Supporting information

Predicting Ion Mobility Collision Cross Sections Using a Deep Neural Network: DeepCCS

Pier-Luc Plante[1, 2, 3], Élina Francovic-Fontaine[1, 2], Jody C. May[4], John A. McLean[4], Erin S. Baker[5], François Laviolette[1], Mario Marchand[1], Jacques Corbeil[1, 2, 3]

[1] Big Data Research Centre, Université Laval, Qc, G1V 0A6, Canada

[2] Centre de Recherche en Infectiologie de l'Université Laval, Axe Maladies Infectieuses et Immunitaires, Centre de Recherche du CHU de Québec-Université Laval, Québec City, G1V 4G2, Canada

[3] Département de médecine moléculaire, Faculté de médecine, Université Laval, Québec City, G1V 0A6, Canada

[4] Department of Chemistry, Center for Innovative Technology, Vanderbilt University, Nashville, Tennessee 37235, United States

[5] Department of Chemistry, North Carolina State University, Raleigh, North Carolina 27695, United States.

*Corresponding Author: Jacques Corbeil (jacques.corbeil@fmed.ulaval.ca)

Table of content

Method supplementary information 1

Outliers description 1

Figure S1 Repeat measurements of the IM spectra and CCS values. 3

Table S1. Possible CNN hyper-parameters values during the random-search cross-validation 4

Table S2. CNN multi-output model performances on the HMDB molecular properties prediction problem 4

Table S3. DeepCCS neural network structure 5

Table S4. CNN structure for HMDB chemical properties prediction 6

Table S5. Effect of repetitive SMILES-ion combination on the single split experiment 8

Table S6. ClassyFire classification at the class level of the datasets used to train and test DeepCCS 9

Table S7. ClassyFire classification at the subclass level of the datasets used to train and test DeepCCS 11

Method supplementary information

The five-fold cross-validation was performed using only the training set, trying a total of 340 different combinations for a total of 1700 different models trained. The complete list of hyper-parameters that were tested and their range are available in **Table S1**. The hyperparameter combination giving the best cross-validation score was kept. Finally, the last maximum pooling layer stride parameter was increased to 2 in order to reduce the total number of learnable parameters in the network. During training, the Adam optimizer was used with a learning rate of $1e^{-4}$ for 150 epochs and a batch size of 2. Further details about the neural network structure can be found in **Table S3** and on the github repositories available at github.com/plpla/DeepCCS and github.com/plpla/DeepCCS_paper.

Outliers description

Five outliers are visible in Figure 5 of the manuscript. Here, we describe each outlier with either a confirmed or highly probable explanation.

A. Name: Methyl behenate
   Class: Lipid
   Ion: M-H
   Reference CCS: 156.1 Å^2
   DeepCCS: 192.7 Å^2
   Explanation: A repeated measurement of methyl behenate using the standardized CCS measurement protocol described by Stow et al. (*Analytical Chemistry* 89(17), 9048-9055, 2017) gives a CCS value of 186.2 Å^2 (**Figure S1-A**). This confirms the reference value error.
B. Name: 1,2-Diacyl-sn-glycero 3-phosphocholine
   Class: Lipid
   Ion: M+H
   Reference CCS: 189.6 Å^2
   DeepCCS: 251.0 Å^2
   Explanation: A similar lipid (PC 34:2), the double bound unspecified form of 1,2-Diacyl-sn-glycero 3-phosphocholine, was measured with a CCS value of 279.5 Å^2 (**Figure S1-B**), a value closer to the predicted value. A low abundance ion signal appears at lower CCS and it is suspected that a similar signal artifact is the source of the reference value error.
C. Name: D-Maltose
   Class: Sugar
   Ion: M-H
   Reference CCS: 205.9 Å^2
   DeepCCS: 169.1 Å^2
   Explanation: A repeated measurement of D-maltose gives a CCS value of 168.8 Å^2 (**Figure S1-C**). This updated measurement is very close to the DeepCCS prediction and confirms the reference value error. Carbohydrates are prone to aggregation, and these multimers readily dissociate during transfer from the IM to the MS stage, resulting in multiple ion signals appearing at higher CCS values (c.f.,

