Open Innovation Platform using Cloud-based Applications and Collaborative Space: A Case Study of Solubility Prediction Model Development

Tsuyoshi Esaki1*, Keiko Kumazawa2, Kazutoshi Takahashi3, Reiko Watanabe4, Tomohide Masuda5, Hirofumi Watanabe6, Yugo Shimizu7, Akitoshi Okada8, Seisuke Takimoto8, Tomohiro Shimada9, Kazuyoshi Ikeda7,10

1 The Center for Data Science Education and Research, Shiga University, 1-1-1 Banba, Hikone, Shiga, 522-8522, Japan
2 Pharmaceutical Discovery Research Laboratories, Teijin Pharma Limited, 4-3-2 Asahigaoka, Hino-shi, Tokyo 191-8512, Japan
3 Research Institute for Bioscience Products and Fine Chemicals, Ajinomoto Co. Inc., 1-1 Suzuki-cho, Kawasaki, Kawasaki, 210-8681, Japan
4 Laboratory of Bioinformatics, AI Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition, 7-6-8, Saito Asagi, Ibaraki, Osaka 567-0085, Japan
5 Pharmaceutical Research Laboratories, Toray Industries, Inc., 6-10-1 Tebira, Kamakura, Kanagawa, 248-8553, Japan
6 Education Center on Computational Science and Engineering, Kobe University, 7-1-48, Minatojimaminamimachi, Chuo-ku, Kobe 650-0047, Japan
7 Division of Physics for Life Functions, Keio University Faculty of Pharmacy, 1-5-30 Shibakoen, Minato-ku, Tokyo 105-8512, Japan.
8 Central Pharmaceutical Research Institute Takatsuki Research Center, Japan Tobacco Inc., 1-1 Murasaki-cho, Takatsuki, Osaka 569-1125, Japan
9 Administration Department, Planning & Control Division, Teijin Pharma Limited, Kasumigaseki Common Gate West Tower 2-1, Kasumigaseki 3, Chiyoda-ku, Tokyo 100-8585, Japan
10 Center for Life Science Technologies, RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan
*E-mail: tsuyoshi-esaki@biwako.shiga-u.ac.jp

(Received December 5, 2019; accepted April 12, 2020; published online April 27, 2020)

Abstract

In recent years, with the emergence of new technologies employing information science, open innovation and collaborative drug discovery research, utilizing biological and chemical experimental data, have been actively conducted. The Young Researcher Association of Chem-Bio Informatics Society (“CBI Wakate”) has constructed an online discussion space using Slack and provided a cloud-based collaborative platform in which researchers have freely discussed specific issues and aimed at raising the level of cross-sectoral communication regarding technology and knowledge. On this platform, we created three channels—dataset, model evaluation and scripts—where participants with different backgrounds co-developed a solution for solubility prediction. In the dataset channel, we exchanged our knowledge and
methodology for calculations using the chemical descriptors for the original dataset and also discussed methods to improve the dataset for pharmaceutical purposes. We have also developed a protocol for evaluating the applicability of solubility prediction models for drug discovery by using the ChEMBL database and for sharing the dataset among users on the cloud. In the model evaluation channel, we discussed the necessary conditions for the prediction model to be used in daily drug discovery research. We examined the effect of these discussions on script development and suggested future improvements. This study provides an example of a new cloud-based open collaboration that can be useful for various projects in the early stage of drug discovery.

Key Words: Solubility, in-silico prediction, Data science, Open innovation, Collaborative drug discovery, Cross-sectoral technology

Area of Interest: In silico drug discovery, emerging new technology
1. Introduction

In recent years, with emergence of new technologies such as artificial intelligence (AI) [1] and data science [2], various researches using information science approaches to bio and chemical experimental data have been actively conducted [3]. The Chem-Bio Informatics (CBI) Society was established in 1991. To date, researchers from many different industries including pharmaceutical, food, and IT companies as well as academic researchers have participated in the activities of the CBI Society [4]. In the CBI field of research, there are increasing demands for developing new methods for solving problems such as discovery of drug development seeds aimed at difficult treatment goals, searching for biomarkers, and production of highly safe drugs and foods. Machine learning methods and AI have already been successful in various fields [5][6]. Several examples have been also reported in this research field that are better than conventional methods [7][8]. These technologies require a large amount of training data, and automated processes in biology and drug discovery chemistry have not progressed at the same rate as in other disciplines. Therefore, it has become necessary to share human resources and information cross-sectionally, using technology in order to attain a broader perspective on these issues.

