International Journal of
Molecular Sciences

MDPI

Article

Comparative Analyses of Medicinal Chemistry and Cheminformatics Filters with Accessible Implementation in Konstanz Information Miner (KNIME)

Sebastjan Kralj [1], Marko Jukič [1,2,*] and Urban Bren [1,2,*]

[1] Laboratory of Physical Chemistry and Chemical Thermodynamics, Faculty of Chemistry and Chemical Engineering, University of Maribor, Smetanova 17, 2000 Maribor, Slovenia; sebastjan.kralj1@um.si
[2] Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Glagoljaška 8, 6000 Koper, Slovenia
* Correspondence: marko.jukic@um.si (M.J.); urban.bren@um.si (U.B.); Tel.: +386-2-22-94-428 (M.J.); +386-2-2294-421 (U.B.)

Abstract: High-throughput virtual screening (HTVS) is, in conjunction with rapid advances in computer hardware, becoming a staple in drug design research campaigns and cheminformatics. In this context, virtual compound library design becomes crucial as it generally constitutes the first step where quality filtered databases are essential for the efficient downstream research. Therefore, multiple filters for compound library design were devised and reported in the scientific literature. We collected the most common filters in medicinal chemistry (PAINS, REOS, Aggregators, van de Waterbeemd, Oprea, Fichert, Ghose, Mozzicconacci, Muegge, Egan, Murcko, Veber, Ro3, Ro4, and Ro5) to facilitate their open access use and compared them. Then, we implemented these filters in the open platform Konstanz Information Miner (KNIME) as a freely accessible and simple workflow compatible with small or large compound databases for the benefit of the readers and for the help in the early drug design steps.

Keywords: high-throughput virtual screening; virtual screening; compound libraries; library design; compound filtering

Citation: Kralj, S.; Jukič, M.; Bren, U. Comparative Analyses of Medicinal Chemistry and Cheminformatics Filters with Accessible Implementation in Konstanz Information Miner (KNIME). *Int. J. Mol. Sci.* **2022**, *23*, 5727. https://doi.org/10.3390/ijms23105727

Academic Editor: Paulino Gomez-Puertas

Received: 20 April 2022
Accepted: 16 May 2022
Published: 20 May 2022

1. Introduction

Combinatorial chemistry (CC), novel library design methodologies, and high-throughput screening (HTS) represent the standard approaches for synthesis and evaluation (searching and selecting) of potential lead compounds in drug design efforts [1]. The combined use of chemical libraries and HTS to sift through large libraries and select desired compounds vastly increases the success rate of drug discovery programs [2,3]. Assays are now performed with libraries consisting of several million compounds: (Pfizer, 4 million [4]; Novartis 1.7 million [5]; Astra Zeneca 4 million [6]). Physical compound libraries and HTS are still regarded as the staple method for identification of leads; however, the advance of computational tools and in silico chemistry means that computer-aided methods have become indispensable in modern drug design efforts. If commercial physical compound libraries include several million molecules, the virtual compound libraries nowadays span from 10^7 to 10^{18} molecules. Nevertheless, such expansion of chemical space is a double-edged sword, as on one hand the probability of finding potential leads when screening larger libraries is greater, but on the other hand, screening of entire libraries even with the aid of computational methods may not be economically viable or even accessible in a timely manner. With the identification of the biological target in the early steps of the drug discovery process and the definition of the binding site, the chemical space adequate for further lead design becomes specific, and such information can be used to tailor compound libraries [7]. The specific nature of molecular recognition and interaction combined with

the fact that drugs must exhibit additional properties such as bioavailability and acceptable toxicity profiles severely narrows the adequate chemical space making drug design a monumental undertaking. Therefore, in a contemporary VS (virtual screening) or HTVS (high-throughput virtual screening) scenario, the database design is essential for efficient downstream calculations and in vitro testing. In order to achieve success in drug design efforts we need to adhere to certain library design guidelines. Libraries should focus the chemical space on the specific problem at hand, the compounds synthesizable and enriched with molecules that have drug-like properties [8].

The main challenge of library construction is to cover as much diversity of the chemical space as possible, while keeping the total number of compounds low to reduce time and money consumption [9,10]. Various molecular filters are often used to increase the hit rates of drug development campaigns [11]. With the hit rate of screening being on average as low as 1%, the simplest and most direct way to increase hit rate is to eliminate molecules with a low probability of becoming leads [12]. Filtering removes both unwanted chemical structures and unwanted chemical properties and is used to tailor the molecular libraries in a target focused manner [1]. The work on molecular filters was pioneered by Chris Lipinski and coworkers, who compared early HTS and combinatorial chemistry drug hits at Pfizer (up to 1994) with a subset of 2245 drugs from the World Drug Index [13]. The aim was to understand the common molecular features of orally available drugs and using an efficient version of the QSAR paradigm for structure permeability as suggested by Van de Waterbeemd et al. [14]. They came to several conclusions on the factors affecting poor absorption and permeation [15]. The main principle behind filtering of libraries is based on the term of drug-likeness. Although the term is often used in different ways by different authors, it generally refers to molecules that have properties or contain functional groups that are consistent with the majority of the known drugs [3,16,17]. The typical drug-like compounds exhibit desirable properties such as oral bioavailability, low toxicity, membrane permeability, and reasonable clearance rates [18]. Drug-like molecules therefore occupy distinct chemical space described by molecular descriptors and assigned cut-off values derived from experience. The first and to this day the most popular filters in use focused on finding effective and orally absorbable compounds [3]. The main goal of such filters was to address ADME (absorption, distribution, metabolism, and excretion) issues. The research on this topic points towards the fact that certain properties such as logP, MW (molecular weight), and number of hydrogen bonding groups correlate with oral bioavailability. This fact has been used to improve the success of finding lead-like molecules with filters that bias the chemical space of libraries, resulting in filters designed for various drug development applications [16,17]. Besides filters for drug-like properties, several filters exist that adopt the same knowledge-based approach in their design but expand beyond the scope of classic drug-like filtering. Filters such as the Ro4 (rule-of-4), designed to focus libraries on protein–protein interaction inhibitors, use descriptor cut-offs that are opposite of what is traditionally defined as drug-like and attest to the universal nature of molecular filters [19].

With preparation of molecular libraries, it is not just a question of what to filter out but when. Rules in the form of filters mean that compounds are discriminated on a pass or fail basis—compounds that pass the rules are considered equal, as are all that breach the rules [20]. Typically, filters are employed in the starting steps of a drug discovery campaign. Applying such filters upfront reduces the number of compounds analyzed in successive steps, speeding up the drug development process. However, this comes at the price of eliminating compounds that could show desirable properties in later phases. This is especially true for stringent filters [21] and for the use of compounds that have conformational flexibility [22]. The application of filters in the later stages avoids the problem of eliminating potential leads, but also causes the computationally intensive tasks to be performed on larger libraries, increasing both the financial and time costs. Moreover, we would like to point out that some authors argue against screening out promiscuous compounds in the early drug discovery [23]. Opponents of filtering point out that any rule-based system of filtering ignores the fact that exceptions exist, and that blind use of such restrictive filters

would eliminate potential drugs such as cyclosporine and erythromycin, where the majority of the drug-like rules break down [3]. Exceptions such as the aforementioned drugs bring up an important topic of distinction between properties of useful lead-like molecules and drugs. Regardless of whether the screening is done upfront of filtering on more diverse libraries or after filtering on more focused libraries, structural changes for lead optimization will usually be necessary [3]. In general structures, lead compounds exhibit less molecular complexity (less MW, fewer number of rings and rotatable bonds) and are less hydrophobic (lower clogP and logD). This indicates that the process of optimizing simple leads into drugs is favorable, supporting the idea of filtering libraries before screening and optimizing them into drugs later [24]. Filtering out "undesirable" molecular species using computational filters thus forms a key element in library preparation and carries an informed decision in defining "favorable" or "undesirable" properties [25]. Thresholds for such properties are often derived from the experience of the pharmaceutical industry [21]. The criteria of "undesirable" structures should always be considered in their suitable scientific context, e.g., the loss of peptidomimetic molecules employing typical rule-based filters such as the Ro5 (Lipinski's rule of five) in the development of a protease inhibitor library would result in a poor hit rate [13,26]. Therefore, we encourage the reader to consider the biological context of the target, the drug discovery campaign, and to employ a plethora of filters to flag compounds for consideration and design in the subsequent drug discovery campaign steps. When using multiple filters in a sequential manner it is generally best to employ the filter that removes the most compounds first to reduce time consumption in later steps. One should also consider which filters will be applied without exceptions and which ones will merely flag the compounds for later assessment. Those that will filter without exceptions should be applied beforehand. A good example of a consecutive filtering protocol is described in the work of Jukič et al., where the library was first filtered for large and small compounds followed by filtering for aggregators, PAINS and REOS [27].

