Journal of Cheminformatics

SOFTWARE

Open Access

The ChemicalToolbox: reproducible, user-friendly cheminformatics analysis on the Galaxy platform

Simon A. Bray[1*], Xavier Lucas[2], Anup Kumar[1] and Björn A. Grüning[1]

Abstract

Here, we introduce the ChemicalToolbox, a publicly available web server for performing cheminformatics analysis. The ChemicalToolbox provides an intuitive, graphical interface for common tools for downloading, filtering, visualizing and simulating small molecules and proteins. The ChemicalToolbox is based on Galaxy, an open-source web-based platform which enables accessible and reproducible data analysis. There is already an active Galaxy cheminformatics community using and developing tools. Based on their work, we provide four example workflows which illustrate the capabilities of the ChemicalToolbox, covering assembly of a compound library, hole filling, protein-ligand docking, and construction of a quantitative structure-activity relationship (QSAR) model. These workflows may be modified and combined flexibly, together with the many other tools available, to fit the needs of a particular project. The ChemicalToolbox is hosted on the European Galaxy server and may be accessed via https://cheminformatics.usegalaxy.eu.

Keywords: Cheminformatics, Protein-ligand docking, QSAR, Galaxy, Molecular dynamics

Introduction

Open-source software packages are now available for a wide range of cheminformatics applications, ranging from downloading [1, 2], manipulating, and processing small molecules [3–5], to protein-ligand docking calculations [6, 7], to quantum chemistry [8]. However, with the growth in the number of applications, the difficulty in combining these tools into easily usable, reproducible analysis workflows increases. Many tools require the user to possess some level of programming skill, or at least ability to use the command line; some also rely on unique file formats. Some tools require compilation of the source code for their use, which not only poses a challenge for computationally inexperienced scientists, but also muddies the waters if another user attempts to reproduce the analysis in another environment [9].

Use of technologies such as Conda [10] and containerization (most notably Docker and Singularity [11–13]) helps to mitigate some of these issues. Conda enables reproducible analyses and simplifies installation, while containerization technologies provide a common working environment across operating systems. However, knowledge of the command line is still required to run software, and the user is responsible for maintaining the thorough records (e.g. through use of a traditional lab book) that are required for full reproducibility of analyses.

Here, we present the ChemicalToolbox, a modular, intuitive platform for cheminformatics analysis, built within the Galaxy system [14, 15]. It combines numerous open-source cheminformatics tools, and integrates them into an intuitive, web-based user interface; requested jobs can then be sent to a high-performace computing (HPC) cluster for execution. Thus, the user has access to a range

*Correspondence: sbray@informatik.uni-freiburg.de
[1] Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, Freiburg, Germany
Full list of author information is available at the end of the article

of useful tools and substantial compute resources, without being exposed directly to the HPC environment, or to the command-line interface used by much cheminformatics software. Tools can be run individually, or combined into workflows, which can then be shared with collaborators. All tools are made publicly available on the European Galaxy server, under the subdomain https://cheminformatics.usegalaxy.eu. As an alternative, the ChemicalToolbox can also be easily installed on personal computers, clusters, and cloud services; once installed, the system can be accessed simultaneously by multiple users, using current standard web browsers.

The ChemicalToolbox provides a range of tools for different applications, as depicted in Fig. 1. Chemical structures can be accessed from online databases such as PubChem [2] and ChEMBL [1]. Manipulation of chemical structures can be performed with OpenBabel [4] and RDKit [3], while calculation of molecular descriptors for QSAR studies may be done using Mordred [16] or PaDEL [17], which rely on RDKit and the Chemical Development Kit (CDK) [5] respectively. Protein-ligand docking may be performed using AutoDock Vina [6] and rDock [7]. Furthermore, the previously published BRIDGE platform [18] extends the core functionality of the ChemicalToolbox into molecular dynamics, providing a suite of tools which draws on the GROMACS [19], AmberTools [20], Parmed [21], and MDAnalysis [22] software. Apart from tools, the Galaxy codebase has been extended to provide features particularly useful for cheminformatics. These include support for a range of filetypes commonly used for reporting chemical structures, including PDB, SMILES, InChI, SMILES, SDF/MOL and MOL2, as well as tools for interconverting between these formats, based on OpenBabel. The most common GROMACS filetypes have also been made available. Another feature integrated directly into the Galaxy codebase is the NGLviewer [23], which may be used for visualization of compounds and macromolecules. Furthermore, apart from the features of the ChemicalToolbox itself, the inherent flexibility of the Galaxy system allows combination of the ChemicalToolbox with existing platforms developed by researchers working in other related areas, such as the Galaxy Genome Annotation project, metabolomics (Workflow4Metabolomics [24], Metaboloflow [25]), proteomics (Galaxy-P [26]), and machine learning—enabling the development of new, transdisciplinary workflows.

