Journal of Pharmaceutical Research International

32(48): 16-30, 2020; Article no.JPRI.65682
ISSN: 2456-9119
(Past name: British Journal of Pharmaceutical Research, Past ISSN: 2231-2919, NLM ID: 101631759)

VLVS: A Knime Based Virtual Library Generation and Screening Workflow

Ismail Hakki Akgün[1*]

[1]Department of Bioengineering, Faculty of Engineering, Ege University, 35100, Bornova, İzmir, Türkiye.

Author contribution

The sole author designed, analyzed, interpreted and prepared the manuscript.

Article Information

DOI:10.9734/JPRI/2020/v32i4831121
Editor(s):
(1) Dr. Giuseppe Murdaca, University of Genoa, Italy.
Reviewers:
(1) Subhash Chandra, Kumaun University, India.
(2) Chittaranjan Baruah, Darrang College (Gauhati University), India.
Complete Peer review History: http://www.sdiarticle4.com/review-history/65682

Original Research Article

Received 02 December 2020
Accepted 07 February 2021
Published 09 February 2021

ABSTRACT

Objective: Tocreate an easy-to-use structure-based screening workflow using KNIME and open source software to prepare and screen virtual libraries in order to discover novel bioactive or drug molecules.
Materials and Methods: In the preparation of the workflow we mentioned in the article KNIME version 4, AutoDock Vina, Pymol were used. KNIME plugins used in this study are RDKit KNIME Integration, ChemAxon/Infocom Marvin Extensions Feature, KNIME Python Integration, Lhasa Metabolism Feature, KNIME Ploty and KNIME JavaScript View. We have used Python3 and libraries in the python scripts through Anaconda installation.
Results: A workflow that can work with a single click after making required adjustments which uses docking as structure-based screening method was created and tested.
Conclusion: With the workflow we have created, it will be possible for researchers especially those who are working in the field of computer-aided drug design/development to create personalized molecule libraries, perform virtual screening, reporting the results in a short time with the least effort.

Keywords: Bioactive molecule design; virtual screening; docking; virtual library; drug design; drug discovery.

**Corresponding author: E-mail: ismail.hakki.akgun@ege.edu.tr;*

1. INTRODUCTION

Virtual molecule libraries can be thought of as a collection of molecules that can theoretically be produced. It is a great advantage that libraries consist of molecules that can be synthesized. Once a possible bioactive molecule has been identified, the continuation of the studies can be ensured as long as the molecules are obtained at high purity and amount. Although there are many physical and virtual molecule libraries [1-5], having unique molecules in their bodies will be an advantage in the race to obtain bioactive molecules for both companies and academic research groups. In the physical and virtual molecule libraries that already exist there are diverse and similar molecules [6] but these libraries allow researchers to screen a relatively similar and narrow regions of huge chemical space [7-9]. Although virtual screening or screening using molecules from existing physical libraries is an advantage for further studies, it poses a disadvantage as the supply chain of the possible bioactive molecule will be dependent on third parties. In these conditions, it becomes important to create virtual or physical molecule libraries with structural diversity using the building blocks that already exist.

Many academic groups and companies have worked in this direction, and some of their efforts are summarized in a mini perspective written by Walters. [9]. For example, scientists in Pfizer have used the information they gained from their syntheses to build the Pfizer Global Virtual Library (PGVL). [10]. Using selected reaction schemes and reagents, the researchers at Lilly obtained about 10 million virtual molecules and used them in virtual screening processes. In addition, the researchers stated in the article they published that the molecules mentioned were synthesizable with the Automated Synthesis Laboratory system called ASL [11]. Some academic groups have also had similar studies. Chevillard and Kolb have built a virtual library of 21 million molecules using virtual reactions and created subets coded as S, consisting of 9994 molecules, M consisting of 99977 molecules, and L, consisting of 999794 molecules. [12]. The number of molecules obtained by the groups in these three examples we have given is quite impressive and represents only a small fraction of the possible sizes of virtual molecule libraries. In addition, it is important to note that the groups used reaction synthesis schemes while creating virtual libraries.

Reaction scheme was defined as “drawing that unambiguously defines the regio- and stereo-chemical outcome of a given synthetic transformation, in general terms, using R groups to show optional substitution.” by the authors of the article which we mentioned above related to Pfizer Global Virtual Library (PGVL). [10] The use of reaction schemes enables the physical availability of the virtual molecules and the reaction conditions to be known. Although there are various methods and software that make it possible to use reaction schemes in virtual reactions, the most striking ones are the Daylight Reaction Smarts [13] and ChemAxon Reactor module [14], which we use indirectly in our article.

Using high-throughput screening methods is one of the most frequently used strategies to discover novel bioactive / drug molecules. [15] While its high precision, speed, minimization of the test method can be counted as the advantages of HTS, its high cost and the need for automation are also its disadvantages. Virtual screening (VS) can be used as a complement to the HTS methods or as a starting point for bioactive or drug molecules discovery process. [16,17] The most commonly used virtual screening methods can be grouped under two main categories. These are Ligand Based Virtual Screening (LBVS) methods and Structure Based Virtual Screening (SBVS) methods. LBVS methods can be defined as the methods which search for similar molecules to a molecule or a group of molecules with known bioactivity. Methods such as ligand base pharmacophore screening, screening of molecular descriptors and properties can be cited as examples of LBVS methods. In SBVS methods, the interactions of molecules with the target structure - mostly proteins - and the scores obtained from these interactions are examined. It is possible to find more information about some studies in which these methods are used to determine the lead molecules to be used in drug development processes in the article prepared by Ma and colleagues. [17]

Docking is one of the most commonly used SBVS method in bioactive / drug molecule discovery process. The docking method examines the interactions of small molecules and proteins at the atomic level in two steps. The first step is to predict the conformation and orientations (poses) that the molecule will have in the active site of the protein. The second step is scoring of the estimated poses. [18] There are

many open source and commercial software that can be used in the docking method. [19] There are many examples that can be given to bioactive or drug molecules determined using the docking method and some of them have been summarized by Ghosh and colleagues [15]. AutoDockVina is one of the most commonly used and cited docking software on SBVS process. [20] Its ease of use, enabling multithreading, successfully predicting docking poses, and being open source make AutoDockVina a popular docking software. There are many successful VS studies in which AutoDockVina is used. For example, Uddin and colleagues designed a series of peptides using the rational design method and selected three of them using AutoDockVina software to be active. It was determined that all three of the selected peptides had better Ki values than the known ligand. [21] In another study, Perryman and colleagues worked to identify new bioactive molecules due to the resistance of *Mycobacterium tuberculosis* to some drugs. 316000 molecules in the NCI library were subjected to virtual screening process using AutoDockVina software and 91 molecules were selected according to the calculated binding energies. After the visual inspection, 16 molecules were selected and two of them were determined as the most active molecules with K_i values of 54 and 59 µm. It has been determined that these two molecules show low structural similarity with the known InhA inhibitor molecules. This is an indication that there may be two new molecular skeletons that can be used in the development of InhA inhibitors. [22]

Konstanz Information Miner (KNIME) is a modular system that allows the creation of workflows that can easily process data consisting of interconnected nodes or modules. KNIME can be easily used in training, research and collaboration projects [23] and enables researchers to make reproducible analyzes. Thanks to its capabilities that can be expanded with add-ons, it has a wide potential for use in life sciences. [24] Next generation sequencing [25], metabolomics analysis [26], QSPR [27], QSAR [28], high content screening [29] and drug discovery studies [30-32] are examples of wide range of uses.