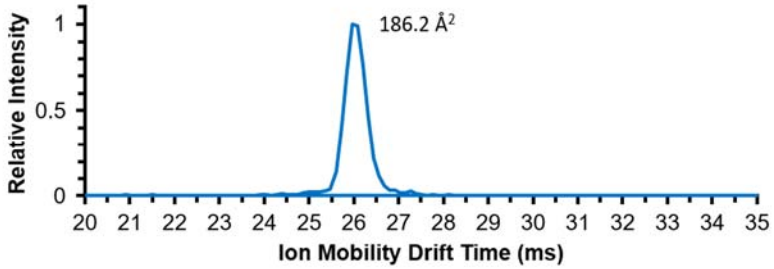
Figure S1-C). It is suspected that the large CCS reported for the reference value is in fact a multimer signal.

D. Name: Sophorose
   Class: Sugar
   Ion: M-H
   Reference CCS: 187.1 Å^2
   DeepCCS: 168.7 Å^2
   Explanation: As was observed with D-maltose (above) it is believed that the large discrepancy between the reference and predicted CCS values for sophorose is due to an aggregate that dissociated into the monomer mass prior to the MS measurement, although no measurements are available to confirm this hypothesis.

E. Name: L-Threonine
   Class: Amino acid
   Ion: M-H
   Reference CCS: 141.4 Å^2
   DeepCCS: 125.1 Å^2
   Explanation: The reference CCS value is higher than other amino acids reported in the reference set. A repeated IM measurement of L-threonine gives a CCS value of 127.6 Å^2 (**Figure S1-D**), which is much closer to the predicted value. This confirms the reference error. L-threonine exhibits several high CCS artifacts in the IM spectrum (c.f., **Figure S1-D**), and these are likely the source of the erroneously high reference CCS.

(A) Methyl Behenate ($C_{23}H_{46}O_2$)

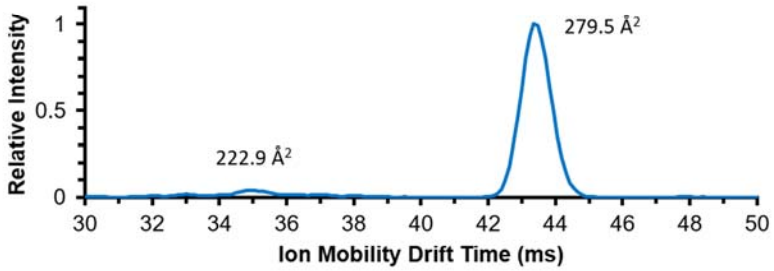
$[M-H]^-$ = 353.3419 Da (measured = 353.3412; 2.1 ppm)



	$^{DT}CCS_{N2}$ (Å²)
1	185.9
2	186.3
3	186.3
4	186.4
5	186.2
6	186.1
7	185.9
Average=	**186.2**
Std. Deviation=	**0.2**

(B) Phosphatidylcholine 34:2 ($C_{42}H_{80}NO_8P$)

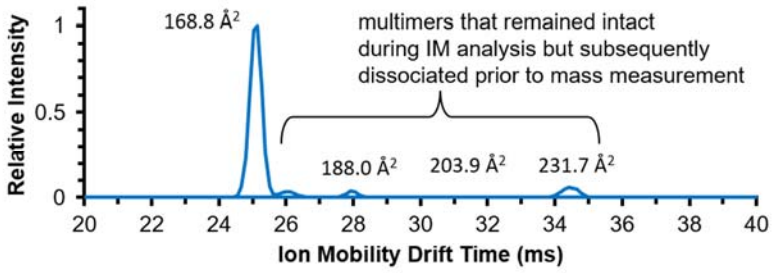
$[M+H]^+$ = 758.5699 Da (measured = 758.5665; 4.5 ppm)



