Pseudo-Natural Products Occur Frequently in Biologically Relevant Compounds

José-Manuel Gally,[†] Axel Pahl,[‡] Paul Czodrowski,[§] Herbert Waldmann[*,†,§]

[†]Department of Chemical Biology, Max-Planck-Institute of Molecular Physiology, Otto-Hahn-Straße 11, 44227 Dortmund, Germany, [‡]Compound Management and Screening Center, Dortmund, Otto-Hahn-Str. 11, 44227 Dortmund, Germany, [§]Faculty of Chemistry and Chemical Biology, Technical University Dortmund, Otto-Hahn-Straße 6, 44227 Dortmund, Germany

Supporting Information Placeholder

ABSTRACT: A new methodology for classifying fragment combinations and characterizing pseudo-natural products (PNPs) is described. The source code is based on open-source tools and is organized as a Python package. Tasks can be executed individually or within the context of scalable, robust workflows. First, structures are standardized and duplicate entries are filtered out. Then, molecules are probed for the presence of predefined fragments. For molecules with more than one match, fragment combinations are classified. The algorithm considers the pair-wise relative position of fragments within the molecule (fused atoms, linkers, intermediary rings), resulting in 18 different possible fragment combination categories. Finally, all combinations for a given molecule are assembled into a fragment combination graph, with fragments as nodes and combination types as edges. This workflow was applied to characterize PNPs in the ChEMBL database via comparison of fragment combination graphs with Natural Product (NP) references, represented by the Dictionary of Natural Products. The Murcko fragments extracted from 2,000 structures previously described were used to define NP-fragments. The results indicate that ca. 23% of the biologically relevant compounds listed in ChEMBL comply to the PNP definition, and that, therefore, PNPs occur frequently among known biologically relevant small molecules. The majority (>95%) of PNPs contains two to four fragments, mainly (>95%) distributed in five different combination types. These findings may provide guidance for the design of new PNPs.

Natural products (NPs) are a rich source of inspiration for drug discovery, and compounds derived from or inspired by natural product structure constitute a major fraction of currently available drugs.[1,2] In light of this proven relevance, design and synthesis of novel bioactive compounds can benefit from the inclusion of structural properties derived from natural products. We have recently introduced the concept of pseudo-natural products[3–5] (PNPs) as novel NP-inspired compound classes which combine the biological relevance of NPs with the efficient exploration of chemical space by fragment-based compound design.[6] In PNPs natural product-derived fragments are combined in unprecedented arrangements which are not available by current biosynthesis pathways. They inherit the biological relevance of NPs, yet explore biologically relevant regions of chemical space not accessed by nature, and it can be expected that PNPs may have novel or different bioactivity and targets compared to the guiding NPs.

Synthesis and biological evaluation of several PNP collections provided proof-of-principle for the concept (**Figure 1**). For instance, the fusion of indole- with morphan fragments and chromane- with tetrahydropyrimidinone fragments respectively resulted in the indomorphan[7] **1**, and chromopynone[8] **5** compound classes, which define novel inhibitors of the glucose transporters GLUT-1 and -3. Moreover, fusion of the indole- and tropane fragments yielded the indotropane compound class, from which Myokinasib[9] **2**, a MLCK1 inhibitor could be identified. Another recent PNP class are indocinchona alkaloids[10] which are obtained by fusion of indole- and cinchona alkaloid fragments. Among this compound class, Azaquindole-1 **3** inhibited the lipid kinase VPS34, thereby suppressing starvation- and rapamycin-induced autophagy. The recombination of pyridine- and dihydropyran fragments led to the pyrano-furo-pyridone[11] PNPs **6**, novel reactive oxygen species inducers and inhibitors of mitochondrial complex I. Li *et al* mimicked the biosynthesis of penilactones by synthesis of the PNP penindolone[12] **4**, constituted of an indole- and two clavatol fragments. Penindolone showed broad-spectrum anti-influenza A activity. Similarly, Yuan *et al.* combined the benzodiazepine- and isoindolinone scaffolds, found in many drugs and NPs and endowed with broad biological activity, into tetracyclic benzodiazepine-fused isoindolinones **7**.[13]

In addition, cheminformatic analysis revealed that large screening libraries are biased towards biogenic molecules that proteins have evolved to recognize, i.e. natural products and related compound classes.[14] This bias may reflect the historical focus of medicinal chemistry on natural products and the resulting synthesis efforts.

These observations suggest that compound classes that match the PNP definition might have been synthesized and biologically evaluated before, without inspiration by the PNP design principle, for instance driven by intuitive inclusion of different natural product structures in medicinal chemistry and chemical biology synthesis programs (**Figure 1**).

For instance, the cycloocta[b]indole compound class[15], which was designed and synthesized following the Biology-Oriented Synthesis principle based on the NP macroline and which targets the mycobacterial phosphatase MptpB, could be described as the result of the fusion of indole- and piperidone fragments. Therefore, in hindsight it was termed indopepenone[3] **8**. Other examples include the carbazopyrrolone-[16] **11** and pyrrofuranolactone-[17] **9** compound classes. In some cases, bioactivity could be detected, e.g. for piperazopyridones[18] **12** (TRPV6 calcium channels inhibitors), diazaspiro alkanes[19] **13** (dopamine D_3 antagonists) and thiazolo-nootkatones[20] **10** (antimicrobial agents).

Hence, PNPs might already constitute a larger fraction of currently available and applied bioactive small molecules. They might have proven to be endowed with biological relevance in general, to display diverse bioactivity and to constitute already widely explored natural product inspired chemotypes in chemical biology and medicinal chemistry research and drug discovery. Such wide application and exploration would validate the PNP principle in a general sense.

In order to explore this possibility, we have analyzed the ChEMBL[21] database, which lists biologically relevant small molecules including their structure and activity, for compounds that conform to the PNP definition. We report that ca. 23% of the biologically relevant compounds listed in ChEMBL and considered in the analysis can be classified as PNPs, and that, therefore, PNPs occur frequently among known biologically relevant small molecules. Based on our analysis, we conclude that the majority (>95%) of PNPs in ChEMBL represent combinations of two to four fragments and five fragment combination types. This finding may provide guidance for the design of new PNPs.

RESULTS AND DISCUSSION

For identification and analysis of the NP fragments and NP fragment combinations, we established an analysis package written in Python 3 and termed it NPFC (natural product fragment combination). NPFC consists of a set of modules and scripts and employs the open-source libraries RDKit[22] (v. 2020.09.1) for processing chemical structures, Pandas[23] (v. 1.1.4) for data handling, Networkx[24] (v. 2.5) for representing fragment connectivity as graphs and Snakemake[25] (v. 5.27.4), for coordinating the different tasks as workflows. The modules and scripts comprise preparation steps including different filtering operations, fragment search, combination and classification steps, establishment of fragment combination graphs and finally PNP annotation (Figure 2).

For the analysis three different data sets were employed, i.e. natural product fragments to be used in the fragment search, data for natural products and the fragment combinations in them and synthetic non-natural compounds listed in ChEMBL to be analyzed for fragment- and fragment combination content.

Preparation of the datasets

For the identification of NP fragments, we employed a set of structurally diverse and biologically-relevant 2,000 NP-derived fragments, identified previously[26], and available in SDF format. The processing of fragments consisted in two main steps, i.e. the preparation of the dataset (load, standardize, deduplicate, see below) and the annotation of fragment combination points. Five structures in the dataset could not be parsed and converted to the RDKit format and were manually curated (see the Supporting Information).

Subsequently, the input file was loaded without errors into a Pandas DataFrame with structures in RDKit format. Next, a set of sequential tasks were applied to the structures for standardization. First, records with empty structures are filtered out. Mixtures are cleared, preferring the largest, non-linear organic compound possible. Structures were further altered using the functionalities of MolVS[27] as implemented within RDKit. Thus, atoms were set to most common isotope only, functional groups were normalized, formal charges were removed whenever possible, the canonical tautomer was enumerated and stereochemistry information was removed. Finally, Murcko scaffolds were extracted from the structures, followed by another round of removing formal charges. None of these steps resulted in the elimination of compounds from the dataset.

Duplicate structures were then filtered out using InChI Key as identity, decreasing the size of the data set to 1,673 entries (84% of the initial dataset). To obtain the best possible depiction for each structure, different methods available in the RDKit were applied to generate 2D coordinates (CoordGen and rdDepictor). A third-party library[28] was employed to score the depictions based on their number of overlapping bonds and atoms and to keep the depiction with the lowest score. Input coordinates were considered as well, when available.

Finally, to identify redundant fragment orientations in combinations, symmetry classes within the structures were annotated as fragment combination points by performing a fragment search of each fragment within itself.[29]

To define the NP chemical space, the Dictionary of Natural Products[30] (DNP, 318,271 records), which is the result of a comprehensive curation and integration of NP structures with known biological origin, was consulted. The corresponding input SDF had first to be converted to UNIX format using the dos2unix utility, before it could be processed. Stereochemical information was absent from the structures.

To scale up the computation on a cluster, the input SDF was split into chunks of 5,000 entries, for a total of 64 chunks, which were then processed almost completely independently. Molecules were converted to RDKit format with minimal losses (0.13% of the initial dataset), with most errors due to incompatibility in aromaticity perception by the RDKit.

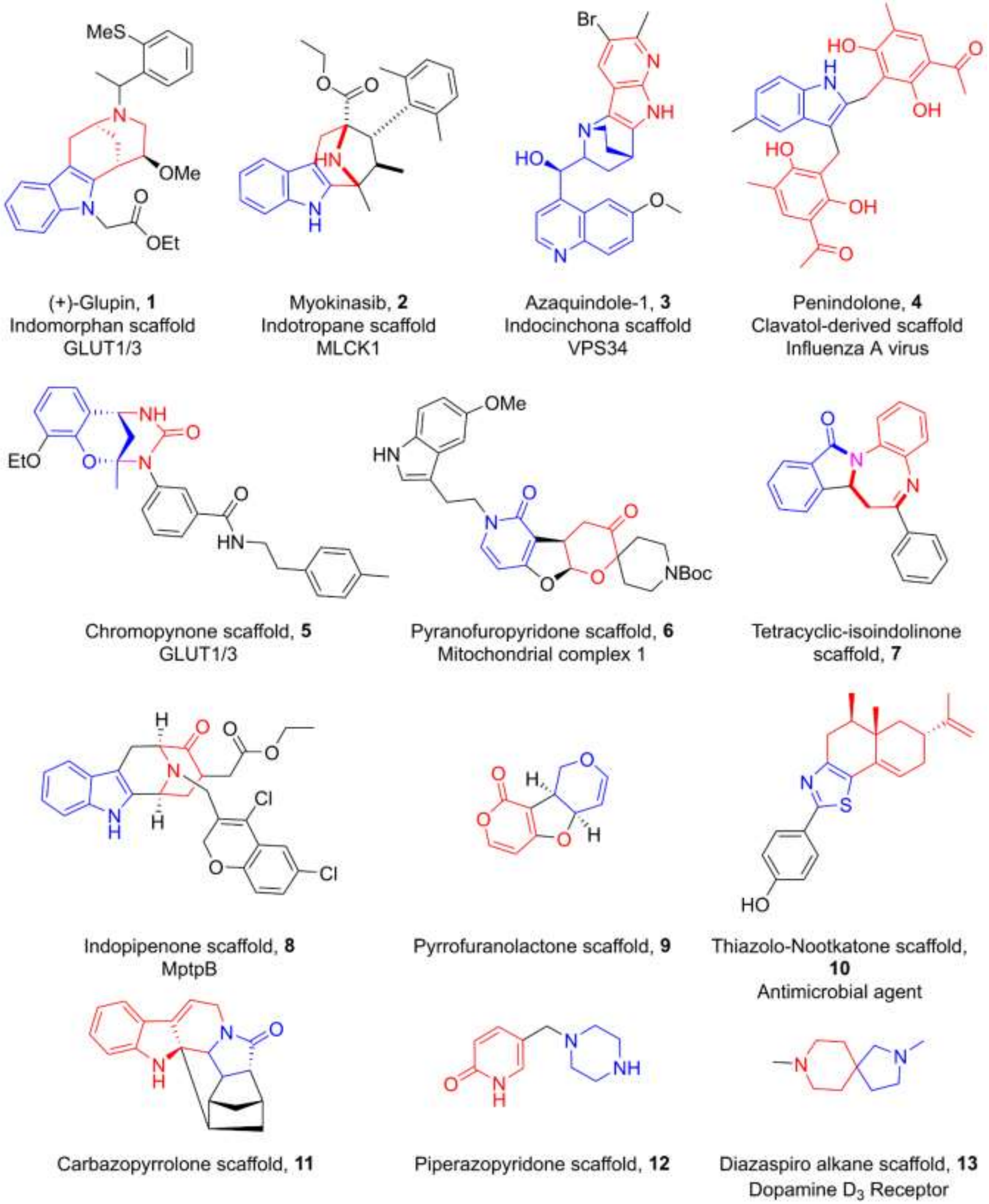
Structures were standardized with the following procedure. First, empty structures were removed (9.15% of the initial dataset). Then, metal atoms were disconnected from the structures and organic non-linear minor compounds were extracted from mixtures, when applicable. To remove sugar units in NPs[31], an in-house script was developed using the RDKit, to detect sugar-like rings and peel them off iteratively, starting from the outer layer on the molecule. A set of filters was then applied to remove unwanted entries, based on the number of heavy atoms ($x \geq 4$, 0.03%), molecular weight ($x \leq 1000.0$ Da, 1.25%), number of rings ($x \geq 1$, 7.26%) and chemical elements (only authorized: H, B, C, N, O, F, P, S, Cl, Br, I, 0.08%). Since the structures were altered from earlier operations, they were explicitly sanitized to update the various computational properties of atoms and bonds and possibly avoid errors downstream. Isotopes were then set to their default, most occurring form and functional groups were normalized. Then, charges were removed on molecules wherever possible, canonical tautomers enumerated and, for consistency, stereochemical information removed when applicable. Finally, the structures were regenerated from SMILES. To avoid longer computational times, due to only a small fraction of problematic structures in the dataset, a timeout was set to 10s per molecule for the entire standardization (1.43% of the data was removed by the timeout).