In this study, based on the issues mentioned above, we prepared a workflow on the KNIME platform which creates virtual product molecule library, enriches product molecules with the help of various medicinal chemistry filters, screens with the docking method, visualizes in a way that allows interactivity, and compiles data and stores them. In addition, an application of the workflow we have applied and the results obtained from this example were discussed.

2. MATERIALS AND METHODS

2.1 Materials

All operations in this study were carried out on the workstation with an i7 processor installed Ubuntu 18.04 on and KNIME version 4, AutoDockVina, Pymol were used. KNIME plugins or software used in this study are RDKit KNIME Integration, ChemAxon/Infocom Marvin Extensions Feature, KNIME Python Integration, Lhasa Metabolism Feature, KNIME Ploty and KNIME JavaScript View. We have used Python3 and libraries in the python scripts through Anaconda installation.

2.2 Methods

2.2.1 Preparation of reactants (section - 1)

First section consists of the workflow which has two separate metanode (Fig. 1). After the reagents are loaded into the workflow in the first metanode, duplicated molecules are removed and a simple reagent code is given to each reagent. After the codes are given to the reagents, all data except molecular structure and reagent code are filtered and the results obtained are transferred to the next metanode. In the second metanode of the section 1 after applying the functional group filter and structure normalization to the reagents, using the RDKit Molecular Descriptor node, SlogP, Total Polar Surface Area (TPSA), Average Molecular Weight (AMW), Number of Lipinski Hydrogen Bond Donor (NumLipinskiHBD), Number of Lipinski Hydrogen Bond Acceptor (NumLipinskiHBA), Number of Rotatable Bonds (NumRotatableBonds) are calculated. Then, molecules that do not meet the conditions of logP<= 3, TPSA <= 60 AMW <300, NumLipinskiHBA<= 3, NumLipinskiHB<= 3 and NumRotatableBonds<= 3 are filtered in accordance with the rule of three (Ro3) [33].

2.2.2 Performing reaction and product idgeneration (section- 2)

In this section filtered reactants are virtually reacted using the reaction scheme provided from the previous section. (Fig. 1) Virtual reactions are

carried out using the RDKit Two Component Reaction node. Subsequently, duplicate product molecules are removed and the remaining is transferred to metanode in which the product codes are generated. According to the coding system that we aim to facilitate the tracking of the product obtained, the codes are created using the reaction code specified by the user at the beginning of the project, and the derived codes of the first and second reagent used in obtaining the product. For example, the generated reaction code of the first member of the table showing the molecules we have obtained in our project is "Reaction_Example_R1_2_R2_12". In this code, "Reaction_Example" part is the reaction code defined by the user at the beginning of the project, "R1_2" is the derived code of the first reagent, "R2_12" is the derived code of the second reagent. The products whose codes have been generated are transferred to the third part where the medicinal chemistry filters will be applied.

2.2.3 Enrichment of libraries using various medicinal chemistry filters (section – 3)

In this section, SlogP, TPSA, AMW, NumLipinskiHBD, NumLipinskiHBA and NumRotatableBonds values were calculated in order to determine the drug-likeness properties of the product molecules (Fig 2). In addition to the features we have mentioned, FractionCsp3 values are also calculated. One of the most important studies conducted to evaluate the similarity of molecules to drug molecules or their potential to become drugs is "Rule of Five (Ro5)" developed by Lipinski. [34] According to the Ro5, the probability of being a drug molecule candidate increases if the molecule of interest has molecular weight less than 500 Da, logP value less than 5, the number of hydrogen bond donors equal to or less than 5, and the number of hydrogen bond acceptors less than or equal to 10. In addition, Veber et al. [35], Egan et al. [36], Ghose et al. [37], and Muegge et al. [38] conducted studies on this topic and suggested a variety of metrics (drug-likeness filters). In the study published by Diana and his colleagues in 2017, they developed a web server named "SwissAdme" that allows scoring the oral bioavailability or drug-likeness of molecules using various metrics such as these. [39] Taking SwissAdme web server as an example we scored the bioavailability and drug-likeness of the products molecules we obtained using the parameters we showed in Table 1 [39]. PAINS [40] and BRENK filters [41] which derived from information obtained from HTS screens are applied to the molecules passing through first two filters of this section. With these filters, product molecules containing functional groups that cause false positive results in bioactivity studies and unwanted pharmacochemical properties are removed. Whether the product molecules will be inhibitors of CYP isoenzymes is another topic that we considered in this section. This consideration was carried out using the WhichCyp 1.2 node developed by Lhasa Limited [42]. The molecules that passed through all the filtration steps we performed are transferred to the fourth section in order to be prepared for and to perform the docking process. Also they are transferred to the fifth section for interactive visualizations and to the sixth section for storage.

2.2.4 Docking preparation, docking and post docking analysis nodes (section - 4)

The processes related to the docking method of the molecules coming from the third section of the workflow are carried out in this fourth section (Fig 3). First, hydrogens are added to the molecules, then their three-dimensional coordinates are created and optimized using RDKit nodes. Before changing the data formats of molecules whose three-dimensional structures have been optimized, they are transferred to the fifth and sixth sections for visualization and storage. After this process, the data format of the molecules is converted to .pdbqt format, which is compatible with the docking software AutoDockVina. In AutoDockVinametanode of this section, there is Python script that creates folders to save the files obtained during the docking process and "docking_result.csv" file where the results will be stored. In the same metanode there is a second Python script that carries out the docking and then compiles the docking results as a post-docking process. This script has been prepared to repeat the docking process three times for each product molecule. Following the docking process, the averages and standard deviations of the RMSD values of the docking poses and the binding energies of each molecule are automatically calculated.In addition to these, the ligand efficiency parameter, which is calculated by dividing the binding energy to the number of heavy atoms (non-hydrogen atoms) in the relevant molecule, which facilitates the selection between possible active molecules, is also calculated with this script. [43] This script writes the compiled results to the "docking results.csv" file, and moves the configuration (to config folder), log (to log folder) and pose (to

docked folder) files to the separate folders generated during the docking process. After combining the data obtained from the docking process and the data previously calculated for each molecule, the results are transferred to the fifth and sixth sections for visualization and storage. The rule-based row filter node which was added to this section allows users to separate molecules with binding energy lower than the specified threshold value. In this way, promising molecules can be easily identified.

