds from marine and 16 compounds from microbial sources. These compounds cover 282 different species of plant origin, 14 different varieties of marine and 8 species of microbial origin and these are represented in Table S1 (supplementary file). In this database, we surveyed all possible natural compounds with anti-obesity potential targeting 219 compounds against Pancreatic Lipase, 91 compounds against suppressing appetite and 102 compounds inhibiting the adipocyte differentiation. The biological activity of these compounds was identified based on their IC50/KI values and is included in this database.

Compounds in AOCD were described by their molecular properties which includes molecular weight, molecular formula, InChl, canonical SMILES, IUPAC name, XLOGP3, number of hydrogen bond donors and acceptors, TPSA, number of rotatable bonds and Lipinski's rule, bioavailability score, BBB permeant, LD50 values and toxicity class. The 2D and 3D structures of the compound can be accessed through SDF or PNG format depending on the relevance. MarvinSketch [29] was employed to elucidate the 3D structure of each compound. The pharmacokinetic properties were predicted using SwissADME [30]. The parameters that satisfied Lipinski's property include MW $\leq$ 500 Dalton, cLogP value less than 5, a maximum of 10 H bond acceptors and 5 H bond donors [31]. Toxicity values were calculated using the ProTox-II server based on the Globally Harmonized System of Classification of Labelling of Chemicals (GHS) [32]. The toxicity classes range from I to VI and the input compound gets classified based on the toxicity level, wherein class I is classified as highly toxic. The class V and VI compounds which possess LD50 values ($2000 < LD50 \leq 5000$) and ($LD50 > 5000$) were found to be acceptable and non-toxic respectively. The web interface for AOCD is depicted in Fig. 1. The accessibility to this database can be searched using accession number, the nomenclature of the compounds, SMILES notation or InChl, PubChem ID, mode of action and based on their origin. Text case is insensitive in the search field. This is available as an open-access database wherein the public can access and download relevant information as per requirement.

2.2. Chemical space diversity

Dataset diversity analysis and visualization were performed using PUMA, Version 1.0 [33]. Chemical space, molecular properties diversity, scaffold diversity and structural fingerprint diversity were analyzed for the compounds in the dataset using PUMA. Based on drug targets, AOCD compounds include three different datasets namely, Pancreatic Lipase (PL), Appetite Suppressant (AS) and Adipogenesis (AD). The number of compounds in each data set is shown in Fig. 2. Prior to the analysis, all compounds were curated by removing duplicates and open babel software was used to assign valences and protonation state of the compounds [34].

2.3. Calculation of chemical descriptor

The chemical descriptors were generated using R package rcdk [35] to compute six molecular properties: molecular weight (MW), hydrogen bond donors (nHBDon), hydrogen bond acceptors (nHBAcc), topological polar surface area (TPSA), number of rotatable bonds (nRotB) and the octanol water partition coefficient (ALogP). The statistics of the six molecular properties were visualized as frequency histogram. The visual representation of the molecular descriptors was analyzed using Principal component analysis (PCA).

2.4. Scaffold diversity

The scaffold definition of molecular core without side chains as suggested by Bemis and Murko was used in this study [36]. PUMA use rcdk to obtain the cyclic Murcko ring systems of all compounds by removing the side chains [33]. Every scaffold (chemotype) is assigned with a unique identifier (ID) and Cyclic System Retrieval curves (CSR) were used to represent the distribution of chemotypes. The quantification of scaffold diversity was calculated by area under the curve (AUC) and the fraction of chemotypes that recover 50% of the molecule on the dataset. Scaled Shannon Entropy (SSE) was calculated to identify the most populated scaffolds.

2.5. Fingerprint diversity

Fingerprint diversity of the tested datasets was computed using three molecular fingerprints options; The Extended Connectivity Fingerprints (ECFP_4) option encodes the circular type of fingerprints with a diameter of four, PubChem represents PubChem's binary substructure fingerprints with 881 bits and MACCS computes the 166- bit MACCS keys [33].

3. Results and discussion

3.1. Database features

AOCD possesses a web interface at https://aocd.swmd.co.in/. Currently, the database holds entries for 349 natural compounds confining to various data fields, predominantly from plant extracts followed by marine and microbial sources. The database delivers an exceptional source of information from the various descriptive fields thereby creating a comprehensive understanding of the natural compounds that target obesity. The categorization for each entry in AOCD includes general information, predicted properties, structural information, toxicity prediction and references. The basic information section provides us with critical details on the accession number, followed by an external link for PubChem Id. Additionally, it contains particulars on the molecular formula and molecular weight of each compound along with the scientific name, origin, class of compounds and biological activity of the compounds retrieved from the literature. The structural information mainly deals with IUPAC name of the compound, SMILES notation or InChl along with the SDF and PNG format for 2D and 3D structures,

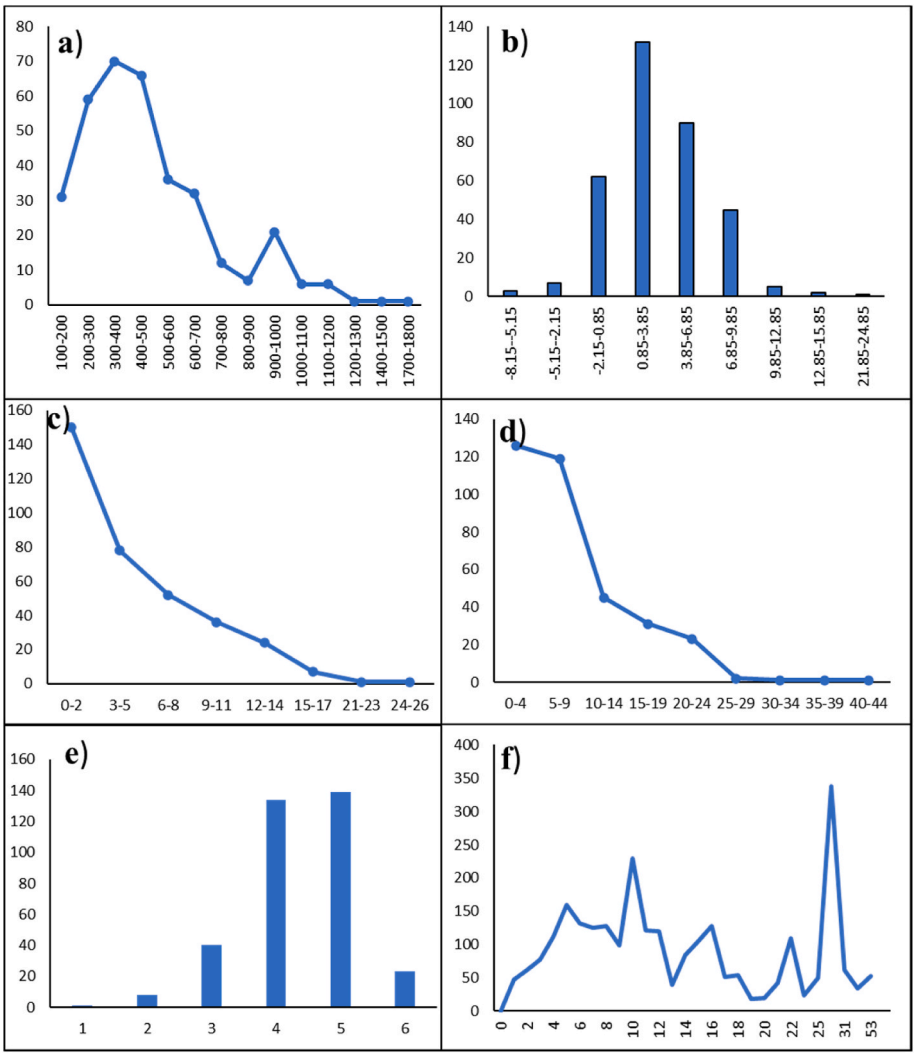

Fig. 3. Molecular properties of 349 compounds in AOCD database. X axis represents (a) molecular weight in daltons, (b) ALogP, (c) hydrogen-bond donors, (d) hydrogen-bond acceptors, (e) class of toxicity and (f) rotatable bonds; Y axis represents No. of compounds.

respectively. The predicted properties generally encompass the pre-computed chemical descriptors of the compounds. It was observed that 226 compounds have a molecular weight less than 500 Daltons, 245 compounds showed cLogP value less than five, 228 compounds resulted in a maximum of 5 H bond donor and 254 compounds exhibited a maximum of 10 H bond acceptor (Fig. 3).

Lipinski's rule is focused to guide the design of drug-like compounds based on specific molecular properties. Further to confirm the oral bioavailability of the compounds, toxicity prediction results suggested that compounds that come under the toxicity class of V can be considered as a drug. Twenty-three compounds that come toxicity class VI are non-toxic compounds. Earlier studies have reported that flavonoids and terpenes classification of compounds were found to have anti-obesity activity [3]. These flavonoids and terpene type of compounds are majorly present in the AOCD database.