For deduplication, a common reference file was used for all chunks. This reference file contained the list of all already known InChI Keys and was updated for each chunk independently, using a lock to avoid simultaneous accesses to the file. Since the stereochemistry was not considered, a large portion of the dataset (28.67%) was found to be duplicates and was therefore filtered out, further decreasing the size of the dataset to 165,467 records (52% of the initial dataset).

To define synthetic compounds, the ChEMBL dataset was downloaded from the official website[32] as a single SDF of 1,941,411 structures, then divided in 389 chunks of 5,000 records. For consistency, the exact same preparation protocol as described above for the DNP was applied, which decreased the total number of entries to 1,668,022 (85.92%), mainly due to filters: duplicates

(9.25%), molecular weight (1.70%), timeout (1.21%), and number of rings (1.06%).

Additionally, NPs were removed from the ChEMBL dataset to better differentiate NPs from synthetic compounds. To achieve this, duplicate structures of DNP inside of ChEMBL were filtered out by means of InChI Key comparison, decreasing the synthetic dataset to 1,632,769 (84.10%) records.



(+)-Glupin, **1**
Indomorphan scaffold
GLUT1/3

Myokinasib, **2**
Indotropane scaffold
MLCK1

Azaquindole-1, **3**
Indocinchona scaffold
VPS34

Penindolone, **4**
Clavatol-derived scaffold
Influenza A virus

Chromopynone scaffold, **5**
GLUT1/3

Pyranofuropyridone scaffold, **6**
Mitochondrial complex 1

Tetracyclic-isoindolinone scaffold, **7**

Indopipenone scaffold, **8**
MptpB

Pyrrofuranolactone scaffold, **9**

Thiazolo-Nootkatone scaffold, **10**
Antimicrobial agent

Carbazopyrrolone scaffold, **11**

Piperazopyridone scaffold, **12**

Diazaspiro alkane scaffold, **13**
Dopamine D_3 Receptor

Figure 1. Examples of pseudo-natural product scaffolds. Red and blue colours denote NP fragments.

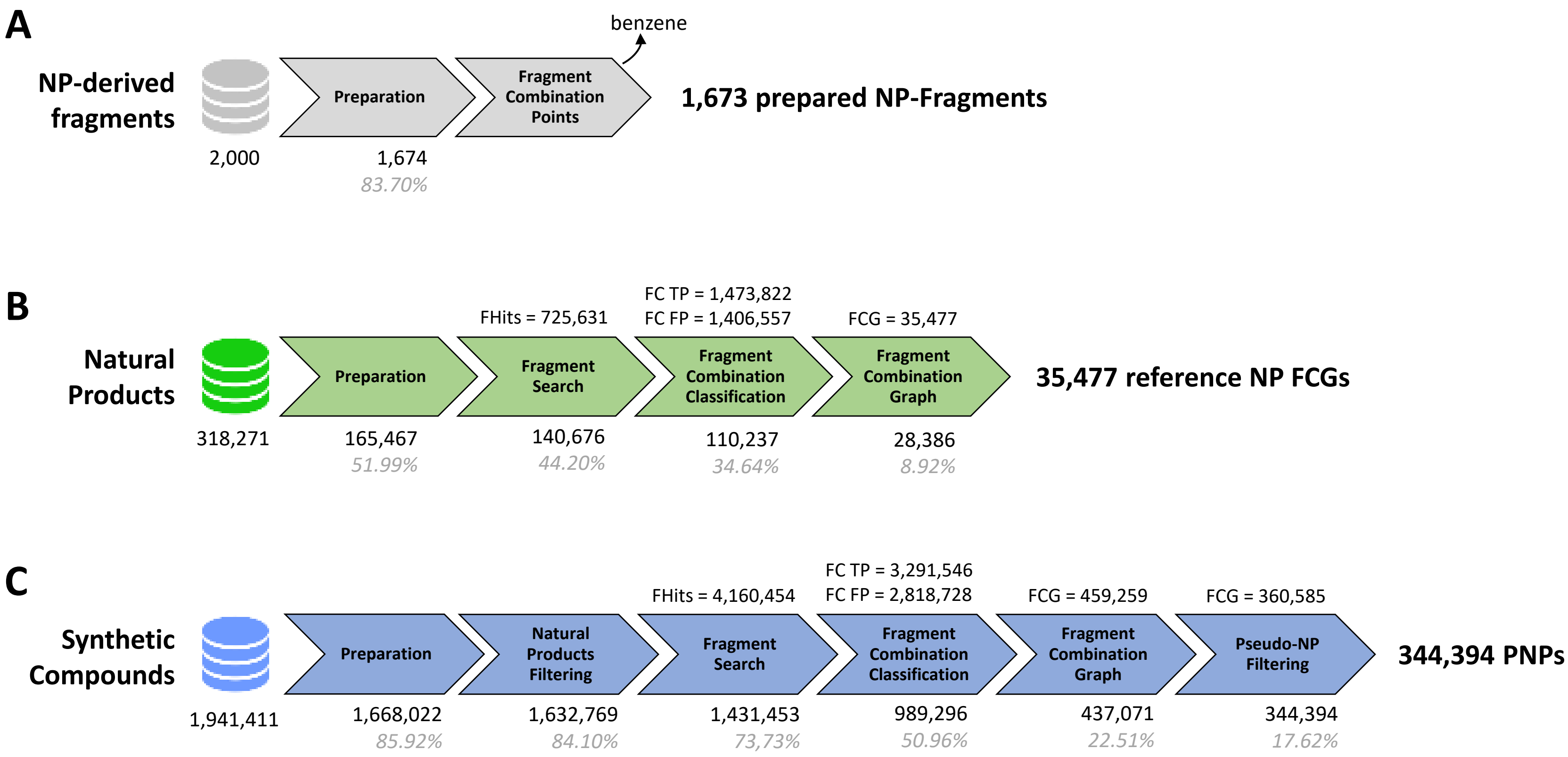


Figure 2. Workflows applied to the different datasets. *A: the workflow for fragments; B: the workflow for natural Products (DNP); C: the workflow for synthetic compounds (ChEMBL). The number of remaining molecules at each step is displayed below the tasks, when changes occur. Below it, the percentage of remaining molecules in regards to the initial number is displayed in grey. Above tasks, the number observed elements is displayed, when different from molecules. FHits: Fragment Hits; FC: Fragment combinations; TP: True Positive; FP: False Positive; Matches: the number of fragment hits; FCG: Fragment Combination Graphs*

Fragment Search

A substructure search was performed to identify all fragments occurring in the NPs (fragment hits). Initial results showed an abundance of the benzene fragment, accounting for respectively 16% and 38% of all fragments hits in DNP and ChEMBL. This high prevalence, in particular for ChEMBL, reflected in subsequent results, introduced a bias in our conclusions (see the Supporting Information for complete results). Benzene was therefore removed from the fragment pool and was not further considered in this study.

Following this procedure, 140,676 molecules (85% of the remaining molecules) were found to contain at least one NP fragment, for a total of 725,631 fragment hits. This high proportion observed for NPs containing NP fragments was expected, especially since the fragments originated from an earlier version of the same database (DNP 18.2).[26]

As for the NPs, the benzene was also removed from the pool of fragments used for the fragment search in ChEMBL which resulted in 4,160,454 fragment hits, for a total of 1,431,453 compounds (73.73% of the initial dataset).

Fragment Combination

The initial classification proposed by Karageorgis *et al.*[7] introduced 5 different connectivity types, i.e. monopodal- and bipodal connections, as well as spiro-, edge- and bridged fusions. To better capture the complexity of NP fragment arrangements, we extended this classification and considered up to 18 fragment combinations and grouped them in categories, types and subtypes, when applicable. For this step, only the fragment atoms found in rings were considered for the classification of combinations. For each molecule, all possible fragment pairs were investigated independently.

The first step of our algorithm for classifying fragment combinations was to identify atoms involved in a fusion between both fragments. If any combination was found, then the combination category was defined as "fusion". The number of fused atoms then determined the type of fusion (see Figure 3 for a graphical depiction):

n == 1 fused atom: spiro

n == 2 fused atoms: edge

3 ≤ n ≤ 5 fused atoms: bridged

n > 5 fused atoms: linker

This resulted in four classes of fusions: fusion spiro (**fs**), fusion edge (**fe**), fusion bridged (**fb**) and fusion linker (**fl**).

The combinations of fragments, that did not have any fused atoms, were categorized as connections. Intermediary rings between both fragments were defined as rings in the molecule, that contained fused atoms with both fragments. The number of identified intermediary rings determined the number of distinct paths that led one fragment to another. This number defined the degree of connection of the combination:

n == 1 path: monopodal (no intermediary ring)

n == 2 paths: bipodal (1 intermediary ring)

n == 3 paths: tripodal (2 intermediary rings)

n > 3 paths: other (>2 intermediary rings)

For instance, a bipodal connection could be described as two fragments having no fused atoms with each other but sharing one intermediary ring between them.

A subtype was defined for bipodal connections and of higher degree as well. For each intermediary ring, the atoms from each fragment, that were also present in the intermediary ring, constituted the fragment connection points (CP). Hence, the number of CPs indicated the interface exposure of each fragment with the intermediary ring considered. For consistency, the same nomenclature used for the type of fusions was applied:

n == 1 CP for any of the two fragments: spiro

n = 2 CP for both fragments: edge

3 ≤ n ≤ 5 CP for any of the two fragments: bridged

n > 5 CP for any of the two fragments: linker

Since there could be multiple intermediary rings, several subtypes could be available. Thus, a priority had to be set to decide the subtype of the connection, considering all intermediary rings. The following order was retained to highlight less common fragment combinations: linker > spiro > bridged > edge.

For fragment combinations with no intermediary rings, a distinction was made between monopodal connections (fragments connected through a linker) and annulated connections, where both fragments belonged to the same ring complex but were separated by more than one ring.

In total, 14 fragment combinations (see Figure 3) were found to be connections: connection monopodal (**cm**), connection annulated (**ca**), connection bipodal spiro (**cbs**), connection bipodal edge (**cbe**), connection bipodal bridged (**cbb**), connection bipodal linker (**cbl**), connection tripodal spiro (**cts**), connection tripodal edge (**cte**), connection tripodal bridged (**ctb**), connection tripodal linker (**ctl**), connection other spiro (**cos**), connection other edge (**coe**), connection other bridged (**cob**), connection other linker (**col**).

In addition to the 18 fragment combination categories described above (fusions and connections), three cases were considered to be false positives (FP). The first class of false positives occurred when two fragments were too far apart from each other in the molecule. A maximum distance threshold, for considering two fragments to have a meaningful combination, was set to 3 intermediary atoms between both fragments. Any combination with more intermediary atoms than the threshold was systematically ignored (connection false positive cut off, **cfc**). These combinations were directly filtered out during classification and were therefore not included in the count of false positives at the end of the computation.

The second class of false positives concerned fragment inclusion. If one fragment completely contained another, then the other fragment was simply a substructure of the first fragment (fusion false positive substructure, **ffs**). In this case, only the larger fragment was retained for the analysis.

The third class of false positives was found for fragment hits with a large proportion of atoms in common. The results indicated that they were not actually fused by synthesis design, but rather overlapping as an artefact of the fragment search. To identify such cases, an arbitrary rule was established, that if the atoms in common between the two fragments constituted a full ring in the molecule, then it was not considered to be a proper fusion, but rather an overlap of the fragments (fusion false positive overlap, **ffo**).

Finally, molecules with at least one valid combination were kept at the end of this step.

For NPs, this step amounted to 110,237 remaining NPs accounting for a total of 2,880,379 fragment combinations. Half of those combinations (48.84%) were actually false positives (ffs or ffo) and the vast majority of the remaining NPs contained at least one ffs combination (84.72%), up to a maximum of 14, whereas about half (47.63%) contained at least one ffo combination (up to 9). Only 12.07 % of the NPs at that stage had no false positive combinations.

