Reduced Set Summary

Moritz E. Beber

2021-06-04

Contents

Intro

The entire, normalized database of compartments, compounds, and reactions at 2.6 GB is larger than what many people are used to. We are interested in creating a reduced set to distribute by default.

Properties

Table 1: The number of chemicals that participate in any recorded reaction and their properties.

Number of Chemicals	with formula	with charge	with SMILES	with mass	with InChI	with InChIKey
41,533	26,985	26,985	26,964	26,922	20,387	20,387

Table 2: The percentage of chemical information in MetaNetX.

Percent of Chemicals	with formula	with charge	with SMILES	with mass	with InChI	with InChIKey
100.00%	64.97%	64.97%	64.92%	64.82%	49.09%	49.09%

Table 3: Chemical formulae that are not fully determined.

Number of Formulae	with `*`	with `R`	with `Z[z]`
26,985	6,533	0	0

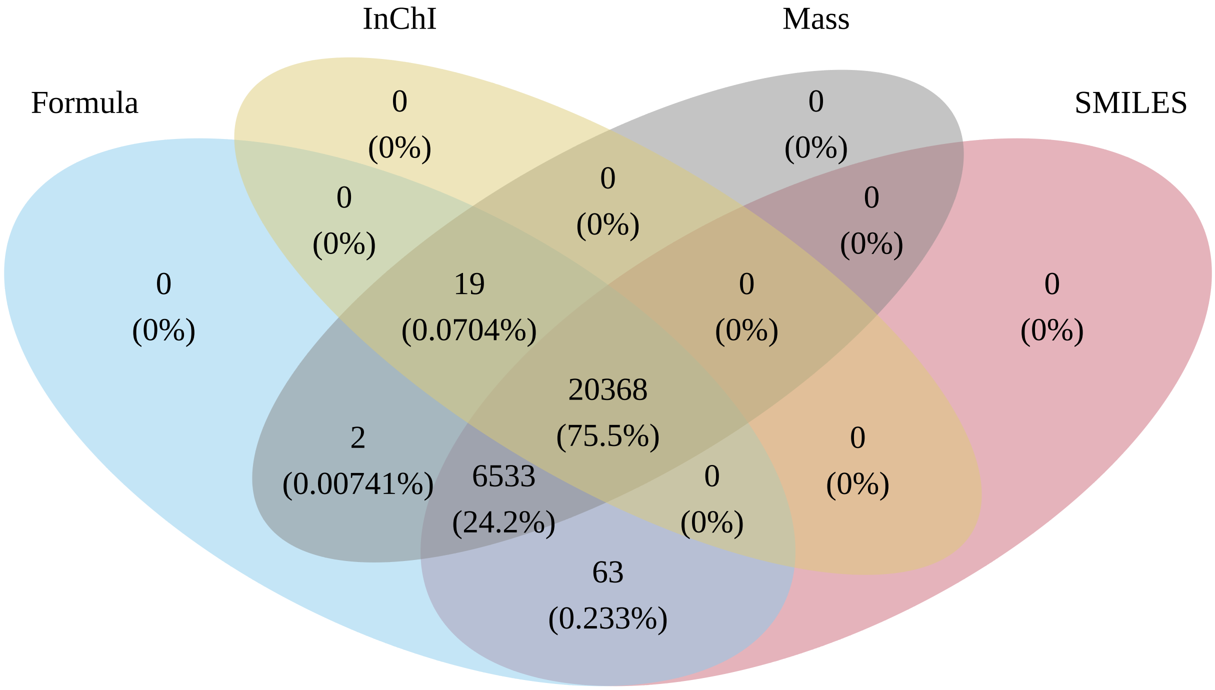

Figure 1: Venn diagram of structural annotation and mass.