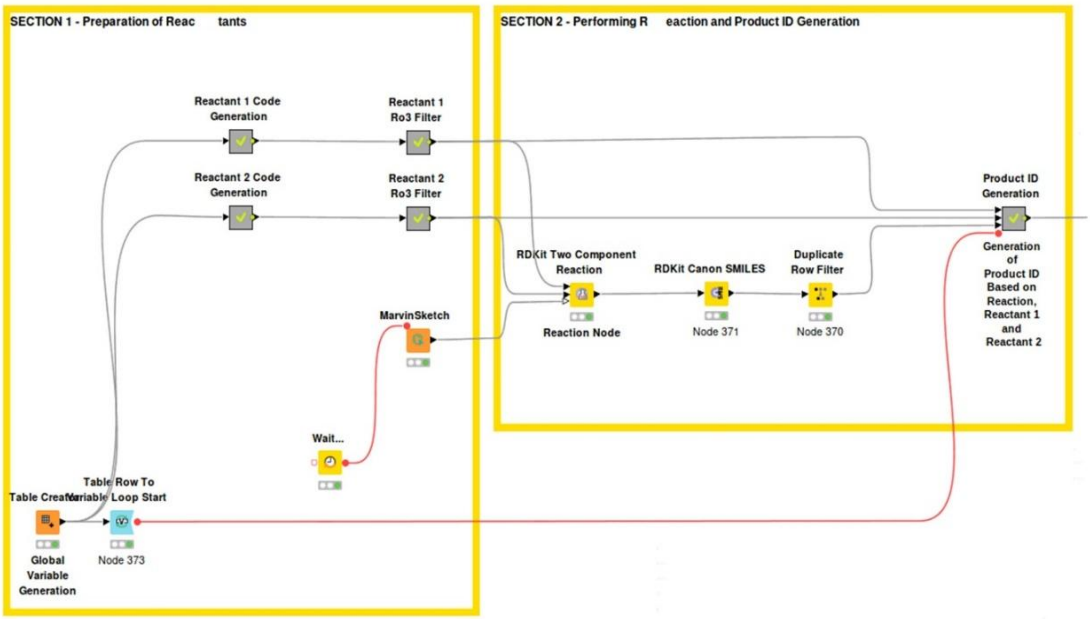


Fig. 1. Section -1 and Section – 2 of the workflow: Preparation of reactants, performing reaction and product idgeneration

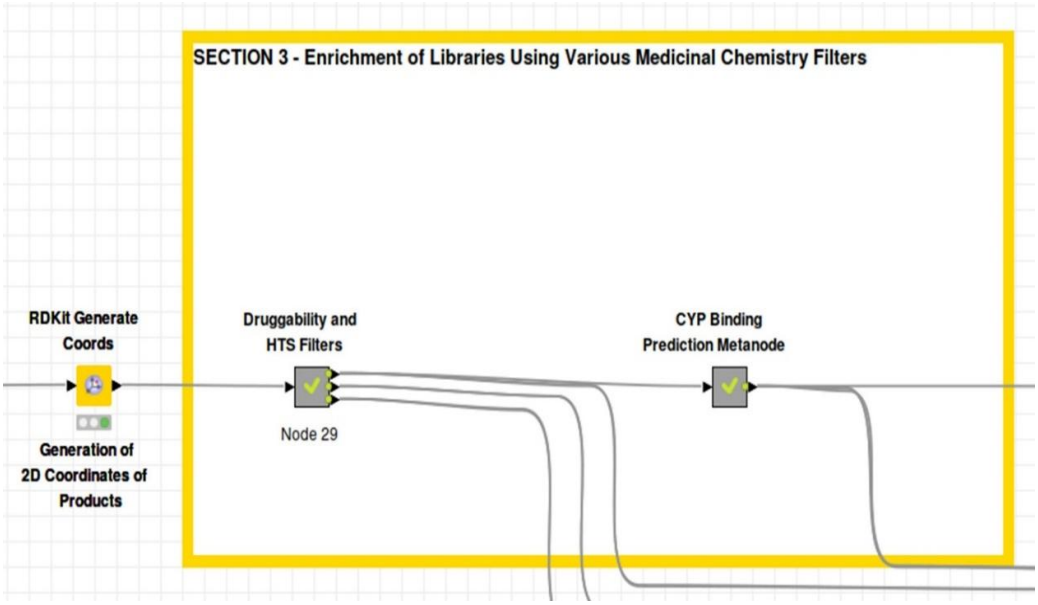


Fig. 2. Section - 3 of the workflow: Enrichment of libraries using various medicinal chemistry filters

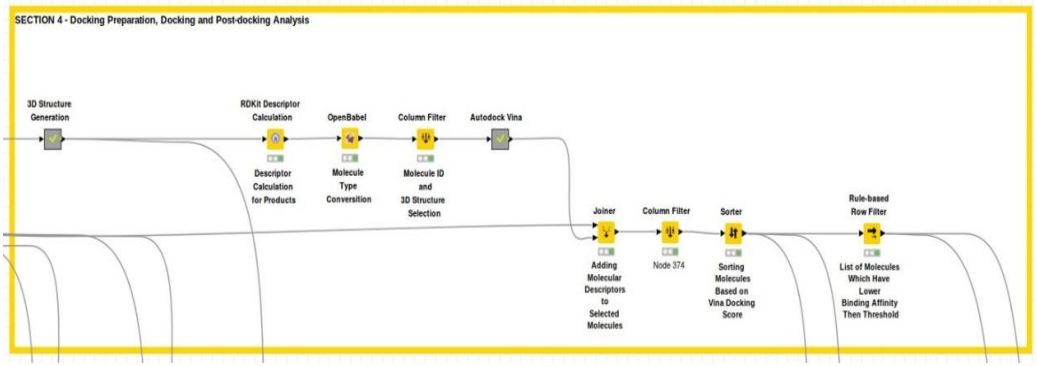


Fig. 3. Section - 4 of the workflow: Docking preparation, docking and post-docking analysis

