# 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 |
|---:|---:|---:|---:|---:|---:|---:|
| 41,531 | 26,974 | 26,974 | 26,953 | 26,911 | 20,376 | 20,376 |

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% | 64.95% | 64.95% | 64.90% | 64.80% | 49.06% | 49.06% |

Table 3: Chemical formulae that are not fully determined.

| Number of Formulae | with `*` | with `R` | with `Z[z]` |
|---:|---:|---:|---:|
| 26,974 | 6,533 | 0 | 0 |

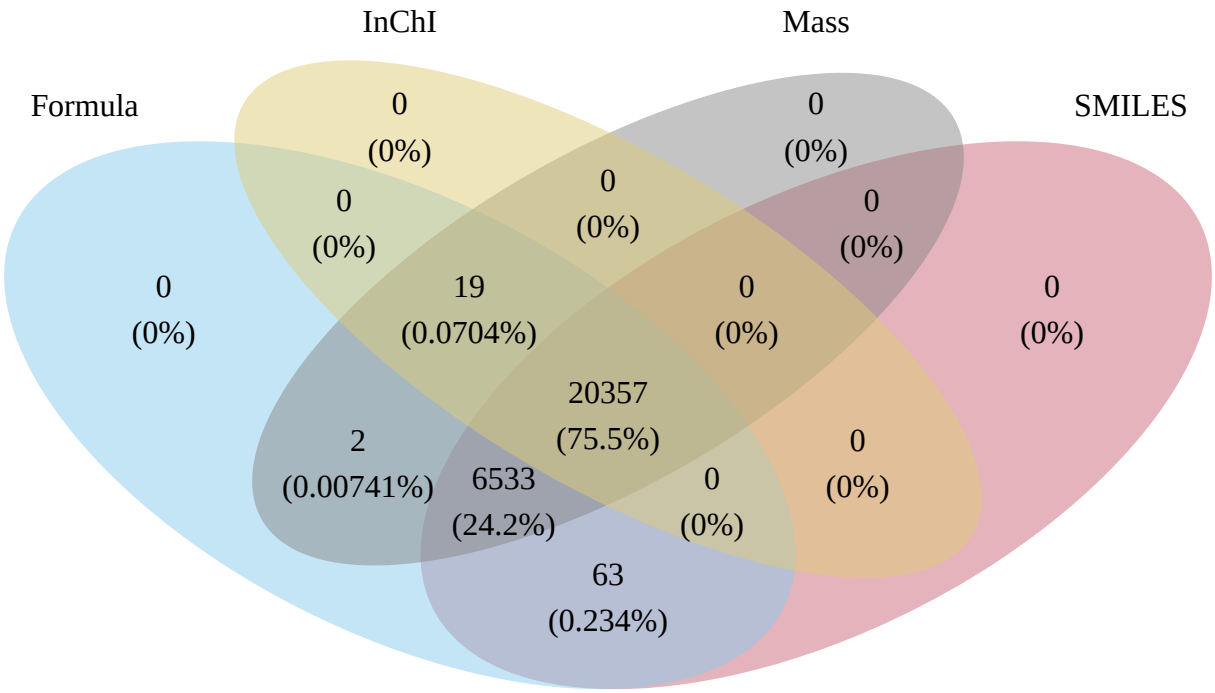


Figure 1: Venn diagram of structural annotation and mass.

