Methods in Pharmacology and Toxicology (2016): 257–296
DOI 10.1007/7653_2014_35

Published online: 22 January 2015

Recent Advances in the Open Access Cheminformatics Toolkits, Software Tools, Workflow Environments, and Databases

Pravin Ambure, Rahul Balasaheb Aher, and Kunal Roy

Abstract

Cheminformatics utilizes various computational techniques to solve a wide variety of drug discovery problems, including drug design and predictive toxicology. These computational exercises employ various toolkits/libraries, workflows, databases, etc. for their applications in lead optimization, virtual screening, chemical database mining, structure-activity/toxicity studies, etc. It is therefore important for such techniques to be freely available. Open-access resources permit free use and redistribution of a product via a free license, while open-source resources also provide source code that can be utilized to modify the product. In order to extract the knowledge from enormous amount of data that accumulates at a staggering rate, open-access or open-source cheminformatics packages also need to be efficient and user-friendly. In this chapter, we record the recent advances in freely available (including both open access and open source) cheminformatics toolkits, software (stand-alone and online applications), workflow environment, and databases. The objective of this chapter is to get the readers acquainted with the freely available resources, so that they can utilize those tools for solving different drug discovery challenges. We will start with the toolkit/libraries such as Chemistry Development Kit (CDK), Open Babel, RDKit, ChemmineR, Indigo, chem^{f}, etc., which provide various functionalities that can aid researchers to develop their own cheminformatics software/applications. Next we will discuss various cheminformatics software tools, including iDrug, PharmDock, DecoyFinder, DemQSAR, Chembench, etc. which have recently been developed with a wide variety of applications. We will further discuss workflow environments, including Konstanz Information Miner (KNIME), Taverna, recent combinations, i.e., CDK-KNIME or CDK-Taverna and their contributions in the cheminformatics field. At the end, we will briefly touch various recent databases, such as QSAR DataBank, VAMMPIRE, CREDO, PubChem3D, MMsINC, etc., and their applications. The open-access resources covered in this chapter would enable the medicinal chemists and cheminformaticians to solve various problems encountered during their research.

Key words Open source, Open access, Cheminformatics, Tool kits, Software, Databases, Stand-alone tools, Online tools, Workflows

1 Introduction

Cheminformatics is an applied field of chemistry that involves the use of different computational resources for solving a variety of problems arising in chemical, pharmaceutical, and allied industries. This field is a combination of chemistry, computer science, and information science that aids in transforming huge raw data into information and this information into knowledge. Cheminformatics has revolutionized various areas including pharmaceutical and chemical research, in taking faster decisions, cutting cost, and hence increasing efficiency.

The availability of cheminformatics resources is based on the provider/source that are commonly categorized into commercially available, open access, or open source. Commercial resources are normally well developed but expensive; thus, their users are limited to those who can afford such costly services. Open-source resources, on the other hand, are freely available (i.e., open access), and their source code is openly accessible for modification or distribution. To best support the majority of the scientific community in resolving the different problems arising in multifarious areas of chemistry, it is essential that cheminformatics resources are openly accessible, which is not only important for resolving problems but also critical for bringing new amendments to overcome shortcomings of current resources and for exploring innovative concept/ideas associated with cheminformatics resources. Developing and managing such freely available cheminformatics resources is often "community driven" and an outcome of a teamwork from numerous contributors. In this chapter, we highlight the advances made in a variety of freely available (both open access and open source) cheminformatics resources, i.e., toolkits, software, workflow, and databases.

2 Cheminformatics Toolkits

Cheminformatics toolkits are a set of libraries comprising of source codes for various algorithms/functions that allow the cheminformaticians to develop their own software applications for possible use in structural similarity searching, virtual screening, database mining, structure-activity relationship analysis, etc. The development of open-source cheminformatics toolkits has started more than a decade ago, and so far many highly functional toolkits have been developed. Some toolkits were developed from the scratch, e.g., Chemistry Development Kit (CDK) [1] and Open Babel [2], while others, such as RDKit [3] and Indigo [4] toolkits, were made open source by donating *in-house* source code under liberal licenses. The development of such cheminformatics toolkits became a part of the Blue Obelisk movement [5, 6] established in 2005 as a response to the lack of Open Data, Open Standards and Open source (ODOSOS) in chemistry.

The main reason that these toolkits are so essential is that their availability aids the development of a next generation of cheminformatics software like Bioclipse [7], Avogadro [8], CDK Descriptor Calculator GUI [9], etc., where there is no need to concern about the low-level details of manipulating and/or handling various algorithms; thus one can focus on providing additional functionality and/or ease of use [6].

In this section, we describe various cheminformatics toolkits (*see* Table 1 for a brief summary) and their recent progress.

2.1 Chemistry Development Kit (CDK)

The CDK is an open-source Java library for structural cheminformatics and bioinformatics. This project was initiated in 2000 by Christoph Steinbeck, Egon Willighagen, and Dan Gezelter, the developers of Jmol [20] and JChemPaint [21]. Till date, it is one of the most active open-source cheminformatics projects that are being carried out with wide support from the scientific community. The number of contributors to this project has increased to 89 in 2014 [22].

CDK toolkit provides many functionalities for developing new software in the cheminformatics field such as various chemical input/output (I/O) file formats, including simplified molecular-input line-entry system (SMILES), Chemical Markup Language (CML), and MIT Design Language (MDL); structure generators; 2D diagram editing and generation; 3D geometry generation; substructure search using exact structures and Smiles ARbitrary Target Specification (SMARTS)-*like* queries; molecular descriptor calculation for quantitative structure-activity relationship (QSAR) study; fingerprint calculation; International Chemical Identifier (InChI) support (via JNI-InChI); etc., and in bioinformatics field, the functionalities include cognate ligand detection, metabolite identification, etc. [1, 23, 24].

At present, the CDK is the source for a number of software projects including JChemPaint [21], SENECA [25], NMRShiftDB [26], PaDEL-Descriptor [27], Jmol [20], JOELib, Nomen, SafeBase, and many more [28]. The functionality of CDK can also be freely accessed through workflow system, such as KNIME and Taverna, which are discussed in the workflow section of this chapter.

The CDK developers regularly perform unit testing, code quality checking, bug fixing, and proper versioning of CDK library. Information about the library functionality, core classes, inheritance hierarchy, and the dependencies among the fundamental classes of the CDK are well described in the literature [1, 10].

CDK is available for Windows, UNIX, and Mac OS and is freely distributed under the GNU Lesser General Public License (LGPL) version 2.0 (v2). In contrast to the more common GNU General Public License (GPL), the LGPL allows the use of the CDK in proprietary software packages.

2.2 RDKit

The RDKit was developed and employed at Rational Discovery during 2000–2006, for building predictive models for absorption, distribution, metabolism, elimination, toxicity, and biological activity. In June 2006, Rational Discovery was shut down, but the

Table 1
Summary of cheminformatics toolkits

Sr. No.	Name	Link	Programming language (wrapper, if any)	Operating system(s)	License	Reference No.
1	Chemistry Development Kit (CDK)	http://sourceforge.net/projects/cdk/	Java	Platform independent	LGPL	[1, 10]
2	RDKit	http://www.rdkit.org/	C++ (Python, Java and C# wrapper)	Mac, Windows, and Linux	BSD	[3]
3	Open Babel	http://openbabel.org/	C++ (Java, .NET platform, Perl, Python, and Ruby wrapper)	Windows, Mac OS X, Linux	GPL	[2]
4	Cinfony	https://code.google.com/p/cinfony/ https://github.com/cinfony/cinfony	Python, Jython	Platform independent	BSD	[11]
5	Small Molecule Subgraph Detector (SMSD)	http://www.ebi.ac.uk/thornton-srv/software/SMSD/	Java	Platform independent	Creative Commons (CC)	[12]
6	Biochemical Algorithms Library (BALL)	http://www.ball-project.org	C++	Windows, Linux, and Mac OS X	LPGL and GPL	[13]

7	Indigo	http://www.ggasoftware.com/opensource/indigo	C++ (Python, Java, and C# wrappers available)	Windows, Linux, and Mac OS X	GPL	[4]
8	jCompoundMapper	http://jcompoundmapper.sourceforge.net/	Java	Platform independent	LPGL	[14]
9	chem^f	https://github.com/stefan-hoeck/chemf http://www.scala-lang.org/	Scala (runs on Java platform)	Platform independent	GPL	[15]
10	Cheminformatics in Python (ChemoPy)	https://code.google.com/p/pychem/downloads/list	Python	Linux and Windows	–	[16]
11	ChemmineR	http://www.bioconductor.org/packages/release/bioc/html/ChemmineR.html	Statistical programming environment "R"	Windows and Mac OS X	Artistic 2.0	[17]
12	Compound-Protein Interaction with R (Rcpi)	http://bioconductor.org/packages/release/bioc/html/Rcpi.html	Statistical programming environment "R"	Windows and Mac OS X	Artistic 2.0	[18]
13	Chemkit	http://wiki.chemkit.org/Main_Page	C++	Windows, Mac, and Linux	BSD	[19]

toolkit was released as open source under BSD license. At present, the open-source development of RDKit is actively contributed by Novartis, which includes the source code donated by Novartis [29].

RDKit offers various functionalities such as different chemical I/O formats, including SMILES/SMARTS, structure data format (SDF), Thor data tree (TDT), Sybyl line notation (SLN), Corina mol2, and Protein Data Bank (PDB); substructure searching; canonical SMILES; chirality support (i.e., R/S or E/Z labeling); chemical transformations (e.g., remove matching substructures); chemical reactions; molecular serialization; similarity/diversity selection; 2D pharmacophores; Gasteiger-Marsili charges; hierarchical subgraph/fragment analysis; Bemis and Murcko scaffold determination; retrosynthetic combinatorial analysis procedure (RECAP) and BRICS implementations; multi-molecule maximum common substructure; feature maps; shape-based similarity; RMSD-based molecule-molecule alignment; shape-based alignment; unsupervised molecule-molecule alignment using Open3-DALIGN algorithm; integration with PyMOL for 3D visualization; functional group filtering; salt stripping; molecular descriptor library; similarity maps; machine learning; etc.

The *Contrib* directory [30] is a part of the standard RDKit distribution that includes source code that has been contributed by the community members, for instance, Local Environment Fingerprints (LEF), a Python source code; plane of best fit (PBF), a C++ source code; matched molecular pair algorithm (mmpa), a Python source code and sample data; a fragment indexing algorithm; and synthetic accessibility (SA) score, a Python source code.

RDKit has an official website, i.e., http://www.rdkit.org/, for online documentation, recent news, Wiki link, and other related information. The core data structures and algorithms are written in C++ programming language, but Python wrapper (generated using Boost.Python) and Java and C# wrapper (generated using SWIG) are also available. Here, wrapper or binding means a thin layer of code that converts a library's existing interface into a user's compatible interface, so that one can use the library in other programming languages. RDKit is supported by Mac, Windows, and Linux operating system [31].

2.3 Open Babel

Open Babel is an open-source chemical toolbox intended to be a cross-platform library that is built to support interconversion between various file formats used in cheminformatics, molecular modeling, and related areas. In addition to the complete, extensible toolkit/libraries, it also offers ready-to-use applications for the development of cheminformatics software [2].

