Editorial



Future Medicinal Chemistry

# Does 'Big Data' exist in medicinal chemistry, and if so, how can it be harnessed?

"...further progress will critically depend on training programs and advances in chemoinformatics, a discipline bridging chemistry and informatics."

**First draft submitted:** 1 August 2016; **Accepted for publication:** 12 August 2016; **Published online:** 15 September 2016

**Keywords:** applicability domain • Big Data • chemoinformatics • education in chemistry and informatics • local and global models • multitask learning • neural networks • virtual chemical spaces

Igor V Tetko*[1,2], Ola Engkvist[3] & Hongming Chen[3]

[1]Helmholtz Zentrum München-German Research Center for Environmental Health (GmbH), Institute of Structural Biology, Ingolstädter Landstraße 1, b. 60w, D-85764 Neuherberg, Germany

[2]BIGCHEM GmbH, Ingolstädter Landstraße 1, b. 60w, D-85764 Neuherberg, Germany

[3]Discovery Sciences, AstraZeneca R&D Gothenburg, Pepparedsleden 1, Mölndal, SE-43183, Sweden

*Author for correspondence:
Tel.: +49 89 3187 3575
Fax: +49 89 3187 3585
itetko@vcclab.org

The term 'Big Data' has gained increasing popularity within the chemistry field and across science broadly in recent years [1]. Chemical databases have seen a dramatic growth over the past decade, with, for example, ChEMBL, REAXYS and PubChem providing hundreds of millions of experimental facts for tens of millions of compounds [1]. Moreover, even larger datasets of experimental measurements are held within in-house data collections at pharma companies [2]. Overall, the total number of entries across these databases is in the range of a billion, $10^9$; however, although this number may seem impressive, it pales into comparison relative to other fields [3], where the amount of data is frequently measured in exabytes, $10^{18}$. Thus, does Big Data really exist within the chemistry field? What are such data within medicinal chemistry specifically and where do the challenges lie in analysis of these data? Big Data refer to data out of the scale of traditional applications, which require efforts beyond the traditional analysis [1]. In this article, we will be discussing how it applies to medicinal chemistry, as well as providing an overview of some of the most important trends in the medicinal chemistry–Big Data field.

## Does Big Data exist in medicinal chemistry?

A dataset could be classified as 'big' if technical resources (speed, memory) are not capable of analyzing the data, using existing methods. Big Data in a field like analysis of particle collision at CERN [3] is driven by physical challenges (hardware, computer speed and physical computer memory required to store and analyze such data), which may be addressed by the development of new and more advanced software.

Medicinal chemistry related data are created and curated in pharmaceutical industry via high-throughput screening (HTS) and drug discovery campaigns and additionally also available in databases sourced from scientific journals, patents etc. For example, AstraZeneca in-house screening database contains over 150 million structure–activity relationship (SAR) data point [2]. The HTS data from pharma companies are usually very sparse and for each screened target there is only a small number of active hits. Further developments are done with a relatively small series of compounds, usually varying from hundreds to thousands of compounds for those series. Specialists who work on these target specific data do not have Big Data in their daily work; traditional modeling algorithm is well enough to handle their datasets.

When the focus is on chemogenomics data, the situation is different. The biggest medicinal chemistry data reservoir, PubChem, currently comprise 91 million chemical structures and 230 million bioactivity data points corresponding to over

FUTURE SCIENCE part of fsg

10K protein targets. The total data size is around 60GBs [4], which is considered 'big' in medicinal chemistry terms, but is in fact still well below the terabyte or even petabyte data comprised by databases such as eBay [5] and Amazon [6]. However, if chemical descriptors (such as structural fingerprints) of compounds are generated for this level of dataset, the total data size will probably be in the realms of the conventional Big Data size. For each specific protein target, the available SAR data will be much less and it would be in the range of hundreds of thousands of data points, or up to a few million data points if inactive compounds from an HTS were taken into account. For building single-target quantitative SAR (QSAR) models, the traditional machine learning algorithms will still be capable of handling this magnitude of data [7]. But, if one wants to use all available chemogenomics data (in databases like PubChem, ChEMBL etc.) to pursue multitask learning (see below) and build one multilabel model to predict multiple target activity simultaneously, traditional algorithms used in chemoinformatics are unlikely to work and it would require huge computer power and a dedicated parallel programming model to solve the problem.

> "The measurements of even some simple properties, such as solubility in water, can be difficult, time consuming and error prone."