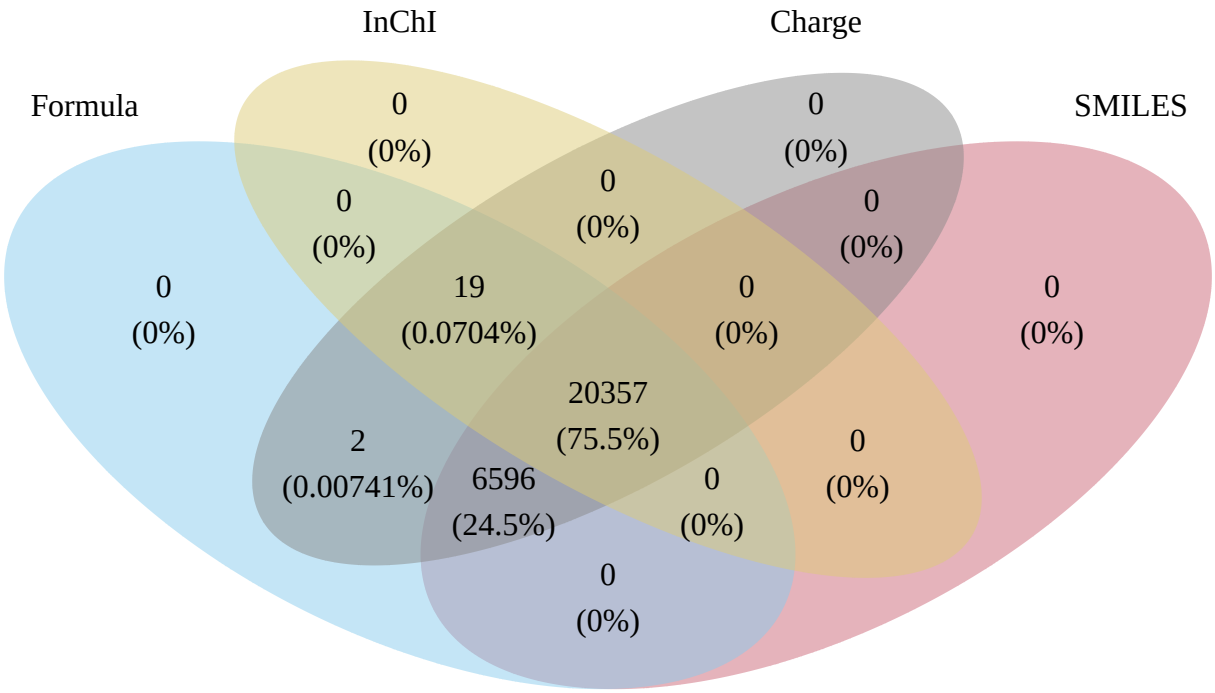


Figure 2: Venn diagram of structural annotation and electric charge.

Table 4: SMILES that are not fully determined.

| Number of SMILES | with `*` | with `R` | with `Z[z]` |
|---|---|---|---|
| 26,953 | 6,596 | 0 | 0 |

Table 5: InChIs that are not fully determined.

| Number of InChIs | with `*` | with `R` | with `Z[z]` |
|---|---|---|---|
| 20,376 | 67 | 0 | 0 |

## Annotation

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

| Identifiers | Unique Namespaces |
|---|---|
| 226,334 | 16 |

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

| Namespace | Frequency |
|---|---|
| bigg.metabolite | 18,063 |
| chebi | 35,033 |
| envipath | 1,580 |
| hmdb | 9,383 |
| kegg.compound | 19,822 |
| kegg.drug | 1,318 |
| kegg.glycan | 812 |
| lipidmaps | 2,832 |
| metacyc.compound | 16,202 |
| metanetx.chemical | 63,561 |
| 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 |
|---|---|
| 386,564 | 16 |

