The benefits of *in silico* modelling to identify possible small molecule drugs and their off-target interactions

Mire Zloh, Stewart B. Kirton
School of Life and Medical Sciences, University of Hertfordshire, College Lane, Hatfield Al10 9AB, United Kingdom.
Correspondence to Mire Zloh (zloh@live.co.uk) and Stewart B. Kirton (s.b.kirton3@herts.ac.uk)

Abstract

The research into the use of small molecules as drugs continues to be a key driver in the development of molecular databases, computer-aided drug design software and collaborative platforms. The evolution of computational approaches is driven by the essential criteria that a drug molecule has to fulfil, from the affinity to targets to minimal side effects while having adequate ADME properties. A combination of ligand- and structure-based drug development approaches are already used to obtain consensus predictions of small molecule activities and their off-target interactions. Further integration of these methods into easy to use workflows informed by systems biology could realise the full potential of available data in the drug discovery and reduce the attrition of drug candidates.

Keywords

Computer-aided drug design, molecular docking, target fishing, off-target interactions

Body of the text

Small molecules as drugs and drug candidates continue to be of interest for pharmaceutical industry [1,2] despite the current backdrop of a global pipeline for new medicines that is running dry across a range of therapeutic areas (see e.g.[3,4]) and a trend towards complex biologics as the next generation of therapeutics, particularly for inflammatory diseases and cancers [5–7]. It could be assumed by world-weary pharmaceutical scientists that the days when small molecules, broadly defined as compounds which adhere to Lipinski's rules of five[8], were the preferred starting point in drug discovery projects are coming to an end. The high failure rates in the later stages of small molecule drug discovery projects, caused by a lack of efficacy and problems with safety profiles [9] give significant cause for concern in the pharmaceutical industry. This has led big pharma to review their discovery pipeline strategies and develop initiatives – such as AstraZeneca's 6 R's framework [10] – to ensure that potential liabilities are removed early in the process.

It would seem counterintuitive that in an age where information is increasingly prevalent and accessible, the failure rate for small molecules at the later stages of drug development projects is still so stubbornly high. Whilst it can be argued that some efficacy and toxicological data relating to small molecules is proprietary, and the inability to access such information can result in unsuitable molecules being progressed through the drug discovery pipeline, it is likely this number is small. If we, as a scientific community, accept this statement it implies that we are not learning enough from past mistakes. It is possible that a number of current projects are progressing molecules which are destined to fail, and that this future failure was predictable in the early stages of the project *via* a thorough and systematic interrogation of the existing and accessible knowledge base. This boils down to the root of the problem being a problem of big data.

The information available for the use in drug discovery processes can be characterized by the "5 Vs" i.e. volume – typically big data concerns datasets; variety – the different nature of the types and forms in which data is acquired; velocity – the rate at which new data relating to a topic is becoming available; variability –inconsistencies with how data is reported, which may hinder subsequent analysis; and veracity – whether the data that is under investigation can be trusted to be accurate.[11] To be successful in identifying novel compounds in the future, it is imperative to overcome challenges of 5 Vs and understand, harness and progress from the wealth of information that is currently available. This colossal task is reliant upon the development of sophisticated *in-silico* systems to help capture, store, curate, analyse and exploit the vast amount of data relating to small molecules. This data can be broadly categorised into non- clinical data i.e. information pertaining to the structure of the molecule, its physicochemical properties and the information from *in vitro* and *in vivo* animal experiments to establish preliminary biological activity and toxicology profiles, and clinical data i.e. data relating to efficacy and toxicology of molecules as a result of clinical trials. Arguably, the greatest gains are to be made by analysing the data at the non-clinical development stage, as identifying patterns at this point in the process, which will ultimately result in failure in clinical trials will help to prioritise resource. The challenges to achieving this arise from the storage, extraction and analysis of the appropriate data.

The physical computing resource for storing, managing and, crucially, analysing the vast amount of freely available small-molecule data is immense, and in the recent past would have proved a financial barrier to a global, information-rich approach to drug discovery. However, the advent of cloud computing, whereby data is stored, managed and processed on remotely networked servers, provides a credible solution to this problem. Cloud computing has already been successfully used in small molecule drug discovery experiments for computer memory intensive activities such as *de novo* drug design and virtual screening[12–14] by allowing the combined processing power of networked

computers to perform complex simulations in an acceptable timeframe, but risks and perception of risk with respect to issues such as information security, data location and disaster recovery[15] continue to limit the widespread adoption of cloud computing, particularly in drug discovery settings where the potential financial rewards for addressing unmet clinical needs can be substantial.

