doi.org/10.1002/minf.202100139

PrepFlow: A Toolkit for Chemical Library Preparation and Management for Virtual Screening

Marion Sisquellas[a] and Marco Cecchini*[a]

Abstract: In the era of big data in Chemistry, the need for automated tools for virtual screening is compelling. Here, we present PrepFlow a toolkit for chemical library preparation and management. Starting from a list of compounds in SMILES or 2D molecular format, PrepFlow outputs a set of 3D molecular structures ready for use in subsequent drug discovery projects. Our development stands out for speed and robustness of execution, the efficient exploitation of HPC resources, and the implementation of an archiving strategy to save computer time, storage, and human intervention. Using a random selection of 600 compounds from available drug banks, we show that the preparation time per ligand on a desktop computer is 6.6 s. Thanks to these performances and the automatic parallelization on HPC, a chemical library of the size of ChEMBL (2 M) was prepared in around 3 days on a computer cluster. PrepFlow is freely distributed at the following link: https://ifm.chimie.unistra.fr/prepflow.

Keywords: drug discovery · chemoinformatics · molecular databases · library preparation · workflow · virtual screening

Innovation costs in the Pharma industry have tremendously increased over the last decades and recently approached 2 billion US dollars with more than 10 years development per molecular entity approved by the FDA.[1] To reduce the attrition rate, virtual high-throughput screening has become routine particularly in the early stage of drug discovery.[2] In fact, an effective filtering of the unfitting compounds reduces the number of bioactivity tests to be performed, alleviating the costs of protein expression and purification, chemical synthesis, and human time. Recently, the integration of molecular modeling techniques with an array of experimental tools led to the development of a novel class of potent and selective androgen receptor inhibitors with an unprecedented mode of action.[3]

A critical step towards successful drug discovery is the selection of the most appropriate chemical library to be screened. At present, several collections are available that can be grossly divided into four families.[4] Commercial libraries like ChemDiv,[5] Enamine[6] or Maybridge[7] offer large amounts of readily accessible compounds that often suffer from low chemical diversity. Academic libraries like the Chimiothèque Nationale from the French National Research Center[8] (70 k) or OPENSCREEN (100 K) from the EU[9] are considerably smaller but they are chemically more diverse and can be freely distributed to not-for-profit institutions. A third group gathers collections of previously approved drugs such as DrugBank,[10] Sigma-Aldrich,[11] or Prestwick Chemical Libraries,[12] which are useful for drug repurposing, and/or databases of known bioactive compounds like ChEMBL[13] (2.1 M). Last, virtual collections of chemical compounds that are synthetically accessible on-demand such as the REAL database of Enamine (15 M) or the GDB databases (24 M–370 M) complete this classification. Considering the size and the number of the currently available chemical databases, there is an obvious need for fast and automated tools for library processing and maintenance in preparation for virtual screening. In addition, recent screening campaigns demonstrated that the larger the number of compounds analyzed in silico, the higher the quality of the identified hits and the lower the false-positive rate, which calls for screening ultra-large chemical libraries to increase the success rate[14][15].

To deal with large chemical databases that grow by 25–30 % every year,[4] preparation steps like standardization of ligands, removal of salts and co-solvents, generation of possible tautomers, calculation of protonation states at the experimental pH, enumeration of stereoisomers, and the determination of 3D coordinates for various conformers must be carried out rapidly and automatically. For this purpose, integrated platforms and workflows have been developed. Some of them like the popular Pipeline Pilot[16] (Accelrys) or LigPrep[17] (Schrödinger) are commercial software. Others are freely distributed to academics and can be combined with available toolkits to improve performances and/or adapt to the user needs. A recent example of open-source software for the preparation of chemical libraries is VSPrep,[18] which is implemented within the pipelining framework KNIME.[19] The integration in KNIME provides straightforward access to a number of bioinformatics and chemo-informatics tools including the Rational Discovery Toolkit (RDKit)[20] and ChemAxon *cxcalc*[21] and allows for

[a] *M. Sisquellas, M. Cecchini*
Institut de Chimie de Strasbourg, UMR7177, CNRS, Université de Strasbourg, F-67083 Strasbourg Cedex, France
E-mail: mcecchini@unistra.fr

Supporting information for this article is available on the WWW under https://doi.org/10.1002/minf.202100139

