Open source drug discovery with Bioclipse

Ola Spjuth[1⋆], Lars Carlsson[2], Jonathan Alvarsson[1], Valentin Georgiev[1], Egon Willighagen[3], and Martin Eklund[1]

[1] Department of Pharmaceutical Biosciences, Uppsala University, P.O. Box 591, SE-751 24 Uppsala, Sweden
[2] AstraZeneca Research and Development, SE-431 83 Mölndal, Sweden
[3] Dept. of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, Maastricht, The Netherlands

Abstract. We present the open source components for drug discovery that has been developed and integrated into the graphical workbench Bioclipse. Building on a solid open source cheminformatics core, Bioclipse has advanced functionality for managing and visualizing chemical structures and related information. The features presented here include QSAR/QSPR modeling, various predictive solutions such as decision support for chemical liability assessment, site-of-metabolism prediction, virtual screening, and knowledge discovery and integration. We demonstrate the utility of the described tools with examples from computational pharmacology, toxicology, and ADME. Bioclipse is used in both academia and industry, and is a good example of open source leading to new solutions for drug discovery.

Background

The use of open source tools in drug discovery is gaining momentum and is an important research field [1–3]. Being able to access, inspect, and extend software reduces time to develop new tools and opens up new possibilities for using algorithms and systems in novel applications. Open source tools are also in many cases easier to integrate into pipelines and workflows, and several open source tools exist for this purpose such as Taverna [4], Knime [5], and Galaxy [6]. Also interesting are systems that integrate software in a workbench, where the user patterns differ from workflow software in that they are more focused on iterative discovery than reproducibility, and with features to operate on various file formats. Examples of such workbenches are Gaggle [7], ISYS [8], and Bioclipse [9, 10].

Bioclipse is an open source workbench and platform for the life sciences. Based on the Eclipse Rich Client Platform (RCP), Bioclipse inherits an advanced plugin architecture and can easily be extended in virtually any direction. Equipped with a scripting engine, all Bioclipse functionality is available from the graphical user interface (GUI) as well as from scripts. This allows for both iterative science with graphical tools, as well as for batch processing where repetitive

⋆ ola.spjuth@farmbio.uu.se

tasks can be automated and entire analyses shared, for example using social sites such as myExperiment [11]. Other core features include a reporting framework for producing printable reports, which can be designed using graphical standalone tools and populated with results in Bioclipse. Also, part of Bioclipse core is the consumption of web services for remote functionality [12].

Bioclipse has an open approach not only to source code but also to data and standards (ODOSOS — Open Data, Open Source, Open Standards), which also are the three pillars the Blue Obelisk movement promotes in the mission towards interoperability in cheminformatics [13, 14]. Data standards and ontologies are of great importance (agreeing on naming and semantics is a non-trivial and tedious process), as well as standards for software interoperability. Bioclipse is built on the OSGi standard [15] for Java module interoperability which is becoming popular in bioinformatics with projects such as Cytoscape [16], and Taverna [4] adopting the standard.

In this paper we summarize a subset of the tools in Bioclipse and demonstrate their usage in drug discovery (see Table 1 for a summary). Applications range from computational toxicology to ADME and metabolism predictions, but also include frameworks for developing and deploying new tools and algorithms. This has lead to a user base that includes both developers, scientists, and teachers on various levels.

Plugin	**Description**	**Reference**
Cheminformatics	Management, editing, and visualization of chemical structures	[9, 10]
QSAR	Functions and graphical tools for setting up Quantitative Structure-Activity Relationship datasets, supporting the open standard QSAR-ML	[17]
Decision Support	Framework and user interface for applying predictive models on chemical structures	[18]
MetaPrint2D	Site-of-metabolism prediction for CYP-mediated biotransformations	[19]
Brunn	Laboratory information management system (LIMS) for drug screening using microplates	[20]
OpenTox	Computational toxicology predictions using the OpenTox infrastructure	[21]
RDF	Framework for working with linked data	[22]
HIVDRC	Prediction of drug susceptibility for HIV proteases and reverse transcriptases	[23]

Table 1. A list of plugins and features for Bioclipse relevant to drug discovery.

Cheminformatics

Cheminformatics in Bioclipse provides core functionality for working with chemical structures and covers loading and saving in various file formats, editing, calculations, and visualizations. Bioclipse uses the Chemistry Development Kit (CDK) [24] for this core cheminformatics functionality. The JChemPaint editor is used for visualization and editing of 2D structures [25] and Jmol is used for depicting 3D structures [26]. For external storage Bioclipse primarily focuses on the MDL molfile, SMILES, InChI, and CML [27] serialization formats. Generation of 3D geometries is performed with Balloon [28].

