NANPDB: A Novel Resource for Natural Products from Northern African Sources

Fidele Ntie-Kang[†,‡,O*], Kiran K Telukunta[§,O], Kersten Döring[§], Conrad V Simoben[†], Aurélien FA Moumbock[‡], Yvette I Malange[‡], Leonel E Njume[‖], Joseph N Yong[‡], Wolfgang Sippl[†] and Stefan Günther[§,▽*]

[†]Department of Pharmaceutical Chemistry, Martin-Luther University of Halle-Wittenberg, Wolfgang-Langenbeck Str. 4, 06120 Halle (Saale), Germany.

[‡]Department of Chemistry and [‖]Chemical and Bioactivity Information Centre, Department of Chemistry, Faculty of Science, University of Buea, P.O. Box 63, Buea, Cameroon.

[§] Institute of Pharmaceutical Sciences, Research Group Pharmaceutical Bioinformatics, Albert-Ludwigs-University Freiburg, Hermann-Herder-Str. 9, 79104 Freiburg, Germany.

[▽]Freiburg Institute for Advanced Studies (FRIAS), University of Freiburg, Albertstraße 19, 79104 Freiburg i.B, Germany.

[O] These authors contributed equally and would like to be regarded as joint first authors

[*] Corresponding authors

ntiekfidele@gmail.com (FNK); stefan.guenther@pharmazie.uni-freiburg.de (SG)

SUPPLEMENTARY DATA

Table S 1: List of biological activities recorded in NANPDB.

Activity	Count actives	% Composition
adipocyte differentiation-inhibition	11	0.6773399
allelopathic	2	0.12315271
antiallergic	1	0.06157635
antiallodynic	1	0.06157635
antiangiogenic	1	0.06157635
anti-asthmatic	1	0.06157635
anticancer	214	13.1773399
anticholinesterase	12	0.73891626
antidiabetic	28	1.72413793
antiepileptic	2	0.12315271
antifeedant	69	4.24876847
antifouling	4	0.24630542
antigenotoxic	1	0.06157635
antiinflammation	99	6.09605911
antiinvasive	2	0.12315271
antimicrobial	419	25.8004926
antimigratory	15	0.92364532
antimutagenic	2	0.12315271
antinociceptive	1	0.06157635
antiobesity	9	0.55418719
antioxidant	191	11.7610837
antiprotozoal	57	3.50985222
antipruritic	1	0.06157635
antiseptic	2	0.12315271
antiulcerogenic	2	0.12315271
binding assays against the following receptors; somatostatin, human B2 bradykinin and neuropeptide gamma receptors	5	0.30788177
bitter principle	1	0.06157635
bradycardia	3	0.18472906
cardiac activity	3	0.18472906
chemo-preventive	2	0.12315271
cytostatic	2	0.12315271
cytotoxic	261	16.0714286
GABA induction	1	0.06157635
germination stimulant	1	0.06157635
haemorrhagic	2	0.12315271
hepatoprotective	21	1.29310345

hypotensive	3	0.18472906
immunomodulation	2	0.12315271
Inhibit various human leukocyte functions	4	0.24630542
inhibition of the Hsp90 machine chaperoning activity	1	0.06157635
inhibition of TNF-alpha release	3	0.18472906
inhibitory activity toward advanced glycation end-products formation	3	0.18472906
p56lck tyrosine inhibition	2	0.12315271
nitric oxide production inhibition	15	0.92364532
insecticidal	5	0.30788177
kinase inhibition	20	1.23152709
methylation	1	0.06157635
molluscidal	39	2.40147783
mutagenicity	2	0.12315271
Na+/K+-ATPase inhibition	8	0.49261084
oxidative burst inhibition	2	0.12315271
parasympatholytic effect	2	0.12315271
phospholipase A2 inhibition	1	0.06157635
phytotoxic activity	20	1.23152709
prevention of lipid peroxidation and inhibition of myeloperoxidase	1	0.06157635
pro-inflammation	2	0.12315271
protective effect against CCl4-induced injury on the human hepatoma cell line (Huh7)	10	0.61576355
spasmolytic	3	0.18472906
urease inhibition	2	0.12315271
vascular activity	1	0.06157635
vasodepression	3	0.18472906
vasorelaxation	19	1.16995074
zootoxicity	5	0.30788177

Notes S1: Notes for completing spreadsheet submission file

Dear potential collaborator,

Kindly read carefully and follow instructions below when filling sample excel file for submission pipeline.