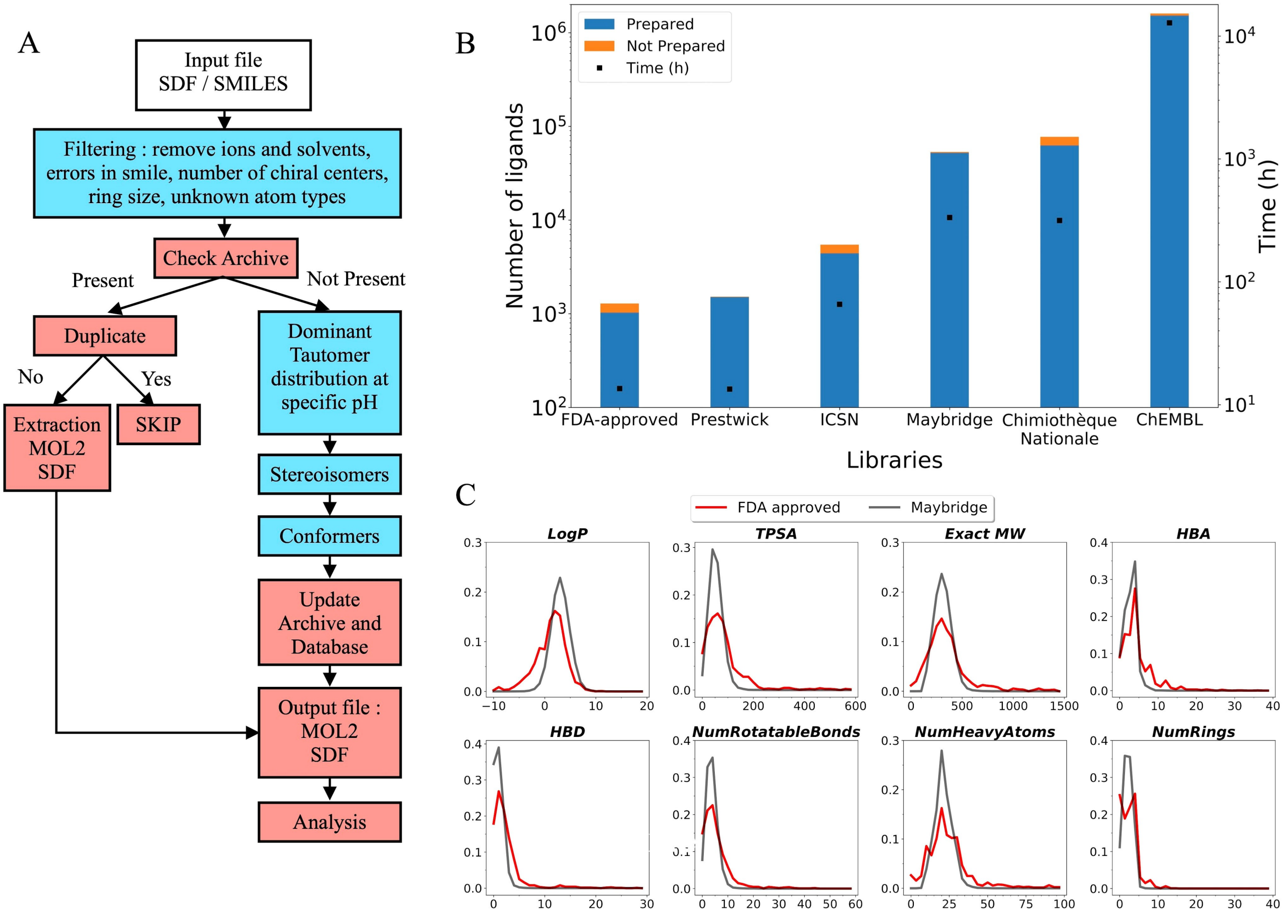


Figure 1. Chemical library preparation and profiling by PrepFlow. (A) Workflow of chemical library preparation by PrepFlow. The input library in 2D format (white box) undergoes several filtering and preparation steps to generate a set of 3D molecular structures. Light blue boxes highlight preparation steps powered by ChemAxon software that are common to existing workflows for library preparation like VSPrep[18] or VFLP.[15] Red boxes correspond to preparation steps specific to PrepFlow that are used to implement an original archiving strategy and rapid chemical profiling. (B) Preparation of six drug banks by PrepFlow. The plot shows the size, the fraction of successfully prepared compounds (blue), and the total preparation time (black squares). (C) Chemical profiling and drug-likeness. The statistical distributions of eight molecular properties, i.e. LogP, the topological polar surface area (TPSA), the exact molecular weight (MW), the number of hydrogen bond acceptors (HBA) or donors (HBD), the number of rotatable bonds, the number of heavy atoms, and the number of rings, for the Maybridge commercial library are compared with those obtained from 1285 compounds from PubChem with the FDA-approved flag.

