Communication

www.molinf.com

molecular informatics

SPECIAL ISSUE

DOI: 10.1002/minf.201600073

BIGCHEM: Challenges and Opportunities for Big Data Analysis in Chemistry

Igor V. Tetko,*[a, b] Ola Engkvist,[c] Uwe Koch,[d] Jean-Louis Reymond,[e] and Hongming Chen[c]

Abstract: The increasing volume of biomedical data in chemistry and life sciences requires the development of new methods and approaches for their handling. Here, we briefly discuss some challenges and opportunities of this fast growing area of research with a focus on those to be addressed within the BIGCHEM project. The article starts with a brief description of some available resources for "Big Data" in chemistry and a discussion of the importance of data quality. We then discuss challenges with visualization of millions of compounds by combining chemical and biological data, the expectations from mining the "Big Data" using advanced machine-learning methods, and their applications in polypharmacology prediction and target de-convolution in phenotypic screening. We show that the efficient exploration of billions of molecules requires the development of smart strategies. We also address the issue of secure information sharing without disclosing chemical structures, which is critical to enable bi-party or multi-party data sharing. Data sharing is important in the context of the recent trend of "open innovation" in pharmaceutical industry, which has led to not only more information sharing among academics and pharma industries but also the so-called "precompetitive" collaboration between pharma companies. At the end we highlight the importance of education in "Big Data" for further progress of this area.

Introduction

The Wikipedia definition "Big Data"[1] is a term for data sets that are so large or complex that traditional data processing applications are inadequate also highlights that not only the size but also the data complexity is very important. In pharmaceutical research area, "Big Data" is emerging from the rapidly growing genomics data thanks to the rapid development of gene sequencing technology. Likewise people start to ask the question if there is Big Data in chemistry?

Over the past decade there actually has been a remarkable increase in the amount of available compound activity and biomedical data.[2–4] The definition of "Big Data" in chemistry is generally not clear. Frequently, the "Big Data" in chemistry refers to considerably larger databases than commonly used ones (in orders of magnitude),[5] which become recently available thanks to the emerging of new experimental techniques such as high throughput screening, parallel synthesis etc.,[3,6] or as access to new chemical information as result of automatic data mining (e.g., patents, literature or in house data collections).[7,8] How to efficiently mining the large scale of data in chemistry becomes an important problem for the future development of the chemical industry including pharmaceutical, agrochemical, biotechnological, fragrances, and general chemical companies.

Big Data collected from literature usually are quite noisy and the reasons are multiple. First of all, this could be due to the biological assay itself, for example the original experiment errors, assay artifacts in screening etc. Secondly, there lack of a standard way for annotating biological endpoints, mode of action and target identifier. Thirdly, errors exist when extracting data values, units and/or chemical name recognition for automatic literature mining. Some actions have been done to address these problems, e.g., im-

[a] *I. V. Tetko*
Helmholtz Zentrum München – German Research Center for Environmental Health (GmbH), Institute of Structural Biology, Ingolstädter Landstraße 1, b. 60w, D-85764, Neuherberg, Germany
Phone: +49 89 3187 3575
Fax: +49 89 3187 3585
**e-mail: itetko@vcclab.org*

[b] *I. V. Tetko*
BIGCHEM GmbH, Ingolstädter Landstraße 1, b. 60w, D-85764, Neuherberg, Germany

[c] *O. Engkvist, H. Chen*
Discovery Sciences, AstraZeneca R&D Gothenburg, Pepparedsleden 1, Mölndal, SE-43183, Sweden

[d] *U. Koch*
Lead Discovery Center GmbH, Otto-Hahn Strasse 15, Dortmund, 44227, Germany

[e] *J.-L. Reymond*
Department of Chemistry and Biochemistry, University of Bern, Freiestrasse 3, 3012, Bern, Switzerland

proving data quality by applying promiscuity filters to clean up the screening data, developing bioassay ontology (BAO) tools to better organize and/or standardize the collected data etc.[9–11]

In order to support data analysis, collection, sharing, and dissemination several large projects, such as ELIXIR (https://www.elixir-europe.org), eTox (http://www.etoxproject.eu), BIGCHEM (http://bigchem.eu) and others, were initiated under the sponsorship of European Commission. Many of these activities are within the core area of chemoinformatics. The BIGCHEM project was recently sponsored by EU Horizon2020 program. The consortium includes academia, big pharma companies, large academic societies (Helmholtz, Fraunhofer) and Small and Medium Enterprises (SMEs). This project mainly aims to develop computational methods specifically for Big Data analysis. Below, we briefly review challenges and opportunities in the Big Data analytics area particularly focusing on several aspects, which are going to be addressed in BIGCHEM project.