2. What is open and innovative collaborative drug research (OICDD) and its goal?

The current drug development requires cross-sectional knowledge and skills. However, there is shortage of human resources and information for effective proceedings on drug discovery. It is necessary to overcome this situation. For this reason, it is necessary to establish a communication space and promote exchanges between scientists (especially active young researchers) in various fields. Moreover, commercial tools have a limited number of studies available for validation while open tools, which assist in communication, are accessible to more scientists. This requires open and innovative collaborative drug discovery research (OICDD) as one of the solutions. Considering this, open innovation is attracting attention, and various collaborations for predicting the properties of compounds in the early stage of drug development have been made [9]. In one application, several competitions (or contests, assessments) were held for protein engineering and hit compound acquisition [10][11]. Informatics approaches are indispensable for drug discovery research in recent years. It is important to create a forum for researchers with diverse knowledge and skills in various fields such as IT and computational science. Therefore, OICDD aims at uncovering specific issues in drug discovery by working with a wide range of participants, using those key technologies. Another objective is to make an objective assessment of the latest technologies. This has been difficult within a single company or organization. However, there are problems with OICDD. Open innovation can often be scaled down over time because it is difficult to generate results that link to the benefits of all participants. Academic researchers are required to publish their new results; however, industry researchers are unable to share their results because of intellectual ownership policies.

Several successful examples of collaborative work between companies and universities or between companies [11] and individuals [12] have been reported. This shows the barrier between companies and academia has become lower [10]. In the CBI field, there is a need for cross-sectional discussion among young researchers (i.e., early-career researchers), and various proposals beyond the research fields based on their innovative ideas are required. The CBI Young Association, “CBI Wakate”, was established in 2018 for the purpose of exchange and development of young
researchers [13]. Thus far, we have held two lectures [14], and provided an online discussion space using Slack [15], a cloud-based communication tool.

In this study, we constructed a system in which participants freely discussed specific issues and aimed at raising the level of cross-sectoral technology and knowledge. This allows experimental researchers to acquire knowledge and technology in information science, and computational researchers can get the chance to gain knowledge and ideas in the field of drug discovery. At the CBI 2019 conference [3], we organized a luncheon seminar called “New Value Creation Project” [16] and discussed the analysis methods and ideas with the audience who did not participate in the Slack. This study proposes a new form of collaboration in which various researchers with different backgrounds construct a prediction model of solubility prediction through their discussions. Thus, we discuss how it affects the improvement of the prediction model. We used several products in this study, and they are summarized in Table 1.

Table 1. Summary of products used in this study

| Product name     | Explanation                                      | How to use                        | Reference |
|------------------|--------------------------------------------------|-----------------------------------|-----------|
| Slack            | Cloud-based communication tool                   | Sharing ideas and discussion      | [15]      |
| KNIME            | Integration tool to create and productionize data science | Preparation of dataset            | [18]      |
| Google Drive     | Cloud-based file exchange tool                   | Sharing dataset                   | [23]      |
| Google Colaboratory | Cloud platform to perform coding of a Python script in Jupyter notebook | Developing and sharing Python script | [28]      |

3. Case Study of Collaborative Platform on Slack

This research project used Slack as a collaborative platform (Figure 1) [17]. Slack is a cloud-based communication tool developed by Slack Technologies. Users can create workspaces, participate in the discussion forums (i.e., channels) for each topic, and communicate among users using a chat tool. In addition, the Slack application programming interface (API) that works with Google Drive and other web services and tools is open to the public and can also incorporate external applications provided by third parties. One of the features of Slack is that it can be used to eliminate barriers between teams and integrate the system (https://slack.com/intl/ja-jp/about). Thus far, it has been difficult for IT engineers, who do not have knowledge about chemistry or biology, to tackle issues of drug discovery. By creating and utilizing a system on Slack, it has been made possible for key players in developing drugs to participate in solving issues in the early stages of drug discovery.

In the dataset channel, the generation of compound sets for model validation was developed using KNIME [18].
Figure 1. Concept of this project. The research topic is solubility prediction. We constructed an environment for sharing ideas using Slack. Next, a meeting via Skype is held to prepare the environment for discussion amongst each Slack channel between core members. Teams constructed by researchers in various majors used Slack to divide their research process into three channels in order to best share knowledge and comment and improve upon each other’s ideas. The channels are as follows: 1) Dataset: Discuss how to import data and prepare several blind datapoints. 2) Model validation: Discuss models usable for experiments. 3) Script: Discuss how to share programming codes. Here, we set a co-development model via Google Colaboratory. In a face-to-face meeting, we present the results about discussion and requirements for each channel. Finally, assessment and feedback are designed. Did we create “New Values”? There are three points: 1) Did this collaboration influence to other channels? 2) Could we find seeds for next collaborations? 3) Did young scientists grow up?

