A Chemoinformatic Analysis on Natural Glycosides with Respect to Biological Origin and Structural Class

Yinliang Chen[#], Yi Liu[#], Nianhang Chen, Yuting Jin, Ruofei Yang, Hucheng Yao, De-Xin Kong[*]

State Key Laboratory of Agricultural Microbiology, Agricultural Bioinformatics Key Laboratory of Hubei Province, College of Informatics, Huazhong Agricultural University, Wuhan, P. R. China.

[#]The first two authors contributed equally.

[*]Corresponding author

Email: dxkong@mail.hzau.edu.cn

Tel: +86-18971629378

Fax: +86-27-8728 0877

Supporting Information

1 Materials and Methods

1.1 Data preparation

The Dictionary of Natural Products (CRC Press, v.25.1) (DNP)[1] is a comprehensive database of natural products (NPs) with known biological source and structural class annotations,[2] covering nearly 270,000 NPs. The NPs' chemical structures, along with their assignations of structural type and biological source were extracted from the DNP database. The structures were standardized using Pipeline Pilot (Version 2016),[3] which included keeping the largest fragment, removing the inorganic compounds, and adding hydrogen atoms. Due to incorrect number of valence bonds of nitrogen atoms, 100 molecular structures that were not parseable by RDKit,[4] were filtered out by a Python script. The NPs' origins were then categorized into kingdoms (animals, plants, bacteria and fungi), and further into different phyla or classes.

1.2 Deglycosylation

Due to its structural complexity, there is no consensus on the definition of sugar moiety. Several algorithms or tools, such as SugarBuster[5] and Sugar Removal Utility (SRU),[6] were developed for glycosyl identification. The former uses specific structural rules to identify cyclic sugars, while the latter recognizes and removes both cyclic and linear sugars.

In this study, SRU was used for the identification and removal of sugar moieties in NPs. The parameters were set as follows:[7] to remove both linear and circular sugar moieties; to remove both terminal and non-terminal sugar moieties; to remove the fragments with fewer than five heavy atoms which got disconnected from the molecule after the removal of sugar moieties; the minimum ratio of the exocyclic oxygen atoms of a circular sugar to the atoms in the sugar ring was set to 0.4; all the other parameters were set as the default values.

Compounds with more than one sugar moiety (connected by glycosidic bonds) and sugars themselves (carbohydrates) were considered as natural glycosides. After removing all sugar moieties, the aglycone with the highest number of heavy atoms was

kept for the subsequent physicochemical property and scaffold analysis.

1.3 Calculation of glycosylation ratios

1.3.1 Glycosylation ratio of DNP

After theoretical deglycosylation by SRU, we calculated the glycosylation ratios of the NPs in DNP, as well as those of NPs with different structural types or biological sources. Some NPs can be assigned to multiple structural types or multiple biological sources. They were counted repetitively and independently in each group.

1.3.2 Glycosylation ratios of compounds in ZINC, ChEMBL, DrugBank

The data preprocessing and deglycosylation procedures of commercially available in-stock compounds (ZINC, v15),[8] biologically active compounds (ChEMBL, v31)[9] and approved drugs (DrugBank, v5.1.9)[10] were the same as those of the DNP. After downloading the structure files from the corresponding databases, Pipeline Pilot was employed to standardize the structures. The curation included keeping the largest fragment, removing the inorganic compounds, and adding hydrogen atoms. SRU was used for the identification and removal of sugar moieties for these databases with the same parameter settings as in DNP analysis.

1.4 Analysis of natural glycosides

1.4.1 Sugar types

We counted the occurrences of different types of sugars, i.e., the circular and linear sugars, the terminal and non-terminal sugars.

The identification of cyclic sugars was performed with SMARTS pattern matching, based on the presence or absence of sugar rings. In this study, circular sugars include furanose, pyranose and heptose, which were determined by the number of atoms in the sugar ring.

According to the definition in SRU, a terminal sugar is defined as a glycosidic substructure that can be removed from the original molecule without creating multiple disconnected structures. In contrast, multiple disconnected structures will be generated when non-terminal sugars are removed from the original molecule. By comparing the deglycosylation results of two different methods (removing only terminal sugars and removing all sugars) of SRU, the terminal and non-terminal sugars were identified. The

deglycosylation was performed iteratively. Thus, all the sugar units in an oligosaccharide or polysaccharide are terminal sugars. The determination of terminal and non-terminal moieties heavily depends on an option named “preservation mode”. This option determines whether a substructure that gets disconnected from the molecule by the removal of a sugar moiety is worth keeping or can get removed along with the sugar. With the default parameter setting, a disconnected structure should contain at least 5 heavy atoms.

It should be kept in mind that the classification is assigned to the sugar units instead of the molecules. Terminal and non-terminal sugars can be presented in a molecule concurrently. For this reason, a natural glycoside can contain more than one type of sugar (furanose, pyranose, heptose, or linear sugar). Such molecules were counted repetitively and independently in each group. For a natural glycoside with multiple sugar units, if all the sugar units were the same type (furanose, pyranose, heptose, or linear sugar), the glycoside was considered to “contain only one type of sugar”.

1.4.2 Glycosyl substitution types

Multiple aglycones (including the smaller ones) may be generated after removing the non-terminal sugar moieties of a glycoside by SRU. Then, a substructure matching of the aglycones over the original glycoside can be made. This process was performed successively according to the aglycones’ size (bigger aglycone first). Finally, the structures of the sugar chains were obtained.