This is not to say that the scientific community guards its data jealously; nothing could be further from the truth. Simple web-browser searches quickly identify a dizzying array of small molecule databases, which contain vast amounts of information that can be harnessed for drug discovery projects. These databases range from large repositories containing top-level information across a number of different areas (e.g. physical, toxicological and spectral properties of small molecules) such as Zinc[16] ChemSpider[17] and PubChem [18] through to those that deal in more detail with compounds that are known or predicted to demonstrate biological activity (CheBI[19], ChemBL[20] DrugBank[21]), those that focus on drug targets and the molecules known to act on them (IUPHAR[22]) and those that consider the metabolites of compounds and the implications of these on patient safety (e.g. Urine Metabolome[23], IMDB Toxin[24]). All these resources contain information, which would reliably inform drug development projects but there are two big data problems, volume and reliability of data, that present a barrier to easy harvesting and applying this information. Although each of the database providers mentioned have developed procedures for automated data acquisition and validation, there is a need for manual inspection of retrieved data due to duplication of data entries and possible variability of results of assays from different research groups. Although a suite of sophisticated web services for these databases is available, which scientists can use to interrogate their databases, it can be difficult for users to extract the exact information they need for their project[25], and another problem of big data – that of variability with respect to how data is stored in each individual repository – can require researchers to spend a lot of time training to become experts in using a number of different web services which again prevent them accessing all relevant information prior to embarking on their project. One way of overcoming these challenges is via the judicious use of web scraping.

In essence, web scraping (also known as web harvesting) is the copying of information from the web to a local repository, usually with the intention of subsequently manipulating and analysing this information. Although a person manually copying and pasting information from the web into a local database is doing webscraping, the term is more usually associated with the automated data collection of 'bots' or 'web crawlers', small pieces of programming code which are designed to harvest and deposit relevant pieces of information with minimal human intervention. Such activities were initially computationally complex and required expert knowledge of the web and scripting languages. However, the development of freely available desktop technologies such as iRobotsoft[26] means that this technique for gathering data is now available to novice, as well as expert, users.

Despite being one of the oldest technologies for extracting data from the internet, web scraping can prove invaluable where web services provided by database moderators do not meet the data capture needs of the user, and could prove invaluable in collating disparate data from the multitude of small molecule databases available prior to embarking on a drug discovery experiment. It is not, however, without its limitations. In order for their databases to be mined by bots the curators need to provide access, which is not always readily forthcoming given concerns over malicious hacking and information security. In addition legal jurisdictions with respect to where the data is stored can prove problematic as different thresholds exist in different parts of the world[27]. These limitations have not completely stymied recent drug discovery initiatives which have combined web-scraping with data-mining technologies to ensure that the latest and most relevant information related to small molecule drug discovery from a range of different sources is being considered at the implementation stage of drug discovery projects (see e.g [28]).

Data mining, sometimes used synonymously with machine learning, is the autonomous analysis of large-scale datasets to generate new information, and will be at the heart of this big data era of small molecule drug discovery. Networked computers processing information (for example the information extracted via web-scraping) will be used to identify previously obfuscated patterns that can directly input into rational drug design and development[29]. This technique is already proving beneficial to investigations into drug repurposing, a relatively low-risk strategy where small molecules that are known to have therapeutic benefits and acceptable safety profiles are investigated to see if they can be applied to other conditions by exploiting drug repurposing databases such as the NCGC Pharmaceutical Collection[30]. A number of drug repurposing studies have already been published to demonstrate the potential of machine learning for exploiting information to identify novel molecular disease targets[31] and repurpose existing medications for the treatment of a range of conditions including lupus[32], neurodegenerative disorders[33] and tuberculosis[34]. As it becomes easier to collate and analyse more data it is expected that this technology can also make significant in-roads into combatting orphan diseases, which although rare still affect up to 350 million people worldwide[35].

The information about small molecules and their properties should be complemented with information that is available about biological targets. The probability of discovering a next blockbuster drug through serendipity has become negligible and various *in silico* screening approaches as well as computer-aided molecule design methods are being developed. The availability of structural information of biomolecules complexed with small molecules with potential bioactivity in Protein Data Bank (PDB) [36] and PDB-REDO [37] is increasing year-on-year. There are currently over 98k models of biomolecules in a complex with a ligand available for the download and possible use in the drug discovery process. One of the most common uses of these target structures is to evaluate a potential binding mode of a small molecule through molecular docking (see e.g. [38]), if the binding sites for those structures are known. The trend in increase of number of available structures in the PDB is considerably higher than the number of publications in the Scopus database that have “docking” and “pdb” words mentioned anywhere in the text of the publications. This is most likely due to two factors, the released structures are either refined structures of previously released complexes with different ligands or published structures are not druggable targets. The *in silico* evaluation of protein druggability can be achieved using protein structural information, their known interactions with FDA approved drugs and knowledge of the human genome [39], albeit this approach provides only information on human targets and neglecting drugs that act on proteins expressed in other organisms.

Since the researchers are often faced with the lack of the structural information on targets of interest alternative strategies are employed, such as homology modelling to use existing structures to obtain structural information about the same targets from different organisms or structures of novel targets. In this way increased number of targets provides additional opportunities to apply docking in an attempt to find therapeutic agents against a wider range of diseases, resulting in almost two-fold higher number of publications that use homology models when compared to those where the docking is carried against a structure from the PDB. As a consequence of the issues mentioned, the information about targets is less suited for automation of the docking process. Furthermore, most of the docking protocols require careful examination of the available target structure which includes pre-processing of models obtained by either x-ray crystallography and NMR spectroscopy. However, web services are being developed, such as SwissDock, that are providing an opportunity to carry out docking against already prepared targets with well-defined binding sites [40]. Albeit this provides an easy access to selected targets and opportunity for direct comparison of the docking results from different research groups as the preparation of targets, scoring functions and algorithms used for docking are always the same, the usefulness of this service is limited by the number of proteins available in such databases.

