

Supporting Information

A cheminformatic insight into the differences between terrestrial and marine originated natural products

Jun Shang^{a,b,c}, Ben Hu^b, Junmei Wang^d, Feng Zhu^a, Yu Kang^a, Dan Li^a, Huiyong Sun^a, De-Xin Kong^{b,*},
Tingjun Hou^{a,c,*}

^aCollege of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China.

^bState Key Laboratory of Agricultural Microbiology and Agricultural Bioinformatics Key Laboratory of Hubei Province, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China

^cState Key Lab of CAD&CG, Zhejiang University, Hangzhou, Zhejiang 310058, China.

^dDepartment of Pharmaceutical Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, United States

Table S1. Information of the 60 physicochemical properties

Descriptors ^a	Description	Statistics					
		Mean	StdDev	Min	Max	Median	Mode
MinEnergy	Minimized energy	36.03	82.92	-17.58	3714.28	22.33	0
		46.88	154.34	-38.24	10896.30	22.82	-1.36
AlogP	The Ghose and Crippen octanol-water partition coefficient	4.07	3.39	-8.81	53.47	3.692	4.37
		3.14	2.84	-19.90	41.42	2.856	2.45
logD	Octanol-water partition coefficient (pH=7.4)	3.68	3.51	-36.32	53.47	3.39	3.81
		2.76	2.97	-22.28	41.42	2.575	2.67
logS	Molecular solubility	-5.89	3.89	-60.29	5.45	-5.315	-4.83
		-4.68	3.15	-45.81	15.74	-4.086	-3.88
MW	Molecular weight	420.70	221.30	32.04	5033.75	379.41	304.47
		396.06	199.32	32.04	3748.60	357.36	264.32
MM	Molecular mass	420.21	221.07	32.03	5030.73	378.28	304.24
		395.77	199.19	32.03	3746.36	357.11	264.14
N_a	Number of heavy (non-hydrogen) atoms	29.06	15.50	2	352	26	24
		28.43	14.22	2	268	26	24
N_b	Number of bonds between heavy atoms	30.71	16.52	1	351	28	24
		30.71	15.63	1	299	28	27
N_H	Number of hydrogens, both implicit and explicit	32.77	20.82	0	376	30	32
		28.88	16.11	0	288	26	22
N_{expH}	Number of explicit hydrogens	32.77	20.82	0	376	30	32
		28.88	16.11	0	288	26	22

N_{expA}	Number of explicit atoms	61.83	35.04	2	728	56	54
		57.32	28.67	5	496	52	39
N_{expB}	Number of explicit bonds	63.48	35.77	1	727	57	56
		59.59	29.79	4	504	54	41
N_{pA}	Number of atoms with a positive charge	1.90E-02	0.16	0	6	0	0
		1.08E-02	0.11	0	3	0	0
N_{nA}	Number of atoms with a negative charge	9.44E-03	0.10	0	3	0	0
		3.82E-03	8.23E-02	0	5	0	0
N_{RB}	Number of bonds in a ring	14.82	10.89	0	171	15	0
		17.20	10.98	0	204	17	6
N_{rotB}	Number of rotatable bonds	7.60	8.52	0	152	5	1
		5.61	6.20	0	120	4	2
N_{aroB}	Number of bonds in aromatic ring systems	4.35	6.55	0	120	0	0
		5.84	7.44	0	120	5	0
N_{bB}	Number of bonds in bridgehead ring systems	1.86	6.53	0	102	0	0
		2.28	6.45	0	88	0	0
N_{R}	Number of base rings	2.64	1.94	0	32	2	2
		3.28	2.09	0	32	3	3
N_{aroR}	Number of base rings that are aromatic	0.77	1.16	0	20	0	0
		1.01	1.28	0	20	1	0
N_{Rasb}	Number of ring assemblies	1.28	0.96	0	20	1	1
		1.44	1.03	0	26	1	1
N_{R3}	Number of rings of size 3	8.88E-02	0.32	0	3	0	0