2.2.5 Interactive dashboards (section - 5)

Section five of the workflow consists of interactive dashboards. (Fig. 4) Dashboards in this section are prepared using KNIME Ploty and KNIME JavaScriptViews nodes. Data used to create interactive graphics are obtained from the parameters calculated in section 3 (SlogP, TPSA, AMW, NumLipinskiHBD, NumLipinskiHBA and NumRotatableBonds, FractionCsp3). In addition to the histograms of these parameters, a radar chart was created to show the optimum oral bioavailability or drug-likeness zone. Since the units of the parameters we use are different, a normalization process has been carried out. The upper and lower boundaries of the optimum zone were established using the values we show in Table 1. The normalization process is carried out so that the normalized value of the lower limit shown in Table 1 is set to 0.33 and the upper limit to 0.67. The same visualization schemes were applied to all of the molecule groups obtained from the filters applied in the third section in order to compare enrichment levels. Different visualization schemes were applied to the results obtained from the docking process. The most important visualization in this section is the three-dimensional interactive graph prepared for molecules with better binding energy than the threshold value set by the user. The average of the binding energies, RMSD value of the poses obtained from the docking process and the ligand efficiency values are used to create the three-dimensional graph. With this graph, different perspectives can be evaluated in selecting the molecule that can be used in further studies. In addition, all visualization dashboards contain interactive tables showing molecules and their properties.

2.2.6 Product molecules save (section - 6)

In the sixth section of the workflow, the data obtained from the third and fourth sections are stored in the csv format using automatically created naming schemes. (Fig. 4) Only the data obtained from the node where the optimized three-dimensional structure of the molecules were created is stored in sdf format.

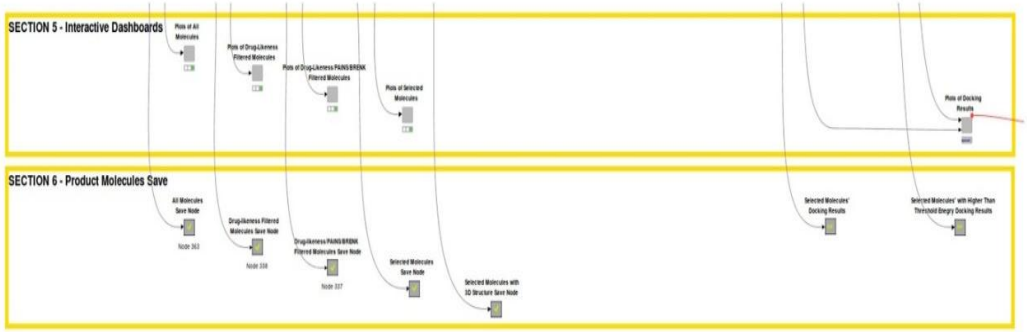


Fig. 4. Section – 5 and Section - 6 of the workflow: Interactive dashboards, product molecules save

Table 1. The upper and lower limits of bioavailability or drug-likeness filter used in section 3

	Lower limit	Upper limit
logP	-0.700	5.000
TPSA (Å)	20.00	130.00
AMW (g/mol - Da)	150.000	500.000
NumLipinskiHBA	0	<= 10
NumLipinskiHBD	0	<= 5
NumRotatableBonds	0	< 9
FractionCsp3	> 0.25	< 1.00

2.2.7 Testing of the workflow

In order to test the workflow we have created, "Acyl chlorides" group containing 317 and "Aromatic Primary Amines" groups containing 2357 molecules from Sigma Selected Subsets were used.

3. RESULTS

All of the 317 acyl chlorides loaded into the workflow were transferred smoothly from the first metanode of the first section "Reaction 1 Code Generator" to the second metanode "Reactant 1 Ro3 Filter". On the other hand, 17 of 2357 aromatic primary amines loaded into workflow were found in the list as duplicates and filtered. In the first metanode of the Reactant 2 2340 molecules were transferred from “Reactant 2 Code Generator” metanode to “Reactant 2 Ro3 Filter” metanode.

The first member of the "Reactant 1 Ro3 Filter" metanode is "Reactant 1 Functional Group Filter" node. Molecules containing more than one acyl chloride functional group were filtered at this node, and the number decreased from 317 to 293. The second member of the "Reactant 1 Ro3 Filter" metanode is "Reactant 1 Structure Normalisation" node. The goal of this node is to remove problematic (uncertain three-dimensional structures, salts, etc.) molecules. It was observed that 287 of the 293 acyl chloride entered this node passed the filter and 6 failed. It was determined that three of the six failed molecules had incorrect three-dimensional structure and three contained more than one fragment. The SlogP, AMW, TPSA, NumLipinskiHBA, NumLipinskiHBD and NumRotatableBonds values were calculated for the 287 acyl chloride passing the first two filters. After the Ro3 filter applied to 287 acyl chloride, the number of molecules dropped to 182. The minimum and maximum values of the parameters before and after the Ro3 filter are shown in Table 2. The remaining 182 acyl chloride were transferred to the section 2 to be used in virtual reaction.

Similar to "Reactant 1 Ro3 Filter" metanode first member of the "Reactant 2 Ro3 Filter" metanode is "Reactant 2 Functional Group Filter" node. Molecules containing more than one primary amine functional group were filtered at this node, and the number decreased from 2340 to 2330. The second member of the "Reactant 2 Ro3 Filter" metanode is "Reactant 2 Structure Normalisation" node as well. It was observed that 1733 of the 2330 aromatic primary amines entering this node passed the filter and 597 failed. It was determined that nearly all of the failed molecules had more than one fragment. The SlogP, AMW, TPSA, NumLipinskiHBA, NumLipinskiHBD and NumRotatableBonds values were calculated for the 1733 aromatic primary amines passing the first two filters. After the Ro3 filter was applied to 1733 aromatic primary amines, the number of molecules dropped to 646. The minimum and maximum values of the parameters before and after the Ro3 filter are shown in Table 3. The remaining 646 aromatic primary amines transferred to the section 2 to be used in virtual reaction.

Considering the number of acyl chloride (182) and aromatic primary amines (646) passing through the first filters, it was seen that we had a chance to obtain 117572 lead like product molecules combinatorially. As a matter of fact, "Reaction Node", the first member of the second section, has produced the number of product molecules we expected. It was checked whether the produced product molecules contain duplicate molecules and eight molecules were determined. After removing the duplicated molecules, the product molecule number dropped to 117568. After the codes of the product molecules were produced and their two-dimensional structures were made uniformed, they were transferred to the third part of the workflow where the medicinal chemistry filters would be applied.