Similarly, the other web services that provide collated information on proteins important in drug discovery, Potential Drug Target Database (PDTD) [41] and Therapeutic Target Database (TDD) [42], have a potential to be useful resources in the drug discovery efforts. However the lack of updates or information on the latest updates is impeding their future use.

Significant progress in improving the relevance of docking results is being made by taking into consideration the flexibility of targets [43] and docking with explicit hydration[44], and these efforts are further enhanced by the development of computational approaches to elucidate modulation of target via potential allosteric binding site [45]. In certain cases, biological activity and the binding affinity cannot be attributed to binding to active and/or allosteric sites. In those cases, the affinity studies should be carried by molecular docking against the whole protein/target surface to obtain a working hypothesis [46]. To ensure that the theoretical studies of interactions between protein and a small molecule provide the most reliable results the docking results against known binding site should be complemented by docking against the whole surface of the target using methods similar to those in the BINDSURF approach [47] and implementation of multi-objective strategies for combining scoring functions [48]. The future of the docking software development should be aiming to incorporate the above enhancements into the process for the hit identification as well as in lead optimization efforts when docking ligands against target databases.

Docking against multiple targets is already employed in target fishing by using reverse docking. This approach is particularly useful if biological activities of sets of molecules determined in cell-based assays are known, but without knowledge of the target protein. The great promise of this approach was demonstrated by predicting targets for 4OH-tamoxifen and vitamin E, where 50% of computationally predicted targets were implicated in the binding of these two molecules [49]. This concept of "target fishing" was developed further via developing a PDTD with the defined binding sites of proteins known to be targets for small molecule therapeutics. The improved algorithms and scoring functions led to the improvement of the target identification results for 4H-tamoxifen and vitamin E [50]. Although promising, this approach has not been utilized fully in most disease areas, the TarFisDock was mainly used in studies to explain activities either of sets of sets of molecules that may have multiple targets [51] or of components of plant extracts [52].. This is mainly due to a complexity of the results obtained, as well as the lack of direct correlation between observed biological activities and corresponding binding affinities to a single possible target. This could be a result of ligands with the activity having affinity for several targets and a possible interplay of cellular pathways affected by these interactions. Therefore, the target selection has to be informed by knowledge-based approaches and using a target pool relevant to the disease of interest [51,53], and further complemented by systems biology computational analysis [54] and network biology [55].

It is apparent that target fishing through molecular docking approaches has some drawbacks, as the activities of small molecules depend on their molecular properties as well as on the target availability. Therefore, other approaches are being developed to address the important issue of target identification that often arises from high throughput screening (HTS) or developing of libraries of analogues. The drug target interactions can be predicted based on the structures of ligands through the comparison to the information derived from known three dimensional structures of protein-drug complexes. An open web server, TargetNet, utilizes naïve Bayes approach in SAR models to evaluate possible affinity of ligands for 623 human proteins [56]. While the ligands are represented as molecular fingerprints for TargetNet to evaluate molecular similarity, the use of 2D structures and fingerprints is a simplification and can lead to biased results, as the outcomes of searches can depend on the methods used to evaluate similarity [57]. This type of a limitation was purposely addressed when developing the PolyPharmacology Browser (PPB), where sets of ligands binding to the same

target are used to develop a consensus voting scheme where six different fingerprints and four fused fingerprints to evaluate similarity of a query molecule to ligands present in ChEMBL database. Moreover, the use of molecular similarity of ligands in quest for target responsible for activity can be expanded by evaluation of their 3D shape and surface feature similarity in addition to using 2D similarity as criteria to predict a possible target. The Chemmapper website [58] allows comparison of a molecular structure of a query molecule to a diverse sets of ligands with known binding in the PDB (7072 structures) and KEGG (5928 structures) datasets, as well as to ligands present in big data bioactivity databases (DrugBank, ChEMBL and BindingDB). However, the results of 2D and 3D similarity comparisons are obtained and have to be analysed separately.

An additional dimension to target identification was introduced by combining 2D and 3D similarity measures between a query molecule and identified ligands as implemented in the SwissTargetPrediction website [59]. Furthermore, the SuperPred webserver takes into consideration 2D, fragment and 3D molecular similarity of a query molecule to all ligands that are associated with the targets in the database [60]. The applicability domain of these two webservers is possibly larger compared to previously mentioned services, as their target databases contain proteins that are not only human but also from other mammals.

However, the above mentioned ligand-based approach to target identification that uses direct comparison of a query structure to a single ligand from the binding databases may not provide a comprehensive overview of the potential interactions that can occur. The development of a pharmacophore approach can take into consideration features in query molecules that should be present for favourable binding to occur. The website Pharmmapper provides a platform for a reverse pharmacophore mapping approach [61,62], that uses over 23k protein structures from the PDB to develop druggable and ligandable pharmacophores in relation to 450 indications and 4800 molecular functions related to these protein structures. Currently, the number of publications citing Pharmmapper in the Scopus search outnumbers publications involving the other above-mentioned approaches, most likely due to the inclusion of targets that are non- mammalian such as bacterial targets being important in a search for therapeutic agents to treat multidrug resistant strains. These are particularly useful features of the Pharmmapper service in addition to providing the information related to function of targets and relevant indications. However, pharmacophore models that are generated for this server are based on a single structure for each target protein. This results in the multiple occurrence of the same protein in the resulting list of targets for a query molecule. An opportunity to consider multiple experimentally determined structures for target proteins is overlooked, as that could introduce some elements of the protein flexibility in target identification and diversify type of ligands that could fit into such pharmacophores.