	$^{DT}CCS_{N2}$ (Å²)
1	278.8
2	280.6
3	279.4
4	279.0
5	279.9
6	279.2
7	279.4
Average=	**279.5**
Std. Deviation=	**0.6**

(C) D-Maltose ($C_{12}H_{22}O_{11}$)

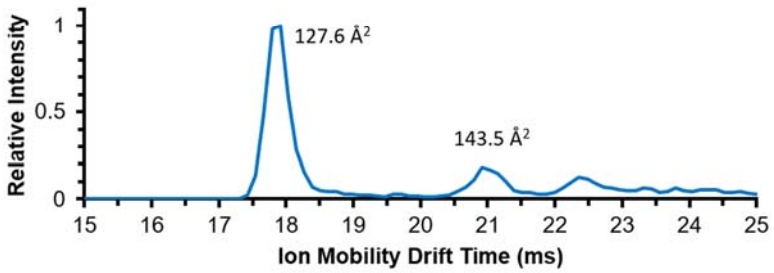
$[M-H]^-$ = 341.1084 Da (measured = 341.1088; 1.2 ppm)



	$^{DT}CCS_{N2}$ (Å²)
1	168.6
2	168.9
3	169.1
4	169.0
5	168.9
6	168.7
7	168.6
Average=	**168.8**
Std. Deviation=	**0.2**

(D) L-Threonine ($C_4H_9NO_3$)

$[M+H]^+$ = 120.0661 Da (measured = 120.0657; 3.0 ppm)



	$^{DT}CCS_{N2}$ (Å²)
1	127.5
2	127.6
3	127.8
4	127.7
5	127.6
6	127.5
7	127.5
Average=	**127.6**
Std. Deviation=	**0.1**

Figure S1. Repeat measurements of the IM spectra and CCS values for **(A)** methyl behenate, **(B)** PC 34:2, **(C)** D-maltose, and **(D)** L-threonine.

Table S1. Possible CNN hyper-parameters values during the random-search cross-validation

Parameter	Values
Max. number of epochs	50, 100 or 150
Batch size	2, 5, 10, 15 or 20
Dropout rate	0.0 to 0.5
Number of convolution layers	Between 1 and 10
Convolution layer width	64, 128, 256 or 384
Convolution filter size	3, 4 or 5
Maximum pooling filter size	2, 3 or 4
Number of dense layers	Between 1 and 10
Dense layers width	64, 128, 256 or 384
Add normalization layer	True or False

Table S2. CNN multi-output model performances on the HMDB molecular properties prediction problem

Property	R^2	Median relative error (%)
Polar surface area	0.9998	0.1670
logS	0.9860	-0.6311
Refractivity	0.9999	0.0995
Polarizability	0.9997	0.2005
logP (ALOGPS)	0.9891	0.6419
logP (Chemaxon)	0.9966	0.1574