Data Repositories

Publicly available databases such as PubChem,[3] BindingDB,[6] and ChEMBL[4] (Table 1) represent examples of large public domain repositories of compound activity data. ChEMBL and BindingDB contain manually extracted data from tens of thousands of articles. PubChem was originally started as a central repository of High Throughput (HTS) screening experiments for the National Institute of Health's (USA) Molecular Libraries Program but also incorporates data from other repositories (e.g., ChEMBL and BindingDB). Commercial databases, such as SciFinder, GOSTAR and Reaxys (Table 1) have accumulated a large amount of data collected from publications and patent data. Similarly to public and commercially available repositories, industry has produced large private collections. For example, more than 150M data points are available as part of AstraZeneca International Bioscience Information System (AZ IBIS) just for experiments performed before 2008.[8] The data quality in databases can significantly vary depending on data source, data acquisition procedures and curation efforts. Accumulated chemical patents represent another rich resource for chemical information. Large-scale text mining has been done on patent corpus to extract useful information. IBM has contributed chemical structures from pre-2000 patents in PubChem.[12] SureChEMBL database[13] was launched in 2014 providing the wealth of knowledge hidden in patent documents and currently contains 17 million compounds extracted from 14 million patent documents.

Under the enormous pressure of developing new drug with more restrained R&D budget, recent years have seen large pharma companies increasingly exploring the so called "open innovation" model for drug discovery research. The collaboration between academics and pharmaceutical industry in terms of compound, data sharing has been largely increased.[18] The examples include AstraZeneca-Sanger Drug Combination prediction challenge to develop better algorithms for treatment of cancer.[19] European Lead Factory[20] is another collaboration effort of seven pharma companies, SMEs and academic partners to create a diverse library of 500k compounds by combining compounds from partners, external users and newly synthesized libraries and to screen these libraries against commercial and public targets. Both academia and industry should benefit from these kind of collaborative efforts, which can result in more chemical and biological data being available in the public domain. More interestingly, even the collaborations between pharmaceutical companies on the so-called "precompetitive" level, which was hardly to imagine ten years ago, has become a trend. These efforts have made sharing of data within each organization become possible and lead to a further increase in the size of "Big Data".[21,22]

Frequent Hitters Analysis

Big Data sometimes also means noisy data. Data coming from HTS experiments could often be contaminated with false positive and false negative results. The errors can appear due to casual problems such as measurements errors, robotic failure, temperature differences, etc., which could be easily addressed with proper experimental protocols (e.g. by repeated measurements). Unfortunately, there are also systematic problems, such as low solubility in water, degradation of compounds or "frequent hitters"

Table 1. Data repositories

Database	Unique compounds	Experimental		Main data types
		facts	data types	
ChEMBL v. 21[4]	1,592,191	13,968,617	1,212,831	PubChem HTS assays and data mined from literature
BindingDB[6]	529,618	1,207,821	6,265	Experimental protein-small molecule interaction data
PubChem[3]	>60M	>157M	>1M	Bioactivity data from HTS assays
Reaxys[14]	>74M	>500M	–	Literature mined property, activity and reaction data
SciFinder (CAS)[15]	>111M	>80M	–	Experimental properties, ^{13}C and ^{1}H NMR spectra, reaction data
GOSTAR[16]	>3M	>24M	>5k	Target-linked data from patents and articles
AZ IBIS[8]	–	>150M	–	AZ *in-house* SAR data points
OCHEM[17]	>600k	>1.2M	>400	Mainly ADMET data collected from literature

(FH). The first two problems can be solved with, e.g. proper analytical procedures, while the latter requires another type of consideration.

FHs are usually referred to as compounds that provide unspecific activity in different assays.[23] Some of these compounds cause nonspecific binding (*e.g.*, reactive compounds) or/and interfere with a particular assay technology (*e.g.*, light quenching, compounds forming micelles, luciferase inhibitors, formation of complexes with tagged proteins for AlphaScreen[24]). Others are promiscuous binders that interact with different targets in a specific, dose-dependent fashion[25] and could constitute up to 99.8% of hits.[26] An analysis of results of these HTS data without filtering nonspecific binders and compounds that interfere with the assay technology could result in a model to predict FHs and not the target activity. Therefore carrying out FH analysis would help to clean screening data and eventually help to build better predictive model.

Various sources for compounds behaving as FH have been proposed such as: chelation, redox activity, membrane disruption, singlet oxygen production, compound fluorescence, cysteine oxidation and non-selective reactivity.[26] It was also estimated that 1–2% of drug-like compounds could self-associate into colloidal aggregates that non-specifically inhibit enzymes and other proteins at a typical screening concentration of 5 uM.[27] Baell and Holloway looked at compounds with activity in multiple assays[11] and found certain substructures, which appeared repeatedly in promiscuous hits, and labeled them "Pan Assay Interference Compounds (PAINS)".

