MetaNetX Chemicals Summary

Moritz E. Beber

2021-06-04

Contents

Intro

A summary of the final database content.

Transformation Steps:

- Deduplicated identifiers per MNX reaction and namespace
- Deduplicated names per MNX reaction and namespace
- Deduplicated InChIs and combined identifiers and names
- Added deprecated MNX identifiers
- Attempted to generate additional InChIs from KEGG MDL MOL blocks (this led to more structures in the past but with MNX 4.2 adds no more information)
- Added additional compounds from eQuilibrator list via PubChem, if they don't exist already (tested by InChI)
- Added more structural information (InChI, SMILES, formula, mass, charge) from either InChI or SMILES if they are present but other information is not

Properties

Table 1: The number of chemicals with properties and sources for them in MetaNetX.

Number of Chemicals	with formula	with charge	with SMILES	with mass	with InChI	with InChIKey
1,043,435	999,995	999,995	999,849	814,242	803,094	803,094

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%	95.84%	95.84%	95.82%	78.03%	76.97%	76.97%

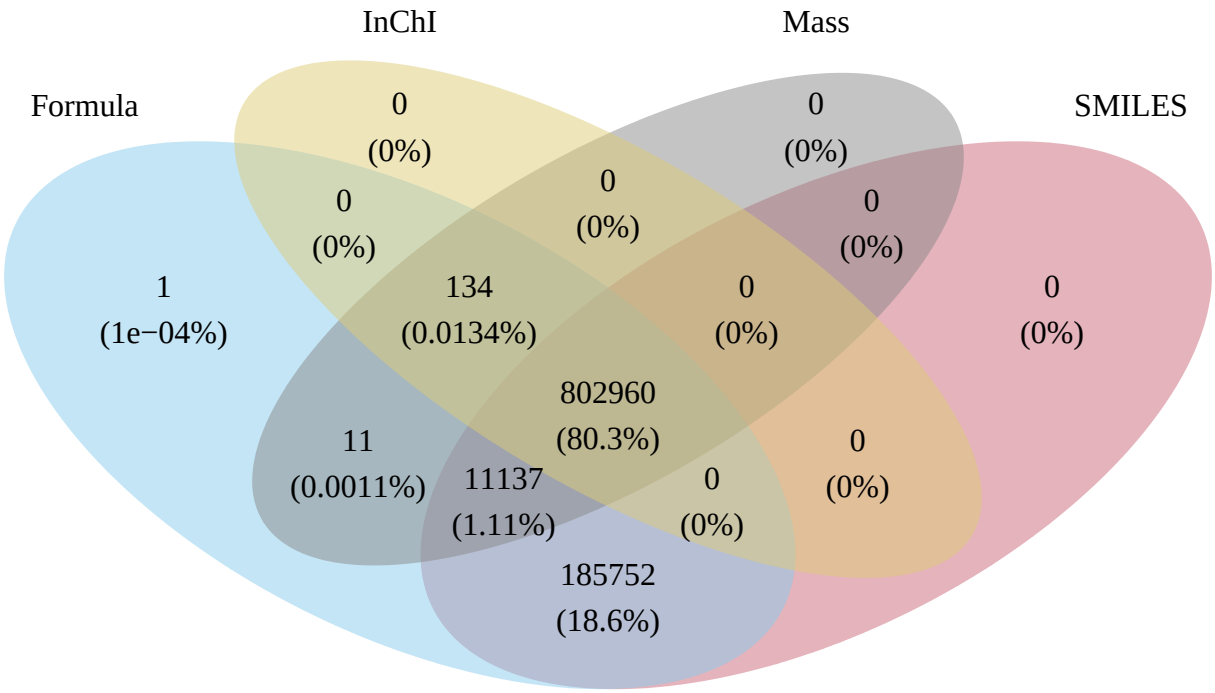

Figure 1: Venn diagram of structural annotation and mass.

