Tetko and Engkvist *J Cheminform (2020) 12:74*
https://doi.org/10.1186/s13321-020-00475-y

Journal of Cheminformatics

EDITORIAL

Open Access

From Big Data to Artificial Intelligence: chemoinformatics meets new challenges

Igor V. Tetko[1,2]* and Ola Engkvist[3]

Abstract

The increasing volume of biomedical data in chemistry and life sciences requires development of new methods and approaches for their analysis. Artificial Intelligence and machine learning, especially neural networks, are increasingly used in the chemical industry, in particular with respect to Big Data. This editorial highlights the main results presented during the special session of the International Conference on Neural Networks organized by "Big Data in Chemistry" project and draws perspectives on the future progress of the field.

The analysis and exploitation of Big Data was the cornerstone of the "Big Data in Chemistry" (BIGCHEM), and of this special issue, which was prepared following the International Conference on Neural Networks (ICANN2019). In total 17 articles, including 15 contributions co-authored by BIGCHEM PhD students and partners, were published in this issue. Its thematic covered many different aspects of the use of Big Data in medicinal chemistry [1, 2] that were actively pursued and advanced during the project. The articles in the issue can be categorized into two main groups.

The first group deals with machine learning methods to improve analysis of large datasets such as those of high-throughput screening (HTS) campaigns. The comparison of structure-based and protein–ligand interaction fingerprints (IFPs) and for the prediction of ligand binding modes for protein kinases were studied by Rodríguez-Pérez et al. [3]. The authors showed that including target-relevant information via IPFs improved predictions of the modes by about 10% compared to the use of traditional atom environment fingerprints. Laufkötter et al. [4] demonstrated that augmenting chemical structure descriptors with bio-activity based fingerprints derived from HTS data provides better performance but, importantly, also superior scaffold hopping capability. Analogously QSAR-derived affinity fingerprints (QAFFP) [5, 6] outperformed classical Morgan fingerprints for scaffold hopping. While Morgan fingerprints due to their robustness and performance for small molecules (see review of David et al. [7]) are frequently used as a gold standard in, e.g., virtual screening and target predictions, they might not be optimal for larger molecules, such as peptides. MinHashed Atom-Pair fingerprints with a diameter of up to four bonds (MAP4) [8] were introduced as a universal fingerprint providing good results for various targets. HTS data are frequently imbalanced with only few active compounds: COVER (conformational oversampling as data augmentation for molecules) generates multiple conformations of molecules, in order to provide an efficient data balancing mechanism for the underrepresented class [9]. All these methodological studies are important to have better models for Big Data.

The second group of articles deals with novel machine-learning algorithms such as the use of generative models (GMs) for molecular *de novo* design in drug discovery. BIGCHEM was one of the originators in this area of research with its pioneering works on applying Recurrent Neural Networks (RNN) with reinforcement learning and variational autoencoders for molecular designs [10, 11] as reviewed elsewhere [12]. LatentGAN represents one of the most advanced

*Correspondence: itetko@vcclab.org
[1] Helmholtz Zentrum München-German Research Center for Environmental Health (GmbH), Institute of Structural Biology, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany
Full list of author information is available at the end of the article

developments of GMs by combining an autoencoder and a generative adversarial neural network [13]. The overfitting can reduce the diversity of autoencoder-generated structures. Generative Examination Networks (GEN) use randomized SMILES and early stopping [14] to prevent this [15]. The effect of the randomized SMILES to improve the quality of GMs is also confirmed with extensive benchmarking based on GDB-13 [16]. Another method to increase the diversity of generated structures was proposed by Blaschke et al. [17] who use memory-assisted reinforcement learning for this purpose. The methods developed in these studies are general ones and can be used to enhance other GMs such as scaffold decoration [18] or Mol-CycleGAN [19]. New types of deep learning algorithms based on Message Passing Neural Networks [20] and Transformer Convolutional Neural Networks (CNN) [21] were also introduced.

There is a significant difference between both groups of articles. The methods used in the first group mainly explore traditional machine learning methods, such as Random Forest, Support Vector Machines, etc. that are based on traditional molecular representations as a vector of descriptors [7]. Those studies could be performed using traditional toolkits with no or little programming effort. Contrary to that, the generative models were based on novel machine deep learning architectures such as CNN, RNN, Long Short Term Memory, Transformers, etc. These methods are more innovative and most of them were introduced, developed and/or implemented by the authors. All of the studies from this second group thus required significant programming skills and expertise in modern toolkits such as TensorFlow, Keras, Pytorch, etc., which are becoming a pre-requisite to get a position in the industry in addition, of course, to excellent knowledge in the basis disciplines. Prospective PhD students, who plan to build their careers in the field of chemoinformatics and Artificial Intelligence (AI), should not overlook these requirements.