To successfully apply filters in HTVS, the selected compound library must use supported data formats, for example, the string representation SMILES (simplified molecular input line entry specification format) or 3D representations such as SDF (structure-data file format) or MOL (MDL Molfile) [28]. In most cases, 3D conformational data are not required for the use of filters, as these filters are usually referred to as "2D filters". Despite the widespread adoption of SMILES for storage and interchange of chemical structures no standard for generating SMILES strings exist. The application of canonical SMILES, which use only a single string per molecule, is recommended to avoid duplication and problems in future filtering. To address issues of specifying isotopism and stereochemistry of a molecule the isomeric SMILES was developed and is useful for scoping the library for stereoisomerism duplicates or to generate stereoisomers and expand the chemical space. A SMILES string can be canonical and isomeric at the same time [29]. The SMILES expansion SMARTS (SMILES arbitrary target specification) allows specification of sub-structural patterns and is used for specification of protonation state, hydrogen count, and ionization states. As both the SMILES and SMARTS format are not an open project and are proprietary, this has resulted in the use of different generation algorithms by software developers, resulting in different SMILES versions for the same compounds. Moves towards the open-source string representations of compounds and standardization have been made with OpenSmiles and InChI [30]. However, with the current state of compound libraries the use of standardized chemical forms is not the norm, and care should be taken when combining such libraries for virtual screening [10]. We recommend the use of Konstanz Information Miner (KNIME) software for standardizing the input format before filtering either from the 3D SDF or the string SMILES representation, in an analogous way performed in the filters provided by this article.

Many filters for compound library design are present in primary scientific literature with some such as Lipinski's rule-of-5 enjoying widespread recognition in the scientific community; however, many filters for drug design do not enjoy the same recognition. To bridge the gap between molecular filters and their accessibility to the public, we sought out

to implement them in an open-access program that allows visual and dataflow programming through a graphical user interface. We therefore collected data on molecular filters, implemented them into existing open-access software, and compared them side by side to benefit the reader in his/her early drug design steps [31].

2. Results

To test and demonstrate the functionality of the filters implemented and their effects on the chemical space, we applied the filtering workflow on a general ZINC database [32]. The database was obtained by accessing the ZINC website (https://zinc.docking.org/tranches/home/ accessed on 21 June 2021) selecting the following parameters (representation "2D", reactivity "standard", purchasability "in-stock") and downloading the SMILES wget command file. The final downloaded library consisted of 9,216,175 compounds (a large non-specific chemical library). Using the KNIME row sampling node, 1% of the total database was sampled and ran through all the filters implemented in KNIME. We then calculated the average values and standard deviations (SD) of several key molecular descriptors using the statistics KNIME node to assess the change in chemical space after filtering (Figures 1–6). The descriptors chosen were a standard basic set most descriptive for initial chemical space assessment; the partition coefficient as SlogP, molecular refractivity (SMR), total polar surface area (TPSA), molecular weight (MW), No. of rotatable bonds, No. of hydrogen bond acceptors (HBA), No. of hydrogen bond donors (HBD), No. of heavy atoms, No. of rings, and the number of atoms C, N, O present in the compounds. We see that filters impact the chemical space of libraries to various degrees. The more specific the filter, the larger the portion removed, since the chemical space on which they are based is far more defined than with general drug-like filters.

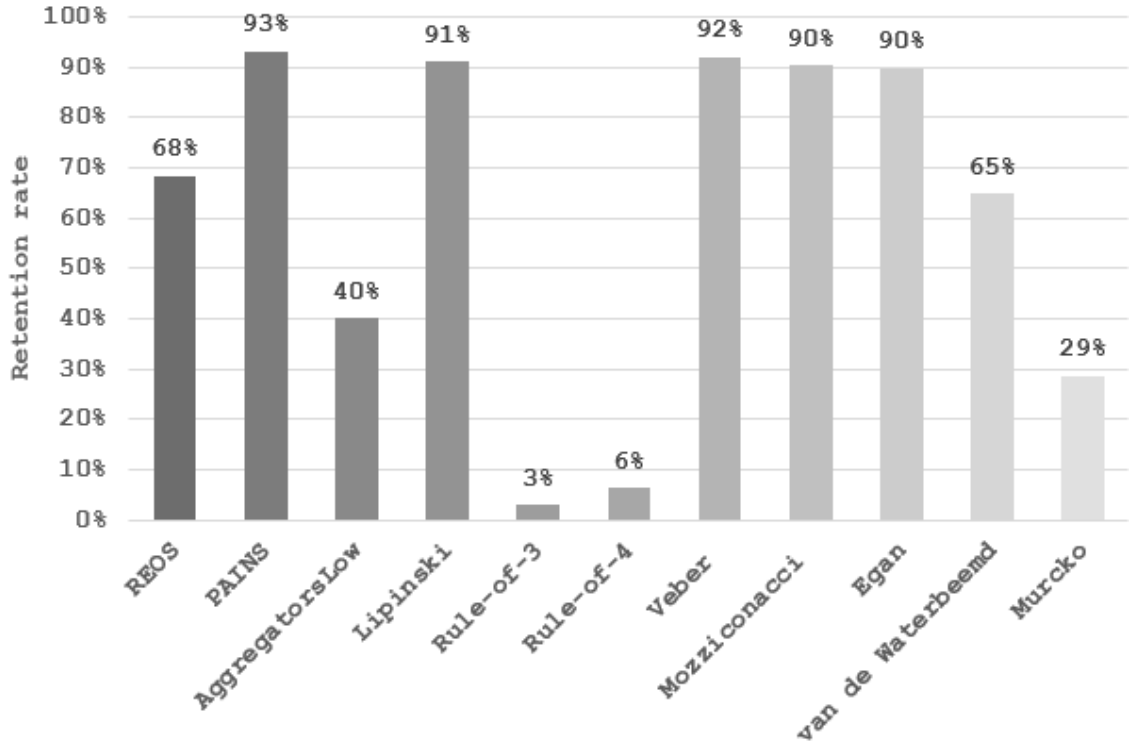

Figure 1. Percentage of compounds passing the described filters. The total number of sampled compounds of the unfiltered library is used as the denominator.

