HHS Public Access

Author manuscript

J Chem Inf Model. Author manuscript; available in PMC 2017 December 07.

Published in final edited form as:

J Chem Inf Model. 2017 February 27; 57(2): 105–108. doi:10.1021/acs.jcim.6b00462.

Chembench: A Publicly-Accessible, Integrated Cheminformatics Portal

Stephen J. Capuzzi[1], **Ian Sang-June Kim**[1], **Wai In Lam**[2], **Thomas E. Thornton**[1], **Eugene N. Muratov**[1], **Diane Pozefsky**[2,*], and **Alexander Tropsha**[1,2,*]

[1]Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, 27599, USA

[2]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Abstract

The enormous increase in the amount of publicly available chemical genomics data and the growing emphasis on data sharing and open science mandates that cheminformaticians make their models publicly available for broad use by the scientific community. Chembench is one of the first publicly-accessible, integrated cheminformatics Web portals. It has been extensively used by researchers from different fields for curation, visualization, analysis, and modeling of chemogenomics data. Since its launch in 2008, Chembench has been accessed more than 1 million times by more than 5K users from a total of 98 countries. We report on the recent updates and improvements that increase the simplicity of use, computational efficiency, accuracy, and accessibility of a broad range of tools and services for computer-assisted drug design and computational toxicology available on Chembench. Chembench remains freely accessible at https://chembench.mml.unc.edu

Graphical Abstract

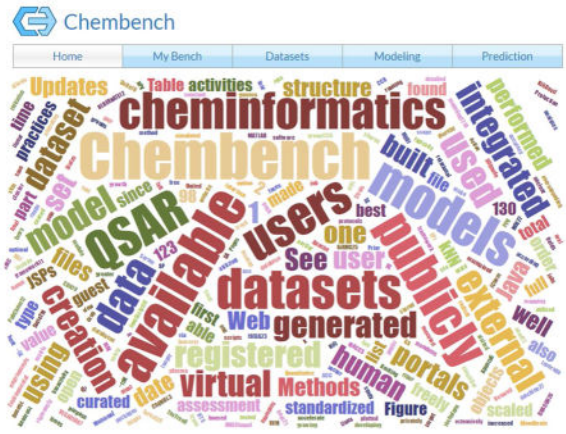
[*]please address the correspondence to: alex_tropsha@unc.edu or pozefsky@cs.unc.edu.

Competing financial interests

The authors declare no competing financial interests.

SUPPORTING INFORMATION

Details about publicly available datasets and predictors are provided in the Supporting Information. For the 86 publicly available datasets on Chembench, the name, size and type (categorical or classification) of the dataset are provided, as well as the descriptors available for modeling, the modelability index, and the date of creation (Table S1). For the 124 publicly available predictors, the name of the model and the underlying dataset, the statistical evaluation of the predictor, the machine learning technique and descriptors used, and the date of creation are all provided (Table S2). All publicly available datasets and predictors are available at https://chembench.mml.unc.edu/.



INTRODUCTION

The scientific community faces an unprecedented growth of chemical genomics data[1] caused by widespread proliferation of high-throughput screening (HTS) technologies in both academia and industry and ever-expanding, public databases of chemical structure and bioactivity, such as PubChem[2,3] and ChEMBL.[4] Quantitative Structure-Activity Relationship (QSAR) models can leverage these data to accelerate chemogenomics research and chemical safety assessment while simultaneously reducing the associated costs.[1,5,6] At the same time, there has been a community-wide push for increased openness in data and model sharing.[1,6] This trend is posed to promote greater cross-disciplinary collaborations and to advance the relevance and rigor of predictive modeling. This evolution in the culture of data science mandates cheminformatics groups to provide the scientific community with the free and open access to QSAR models.

Currently, freely-accessible QSAR models are typically shared through standalone software applications. Such examples include EPI Suite™ from the United States Environmental Protection Agency[7], VEGA-QSAR[8], and ToxTree.[9] While these standalone applications provide access to some QSAR models, new models cannot be developed by the user, thereby limiting the scope of available models. The optimal method for both developing and sharing QSAR models is the creation of freely-accessible, integrated cheminformatics Web portals.