Open Babel 2.3 can perform reading, writing, and interconversion of over 111 chemical file formats that includes reading and writing of 82 and 85 file formats, respectively. These encompass common formats used in cheminformatics (SMILES, InChI, MOL, MOL2), I/O files from a variety of computational chemistry packages (GAMESS, Gaussian, MOPAC), crystallographic file formats (CIF, ShelX), reaction formats (MDL RXN), file formats used by molecular dynamics simulation (Amber) and docking packages (AutoDock), formats used by 2D drawing packages (ChemDraw), 3D viewers (Chem3D, Molden), chemical kinetics, and thermodynamics (ChemKin, Thermo). It supports filtering and searching molecule files using Daylight SMARTS pattern matching and computes group contribution descriptors such as LogP, polar surface area (PSA). and molar refractivity (MR). It also provides extensible molecular fingerprinting and molecular mechanics functions.

File formats are employed as "plug-ins" which aid users to contribute new file formats. Further, depending on the file format, Open Babel can extract additional information besides molecular structure. For instance, property fields can be read from SDF files, unit cell information can be extracted from CIF files, and vibrational frequencies can be extracted from computational chemistry log files. For each file format, multiple choices/options can be chosen to read or write in a particular format.

Open Babel has its origin in a version of OELib, which was released as open-source software by OpenEye Scientific under the GNU GPL v2. In 2001, OpenEye decided to rewrite OELib *in-house* as the proprietary OEChem library, and the existing code from OELib was released as the new Open Babel project. Since then, Open Babel has been developed and substantially extended as an open-source project with extensive international collaborations. Up to November 2014, it has over 324,780 downloads [32] and more than 476 citations [33] and has been utilized by over 40 software projects. While the majority of the Open Babel library is written in C++, its bindings have been developed for other programming languages, including Java, .NET platform, Perl, Python, and Ruby. These can be automatically generated from the C++ header files using the SWIG tool. Open Babel supports a wide variety of C++ compilers (MSVC, GCC, Intel Compiler, MinGW, Clang), operating systems (Windows, Mac OS X, Linux), and platforms (32-bit, 64-bit). OpenEye scientific has also provided a set of cheminformatics and modeling toolkits [34] that are freely accessible under the free public domain research license [35].

In summary, Open Babel offers a solution to deal with the growing number of chemical file formats along with various functionalities like conformer searching, 2D depiction, filtering, batch conversion, and substructure and similarity search. For software developers, it can be used as a library to handle data in ample of areas such as organic chemistry, drug design, molecular modeling, and computational chemistry [2, 36].

Scripting languages like Python are highly popular since they allow rapid writing of scripts within a few lines of code and are well suited for common programming tasks in cheminformatics. For the same reason, a Python wrapper for Open Babel called Pybel [37] has been made available. It is an open-source, cross-platform Python module that provides the functionality of the Open Babel toolkit to Python programmers.

2.4 Cinfony

In the present scenario, the most active open-source cheminformatics toolkits under development comprise of Open Babel, the CDK, and the RDKit. All of these toolkits share the same core functionality although the implementation details and the chemical model employed may differ. However, these toolkits are independently developed, and therefore, each has certain specific functionalities, e.g., these toolkit support different sets of file formats and force fields and represent various molecular fingerprints and descriptors in different ways. There are also features in each of the toolkits that are not shared by the others.

To this end, Cinfony is a Python module which provides a common interface for Open Babel, RDKit, and CDK through a simple and robust method to pass the chemical models among these toolkits. It is an extension of Pybel (*discussed above*), a Python module that only provides access to Open Babel. It allows interoperability at the application programming interface (API) level, which has the advantage of not requiring any changes to the existing software. It is platform independent and hence supported by all operating systems such as Mac, Windows, and Linux operating system and is released as open source under the BSD license. The details about the barriers, interoperability, implementation, and performance are well discussed in the literature [11].

2.5 Small Molecule Subgraph Detector (SMSD)

The chemical similarity determination between molecules is widely used to compute chemical diversity and for clustering analysis of similar molecules. This is a highly useful concept in cases like discovering new *drug-like* molecules. Maximum common subgraph (MCS) is one of the recent methods that overcome nearly all the shortcomings posed by descriptor- or fragment-based similarity searches. All of the early MCS algorithms lacked chemical knowledge to rank the MCS solutions, which led to the development of SMSD toolkit [12].

SMSD is a chemically sensitive and robust tool, which uses a combination of various graph-matching algorithms (i.e., CDKMCS, MCS+, and VF+ Lib [12]) for finding the MCS among small molecules. It can generate bond-sensitive and bond -insensitive MCS and ranks the solutions according to minimal fragments, bond breaking energy, and stereochemical matches. It also overcomes the disadvantages of MCS algorithm coded in the CDK toolkit (CDKMCS) by first using the atom and bond count filter to discriminate between two dissimilar structures before performing the MCS search and, secondly, by using the VF+ Lib and MCS+ method. The reported disadvantages of CDKMCS include: (a) it may treat two chemically nonidentical molecules as identical because it works on the maximum common induced subgraph (MCIS) principle, and (b) the runtime is high if two graphs are large with few dissimilar edges.

SMSD is a combination of various algorithms (i.e., CDKMCS, MCS+, and VF+ Lib). The choice of using which algorithm is completely based on the complexity of the input molecules. CDKMCS is used first for molecules whose bond count and atom count are not equal. If the solution is not computed within a limited set time, then it is passed to the MCS+ algorithm, which starts the search from the scratch. If the *d*-edge count (those edges that do not share similar bond types) is greater than 99,999, then VF+ Lib is used to find MCS, which is very efficient in handling medium- to large-sized graphs. The MCS solutions are then passed to the chemical post-filters, which ranks the solutions in a chemically meaningful way. The three filters are applied in the following orders: (a) specific matching of the chemical functional groups, bond types, and stereochemistry of molecules are identified and matched; (b) the resulting solutions are sorted in ascending order of the total bond breaking energy required by this MCS match (i.e., lowest energy is highest ranked); and (c) the top solutions are selected based on the above two steps (in the case that the matched part of the molecule is detached from the reference structure, the solutions are sorted again in decreasing order based on the number of fragments generated). In such a way, one can get chemically relevant MCS solutions computed in polynomial time.

SMSD can be applied in a variety of bioinformatics and cheminformatics areas for exhaustive MCS matching. For example, it can be employed to analyze metabolic networks by matching the reactants with the products of the reactions. It can also be used to detect the MCS/substructure in small molecules reported by metabolomics experiments, as well as to screen for *drug-like* compounds with similar substructures.

SMSD is a Java-based toolkit, which is platform independent and is released under Creative Commons (CC) license [38]. This toolkit is freely available at www.ebi.ac.uk [12].

2.6 Biochemical Algorithms Library (BALL) 1.3

BALL is written in C++ with a specific purpose of significantly reducing the development time of building derived computational applications while ensuring stability and errorless implementation in the computational biology and molecular modeling field. It provides an extensive set of data structures along with classes for molecular mechanics (MM), file I/O, comparison and analysis of protein structures, advanced solvation methods, and visualization. BALL has been designed to be robust, easy to use, and also simply extensible due to its object-oriented programming background. In 2010, a new version, i.e., BALL 1.3, was released, and this version showed significant improvements in functionality compared to the previous version that had been released in 1999 [13].

BALL handles a variety of molecular structure formats. The previously published version only supported the PDB and MOL2 molecular file formats, while version 1.3 additionally reads and writes MOL, HIN, XYZ, KCF, and SD files. It also supports a variety of other data sources, such as DCD, DSN6, GAMESS, JCAMP, SCWRL, and TRR. Along with this, the new version also offers functionality for generating and editing molecules. For proteins, DNA, and RNA, BALL 1.3 can automatically deduce much of the missing information such as connectivity or bond order or missing hydrogen from an extensible fragment database. Both fragment database and rotamer library have been significantly improved in the latest version. A rotamer library allows the user to easily determine the most likely side-chain conformation of a protein residue or to switch between various rotameric states. The new features also include a kekulizer (an aromaticity processor), a secondary structure predictor, and hydrogen bond detection.

The previous version of BALL has provided two force field classes, i.e., CHARMM and AMBER. In the 1.3 version, an implementation of the Merck Molecular Force Field (MMFF94) has been included that allows handling of almost all types of organic compounds. The energy minimization functions have been extended via providing standard methods like steepest descent and conjugate gradient and the well-known methods, i.e., L-BFGS and shifted L-VMM algorithms [13].

BALL is freely available and supported by all major operating systems, including Linux, Windows, and Mac OS X. Previously, BALL was distributed as a commercial product, but now it is released open source under the GNU Lesser General Public License (LGPL), and parts of the code are released under the GNU GPL. The source code and binary packages are available from the project website at http://www.ball-project.org [13].

2.7 Indigo

In 2009, GGA software services released a toolkit titled "Indigo" and related software under the terms of GNU GPL. It includes some unique algorithms developed by GGA, along with some standard well-known algorithms.

The main features of Indigo are:

- Supports commonly used and popular chemistry formats: molfiles/Rxnfiles v2000 and v3000, SDF, RDF, SMILES, SMARTS, and SMIRKS
- Supports tetrahedral and *cis-trans* stereochemistry
- Molecule and reaction rendering to PNG, SVG, and PDF files
- Molecule and reaction depiction
- Aromatization and kekulization

- Canonical (isomeric) SMILES computation
- Exact and substructure matching for molecules and reactions
- Support matching and highlighting
- Matching of tautomers and resonance structures
- Computing molecule and reaction fingerprints
- Similarity search
- Maximum common substructure (MCS) algorithm
- R-group deconvolution and scaffold detection

It is a C++ language-based library, majorly focused on performance and essential chemical features. The high-level wrappers or bindings are built around it for Python, Java, and C# language. This library also allows multi-threaded use. All binaries are supported by all the essential operating systems, i.e., Windows, Linux, and Mac OS X, both 32-bit and 64-bit [39]. The JAVA GUI utilities, command-line utilities, KNIME nodes, and documentation material are available on their official website [4]. A commercial license version is also available for receiving ongoing support and maintenance and for clients who like to include Indigo as a component in their proprietary software product [4].

2.8 jCompoundMapper

jCompoundMapper [14] is an open-source Java library for the encoding of chemical graphs as fingerprints. It offers a variety of topological (e.g., radial atom environments, extended connectivity fingerprints, depth-first search fingerprints, or autocorrelation vectors) and geometrical (e.g., 2-point and 3-point encodings or geometrical atom environments) fingerprints. It is based on CDK, which offers the basic functionality for parsing, typing, and graph algorithms for molecular data and also provides several fingerprint functionalities. But unlike CDK, jCompoundMapper focuses on exact definition of its encoding and provides functionality to export the fingerprints or pairwise similarity matrices to formats of popular machine learning toolboxes such as comma-separated value (*csv*) format, LIBSVM format (sparse and matrix), and WEKA ARFF. Hence, various data mining libraries can be directly applied on the output files.

