RESEARCH ARTICLE

SANCDB: an update on South African natural compounds and their readily available analogs

Bakary N'tji Diallo[1], Michael Glenister[1], Thommas M. Musyoka[1], Kevin Lobb[1,2] and Özlem Tastan Bishop[1*]

Abstract

Background: South African Natural Compounds Database (SANCDB; https://sancdb.rubi.ru.ac.za/) is the sole and a fully referenced database of natural chemical compounds of South African biodiversity. It is freely available, and since its inception in 2015, the database has become an important resource to several studies. Its content has been: used as training data for machine learning models; incorporated to larger databases; and utilized in drug discovery studies for hit identifications.

Description: Here, we report the updated version of SANCDB. The new version includes 412 additional compounds that have been reported since 2015, giving a total of 1012 compounds in the database. Further, although natural products (NPs) are an important source of unique scaffolds, they have a major drawback due to their complex structure resulting in low synthetic feasibility in the laboratory. With this in mind, SANCDB is, now, updated to provide direct links to commercially available analogs from two major chemical databases namely Mcule and MolPort. To our knowledge, this feature is not available in other NP databases. Additionally, for easier access to information by users, the database and website interface were updated. The compounds are now downloadable in many different chemical formats.

Conclusions: The drug discovery process relies heavily on NPs due to their unique chemical organization. This has inspired the establishment of numerous NP chemical databases. With the emergence of newer chemoinformatic technologies, existing chemical databases require constant updates to facilitate information accessibility and integration by users. Besides increasing the NPs compound content, the updated SANCDB allows users to access the individual compounds (if available) or their analogs from commercial databases seamlessly.

Keywords: SANCDB, Natural products, Commercial analogs, Drug discovery

Introduction

Throughout history, natural products (NPs) have benefited mankind in food, pesticides, cosmetic products and as drugs [1]. NP research has, especially, been a growing field in modern drug discovery as they offer unique chemical scaffolds [2], hence a greater structural diversity than synthetic ones, and they cover a large area of chemical space [3]. Between 2010 and 2019, NPs contributed to 25–33% of approved small molecules [4], and it is estimated that they represent 35% of medicines [5]. Some of the notable approved drugs, either from pure or derived NPs, include lefamulin, the aminoglycoside antibiotic plazomicin; tafenoquine succinate, an antimalarial agent; and aplidine, an anticancer agent [4]. Given the high interest in NP research, over 120 NP databases and collections have been developed, and are continuously being updated [1, 3] with new information and functionalities.

The South African Natural Compounds Database (SANCDB; https://sancdb.rubi.ru.ac.za/) is a fully

---

*Correspondence: O.tastanbishop@ru.ac.za
[1] Research Unit in Bioinformatics (RUBi), Department of Biochemistry and Microbiology, Rhodes University, Makhanda/Grahamstown 6140, South Africa
Full list of author information is available at the end of the article

Diallo *et al. J Cheminform* (2021) 13:37

Page 2 of 14

referenced database of NPs derived from sources within South Africa [6]. The database and website were established by the Research Unit in Bioinformatics (RUBi) in 2015. The main content of the database is a set of NP chemical structures in different chemical formats linked to their primary literature reference. Since its creation, the database has attracted significant interest in diverse domains including NP research, drug discovery, cheminformatics and machine learning with up to 52 citations. Besides its use in hit identification in drug discovery studies, SANCDB content has also served as training data for machine learning models [7, 8]; training data for a NPs likeness scorer (NaPLeS) [7]; and for NP-Scout, a machine learning approach for the identification of NPs [9]. Similarly, the database compounds have served to train STarFish, a target fishing model for NPs [8]. SANCDB was also utilized as an intermediate node for data integration into larger or more specialized information systems. An example is Natural Product Activity and Species Source (NPASS), which is a NP activity and species source database built on some NPs data resources, including SANCDB [10]. Similarly, SANCDB data has also been used in the COCONUT (COlleCtion of Open Natural ProdUcTs) database [11].

The usage of the South African natural compounds as drugs is yet to be reported. However, the search for potential hits with activity against infectious agents and cancer is on the rise, and may lead to the identification of drugs in the near future [12]. In terms of drug discovery studies, the database has been used for identification of hit compounds against the active (orthosteric) site of various biological drug targets in diseases including malaria, trypanosomiasis and severe acute respiratory syndrome Corona Virus 2 (SARS-COV-2) [13–18]. Additionally, potential allosteric modulators such as 20(29)-lupene-3β-isoferulate (SANC00518) for human Hsp90α [14]; discorhabdin N (SANC00132), for human Hsp72 and Hsc70 [15]; gordonoside A (SANC00456) for *Plasmodium falciparum* Prolyl tRNA synthetase [19] have also been identified from SANCDB.

Here, we present an updated version of SANCDB with a number of new features. Firstly, over the last five years, since the inception of SANCDB, considerable NP research has been performed in the country. For instance, in 2019, Fantoukh et al. isolated 11 compounds from *Aspalathus linearis* [20], and Awolola et al. reported four compounds from the genus *Ficus* [21]. A variety of flavonoids, proanthocyanidins, ellagitannins, oligosaccharides and quinic acid derivatives were isolated from *Myrothamnus flabellifolia Welw* in 2016 [22]. Thus, the updated SANCDB has curated such continuously growing information. Secondly, in the updated version of the database, we provide links to commercially available

analogs for each compound to increase the search space and the availability of physical compounds. We hope that this unique feature of SANCDB will later be implemented by other NPs databases. Most of the time NPs' physical availability is either limited [1] or very expensive due to required isolation methods [5, 23]; and it has been well demonstrated that they can serve as good start points for the development of synthesizable analogs which may lead to effective drugs [24]. Finally, we include a scaffold analysis for all compound entries; this was undertaken to determine the database chemical diversity, which is an important aspect in the exploration of potential hit compounds [25]. Compound classification in SANCDB is also revisited.

Methods

Update of the database

NPs isolated from South African sources were searched in the literature, and uploaded through the earlier pipeline described in the preceding publication [6]. Elsevier's abstract and citation database (Scopus) Application Programming Interface (API) was used to identify more references. This allowed access to the scholarly databases indexed by Scopus [26] for collection, parsing and extraction of organized literature references. From the current set of references in SANCDB, we retrieved the list of all authors using the reference Digital Object Identifier (DOI). Using the Scopus API [26], a list of all publications associated with each author was retrieved. Both redundant publications and those in which none of the authors had a South African affiliation were removed as a pre-filtering step. From the remaining list of publications, the abstracts, and if available, the full text of the articles, were retrieved. Keywords "South Africa", "compound" and "isolate" were then searched in the abstracts and full texts, removing documents not presenting any of these keywords. The resulting list of publications was then searched in SciFinder [27] to find each compound's Chemical Abstracts Service (CAS) [27] number. References were then checked to confirm that sources were indeed from South Africa. Using an updated pipeline described in the original publication [6], new compounds were uploaded into the database. Compound sources were mapped to their genera, families and kingdoms using pygbif [28] a Python client for the Global Biodiversity Information Facility (GBIF) [29] API. PubChemPy [30] was used to retrieve additional information on chemical compounds utilizing their CAS number as query. Compound IDs for different databases (ChEMBL [31], DrugBank [32], ZINC [33], PubChem [34]) were automatically retrieved from PubChem [34]. Besides the previously assigned compound classification generated manually, additional automated classification based on

Diallo *et al. J Cheminform*   (2021) 13:37

Page 3 of 14

ClassyFire [35] was also included. To allow diverse usage of the compounds in docking studies, structures were prepared in different ready-to-dock formats, viz Auto-Dock *pdbqt* and Schrödinger Maestro format [36]. All aromatic compound structure depictions were standardized to the Kekulé form.

Analogs

Analogs for each SANCDB entry were extracted from the Mcule [37] and Molport [38] databases. The set of purchasable compounds in SMILES format from each of these databases (Version October 2019, latest version at that date) was downloaded. Similarity scores (Tanimoto coefficient) were computed using OpenBabel (Version 2.3) [39] fingerprint FP2 [40], a path-based fingerprint which indexes compounds linear fragments up to seven atoms. With the set of indexed fragments, a hash number from zero to 1020 was used to set a bit in a 1024-bit vector. A Tanimoto coefficient of 0.6 or greater was used as cut-off for analog identification. These steps have been incorporated into an automated update pipeline in the backend which fetches updated analog data from the respective Mcule and MolPort database APIs monthly. The front end of the database has also received updates and additions to integrate information about the newly added compounds and analogs.

Chemoinformatic analysis

SANCDB compounds' scaffolds were calculated through the Bemis-Murcko decomposition of molecules [3, 25, 41–45]. Scopy [46] was utilized for scaffold calculations. Unique scaffolds from each molecule were assembled with their respective frequencies to generate the molecule cloud [46, 47].

An analysis of compound distributions into drug-like, extended drug-like, lead-like, fragment-like, protein–protein inhibitor-like (PPI-like) subsets was performed (Table 1), using conditions defined previously [48].

To assess the coverage of SANCDB chemical space by the analogs, Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE)

were applied on compounds' non-normalized Molecular Quantum Numbers (MQN) [49] for dimensionality reduction. MQN descriptors were computed using rdkit.Chem.rdMolDescriptors module of RDKit [50]. T-SNE is a variant of Stochastic Neighbor Embedding calculating similarity between two points in the low-dimensional space using a Student-t distribution [51]. The method has been used for chemical space analysis [52–54]. PCA and t-SNE implementations in scikit-learn were used [55, 56]. A learning rate of 100 and perplexity of 50 were used for t-SNE. Other parameters were kept to default. The 42 dimensions in MQN property space were reduced to two dimensions with both methods. The 3D structures of the identified analogs were prepared using OpenBabel [39] and resulting geometries minimized in RDKit [50] using the Merck Molecular Force Field (MMFF94) [57]. Moreover, the AutoDock ready to dock files '*pdbqt*' were prepared using AutoDock 4 utilities [58]. The data was analyzed in the Jupyter Notebook [59] environment using the Python module pandas [60, 61] and pandas-profiling [62]. The descriptive statistics and plots were done in the same environment.