Table 2. The upper and lower limits of parameters of acyl chlorides used in the example workflow. Number of the molecules written in parenthesis

	Before Ro3 filter applied (287)		Before Ro3 filter applied (182)	
	Lower limit	Upper limit	Lower limit	Upper limit
logP	-0.075	8.574	-0.075	2.991
TPSA (Å)	17.07	103.35	17.07	51.21
AMW (g/mol - Da)	78.498	510.571	78.498	284.455
NumLipinskiHBA	1	7	1	3
NumLipinskiHBD	0	1	0	1
NumRotatableBonds	0	20	0	3

Table 3. The upper and lower limits of parameters of aromatic primary amines used in the example workflow. Number of the molecules written in paranthesis

	Before Ro3 filter applied (1733)		Before Ro3 filter applied (646)	
	Lower limit	Upper limit	Lower limit	Upper limit
logP	-5.622	8.073	-0.008	2.997
TPSA (Å)	26.02	702.02	26.02	59.14
AMW (g/mol - Da)	83.094	1620.693	83.094	298.909
NumLipinskiHBA	1	43	1	3
NumLipinskiHBD	2	25	2	3
NumRotatableBonds	0	40	0	3

In the first step of the third section, the SlogP, AMW, TPSA, NumLipinskiHBA, NumLipinskiHBD and NumRotatableBonds and FractionCsp3 parameters of the product molecules were calculated. After the oral bioavailability and drug-likeness filters which we have adapted from the SwissAdme server that we mentioned in the Method section above, the number of product molecules decreased from 117568 to 25256 (4.65 fold enrichment), and after PAINS and BRENK filters to 13121 (8.96 fold enrichment). After the CYP filter, this number dropped to 565 (208.08 fold enrichment). The minimum and maximum values of the parameters of all product molecules were shown in the Table 4 and their radar chart in Fig. 5. After optimizing the three-dimensional structures of the remaining 565 product molecules, docking was carried out against Adenosine A2a receptor. Adenosine A2a receptor was chosen randomly because it was the first target listed in the DUD.E database to test the workflow.[44] As a result of the docking process, the best bonding energy was calculated as -9.7 (Reaction_Example_R1_241_R2_1308) and the worst binding energy was calculated as -4.4 kcal/mol (Reaction_Example_2_R1_99_R2_2283).
Binding poses of these molecules are shown in Fig. 6. 13 product molecules were identified with a calculated binding energy equal to or better than -9.0 kcal / mol. The top three structures and properties of these molecules are shown in Table 5. It has been observed that 10 of the mentioned molecules are derived from R1_241 and 3 of them are derived from R1_197 (Fig. 7).

It was aimed to determine whether the 13 molecules we selected were supplied by any vendor or not by searching the MolPort database (Similarity threshold wad set to 1). Only the record for the product molecules with the code Reaction_Example_R1_197_R2_1052 and Reaction_Example_R110_197_R2_78 was found. Also, it was examined whether there are bioactivity record of the selected molecules in the CHEMBL database and no record was found for any of them (Similarity> = 100%). In addition to these two searches, the SureCHEMBL patent database was searched for selected molecules and records were found for Reaction_Example_R1_197_R2_1052 and Reaction_Example_R110_197_R2_78 molecules or their highly similar structures. This has shown that the 11 molecules we chose in the last step are highly specific and not previously studied molecules.

Table 4. The upper and lower limits of parameters of after each medicinal chemistry filters applied in section 3. Number of the molecules written in parenthesis

	All product molecules (117568)		Product molecules after oral bioavailability filter applied (25256)		product molecules after PAINS and BRENKfiltersapplied (13121)		Product molecules after CYPfilters applied (565)	
	Lower limit	Upper limit	Lower limit	Upper limit	Lower limit	Upper limit	Lower limit	Upper limit
logP	-0.479	5.593	-0.468	5.000	-0.089	5.000	-0.089	4.510
TPSA (Å)	29.10	96.36	29.10	93.20	29.10	93.20	29.10	89.02
AMW (g/mol)	125.131	546.903	151.169	496.794	151.169	469.055	151.169	371.360
NumLipinskiHBA	2	6	2	6	2	6	2	6
NumLipinskiHBD	1	3	1	2	1	2	1	2
NumRotatableBonds	1	7	1	7	1	7	1	6
FractionCSP3	0.000	0.812	0.263	0.812	0.263	0.812	0.263	0.750

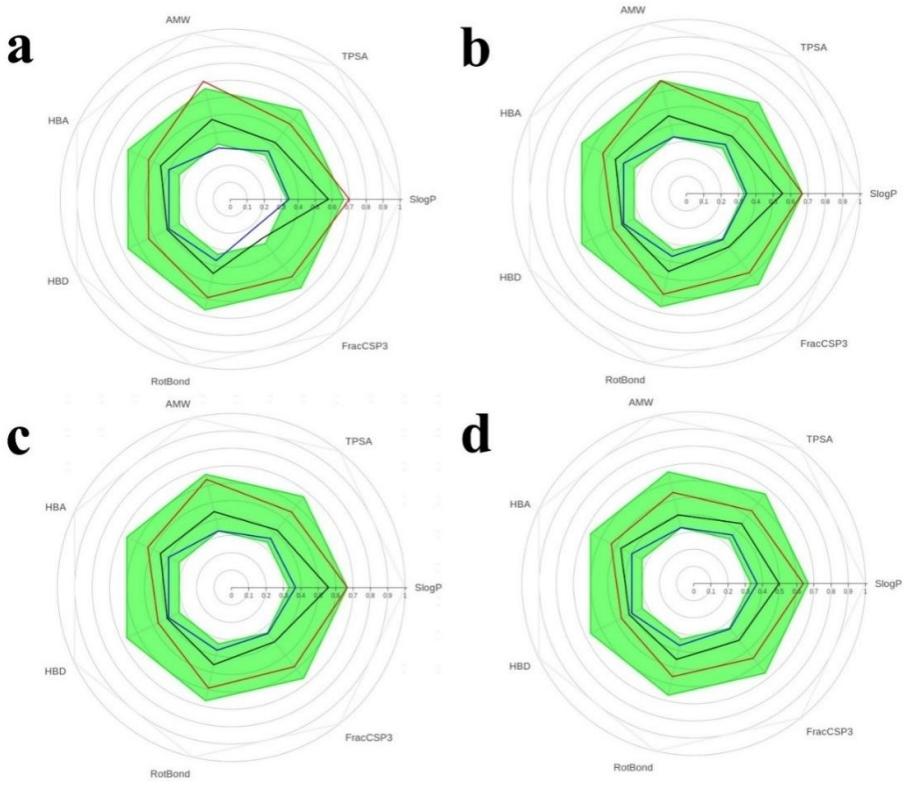


Fig. 5. Radar plots of the physicochemical parameters of product molecules a) all product molecules generated b) drug-likeness filtered product molecules, c) drug-likeness / PAINS / BRENK filtered product molecules d) selected (drug-likeness / PAINS / BRENK / CYP isoenzyme inhibitor filtered) product molecules (Blue line: minimums of normalized physicochemical property values of all product molecules, black line: averages of normalized physicochemical property values of all product molecules, red line: maximums of normalized physicochemical property values of all product molecules, green area: Suitable physicochemical space for oral bioavailability)