Chembench[10] is one of the first publicly-accessible, integrated cheminformatics portals, along with OCHEM[11]. Chembench has been used for research in chemical genomics, drug discovery, computational toxicology, and other fields. Since its creation in 2008, Chembench has been extensively used for the development and distribution of QSAR models, as well as a teaching tool. To date, Chembench has been visited over one million times. Chembench has more than 550 active registered users from 38 different countries, with an additional ~5K individual non-registered users from a total of 98 countries (Figure 1). In order to keep Chembench as a state-of-the-art Web portal during this time of growing interest in open science and molecular modeling, constant updates and upgrades must be made. Here, we aim to report on the updates and improvements made in the recent years to increase the simplicity, speed, reliability, and openness of Chembench.

METHODS

Chembench is a Java-based system, utilizing Java Server Pages (JSPs) with JavaScript[12] at the front end of the website. The interface between data located on the JSPs and Java objects is maintained by the Apache Struts 2 framework.[13] HIBERNATE[14] provides the framework for mapping the Java objects to a relational database. Chembench is freely accessible at https://chembench.mml.unc.edu/.

Prior to QSAR modeling, the datasets are automatically curated following the protocols developed earlier in our group.[15–17] For structural standardization and compound visualization J Chem Suite[18] is used. Visualization scripts are executed using the R environment.

The models are built and rigorously validated according to best practices of QSAR modeling.[19,20] Chembench implements several chemical descriptor generation packages – CDK,[21] DRAGON,[22] MACCS keys,[23] MOE,[24] and ISIDA.[25] Molconn Z descriptors[26] are no longer supported; however, descriptors files for archived datasets are still available for download. In addition, users can also upload their datasets characterized by any descriptors precomputed outside of Chembench. The following machine learning algorithms are supported by Chembench for both continuous and classification model building: random forest,[27] support vector machine,[28] k-nearest neighbors (kNN) with genetic algorithm (GA) or simulated annealing (SA) descriptor selection.[29] Predictors are built using the scikit-learn package from Python.[30] Single compounds for prediction can be sketched using JSME, a free molecule editor written in JavaScript.[31] Calculations are performed on the KillDevil 800-node Beowulf Linux cluster housed at UNC-Chapel Hill.

CHEMBENCH ENVIRONMENT

Chembench facilitates cheminformatics analyses via four modules described below: My Bench; Datasets; Modeling; and Prediction. Each module can be utilized individually or integrated as part of an integrated study design.

My Bench

Every dataset, predictor (QSAR model), and prediction created by a register user on Chembench is privately stored and available for personal download. After receiving approval from the Chembench management team, registered users have the option to make all datasets, predictors, and predictions publicly available. Both registered and guest (non-registered) users are able to download all publicly available datasets and predictors. Curated datasets downloaded from Chembench contain, among other files, the standardized structure file and generated descriptor matrices. Predictors, when downloaded, contain the non-overlapping training and test sets used in each fold during cross-validation and the underlying Python scripts used for model building. Guest users are not able to download datasets and predictors associated with proprietary descriptors. Users can track the progress of all running jobs using the job queue feature.

Datasets

Chembench facilitates the creation of datasets for model building and validation. Modeling datasets can be used for either predictor generation or virtual screening, while prediction datasets are used exclusively for virtual screening. Users have the option to upload proprietary descriptors; otherwise, available descriptors (see Methods) are automatically generated from an uploaded structure file. The modelability index (MODI) of each dataset is calculated automatically.[32] Structures are standardized following our chemical data curation workflow,[16] and a chemical similarity heat maps, using either Tanimoto similarity[33] or Mahalanobis distance measure,[34] can be generated as an option. Rigorous external validation is an inherent part of model building. For this purpose, specific set of compounds can be selected for external validation or, as other options, random external set or n-fold external cross-validation can be used.[19] In order to promote best practices of QSAR modeling,[19] Chembench will automatically warn the user if the modeling dataset is too small (less than 40 compounds) for rigorous QSAR modeling. After the dataset has been created, the user can view chemical structures, examine the heatmap and a histogram of activities, as well as investigate each generated descriptor type and examine possible errors during calculation. Currently, there are 86 publicly available datasets on Chembench, including, for example, datasets of human skin sensitization,[35,36] P-glycoprotein substrates,[37] chemical toxicants tested against *Tetrahymena pyriformis,*[38] and blood–brain barrier permeability.[39] A full list of publicly available datasets on Chembench can be found in the Supporting Information.