Another big challenge in medicinal chemistry, where Big Data can have an impact, is related to the question of which molecule to synthesize next in a drug discovery project. To identify the optimal candidate for synthesis, large virtual chemical spaces need to be explored, which clearly is a Big Data-related problem. So return to our initial question, we can conclude that Big Data does exist in medicinal chemistry and there are a number of challenges associated with this, depending on which aspect of the field is under focus.

## Is Big Data really useful for prediction?

Let us now consider an example of how data size can make a difference in property predictions. In 2014 IVT published a melting point (MP) model based on approximately 50k compounds [8], which was succeeded by approximately 275k compounds [9] model in 2016. The 'large' set of 50 k compounds was processed by On-line Chemical Database and Modeling Environment (OCHEM) [10,11] using the same approaches applied in multiple previous studies. The latter set was considered Big Data since we could not use the previous tools without changes. Among other things, we had to implement parallel calculations using a support vector machines method, solve problems of storage of very large data matrices using sparse format (including calculations with matrices incorporating >0.2 trillion entries), as well as account for several other technical challenges. Were the results worth our efforts? Yes. The model developed with Big Data was more accurate and calculated, for example, the lowest published error for Bergström set of drugs [12]. Moreover, prediction of water solubility using Yalkowsky's General Solubility Equation, which is based on MP and octanol/water partition coefficient (logP) [13], was also significantly improved compared with using the model developed with 50k [9]. Considering that the Big Data set mainly contained data automatically mined from patent literature [9], it also proved feasibility and success of a developed fully automatic data extraction technology.

## Multitask learning & deep learning for Big Data

Data collection is always a challenging task. The measurements of even some simple properties, such as solubility in water, can be difficult, time consuming and error prone [14]. However, many physicochemical and biological properties are strongly interrelated, for example, the water solubility depends on logP and MP as shown by the General Solubility Equation [13]. One strategy could be to explore these relationships by developing models for several properties simultaneously [15]. This multitask learning concept is especially attractive for building polypharmacology prediction models, since many protein targets are interrelated due to the intrinsic similarity in sequence or interaction pattern of binding pockets. One recent study shows that massive multitask networks obtain predictive accuracies significantly better than single-task methods when the different outcomes are related to each other [16]. Deep learning methods are also thought to help in addressing this issue (see [17,18] for a review of these approaches in drug discovery). These new and promising approaches have already been used to achieve superhuman accuracy in recognizing Chinese characters [19] and to develop a computer program (AlphaGo), which was capable of beating 18-time world champion Lee Sedol at the ancient and complicated game, Go [20]. An important milestone for deep learning and multitask learning was their performance in the TOX21 challenge, where the combination of them provided overall best accuracy using the area under the curve performance measure [21]. Interestingly, the best balanced accuracy was calculated using Associative Neural Networks [22], which

were developed using traditional 'shallow' neural networks [23]. This method also contributed the top-ranked model for the ToxCast challenge [24]. Probably, their combination, also known as 'deep Associative Neural Networks', can contribute even better models. Deep learning has already demonstrated advantages in combining tasks and data by simultaneous analysis of 259 datasets totaling >40 M data points from public databases [16]. Large-scale chemogenomics modeling is currently an active research field with several important publications recently from Jansen Pharmaceuticals [25–28]. The approaches may still need to be optimized to learn imbalanced datasets, data weighted by measurements accuracies, to identify methods for optimal combination of qualitative and quantitative data as well as use of unsupervised data [17]. Thus, instead of filtering data by removing low-quality records coming from less reliable experiments, one may develop better global models by using all data. A Horizon2020-funded research project is currently working on addressing the current limitations of large-scale chemogenomics modeling [29]. The data of different quality could be weighted by their experimental or estimated accuracies. In the future, the machine learning methods may also be merged with systems biology approaches to predict pharmacokinetic/pharmacodynamic (PK/PD) and/or to better use *in vitro* measurements to estimate *in vivo* toxicity, which remains a challenge for traditional methods [24]. Another important direction is the optimal use of global models to create highly accurate local models based on additional data, as was demonstrated in our study that looked at predicting the logP of Pt complexes using computational methods [30]. The development of such approaches is important to improve global models for new compound series using few high accuracy measurements. These tasks are typical among the medicinal chemistry field and this is one of the application areas where the use of Big Data is highly required.