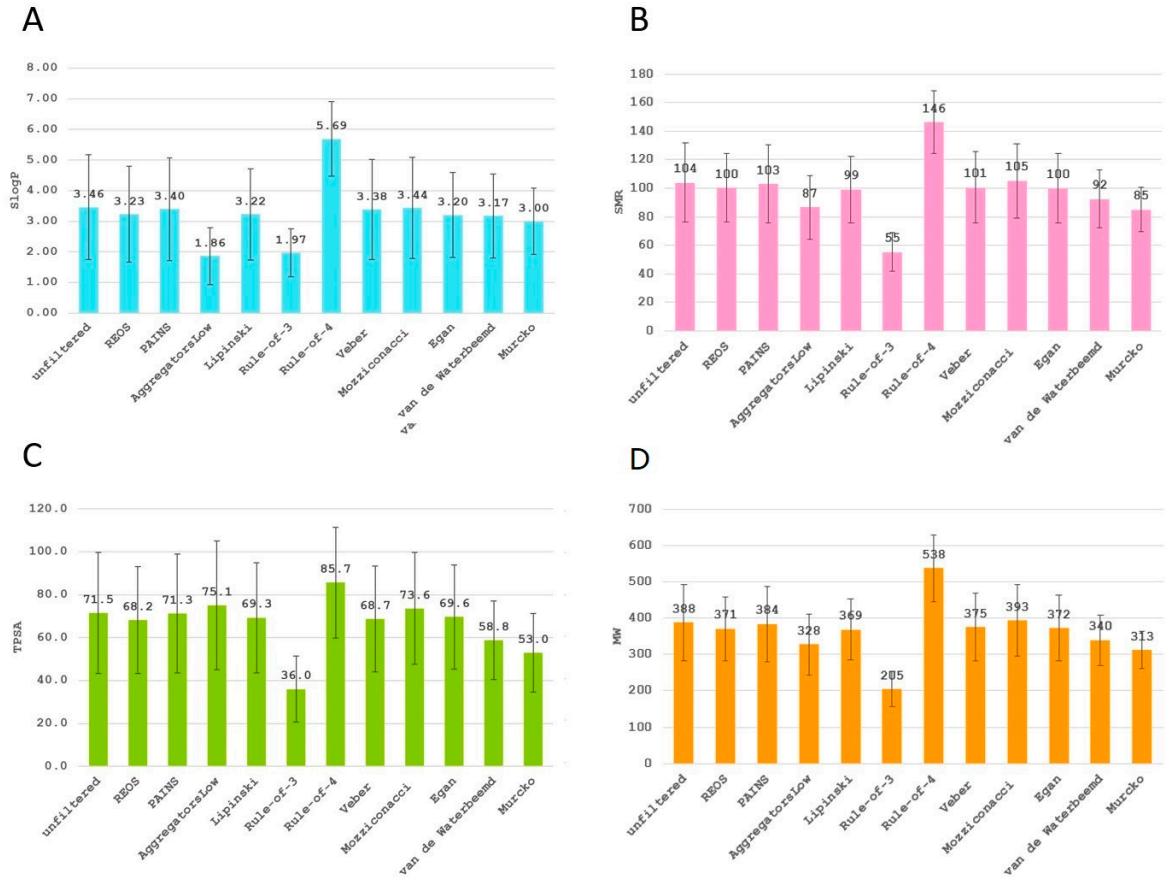

Figure 2. (**A**) The average descriptor value with SD of SlogP for the unfiltered and post filtering libraries. The majority of the filters have values close to the average, with AggregatorsLow and Rule-of-3 scoring lower, since they have strict cut-off values for SlogP. The partition coefficient is used to assess the lipophilicity of a drug and its ability to cross cell membranes. (**B**) The average descriptor value with SD of SMR for the unfiltered and post filtering libraries. The clear outliers are the Ro3 and Ro4, which strictly define the chemical space. (**C**) The average descriptor value with SD of TPSA for the unfiltered and post filtering libraries. The Ro3 with molecules small in size scores lower than the average, with the Ro4 being slightly above the average, but not as significantly as in the previous graphs. (**D**) The average descriptor value with SD of MW for the unfiltered and filtered libraries. The molecular weight descriptor is a very common descriptor used for cut-off values. As most of the filters aim at drug-like molecules except for the Ro4 and Ro3, the average weights are very similar.

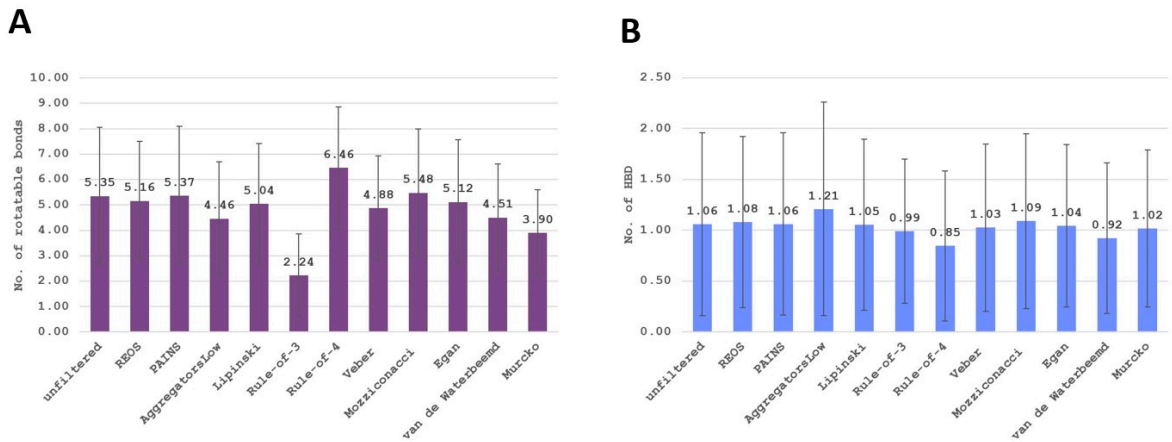

Figure 3. *Cont.*

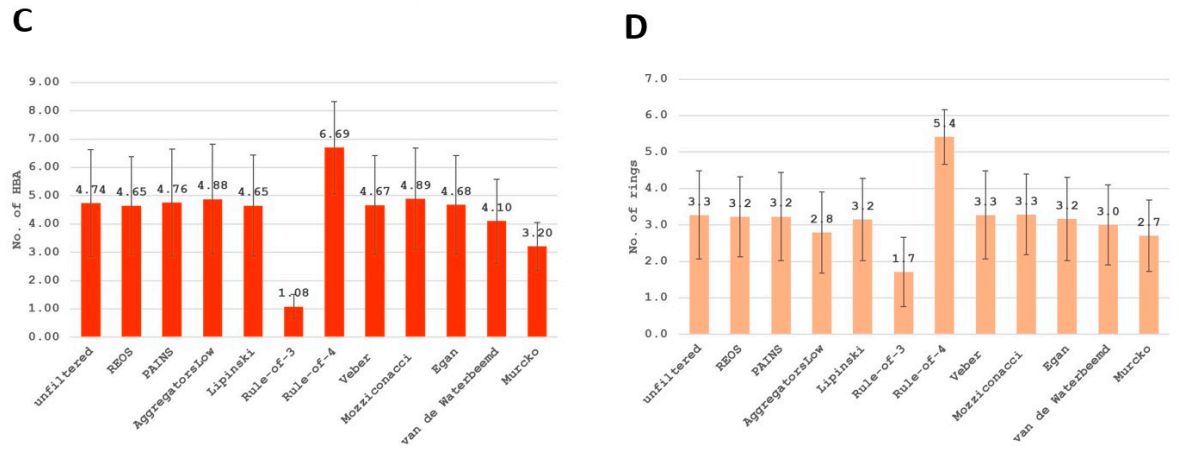

Figure 3. (**A**) The average descriptor value with SD of the No. of rotatable bonds for the unfiltered and filtered libraries. The graph bears resemblance to the other graphs where Ro3 and Ro4 stand out, with the other filters having average values close to the unfiltered library. (**B**) The average descriptor value with SD of the No. of HBD. We see that molecules that pass the aggregators filter have a slightly higher value of hydrogen bond donors. What is interesting is also the fact that the Ro4 scores lower than the average, despite having molecules that are larger and contain more N and O atoms which are usually involved in hydrogen bonding (Figure 4). (**C**) The average descriptor value with SD of the No. of HBA before and after filtering of the library. The Ro3 filter has a significantly lower value as its aim is to find the starting fragments from which the molecule is built. This usually leaves space for the attachment of desired functional groups to the fragment, but as a result the number of HBA is lower. (**D**) The average descriptor value with SD of the No. of rings present before and after filtering of the library. As Ro4 shifts the chemical space towards larger molecules the number of rings increases as well. The opposite happens with Ro3 where the small molecular weight does not allow for a large number of rings to be present.