		8.48E-02	0.31	0	6	0	0
N_{R4}	Number of rings of size 4	7.56E-03	8.73E-02	0	2	0	0
		1.33E-02	0.12	0	5	0	0
		0.66	0.88	0	10	0	0
N_{R5}	Number of rings of size 5	0.71	0.92	0	16	0	0
		1.66	1.53	0	28	1	1
N_{R6}	Number of rings of size 6	2.30	1.78	0	25	2	3
		4.99E-02	0.26	0	5	0	0
N_{R7}	Number of rings of size 7	6.91E-02	0.27	0	4	0	0
		1.37E-02	0.13	0	3	0	0
N_{R8}	Number of rings of size 8	1.14E-02	0.11	0	2	0	0
		0.16	0.38	0	4	0	0
N_{R9+}	Number of rings of size 9 or bigger	8.40E-02	0.32	0	8	0	0
		40.46	24.39	0	518	37	38
N_c	Number of chains	36.60	19.52	2	320	33	28
		11.15	7.47	0	108	11	1
N_{Casb}	Number of chain assemblies	12.28	7.22	1	124	12	12
		4.92E-03	0.14	0	7	0	0
N_{steA}	Number of stereo atoms	1.18E-02	0.23	0	15	0	0
		1.78	2.04	0	59	1	0
N_{steB}	Number of stereo bonds	1.36	1.65	0	37	1	0
		5.19	3.84	0	63	4	3
N_{douB}	Number of double bonds	5.68	4.22	0	84	5	4

N_{triB}	Number of triple bonds	4.78E-02	0.33	0	6	0	0
		3.09E-02	0.25	0	7	0	0
N_{Hacc}	Number of heteroatoms with one or more lone pairs	5.52	4.75	0	131	4	4
		6.04	4.76	0	104	5	4
N_{Hdon}	Number of heteroatoms with one or more attached Hydrogen atoms	2.35	2.94	0	73	2	1
		2.37	2.83	0	57	2	1
N_{HaccL}	Number of hydrogen bond acceptor defined by Lipinski equal to the number of N Plus O Count ($N_{\text{N+O}}$)	5.85	4.98	0	132	5	4
		6.23	4.89	0	104	5	4
N_{HdonL}	Number of hydrogen bond donor defined by Lipinski close to the number of H donors (N_{Hdon})	2.42	3.02	0	73	2	1
		2.41	2.91	0	57	2	1
V_{M}	Molecular volume	294.11	155.99	25.38	3535.64	263.76	238.04
		273.89	133.80	25.03	2261.05	245.93	180.76
SA	Molecular surface area	419.90	222.05	43.13	5519.49	375.31	253.17
		391.93	191.89	49.49	3445.90	350.06	340.06
PSA	Molecular polar surface area	92.87	81.87	0	2250.36	72.83	20.23
		96.39	80.06	0	1743.71	79.15	46.53
$f\text{PSA}$	Molecular fractional polar surface area	0.22	0.13	0	1	0.20	0
		0.25	0.12	0	1	0.23	0
SASA	Molecular solvent accessible surface area	649.02	284.33	146.33	6554.03	592.63	559.08
		602.61	237.36	178.16	4430.80	551.05	455.30
PSASA	Molecular polar solvent accessible surface area	152.04	124.35	0	3286.92	122.57	52.15
		156.27	123.48	0	2739.85	128.95	85.02
$f\text{PSASA}$	Molecular fractional polar solvent	0.23	0.13	0	0.945	0.213	0