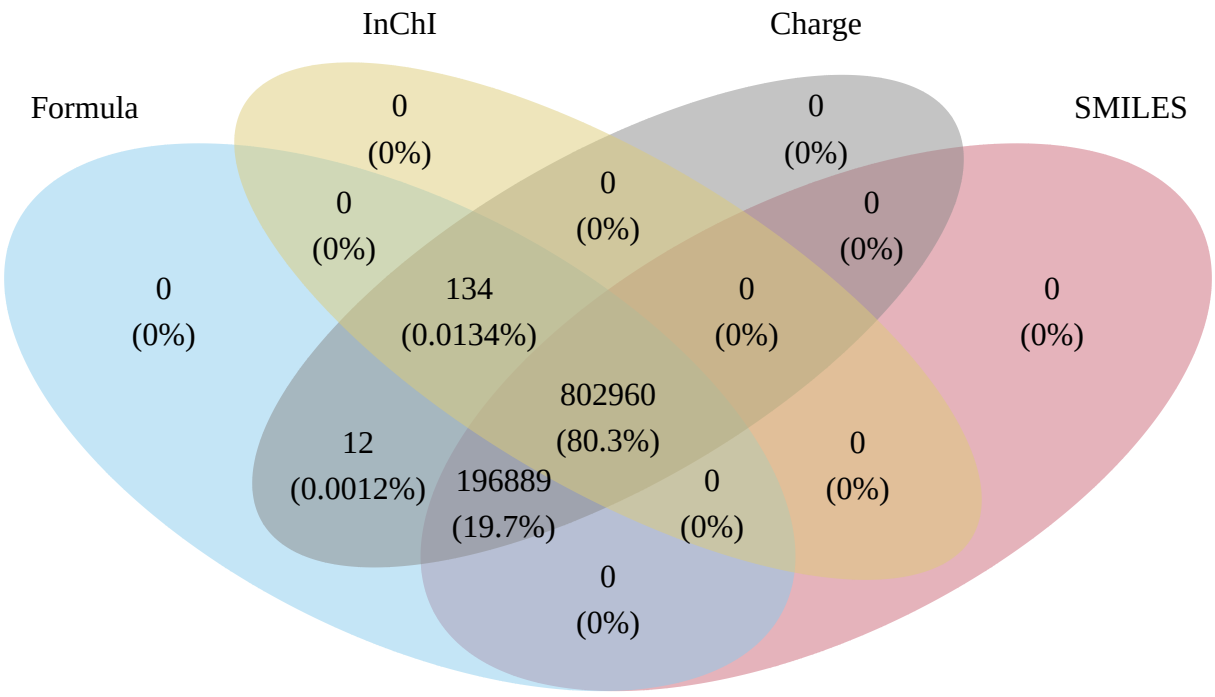

Figure 2: Venn diagram of structural annotation and electric charge.

Table 3: Chemical formulae that are not fully determined.

Number of Formulae	with *	with R	with Z[z]
999,995	11,063	0	0

Table 4: InChIs that are not fully determined.

Number of InChIs	with *	with R	with Z[z]
803,094	1,933	0	0

Table 5: SMILES that are not fully determined.

Number of SMILES	with *	with R	with Z[z]
999,849	196,889	0	0

There are 0 duplicated InChIs.

Annotation

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

Identifiers	Unique Namespaces
2,462,968	18

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

Namespace	Frequency
bigg.metabolite	18,217
chebi	134,607
envipath	12,306
hmdb	195,008
kegg.compound	37,346
kegg.drug	22,294
kegg.environ	1,728
kegg.glycan	22,084
lipidmaps	43,085
metacyc.compound	20,296
metanetx.chemical	1,094,145
pubchem.compound	13
reactome	5,526
rhea.generic	1,494
rhea.polymer	228
sabiork.compound	8,944
seed.compound	67,990
slm	777,657