The toxicity section displays the LD50 values and class of toxicity of the compounds. It was found that 185 compounds showed median lethal dose (LD50) values above 2000 which comes under toxicity class 5 and 6 according to the GHS system. Anti-obesity compounds from the AOCD database comes under the different classifications of compounds (Fig. 4). The structural diversity showed overall flavonoids are more, specifically in the PL dataset followed by terpenes, hydrolase, prenol lipids, etc. The bibliography section displays the relevant references to the respective compounds. A total of 197 literature studies were text mined for the reference section.

3.2. Calculation of chemical descriptor

The Platform for unified molecular analysis software was used for the calculation of molecular descriptors. The statistical difference between the datasets were determined using non-parametric Wilcoxon rank sum test. The statistical distribution of six molecular properties: molecular weight (MW), topological polar surface area (TPSA), hydrogen bond donors (nHBDon), hydrogen bond acceptors (nHBAcc), octanol-water partition coefficient (ALogP) and number of rotatable bonds (nRotB) which have pharmaceutical relevance are provided in Table 1.

Molecular size was represented by MW, which indicated that PL datasets have the highest average MW value of 512.97 and largest MW of 1729.47 when compared to other dataset values. ALogP, HBDon and HBAcc determine the hydrophilicity of the datasets. It was noticed that PL dataset has the highest mean values considering nHBAcc and nHBDon properties followed by AS and AD. AD showed the highest mean value for octanol/water partition coefficient property followed by PL and AS datasets. The flexibility of the tested dataset compounds was represented using TPSA, and the nRotB. PL dataset showed highest value of calculated flexibility parameters. There is no statistical significance between MW of PL datasets when compared to AS or AD datasets. In addition to this, significant difference between AS and AD datasets was

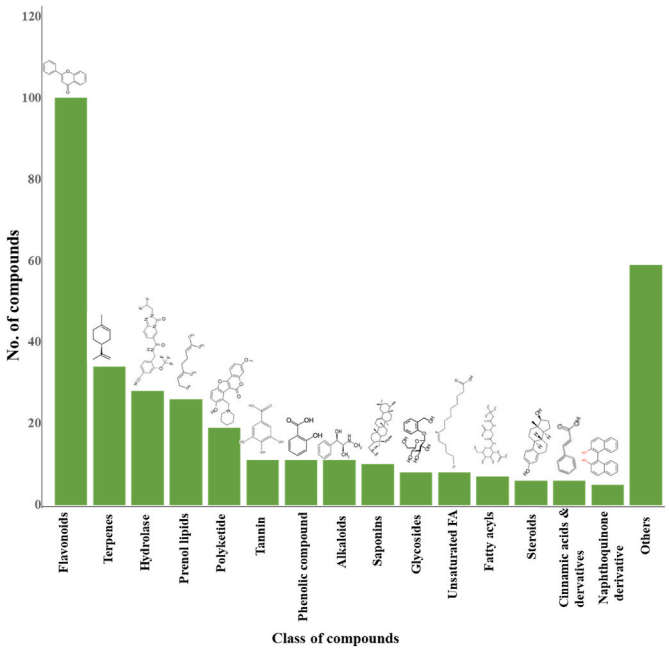

Fig. 4. Different class of compounds reported in AOCD database.

observed. The statistical significance difference between PL and AD datasets was observed in all calculated hydrophilic descriptors. Regarding the molecular flexibility, AD dataset showed significant difference when compared to AS or PL datasets. The molecular properties of the tested datasets were plotted as frequency histograms as shown in Fig. 5.

Principal component analysis facilitates the two-dimensional mode of visualization of database properties as depicted in Fig. 6. PCA of anti-obesity compound database was represented by three principal components (PC1, PC2 and PC3) which resulted in 95% variability and the summary of PCA loadings is given in Table 2. Interestingly, it was noted that from Table 2, six molecular descriptors were found to be positive to PC1, which indicates rise in the descriptor values along x-axis. PC2 is mostly associated with molecular weight and PC3 is associated with TPSA properties.

In addition to chemical space visualization, outliers have been recognized from each dataset of the AOCD database (Fig. 6). Five outliers are identified from the PL dataset, nine outliers are recognized from AS dataset and six outliers from the AD dataset.

PL dataset contains five outliers, which includes AO240 Anthocyanin1, AO224-(6S,9R,12R)-1-((R)-1-((R)-2-((S)-1-((S)-5-amino-2-((R)-2-amino-3-mercaptopropanamido)-5-oxopentanoyl) pyrrolidine-2-carboxamido)-3-(1H-imidazole-5-yl) propanoyl) pyrrolidin-2-yl)-6-(3-amino-3-oxopropyl)-9-((R)-1-hydroxyethyl)-12-(mercaptomethyl)-1,4,7,10-tetraoxo-2,5,8,11-tetraazatridecan-13-oic acid, AO159- Punicalagin, AO154- Platycodin D and AO327 Triolein (Fig. 7).

PL dataset containing anthocyanin 1, a flavonoid isolated from *Morus austrails poir*, was reported to prevent obesity [37]. Triolein from marine algae *Caulerpa taxifolia* involved in the inhibition of lipase enzyme [38]. Another PL dataset outlier (6S,9R,12R)-1-((R)-1-((R)-2-((S)-1-((S)-5-amino-2-((R)-2-amino-3-mercaptopropanamido)-5-oxopentanoyl) pyrrolidine-2-carboxamido)-3-(1H-imidazole-5-yl) propanoyl) pyrrolidin-2-yl)-6-(3-amino-3-oxopropyl)-9-((R)-1-hydroxyethyl)-12-(mercaptomethyl)-1,4,7,10-tetraoxo-2,5,8,11-tetraazatridecan-13-oic acid of hydrolase classification was extracted from *Curcuma amada* [39]. Platycodin D extracted from *Platycodon grandiflorum* which belongs to the class of terpenes exhibited anti-obesity effects. Punicalagin from *Punica granatum* belongs to the class of tannins was reported to have anti-obesity effects [40].

AS dataset has recognized nine outliers which includes AO160-Punicalin, AO101- Ginsenoside Rb1, AO063 – Corilagin, AO193- vitisin A, AO163- Quercetin 3,7 diglucoside, AO276- Levan n, AO016- Annohexocin, AO154 Platycodin D and AO159 Punicalagin (Fig. 8). AS dataset contains Punicalin, a class of tannin, isolated from *Punica granatum* reported for anti-obesity effect [40]. Ginsenoside Rb1 extracted from *Panax quinquefolium* was also found to have anti-obesity activity [41]. Another AS outlier Corilagin, a tannin class of compound obtained from *Geranii herba* reported for anti-viral, anti-obesity, anti-inflammatory activities [42]. Vitisin, a flavonoid class of compound, isolated from *Vitis vinifera* found to possess anti-inflammatory, neuroprotective, and anti-cholesterolemic activities [43]. Quercetin 3, 7 diglucoside extracted from *Taraxacum officinale* which belongs to the class of flavonoid has anti-inflammatory, antioxidant and anti-obesity activities [44]. Levan, a polysaccharide, obtained from *Lolium perenne* reported to prevent obesity [45]. Annohexocin isolated from *Annona muricata* which belongs to the class of polyketide reported for the anti-obesity effect (Elekofehinti 2020).

Table 1
Statistical distribution of the calculated chemical descriptors.

Chemical descriptors	Dataset	Min	1st Qu[a]	Median	Mean	3rdQu[b]	Max	Std.Dev[c]
MW	PL	123.03	317.10	444.23	512.97	626.36	1729.47	271.52
	AS	123.03	280.69	332.19	413.90	594.44	1224.57	216.58
	AD	0	235.03	314.13	374.96	447.49	1422.65	218.65
TPSA	PL	0	66.76	117.76	152.72	218.29	695.41	119.16
	AS	9.23	46.82	99.28	116.52	136.68	510.94	97.55
	AD	0	41.58	86.93	104.20	129.82	565.05	94.66
nRotB	PL	0	3	6	7.36	10	53	6.66
	AS	0	2	5	8.69	14.5	26	8.94
	AD	0	1	3.5	4.91	7.75	31	5.18
nHBDon	PL	0	1	4	5.12	8	25	4.83
	AS	0	1	3	4.12	5	17	3.85
	AD	0	1	2.5	3.58	5	22	3.73
nHBAcc	PL	0	4	7	9.25	12	43	7.07
	AS	1	3	6	6.91	8	30	5.72
	AD	0	3	4.5	6.09	7.75	35	5.67
ALogP	PL	−11.19	−2.71	−0.94	−0.92	1.24	8.93	2.78
	AS	−8.78	−4.51	−1.303	−2.39	0.13	2.53	3.30
	AD	−11.12	−1.194	0	0.15	1.36	8.55	1

MW: molecular weight, TPSA: topological polar surface area, nRotB: number of rotatable bonds, nHBAcc: number of hydrogen bond acceptors, nHBDon: number of hydrogen bond donors, AlogP: octanol/water partition coefficient.

[a] First quarter.

[b] Third quarter.

[c] Standard deviation.