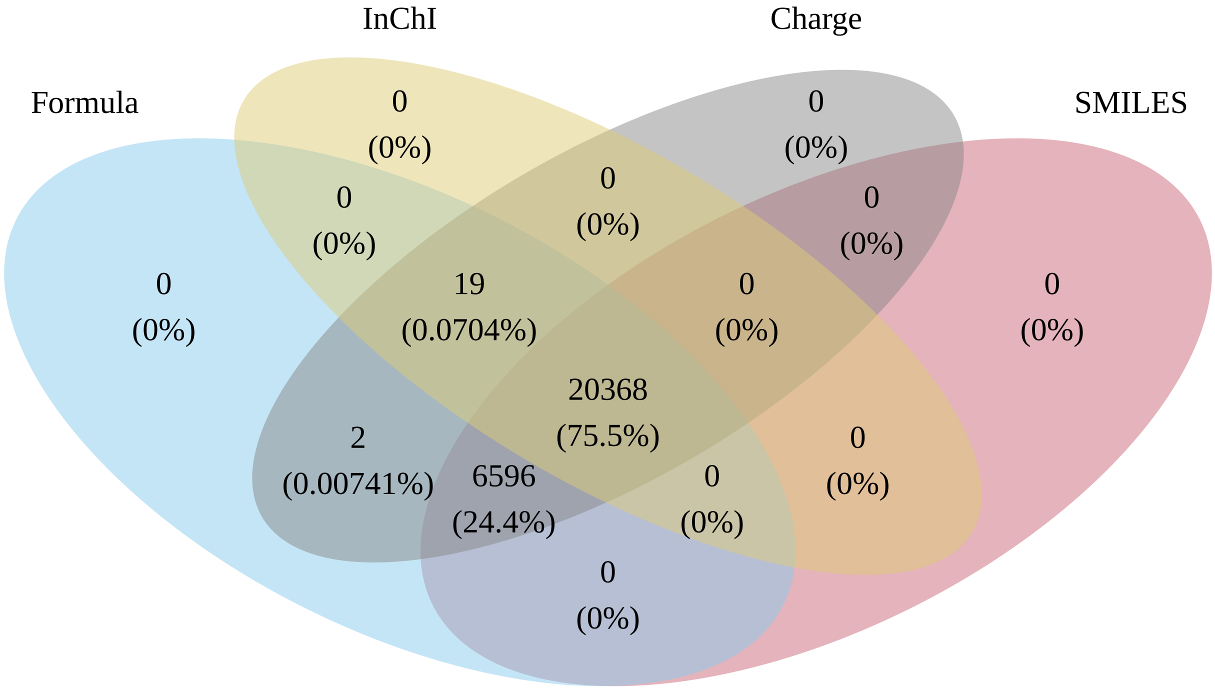

Figure 2: Venn diagram of structural annotation and electric charge.

Table 4: SMILES that are not fully determined.

Number of SMILES	with `*`	with `R`	with `Z[z]`
26,964	6,596	0	0

Table 5: InChIs that are not fully determined.

Number of InChIs	with `*`	with `R`	with `Z[z]`
20,387	0	0	0

Annotation

Table 6: Overall number of identifiers and of unique source namespaces.

Identifiers	Unique Namespaces
226,427	16

Table 7: Number of identifiers per source namespace. Identifiers are deduplicated compared to raw tables.

Namespace	Frequency
bigg.metabolite	18,063
chebi	35,031
envipath	1,586
hmdb	9,383
kegg.compound	19,822
kegg.drug	1,318
kegg.glycan	810
lipidmaps	2,838
metacyc.compound	16,202
metanetx.chemical	63,646
reactome	2,039
rhea.generic	1,490
rhea.polymer	195
sabiork.compound	8,899
seed.compound	43,274
slm	1,831

Names

Table 8: Overall number of names and of unique source namespaces.