The non-specific binding, however, can be also important and be extensively exploited by nature. A significant overlap between PAINS substructures and natural products for quinones and catechols[28] indicate that these scaffolds were selected by evolution for their shotgun properties. Other substructures such as 2-amino thiazoles have been shown to be FHs in the sense of promiscuous binders, but are also present in marketed drugs.[29] An application of alerts developed by chemical providers to flag problematic compounds found that drugs are two-three folds enriched with such alerts as compared to the screening libraries.[30] Thus, a blind exclusion of "undesired" compounds may result in a significant risk to miss potentially interesting compounds and thus throw the baby out with the bathwater.

In order to develop FH filters it is important to find assays, which use similar technology. To be able to better analyze and compare different HTS data the Bio-Assay Ontology (BAO) concept was developed.[10,31] It has been used to both annotate HTS in PubChem and in an industrial setting.[9] The use of BAO makes it easier to group assays according to the used technologies and to identify relevant FHs. The identified catalogue of FH substructures can be very useful to remove chemical matter in future HTS campaigns that will specifically interfere with the used assay technology. Thus, the information about the mechanism of action of FHs will be important to design and correctly interpret screening campaigns.[24,32]

It should be noted that even the best BAO and best methods of standardization of experiments will never fully address the problem of heterogeneity and complexity of biological and chemical data. By no means experiments performed in mouse and in rats can be combined into one single "activity" column associated to a single, standardized structure for all possible experiments. Such merge can be performed only depending on the conditions of experiments, endpoint and, importantly, properties of compounds. For example, even for simple lipophilicity property logP shake-flask values can be merged with logD values obtained from HPLC experiments only for compounds non ionized under the pH of the experiment but not for all possible structures.

Data Visualization and Exploration of Chemical Space

The visualization and compact representation of millions of compounds (such as >110M compounds in SciFinder, see Table 1), which is usually the first step of data analysis, represents significant challenge in Big Data analysis. It is usually done by projecting large compound collections into a low dimensional space, amenable to visual inspection and intuitive analysis by the human brain. It could help to detect chemical entities with novel chemical scaffolds and physicochemical properties (*e.g.*, for compound library design), to compare different libraries or to identify regions of chemical space that possess certain pharmacological profile.[33] Exemplary approaches such as principle component analysis (PCA),[34] Generative Topographic Mapping (GTM),[35] Kohonen networks,[36] Diffusion Maps,[37] and interactive maps obtained by projection of high-dimensional descriptor spaces,[38,39] are promising techniques in this context. Such visualization methods can be also used to interpret structure-activity relationships.[40] For example, in the "Stargate" version of GTM, latent space links two different initial spaces – one defined by molecular descriptor and another one by experimental activities.[41] This allows, on one hand, to predict the whole pharmacological profile for one particular chemical structure and, on the other hand, to identify new structures corresponding to the given profile. Another example of exploring chemical space is using a so called ChemGPS approach to represent and navigate through drug-like[42] and pharmacokinetic[43] chemical space based on PCA components extracted from molecular 2D descriptors. Its variant ChemGPS-NP[44] characterize the natural product space in particular. It has been shown that the accuracy of describing molecules in ChemGPS-NP defined space is similar to the accuracy of structural fingerprints in retrieving bioactive molecules.[45]

Beside the space represented by known and available chemical structures, the chemical space composed by virtual compounds is much bigger. The number of potential

molecular structures, which could theoretically be enumerated, is vast. For example, the database GDB-17 contains 166.4 billion molecules that are possible combinations of up to 17 atoms of C, N, O, S and halogens following simple rules of chemical stability and synthetic feasibility.[46,47] Although GDB-17 is already very large, it would be many orders of magnitude larger if extended to 20–30 heavy atoms, which is the average size of drug-like molecules.[48] These data sets raise new challenges even for traditional profiling of chemical compound collections, which is used to identify chemicals with favorable properties (e.g., Lipinski's rule, non-toxic, etc.). Even a fast algorithm, which is able to process 100,000 molecules per minute, will require >1,000 days ~ 3 years of calculations on one core to annotate the full GDB-17. If the model supports efficient parallelization, it could be executed on, e.g., the supercomputer of Leibniz Supercomputing Centre with 241,000 cores. In this case the calculation time can be theoretically decreased to ten minutes. Instead of a brute force approach one can rely on, e.g. sequential triaging scheme that eliminates undesired regions, such as low solubility or low prediction accuracy due to limited applicability domain of model,[49] by very fast algorithms first and then applies more compute expensive methods on smaller subspaces. Thus novel approaches or workflows are needed to efficiently search through this enormous chemical space.