Results and discussions

Database design and website interface

SANCDB uses a MySQL database, managed by the Django framework as described previously [6]. The appearance of the site has seen numerous minor changes such as removing clutter from page elements and menus, using Bootstrap to standardize elements, and adjusting text sizes to give the site a more modern and neater appearance. Redundant and unused JavaScript and CSS libraries have been removed. This reduced the initial page load by 0.5 MB and improved maintainability of the SANCDB codebase.

A similarity search was added to the database search functionalities. The database originally only had a substructure search function. The user can now do a similarity or a substructure search. The query is searched in the database using Open Babel FP2 fingerprints and all

Table 1 Molecular properties and conditions used to determine compound subsets

Molecular properties	Conditions
Drug-like	MW ≤ 500 & MW ≥ 150 & logP ≤ 5 & nHD ≤ 5 & nHA ≤ 10
Extended drug-like	Druglike & nRot ≤ 7 & TPSA < 150
Lead-like	MW ≥ 250 & MW ≤ 350 & nRot ≤ 7 & logP ≤ 3.5
Fragment-like	nHA ≥ 3 & MW ≤ 300 & nHD ≤ 3 & logP ≤ 3
PPI-like	nRing ≥ 4 & MW > 400 & nHA > 4 & logP > 4

MW Molecular weight, *logP* lipophilicity, *nHA* number of hydrogen bond acceptor, *nHD* number of hydrogen bond donor, *TPSA* total polar surface area, *nRot* number of rotatable bonds and *nRing* number of rings

Diallo *et al. J Cheminform*   (2021) 13:37

Page 4 of 14

compounds having at least 0.6 Tanimoto similarity are returned.

Compound summary page and analogs

As shown in the database schema in Fig. 1, updates featuring the molecular mass and additional compound identification parameters were implemented within the compounds table. Further, a new table was created to link analogs to the respective compounds. Each compound now has the option to view and download commercially available analogs. A total of 380,206 commercially available analogs were added from the Mcule [37] and Mol-Port [38] databases (at the time of writing). A user can see a list of analogs ordered by similarity score for any compound in SANCDB. Each analog links to its respective entry on the vendor's website.

From the compound summary page, for each compound there are also links to ChEMBL, DrugBank, PubChem, and ZINC databases, if available. Data points for molecular mass and, when available, a link to Chem-Spider have been added to the interface. Standardized compound classifications according to ClassyFire, are

Fig. 1 Database schema detailing additions to the existing SANCDB database, compound summary and analog pages. **a** Basic compound information (Name, Formula, SMILES…). **b** Allows users to view 2D and 3D depictions of the compound. **c** Structure download options and navigation to page listing commercially available analogs. **d** References, classifications, other names, sources organisms and uses. Analogs are updated monthly using an automated pipeline to access external database APIs, calculate similarity scores and insert into the local SQL database
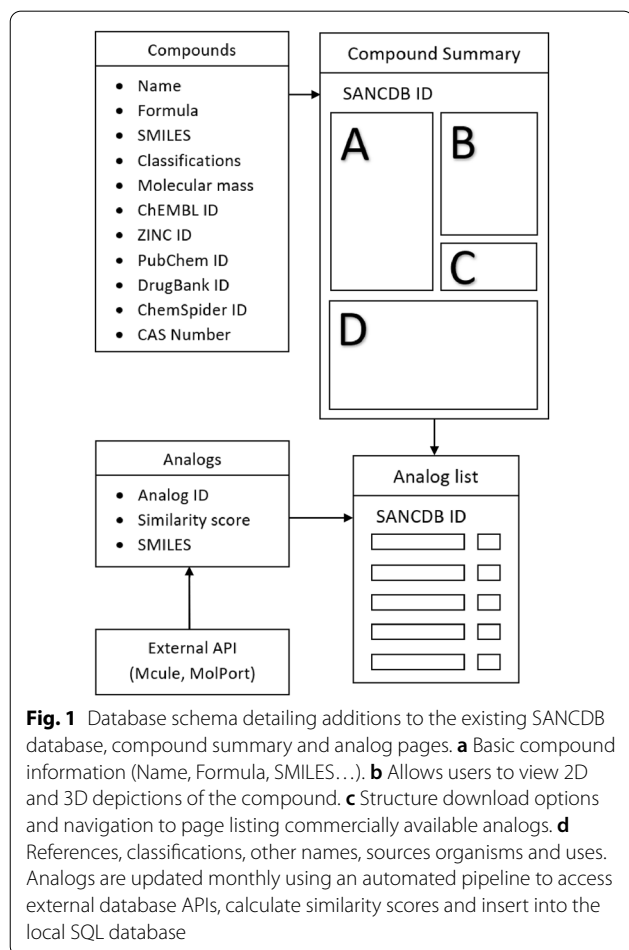
displayed in addition to all existing classifications. Previously, compound structures for MOL2, PDB, minimized PDB, SDF and SMILES were available. Subsequently, Maestro and PDBQT files have been added to these options for existing and new compounds.
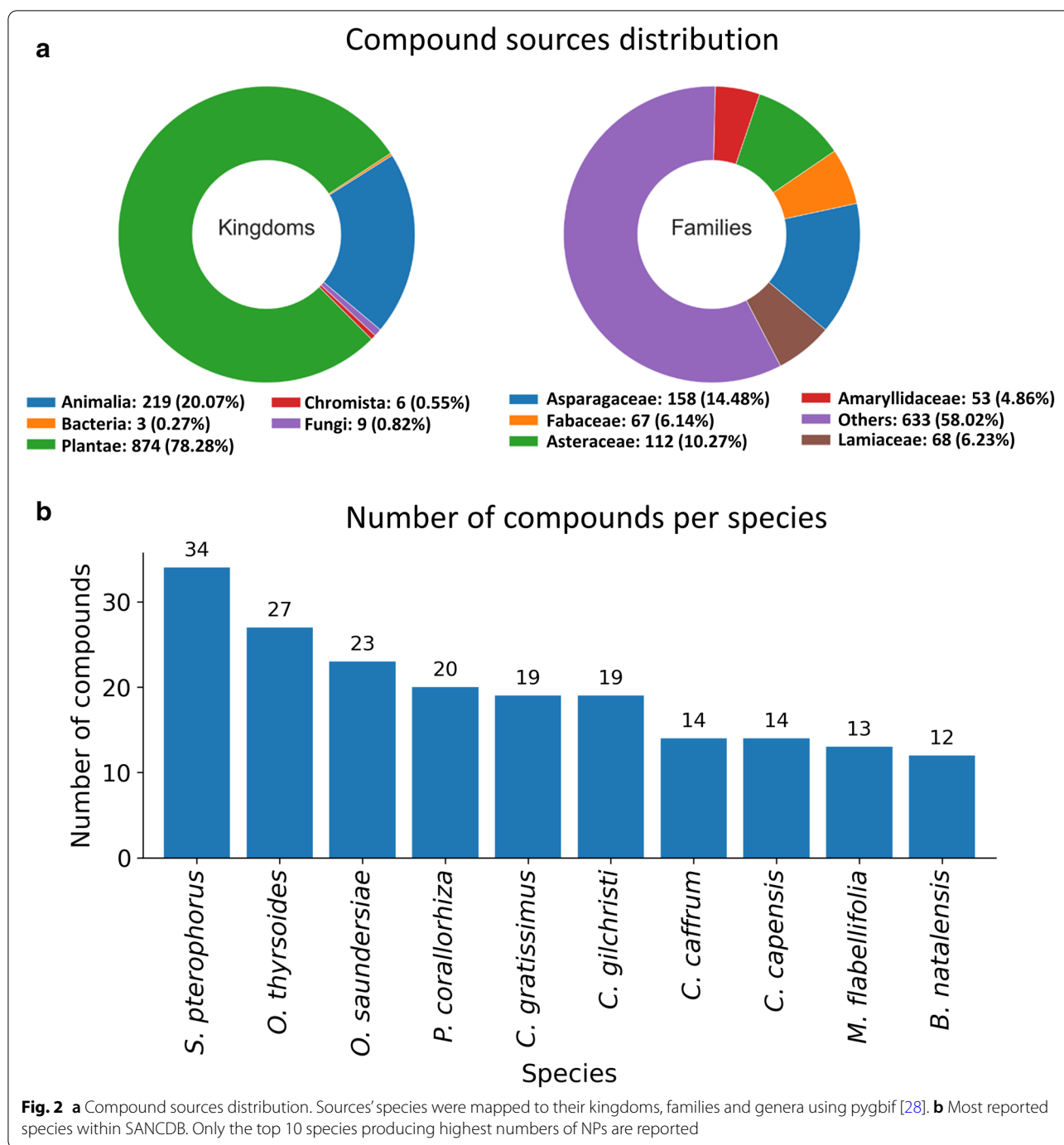
Database content

The database initially contained 600 compounds updated here to 1012 compounds from 359 literature references comprising mainly journal articles. Cumulatively, compounds had been isolated from 321 sources, distributed among five biological kingdoms (fungi, bacteria, plants etc.), 104 distinct families and 187 genera. The distribution of the sources of compounds according to their families and kingdoms is shown in Fig. 2a. Plants represented the majority of the sources counting for 854 compounds (78.3%) followed by animals 219 (20.1%), fungi 9 (0.8%), chromista 6 (0.5%) and finally bacteria with 3 compounds (0.3%). This distribution is similar to some of the NPs databases where plants are the primary sources [48]. It is interesting to note the low proportion of compounds, isolated from bacterial and from microbial sources, given that they are major sources of NPs [3]. *Streptomyces sp.* with three isoflavones was the only bacteria species source reported in the entire database. *Clathrina aff reticulum*, *Eurotium rubrum*, *Termitomyces microcarpus* and *Fusarium proliferatum* were the only four fungi. Yet, microbes remain sources for major classes of antibiotics [3]. This low microbial source proportion was also consistent with other NPs databases [48, 63], and some databases only focus on plants [64–66] probably because of their abundance as NP sources. The potential of NPs from microbes in South Africa may be under-explored. A similar observation was made in the source distribution of the Brazilian compound database NuBBE [48]. Plants were found to be the major producer of NPs compared to other sources (animals, fungi, bacteria) [67]. This may be explained by plant uses being more documented in traditional medicines and having easier accessibility than other sources.