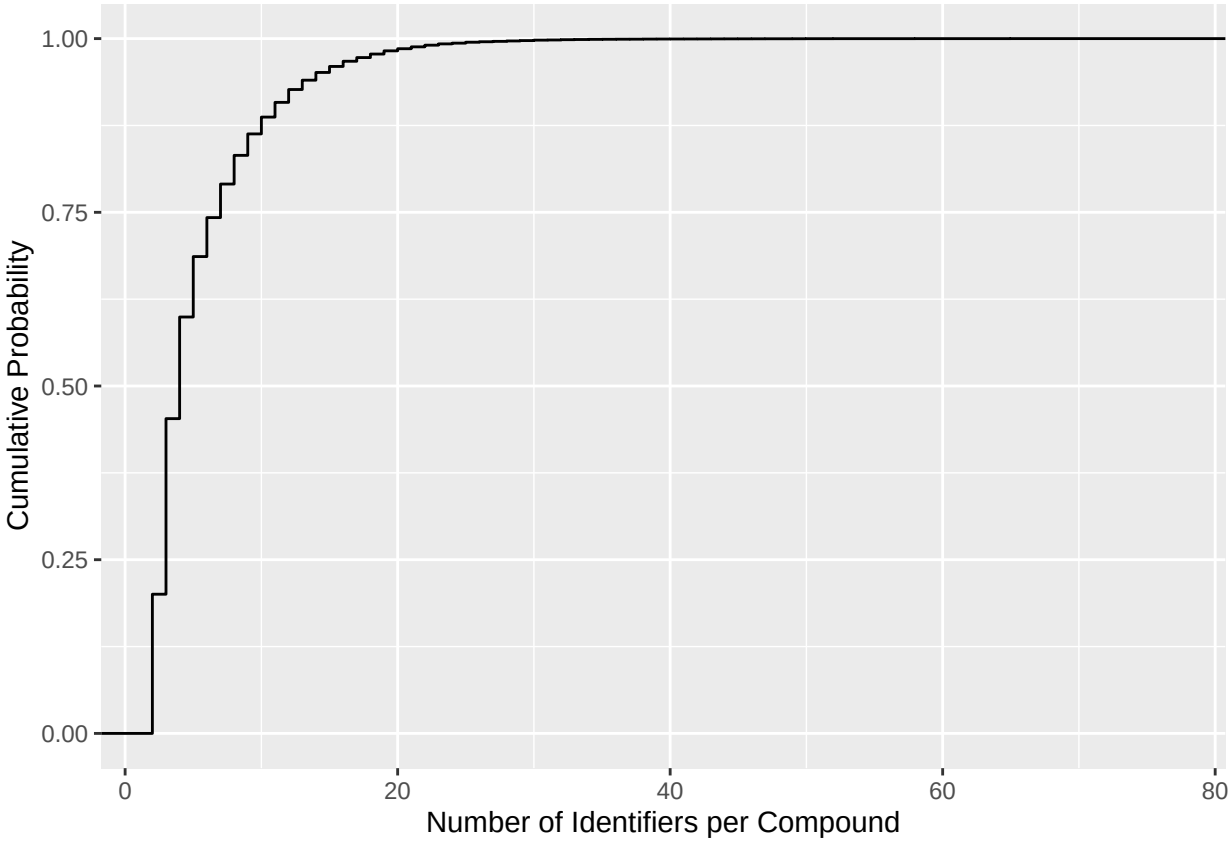


Figure 3: The empirical cumulative distribution function (eCDF) of the number of distinct identifiers per compound.