A number of other workflow management systems are commonly used in cheminformatics; the most prominent are Pipeline Pilot [27] and KNIME [28, 29]. Pipeline Pilot is a workflow management software developed by Accelrys Enterprise Platform and published as a proprietary application. It offers tools bundled into 'component

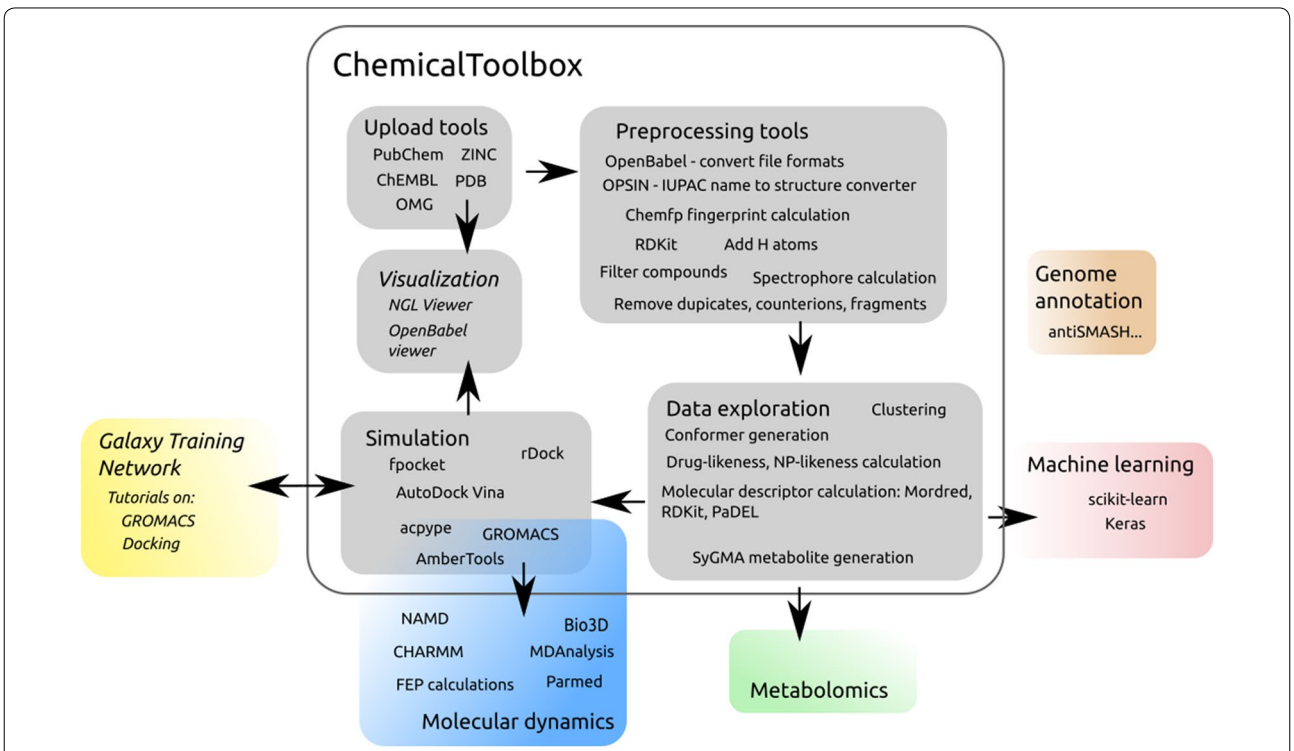

Fig. 1 Tools and visualizations available via the ChemicalToolbox. Colored boxes represent other related Galaxy communities, each with their own tools and workflows which can be flexibly combined with those of the ChemicalToolbox