Modeling

Chembench allows for the generation of statistically validated QSAR models of target endpoints for either personally uploaded or publicly available modeling datasets. Generated descriptors (See Methods) or externally uploaded descriptors, if applicable, are available for use in the predictor. Descriptors can be range scaled, auto scaled, or left unscaled. Users can manually set the maximum allowed descriptor cross-correlation. For each pair of descriptors, if the correlation coefficient is above the maximum, one of the two will be chosen randomly and removed; descriptors with zero variance across compounds will be automatically removed as well. It should be noted that if the descriptor type cannot be generated due to incompatible chemical structures (see Datasets), then this descriptor type will be unavailable for use in QSAR modeling. After the model has been built, the robustness of the predictor can be probed through a detailed assessment of external validation statistics. For models built with continuous data, the linear regression is plotted, and the Q^2, RMSE, and MAE are calculated. For models built with categorical data, a confusion matrix is generated from which specificity (SP), sensitivity (SE), correct classification rate (CCR), accuracy (ACC), negative predictive value (NPV), and positive predictive value (PPV) are calculated. All external validation results can be download as *.csv files. For all models, regardless of the machine learning algorithm, y-randomization[40] is performed with a corresponding statistical evaluation. For random forest, the individual trees can be investigated with the selected descriptors displayed. Additionally, feature selection is performed, and the most important descriptors are ranked. For SVM, a matrix search is used and the gamma parameter of each radical basis function (RBF) kernel can be identified. For kNN, the number of k nearest neighbors and descriptors used for each model

can be probed. All users can download publicly available models from Chembench, while only registered users can save, store, and download their personal models on Chembench. Currently, there are 124 publicly available predictors on Chembench that can either be downloaded or used for virtual screening (See Prediction), including, for example, predictors of the human intestinal transporter inhibition,[41] human oral bioavailability,[42] human plasma protein binding,[43] stress response and nuclear receptor signaling toxicity assays.[44] A full list of publicly available predictors can be found in the Supporting Information.

Prediction

Chembench possesses several prediction modalities for single compounds, batches of multiple compounds, and virtual chemical libraries. For instance, a single compound can be sketched using JSME[31] or its SMILES uploaded. Additionally, Chembench has integrated several publicly available chemical libraries, such as the DrugBank[45] and the ZINC lead-like library,[46] that can be used for virtual screening. A user can also upload a specific library of interest or a batch of compounds (See Datasets). Then, the specific activity or spectrum of activities of the compound(s) or the virtual library can be predicted by selecting the desired predictor(s). The user has the ability to predict one or more endpoints for one or more compounds. The applicability domain threshold,[19] if selected, can be manually set. Non-registered users, using publicly available predictors, have the ability to predict single compounds and publicly available chemical libraries. In order to encourage registration, non-registered users cannot perform batch predictions of multiple compounds. It should be noted that models built with user-uploaded, proprietary descriptors cannot be used in the “Predict a Single Compound” component of Chembench, as these descriptors cannot be automatically generated. Moreover, only random forest (RF) models are compatible with the “Predict a Single Compound” component in order to accelerate the speed of the prediction. For the prediction of a single compound using kNN or SVM models, the compound should be uploaded as a dataset, and the “Predict a Dataset” function should be used.

CONCLUSIONS

Chembench implements the best practices of QSAR modeling and validation,[19] and all publicly available models are fully compliant with OECD principles for the validation of (Q)SAR models.[47] Chembench provides a variety of cheminformatics and data science-related services including data curation, standardization, and visualization; descriptor generation; development, rigorous external validation, and interpretation of QSAR models; prediction of a single property or activity profile for compound(s) of interest or prepared virtual screening libraries; targeted design of novel compounds with desired activity profile; etc. While Chembench in its current form is useful for both expert and beginner modellers, it is constantly being updated to meet the needs of the scientific community. Updates in progress are fragment-based structural interpretation of QSAR models, implementation of SiRMS[48] descriptors, and additional datasets for download. Chembench was one of the first cheminformatics Web portals and, since its creation in 2008, continues to position itself as the gold-standard of publicly-accessible, integrated cheminformatics portals. Chembench, along with similar cheminformatics portals such as OCHEM,[11] promotes the principles of both open science and data and model sharing[49] in the era of Big Data.[1] The continuing