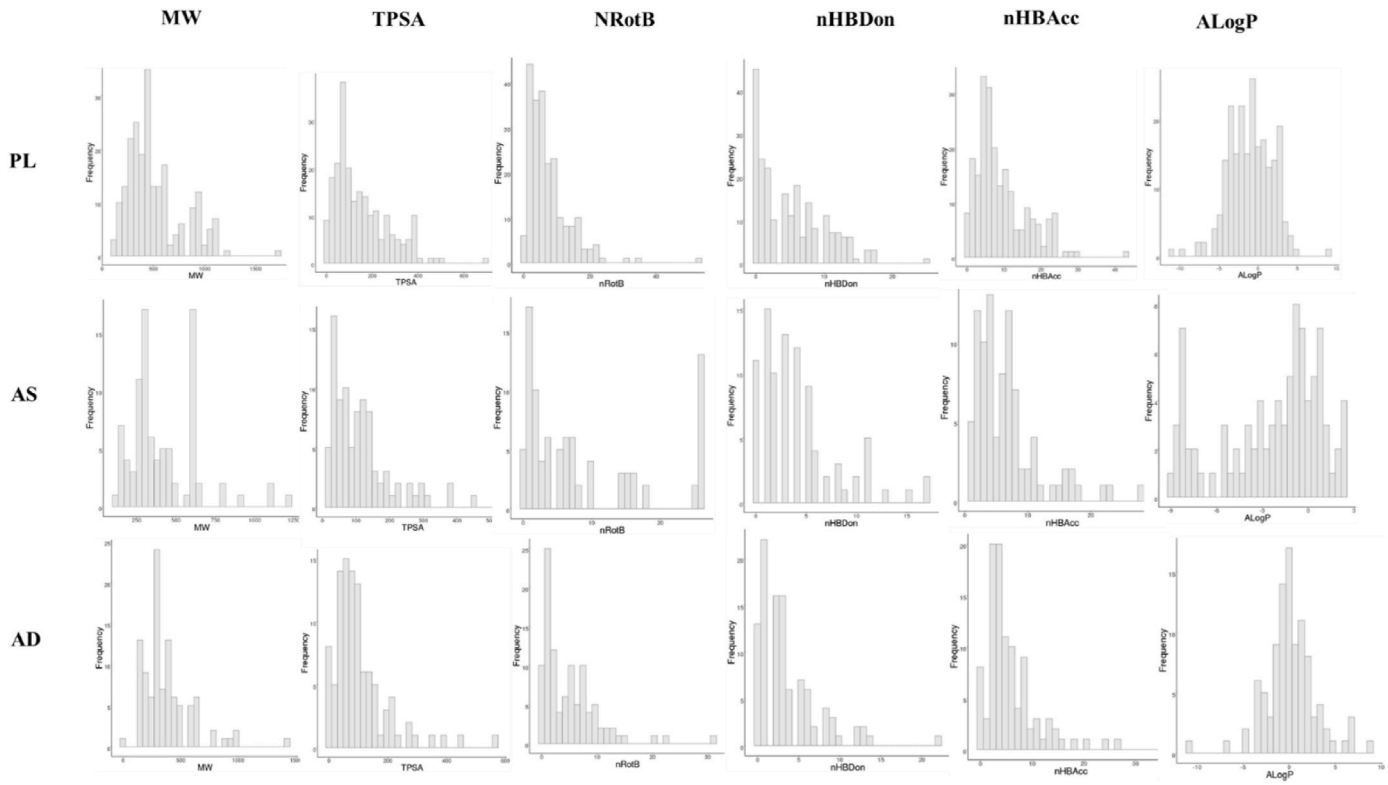

Fig. 5. Frequency distribution of chemical descriptors of tested datasets, including PL, AS and AD datasets.

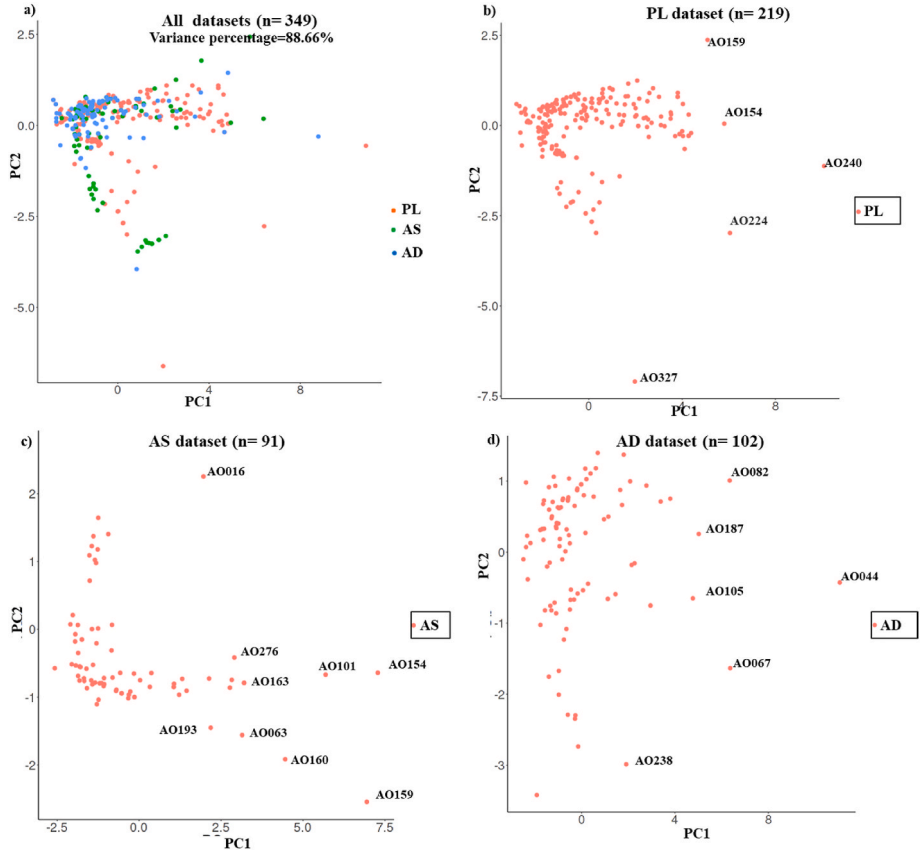

Fig. 6. Principal component analysis (PCA) for all the tested datasets **(a)**; PL, AS and AD datasets with outlier's **(b), (c)** and **(d)**.

Table 2
PCA contribution of six Molecular properties to each Principal component.

Chemical descriptors	PC1	PC2	PC3
MW	0.43	−0.02	−0.51
TPSA	0.46	0.21	−0.03
nRotB	0.21	−0.82	−0.32
nHBDon	0.44	0.29	0.09
nHBAcc	0.46	0.18	−0.04
ALogP	0.35	0.37	−0.78
Proportion of variance	89.69	97.89	99.12

PC1- First Principal component; PC2- Second Principal component; PC3- Third Principal Component.

AD dataset contains six outliers which includes AO044- Capsicoside, AO238- 9,12,15 Octadecatrienoic acid, AO067-Crocin, AO105- Ginsenoside Re, AO187- Theaflavin 3,3′-digallate, AO082- Ellagitannins (Fig. 9). Capsicoside, an outlier of AD dataset isolated from *Capsicum annum* which belongs to the class of steroids involved in the treatment of obesity [46]. 9,12,15 Octadecatrienoic acid belongs to the class of flavonoid isolated from *Portulaca oleracea* reported for obesity [47]. Crocin, another AD outlier, isolated from *Crocus Sativus* which belongs to prenol Lipid was found to possess anti-oxidant and anti-obesity activities [48]. Ginsenoside, a glycoside from *Panax ginseng* was found to possess anti-diabetic, anti-microbial, anti-cardiovascular, anti-inflammatory, anti-oxidant and anti-obesity activities [49]. Ellagitannins, a flavonoid obtained from *Camellia sinensis*, were reported for anti-obesity effect [50].

Pairwise comparison of inter- and intra-distances of dataset compounds was computed with six molecular properties using Euclidean distance as shown in Fig. 10. The highest inter-distance was found between AS and AD datasets, followed by PL and AD and finally between PL and AD. Further, PL dataset possessed the highest intra-distance diversity based on the calculated molecular properties. Euclidean distance computed that PL datasets possessed the highest intra-distance diversity with highest number of compounds (n = 219) when compared to the other datasets in database. The highest number of scaffolds were found in the PL dataset which can be indicated by the effect of the size of the datasets.

3.3. Scaffold diversity of AOCD database

The statistical scaffold parameters that summarize compound number (M), unique scaffold (N), number of chemotypes containing only one compound (NSING), the fraction of chemotypes and singletons relative to the number of molecules in the data set (FNM and FNSING. M, respectively) is represented in Table 3. A total of 104 scaffolds were generated from all datasets, which were further divided into 47 scaffolds for PL datasets, 27 scaffolds for AS datasets and 30 scaffolds for AD datasets (Fig. 11a). PL dataset has the highest scaffold number when compared to AD followed by AS. Eleven common scaffolds were found to be common among all three datasets. In addition to this, PL and AS datasets possessed the highest number of common identical scaffolds (Fig. 11b).