There are other platforms that deal with the use of big data in target identification for small molecules that show activity in cell-based and whole organism assays (see [57] for a comprehensive list). However, all of those approaches have their limitations leading to an uncertainty of target prediction for each query molecule. This may be overcome by merging the best features of already developed approaches into a single platform and proposing a census scoring of each solution obtained by different methods.

It is interesting that these relatively recent efforts in developing methods for target fishing did not result in the expected progress of identifying targets for molecules with known activity and the number of publications does not reflect the promise that these methods offer. There are some examples of useful applications that arose from application of one or a combination of several methods that led to, for example, the discovery and experimental validation of novel mechanisms of inhibition in *Magnaporthe oryzae* by a chalcone-based inhibitor [63] and identification of ten

phytochemicals from *Rhazya stricta* that may have good anticancer activities [64]. There are also some examples where these services enabled scaffold hopping in identification of novel inhibitors for treatment of type 2 diabetes [65] and small lung cancer [66]. The experimental validation of predicted targets is becoming one of the most important criteria for consideration of manuscripts, which is not always an easy undertaking, thus the number of studies that report successful target identification is relatively low.

Despite the discouraging outcome for discovery of hits and development of lead molecules, the applications of the methods described above and similar ones have boomed in three other fields: evaluating polypharmacology of small molecules, off-target interactions prediction and drug repurposing.

The initial efforts in predicting the side effects of small molecules was via evaluation of their affinity towards protein structures using the inverse docking procedure [67]. This has led to development of a number of webservers employing structure-based approaches. Virtualtoxlab tool is one of the first server implementations that provided prediction of toxic potential for small molecules via evaluating their interactions with 16 proteins. This set of proteins are known or suspected to trigger adverse effects and include 10 nuclear hormone receptors (NHR). These predictions are made by a combination of flexible docking with multi-dimensional QSAR [68,69]. It may appear that such predictions are not important as the comprehensive *in vitro* study of 615 drugs did not find a significant number of drugs that interact with the NHR. The most likely reason for absence of interactions of drugs with NHR is that drug candidates with a potential to exert such interactions were discontinued during the drug development process. Although these experiments did not indicate problems related to NSAID toxicity, Virtualtoxlab results revealed that diclofenac and celecoxib may have affinity for thyroid hormone receptor β. This receptor is implicated in hypothyroidism associated with increased heart muscle stiffness and therefore inhibition may lead to an increased risk of myocardial infarction. These *in silico* findings were confirmed by *ex vivo* studies, indicating the importance of prediction results [70], especially if these are used in early drug discovery and development stages. Opensource developments in this areas are also prominent, where a website service “Endocrine Disruptome” provides a similar prediction of small molecule affinities to 12 NHRs [71], allowing toxicity predictions in resource-challenged environments. Moreover, their docking interface for target systems (DoTS) is freely available for implementation on secure servers, thus addressing potential intellectual property issues that may arise if publicly available servers are used.

While significant progress has been made in utilizing *in silico* methods to predict drug-target interactions (see e.g. [72]) and enable drug repurposing (see e.g. [73], utilising not only standard methods and ideas, but also innovative concepts, approaches and algorithms, there are intrinsic problems that are related to how the research in this area is funded and how these efforts are rewarded. Whilst the pharmaceutical industry is engaged in data sharing and there are excellent examples of developing collaborative platforms for opensource drug discovery [34], their need to protect their intellectual property and not share all the data and software applications is understandable. However, due to the lack of appropriate funding, the efforts of the academic community often result in projects that are short-lived delivering sometimes ingenious software solutions that are frequently not finished and/or difficult to integrate into other relevant software platforms as the file formats and data storage are not standardized. More often than not, as a drive to meet criteria for academic promotions and to achieve quick wins, software solutions or web services are developed that appear redundant and do not significantly contribute to furthering the progress in the field. Furthermore, the efforts in developing collaborative drug discovery projects appear not to be unified as the researchers have to make a choice where to direct their efforts when

there many platforms available, e.g. Open PHACTS [74], Online chemical modelling environment [75], Collaborative Drug Discovery Vault [76], in addition to platforms dedicated to specific therapeutic areas such as SysBorg 2.0 for open source drug discovery platform to fight tubercolosis [77] .

Future perspectives

In order to have a true impact on small-molecule drug discovery projects, the acquisition and processing of pertinent big data will need to be standardized and automated. However, a dichotomy exists whereby such automation will need to be generalizable enough to broadly apply to any drug discovery project whilst also being customisable enough to address the nuances of each therapeutic target including not only proteins but also oligonucleotides and other validated druggable targets. Additionally, these automated analysis frameworks would need to be accessible to non-experts in order to ensure that barriers to their adoption don't exist.