	accessible surface area	0.26	0.13	0	0.966	0.238	0
V_{SA}	Molecular solvent accessible volume	565.40	242.31	129.38	5642.69	518.31	476.19
		523.66	205.32	155.27	3866.15	480.57	405.41
N_{N+O}	Count of nitrogen plus oxygen	5.85	4.98	0	132	5	4
		6.23	4.89	0	104	5	4
CHBA	Count of hydrogen bond acceptor using predefined substructure queries	3.12	2.87	0	59	2	2
		3.61	3.07	0	56	3	2
CHBD	Count of hydrogen bond donor using predefined substructure queries	1.86	2.71	0	65	1	0
		1.88	2.62	0	57	1	0
Corg	Organic count	29.06	15.51	2	352	26	24
		28.43	14.22	2	268	26	24
C_H	Hydrogen count	32.77	20.82	0	376	30	32
		28.88	16.11	0	288	26	22
C_C	Carbon count	22.82	11.95	0	219	21	20
		22.12	10.54	0	176	20	20
C_N	Nitrogen count	0.99	2.04	0	60	0	0
		0.60	1.64	0	48	0	0
C_O	Oxygen count	4.86	4.17	0	117	4	2
		5.63	4.36	0	104	5	4
C_F	Florine count	3.70E-03	0.12	0	12	0	0
		2.93E-03	0.12	0	16	0	0
C_P	Phosphorus count	8.47E-03	0.10	0	3	0	0
		5.63E-03	0.10	0	6	0	0

C_S	Sulphur count	0.11	0.53	0	20	0	0
		4.68E-02	0.32	0	8	0	0
C_{Cl}	Chlorine count	8.79E-02	0.45	0	11	0	0
		2.39E-02	0.22	0	10	0	0
C_{Br}	Bromine count	0.17	0.66	0	8	0	0
		2.39E-03	7.09E-02	0	8	0	0
C_I	Iodine count	8.56E-03	0.13	0	4	0	0
		2.27E-04	1.70E-02	0	2	0	0

^aIn every property, the first row is the property of the molecules in MNPs, and the second row is that in TNPs.

Table S2. Statistics of numbers (N) and percentages (P) of the compounds satisfying different numbers of Ro5 rules^a

Databases	Statistics of Ro5							
	Ro5_4_N	Ro5_3_N	Ro5_2_N	Ro5_1_N	Ro5_0_N	Total number	Ro5_4P	Ro5_3P
CMC	6206	1014	527	407	8	8162	76.04%	12.42%
CNPD	32472	9474	5464	4603	429	52442	61.92%	18.07%
DMNP_origin	18285	8121	4954	3693	830	35883	50.96%	22.63%
MNPs	18161	7906	4296	1868	706	32937	55.14%	24.00%
DNP_origin	86618	25193	17404	20643	1751	151609	57.13%	16.62%
TNPs	87010	23676	13446	6440	1499	132071	65.88%	17.93%
MDDR	84561	28790	18584	3859	104	135898	62.22%	21.19%
TCMCD	33554	11561	7893	9743	515	63266	53.04%	18.27%
TCMD	12084	3859	2831	3884	139	22797	53.01%	16.93%

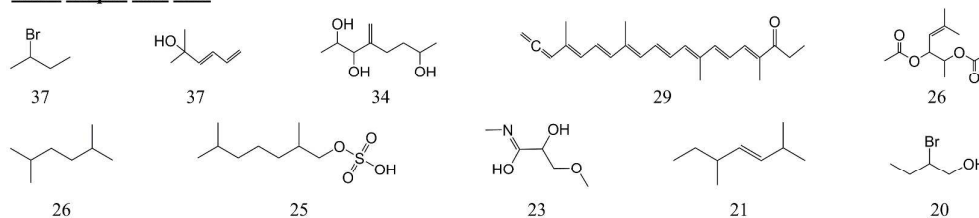
	a1_N	b1_N	c1_N	d1_N	a1_P	b1_P	c1_P	d1_P
CMC	7361	7006	7662	7298	90.19%	85.84%	93.87%	89.41%
CNPD	44315	40907	45496	43123	84.50%	78.00%	86.75%	82.23%
DMNP_origin	29487	25496	30658	25463	82.18%	71.05%	85.44%	70.96%
MNPs	29060	24915	30132	22715	88.23%	75.64%	91.48%	68.96%
DNP_origin	119267	107749	124022	126464	78.67%	71.07%	81.80%	83.41%
TNPs	118135	105359	121862	107044	89.45%	79.77%	92.27%	81.05%
MDDR	123731	103543	131677	106690	91.05%	76.19%	96.89%	78.51%
TCMCD	48668	43390	50438	51932	76.93%	68.58%	79.72%	82.09%