Table S3. DeepCCS neural network structure

Layer (type)	Output Shape	Param #	Connected to
smile (InputLayer)	(None, 250, 37)	0	
conv1d_1 (Conv1D)	(None, 247, 64)	9536	smile[0][0]
conv1d_2 (Conv1D)	(None, 244, 64)	16448	conv1d_1[0][0]
max_pooling1d_1 (MaxPooling1D)	(None, 243, 64)	0	conv1d_2[0][0]
conv1d_3 (Conv1D)	(None, 240, 64)	16448	max_pooling1d_1[0][0]
max_pooling1d_2 (MaxPooling1D)	(None, 239, 64)	0	conv1d_3[0][0]
conv1d_4 (Conv1D)	(None, 236, 64)	16448	max_pooling1d_2[0][0]
max_pooling1d_3 (MaxPooling1D)	(None, 235, 64)	0	conv1d_4[0][0]
conv1d_5 (Conv1D)	(None, 232, 64)	16448	max_pooling1d_3[0][0]
max_pooling1d_4 (MaxPooling1D)	(None, 231, 64)	0	conv1d_5[0][0]
conv1d_6 (Conv1D)	(None, 228, 64)	16448	max_pooling1d_4[0][0]
max_pooling1d_5 (MaxPooling1D)	(None, 227, 64)	0	conv1d_6[0][0]
conv1d_7 (Conv1D)	(None, 224, 64)	16448	max_pooling1d_5[0][0]
max_pooling1d_6 (MaxPooling1D)	(None, 112, 64)	0	conv1d_7[0][0]
flatten_1 (Flatten)	(None, 7168)	0	max_pooling1d_6[0][0]
adduct (InputLayer)	(None, 4)	0	
concatenate_1 (Concatenate)	(None, 7172)	0	flatten_1[0][0], adduct[0][0]
dense_1 (Dense)	(None, 384)	2754432	concatenate_1[0][0]
dense_2 (Dense)	(None, 384)	147840	dense_1[0][0]
dense_3 (Dense)	(None, 1)	385	dense_2[0][0]

Table S4. CNN structure for HMDB chemical properties prediction

Layer	(type)	Output Shape	Param #	Connect to
smiles	(InputLayer)	(None, 250, 58)	0	
conv1d_1	(Conv1D)	(None, 247, 64)	14912	smiles[0][0]
conv1d_2	(Conv1D)	(None, 244, 64)	16448	conv1d_1[0][0]
max_pooling1d_1	(MaxPooling1D)	(None,243, 64)	0	conv1d_2[0][0]
conv1d_3	(Conv1D)	(None, 240, 64)	16448	max_pooling1d_1[0][0]
max_pooling1d_2	(MaxPooling1D)	(None, 239, 64)	0	conv1d_3[0][0]
conv1d_4	(Conv1D)	(None, 236, 64)	16448	max_pooling1d_2[0][0]
max_pooling1d_3	(MaxPooling1D)	(None, 235 , 64)	0	conv1d_4[0][0]
conv1d_5	(Conv1D)	(None, 232, 64)	16448	max_pooling1d_3[0][0]
max_pooling1d_4	(MaxPooling1D)	(None, 231 , 64)	0	conv1d_5[0][0]
conv1d_6	(Conv1D)	(None, 228, 64)	16448	max_pooling1d_4[0][0]
max_pooling1d_5	(MaxPooling1D)	(None,227, 64)	0	conv1d_6[0][0]
conv1d_7	(Conv1D)	(None, 223, 64)	16448	max_pooling1d_5[0][0]
max_pooling1d_6	(MaxPooling1D)	(None, 112, 64)	0	conv1d_7[0][0]
flatten_1	(Flatten)	(None, 7168)	0	max_pooling1d_6[0][0]
dense_1	(Dense)	(None, 384)	2752896	flatten_1[0][0]
dense_3	(Dense)	(None, 384)	2752896	flatten_1[0][0]

dense_5	(Dense)	(None, 384)	2752896	flatten_1[0][0]
dense_7	(Dense)	(None, 384)	2752896	flatten_1[0][0]
dense_9	(Dense)	(None, 384)	2752896	flatten_1[0][0]
dense_11	(Dense)	(None, 384)	2752896	flatten_1[0][0]
dense_2	(Dense)	(None, 384)	147840	dense_1[0][0]
dense_4	(Dense)	(None, 384)	147840	dense_3[0][0]
dense_6	(Dense)	(None, 384)	147840	dense_5[0][0]
dense_8	(Dense)	(None, 384)	147840	dense_7[0][0]
dense_10	(Dense)	(None, 384)	147840	dense_9[0][0]
dense_12	(Dense)	(None, 384)	147840	dense_11[0][0]
polar_surface_area	(Dense)	(None, 1)	385	dense_2[0][0]
logs	(Dense)	(None, 1)	385	dense_4[0][0]
refractivity	(Dense)	(None, 1)	385	dense_6[0][0]
polarizability	(Dense)	(None, 1)	385	dense_8[0][0]
logp_alogps	(Dense)	(None, 1)	385	dense_10[0][0]
logp_chemaxon	(Dense)	(None, 1)	385	dense_12[0][0]