The distribution of chemotypes were analyzed by computing CSR curves (Fig. 11c). The CSR curve is obtained by plotting a fraction of scaffolds on the x-axis and a fraction of compounds on the y-axis. Quantification of the CSR curve (Fig. 11c) using the area under the curve (AUC) and a fraction of chemotypes required to acquire half of the compounds (F50) can determine more diverse datasets based on their scaffolds. AS dataset possessed the highest scaffold diversity with AUC and F50 values of 0.78 and 0.07, respectively, followed by AD and PL datasets. Earlier studies have reported that the highest F50 value along with AUC value close to 0.5 possessed the highest scaffold diverse dataset [33]. Findings from the CSR curve showed that AS dataset curve is close to the diagonal with the highest F50 values which imply the highest scaffold diversity when compared to AD and PL datasets.

Scaffold diversity for all datasets was described using CSR curves, while the most populated frequent scaffold employs Scaled Shannon Entropy (SSE) [51]. The most frequent scaffold was calculated using Scaled Shannon entropy (SSE). These SSE values vary between 0 and 1, where 0 indicates most of the compounds in the dataset share the same scaffold and 1 implies all scaffolds consist of the same number of compounds. Scaled Shannon Entropy (SSE) for the most populated scaffold (10–60 scaffolds) is given in Table 4.

When comparing the scaffold diversity for the tested datasets using SSE at the level of 10 scaffolds, the PL dataset is more diverse with SSE value of 0.96 followed by AD and AS datasets. The most populated scaffold (SSE10) for each dataset such as PL, AS and AD is represented in

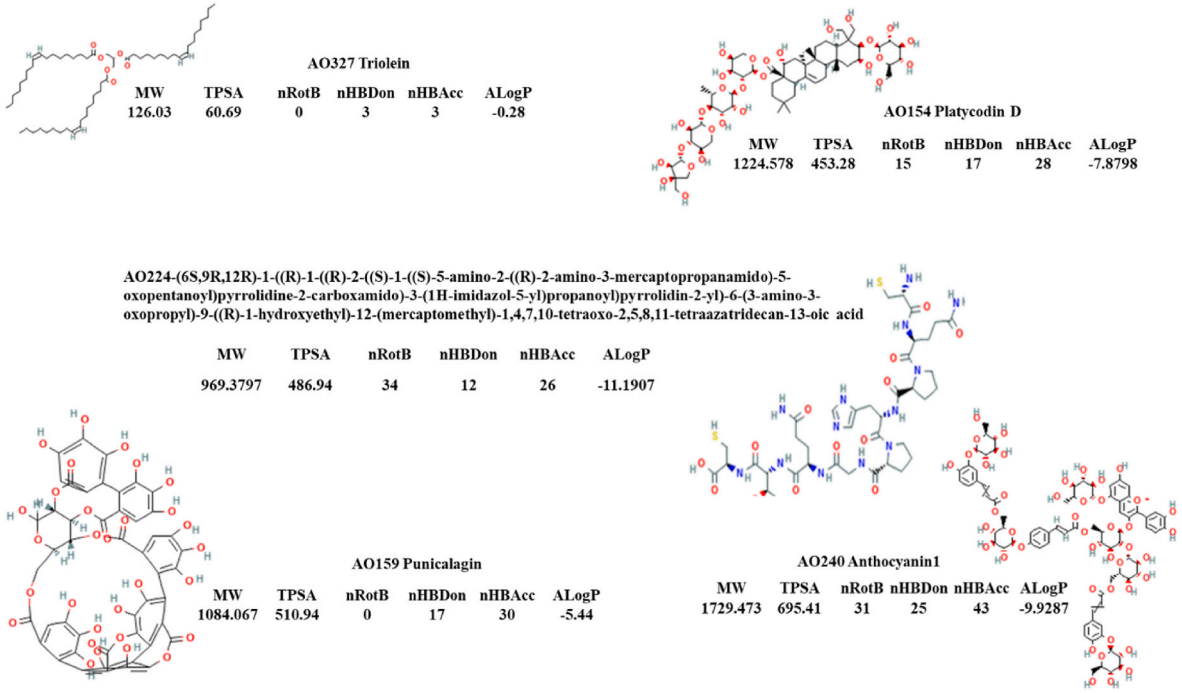

Fig. 7. Identified Outlier's in PL dataset representing chemical structure along with its molecular descriptors.

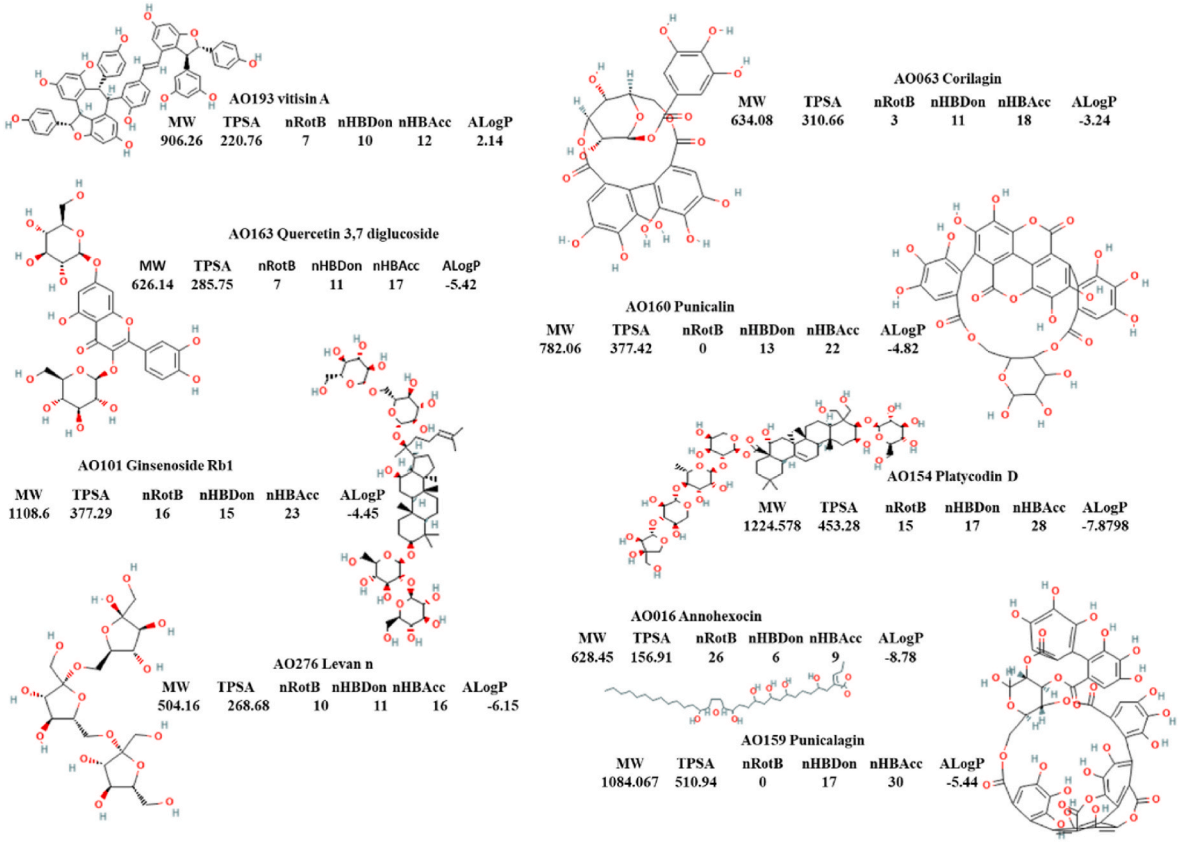

Fig. 8. Identified Outlier's in AS dataset representing chemical structure along with its molecular descriptors.