Regarding families, Asparagaceae (158–14.48%), Asteraceae (112–10.27%), Lamiaceae (68–6.23%), Fabaceae (67–6.14%) and Amaryllidaceae (53–4.86%) were the top family sources. Regarding genera, *Ornithogalum* (62–5.7%), *Senecio* (58–5.3%), *Eucomis* (39–3.6%), *Salvia* (39–3.6%) and *Plocamium* (38–3.5%) were the top ones.

The top five species with the highest number of isolated compounds were *Senecio pterophorus* (34–3.1%), *Ornithogalum thyrsoides* (27–2.4%), *Ornithogalum saundersiae* (23–2.1%), *Plocamium corallorhiza* (20–1.8%) and *Cephalodiscus gilchristi* (19–1.7%) as shown in Fig. 2b. *Seneccio pterophorus*, is an Asteraceae producing macrocyclic diester pyrrolizidine alkaloids known

Fig. 2 a Compound sources distribution. Sources' species were mapped to their kingdoms, families and genera using pygbif [28]. **b** Most reported species within SANCDB. Only the top 10 species producing highest numbers of NPs are reported

to be hepatotoxic, carcinogenic, genotoxic and teratogenic [68]. Two species of Ornithogalum plant from the Asparagaceae family: *thyrsoides* (27 compounds) and *saundersiae* (23 compounds) were the second and third sources with the highest number of isolated compounds, respectively. The *thyrsoides* species is widespread in South Africa, Western Cape and known for its cytotoxicity against HL-60 human promyelocytic leukemia cells

[69, 70]. *Ornithogalum saundersiae* is an ornamental flower from Mpumalanga, KwaZulu-Natal, and Swaziland, toxic to cattle [71, 72]. Also known as the tube worm [73], *Cephalodiscus gilchristi* is a cephalodiscidae containing highly potent alkaloids against lymphocytic leukemia [74]. This marine invertebrate also produces cephalostatin 1, a potent cell growth-inhibiting compound [73]. Finally, *Plocamium corallorhiza* is a red algae

Diallo *et al. J Cheminform*   (2021) 13:37

Page 6 of 14

in the family Plocamiaceae, abundant in South Africa known for its halogenated monoterpenes [75, 76]. A common characteristic among these top sources is that they are naturally widespread. Thus, they may be more accessible for compound isolation and extraction studies and the source of more compounds as a result. Information in the database may contribute to biodiversity conservation. Indeed, the above information may contribute to finding NPs source hotspots that conservation efforts may prioritize. None of these species was found in the South African National Biodiversity Institute (SANBI) list of endangered species [77].
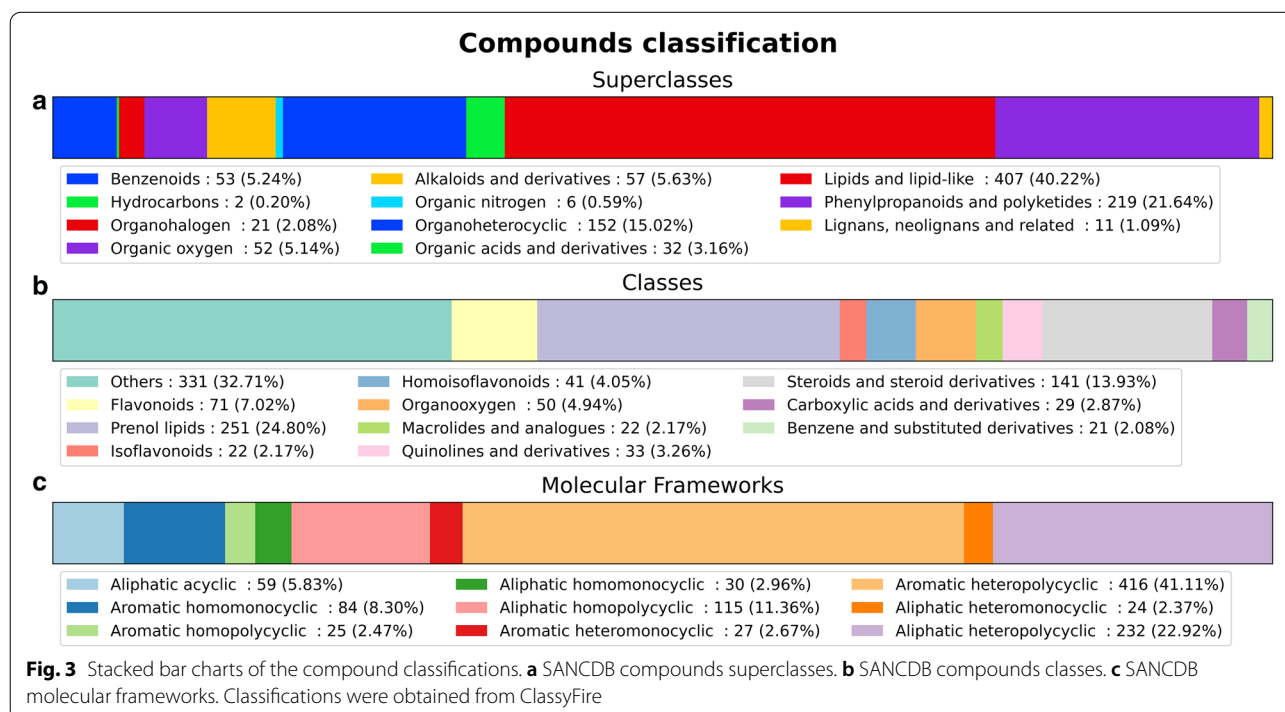
Compounds classification

NPs classification is useful to assess their diversity, and can be done via different schemes [48, 78]. Here, Classy-Fire, which performs a hierarchical classification using structural patterns into kingdoms, superclasses, classes and subclasses, was utilized [20]. We noted 11 superclasses out of the 26 ClassyFire organic compound superclasses; 77 classes out of 764 ClassyFire classes [35]; and 124 subclasses. These numbers indicate database diversity, and therefore potentiality for a variety of biological activities. The distribution of compounds superclasses is shown in Fig. 3a, classes in Fig. 3b and molecular frameworks in Fig. 3c. Top compound classes were the phenol lipids (251–24.8%), the steroids and steroid derivatives (141–13.9%), the flavonoids (71–7%), the organooxygen

compounds (50–4.9%) and the homoisoflavonoids (41–4.1%).

Comparatively, SANCDB has a similar distribution to other compound databases. The Integrated Ethiopian Traditional Herbal Medicine and Phytochemicals Database (ETM-DB) [79] compounds classification was also done using ClassyFire, and shares the same top three superclasses. ETM-DB with 3930 compounds has 22 superclasses and 200 classes. The 500 Pan-African Natural Products Library (p-ANAPL) compounds are distributed across 30 classes [80] while the Nuclei of Bioassays, Ecophysiology and Biosynthesis of Natural Products Database (NuBBE) database has 14 classes with 2147 compounds [48]. NuBBE and p-ANAPL use however a different classification scheme. SANCDB contains 13.93% of steroids and derivatives, and was previously shown to have the highest rates of steroids compared to other NPs databases [81].

The database was rich in polycyclic compounds in Fig. 3c. Indeed, the molecular framework distribution showed that only 59 (5.8%) of the compounds were acyclic. Nine distinct types of molecular frameworks were found in SANCDB. The most common frameworks were the aromatic heteropolycyclic (416–41.1%), the aliphatic heteropolycyclic (232–22.9%) and the aliphatic homopolycyclic (115–11.4%). Molecular framework is similar to the concept of scaffold [44] and describes compounds according to their aliphaticity/aromaticity, ring count, and the diversity of atom types [35]. It is an important

Fig. 3 Stacked bar charts of the compound classifications. **a** SANCDB compounds superclasses. **b** SANCDB compounds classes. **c** SANCDB molecular frameworks. Classifications were obtained from ClassyFire

Diallo *et al. J Cheminform*   (2021) 13:37

Page 7 of 14

feature to assess compounds libraries' diversity in screening campaigns [26].

Compounds activities

From the source reference, where significant biological activity of isolated compounds was determined, the information was recorded and standardized to avoid duplication. The distribution of biological activities for the 318 compounds showed anticancer (158–31.6%), antibacterial (61–12.2%), AChE inhibitor (38–7.6%), antimalarial, (34–6.8%) and antiproliferative (20–4.0%) as most common activities (Fig. 4). We noted 59 distinct activity types. Some compounds showed a variety of activities (>6 different activities): Combretastatin A-1, Ouabain, Acovenoside A, Isoorientin and Quercetin. They were also found to be associated with at least 15 predicted targets in ChEMBL [31] with 90% confidence. They can be a good starting point for multi-target inhibitors. It is noteworthy that activities assignment was not standardized. Recorded activities could be at the molecular, cellular, tissue or disease level. Also, recording of activities was limited to only the reference in the database. Furthermore, only compounds showing a significant level of activity were recorded, limiting the number of assigned biological activities to 318 out of the 1012 in the database.