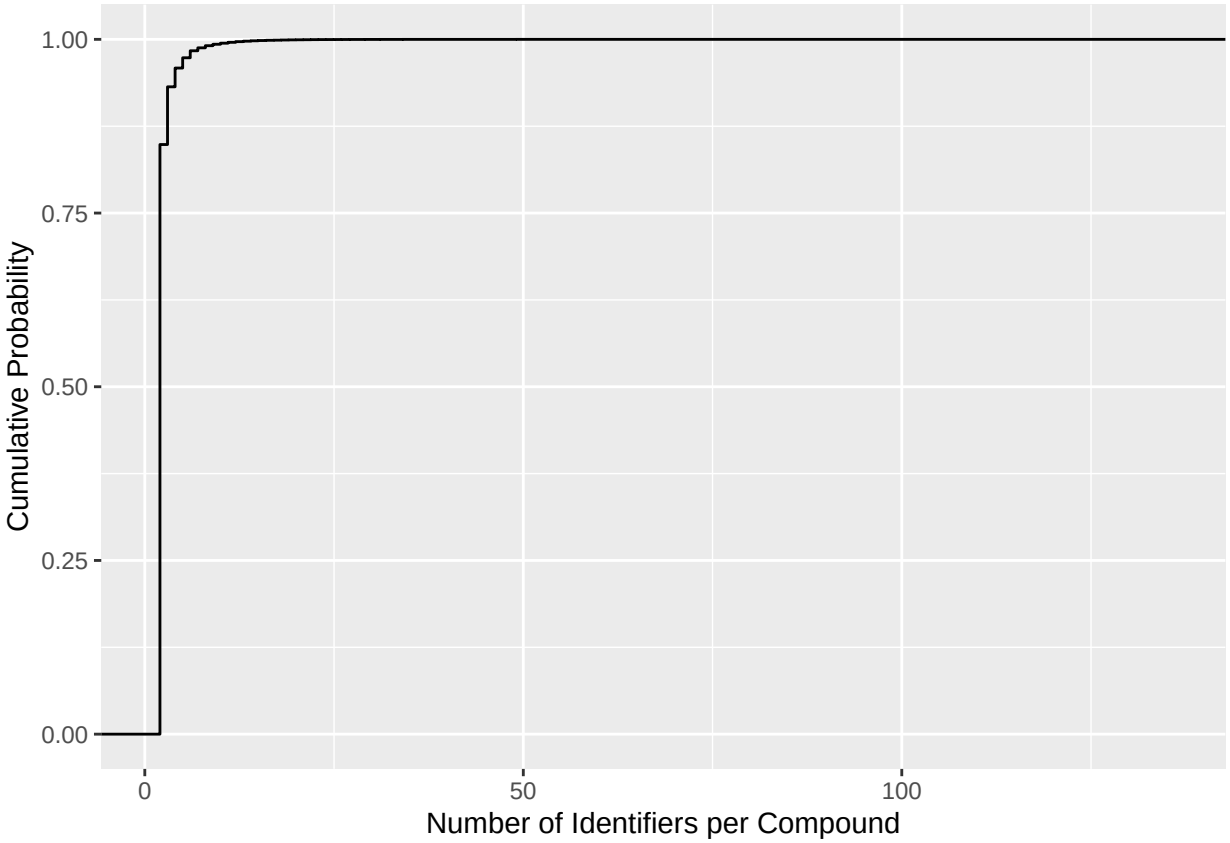

Figure 3: The empirical cumulative distribution function (eCDF) of the number of distinct identifiers per compound.

Names

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

Names	Unique Namespaces
4,296,792	18

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

Namespace	Frequency
bigg.metabolite	18,017
chebi	362,610
envipath	12,375
hmdb	1,355,206
kegg.compound	49,563
kegg.drug	31,027
kegg.environ	2,052
kegg.glycan	22,283
lipidmaps	122,097

Namespace	Frequency
metacyc.compound	45,598
metanetx.chemical	3
pubchem.compound	609
reactome	14,171
rhea.generic	1,494
rhea.polymer	228
sabiork.compound	13,669
seed.compound	113,807
slm	2,131,983