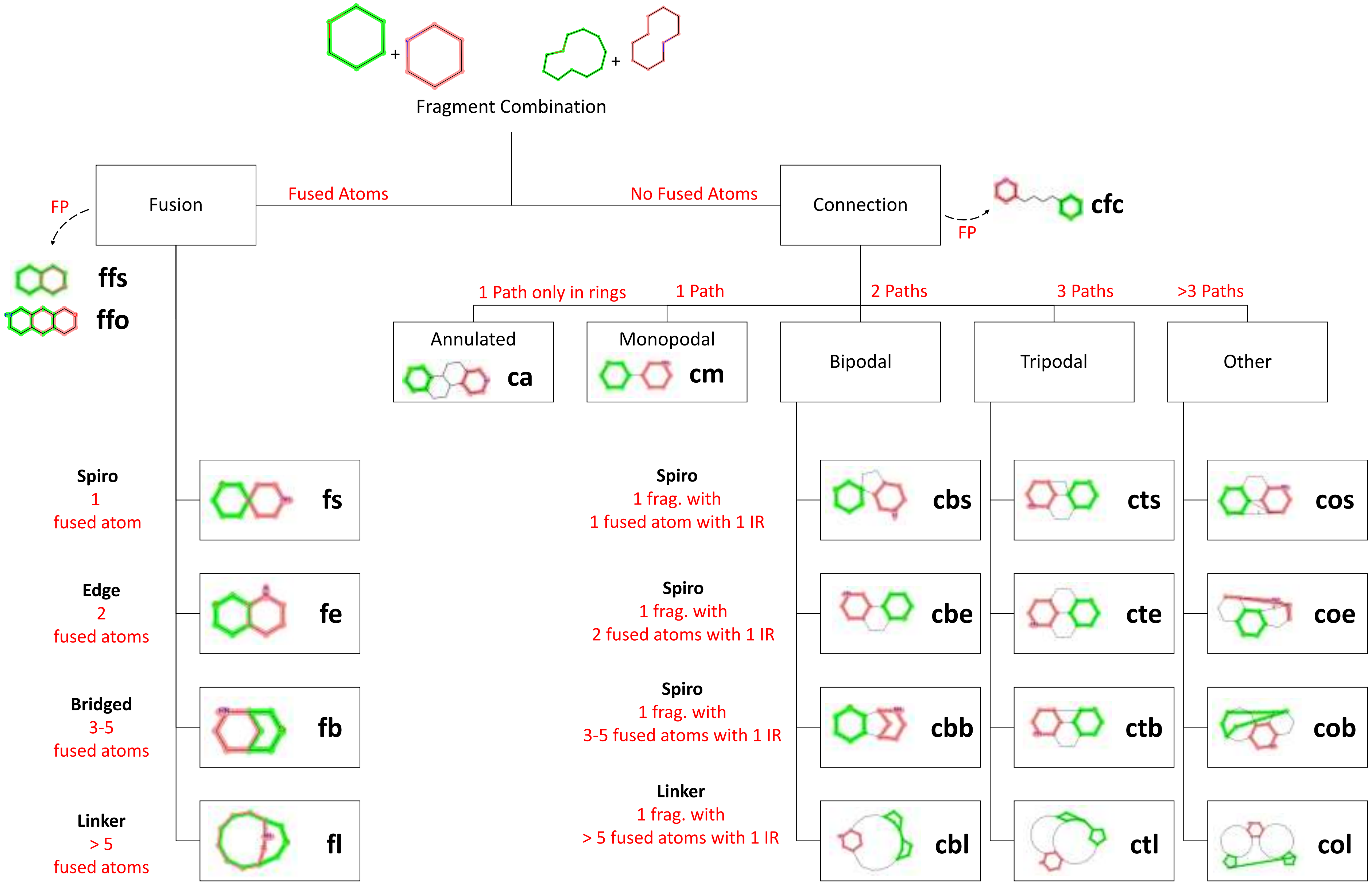


Figure 3. Decision tree applied for classification of fragment combinations. Fragments are highlighted in red and green; fused atoms and bonds are highlighted in both red and green; IR: Intermediary Ring, defined as a ring directly located between both fragments.

For ChEMBL, this step resulted in 989.296 molecules (50.96% of the initial dataset) accounting for 6,110,274 fragment combinations. Also, in this case, only less than half of the identified combinations (46.13%) constituted valid fragment combinations. The majority of synthetic compounds (68.67%) contained at least one false positive combination, with **ffs** combinations being much more represented than **ffo** combinations (respectively 66.55% and 13.76% of remaining compounds having at least one of those).

Fragment Combination Graph

To represent the entire fragment connectivity, fragment combinations were assembled into fragment combination graphs (FCG), with fragment types as nodes and combinations as edges.

In case of false positive combinations ffs (one fragment was a substructure of another), only the larger fragment was considered for the analysis, hence avoiding redundant edges. Moreover, fragment graphs that contained overlapping fragments were considered to represent alternative fragment connectivities and were therefore split up into different FCGs. To avoid a combinatorial explosion of the number of graphs due to overlaps, a maximum threshold of 5 overlaps (**ffo** combinations) was set per molecule. Structures above this threshold were filtered out. In addition, graphs containing disconnected subgraphs were separated into new entries as well.

Only graphs containing at least one valid fragment combination were further considered, decreasing the final number of NPs in the dataset to 28,386 structures (8.92% of the initial size), for a total of 35,477 fragment combination graphs.

For ChEMBL, the assembly of FCGs resulted in further reduction of the dataset size by half (437,071 structures, 22.51% of the initial dataset). The relatively low amount of **ffo** combinations resulted in most molecules containing only 1 FCG (86.24%), for a total of 459,259 graphs

Pseudo-Natural Product Filtering

A PNP was defined as a synthetic compound with a NP fragment connectivity not found in any known NP. Hence, each FCG of a probed molecule was compared with each graph of every NP. In case that the probed molecule's FCG differed from each NP graph by at least one fragment combination, the FCG was labelled as PNP. In contrast, if an NP graph is found to contain the same fragment combinations as the probed molecule graph, the latter was labelled as NP-like because it represents fragment combinations found in a NP. This can lead to a situation that one molecule can contain graphs with PNP- and NP-like character at the same time. If at least one graph of the probed molecule was labelled as PNP, then the molecule was identified as a PNP. Otherwise, it was labelled as NP-like. To prioritize innovative fragment combinations rather than repetitions, the number of occurrences of a fragment combination was not considered during comparison. Indeed, a molecule with 3 repeated fragment combinations would be matched with a molecule containing only 1 of the same fragment combinations.

The pair-wise comparison of the 459,259 FCGs from ChEMBL with each of the 35,477 FCGs from the DNP resulted in the identification of 344,334 PNPs (example in **Figure 4**). This accounted for 78.80% of the remaining synthetic compounds, i.e. containing at least two NP-derived fragments involved in at least one meaningful combination, and 17.62% of the whole ChEMBL dataset.

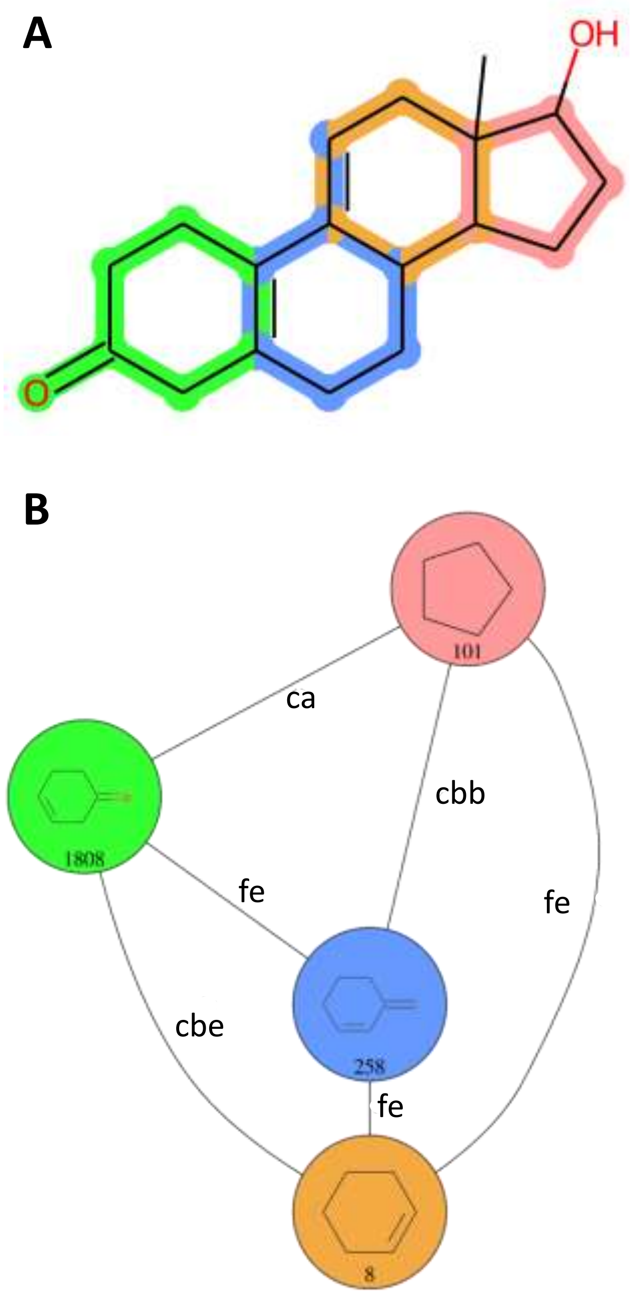


Figure 4. Example of a pseudo-natural product (CHEMBL2311179). A: structure with highlighted NP-derived fragments. Fused atoms and bonds are highlighted with both colors of each fragment. B: Corresponding fragment combination graph, with nodes annotated with fragment structures and IDs, and edges annotated with fragment combination classes.

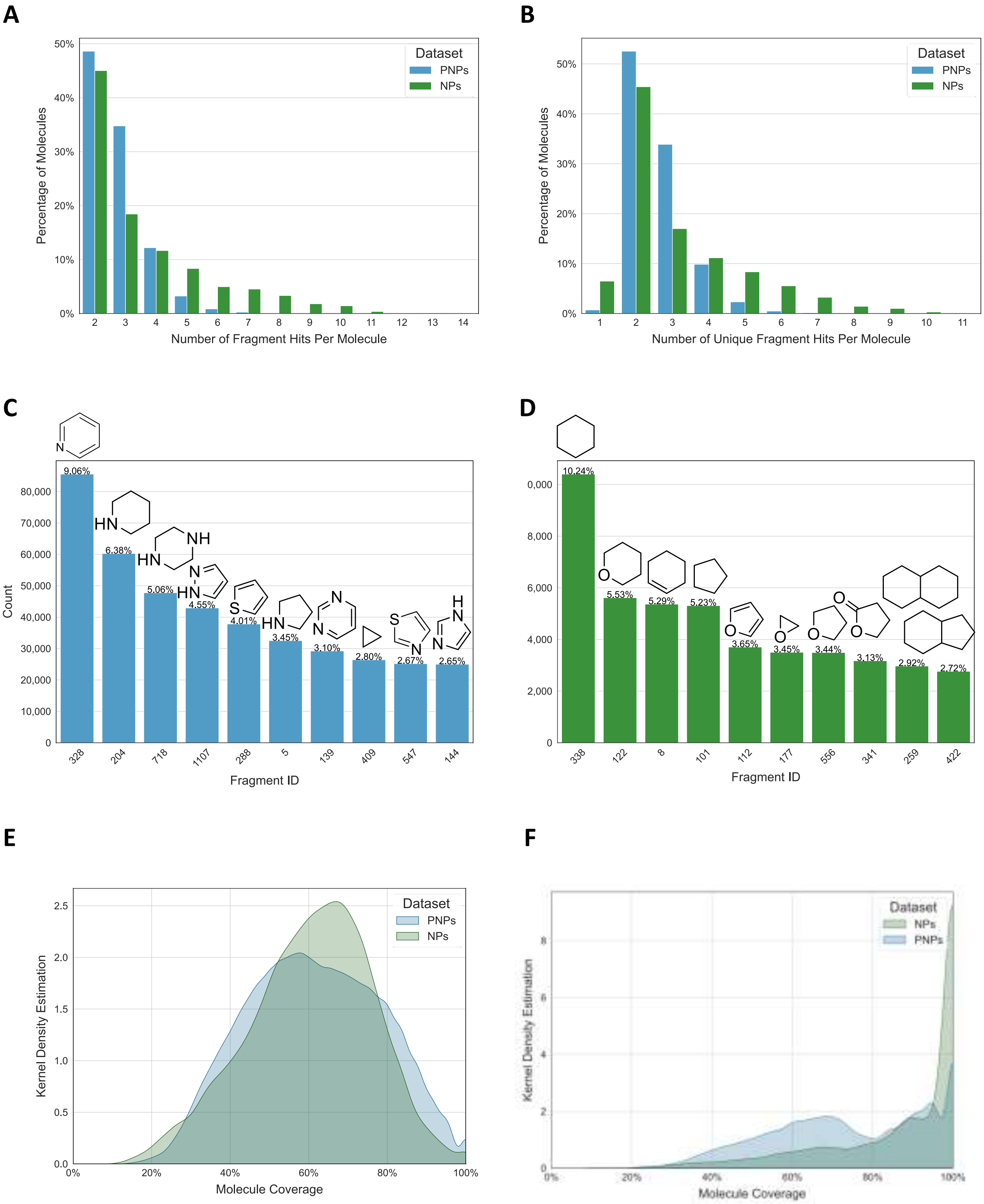


Figure 5. Comparison of PNPs and NPs. A: Number of fragments per molecule; B: Number of unique fragments per molecule (each fragment type is counted only once); C: Ten most occurring fragments in PNPs; D: Ten most frequently occurring fragments in NPs; E: Molecule coverage by fragments per molecule, defined as the ratio of the number of heavy atoms found in fragments to the total number of heavy atoms of the molecule; F: Molecule coverage by fragment per molecule, while considering only rings and linkers in molecules.

Comparison of NPs and PNPs

The connectivity of most PNP NP-fragments could be captured with only 1 FCG per molecule (1.05 in average), whereas more FCGs per molecules were required for NPs (1,25 in average). This observation was consistent with NPs having proportionally three times more **ffo** combinations than synthetic compounds and *a fortiori* PNPs. For analysing results, all FCGs of a molecule were merged into one, while common parts where only considered once.