The number of sugar units in each sugar chain was counted using SMARTS pattern matching. This way, the substitution types (mono-, oligo- or polysaccharide) in the original glycosides can be detected. In this study, a glycoside can be substituted in 1-7 sites, which were calculated as the number of the generated sugar chains. We analyzed the glycosides with one or two substitution sites, which accounted for 98.73% of all the glycosides.

1.4.3 Glycosidic bonds

Glycosides are composed of sugars or sugar derivatives and aglycones connected through glycosidic bonds. Depending on the atom forming the glycosidic bond,

glycosides can be classified into O-, C-, N-, S-glycosides.[11] SMARTS pattern matching rules were designed to identify glycosidic bonds and explore the distribution of the glycosidic bonds over NPs from different biological sources.

1.5 Distribution of natural glycosides in biological sources or structural types

In order to explore the distribution of natural glycosides in different biological sources, the number and proportion of natural glycosides from animals, plants, bacteria and fungi were calculated according to the annotations in DNP. Then, we categorized these natural glycosides into phyla or classes. Their glycosylation ratios were calculated.

We also counted the glycosylation ratios of NPs of various structural classes/subclasses. The structural classes/subclasses were assigned according to the annotations in DNP. For molecules with more than one biological source or structural class, they were counted repetitively and independently in each group.

1.6 Analysis of glycosyl substitution sites

Flavonoids and terpenoids are two major structural types of NPs. Both of them have high glycosylation ratios. Therefore, their glycosyl substitution profiling was analyzed. According to the annotations in DNP, flavonoids were further classified into anthocyanidins, dihydroflavonols, flavanones, flavans, flavanols and leucoanthocyanidins, flavones and flavonols, isoflavonoids. Due to the structural diversity of terpenoids, there are no classical scaffolds for glycosyl substitution analysis. The most common Murcko scaffolds were generated for the following analysis. Smaller scaffolds with fewer than 10 heavy atoms, such as the benzene ring and the naphthalene, were not suitable for glycosylation site analysis and were thus discarded. Here, we termed them as "simple Murcko scaffold". Then, glycosyl substitution sites in the subclasses of flavonoids and terpenoids were detected by the "*Generate RGroups*" component of Pipeline Pilot. The number of glycosyl substitutions occurred in different substitution sites was counted.

1.7 Physicochemical property analysis

We calculated 19 physicochemical properties of glycosides, aglycones and non-glycosides to investigate the differences in physicochemical properties among them. These properties were molecular weight (MW), hydrogen bond acceptor (HBA),

hydrogen bond donor (HBD), octanol-water partition coefficient (AlogP), topological polar surface area (TPSA), number of rotatable bonds (NumRotatableBonds) (conjugated single bonds were not considered), number of heavy atoms (NumHeavyAtoms), number of aromatic rings (NumAromaticRings), number of aliphatic rings (NumAliphaticRings), number of rings (NumRings), fraction of Csp3 atoms (FractionCsp3), number of nitrogens and oxygens (NOCount), number of NHs or OHs (NHOHCount), number of carbon atoms (NumCAtoms), number of oxygen atoms (NumOAtoms), number of nitrogen atoms (NumNAtoms), number of sulphur atoms (NumSAtoms), number of heteroatoms (NumHeteroatoms) and number of chiral centers (NumChiralCenters). All of the properties were calculated using RDKit, except the NumChiralCenters that was calculated in Molecular Operating Environment (MOE, Version 2019.01).[12] Monosaccharides and oligosaccharides were excluded from this calculation because they do not have an aglycone.

Furthermore, we analyzed the glycosylation ratios of the NPs with different property ranges (four properties: MW, AlogP, HBA and HBD). Principal component analysis (PCA) was performed based on the 19 physicochemical properties to present the distribution of the glycosides, the aglycones and the non-glycosides in the chemical property space.

1.8 Scaffold analysis

1.8.1 Murcko scaffold generation

Murcko scaffold refers to the ring systems and linkers between the ring systems in a molecule, while the side chains of the rings and the linkers are removed.[13] Exocyclic double bonds and the double bond directly attached to the linkers are kept. To explore the relationship between structure and glycosylation level, Murcko scaffolds were generated by RDKit. Only scaffolds with biological origin information were kept for the subsequent analyses. Occurring frequencies and glycosylation ratios were calculated for each scaffold. Then, the structural characteristics of dominant scaffolds and scaffolds with high glycosylation levels were investigated. Lastly, glycosylation rates of unique aglycone scaffolds from animals, plants, bacteria and fungi were compared.

1.8.2 Chemical space visualization

The SAR Map is originally designed for high-throughput screening data analysis. In this map, similar scaffolds lie near to each other or are clustered together. To visualize the distribution of scaffolds from different biological sources and their corresponding glycosylation ratios in chemical space, we generated a SAR Map of the aglycone scaffolds with biological source information using Data Miner.[14] Data Miner uses the K-dissimilarity selection clustering algorithm (also known as OptiSim)[15-17] first to select a diverse and representative subset of the original dataset based on UNITY fingerprints only. Any singletons (compounds that do not have neighbors within a set radius) are represented as points around the edge of the nonlinear mapping (NLM) plot[18] and listed as clusters of one compound. All the parameters were set to default. In the SAR Map, the larger the cluster, the greater the number of scaffolds it contains. We used color and shape to present the glycosylation rates. Triangles represent scaffolds with glycosylation ratios of 0-30%. Squares represent scaffolds with glycosylation ratios of 30%-70%. Stars represent scaffolds with glycosylation ratios of 70%-100%. The color changes from blue to magenta, indicating a low to high glycosylation rate.