local installations of the software, which grants for comfort and flexibility of execution while preserving the confidentiality of information. Another recent example is VirtualFlow[15] that was developed as a highly automated and versatile open-source platform for screening of big data. Made of two application tools written in Bash, ligand preparation is accomplished by VFLP (Virtual Flow Ligand Preparation) that was designed to leverage parallelization in a HPC environment. Particularly original is the attention that was paid to handle errors during execution, i.e. unexpected terminations, input/output problems, etc., and the effort to grant versatility on computer clusters operated with different job schedulers. Albeit the ligand preparation steps in VFLP are not novel, the improved performances in automation, distribution and execution opened to the preparation and docking of a chemical library of 1.4 billion compounds, yielding a set of structurally diverse inhibitors of protein-protein interactions with sub-micromolar binding affinities.[15]

In this context, we have developed PrepFlow a new toolkit for chemical-library preparation and management for virtual screening. Starting from a chemical library in SMILES or 2D molecular format, PrepFlow outputs a set of 3D molecular structures that are ready for use in subsequent drug discovery projects. Its workflow is composed of several steps that include: standardization; filtering; tautomer, stereoisomer and conformer enumeration; and the generation of 3D coordinates with partial atomic charges (Figure 1A). PrepFlow is coded in Python3[22] and requires the only addition of three libraries, i.e. NumPy,[23] pandas[24] and RDKit[20] which are freely distributed. To carry out the preparation steps, three tools from ChemAxon Marvin[21,25] are invoked, i.e. *cxcalc*, *standardize* and *molconvert*.

Several preparation steps in PrepFlow are in common with VSPrep.[18] Chemical information is inputted in standard molecular formats, e.g. SMILES[26] or SDF,[27] which are converted into unique SMILES[28] using ChemAxon *molconvert*.[25] After filtering to remove ions and/or solvent molecules and exclude compounds with too many unassigned chiral centers (>3), the dominant tautomeric forms at the given pH as well as possible stereoisomers and conformers are enumerated. 3D atomic coordinates and partial atomic charges are then generated using various ChemAxon tools; see *Methods* for details. Importantly and unlike VSPrep,[18] PrepFlow implements an archiving strategy that avoids preparing duplicates. Molecular archiving is materialized via the combination of a single file containing the list of previously prepared compounds, i.e. the ARCHIVE, and a storage unit containing 3D molecular structures with topological connectivity, i.e. the DATABASE, which are readily accessible by the user. Every time a new library is processed, PrepFlow checks whether any new chemical entity in input is part of the ARCHIVE by rapid InChiKey comparison and if so, it skips the preparation steps (Figure 1A, red boxes). Since new compounds are constantly deposited in both public and proprietary chemical databases, this feature allows for efficient incremental updates while saving computer resources and human intervention.

Another useful feature of PrepFlow is that library preparation can be efficiently distributed over HPC resources using the same archiving strategy. In the current implementation, the chemical library in input is copied to the master node of the HPC cluster along with the ARCHIVE. The chemical collection is pruned to remove duplicates, i.e. molecules that are present in the ARCHIVE, and split over the available compute resources. For this purpose, using an appropriate user-supplied header for the job scheduler operating on the master node (e.g. slurm), PrepFlow prepares and submits automatically batches of independent jobs that materialize an efficient library preparation in parallel; see SI for an example of slurm header. Once all compounds are processed, PrepFlow collects information from the individual nodes (i.e. 3D coordinates in MOL2/SDF formats) and updates both the ARCHIVE and the DATABASE.

Last, using ChemAxon *cxcalc* PrepFlow provides a synthetic overview of the chemical information stored in the library (i.e. chemical profiling) by displaying the statistical distributions of key molecular properties like LogP, the topological polar surface area (TPSA), the exact molecular weight (MW), the number of hydrogen bond acceptors (HBA) or donors (HBD), the number of rotatable bonds, the number of heavy atoms, and the number of rings. These features are calculated at no additional cost and are relevant to quantify the drug-likeness of the compounds stored in the library, e.g. by comparing with distributions obtained from a database of approved drugs (see below).