Names	Unique Namespaces
386,597	16

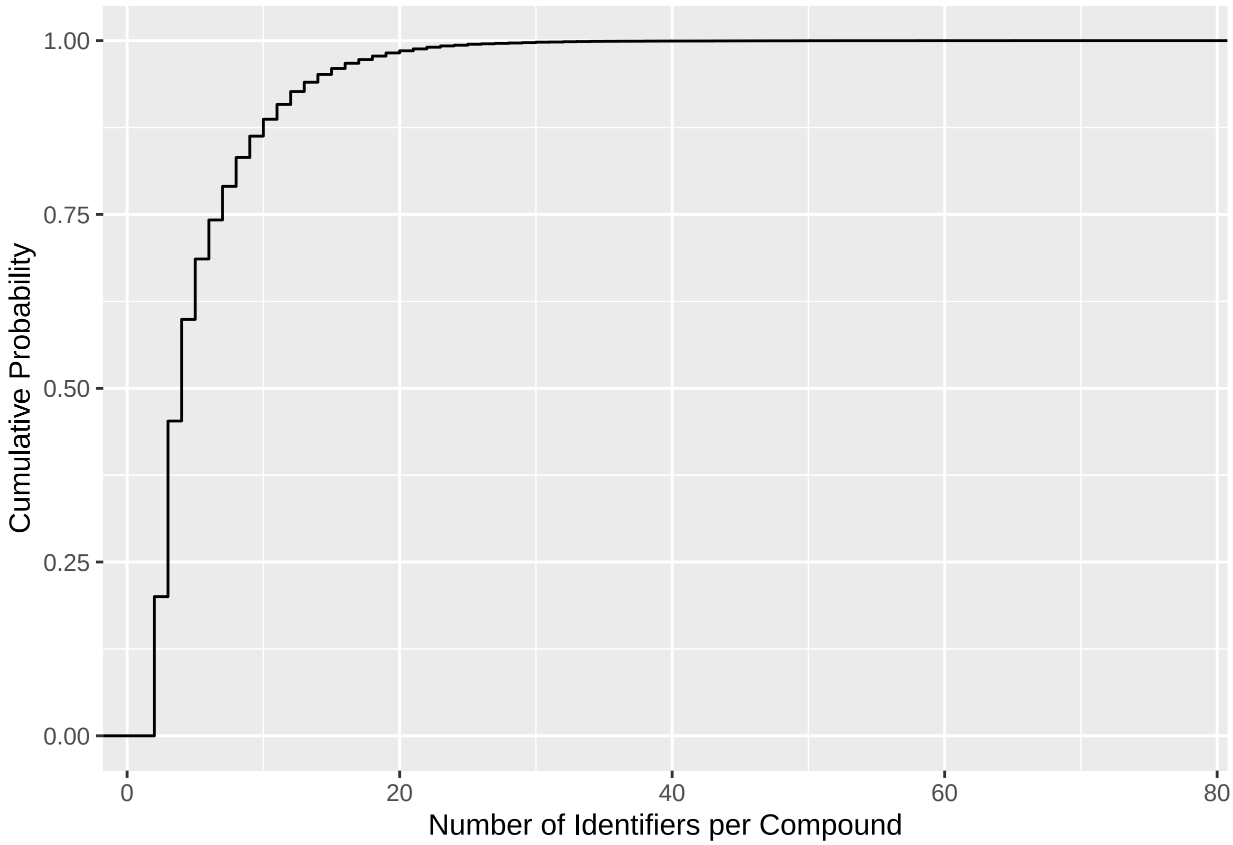

Figure 3: The empirical cumulative distribution function (eCDF) of the number of distinct identifiers per compound.