Besides this basic cheminformatics functionality, Bioclipse also supports chemical names. There is search functionality to find chemicals in ChemSpider and PubChem, and with the OPSIN [29] plugin handling IUPAC names is trivial. Further support is provided by a plugin for the Chemical Resolver Identifier (`cactus.nci.nih.gov/chemical/structure`).

Decision support

Predictive modeling is a core component in drug safety for assessing chemical liabilities, and aims at predicting the toxicity of novel chemical structures in silico [30]. Several approaches exist for this, including database lookups, structural alerts, and QSAR, where most are based on the assumption that similar structures are similar in response (e.g. activity or property). While this hypothesis is not valid in all cases, the process is fast and used extensively in drug discovery [31]. Structural alerts are chemical substructures or structural patterns which have been linked to a response, and are commonly defined manually by experts in the field.

Modeling with QSAR/QSPR (Quantitative Structure-Activity/Property Relationships) is another common method in ligand-based drug discovery. By relating chemical structures to a response using mathematical or statistical methods, predictions can be made for novel compounds [32]. One key factor is the numerical representation of chemical structures (called descriptors) for the analyses, which should capture the chemistry and allow for inter/extrapolation in chemical space. Such descriptors can for example be based on physicochemical properties [33], chemical fingerprints [34, 35], or chemical fragments [36].

A core component for drug discovery in Bioclipse is the Decision Support system, providing a platform for accessing computational models, such as database lookups, structural alerts, and QSARs from a unified GUI [18]. The use case is: given a chemical structure, present in a condensed manner the relevant information from as many sources as possible, including predictive models. Bioclipse Decision Support presents several general properties which are important for drug discovery:

Predict on a local computer Bioclipse Decision Support can execute simultaneous (predictive) models for novel compounds on the local computer, without requiring an internet connection. This is important as many drug discovery projects

are reluctant or prohibited (for security reasons) to send information about drug leads over untrusted networks. It also means that predictions can be made on laptops while traveling, and enables fast response without the latency of networks and remote servers. However, offline predictions are naturally restricted to the data and models that can be kept on the local computer.

Interpretable results Predictions in the domain of drug discovery can be done for filtering purposes, where the objective is to rapidly predict a property (such as logD, logP, or solubility) and prioritize between a large set of structures. But also for explanatory purposes and for decision support on how to make favorable structural modifications [37]. Bioclipse Decision Support provides several means for interpretations of models. In the example of a structural alert, the corresponding substructure is highlighted in the query molecule. The chemical structures of identified exact matches and near neighbors in databases can be displayed, and for QSAR models it is possible to get a graphical interpretation of nonlinear models [38, 39].

Fast predictions With a focus on fast response-time from predictions it is possible to execute predictions for a molecule, make changes to the chemical structure, rerun the predictions, and get updated predictions in near-real time. This allows for testing how predictions would be affected by different hypotheses regarding structural modifications.

Computational toxicology

There are several features in Bioclipse for computational toxicology. Bioclipse Decision Support has been used to present models based on open data for the endpoints Ames mutagenicity [43], carcinogenicity [44], and AhR inhibition [45]. Structural alerts were included that have been linked to toxicity, also sometimes referred to as toxicophores [43, 46]. Experiences with multiple simultaneous models is very illustrative for the multi-objective optimization problem that drug discovery constitutes. One example of various hypotheses-trying on structural modifications is available in Figure 1.

Another feature in Bioclipse is the integration of predictive toxicity models from the OpenTox project [21, 47]. Here the Bioclipse Decision Support was extended to present the online predictive services of OpenTox alongside existing models in Bioclipse, allowing users to easily access all available OpenTox models. We believe that the combination of models running on the local computer with optional remote services in a graphical workbench gives a very flexible workbench for predictive toxicology.