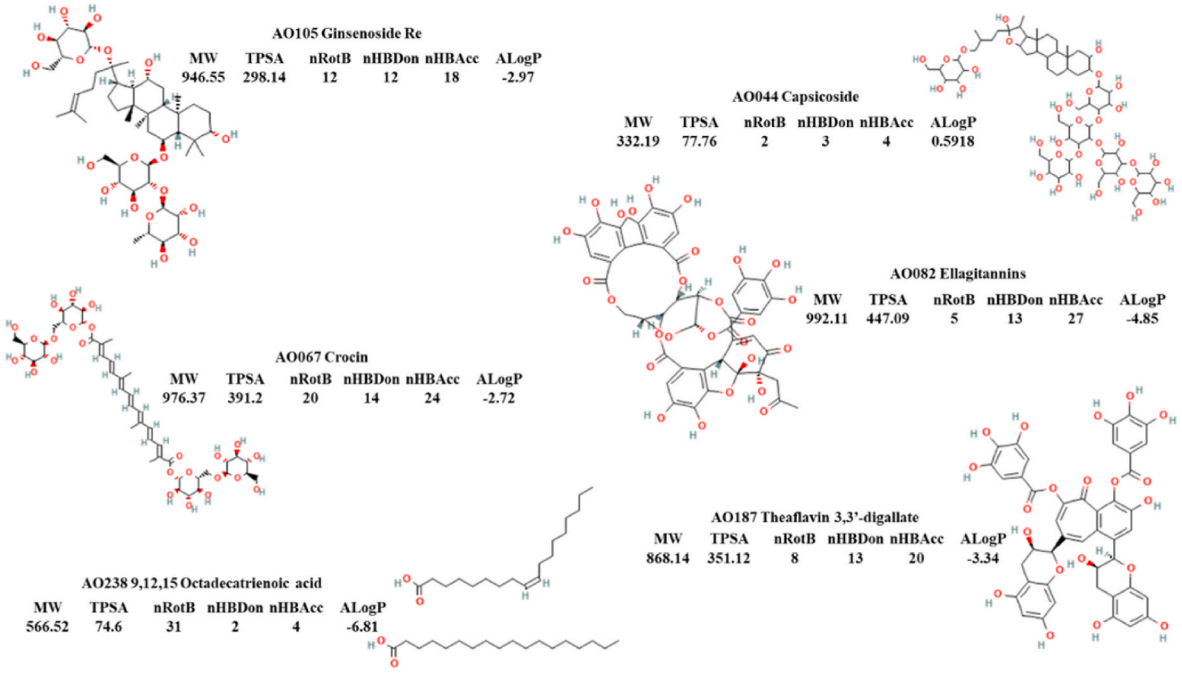

Fig. 9. Identified Outlier's in AD dataset representing chemical structure along with its molecular descriptors.

Fig. 12, Figs. 13 and 14, respectively.

One of the most common scaffolds among the three datasets is hydroxy citric acid with scaffold ID 1, which has a scaffold frequency of 15 for PL datasets, 12 for AS datasets and 11 for AD datasets. This scaffold accounts for 31.9%, 44.4%, and 66.6% of the total number of compounds for PL, AS and AD datasets, respectively. Another scaffold gallocatechin is found to be common in PL (with a scaffold frequency of 5) and AS datasets (with a scaffold frequency of 3). In addition, gallic acid and Myricetin share a common scaffold among PL and AD datasets. Further, Daidzein is found to be common in AS and AD with scaffold

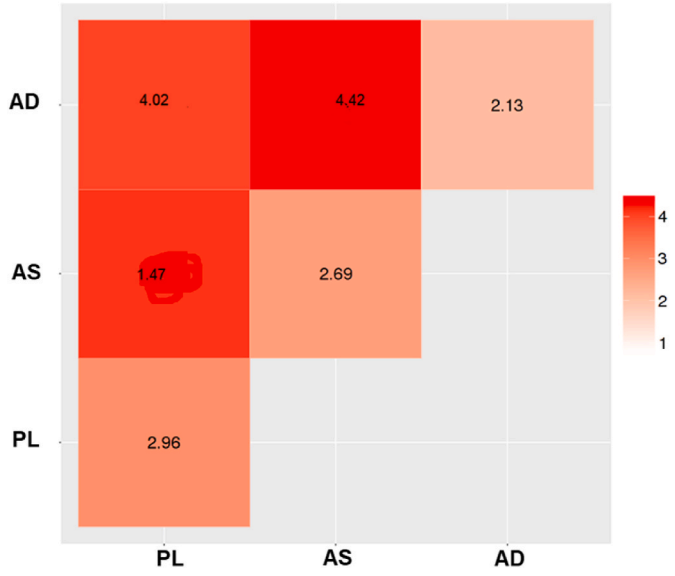

Fig. 10. Euclidean distance correlation matrix computed with six molecular properties for the compounds in Dataset.

frequencies of 2 and 3, respectively. The most common scaffold hydroxy citric acid among the three datasets exhibited anti-inflammation, anti-oxidant, anti-cancer and anti-obesity activities [52–54].

3.4. Fingerprint diversity

The distribution of three fingerprint diversities including ECFP_4, PubChem (881-Bits) and MACCS keys (166-Bits) was computed and represented in cumulative distribution functions (CDFs) (Fig. 15 and Table 5).

The results suggested that all three datasets followed the same path and it is difficult to distinguish the fingerprint diversity based on ECFP_4 (Fig. 15a). Thus, the fingerprint diversity of ECFP_4 for tested dataset showed no significant difference as all the datasets followed same mathematical path. According to MACCS (166-Bits) of similarity matrix, the PL dataset showed less diversity with a similarity median of 0.44 followed by AS and AD datasets (Fig. 15b). Further, PubChem (881-Bits) pairwise similarity data suggested that AD is less diverse with similarity median of 0.40 when compared to AS and PL datasets (Fig. 15c). Thus, based on the chemoinformatics study performed using PUMA for AOCD database, it was revealed that PL dataset compounds exhibited the highest diversity when compared to AS and AD datasets which is in agreement with the size of PL dataset.

Table 3
Scaffold counts and CSR curve parameters, including AUC and F_{50} for PL, AS and AD datasets.

Datasets	M	N	FNM	NSING	FNSING.M	FNSING.N	AUC	F50
PL	219	47	0.21	30	0.13	0.63	0.83	0.06
AS	91	27	0.29	16	0.17	0.59	0.78	0.07
AD	102	30	0.29	16	0.15	0.53	0.78	0.06

M: Total number of compounds in a dataset, N: Total number of scaffolds in a dataset, NSING: Number of scaffolds with only one compound, AUC: area under CSR curve, F50: fraction of chemotypes required to acquire half of the compounds, PL: Pancreatic Lipase, AS: Appetite Suppressant and AD: Adipogenesis.

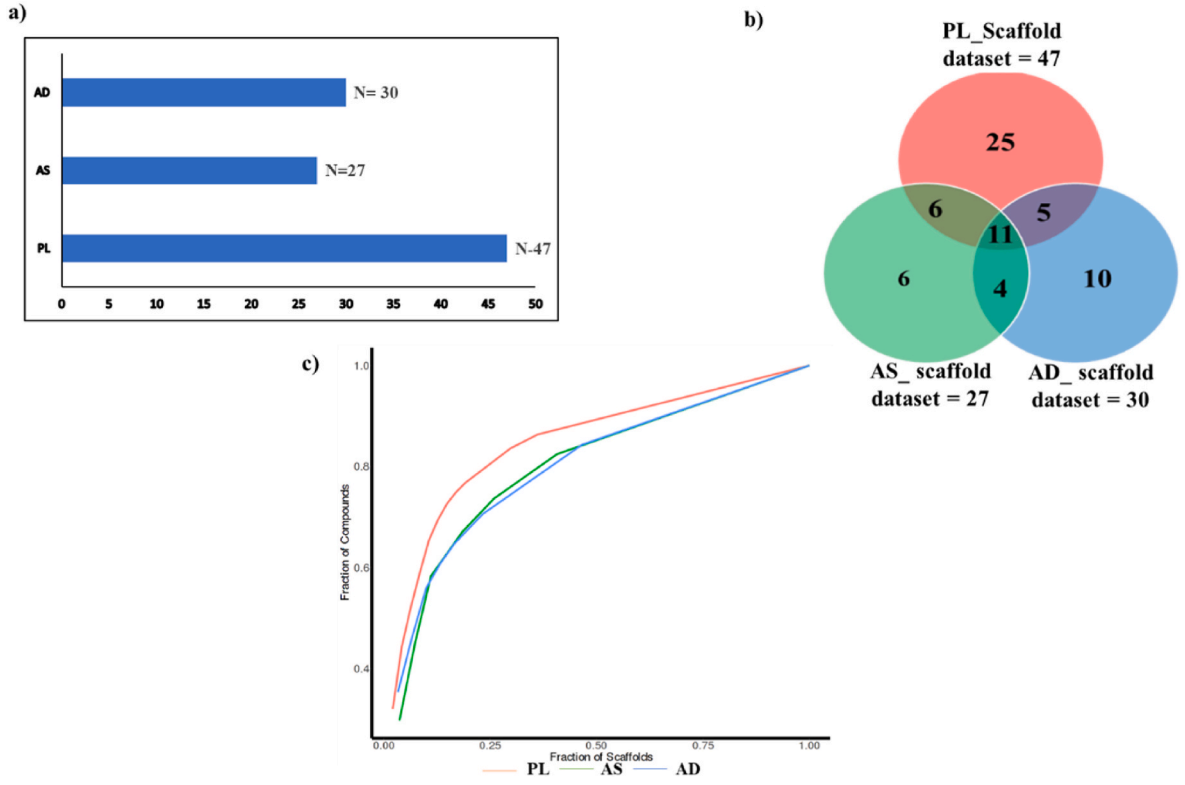

Fig. 11. The number of Scaffolds generated based on Bemis and Murko system **(a)**; Venn diagram describing the common scaffolds among tested datasets **(b)** and CSR curves **(c)**.

Table 4
Scaled Shannon Entropy (SSE) for the most populated scaffold (10–60 scaffolds).

Datasets	SSE10	SSE20	SSE30	SSE40	SSE50	SSE60
PL	0.96	0.96	0.94	0.92	0.92	0.92
AS	0.89	0.87	0.88	NA	NA	NA
AD	0.90	0.88	0.89	0.9	0.91	NA

NA: Not Applicable.