Commercially available analogs

A first evaluation of SANCDB NPs availability showed that only about 30% of SANCDB was readily purchasable. 316 were obtainable on MolPort [38] while 327 on Mcule [37]. A previous assessment of NPs commercial availability (in the ZINC subset of readily purchasable compounds) showed that only 10% were purchasable [3]. In general, NPs are insufficiently covered in commercial databases [3]. Additionally, 118 SANCDB compounds
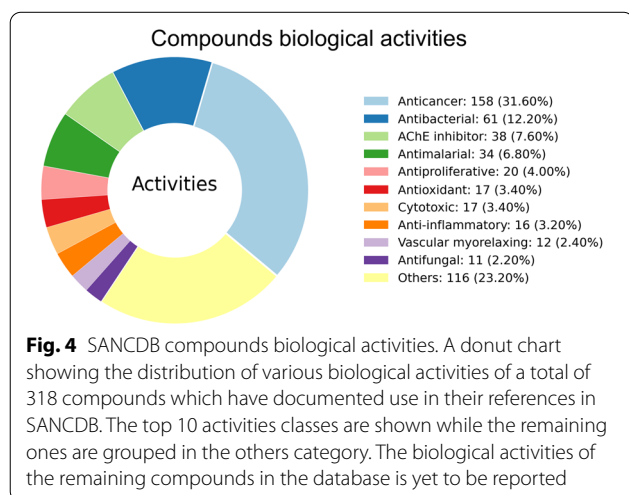
had synthetic accessibility [82] score greater than six (see Additional file 1: Fig. S1). They may thus pose synthetic challenges [82].
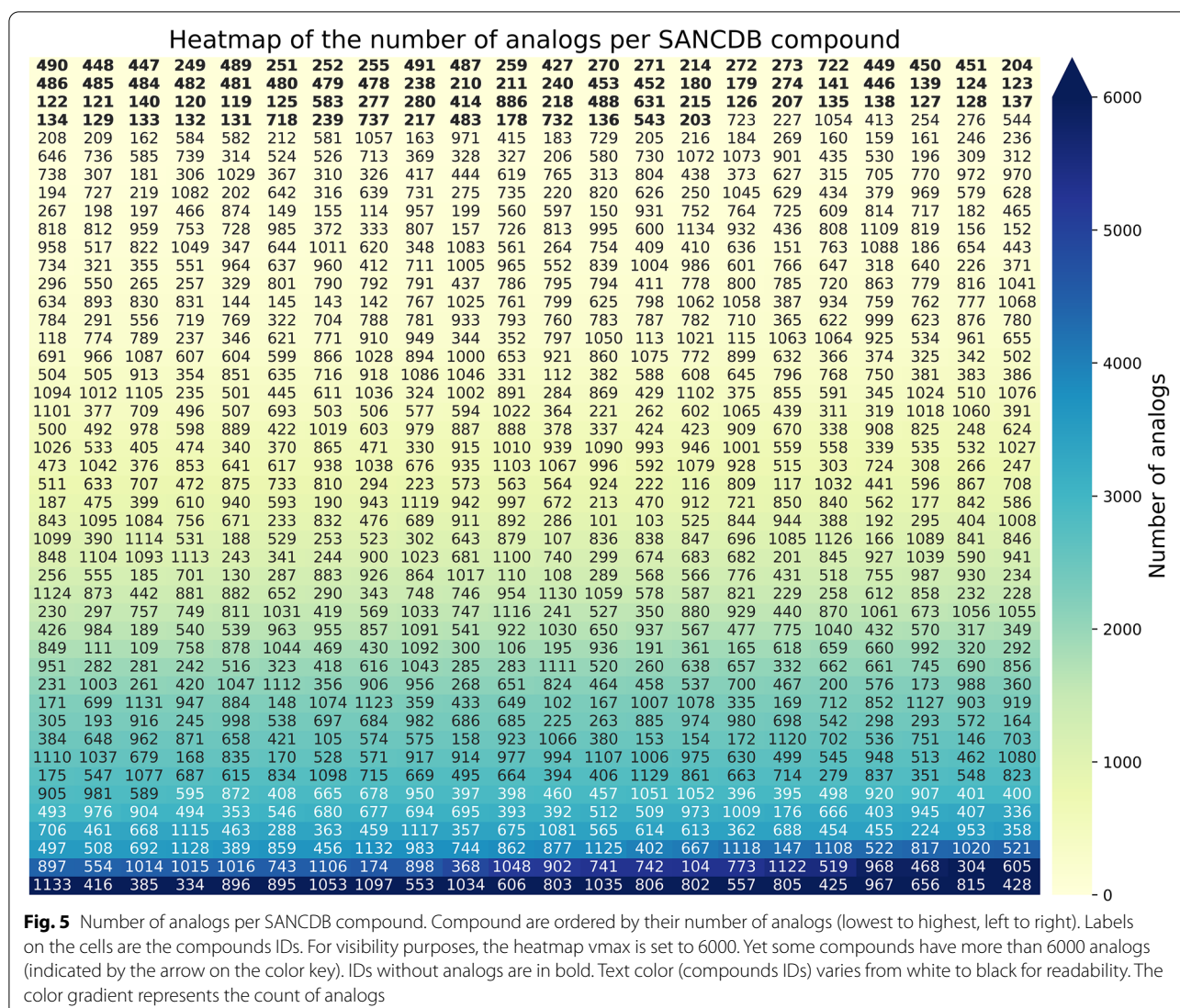
In the updated version of SANCDB, a total of 380,206 unique commercially available analogs were added with an average of 1487 analogs per compound (at the time of writing). 232,747 analogs were retrieved from Mcule and 141,320 from Molport. Each compound analogs' SMILES and Tanimoto similarity scores are available for download. The downloaded Molport [38] and Mcule [37] databases contained 7,597,214 and 9,884,200 compounds, respectively in October 2019. The distribution of the number of analogs per compound is shown in Fig. 5. Frequencies ranged from 42,224 to zero analogs. SANC00428, SANC00815, SANC00656, SANC00967 and SANC00425 have the highest number of analogs with 42,224, 29,045, 27,823, 26,638, 24,993 analogs respectively. These compounds have low molecular weight (MW). All compounds with more than 10,000 analogs had a MW below 300 Da. However, there was no correlation between the number of analogs and the MW with a negligible Pearson correlation of -0.09 (see Additional file 1: Figs. S2 and S3).

No analog was found for 70 compounds using a Tanimoto similarity coefficient threshold as low as 0.6. SciFinder [27] database was used to manually retrieve analogs for these compounds. As SciFinder [27] search was per similarity interval (e.g. 0.75–0.79 similarity interval), only the first interval having analogs, starting from the highest, was considered. The number of analogs for these compounds ranged from one to 29 with 23 compounds having only one analog [27]. The number of analogs per compounds and their respective similarity thresholds are available in Additional file 1: Table S1. Also, only 43% (442) of the compounds had more than 1000 analogs in Molport [38] and Mcule [37] datasets. This indicates a low coverage of NPs availability considering the size of the datasets used (Molport and Mcule with 7,597,214 and 9,884,200 respectively) and the low Tanimoto similarity coefficient cutoff used (0.6).

Analogs covered SANCDB chemical space regions (Fig. 6). SANCDB compounds formed a cluster overlapped with analogs which extend by decreasing similarity score. Analogs with similarity values in the range (0.6, 0.7) were the most isolated. Some SANCDB compounds without analogs occupied a small isolated cluster (zoomed region of the plot). Further analysis showed that they corresponded to compounds with zero analogs. The t-SNE visualization showed a less dense cloud, thus indicating further separation between compounds (see Additional file 1: Fig. S4).

SANCDB analogs can expand initial drug discovery projects. Analog information is important during

Fig. 4 SANCDB compounds biological activities. A donut chart showing the distribution of various biological activities of a total of 318 compounds which have documented use in their references in SANCDB. The top 10 activities classes are shown while the remaining ones are grouped in the others category. The biological activities of the remaining compounds in the database is yet to be reported