As such, modular workflow interfaces seem set to play (an even bigger) role in the near future. Although competitors are emerging into this market all the time, there are two main pieces of software which have influenced drug discovery over the past twenty years, namely the proprietary Pipeline Pilot[78] which was launched in 1999 and the Open Source KNIME[79] which followed in 2004. Both of these tools allow users to build general architectures from a library of pre-designed modules in a "drag and drop" style to facilitate the automated processing and filtering of data. They do this whilst retaining sufficient fine control to allow these architectures to easily become custom-tailored by giving users the ability to modify threshold values for parameters in each of the modules to suit the needs of their project.

These architectures can incorporate well-established models for predicting not only potential activity of small molecules but also physicochemical and toxicological properties such as aqueous solubility[80–82], plasma-protein binding[83], blood-brain-barrier permeation[84–86] cytochrome P450 isoform specificity[87] [88] and off-target interactions [50,56–62] in order to remove potential liabilities at an early stage of the process. It is desirable that in the future calculated properties could be compared against a database of experimental values for such properties harvested from the web to highlight limitations of any model and guard against identifying false negatives as part of the screening process. It is also possible that the workflow could then be extended to use machine learning to address the weaknesses of a model, based on the outcome of these comparisons and automatically build, test and validate iterative models based on the information it receives [89]. Although we have not reached this stage yet, there are again a significant number of publications in the scientific literature, which show the positive impact of workflow interfaces in a diverse range of therapeutic areas including the identification of molecules with the potential to treat cancer [90][91,92], and those with potential to act as anti-inflammatory compounds [93]. Given the scalability of such interfaces, it does not seem unreasonable to expect their prominence to grow in small molecule drug development in response to the challenges and opportunities presented by big data.

Executive summary

Small molecules continue to attract the interest of pharmaceutical companies and academic research groups as drugs and drug candidates.

Increased availability of information on molecular properties of small molecules, their biological activities and targets drive the development of methods for processing big data.

The optimal use of small molecule information can be achieved in conjunction with the structural information available on the biological targets.

Publicly available web servers can be used to predict potential targets for small molecules and off-target interactions via a combination of ligand-based and structure-based methods.

Standardization of the data storage in the databases and incorporation of well-established methods for prediction of activities, molecular properties, side effects and network biology into easy-to-use workflows can provide a basis for the next paradigm in drug discovery.

Financial and competing interests disclosure

The authors declare no financial and competing interests

References:

ADDIN ZOTERO_BIBL {"custom":[]} CSL_BIBLIOGRAPHY 1. Small molecules continue to go from strength to strength [Internet]. Available from: https://www.manufacturingchemist.com/technical/article_page/Small_molecules_continue_to_go_from_strength_to_strength/122642.

2. Munk, Stephen A. Growing Demand for Small-Molecule CDMO Services [Internet]. Available from: http://www.americanpharmaceuticalreview.com/Featured-Articles/182922-Growing-Demand-for-Small-Molecule-CDMO-Services/.

3. Dean B, Moller H-J, Svensson TH, *et al.* Problems and solutions to filling the drying drug pipeline for psychiatric disorders: a report from the inaugural 2012 CINP Think Tank. *Int. J. Neuropsychopharmacol.* 17(1), 137–148 (2014).

4. Spellberg B, Bartlett J, Wunderink R, Gilbert DN. Novel Approaches Are Needed to Develop Tomorrow's Antibacterial Therapies. *Am. J. Respir. Crit. Care Med.* 191(2), 135–140 (2015).

5. Dulai PS, Sandborn WJ. Next-Generation Therapeutics for Inflammatory Bowel Disease. *Curr. Gastroenterol. Rep.* 18(9), 51 (2016).

6. Kintzing JR, Filsinger Interrante MV, Cochran JR. Emerging Strategies for Developing Next-Generation Protein Therapeutics for Cancer Treatment. *Trends Pharmacol. Sci.* 37(12), 993–1008 (2016).

7. Choi SH, Shah K. Engineered Bifunctional Proteins and Stem Cells: Next Generation of Targeted Cancer Therapeutics. *Discov. Med.* 22(120), 157–166 (2016).

8. Lipinski CA. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* 1(4), 337–341 (2004).

9. Arrowsmith J, Miller P. Trial Watch: Phase II and Phase III attrition rates 2011-2012. *Nat. Rev. Drug Discov.* 12(8), 569–569 (2013).

10. Cook D, Brown D, Alexander R, *et al.* Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discov.* 13(6), 419–431 (2014).

11. Hilbert M. Big Data for Development: A Review of Promises and Challenges. *Dev. Policy Rev.* 34(1), 135–174 (2016).

12. Capuccini M, Ahmed L, Schaal W, Laure E, Spjuth O. Large-scale virtual screening on public cloud resources with Apache Spark. *J. Cheminformatics*. 9(1), 15 (2017).

13. Jaghoori MM, Bleijlevens B, Olabarriaga SD. 1001 Ways to run AutoDock Vina for virtual screening. *J. Comput. Aided Mol. Des.* 30(3), 237–249 (2016).