need for Chembench and the high quality of services it provides are supported by more than 1 million visits and more 550 registered and ~5K unregistered users from a total of 98 countries as of today. Chembench is freely accessible at https://chembench.mml.unc.edu/.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors appreciate the financial support from NIH (GM096967 and GM66940). We are grateful to Chemical Computing Group, Kode Chemoinformatics, eduSoft, ChemAxon, and Sunset Molecular for their software licenses. We also thank Steven Fishback, Hugh Crissman, and UNC Information Technology Services for their support and are grateful to former, current, and future members of the Molecular Modeling Lab as well as students taking introductory course on molecular modeling taught by AT for their input and help in developing and testing Chembench.

References

1. Tetko IV, Engkvist O, Koch U, Reymond JL, Chen H. BIGCHEM: Challenges and Opportunities for Big Data Analysis in Chemistry. Mol Inform. 2016; 35:615–621. [PubMed: 27464907]
2. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH. PubChem Substance and Compound Databases. Nucleic Acids Res. 2015; 44:D1202–13. [PubMed: 26400175]
3. Wang Y, Suzek T, Zhang J, Wang J, He S, Cheng T, Shoemaker BA, Gindulyte A, Bryant SH. PubChem BioAssay: 2014 Update. Nucleic Acids Res. 2014; 42:D1075–82. [PubMed: 24198245]
4. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. Nucleic Acids Res. 2012; 40:D1100–7. [PubMed: 21948594]
5. Fourches, D. Cheminformatics: At the Crossroad of Eras. In: Gorb, L.Kuz'min, V., Muratov, E., editors. Application of Computational Techniques in Pharmacy and Medicine. Springer; New York: 2014. p. 539-546.
6. Ekins S, Clark AM, Swamidass SJ, Litterman N, Williams AJ. Bigger Data, Collaborative Tools and the Future of Predictive Drug Discovery. J Comput Aided Mol Des. 2014; 28:997–1008. [PubMed: 24943138]
7. U.S. Environmental Protection Agency. [accessed Jul 13: 2016] EPI Suite v 4.11. http://www.epa.gov/opptintr/exposure/pubs/episuitedl.htm
8. Istituto di Ricerche Farmacologiche Mario Negri Milano. [accessed Jul 13: 2016] VEGA-QSAR. http://www.vega-qsar.eu/index.php
9. Patlewicz G, Jeliazkova N, Safford RJ, Worth AP, Aleksiev B. An Evaluation of the Implementation of the Cramer Classification Scheme in the Toxtree Software. SAR QSAR Environ Res. 2008; 19:495–524. [PubMed: 18853299]
10. Walker T, Grulke CM, Pozefsky D, Tropsha A. Chembench: A Cheminformatics Workbench. Bioinformatics. 2010; 26:3000–3001. [PubMed: 20889496]
11. Sushko I, Novotarskyi S, Körner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY, Todeschini R, Varnek A, Marcou G, Ertl P, Potemkin V, Grishina M, Gasteiger J, Schwab C, Baskin II, Palyulin VA, Radchenko EV, Welsh WJ, Kholodovych V, Chekmarev D, Cherkasov A, Aires-de-Sousa J, Zhang Q-Y, Bender A, Nigsch F, Patiny L, Williams A, Tkachenko V, Tetko IV. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. J Comput Aided Mol Des. 2011; 25:533–554. [PubMed: 21660515]
12. McPherson, S. [accessed May 7: 2015] JavaServer Pages: A Developer's Perspective. http://www.oracle.com/technetwork/articles/javase/jsp-135132.html
13. Roughley, I. Starting Struts 2. https://www.infoq.com/minibooks/starting-struts2