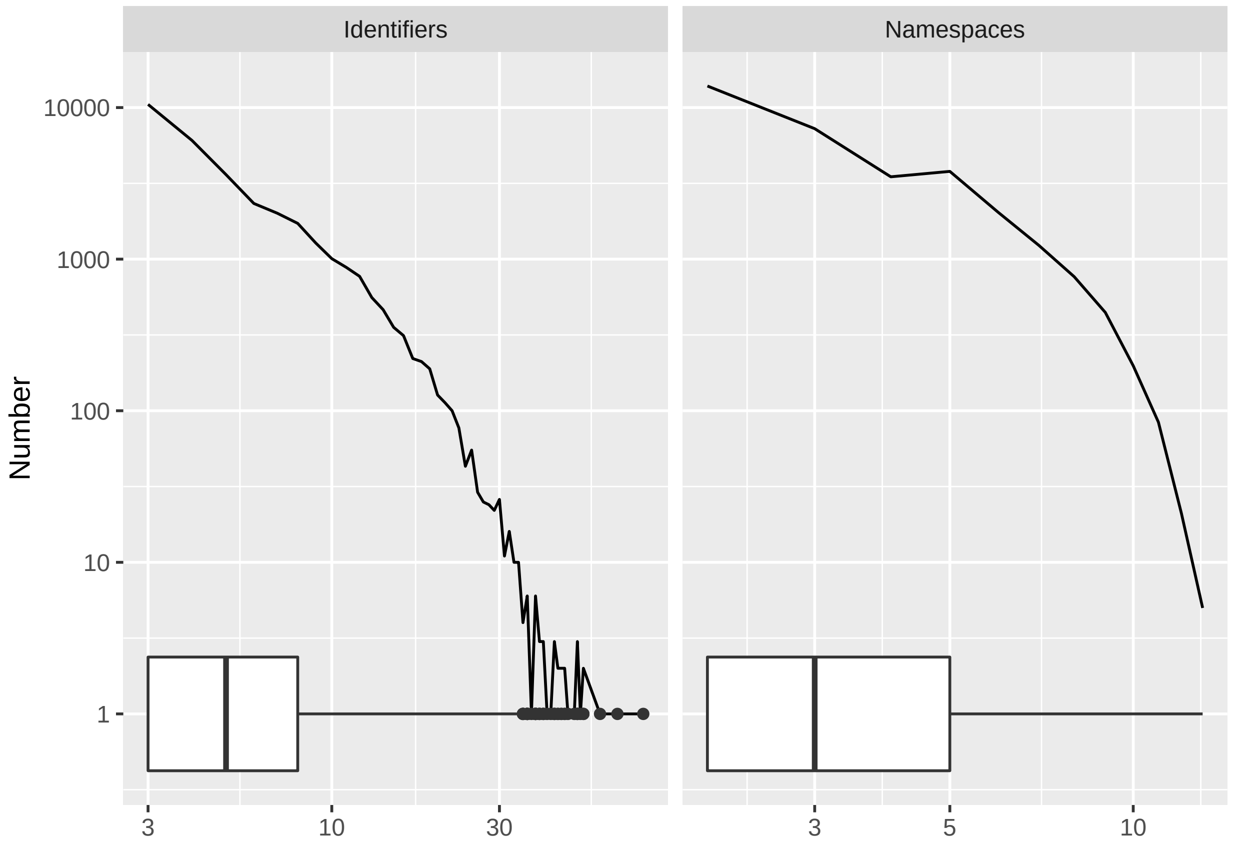

Figure 4: The number of distinct identifiers per compound and the number of unique source namespaces per compound. Only compounds that have more than two identifiers are included.

Table 9: Number of names per source namespace. Names are deduplicated compared to raw tables.

Namespace	Frequency
bigg.metabolite	17,863
chebi	102,619
envipath	1,582
hmdb	77,980
kegg.compound	28,491
kegg.drug	2,168
kegg.glycan	925
lipidmaps	9,308
metacyc.compound	37,227
metanetx.chemical	9
reactome	8,469
rhea.generic	1,490
rhea.polymer	195
sabiork.compound	13,594
seed.compound	80,667
slm	4,010