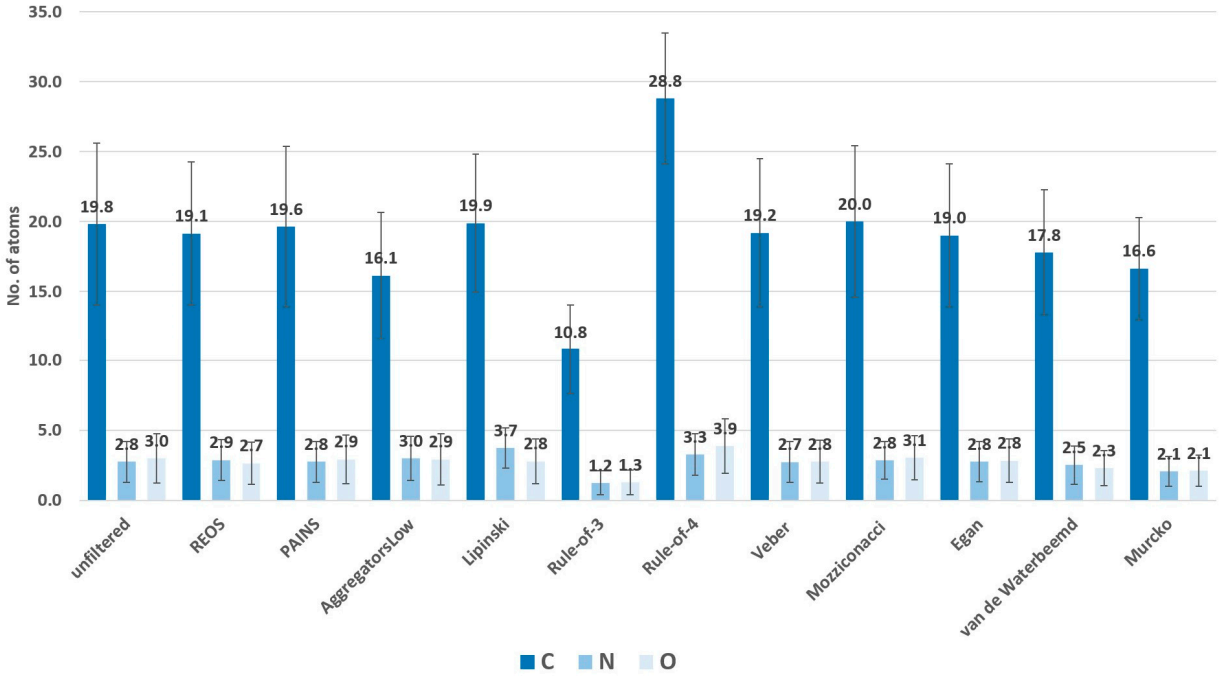

Figure 4. The average number of C, N, and O atoms present in the compounds. The majority of the libraries are within the average values of the unfiltered, with Rule-of-4 having higher values since the filter retains large molecules that are better suited for inhibiting protein–protein interactions. Rule-of-3 scores lower as it retains smaller compounds suitable for fragment-based drug design.

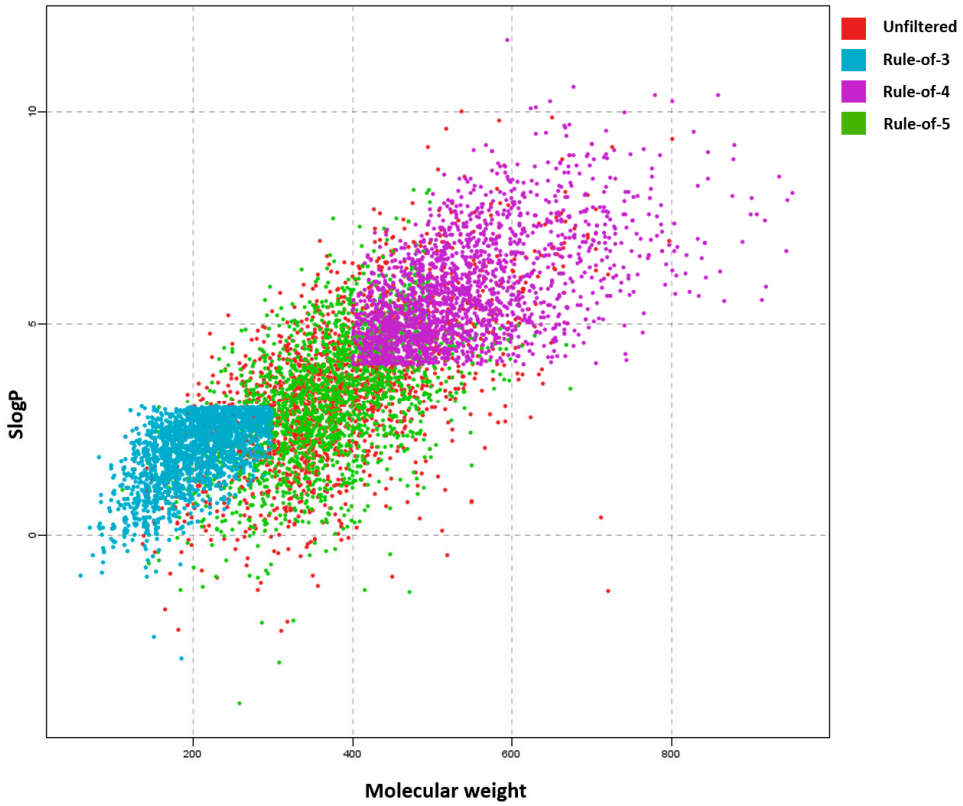

Figure 5. A 2D scatter plot of SlogP and molecular weight for compounds that passed individual filters. We can see the impact of using filters on the chemical space as the Rule-of-3 (blue) is totally separated from the Rule-of-4 (purple) group. This is due to the strict molecular weight cut-off and the SlogP cut-off as we see horizontal and vertical lines indicating where the cut-offs are. Since Lipinski's Rule-of-5 (green) allows one rule break per compound, we do not observe such strict horizontal lines and the chemical space after filtering is still very similar to the unfiltered (red) library.

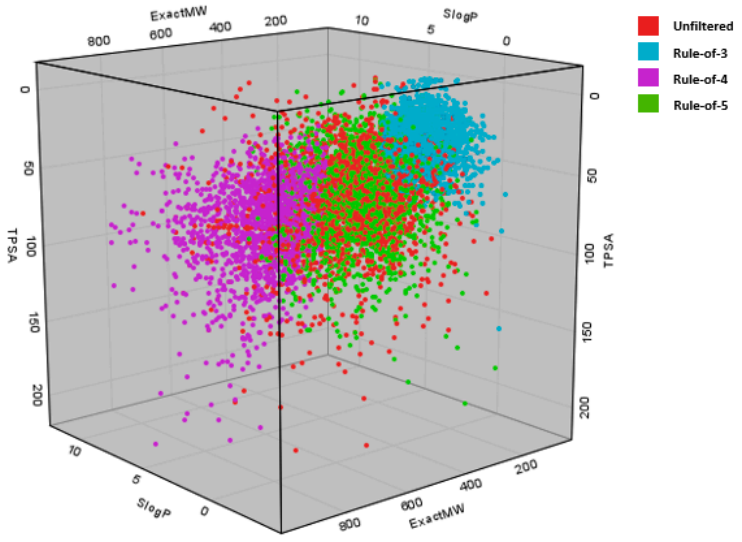

Figure 6. A 3D scatterplot of exact molecular weight, SlogP, and total polar surface area (TPSA) for the unfiltered (red) and filtered Rule-of-3 (Blue), Rule-of-4 (purple), and Rule-of-5 (green). We can see how the space occupied by compounds changes drastically; similar observations can be made as in the case of the 2D plot in Figure 5.