TCMD	17214	15428	17681	19136	75.51%	67.68%	77.56%	83.94%
------	-------	-------	-------	-------	--------	--------	--------	--------

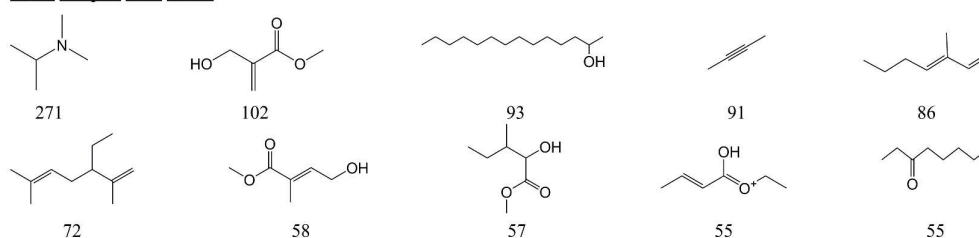
	Ro5_3a0_N	Ro5_3b0_N	Ro5_3c0_N	Ro5_3d0_N	Ro5_3a0_P	Ro5_3b0_P	Ro5_3c0_P	Ro5_3d0_P
CMC	79	272	66	597	7.79%	26.82%	6.51%	58.88%
CNPD	341	2069	873	6191	3.60%	21.84%	9.21%	65.35%
DMNP_origin	191	1381	581	5968	2.35%	17.01%	7.15%	73.49%
MNPs	130	1341	440	5995	1.64%	16.96%	5.57%	75.83%
DNP_origin	1028	6570	2925	14670	4.08%	26.08%	11.61%	58.23%
TNPs	618	6249	2054	14755	2.61%	26.39%	8.68%	62.32%
MDDR	1493	10362	519	16416	5.19%	35.99%	1.80%	57.02%
TCMCD	408	3038	1136	6979	3.53%	26.28%	9.83%	60.37%
TCMD	145	1017	479	2218	3.76%	26.35%	12.41%	57.48%

^aRo5_4/3/2/1/0 represents the number of molecules satisfying 4/3/2/1/0 Lipinski's Rule-of-5; a1/b1/c1/d1 represents the number of molecules satisfying the a/b/c/d rule, which represents number of hydrogen bond acceptors (NHA) ≤ 10 , molecular weight ≤ 500 , number of hydrogen bond donors (NHD) ≤ 5 or calculated octanol-water partition coefficient (ClogP) ≤ 5 , respectively; Ro5_3a0/b0/c0/d0 represents the number of molecules from Ro5_3 that do not satisfy the a/b/c/d rule.

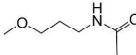
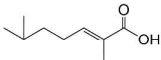
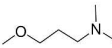
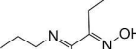
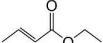
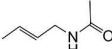
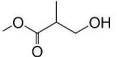
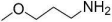
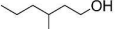
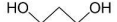
MNPs unique: total 7919



TNPs unique: total 18575



MNPs/TNPs common: total 2269

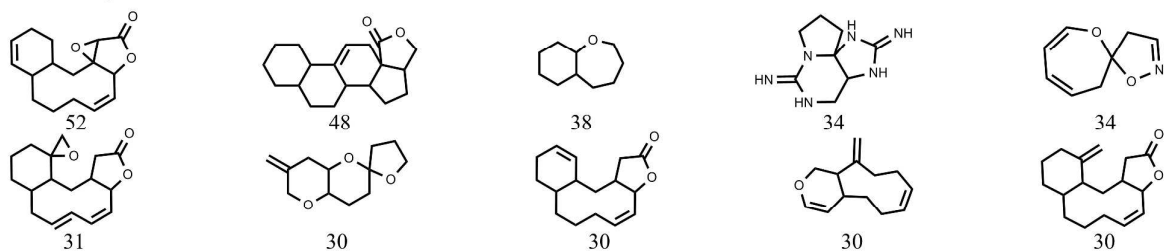
MNPs/TNPs equal*		MNPs/TNPs descend		MNPs/TNPs ascend	
	54.111%:54.293%		60:1, 265		1:148, 1/34
	9.158%:9.095%		40:1, 177		2:196, 1/22
	1.051%:1.041%		32:1, 141		29:2311, 1/18
	0.655%:0.669%		28:1, 124		6:409, 1/15
	0.271%:0.273%		55:2, 121		3:174, 1/13
	0.170%:0.187%		64:3, 94		1:54, 1/12
	0.081%:0.079%		4585:238, 85		1:54, 1/12
	0.050%:0.054%		19:1, 84		1:52, 1/12
	0.044%:0.047%		17:1, 75		1:52, 1/12
	0.022%:0.023%		17:1, 75		2:102, 1/12