14. Chang K-W, Tsai T-Y, Chen K-C, *et al.* iSMART: An Integrated Cloud Computing Web Server for Traditional Chinese Medicine for Online Virtual Screening, de novo Evolution and Drug Design. *J. Biomol. Struct. Dyn.* 29(1), 243–250 (2011).

15. Brender N, Markov I. Risk perception and risk management in cloud computing: Results from a case study of Swiss companies. *Int. J. Inf. Manag.* 33(5), 726–733 (2013).

16. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* 52(7), 1757–1768 (2012).

17. ChemSpider | Search and share chemistry [Internet]. Available from: http://www.chemspider.com/.

18. The PubChem Project [Internet]. Available from: https://pubchem.ncbi.nlm.nih.gov/.

19. Hastings J, de Matos P, Dekker A, *et al.* The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.* 41(Database issue), D456-463 (2013).

20. Bento AP, Gaulton A, Hersey A, *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42(Database issue), D1083-1090 (2014).

21. Law V, Knox C, Djoumbou Y, *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* 42(Database issue), D1091-1097 (2014).

22. Sharman JL, Benson HE, Pawson AJ, *et al.* IUPHAR-DB: updated database content and new features. *Nucleic Acids Res.* 41(D1), D1083–D1088 (2013).

23. Bouatra S, Aziat F, Mandal R, *et al.* The Human Urine Metabolome. *PLOS ONE*. 8(9), e73076 (2013).

24. Wishart DS, Jewison T, Guo AC, *et al.* HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res.* 41(Database issue), D801-807 (2013).

25. Glez-Peña D, Lourenço A, López-Fernández H, Reboiro-Jato M, Fdez-Riverola F. Web scraping technologies in an API world. *Brief. Bioinform.* 15(5), 788–797 (2014).

26. IRobotSoft -- Visual Web Scraping and Web Automation [Internet]. Available from: http://irobotsoft.com/.

27. Hirschey J. Symbiotic Relationships: Pragmatic Acceptance of Data Scraping. *Berkeley Technol. Law J.* [Internet]. 29(4) (2014). Available from: http://scholarship.law.berkeley.edu/btlj/vol29/iss4/16.

28. McEntire R, Szalkowski D, Butler J, *et al.* Application of an automated natural language processing (NLP) workflow to enable federated search of external biomedical content in drug discovery and development. *Drug Discov. Today*. 21(5), 826–835 (2016).

29. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today*. 20(3), 318–331 (2015).

30. Huang R, Southall N, Wang Y, *et al.* The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Sci. Transl. Med.* 3(80), 80ps16 (2011).

31. Harnie D, Saey M, Vapirev AE, *et al.* Scaling machine learning for target prediction in drug discovery using Apache Spark. *Future Gener. Comput. Syst.* 67, 409–417 (2017).

32. Toro-Domínguez D, Carmona-Sáez P, Alarcón-Riquelme ME. Support for phosphoinositol 3 kinase and mTOR inhibitors as treatment for lupus using in-silico drug-repurposing analysis. *Arthritis Res. Ther.* 19, 54 (2017).

33. Dovrolis N, Kolios G, Spyrou G, Maroulakou I. Laying in silico pipelines for drug repositioning: a paradigm in ensemble analysis for neurodegenerative diseases. *Drug Discov. Today*. 22, 805–813 (2017).

34. Ekins S, Spektor AC, Clark AM, Dole K, Bunin BA. Collaborative drug discovery for More Medicines for Tuberculosis (MM4TB). *Drug Discov. Today*. 22(3), 555–565 (2017).

35. Mertz L. Turning the Unknown into Known: Data Mining Is Increasingly Used to Prospect for Rare-Disease Biology and Treatments. *IEEE Pulse*. 8(1), 28–32 (2017).

36. Berman HM, Westbrook J, Feng Z, *et al.* The Protein Data Bank. *Nucleic Acids Res.* 28(1), 235–242 (2000).

37. Joosten RP, Salzemann J, Bloch V, *et al.* PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *J. Appl. Crystallogr.* 42(3), 376–384 (2009).

38. Ferreira LG, dos Santos RN, Oliva G, Andricopulo AD. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules*. 20(7), 13384–13421 (2015).

39. Kim B, Jo J, Han J, Park C, Lee H. In silico re-identification of properties of drug target proteins. *BMC Bioinformatics*. 18(7), 248 (2017).

40. Grosdidier A, Zoete V, Michielin O. Blind docking of 260 protein-ligand complexes with EADock 2.0. *J. Comput. Chem.* 30(13), 2021–2030 (2009).

41. Gao Z, Li H, Zhang H, *et al.* PDTD: a web-accessible protein database for drug target identification. *BMC Bioinformatics*. 9, 104 (2008).

42. Yang H, Qin C, Li YH, *et al.* Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res.* 44(D1), D1069-1074 (2016).

43. Wong CF. Flexible receptor docking for drug discovery. *Expert Opin. Drug Discov.* 10(11), 1189–1200 (2015).

44. Forli S, Huey R, Pique ME, Sanner MF, Goodsell DS, Olson AJ. Computational protein-ligand docking and virtual drug screening with the AutoDock suite. *Nat. Protoc.* 11(5), 905–919 (2016).