## Challenges of exploring virtual chemical spaces

The global models developed with technologies described in the previous section can be particularly useful for searching virtual chemical spaces, which is basically a Big Data problem. In a drug discovery project, the constantly posted question by medicinal chemist is which molecule to synthesize next. Due to the vastness of the chemical space even to enumerate the chemical space around the existing prioritized compounds and to score the compounds for synthetic feasibility, ADME and on-target as well as off-target potency is a true Big Data problem. If all available information is taking into account with the latest machine learning algorithms like multitask learning where all models are trained simultaneously much larger computing resources are needed in comparison to standard single QSAR models. Additionally it is desirable that all models can be automatically updated every time new experimental data are uploaded. A specific example would be to build models to predict off-targets for each molecule proposed for synthesis. To train multitask models for the whole genome would be several thousand models with up to more than one million data points per model if HTS data are used for training the models. Thus the enumeration of chemical space, as well as the building and updating of models, are all Big Data problems that are highly relevant for medicinal chemistry.

> "The development and use of methods to advance analysis of Big Data requires adequately trained specialists."

As an example, the chemical universe database GDB17 enumerated >166 billion compounds containing up to 17 atoms [31] while the total space of drug-like molecules is estimated to be about $10^{60}$. For medicinal chemists these virtual spaces can be used to identify new drug-like molecules with favorable properties, for example, promising ADME/T properties that are conducive toward drug development. This is a highly challenging task. Even an annotation of GDB17 using a fast prediction model, which would calculate 100,000 molecules per minute, would require more than 3 years of computing time on a single processor [1]. Moreover, the straightforward prediction of all these compounds can be of a limited value. The numbers of existing experimental measured data points vary from hundreds (complex biological properties such as oral bioavailability) to hundreds of thousands of measurements (simple physicochemical properties, such as MP). The use of multilearning of several properties can help to enlarge experimental space, but even in this case the developed models would still need to extrapolate from one experimental measurement to hundreds of thousands or even hundreds of millions of compounds for prediction of GDB17. The extrapolations, of course, cannot be reliable for all compounds. The applicability domain (AD) methods [32] can help to identify subsets of molecules in the twilight drug-like zone, in other words, molecules with reliable predictions but also with properties favorable for drug development (high solubility, low toxicity and so on). Although many AD methods exist [32], a model based on compounds that are solid at a room

temperature failed to identify compounds, which have MP below 0°C, using a state-of-the art method for AD definition [8]. This result indicates that the AD methods need further development to be reliably used for analysis of large virtual chemical spaces. To better estimate the AD and accordingly the confidence in predictions, the method of conformal prediction has been pioneered in the drug discovery field by AstraZeneca and collaborators [33]. Conformal predictors were originally developed at Royal Holloway, University of London, UK.

A prominent example of medicinal chemistry-related application of Big Data in pharmaceutical industry is the design and utilization of the so-called virtual library, which is constructed on compiled large number of organic chemical reactions and available chemical reagents. In AstraZeneca, such a virtual library was constructed using synthetic protocols extracted from in-house corporate electronic laboratory notebook to enable virtual screening in this huge chemical space (can reach $10^{15}$ products in theory) via 2D structural fingerprint [34]. Similar systems have been developed in other pharmaceutical companies. BI-Claim from Boehringer Ingelheim uses in-house combinatorial library generation protocols and commercial reagents to generate virtual libraries [35]. The system could theoretically enumerate $5 \times 10^{11}$ virtual chemical structures and the similarity searching can be carried out via Ftrees-Fragment Spaces. One application of this platform on drug discovery project has been reported [36] that virtual screening on the combinatorial libraries via Ftrees-Fragment Spaces led to the identification of two new structural classes of GPR119 agonists with submicromolar *in vitro* potencies. Researchers from Pfizer reported on the Pfizer Global Virtual Library (PGVL) of synthetically feasible compounds, which makes use of over 1200 combinatorial reactions and can theoretically enumerate $10^{14}$–$10^{18}$ virtual compounds [37]. A custom desktop software package, PGVL-Hub, has been developed to enable the similarity search on interested queries and design-focused libraries [38]. The impact of PGVL-Hub has been applied [39] in the discovery of novel Chk1 inhibitors, where two lead compounds were obtained through two rounds of focused library design using PGVL-Hub based on one initial HTS hit. Recently researchers of Lilly reported using their own virtual library platform, Proximal Lilly Collection, to carry out near neighbor search, focused library design and virtual screening. To develop selective hRIO2 Kinase Inhibitors, an similarity search on Proximal Lilly Collection based on an old anti-inflammatory drug diphenpyramide was done and led to the identification of one follow-up compound with tenfold increment on its potency [40].