collections'; two of which, the Chemistry and ADMET collections, provide similar functionality to the ChemicalToolbox. Pipeline Pilot is known for its user-friendly interface and ease of use for new users [30]. However, its proprietary nature makes reproducible research and sharing data very difficult or impossible, and the cost of purchasing a license is prohibitive for many researchers. KNIME, like the ChemicalToolbox, is open-source and free-of-charge, and also leverages well-known open-source software such as the CDK [5, 31] and RDKit in its extensions. KNIME 'nodes' are analogous to Galaxy tools, and are assembled into workflows in a similar manner. However, unlike the ChemicalToolbox, the free version of KNIME is not scalable for usage with an HPC or cloud environment; for this, a commercial license for KNIME Server must be purchased. Furthermore, the experience of using KNIME is comparable to programming with a graphical interface; KNIME describes its workflows as a 'graphic equivalent to a script'. By contrast, the ChemicalToolbox explicitly aims for accessibility to users without programming experience, as the majority of life scientists do not possess these skills.

Offering a cheminformatics toolbox as part of Galaxy has a number of advantages. Firstly, the Galaxy platform is a well-developed, mature project, and while originally developed for genomics research, it is fundamentally agnostic regarding the field of research. The ChemicalToolbox allows chemists to also access the features provided by the Galaxy platform, including a curated body of training material provided by the Galaxy Training Network [32]. Secondly, all ChemicalToolbox tools can be used via the European Galaxy server, which provides free access to generous computational resources for computational analysis, based on the de.NBI cloud [33] and the ELIXIR network [34]. However, the flexibility of the Galaxy system also allows users to download the ChemicalToolbox and run it locally or on their own server. There is already a small but active Galaxy computational chemistry community, constantly maintaining and contributing tools.

Implementation

While the ChemicalToolbox is primarily available via the European Galaxy instance, it has been designed as a dynamic cheminformatics platform, which can be implemented in diverse working environments and architectures. As it is built on top of the Galaxy framework, the ChemicalToolbox can be configured to run on diverse compute clusters, e.g. Kubernetes [35], TORQUE [36], DRMAA [37], Condor [38], or Pulsar [39]. This scalability allows users to perform compute-intensive cheminformatics calculations, including filtering, converting, and calculating hundreds of physicochemical properties and descriptors for many millions of compounds in a matter of hours.

Any software tool that is parameterizable and can be executed through a terminal command line can be wrapped as a Galaxy tool and included into the ChemicalToolbox, regardless of the programming language used for the implementation of the algorithm. Using the Galaxy ToolShed, each tool can be installed through the user's web browser by clicking on the required software—analogous to the 'app stores' provided by companies such as Apple or Microsoft. Moreover, the associated dependencies are automatically downloaded, compiled, and made accessible within the Galaxy environment. As the Galaxy ToolShed supports tool dependency versioning, the ChemicalToolbox is able to keep track of tool versions, enabling reproducibility and maintaining software provenance over time. Tool execution triggers creation of a Conda environment or download of a container with all software requirements installed, all with the specified versions. When executing outdated workflows in the ChemicalToolbox, the user is notified about newer versions of the tools and is allowed to choose specific versions for execution.

Many kinds of calculations in computational chemistry can be easily parallelized; an example is protein-ligand docking, where each of thousands of compounds in a library needs to be assessed individually. In the ChemicalToolbox, this is achieved by the use of collections. A Galaxy collection allows related files to be grouped together and processed identically. Input files (for example, a docking library in SDF format) are split according to defined parameters (the SDF delimiter), and when the AutoDock Vina or rDock tool is run on the resulting collection, docking is performed for each element of the collection separately and in parallel. Such a parallelization process is carried out automatically in the background, and can be easily parameterized and scaled-up by the server administrator responsible for maintaining the ChemicalToolbox as a suitable platform for high-performance computing.