One of the major differences of these new methods is their ability to infer statistical dependencies directly from chemical structures, which can be represented as text, e.g. SMILES [21, 22], chemical graphs [20], or 3D images [23]. The benchmarking studies [20, 22] show that such methods can achieve similar or better performances compared to traditional methods for classical tasks such as QSAR[24] but at the same time allow for intuitive interpretation of models [21]. Moreover, they can be used to address very different tasks, such as the aforementioned generation of molecules with desired properties or/and the prediction of single step (retro) synthesis [25, 26], or even complete retro-synthesis [27, 28] that could not be achieved with traditional methods. All these approaches are part of the emerging area of AI, which is going to drive the future of chemoinformatics.

AI is fast becoming a ubiquitous part of modern life, and is also increasingly employed in the pharmaceutical industry to automate key steps in drug design. Compared to Big Data challenges, "how to best analyse the Big Data" [1, 2], the future progress in this field is linked to the need for explainable "chemistry aware" methods. Such method should allow the elucidation of the molecular basis of compound activity, or to directly suggest new compounds with improved properties, or to optimise routes for synthesis supported with chemical knowledge. These and other topics, such as interpretable deep learning, use of knowledge elicitation from human experts, machine learning and molecular dynamics, language and quantum chemistry based retro-synthesis prediction, scalable multi-objective synthesis route optimization, methods for scaffold hopping, uncertainty estimation of AI methods, etc., will be investigated within the "Advanced machine learning for Innovative Drug Discovery" (AIDD, http://ai-dd.eu). This project will employ 16 fellows starting January 2021, who will get training and full support in theoretical and practical skills from their supervisors and via various network activities. While not being a direct continuation of BIGCHEM, AIDD will definitely contribute to the further development of the successful methods originated from the previous network.

The advance in this field critically depends on the availability of open source software, which is important for sustainable progress and sharing results. A distinct feature of BIGCHEM was the voluntary decision of its several partners to release the source code for its methodological developments, which dramatically boosted the respective research areas. For example the publicly available source code [29] from the REINVENT [10] article was forked 75 times since its publication, which definitely contributed to a rigorous validation, and a wide acceptance of the published results by the scientific community. The same principle will be widened in the AIDD, where all partners have agreed to release the source codes of their individual projects to improve their dissemination.

In summary, this special issue comprises a carefully selected collection of articles in Big Data, most of which were contributed by BIGCHEM partners or reported during the ICANN19 (http://e-nns.org/icann2019). Considering the great success of the project, which contributed about 70 publications that were cited nearly 1000 times in 2020 alone (https://scholar.google.com/citations?user=eLncF6MAAAAJ) as well as of the ICANN19, which was attended by all-time record of 500 participants and resulted in five volumes of proceedings [30], we believe that this special issue will be of a great interest to the readers of the journal.

Acknowledgements
This study was partially funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Innovative Training Network European Industrial Doctorate grant agreement No. 676434, "Big Data in Chemistry" (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 fruitful collaboration during the project. IVT is CEO and founder of BIGCHEM GmbH, which licenses the On-line Database and Chemical Modelling Environment (http://ochem.eu), that was used in several studies reported in this editorial. OE declares that he has no actual or potential conflicts of interests.

Author details
[1] Helmholtz Zentrum München-German Research Center for Environmental Health (GmbH), Institute of Structural Biology, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany. [2] BIGCHEM GmbH, Valerystr. 49, 85716 Unterschleißheim, Germany. [3] Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden.

Received: 18 November 2020 Accepted: 18 November 2020
Published online: 18 December 2020