Structure-Activity Relationship Modeling

Although a plethora of machine learning algorithms is available for SAR studies[50,51] there is an increasing need for robust and efficient computational methods, which are able to cope with very large and heterogeneous datasets. The current methods already allow to build predictive models from hundreds of thousands of compounds and high-dimensional descriptors with data matrices of >0.2 trillion entries.[5] Advances in this field can be also expected from data fusion methods, which simultaneously model several related properties.[52] The simultaneous modeling of such incompatible data by exploring inter-correlation between different properties, e.g. tissue/air partitioning coefficients in humans and in rats, has already successfully contributed models with improved accuracy compared to those built with any single activity data.[52] Numerous methods have been developed to predict compound polypharmacology.[53–55] Prediction of the binding affinity of ligands to multiple proteins allows to anticipate potential selectivity issues, discover beneficial multi-target activities as early as possible in the drug discovery process,[56] or make target deconvolution for phenotypic screening.[57] Most of these methods rely on building single target model individually, one future development could be to use all available chemogenomics data to pursue multi-task learning and build one multi-label model to predict multiple target activity simultaneously. A recent study shows that massive multitask networks obtain predictive accuracies significantly better than single-task methods.[58] Probabilistic matrix factorization (PMF) methods have been found particularly useful in building multi-task model.[59,60] Further injection of ligand and protein information into PMF method as side information may further improve the prediction accuracy.[61] However in Big Data setting, this would require huge computer power and dedicated parallel programming model. Recently deep learning technology has gained large attention in public media. In 2015 the deep learning models achieved accuracy of human brain for handwritten Chinese character recognition[62], while in 2016 a deep-learning network won Go[63] tournament against the human champion. Moreover, recent announcement of Google Cloud Platform[64] has made possible the use of technologies staying behind the best implementation of machine learning algorithms by non-experts. The deep learning neural network technology, which is able to efficiently deal with high-dimensional and complex data, has also been applied in chemoinformatics area[51,65,66] and is expected to further contribute to the progress of this direction of studies.

Another important question is "does more data contribute better models"? A consensus model to predict melting point (MP), which was developed with N = 275k measurements, calculated RMSE = 31 ± 1 °C for Bergström data set[67] of drugs (N = 277).[5] This result is almost 15 °C improvement compared to the results of the original study[67] and 3 °C improvement compared to the model developed with N = 47k molecules.[68] It should be noticed that models were developed using different descriptors and machine learning methods, which could contribute to the difference in their performances. To exclude influence of these factors, we used exactly the same protocols from ref[5] to develop a model using Bergström data only. The developed consensus model calculated RMSE = 50 ± 1 °C confirming, on one hand, that the improvement in the prediction accuracy for MP was contributed by an increase of the training set size and, on the other hand, suggesting that modern automatic text patent mining, which was used to contribute >80 % of data of the 275k set, produced data of excellent quality.

De Novo Design

De novo design aims at generating new chemical entities with drug-like properties and desired biological activities in a directed fashion. Comparing with normal virtual screening or HTS, which search for active compounds in physically available compound database, de novo design tries to generate hypothetical candidate compounds *in silico*. There are mainly two type of methods for making de Novo molecular design, one is the based on the similarity to known active compounds, i.e., ligand based De Novo design, and the other type is based on protein 3D structure to generate new compounds, i.e., structure based De Novo design. Here we mainly discuss ligand based methods, structure based methods can be found in elsewhere.[69]

One way for doing de Novo design is to search the large virtual compound database such as the GDB to get de novo hits. In order to search vast virtual chemical space, one would need integrated workflows combining efficient search and multiparameter optimization strategies to filter out molecules with sub-optimal profiles as early as possible. For example physicochemical and synthetic feasibility filters can be frontloaded to trim down the number of compounds. Ruddigkeit et al.[70] was able to search entire GDB17 database with a workflow combining MQN 2D structure fingerprint with ROCS shape matching method. Another strategy is reaction-driven, fragment-based de novo design. Based on known chemical reactions and commercially available build blocks, chemically diverse and synthetically feasible compounds are generated via normally multistep and multi-parameter optimization process searching for candidate compounds which satisfy to certain property profile. These reaction-based methods have been successfully applied to design de novo bioactive compounds.[71–73]

The third strategy to provide an intelligent search of new compounds is to generate structures, which are sufficiently new but still within the chemical spaces covered by the models. A group of these methods, which is known as "inverse QSAR", has received a boost during recent year due to increasing computational power and new theoretical developments. A set of linear constrained Diophantine equations was used by Faulon et al.[74] to exhaustively enumerate new compounds. Wong et al.[75] used kernel methods to map training compounds from input space to the kernel feature space. In this space the authors generated new data points, which were used to recover the chemical structures. Actually, this approach is similar to that of aforementioned "Stargate" GTM,[41] with an exception that the former algorithm does not use supervised learning to create maps. In another approach Funatsu et al.[76] used Gaussian models and Bayesian inference to exhaustively fill a target region of the model space with new structures. Thus, these methods propose novel chemical structures while still staying within the chemical space of the QSAR models.