Results

Here we present a number of case studies which demonstrate the capabilities of the ChemicalToolbox. For each case study, tools are chained together to form a 'workflow', which in the Galaxy system can be used much like an individual tool, thus enabling the flexible creation and combination of new functionalities as desired. Each of the workflows is published online under https://usegalaxy.eu/workflows/list_published and labelled with the 'cheminformatics' tag, as are sample Galaxy histories for each of the workflows under https://usegalaxy.eu/histo

ries/list_published. Simplified schematic diagrams of the workflows are provided in Additional file 1, together with individual links to each workflow and history.

Hole filling and library optimization

The correct choice of chemical libraries is a crucial step in high-throughput virtual screening [40]. By using larger libraries, the chances of identifying hits increase, [41] along with the complexity and resources required for proper storage and testing. Moreover, it has been estimated that the chemical space contains more than 10^{60} molecules, a number impossible to handle currently or in the near future [42]. As a consequence, pre-filtered and focused libraries are commonly used in drug discovery, at the risk of exploring a minute portion of the chemical space (from hundreds to millions of compounds) and leaving large regions of the chemical space unexplored. As a result, hole filling and library optimization have assumed a major role in the fields of cheminformatics and drug discovery.

Here we demonstrate a ChemicalToolbox workflow which can be used to optimize a compound library using hole-filling. Downloading all drugs listed on the Therapeutic Target Database [43] (TTD) provides a small library of around 20,000 compounds. For the purpose of this workflow, our aim is to 'top-up' this library to 50,000, ensuring that added compounds are located in more sparsely occupied regions of the chemical space. Initially, we download the entirety of the PubChem database, which serves as the source for the new molecules, before calculating molecular fingerprints (using the Chemfp library [44]) for both PubChem and TTD compounds. Taylor-Butina clustering [45] is then performed on the TTD and singletons are identified, i.e. clusters which contain only a single molecule; these are used as seeds for expansion of the compound library. We then perform a similarity search to identify PubChem compounds within a distance threshold of the TTD singletons just found, which yields a total of around 2 million. In order to select compounds evenly, we perform Taylor-Butina clustering once again on our pool of 2 million molecules. A single compound is then selected from each of 30,000 different clusters, and added to the compound library, topping it up to 50,000.

Ligand library preparation

The preparation of ligand libraries is an important aspect of in silico high-throughput virtual screening, where small molecules are systematically tested in the catalytic or binding site of a protein (for example, via protein-ligand docking) aiming at the selection of candidate compounds with specific structural and physicochemical features. We provide a ChemicalToolbox workflow which offers an efficient solution for the large-scale management of data sets containing millions of molecules.

Initially, the workflow queries several freely available databases (including PubChem, ChEMBL and ZINC [46]) and automatically loads and converts all molecules to canonical SMILES for uniformity using OpenBabel. A specialist tool is used to extract all structures from the PubChem FTP site, while a general download tool can be used to access the other databases. After concatenating the resulting SMILES files and removing counterions and fragments, a final, cleaned dataset of almost 200 million unique compounds in the SMILES format was obtained (databases accessed on 04.10.2019). It is worth mentioning that the ChemicalToolbox has been specifically designed to automatically handle many format files (SDF and SMILES in the present workflow) encoding from a few hundreds or thousands up to many millions of molecules.

Protein-ligand docking

A common aim in cheminformatics is assessing the interactions of compounds with a protein. Protein-ligand docking involves estimating the interaction energy and the optimal recognition pose of a given ligand in complex with a protein [47, 48]. The ChemicalToolbox contains a number of tools which can be used for protein-ligand docking, including docking software AutoDock Vina and rDock. The fpocket tool can also be used for automatic identification of pockets which are suitable for docking [49].