The large majority of PNPs (95.58%) contains 2-4 NP fragments, whereas this number ranged more widely from 2 to 7 fragments for NPs (99.53%, **Figure 5**A). These values did not vary for PNPs, when considering every fragment type only once per molecule (96.29% had 2-4 fragments, **Figure 5**B). The impact on NPs was much more noticeable, with 97.21% having 1-7 fragments, indicating that NPs had more repeated fragments than PNPs. The number of occurrences of each fragment type was measured for both datasets (1.05 ± 0.18 for PNPs and 1.15 ± 0.18 for NPs) and validated this assumption.

The 10 most abundant fragments observed in PNPs (**Figure 5**C) accounted for 43.28% of all fragment hits for that dataset and were mostly (7 of 10) constituted of nitrogen-containing single rings, with half of them being aromatic. For NPs (**Figure 5**D), the top 10 fragments represented 45.60% of all fragment hits, and consisted mostly (9 of 10) in non-aromatic rings. Half of the most occurring fragments contained only carbon, and the only heteroatom was oxygen.

To assess how much of the molecular topology was captured with our fragment connectivity description, the molecule coverage by fragments was calculated as the percentage of heavy atoms found in fragments compared to the total number of heavy atoms of the molecule (**Figure 5**E). The molecule coverage by fragments was very similar for PNPs and NPs (61% ± 17 and 60% ± 16 respectively in average, with standard deviations as margins). Since NPs frequently have multiple substituents and side chains, we computed the molecule coverage while considering only rings and linkers in molecules (**Figure 5**F). This resulted in an increase of the molecule coverage for PNPs (74% ± 19) and to a stronger extent for NPs (88% ± 17).

The 18 fragment combination categories were identified with different frequencies for the two datasets. For NPs, fragments were found to be involved in 117,956 combinations in total, which were divided into 7 main classes (**Figure 6**), i.e. fusion edge (**fe**, 37.60% of all combinations), connection monopodal (**cm**, 22.35%), connection bipodal edge (**cbe**, 13.63%), fusion bridged (**fb**, 7.88%), connection annulated (7.16%), fusion spiro (**fs**, 4.42%), connection bipodal bridged (**cbb**, 3.17%), and connection bipodal spiro (**cbs**, 2.95%). For PNPs, the fragment combinations (721,589 in total) most frequently were connection monopodal (**cm**, 76.60%) and fusion edge (**fe**, 15.68%). Other significant combinations observed were fusion spiro (**fs**, 2.07%), connection bipodal edge (**cbe**, 1.85%) and fusion bridged (**fb**, 1.81%), indicating that 98.01% of all PNPs analysed here were represented by only five different types of combinations.

Given this high proportion of PNPs identified in the dataset (78.80%), we investigated whether the synthesis and biological analysis of PNPs in the literature was rather a new trend, or whether such compounds have historically been reported in comparable numbers. Structures were annotated with the earliest publication dates available, as indicated in ChEMBL (Figure 7A). Considering our dataset of NP-derived fragments and using the DNP as NP reference, the results clearly indicate that preparation of PNPs has been consistently described over the last 45 years. Remarkably, the percentage of PNPs in published structures per year in ChEMBL increased from 9.81% (± 3.08) in average for 1976-2000 to 18.26% (± 4.04) for 2000-2018 (Figure 7B).

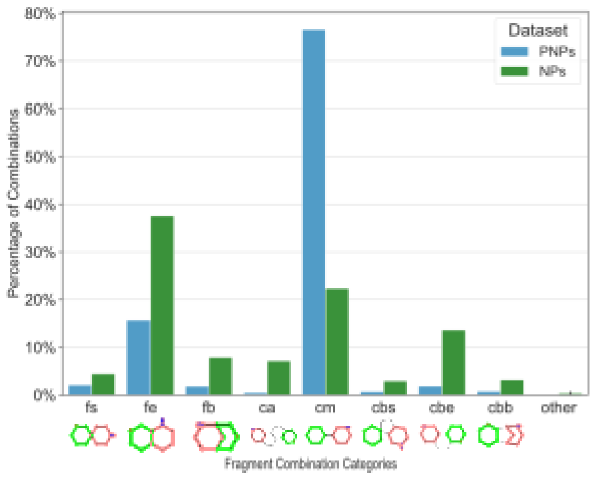


Figure 6. Number of fragment combinations by category. Structures below the bars are simple, manually drawn examples of combinations of two fragments (red and green); fs: fusion spiro, fe: fusion edge, fb: fusion bridged, ca: connection annulated, cm: connection monopodal, cbs: connection bipodal spiro, cbe: connection bipodal edge, cbb: connection bipodal bridged, other: aggregation of all other categories with less than 1% of the total number of combinations.

To investigate whether the same fragments were connected in the same orientation in synthetic compounds and natural products, fragment connection points were considered in addition to the fragment types (nodes) and combination classes (edges), when annotating PNPs. This resulted in a significant increase in the proportion of PNPs, i.e. from 78.80% to 89.49% of the remaining synthetic compounds (Figure 8), suggesting that the orientation of fragments vary significantly in combinations across synthetic compounds and NPs.

PNP Scaffolds

The fragment connectivity of PNPs, represented as FCGs, was used to define molecular scaffolds. The fragment and combination types together with the fragment connection points information allowed to identify 117,184 unique scaffolds. 25.96% of these scaffolds were represented by only one PNP, whereas more than half of them (52.65%) accounted for small collections of 1-5 compounds and 94.76% contained up to 100 members (Figure 9). For comparison, Murcko scaffolds were extracted from the PNPs as well, amounting to 196,321 unique Murcko scaffolds, which are mostly singletons (73.19%). 91.50% of the unique Murcko scaffold types consisted in collections of up to 10 PNP structures, whereas 98.79% contained up to 100 compounds. These data suggest that the PNPs identified in the ChEMBL database consist of a large number of smaller collections of different compound classes, and not of a few large compound libraries.

CONCLUSION

We have developed a computational method for preprocessing molecular structures, classifying fragment combinations and characterizing pseudo-natural products (pseudo-NPs) using a graph approach.

To this end, a set of 2,000 biologically relevant and diverse NP-derived fragments, previously described by Over *et al.*[26], was used to define the fragment space. During the preparation of the fragments, Murcko scaffolds were extracted. This frequently resulted in the generation of the benzene ring as a fragment, which was found to be overrepresented in ChEMBL. The benzene ring was therefore removed from the fragment pool, yielding a total of 1,673 NP fragments.

The Dictionary of Natural Products (DNP) was used to define the chemical space represented by Natural Products (NPs). The 1,673 NP fragments were searched within each of the NP structures after preparation. Fragment combinations were classified using the relative position of fragment pairs within the molecules, resulting in 18 different possible categories, which ultimately might be employed to design synthesis targets and efforts. Fragment connectivity was represented as one, or possibly several, alternate graphs of fragments (nodes) and combinations (edges).

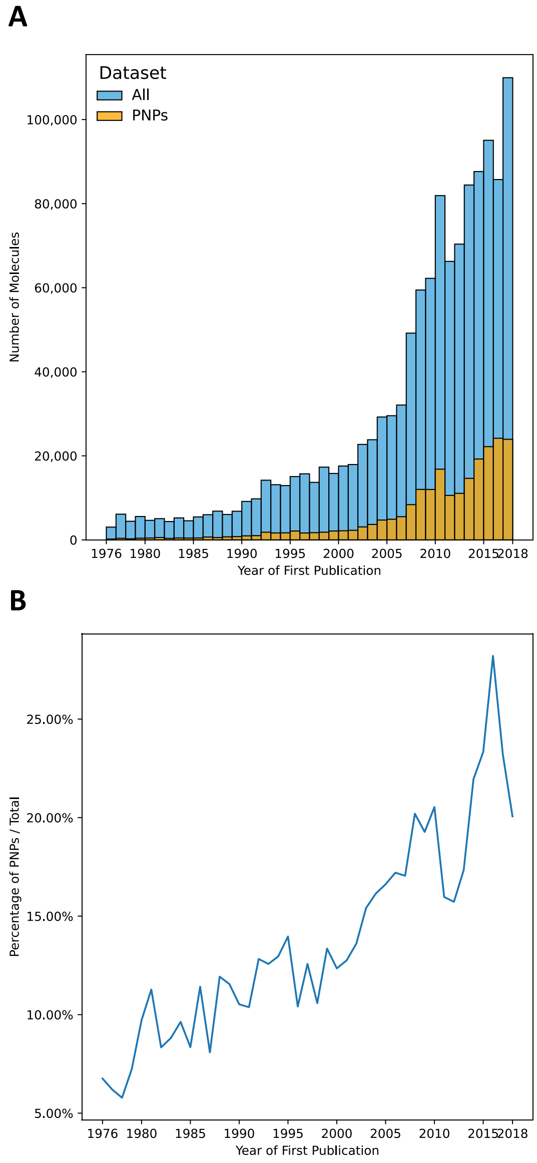


Figure 7. A. Date of first publication of structures in ChEMBL 26. The dataset All refers to all structures in ChEMBL with a document annotation and a valid publication date (1,770,906 records); for PNPs: 224,662 of 344,394 structures (65.23%); B. Percentage of PNPs of total of published structures in ChEMBL 26 per year. No data was available after the year 2018. The data for 1974 consisted only of 2 records and was therefore not considered. For determining the first publication date of compounds, the publication information from duplicate structures filtered out during preparation was considered, when available.

The same protocol was applied to the ChEMBL database, used to represent synthetic compounds, with the additional step of filtering NPs from the dataset using standardized molecule identity (InChI Key) before fragment search. Fragment combination graphs from synthetic compounds were then compared with the graphs obtained previously from the NPs to identify PNPs, i.e. synthetic compounds containing NP-derived fragments in combinations, that could not be found in any NP structure.

A high percentage of the synthetic compounds remaining at the end of the pipeline matched the pseudo-NP criteria (78.80%). These PNPs usually contained 2-4 fragments, involved in 5 main types of combinations (**cm**, **fe**, **fs**, **cbe** and **fb**). The orientation of the fragments within the combinations was investigated as well, while considering symmetry, by adding another criterion to the comparison of fragment combination graphs. Results indicated that the orientation of fragments varied in synthetic compounds compared to NPs, which led to an increase of the rate of PNPs to 88.98% of synthetic compounds remaining at the end of the pipeline.

These results demonstrate that PNPs as defined by us have, in fact, been synthetized for at least 45 years, and it can be concluded that they occur frequently among biologically relevant small molecules. The frequency of their occurrence in biologically relevant compounds is testimony of their biological relevance and validates the PNP design principle as a successful and actually historically proven general concept for the discovery of new bioactive chemical matter and the exploration of biologically relevant chemical space.

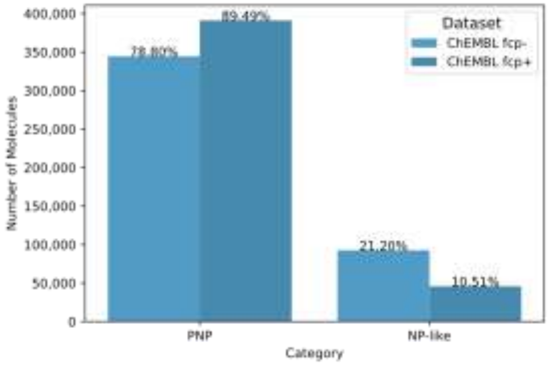


Figure 8. Number of pseudo-natural products (PNPs) in ChEMBL, with or without considering fragment combination points during PNP annotation (fcp -/+).

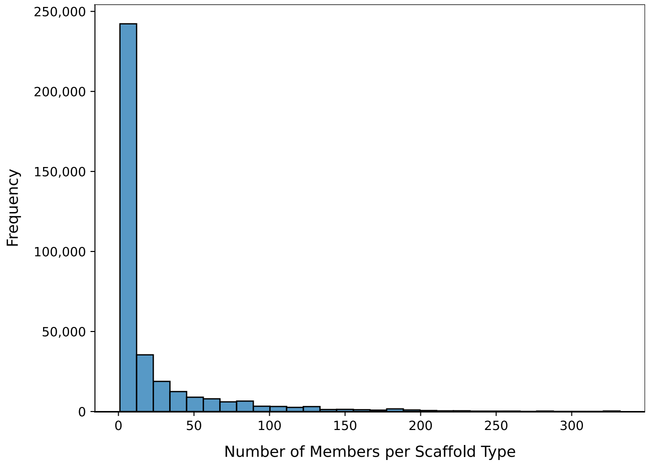


Figure 9. Distribution of the number of members per scaffold type in PNPs.

ASSOCIATED CONTENT

Supporting Information

The structures of the set of 1,673 NP-derived fragments obtained after preparation in SDF format as well as a molecular grid in PDF, with molecules annotated with fragment connection labels. A summary of the results obtained using benzene as a NP-derived fragment.

The Supporting Information is available free of charge on the ACS Publications website.

DATA AND SOFTWARE AVAILABILITY

The npfc package is freely available at https://github.com/mpimp-comas/npfc. The installation guidelines on the repository page describe the installation process with all required dependencies. The initial NP-derived fragments can be downloaded from the Supporting Information of reference 26. The ChEMBL 26 dataset is accessible using the link from reference 32. A commercial agreement was necessary for us to use the Dictionary of Natural Products.

AUTHOR INFORMATION

Corresponding Author

Herbert Waldmann.
Email: herbert.waldmann@mpi-dortmund.mpg.de

Present addresses

[†]Department of Chemical Biology, Max-Planck-Institute of Molecular Physiology, Otto-Hahn-Straße 11, 44227 Dortmund Germany

[‡]Compound Management and Screening Center, Dortmund, Otto-Hahn-Str. 11, 44227 Dortmund, Germany

[§]Faculty of Chemistry and Chemical Biology, Technical University Dortmund, Otto-Hahn-Straße 6, 44227 Dortmund, Germany

Author Contributions

H.W., P.C., A.P. and J.M.G. conceived and designed this project. The conception and the development of the chemoinformatics tool as well as its applications were performed by J.M.G., with monitoring from A.P. The results were analyzed and discussed by all authors. All authors contributed to the manuscript.