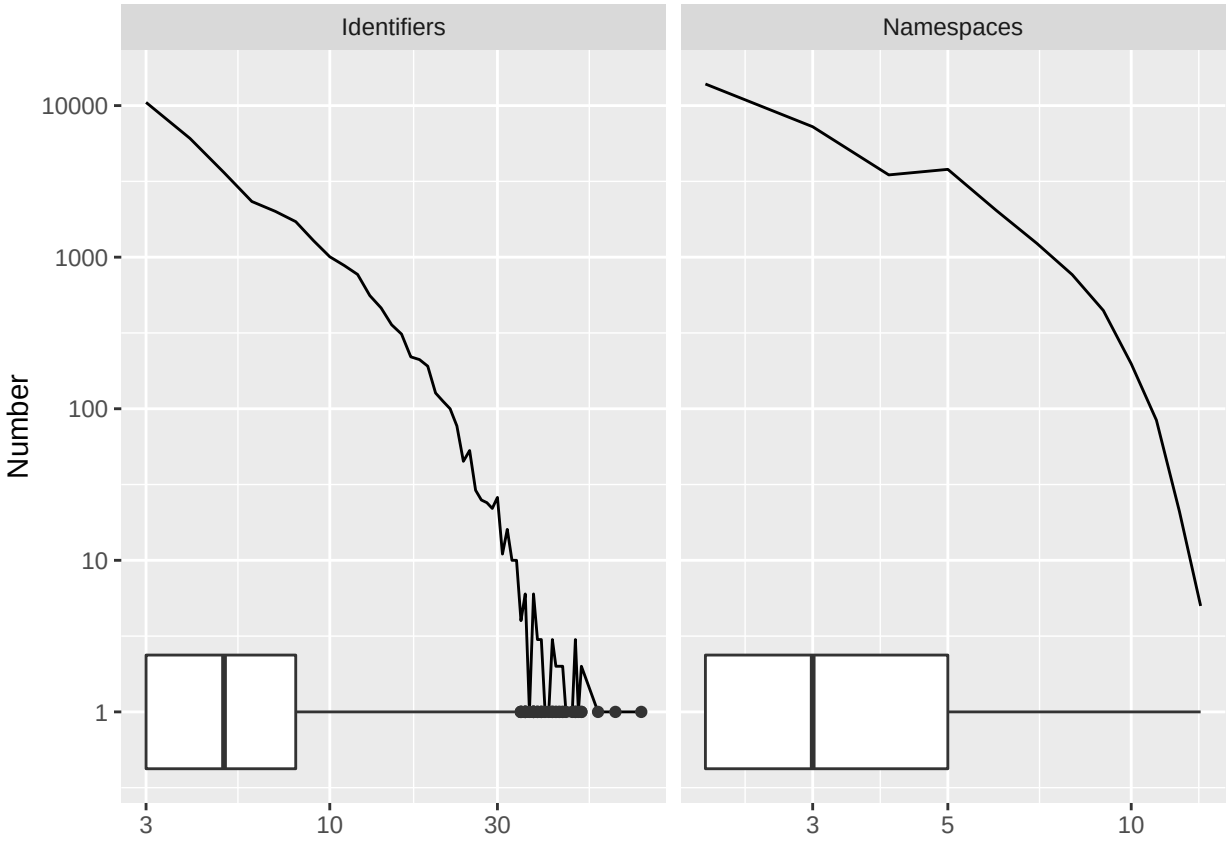


Figure 4: The number of distinct identifiers per compound and the number of unique source namespaces per compound. Only compounds that have more than two identifiers are included.

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

| Namespace | Frequency |
|---|---|
| bigg.metabolite | 17,863 |
| chebi | 102,621 |
| envipath | 1,576 |
| hmdb | 77,980 |
| kegg.compound | 28,491 |
| kegg.drug | 2,168 |
| kegg.glycan | 927 |
| lipidmaps | 9,285 |
| metacyc.compound | 37,226 |
| metanetx.chemical | 2 |
| reactome | 8,469 |
| rhea.generic | 1,490 |
| rhea.polymer | 195 |
| sabiork.compound | 13,594 |
| seed.compound | 80,667 |
| slm | 4,010 |