45. Lu S, Huang W, Zhang J. Recent computational advances in the identification of allosteric sites in proteins. *Drug Discov. Today*. 19(10), 1595–1600 (2014).

46. Lauinger IL, Vivas L, Perozzo R, *et al.* Potential of Lichen Secondary Metabolites against Plasmodium Liver Stage Parasites with FAS-II as the Potential Target. *J. Nat. Prod.* 76(6), 1064–1070 (2013).

47. Sánchez-Linares I, Pérez-Sánchez H, Cecilia JM, García JM. High-Throughput parallel blind Virtual Screening using BINDSURF. *BMC Bioinformatics*. 13(14), S13 (2012).

48. Gu J, Yang X, Kang L, Wu J, Wang X. MoDock: A multi-objective strategy improves the accuracy for molecular docking. *Algorithms Mol. Biol.* 10, 8 (2015).

49. Chen Y z., Zhi D g. Ligand–protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins Struct. Funct. Bioinforma.* 43(2), 217–226 (2001).

50. Li H, Gao Z, Kang L, *et al.* TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic Acids Res.* 34(suppl_2), W219–W224 (2006).

51. Erić S, Ke S, Barata T, *et al.* Target fishing and docking studies of the novel derivatives of aryl-aminopyridines with potential anticancer activity. *Bioorg. Med. Chem.* 20(17), 5220–5228 (2012).

52. Jeong C-H, Bode AM, Pugliese A, *et al.* [6]-Gingerol Suppresses Colon Cancer Growth by Targeting Leukotriene A_4 Hydrolase. *Cancer Res.* 69(13), 5584–5591 (2009).

53. Scafuri B, Marabotti A, Carbone V, Minasi P, Dotolo S, Facchiano A. A theoretical study on predicted protein targets of apple polyphenols and possible mechanisms of chemoprevention in colorectal cancer. *Sci. Rep.* 6, srep32516 (2016).

54. Dougherty BV, Moutinho Jr TJ, Papin J. Accelerating the Drug Development Pipeline with Genome-Scale Metabolic Network Reconstructions. *Syst. Biol.* 6 (2017).

55. Hopkins AL. Network pharmacology. *Nat. Biotechnol.* 25(10), 1110–1111 (2007).

56. Yao Z-J, Dong J, Che Y-J, *et al.* TargetNet: a web service for predicting potential drug–target interaction profiling via multi-target SAR models. *J. Comput. Aided Mol. Des.* 30(5), 413–424 (2016).

57. Awale M, Reymond J-L. The polypharmacology browser: a web-based multi-fingerprint target prediction tool using ChEMBL bioactivity data. *J. Cheminformatics*. 9(1), 11 (2017).

58. Gong J, Cai C, Liu X, *et al.* ChemMapper: a versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method. *Bioinformatics*. 29(14), 1827–1829 (2013).

59. Gfeller D, Grosdidier A, Wirth M, Daina A, Michielin O, Zoete V. SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic Acids Res.* 42(W1), W32–W38 (2014).

60. Nickel J, Gohlke B-O, Erehman J, *et al.* SuperPred: update on drug classification and target prediction. *Nucleic Acids Res.* 42(W1), W26–W31 (2014).

61. Liu X, Ouyang S, Yu B, *et al.* PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. *Nucleic Acids Res.* 38(suppl_2), W609–W614 (2010).

62. Wang X, Shen Y, Wang S, *et al.* PharmMapper 2017 update: a web server for potential drug target identification with a comprehensive target pharmacophore database. *Nucleic Acids Res.* [Internet]. Available from: https://academic.oup.com/nar/article/doi/10.1093/nar/gkx374/3791213/PharmMapper-2017-update-a-web-server-for-potential.

63. Chen H, Wang X, Jin H, Liu R, Hou T. Discovery of the molecular mechanisms of the novel chalcone-based Magnaporthe oryzae inhibitor C1 using transcriptomic profiling and co-expression network analysis. *SpringerPlus*. 5(1), 1851 (2016).

64. Obaid AY, Voleti S, Bora RS, *et al.* Cheminformatics studies to analyze the therapeutic potential of phytochemicals from Rhazya stricta. *Chem. Cent. J.* 11(1), 11 (2017).

65. Li S, Xu H, Cui S, *et al.* Discovery and Rational Design of Natural-Product-Derived 2-Phenyl-3,4-dihydro-2H-benzo[f]chromen-3-amine Analogs as Novel and Potent Dipeptidyl Peptidase 4 (DPP-4) Inhibitors for the Treatment of Type 2 Diabetes. *J. Med. Chem.* 59(14), 6772–6790 (2016).

66. Hao Y, Wang X, Zhang T, *et al.* Discovery and Structural Optimization of N5-Substituted 6,7-Dioxo-6,7-dihydropteridines as Potent and Selective Epidermal Growth Factor Receptor (EGFR) Inhibitors against L858R/T790M Resistance Mutation. *J. Med. Chem.* 59(15), 7111–7124 (2016).

67. Chen YZ, Ung CY. Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand–protein inverse docking approach. *J. Mol. Graph. Model.* 20(3), 199–218 (2001).