Firstly, a protein structure and a compound library are created, either uploaded by the user or downloaded directly from online databases such as the PDB or ChEMBL. These can be processed using the Filter tool, which can apply either a commonly-used ruleset, such as Lipinski's rule-of-five [50], or a set of user-defined properties. In this case, we use two very different systems as illustrative examples: the Hsp90 chaperone protein (structure published under PDB accession code 2brc [51]) and the β_2-adrenergic receptor (structure published under PDB accession code 3pds [52]). Identification of a binding site allows the definition of a 3D box which is searched (using AutoDock Vina, though rDock is also available) to find a variety of possible binding positions for each of the compounds in the library. Results can be extracted from the output SD files and plotted, or used for further analysis.

Machine learning for predicting small molecule protein interactions

The Galaxy platform contains tools from multiple disciplines, which offers the opportunity to conduct interdisciplinary analyses. Recently, a suite of statistical and machine learning tools has been made available. This allows the development of quantitative structure-activity relationship (QSAR) models in the ChemicalToolbox.

As an illustrative example, we have published a Galaxy workflow for constructing a random forest classifier for predicting the activity of compounds as agonists of the estrogen receptor alpha signaling (ERα) pathway. Data are downloaded directly from the relevant PubChem assay, which forms part of the Tox21 program [53]. Initially, tools based on OpenBabel are used to remove counterions or small fragments from the compound library, as well as any duplicated molecules. For the remaining 7459 compounds, over 1800 two- and three-dimensional molecular descriptors are calculated using the Mordred tool [16] and 21 selected as features for building the classification model. A training/test split of 0.7/0.3 was used and a classification model built using the random forest method (in this case, the number of trees used by the classifier is 100) based on the descriptor values calculated for the training data. The random forest algorithm is applied using the implementation published as part of the scikit-learn Python library [54]. Aside from generation of a model that can be applied to new data, the effectiveness of the model can be tested and the results visualized in the form of a ROC curve, precision, recall and f-score plots, and confusion matrix. Here, an AUC value of 0.72 is achieved, which is reasonable considering the simple approach to feature and parameter selection applied here.

Training material

In addition to publishing the workflows described above, we have also created online tutorials providing an introduction to the features of the ChemicalToolbox, made available via the Galaxy Training Network [32], which already provides a range of introductory and advanced training material for analysis on the Galaxy platform. These tutorials may be found under https://training.galaxyproject.org/training-material/topics/computational-chemistry. For example, the tutorial on protein-ligand docking follows the workflow described above, using a small library of ligands downloaded from ChEMBL and docking them to the Hsp90 protein using AutoDock Vina. In addition, the tutorial guides the user through several other analyses of the compound library, using OpenBabel-based tools to visualize compounds and convert between different formats as required, and performing Taylor-Butina clustering based on calculated chemfp fingerprints.

The Galaxy computational chemistry community has developed a number of other more specialized tutorials, mainly focusing on molecular dynamics simulation and analysis. Other tutorials cover free energy perturbation and the application of machine learning to cheminformatics.

Conclusions

We have prepared the infrastructure and software for the ChemicalToolbox, a Galaxy-based cheminformatics webserver available via https://cheminformatics.usegalaxy.eu, and made a number of workflows available which demonstrate its capabilities, together with accompanying online introductory tutorials. Such a project can by its nature never be complete or comprehensive; new scientific advances will always result in the development of new software and new analytical approaches. However, the ChemicalToolbox is already sufficiently developed to be used to perform novel and interesting analyses, as well as for pedagogical purposes. We hope that the work published so far will provide a useful resource for chemists and cheminformaticians alike. With this publication, we hope to grow the Galaxy computational chemistry community further and to provide an impetus for further development of the ChemicalToolbox.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13321-020-00442-7.

Additional file 1. Additional figures.

Acknowledgements

The authors would like to thank the following for discussions and instructive comments: Rolf Backofen, Chris Barnett, Leo Biscassi, Stefan Bleher, Andrew Dalke, Anika Erxleben, Stefan Günther, Gregory von Kuster, Hitesh Patel and Tharindu Senapathi. We also thank Tim Dudgeon and Rachel Skyner for discussion regarding the docking workflow and Alireza Khanteymoori for discussion regarding the machine learning workflow. Finally, we thank the entire Freiburg Galaxy Team for their support and helpful comments, particularly Joachim Wolff for reading and providing feedback on the paper.