4. Preparing Slack Channels According to Roles

An open discussion space called “CBI Wakate” was created in Slack (Figure 2). There are three different channels (dataset, model evaluation, and script). The CBI Wakate Slack is available from the URL (https://slack.com/). If any users are interested in viewing the Slack workspace, they may contact the authors of this paper, and they will be invited.
Figure 2. A snapshot of CBI Wakate Slack front page.
The left side of the front page shows channels (yellow square). The selected channel is shown in the center of page as topics (green circle). New opinions and requests are written in the place. The right side is threads which shows the discussion in each topic (blue square).

4.1 Dataset Channel

In the dataset channel, shown in Figure 3, the dataset used for solubility prediction was discussed amongst the users. A total of 6 participants, consisting of 4 academic researchers (university and national research institute) and 2 industry researchers. Four of the participants focused on computational chemistry (i.e., "Dry" research) such as bioinformatics and simulation, while the other two were based on “Wet” research in the fields of biology or organic synthesis. Thus, in this dataset channel, we prepared an open space where participants with various backgrounds, including researchers who do not specialize in in-silico research, can discuss beyond each organization.
Figure 3. A Snapshot of dataset channel in the CBI Wakate Slack. Several key topics related to generating datasets suggested by different users are shown in the center of the snapshot. A specific discussion for each topic took place in the thread on the right side of the snapshot. The language used in discussions is in Japanese.

We created threads for discussing and exchanging opinions amongst the participants about five particularly important topics. The list of topics for the threads is as follows:

i. Compound datasets used for the solubility prediction
ii. How to generate datasets from compound structure data
iii. Accuracy of descriptor calculation and 3D structure optimization
iv. How to share datasets with Script channel
v. What is important for the validation dataset?

i) First, we discussed the aqueous solubility prediction methods among the participants described in the previous study [19] on the Slack in order to understand how to collect the training data for this prediction model. ii) Next, we investigated the method of descriptor calculation used in the original paper [19], and also discussed the descriptor calculation method commonly used in the input file for predicting logS value from SMILES, and exchanged the information and exchanged the information. iii) Using this process, the experiences, knowledge, and skills of each researcher were exchanged amongst the participants regarding the accuracy of typical descriptor calculation software and 3D structure optimization. Through discussion, we noticed that Dragon and Pipeline Pilot [20] can be used as representative descriptor calculations for commercial tools, and RDKit [21] can be used as that for free tools. Furthermore, computational robustness is as important as its accuracy, especially in the preprocessing of 3D structure optimization prior to generating drug-like compound datasets. Therefore, we concluded that it is important to select appropriate tools for
descriptor calculation based on the above findings and past benchmarks from the previous studies [22]. iv) In addition, the technical solutions on how to utilize the results of the discussion in "dataset channel" into "script channel" were investigated. As a technical solution, a method utilizing cloud-based file exchange software such as Google Drive [23] or Dropbox [24] was proposed. Actually, we created some files for this channel on Dropbox. The files and protocols uploaded on the shared file are as shown below (Figure 4).

Figure 4. A snapshot of dataset files for dataset channel on Google Drive.

Compound datasets for evaluation were created with KNIME and saved to Google Drive. The ChEMBL database (version 22) was used for generation of evaluation compound datasets. Three different datasets, consisting of small molecule drugs (chembl_22_drugs.sdf), preclinical test compounds (chembl_22_candidates.sdf), and bioactive compounds (chembl_22_chembl.sdf), have been stored. Original dataset (13321_2018_308_MOESM1_ESM.csv) is also stored. These dataset files in Google Drive can be read from scripts developed on Google Colaboratory.

v) It was pointed out during the discussion in the model validation channel that it is important to verify the training dataset using a dataset based on known drug discovery data such as ChEMBL [25] data. We implemented a questionnaire function into Slack to ask the participants which database to use to create the evaluation dataset for this channel, and the size of the evaluation dataset. As a result, ChEMBL was selected for the evaluation dataset, and the number of compounds was decided to be 3000 or more. Therefore, we used KNIME [18], a data integration software, to extract information from the ChEMBL database and create three different datasets (small molecule drugs, clinical candidates, and bioactive compounds) for evaluation.