4. Conclusion

Natural plant-based compounds provide a vast pool of inhibition with the ability to be developed into clinical products. Thus, in this AOCD database, we surveyed all possible bio-active compounds with anti-obesity properties targeting pancreatic lipase, suppressing appetite, and adipocyte differentiation. The database delivers an exceptional source of information from various descriptive fields, thereby creating a comprehensive understanding of the natural compounds that target obesity. The AOCD database was divided into three datasets based on the mechanism of action as PL, AS, and AD datasets. Principal component analysis was used for visualization of chemical space based on six molecular properties. Based on the molecular properties, Euclidean distances for datasets were computed to analyze the inter- and intra-dataset distance. It was suggested that the PL dataset exhibited the highest dataset diversity of compounds when compared to AS and AD datasets. Scaffold diversity analysis also revealed that the PL dataset showed the highest scaffolds among other datasets of compounds. The pairwise similarity distribution data computed by PubChem (881-Bits) showed that PL datasets has the highest fingerprint diversity when compared to other datasets of compounds in AOCD database. Thus, the AOCD database will keep expanding with more and more information relevant to molecular interactions with additional chemical descriptors.

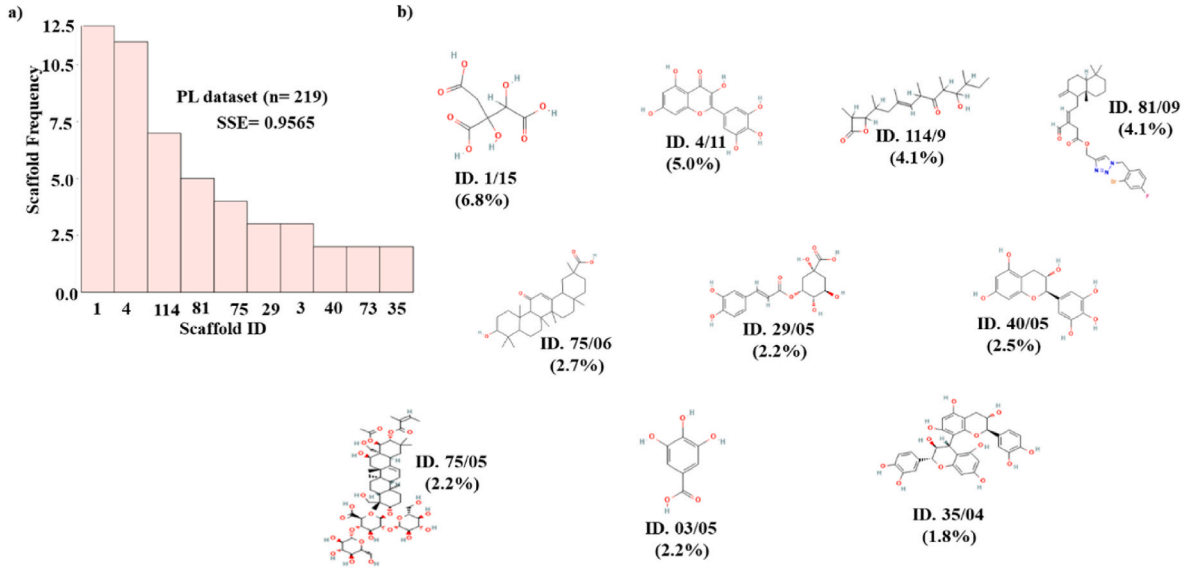

Fig. 12. Scaffold frequency histogram for PL dataset **(a)**; The chemical structure of the most populated 10 scaffolds (SSE10) in PL dataset. The number under each structure indicates the scaffold ID/frequency of the corresponding scaffold in each data set with percentage of scaffolds in each dataset **(b).**

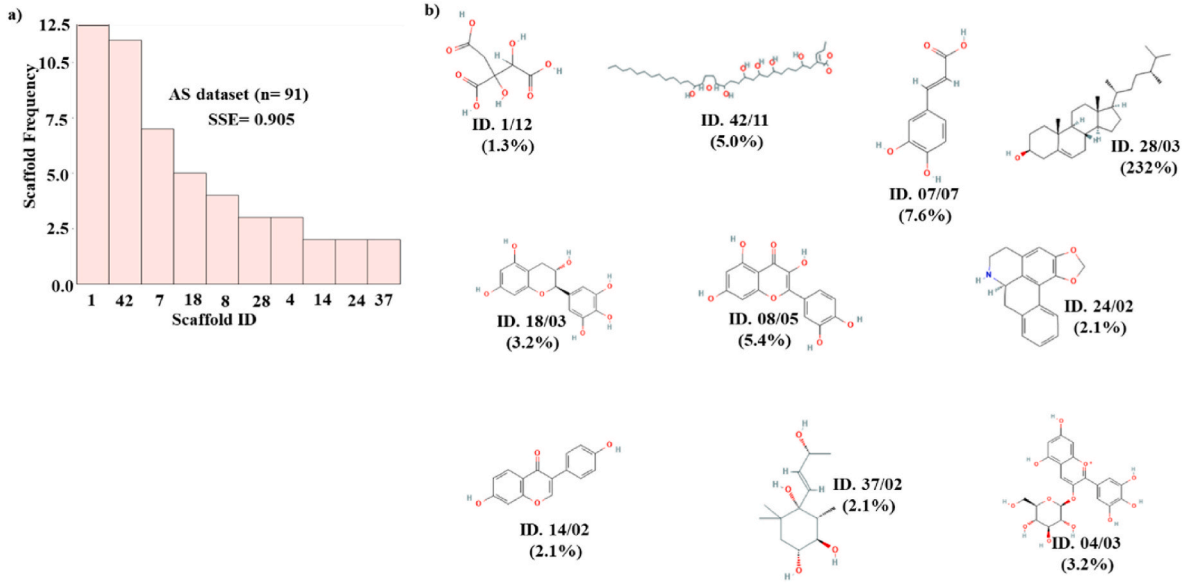

Fig. 13. Scaffold frequency histogram for AS dataset **(a)**; The chemical structure of the most populated 10 scaffolds (SSE10) in AS dataset. The number under each structure indicates the scaffold ID/frequency of the corresponding scaffold in each data set with percentage of scaffolds in each dataset **(b).**

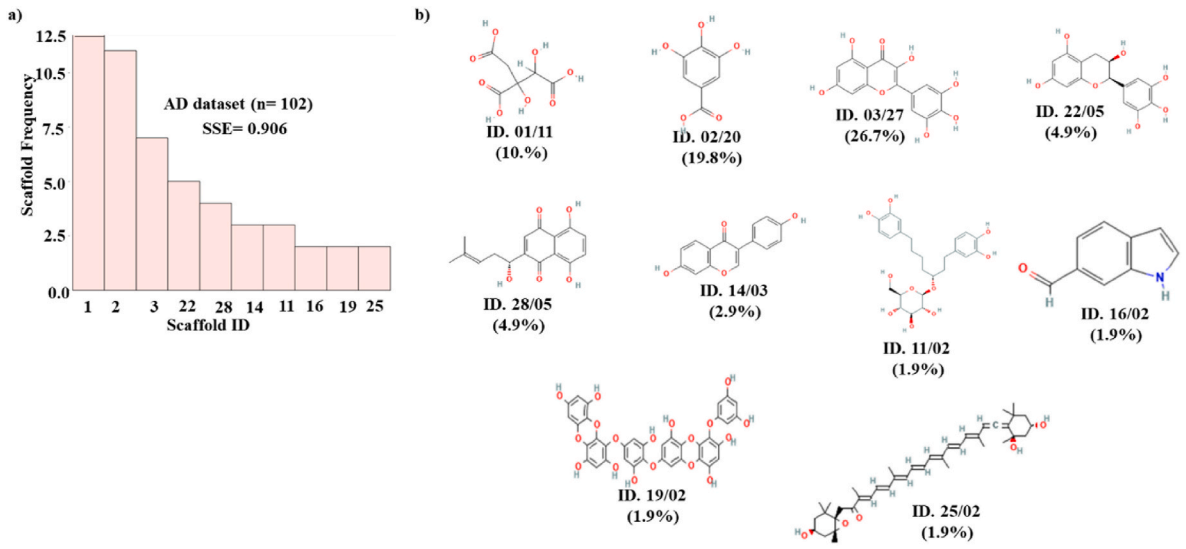

Fig. 14. Scaffold frequency histogram for AD dataset **(a)**; The chemical structure of the most populated 10 scaffolds (SSE10) in AD dataset. The number under each structure indicates the scaffold ID/frequency of the corresponding scaffold in each data set with percentage of scaffolds in each dataset **(b).**