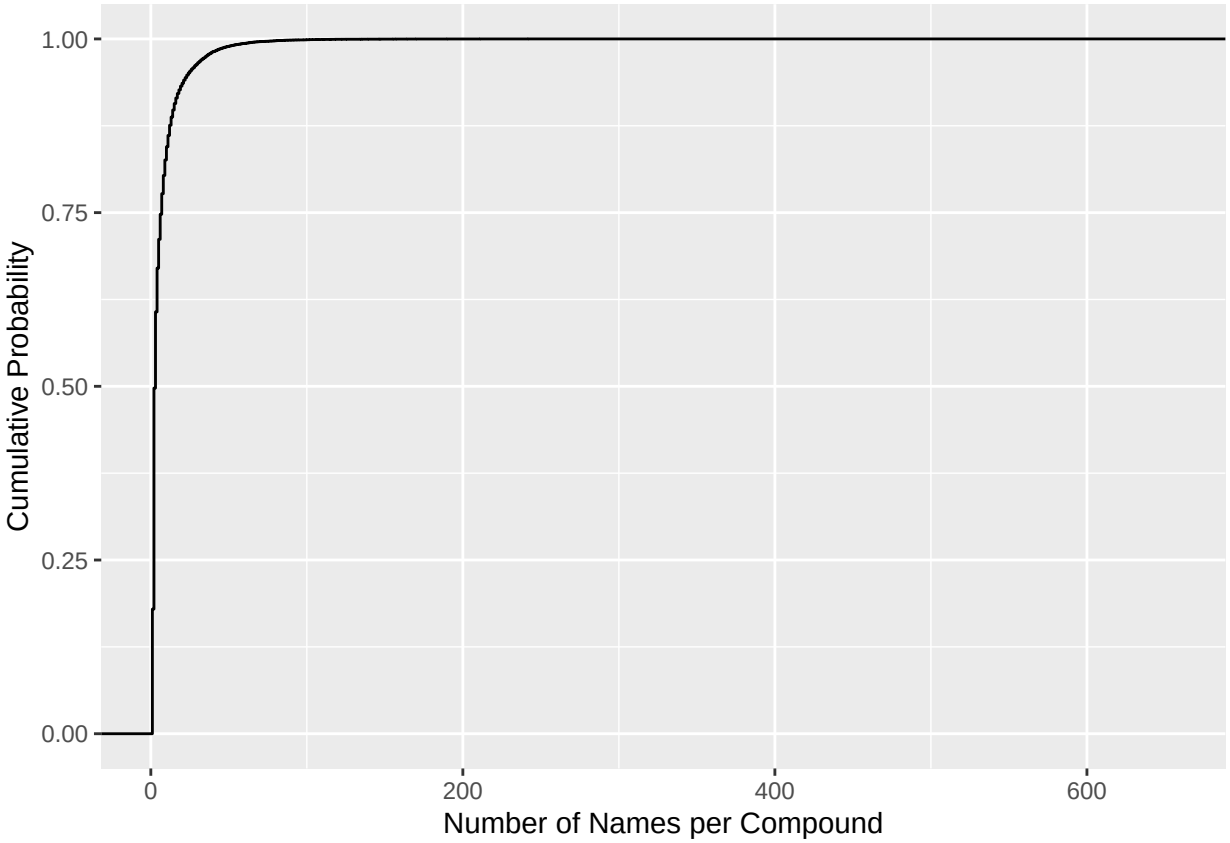


Figure 5: The empirical cumulative distribution function (eCDF) of the number of distinct names per compound.

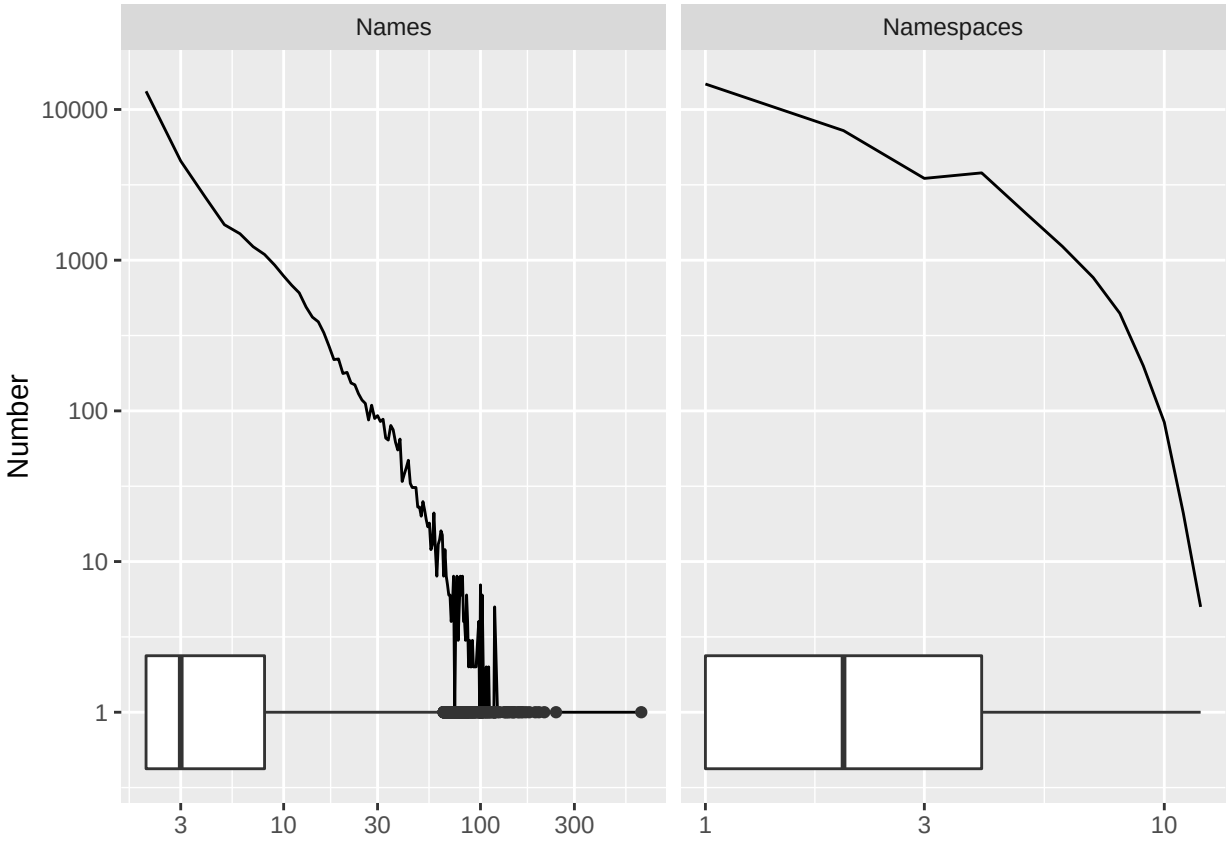


Figure 6: The number of names per compound and the number of unique source namespaces per compound. Only compounds that have more than one name are included.