This library is built from various implementations of literature fingerprints and descriptors used in comparison studies, and its algorithms can be parameterized with various options to adapt the encodings, for instance, by applying a custom labeling function; adjusting the search depth, the distance cutoff, or the geometrical scaling factor; etc.

jCompoundMapper is platform independent and is released under the LPGL license. It features a command-line interface but can also be used as a Java application programming interface (API). The access via the API or the binary using the command-line interface enables the user to utilize the library for batch processing. The source code and an executable library are available at SourceForge [40].

2.9 chem^f

Chem^f is a chemistry toolkit, a first of its kind being built using a functional programming language named "*Scala*" [15]. *Scala* is a modern multi-paradigm open-source programming language that is fully object-oriented with a strong support for typical concepts from functional programming such as higher-order functions, type inference, and pattern matching. It has one of the most expressive type systems [41].

Most of the freely available and widely used cheminformatics toolkits, such as the CDK or Open Babel, are written in object-oriented languages using typical imperative concepts such as mutable data structures and opaque methods to implement chemical entities and algorithms. The developers of the Chem^f toolkit have discussed the advantages of functional programming compared to the imperative programming languages and reported an example for justification, i.e., the comparison between CDK's and Chem^f's SMILES parser [15].

Functional programming languages were designed with referential transparency in mind, and they encourage a more declarative style of programming without the control statements and value assignments that are typically found in imperative languages. An expression in a program is called referentially transparent when it performs calculation of the result just from its input parameters. Functional programming significantly facilitates writing referentially transparent functions and using immutable data structures. The methods and objects written in most of the languages (other than functional languages) are a priori unsafe to be used in parallel computations (i.e., multi-threaded operations) unless their documentation explicitly states differently. More detailed information about the functional programming, *Scala* language, and some examples illustrating its basic syntax and comparison with other open-source toolkits are provided in the freely accessible literature [15].

In summary, chem^f is an open-source toolkit written in Scala language and is released under GNU GPL and is currently under development [42].

2.10 Cheminformatics in Python (ChemoPy)

ChemoPy is an open-source package for computing the commonly used structural and physicochemical features. It depends on several other packages that are Pybel, RDKit, Open Babel, and MOPAC, in order to provide its complete functionalities. It calculates about 16 feature groups composed of 19 various features that in all comprises of around 1,135 descriptors. Additionally, it offers seven types of molecular fingerprint systems that include topological fingerprints, electro-topological state fingerprints, MACCS keys, FP4 keys, atom-pair fingerprints, topological torsion fingerprints, and Morgan/circular fingerprints. Some of these features and fingerprints are derived using Open Babel and RDKit. Using MOPAC, ChemoPy computes a large number of 3D molecular descriptors. Interestingly, ChemoPy is reported as the first open-source package computing a large number of molecular features based on the MOPAC optimization.

The ChemoPy toolkit encloses several modules and functions manipulating drug molecules. For instance, to obtain the molecular structures easily, ChemoPy provides a downloadable module to get molecular structures from four databases (i.e., KEGG, PubChem, DrugBank, and CAS). Further, ChemoPy can compute a large number of 2D and 3D descriptors and offers two ways to calculate these molecular descriptors. One way is to utilize the built-in modules, which consist of 19 modules responding to the calculation of descriptors from 16 feature groups, and the second way is to call the PyChem2d or PyChem3d class by importing the pychem module [43], which encapsulates commonly used descriptor calculation methods. Here, PyChem2d and PyChem3d are responsible for the calculation of 2D and 3D molecular descriptors, respectively.

The developers of ChemoPy have recommended utilizing this toolkit to analyze and represent the drugs or ligand molecules under investigation and suggested that this package will be helpful when exploring questions concerning drug activity, ADME/T, and drug-target interactions [16].

ChemoPy is written solemnly in Python language. It is supported by Linux and Windows operating systems. New extensions or functionalities can be implemented easily without cumbersome or time-consuming modifications in the source code because of the modular structure of ChemoPy [16].

2.11 ChemmineR

ChemmineR is a cheminformatics package for analyzing drug-like small molecule data in the popular statistical programming environment R. The first version of this package was published in 2008 [17]. It comprised of functions for 2D structural similarity comparisons between compounds, similarity searching against compound databases, functions for clustering entire compound libraries, and visualizing the clustering results.

The recent version of ChemmineR released in 2013 has additional utilities and add-on packages, including functions for efficient processing of large numbers of small molecules, physicochemical/structural property predictions, structural similarity searching, classification, and clustering of compound libraries with a wide spectrum of algorithms, including mismatch tolerant MCS search algorithm [44] used for pairwise compound comparisons. Accelerated compound similarity searching is now enabled with *eiR* add-on package [45]. The current version of ChemmineR also integrates a subset of cheminformatics functionalities implemented in the Open Babel C++ library. These utilities can be enabled by installing the ChemmineOB package and the Open Babel software. ChemmineR can automatically detect ChemmineOB and make use of its additional utilities. Streaming functionality allows processing of millions of molecules using *sdfStream* function. The recent addition also includes fast and memory-efficient fingerprint search, which supports the use of atom pair or PubChem fingerprints, and improved SMILES support via new SMIset object class and SMILES import/export functions.

ChemmineR is freely available from the Bioconductor official website (http://www.bioconductor.org/packages/release/bioc/html/ChemmineR.html). It is distributed under Artistic 2.0 license and is available for both Windows and Mac OS X operating systems [17].

2.12 Compound-Protein Interaction with R (Rcpi)

Rcpi toolkit provides a freely available R/Bioconductor package focusing on integrating bioinformatics and cheminformatics into a molecular informatics platform for drug discovery. It aims at providing a complete toolkit for complex molecular representation from small molecules and proteins and more complex interactions, including protein-protein and compound-protein interactions [18].

The functionalities provided by Rcpi toolkit can be divided into four groups as follows:

(a) For small molecules:

- It calculates more than 300 molecular descriptors, including constitutional, topological, geometrical, electronic, hybrid, and molecular property descriptors.
- It calculates ten types of molecular fingerprints, including standard and extended Daylight fingerprints, graph fingerprints based on simple connectivity, hybridization fingerprints only based on hybridization state, FP4 keys, E-state fingerprints, MACCS keys, PubChem fingerprints, KR fingerprints defined by Klekota and Roth, short path fingerprints, etc.
- It can compute parallelized pairwise similarity derived by fingerprints and five types of similarity measures within a list of small molecules.
- It also perform parallelized chemical similarity search with selected similarity metrics and MCS search between a query molecule and a molecular database.

(b) For protein sequences:

- It computes a large number of commonly used structural and physicochemical descriptors, such as amino acid composition descriptors; autocorrelation descriptors; composition, transition, and distribution descriptors; conjoint triad descriptors; quasi-sequence-order descriptors; and pseudo amino acid composition descriptors; etc.
- It calculates six types of generalized scale-based descriptors for proteochemometric (PCM) modeling, such as generalized scale-based descriptors derived by principal component analysis, amino acid properties, molecular descriptors, factor analysis, and multidimensional scaling, and generalized BLOSUM/PAM matrix-derived descriptors.
- It computes profile-based protein features based on position-specific scoring matrix (PSSM).
- It also performs parallelized similarity derived by protein sequence alignment and Gene Ontology (GO) semantic similarity measures between a list of protein sequences/ GO terms/Entrez Gene IDs.

(c) For interaction data:

By combining various types of descriptors for drugs and proteins, interaction descriptors representing protein-protein or compound-protein interactions could be conveniently generated with Rcpi, including:

- Two types of compound-protein interaction descriptors
- Three types of protein-protein interaction descriptors

(d) Several useful secondary utilities are also included in Rcpi that are as follows:

- Parallelized molecule and protein sequence retrieval from several online databases, such as PubChem, ChEMBL, KEGG, DrugBank, UniProt, RCSB PDB, etc.
- Molecular reading/writing in SMILES/SDF formats for small molecules and FASTA/PDB formats for proteins
- Molecular format conversion between around 140 types of molecular file formats defined by Open Babel

It is recommended to use Rcpi to analyze and represent various complex molecular data under study as well as to explore various queries concerning structure, functions, and interactions of such molecules in system biology perspective. Rcpi is freely available from the Bioconductor official website (http://bioconductor.org/packages/release/bioc/html/Rcpi.html) and is released under Artistic 2.0 license. It is available for Windows and Mac OS

X operating systems. Users can conveniently apply various statistical machine learning methods in R to solve various problems in drug discovery and computational biology [18].

2.13 Chemkit

Chemkit is an open-source library developed by Kyle Lutz [19], which supports molecular modeling, cheminformatics, and visualization functionalities. The key features provided by Chemkit include I/O chemical file formats (pdb, cml, cjson, cube, fhz, fps, inchi, mol, mol2, sdf, smi, etc.), access to the Web resources (e.g., Protein Data Bank, PubChem), calculation of various molecular descriptors, automatic atom typing and topology building, and visualization based on OpenGL.

The Chemkit library is written in C++ language, and it uses Qt framework for graphics. It is released under the BSD license. It is a cross platform library and is supported by Windows, Mac, and Linux operating systems.

Other toolkits that are not recently updated and/or not actively maintained but are worth mentioning include Chemfp 1.1 (http://chemfp.com/; Python Library), Chemical Descriptor Library (http://cdelib.sourceforge.net/doc/index.html; a C++ library), PerlMol–Perl Modules for Molecular Chemistry (http://www.perlmol.org/; collection of Perl scripts), MayaChemTools (http://www.mayachemtools.org/; collection of Perl scripts), JOELib/JOELib2 (http://sourceforge.net/projects/joelib/; Java library), and mx-Java (https://code.google.com/p/mx-Java/; Java Library).

3 Cheminformatics Software

Software is a set of machine-readable instructions that directs a computer's processor to perform specific functions. It is usually written in high-level programming languages that are easier and more efficient for humans to understand and employ than the machine language. Cheminformatics software provides various ready-to-use cheminformatics functionalities like virtual screening, chemical structural editor, QSAR model development, molecular dynamics packages, etc. that are most often user-friendly tools comprising of graphical user interface (GUI).

Software can be commercial, open access, or open source. They can be developed with or without the use of external libraries/toolkits. Existing libraries/toolkits may help reduce the software development time because source codes for various algorithms are already well defined in terms of efficiency and are ready to use. However, in many cases, the developers prefer not to use external toolkits to avoid the usage of inefficient or time-consuming algorithms, but to develop their own more efficient algorithms.

In this section, we have divided the cheminformatics software tools into two parts: stand-alone software and online tools.

The stand-alone software is a tool/application that can work offline and does not require another software package to run. The online tool means a Web-based platform-independent software tool that runs online, making the facilities available to users over the Internet. Here, both open-access and open-source software are discussed, including computer as well as mobile applications. The brief summary of cheminformatics software and their home page, supporting operating system, and programming languages are listed in Table 2.

3.1 Stand-Alone Software Tools

3.1.1 Molpher

Molpher [46] is an open-source software framework for the systematic exploration of the chemical space. When a source/target molecular pair is given as an input entry, Molpher identifies the structural neighborhood through a process known as molecular morphing. The molecular morphing process produces a path in the chemical space by an iterative application of morphing operators, which represent a structural change such as the addition or removal of an atom or a bond. This path consisting of molecules called as morphs and its surroundings constitutes a virtual chemical library focused on a mechanistic class of compounds given by the characteristics of the source/target pair. Although Molpher is written in C++ language, it uses Boost C++ libraries [47] for standard tasks and employs open-source cheminformatics RDKit [3] for chemical functionalities. It could be easily incorporated into any computational drug design pipeline and thus is highly useful for either discovery of novel drugs or as new tool for chemical biology [46].