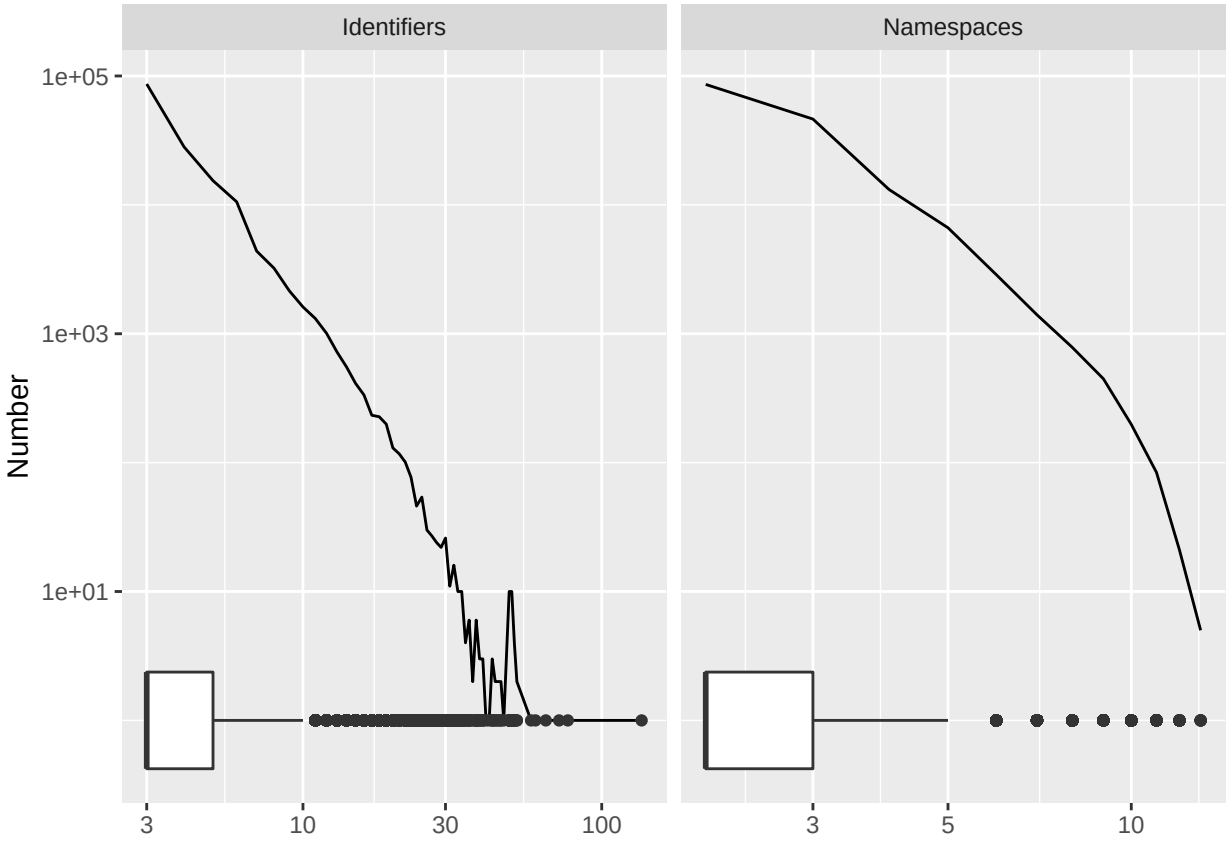

Figure 4: The number of identifiers per compound and the number of unique source namespaces per compound. Only compounds that have more than two identifiers are included.