## Training in Big Data analytics

The development and use of methods to advance analysis of Big Data requires adequately trained specialists [1]. In this respect chemoinformatics specialists, who receive education both in chemistry and in informatics, will play a leading role in the further development of this field. Funding provided by the EU commission to educational programs, such as Marie Skłodowska-Curie Innovative Training Network European Doctorate 'Big Data in Chemistry' [41], is also important in developing specialized training programs that closely match the requirements of industry with proposed theoretical and practical training.

## Conclusion

While one may question if Big Data accurately describes the datasets handled within the medicinal chemistry field, there is no denying that there is a demand for Big Data approaches that are capable of analyzing the increasing volumes of data in this field. Techniques and methods that enable the exploration of virtual chemical spaces and (deep) learning of several properties simultaneously are expected to allow medicinal chemists to efficiently exploit Big Data. Last but not least, further progress will also critically depend on training programs and advances in chemoinformatics, a discipline bridging chemistry and informatics.

### Acknowledgements

This 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.

### Financial & competing interests disclosure

The project leading to this article has received funding from the European Union's Horizon2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 676434, 'Big Data in Chemistry' ('BIGCHEM'). IV Tetko is CEO and founder of BigChem GmbH, which licenses the OCHEM [10]. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

## References


1 Tetko IV, Engkvist O, Koch U, Reymond JL, Chen H. BIGCHEM: challenges and opportunities for Big Data analysis in chemistry. *Mol. Inform.* doi:10.1002/minf.201600073 (2016) (Epub ahead of print).

2 Muresan S, Petrov P, Southan C *et al.* Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data. *Drug Discov. Today* 16(23–24), 1019–1030 (2011).

3 Mondal K. Design issues of Big Data parallelisms. *Adv. Intell. Syst. Comput.* 434, 209–217 (2016).

4 PubChem BioAssay Database. http://pubchem.ncbi.nlm.nih.gov

5 Inside eBay's 90PB data warehouse. www.itnews.com.au/news/inside-ebay8217s-90pb-data

6 How Amazon Works. http://money.howstuffworks.com/amazon1.htm

7 Mervin LH, Afzal AM, Drakakis G, Lewis R, Engkvist O, Bender A. Target prediction utilising negative bioactivity data covering large chemical space. *J. Cheminform.* 7, 51 (2015).

8 Tetko IV, Sushko Y, Novotarskyi S *et al.* How accurately can we predict the melting points of drug-like compounds? *J. Chem. Inf. Model.* 54(12), 3320–3329 (2014).

9 Tetko IV, Lowe D, Williams AJ. The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from PATENTS. *J. Cheminform.* 8, 2 (2016).

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

11 Online Chemical Database. http://ochem.eu

12 Bergstrom CA, Norinder U, Luthman K, Artursson P. Molecular descriptors influencing melting point and their role in classification of solid drugs. *J. Chem. Inf. Comput. Sci.* 43(4), 1177–1185 (2003).

13 Ran Y, Yalkowsky SH. Prediction of drug solubility by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* 41(2), 354–357 (2001).

14 Balakin KV, Savchuk NP, Tetko IV. *In silico* approaches to prediction of aqueous and DMSO solubility of drug-like compounds: trends, problems and solutions. *Curr. Med. Chem.* 13(2), 223–241 (2006).

15 Varnek A, Gaudin C, Marcou G, Baskin I, Pandey AK, Tetko IV. Inductive transfer of knowledge: application of multi-task learning and feature net approaches to model tissue-air partition coefficients. *J. Chem. Inf. Model.* 49(1), 133–144 (2009).

16 Ramsundar B, Kearnes S, Riley P, Webster D, Konerding D, Pande V. Massively multitask networks for drug discovery. *ArXiv e-prints* 1502.02072 (2015).

17 Baskin I, Winkler D, Tetko IV. A renaissance of neural networks in drug discovery. *Expert Opin. Drug Discov.* 11(8), 785–795 (2016).