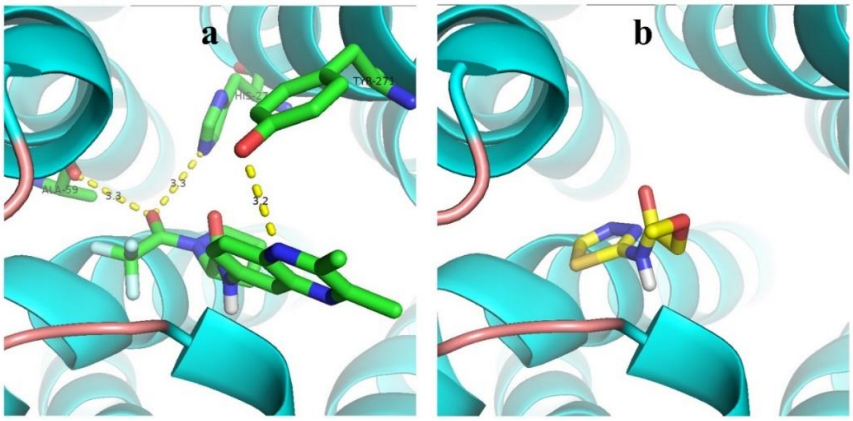


Fig. 6. Binding poses of product molecules a) Reaction_example_R1_241_R2_1308 (-9.7 kcal / mol) and b) Reaction_example_2_R1_99_R2_2283 (-4.4 kcal / mol)

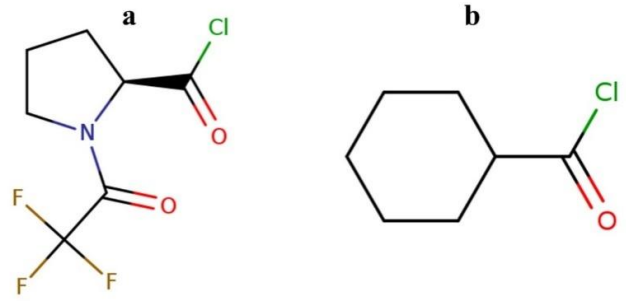


Fig. 7. Structures of reactant a) R1_241 - (2S)-1-(trifluoroacetyl)-2-pyrrolidinecarbonyl chloride - (Cas No: 36724-68-2) and b) R1_197 - cyclohexanecarbonyl chloride - (Cas No: 2719-27-9)

As we mentioned above, we created some of the filters we use in our workflow by taking SwissAdme server as an example. We have reviewed the ADME properties of the final product molecules using the SwissAdme server to create a consensus. When the data obtained were examined, it was observed that all of the remaining 11 molecules fit the oral bioavailability and drug-likeness filters determined by the server. However, it was also observed that the probability of being a CYP enzyme inhibitor was evaluated differently from our model. In order to make a consensus between the results obtained from the server and the results produced by the workflow, molecules that were not marked as inhibitors for all CYP isoenzymes were selected. At the end of this process, it was determined that Reaction_Example_R1_241_R2_910, Reaction_Example_R1_241_R2_931 and Reaction_Example_R1_241_R2_1739 molecules were not marked as inhibitors for any CYP isoenzymes. All the processes performed show that the molecules coded as eaction_Example_R1_241_R2_910, Reaction_Example_R1_241_R2_931 and Reaction_Example_R1_241_R2_1739 (Fig. 8) are suitable lead molecules that can be used in the processes of Adenosine A2a receptor ligand development studies.

When the reagent code generation nodes we performed at the beginning of our workflow are examined, it was determined that the molecules required to synthesize three molecules are R1_241: (2S) -1- (trifluoroacetyl) -2-pyrrolidinecarbonyl chloride (CAS No: 36724-68-2), R2_910: 2-amino-N-isopropylbenzamide (CAS No: 30391-89-0), R2_931: N- (2-aminophenyl) acetamide (CAS No: 34801-09-7), R2_1739: 4- (4,5-dihydro-1H-imidazol-2-yl) It has (CAS No: 61033-71-4).

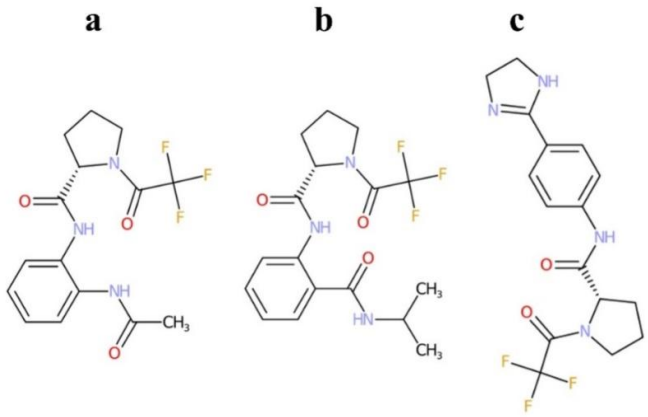
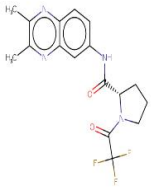
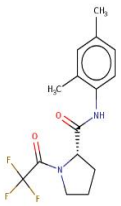
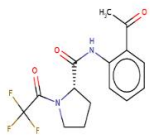


Fig. 8. Structure of selected molecules a) Reaction_Example_R1_241_R2_910, b) Reaction_example_R1_241_R2_931 and c) Reaction_example_R1_241_R2_1739 at the end of the all analysis

Table 5. Structures and properties of three products molecules that have better binding energy than selected threshold value of -9.0 kcal / mol

Product_id	Structure of product molecule	logP	TPSA (Å)	AMW (g/mol, Da)	HBA	HBD	NRB	FCsp3	Dock_avg (-kcal/mol)	Dock_stdDev	RMSD	RMSD_StdDev	LE (- kcal / mol / HA)
Reaction_example_R1_241_R2_1308		2.74	75.19	366.34	6	1	2	0.41	-9.7	0	0.01	0.01	-0.37
Reaction_example_R1_241_R2_1020		2.80	49.41	314.31	4	1	2	0.47	-9.3	0	0.02	0.01	-0.42
Reaction_example_R1_241_R2_1089		2.38	66.48	328.29	5	1	3	0.40	-9.3	0	0.52	0.43	-0.40

4. DISCUSSION

In this study, we focused on developing a KNIME-based workflow that can be used in bioactive or drug molecule development processes, which is simple to use and suitable for customization. After determining the protein target that the end user wants to work with, making the necessary adjustments in the workflow, we aimed to make the whole process proceed on its own with just one click. The workflow is divided into six sections and the tasks and outputs of each are explained. The workflow we have created can be used as a whole from the start to finish, and can be used as separate parts as long as the inputs are compatible with sections or metanodes. In the docking process, which is the structure-based virtual screening method used in the fourth part, AutoDockVina software has been chosen in terms of speed and ease of use. The script used in the relevant metanode is in a mono block structure and has been prepared for a workstation with Vina and Pymol installed in the Linux system. However, with simple changes in the script, it can be used in workstations using different operating systems such as Windows and Mac.