-On the 1st sheet of the spreadsheet, you will find examples of how the spreadsheet is completed, shown for some selected organisms with family names beginning with letter C (a few examples that we have completed). You can assess some of the reference papers or request them from us, in case you cannot assess them. On the 2nd sheet of the spreadsheet, you are to provide information for data entries you plan to submit to NANPDB. We kindly suggest that you first ensure that such data do not currently exist in NANPDB before spending useful efforts. The following notes would be of help. The examples were provided to assist you in the case of difficulty. The following are few examples:

- **Compound Code:** The first three characters of the family of the organism from which compound was identified, followed by an underscore and a five-code number (in chronological order), e.g. the first compound from the family Cactaceae should be CAC_00001, the second compound CAC_00002, etc. You should endeavor to maintain the same order (or numbering) in publications, to ease the work. In your case, I would suggest you use a different coding according to your initials, e.g. the first compound code from James Brown could be JB_00001

- **Compound Name:** It must begin with lowercase letters. Greek letters must be written out, not the Greek characters, e.g. α as alpha, Δ as Delta, etc. By looking through the work of family C, you will have an idea of what I did. N.B: To assist you, copying from the pdf document and pasting on WordPad, then copying from WordPad/NotePad and pasting on the spreadsheet will help. However, you must be careful, since copying and pasting compound names can be tricky. You must always check at the end that the compound names are correct. This applies to all other information you copy from pdf documents.

- **PubChem ID:** PubChem IDs of the compounds (obtainable from https://pubchem.ncbi.nlm.nih.gov/). You simply type (or copy and paste) the names of the compounds on this site and get the PCID code. Verify that the ones I did for families C are correct.

- **Canonical Smiles Code:** leave that to us. The canonical smiles will be generated later using OpenBabel software. If you provide a PubChem ID of a molecule structure file, our database team will use this to generate the SMILES strings.

- **Non-Canonical Smiles Code:** Leave that to us.

- **Source Species (species name):** Most often, this information is available from the title or the abstract. You have to read carefully, as in some cases, several species are used in a single study. In other cases, endophytes or symbiotic organisms growing on plants are studied. You must be certain about which organism was the source of the compound(s) in this case. Reading the experimental section will sometimes help clarify.

- **Known Uses of Source Species:** Sometimes given in the introduction (sometimes detailed, sometimes not). When several papers report work on the same species, you can get a more comprehensive summary only after reading through all the reports.

- **Part of Organism Studied:** Easily found in the title, introduction, or experimental section.

- **Kingdom:** Animalia, Plantae, Fungi, etc.

- **Family:** Easily found in the title, introduction, and on the list of keywords or experimental section. Be careful! Family names could be tricky, depending on the classification system and how old the publication is. Taking the names I adopted in the naming of the pdf files may sound simpler.

- **Source Availability (Country):** Often available from the experimental description of the plant collection. When this is not available, indicate by " Not specified".

- **Source Reference Number:** Often available from the experimental description of the plant collection. When this is not available, indicate by " No reference".

N.B: Sometimes, the Source Availability and Source Reference Number are reported elsewhere. I have access to most journal articles and can easily check this myself. In such cases, specify to me in an email which of source references should be checked and provide the links to the requested papers.

- **Compound Class:** Write with uppercase letters, e.g. Terpenoid, Steroid, Coumarin, Lignan, Phenolic, Flavonoid, etc. A basic knowledge of natural product chemistry is required.