Availability and requirements

Project name: ChemicalToolbox. Project home page: https://cheminformatics.usegalaxy.eu. Operating system(s): The ChemicalToolbox can be hosted on a Linux or OSX operating system; in addition, it is possible to run the ChemicalToolBox in a container on all operating systems that support OCI-approved containers, by means of the Galaxy Docker Project, which we recommend to users who wish to install on a Windows system (https://github.com/bgruening/docker-galaxy-stable). It is accessible for the user via a web browser on any operating system. Programming language: Python. Other requirements: dependencies for individual tools (e.g. RDKit, OpenBabel, GROMACS) installed via Conda or Singularity. License: The Galaxy software is released under the

Authors' contributions
SAB, XL and BAG planned the project and developed the software. SAB developed the training material and the workflows presented in the paper. AK integrated the visualizations and provided advice on use of Galaxy machine learning tools. BAG supervised the project. All authors contributed to writing the paper. All authors read and approved the final manuscript.

Funding
This work was supported by funding from the following organizations: S.A.B. was funded by the European Open Science Cloud (EOSC-Life) (Grant No. 824087); X.L. was funded by the Federal State of Baden Württemberg (Ministry of Arts and Science) through the "Zukunftsoffensive IV (ZO IV) Juniorprofessoren-Programm"; A.K. was funded by the German Federal Ministry of Education and Research [031 A538A de.NBI-RBC]; and B.A.G was funded by the German Research Foundation for the Collaborative Research Center 992 Medical Epigenetics [SFB 992/1 2012 and SFB 992/2 2016].

Availability of data and materials
All Galaxy tools are available via the Galaxy ToolShed under https://toolshed.g2.bx.psu.edu and may be executed on the publicly available Galaxy server https://cheminformatics.usegalaxy.eu. Example histories and workflows are available on https://cheminformatics.usegalaxy.eu via the links provided in the article text.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, Freiburg, Germany. [2] Roche Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Grenzacherstrasse 124, Basel, Switzerland.