Since the interactive visualization components used in the fifth part work with the website logic, they can be displayed on remote screens with appropriate adjustments. The data storage metanodes we mentioned in the sixth section are important in terms of compiling and storing the results obtained. When the end users want to make changes to the file names, they can change the patterns we recommend in our workflow by making the necessary changes in the "String Manipulation" nodes.
The use of open source software in our workflow is also an important feature. In this way, we think that the use of workflow will be an advantage, especially since it will not bring a financial burden to academic groups. At this point, it will be beneficial for the end user to review the license terms while making the necessary installations and adjustments.

We have managed to produce a total of 117572 product molecules after performing the necessary filtrations with 317 acyl chloride and 2357 aromatic primary amine, which we chose to test the workflow, thanks to the applied medicinal chemistry filters 565 of them were found to be bioactive or capable of being used as bioactive / drug molecules. At the end of the docking virtual screening process, it was determined that 13 molecules could have better binding energy than -9.0 kcal / mol.

Whether there are companies producing these molecules, whether bioactivity screens have been carried out before and whether they are the subject of patents were examined and no records were found in the databases (MolPort, CHEMBL, SureCHEMBL) about 11 of these 13 molecules. Eleven were repeatedly uploaded to the SwissAdme server and three of them were found to be unlikely to be inhibitors of CYP isoenzymes by this server. Until the end of the process where we identified the possible bioactive 13 molecules, the processes were carried out with a single click without any intervention. This process took approximately three hours at the workstation we mentioned in the material section.

Some aspects of our workflow are open to improvement. For example, the metanode, in which CYP isoenzyme inhibitors are tried to be identified, takes almost half of the entire execution time. We continue to work on to improve the node where CYP inhibitors are tried to be determined with a more effective model. Also, the component in which all product molecules obtained is tried to be visualized has a long execution time. This situation is related to the creation of a table of many molecules and can be accelerated by reducing the number of molecules to be shown in the table.

VSPrep is one of the good examples to use KNIME software and plugins to used to generate and standardize virtual libraries. [32] Gally and his colleagues created a workflow named as VSPrep to prepare small molecules for virtual screening processes. In the first step of the workflow, the molecules were standardized, then the duplicate molecules were removed. After some filtration processes related to stereochemistry and tautomerism, conformers of the molecules were generated. The conformers created were recorded to be used in virtual screening processes. VSPrep and the workflow we discuss in our article have some commonalities like creating codes for molecules, generating three dimensional structure and saving structures for further studies. There are some differences as well as similar aspects. For example, while we focused on building a personal library in our workflow, VSPrep focused on curation of existing libraries. TeachOpenCADD KNIME workflow is a very comprehensive workflow aimed at teaching

computer-aided design work to beginners. [30] TeachOpenCADD consists of eight sub-workflows where data is transferred to the workflow, various filters are applied, the similarities of the molecules of interest are examined and similar ones are searched and possible activities are evaluated. Unlike our study, TeachOpenCADD uses ligand-based screening methods to identify new potential bioactive molecules. One of the studies recently performed using KNIME was carried out by Tuerkova and Zdrazil. [45] The authors examined the molecules that can be used in the treatment of rare diseases and COVID-19 with the drug reproposing approach using the ligand-based screening method. They created an integrated virtual screening workflow on KNIME by using the Application Programming Interfaces (API) services of the servers and other plugins in their workflows. In our study, no API was used, and our update plans include changes to screen the databases of patents, bioactivity and suppliers for possible active molecules or raw materials using existing APIs. Gonzalez and colleagues have described a workflow to obtain a series of lactam derivative molecules using KNIME software and plugins. [46] As in our study, possible building block molecules were loaded into the workflow, filtered according to the Ro3 rule, and then possible molecules were obtained. Then, possible intermolecular cyclization schemes were defined and applied to the obtained molecules. Although a single-step reaction is defined in our study, this number can be easily increased by increasing the number of reassembly nodes used or by using Python scripts.

5. CONCLUSION

The production of a large number of molecules, including various physicochemical parameters, using a simple reaction scheme and reactants in such a short time, and the fact that they have been recorded in a structure that can be transferred to personal databases, shows the success of the workflow. With the help of our workflow and the encouraging results we have obtained from the sample library we generated, it will be possible for smaller companies, enterprises and academic groups to apply the method we mentioned in the introduction section of this paper about generation of novel bioactive or drug molecule studies that have been used by pharmaceutical companies and some academic groups.

DISCLAIMER

The products used for this research are commonly and predominantly use products in our area of research and country. There is absolutely no conflict of interest between the authors and producers of the products because we do not intend to use these products as an avenue for any litigation but for the advancement of knowledge. Also, the research was not funded by the producing company rather it was funded by personal efforts of the authors.

CONSENT

It's not applicable.

ETHICAL APPROVAL

It's not applicable.

ACKNOWLEDGEMENT

Readers who want to reach or use the workflow mentioned in our article are kindly requested to contact the author.

COMPETING INTERESTS

Author has declared that no competing interests exist.