Data Sharing and Data Security

Even large pharma companies can accumulate only limited amounts of relevant property information. As it was mentioned before, sharing data collected by different organizations offers the opportunity to develop computational models on a much broader data basis, thereby increasing model robustness, accuracy and coverage of chemical space.[77,78] The development of approaches to predict ADME/T properties in a collaborative manner is becoming a part of future pharma R&D strategies. Recently, AstraZeneca and Bayer made the efforts to compare their entire compound collection in a secure manner,[22] while AstraZeneca and Roche started a data sharing consortium on the topic of matched molecular pairs to improve metabolism, pharmacokinetics, and safety of their compounds through MedChemica.[21] Moreover, AstraZeneca has already donated some of its ADMETox data to CheMBL.[79] However, collaborative efforts in this field are generally not straightforward. The intellectual property aspects associated with private compound collections and associated data might be relevant for ongoing drug discovery efforts. Secure multiparty computation methods based on modern encryption theory[80,81] provide ways to develop models without the need to share molecular structures or proprietary molecular representations. These methods are compute-intense and bandwidth-demanding but fast development of Internet and increasing computational power of computers is making them applicable to real world problems.[82,83]

Training Big Data Scientists – The Chemoinformaticians

The "Big Data" challenges require professionally trained experts, "data scientists in chemistry" – the chemoinformaticians, who can cope with the complexity and diversity of problems in this field of scientific discovery. Traditional "data scientists" coming from computer science field, as well as computational chemists with little knowledge in computer science, are very unlikely to have sufficient knowledge and expertise to address both chemoinformatics questions and will need additional training. Important questions in this regard are following ones: How should one balance chemistry and computer science training? How should one ensure a high level of scientific expertise and, at the same time, a practically oriented mindset? Which new and rapidly developing methodologies should be considered? How should one prepare trainees to work at the interface between computing, chemistry, and pharmaceutical research? These questions can be answered only during close interactions of academic partners and the end-users and tight involvement of industrial partners in targeted research trainings. In this respect the training programs, such as offered through Marie Skłodowska-Curie Actions, provide generous funding support by means of Innovative Training Networks, which foster and promote such type of interactions.

Conclusions

Both industry and academic partners share high expectations from "Big Data" in chemistry, which is a new emerging area of research on the borders of several disciplines. The advance in this area requires development of new computational approaches and more importantly education of scientists, who will further progress this field.

Conflict of Interest

IVT is CEO and founder of BigChem GmbH, which licenses the OCHEM [17].

Acknowledgements

The project leading to this article has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 676434, "Big Data in Chemistry" ("BIGCHEM", http://bigchem.eu). The article reflects only the authors' view and neither the European Commission nor the Research Executive Agency are responsible for any use that may be made of the information it contains. The authors thank BIGCHEM partners for their comments and suggestions, which were important to improve this manuscript. IVT is CEO and founder of BigChem GmbH, which licenses the OCHEM.[17] The other authors declared that they have no actual or potential conflicts of interests.