Received: 4 February 2020 Accepted: 16 May 2020
Published online: 01 June 2020

References

1. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E et al (2016) The ChEMBL database in 2017. Nucleic Acids Res 45(D1):945–954
2. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA et al (2015) PubChem substance and compound databases. Nucleic Acids Res 44(D1):1202–1213
3. Landrum G (2019) RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org. Accessed 23 Jan 20.
4. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) OpenBabel: an open chemical toolbox. J Cheminform 3(1):33
5. Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, Kuhn S, Pluskal T, Rojas-Chertó M, Spjuth O et al (2017) The Chemistry Development Kit (CDK) v2. 0: atom typing, depiction, molecular formulas, and substructure searching. J Cheminform 9(1):33
6. Trott O, Olson AJ (2009) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31(2):455–461
7. Ruiz-Carmona S, Alvarez-Garcia D, Foloppe N, Garmendia-Doval AB, Juhos S, Schmidtke P, Barril X, Hubbard RE, Morley SD (2014) rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. PLoS Comput Biol 10(4):1003571
8. Turney JM, Simmonett AC, Parrish RM, Hohenstein EG, Evangelista FA, Fermann JT, Mintz BJ, Burns LA, Wilke JJ, Abrams ML et al (2012) Psi4: an open-source ab initio electronic structure program. Wiley Interdiscip Rev Comput Mol Sci 2(4):556–565
9. Taschuk M, Wilson G (2017) Ten simple rules for making research software more robust. PLoS Comput Biol 13(4)
10. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Köster J (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods 15(7):475
11. Merkel D (2014) Docker: lightweight Linux containers for consistent development and deployment. Linux J 2014(239):2
12. Boettiger C (2015) An introduction to Docker for reproducible research. ACM SIGOPS Oper Syst Rev 49(1):71–79
13. Kurtzer GM, Sochat V, Bauer MW (2017) Singularity: scientific containers for mobility of compute. PloS ONE 12(5):0177459
14. Blankenberg D, Kuster GV, Bouvier E, Baker D, Afgan E, Stoler N, Taylor J, Nekrutenko A (2014) Dissemination of scientific software with Galaxy ToolShed. Genome Biol 15(2):403
15. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Eberhard C, Grüning B, Guerler A, Hillman-Jackson J, Von Kuster G, Rasche E, Soranzo N, Turaga N, Taylor J, Nekrutenko A, Goecks J (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res 44(W1):3–10
16. Moriwaki H, Tian Y-S, Kawashita N, Takagi T (2018) Mordred: a molecular descriptor calculator. J Cheminform 10(1):4
17. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32(7):1466–1474
18. Senapathi T, Bray S, Barnett CB, Grüning B, Naidoo KJ (2019) Biomolecular Reaction & Interaction Dynamics Global Environment (BRIDGE). Bioinformatics 35(18):3508–3509
19. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E (2015) GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 1–2:19–25
20. Case D, et al. (2018) AmberTools Manual 2018. University of California, San Francisco. University of California. http://ambermd.org/doc12/Amber18.pdf. Accessed 23 Jan 20.
21. Swails J, Hernandez C, Mobley D, Nguyen H, Wang L, Janowski P (2016) ParmEd: Cross-program parameter and topology file editor and molecular mechanical simulator engine. https://parmed.github.io/ParmEd/html/index.html. Accessed 23 Jan 20.
22. Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O (2011) MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. J Comput Chem 32(10):2319–2327
23. Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlić A, Rose PW (2018) NGL viewer: web-based molecular graphics for large complexes. Bioinformatics 34(21):3755–3758
24. Guitton Y, Tremblay-Franco M, Corguillé GL, Martin J-F, Pétéra M, Roger-Mele P, Delabrière A, Goulitquer S, Monsoor M, Duperier C, Canlet C, Servien R, Tardivel P, Caron C, Giacomoni F, Thévenot EA (2017) Create, run, share, publish, and reference your LC–MS, FIA–MS, GC–MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 galaxy online infrastructure for metabolomics. Int J Biochem Cell Biol 93:89–101
25. van Rijswijk M, Beirnaert C, Caron C, Cascante M, Dominguez V, Dunn WB, Ebbels TMD, Giacomoni F, Gonzalez-Beltran A, Hankemeier T, Haug K, Izquierdo-Garcia JL, Jimenez RC, Jourdan F, Kale N, Klapa MI, Kohlbacher O, Koort K, Kultima K, Corguillé GL, Moschonas NK, Neumann S, O'Donovan C, Reczko M, Rocca-Serra P, Rosato A, Salek RM, Sansone S-A, Satagopam V, Schober D, Shimmo R, Spicer RA, Spjuth O, Thévenot EA, Viant MR, Weber RJM, Willighagen EL, Zanetti G, Steinbeck C (2017) The future of metabolomics in ELIXIR. F1000Research 6:1649
26. Stewart PA, Kuenzi BM, Mehta S, Kumar P, Johnson JE, Jagtap P, Griffin TJ, Haura EB (2019) The Galaxy platform for reproducible affinity proteomic mass spectrometry data analysis. In: Methods in molecular biology. Springer, New York, p. 249–61
27. Accelrys: BIOVIA Pipeline Pilot. 2019. https://www.3dsbiovia.com/products/collaborative-science/biovia-pipeline-pilot. Accessed 23 Jan 20.
28. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Thiel K, Wiswedel B (2009) KNIME—the Konstanz Information Miner: version 2.0 and beyond. ACM SIGKDD Explor Newsl 11(1):26–31
29. KNIME: Konstanz Information Miner. 2020. https://www.knime.com/. Accessed 31 Mar 20.
30. Warr WA (2012) Scientific workflow systems: Pipeline Pilot and KNIME. J Comput Aided Mole Des 26(7):801–804
31. Beisken S, Meinl T, Wiswedel B, de Figueiredo LF, Berthold M, Steinbeck C (2013) KNIME-CDK: workflow-driven cheminformatics. BMC Bioinform 14(1):257