* The total frequencies of chain assemblies in MNPs and TNPs are 367156 and 1621518, respectively.

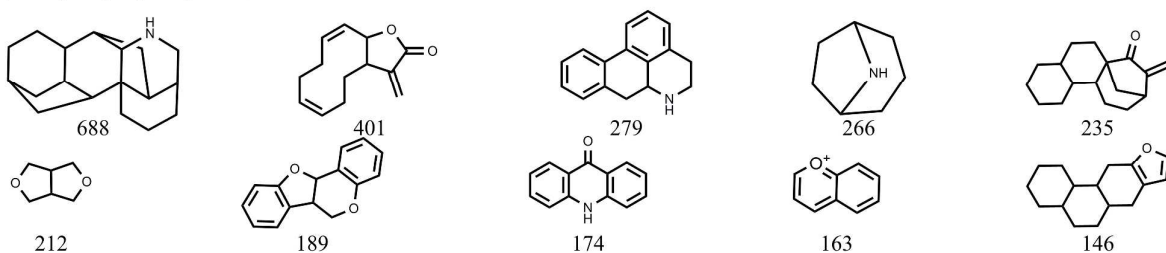
Figure S1. Unique and common chain assemblies of MNPs and TNPs. In the unique part, the number for each fragment is its frequency. In the common part, “equal” means that the fragments in this column share almost the same percentages in TNPs and MNPs, and the two numbers are their percentages in MNPs and TNPs, respectively; “descend” means that the fragments in this list appear more frequently in MNPs, and their frequencies of MNPs *versus* TNPs are shown before the comma and their frequency percentage ratios of MNPs *versus* TNPs are shown after the comma; “ascend” means that the fragments in this list appear

more frequently in TNPs, and their frequencies of MNPs *versus* TNPs are shown before the comma and their frequency percentage ratios of MNPs *versus* TNPs are shown after the comma.