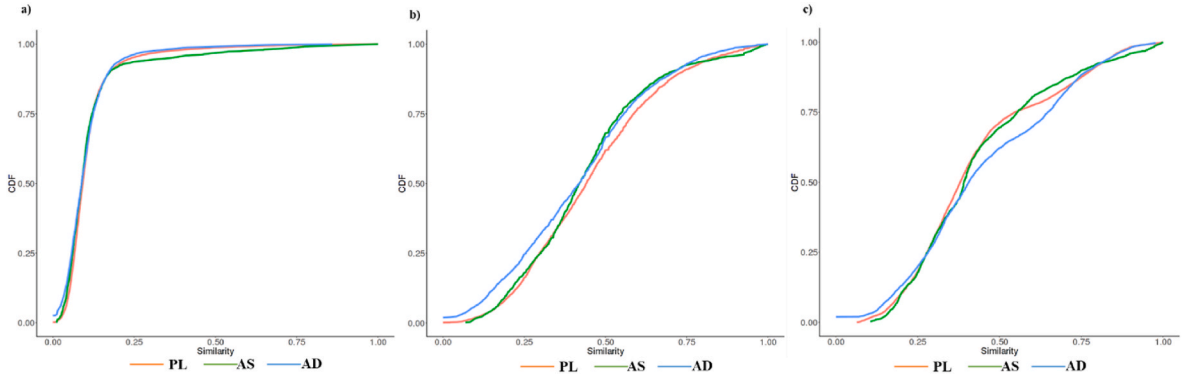

Fig. 15. Cumulative distribution functions (CDFs) of the similarity regarding pairwise values computed using extended connectivity fingerprints with a diameter of 4 **(a)**; MACCS keys **(b)** and PubChem **(c).**

Table 5
Statistical distribution of the datasets pairwise similarity data computed using Tanimoto coefficient, extended connectivity fingerprints with a diameter of 4 (ECFP_4), MACCS (166-Bits) and PubChem (881-bits).

Fingerprint	Dataset	Min	1st Qu[a]	Median	Mean	3rd Qu[b]	Max	Std. dev[c]
ECFP_4	PL	0	0.06	0.09	0.11	0.12	1	0.09
	AS	0	0.06	0.08	0.11	0.12	1	0.12
	AD	0	0.05	0.08	0.10	0.12	1	0.08
MACCS (166-Bits)	PL	0	0.29	0.44	0.45	0.58	1	0.20
	AS	0.06	0.3	0.41	0.44	0.54	1	0.19
	AD	0	0.25	0.41	0.41	0.55	1	0.21
PubChem (881-Bits)	PL	0.05	0.27	0.37	0.43	0.54	1	0.21
	AS	0.06	0.28	0.39	0.43	0.55	1	0.20
	AD	0	0.28	0.40	0.45	0.66	1	0.22

[a] First quarter.
[b] Third quarter.
[c] Standard deviation.

The users can contribute pertinent information to the authors through email. Periodic upgradation of the database happens based on current trends and research. Chemoinformatic analysis pertaining to drug discovery from this database can be constructively utilized for studies such as in silico approach, pharmacophore search, molecular docking, dynamic studies, and so on.

Funding

This work was funded by Sri Ramachandra Institute of Higher Education and Research for providing research support under grant of Founder-Chancellor Shri. N.P.V. Ramasamy Udayar Research Fellowship (U02B190575).

Author contributions

Lavanya Prabhakar: Carried out Research work and wrote the manuscript. Dicky John Davis G: Supervisor of the Research work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Anti-obesity compound database provides comprehensive information on bioactive compounds from natural sources which is available online at https://aocd.swmd.co.in/.

Acknowledgement

The authors acknowledge Sri Ramachandra Institute of Higher Education and Research for providing research support under grant of Founder-Chancellor Shri. N.P.V. Ramasamy Udayar Research Fellowship.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmgm.2023.108414.

References

[1] Z.I. Noor, D. Ahmed, H.M. Rehman, M.T. Qamar, M. Froeyen, S. Ahmad, In vitro antidiabetic, anti-obesity and antioxidant analysis of ocimum basilicum aerial biomass and in silico molecular docking simulations with alpha-amylase and lipase enzymes, Biology 8 (2019) 1–20, https://doi.org/10.3390/biology8040092.
[2] C. Fu, Y. Jiang, J. Guo, Z. Su, Natural products with anti-obesity effects and different mechanisms of action, J. Agric. Food Chem. 64 (2016) 9571–9585, https://doi.org/10.1021/acs.jafc.6b04468.
[3] J.W. Yun, Possible anti-obesity therapeutics from nature - a review, Phytochemistry 71 (2010) 1625–1641, https://doi.org/10.1016/j.phytochem.2010.07.011.
[4] A.L. Harvey, Natural products in drug discovery, Drug Discov. Today 13 (2008) 894–901, https://doi.org/10.1016/j.drudis.2008.07.004.
[5] D.J. Newman, G.M. Cragg, Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019, J. Nat. Prod. 83 (2020) 770–803, https://doi.org/10.1021/acs.jnatprod.9b01285.
[6] J. Shen, X. Xu, F. Cheng, H. Liu, X. Luo, J. Shen J, et al., Virtual screening on natural products for discovering active compounds and target information, Curr. Med. Chem. 10 (21) (2005) 2327–2342, https://doi.org/10.2174/0929867033456729.
[7] M. Lahlou, The success of natural products in drug discovery, Pharmacol. Pharm. 4 (3) (2013) 17–31, https://doi.org/10.4236/pp.2013.43A003.
[8] A.G. Atanasov, S.B. Zotchev, V.M. Dirsch, I.E. Orhan, M. Banach, J.M. Rollinger, et al., Natural products in drug discovery: advances and opportunities, Nat. Rev. Drug Discov. 20 (3) (2021) 200–216, https://doi.org/10.1038/s41573-020-00114-z.
[9] G.D.J. Davis, A.H.R. Vasanthi, QSAR based docking studies of marine algal anticancer compounds as inhibitors of protein kinase B (PKBβ), Eur. J. Pharmaceut. Sci. 76 (2015) 110–118, https://doi.org/10.1016/j.ejps.2015.04.026.
[10] P. Lavanya, G.D.J. Davis, S.R. Singh, A. Murali, Computational analysis of marine algal compounds for obesity managament against pancreatc lipase, J. Biomol. Struct. Dyn. (2022) 1–11, https://doi.org/10.1080/07391102.2022.2074139.
[11] D.J. Newman, Natural products as leads to potential drugs: an old process or the new hope for drug discovery, J. Med. Chem. 51 (2008) 2589–2599, https://doi.org/10.1021/jm0704090.
[12] G.D.J. Davis, A.H.R. Vasanthi, Seaweed metabolite database (SWMD): a database of natural compounds from marine algae Bioinformation, Bioinformation 5 (2011) 361–364, https://doi.org/10.6026/97320630005361.
[13] F.M. Afendi, T. Okada, M. Yamazaki, A. Hirai-Morita, Y. Nakamura, K. Nakamura, et al., KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research, Plant Cell Physiol. 53 (2) (2012) 1–12, https://doi.org/10.1093/pcp/pcr165.
[14] R. Xue, Z. Fang, M. Zhang, Z. Yi, C. Wen, T. Shi, TCMID: traditional Chinese medicine integrative database for herb molecular mechanism analysis, Nucleic Acids Res. 41 (2013) 1089–1095, https://doi.org/10.1093/nar/gks1100.
[15] J. Gu, Y. Gui, L. Chen, G. Yuan, X. Xu, CVDHD: a cardiovascular disease herbal database for drug discovery and network pharmacology, J. Cheminf. 5 (1) (2013) 5–10. http://www.jcheminf.com/content/5/1/51.
[16] S. Pathania, S.M. Ramakrishnan, G. Bagler, Phytochemica: A Platform to Explore Phytochemicals of Medicinal Plants, Database, 2015, pp. 1–8, https://doi.org/10.1093/database/bav075.
[17] I. Ray, A. Bhattacharya, R. K De, OCDD: an obesity and co-morbid disease database, BioData Min. 10 (1) (2017) 1–11, https://doi.org/10.1186/s13040-017-0153-5.
[18] A. Madariaga-Mazón, J.J. Naveja, J.L. Medina-Franco, K.O. Noriega-Colima, K. Martinez-Mayorga, DiaNat-DB: a molecular database of antidiabetic compounds from medicinal plants, RSC Adv. 11 (9) (2021) 5172–5178, https://doi.org/10.1039/D0RA10453A.
[19] M. Sorokina, P. Merseburger, K. Rajan, M.A. Yirik, C. Steinbeck, COCONUT online: collection of open natural products database, J. Cheminf. 13 (1) (2021) 1–13, https://doi.org/10.1186/s13321-020-00478-9.
[20] M. Karthikeyan, B.S. Karthikeyan, R.P. Vivek-Ananth, R.P. Bharath Chand, S. R. Aparna, P. Mangalapandi, A. SamalK, IMPPAT: a curated database of Indian medicinal plants, phytochemistry and therapeutics, Sci. Rep. 8 (2018) 1–17, https://doi.org/10.1038/s41598-018-22631-z.
[21] E.G. Lyakhova, A.I. Kalinovsky, S.A. Kolesnikova, V.E. Vaskovsky, V.A. Stonik, Halogenated diterpenoids from the red alga Laurencia nipponica, Phytochemistry 65 (2004) 2527–2532, https://doi.org/10.1016/j.phytochem.2004.07.005.
[22] K. Martinez-Mayorga, A. Madariaga-Mazon, J.L. Medina-Franco, G. Maggiora, The impact of chemoinformatics on drug discovery in the pharmaceutical industry, Expet Opin. Drug Discov. 15 (2020) 293–306, https://doi.org/10.1080/17460441.2020.1696307.
[23] J.L. Medina-Franco, Interrogating novel areas of chemical space for drug discovery using chemoinformatics, Drug Dev. Res. 73 (2012) 430–438, https://doi.org/10.1002/ddr.21034.
[24] A.H. Al SharieT, El-Elimat, Y.O. Al Zu'bi, A.J. Aleshawi, J.L. Medina-Franco, Chemical space and diversity of seaweed metabolite database (SWMD): a cheminformatics study, J. Mol. Graph. Model. 100 (2020) 1–13, https://doi.org/10.1016/j.jmgm.2020.107702.
[25] A. Ganesan A, The impact of natural products upon modern drug discovery, Curr. Opin. Chem. Biol. 12 (2008) 306–317, https://doi.org/10.1016/j.cbpa.2008.03.016.
[26] B.F. Begam, J.K. Satheesh, A study on cheminformatics and its applications on modern drug discovery, Procedia Eng. 38 (2012) 1264–1275, https://doi.org/10.1016/j.proeng.2012.06.156.
[27] E.S. Maida, E.B. Balaji, S. Divakar, G. Geetha, Ligand based pharmacophore modeling, virtual screening and molecular docking studies to design novel pancreatic lipase inhibitors, Int. J. Pharm. Pharmaceut. Sci. 9 (4) (2017) 1–11. http://www.ncbi.nlm.nih.gov/pubmed/25525429.
[28] M.D. Mashabela, P. Masamba, A.P. Kappo, Metabolomics and chemoinformatics in agricultural biotechnology research: complementary probes in unravelling new metabolites for crop improvement, Biology 11 (8) (2022) 1–23, https://doi.org/10.3390/biology11081156.
[29] F. Csizmadia, JChem: java applets and modules supporting chemical database handling from web browsers, J. Chem. Inf. Comput. Sci. 40 (2000) 323–324, https://doi.org/10.1021/ci9902696.
[30] A. Daina, O. Michielin, V. Zoete SwissADME, A free web tool to evaluate pharmacokinetics , drug- likeness and medicinal chemistry friendliness of small molecules, Nat Publ Gr (2017) 1–13, https://doi.org/10.1038/srep42717.
[31] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev. 46 (2001) 3–26, https://doi.org/10.1016/s0169-409x(00)00129-0.
[32] P. Banerjee, A.O. Eckert, A.K. Schrey, R. Preissner, ProTox-II: a webserver for the prediction of toxicity of chemicals, Nucleic Acids Res. 46 (2018) W257–W263, https://doi.org/10.1093/nar/gky318.
[33] M. Gonzalez-Medina, L. Medina-Franco, Platform for unified molecular analysis : PUMA, J. Chem. Inf. Model. 17 (2017) 1735–1740, https://doi.org/10.1021/acs.jcim.7b00253.
[34] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, Open Babel: an open chemical toolbox, J. Cheminf. 7 (2010) 1–34, https://doi.org/10.1186/1758-2946-3-33.
[35] R. Guha, Chemical informatics functionality in R, J. Stat. Software 18 (2007) 1–16, https://doi.org/10.18637/jss.v018.i05.
[36] M. González-Medina, J.L. Medina-Franco, Chemical diversity of cyanobacterial compounds: a chemoinformatics analysis, ACS Omega 4 (2019) 6229–6237, https://doi.org/10.1021/acsomega.9b00532.
[37] T. Wu, Q. Tang, Z. Gao, Z. Yu, H. Song, X. Zheng, Blueberry and mulberry juice prevent obesity development in C57bl/6 mice, PLoS One 8 (2013) 2–8, https://doi.org/10.1371/journal.pone.0077585.
[38] N. Bitoua, M. Ninomiyaa, T. Tsujitab, H. Okudac H, Screening of lipase inhibitors from marine algae, Lipids 34 (1999) 441–445, https://doi.org/10.1016/B978-0-08-100596-5.21486-8%0A.
[39] C. Heinis, T. Rutherford, S. Freund, G. Winter, Phage-encoded combinatorial chemical libraries based on bicyclic peptides, Nat. Chem. Biol. 5 (2009) 502–507, https://doi.org/10.1038/nchembio.184.
[40] M.N. Al-Muammar, F. Khan F, Obesity: the preventive role of the pomegranate (Punica granatum), Nutrition 28 (2012) 595–604, https://doi.org/10.1016/j.nut.2011.11.013.
[41] N. Karu, R. Reifen, Z. Kerem, Weight gain reduction in mice fed Panax ginseng saponin, a pancreatic lipase inhibitor, J. Agric. Food Chem. 55 (2007) 2824–2828, https://doi.org/10.1021/jf0628025.