3.1.2 PharmDock

PharmDock [48] is a pharmacophore-based docking program that combines pose sampling and ranking, which are based on optimized protein-based pharmacophore models, with local optimization using an empirical scoring function. The testing of PharmDock for ligand pose prediction, binding affinity estimation, compound ranking, and virtual screening yielded comparable or better performance as compared to other existing and widely used docking programs [49, 50]. This docking program comes with an easy-to-use GUI within PyMOL [48].

3.1.3 VHELIBS

VHELIBS (validation helper for ligands and binding sites) is a software tool for assessing the quality of ligands and binding sites in the crystallographic models from the PDB/PDB_REDO for the non-crystallographers (i.e., users with little or no crystallography knowledge). It allows the users to check how the ligand and binding site coordinates fit to the electron density map (ED) and to validate the protein structures prior their use for the drug discovery purposes [51].

3.1.4 MOLE 2.0

MOLE 2.0 [52] is an advanced software tool for analyzing the molecular channel and pores of the biomolecular surface. MOLE 2 also estimates the physicochemical properties of the identified

Table 2
Summary of cheminformatics software

Sr. No.	Stand-alone software/online tools	Project home page	Operating system(s)	Programming language
A	*Stand-alone software*			
1	*Molpher*	http://siret.cz/molpher/	MS Windows (client and server), Linux (server)	C++
2	*PharmDock*	http://people.pharmacy.purdue.edu/~mlill/software/pharmdock	Linux	C, Python
3	*VHELIBS*	http://urvnutrigenomica-ctns.github.com/VHELIBS/	Platform independent, Linux (server)	Python, Java
4	*MOLE 2.0*	http://mole.chemi.muni.cz	Mac OS, Linux, Windows	C#
5	*FragVLib*	http://www.unc.edu/~raed/FragVLib.zip	MS Windows, Linux	C++
6	*TB Mobile (iOS)*	https://itunes.apple.com/us/app/tbmobile/id567461644?mt=8	iOS	Objective-C programming
7	*TB Mobile (Android)*	http://play.google.com/store/apps/details?id=com.mmi.android.tbmobile	Android	Java
8	*JSME*	http://peter-ertl.com/jsme/	Platform independent	JavaScript
9	*CheS-Mapper*	http://ches-mapper.org	Cross-platform	Java with Java Web Start support (can be started from a Web browser)
10	*ScreeningAssistant2 (SA2)*	http://sa2.sourceforge.net/	Platform independent	JAVA/SQL
11	*LipidMapsTool*	www.lipidmaps.org/downloads/	Platform independent	Perl
12	*DecoyFinder*	http://urvnutrigenomica-ctns.github.io/DecoyFinder/	Linux, Windows	Python

13	*Open Molecule Generator (OMG)*	http://sourceforge.net/p/openmg	Linux 64 bits, Linux 32 bits, Mac OS X	Java, C
14	*mol2chemfig*	http://chimpsky.uwaterloo.ca/mol2chemfig/	Linux, Windows, Mac	Python 2.7
15	*LICSS*	http://code.google.com/p/excel-cdk/	Windows (XP, Vista or Windows7); Microsoft Excel for Windows (1997–2010)	VBA, Java, C++
16	*Avogadro*	http://avogadro.openmolecules.net/	Cross-platform	C++, Python
17	*MyChemise*	http://www.knalltundstinkt.de/englische%20Version/knalltundstinktE.html	Web based, Windows	Java
18	*Open3DALIGN*	http://open3dalign.org/	Windows 32/64-bit, Linux 32/64-bit, Solaris x86 32/64-bit, FreeBSD 32/64-bit, Intel Mac OS X 32/64-bit	C, BLAS and LAPACK libraries
19	*mpAD4*	http://autodock.scripps.edu/downloads/multilevel-parallel-autodock4.2	Platform independent	C++
20	*DemQSAR*	http://agknapp.chemie.fu-berlin.de/dempred/	Platform-independent	Java
21	*Shape*	http://sourceforge.net/projects/shapega	Linux	Java 1.5 or higher
22	*OrChem*	http://orchem.sourceforge.net/	Platform independent	Java 1.5 or higher; *Database system:* Oracle 11 g (with JRE 1.5)
23	*PaDEL-Descriptor*	http://padel.nus.edu.sg/software/padeldescriptor/	Platform independent	Java
24	*QSARINS and QSARINS-Chem*	http://www.qsar.it/	Windows 2000 or more recent version	C++

(continued)

Table 2 (continued)

Sr. No.	Stand-alone software/online tools	Project home page	Operating system(s)	Programming language
25	*VEGANIC*	http://www.vega-qsar.eu/download.html	Platform independent	Java
26	*OECD QSAR toolbox*	http://www.oecd.org/chemicalsafety/risk-assessment/theoecdqsartoolbox.htm	Windows XP	–
27	*ChemAxon Software tools*	http://www.chemaxon.com/free-software/	Platform independent	Java
28	*DTC-Lab Software tools*	http://teqip.jdvu.ac.in/QSAR_Tools/	Platform independent	Java
B	*Online tools*			
1	*iDrug*	http://lilab.ecust.edu.cn/idrug	Web-based platform independent	Java, Python
2	*OCHEM*	http://ochem.eu	Web-based platform independent	Java
3	*Chembench*	http://chembench.mml.unc.edu	Web-based platform independent	Java
4	*Spectral Game*	http://spectralgame.com	Web-based platform independent	HTML, PHP, JavaScript, JAVA (JSpecView)
5	*ACD/I-Lab tool*	http://www.acdlabs.com/resources/ilab/	Web-based platform independent	–

channels, i.e., hydropathy, hydrophobicity, polarity, charge, and mutability; this feature was absent in the previously developed related software tools, such as MOL 1.x [53], MolAxis [54, 55], and CAVER 3.0 [56]. The estimated physicochemical properties of the identified channels in the selected biomacromolecules corresponded well with the known functions of the respective channels. Thus, the predicted physicochemical properties by MOLE 2.0 provide useful information about the potential functions of identified channels [52].

3.1.5 FragVLib

FragVLib is a free database mining software for performing similarity search across database(s) of ligand-receptor complexes for identifying binding pockets which are similar to that of a target receptor. The methodology employed relies on the graph representation of interfacial atoms for the ligand-receptor complex. The interfacial atoms are defined as nodes, and the distances between them are represented by edges connecting these nodes. The search is based on 3D geometric and chemical similarity of the atoms forming the binding pocket. For each match identified, the ligand fragments corresponding to that binding pocket are extracted, and thus the formed virtual library of fragments (FragVLib) is useful to the structure-based drug design [57].

3.1.6 TB Mobile

TB Mobile is a mobile application (app) that provides the useful functionality for viewing and manipulating data about the antitubercular compounds with activity against *Mycobacterium tuberculosis* (Mtb), their targets, pathways, and other related information available in the Collaborative Drug Discovery (CDD) database. The app enables the similarity searching to identify the potential targets and to retrieve the active compounds. The molecules can be copied to the clipboard, opened with other apps, and bookmarked and exported. TB Mobile may assist the researchers as part of their workflow in identifying the potential targets for hits generated from phenotypic screenings and in hit prioritizations. The TB Mobile app is freely available from the Apple iTunes App Store and Google Play [58].

3.1.7 JSME

JSME is the free molecular editor (JSME) written in the JavaScript. The actual molecule editing Java code of the JME editor was translated into a JavaScript with the help of the Google Web Toolkit compiler and a custom library that emulates a subset of the GUI features of the Java runtime environment (JRE). In this process, the editor performance was enhanced by the additional functionalities of a substituent menu; copy/paste, drag and drop, and undo/redo capabilities; and an integrated help as compared to the previously available JME applet. The editor supports a molecule editing on the touch devices, such as iPhone, iPad, Android phones, and tablets in addition to the desktop computers. This new editor is easy to use and easy to be incorporated into the Web pages [59].

3.1.8 CheS-Mapper

CheS-Mapper (Chemical Space Mapper) is a 3D molecular viewer software tool to visualize and explore the chemical datasets. It divides a large dataset into clusters of similar compounds and consequently arranges them in the 3D space, such that their spatial proximity reflects their similarity. The tool detects the subgroups (clusters) within the data and can be employed to analyze the data to find the possible structure-activity relationship (SAR) information. This tool also calculates different kind of features, such as structural fragments as well as quantitative chemical descriptors [60].

3.1.9 ScreeningAssistant2 (SA2)

ScreeningAssistant2 (SA2) is an open-source Java software dedicated to the storage and analysis of small to very large chemical libraries. SA2 stores the unique molecules in a MySQL database and encapsulates several cheminformatics methods such as provider's management, interactive visualization, scaffold analysis, diverse subset creation, descriptors calculation, substructure/ SMART search, similarity search, and filtering. It facilitates the management of chemical libraries through an intuitive and interactive graphical interface and provides a set of advanced methods to analyze and exploit their content. Thus, it is useful for removing a variety of classes of compounds that are likely to be characterized as false positives in biochemical screening [61].

3.1.10 LipidMapsTool

LipidMapsTool is a software package for the template-based combinatorial enumeration of virtual compound libraries for lipids. A set of command-line scripts is used to enumerate all possible structures corresponding to the specified lipid abbreviations without any additional input requirements from the user. The virtual libraries are enumerated for the specified lipid abbreviations by using matching lists of predefined templates and chain abbreviations, instead of core scaffolds and lists of R-groups provided by the user. This tool is capable of generating large virtual compound libraries for lipids with minimal input from the user [62].

3.1.11 DecoyFinder

DecoyFinder [63] is a Python-based GUI application for the building of target-specific decoy sets. It selects a set of decoys for a target from a compound database based on a given collection of active ligands. The algorithm for decoy selection implemented in DecoyFinder is similar to that used to construct the DUD database [64, 65] and other benchmarks [66]. The MACCS fingerprints [67] and five physical descriptors are calculated for each active and potential decoy molecule using the Open Babel toolbox [2]. The Decoys are selected if they are similar to the active ligands according to five physical descriptors (molecular weight, number of rotational bonds, total hydrogen bond donors, total hydrogen bond acceptors, and the octanol-water partition coefficient) without being chemically similar to any of the active ligands used as an input (according to the Tanimoto coefficient between MACCS fingerprints). This is the first application designed to build the target-specific decoy sets [63].

3.1.12 Open Molecule Generator (OMG)

OMG is the first open-source structure generator, which produces all the non-isomorphic chemical structures that match with a given elemental composition. Also, this structure generator accepts additional input as one or multiple nonoverlapping prescribed substructures to drastically reduce the number of possible chemical structures. OMG relies on a modified version of the Canonical Augmentation Path approach, which grows intermediate chemical structures by adding bonds and checks that at each step only unique molecules are produced. When OMG was compared with the commercially available structure generator such as MOLGEN [68], the results obtained, i.e., the number of molecules generated, were identical for elemental compositions having only C, O, and H. The major advantage of OMG is that it is an open-source software; thus, the user can understand the functioning of software and also can customize the software according to his/her requirements. This structure generator would be useful to many fields, especially to the metabolomics area, where identifying the unknown metabolites is still a major bottleneck [69].