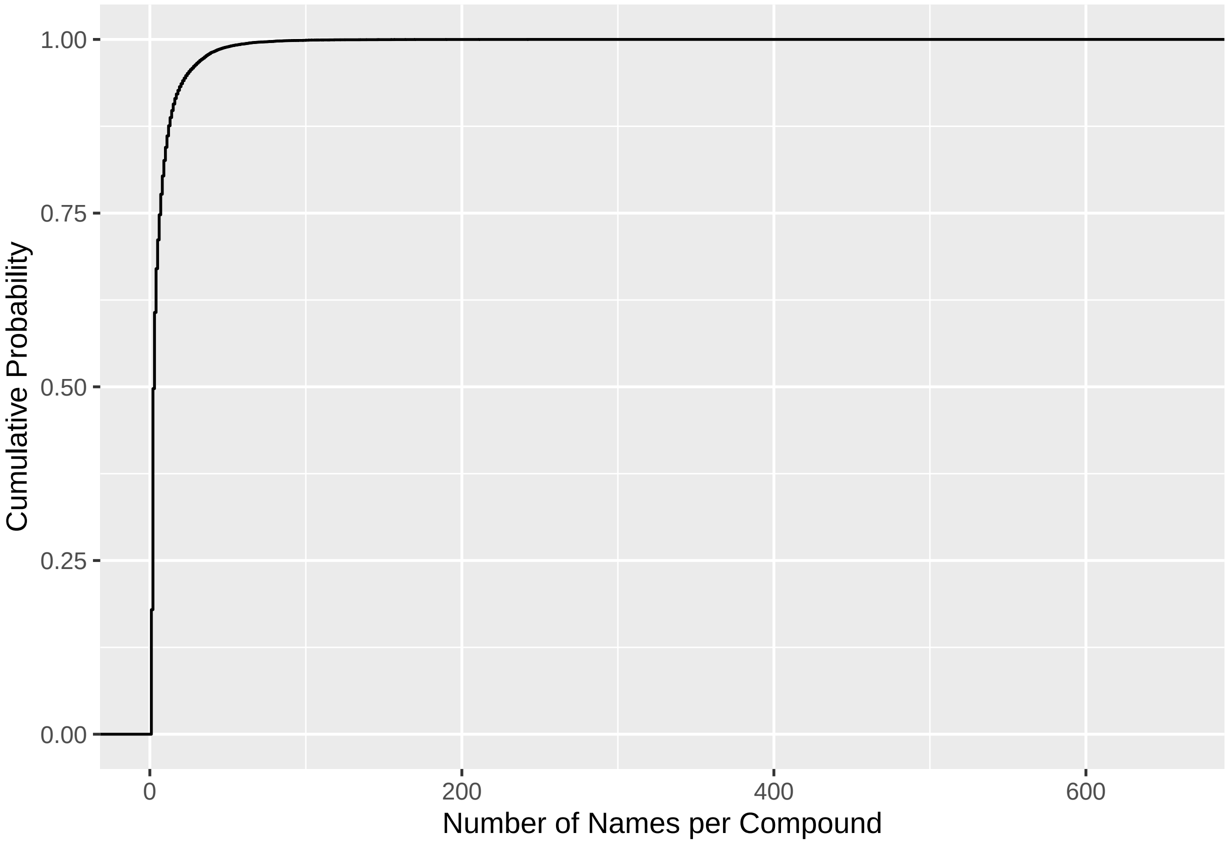

Figure 5: The empirical cumulative distribution function (eCDF) of the number of distinct names per compound.

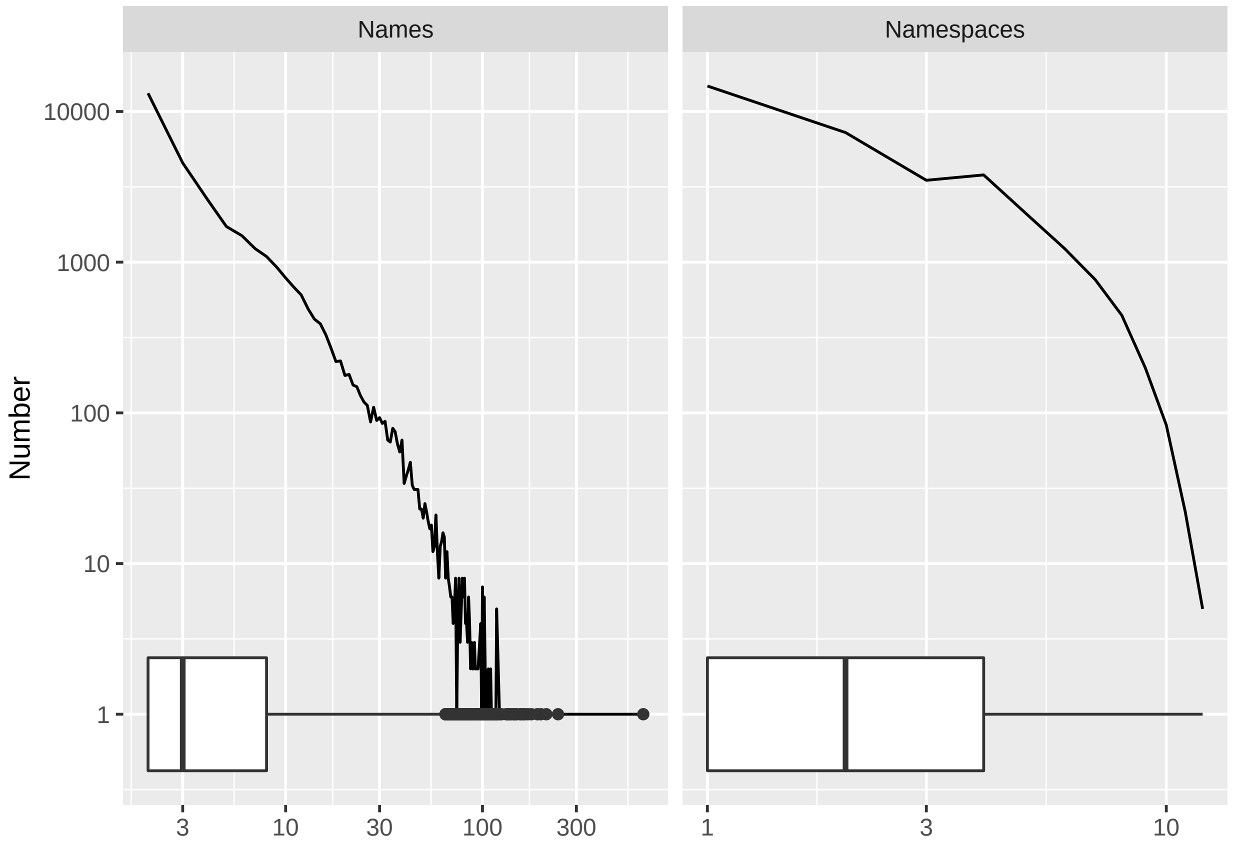

Figure 6: The number of names per compound and the number of unique source namespaces per compound. Only compounds that have more than one name are included.