To demonstrate the use of PrepFlow, six different chemical libraries have been prepared (Figure 1B). The queried datasets include: two commercial libraries, i.e. the Prestwick Chemical Library[12] (1.5k) and Maybridge[7] (50k); two academic libraries, i.e. the drug bank of the Institut de Chimie de Substances Naturelles ICSN (5k) and the French national library "Chimiothèque Nationale"[8] (70 k); the large and popular database of bioactive compounds with drug-like properties ChEMBL[13] (2 M); and a small collection of FDA-approved drugs (1.3 k). These libraries were chosen because of their size, chemical diversity, and relevance for drug discovery. The exploration of increasingly larger fractions of the chemical space helped PrepFlow becoming general, efficient and error proof. The latter aspect was greatly improved by the automatic renaming of ligands with their InChiKey identifier as well as specialized filters to resist compounds producing errors or whose preparation was exceedingly long; see *Methods*. The size of each library, the fraction of successfully prepared compounds, and the preparation time by PrepFlow are shown in Figure 1B. These results show that PrepFlow was successful in preparing >80% of each library independently of size and chemical complexity; see SI for details on failures. To benchmark the performances of PrepFlow, a dataset of 600 compounds was built by random extraction of 100 compounds per library and prepared on three different computing architectures, i.e. one desktop computer (*Lab computer*, Intel(R) Core(TM) i7-4770 K CPU @ 3.50GHz 8 cores) and two HPC centers, the Mésocentre de Calcul of the University of Strasbourg (*HPC1*, Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz 16 cores) and the French national supercomputer GENCI-IDRIS (*HPC2*, Intel(R) Xeon(R) Gold 6248 CPU @ 2.50 GHz 40 cores). The results in Table 1 show that the average preparation time per ligand on two CPU cores on the fastest machine (i.e. *Lab computer* @ 3.50 GHz) was 6.64 s, which grants for a production rate of >13 k compounds per day. When running on HPC resources, the preparation time per ligand increased by a factor of 3–4 dependently on the architecture (Table 1) due to the lower frequency of the processors (~2.50 GHz), the older manufacturing generation, and possibly communication bottlenecks between the master node and the slaves. Nonetheless, the exploitation of compute resources in parallel allowed the preparation of a chemical library of the size of ChEMBL (2,1 M) in 6.6 days on a mid-size commodity cluster (e.g. *HPC1*) and 2.8 days on a supercomputer (e.g. *HPC2*); see Table 1. We note that the same preparation on a faster and newer-generation desktop computer (8 cores) would have taken 37.3 days, i.e. 1.5 months. Hence, the automatic parallelization by PrepFlow on HPC resources along with the implementation of tools for handling errors that allow for robust exploitation of parallel computing resources opens to the treatment of big data in Chemistry.

Table 1. Benchmark of PrepFlow. Computational performances were evaluated by preparing a library of 600 compounds on three different computing architectures; i.e. one desktop computer (*Lab computer*) embedded with Intel(R) Core(TM) i7-4770 K CPU @ 3.50GHz with 8 cores, the Mésocentre de Calcul in Strasbourg (*HPC1*) powered by Intel(R) Xeon(R) CPUs E5-2640 v3 @ 2.60GHz with 16 cores, and the French national supercomputer GENCI-IDRIS (*HPC2*) featuring Intel(R) Xeon(R) Gold 6248 CPUs @ 2.50GHz with 40 cores. The preparation time per ligand was extracted from the median of the distribution; see SI for details. To illustrate the advantage of library preparation on distributed computing, the preparation time per ligand on the three computing architectures is compared. The analysis shows that despite lower absolute performances, the exploitation of parallel computing provides a significant advantage and opens to large-scale chemical library handling for drug discovery.