MNPs unique: total 3456

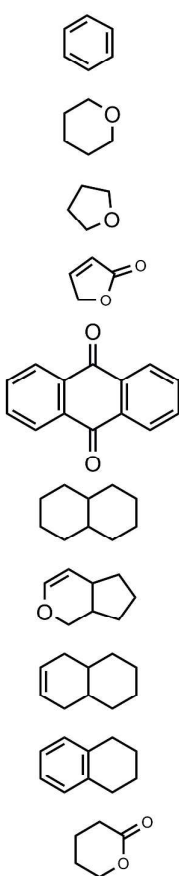


TNPs unique: total 15697



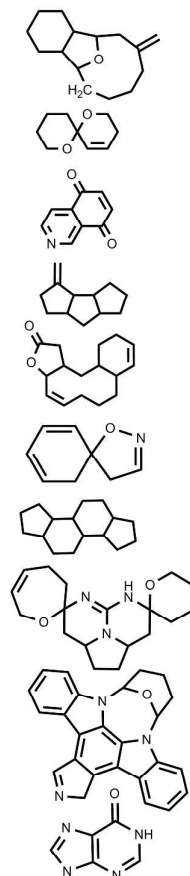
MNPs/TNPs common: total 1412

MNPs/TNPs equal*



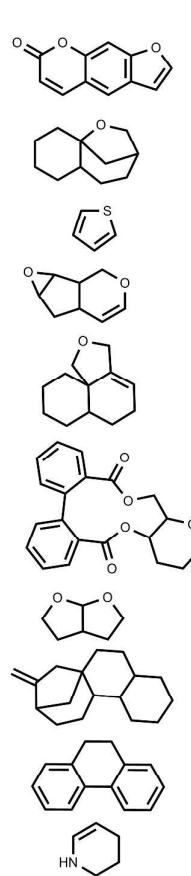
25.578%:27.245%
4.040%:3.899%
1.601%:1.621%
1.109%:1.174%
0.478%:0.507%
0.310%:0.294%
0.203%:0.196%
0.184%:0.168%
0.118%:0.128%
0.130%:0.127%

MNPs/TNPs descend



32:1, 144
29:1, 131
27:1, 122
27:1, 122
26:1, 117
95:4, 107
21:1, 95
18:1, 81
34:2, 77
17:1, 77

MNPs/TNPs ascend

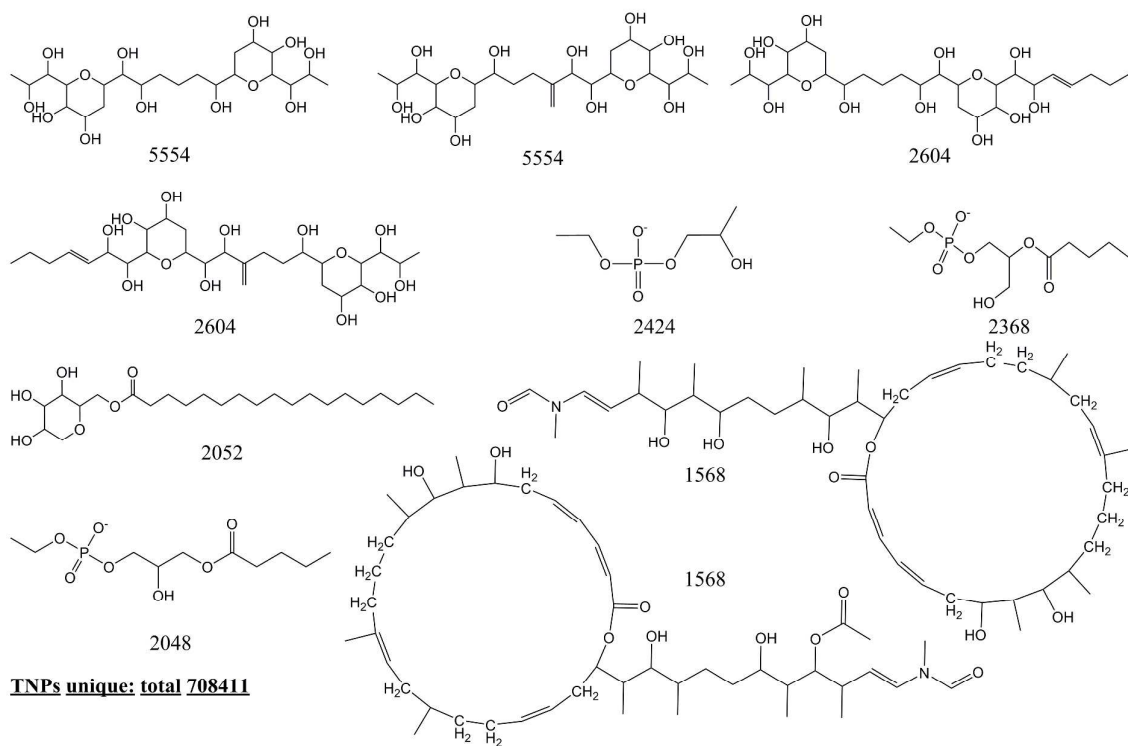


1:275, 1/61
2:482, 1/54
2:370, 1/41
1:134, 1/30
1:127, 1/28
2:200, 1/22
1:91, 1/20
3:269, 1/20
3:240, 1/18
1:80, 1/18