Diallo *et al. J Cheminform* (2021) 13:37

Page 8 of 14

Heatmap of the number of analogs per SANCDB compound

```
490  448  447  249  489  251  252  255  491  487  259  427  270  271  214  272  273  722  449  450  451  204
486  485  484  482  481  480  479  478  238  210  211  240  453  452  180  179  274  141  446  139  124  123
122  121  140  120  119  125  583  277  280  414  886  218  488  631  215  126  207  135  138  127  128  137
134  129  133  132  131  718  239  737  217  483  178  732  136  543  203  723  227 1054  413  254  276  544
208  209  162  584  582  212  581 1057  163  971  415  183  729  205  216  184  269  160  159  161  246  236
646  736  585  739  314  524  526  713  369  328  327  206  580  730 1072 1073  901  435  530  196  309  312
738  307  181  306 1029  367  310  326  417  444  619  765  313  804  438  373  627  315  705  770  972  970
194  727  219 1082  202  642  316  639  731  275  735  220  820  626  250 1045  629  434  379  969  579  628
267  198  197  466  874  149  155  114  957  199  560  597  150  931  752  764  725  609  814  717  182  465
818  812  959  753  728  985  372  333  807  157  726  813  995  600 1134  932  436  808 1109  819  156  152
958  517  822 1049  347  644 1011  620  348 1083  561  264  754  409  410  636  151  763 1088  186  654  443
734  321  355  551  964  637  960  412  711 1005  965  552  839 1004  986  601  766  647  318  640  226  371
296  550  265  257  329  801  790  792  791  437  786  795  794  411  778  800  785  720  863  779  816 1041
634  893  830  831  144  145  143  142  767 1025  761  799  625  798 1062 1058  387  934  759  762  777 1068
784  291  556  719  769  322  704  788  781  933  793  760  783  787  782  710  365  622  999  623  876  780
118  774  789  237  346  621  771  910  949  344  352  797 1050  113 1021  115 1063 1064  925  534  961  655
691  966 1087  607  604  599  866 1028  894 1000  653  921  860 1075  772  899  632  366  374  325  342  502
504  505  913  354  851  635  716  918 1086 1046  331  112  382  588  608  645  796  768  750  381  383  386
1094 1012 1105 235  501  445  611 1036  324 1002  891  284  869  429 1102  375  855  591  345 1024  510 1076
1101 377  709  496  507  693  503  506  577  594 1022  364  221  262  602 1065  439  311  319 1018 1060  391
500  492  978  598  889  422 1019  603  979  887  888  378  337  424  423  909  670  338  908  825  248  624
1026 533  405  474  340  370  865  471  330  915 1010  939 1000  993  946 1001  559  558  339  535  532 1027
473 1042  376  853  641  617  938 1038  676  935 1103 1067  996  592 1079  928  515  303  724  308  266  247
511  633  707  472  875  733  810  294  223  573  563  564  924  222  116  809  117 1032  441  596  867  708
187  475  399  610  940  593  190  943 1119  942  997  672  213  470  912  721  850  840  562  177  842  586
843 1095 1084 756  671  233  832  476  689  911  892  286  101  103  525  844  944  388  192  295  404 1008
1099 390 1114 531  188  529  253  523  302  643  879  107  836  838  847  696 1085 1126  166 1089  841  846
848 1104 1093 1113 243  341  244  900 1023  681 1100  740  299  674  683  682  201  845  927 1039  590  941
256  555  185  701  130  287  883  926  864 1017  110  108  289  568  566  776  431  518  755  987  930  234
1124 873  442  881  882  652  290  343  748  746  954 1130 1059  578  587  821  229  258  612  858  232  228
230  297  757  749  811 1031  419  569 1033  747 1116  241  527  350  880  929  440  870 1061  673 1056 1055
426  984  189  540  539  963  955  857 1091  541  922 1030  650  937  567  477  775 1040  432  570  317  349
849  111  109  758  878 1044  469  430 1092  300  106  195  936  191  361  165  618  659  660  992  320  292
951  282  281  242  516  323  418  616 1043  285  283 1111  520  260  638  657  332  662  661  745  690  856
231 1003  261  420 1047 1112  356  906  956  268  651  824  464  458  537  700  467  200  576  173  988  360
171  699 1131  947  884  148 1074 1123  359  433  649  102  167 1007 1078  335  169  712  852 1127  903  919
305  193  916  245  998  538  697  684  982  686  685  225  263  885  974  980  698  542  298  293  572  164
384  648  962  871  658  421  105  574  575  158  923 1066  380  153  154  172 1120  702  536  751  146  703
1110 1037 679  168  835  170  528  571  917  914  977  994 1107 1006  975  630  499  545  948  513  462 1080
175  547 1077  687  615  834 1098  715  669  495  664  394  406 1129  861  663  714  279  837  351  548  823
905  981  589  595  872  408  665  678  950  397  398  460  457 1051 1052  396  395  498  920  907  401  400
493  976  904  494  353  546  680  677  694  695  393  392  512  509  973 1009  176  666  403  945  407  336
706  461  668 1115  463  288  363  459 1117  357  675 1081  565  614  613  362  688  454  455  224  953  358
497  508  692 1128  389  859  456 1132  983  744  862  877 1125  402  667 1118  147 1108  522  817 1020  521
897  554 1014 1015 1016  743 1106  174  898  368 1048  902  741  742  104  773 1122  519  968  468  304  605
1133 416  385  334  896  895 1053 1097  553 1034  606  803 1035  806  802  557  805  425  967  656  815  428
```

Color key (right): Number of analogs — 0, 1000, 2000, 3000, 4000, 5000, 6000

Fig. 5 Number of analogs per SANCDB compound. Compound are ordered by their number of analogs (lowest to highest, left to right). Labels on the cells are the compounds IDs. For visibility purposes, the heatmap vmax is set to 6000. Yet some compounds have more than 6000 analogs (indicated by the arrow on the color key). IDs without analogs are in bold. Text color (compounds IDs) varies from white to black for readability. The color gradient represents the count of analogs

hit optimization in drug discovery process. Screening hits identified from SANCDB can be further optimized through their analogs [13, 16, 17, 19, 83]. Additionally, more potent analogs of the potential allosteric modulators identified in SANCDB [14, 15, 84] may further enhance allosteric modulation of these compounds on their targets.
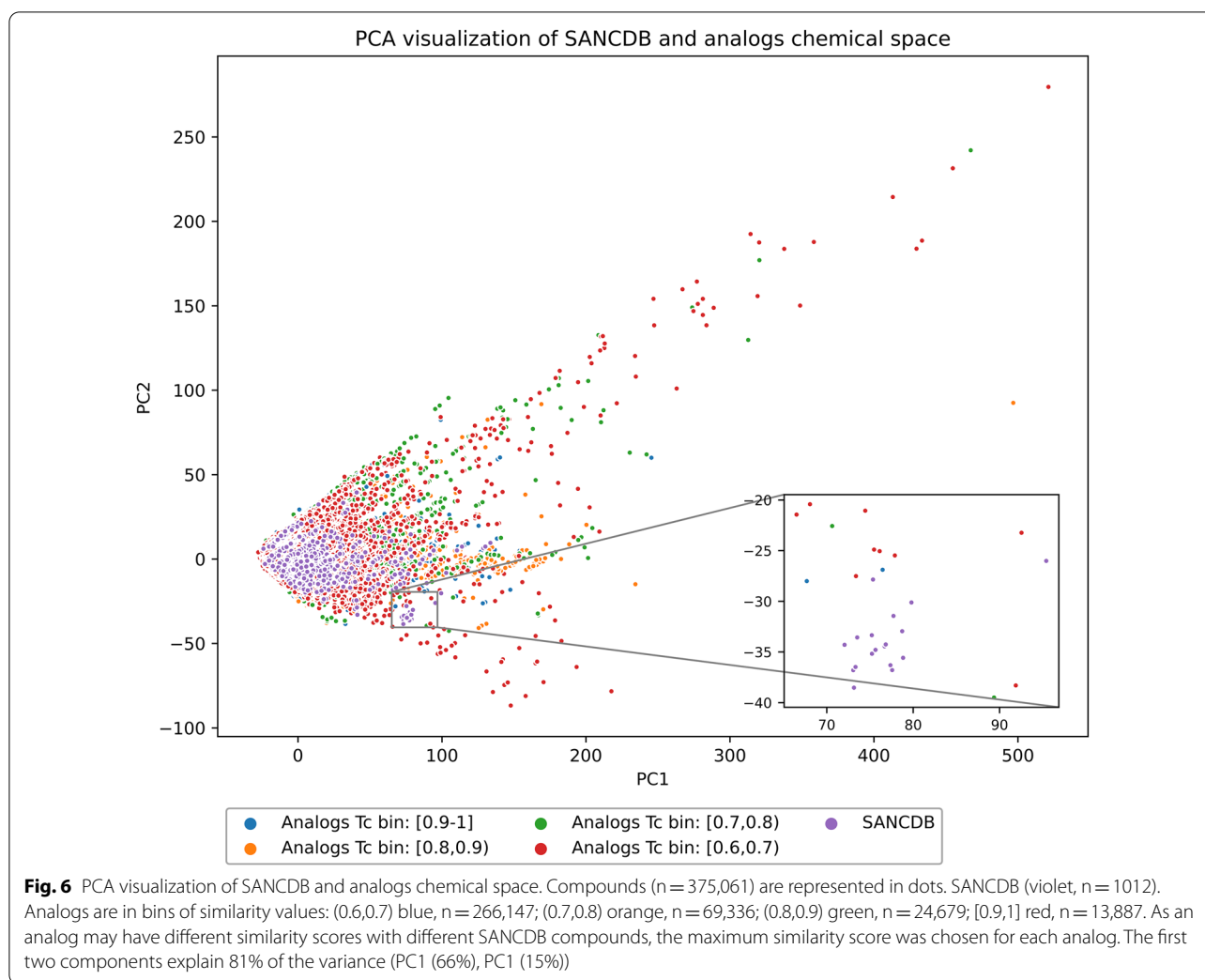
Scaffolds and compounds subsets

SANCDB compounds were analyzed in terms of their scaffolds and with regard to different subsets of chemical compounds relevant to drug discovery. NPs scaffolds are of interest as they are rich in sp^3-configured centers while synthetic scaffolds are generally flatter [80, 85]. They also often serve as the basis for synthetic modifications of drug-like compounds [65]. Scaffold diversity is ideal for

screening libraries as virtual screening also aims to find new scaffolds [66].

The molecule cloud visualization in Fig. 7 highlights top-ranked scaffolds. It also helps assess the diversity of scaffolds and their structural features, allowing the reader a rapid overview of the most common scaffolds [47]. However, a drawback may be the less visible of the less common scaffolds which may still be of interest. All scaffolds and their count are presented in Additional file 1: Table S2.