- **Compound Subclass:** Write with lowercase letters, e.g. sesquiterpenoid, sterol, flavane, etc. A basic knowledge of natural product chemistry is required, a clear difference between compound classes and subclasses must be made, e.g. triterpene is a subclass,

while terpenoid is a major class. It is sometimes easy to make mistakes by leaving the final letter "s" in compound classes in the title, e.g. the compound subclass can be sesquiterpene lactone, not sesquiterpene lactones. Such information is sometimes provided in the abstract, but if you doubt, ask us.

- **Known Biological Activity:** Sometimes provided in the title, abstract or results, and discussion. Be careful! Sometimes the title could be misleading, e.g. Cytotoxic sesquiterpenes from ..... This does not mean that:

    1. All isolated compounds are sesquiterpenes, careful reading may prove this wrong.

    2. All compounds have the biological activity Cytotoxic. By reading the text of results (sometimes results are presented in tables), you will be careful to know which compounds actually exhibited this activity.

- **Mode of Action:** Sometimes provided in the title, abstract or results and discussion, e.g. a compound may be an antidiabetic and inhibition of alpha-glucosidase activity. In such a case, the biological activity is "antidiabetic", while the mode of action is "inhibition of alpha-glucosidase activity".

- **Source Country:** Get information from data on plant collection (experimental), not from author affiliation, as this could be misleading. Authors from Egypt may have studied a plant collected from Saudi Arabia, for example.

- **Place of Collection:** Such information is often available from data on plant collection (experimental). The specific location is important here, e.g. 20 km from Cairo, West of Rabat, a village market in Algiers, etc.

- **GPS Coordinates:** Sometimes this is provided in the experimental section. If not, simply specify " Unavailable".

- **Collection Date:** The writing of dates must be uniform, so in case the date of collection was 23 September 2013, it is written as 9/23/2013. In case the date is not completely written in the article, e.g. "in September 2013", the first date of that month is used, so we write 9/1/2013 but include a comment on column Z = Comment 2 (Collection Date and Place) like this "Reported in literature source as September 2013.". In the case where only the year is mentioned, e.g. collected in 2013, we use the first day of that year 1/1/2013. When nothing is mentioned, we take the collection date as exactly 1 year before the date of submission of the article and we write a comment on column Y as "Data for sample collection is not reported in literature source. The data curator estimated as 1 year before submission of the publication." Of course, if the exact date is reported in the literature source, use that date in format DD/MM/YYYY and leave no comment in column Z.

- **Author(s):** All author names for all papers must be included, following one unique format, e.g. Asia NS, Maribo GHK, Taiwoo RT, Su W. This will permit the filtering for author names and include all authors in the search, not just the first author. Author names are written as Surnames followed by abbreviations of all other names, e.g. Sara Hoet is Hoet S, not Sara H. Also carefully check author names as you do copy and paste, e.g. Ganfon H is not Ganfona H. The last letter a must have infiltrated from institutional superscripts in the header.

- **Reference Type:** This could be Journal article, MSc thesis, PhD thesis, Conference report, etc.

- **Reference:** This must be unique; Journal name,year,volume(issue),page numbers. Please, no spaces! Examples are shown in Families C. No abbreviations of journal names.

- **Title:** Be uniform! Only first word in Title, proper nouns and genus names are capitalized, e.g. only the words Antiinflammatory, Opuntia, Ker-Gawl, Haw. and Egypt are capitalized in the following title: Antiinflammatory flavonoids from Opuntia dillenii (Ker-Gawl) Haw. flowers growing in Egypt.

- **Comment 1 (Source Organism):** These could be comments on, for example, alternative names of species, e.g. the species Callyspongia siphonella is also known as Siphonochalina siphonella.

- **Comment 2 (Collection Date and Place):** See previous comments on collection dates.

- **Comment 3 (Compound Type):** Here we specify if this compound, scaffold, compound type, etc. was isolated from the species (genus, family, or natural source) for the first time, etc.

- **Comment 4 (PMID):** The PubMed ID of the reference is obtainable from the official site http://www.ncbi.nlm.nih.gov/pubmed. If you can't find this (due to obvious slow connections in Cameroon), leave it to me.