	Preparation of one ligand		ChEMBL	
	Reference time (s) (1 CPU core)	Optimized time (s) (2 CPU cores)	Maximal number of jobs running in parallel (2 CPU cores)	Theoretical preparation time (days)
Lab computer	10.47	6.64	4	37.3
HPC1	26.99	23.99	81	6.65
HPC2	22.93	15.07	120	2.82

To increase the success rate of drug discovery, it is important to assess the drug-likeness of the library to be screened. For this purpose, rapid chemical profiling by PrepFlow is useful to compare the chemical content of a given library with that of a database of approved drugs, e.g. by FDA. Molecular features of the commercial library Maybridge were evaluated and compared against 1285 compounds extracted from PubChem with the FDA-approved flag. Figure 1C shows the distributions of eight molecular properties that are typically used to predict drug oral bioavailability.[29] The comparison highlights similar distributions overall with a few notable exceptions. First, analysis of LogP indicates that compounds from Maybridge are significantly more lipophilic, i.e. $\mathrm{LogP} > 0$, suggesting that this library might be suboptimal for targeting solvent-exposed or polar binding sites. Second, the statistical distributions of the molecular weight and the number of hydrogen bond donors and acceptors indicate that Maybridge strictly follows the Lipinski's rule of five (i.e. $\mathrm{MW} < 500$ Da, $\mathrm{HBD} < 5$, $\mathrm{HBA} < 10$, and $\mathrm{LogP} < 5$) even more than the FDA-approved drugs, so that this collection is well suited to search for potentially improvable hits. Third, the distribution of the topological polar surface area (TPSA) indicates that the vast majority of Maybridge compounds cope with the criterion that cell permeability requires $\mathrm{TPSA} < 140$ Å^2.[30] Consistent with the analysis of LogP, this observation indicates that Maybridge is possibly too conservative on the molecular polarity as FDA-approved drugs feature a significant fraction of chemical entities with TPSA between 150–200 Å. Last, the statistical distribution of the number of rings is different in the two libraries, indicating scarceness of Maybridge compounds with a number of rings between 5 and 13. Although this comparison is not directly related to the quality of the hits one can find by screening, it provides an estimate of the probability that the identified hits will be taken further on the drug discovery process. Finally, starting with a series of known actives for a given receptor, chemical profiling by PrepFlow can be used to build targeted chemical libraries, e.g. by extracting compounds from larger databases with consistent molecular properties.

The present application note reports on the development of PrepFlow a toolkit for chemical library preparation, management, and rapid profiling. Our software stands out for the speed and robustness of execution, the efficient exploitation of HPC resources, and the implementation of an archiving strategy that saves computer resources and minimizes human intervention. PrepFlow is freely accessible to academics and not-for-profit institutions at https://ifm.chimie.unistra.fr/prepflow.

Methods

Input. A list of molecules in a structure-data file (SDF) or smile (SMILES) format must be provided as input. These files formats are widely used for importing and exporting chemical information and are the common standard to distribute academic and commercial libraries. In the initial steps of the preparation, the primary chemical entities are converted into unique SMILES using ChemAxon *molconvert*, while the secondary entities, i.e. ions and/or solvents, are removed using ChemAxon *standardize*. In addition, molecules are renamed using their InChiKey as determined by ChemAxon *molconvert* that provides a unique identifier, useful for rapid string-matching comparison when checking for duplicates. At this stage, a lookup table linking the original molecular name with the InChiKey is also generated, which is useful for compound identification and purchasing.

Filtering. In order to keep library preparation rapid and robust, a series of filters have been implemented. Specifically, chemical entities that contain unknown or exotic atoms that do not belong to the following list ['H', 'B', 'C', 'N', 'O', 'P', 'S', 'F', 'Cl', 'Br', 'I'] are discarded. Also, compounds featuring too many (>3) unassigned chiral centers, too large cyclic structures (>7-membered rings), or too many (>20) double bonds are filtered out. These filters reduce the occurrence of errors and remove excessively long time lags due to the conformational complexity of the treated compounds. All thresholds are set as defaults and can be modified by the user.

Preparation of ligands. All compounds that survive the initial filtering are converted into 3D molecular structures. For this purpose, all stereoisomers of the dominant tautomeric form(s) must be enumerated. To this aim, ChemAxon *cxcalc dominanttautomerdistribution* is used to determine the tautomers distribution at a given pH. All tautomers and protonation isomers with a predicted probability < 10 % are discarded and all stereoisomers of the dominant tautomeric forms enumerated using ChemAxon *cxcalc stereoisomers* whenever relevant, i.e. in the presence of unassigned chiral centers. Multiple 3D structures per stereoisomer are then generated using ChemAxon *cxcalc conformers* and only the coordinates corresponding to the lowest-energy conformers based on the Merck Molecular Force Field (MMFF94) are kept.