In SANCDB, about half of the compounds presented a unique scaffold. Indeed, 501 unique scaffolds were identified from the 1012 compounds, indicating their diversity. 59 compounds did not present a scaffold as they were purely aliphatic. Scaffold frequencies followed a "long tail" distribution (see Additional file 1: Fig. S5) common in compound datasets [47]. This shows the high number
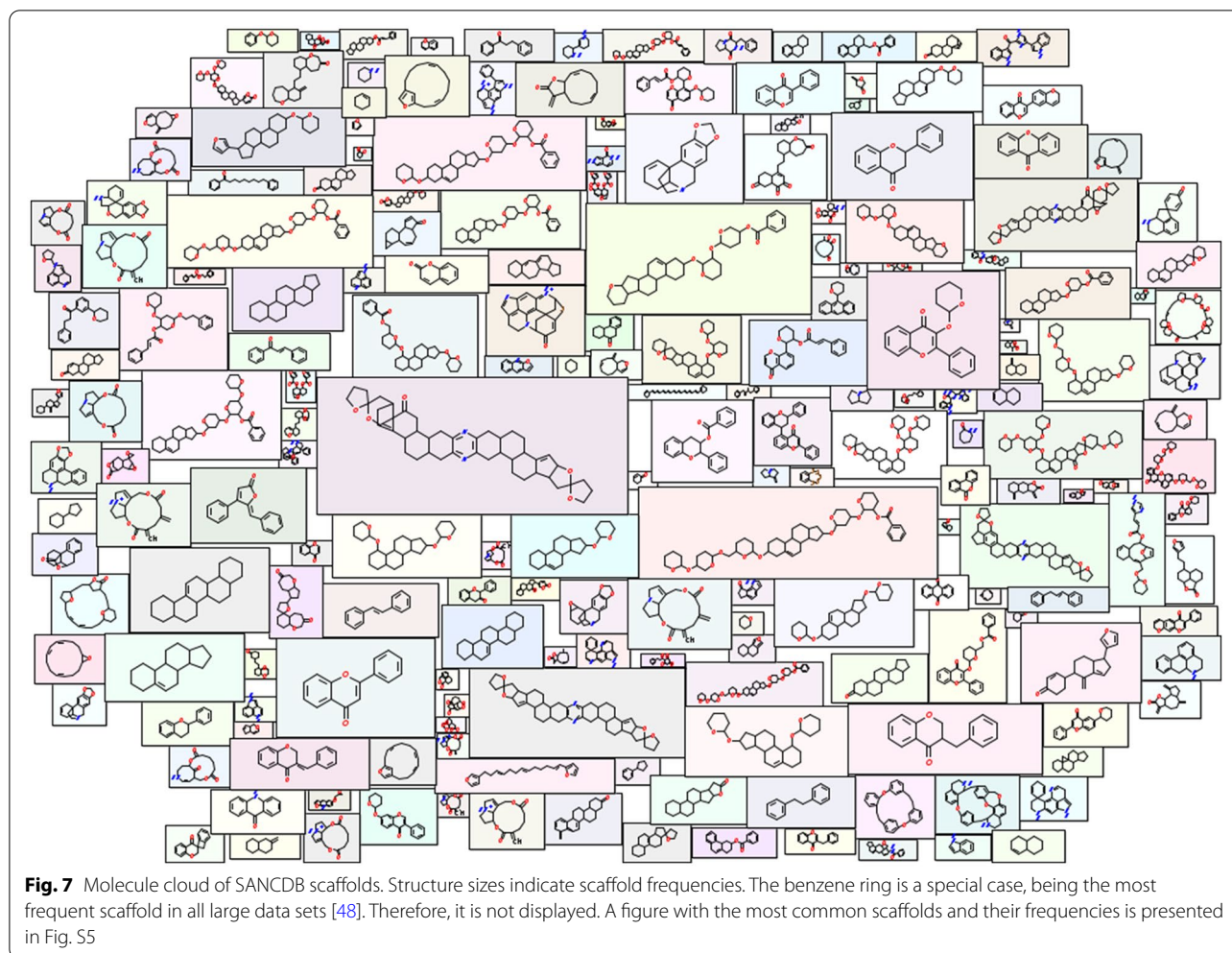
Diallo *et al. J Cheminform* (2021) 13:37

Page 9 of 14

Fig. 6 PCA visualization of SANCDB and analogs chemical space. Compounds (n = 375,061) are represented in dots. SANCDB (violet, n = 1012). Analogs are in bins of similarity values: (0.6,0.7) blue, n = 266,147; (0.7,0.8) orange, n = 69,336; (0.8,0.9) green, n = 24,679; [0.9,1] red, n = 13,887. As an analog may have different similarity scores with different SANCDB compounds, the maximum similarity score was chosen for each analog. The first two components explain 81% of the variance (PC1 (66%), PC1 (15%))

of singletons, compounds containing a unique scaffold in the entire database, highlighting this latter diversity.

Most common scaffolds were flavonoids, already known to be common in NP datasets [44]. Interestingly, they were only the third most represented in the distribution of compound classes in Fig. 3. This may be related to a structural diversity in the first two categories compared to flavonoid which may be more homogeneous. Structures of the top 10 scaffolds are represented in Additional file 1: Fig. S6. The chromane 3-Benzylchroman-4-one was the most common scaffold with 30 compounds. Its structure presents a bicycle consisting of a 3,4-dihydro-1-ben-zopyran, with a ketone group which is a structural alert. Structural alerts are high reactive groups which may cause toxicity [86]. The compound is known for human monoamine oxidase B inhibition [87]. Chromane scaffolds are promiscuous in NPs [67] and known for their anticancer activity [88] which may also be related with the database richness in anticancer compounds in Fig. 4.

The second most abundant was flavone found in 21 compounds. Its prodrug aminoflavone reached phase 2 clinical trials for breast cancer treatment [87]. The related compound in the database may present similar activity. Finally, flavanone was the third most common scaffold with 14 compounds. This scaffold also presented a ketone group as a structural alert.

Over half of the database was drug-like or extended drug-like compounds, shown in Fig. 8. We noted a minor difference (41 compounds) between the drug-like and extended drug-like compounds, with the latter having only two more conditions to the drug-like category. PPI-like and fragments-like subsets represent the most stringent conditions for the database compounds, hence showing the least number of compounds. NP datasets may have a low proportion of fragments due to their polycyclic nature. As shown by the compound classification, SANCDB was rich in polycyclic compounds (~75% of the compounds see Fig. 3c). Thus, a low proportion of

Diallo *et al. J Cheminform* (2021) 13:37

Page 10 of 14

Fig. 7 Molecule cloud of SANCDB scaffolds. Structure sizes indicate scaffold frequencies. The benzene ring is a special case, being the most frequent scaffold in all large data sets [48]. Therefore, it is not displayed. A figure with the most common scaffolds and their frequencies is presented in Fig. S5

fragments is expected. PPI-like compounds (MW > 400 and logP > 4) represented the smallest set with only 78 compounds. This contrasts with the high proportion of

Fig. 8 Proportions of compounds in each subset (drug-like, extended drug-like, fragment-like, lead-like, PPI-like). The count of each subset and their related proportions on the x-axis are reported. The green area indicates compounds fitting into that category. In the PAINS group, it indicates compound free of PAINS patterns

polycyclic compounds in the database. Given these low proportions of fragment-like and PPI-like compounds, the database may not be suitable for fragment-based drug discovery or protein–protein inhibition.

These proportions were similar to those observed in many other NPs databases with drug-like and extended drug-like being more than 50% while PPI-like, fragments-like and PAINS have low proportions [45].

These subsets fit different contexts in drug discovery. For example, fragments can easily be found in early stage screening to identify potent chemotypes for latter optimization. PPI-like are ideal candidates to block protein–protein interaction. The distribution of the different subsets can also be a good indicator of the ideal context for a dataset. For example, for a database enriched in fragments, fragment-based drug discovery approaches might be ideal. Small molecule databases for screening such as ZINC [33] are often subdivided into subsets. PAINS patterns are used to filter out frequent hitters in screening [89]. Hence, the various SANCDB subsets
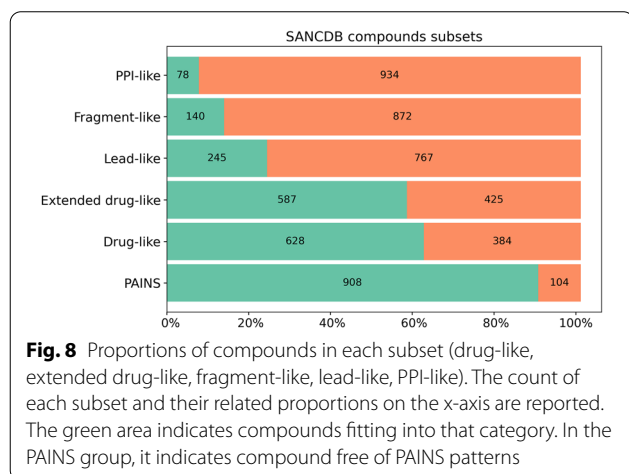
can be used to establish custom-made virtual screening experiments based on user's specific demands.

Conclusions

NPs remain an integral component of the drug discovery process. Hitherto, a large proportion of approved drugs have been derived from NPs [5]. This has inspired the establishment of numerous databases which contain diverse chemical classes of NPs to facilitate bioprospecting of important leads for biomedical and chemical research [1, 3]. Since the establishment of SANCDB in 2015, its usage as a source of data for both in silico screening and machine learning has been on the rise. Thus, to maintain its relevance, the current work aimed to update the database with additional compounds isolated from South African natural resources, as well as to add new functionalities aimed at providing a larger chemical space for hit exploration. To this end, the updated fully referenced relational database contains more than 1000 unique compounds from South Africa. A classification and scaffold analysis showed a diverse chemical representation with 501 unique NP scaffolds. The chemical diversity of a database is an indicator of how useful it can be for hit identification [25]. The database dataset is freely accessible and is downloadable in different chemical formats including ready to dock ones using either AutoDock [58] or Maestro from Schrödinger [37]. In consideration of the universally acknowledged limitations of NPs as a result of their complex structural organization, the current update also includes the incorporation of readily available analogs from two main commercial chemical databases (MolPort [38] and Mcule [37]). In comparison to other existing NP databases, this feature is present only in SANCDB and will provide users with a larger chemical library of compounds for both chemoinformatic and bioscreening studies. The analogs from the different databases are constantly updated via an automated pipeline making it more reliable. Analogs have been linked to their sources on Mcule [37] and MolPort [38], allowing users to obtain compounds for in vitro screening seamlessly.

Abbreviations

API: Application Programming Interface; AChE: Acetylcholinesterase; CAS: Chemical Abstracts Service; DOI: Digital Object Identifier; ETM-DB: Integrated Ethiopian Traditional Herbal Medicine and Phytochemicals Database; GBIF: Global Biodiversity Information Facility; MMFF94: Merck Molecular Force Field; MQN: Molecular Quantum Numbers; MW: Molecular weight; NaPLeS: Natural products likeness scorer; NP: Natural Product; NPASS: Natural Product Activity and Species Source; nHA: Number of hydrogen bond acceptor; nHD: Number of hydrogen bond donor; nRing: Number of rings; nRot: Number of rotatable bonds; NuBBE: Nuclei of Bioassays, Ecophysiology and Biosynthesis of Natural Products Database; p-ANAPL: Pan-African Natural Products Library; PPI-like: Protein–protein inhibitor like; RUBi: Research Unit in Bioinformatics; SANBI: South African National Biodiversity Institute; SANCDB: South African natural

compound database; TPSA: Total Polar Surface Area; t-SNE: T-Distributed Stochastic Neighbor Embedding.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-021-00514-2.