1.9 Code availability

The source code used in this analysis is available from https://github.com/ylchen0622/A-Chemoinformatic-Analysis-on-Natural-Glycosides.

Abbreviations

AlogP: octanol-water partition coefficient

DNP: Dictionary of Natural Products

FractionCsp3: fraction of Csp3 atoms

HBA: hydrogen bond acceptor

HBD: hydrogen bond donor

MW: molecular weight

NHOHCount: number of NHs or OHs

NLM: nonlinear mapping

NOCount: number of nitrogens and oxygens

NumAliphaticRings: number of aliphatic rings

NumAromaticRings: number of aromatic rings

NumCAtoms: number of carbon atoms

NumHeavyAtoms: number of heavy atoms

NumHeteroatoms: number of heteroatoms

NumNAtoms: number of nitrogen atoms

NumOAtoms: number of oxygen atoms

NumRings: number of rings

NumRotatableBonds: number of rotatable bonds

NumSAtoms: number of sulphur atoms

PCA: principal component analysis

SAR: structure-activity relationship

SRU: Sugar Removal Utility

TPSA: topological polar surface area

2 Supplementary Figures

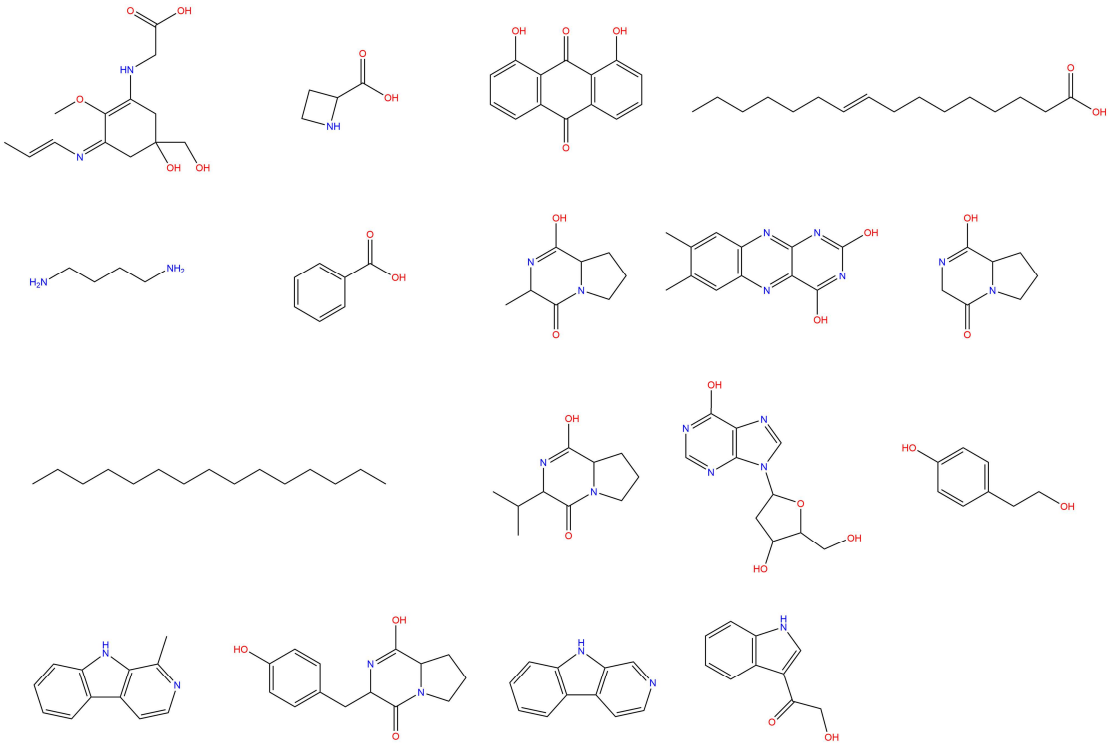

Figure S1. The 17 natural products occurring in all four kingdoms: animals, plants, bacteria and fungi.