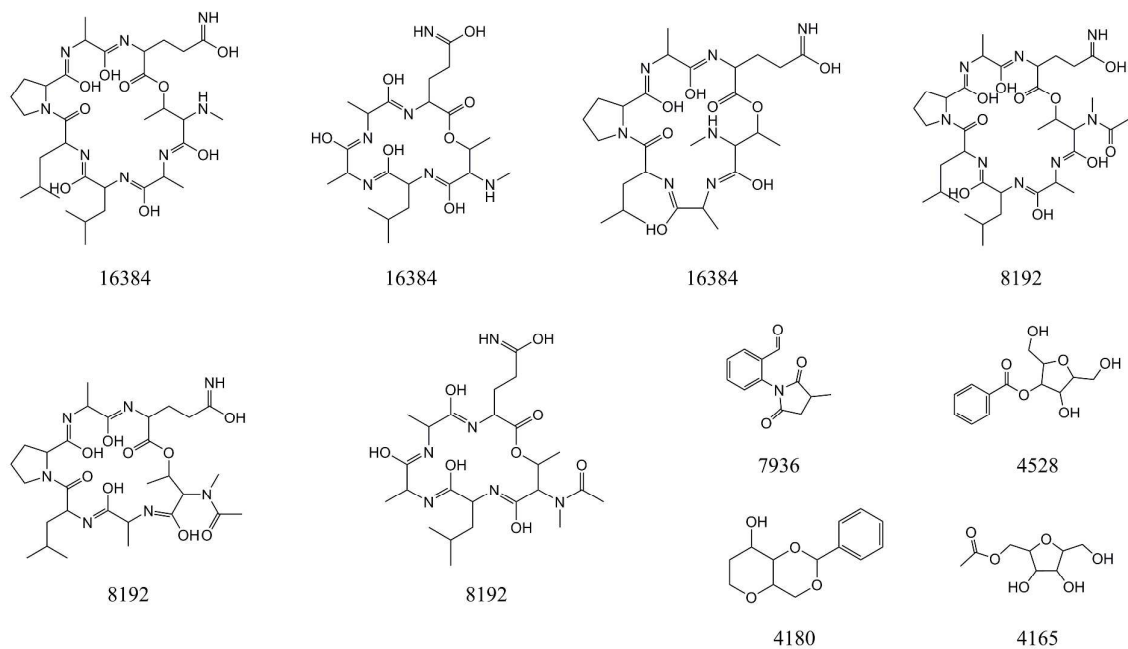
* The total frequencies of ring assemblies in MNPs and TNPs are 42279 and 190454, respectively.

Figure S2. Unique and common ring assemblies of MNPs and TNPs.

MNPs unique: total 131872



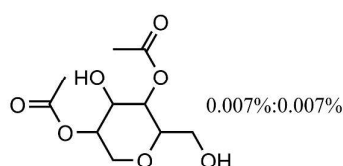
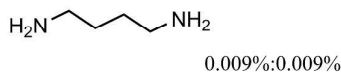
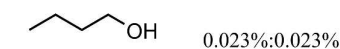
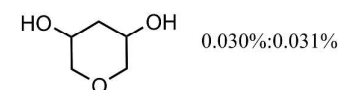
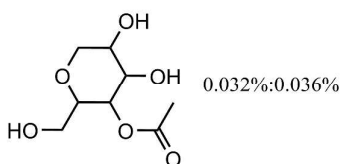
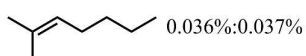
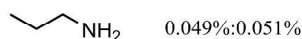
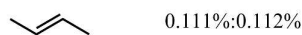
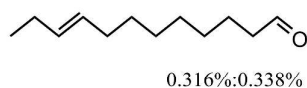
TNPs unique: total 708411