3. Discussion

The REOS (rapid elimination of swill) filter removed 32% of the compounds from the tested ZINC database subset but left the chemical space unaffected when compared to the original dataset, as it has no cutoffs on the investigated descriptors. This holds true for the functional group filter for PAINS (pan-assay interference compounds) as well; however, only 7% of compounds were removed. This is due to the fact that the filter is not as broad and tries to remove certain problematic moieties that were not captured in previously developed functional group filters (e.g., REOS). The aggregator filter, designed by Irwin et al. compares similarity of a library to 12.600 known aggregators (http://advisor.bkslab.org/rawdata/ accessed on 10 February 2022), (set to the most stringent cutoff of "low" similarity to known aggregators) removes ~60% of the database and significantly lowers the SlogP value, as it uses this descriptor as a cutoff to determine the aggregation propensity. The number of rings is also lower after filtering for aggregators implicating that the presence of rings might be involved in aggregation. However, this descriptor is not used as a cutoff, but is indirectly correlated with the properties of aggregators. The average value of rings for the dataset of known aggregators is 3.6 ± 1.03, which is slightly above the average of 3.3 ± 1.2 for the general database, meaning that compounds with rings would likely score higher in the Tanimoto coefficient comparison and get filtered out. The Ro3 (rule of three) and Ro4 (rule of four) filters are the most stringent filters as they define the most specific chemical space, filtering out 97% and 94% of the database, respectively. Despite their similarity in the filtered-out percentage, they operate in opposite ways. The Ro3 represents a strict filter designed to support "hit identification" and "fragment-based" drug research and only accepts molecules with a molecular weight of less than 300. It supports the paradigm that small compounds still capture the desired chemical space yet leave a lot of space for future compound optimization towards leads. The Ro4 attempts to capture the protein–protein interaction inhibitor chemical space and retains molecules with molecular weight above 400, as such larger molecules are able to form multiple interactions. Morelli et al. designed the filter with the aim of establishing guidelines for druggable protein–protein inhibitors, since these most often break traditional property filter rules. Beside the high MW cutoff, Ro4 retains only compounds containing multiple rings and is often above average in the descriptor value graphs (Figures 2–4). The Veber and Egan filters remove a small fraction of molecules with 7.9% and 10.3%, respectively, as they both apply only two filtering rules with a mild cut-off value. The Veber filter tries to capture molecules with good oral bioavailability properties. With just two cut-offs that focus strictly on oral bioavailability, it filters out 8% of the dataset. Another bioavailability and membrane permeability filter is the Egan filter which filters out 10% of the dataset. The molecules score lower in average descriptor values across all the examined descriptors, with both the Egan and Veber filters supporting the notion that smaller compounds are more membrane permeable and show greater bioavailability. The Mozziconacci filter, a filter for drug-like properties, applies five descriptor cutoff rules. All five descriptors used are different from the classical Rule-of-5 descriptors. The Lipinski Rule-of-5 is a set of four rules (logP, MW, and H-bond donor and acceptor cut-offs) for drug-likeness and oral bioavailability derived from a subset of 2245 drugs. It removes a similar share of the data set as well with the Lipinski filter removing 9% and the Mozziconacci filter 10%. Despite both being drug-like filters placing the filters in a chain-like matter, with the Mozziconaci filter placed after Lipinski, we filter out an additional 9% of the total dataset. This means that the drug-like definition of both filters is very different and may be used in conjunction for strict drug-like filtering. Despite only two descriptor rules for the passing of the blood–brain barrier, the Van de Waterbeemd filter removes 35% of the molecules from the database, in large part due to the small TPSA cutoff value, which is reflected in a reasonably low average TPSA descriptor value (Figure 2).The Murcko filter, due to its specificity (determining compounds with central nervous system (CNS) activity), filters out 71% percent of the database using five cut-offs. Low descriptor values for TPSA and molecular weight can also be observed

as with the Veber and Egan filters, since these molecules must be smaller in order to pass the blood–brain barrier [33].

4. Materials and Methods

To facilitate open access use of various filters for drug design, we decided to implement the described filters into a single unit, where researchers could access various filters or combine them to a multi-filter to speed up their own drug development efforts. The first step incorporated a thorough search of the literature for information on molecular filters with the aim of defining, implementing, and sorting them as clearly as possible for the end user. Filters described were sorted into one of the two groups; filters that filter out based on the presence of functional groups and filters that filter out based on physiochemical properties.

Filters designed to exclude compounds based on the presence of functional groups most often aim to remove compounds that are reactive toward protein targets. The most common such functional groups are Michael acceptors, ketones, aldehydes, and suicide inhibitors. Such compounds would likely be false HTS positives and would increase time and money expenses spent on screening. Removing reactive functionality is based on the premise that covalent interactions are not desired for drug design except for specific cases [15]. Besides filtering for compounds with reactive species, functional group filters aim to remove optically interfering components, aggregators, fluorescent compounds, firefly luciferase inhibitors, redox cycling compounds, oxidizers, cytotoxic compounds, compounds with quenching ability, and surfactant-like compounds, all of which would frequently appear as false positives in the screening tests. Several filters fall under this category, with their properties described in Table 1 [34,35]. Some filters, although classified as functional group filters, do possess some additional property filters making them hybrid filters. We collected all filters present in the literature and added a brief description with the cut-off values on which the filter is based (Tables 1 and 2).

Table 1. The most common functional group filters described in the scientific literature presented in alphabetical order.

Name/Reference	Description	Features/Cutoff Values
Aggregators [36]	Tanimoto coefficient similarity search to a database of known aggregators.	Tanimoto coefficient similarity $\geq$ 0.85 or SlogP > 5 (high similarity), Tanimoto coefficient similarity $\geq$ 0.5 and SlogP > 3 (medium similarity), Tanimoto coefficient similarity < 0.85 and SlogP $\leq$ 3 (low similarity)
Ely Lilly Rules [37]	A set of 275 rules, developed over an 18-year period, used to identify compounds that may interfere with biological assays, allowing their removal from screening sets.	Reasons for rejection of compounds: reactivity, interference with assay measurements (fluorescence, absorbance, quenching), instability and lack of druggability (lacking both oxygen and nitrogen)
Muegge method [18,38]	Bioavailability prediction rules dubbed the Muegge method. Pharmacophore filter developed by analyzing known drug databases, with four functional molecular motifs determined to be important in drug-like molecules:	Primary, secondary, and tertiary amines are considered pharmacophore points but not pyrrole, indole, thiazole, isoxazole, other azoles, or diazines. Compounds with more than one carboxylic acid are dismissed. Compounds without a ring structure are dismissed. Intracyclic amines that occur in the same ring are fused and count as only one pharmacophore point.

Table 1. *Cont.*

Name/Reference	Description	Features/Cutoff Values
PAINS [39]	Removal of frequent hitters (promiscuous compounds) by identifying sub-structural features not recognized by filters commonly used to identify reactive compounds.	Functional groups such as rhodanines, phenolic Mannich bases, hydroxyphenylhydrazones, alkylidene barbiturates, alkylidene heterocycles, 1,2,3-aralkylpyrroles, activated benzofurazans, 2-amino-3-carbonylthiophenes, catechols, and quinones do not pass the filters.
REOS [1] [3,40,41]	Seven property filters (similar to the PATTY rules in program developed at Merck)	H-bond donor ≤ 5, H-bond acceptors ≤ 10, $-2 \leq$ Formal charge $\leq +2$, Number of rotatable bonds ≤ 8, $200 \leq$ Molecular weight ≤ 500, $20 \leq$ number of heavy atoms ≤ 50, $-2 \leq \log P \leq 5$
	Functional group filters for the removal of problematic structures dubbed REOS (rapid elimination of swill; program developed at Vertex).	Reactive, toxic and other undesirable moieties such as nitro groups, preoxides, triflates, aldehydes, acetals, etc.

[1] REOS is a hybrid filter which combines a set of functional group filters with property filters. As the REOS filter can be combined with other (property) filtering schemes, the property filtering part can be omitted and only functional group filters employed. As implemented in KNIME, the user can also specify the maximum quantity for each of the functional group rules, tuning the filter to the needs of the individual research scenario. REOS moieties in the SMARTS format can be found inside the KNIME workflow "REOS substructures" node.