- **Comment 5 (Literature Source):** Comments on literature source could include personal communication with authors of a paper, names of journals that have changed, etc., e.g. Natural Product Research was "Formerly Natural Product Letters".

- **Comment 6 (Data Curator):** This could be James Brown, 2015

- **Comment 7 (Literature Source Link):** You can provide an internet link or URL where the literature reference could be assessed. Alternatively, you could provide a doi for the reference.

- **Source species alternative names:** Provide alternative names (scientific or common names) of the source organism from which the compound was identified.
- **Link to Taxonomy data for source species:** Such information could be a link from the NCBI Taxonomy database (http://www.ncbi.nlm.nih.gov/taxonomy), WoRMS database (http://www.marinespecies.org), MycoBank (http://www.mycobank.org), Tropicos (http://www.tropicos.org/), etc.
- **Link to Wikipedia (source species):** You can provide a Wikipedia link where additional information about the source species could be retrieved.
- **Additional comments:** Depends on the contributor
- **Repetitions:** When we note that the same compound is repeated (from compound names or synonyms, we include each synonym as a separate entry but include a comment under canonical SMILES generation. This is a reason to avoid wasting time in generating the SMILES for the same compound (essentially the same entity many times). We do the same when the same compound is isolated from different species of for the same species but under different studies.

P.S:

On the third page of the Excel submission file, kindly provide additional useful notes or any explanations that you think the database worker may find useful.

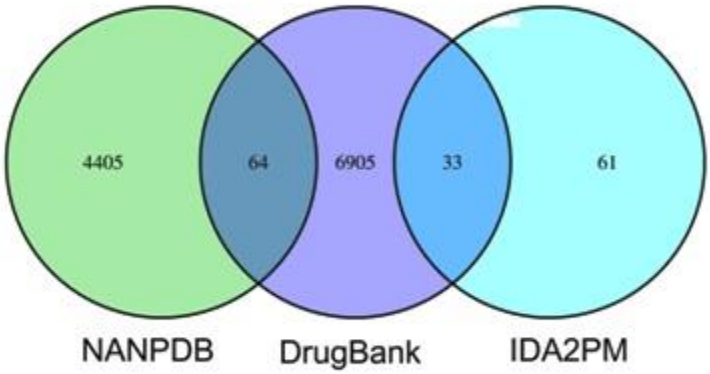

Fig. S1: Venn diagram showing the intersection of NANPDB with known drugs.

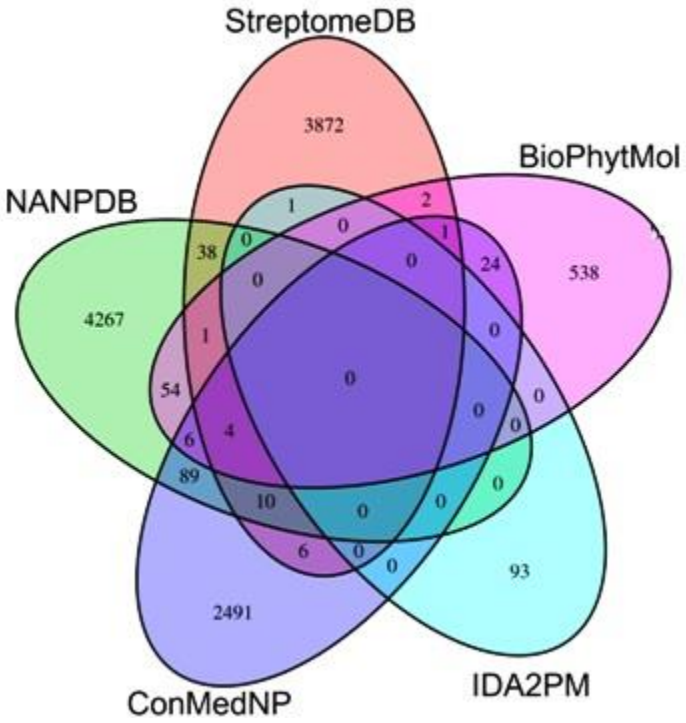

Fig. S2: Venn diagram showing the intersection of NANPDB with NPs and known drugs.
(.PNG)