Running on HPC. When running on HPC, in addition to the files required for a standard execution, an ARCHIVE (not mandatory) and the header of the job scheduler operating on the master node (e.g. SLURM) must be supplied. On HPC, library preparation is divided into two steps: i. removing of duplicates, i.e. chemical entities that are already listed in the ARCHIVE; and ii. preparation of ligands in batches, which are distributed and run in parallel over multiple nodes. By default, PrepFlow distributes 1500 ligands per compute node, but this number is user-defined. Once all batches are complete, the output is collected, merged, and stored in the DATABASE.

Output. During library preparation, several output files are produced: i. A log file is generated at each step of the preparation, which is useful to monitor the runtime execution and spot errors; ii. A summary file displaying the command line used for the preparation, the total execution time, and the name and status (i.e. DONE/SKIP/PASS/ERROR) of all compounds is outputted at the end of the execution; iii. A series of MOL2 files containing the 3D coordinates of the dominant tautomers/conformers/stereoisomers per compound with Gasteiger partial atomic charges[31] computed by RDKit are generated; iv. A summary file providing an overview of the prepared libraries, the number of ligands per library, and the date of the last modification per library is generated/updated in the DATABASE; v. The statistical distributions of several molecular features determined by ChemAxon *cxcalc* are computed and plotted for chemical profiling.

Acknowledgements

We are grateful to Dr. Adrien Cerdan and Gilberto Pereira for a critical reading of the manuscript and Katia Galentino for invaluable help on testing and improving the functionalities of PrepFlow. This work received funding from the French National Research Agency (ANR-18-CE11-0015) and was granted access to the HPC resources of IDRIS under the allocation 2020-[076644] attributed by GENCI and the University of Strasbourg (Mésocentre de Calcul). The computing resources in Strasbourg were funded in part by the Equipex Equip@Meso project (Programme Investissements d'Avenir) and the CPER Alsacalcul/Big Data.

Conflict of Interest

None declared.

Data Availability Statement

The data that supports the findings of this study are available in the supplementary material of this article.

References

[1] J. A. DiMasi, H. G. Grabowski, R. W. Hansen, *J. Health Econ.* **2016**, *47*, 20-33.
[2] J. J. Irwin, B. K. Shoichet, *J. Med. Chem.* **2016**, *59*, 4103–4120.
[3] F. Ban, K. Dalal, H. Li, E. LeBlanc, P. S. Rennie, A. Cherkasov, *J. Chem. Inf. Model.* **2017**, *57*, 1018–1028.
[4] D. Rognan, P. Bonnet, *M S-Med. Sci.* **2014**, *30*, 1152–1160.
[5] ChemDiv – a vast collection of diverse screening compounds and small molecule libraries for accelerated discovery chemistry (https://www.chemdiv.com/).
[6] Enamine (https://enamine.net).
[7] Maybridge – Screening Compounds and Libraries for Hit Identification (www.thermofisher.com).
[8] ChemBioFrance – Infrastructure de recherche (https://chembiofrance.cn.cnrs.fr).
[9] P. Brennecke, D. Rasina, O. Aubi, K. Herzog, J. Landskron, B. Cautain, F. Vicente, J. Quintana, J. Mestres, B. Stechmann, B. Ellinger, J Brea, J. L. Kolanowski, R. Pilarski, M. Orzaez, A. Pineda-Lucena, L. Laraia, F. Nami, P. Zielenkiewicz, K. Paruch, E. Hansen, J. P. von Kries, M. Neuenschwander, E. Specker, P. Bartunek, S. Simova, Z. Lesnikowski, S. Krauss, L. Lehtiö, U. Bilitewski, M. Brönstrup, K. Taskén, A. Jirgensons, H. Lickert, M. H. Clausen, J. H. Andersen, M. J. Vicent, O. Genilloud, A. Martinez, M. Nazaré, W. Fecke, P. Gribbon, *SLAS Discov* **2019**, *24*, 398–413.
[10] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, *Nucleic Acids Res.* **2018**, *46*, D1074–D1082.
[11] Sigma-Aldrich – Library of Pharmacologically Active Compounds (https://www.sigmaaldrich.com).
[12] Prestwick Chemical Library® *Prestwick Chemical Libraries* (https://www.prestwickchemical.com).
[13] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit, A. R. Leach, *Nucleic Acids Res.* **2017**, *45*, D945–D954.