Table 2. The most common property filters described in the scientific literature.

Name/Reference	Description	Features/Cutoff Values
Egan [42]	Set of rules designed by analyzing the data on compounds both well and poorly absorbed in humans with multivariate statistics. Two descriptors (AlogP and PSA) were chosen for inclusion when determining membrane permeability. Compounds that pass exhibit good bioavailability.	AlogP ≤ 5.88, polar surface area ≤ 131.6 Å^2
Fichert [43]	Rules for structure-permeability based on a set of 41 small drug-like molecules. LogD is the main property that determines permeability, with structures passing this filter being highly permeable in the Cacao-2 model.	Molecular weight ≤ 500, $0 \leq \log D \leq 3$
Ghose [44]	A set of rules for drug-likeness derived from characterizing 6304 compounds taken from the Comprehensive Medicinal Chemistry Database.	$180 \leq$ molecular weight ≤ 480, $40 \leq$ molecular refractivity ≤ 130, $-0.4 \leq \text{ClogP} \leq 5.6$, $20 \leq$ number of atoms ≤ 70
Lee filter [45]	Analysis of natural products to determine potential appealing scaffolds for future drug design. Pharmacophoric properties of natural products, trade drugs, and virtual combinatorial library were assessed, finding key properties and several scaffolds which could work as building blocks.	MW mean ~356 LogP mean ~2.1
Lipinski (Rule-of-5) [13]	A set of four rules for drug-likeness and oral bioavailability derived from a subset of 2245 drugs from the World Drug Index. The rules aim to address the ADME issues.	Molecular weight ≤ 500, $\log P \leq 5$, H-bond donors ≤ 5, H-bond acceptors ≤ 10
Mozziconacci [46]	Filter developed by Mozziconacci after analyzing 15 freely available chemical libraries (2 million compounds). Drug-likeness was examined using common chemical features and based on the successive filters were designed to extract the drug-like subset.	Rotatable bonds ≤ 15, number of rings ≤ 6, oxygen atoms ≥ 1, nitrogen atoms ≥ 1, halogen atoms ≤ 7
Murcko filter [33,47]	Rules for determining CNS activity, joining 7 property descriptors (Rule-of-5 with the addition of rotatable bonds, aromatic density, and a measure for branching) and 166 fingerprint descriptors to determine presence or absence of functional groups.	MW 200–540, $\log P$ 0–5.2, H-bond acceptors ≤ 4, H-bond donor ≤ 3, rotatable bonds ≤ 7, branching behavior 3.4–12.2, aromatic rings < 3
Oprea Lead-Like [1,24]	A set of rules based on lead-like vs. drug-like comparison after examination of several commercially available databases. The rules aim to maintain focus towards effective and orally absorbable compounds. Beside the properties chosen based on the Rule-of-5, additional properties were chosen to better reflect molecular complexity of a library and the rigidity of a molecule.	Molecular weight < 450, $-3.5 \leq \log P < 4.5$, $-4 \leq \log D \leq 4$, number of rings ≤ 4, nonterminal single bonds ≤ 10, H-bond donor ≤ 5, H-bond acceptor ≤ 8

Table 2. *Cont.*

Name/Reference	Description	Features/Cutoff Values
Rule-of-3 [48]	Rules designed to support "fragment-based" drug research. Hits obtained using this filter can be useful for fragment libraries used to generate potential leads. Fragment libraries are useful for sampling chemical diversity or targeting specific interactions.	Molecular weight $\leq$ 300, logP $\leq$ 3, H-bond donor $\leq$ 3, H-bond acceptors $\leq$ 3 rotatable bonds $\leq$ 3
Rule-of-4 [19]	A set of rules derived from analyzing the 2P2I database that contains protein–protein interaction inhibitors with the aim of establishing guidelines for druggable protein–protein inhibitors, since these most often break traditional property filter rules.	Molecular weight $\geq$ 400, logP $\geq$ 4, number of rings $\geq$ 4, H-bond acceptors $\geq$ 4
van de Waterbeemd [49,50]	Physiochemical properties for estimation of blood–brain barrier crossing of compounds. Rules were derived by examination of lipophilicity, H-bonding capacity, and molecular shape and size descriptors of marketed CNS and CNS-inactive drugs.	Molecular weight $\leq$ 450, polar surface area $\leq$ 90 Å^2
Veber [51]	Two rules to meet the criteria for oral bioavailability derived after studying bioavailability measurements in rats for of over 1100 drug candidates at GlaxoSmithKline.	Rotatable bonds $\leq$ 10, polar surface area $\leq$ 140 Å^2

The other group of filters consists of classical property filters designed to bias the chemical space of filtered libraries into a predetermined and desired direction. As stated above, the majority of such filters aim to define and narrow the scope of the library towards the drug-like paradigm. Property filters eliminate the extrema of undesired properties present in the libraries [1]. The extrema are determined from distributions in databases of desired compounds (e.g., databases of approved drugs).

After a careful analysis of the primary filter literature and the implementation of filters in existing bioinformatics software packages, KNIME was chosen as an open and accessible platform for the implementation of examined filters. Its intuitive workflow design, supported by a graphical interface, and its ability for large scale HTVS with the KNIME server makes it perfect for the integration in the established drug design workflows of users, be it ligand or structure-based drug design. KNIME allows users to create visual data flows, or pipelines, where data traverse multiple user-selected nodes. These nodes represent an essential part of KNIME, with each node possessing unique data processing capabilities, where the input and output of each node can transparently be analyzed (Figure 7) [14].

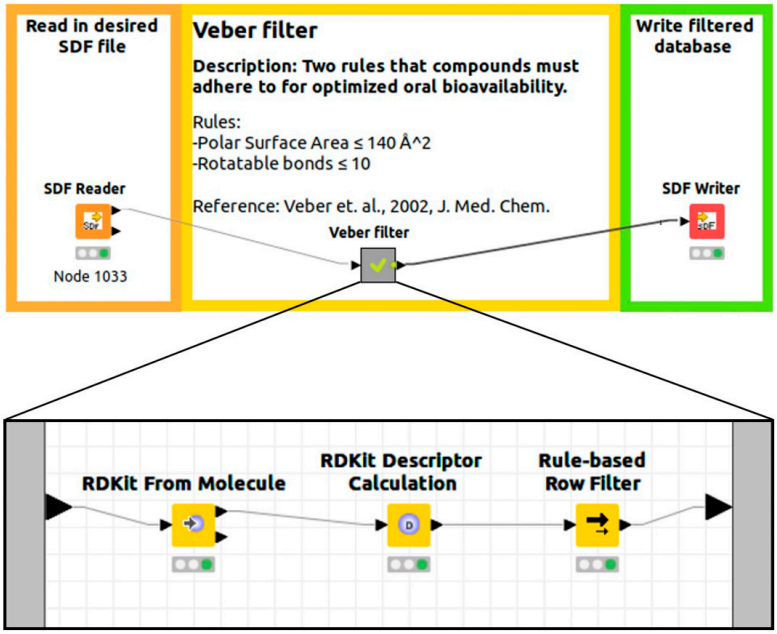

Figure 7. KNIME workflow example of our Veber filter implementation for effective design of compound libraries. Black lines represent the expanded meta node that contains sub-nodes [51].

The workflows were created using KNIME version 4.2.3 (available at http://knime.org accessed on 17 November 2020). Additional expansion nodes from RDKit, MOE extensions, and Vernalis KNIME were used for the final version of the workflow alongside the default KNIME nodes. All the mentioned nodes are distributed as KNIME community extensions accessible to everyone in their full functionality. All nodes and workflows are open and editable by the user if he/she wishes to change certain parameters or develop novel filters. Experienced users can expand the meta nodes and delete redundant steps in the process (e.g., duplicate generation of the canonical SMILES in the linked workflow) when combining several filters for their drug design, which would result in even faster workflows (Figure 8). The node output can be edited to produce various outputs ranging from text and table formats to chemical library formats suitable for further drug design.