References

[1] Big Data. https://en.wikipedia.org/wiki/Big_data (10 June 2016).
[2] B. Chen, A. J. Butte, *Clin. Pharmacol. Ther.* **2016**, *99*, 285 – 297.
[3] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, S. H. Bryant, *Nucleic Acids Res.* **2016**, *44*, D1202 – 1213.
[4] G. Papadatos, A. Gaulton, A. Hersey, J. P. Overington, *J. Comput. Aided. Mol. Des.* **2015**, *29*, 885 – 896.
[5] I. V. Tetko, D. Lowe, A. J. Williams, *J. Cheminform.* **2016**, *8*, 2.
[6] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, J. Chong, *Nucleic Acids Res.* **2016**, *44*, D1045 – 1053.
[7] N. Schneider, D. M. Lowe, R. A. Sayle, M. A. Tarselli, G. A. Landrum, *J. Med. Chem.* **2016**, *59*, 4385 – 4402.
[8] S. Muresan, P. Petrov, C. Southan, M. J. Kjellberg, T. Kogej, C. Tyrchan, P. Varkonyi, P. H. Xie, *Drug Discov. Today* **2011**, *16*, 1019 – 1030.
[9] L. Zander Balderud, D. Murray, N. Larsson, U. Vempati, S. C. Schurer, M. Bjareland, O. Engkvist, *J. Biomol. Screen.* **2015**, *20*, 402 – 415.
[10] S. Abeyruwan, U. D. Vempati, H. Kucuk-McGinty, U. Visser, A. Koleti, A. Mir, K. Sakurai, C. Chung, J. A. Bittker, P. A. Clemons, S. Brudz, A. Siripala, A. J. Morales, M. Romacker, D. Twomey, S. Bureeva, V. Lemmon, S. C. Schurer, *J. Biomed. Semantics.* **2014**, *5*, S5.
[11] J. B. Baell, G. A. Holloway, *J. Med. Chem.* **2010**, *53*, 2719 – 2740.
[12] IBM Contributes Data to the National Institutes of Health to Speed Drug Discovery and Cancer Research Innovation. http://www.prnewswire.com/news-releases/ibm-contributes-data-to-the-national-institutes-of-health-to-speed-drug-discovery-and-cancer-research-innovation-135275888.html (10 June 2016).
[13] G. Papadatos, M. Davies, N. Dedman, J. Chambers, A. Gaulton, J. Siddle, R. Koks, S. A. Irvine, J. Pettersson, N. Goncharoff, A. Hersey, J. P. Overington, *Nucleic Acids Res.* **2016**, *44*, D1220 – 1228.
[14] Chemical Data – Reaxys. http://www.elsevier.com/solutions/reaxys (10 June 2016).
[15] SciFinder – A CAS solution. http://www.cas.org/products/scifinder (10 June 2016).
[16] GOSTAR – GVK BIO Online Structure Activity Relationship Database. http://www.gostardb.com (10 June 2016).
[17] I. Sushko, S. Novotarskyi, R. Korner, A. K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V. V. Prokopenko, V. Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, I. I. Baskin, V. A. Palyulin, E. V. Radchenko, W. J. Welsh, V. Kholodovych, D. Chekmarev, A. Cherkasov, J. Aires-de-Sousa, Q. Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko, I. V. Tetko, *J. Comput. Aided. Mol. Des.* **2011**, *25*, 533 – 554.
[18] Open Innovation Case Study: Pfizer's Centers for Therapeutic Innovation. http://www.bioendeavor.net/CommonData/NewsFiles/Pfizer.pdf (6 May 2016).
[19] AstraZeneca-Sanger Drug Combination Prediction DREAM Challenge. https://www.synapse.org/ – !Synapse:syn4231880/wiki/235645 (10 June 2016).
[20] European Lead Factory. http://www.europeanleadfactory.eu (10 June 2016).
[21] Roche and AstraZeneca launch medicinal chemistry data-sharing consortium to further accelerate drug discovery. http://www.roche-nutley.com/home/press-releases/june-26–2013.html (10 June 2016).
[22] T. Kogej, N. Blomberg, P. J. Greasley, S. Mundt, M. J. Vainio, J. Schamberger, G. Schmidt, J. Huser, *Drug Discov. Today* **2013**, *18*, 1014 – 1024.
[23] O. Roche, P. Schneider, J. Zuegge, W. Guba, M. Kansy, A. Alanine, K. Bleicher, F. Danel, E. M. Gutknecht, M. Rogers-Evans, W. Neidhart, H. Stalder, M. Dillon, E. Sjogren, N. Fotouhi, P. Gillespie, R. Goodnow, W. Harris, P. Jones, M. Taniguchi, S. Tsujii, W. von der Saal, G. Zimmermann, G. Schneider, *J. Med. Chem.* **2002**, *45*, 137 – 142.
[24] K. Schorpp, I. Rothenaigner, E. Salmina, J. Reinshagen, T. Low, J. K. Brenke, J. Gopalakrishnan, I. V. Tetko, S. Gul, K. Hadian, *J. Biomol. Screen.* **2014**, *19*, 715 – 726.
[25] P. Schneider, M. Rothlisberger, D. Reker, G. Schneider, *Chem. Commun.* **2016**, *52*, 1135 – 1138.
[26] J. L. Dahlin, J. W. Nissink, J. M. Strasser, S. Francis, L. Higgins, H. Zhou, Z. Zhang, M. A. Walters, *J. Med. Chem.* **2015**, *58*, 2091 – 2113.
[27] B. Y. Feng, B. K. Shoichet, *Nat. Protoc.* **2006**, *1*, 550 – 553.
[28] J. B. Baell, *J. Nat. Prod.* **2016**, *79*, 616 – 628.
[29] S. M. Devine, M. D. Mulcair, C. O. Debono, E. W. Leung, J. W. Nissink, S. S. Lim, I. R. Chandrashekaran, M. Vazirani, B. Mohanty, J. S. Simpson, J. B. Baell, P. J. Scammells, R. S. Norton, M. J. Scanlon, *J. Med. Chem.* **2015**, *58*, 1205 – 1214.
[30] I. Sushko, E. Salmina, V. A. Potemkin, G. Poda, I. V. Tetko, *J. Chem. Inf. Model.* **2012**, *52*, 2310 – 2316.
[31] BioAssayOntology. http://bioassayontology.org/ (10 June 2016).
[32] J. K. Brenke, E. S. Salmina, L. Ringelstetter, S. Dornauer, M. Kuzikov, I. Rothenaigner, K. Schorpp, F. Giehler, J. Gopalakrishnan, A. Kieser, S. Gul, I. V. Tetko, K. Hadian, *J. Biomol. Screen.* **2016**, *21*, 596–607.
[33] M. Reutlinger, G. Schneider, *J. Mol. Graph. Model.* **2012**, *34*, 108 – 117.
[34] M. Feher, J. M. Schmidt, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218 – 227.
[35] H. A. Gaspar, G. Marcou, D. Horvath, A. Arault, S. Lozano, P. Vayer, A. Varnek, *J. Chem. Inf. Model.* **2013**, *53*, 3318 – 3325.
[36] P. Schneider, K. Stutz, L. Kasper, S. Haller, M. Reutlinger, F. Reisen, T. Geppert, G. Schneider, *Pharmaceuticals* **2011**, *4*.