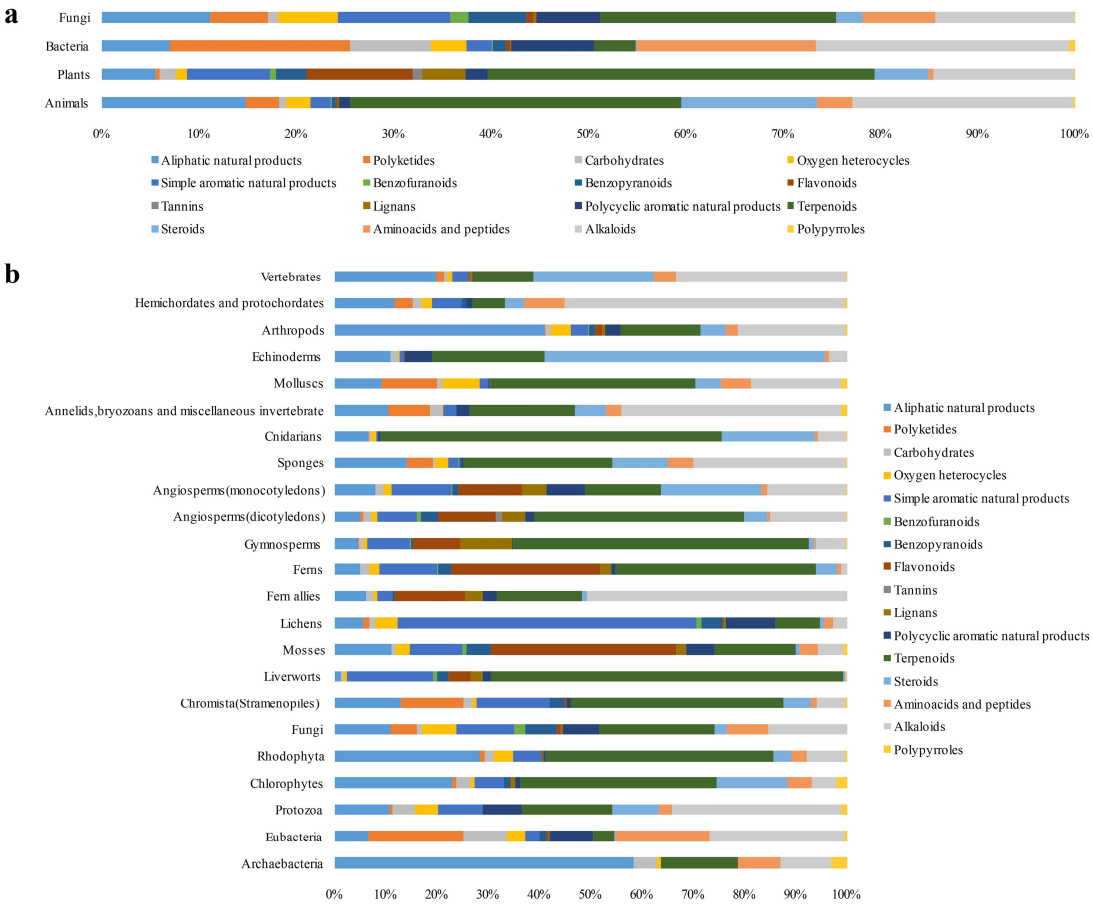

Figure S2. The distribution of natural products of different structural types in: (a) animals, plants, bacteria, fungi; (b) phyla or classes.

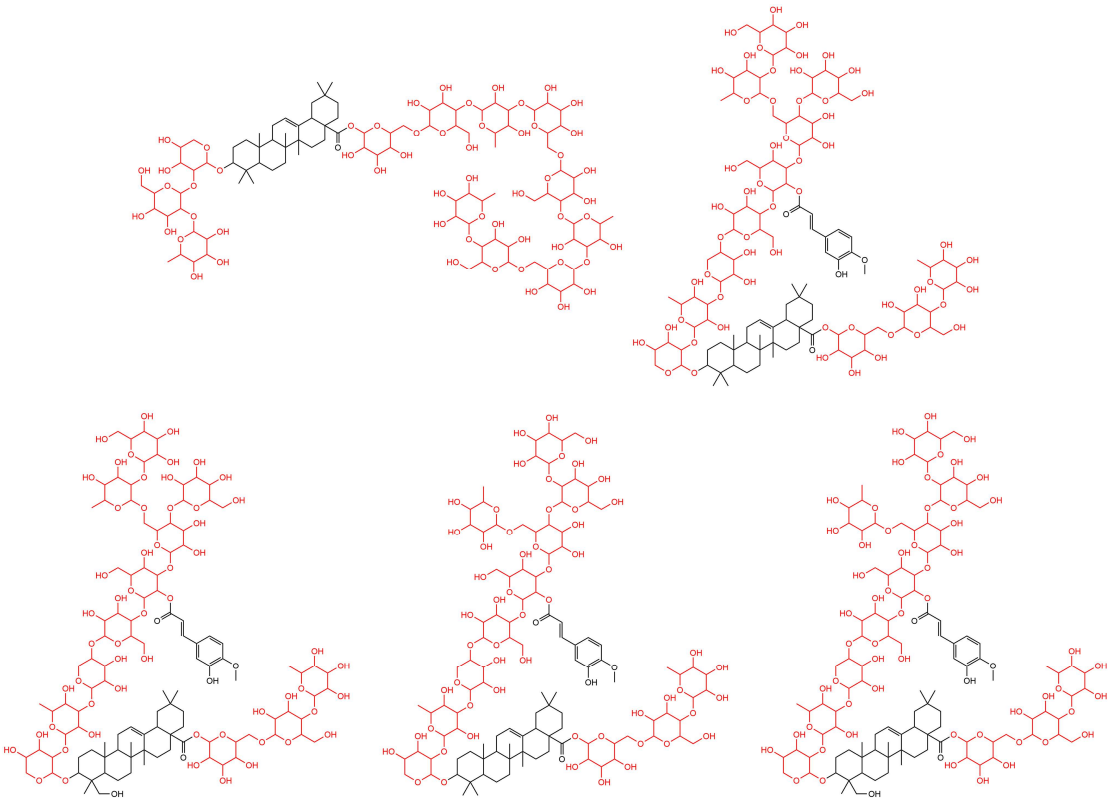

Figure S3. Five natural glycosides with 12 sugar units in DNP.