3.1.13 mol2chemfig

This tool is written in the Python language to convert a large number of structures in molfile or SMILES format into the LATEX source code. Its output is written in the syntax defined by the chemfig TEX package, which allows for flexible and concise description of chemical structures and reaction mechanisms. The program is freely available through a Web interface. It also can be locally installed on the user's computer, and the source code is accessible from the home page [70].

3.1.14 LICSS

LICSS is the lightweight Excel-based chemical spreadsheet (open-source software) which stores structures as SMILES strings. Chemical operations are carried out by calling Java code modules, which uses the CDK, JChemPaint, and OPSIN libraries to provide the cheminformatics functionality. The compounds in the sheets may be visualized (individually or combined), and the sheets may be searched by substructure or similarity. The descriptors available in CDK may also be calculated for all the compounds, and various cheminformatics operations such as fingerprint calculation, Sammon mapping, clustering, and R-group table creation may be carried out. It can be used suitably on a sheet containing thousands of compounds without compromising the normal performance of Microsoft Excel [71].

3.1.15 Avogadro

The Avogadro library is an advanced semantic chemical editor, visualization, and analysis platform. This framework provides a code library and application programming interface (API) with the three-dimensional visualization capabilities and has direct applications for research and education in the fields of chemistry, physics, materials science, and biology. Moreover, this application also provides a rich graphical interface using dynamically loaded plug-in through the library itself. The application and library can be extended by implementing a plug-in module in C++ or Python to explore different visualization techniques, build/manipulate molecular structures, and interact with the other programs [8].

3.1.16 MyChemise

My Chemical Structure Editor (MyChemise) is a new 2D structure editor designed as a Java applet, which enables the direct creation of structures in the Internet using a Web browser. It has a morphing module, which allows the creation of different types of presentation for dynamic visualization, for example, clear and simple illustration of molecule vibrations and reaction sequences. Thus, this new 2D drawing program has a versatile way of creating structural images [72].

3.1.17 Open3DALIGN

Open3DALIGN is an open-source software tool which is capable of carrying out conformational searches and multi-conformational, unsupervised rigid-body alignment of 3D molecular structures. The multiple alignment paradigms (i.e., atom-based, pharmacophore-based and mixed) are implemented in the methodology. The mixed and atom-based superposition algorithms give rise to the most consistent and well-ordered alignments, which are particularly suitable for the 3D-QSAR techniques. The high computational performance, the unsupervised nature of the alignment algorithms, and its scriptable interface make Open3DALIGN an ideal component of automated cheminformatics workflows. The Open3DALIGN tool is written in C language and linked to high-performance BLAS and LAPACK libraries with parallel algorithms implemented for high computational performance using the multiprocessor architectures [73].

3.1.18 mpAD4

Norgan et al. [74] reported the multilevel parallelized AutoDock 4.2 (mpAD4) software which was built using system-level (MPI) and node-level (OpenMP) parallelization to facilitate the application of this docking software on MPI-enabled systems and multithread the execution of individual docking jobs. The multithreading of AutoDock's Lamarkian Genetic Algorithm with OpenMP increases the speed of individual docking jobs; when combined with MPI parallelization, it can significantly reduce the execution time of virtual screenings. This multilevel parallelized AutoDock 4.2 software speeds up the execution of certain molecular docking workloads and allows the user to optimize the degree of system-level (MPI) and node-level (OpenMP) parallelization to best fit both the workloads and the computational resources.

3.1.19 DemQSAR

DemQSAR is a stand-alone Java application tool that can be used to predict the human volume of distribution (VD_{ss}) and human clearance (CL). DemQSAR integrates the open-source CDK library to compute various molecular descriptors and fingerprints, and thus the QSAR models can be built without any additional software. DemQSAR incorporates two state-of-the-art feature selection strategies: embedded Lasso and recursive feature elimination (RFE). The appropriate quality measures are computed automatically depending upon whether the analysis being performed is classification based or regression based. In addition to the predicted VD_{ss} and CL values, 2D images, SMILES codes, molecular formula, and molecular weights are also computed for the uploaded compounds. Due to its fully automated approach and good predictive power, DemQSAR is an attractive tool for many QSAR/QSPR tasks [75].

3.1.20 Shape

Shape software package is used for predicting 3D conformation of carbohydrates up to a considerable size, which covers most of the known biologically active oligosaccharide compounds. Because detailed experimental three-dimensional structures of carbohydrates are often difficult to acquire, software, such as Shape, is an alternative and attractive method for the prediction of oligosaccharide conformations. The predictions of Shape agreed well with experimental data as well as with other published conformation prediction studies [76].

3.1.21 OrChem

OrChem is an open-source extension of the Oracle 11G database platform. It added the registration and indexing of chemical structures functions to support the fast substructure and similarity searching. Its cheminformatics functionality is provided by the CDK toolkit. OrChem provides similarity searching with response times in the order of seconds for databases with millions of compounds, depending on a given similarity cutoff [77].

3.1.22 PaDEL-Descriptor

PaDEL-Descriptor is a free and open-source software for calculating the molecular descriptors and fingerprints. This software calculates 797 descriptors (663 1D and 2D descriptors, 134 3D descriptors) and ten types of fingerprints. It has both graphical user interface and command-line interface that function on all major platforms (Windows, Linux, MacOS). PaDEL-Descriptor supports more than 90 different molecular file formats and is multi-threaded [27].

3.1.23 QSARINS and QSARINS-Chem

QSARINS (QSAR-INSUBRIA) is a free software for the development, analysis, validation, and application of QSAR MLR models according to the OECD principles. The updated version of QSARINS, i.e., QSARINS-Chem, comprises several datasets of environmental pollutants and their corresponding endpoints (i.e., physicochemical properties and biological activities). Those chemicals can be accessed by querying CAS number, SMILES string, compound names, etc., and the developed models can be downloaded in the QSAR model reporting format (QMRF) [78].

3.1.24 VEGA Non-Interactive Client (VEGANIC)

VEGA [79] (**V**irtual models for property **E**valuation of chemicals within a **G**lobal **A**rchitecture) is an open-source platform that provides the valid QSAR models to be used especially under the European legislation for chemical substances (REACH). VEGANIC is a software under VEGA platform in which one can execute all the VEGA models on the local machine without sending any information to the server. It is freely available for download from the VEGA website [79], and the endpoints and properties such as BCF (bioconcentration factor), mutagenicity, carcinogenicity, developmental toxicity, skin sensitization, LC_{50} aquatic toxicity, biodegradability, and LogP can be determined.

3.1.25 OECD QSAR Toolbox

OECD QSAR toolbox [80] is a software application intended to be used by governments, chemical industries, and other stakeholders in filling the gaps in (eco)toxicity data needed for assessing the hazards of chemicals. The key features of this toolbox include identification of relevant structural characteristics and potential mechanism or mode of action of a target chemical, identification of other chemicals that have the same structural characteristics and/or mechanism or mode of action, filling the data gap(s), using existing experimental data, and systematically grouping of chemicals into categories according to the presence or potency of a particular effect for all members of the category.

3.1.26 ChemAxon Software Tools

ChemAxon is a software company specializing in developing programming interfaces and end-user applications for cheminformatics and life science research. It provides some of the noncommercial and academic free version tools such as MarvinSketch, MarvinView, MarvinSpace, Marvin JS, JChem Base/JChem Cartridge, MolConverter, and JChem for Excel applications [81].

3.1.27 DTC Lab Software Tools

The Drug Theoretics and Cheminformatics laboratory (DTC Lab., Jadavpur University, India) has developed several open-access cheminformatics tools, mostly dealing with QSAR studies. All the stand-alone tools are built using Java language and can be freely downloaded from the official website [82]. These basic QSAR tools such as MLR plus Validation (*for performing MLR and computing all validation parameters*), dataset division GUI (*dataset division* *into training and test sets using various algorithms*), data pretreatment GUI (*to remove constant and intercorrelated descriptors prior to model development*), modified K-medoid (*clustering method*), AD-MDI and Euclidean (*to define applicability domain*), genetic algorithm, stepwise MLR and MLR best subset selection (*variable selection methods*), etc., are very helpful when performing QSAR studies; manuals and sample input files of these software tools are also provided.

3.2 Online Tools

3.2.1 iDrug

*i*Drug is a versatile Web-based server for pharmacophore and similarity-based virtual screening and target identification to facilitate computational drug discovery. It provides ready-to-access compounds and pharmacophore target databases (such as ZINC, NCI, PharmTargetDB) for virtual screening and target identification. Different modules such as Cavity (detects and scores potential binding sites of a protein), Pocket v.2 (derives pharmacophore models based on a given receptor of complex structure), PharmMapper (pharmacophore mapping), SHAFTS (3D similarity calculation), Cyndi (molecular conformation generation), and Pybel (Python wrapper for the Open Babel cheminformatics toolkit) have been incorporated and can work together as a pipeline. Different molecular design processing tasks can be submitted and visualized simply in one browser without installing locally any standalone modeling software. It provides a novel, fast, and reliable tool for conducting drug design experiments [83].

3.2.2 OCHEM

OCHEM is a Web-based platform that provides the tools for automation of typical steps necessary to create a predictive QSAR/QSPR model. The platform consists of two major subsystems: a database of experimental measurements and a modeling framework. The database contains almost 10,000 data points for the density, bubble point, and azeotropic behavior of binary mixtures. The OCHEM has the features that allow the reading and uploading of data for binary nonadditive mixtures, creating special descriptors for mixtures, and validating models. It is a useful Web-based tool for the modeling and prediction of mixtures of chemical compounds [84].

3.2.3 Chembench

Chembench [85] is a Web-based tool for analyzing the experimental chemical structure-activity data (QSAR modeling and prediction). It provides a broad range of tools for data visualization and embeds a rigorous workflow for creating and validating the predictive QSAR models and using them for virtual screening of chemical libraries to prioritize the compound selection for drug discovery and/or chemical safety assessment. It supports model building with kNN [86] and random forest [87] techniques. User may predict a specific activity or a spectrum of activities for a virtual chemical library or a single compound (available libraries include NCI diversity set (http://dtp.nci.nih.gov/branches/dscb/diversity_explanation.html), DrugBank [88], ChEMBL (http://www.ebi.ac.uk/chembldb/), and Wombat [89]); user may also upload his own library [85].