Table S2: List of Journals Consulted in Constructing The Northern African Natural Products Database (NANPDB)

Journal type	List
International	*Acta Crystallographica, African Journal of Health Sciences, Asian Journal of Traditional Medicine, Arabian Journal of Chemistry, Crystallographica, Asia-Pacific Journal of Tropical Medicine, Biochemical Systematics and Ecology, Bioorganic and Medicinal Chemistry, Bioorganic and Medicinal Chemistry Letters, Bioscience Biotechnology and Biochemistry, Biological and Pharmaceutical Bulletin, BMC Complementary and Alternative Medicine, BMC Research Notes, Boletín Latinoamericano y del Caribe de Plantas Medicinales y Aromáticas, Bulletin of the Chemical Society of Ethiopia, Carbohydrate Research, Chemistry and Biodiversity, Chemical and Pharmaceutical Bulletin, Chemistry of Natural Compounds, Chinese Chemical Letters, Evidence Based Complementary and Alternative Medicine, Fitoterapia, Helvetica Chimica Acta, Inflammopharmacology, Journal of Natural Products, Journal of Asian Natural Products Research, Journal of Ethnopharmacology, Journal of Medicinal Plants Research, Journal of the American Chemical Society, Journal of the American Oil Chemistry Society, Journal of Organic Chemistry, Journal of Pharmacognosy and Phytotherapy, Malaria Journal, Molecules, Natural Product Communications, Natural Product Letters, Natural Product Research, Natural Product Science, Organic and Medicinal Chemistry Letters, Pakistani Journal of Medical Science, Parasitology Research, Pharmaceutical Biology, Pharmacologia, Pharmacologyonline, Pharmazie, Phytochemical Analysis, Phytochemistry, Phytochemistry Letters, Phytopharmacology, Pharmaceutical Biology, Phytotherapy Research, Phytomedicine, Phytomedicine Reserach, Planta Medica, Planta Medica Letters, PLoS One, Pure and Applied Chemistry, Rasayan Journal of Chemistry, Records of Natural Products, Research Journal in Phytochemistry, Research Journal in Medicinal Plants, RSC Advances, South African Journal of Botany, Tetrahedron, Tetrahedron Letters* and *Zeitschrift für Naturforschung*.