Notes

The authors declare no competing financial interests.

ACKNOWLEDGMENT

The authors would like to thank current and former members of the Waldmann group at the MPI Dortmund and COMAS, in particular Dr. Claude Ostermann and Dr. Malte Gersch for their helpful comments on this study.

REFERENCES

(1) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83* (3), 770–803. https://doi.org/10.1021/acs.jnatprod.9b01285.

(2) Cottens, S.; Kallen, J.; Schuler, W.; Sedrani, R. Derivation of Rapamycin: Adventures in Natural Product Chemistry. *Chimia* **2019**, *73* (7), 581–590. https://doi.org/10.2533/chimia.2019.581.

(3) Karageorgis, G.; Foley, D. J.; Laraia, L.; Waldmann, H. Principle and Design of Pseudo-Natural Products. *Nat. Chem.* **2020**, *12* (3), 227–235. https://doi.org/10.1038/s41557-019-0411-x.

(4) Cremosnik, G. S.; Liu, J.; Waldmann, H. Guided by Evolution: From Biology Oriented Synthesis to Pseudo Natural Products. *Nat. Prod. Rep.* **2020**, *37* (11), 1497–1510. https://doi.org/10.1039/d0np00015a.

(5) Karageorgis, G.; Foley, D. J.; Laraia, L.; Brakmann, S.; Waldmann, H. Pseudo Natural Products-Chemical Evolution of Natural Product Structure. *Angew. Chem. Int. Ed Engl.* **2021**. https://doi.org/10.1002/anie.202016575.

(6) Erlanson, D. A.; Fesik, S. W.; Hubbard, R. E.; Jahnke, W.; Jhoti, H. Twenty Years on: The Impact of Fragments on Drug Discovery. *Nat. Rev. Drug Discov.* **2016**, *15* (9), 605–619. https://doi.org/10.1038/nrd.2016.109.

(7) Ceballos, J.; Schwalfenberg, M.; Karageorgis, G.; Reckzeh, E. S.; Sievers, S.; Ostermann, C.; Pahl, A.; Sellstedt, M.; Nowacki, J.; Carnero Corrales, M. A.; Wilke, J.; Laraia, L.; Tschapalda, K.; Metz, M.; Sehr, D. A.; Brand, S.; Winklhofer, K.; Janning, P.; Ziegler, S.; Waldmann, H. Synthesis of Indomorphan Pseudo-Natural Product Inhibitors of Glucose Transporters GLUT-1 and -3. *Angew. Chem. Int. Ed Engl.* **2019**, *58* (47), 17016–17025. https://doi.org/10.1002/anie.201909518.

(8) Karageorgis, G.; Reckzeh, E. S.; Ceballos, J.; Schwalfenberg, M.; Sievers, S.; Ostermann, C.; Pahl, A.; Ziegler, S.; Waldmann, H. Chromopynones Are Pseudo Natural Product Glucose Uptake Inhibitors Targeting Glucose Transporters GLUT-1 and -3. *Nat. Chem.* **2018**, *10* (11), 1103–1111. https://doi.org/10.1038/s41557-018-0132-6.

(9) Schneidewind, T.; Kapoor, S.; Garivet, G.; Karageorgis, G.; Narayan, R.; Vendrell-Navarro, G.; Antonchick, A. P.; Ziegler, S.; Waldmann, H. The Pseudo Natural Product Myokinasib Is a Myosin Light Chain Kinase 1 Inhibitor with Unprecedented Chemotype. *Cell Chem. Biol.* **2019**, *26* (4), 512-523.e5. https://doi.org/10.1016/j.chembiol.2018.11.014.

(10) Foley, D. J.; Zinken, S.; Corkery, D.; Laraia, L.; Pahl, A.; Wu, Y.-W.; Waldmann, H. Phenotyping Reveals Targets of a Pseudo-Natural-Product Autophagy Inhibitor. *Angew. Chem. Int. Ed.* **2020**, *59* (30), 12470–12476. https://doi.org/10.1002/anie.202000364.

(11) Christoforow, A.; Waldmann, H.; Wilke, J.; Binici, A.; Pahl, A.; Ostermann, C.; Sievers, S. Design, Synthesis and Phenotypic Profiling of Pyrano-Furo-Pyridone Pseudo Natural Products. *Angew. Chem. Int. Ed Engl.* **2019**. https://doi.org/10.1002/anie.201907853.

(12) Wu, G.; Yu, G.; Yu, Y.; Yang, S.; Duan, Z.; Wang, W.; Liu, Y.; Yu, R.; Li, J.; Zhu, T.; Gu, Q.; Li, D. Chemoreactive-Inspired Discovery of Influenza A Virus Dual Inhibitor to Block Hemagglutinin-Mediated Adsorption and Membrane Fusion. *J. Med. Chem.* **2020**, *63* (13), 6924–6940. https://doi.org/10.1021/acs.jmedchem.0c00312.

(13) Yuan, S.; Yue, Y.-L.; Zhang, D.-Q.; Zhang, J.-Y.; Yu, B.; Liu, H.-M. Synthesis of New Tetracyclic Benzodiazepine-Fused Isoindolinones Using Recyclable Mesoporous Silica Nanoparticles. *Chem. Commun. Camb. Engl.* **2020**, *56* (77), 11461–11464. https://doi.org/10.1039/d0cc04875e.

(14) Hert, J.; Irwin, J. J.; Laggner, C.; Keiser, M. J.; Shoichet, B. K. Quantifying Biogenic Bias in Screening Libraries. *Nat. Chem. Biol.* **2009**, *5* (7), 479–483. https://doi.org/10.1038/nchembio.180.

(15) Nören-Müller, A.; Wilk, W.; Saxena, K.; Schwalbe, H.; Kaiser, M.; Waldmann, H. Discovery of a New Class of Inhibitors of Mycobacterium Tuberculosis Protein Tyrosine Phosphatase B by Biology-Oriented Synthesis. *Angew. Chem. Int. Ed Engl.* **2008**, *47* (32), 5973–5977. https://doi.org/10.1002/anie.200801566.

(16) Carné-Carnavalet, B. de; Krieger, J.-P.; Folléas, B.; Brayer, J.-L.; Demoute, J.-P.; Meyer, C.; Cossy, J. Diastereodivergent Pictet–Spengler Cyclization of Bicyclic N-Acyliminium Ions: Controlling a Quaternary Stereocenter. *Eur. J. Org. Chem.* **2015**, *2015* (6), 1273–1282. https://doi.org/https://doi.org/10.1002/ejoc.201403469.

(17) Bartlett, M. J.; Turner, C. A.; Harvey, J. E. Pd-Catalyzed Allylic Alkylation Cascade with Dihydropyrans: Regioselective Synthesis of Furo[3,2-c]Pyrans. *Org. Lett.* **2013**, *15* (10), 2430–2433. https://doi.org/10.1021/ol400902d.

(18) Cunha, M. R.; Bhardwaj, R.; Carrel, A. L.; Lindinger, S.; Romanin, C.; Parise-Filho, R.; Hediger, M. A.; Reymond, J.-L. Natural Product Inspired Optimization of a Selective TRPV6 Calcium Channel Inhibitor. *RSC Med. Chem.* **2020**, *11* (9), 1032–1040. https://doi.org/10.1039/d0md00145g.

(19) Reilly, S. W.; Griffin, S.; Taylor, M.; Sahlholm, K.; Weng, C.-C.; Xu, K.; Jacome, D. A.; Luedtke, R. R.; Mach, R. H. Highly Selective Dopamine D3 Receptor Antagonists with Arylated Diazaspiro Alkane Cores. *J. Med. Chem.* **2017**, *60* (23), 9905–9910. https://doi.org/10.1021/acs.jmedchem.7b01248.

(20) Alkhaibari, I.; Raj Kc, H.; Alnufaie, R.; Gilmore, D. F.; Alam, M. A. Synthesis of Chimeric Thiazolo-Nootkatone Derivatives as Potent

Antimicrobial Agents. *ChemMedChem* **2021**. https://doi.org/10.1002/cmdc.202100230.

(21) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45* (Database issue), D945–D954. https://doi.org/10.1093/nar/gkw1074.

(22) RDKit: Open-source cheminformatics http://www.rdkit.org/ (accessed 2021 -03 -15).

(23) McKinney, W. Data Structures for Statistical Computing in Python; Austin, Texas, 2010; pp 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

(24) Hagberg, A.; Swart, P.; S Chult, D. *Exploring Network Structure, Dynamics, and Function Using Networkx*; LA-UR-08-05495; LA-UR-08-5495; Los Alamos National Lab. (LANL), Los Alamos, NM (United States), 2008.

(25) Mölder, F.; Jablonski, K. P.; Letcher, B.; Hall, M. B.; Tomkins-Tinch, C. H.; Sochat, V.; Forster, J.; Lee, S.; Twardziok, S. O.; Kanitz, A.; Wilm, A.; Holtgrewe, M.; Rahmann, S.; Nahnsen, S.; Köster, J. Sustainable Data Analysis with Snakemake. *F1000Research* **2021**, *10*, 33. https://doi.org/10.12688/f1000research.29032.1.

(26) Over, B.; Wetzel, S.; Grütter, C.; Nakai, Y.; Renner, S.; Rauh, D.; Waldmann, H. Natural-Product-Derived Fragments for Fragment-Based Ligand Discovery. *Nat. Chem.* **2013**, *5* (1), 21–28. https://doi.org/10.1038/nchem.1506.

(27) Standardization — MolVS 0.1.1 documentation https://molvs.readthedocs.io/en/latest/guide/standardize.html (accessed 2020 -01 -21).

(28) pdbeccdutils · master · pdbe / ccdutils https://gitlab.ebi.ac.uk/pdbe/ccdutils/-/tree/master/pdbeccdutils (accessed 2021 -03 -11).

(29) RDKit mailing list - symmetry class https://sourceforge.net/p/rdkit/mailman/message/27897393/ (accessed 2021 -03 -11).

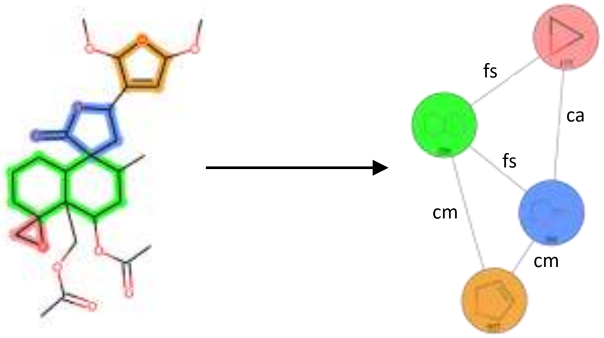
(30) Dictionary of Natural Products http://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml (accessed 2021 -02 -08).

(31) Schaub, J.; Zielesny, A.; Steinbeck, C.; Sorokina, M. Too Sweet: Cheminformatics for Deglycosylation in Natural Products. *J. Cheminformatics* **2020**, *12* (1), 67. https://doi.org/10.1186/s13321-020-00467-y.

(32) Index of /pub/databases/chembl/ChEMBLdb/releases/chembl_26/ http://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_26/ (accessed 2021 -02 -10).

Authors are required to submit a graphic entry for the Table of Contents (TOC) that, in conjunction with the manuscript title, should give the reader a representative idea of one of the following: A key structure, reaction, equation, concept, or theorem, etc., that is discussed in the manuscript. Consult the journal's Instructions for Authors for TOC graphic specifications.

Insert Table of Contents artwork here



Pseudo-Natural Products Occur Frequently in Biologically Relevant Compounds

José-Manuel Gally,† Axel Pahl, ‡ Paul Czodrowski, § Herbert Waldmann*,†,§

†Department of Chemical Biology, Max-Planck-Institute of Molecular Physiology, Otto-Hahn-Straße 11, 44227 Dortmund, Germany,

‡Compound Management and Screening Center, Dortmund, Otto-Hahn-Str. 11, 44227 Dortmund, Germany

§Faculty of Chemistry and Chemical Biology, Technical University Dortmund, Otto-Hahn-Straße 6, 44227 Dortmund, Germany

Supporting Information

TABLE OF CONTENTS

I. List of files included in SI 15

II. Manual curation of structural errors in the NP fragments 15

III. Preparation of the datasets 16

IV. Impact of the benzene fragment on the results 26

V. Computational time 27

TABLE OF FIGURES

Figure S1 Manual curation of the five NP-derived fragments failing the RDKit conversion. 15
Figure S2. Results of the preparation of the NP-derived fragments dataset. 18
Figure S3. Categories of filtered entries in the NP-derived fragments dataset. 19
Figure S4. Results of the preparation of the DNP dataset. 20
Figure S5. Categories of errors raised by entries in the DNP dataset. 22
Figure S6. Categories of filtered entries in the DNP dataset. 21
Figure S7. Results of the preparation of the ChEMBL dataset. 23
Figure S8. Categories of errors raised by entries in the ChEMBL dataset. 25
Figure S9. Categories of filtered entries in the ChEMBL dataset. 24
Figure S10. Results for the NPFC workflows including benzene in the NP-fragment. 26

TABLE OF TABLES

Table S1. Possible error cases in NPFC. 16
Table S2. Filters implemented in NPFC. 17
Table S3. Number of molecules and entries per step including benzene 26

I. List of files included in the Supporting Information

- This document
- The prepared 1,673 NP-derived fragments in SDF format, generated with RDKit
- The prepared 1,673 NP-derived fragments in PDF format, with FCP annotations
- The structures from the manuscript in SMILES format

The NP-derived fragment SDF contains different properties:

- idm: molecule identifier (inherited from the "Cluster" property of the original SDF from Over
- inchikey: the molecule InChI Key, computed with RDKit
- _fcp_labels: the fragment combination point labels enabling the unambiguous orientation of the fragment when symmetry centres are present.
- Num_symmetry_groups: the number of different groups of atoms regrouped within the same symmetry group (i.e. 1a, 1b, etc.)