3.2.4 Spectral Game

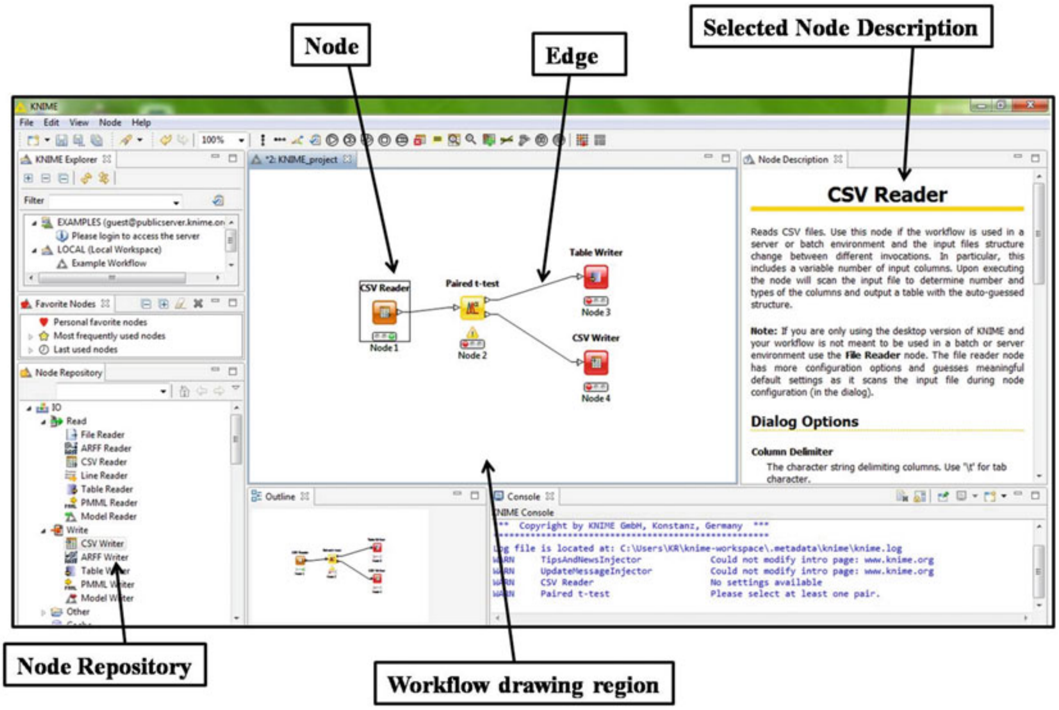
Spectral Game is a Web-based game where players try to match molecules to various forms of interactive spectra, including 1D/2D NMR, mass spectra, and infrared spectra. Player earns one point for every correct answer, and the play continues until player supplies the incorrect answer. The game is usually played by using a Web browser interface. The spectra are uploaded as Open Data to ChemSpider in JCAMP-DX format and are used for the problem sets together with structures extracted from the website. The spectra are displayed using JSpecView, an open-source spectrum viewing applet, which affords zooming and integration. The application of the game is also utilized for the teaching of proton NMR spectroscopy [90].

3.2.5 ACD/I-Lab Tool

ACD/I-Lab [91] is a Web-based prediction engine which provides structure-based predictions of the following properties: free basic physicochemical properties and IUPAC naming capabilities for structures containing 50 atoms or less, advanced physicochemical properties (absolve, boiling point/vapor pressure, adsorption coefficient/BCF, Log*P*, log*D*, pK_a, solubility, etc.), ADME characteristics (bioavailability, absorption, active transport, plasma binding, Vd, Pgp inhibitors, and Pgp substrates), toxicity hazards (AMES test, genotoxicity, aquatic toxicity, health effects, endocrine disruption, MRDD), NMR spectra and chemical shifts for ^{13}C, and systematic chemical nomenclature and structure generation (IUPAC, index names). Several of these predictions are free, while others require licenses.

4 Workflows

Workflow systems in various fields including cheminformatics allow users/scientists to define, manage, and execute time-consuming processes in succession and/or to perform recurring task effectively. A workflow process is logically carried out using a GUI (*see* Fig. 1) where various types of nodes or software components are available for connection through edges or pipes that define the workflow process. Here, nodes are defined by three parameters, i.e., (1) input metadata, (2) algorithms or user-defined parameters or rules, and (3) output metadata. Further, nodes can be connected together only if the output of the previous node represents the mandatory input requirements of the subsequent node. Edges or pipes are the visual representation that provides information about



CSV Reader

Reads CSV files. Use this node if the workflow is used in a server or batch environment and the input files structure change between different invocations. In particular, this includes a variable number of input columns. Upon executing the node will scan the input file to determine number and types of the columns and output a table with the auto-guessed structure.

Note: If you are only using the desktop version of KNIME and your workflow is not meant to be used in a batch or server environment use the **File Reader** node. The file reader node has more configuration options and guesses meaningful default settings as it scans the input file during node configuration (in the dialog).

Dialog Options

Column Delimiter

The character string delimiting columns. Use "\t" for tab character.

```
***  Copyright by KNIME GmbH, Konstanz, Germany  ***
****************************************************
Log file is located at: C:\Users\KR\knime-workspace\.metadata\knime\knime.log
WARN    TipsAndNewsInjector         Could not modify intro page: www.knime.org
WARN    UpdateMessageInjector       Could not modify intro page: www.knime.org
WARN    CSV Reader                  No settings available
WARN    Paired t-test               Please select at least one pair.
```

Fig. 1 Basic components of a workflow system (a snapshot of KNIME workflow was taken for demonstration)

the direction of execution process and are also used for dividing the workflow process [92].

Workflows are increasingly employed and are very useful in cheminformatics since numerous recurring tasks can be automated, which in many cases significantly reduces the manual attention for time-consuming multiple step processes, such as virtual screening via various filters, docking of huge libraries, QSAR studies, etc. There are several commercially well-established workflow systems that are developed for the cheminformatics field such as Pipeline Pilot [93] from Accelrys (*now BIOVIA Pipeline Pilot*) and the InforSense Knowledge Discovery Environment (KDE) from InforSense [94]. Initially, Pipeline Pilot was mainly employed in the cheminformatics field but later extended its functionality to other scientific fields such as next-generation sequencing and imaging. In 2009, the InforSense organization was acquired by IDBS and the KDE has further made progress in the translational medicine field since then [95].

4.1 Open-Source Workflows

In the open-source community, Konstanz Information Miner (KNIME) and Taverna have made a significant impact and offer a wide applicable domain. Cancer Grid (CaGrid) [96] is another open-source workflow system which is an extension of Taverna. *myExperiment* [97] is a collaborative environment where scientists can safely publish their workflows and is the largest freely accessible public repository of scientific workflows. Many examples of KNIME and Taverna workflows that are highly useful for various cheminformatics tasks can be found in the repository.

In this section, we will discuss in detail the recent developments in the above mentioned open-source workflows, especially the new functionalities/plug-in added in these workflow systems.

4.1.1 Taverna

Taverna workflow management system was created by myGrid team (http://www.mygrid.org.uk/) and is currently funded through FP7 projects BioVeL (http://www.biovel.eu/), SCAPE (http://www.scape-project.eu/), and Wf4Ever (http://www.wf4ever-project.org/). It is licensed under the Lesser General Public License (LGPL) Version 2.1. Taverna was initially used in bioinformatics but is now employed in other fields too, including cheminformatics. Recently, the CDK toolkit was combined to Taverna workflow system [98] to improve the handling of various cheminformatics functionalities.

CDK-Taverna: The integration of CDK with Taverna was started in 2005 to extend the functionalities of Taverna in the cheminformatics domain. Both the CDK and Taverna, as well as combined technology, are open source.

The CDK-Taverna 1.0 plug-in provides 164 different workers (note that in case of Taverna, nodes are called workers). These include workers for I/O of various chemical files and line notation formats, databases I/O, and clustering methods and for computing descriptors for atoms, bonds, and molecules. The miscellaneous workers comprise of a substructure filter, a reaction enumerator, an aromaticity detector, etc. The complete list of workers with short description is provided at http://cdk.sourceforge.net/cdk-taverna/workers.html. The architecture of Taverna does not allow the "*loop*" function to control huge data entries to process them one by one. Combining Taverna with CDK allowed workers to act like loops and permit workflows to process large datasets using Taverna's iteration-and-retry mechanism. To allow storage and fast retrieval of up to a million molecules without running into memory limitations, the CDK-Taverna 1.0 supports database system using the PostgreSQL relational database management system (RDBMS) with the open-source Pgchem::tigress extension [98].

The version 2.0 of CDK-Taverna was further developed to extend its usability and strength; CDK-Taverna 2.0 not only made improvements in all workers but also major improvements to the whole platform through a complete setup on the basis of Taverna 2.2 and CDK 1.3.5. This improved version now provides 192 different workers and supports 64-bit computing and multi-core usages by paralleled threads allowing fast in-memory processing and analysis of huge number of molecules, which benefits workers associated with molecular descriptor calculation or machine learning.

The data analysis abilities are also extended with newly added workers that offer access to the open-source WEKA library for clustering and machine learning in addition to dataset division into training and test sets. CDK-Taverna 1.0 offered basic functions for combinatorial chemistry-related reaction enumeration; it supported the use of only two reactants, a single product and one generic group per reactant. As a comparison, the advanced reaction enumeration options employed by CDK-Taverna 2.0 incorporated significant improvements including multi-match detection, no limitation for number of reactants, products or generic groups, and variable R-groups, ring sizes, and atom definitions. This version also provided two groups of workers to compute natural product (NP)-*likeness* core for small molecules.

The CDK-Taverna 2.0 plug-in is built in Java (platform independent) and supported by Windows, Linux, and Mac OS/X (32- and 64-bit). It is released under the GNU LGPL. For its installation, it takes advantage of the plug-in detection manager of Taverna. The plug-in is available at http://www.cdk-taverna.de/plug-in/ and the user can select the desired version. The CDK-Taverna plug-in uses Maven 2 as a build system. It also uses other open-source components such as Bioclipse for visualization of workflow results and Pgchem::tigress as an interface to the database back end for storage of large datasets. A Wiki page is available for this version, which provides general information about the project, documentation, and installation procedure [99, 100].

CaGrid: Cancer Grid (CaGrid) is a workflow system, which prefers to employ and extend Taverna due to a number of benefits including integration with the Web services technology, plug-in architecture which provides an easy integration of third party extensions, and a wide scientific community base. CaGrid is the underlying infrastructure of Cancer Biomedical Informatics Grid (caBIG) and is built on the Globus Toolkit Grid middleware. CaGrid consists of Web services as virtualized access points of data and analytical resources related to cancer detection, diagnosis, treatment, and prevention [96].

Biological Data Interactive Clustering Explorer (BioDICE) Taverna Plug-In: BioDICE is a recent Taverna plug-in [101] that offers clustering analysis capacity and provides visualization of multidimensional biological datasets. The Self-organizing map (SOM) is a well-known unsupervised method for data visualization and clustering. The core algorithm in BioDICE is a fast learning SOM (FLSOM), an improved version of the SOM algorithm that belongs to the category of the emergent self-organizing maps (ESOM). BioDICE is the first Taverna component performing SOM clustering with U-Matrix visualization. Other Taverna plug-ins, i.e., CDK or RapidMiner, were lacking these functionalities; hence BioDICE filled such a gap. The BioDICE plug-in and its documentation, tutorial, workflow, and dataset examples are available at http://biolab.pa.icar.cnr.it/biodice.html.

4.1.2 KNIME

KNIME was developed by a team of software engineers led by Michael Berthold at the University of Konstanz, Germany. It is licensed under GNU GPL. Initially, KNIME entered the market as a data mining tool but rapidly gained popularity in the cheminformatics community. Further combination of KNIME with CDK [102] extended a large amount of cheminformatics functionalities.