18 Gawehn E, Hiss JA, Schneider G. Deep learning in drug discovery. *Mol. Inf.* 35(1), 3–14 (2016).

19 96.7% recognition rate for handwritten Chinese characters using AI that mimics the human brain. http://phys.org/news/2015-09-recognition-handwritten

20 Borowiec S. AlphaGo seals 4–1 victory over Go grandmaster Lee Sedol. *The Guardian* 15 March (2016). www.theguardian.com/technology/2016/mar/15

21 Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.* 3, 80 (2016).

22 Tetko IV. Associative neural network. *Methods Mol. Biol.* 458, 185–202 (2008).

23 Abdelaziz A, Spahn-Langguth H, Werner-Schramm K, Tetko IV. Consensus modeling for HTS assays using *in silico* descriptors calculates the best balanced accuracy in Tox21 challenge. *Front. Environ. Sci.* 4, 2 (2016).

24 Novotarskyi S, Abdelaziz A, Sushko Y, Korner R, Vogt J, Tetko IV. ToxCast EPA *in vitro* to *in vivo* challenge: insight into the rank-i model. *Chem. Res. Toxicol.* 29(5), 768–775 (2016).

25 Simm J, Arany A, Zakeri P *et al.* Macau: scalable bayesian multi-relational factorization with side information using MCMC. *ArXiv e-prints* 1509.04610 (2015). https://arxiv.org/abs/1509.04610

26 Zawbaa HM, Szlek J, Grosan C, Jachowicz R, Mendyk A. Computational intelligence modeling of the macromolecules release from PLGA microspheres – focus on feature selection. *PLoS ONE* 11(6), e0157610 (2016).

27 Arany A, Simm J, Zakeri P *et al.* Highly scalable tensor factorization for prediction of drug–protein interaction type. *ArXiv e-prints* 1512.00315 (2015). https://arxiv.org/pdf/1512.00315.pdf

28 Unterthiner T, Mayr A, Klambauer G *et al.* Deep learning as an opportunity in virtual screening. Presented at: *NIPS 2014 Deep Learning and Representation Learning Workshop.* Montreal, Canada, 8–13 December 2014.

29 Exascalable Compound Activity Prediction Engines: www.excape-h2020.eu

30 Tetko IV, Jaroszewicz I, Platts JA, Kuduk-Jaworska J. Calculation of lipophilicity for Pt(II) complexes: experimental comparison of several methods. *J. Inorg. Biochem.* 102(7), 1424–1437 (2008).

31 Ruddigkeit L, Van Deursen R, Blum LC, Reymond JL. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* 52(11), 2864–2875 (2012).

32 Tetko IV, Bruneau P, Mewes HW, Rohrer DC, Poda GI. Can we estimate the accuracy of ADME-Tox predictions? *Drug Discov. Today* 11(15–16), 700–707 (2006).

33 Norinder U, Carlsson L, Boyer S, Eklund M. Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. *J. Chem. Inf. Model.* 54(6), 1596–1603 (2014).

34 Vainio MJ, Kogej T, Raubacher F. Automated recycling of chemistry for virtual screening and library design. *J. Chem. Inf. Model.* 52(7), 1777–1786 (2012).

35 Lessel U, Wellenzohn B, Lilienthal M, Claussen H. Searching fragment spaces with feature trees. *J. Chem. Inf. Model.* 49(2), 270–279 (2009).

36 Wellenzohn B, Lessel U, Beller A, Isambert T, Hoenke C, Nosse B. Identification of new potent GPR119 agonists by combining virtual screening and combinatorial chemistry. *J. Med. Chem.* 55(24), 11031–11041 (2012).

37 Peng Z. Very large virtual compound spaces: construction, storage and utility in drug discovery. *Drug Discov. Today Technol.* 10(3), e387–394 (2013).

38 Peng Z, Yang B, Mattaparti S *et al.* PGVL Hub: An integrated desktop tool for medicinal chemists to streamline design and synthesis of chemical libraries and singleton compounds. *Methods Mol. Biol.* 685 295–320 (2011).

39 Teng M, Zhu J, Johnson MD *et al.* Structure-based design and synthesis of (5-arylamino-2H-pyrazol-3-yl)-biphenyl-2′,4′-diols as novel and potent human CHK1 inhibitors. *J. Med. Chem.* 50(22), 5253–5256 (2007).

40 Nicolaou CA, Watson IA, Hu H, Wang J. The proximal lilly collection: mapping, exploring and exploiting feasible chemical space. *J. Chem. Inf. Model.* 56(7), 1253–1266 (2016).

41 BigChem. http://bigchem.eu