Table S5. Effect of repetitive SMILES-ion combination on the single split experiment

Dataset	Single split		Single split no repetitions	
	R^2	Median relative error (%)	R^2	Median relative error (%)
Global	**0.976 (0.001)**	**2.67 (0.18)**	**0.976 (0.001)**	**2.67 (0.18)**
MetCCS Agilent pos.	0.960 (0.005)	2.02 (0.24)	0.957 (0.006)	2.50 (0.12)
MetCCS Agilent neg.	0.969 (0.005)	3.11 (0.49)	0.978 (0.005)	2.87 (0.53)
Astarita pos.	0.901 (0.013)	4.86 (0.30)	0.901 (0.011)	5.05 (0.33)
Astarita neg.	0.955 (0.006)	3.13 (0.48)	0.961 (0.005)	3.11 (0.44)
Baker	0.954 (0.006)	2.43 (0.11)	0.948 (0.008)	2.64 (0.11)
McLean	0.995 (0.001)	1.49 (0.14)	0.992 (0.001)	1.71 (0.20)
CBM 2018	0.930 (0.010)	2.26 (0.28)	0.930 (0.010)	2.26 (0.28)
Repetitives	-	-	0.960 (0.005)	2.69 (0.34)

Table S6. ClassyFire classification at the class level of the datasets used to train and test DeepCCS

Class	Number	Class	Number
Flavonoids	17	Tetracyclines	1
Ergoline and derivatives	1	Organic phosphonic acids and derivatives	1
Pyrimidine nucleotides	18	Nucleoside and nucleotide analogues	1
('Unknown	2	Harmala alkaloids	1
6,7-benzomorphans	1	Benzothiadiazoles	1
Strychnos alkaloids	1	Keto acids and derivatives	7
Fatty Acyls	92	Diarylheptanoids	1
Glycerophospholipids	38	Pyridine nucleotides	2
Isoflavonoids	6	Glycerolipids	3
Organofluorides	3	Organic phosphoric acids and derivatives	5
Diazanaphthalenes	2	Piperidines	7
Pyridines and derivatives	22	Diazines	12
Imidazothiazoles	1	Tetrahydroisoquinolines	1
Triazines	3	Carboxylic acids and derivatives	211
Thioethers	1	Lactones	3
Organic sulfuric acids and derivatives	2	Benzothiopyrans	5
Pyrrolopyrazines	1	Azoles	14
Anthracenes	3	Tetrapyrroles and derivatives	5
Benzofurans	1	Organic sulfonic acids and derivatives	1
Organooxygen compounds	135	Purine nucleosides	14
Piperazinoazepines	2	Carboximidic acids and derivatives	2
Indanes	1	Cycloheptathiophenes	1
5'-deoxyribonucleosides	4	Pyrimidine nucleosides	9
Dibenzocycloheptenes	5	Oxazinanes	1
Ribonucleoside 3'-phosphates	1	Benzene and substituted derivatives	134
Phenols	21	Imidazopyrimidines	23
Indenes and isoindenes	1	Cinnamaldehydes	1
Biotin and derivatives	1	Lactams	1
Benzodioxoles	6	Phenanthrenes and derivatives	5
Tetralins	1	Dithiolanes	1