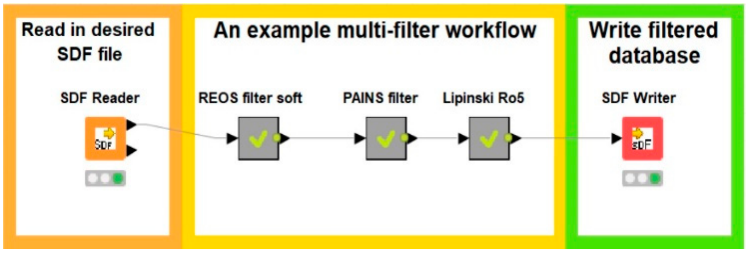

Figure 8. An example of combining meta nodes to form a complex drug design workflow.

We implemented 11 filters (REOS, PAINS, Aggregators, Rule-of-5, Rule-of-4, Rule-of-3, Veber filter, Mozziconacci filter, Egan filter, Van de Waterbeemd filter, Murcko filter) into our multi-filter KNIME workflow accessible at public repository (https://gitlab.com/Jukic/knime_medchem_filters/ accessed on 15 January 2022). The PAINS and REOS filter are both based on the RDKit substructure counter and compare the substructures present in the input database with a list of problematic functional groups. A rule-based row filter removes the hits from the database. The aggregation propensity detection filter, called the "aggregator filter", evaluates the aggregation propensity based on the similarity calculated by Tanimoto coefficients of given molecules to a database containing known aggregators [15]. The user can personally control how strict the filter is with the low, medium, and high propensity filters provided. The remaining filters are knowledge-based rule-based filters that, when expanded, can often be modified by the user to suit his or her own needs. The filters are simple property counting filters that firstly calculate descriptor values using the RDKit Descriptor calculator node or the molecule properties (Mozziconacci) and then employ the rule-based row filters. The exception being the Rule-of-5 which allows one rule break, to incorporate the filter consisting of rule engines that assign the value of 1 for each rule break, with the math formula summing up all the values and the final rule-based row filter comparing the value to see. The impact of strict cut-offs that define specific chemical spaces and milder filters such as the Lipinski's Rule-of-5 which allow a rule break can be seen in Figures 5 and 6.

5. Conclusions

After analyzing and implementing several molecular medicinal chemistry filters and testing the created workflows, we conclude that compound filters are essential for modern computer aided drug design (CADD). They provide the researcher with a simple, fast, and robust way to enrich the chemical space and to reduce the time associated with post-filtering methods. They are also easy to use and can be customized to particular preferences of the studied chemical space. However, the user must be aware of the properties used for filtering, as some, such as REOS and PAINS, were not designed with covalent chemistry in mind. In such cases, it is better to flag the compounds for a later evaluation. We firmly believe that this article provides medicinal chemistry community with a handful of useful

workflows for novel drug design, identification, and HTVS, as well as with a good initial overview of compound filtering in drug discovery.

Author Contributions: Conceptualization, methodology, review and editing, S.K., M.J. and U.B.; original draft preparation, data curation, visualization, S.K. and M.J.; supervision, resources, U.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Slovenian Research Agency (ARRS) program and project grants P2-0046, P1-0403, L2-3175, J1-1715, and J1-2471 as well as by the Slovenian Ministry of Education and Science research project grants ELIXIR-SI and HPC RIVR.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We gratefully acknowledge the support of NVIDIA Corporation with the donation of GPU hardware that was used in this research.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

CC	combinatorial chemistry
HTS	high-throughput screening
VS	virtual screening
HTVS	high-throughput virtual screening
ADME	absorption, distribution, metabolism, and excretion
MW	molecular weight
Ro4	Rule-of-4
Ro5	Rule-of-5
Ro3	Rule-of-3
SDF	structure-data file format
MOL	MDL Molfile
SMILES	simplified molecular input line entry specification format
SMARTS	SMILES arbitrary target specification
KNIME	Konstanz Information Miner
SD	standard deviations
SMR	molecular refractivity
TPSA	total polar surface area
MW	molecular weight
HBA	No. of hydrogen bond acceptors
HBD	No. of hydrogen bond donors
REOS	rapid elimination of swill
PAINS	pan-assay interference compounds
CNS	central nervous system

References

1. Oprea, T.I. Property Distribution of Drug-Related Chemical Databases. *J. Comput. Aided Mol. Des.* **2000**, *14*, 251–264. [CrossRef] [PubMed]
2. Liu, R.; Li, X.; Lam, K.S. Combinatorial Chemistry in Drug Discovery. *Curr. Opin. Chem. Biol.* **2017**, *38*, 117–126. [CrossRef] [PubMed]
3. Walters, W.P.; Murcko, M.A. Prediction of "Drug-Likeness". *Adv. Drug Deliv. Rev.* **2002**, *54*, 255–271. [CrossRef]
4. Bakken, G.A.; Bell, A.S.; Boehm, M.; Everett, J.R.; Gonzales, R.; Hepworth, D.; Klug-McLeod, J.L.; Lanfear, J.; Loesel, J.; Mathias, J.; et al. Shaping a Screening File for Maximal Lead Discovery Efficiency and Effectiveness: Elimination of Molecular Redundancy. *J. Chem. Inf. Model.* **2012**, *52*, 2937–2949. [CrossRef] [PubMed]
5. Njoroge, M.; Njuguna, N.M.; Mutai, P.; Ongarora, D.S.B.; Smith, P.W.; Chibale, K. Recent Approaches to Chemical Discovery and Development against Malaria and the Neglected Tropical Diseases Human African Trypanosomiasis and Schistosomiasis. *Chem. Rev.* **2014**, *114*, 11138–11163. [CrossRef]