[14] J. Lyu, S. Wang, T. E. Balius, I. Singh, A. Levit, Y. S. Moroz, M. J. O'Meara, T. Che, E. Algaa, K. Tolmachova, A. A. Tolmachev, B. K. Shoichet, B. L. Roth, J. J. Irwin, *Nature* **2019**, *566*, 224–229.

[15] C. Gorgulla, A. Boeszoermenyi, Z.-F. Wang, P. D. Fischer, P. W. Coote, K. M. Padmanabha Das, Y. S. Malets, D. S. Radchenko, Y. S. Moroz, D. A. Scott, K. Fackeldey, M. Hoffmann, I. Iavniuk, G. Wagner, H. Arthanari, *Nature* **2020**, *580*, 663–668.

[16] Pipeline Pilot – BIOVIA – Dassault Systèmes® (https://www.3ds.com).

[17] Schrödinger | Schrödinger is the scientific leader in developing state-of-the-art chemical simulation software for use in pharmaceutical, biotechnology, and materials research (https://www.schrodinger.com/).

[18] J.-M. Gally, S. Bourg, Q.-T. Do, S. Aci-Sèche, P. Bonnet, *Mol. Inform.* **2017**, *36*, 1700023.

[19] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, Peter Ohl, C. Sieb, K. Thiel, B. Wiswedel, in *Data Analysis, Machine Learning and Applications,* (Eds.: C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker), Springer, Heidelberg, **2008**, pp 319–326.

[20] G. Landrum, RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling, **2006**.

[21] ChemAxon, Calculator Plugins were used for structure property prediction and calculation. Calculator, Version 19.4.0., 2019, ChemAxon, **2019** (http://www.chemaxon.com).

[22] G. Van Rossum, F. L. Drake, in *Python 3 Reference Manual*, CreateSpace, Scotts Valley, **2009**.

[23] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, *Nature* **2020**, *585*, 357–362.

[24] W. McKinney, in *Proceedings of the 9th Python in Science Conference.*, (Eds.: S. van der Walt, J. Millman), **2010**, pp. 56–61.

[25] Chemaxon, Marvin was used for drawing, and characterizing chemical structures. MarvinSketch, Version 19.4.0., ChemAxon, **2019** (http://www.chemaxon.com).

[26] D. Weininger, *J. Chem. Inf. Comp. Sci.* **1988**, *28*, 31–36.

[27] A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, J. Laufer, *J. Chem. Inf. Comp. Sci.* **1992**, *32*, 244–255.

[28] D. Weininger, A. Weininger, J. L. Weininger, *J. Chem. Inf. Comp. Sci.* **1989**, *29*, 97–101.

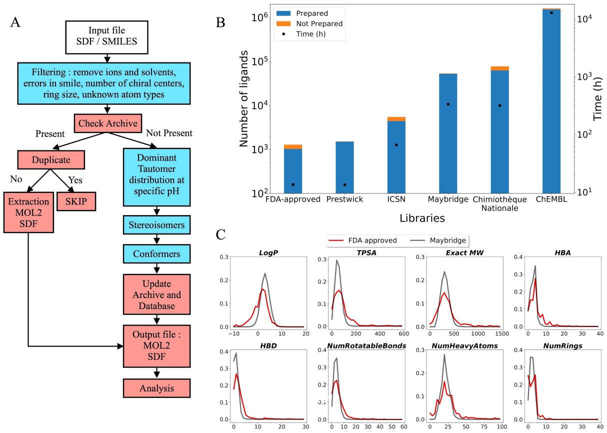
[29] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Deliver. Rev.* **1997**, *23*, 3–25.

[30] H. Pajouhesh, G. R. Lenz, *Neurotherapeutics* **2005**, *2*, 541–553.

[31] J. Gasteiger, M. Marsili, *Tetrahedron* **1980**, *36*, 3219–3228.

Received: May 27, 2021
Accepted: August 13, 2021
Published online on [redacted], [redacted]

APPLICATION NOTE



*M. Sisquellas, M. Cecchini**

1 – 7

PrepFlow: A Toolkit for Chemical Library Preparation and Management for Virtual Screening