14. King G, Bauer C, Anderson MR, Bernard E, Eberole S. HIBERNATE - Relational Persistence for Idiomatic Java. HIBERNATE Community Doc. 2009
15. Fourches D, Muratov E, Tropsha A. Curation of Chemogenomics Data. Nat Chem Biol. 2015; 11:535. [PubMed: 26196763]
16. Fourches D, Muratov E, Tropsha A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. J Chem Inf Model. 2010; 50:1189–1204. [PubMed: 20572635]
17. Fourches D, Muratov EN, Tropsha A. Trust, But Verify II: A Practical Guide to Chemogenomics Data Curation. J Chem Inf Model. 2016; 56:1243–1252. DOI: 10.1021/acs.jcim.6b00129 [PubMed: 27280890]
18. [accessed May 7: 2016] ChemAxon. Version 16.5.30.0. http://www.chemaxon.com/jchem/doc/user
19. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol Inform. 2010; 29:476–488. [PubMed: 27463326]
20. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuz'min VE, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, Tropsha A. QSAR Modeling: Where Have You Been? Where Are You Going To? J Med Chem. 2014; 57:4977–5010. [PubMed: 24351051]
21. [accessed May 7: 2016] The Chemistry Development Kit. CDK Descriptor Calculator (v1.4.6). http://www.rguha.net/code/java/cdkdesc.html
22. Kode. [accessed May 7: 2016] DRAGON 7.0. https://chm.kode-solutions.net/products_dragon.php
23. Accelrys. MACCS structural keys.
24. Chemical Computing Group. [accessed May 7: 2016] QuaSAR-Descriptor. http://www.chemcomp.com/journal/descr.htm
25. Ruggiu F, Marcou G, Varnek A, Horvath D. ISIDA Property-Labelled Fragment Descriptors. Mol Inform. 2010; 29:855–868. [PubMed: 27464350]
26. EduSoft. [accessed Jul 13: 2016] Molconn-Z version 4.0. http://www.edusoft-lc.com/molconn/
27. Breiman L. Random Forests. Mach Learn. 2001; 45:5–32.
28. Vapnik, VN. The Nature of Statistical Learning Theory. Springer New York; New York, NY: 2000.
29. Zheng, Tropsha. Novel Variable Selection Quantitative Structure--Property Relationship Approach Based on the K-Nearest-Neighbor Principle. J Chem Inf Comput Sci. 2000; 40:185–194. [PubMed: 10661566]
30. Python. [accessed Aug 1: 2016] scikit-learn. http://scikit-learn.org/stable/
31. Bienfait B, Ertl P. JSME: A Free Molecule Editor in JavaScript. J Cheminform. 2013; 5:24. [PubMed: 23694746]
32. Golbraikh A, Muratov E, Fourches D, Tropsha A. Data Set Modelability by QSAR. J Chem Inf Model. 2014; 54:1–4. [PubMed: 24251851]
33. Tanimoto, T. IBM Internal Report. Armonk: IBM Corp; 1957.
34. Mahalanobis P. On the Generalised Distance in Statistics. Proc Natl Inst Sci India. 1936; 2:49–55.
35. Alves VM, Muratov E, Fourches D, Strickland J, Kleinstreuer N, Andrade CH, Tropsha A. Predicting Chemically-Induced Skin Reactions. Part I: QSAR Models of Skin Sensitization and Their Application to Identify Potentially Hazardous Compounds. Toxicol Appl Pharmacol. 2015; 284:262–272. [PubMed: 25560674]
36. Alves VM, Muratov EN, Fourches D, Strickland J, Kleinstreuer N, Andrade CH, Tropsha A. Predicting Chemically-Induced Skin Reactions. Part II: QSAR Models of Skin Permeability and the Relationships between Skin Permeability and Skin Sensitization. Toxicol Appl Pharmacol. 2015; 284:273–280. [PubMed: 25560673]
37. de Lima PC, Golbraikh A, Oloff S, Xiao Y, Tropsha A. Combinatorial QSAR Modeling of P-Glycoprotein Substrates. J Chem Inf Model. 2006; 46:1245–1254. [PubMed: 16711744]
38. Zhu H, Tropsha A, Fourches D, Varnek A, Papa E, Gramatica P, Oberg T, Dao P, Cherkasov A, Tetko IV. Combinatorial QSAR Modeling of Chemical Toxicants Tested against Tetrahymena Pyriformis. J Chem Inf Model. 2008; 48:766–784. [PubMed: 18311912]