6. Morgan, P.; Brown, D.G.; Lennard, S.; Anderton, M.J.; Barrett, J.C.; Eriksson, U.; Fidock, M.; Hamrén, B.; Johnson, A.; March, R.E.; et al. Impact of a Five-Dimensional Framework on R&D Productivity at AstraZeneca. *Nat. Rev. Drug Discov.* **2018**, *17*, 167–181. [CrossRef]
7. Náray-Szabó, G. Analysis of Molecular Recognition: Steric Electrostatic and Hydrophobic Complementarity. *J. Mol. Recognit.* **1993**, *6*, 205–210. [CrossRef]
8. Walters, W.P.; Stahl, M.T.; Murcko, M.A. Virtual Screening—An Overview. *Drug Discov. Today* **1998**, *3*, 160–178. [CrossRef]
9. Hajduk, P.J.; Galloway, W.R.J.D.; Spring, D.R. A Question of Library Design. *Nature* **2011**, *470*, 42–43. [CrossRef]
10. Kralj, S.; Jukič, M.; Bren, U. Commercial SARS-CoV-2 Targeted, Protease Inhibitor Focused and Protein–Protein Interaction Inhibitor Focused Molecular Libraries for Virtual Screening and Drug Design. *IJMS* **2021**, *23*, 393. [CrossRef]
11. Macarron, R.; Banks, M.N.; Bojanic, D.; Burns, D.J.; Cirovic, D.A.; Garyantes, T.; Green, D.V.S.; Hertzberg, R.P.; Janzen, W.P.; Paslay, J.W.; et al. Impact of High-Throughput Screening in Biomedical Research. *Nat. Rev. Drug Discov.* **2011**, *10*, 188–195. [CrossRef] [PubMed]
12. Thorpe, D.S.; Edith Chan, A.W.; Binnie, A.; Chen, L.C.; Robinson, A.; Spoonamore, J.; Rodwell, D.; Wade, S.; Wilson, S.; Ackerman-Berrier, M.; et al. Efficient Discovery of Inhibitory Ligands for Diverse Targets from a Small Combinatorial Chemical Library of Chimeric Molecules. *Biochem. Biophys. Res. Commun.* **1999**, *266*, 62–65. [CrossRef] [PubMed]
13. Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **2001**, *46*, 3–26. [CrossRef]
14. van De Waterbeemd, H.; Camenisch, G.; Folkers, G.; Raevsky, O.A. Estimation of Caco-2 Cell Permeability Using Calculated Molecular Descriptors. *Quant. Struct.-Act. Relat.* **1996**, *15*, 480–490. [CrossRef]
15. Oprea, T. Virtual Screening in Lead Discovery: A Viewpoint. *Molecules* **2002**, *7*, 51–62. [CrossRef]
16. Walters, W.P.; Murcko, A.A.; Murcko, M.A. Recognizing Molecules with Drug-like Properties. *Curr. Opin. Chem. Biol.* **1999**, *3*, 384–387. [CrossRef]
17. Lipinski, C.A. Drug-like Properties and the Causes of Poor Solubility and Poor Permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235–249. [CrossRef]
18. Muegge, I. Pharmacophore Features of Potential Drugs. *Chemistry* **2002**, *8*, 1976–1981. [CrossRef]
19. Morelli, X.; Bourgeas, R.; Roche, P. Chemical and Structural Lessons from Recent Successes in Protein–Protein Interaction Inhibition (2P2I). *Curr. Opin. Chem. Biol.* **2011**, *15*, 475–481. [CrossRef]
20. Bickerton, G.R.; Paolini, G.V.; Besnard, J.; Muresan, S.; Hopkins, A.L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4*, 90–98. [CrossRef]
21. Charifson, P.S.; Walters, W.P. Filtering Databases and Chemical Libraries. *J. Comput. Aided Mol. Des.* **2002**, *16*, 311–323. [CrossRef] [PubMed]
22. Lumley, J.A. Compound Selection and Filtering in Library Design. *QSAR Comb. Sci.* **2005**, *24*, 1066–1075. [CrossRef]
23. Senger, M.R.; Fraga, C.A.M.; Dantas, R.F.; Silva, F.P. Filtering Promiscuous Compounds in Early Drug Discovery: Is It a Good Idea? *Drug Discov. Today* **2016**, *21*, 868–872. [CrossRef] [PubMed]
24. Oprea, T.I.; Davis, A.M.; Teague, S.J.; Leeson, P.D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308–1315. [CrossRef]
25. Olah, M.M.; Bologa, C.G.; Oprea, T.I. Strategies for Compound Selection. *Curr. Drug Discov. Technol.* **2004**, *1*, 211–220. [CrossRef]
26. Lenci, E.; Trabocchi, A. Peptidomimetic Toolbox for Drug Discovery. *Chem. Soc. Rev.* **2020**, *49*, 3262–3277. [CrossRef]
27. Jukič, M.; Janežič, D.; Bren, U. Ensemble Docking Coupled to Linear Interaction Energy Calculations for Identification of Coronavirus Main Protease (3CLpro) Non-Covalent Small-Molecule Inhibitors. *Molecules* **2020**, *25*, 5808. [CrossRef]
28. Dalby, A.; Nourse, J.G.; Hounshell, W.D.; Gushurst, A.K.I.; Grier, D.L.; Leland, B.A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244–255. [CrossRef]
29. Hähnke, V.D.; Kim, S.; Bolton, E.E. PubChem Chemical Structure Standardization. *J. Cheminform.* **2018**, *10*, 36. [CrossRef]
30. Heller, S.R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* **2015**, *7*, 23. [CrossRef]
31. Zhu, T.; Cao, S.; Su, P.-C.; Patel, R.; Shah, D.; Chokshi, H.B.; Szukala, R.; Johnson, M.E.; Hevener, K.E. Hit Identification and Optimization in Virtual Screening: Practical Recommendations Based on a Critical Literature Analysis: Miniperspective. *J. Med. Chem.* **2013**, *56*, 6560–6572. [CrossRef] [PubMed]
32. Irwin, J.J.; Shoichet, B.K. ZINC—A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182. [CrossRef] [PubMed]
33. Ajay; Bemis, G.W.; Murcko, M.A. Designing Libraries with CNS Activity. *J. Med. Chem.* **1999**, *42*, 4942–4951. [CrossRef] [PubMed]
34. Thorne, N.; Auld, D.S.; Inglese, J. Apparent Activity in High-Throughput Screening: Origins of Compound-Dependent Assay Interference. *Curr. Opin. Chem. Biol.* **2010**, *14*, 315–324. [CrossRef] [PubMed]
35. Rishton, G.M. Nonleadlikeness and Leadlikeness in Biochemical Screening. *Drug Discov. Today* **2003**, *8*, 86–96. [CrossRef]
36. Irwin, J.J.; Duan, D.; Torosyan, H.; Doak, A.K.; Ziebart, K.T.; Sterling, T.; Tumanian, G.; Shoichet, B.K. An Aggregation Advisor for Ligand Discovery. *J. Med. Chem.* **2015**, *58*, 7076–7087. [CrossRef]
37. Bruns, R.F.; Watson, I.A. Rules for Identifying Potentially Reactive or Promiscuous Compounds. *J. Med. Chem.* **2012**, *55*, 9763–9772. [CrossRef]

38. Muegge, I.; Heald, S.L.; Brittelli, D. Simple Selection Criteria for Drug-like Chemical Matter. *J. Med. Chem.* **2001**, *44*, 1841–1846. [CrossRef]
39. Baell, J.B.; Holloway, G.A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719–2740. [CrossRef]
40. Walters, W.P.; Namchuk, M. Designing Screens: How to Make Your Hits a Hit. *Nat. Rev. Drug Discov.* **2003**, *2*, 259–266. [CrossRef]
41. Bush, B.L.; Sheridan, R.P. PATTY: A Programmable Atom Type and Language for Automatic Classification of Atoms in Molecular Databases. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756–762. [CrossRef]
42. Egan, W.J.; Merz, K.M.; Baldwin, J.J. Prediction of Drug Absorption Using Multivariate Statistics. *J. Med. Chem.* **2000**, *43*, 3867–3877. [CrossRef] [PubMed]
43. Fichert, T.; Yazdanian, M.; Proudfoot, J.R. A Structure-Permeability Study of Small Drug-like Molecules. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 719–722. [CrossRef]
44. Ghose, A.K.; Viswanadhan, V.N.; Wendoloski, J.J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55–68. [CrossRef] [PubMed]
45. Lee, M.L.; Schneider, G. Scaffold Architecture and Pharmacophoric Properties of Natural Products and Trade Drugs: Application in the Design of Natural Product-Based Combinatorial Libraries. *J. Comb. Chem.* **2001**, *3*, 284–289. [CrossRef] [PubMed]
46. Mozziconacci, J.C.; Arnoult, E.; Baurin, N.; Marot, C. Preparation of a Molecular Database from a Set of 2 Million Compounds for Virtual Screening Applications: Gathering, Structural Analysis and Filtering. In Proceedings of the 9th Electronic Computational Chemistry Conference, World Wide Web, 1–31 March 2003.
47. Darvas, F.; Keseru, G.; Papp, A.; Dorman, G.; Urge, L.; Krajcsi, P. In Silico and Ex Silico ADME Approaches for Drug Discovery. *CTMC* **2002**, *2*, 1287–1304. [CrossRef]
48. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A "rule of Three" for Fragment-Based Lead Discovery? *Drug Discov. Today* **2003**, *8*, 876–877. [CrossRef]
49. van de Waterbeemd, H.; Camenisch, G.; Folkers, G.; Chretien, J.R.; Raevsky, O.A. Estimation of Blood-Brain Barrier Crossing of Drugs Using Molecular Size and Shape, and H-Bonding Descriptors. *J. Drug Target.* **1998**, *6*, 151–165. [CrossRef]
50. van de Waterbeemd, H. Physicochemical Approaches to Drug Absorption. In *Methods and Principles in Medicinal Chemistry*; van de Waterbeemd, H., Testa, B., Eds.; Wiley: Hoboken, NJ, USA, 2008; Volume 40, pp. 69–99; ISBN 978-3-527-32051-6.
51. Veber, D.F.; Johnson, S.R.; Cheng, H.-Y.; Smith, B.R.; Ward, K.W.; Kopple, K.D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615–2623. [CrossRef]