ADME predictions

ADME (absorption, distribution, metabolism, and excretion) describes the disposition of a drug within an organism, and is well recognized as an important element in small molecule drug discovery and development [50]. Of primary aim in

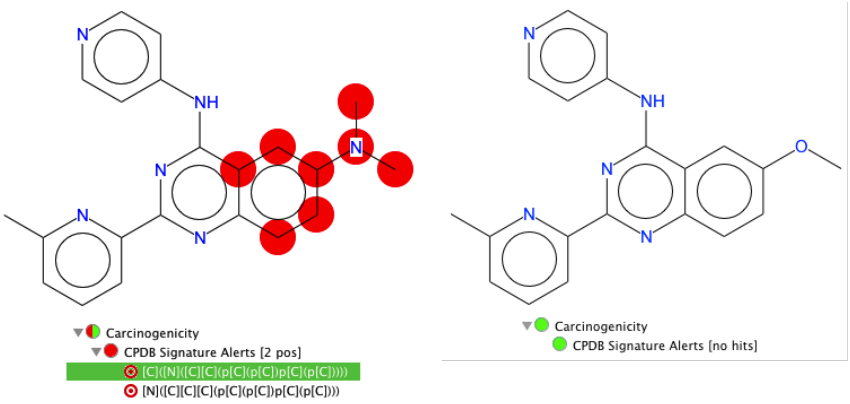

Fig. 1. (a) The compound TCMDC-135308 was selected from the Tres Cantos Antimalarial Compound Set (TCAMS) - a collection of compounds that inhibits growth of Plasmodium falciparum by at least 80 % at a concentration of 2 νM [40] - and which also is similar (Tanimoto 0.915) to quinazoline 3d [41], a potent human-TGFβ1 inhibitor. The compound was accessed from the ChEMBL Neglected Tropical Disease Database (https://www.ebi.ac.uk/chemblntd) and imported into Bioclipse Decision Support, and we note a structural alert for carcinogenicity with the corresponding atoms highlighted. (b) Replacing the dimethylamino group of TCMDC-135308 by a methoxy group shows no remaining carcinogenicity alerts. Example adapted from [42].

ADME modeling is to promote candidates presenting ADME-properties within certain intervals, such as the well known (but debated) Lipinski rule of having an octanol-water partition coefficient (logP) of less than 5 [51]. Several ADME models for Bioclipse Decision Support are available, including water solubility [52], Blood-Brain-Barrier Penetration [53], and P-Glycoprotein (PgP) Inhibition [54].

It is estimated that over 70 % of marketed small molecule drugs are eliminated primarily by CYP enzymes [55]. MetaPrint2D is a tool for predicting site-of-metabolism for CYP-mediated biotransformations, and which has been integrated into Bioclipse [19]. By describing a large training set of biotransformations with circular fingerprints [56], we achieved fast and accurate predictions. The speed opens up for making site-of-metabolism predictions early available in the drug discovery process.

Screening

Drug discovery encompasses various forms of screening, such as high throughput screening [58], and virtual screening [59]. Brunn is a laboratory information system for high throughput screening [20] on microtiter plates which is based on Bioclipse. It provides the user with a point-and-click interface for managing a multitude of plate layouts. Excel like formulæ can be defined in order to calcu-

late values for wells such as for example survival index or other score functions. Coefficient of Variation (CV %) is calculated for replicates as an help in identifying positional variability, and other calculations can be manually created. Brunn handles dilution series and produces dose response curves with IC_{50} values which are presented in printable reports using the reporting system in Bioclipse.

In a recent study, predictive models for cancer growth inhibition were built on the U251 dataset of the National Cancer Institute Developmental Therapeutics Program (NCI60 cell line panel), together with mutagenicity models based on Ames data [43], using the statistical language R integrated in Bioclipse [60]. Approved compounds in drugbank [61] were then prioritized using these models for inhibiting cancer growth and with low predicted mutagenicity profile. The results showed that indeed a number of anti-cancer drugs were present among the top ranked drugs.