Finally, we concluded that it is important to use private or blinded data (i.e., unknown data that has not yet been released) for model accuracy validation. As a result of our discussions among researchers from pharmaceutical companies and academia, it was suggested that not only in-house data in pharmaceutical companies but also outcomes of national research projects could be used as blind datasets. However, it is claimed that it is difficult to share in-house data with external people for confidentiality reasons. It will be a future discussion of whether blinded datasets can be used for evaluation purposes.
4.2 Model Validation Channel

In the model validation channel, we aimed to have an interactive discussion with other channel participants about which points are important for practical use and in which situations and models are expected to be used in the daily drug discoveries.

There were a total of eight members who participated in this channel, of which three were academic researchers and five were industrial researchers.

The list of issues for the threads is as follows:

i. What indices are used for mode validation?

ii. Is it important to show the contribution of the descriptors to the prediction results?

iii. How to evaluate “chemical space”?

In the CBI annual meeting 2019, we further discussed topics i and iii.

For topic i, we talked about specific indices such as accuracy, coefficient of determination (R²), root mean squared error (RMSE), and Cohen’s kappa. In addition, we discussed the changing needs along with the drug discovery process and the influence of the biased dataset. The view championing the importance of fulfilling the medicinal chemist’s needs, such as unthought compound design, intuitive visualization, and so on, was attracting interest. The participants shared their experiences about specific medicinal chemist’s needs or the situations that arise during the drug discovery process and the solutions they offered to each.

As one of the examples of the feedback for the medicinal chemists, many participants commented on the importance of offering information about the contribution of the descriptors for the prediction results. Not only the accuracy of the models, but also the way of visualizing the results affected whether the medicinal chemists used the models or not. We could deepen trust with medicinal chemists through obtaining compounds with good potency by suggesting a new promising design that is unexpected for medicinal chemists.

Regarding topic iii, we discussed the objectives of evaluating chemical space. From the point of view of the proper dataset for drug discovery, both ChEMBL and approved drugs are good reference datasets. Both fingerprint and physicochemical parameters, like molecular weight and logP, are used for principal component analysis (PCA). We also discussed problems for evaluating the chemical space such as:

- Approved drugs have good ADMET properties and are not fit for early drug discovery processes.
- PC1 and PC2 have a low contribution ratio when using fingerprint for PCA.

As for the general remarks, a lot of industrial researchers stated their concerns about chasing accuracy of the models from the dataset with relatively small numbers. They proposed we should put much effort to share the examination process. We would like to have continuous discussions for fostering new value creation, which is useful to all the participants.

4.3 Script Channel

In the script channel, we discussed the development of a script for solubility predictions. One of our goals is to share the constructed scripts for researchers who are inexperienced in the informatics field. Sharing scripts will be helpful for them. In decades, deep learning has come under the pharmaceutical research [26]. We selected Python as the programing language and used it in the Jupyter notebook [27] for this collaboration because of the ease of performing deep learning using these tools. Furthermore, Google Colaboratory [28] was employed as the platform for Jupyter
notebook in order to create an open viewing environment and allow scripts to be modified freely by all collaborators (Figure 5, 6). The features of Google Colaboratory are as follows.

Figure 5. Top page of Google Colaboratory.
Prepared data which included the solubility values and calculated descriptors of compounds were imported to Google Colaboratory.

Figure 6. Prediction Result for Solubility Values.
The statistic scores of cross validation and external test set were shown on the top of scatter plot. In the scatter plot, x-axis and y-axis show observed and predicted values, respectively.
Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely on the cloud. With Colaboratory it is possible to write and execute code, save and share analyses, and access powerful computing resources, all for free from a browser. The scripts are shared among collaborators using browsers because source codes and the results are stocked in Google Drive. We considered this interface was proper for our collaboration and decided to select this.

There were four participants in the script channel. All of them were academic researchers. Three of them specialized in bioinformatics, and one in molecular simulation and cheminformatics.

Threads were created for four important topics. The list of these threads are as follows:

i. Programming languages and interfaces
ii. How to share the csv data used in this collaboration?
iii. How to share the datasets with the dataset channel
iv. How to share the requests and draft with the model validation channel

At first, to determine the proper formats of scripts which could be shared for collaborators, we discussed the programming languages and interfaces. Rstudio or R AnalyticFlow was used for R users, and Jupyter Notebook or Spyder was selected for Python users. Additionally, we discussed analytical tools such as workflows using KNIME. The second was about the way to store the csv files used for collaborations. In this collaboration, the csv file which was used for analysis was uploaded to the script channel and shared because the file size was not so large (16.8 mega bytes).