Local journals	*Bulletin of Pharmaceutical Science of Assiut University*

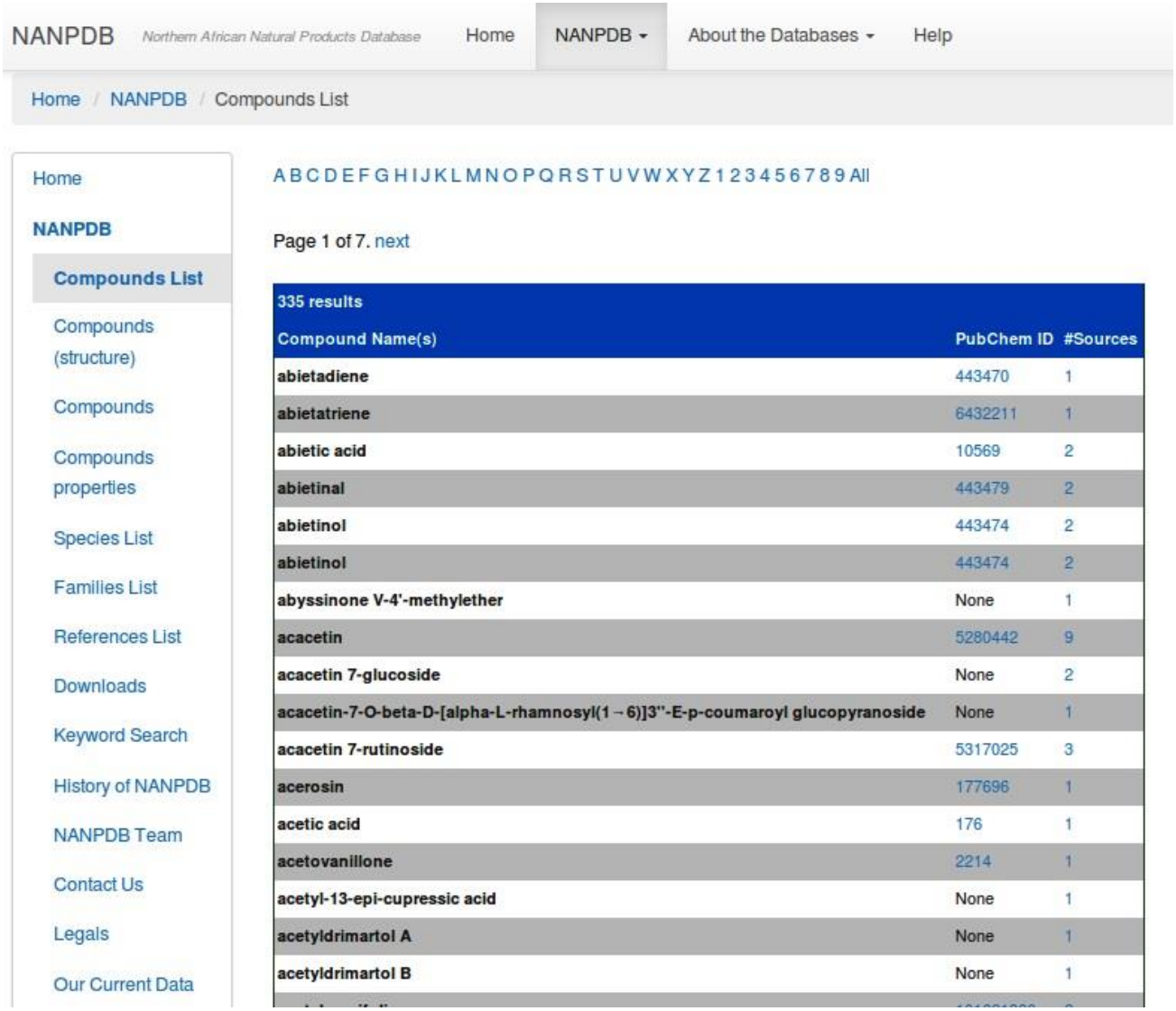

Fig. S3: The graphical user interface of NANPDB. Additional information is available at http://african-compounds.org/nanpdb/help/

Table S3 Comparison of ro5 parameters of NANPDB with approved drugs and five other NP datasets.

Dataset	MW			clogP			HBA		
	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg
NANPDB	1,385.51	55.08	419.79	20.05	-9.88	2.15	53	0	9.06
IDA2PM	1,626.24	17.03	370.13	16.71	-12.49	2.16	64	0	6.98
DRUGBANK	1,150.19	17.03	341.10	17.84	-16.04	1.56	69	0	7.23
NuBBE	1,171.04	86.09	369.45	18.19	-5.31	3.22	38	0	6.29
StreptomeDB 2.0	1,473.69	17.03	514.75	19.21	-18.67	1.03	71	0	12.51
ConMedNP	1,439.60	84.16	426.70	22.29	-6.80	3.86	70	0	5.85
BioPhytMol	1,084.73	74.08	347.43	13.45	-5.41	3.92	39	0	4.73
NPACT	1,383.53	106.12	441.94	18.46	-6.32	3.19	51	0	8.20

Dataset	HBD			NRB			NLV		
	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg
NANPDB	21	0	3.02	50	0	7.81	3	0	0.83