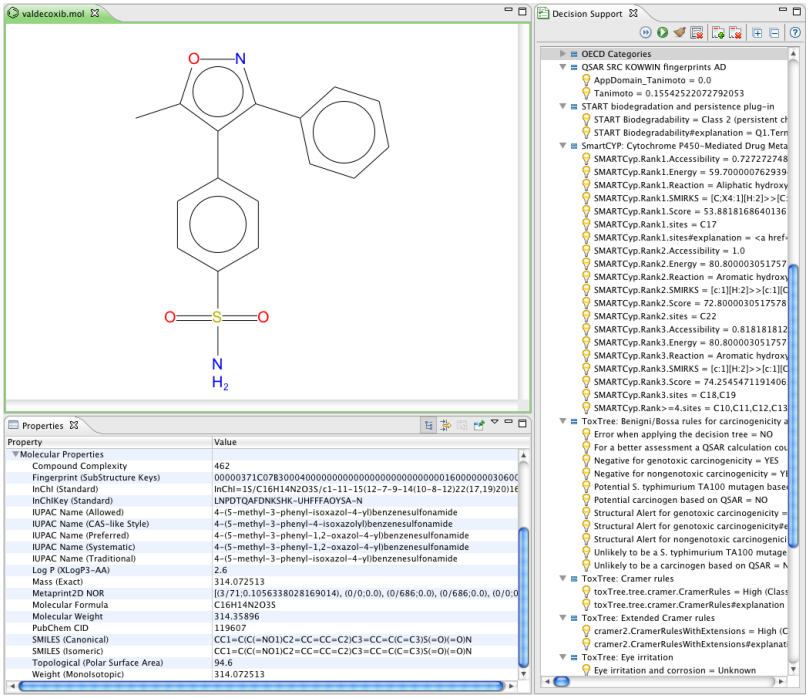

Fig. 2. Computational toxicology predictions using the OpenTox models. For example we see the results of tools such as SmartCyp [48] and ToxTree [49]; these predictions are accessed remotely over the internet and via the OpenTox infrastructure [21].

Knowledge discovery

An important feature of a workbench for drug discovery is the possibility to discover and aggregate the wealth of information that can be found on public websites: for decision making it is important to know what has already been done before. Bioclipse provides functionality to search for structures in PubChem and ChemSpider, but it also allows for searching the internet for drug-properties using Semantic Web technologies. This is an increasingly used web technology, used for example in drug discovery by the IMI-funded Open PHACTS project [62] and in safety by the EU/Colipa-funded ToxBank project for alternative testing (`http://www.toxbank.net/`). Using this approach various databases can be seamlessly searched, including those provided by the Linked Open Drug Data community of the W3C Health Care and Life Sciences interest group, such as ChEMBL [63], ChEBI [64], DrugBank [61], SIDER [65], and DBPedia [66]. An important difference with regular web databases is that semantic web technology allows for following of active links between databases, very much like clicking hyperlinks in web pages.

The *linked data* nature of this approach allows the web to be crawled for information about those drugs. The search can be initiated by e.g. a chemical

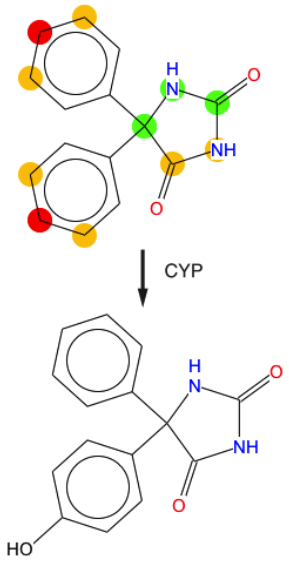

Fig. 3. The MetaPrint2D prediction for the drug Phenytoin (top structure); atoms with high probability of being site-of-metabolism are colored in red. We see that the predictions are in concordance with the results published in [57], where it is noted that CYP facilitates 4-hydroxylation of phenytoin to yield 5-(-4-hydroxyphenyl)-5-phenylhydantoin (HPPH, bottom structure).

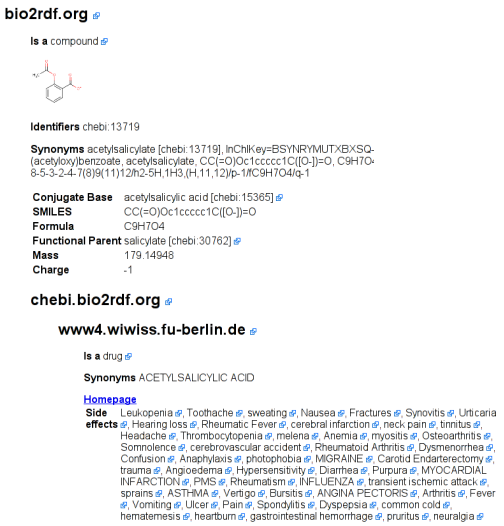

Fig. 4. Search results for aspirin found by following Linked Open Drug Data links between databases. The top half shows properties found in the ChEBI database as shared by Bio2RDF. The bottom half shows side effects from the SIDER database, redistributed as Linked Data by the Free University of Berlin.

structure, for example a SMILES string or MDL molfile, or from a drug name resolved with the Chemical Resolver Identifier discussed earlier. This will results in a few calls out to databases after which the semantic web links will be followed and further information about the seeding drug identified. Common ontologies are recognized and used to extract specific data; for example, the Cheminformatics Ontology (CHEMINF) [67] can be used to extract drug properties. Using this approach, online databases can be queried for potentially interesting information, including pharmacology, ADME properties, interactions with other drugs and target, etc. (see Figure 4), and the result may serve as initial information for more in-depth studies.