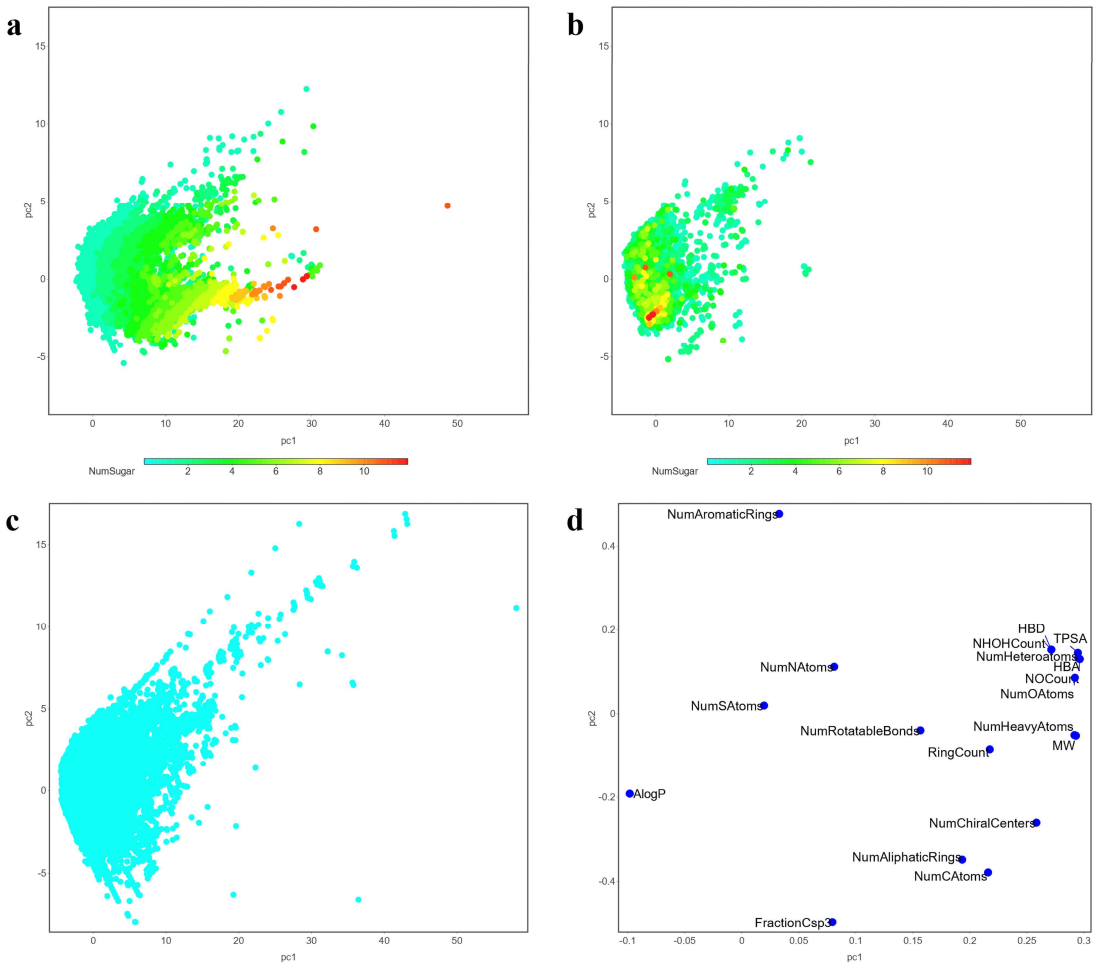

Figure S4. Principal component analysis (PCA) based on 19 physicochemical properties for (a) glycosides, (b) aglycones and (c) non-glycosides. (d) The loadings plot of PC1 and PC2, which explain 55.54% and 15.50% of the total variance, respectively. The dots in panels (a) and (b) were colored according to the number of sugar units in corresponding natural glycosides.