Third, we had aimed to describe the chemical spaces of collected compounds. This idea was taken from contributors on the dataset channel. Thus, we visualized the chemical space of collected data using the two-dimensional scatter plot of molecular weight and logP. Furthermore, we obtained several suggestions as follows: using principal component analysis to find outliers (Figure 7), and coloring the compounds using the numbers satisfying Lipinski’s rule of five [29].

**Figure 7.** A plot image of PCA. All descriptors of compounds were used for PCA. Colored circles represented each compound. The values of solubility were colored from blue (low solubility) to red (high solubility).

Finally, two objectives were laid out for the model validation channel. One is “In the constructed classification models, accuracy and Cohen’s kappa were used for validating the performance of the models. The accuracy scores in each label class should be used to express the results for medicinal chemists” and “When making use of the three-class system (low, moderate,
and high) for solubility, it is important to avoid predicting that low solubility compounds will wind up in the high category, or high solubility compounds into low category.” In the early stages of drug discovery, it is acceptable to regard low or high solubility compounds as moderate ones. This means that it is necessary to retain ambiguous compounds in the moderate category that are difficult to clearly classify into the low or high categories. Thus, we have started to construct a script to distinguish three categories of solubility degree in order to meet the requirements.

In the script channel, we incorporated requests from other channels and discussed how their requests might be realized. These requests collected from several various viewpoints are important and we have considered that these are hurdles to be overcome for really a usable prediction model in drug discovery.

4.4 Future Work

To lead collaborations that develop into successful projects, two aims are important as follows: 1) Development of models that can be participated in by various researchers (for informatics researchers), and 2) Proposal of tools that can be used for drug discovery (for experimental researchers). Our model must continue developing and improving in order to become a useful tool for experimental researchers. Our script constructed in this collaboration is one such tool that is freely available via the Slacks. Using this tool, various researchers can access and reproduce our models. It would be very helpful to collect more feedbacks that could improve the performance of this script.

5. Conclusion

Using Slack as a collaborative framework for research, we created three channels (Dataset, Model Validation, and Script), assigned our collaborators appropriately to each of them, and then started a discussion about each topic. In our collaboration, we discovered several important perspectives revealed by the collaborators interacting in the three channels. Furthermore, we implemented the feedback into a Python script by reflecting the requests and comments and finally revealed new tasks for future improvement. The collected requests involved hints that were necessary for building our model into a more usable one for researchers in the early phase of drug discovery. For example, one such innovation involved visualizing the chemical spaces of collected compounds colored using the numbers satisfying Lipinski’s rule of five. Another use of this feedback is avoiding misprediction of solubility to an extreme category, such as placing low-solubility compounds into a high-solubility category, and vice versa. Our model must be further developed and improved to realize a useful tool for experimental researchers.

From this study, we have reaffirmed that there is, as of now, no environment to sustain collaborations between informatics and pharmaceutical researchers. Statistical or technical methods developed in the field of informatics have not been effectively used for drug discovery. Therefore, new discoveries on potential seeds, drug targets, or biomarkers are not being accelerated because informatics researchers are not aware of the requirements of pharmaceutical researchers. Effective drug discovery requires the cooperation and preparation of an environment for continuous collaborations. Therefore, our proposal is probably one of OICDD's solutions. To solve and continue the OICDD's issues, we first need to prepare an environment for easy collaboration. It is also important to continue to fund research, organize seminars, and plan contests to establish open an innovation and collaboration.
This study is a beginning of a new era of collaboration and shows one of the means by which constructed frames of collaboration will be helpful for various projects, in finding new pharmaceutical products and, accelerating drug discovery.

Acknowledgements

We would like to thank all attendees of the CBI 2019 luncheon seminar and thank Yukio Tada and Akihiko Konagaya for their advice on this research. We would like to thank the editor and the anonymous reviewers for comments on earlier version of this paper.