Additional file 1: Fig S1. Distribution of SAscore for SANCDB compounds. X-axis represents SAscores and y-axis quantifies the corresponding probability densities. **Fig. S2.** Scatter plot of compounds molecular weight (MW) versus analogs count. X-axis and y-axis correspond to MW (Dalton) and the number of analogs respectively. **Fig. S3.** Scatter plot of compounds molecular weight (MW) versus analogs count. **Fig. S4.** t-SNE visualization of SANCDB and analogs chemical space. **Fig. S5.** Histogram and kernel density distribution of the scaffolds count. **Fig. S6.** Top 10 SANCDB scaffolds structures and their counts. **Table S1.** SANCDB analogs from Sci-finder for compounds without analogs on Mcule and Molport chemical databases. **Table S2.** A summary of all scaffold structures in SANCDB database.

Availability of data and materials

All data generated or analysed during this study are included in this published article. Supplementary information accompanies this article. SANCDB is freely available at https://sancdb.rubi.ru.ac.za/.

Declarations

Competing interests

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Diallo *et al. J Cheminform* (2021) 13:37

Page 12 of 14

Author details
[1]Research Unit in Bioinformatics (RUBi), Department of Biochemistry and Microbiology, Rhodes University, Makhanda/Grahamstown 6140, South Africa. [2]Department of Chemistry, Rhodes University, Makhanda/Grahamstown 6140, South Africa.

References
1. Sorokina M, Steinbeck C (2020) Review on natural products databases: where to find data in 2020. J Cheminform 12:1–51. https://doi.org/10.1186/s13321-020-00424-9
2. Mishra BB, Tiwari VK (2011) Natural products: an evolving role in future drug discovery. Eur J Med Chem 46:4769–4807. https://doi.org/10.1016/j.ejmech.2011.07.057
3. Chen Y, De Bruyn KC, Kirchmair J (2017) Data resources for the computer-guided discovery of bioactive natural products. J Chem Inf Model 57:2099–2111. https://doi.org/10.1021/acs.jcim.7b00341
4. Newman DJ, Cragg GM (2020) Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. J Nat Prod 83:770–803. https://doi.org/10.1021/acs.jnatprod.9b01285
5. Calixto JB (2019) The role of natural products in modern drug discovery. An Acad Bras Cienc. https://doi.org/10.1590/0001-3765201920190105
6. Hatherley R, Brown DK, Musyoka TM et al (2015) SANCDB: A South African natural compound database. J Cheminform. https://doi.org/10.1186/s13321-015-0080-8
7. Sorokina M, Steinbeck C (2019) Naples: a natural products likeness scorer—web application and database. J Cheminform 11:1–7. https://doi.org/10.1186/s13321-019-0378-z
8. Cockroft NT, Cheng X, Fuchs JR (2019) STarFish: a stacked ensemble target fishing approach and its application to natural products. J Chem Inf Model 59:4906–4920. https://doi.org/10.1021/acs.jcim.9b00489
9. Chen Y, Stork C, Hirte S, Kirchmair J (2019) NP-Scout: machine learning approach for the quantification and visualization of the natural product-likeness of small molecules. Biomolecules 9:43. https://doi.org/10.3390/biom9020043
10. Zeng X, Zhang P, He W et al (2018) NPASS: Natural product activity and species source database for natural product research, discovery and tool development. Nucleic Acids Res. https://doi.org/10.1093/nar/gkx1026
11. Sorokina M, Merseburger P, Rajan K et al (2021) COCONUT online: collection of open natural products database. J Cheminform 13:2. https://doi.org/10.1186/s13321-020-00478-9
12. Veale CGL, Müller R (2020) Recent highlights in anti-infective medicinal chemistry from South Africa. ChemMedChem 15:809–826. https://doi.org/10.1002/cmdc.202000086
13. Musyoka TM, Kanzi AM, Lobb KA, Tastan Bishop Ö (2016) Structure based docking and molecular dynamic studies of plasmodial cysteine proteases against a south african natural compound and its analogs. Sci Rep 6:23690. https://doi.org/10.1038/srep23690
14. Penkler DL, Atilgan C, Tastan Bishop Ö (2018) Allosteric modulation of human hsp90α conformational dynamics. J Chem Inf Model 58:383–404. https://doi.org/10.1021/acs.jcim.7b00630
15. Amusengeri A, Tastan Bishop Ö (2019) Discorhabdin N, a South African natural compound, for Hsp72 and Hsc70 allosteric modulation: combined study of molecular modeling and dynamic residue network analysis. Molecules. https://doi.org/10.3390/molecules24010188
16. Musyoka T, Özlem TB (2019) South African abietane diterpenoids and their analogs as potential antimalarials: novel insights from hybrid computational approaches. Molecules 24:4036. https://doi.org/10.3390/molecules24224036
17. Kimuda MP, Laming D, Hoppe HC, Tastan Bishop O (2019) Identification of novel potential inhibitors of pteridine reductase 1 in *Trypanosoma brucei* via computational structure-based approaches and in vitro inhibition assays. Molecules. https://doi.org/10.3390/molecules24010142
18. Karki N, Verma N, Trozzi F et al (2021) Predicting potential sars-cov-2 drugs-in depth drug database screening using deep neural network framework ssnet, classical virtual screening and docking. Int J Mol Sci 22:1–16. https://doi.org/10.3390/ijms22031392
19. Nyamai DW, Tastan Bishop Ö (2020) Identification of selective novel hits against plasmodium falciparum prolyl tRNA synthetase active site and a predicted allosteric site using in silico approaches. Int J Mol Sci 21:3803. https://doi.org/10.3390/ijms21113803
20. Fantoukh OI, Dale OR, Parveen A et al (2019) Safety assessment of phytochemicals derived from the globalized South African Rooibos Tea (*Aspalathus linearis*) through interaction with CYP, PXR, and P-gp. J Agric Food Chem 67:4967–4975. https://doi.org/10.1021/acs.jafc.9b00846
21. Awolola GV, Sofidiya MO, Baijnath H et al (2019) The phytochemistry and gastroprotective activities of the leaves of *Ficus glumosa*. South African J Bot 126:190–195. https://doi.org/10.1016/j.sajb.2019.01.015
22. Engelhardt C, Petereit F, Lechtenberg M et al (2016) Qualitative and quantitative phytochemical characterization of *Myrothamnus flabellifolia* Welw. Fitoterapia 114:69–80. https://doi.org/10.1016/j.fitote.2016.08.013
23. Bernardini S, Tiezzi A, Laghezza Masci V, Ovidi E (2018) Natural products for human health: an historical overview of the drug discovery approaches. Nat Prod Res 32:1926–1950. https://doi.org/10.1080/14786419.2017.1356838
24. Cragg GM, Newman DJ (2013) Natural products: a continuing source of novel drug leads. Biochim Biophys Acta 1830:3670–3695. https://doi.org/10.1016/j.bbagen.2013.02.008
25. Garcia-Castro M, Zimmermann S, Sankar MG, Kumar K (2016) Scaffold diversity synthesis and its application in probe and drug discovery. Angew Chemie Int Ed 55:7586–7605. https://doi.org/10.1002/anie.201508818
26. Elsevier (2010) Elsevier Developer Portal. In: Elsevier.com. https://dev.elsevier.com/tecdoc_text_mining.html. Accessed 11 Jul 2020
27. CAS (2015) SciFinder - A CAS Solution. In: Publication. http://www.cas.org/products/scifinder. Accessed 24 Oct 2015.
28. Chamberlain S pygbif 0.4.0 documentation—pygbif 0.4.0 documentation. https://pygbif.readthedocs.io/en/latest/index.html. Accessed 4 Jun 2020
29. GBIF. https://www.gbif.org/. Accessed 4 Jun 2020
30. Swain M (2014) PubChemPy: A way to interact with PubChem in Python
31. Mendez D, Gaulton A, Bento AP et al (2019) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 47:D930–D940. https://doi.org/10.1093/nar/gky1075
32. Lo EJ, Iynkkaran I, Li C et al (2017) DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 46:D1074–D1082. https://doi.org/10.1093/nar/gkx1037
33. Sterling T, Irwin JJ (2015) ZINC 15—ligand discovery for everyone. J Chem Inf Model 55:2324–2337. https://doi.org/10.1021/acs.jcim.5b00559
34. Kim S, Chen J, Cheng T et al (2019) PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 47:D1102–D1109. https://doi.org/10.1093/nar/gky1033
35. Djoumbou Feunang Y, Eisner R, Knox C et al (2016) ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminform 8:1–20. https://doi.org/10.1186/s13321-016-0174-y
36. Release S (2017) 1: Maestro. Schrödinger LLC, New York, p 2017
37. Kiss R, Sandor M, Szalai FA (2012) http://Mcule.com: a public web service for drug discovery. J Cheminform. DOI:https://doi.org/10.1186/1758-2946-4-s1-p17
38. Easy compound ordering service - MolPort. https://www.molport.com/shop/index. Accessed 11 Jul 2020
39. O'Boyle NM, Banck M, James CA et al (2011) Open babel: an open chemical toolbox. J Cheminform 3:33. https://doi.org/10.1186/1758-2946-3-33
40. Molecular fingerprints and similarity searching—Open Babel v2.3.0 documentation. http://openbabel.org/docs/dev/Features/Fingerprints.html. Accessed 20 Sep 2020
41. Pilón-Jiménez BA, Saldívar-González FI, Díaz-Eufracio BI, Medina-Franco JL (2019) BIOFACQUIM: a Mexican compound database of natural products. Biomolecules. https://doi.org/10.3390/biom9010031
42. Kearney SE, Zahoránszky-Kőhalmi G, Brimacombe KR et al (2018) Canvass: a crowd-sourced, natural product screening library for exploring biological space. ACS Cent Sci. https://doi.org/10.26434/CHEMRXIV.7172369.V2
43. Sánchez-Cruz N, Pilón-Jiménez BA, Medina-Franco JL (2020) Functional group and diversity analysis of BIOFACQUIM: a Mexican natural product database. F1000Res 8:2071. https://doi.org/10.12688/f1000research.21540.2