[37] N. V. Kireeva, S. I. Ovchinnikova, I. V. Tetko, A. M. Asiri, K. V. Balakin, A. Y. Tsivadze, *ChemMedChem* **2014**, *9*, 1047 – 1059.
[38] L. Ruddigkeit, M. Awale, J. L. Reymond, *J. Cheminform.* **2014**, *6*, 27.
[39] M. Awale, J. L. Reymond, *J. Chem. Inf. Model.* **2015**, *55*, 1509 – 1516.
[40] M. Reutlinger, W. Guba, R. E. Martin, A. I. Alanine, T. Hoffmann, A. Klenner, J. A. Hiss, P. Schneider, G. Schneider, *Angew. Chem. Int. Ed. Engl.* **2011**, *50*, 11633 – 11636.
[41] H. A. Gaspar, Baskin, Il, G. Marcou, D. Horvath, A. Varnek, *J. Chem. Inf. Model.* **2015**, *55*, 2403 – 2410.
[42] T. I. Oprea, J. Gottfries, *J. Comb. Chem.* **2001**, *3*, 157 – 166.
[43] T. I. Oprea, I. Zamora, A. L. Ungell, *J. Comb. Chem.* **2002**, *4*, 258 – 266.
[44] J. Larsson, J. Gottfries, S. Muresan, A. Backlund, *J. Nat. Prod.* **2007**, *70*, 789 – 794.
[45] R. Buonfiglio, O. Engkvist, P. Varkonyi, A. Henz, E. Vikeved, A. Backlund, T. Kogej, *J. Chem. Inf. Model.* **2015**, *55*, 2375 – 2390.
[46] J. L. Reymond, *Acc. Chem. Res.* **2015**, *48*, 722 – 730.
[47] L. Ruddigkeit, R. van Deursen, L. C. Blum, J. L. Reymond, *J. Chem. Inf. Model.* **2012**, *52*, 2864 – 2875.
[48] P. Ertl, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374 – 380.
[49] I. Sushko, S. Novotarskyi, R. Körner, A. K. Pandey, V. V. Kovalishyn, V. V. Prokopenko, I. V. Tetko, *J. Chemom.* **2010**, *24*, 202 – 208.
[50] A. Varnek, I. Baskin, *J. Chem. Inf. Model.* **2012**, *52*, 1413 – 1437.
[51] I. I. Baskin, D. A. Winkler, I. V. Tetko, *Expert Opin. Drug Discov.* **2016**, *11*, 785 – 795.
[52] A. Varnek, C. Gaudin, G. Marcou, I. Baskin, A. K. Pandey, I. V. Tetko, *J. Chem. Inf. Model.* **2009**, *49*, 133 – 144.
[53] A. Anighoro, J. Bajorath, G. Rastelli, *J. Med. Chem.* **2014**, *57*, 7874 – 7887.
[54] P. Schneider, G. Schneider, *J. Med. Chem.* **2016**, *59*, 4077 – 4086.
[55] L. H. Mervin, A. M. Afzal, G. Drakakis, R. Lewis, O. Engkvist, A. Bender, *J. Cheminform.* **2015**, *7*, 51.
[56] G. Rastelli, L. Pinzi, *Front. Pharmacol.* **2015**, *6*, 157.
[57] J. Lee, M. Bogyo, *Curr. Opin. Chem. Biol.* **2013**, *17*, 118 – 126.
[58] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, V. Pande, *ArXiv e-prints* **2015**, 1502.02072.
[59] M. Gonen, *Bioinformatics* **2012**, *28*, 2304 – 2310.
[60] J. C. Costello, L. M. Heiser, E. Georgii, M. Gonen, M. P. Menden, N. J. Wang, M. Bansal, M. Ammad-ud-din, P. Hintsanen, S. A. Khan, J. P. Mpindi, O. Kallioniemi, A. Honkela, T. Aittokallio, K. Wennerberg, N. D. Community, J. J. Collins, D. Gallahan, D. Singer, J. Saez-Rodriguez, S. Kaski, J. W. Gray, G. Stolovitzky, *Nat. Biotechnol.* **2014**, *32*, 1202 – 1212.
[61] J. Simm, A. Arany, P. Zakeri, T. Haber, J. K. Wegner, V. Chupakhin, H. Ceulemans, Y. Morea, *ArXiv e-prints* **2015**, 1509.04610.
[62] 96.7% recognition rate for handwritten Chinese characters using AI that mimics the human brain. http://phys.org/news/2015-09-recognition-handwritten-chinese-characters-ai.html (10 June 2016).