[42] Y. Liu, M. Sun, H. Yao, Y. Liu, R. Gao, Herbal Medicine for the Treatment of Obesity : an Overview of Scientific Evidence from 2007 to 2017, Hindawi, 2017, https://doi.org/10.1155/2017/8943059.
[43] S. Eom, M. Lee, E. Lee, Y. Kim, T.H. Kim, Pancreatic lipase inhibitory activity of phlorotannins isolated from Eisenia bicyclis, Phyther. Res. (2012) 10–13, https://doi.org/10.1002/ptr.4694.
[44] C.I. Gamboa-gómez, N.E. Rocha-guzmán, J.A. Gallegos-infante, M.R. Moreno-jiménez, B.D. Vázquez-cabral, R.F. González-laredo, Plants with potential use on obesity and its complications for type 2 diabetes and cardiovascular diseases , where the central and causal compo- and O ' Rahilly, EXCLI J 14 (2015) 809–831, https://doi.org/10.17179/excli2015-186%0A.
[45] J.S. Oh, S.R. Lee, K.T. Hwang, G.E. Ji, The anti-obesity effects of the dietary combination of fermented red ginseng with levan in high fat diet mouse model, Phyther Res 28 (2014) 617–622, https://doi.org/10.1002/ptr.5042.
[46] J. Sung, J. Lee, G. Capsicoside, A furostanol saponin from pepper (Capsicum annuum L.) seeds, suppresses adipogenesis through activation of AMP-activated protein kinase in 3T3-L1 cells, J. Funct.Foods 20 (2016) 148–158, https://doi.org/10.1016/j.jff.2015.10.024.
[47] Q.A. Abdul, R.J. Choi, H.A. Jung, J.S. Choi, Health benefit of fucosterol from marine algae: a review, J. Sci. Food Agric. 96 (2016) 1856–1866, https://doi.org/10.1002/jsfa.7489.
[48] M. Mashmoul, A. Azlan, H. Khaza'Ai, B.N.M. Yusof, S.M. Noor, Saffron: a natural potent antioxidant as a promising anti-obesity drug, Antioxidants 2 (2013) 293–308, https://doi.org/10.3390/antiox2040293.
[49] Z.A. Ratan, M.F. Haidere, Y.H. Hong, S.H. Park, J. Lee, J. Lee, J. Youl, Cho3Pharmacological potential of ginseng and its major component ginsenosides, J. Ginseng Res. 45 (2021) 199–210, https://doi.org/10.1016/j.jgr.2020.02.004.
[50] M. Nakai, Y. Fukui, S. Asami, Y. Toyoda-Ono, T. Iwashita, H. Shibata, Inhibitory effects of oolong tea polyphenols on pancreatic lipase in vitro, J. Agric. Food Chem. 53 (2005) 4593–4598, https://doi.org/10.1021/jf047814+.
[51] F.I. Saldívar-González, M. Valli, A.D. Andricopulo, V. Da Silva Bolzani, J. L. Medina-Franco JL, Chemical space and diversity of the NuBBE database: a chemoinformatic characterization, J. Chem. Inf. Model. 59 (2019) 74–85, https://doi.org/10.1021/acs.jcim.8b00619.
[52] M. Goudarzvand, S. Afraei, S. Yaslianifard, S. Ghiasy, G. Sadri, M. Kalvandi, Hydroxycitric acid ameliorates inflammation and oxidative stress in mouse models of multiple sclerosis, Neural Regen Res 11 (2016) 1610–1616, https://doi.org/10.4103/1673-5374.193240.
[53] M. Ibuki, C. Shoda, Y. Miwa, A. Ishida, K. Tsubota, T. Kurihara, Therapeutic effect of Garcinia cambogia extract and hydroxycitric acid inhibiting hypoxia-inducible factor in a murine model of age-related macular degeneration, Int. J. Mol. Sci. 20 (2019) 1–20, https://doi.org/10.3390/ijms20205049.
[54] M. Tomar, R.P. Rao, P. Dorairaj, A. Koshta, S. Suresh, M. Rafiq, A clinical and computational study on anti-obesity effects of hydroxycitric acid, RSC Adv. 9 (2016), https://doi.org/10.1039/c9ra01345h, 18578–88.