('Unclassified	54	Organic carbonic acids and derivatives	2
Cinnamic acids and derivatives	6	Peptidomimetics	7
Yohimbine alkaloids	2	Stilbenes	2
Benzazepines	8	Benzoxazoles	1
Purine nucleotides	30	Morphinans	12
Benzoxazepines	2	Indoles and derivatives	36
Polypeptides	7	Naphthalenes	8
Organonitrogen compounds	29	Coumarins and derivatives	3
Pyrroles	3	Thienodiazepines	1
Phenol ethers	10	Quinolines and derivatives	6
Organic oxoanionic compounds	2	Benzothiazepines	3
Isoquinolines and derivatives	1	Benzimidazoles	2
Benzocycloheptapyridines	2	Dihydrofurans	1
Phenylpropanoic acids	5	Flavin nucleotides	2
Imidazole ribonucleosides and ribonucleotides	1	Amaryllidaceae alkaloids	1
Organic dithiophosphoric acids and derivatives	1	Benzodiazepines	32
Phthalide isoquinolines	1	Non-metal oxoanionic compounds	1
Pyridopyrimidines	2	Cinchona alkaloids	2
Macrolactams	1	Steroids and steroid derivatives	54
Benzothiepins	1	Hydroxy acids and derivatives	7
Diazinanes	10	(5'->5'-dinucleotides	6
Linear 1,3-diarylpropanoids	3	Tropane alkaloids	1
Benzoxepines	2	Pteridines and derivatives	10
Sphingolipids	5	Benzopyrans	2
Prenol lipids	22	Benzothiazines	16

Table S7. ClassyFire classification at the subclass level of the datasets used to train and test DeepCCS

Subclass	Number	Subclass	Number
Pterins and derivatives	8	Fatty aldehydes	1
Benzylethers	2	1-hydroxy-2-unsubstituted benzenoids	2
Chalcones and dihydrochalcones	1	Pheniramines	1
Unknown	132	Retinoids	3
Delta valerolactones	1	Pyridinecarboxylic acids and derivatives	7
Short-chain keto acids and derivatives	3	Quinoline carboxylic acids	2
Dibenzoxazepines	2	Phosphosphingolipids	1
Anilides	9	Phenylpiperidines	3
Cyclopyrrolones	1	Piperidinecarboxylic acids and derivatives	2
Gamma butyrolactones	2	Fatty acid esters	9
Phenylbutylamines	4	Fentanyls	1
Quinone and hydroquinone lipids	6	Steroid esters	1
Purine nucleotide sugars	4	Purine ribonucleotides	18
Diphenylmethanes	21	Carbazoles	3
Pyridine carboxaldehydes	3	Imidazoles	8
Hydroxyindoles	2	Carbonyl compounds	19
Pregnane steroids	4	Beta hydroxy acids and derivatives	4
Sulfated steroids	2	Indolyl carboxylic acids and derivatives	5
Monoterpenoids	4	Aminoquinolines and derivatives	2
Organic pyrophosphates	2	N-phenylureas	2
Benzodifurans	1	Linear diarylheptanoids	1
Fatty acids and conjugates	53	Unclassified	54
Dicarboxylic acids and derivatives	3	Gamma-keto acids and derivatives	2
Terpene glycosides	2	Pyrazoles	3
Galanthamine-type amaryllidaceae alkaloids	1	Aniline and substituted anilines	2
Amino acids, peptides and analogues	195	1-benzopyrans	2
Hydroxypyridines	1	Beta lactams	1
Androstane steroids	6	Carboxylic acid derivatives	5
O-methylated flavonoids	5	Styrenes	1
Diterpenoids	2	Hydroxysteroids	8