IDA2PM	19	0	2.17	52	0	6.81	4	0	0.45
DRUGBANK	25	0	2.65	62	0	6.96	4	0	0.42
NuBBE	16	0	1.59	30	0	5.99	3	0	0.54
StreptomeDB2.0	29	0	3.84	66	0	12.72	4	0	1.29
ConMedNP	37	0	2.39	56	0	5.51	4	0	0.71
BioPhytMol	17	0	1.12	51	0	5.78	4	0	0.46
NPACT	18	0	2.54	44	0	8.41	4	0	0.79

MW: Molecular weight; clogP: calculated logarithm of *n*-octanol/water partition coefficient; HBA: number of hydrogen bond acceptors; HBD: number of hydrogen bond donors; NRB: number of rotatable single bonds; NLV: number of violations of Lipinski rules; Max: maximum value; Min: minimum value; Avg: mean value.

Table S4 Maximum, minimum and mean predictions for selected toxicity parameters for NANPDB.

Parameter	**Maximum**	**Minimum**	**Mean**
Human maximum tolerated dose (log mg/kg/day)[a]	2.21	-1.14	0.34
Rat oral acute toxicity (mol/kg)	4.11	0.49	2.23

Rat oral chronic toxicity (log mg/kg_bw/day)	3.84	-2.57	1.63
T. pyriformis toxicity (log μg/L)[b]	2.57	-1.70	0.88
Minnow toxicity (log mM)[c]	5.28	-4.70	1.25

[a] ≤ 0.477 log mg/kg/day is considered to be low, while > 0.477 log mg/kg/day is considered to be high; [b] < -0.5 log μg/L is considered to be toxic; [c] log LC_{50} values < -0.3 indicate high acute toxicity.

MRTD is considered low when ≤ 0.477 log mg/kg/day and high > 0.477 log mg/kg/day. Based on the mean value of 0.34 log mg/kg/day, the predictions showed that about 60% of NANPDB compounds showed low MRTD, being much lower than the threshold value of 0.477 log mg/kg/day. The oral rat acute toxicity expresses the toxic potency of a compound in terms of the lethal dosage values (LD_{50} in mol/kg), i.e. the amount of a compound administered as a single dose, which causes the death of 50% of a group of test animals. For toxicity against the protozoan *T. pyriformis*, a compound with a predicted $pIGC_{50}$ value (negative logarithm of the concentration required to inhibit 50% growth in log μg/L) < -0.5 log μg/L is considered to be toxic. This corresponds to only about 1% of NANPDB compounds. For fish (Fathead Minnows), an equivalent lethal concentration value (LC_{50}), representing the concentration of a molecule necessary to cause the death of 50% of experimentally tested Flathead Minnows, LC_{50}< 0.5 mM (i.e. log LC_{50}< -0.3) are regarded to cause high acute toxicity. This corresponds to only about 6% of NANPDB compounds. In both cases, the mean $pIGC_{50}$ and log LC_{50} values (0.88 log μg/L and 1.25 log mM respectively, Table 5) are both far from the threshold values of -0.5 log μg/L and -0.3 log mM, respectively.