[63] AlphaGo seals 4-1 victory over Go grandmaster Lee Sedol. https://www.theguardian.com/technology/2016/mar/15/google-alphago-seals-4-1-victory-over-grandmaster-lee-sedol (10 June 2016).
[64] Google Cloud Platform. https://cloud.google.com/ (10 June 2016).
[65] E. Gawehn, J. A. Hiss, G. Schneider, *Mol. Inf.* **2016**, *35*, 3-14.
[66] A. Mayr, G. Klambauer, T. Unterthiner, S. Hochreiter, *Frontiers Environ. Sci.* **2016**, *3*.
[67] C. A. Bergstrom, U. Norinder, K. Luthman, P. Artursson, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1177 – 1185.
[68] I. V. Tetko, Y. Sushko, S. Novotarskyi, L. Patiny, I. Kondratov, A. E. Petrenko, L. Charochkina, A. M. Asiri, *J. Chem. Inf. Model.* **2014**, *54*, 3320 – 3329.
[69] G. Schneider, *De novo Molecular Design*, Wiley, **2013**.
[70] L. Ruddigkeit, L. C. Blum, J. L. Reymond, *J. Chem. Inf. Model.* **2013**, *53*, 56 – 65.
[71] H. M. Vinkers, M. R. de Jonge, F. F. Daeyaert, J. Heeres, L. M. Koymans, J. H. van Lenthe, P. J. Lewi, H. Timmerman, K. Van Aken, P. A. Janssen, *J. Med. Chem.* **2003**, *46*, 2765 – 2773.
[72] G. Schneider, T. Geppert, M. Hartenfeller, F. Reisen, A. Klenner, M. Reutlinger, V. Hahnke, J. A. Hiss, H. Zettl, S. Keppner, B. Spankuch, P. Schneider, *Future Med. Chem.* **2011**, *3*, 415 – 424.
[73] M. Hartenfeller, H. Zettl, M. Walter, M. Rupp, F. Reisen, E. Proschak, S. Weggen, H. Stark, G. Schneider, *PLoS Comput. Biol.* **2012**, *8*, e1002380.
[74] C. J. Churchwell, M. D. Rintoul, S. Martin, D. P. Visco, Jr., A. Kotu, R. S. Larson, L. O. Sillerud, D. C. Brown, J. L. Faulon, *J. Mol. Graph. Model.* **2004**, *22*, 263 – 273.
[75] W. W. Wong, F. J. Burkowski, *J. Cheminform.* **2009**, *1*, 4.
[76] T. Miyao, H. Kaneko, K. Funatsu, *J. Chem. Inf. Model.* **2016**, *56*, 286 – 299.
[77] R. Chaguturu, *Collaborative innovation in drug discovery: Strategies for public and private partnerships*, **2014**.
[78] I. V. Tetko, R. Abagyan, T. I. Oprea, *J. Comput. Aided. Mol. Des.* **2005**, *19*, 749 – 764.
[79] Experimental in vitro DMPK and physicochemical data on a set of publicly disclosed compounds. http://dx.doi.org/10.6019/CHEMBL3301361 (6 June 2016).
[80] A. F. Karr, J. Feng, X. Lin, A. P. Sanil, S. S. Young, J. P. Reiter, *J. Comput. Aided. Mol. Des.* **2005**, *19*, 739 – 747.
[81] R. Cramer, I. Damgaard, J. B. Nielsen, *Secure multiparty computation : an information-theoretic approach*, Cambridge University Press, Cambridge ; New York, **2015**.
[82] R. Bendlin, I. Damgård, C. Orlandi, S. Zakarias, in *Proceedings of the 30th Annual international conference on Theory and applications of cryptographic techniques: advances in cryptology*, Springer-Verlag, Tallinn, Estonia, **2011**, pp. 169 – 188.
[83] R. Agrawal, R. Srikant, *SIGMOD Record (ACM Special Interest Group on Management of Data)* **2000**, *29*, 439 – 450.

Received: May 19, 2016
Accepted: July 6, 2016
Published online: July 28, 2016