Diallo *et al. J Cheminform*   (2021) 13:37

Page 13 of 14

44. Singh N, Guha R, Giulianotti MA et al (2009) Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. J Chem Inf Model 49:1010–1024. https://doi.org/10.1021/ci800426u

45. Saldívar-Gonzaíez FI, Valli M, Andricopulo AD et al (2019) Chemical space and diversity of the NuBBE database: a chemoinformatic characterization. J Chem Inf Model. https://doi.org/10.1021/acs.jcim.8b00619

46. The Scopy's documentation—Scopy 1.2.3 documentation. https://scopy.iamkotori.com/index.html. Accessed 10 Aug 2020

47. Ertl P, Rohde B (2012) The molecule cloud—compact visualization of large collections of molecules. J Cheminform 4:1. https://doi.org/10.1186/1758-2946-4-12

48. Pilon AC, Valli M, Dametto AC et al (2017) NuBBEDB: An updated database to uncover chemical and biological information from Brazilian biodiversity. Sci Rep. https://doi.org/10.1038/s41598-017-07451-x

49. Nguyen KT, Blum LC, Van Deursen R, Reymond JL (2009) Classification of organic molecules by molecular quantum numbers. ChemMedChem 4:1803–1805. https://doi.org/10.1002/cmdc.200900317

50. Landrum G (2016) RDKit: open-source cheminformatics software

51. Van Der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2625

52. Janssen AP, Grimm SH, Wijdeven MRH et al (2018) Drug discovery maps, a machine learning model that visualizes and predicts kinome—inhibitor interaction landscapes. J Chem Inf Model. https://doi.org/10.1021/acs.jcim.8b00640

53. Naveja JJ, Medina-Franco JL (2019) Finding constellations in chemical space through core analysis. Front Chem 7:510. https://doi.org/10.3389/fchem.2019.00510

54. Yosipof A, Guedes RC, García-Sosa AT (2018) Data mining and machine learning models for predicting drug likeness and their disease or organ category. Front Chem 6:162. https://doi.org/10.3389/fchem.2018.00162

55. Pedregosa F, Varoquaux GG, Gramfort A et al (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830

56. sklearn.manifold.TSNE—scikit-learn 0.23.1 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html. Accessed 17 Jul 2020

57. Tosco P, Stiefl N, Landrum G (2014) Bringing the MMFF force field to the RDKit: Implementation and validation. J Cheminform 6:37. https://doi.org/10.1186/s13321-014-0037-3

58. Morris GM, Huey R, Lindstrom W et al (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 30:2785–2791. https://doi.org/10.1002/jcc.21256

59. Kluyver T, Ragan-Kelley B, Pérez F, et al (2016) Jupyter Notebooks-a publishing format for reproducible computational workflows. In: ELPUB. pp 87–90

60. McKinney W (2011) pandas: a Foundational python library for data analysis and statistics. Python High Perform Sci Comput 1–9.

61. McKinney W (2010) Data structures for statistical computing in Python. Proc 9th Python Sci Conf. https://doi.org/10.25080/majora-92bf1922-00a

62. Brugman S (2019) pandas-profiling: exploratory data analysis for Python

63. Ntie-Kang F, Telukunta KK, Döring K et al (2017) NANPDB: a resource for natural products from northern African sources. J Nat Prod 80:2067–2076. https://doi.org/10.1021/acs.jnatprod.7b00283

64. Ntie-Kang F, Zofou D, Babiaka SB et al (2013) AfroDb: a select highly potent and diverse natural product library from African medicinal plants. PLoS ONE. https://doi.org/10.1371/journal.pone.0078085

65. Ntie-Kang F, Onguéné PA, Scharfe M et al (2014) ConMedNP: a natural product library from Central African medicinal plants for drug discovery. RSC Adv 4:409–419. https://doi.org/10.1039/c3ra43754j

66. Ntie-Kang F, Mbah JA, Mbaze LMa et al (2013) CamMedNP: Building the Cameroonian 3D structural natural products database for virtual screening. BMC Complement Altern Med. https://doi.org/10.1186/1472-6882-13-88

67. Ertl P, Schuhmann T (2020) Cheminformatics analysis of natural product scaffolds: comparison of scaffolds produced by animals, plants, fungi and bacteria. Mol Inform. https://doi.org/10.1002/minf.202000017

68. Castells E, Mulder PPJ, Pérez-Trujillo M (2014) Diversity of pyrrolizidine alkaloids in native and invasive Senecio pterophorus (Asteraceae): Implications for toxicity. Phytochemistry 108:137–146. https://doi.org/10.1016/j.phytochem.2014.09.006

69. Kuroda M, Ori K, Mimaki Y (2006) Ornithosaponins A-D, four new polyoxygenated steroidal glycosides from the bulbs of Ornithogalum thyrsoides. Steroids 71:199–205. https://doi.org/10.1016/j.steroids.2005.10.001

70. Ornithogalum thyrsoides | PlantZAfrica. http://pza.sanbi.org/ornithogalum-thyrsoides. Accessed 15 Jun 2020

71. Ornithogalum saundersiae | PlantZAfrica. http://pza.sanbi.org/ornithogalum-saundersiae. Accessed 15 Jun 2020

72. Iguchi T, Kuroda M, Naito R et al (2019) Cholestane glycosides from Ornithogalum saundersiae bulbs and the induction of apoptosis in HL-60 cells by OSW-1 through a mitochondrial-independent signaling pathway. J Nat Med 73:131–145. https://doi.org/10.1007/s11418-018-1252-4

73. Davies-Coleman M, Veale C (2015) Recent advances in drug discovery from South African marine invertebrates. Mar Drugs 13:6366–6383. https://doi.org/10.3390/md13106366

74. Pettit GR, Kamano Y, Dufresne C et al (1989) Isolation and structure of the unusual Indian Ocean Cephalodiscus gilchristi components, cephalostatins 5 and 6. Can J Chem 67:1509–1513. https://doi.org/10.1139/v89-231

75. Knott MG, Mkwananzi H, Arendse CE et al (2005) Plocoralides A-C, polyhalogenated monoterpenes from the marine alga *Plocamium corallorhiza*. Phytochemistry 66:1108–1112. https://doi.org/10.1016/j.phytochem.2005.03.029

76. Mann MGA, Mkwananzi HB, Antunes EM et al (2007) Halogenated monoterpene aldehydes from the South African marine alga *Plocamium corallorhiza*. J Nat Prod 70:596–599. https://doi.org/10.1021/np060547c

77. SANBI (2019) Threatened Species Programme | SANBI Red List of South African Plants. In: South African Natl Biodivers Inst http://redlist.sanbi.org/stats.php. Accessed 9 Jul 2020

78. Banerjee P, Erehman J, Gohlke BO et al (2015) Super natural II-a database of natural products. Nucleic Acids Res. https://doi.org/10.1093/nar/gku886

79. Bultum LE, Woyessa AM, Lee D (2019) ETM-DB: integrated Ethiopian traditional herbal medicine and phytochemicals database. BMC Complement Altern Med. https://doi.org/10.1186/s12906-019-2634-1

80. Ntie-Kang F, Amoa Onguéné P, Fotso GW et al (2014) Virtualizing the p-ANAPL Library: a step towards drug discovery from African medicinal plants. PLoS ONE 9:e90655. https://doi.org/10.1371/journal.pone.0090655

81. Chen Y, Garcia De Lomana M, Friedrich NO, Kirchmair J (2018) Characterization of the chemical space of known and readily obtainable natural products. J Chem Inf Model 58:1518–1532. https://doi.org/10.1021/acs.jcim.8b00302

82. Ertl P, Schuffenhauer A (2009) Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. J Cheminform 1:8. https://doi.org/10.1186/1758-2946-1-8

83. Elkhattabi L, Charoute H, Saile R, Barakat. (2020) A computational approach revealed potential affinity of antiasthmatics against receptor binding domain of 2019n-cov spike glycoprotein. https://doi.org/10.26434/chemrxiv.12115638.v1

84. Nyamai DW, Tastan Bishop Ö (2019) Aminoacyl tRNA synthetases as malarial drug targets: a comparative bioinformatics study. Malar J 18:1–27. https://doi.org/10.1186/s12936-019-2665-6

85. Meyers J, Carter M, Mok NY, Brown N (2016) On the origins of three-dimensionality in drug-like molecules. Future Med Chem 8:1753–1767. https://doi.org/10.4155/fmc-2016-0095

86. Limban C, Nuţă DC, Chiriţă C et al (2018) The use of structural alerts to avoid the toxicity of pharmaceuticals. Toxicol Reports 5:943–953

87. CHEMBL1766622 Compound Report Card. https://www.ebi.ac.uk/chembldb/index.php/compound/inspect/CHEMBL2079699. Accessed 3 Sep 2020

88. Simon L, Abdul Salam AA, Madan Kumar S et al (2017) Synthesis, anticancer, structural, and computational docking studies of 3-benzyl-chroman-4-one derivatives. Bioorganic Med Chem Lett 27:5284–5290. https://doi.org/10.1016/j.bmcl.2017.10.026

Diallo *et al. J Cheminform*   *(2021) 13:37*

Page 14 of 14

89. Baell JB, Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J Med Chem 53:2719–2740. https://doi.org/10.1021/jm901137j

Publisher's Note