Purines and purine derivatives	23	Phenylpropanes	3
Benzenediols	9	Fatty acyl thioesters	16
Triradylcglycerols	1	Vitamin D and derivatives	1
Purine deoxyribonucleotides	6	Benzylisoquinolines	1
Pyrimidine nucleotide sugars	5	Benzenesulfonamides	17
Indolines	1	Cholestane steroids	3
Trifluoromethylbenzenes	2	1,3, 5-triazines	2
Bipyridines and oligopyridines	1	O-methylated isoflavonoids	3
Benzoic acids and derivatives	31	Flavones	5
Purine 2-deoxyribonucleosides	3	Alloxazines and isoalloxazines	1
Medium-chain keto acids and derivatives	2	Quinolones and derivatives	1
Pyridoxamines	2	Aryl thioethers	1
Phosphate esters	5	Pyrimidine ribonucleotides	6
Halobenzenes	10	Aminophenyl ethers	1
1-ribosyl-imidazolecarboxamides	1	5-deoxy-5-thionucleosides	3
Benzyl alcohols	2	Methoxybenzenes	4
Isoprenoid phosphates	2	1-benzothiopyrans	5
Substituted pyrroles	3	Sulfanilides	1
Benzylpiperidines	1	Pyridoindoles	1
Indolecarboxylic acids and derivatives	1	Dibenzothiazepines	1
Indoloquinolines	1	Flavonoid glycosides	3
Fatty acyl glycosides	1	Phenoxy compounds	2
Piperazines	10	Methoxyphenols	8
Bile acids, alcohols and derivatives	23	Alpha-halocarboxylic acids and derivatives	1
Indoles	6	Pyrimidine 2-deoxyribonucleosides	3
Cinnamic acids	2	Glycerophosphoglycerols	2
Hydroxycinnamic acids and derivatives	4	Tryptamines and derivatives	15
Amines	21	Glycerophosphocholines	19
Hydroxycoumarins	3	Pyrrolidinylpyridines	1
Carboximidic acids	2	Dibenzodiazepines	2
Porphyrins	3	Quaternary ammonium salts	6
Biflavonoids and polyflavonoids	1	Lysergic acids and derivatives	1
Anthraquinones	1	Steroid lactones	1
Xylenes	3	Lineolic acids and derivatives	6
Anisoles	5	Butyrophenones	2
1, 4-benzodiazepines	27	Thiazoles	3
Eicosanoids	6	Cyclohexylphenols	1

Tricarboxylic acids and derivatives	6	Dithiophosphate O-esters	1
Flavans	3	Pyrimidines and pyrimidine derivatives	12
Ureas	2	Short-chain hydroxy acids and derivatives	1
Furanones	1	Phenylacetamides	1
Phenylpyruvic acid derivatives	1	Alpha hydroxy acids and derivatives	2
Glycerophosphoethanolamines	13	Steroidal glycosides	1
Triterpenoids	1	Tetracarboxylic acids and derivatives	1
Carbohydrates and carbohydrate conjugates	107	Naphthoylindoles	1
Tyrosols and derivatives	2	Lipoamides	1
Glycerophosphates	2	Non-metal pyrophosphates	1
Benzenesulfonic acids and derivatives	1	Nicotinamide nucleotides	1
Cyclic pyrimidine nucleotides	1	Benzoylindoles	1
Ceramides	2	Bilirubins	2
Sesquiterpenoids	2	Dibenzothiepins	1
Naphthoquinones	2	Pyrimidine deoxyribonucleotides	6
Nicotinic acid nucleotides	1	Benzodiazines	2
Biphenyls and derivatives	2	Dibenzoxepines	2
Glycerophosphoserines	2	2,6-dimethyl-3-benzazocines	1
Benzylamines	1	Cyclic purine nucleotides	2
Rotenoids	1	Diradylglycerols	1
Hydropyridines	4	Pyridoxines	2
Alcohols and polyols	8	Estrane steroids	4
Guanidines	2	Isoflav-2-enes	2
Dibenzazepines	7	Organosulfonic acids and derivatives	1
Phenothiazines	14	Morpholines	1
Glycosylglycerols	1	Phenethylamines	5
Hybrid peptides	7	Aminotriazines	1
Organic phosphonic acids	1	Arylsulfates	2
Glycosphingolipids	2		