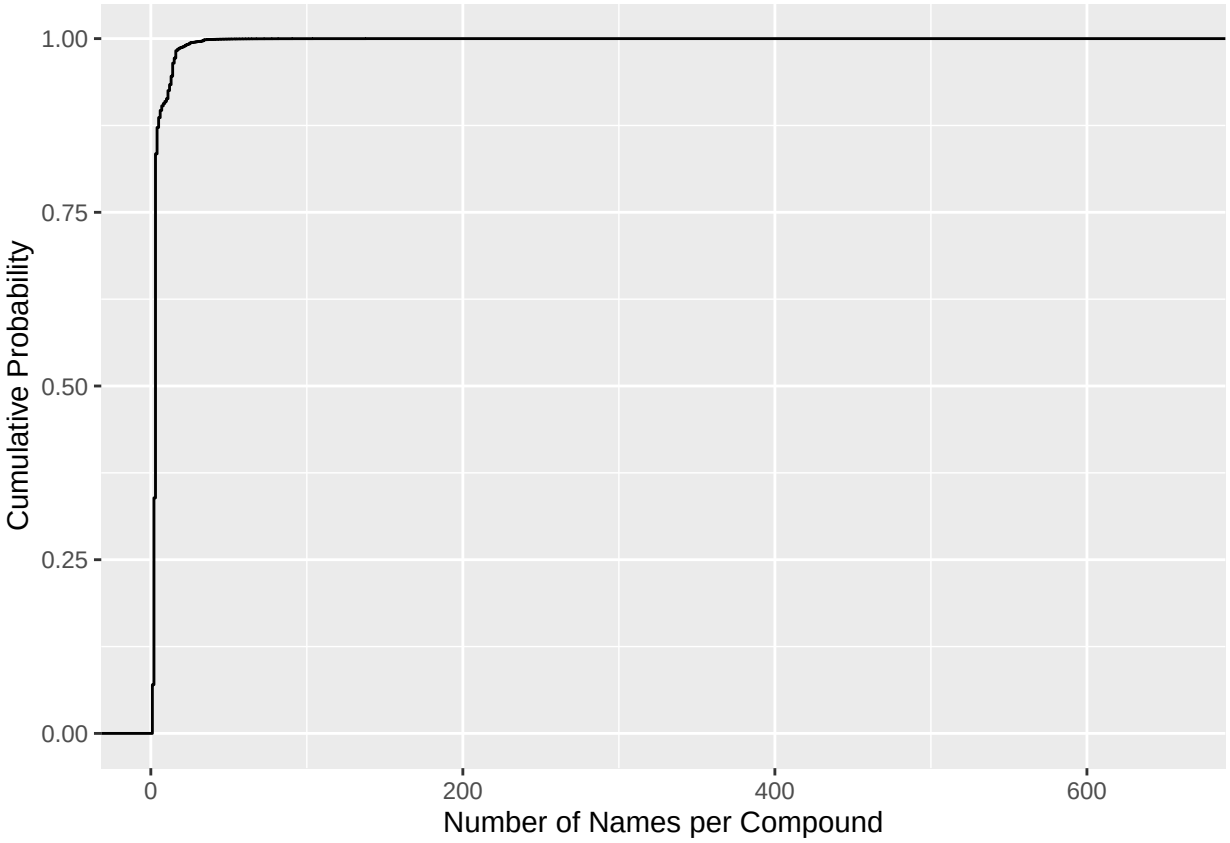

Figure 5: The empirical cumulative distribution function (eCDF) of the number of distinct names per compound.

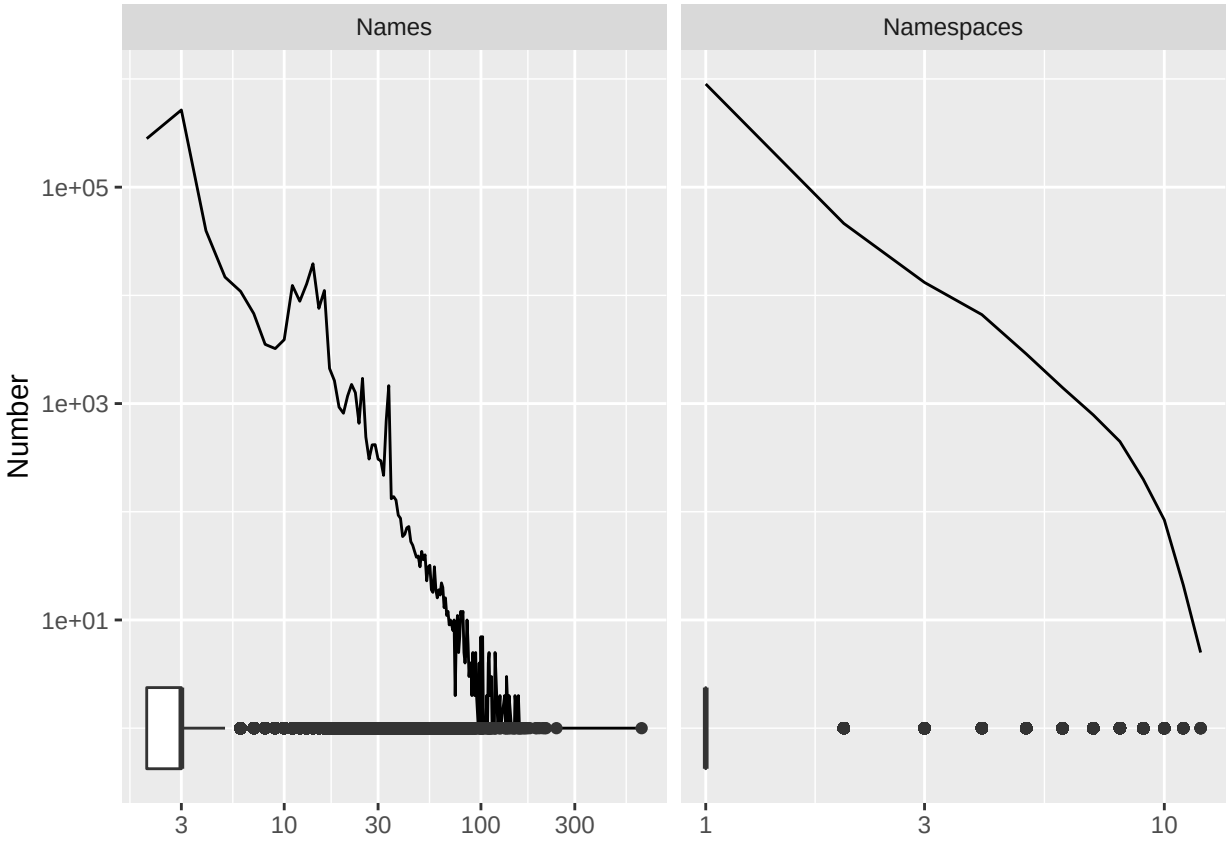

Figure 6: The number of names per compound and the number of unique source namespaces per compound. Only compounds that have more than one name are included.