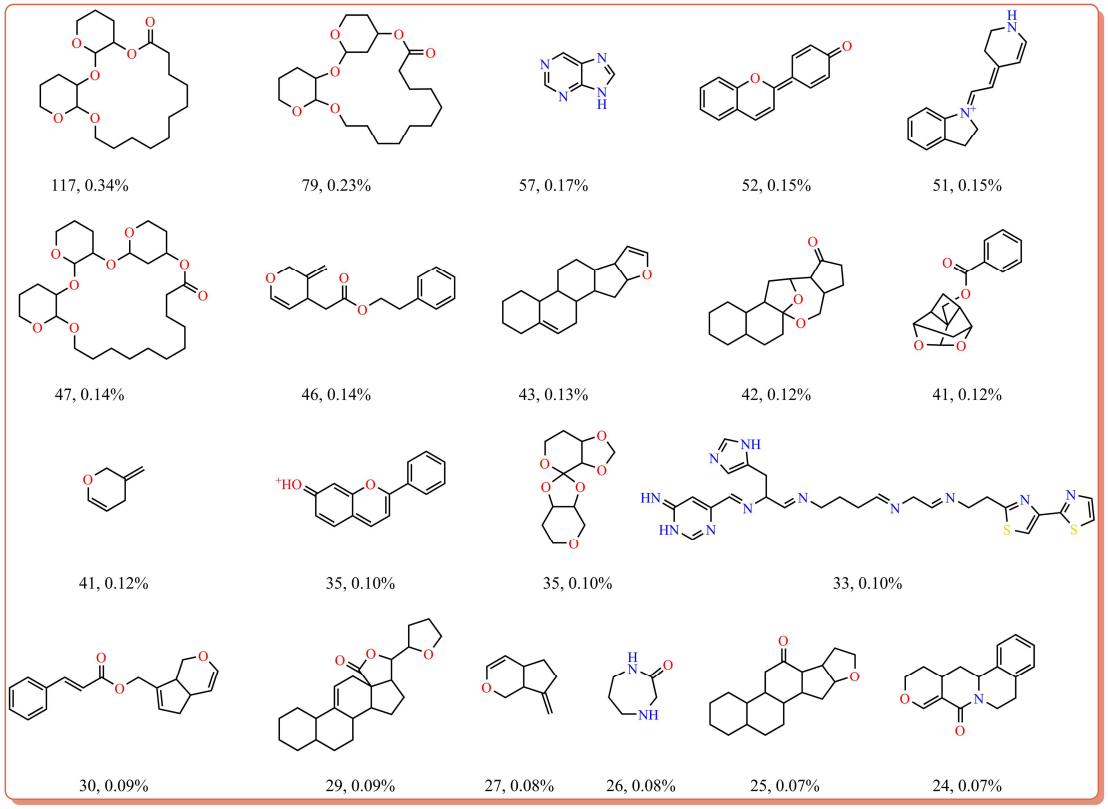

Figure S5. Top 20 aglycone scaffolds (order by frequency) whose glycosylation ratio is 100%. The numerical value is the number of this scaffold. The percentage represents the proportion of the scaffold over all the aglycone scaffolds with biological source information.

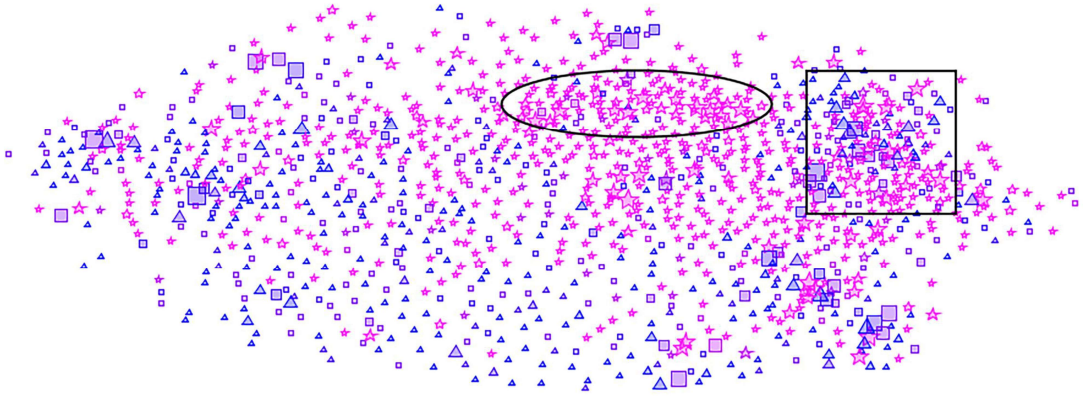

Figure S6. The SAR Map of aglycone scaffolds with biological source information. SAR Map is a non-linear mapping (NLM) of chemical structure in two dimensions space. Triangles represent scaffolds with glycosylation ratios of 0-30%. Squares represent scaffolds with glycosylation ratios of 30%-70%. Stars represent scaffolds with glycosylation ratios of 70%-100%. The color changes from blue to magenta, indicating a low to high glycosylation rate.

3 Supplementary Tables

Table S1. Distribution of the glycosides of different structural types in different phyla or classes