32. Batut B, Hiltemann S, Bagnacani A, Baker D, Bhardwaj V, Blank C, Bretaudeau A, Brillet-Guéguen L, Čech M, Chilton J et al (2018) Community-driven data analysis training for biology. Cell Syst 6(6):752–758
33. German Network for Bioinformatics Infrastructure: de.NBI cloud. 2020. https://www.denbi.de/cloud. Accessed 31 Mar 20.
34. ELIXIR network: ELIXIR. 2020. https://elixir-europe.org/. Accessed 31 Mar 20.
35. Kubernetes: Production-Grade Container Orchestration. 2020. https://kubernetes.io/. Accessed 31 Mar 20.
36. Adaptive Computing: QUEue Manager (TORQUE). 2013. http://www.adaptivecomputing.com/products/torque. Accessed 23 Jan 20.
37. Troger P, Rajic H, Haas A, Domagalski P (2007) Standardization of an API for distributed resource management systems. In: Seventh IEEE international symposium on cluster computing and the grid (CCGrid 2007). IEEE, Rio de Janeiro
38. Tannenbaum T, Wright D, Miller K, Livny M (2001) Condor—a distributed job scheduler. In: Sterling T (ed) Beowulf cluster computing with Linux. MIT Press, Cambridge
39. Chilton J. Pulsar. 2019. https://github.com/galaxyproject/pulsar. Accessed 23 Jan 20.
40. Kumar V, Krishna S, Siddiqi MI (2015) Virtual screening strategies: recent advances in the identification and design of anti-cancer agents. Methods 71:64–70
41. Lyu J, Wang S, Balius TE, Singh I, Levit A, Moroz YS, O'Meara MJ, Che T, Algaa E, Tolmachova K, Tolmachev AA, Shoichet BK, Roth BL, Irwin JJ (2019) Ultra-large library docking for discovering new chemotypes. Nature 566(7743):224–229
42. Bohacek RS, McMartin C, Guida WC (1996) The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev 16(1):3–50
43. Li YH, Yu CY, Li XX, Zhang P, Tang J, Yang Q, Fu T, Zhang X, Cui X, Tu G et al (2017) Therapeutic Target Database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. Nucleic Acids Res 46(D1):1121–1127
44. Dalke A (2013) The FPS fingerprint format and chemfp toolkit. J Cheminform 5(1):36
45. Butina D (1999) Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: a fast and automated way to cluster small and large data sets. J Chem Inf Comput Sci 39(4):747–750
46. Sterling T, Irwin JJ (2015) ZINC 15-ligand discovery for everyone. J Chem Inf Model 55(11):2324–2337
47. Pagadala NS, Syed K, Tuszynski J (2017) Software for molecular docking: a review. Biophys Rev 9(2):91–102
48. Sousa SF, Fernandes PA, Ramos MJ (2006) Protein-ligand docking: current status and future challenges. Proteins Struct Funct Bioinform 65(1):15–26
49. Le Guilloux V, Schmidtke P, Tuffery P (2009) Fpocket: an open source platform for ligand pocket detection. BMC Bioinform 10(1):168
50. Lipinski CA (2004) Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov Today Technol 1(4):337–341
51. Cheung K-MJ, Matthews TP, James K, Rowlands MG, Boxall KJ, Sharp SY, Maloney A, Roe SM, Prodromou C, Pearl LH, Aherne GW, McDonald E, Workman P (2005) The identification, synthesis, protein crystal structure and in vitro biochemical evaluation of a new 3,4-diarylpyrazole class of Hsp90 inhibitors. Bioorg Med Chem Lett 15(14):3338–3343
52. Rosenbaum DM, Zhang C, Lyons JA, Holl R, Aragao D, Arlow DH, Rasmussen SGF, Choi H-J, DeVree BT, Sunahara RK, Chae PS, Gellman SH, Dror RO, Shaw DE, Weis WI, Caffrey M, Gmeiner P, Kobilka BK (2011) Structure and function of an irreversible agonist-β_2 adrenoceptor complex. Nature 469(7329):236–240
53. National Center for Advancing Translational Sciences: Tox21 Data Challenge 2014. 2014. https://tripod.nih.gov/tox21/challenge. Accessed 23 Jan 20
54. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

Publisher's Note