68. Vedani A, Dobler M, Spreafico M, Peristera O, Smiesko M. VirtualToxLab - In silico prediction of the toxic potential of drugs and environmental chemicals: Evaluation status and internet access protocol. *Altex*. 24(3), 153–161 (2007).

69. Smieško M, Vedani A. VirtualToxLab: Exploring the Toxic Potential of Rejuvenating Substances Found in Traditional Medicines [Internet]. In: *In Silico Methods for Predicting Drug Toxicity*. Benfenati E (Ed.). . Springer New York, 121–137 (2016) [cited 2017 Mar 8]. Available from: http://dx.doi.org/10.1007/978-1-4939-3609-0_7.

70. Zloh M, Perez-Diaz N, Tang L, Patel P, Mackenzie LS. Evidence that diclofenac and celecoxib are thyroid hormone receptor beta antagonists. *Life Sci.* 146, 66–72 (2016).

71. Kolšek K, Mavri J, Sollner Dolenc M, Gobec S, Turk S. Endocrine Disruptome—An Open Source Prediction Tool for Assessing Endocrine Disruption Potential through Nuclear Receptor Binding. *J. Chem. Inf. Model.* 54(4), 1254–1267 (2014).

72. Cheng T, Hao M, Takeda T, Bryant SH, Wang Y. Large-Scale Prediction of Drug-Target Interaction: a Data-Centric Review. *AAPS J.* , 1–12 (2017).

73. March-Vila E, Pinzi L, Sturm N, *et al.* On the Integration of In Silico Drug Design Methods for Drug Repurposing. *Front. Pharmacol.* [Internet]. 8 (2017). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5440551/.

74. Blomberg N, Ecker GF, Kidd R, Mons B, Williams-Jones B. Knowledge driven drug discovery goes semantic. *MedChemComm*. 2(8), 72–76 (2011).

75. Sushko I, Novotarskyi S, Körner R, *et al.* Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided Mol. Des.* 25(6), 533–554 (2011).

76. Hohman M, Gregory K, Chibale K, Smith PJ, Ekins S, Bunin B. Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. *Drug Discov. Today*. 14(5), 261–270 (2009).

77. Open Source Drug Discovery [Internet]. Available from: http://www.osdd.net/home.

78. BIOVIA Pipeline Pilot | Scientific Workflow Authoring Application for Data Analysis [Internet]. Available from: http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/.

79. KNIME | Open for Innovation [Internet]. Available from: https://www.knime.org/.

80. Ran Y, Jain N, Yalkowsky SH. Prediction of Aqueous Solubility of Organic Compounds by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* 41(5), 1208–1217 (2001).

81. Ali J, Camilleri P, Brown MB, Hutt AJ, Kirton SB. Revisiting the General Solubility Equation: In Silico Prediction of Aqueous Solubility Incorporating the Effect of Topographical Polar Surface Area. *J. Chem. Inf. Model.* 52(2), 420–428 (2012).

82. Delaney JS. Predicting aqueous solubility from structure. *Drug Discov. Today*. 10(4), 289–295 (2005).

83. Gleeson MP. Plasma Protein Binding Affinity and Its Relationship to Molecular Structure: An In-silico Analysis. *J. Med. Chem.* 50(1), 101–112 (2007).

84. Abbott NJ. Prediction of blood–brain barrier permeation in drug discovery from in vivo, in vitro and in silico models. *Drug Discov. Today Technol.* 1(4), 407–416 (2004).

85. Narayanan R, Gunturi SB. In silico ADME modelling: prediction models for blood–brain barrier permeation using a systematic variable selection method. *Bioorg. Med. Chem.* 13(8), 3017–3028 (2005).

86. Clark DE. In silico prediction of blood–brain barrier permeation. *Drug Discov. Today*. 8(20), 927–933 (2003).

87. Gleeson MP, Davis AM, Chohan KK, *et al.* Generation of in-silico cytochrome P450 1A2, 2C9, 2C19, 2D6, and 3A4 inhibition QSAR models. *J. Comput. Aided Mol. Des.* 21(10–11), 559–573 (2007).

88. Matsson P, Bergström CAS. Computational modeling to predict the functions and impact of drug transporters. *Silico Pharmacol.* 3(1), 8 (2015).

89. Muegge I, Bentzien J, Mukherjee P, Hughes RO. Automatically updating predictive modeling workflows support decision-making in drug design. *Future Med. Chem.* 8(14), 1779–1796 (2016).

90. Hinderlich S, Neuenschwander M, Wratil P-R, *et al.* Small molecules targeting human N-acetylmannosamine kinase. *ChemBioChem*. , n/a-n/a (2017).

91. Shao Z, Xu P, Xu W, *et al.* Discovery of novel DNA methyltransferase 3A inhibitors via structure-based virtual screening and biological assays. *Bioorg. Med. Chem. Lett.* 27(2), 342–346 (2017).

92. Zou F, Yang Y, Ma T, Xi J, Zhou J, Zha X. Identification of novel MEK1 inhibitors by pharmacophore and docking based virtual screening. *Med. Chem. Res.* 26(4), 701–713 (2017).

93. Bai R, Shi Q, Liang Z, *et al.* Development of CXCR4 modulators by virtual HTS of a novel amide-sulfamide compound library. *Eur. J. Med. Chem.* 126, 464–475 (2017).