	Animals								Plants									Bacteria			Fungi	
	Sponges	Cnidarians	Annelids, bryozoans and miscellaneous invertebrate phyla	Molluscs	Echinoderms	Arthropods	Hemichordates and protochordates	Vertebrates	Chlorophytes	Rhodophyta	Liverworts	Lichens	Fern allies	Ferns	Gymnosperms	Angiosperms (dicotyledons)	Angiosperms (monocotyledons)	Archaebacteria	Eubacteria	Fungi	Chromista (Stramenopiles)
Aliphatic NPs	137/1293 (10.6%)	25/334 7.49%	25/56 44.64%	5/161 3.11%	117/174 67.24%	6/1055 0.57%	21/171 12.28%	2/198 1.01%	29/142 20.42%	22/564 3.9%	4/22 18.18%	13/50 26.0%	7/45 15.56%	10/85 11.76%	1/201 0.5%	1035/5751 18.0%	146/869 16.8%	16/77 20.78%	183/994 18.41%	159/2063 7.71%	36/262 13.74%
Polyketides	14/464 3.02%	9/16 56.25%	—	3/194 1.55%	—	—	2/51 3.92%	—	—	7/19 36.84%	—	—	—	1/1 100.0%	—	1/590 0.17%	—	—	778/2828 27.51%	38/961 3.95%	52/254 20.47%
Oxygen heterocycles	4/230 1.74%	—	—	—	1/2 50.0%	—	—	—	—	1/76 1.32%	3/13 23.08%	8/38 21.05%	5/6 83.33%	15/33 45.45%	8/33 24.24%	148/1242 11.92%	17/165 10.3%	—	14/561 2.5%	18/1266 1.42%	—
Simple aromatic NPs	—	—	—	—	—	5/90 5.56%	—	—	2/35 5.71%	—	16/290 5.52%	15/515 2.91%	6/22 27.27%	90/194 46.39%	97/356 27.25%	2953/9077 32.53%	306/1283 23.85%	—	37/394 9.39%	79/2117 3.73%	—
Benzofuranoids	—	—	—	—	—	2/4 50.0%	—	—	—	—	—	—	—	—	2/4 50.0%	89/765 11.63%	1/33 3.03%	—	—	22/383 5.74%	—
Benzopyranoids	—	—	—	—	—	2/30 6.67%	—	—	—	—	2/35 5.71%	4/37 10.81%	2/3 66.67%	5/46 10.87%	3/17 17.65%	498/4011 12.42%	43/137 31.39%	—	49/199 24.62%	19/1177 1.61%	—
Flavonoids	—	—	—	—	—	9/33 27.27%	—	—	1/2 50.0%	2/2 100.0%	66/78 84.62%	—	39/101 38.61%	226/498 45.38%	166/398 41.71%	5005/12970 38.59%	591/1363 43.36%	—	17/66 25.76%	10/140 7.14%	1/5 20.0%
Tannins	—	—	—	—	—	1/3 33.33%	—	—	—	—	—	—	—	1/1 100.0%	1/2 50.0%	585/1312 44.59%	4/5 80.0%	—	—	—	—
Lignans	—	—	—	—	—	3/15 20.0%	—	—	—	—	14/41 34.15%	—	9/25 36.0%	22/35 62.86%	93/434 21.43%	1203/5284 22.77%	95/520 18.27%	—	2/19 10.53%	10/67 14.93%	—
Polycyclic aromatic NPs	—	—	—	—	—	21/74 28.38%	—	—	—	—	—	—	—	5/13 38.46%	—	394/2062 19.11%	146/831 17.57%	—	342/1305 26.21%	44/1346 3.27%	—
Terpenoids	117/2659 4.4%	81/3397 2.38%	—	3/707 0.42%	275/349 78.8%	10/402 2.49%	—	—	3/237 1.27%	1/887 0.11%	5/1190 0.42%	—	11/122 9.02%	205/674 30.42%	82/2479 3.31%	11867/47250 25.12%	324/1648 19.66%	8/20 40.0%	126/629 20.03%	206/4217 4.88%	1/848 0.12%
Steroids	78/993 7.85%	74/929 7.97%	—	—	479/875 54.74%	39/130 30.0%	—	11/236 4.66%	31/86 36.05%	1/70 1.43%	—	—	1/7 14.29%	25/70 35.71%	—	2869/5227 54.89%	1704/2164 78.74%	—	2/18 11.11%	12/456 2.63%	—
Aminoacids and peptides	18/460 3.91%	—	—	1/106 0.94%	—	3/61 4.92%	7/116 6.03%	2/43 4.65%	—	1/57 1.75%	—	—	—	—	—	73/579 12.61%	8/121 6.61%	—	162/2799 5.79%	19/1526 1.25%	—
Alkaloids	72/2736 2.63%	18/289 6.23%	2/227 0.88%	3/313 0.96%	6/55 10.91%	25/539 4.64%	20/797 2.51%	2/333 0.6%	1/29 3.45%	3/150 2.0%	—	—	5/373 1.34%	4/19 21.05%	29/257 11.28%	993/17266 5.75%	149/1713 8.7%	3/13 23.08%	400/3969 10.08%	43/2883 1.49%	4/111 3.6%
Polypyrroles	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	6/62 9.68%	4/12 33.33%	2/4 50.0%	13/82 15.85%	—	—

For the groups with more than 50 compounds, glycosylation ratios are indicated by a color gradient, ranging from dark blue (minimum glycosylation ratio) to dark green (maximum glycosylation ratio).

Table S2. Average values for physicochemical properties of glycosides, aglycones and non-glycosides from different biological sources