The npfc package is available at https://github.com/mpimp-comas here additional documentation is available on the package itself (API and overall architecture). The "_fcp_labels" property in the provided SDF is pickled and encoded as base64 strings. Please refer to the npfc.utils.decode_object to convert them back into Python dictionaries.

II. Manual curation of structural errors in the NP fragments

Five structures could not be parsed with RDKit directly using the input SDF. Rather than just discarding them, the structures were fixed using MarvinSketch 19.2.0 (Figure S1).

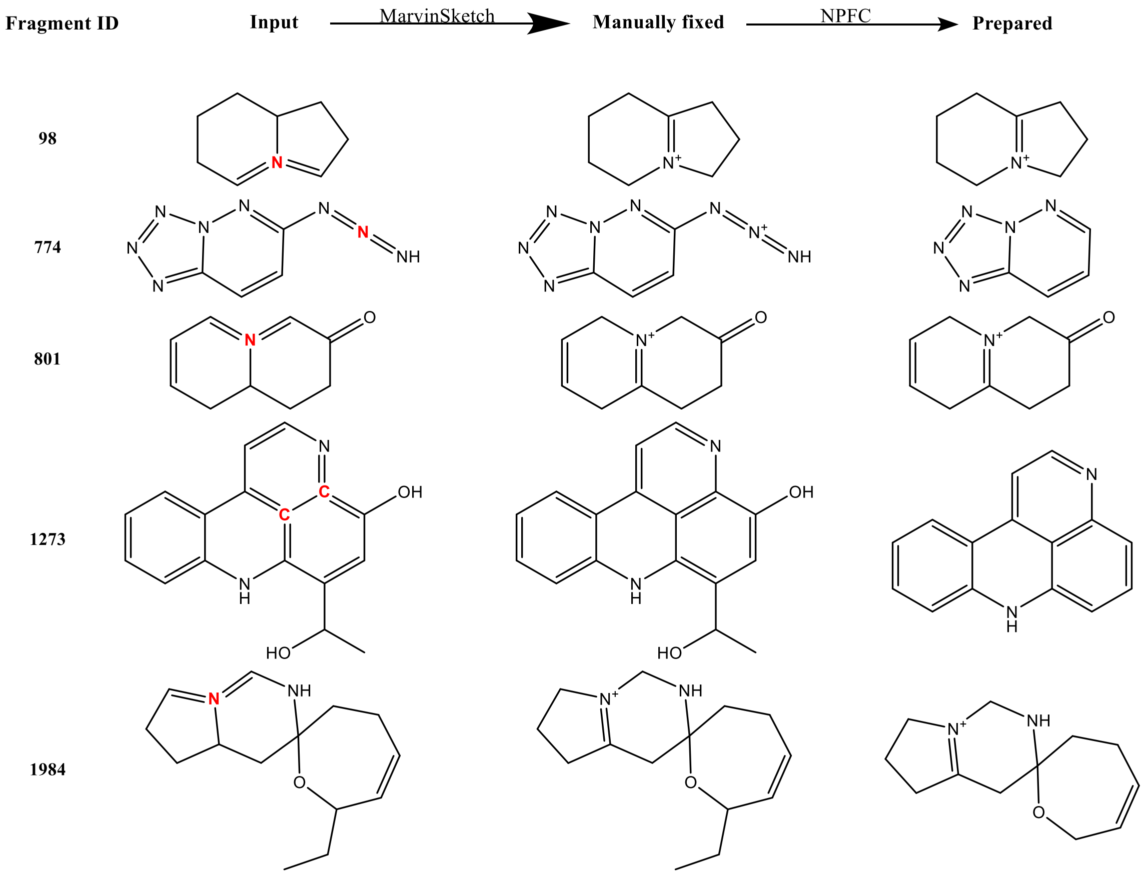


Figure S1 Manual curation of the five NP-derived fragments failing the RDKit conversion. Atoms with valence errors are highlighted in red in the input structures.

III. Preparation of the datasets

The preparation of the datasets consisted of 5 steps:

1. Split the data into chunks (optional)
2. Load the SD File(s) into Pandas DataFrame(s) with RDKit molecules
3. Standardize the structures
4. Removing duplicate entries
5. Depict molecules

During these steps, molecules might fail the process and be discarded (Table S1).

Table S1. Possible error cases in NPFC.

Step	Label	Description
load	cannot_load	The molecule could not be parsed into an RDKit Mol object.
standardize	initiate_mol	The RDKit Mol object could not be initiated from its pickled state.
	disconnect_metal	An error occurred when when disconnecting metal atoms from the RDKit Mol object.
	sanitize	An error occurrend when updating the RDKit Mol object internal properties.
	clear_isotopes	An error occurred when setting the RDKit Mol object atoms to their default isotope.
	normalize	An error occurred when normalizing functional groups of the RDKit Mol object.
	uncharge	An error occurred when removing formal charges from the RDKit Mol object atoms.
	canonicalize	An error occurred when computing the canonical tautomer for the RDKit Mol object.
	clear_stereo	An error occurred when removing the stereochemistry labels from the RDKit Mol object.
	empty_final	The RDKit Mol object was empty at the end of the standardization process.

Similarly, molecules could be filtered out if they did not respect certain criteria (Table S2).

Table S2. Filters implemented in NPFC.

Step	Label	Description
standardize	empty	The RDKit Mol object did not contain any atom.
	num_heavy_atoms	The number of heavy atoms of the RDKit Mol object was outside the expected range. *num_heavy_atoms > 3*
	molecular_weight	The molecular weight (Da) of the RDKit Mol object was outside the expected range. *molecular_weight <= 1000.0*
	num_rings	The number of rings of the RDKit Mol object was outside the expected range. *num_rings > 0*
	elements	The RDKit Mol object contained atoms outside of the expected elements. *elements in H, B, C, N, O, F, P, S, Cl, Br, I.*
	timeout	The standardization process of the RDKit Mol object lasted longer than the timeout limit. *timeout = 10s (per molecule)*
deduplicate	duplicate	The molecule was found to be a duplicate of another, already registered entry. Identity check is performed via InChI Key comparison.

Default values are displayed in italic.

For fragments, only the filters on empty and duplicate structures were enabled, as well as the timeout limit (mandatory).

NP-derived fragments

The 2,000 fragments from Over *et al.* were prepared after five of them were manually curated (see section II). Results are represented by Figure S2 (overall) and Figure S3 (filtered entries).

Natural Products (DNP)

The results of the preparation of the 318,271 records of the DNP dataset are represented by Figure S4 (overall), Figure S5 (filtered entries) and Figure S6 (errors).

Synthetic Compounds (ChEMBL)

The preparation of the 1,941,411 records of the ChEMBL26 dataset are represented by Figure S7 (overall), Figure S8 (filtered entries) and Figure S9 (errors).

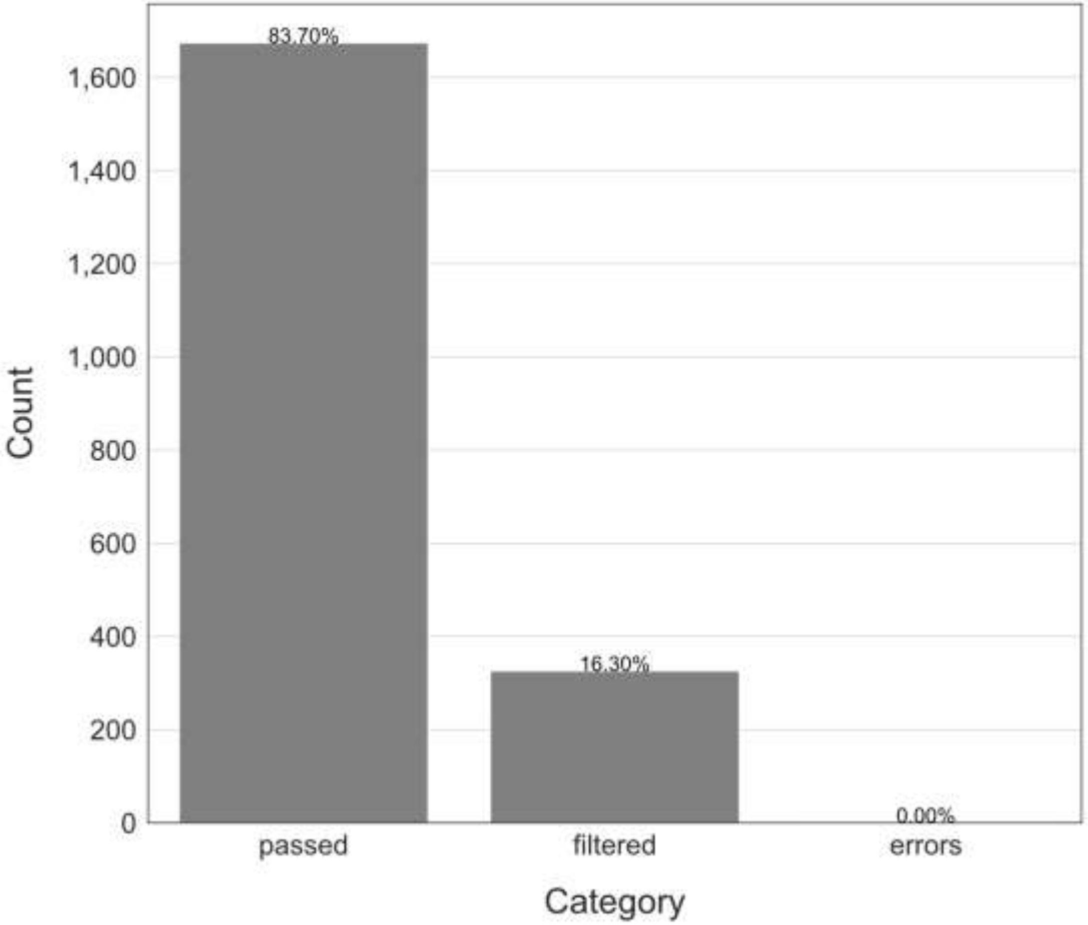


Figure S2. Results of the preparation of the NP-derived fragments dataset. Percentages are relative to the initial number of entries (2,000)

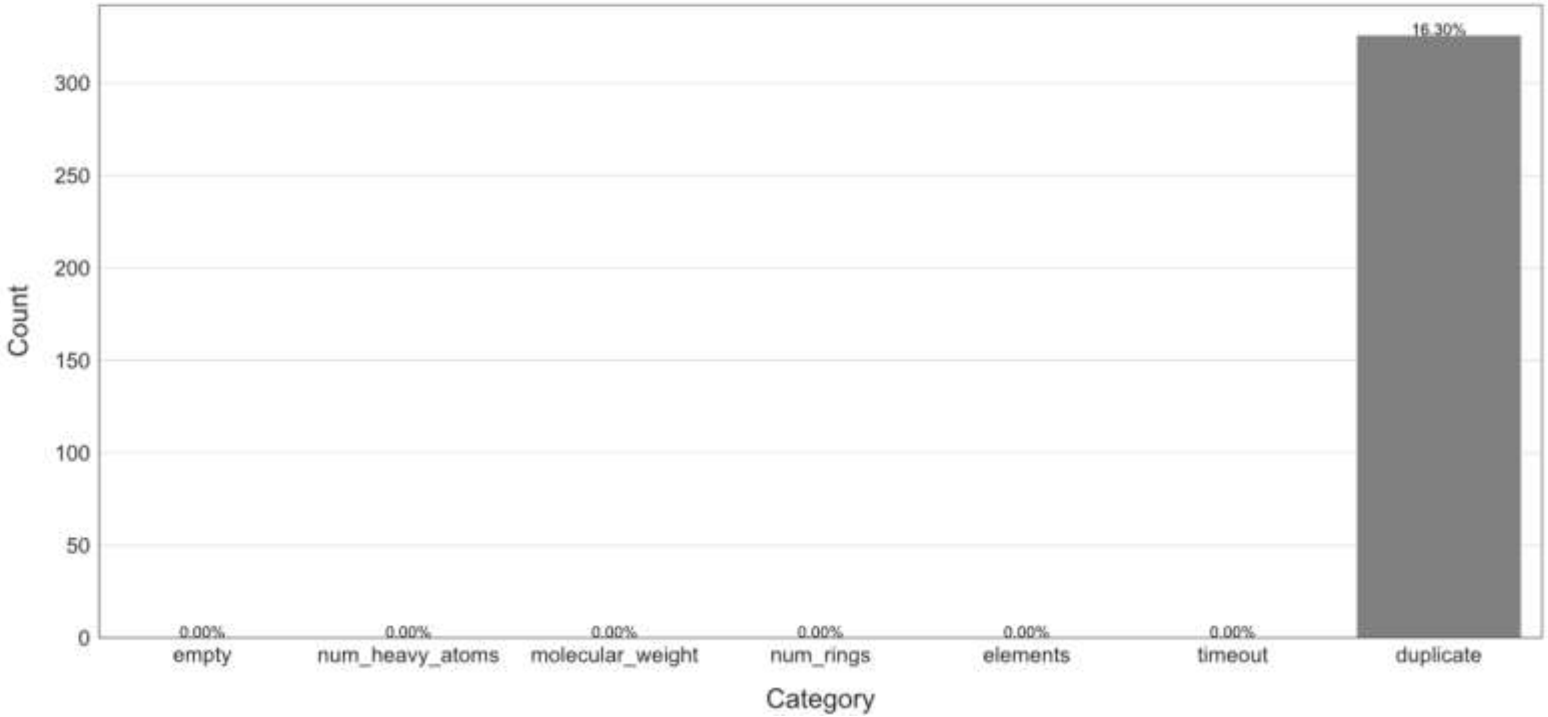


Figure S3. Categories of filtered entries in the NP-derived fragments dataset. Percentages are relative to the initial number of entries (2,000)