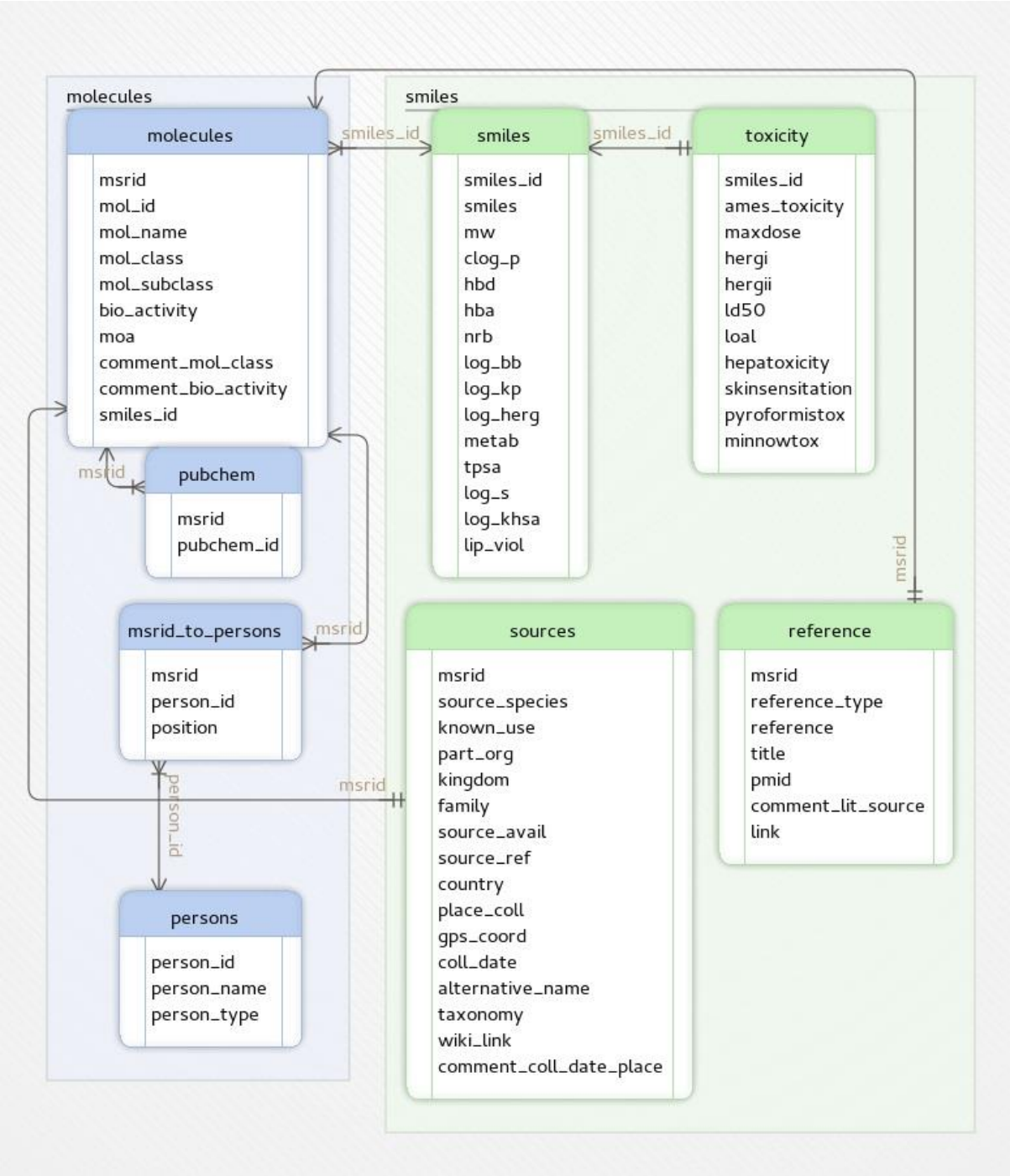

Figure S4 Detailed database scheme showing the main SQL tables and the connection tables. (A) the *molecules* table contains information about the molecule ID, molecule name, PubChem ID, SMILES string, molecule class, and subclass, known biological activities and modes of action, as well as additional comments and computed physicochemical properties data, e.g. molecular weight, *c*log*P*, etc.; (B) the

sources table includes the species names, their known uses, part of species where the molecule was identified, the species kingdom and family, where a sample of the species has been stored and the reference number, the country, locality and GPS coordinates of the venue of collection, collection date, alternative names of source species, links to taxonomic data of species and link to Wikipedia; (C) the *reference* table contains information about the reference types, the references, the titles and PubMed IDs of the literature references as well as additional comments and the links to the references; (D) the *persons* table contains information about the person name and type (author of literature reference or data curator); (E) the SMILES table prevents redundancy in the molecules table; (F) the toxicity table contains toxicity prediction data. In all entries, the Molecule-Source-Reference ID (*msrid*) is a unique entry code that serves as a unique entry and connection between all related data in the four tables (A to D) and related to the last two tables (E and F) *via* the *smiles id* contained in a connection table.