39. Zhang L, Zhu H, Oprea TI, Golbraikh A, Tropsha A. QSAR Modeling of the Blood-Brain Barrier Permeability for Diverse Organic Compounds. Pharm Res. 2008; 25:1902–1914. [PubMed: 18553217]
40. Kuz'min VE, Muratov EN, Artemenko AG, Varlamova EV, Gorb L, Wang J, Leszczynski J. Consensus QSAR Modeling of Phosphor-Containing Chiral AChE Inhibitors. QSAR Comb Sci. 2009; 28:664–677.
41. Sedykh A, Fourches D, Duan J, Hucke O, Garneau M, Zhu H, Bonneau P, Tropsha A. Human Intestinal Transporter Database: QSAR Modeling and Virtual Profiling of Drug Uptake, Efflux and Interactions. Pharm Res. 2013; 30:996–1007. [PubMed: 23269503]
42. Kim MT, Sedykh A, Chakravarti SK, Saiakhov RD, Zhu H. Critical Evaluation of Human Oral Bioavailability for Pharmaceutical Drugs by Using Various Cheminformatics Approaches. Pharm Res. 2014; 31:1002–1014. [PubMed: 24306326]
43. Zhu XW, Sedykh A, Zhu H, Liu SS, Tropsha A. The Use of Pseudo-Equilibrium Constant Affords Improved QSAR Models of Human Plasma Protein Binding. Pharm Res. 2013; 30:1790–1798. [PubMed: 23568522]
44. Capuzzi SJ, Politi R, Isayev O, Farag S, Tropsha A. QSAR Modeling of Tox21 Challenge Stress Response and Nuclear Receptor Signaling Toxicity Assays. Front Environ Sci. 2016; 4:3.
45. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets. Nucleic Acids Res. 2008; 36:D901–6. [PubMed: 18048412]
46. Sterling T, Irwin JJ. ZINC 15--Ligand Discovery for Everyone. J Chem Inf Model. 2015; 55:2324–2337. [PubMed: 26479676]
47. OECD Principles for the Validation, For Regulatory Purposes, of (Quantitative) Structure–Activity Relationship Models. Organisation for Economic Co-operation and Development; Paris: 2004. http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf [accessed June 13: 2016]
48. Kuz'min VE, Artemenko aG, Muratov EN. Hierarchical QSAR Technology Based on the Simplex Representation of Molecular Structure. J Comput Aided Mol Des. 2008; 22:403–421. [PubMed: 18253701]
49. Tetko IV, Maran U, Tropsha A. Public (Q)SAR Services, Integrated Modeling Environments, and Model Repositories on the Web: State of the Art and Perspectives for Future Development. Mol Inform. 2016; 0:1–14.

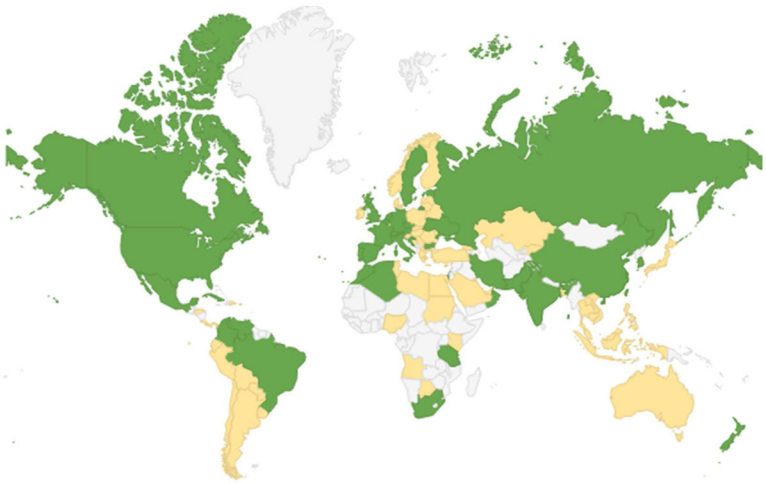


Figure 1.
World map of Chembench users. Countries with registered and guest users are shown in green and yellow, respectively.