In addition, these semantic web approaches can also be used to simply mine a single database, such as the ChEMBL database [63]. We used this approach earlier to search and aggregate chemical structures and matching properties. We previously published a study where this approach was used to perform a substructure mining to find molecular fragments specific for particular protein families, and to create QSAR model from these structures for predicting IC_{50} [22].

Discussion

In this article we have described many of the features in Bioclipse that are used in drug discovery. During the years, many of these tools have matured and

gained acceptance in both academic and industrial settings. Examples include the MetaPrint2D and Decision Support features which now form part of AstraZeneca R&D's toolbox for computational toxicology and ADME predictions. Another example is the development and adoption of Bioclipse as a workbench for predictive toxicology in the OpenTox community. The latter is also a good example of the power of open source and open standards, where it was possible, without substantial effort, to integrate and consume remote predictive OpenTox services from within Bioclipse. There are many other features in Bioclipse that are more distantly related to drug discovery, such as the bioinformatics features for working with sequences (DNA, RNA, protein), the statistical analysis framework building on R, and many other unpublished features - see the Bioclipse website (www.bioclipse.net) and wiki for more extensive feature listings.

It has been argued that the prime benefit of open source is the reduction of R&D costs, but there are also other substantial advantages. For example, the time for implementing new algorithms can be reduced substantially when building on existing open source frameworks and libraries. Another implication is that users can locate and fix bugs themselves, which can save a lot of time. The level of trust in predictions is higher when scientists can inspect the code. Also, the voluntary participation of fellow researchers can improve tool quality and e.g. identify failing use cases that otherwise would have gone unnoticed. The reuse of code in another setting with slight modifications can also open up for novel applications, benefitting the original implementor.

One obstacle with open source drug discovery is the increasing amount of incompatible software tools being developed. In order for open source drug discovery to reach its full potential, the burden of integrating such tools and data into larger frameworks needs to be reduced. Equally important is the simple provisioning of tools for end-users, who normally do not possess extensive computer skills. Bioclipse addresses both of these issues with an extensible framework where new algorithms, visualizations, and other tools can easily be integrated and provisioned via a software update function for downloading the latest features — without requiring them to be located on a central server. The implementations are then exposed via graphical user interfaces (such as editors and wizards) that scientists are used to work with, hiding technical solutions.

There are today hundreds of software frameworks and tools that are used as standalone packages. Not all are suitable for integration in a workbench like Bioclipse, but a general advancement of standards for both data and software would enable more interoperable solutions and reduce data loss in file format conversions. The Bioclipse project takes active part in such working groups when possible, such as the Blue Obelisk [13] and OpenTox [47]. The adoption of the plugin-architecture OSGi by widely used open source software in bioinformatics, such as Taverna, Knime, and Cytoscape ensures the future compatibility of Bioclipse with other tools, and opens up for the idea of a marketplace for computational drug discovery where the best available methods are easily available, to the benefit of everyone.

Acknowledgements

The authors would like to acknowledge the Bioclipse developers and the OpenTox Project, especially Roman Affentrager who developed the original Malaria use case for Bioclipse Decision Support. Financial support was provided by Swedish VR-2011-6129 and the Swedish strategic research program eSSENCE.

Bibliography

[1] DeLano, W. L. The case for open-source software in drug discovery. *Drug Discov Today* **2005**, *10*, 213–217.

[2] Edwards, A. Open-source science to enable drug discovery. *Drug Discov Today* **2008**, *13*, 731–733.

[3] Geldenhuys, W. J.; Gaasch, K. E.; Watson, M.; Allen, D. D.; Van der Schyf, C. J. Optimizing the use of open-source software applications in drug discovery. *Drug Discov Today* **2006**, *11*, 127–132.

[4] Oinn, T.; Addis, M.; Ferris, J.; Marvin, D.; Senger, M.; Greenwood, M.; Carver, T.; Glover, K.; Pocock, M. R.; Wipat, A.; Li, P. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **2004**, *20*, 3045–3054.