References

1. Tetko IV, Engkvist O, Koch U, Reymond JL, Chen H (2016) BIGCHEM: challenges and opportunities for Big Data analysis in chemistry. Mol Inform 35(11–12):615–621
2. Tetko IV, Engkvist O, Chen H (2016) Does "Big Data" exist in medicinal chemistry, and if so, how can it be harnessed? Future Med Chem 8(15):1801–1806
3. Rodríguez-Pérez R, Miljković F, Bajorath J (2020) Assessing the information content of structural and protein–ligand interaction representations for the classification of kinase inhibitor binding modes via machine learning and active learning. J Cheminform 12(1):36
4. Laufkötter O, Sturm N, Bajorath J, Chen H, Engkvist O (2019) Combining structural and bioactivity-based fingerprints improves prediction performance and scaffold hopping capability. J Cheminform 11(1):54
5. Cortés-Ciriano I, Škuta C, Bender A, Svozil D (2020) QSAR-derived affinity fingerprints (part 2): modeling performance for potency prediction. J Cheminform 12(1):41
6. Škuta C, Cortés-Ciriano I, Dehaen W, Kříž P, van Westen GJP, Tetko IV, Bender A, Svozil D (2020) QSAR-derived affinity fingerprints (part 1): fingerprint construction and modeling performance for similarity searching, bioactivity classification and scaffold hopping. J Cheminform 12(1):39
7. David L, Thakkar A, Mercado R, Engkvist O (2020) Molecular representations in AI-driven drug discovery: a review and practical guide. J Cheminform 12(1):56
8. Capecchi A, Probst D, Reymond JL (2020) One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J Cheminform 12(1):43
9. Hemmerich J, Asilar E, Ecker GF (2020) COVER: conformational oversampling as data augmentation for molecules. J Cheminform 12(1):18
10. Olivecrona M, Blaschke T, Engkvist O, Chen H (2017) Molecular de-novo design through deep reinforcement learning. J Cheminform 9(1):48
11. Blaschke T, Olivecrona M, Engkvist O, Bajorath J, Chen H (2018) Application of generative autoencoder in De Novo molecular design. Mol Inform 37(1–2):1700123
12. Engkvist O, Arús-Pous J, Bjerrum EJ, Chen H: Chapter 13 Molecular De Novo Design Through Deep Generative Models. Artificial Intelligence in Drug Discovery. The Royal Society of Chemistry; 2021. pp. 272–300.
13. Prykhodko O, Johansson SV, Kotsias P-C, Arús-Pous J, Bjerrum EJ, Engkvist O, Chen H (2019) A de novo molecular generation method using latent vector based generative adversarial network. J Cheminform 11(1):74
14. Tetko IV, Livingstone DJ, Luik AI (1995) Neural network studies. 1. Comparison of overfitting and overtraining. J Chem Inf Comput Sci 35(5):826–833
15. van Deursen R, Ertl P, Tetko IV, Godin G (2020) GEN: highly efficient SMILES explorer using autodidactic generative examination networks. J Cheminform 12(1):22
16. Arús-Pous J, Johansson SV, Prykhodko O, Bjerrum EJ, Tyrchan C, Reymond J-L, Chen H, Engkvist O (2019) Randomized SMILES strings improve the quality of molecular generative models. J Cheminform 11(1):71
17. Blaschke T, Engkvist O, Bajorath J, Chen H (2020) Memory-assisted reinforcement learning for diverse molecular de novo design. J Cheminform 12(1):68
18. Arús-Pous J, Patronov A, Bjerrum EJ, Tyrchan C, Reymond J-L, Chen H, Engkvist O (2020) SMILES-based deep generative scaffold decorator for de-novo drug design. J Cheminform 12(1):38
19. Maziarka Ł, Pocha A, Kaczmarczyk J, Rataj K, Danel T, Warchoł M (2020) Mol-CycleGAN: a generative model for molecular optimization. J Cheminform 12(1):2
20. Withnall M, Lindelöf E, Engkvist O, Chen H (2020) Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction. J Cheminform 12(1):1
21. Karpov P, Godin G, Tetko IV (2020) Transformer-CNN: Swiss knife for QSAR modeling and interpretation. J Cheminform 12(1):17
22. Tetko IV, Karpov P, Bruno E, Kimber TB, Godin G: Augmentation is what you need! Artificial neural networks and machine learning—ICANN 2019: Workshop and Special Sessions: 17th–19th September 2019 2019; Münich. Springer International Publishing. pp. 831–835.
23. Iqbal J, Vogt M, Bajorath J (2020) Activity landscape image analysis using convolutional neural networks. J Cheminform 12(1):34
24. Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, Oprea TI, Baskin II, Varnek A, Roitberg A et al (2020) Correction: QSAR without borders. Chem Soc Rev 49(11):3716
25. Tetko IV, Karpov P, Van Deursen R, Godin G (2020) State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. Nat Comm 11(1):1–11
26. Karpov P, Godin G, Tetko IV: A Transformer Model for Retrosynthesis. In: *Artificial Neural Networks and Machine Learning—ICANN 2019: Workshop and Special Sessions: 17th–19th September 2019 2019; Münich*. Springer International Publishing. pp. 817–830.
27. Thakkar A, Bjerrum EJ, Engkvist O, Reymond J-L: Neural network guided tree-search policies for synthesis planning. Artificial neural networks and machine learning—ICANN 2019: workshop and special sessions: 17th–19th September 2019 2019; Münich. Springer International Publishing: 721-724.
28. Genheden S, Thakkar A, Chadimová V, Reymond J-L, Engkvist O, Bjerrum E (2020) AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. J Cheminform 12(1):70
29. REINVENT [https://github.com/MarcusOlivecrona/REINVENT]
30. Tetko IV, Theis F, Karpov P, Kůrková V (2019) Artificial Neural Networks and Machine Learning—ICANN 2019: 28th International Conference on Artificial Neural Networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). volumes 11727–11731 LNCS

Publisher's Note