Physicochemical Property	Animals			Plants			Bacteria			Fungi			All		
	Glycoside	Aglycone	Non-Glycoside	Glycoside	Aglycone	Non-Glycoside	Glycoside	Aglycone	Non-Glycoside	Glycoside	Aglycone	Non-Glycoside	Glycoside	Aglycone	Non-Glycoside
HBA	16.46	5.26	5.22	15.94	5.26	5.58	17.26	9.35	9.65	13.02	5.64	6	16.02	5.55	5.92
HBD	8.39	3.31	2.02	8.41	2.96	1.91	8.57	5.2	4.51	7.23	3.31	2.51	8.4	3.15	2.2
AlogP	1.78	5.19	4.58	-0.37	3.22	3.75	1.53	2.83	3.49	1.92	4.33	3.2	-0.09	3.31	3.78
TPSA	258.08	93.38	80.08	251.31	89.64	83.29	266.26	154.65	153.9	212.06	97.32	95.23	252.12	94.58	90.07
MW	845.55	462.13	420.74	738.97	365.41	394.65	814.01	505.25	518.72	681.38	420.61	382.45	748.84	381.14	407.84
NumRotatableBonds	16.65	9.55	6.95	9.55	2.88	4.55	12.54	6.78	8.11	16.2	10.4	4.8	10.23	3.61	5.18
NumHeavyAtoms	58.52	32.62	29.1	51.86	26.27	28.38	56.94	35.73	36.67	47.68	29.92	27.37	52.5	27.32	29.11
NumAromaticRings	0.26	0.18	0.57	1.12	0.97	1.13	1.14	0.99	1.21	0.6	0.54	0.96	1.07	0.93	1.04
RingCount	5.34	3.02	2.75	5.85	3.5	3.63	4.89	2.97	2.91	3.36	2.05	3.01	5.71	3.41	3.4
FractionCsp3	0.86	0.79	0.66	0.7	0.55	0.56	0.71	0.59	0.52	0.72	0.61	0.53	0.71	0.56	0.57
NOCount	16.46	5.26	5.22	15.94	5.26	5.58	17.26	9.35	9.65	13.02	5.64	6	16.02	5.55	5.92
NHOHCount	8.39	3.31	2.02	8.41	2.96	1.91	8.57	5.2	4.51	7.23	3.31	2.51	8.4	3.15	2.2
NumCAtoms	40.54	26.26	20.84	29.82	15.79	16.75	33.83	21.19	20.65	31.42	21.33	16.15	30.67	16.79	17.6
NumOAtoms	15.76	4.75	4.16	15.66	4.98	5.08	14.74	7.4	6.85	12.43	5.13	5.26	15.55	5.15	5.15
NumNAtoms	0.49	0.42	0.72	0.06	0.06	0.26	1.97	1.57	2.3	0.34	0.34	0.52	0.22	0.19	0.51
NumSAtoms	0.36	0.14	0.1	0.02	0.02	0.02	0.07	0.03	0.12	0.08	0.04	0.06	0.04	0.02	0.04
NumAliphaticRings	5.08	2.83	2.18	4.73	2.53	2.51	3.75	1.97	1.7	2.75	1.51	2.04	4.65	2.49	2.36
NumHeteroatoms	16.87	5.46	5.6	15.97	5.27	5.65	17.57	9.56	10	13.14	5.7	6.14	16.08	5.59	6.06
NumChiralCenters	17.81	7.15	4.60	15.58	5.08	4.35	14.61	6.40	4.88	11.35	4.61	3.75	15.56	5.27	4.38

Table S3. Loadings of the first two components in PCA analysis

Parameter	PC1	PC2
NumChiralCenters	0.2584	-0.2605
HBA	0.2964	0.1300
HBD	0.2715	0.1524
AlogP	-0.0981	-0.1915
TPSA	0.2948	0.1450
NumRotatableBonds	0.1567	-0.0394
NumHeavyAtoms	0.2917	-0.0506
MW	0.2931	-0.0519
NumAromaticRings	0.0329	0.4768
RingCount	0.2175	-0.0844
FractionCsp3	0.0795	-0.5003
NOCount	0.2964	0.1300
NHOHCount	0.2715	0.1524
NumCAtoms	0.2160	-0.3787
NumOAtoms	0.2921	0.0855
NumNAtoms	0.0811	0.1118
NumSAtoms	0.0196	0.0197
NumAliphaticRings	0.1933	-0.3483
NumHeteroatoms	0.2951	0.1332

Table S4. The numbers of aglycone Murcko scaffolds from different biological sources

Biological Source	Total[a]	Unique[b]	Proportion
Animals	172	98	56.98%
Plants	2386	2266	94.97%
Bacteria	491	427	86.97%
Fungi	183	100	54.64%

[a]Total number of aglycone scaffolds without reduplicated structures from animals, plants, bacteria and fungi.

[b]The number of unique aglycone scaffolds that appear in only one biological source.

References

1. Dictionary of Natural Products (DNP), http://dnp.chemnetbase.com, (accessed Nov 30, 2022).
2. Y. Chen, C. de Bruyn Kops and J. Kirchmair, *J. Chem. Inf. Model.*, 2017, **57**, 2099-2111.
3. Pipeline Pilot 2016, http://accelrys.com/, (accessed February 28th, 2023).
4. RDKit: Open-Source Cheminformatics Software, https://www.rdkit.org, (accessed Nov 30, 2022).
5. Y. Chen, M. Garcia de Lomana, N. O. Friedrich and J. Kirchmair, *J. Chem. Inf. Model.*, 2018, **58**, 1518-1532.
6. J. Schaub, A. Zielesny, C. Steinbeck and M. Sorokina, *J. Cheminform.*, 2020, **12**, 1-20.
7. J. Schaub, A. Zielesny, C. Steinbeck and M. Sorokina, *Biomolecules*, 2021, **11**, 486.
8. T. Sterling and J. J. Irwin, *J. Chem. Inf. Model.*, 2015, **55**, 2324-2337.
9. A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit and A. R. Leach, *Nucleic Acids Res.*, 2017, **45**, D945-D954.
10. D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox and M. Wilson, *Nucleic Acids Res.*, 2018, **46**, D1074-d1082.
11. W. Schwab, T. Fischer and M. Wüst, *Eng. Life Sci.*, 2015, **15**, 376-386.
12. Molecular Operating Environment (MOE), 2019.01; Chemical Computing Group ULC, 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2019.
13. G. W. Bemis and M. A. Murcko, *J. Med. Chem.*, 1996, **39**, 2887-2893.
14. DataMiner 1.6, http://www.tripos.com/, (accessed Nov 30, 2022).
15. R. D. Brown and Y. C. Martin, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 572-584.

16. R. D. Clark, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 1181-1188.

17. R. D. Clark, D. E. Patterson, F. Soltanshahi, J. F. Blake and J. B. Matthew, *J. Mol. Graph. Model.*, 2000, **18**, 404-411.

18. J. W. Sammon, *IEEE Trans. Comput.*, 1969, **18**, 401–409.