KNIME-CDK: The KNIME workflow platform supports a wide range of functionality and has a large number of active users in the cheminformatics community. Thus, CDK is combined with KNIME to wrap the CDK's core functionality and released to the users. This KNIME-CDK plug-in, similar to the CDK-Taverna plug-in, is open source and community driven.

KNIME-CDK [102] consists of various functions which include molecule conversion to/from commonly used formats, generation of signatures, fingerprints, and molecular properties. The plug-in recognizes molecules in CML, SDFile, MDL Mol, InChI, and SMILES formats via the *Molecule to CDK* node and can write SDFiles via the *CDK to Molecule* node. All other operations are performed on the internal CDK molecule representation that includes generation of coordinates, hydrogen manipulator, structure sketcher, atom signatures, common fingerprints (e.g., MACCS and Pubchem), 2D and 3D descriptor values (e.g., XLogP and Lipinski's rule of five), chemical name lookup via OPSIN, and substructure search. It can also be utilized in the management and analysis of chemical libraries through descriptors, conformer analysis via RMSD, and NMR spectra prediction. The KNIME preference page contains a CDK tab to set global visualization preferences, and a renderer is provided to draw the molecules using the JChemPaint library. The different routes employed in the workflow can be run in parallel and the nodes are always run multi-threaded.

KNIME-CDK plug-in has been developed in Java (platform independent) and installed via the KNIME update mechanism. It is build under GNU LGPL license.

5 Databases

A database is a collection of systematically organized or structured repository of indexed information that allows easy retrieval, updating, analysis, and output of data. Freely accessible cheminformatics databases include the databases of chemical structures, proteins, QSAR models, drugs, biological targets, bioactive molecules, etc. The widely used cheminformatics databases, their official links, and a brief description about each database are illustrated in Table 3. Each of these tabulated databases has its own application in specific areas.

Table 3
List of cheminformatics databases, their links and brief description for each database

Sr. No.	Name of database	Database link	Description (reference)
1	QSAR DataBank	http://qsardb.org/repository/	QSAR DataBank [103]
2	VAMMPIRE	http://vammpire.pharmchem.uni-frankfurt.de/vammpire/	Structure-based drug design and optimization [104]
3	PubChem3D	https://pubchem.ncbi.nlm.nih.gov/	Open repository for small molecules and their experimental biological activity [105]
4	MMsINC	http://mms.dsfarm.unipd.it/MMsINC/search/	Chemical structures database [106]
5	CREDO	http://marid.bioc.cam.ac.uk/credo	Protein-ligand interaction database [107]
6	ChemBank	http://chembank.broadinstitute.org/	Small molecule database [108]
7	DrugBank	http://www.drugbank.ca/	Drugs and target information [88]
8	ChemDB	http://cdb.ics.uci.edu	Small molecule database [109]
9	ChemMine	http://chemminedb.ucr.edu/	Compound mining database for chemical genomics [110]
10	National Cancer Institute (NCI) 3D database	http://www.cancer.gov/cancertopics/pdq/cancerdatabase	Anticancer drug database
11	ZINC	http://zinc.docking.org	A free database of commercially available small molecules [113]
12	ChEMBL	https://www.ebi.ac.uk/chembl/	Bioactive molecules with drug-like properties
13	Therapeutic Target Database (TTD)	http://xin.cz3.nus.edu.sg/group/ttd/ttd.asp	Drug database

(continued)

Table 3
(continued)

Sr. No.	Name of database	Database link	Description (reference)
14	PharmGKB	http://www.pharmgkb.org/	Pharmacogenomics knowledge resource
15	STITCH (search tool for interactions of chemicals)	http://stitch.embl.de/	Resource to explore known and predicted interactions of chemicals and proteins
16	SuperTarget	http://bioinf-apache.charite.de/supertarget_v2/	Drugs and target proteins database
17	ChemSpider	http://www.chemspider.com/About.aspx	Chemical structure database [112]

The **QsarDB** [103] could be used to solve everyday QSAR and predictive modeling problems, including applications in the field of predictive toxicology. The utility and benefit of QsarDB can also be applied to a wide variety of other endpoints.

VAMMPIRE [104], a matched molecular pair database, provides valuable information for structure-based lead optimization and for fundamental studies such as understanding protein-ligand interactions.

PubChem3D [105], an addition to the existing contents of PubChem, provides a new dimension to its ability to search, subset, export, visualize, and analyze chemical structures and associated biological data.

MMsINC [106] is a database of nonredundant, richly annotated, and biomedically relevant chemical structures. It has been created to support chemo-centric approach to relate protein pharmacology by ligand chemistry.

CREDO [107] is a novel and comprehensive publicly available database of protein-ligand interactions, which uses contacts as structural interaction fingerprints, implements novel features, and is completely scriptable through its application programming interface.

ChemBank [108] consists of freely available data derived from small molecules and small molecule screens, which can be used to guide chemists in synthesizing novel compounds or libraries and aid biologists in identifying small molecules that block the specific biological pathways of the target protein.

The **DrugBank** database [88] contains the information of both the drug (i.e., chemical, pharmacological, and pharmaceutical compound) and drug targets (i.e., sequence, structure, and pathway). The different categories of drugs include FDA-approved small molecule drugs, FDA-approved biotech (protein/peptide) drugs, nutraceuticals, and experimental drugs.

ChemDB [109] and **ChemMine** [110] are public databases of small molecules and for chemical genomics, respectively. The ChemMine database, a compound mining database, facilitates drug and agrochemical discovery and chemical genomics screening.

The **NCI DIS** 3D database [111] is a collection of 3D structures of drugs. It was built and maintained by the Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Rockville, USA. This 3D database is being used for identifying 3D pharmacophoric features in order to discover novel active anticancer molecules.

ChemSpider [112] is a free online chemical structure database owned by Royal Society of Chemistry. It provides fast access to over 32 million structures, properties, and associated information. By combining and integrating compounds from varied data sources, ChemSpider facilitates the discovery of chemical data from a single online search. It offers both text and structure search for the query compound and provides a unique service to improve this information by curation and annotation.

In summary, the information associated with all these databases is a valuable resource of small molecules/drugs/proteins for medicinal chemists, biologists, cheminformaticians, and bioinformaticians for their exploitation in their respective researches.

6 Conclusion

This chapter highlights freely available cheminformatics resources including toolkits, software, workflow, and databases. The readers will get acquainted to the functionalities as well as the recent advances of these open-source toolkits, workflow systems, ready-to-use software, and freely accessible databases. The information provided would assist the cheminformaticians, including programmers/developers, to explore and utilize these freely available resources to resolve different computational challenges in the cheminformatics field and further contribute in new advancement for the next-generation cheminformatics resources.

References

1. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E (2003) The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bio-informatics. J Chem Inf Comput Sci 43:493–500
2. O'boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. J Cheminform 3:33
3. Landrum G (2013) RDKit: cheminformatics and machine learning software. rdkit.org
4. http://ggasoftware.com/opensource/indigo
5. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner J, Willighagen EL (2006) The blue obelisk interoperability in chemical informatics. J Chem Inf Model 46:991–998
6. O'Boyle NM, Guha R, Willighagen EL, Adams SE, Alvarsson J, Bradley J-C, Filippov IV, Hanson RM, Hanwell MD, Hutchison GR (2011) Open data, open source and open standards in chemistry: the Blue Obelisk five years on. J Cheminform 3:37
7. Spjuth O, Helmus T, Willighagen EL, Kuhn S, Eklund M, Wagener J, Murray-Rust P, Steinbeck C, Wikberg JES (2007) Bioclipse: an open source workbench for chemo- and bioinformatics. BMC Bioinformatics 8:59
8. Hanwell MD, Curtis DE, Lonie DC, Vandermeersch T, Zurek E, Hutchison GR (2012) Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. J Cheminform 4:17
9. Guha R (2006) CDK descriptor calculator GUI. http://www.rguha.net/code/java/cdkdesc.html
10. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL (2006) Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bioinformatics. Curr Pharm Des 12:2111–2120
11. O'Boyle NM, Hutchison GR (2008) Cinfony—combining Open Source cheminformatics toolkits behind a common interface. Chem Cent J 2:24
12. Rahman SA, Bashton M, Holliday GL, Schrader R, Thornton JM (2009) Small Molecule Subgraph Detector (SMSD) toolkit. J Cheminform 1:12
13. Hildebrandt A, Dehof AK, Rurainski A, Bertsch A, Schumann M, Toussaint NC, Moll A, Stockel D, Nickels S, Mueller SC (2010) BALL-biochemical algorithms library 1.3. BMC Bioinformatics 11:531
14. Hinselmann G, Rosenbaum L, Jahn A, Fechner N, Zell A (2011) jCompoundMapper: an open

source Java library and command-line tool for chemical fingerprints. J Cheminform 3:3

15. Hock S, Riedl R (2012) chemf: a purely functional chemistry toolkit. J Cheminform 4:1–19
16. Cao D-S, Xu Q-S, Hu Q-N, Liang Y-Z (2013) ChemoPy: freely available python package for computational biology and chemoinformatics. Bioinformatics 29:1092–1094
17. Cao Y, Charisi A, Cheng L-C, Jiang T, Girke T (2008) ChemmineR: a compound mining framework for R. Bioinformatics 24:1733–1734
18. Cao D-S, Xiao N, Xu Q-S, Chen AF (2014) Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds, and their interactions. Bioinformatics. doi:10.1093/bioinformatics/btu1624
19. http://wiki.chemkit.org/Main_Page
20. Herraez A (2006) Biomolecules in the computer: Jmol to the rescue. Biochem Mol Biol Educ 34:255–261
21. Krause S, Willighagen E, Steinbeck C (2000) JChemPaint—using the collaborative forces of the internet to develop a free editor for 2D chemical structures. Molecules 5:93–98
22. https://github.com/cdk/cdk/blob/master/AUTHORS
23. Bashton M, Nobeli I, Thornton JM (2006) Cognate ligand domain mapping for enzymes. J Mol Biol 364:836–852
24. Rojas-Cherto M, Kasper PT, Willighagen EL, Vreeken RJ, Hankemeier T, Reijmers TH (2011) Elemental composition determination based on MSn. Bioinformatics 27:2376–2383
25. Steinbeck C (2001) SENECA: a platform-independent, distributed, and parallel system for computer-assisted structure elucidation in organic chemistry. J Chem Inf Comput Sci 41:1500–1507
26. Steinbeck C, Kuhn S (2004) NMRShiftDB—compound identification and structure elucidation support through a free community-built web database. Phytochemistry 65:2711–2717
27. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474
28. http://cdk.sourceforge.net/old_web/software.html
29. http://www.rdkit.org/docs/Overview.html
30. http://www.rdkit.org/docs/Overview.html#the-contrib-directory
31. http://rdkit.org/RDKit_Docs.current.pdf
32. http://sourceforge.net/projects/openbabel/files/stats/timeline?dates=2001-11-25+to+2014-11-14
33. http://scholar.google.co.in/scholar?hl=en&as_sdt=0,5&q=openbabel
34. http://www.eyesopen.com/toolkits
35. http://www.eyesopen.com/academic
36. http://openbabel.org/
37. O'Boyle NM, Morley C, Hutchison GR (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. Chem Cent J 2:5
38. http://creativecommons.org/licenses/by/3.0/
39. Pavlov D, Rybalkin M, Karulin B, Kozhevnikov M, Savelyev A, Churinov A (2011) Indigo: universal cheminformatics API. J Cheminform 3:P4
40. http://jcompoundmapper.sourceforge.net/
41. http://www.scala-lang.org/
42. https://github.com/stefan-hoeck/chemf
43. Jarvis RM, Broadhurst D, Johnson H, O'Boyle NM, Goodacre R (2006) PYCHEM: a multivariate analysis package for python. Bioinformatics 22:2565–2566
44. Wang Y, Backman TWH, Horan K, Girke T (2013) fmcsR: mismatch tolerant maximum common substructure searching in R. Bioinformatics 29:2792–2794
45. Cao Y, Jiang T, Girke T (2010) Accelerated similarity searching and clustering of large compound sets by geometric embedding and locality sensitive hashing. Bioinformatics 26:953–959
46. Hoksza D, Skoda P, Vorsilak M, Svozil D (2014) Molpher: a software framework for systematic chemical space exploration. J Cheminform 3:32
47. Schling B (2011) The boost C++ libraries. XML Press, Laguna Hills, CA
48. Hu B, Lill MA (2014) PharmDock: a pharmacophore-based docking program. J Cheminform 6:1–14
49. Plewczynski D, Lazniewski M, Augustyniak R, Ginalski K (2011) Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. J Comput Chem 32:742–755
50. Li X, Li Y, Cheng T, Liu Z, Wang R (2010) Evaluation of the performance of four molecular docking programs on a diverse set of protein–ligand complexes. J Comput Chem 31:2109–2125
51. Cereto-Massague A, Ojeda MJ, Joosten RP, Valls C, Mulero M, Salvado MJ, Arola-Arnal