[5] Warr, W. A. Scientific workflow systems: Pipeline Pilot and KNIME. *J Comput Aided Mol Des* **2012**,

[6] Giardine, B.; Riemer, C.; Hardison, R. C.; Burhans, R.; Elnitski, L.; Shah, P.; Zhang, Y.; Blankenberg, D.; Albert, I.; Taylor, J.; Miller, W.; Kent, W. J.; Nekrutenko, A. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **2005**, *15*, 1451–5.

[7] Shannon, P. T.; Reiss, D. J.; Bonneau, R.; Baliga, N. S. The Gaggle: an open-source software system for integrating bioinformatics software and data sources. *BMC Bioinformatics* **2006**, *7*, 176.

[8] Siepel, A.; Farmer, A.; Tolopko, A.; Zhuang, M.; Mendes, P.; Beavis, W.; Sobral, B. ISYS: a decentralized, component-based approach to the integration of heterogeneous bioinformatics resources. *Bioinformatics* **2001**, *17*, 83–94.

[9] Spjuth, O.; Helmus, T.; Willighagen, E. L.; Kuhn, S.; Eklund, M.; Wagener, J.; Murray-Rust, P.; Steinbeck, C.; Wikberg, J. E. S. Bioclipse: an open source workbench for chemo- and bioinformatics. *BMC Bioinformatics* **2007**, *8*, 59.

[10] Spjuth, O.; Alvarsson, J.; Berg, A.; Eklund, M.; Kuhn, S.; Mäsak, C.; Torrance, G.; Wagener, J.; Willighagen, E. L.; Steinbeck, C.; Wikberg, J. E. S. Bioclipse 2: a scriptable integration platform for the life sciences. *BMC Bioinformatics* **2009**, *10*, 397.

[11] Goble, C. A.; Bhagat, J.; Aleksejevs, S.; Cruickshank, D.; Michaelides, D.; Newman, D.; Borkum, M.; Bechhofer, S.; Roos, M.; Li, P.; De Roure, D. myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res* **2010**, *38 Suppl*, W677–82.

[12] Wagener, J.; Spjuth, O.; Willighagen, E. L.; Wikberg, J. E. S. XMPP for cloud computing in bioinformatics supporting discovery and invocation of asynchronous web services. *BMC Bioinformatics* **2009**, *10*, 279.

[13] Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. L. The Blue Obelisk-interoperability in chemical informatics. *J Chem Inf Model* **2006**, *46*, 991–998.

[14] O'Boyle, N. M. et al. Open Data, Open Source and Open Standards in chemistry: The Blue Obelisk five years on. *J Cheminform* **2011**, *3*, 37.
[15] OSGi Alliance. `http://www.osgi.org`.
[16] Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **2003**, *13*, 2498–2504.
[17] Spjuth, O.; Willighagen, E. L.; Guha, R.; Eklund, M.; Wikberg, J. E. Towards interoperable and reproducible QSAR analyses: Exchange of datasets. *J. Cheminform.* **2010**, *2*, 5.
[18] Spjuth, O.; Eklund, M.; Ahlberg Helgee, E.; Boyer, S.; Carlsson, L. Integrated decision support for assessing chemical liabilities. *J Chem Inf Model* **2011**, *51*, 1840–7.
[19] Carlsson, L.; Spjuth, O.; Adams, S.; Glen, R. C.; Boyer, S. Use of historic metabolic biotransformation data as a means of anticipating metabolic sites using MetaPrint2D and Bioclipse. *BMC Bioinformatics* **2010**, *11*, 362.
[20] Alvarsson, J.; Andersson, C.; Spjuth, O.; Larsson, R.; Wikberg, J. E. S. Brunn: an open source laboratory information system for microplates with a graphical plate layout design process. *BMC Bioinformatics* **2011**, *12*, 179.
[21] Willighagen, E. L.; Jeliazkova, N.; Hardy, B.; Grafström, R. C.; Spjuth, O. Computational toxicology using the OpenTox application programming interface and Bioclipse. *BMC Res Notes* **2011**, *4*, 487.
[22] Willighagen, E.; Alvarsson, J.; Andersson, A.; Eklund, M.; Lampa, S.; Lapins, M.; Spjuth, O.; Wikberg, J. Linking the Resource Description Framework to cheminformatics and proteochemometrics. *Journal of Biomedical Semantics* **2011**, *2*, S6+.
[23] Spjuth, O.; Eklund, M.; Lapins, M.; Junaid, M.; Wikberg, J. E. S. Services for prediction of drug susceptibility for HIV proteases and reverse transcriptases at the HIV drug research centre. *Bioinformatics* **2011**, *27*, 1719–20.
[24] Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. Recent developments of the Chemistry Development Kit (CDK) - an open-source Java library for chemo- and bioinformatics. *Curr Pharm Des* **2006**, *12*, 2111–2120.
[25] Krause, S.; Willighagen, E.; Steinbeck, C. JChemPaint - Using the Collaborative Forces of the Internet to Develop a Free Editor for 2D Chemical Structures. *Molecules* **2000**, *5*, 93–98.
[26] Jmol: an open-source Java viewer for chemical structures in 3D. `http://www.jmol.org/`.
[27] Murray-Rust, P.; Rzepa, H. S. Chemical markup, XML and the World-Wide Web. 2. Information objects and the CMLDOM. *J Chem Inf Comput Sci* **2001**, *41*, 1113–1123.
[28] Vainio, M. J.; Johnson, M. S. Generating conformer ensembles using a multiobjective genetic algorithm. *J Chem Inf Model* **2007**, *47*, 2462–2474.
[29] Lowe, D. M.; Corbett, P. T.; Murray-Rust, P.; Glen, R. C. Chemical name to structure: OPSIN, an open source solution. *Journal of chemical information and modeling* **2011**, *51*, 739–753.