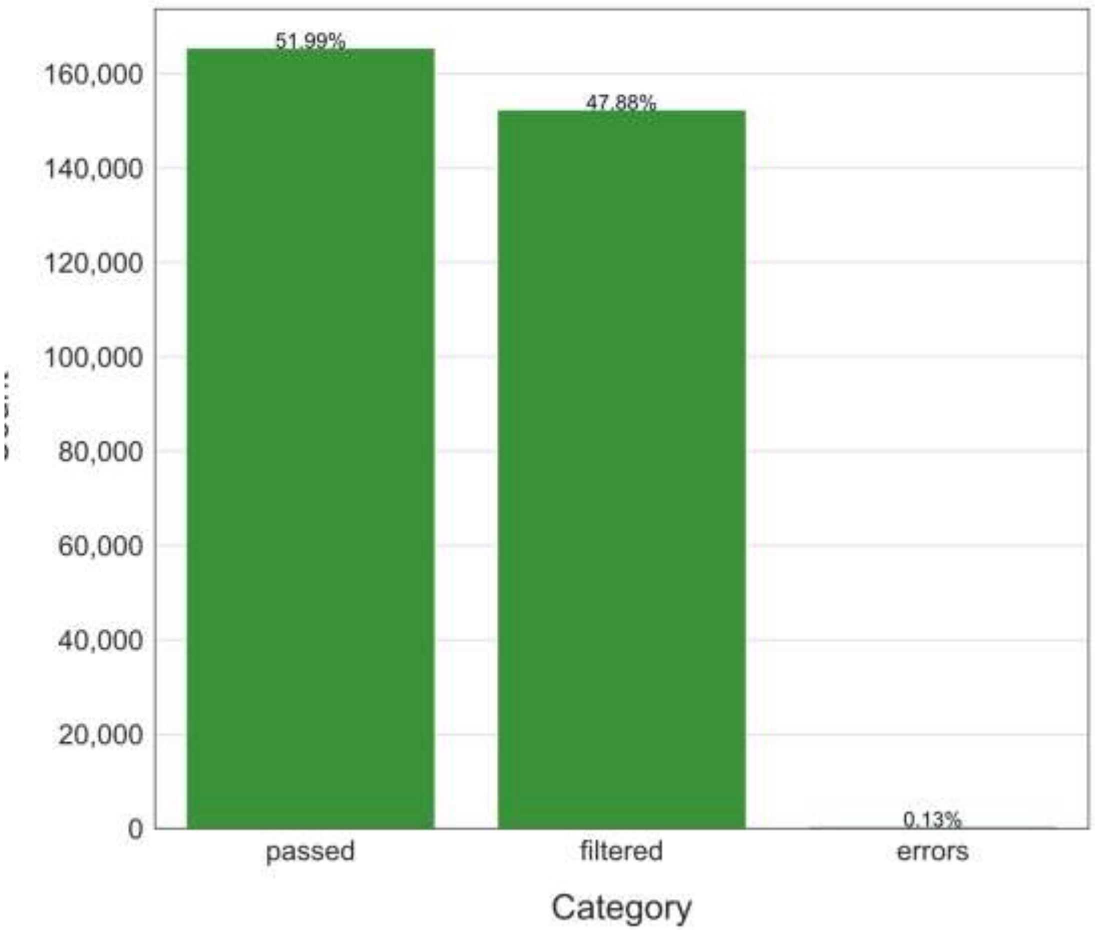


Figure S4. Results of the preparation of the DNP dataset. Percentages are relative to the initial number of entries (318,271)

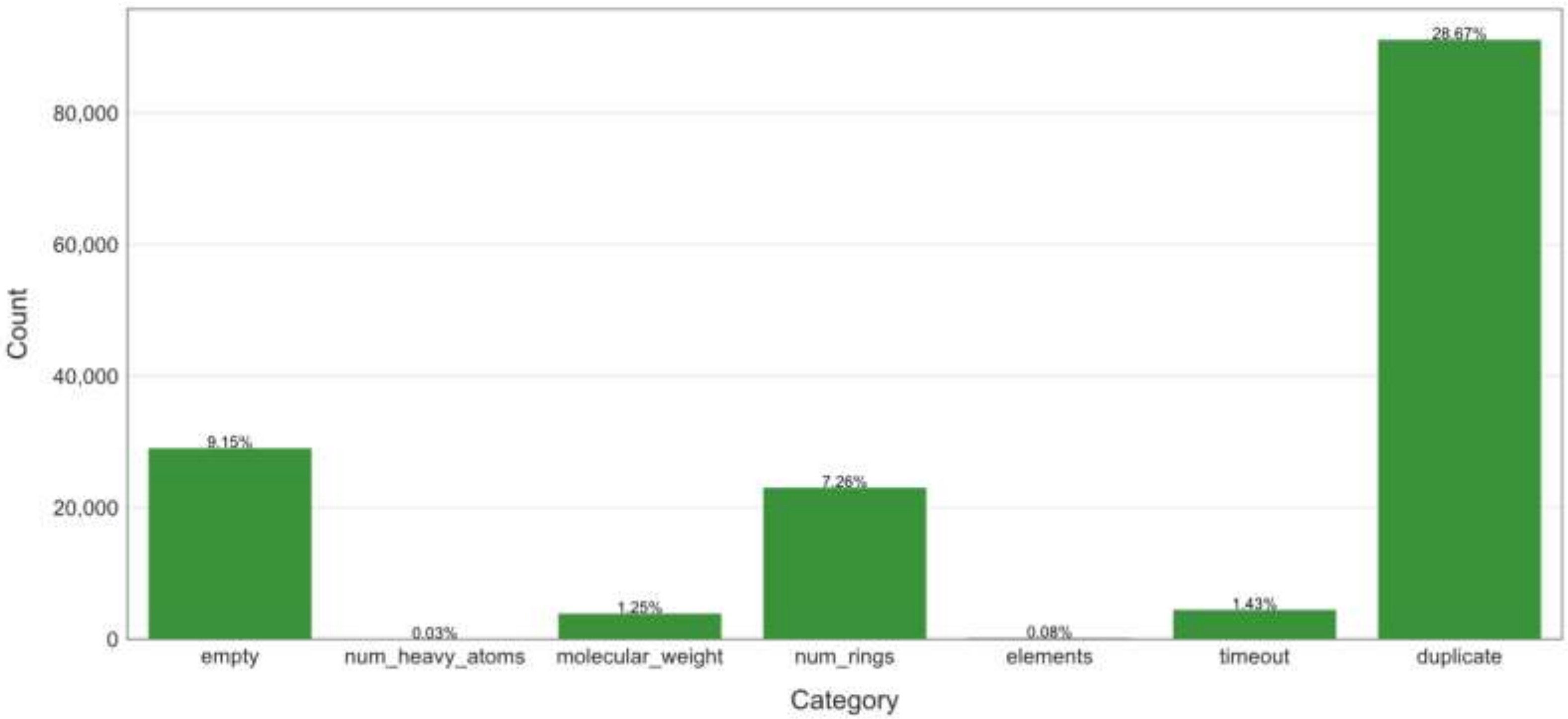


Figure S5. Categories of filtered entries in the DNP dataset. Percentages are relative to the initial number of entries (318,271)

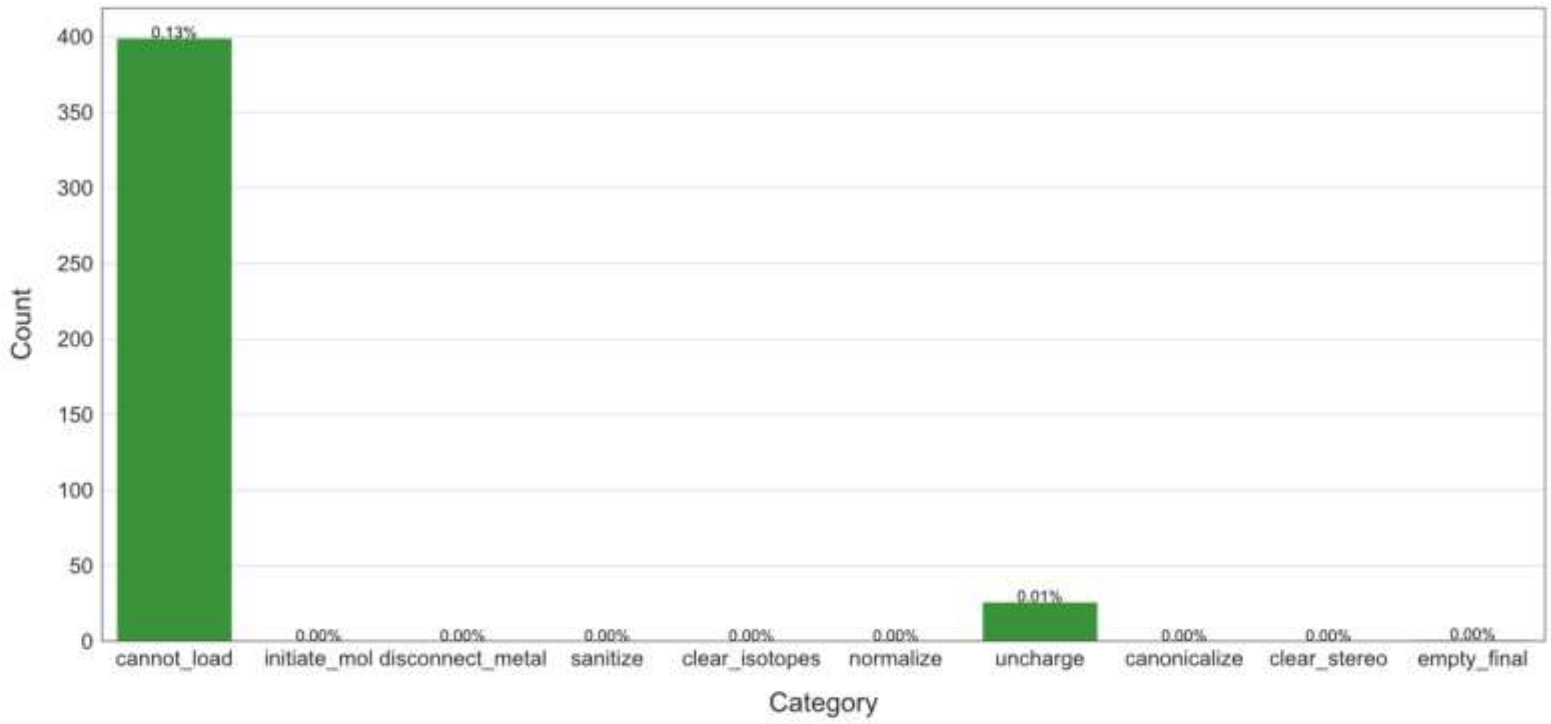


Figure S6. Categories of errors raised by entries in the DNP dataset. Percentages are relative to the initial number of entries (318,271)

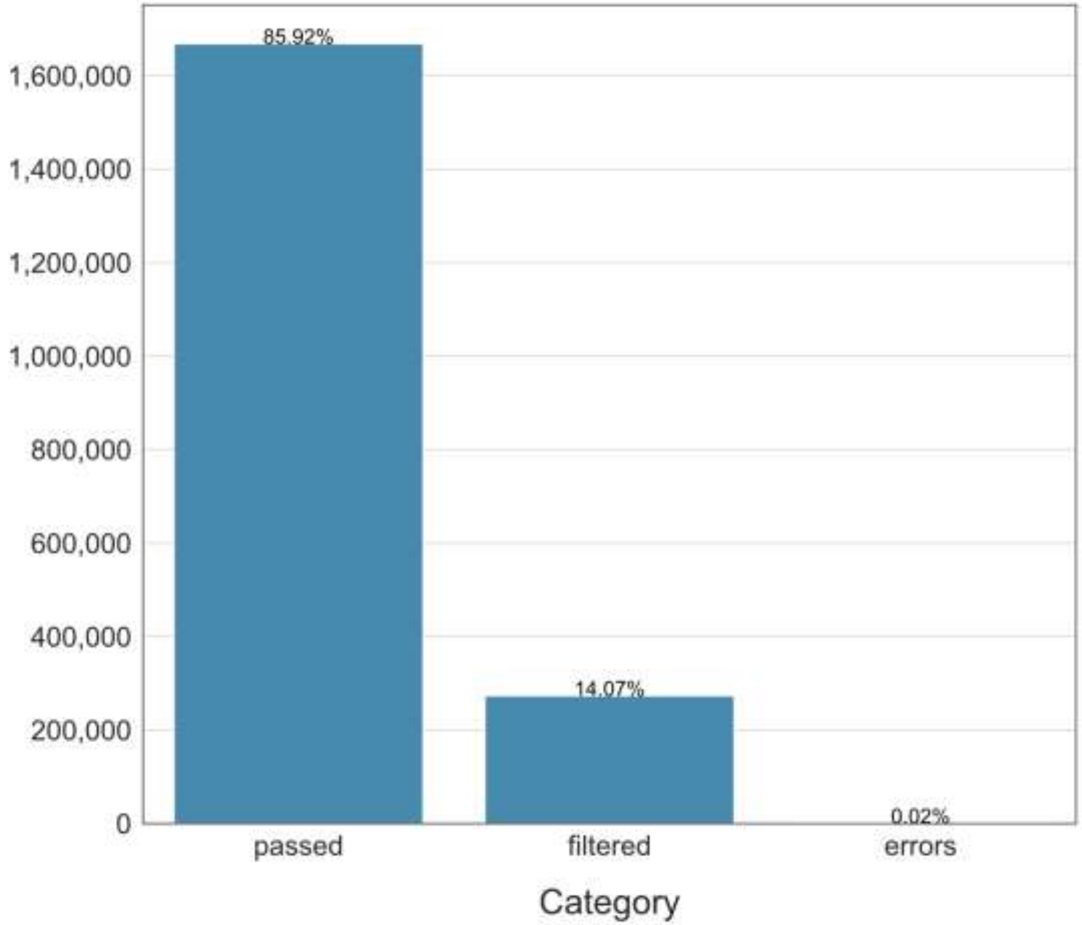


Figure S7. Results of the preparation of the ChEMBL dataset. Percentages are relative to the initial number of entries (1,941,411)