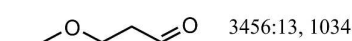
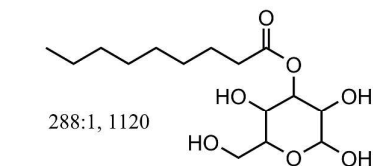
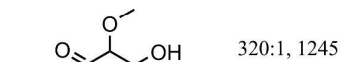
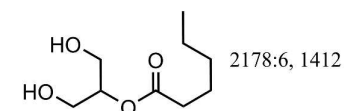
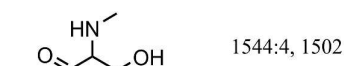
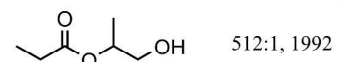
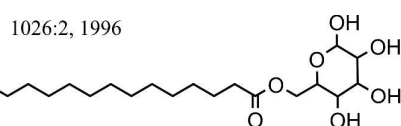
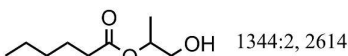
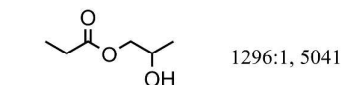
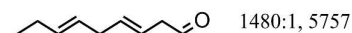
(A)

MNPs/TNPs common: total 10251

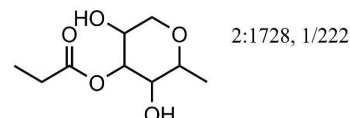
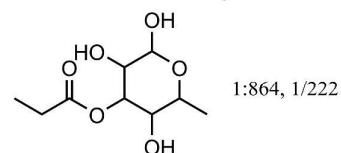
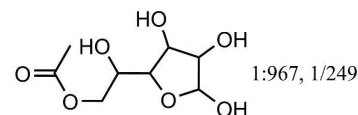
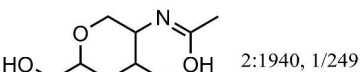
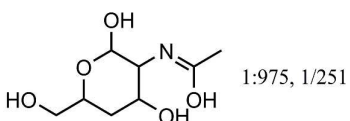
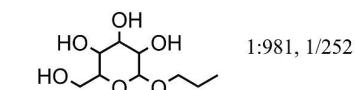
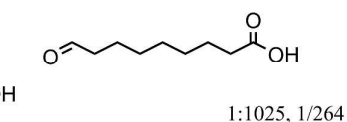
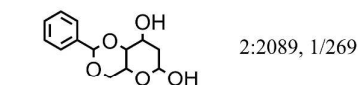
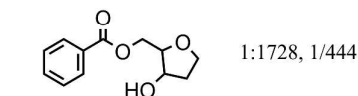
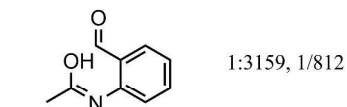
MNPs/TNPs equal*



MNPs/TNPs descend



MNPs/TNPs ascend



* The total frequencies of RECAP fragments in MNPs and TNPs are 1777882 and 6915803, respectively.

(B)

Figure S3. (A) unique and (B) common RECAP fragments of MNPs and TNPs.

Chemical structures of compounds 30, 31, 32, 42, 52, and 58 are shown. Compound 30 is a benzyl-protected nucleoside derivative. Compound 31 is a complex polycyclic molecule with multiple oxygen atoms. Compound 32 is a long-chain molecule with multiple double bonds and a terminal epoxide. Compound 42 is a complex polycyclic molecule with multiple oxygen atoms. Compound 52 is a complex polycyclic molecule with multiple oxygen atoms. Compound 58 is a simple cyclic ether.

363 322 283 233 214

202 175 175 138 137

MNP/TPNs equal*		MNPs/TPNs descend		MNPs/TPNs ascend	
	0.631%:0.598%		34:1, 144		1:155, 1/37
	0.440%:0.460%		32:1, 135		6:898, 1/35
	0.364%:0.333%		27:1, 114		3:251, 1/20
	0.336%:0.309%		26:1, 110		7:439, 1/15
	0.271%:0.300%		22:1, 93		1:58, 1/14
	0.225%:0.220%		19:1, 80		3:172, 1/14
	0.218%:0.203%		33:2, 70		15:797, 1/13
	0.208%:0.191%		16:1, 68		1:52, 1/12
	0.170%:0.176%		31:2, 66		4:202, 1/12
	0.156%:0.152%		15:1, 63		1:50, 1/12

Figure S4. Unique and common Murcko frameworks of MNPs and TNPs.