REFERENCES

1. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, et al. The ChEMBL bioactivity database: an update. Nucleic Acids Res. 2014;42(D1):D1083-D90.
2. Li Q, Cheng T, Wang Y, Bryant SH. PubChem as a public resource for drug discovery. Drug Discov Today. 2010;15(23-24):1052-7.
3. Sterling T, Irwin JJ. ZINC 15–ligand discovery for everyone. J Chem Inf Model. 2015;55(11):2324-37.
4. Pence HE, Williams A. Chemspider: An online chemical ınformation resource. J Chem Educ. 2010;87(11):1123-4.
5. Huggins DJ, Venkitaraman AR, Spring DR. Rational methods for the selection of diverse screening compounds. ACS Chem Biol. 2011;6(3):208-17.
6. Voigt JH, Bienfait B, Wang S, Nicklaus MC. Comparison of the NCI open database with seven large chemical

structural databases. J Chem Inf Comput Sci. 2001;41(3):702-12.
7. Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: A molecular modeling perspective. Med Res Rev. 1996;16(1):3-50.
8. Lewell XQ, Judd DB, Watson SP, Hann MM. Recap retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. J Chem Inf Comput Sci. 1998;38(3):511-22.
9. Walters WP. Virtual chemical libraries: miniperspective. J Med Chem. 2018;62(3):1116-24.
10. Hu Q, Peng Z, Sutton SC, Na J, Kostrowicki J, Yang B et al. Pfizer Global Virtual Library (PGVL): a chemistry design tool powered by experimentally validated parallel synthesis information. ACS Comb Sci. 2012;14(11):579-89.
11. Nicolaou CA, Watson IA, Hu H, Wang J. The proximal lilly collection: Mapping, exploring and exploiting feasible chemical space. J Chem Inf Model. 2016;56(7):1253-66.
12. Chevillard F, Kolb P. SCUBIDOO: A large yet screenable and easily searchable database of computationally created chemical compounds optimized toward high likelihood of synthetic tractability. J Chem Inf Model. 2015;55(9):1824-35.
13. Available:https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html
14. Pirok G, Máté N, Varga J, Szegezdi J, Vargyas M, Dóránt S, et al. Making “real” molecules in virtual space. J Chem Inf Model. 2006;46(2):563-8.
15. Ghosh S, Nie A, An J, Huang Z. Structure-based virtual screening of chemical libraries for drug discovery. Curr Opin Chem Biol. 2006;10(3):194-202.
16. Bajorath J. Integration of virtual and high-throughput screening. Nat Rev Drug Discov. 2002;1(11):882-94.
17. H Ma X, Zhu F, Liu X, Shi Z, X Zhang J, Y Yang S et al. Virtual screening methods as tools for drug lead discovery from large chemical libraries. Curr Med Chem. 2012;19(32):5562-71.
18. Meng X-Y, Zhang H-X, Mezei M, Cui M. Molecular docking: A powerful approach for structure-based drug discovery. Curr Comput Aided Drug Des. 2011;7(2):146-57.
19. Pagadala NS, Syed K, Tuszynski J. Software for molecular docking: A review. Biophys Rev. 2017;9(2):91-102.
20. Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. J Comput Chem. 2010;31(2):455-61.
21. Uddin MZ, Li X, Joo H, Tsai J, Wrischnik L, Jasti B. Rational design of peptide ligands based on knob–socket protein packing model using CD13 as a prototype receptor. ACS Omega. 2019;4(3):5126-36.
22. Perryman AL, Yu W, Wang X, Ekins S, Forli S, Li S-G et al. A virtual screen discovers novel, fragment-sized inhibitors of mycobacterium tuberculosis InhA. J Chem Inf Model. 2015;55(3):645-59.
23. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME-the Konstanz information miner: Version 2.0 and beyond. SIGKDD Explor. 2009;11(1):26-31.
24. Fillbrunn A, Dietz C, Pfeuffer J, Rahn R, Landrum GA, Berthold MR. KNIME for reproducible cross-domain analysis of life science data. J Biotechnol. 2017;261:149-56.
25. Jagla B, Wiswedel B, Coppée J-Y. Extending KNIME for next-generation sequencing data analysis. Bioinformatics. 2011;27(20):2907-9.
26. Verhoeven A, Giera M, Mayboroda OA. KIMBLE: A versatile visual NMR metabolomics workbench in KNIME. Anal Chim Acta. 2018;1044:66-76.
27. Falcón-Cano G, Molina C, Cabrera-Pérez MA. ADME prediction with KNIME: In silico aqueous solubility consensus model based on supervised recursive random forest approaches. ADMET and DMPK. 2020;8(3):251-73.
28. Kausar S, Falcao AO. An automated framework for QSAR model building. J Cheminform. 2018;10(1):1.
29. Stöter M, Niederlein A, Barsacchi R, Meyenhofer F, Brandl H, Bickle M. CellProfiler and KNIME: Open source tools for high content screening. Target Identification and Validation in Drug Discovery: Springer. 2013;105-22.
30. Sydow D, Wichmann M, Rodríguez-Guerra J, Goldmann D, Landrum G, Volkamer A. TeachopenCADD-KNIME: A teaching platform for computer-aided drug design using KNIME workflows. J Chem Inf Model. 2019;59(10):4083-6.

31. Nicola G, Berthold MR, Hedrick MP, Gilson MK. Connecting proteins with drug-like compounds: Open source drug discovery workflows with BindingDB and KNIME. Database;2015.
32. Gally JM, Bourg S, Do QT, Aci-Sèche S, Bonnet P. VSPrep: A general KNIME workflow for the preparation of molecules for virtual screening. Mol Inform. 2017;36(10):1700023.
33. Congreve M, Carr R, Murray C, Jhoti H. A'rule of three'for fragment-based lead discovery? Drug Discov Today. 2003;19(8):876-7.
34. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 1997;23(1-3):3-25.
35. Veber DF, Johnson SR, Cheng H-Y, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem. 2002;45(12):2615-23.
36. Egan WJ, Merz KM, Baldwin JJ. Prediction of drug absorption using multivariate statistics. J Med Chem. 2000;43(21):3867-77.
37. Ghose AK, Viswanadhan VN, Wendoloski JJ. 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(1):55-68.
38. Muegge I, Heald SL, Brittelli D. Simple selection criteria for drug-like chemical matter. J Med Chem. 2001;44(12):1841-6.
39. Daina A, Michielin O, Zoete V. SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep. 2017;7:42717.
40. Baell JB, Holloway GA. 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(7):2719-40.
41. Brenk R, Schipani A, James D, Krasowski A, Gilbert IH, Frearson J, et al. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. ChemMedChem. 2008;3(3):435.
42. Rostkowski M, Spjuth O, Rydberg P. WhichCyp: Prediction of cytochromes P450 inhibition. Bioinformatics. 2013;29(16):2051-2.
43. Hopkins AL, Groom CR, Alex A. Ligand efficiency: A useful metric for lead selection. Drug Discov Today. 2004;9(10):430-1.
44. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J Med Chem. 2012;55(14):6582-94.
45. Tuerkova A, Zdrazil BJJoc. A ligand-based computational drug repurposing pipeline using KNIME and Programmatic Data Access: Case studies for rare diseases and COVID-19. J. Cheminformatics. 2020;12(1):1-20.
46. Saldívar-González FI, Huerta-García CS, Medina-Franco JLJJoc. Chemoinformatics-based enumeration of chemical libraries: A tutorial. J. Cheminformatics. 2020;12(1):1-25.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/65682