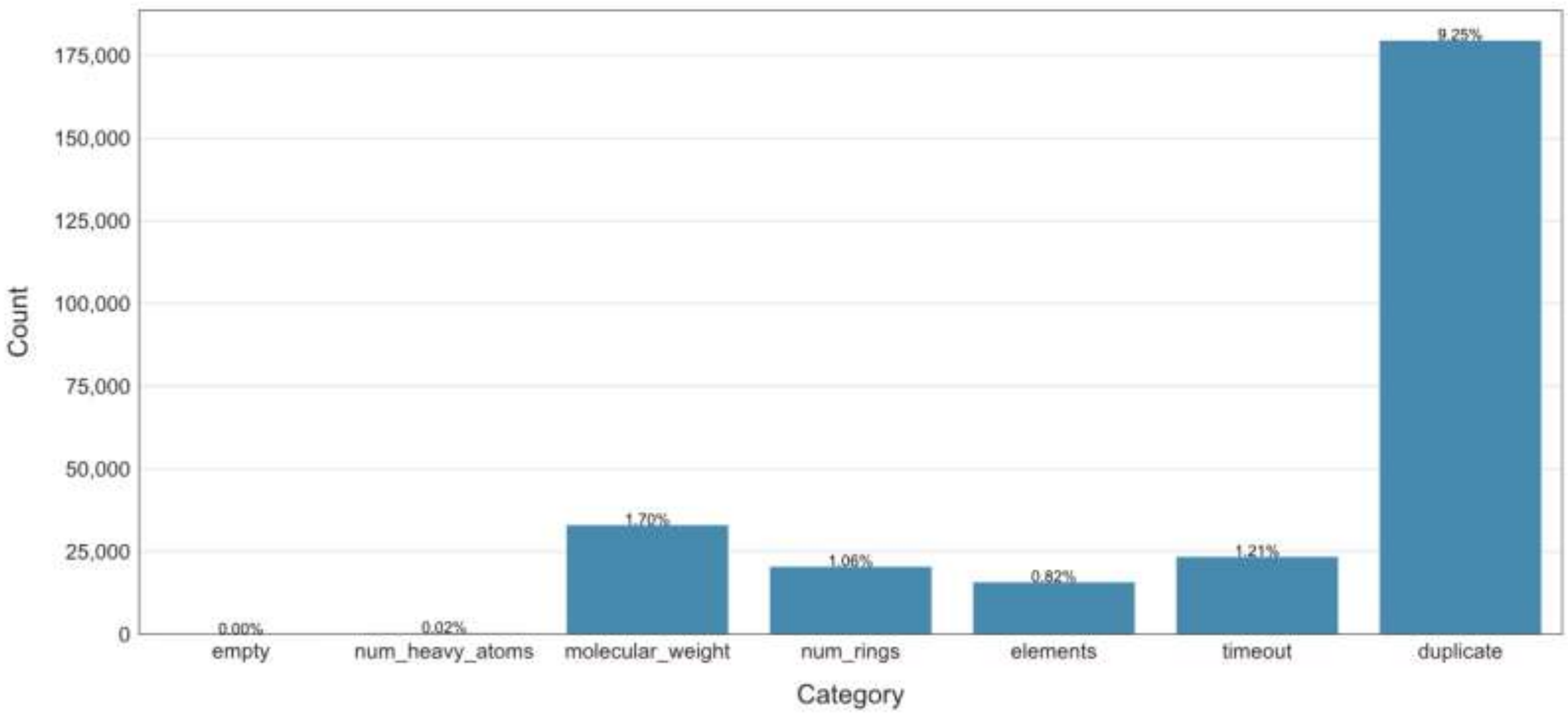


Figure S8. Categories of filtered entries in the ChEMBL dataset. Percentages are relative to the initial number of entries (1,941,411)

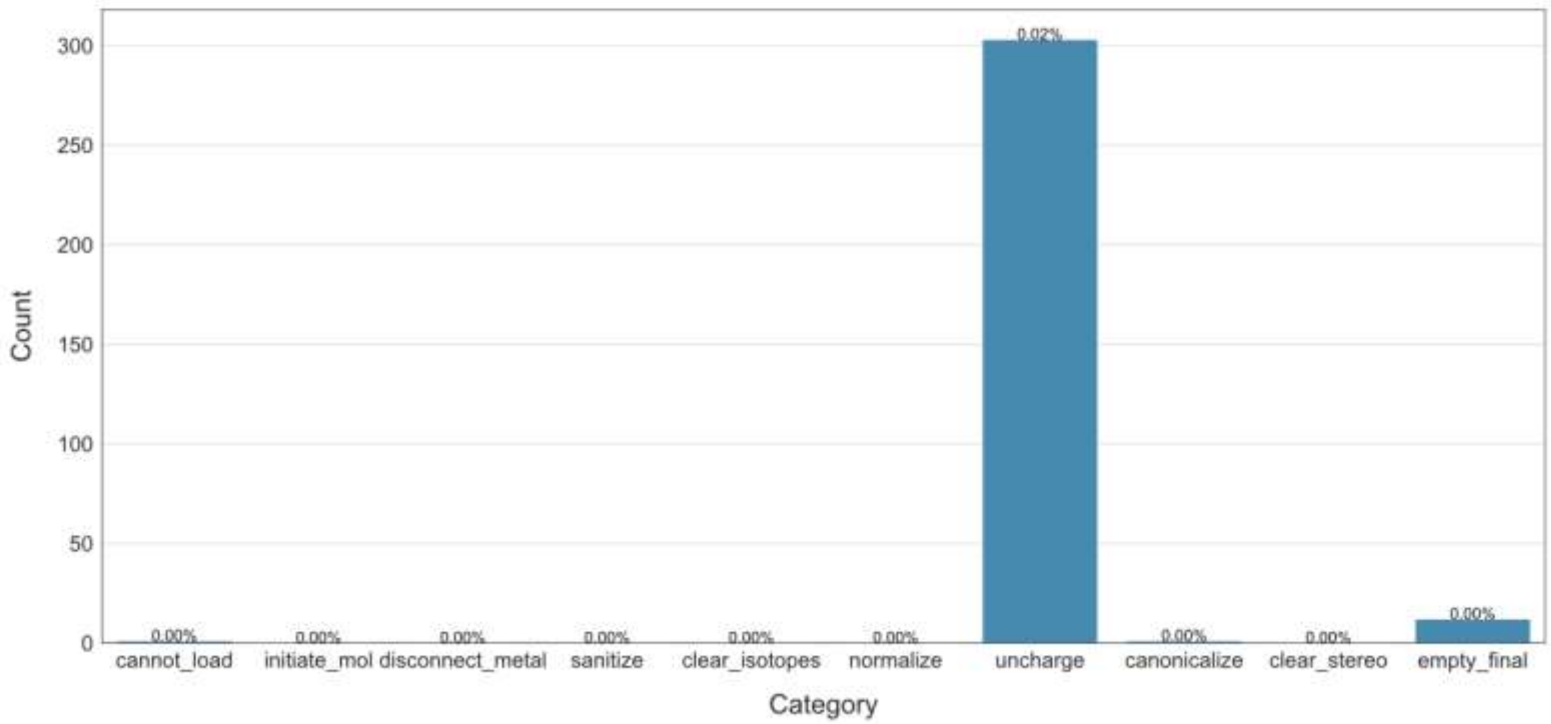


Figure S9. Categories of errors raised by entries in the ChEMBL dataset. Percentages are relative to the initial number of entries (1,941,411)

IV. Impact of the benzene fragment on the results

The data below describes the results obtained during our first attempt of the NPFC approach, while including the benzene as a Natural Product (NP) fragment. The datasets, software versions and computational resources were the same as used for the main manuscript.

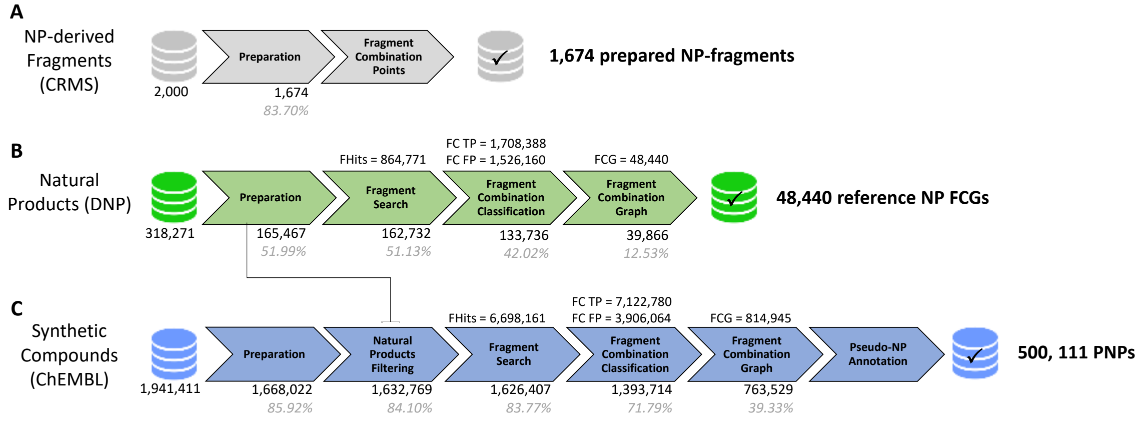


Figure S10. Results for the NPFC workflows including benzene in the NP-fragment. A: results for fragments; B: results for Natural Products; C: results for Synthetic Compounds. The number of remaining molecules at each step is displayed below the tasks, when changes occur. Below it, the percentage of remaining molecules in regards to the initial number is display in grey. Above tasks, the number observed elements is displayed, when different from molecules. Fhits: Fragment Hits; FC: Fragment combinations; TP: True Positive; FP: False Positive; Matches: the number of fragment hits; FCG: Fragment Combination Graphs.

During our first attempt at the NPFC project, we included all fragments obtained from the Over *et al.* dataset. Since the Murcko scaffolds were extracted from the structures (using RDKit), the benzene ring was obtained as an NP-derived fragment on its own and was found to strongly impact the results (Table S3). Indeed, the benzene fragment was found to be part of 37.34% of all remaining NPs at the end of the workflow and represented 26.73% of all fragment combinations at this stage (data not shown). The impact observed for PNPs was even stronger, with the benzene ring found in 63.39% of the structures and representing 44.11% of all fragment combinations.

Table S3. Number of molecules and entries per step including benzene

Step	DNP		ChEMBL	
	Number of molecules incl. benzene (% of total)	Number of entries incl. benzene (% of total)	Number of molecules incl. benzene (% of total)	Number of entries incl. benzene (% of total)
FS	81,656 (50.18%)	139,140 (16.09%)	1,431,740 (88.03%)	2,537,707 (37.89%)
FCC	63,922 (47.80%)	354,169 (10.95%)	1,212,409 (86.99%)	4,918,570 (44.60%)
FCG	14,886 (37.34%)	16,920 (34.93%)	530,266 (69.45%)	557,805 (68.45%)
PNP	N/A		317,005 (63.39%)	333,629 (62.54%)

FS: Fragment Search; FCC: Fragment Combination Classification; FCG: Fragment Combination Graph; PNP: Pseudo-Natural Product. The nature of the entries varies depending on the step, i.e. FS: fragment hits; FCC: fragment combinations; FCG/PNP: fragment combination graphs. Percentages relative to the total of molecules/entries found at each step (with or without benzene) are displayed in parenthesis.

V. Computational time

Computations were performed on a local cluster of 4 nodes of type Intel® Xeon® Gold 6136, with 12 cores at 3.70 GHZ each (24 threads). Elapsed time was estimated to approx. 90s for the fragments, 40min for the DNP and 4h for the ChEMBL. Cumulated computational time (when considering the running time for all tasks sequentially) was 1min 22s for the fragments, 2 days 11h 32min for the DNP and 16 days 11h 50min for the ChEMBL. Most time-consuming steps were the standardization of the structures (in particular the tautomer canonicalization), the fragment search and to a lower extent, the fragment combination classification (see supporting information for more details). No exhaustive benchmarking was performed.