JOURNAL OF CHEMICAL INFORMATION AND MODELING

EDITORIAL

pubs.acs.org/jcim

Data-Driven High-Throughput Prediction of the 3-D Structure of Small Molecules: Review and Progress. A Response to the Letter by the Cambridge Crystallographic Data Centre

ABSTRACT: A response is presented to sentiments expressed in "Data-Driven High-Throughput Prediction of the 3-D Structure of Small Molecules: Review and Progress. A Response from The Cambridge Crystallographic Data Centre", recently published in the Journal of Chemical Information and Modeling,[1] which may give readers a misleading impression regarding significant impediments to scientific research posed by the CCDC.

The core of the problem is the statement in the CCDC's license that "licensee will not redistribute derived data without prior permission of the CCDC". While it may be understandable that restrictions ought to be placed on the redistribution of the Cambridge Structural Database (CSD) itself, posing restrictions on all "derived data" is indeed problematic and creates significant impediments for academic research and science.

In the specific case of the article in question,[2] a small molecule 3-D structure predictor and Web server (COSMOS) was constructed using parameters (e.g., bond length and angles) of molecular fragments from the CSD. The CCDC vigorously intervened to prevent distribution of such a tool. The statement in the CCDC's letter that "express permission was immediately granted" is simply false. A dozen librarians and other staff from the University of California (UC) had to intervene under the threat of losing a system-wide license to the CSD. As a temporary solution, UC Irvine reluctantly agreed to separate the UCI license from the system-wide UC license and place major restrictions on the COSMOS Web server, such as not allowing the prediction of more than 100 molecules at a time or periodically reporting to the CCDC the list of users of the COSMOS Web server.

Since most of the useful data in the CSD is derived from the literature, parameters derived from the CSD are inherently parameters derived from the literature. In fact, virtually all 3-D structure predictors for small molecules use parameters such as the length of single, double, or triple carbon−carbon bonds. Should all such predictors be considered as "derived" from the CSD and thus subject to the CCDC restrictive practices? More importantly, and beyond the issue of the definition of what it really means to be "derived" from the CSD, trying to prevent or restrict the free distribution and use of 3-D structure prediction tools and servers within the academic research community is a major obstacle and impediment to research in this area. In the high-throughput data era, data mining tools that allow scientists to process large amounts of data are critical for the progress of science. This is obvious in sciences as diverse as astronomy, physics, and biology where for instance high-throughput tools for annotating entire genomes or predicting the 3-D structure of proteins are freely distributed for academic research purposes. The circulation of these tools is essential for assessing their quality, comparing them to each other, improving them, and ultimately for advancing our understanding of the data they are meant to interpret. Accurate prediction of 3-D structures is central to chemistry and drug discovery; thus, any restrictions in this area impact not only the scientists that are involved but ultimately all the tax payers.

As history shows, those who stand in the way of democracy and scientific progress end up losing over the long run. The reactionary attitude of the CCDC staff has started to backfire by energizing academic laboratories around the world to find alternative solutions around the CCDC. There are already several efforts (e.g., Crystallography Open Database[3] and CrystalEye[4]) to produce large, freely available databases of crystallographic structures using the same main source as the CSD—publicly available data. Furthermore, quantum mechanical methods have now reached the level of speed and accuracy where they can be used on computer clusters to obtain accurate structures for large numbers of molecules and fragments. Accordingly, we expect a version of COSMOS that does not use any fragments "derived" from the CSD to be available in the near future. The CCDC staff should want to be at the frontiers of science and revise its outdated licensing terms to allow and encourage greater use of its data in different and creative ways, thereby helping science rather than standing in its way.

PierreBaldi*

Department of Computer Science, University of California, Irvine, Irvine, California 92697-3435, United States

AUTHOR INFORMATION

Corresponding Author

*pfbaldi@ics.uci.edu.

REFERENCES

(1) Groom, C. R. Data-driven high-throughput prediction of the 3-D structure of small molecules: Review and progress. A response from The Cambridge Crystallographic Data Centre. *J. Chem. Inf. Model.* **2011**, *51*, 2787–2787.

(2) Andronico, A.; Randall, A.; Benz, R. W.; Baldi, P. Data-driven high-throughput prediction of the 3-D structure of small molecules: Review and progress. *J. Chem. Inf. Model.* **2011**, *51*, 760–776.

(3) Gražulis, S.; Chateigner, D.; Downs, R. T.; Yokochi, A. F. T.; Quirós, M.; Lutterotti, L.; Manakova, E.; Butkus, J.; Moeck, P.; Le Bail, A. Crystallography Open Database: An open-access collection of crystal structures. *J. Appl. Crystallogr.* **2009**, *42* (4), 726–729.

(4) CrystalEye. University of Cambridge. http://wwmm.ch.cam.ac.uk/crystaleye/index.html.

Published: November 22, 2011