[30] Dearden, J. C. In silico prediction of drug toxicity. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 119–127.
[31] Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350–4358.
[32] Benigni, R. Chemical structure of mutagens and carcinogens and the relationship with biological activity. *J. Exp. Clin. Cancer Res.* **2004**, *23*, 5–8.
[33] Akamatsu, M. Importance of physicochemical properties for the design of new pesticides. *J Agric Food Chem* **2011**, *59*, 2909–17.
[34] Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 379–386.
[35] Xue, L.; Stahura, F. L.; Bajorath, J. Similarity search profiling reveals effects of fingerprint scaling in virtual screening. *J Chem Inf Comput Sci* **2004**, *44*, 2032–9.
[36] Varnek, A. Fragment descriptors in structure-property modeling and virtual screening. *Methods Mol Biol* **2011**, *672*, 213–43.
[37] Guha, R. On the interpretation and interpretability of quantitative structure-activity relationship models. *J Comput Aided Mol Des* **2008**, *22*, 857–71.
[38] Carlsson, L.; Helgee, E. A.; Boyer, S. Interpretation of nonlinear QSAR models applied to Ames mutagenicity data. *J Chem Inf Model* **2009**, *49*, 2551–8.
[39] Hasegawa, K.; Funatsu, K. Non-linear modeling and chemical interpretation with aid of support vector machine and regression. *Curr Comput Aided Drug Des* **2010**, *6*, 24–36.
[40] Gamo, F.-J.; Sanz, L. M.; Vidal, J.; de Cozar, C.; Alvarez, E.; Lavandera, J.-L.; Vanderwall, D. E.; Green, D. V. S.; Kumar, V.; Hasan, S.; Brown, J. R.; Peishoff, C. E.; Cardon, L. R.; Garcia-Bustos, J. F. Thousands of chemical starting points for antimalarial lead identification. *Nature* **2010**, *465*, 305–10.
[41] Gellibert, F.; Fouchet, M.-H.; Nguyen, V.-L.; Wang, R.; Krysa, G.; de Gouville, A.-C.; Huet, S.; Dodic, N. Design of novel quinazoline derivatives and related analogues as potent and selective ALK5 inhibitors. *Bioorg Med Chem Lett* **2009**, *19*, 2277–81.
[42] Willighagen, E.; R., A.; Grafström, R. C.; Hardy, B.; Jeliazkova, N.; Spjuth, O. *Open Source Software in Life Science Research: Practical Solutions to Common Challenges in the Pharmaceutical Industry and Beyond*; Biohealthcare Publishing Ltd, 2012.
[43] Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.* **2005**, *48*, 312–320.
[44] Fitzpatrick, R. B. CPDB: Carcinogenic Potency Database. *Med. Ref. Serv. Q.* **2008**, *27*, 303–311.
[45] Denison, M. S.; Nagy, S. R. Activation of the aryl hydrocarbon receptor by structurally diverse exogenous and endogenous chemicals. *Annu. Rev. Pharmacol. Toxicol.* **2003**, *43*, 309–34.