References

[1] AI (wikipedia), https://en.wikipedia.org/wiki/Artificial_intelligence (accessed Nov 1, 2019)
[2] Data Science (wikipedia), https://en.wikipedia.org/wiki/Data_science (accessed Nov 1, 2019)
[3] CBI Annual Meeting 2019, https://cbi-society.org/taikai/taikai19/index_e.html (accessed Nov 1, 2019)
[4] The Chem-Bio Informatics Society (CBI), https://cbi-society.org/home/english/about_e.html (accessed Nov 1, 2019)
[5] Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; et al. Mastering the game of Go with deep neural networks and tree search., *Nature*. 2016, 529, 484–489. doi: 10.1038/nature16961
[6] Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; et al. Mastering the game of Go without human knowledge., *Nature*. 2017, 550, 354–359. doi: 10.1038/nature24270
[7] Kaggel, https://www.kaggle.com/competitions (accessed Nov 1, 2019)
[8] Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; et al. Protein structure prediction using multiple deep neural networks in CASP13, *Proteins*. 2019, 87, 1141–1148. doi: 10.1002/prot.25834
[9] Reichman, M.; Simpson, P. B. Open innovation in early drug discovery: roadmaps and roadblocks, *Drug Discov Today*. 2016, 21, 779–788. doi: 10.1016/j.drudis.2015.12.008
[10] Chiba, S.; Ishida, T.; Ikeda, K.; Mochizuki, M.; Teramoto, R.; et al. An iterative compound screening contest method for identifying target protein inhibitors using the tyrosine-protein kinase Yes., *Sci Rep*. 2017, 7(1), 12038. doi:10.1038/s41598-017-10275-4
[11] Almagro, J. C.; Teplyakov, A.; Luo, J.; Sweet, R. W.; Kodangattil, S.; et al. Second antibody modeling assessment (AMA-II), *Proteins*. 2014, 82, 1553–62. doi:10.1002/prot.24567.
[12] Maria, A. G.; Timothy, G.; Abdelaziz, M.; Alan, D. P.; Marta, P. N.; et al. Palkowitz, Marta Paineiro-Nunez, and Ian Watson, Open Innovation Drug Discovery (OIDD): A Potential Path to Novel Therapeutic Chemical Space, *Curr Top Med Chem*. 2014, 14, 294–303. doi: 10.2174/1568026613666131127125858
[13] CBI Wakate, https://wakate.cbi-society.info/wakate/ (accessed Nov 1, 2019)
[14] Wakate Lecture, https://cbi-society.org/home/documents/society_journal/CBI_2019_7_1.pdf. (accessed Nov 1, 2019)
[15] Slack, https://slack.com/intl/ja-jp/about. (accessed Nov 1, 2019)
[16] Wakate Luncheon Seminar, https://cbi-society.org/taikai/taikai19/LS/LS-09_Wakate.pdf. (accessed Nov 1, 2019)
[17] Kumparak, Greg (February 5, 2015). "Slack's Co-Founders Take Home The Crunchie For
 Founder Of The Year". TechCrunch.com. Retrieved October 13, 2018., https://techcrunch.com/2015/02/05/slacks-co-founders-take-home-the-crunchie-for-founder-of-the-year/ (accessed Nov 1, 2019)

[18] The Konstanz Information Miner (KNIME), https://www.knime.com/ (accessed Nov 1, 2019)

[19] Wang, J.; Hou, T.; Xu, X. Aqueous solubility prediction based on weighted atom type counts and solvent accessible surface areas. *J. Chem. Inf. Model.* **2009**, *49*, 571–581. doi: 10.1021/ci800406y

[20] Dassault Systèmes BIOVIA, *Pipeline Pilot*, Release 2018, San Diego: Dassault Systèmes, 2018.

[21] RDKit, Open-source cheminformatics; http://www.rdkit.org (accessed Nov 1, 2019)

[22] Hawkins, P. C. D. Conformation Generation: The State of the Art., *J. Chem. Inf. Model.* **2017**, *57*(8), 1747–1756. doi: 10.1021/acs.jcim.7b00221

[23] Google Drive, https://www.google.com/drive/ (accessed Nov 1, 2019)

[24] Dropbox, https://www.dropbox.com/about (accessed Nov 1, 2019)

[25] Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **2017**, *45*(D1), D945–D954. doi:10.1093/nar/gkw1074

[26] Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery., *Drug Discov Today*. **2018**, *23*, 1241–1250. doi: 10.1016/j.drudis.2018.01.039

[27] Jupyter Notebook, https://jupyter.org/about (accessed Nov 1, 2019)

[28] Google Colaboratory, https://colab.research.google.com/ (accessed Nov 1, 2019)

[29] Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3–25. doi: 10.1016/S0169-409X(96)00423-1