A, Arola L, Garcia-Vallve S, Pujadas G (2013) The good, the bad and the dubious: VHELIBS, a validation helper for ligands and binding sites. J Cheminform 5:36

52. Sehnal D, Varekova RS, Berka K, Pravda L, Navratilova V, Banas P, Ionescu C-M, Otyepka M, Koca J (2013) MOLE 2.0: advanced approach for analysis of biomacromolecular channels. J Cheminform 5:39
53. Petrek M, Kosinova P, Koca J, Otyepka M (2007) MOLE: a Voronoi diagram-based explorer of molecular channels, pores, and tunnels. Structure 15:1357–1363
54. Yaffe E, Fishelovitch D, Wolfson HJ, Halperin D, Nussinov R (2008) MolAxis: efficient and accurate identification of channels in macromolecules. Proteins 73:72–86
55. Yaffe E, Fishelovitch D, Wolfson HJ, Halperin D, Nussinov R (2008) MolAxis: a server for identification of channels in macromolecules. Nucleic Acids Res 36:W210–W215
56. Chovancova E, Pavelka A, Benes P, Strnad O, Brezovsky J, Kozlikova B, Gora A, Sustr V, Klvana M, Medek P (2012) CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. PLoS Comput Biol 8:e1002708
57. Khashan R (2012) FragVLib a free database mining software for generating "Fragment-based Virtual Library" using pocket similarity search of ligand-receptor complexes. J Cheminform 4:1–6
58. Ekins S, Clark AM, Sarker M (2013) TB Mobile: a mobile app for anti-tuberculosis molecules with known targets. J Cheminform 5:13
59. Bienfait B, Ertl P (2013) JSME: a free molecule editor in JavaScript. J Cheminform 5:24
60. Gutlein M, Karwath A, Kramer S (2012) CheS-Mapper—chemical space mapping and visualization in 3D. J Cheminform 4:7
61. Le Guilloux V, Arrault A, Colliandre L, Bourg SP, Vayer P, Morin-Allory L (2012) Mining collections of compounds with Screening Assistant 2. J Cheminform 4:1–16
62. Sud M, Fahy E, Subramaniam S (2012) Template-based combinatorial enumeration of virtual compound libraries for lipids. J Cheminform 4:23
63. Cereto-Massague A, Guasch L, Valls C, Mulero M, Pujadas G, Garcia-Vallve S (2012) DecoyFinder: an easy-to-use python GUI application for building target-specific decoy sets. Bioinformatics 28:1661–1662
64. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49:6789–6801
65. Irwin JJ (2008) Community benchmarks for virtual screening. J Comput Aided Mol Des 22:193–199
66. Wallach I, Lilien R (2011) Virtual decoy sets for molecular docking benchmarks. J Chem Inf Model 51:196–202
67. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 42:1273–1280
68. Kerber A, Laue R, Gruner T, Meringer M (1998) MOLGEN 4.0. MATCH Commun Math Comput Chem 37:205–208
69. Peironcely JE, Rojas-Cherto M, Fichera D, Reijmers TH, Coulier L, Faulon J-L, Hankemeier T (2012) OMG: open molecule generator. J Cheminform 4:21
70. Brefo-Mensah EK, Palmer M (2012) mol2chemfig, a tool for rendering chemical structures from molfile or SMILES format to LATEX code. J Cheminform 4:24
71. Lawson KR, Lawson J (2012) LICSS—a chemical spreadsheet in microsoft excel. J Cheminform 4:1–7
72. Wilhelm J-H (2011) MyChemise: a 2D drawing program that uses morphing for visualisation purposes. J Cheminform 3:53
73. Tosco P, Balle T, Shiri F (2011) Open3DALIGN: an open-source software aimed at unsupervised ligand alignment. J Comput Aided Mol Des 25:777–783
74. Norgan AP, Coffman PK, Kocher J-P, Katzmann DJ, Sosa CP (2011) Multilevel parallelization of AutoDock 4.2. J Cheminform 3:12
75. Demir-Kavuk O, Bentzien J, Muegge I, Knapp E-W (2011) DemQSAR: predicting human volume of distribution and clearance of drugs. J Comput Aided Mol Des 25:1121–1133
76. Jimmy R, Laurence M, Serge P (2009) Shape: automatic conformation prediction of carbohydrates using a genetic algorithm. J Cheminform 1:1–7
77. Rijnbeek M, Steinbeck C (2010) OrChem: an open source chemistry search engine for Oracle. J Cheminform 2:P28
78. Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S (2013) QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. J Comput Chem 34:2121–2132

79. http://www.vega-qsar.eu/index.php
80. http://www.oecd.org/chemicalsafety/risk-assessment/theoecdqsartoolbox.htm
81. http://www.chemaxon.com/free-software/
82. http://teqip.jdvu.ac.in/QSAR_Tools/
83. Wang X, Chen H, Yang F, Gong J, Li S, Pei J, Liu X, Jiang H, Lai L, Li H (2014) iDrug: a web-accessible and interactive drug discovery and design platform. J Cheminform 6:1–8
84. Oprisiu I, Novotarskyi S, Tetko IV (2013) Modeling of non-additive mixture properties using the Online CHEmical database and Modeling environment (OCHEM). J Cheminform 5:4
85. Walker T, Grulke CM, Pozefsky D, Tropsha A (2010) Chembench: a cheminformatics workbench. Bioinformatics 26:3000–3001
86. Zhang L, Zhu H, Oprea T, Golbraikh A, Tropsha A (2008) QSAR modeling of the blood-brain barrier permeability for diverse organic compounds. Pharm Res 25:1902–1914
87. Breiman L (2001) Random forests. Mach Learn 1:5–32
88. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 34:D668–D672
89. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Mracec M (2004) WOMBAT: world of molecular bioactivity. Chemoinf. Drug Disc., Wiley-VCH, New York, 223–239
90. Bradley J-C, Lancashire RJ, Lang ASID, Williams AJ (2009) The Spectral Game: leveraging Open Data and crowdsourcing for education. J Cheminform 1:1–10
91. http://www.acdlabs.com/resources/ilab/
92. Tiwari A, Sekhar AKT (2007) Workflow based framework for life science informatics. Comput Biol Chem 31:305–319
93. http://accelrys.com/products/pipeline-pilot/
94. http://www.inforsense.com/
95. Warr WA (2012) Scientific workflow systems: Pipeline pilot and KNIME. J Comput Aided Mol Des 26:1–4
96. Tan W, Madduri R, Nenadic A, Soiland-Reyes S, Sulakhe D, Foster I, Goble CA (2010) CaGrid Workflow Toolkit: a taverna based workflow tool for cancer grid. BMC Bioinformatics 11:542
97. http://www.myexperiment.org/
98. Kuhn T, Willighagen EL, Zielesny A, Steinbeck C (2010) CDK-Taverna: an open workflow environment for cheminformatics. BMC Bioinformatics 11:159
99. http://cdktaverna2.ts-concepts.de/wiki/index.php?title=Main_Page
100. Truszkowski A, Jayaseelan KV, Neumann S, Willighagen EL, Zielesny A, Steinbeck C (2011) New developments on the cheminformatics open workflow environment CDK-Taverna. J Cheminform 3:54
101. Fiannaca A, La Rosa M, Di Fatta G, Gaglio S, Rizzo R, Urso A (2014) The BioDICE Taverna plugin for clustering and visualization of biological data: a workflow for molecular compounds exploration. J Cheminform 6:1–6
102. Beisken S, Meinl T, Wiswedel B, de Figueiredo LF, Berthold MR, Steinbeck C (2013) KNIME-CDK: workflow-driven cheminformatics. BMC Bioinformatics 14:257
103. Ruusmann V, Sild S, Maran U (2014) QSAR DataBank—an approach for the digital organization and archiving of QSAR model information. J Cheminform 6:25
104. Weber J, Achenbach J, Moser D, Proschak E (2013) VAMMPIRE: a matched molecular pairs database for structure-based drug design and optimization. J Med Chem 56:5203–5207
105. Bolton E, Chen J, Kim S, Han L, He S, Shi W, Simonyan V, Sun Y, Thiessen PA, Wang J (2011) PubChem3D: a new resource for scientists. J Cheminform 3:32
106. Masciocchi J, Frau G, Fanton M, Sturlese M, Floris M, Pireddu L, Palla P, Cedrati F, Rodriguez P, Moro S (2009) MMsINC: a large-scale chemoinformatics database. Nucleic Acids Res 37:D284–D290
107. Schreyer A, Blundell T (2009) CREDO: a protein-ligand interaction database for drug discovery. Chem Biol Drug Des 73:157–167
108. Seiler KP, George GA, Happ MP, Bodycombe NE, Carrinski HA, Norton S, Brudz S, Sullivan JP, Muhlich J, Serrano M (2008) ChemBank: a small-molecule screening and cheminformatics resource database. Nucleic Acids Res 36:D351–D359
109. Chen J, Swamidass SJ, Dou Y, Bruand J, Baldi P (2005) ChemDB: a public database of small molecules and related chemoinformatics resources. Bioinformatics 21:4133–4139

110. Girke T, Cheng L-C, Raikhel N (2005) ChemMine. A compound mining database for chemical genomics. Plant Physiol 138:573–577

111. Milne GWA, Nicklaus MC, Driscoll JS, Wang S, Zaharevitz D (1994) National Cancer Institute drug information system 3D database. J Chem Inf Comput Sci 34:1219–1224

112. Pence HE, Williams A (2010) ChemSpider: an online chemical information resource. J Chem Educ 87:1123–1124

113. Irwin JJ, Shoichet BK (2005) ZINC—a free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182