[46] Ashby, J. Fundamental structural alerts to potential carcinogenicity or non-carcinogenicity. *Env. Mutagen.* **1985**, *7*, 919–921.
[47] Hardy, B. et al. Collaborative development of predictive toxicology applications. *Journal of Cheminformatics* **2010**, *2*, 7+.
[48] Rydberg, P.; Vasanthanathan, P.; Oostenbrink, C.; Olsen, L. Fast prediction of cytochrome P450 mediated drug metabolism. *ChemMedChem* **2009**, *4*, 2070–9.
[49] Patlewicz, G.; Jeliazkova, N.; Safford, R. J.; Worth, A. P.; Aleksiev, B. An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. *SAR QSAR Environ Res* **2008**, *19*, 495–524.
[50] Caldwell, G. W.; Yan, Z.; Tang, W.; Dasgupta, M.; Hasting, B. ADME optimization and toxicity assessment in early- and late-phase drug discovery. *Curr. Top. Med. Chem.* **2009**, *9*, 965–980.
[51] Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **2001**, *46*, 3–26.
[52] Wang, J.; Krudy, G.; Hou, T.; Zhang, W.; Holland, G.; Xu, X. Development of reliable aqueous solubility models and their application in druglike analysis. *J Chem Inf Model* **2007**, *47*, 1395–404.
[53] Li, H.; Yap, C. W.; Ung, C. Y.; Xue, Y.; Cao, Z. W.; Chen, Y. Z. Effect of selection of molecular descriptors on the prediction of blood-brain barrier penetrating and nonpenetrating agents by statistical learning methods. *J Chem Inf Model* **2005**, *45*, 1376–84.
[54] Chen, L.; Li, Y.; Zhao, Q.; Peng, H.; Hou, T. ADME evaluation in drug discovery. 10. Predictions of P-glycoprotein inhibitors using recursive partitioning and naive Bayesian classification techniques. *Mol Pharm* **2011**, *8*, 889–900.
[55] Wienkers, L. C.; Heath, T. G. Predicting in vivo drug interactions from in vitro drug discovery data. *Nat Rev Drug Discov* **2005**, *4*, 825–33.
[56] Glen, R. C., Robert C; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **2006**, *9*, 199–204.
[57] Gonzalez, F. J.; H, T. R. *Goodman & Gilman's The Pharmacological Basis of Therapeutics*; McGraw-Hill, 2005.
[58] Hertzberg, R. P.; Pope, A. J. High-throughput screening: new technology for the 21st century. *Current Opinion in Chemical Biology* **2000**, *4*, 445–451.
[59] Walters, W.; Stahl, M. T.; Murcko, M. A. Virtual screening—an overview. *Drug Discovery Today* **1998**, *3*, 160–178.
[60] Spjuth, O.; Georgiev, V.; Carlsson, L.; Willighagen, E.; Alvarsson, J.; Berg, A.; Wikberg, J. E. S.; Eklund, M. Bioclipse-R: Integrating management and visualization of life science data with statistical analysis. *Submitted*
[61] Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res* **2011**, *39*, D1035–41.

[62] Williams, A. J.; Harland, L.; Groth, P.; Pettifer, S.; Chichester, C.; Willighagen, E. L.; Evelo, C. T.; Blomberg, N.; Ecker, G.; Goble, C.; Mons, B. Open PHACTS: semantic interoperability for drug discovery. *Drug Discov Today* **2012**,

[63] Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **2012**, *40*, D1100–7.

[64] Degtyarenko, K.; de Matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; McNaught, A.; Alcántara, R.; Darsow, M.; Guedj, M.; Ashburner, M. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research* **2008**, *36*, D344–D350.

[65] Kuhn, M.; Campillos, M.; Letunic, I.; Jensen, L. J.; Bork, P. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology* **2010**, *6*.

[66] Samwald, M.; Jentzsch, A.; Bouton, C.; Kallesøe, C. S.; Willighagen, E.; Hajagos, J.; Marshall, M. S.; Prud'hommeaux, E.; Hassenzadeh, O.; Pichler, E.; Stephens, S. Linked open drug data for pharmaceutical research and development. *J Cheminform* **2011**, *3*, 19.

[67] Hastings, J.; Chepelev, L.; Willighagen, E.; Adams, N.; Steinbeck, C.; Dumontier, M. The chemical information ontology: provenance and disambiguation for chemical data on the biological semantic web. *PLoS One* **2011**, *6*, e25513.