Members of our early career panel highlight key research articles on the theme of computer-aided drug design

Filip Miljković*,¹ & Rajan Chaudhari²

¹Data Science & Artificial Intelligence, Imaging & Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Gothenburg, Sweden

²Department of Experimental Therapeutics, The University of Texas MD Anderson Cancer Center, Houston, TX 77054, USA

*Author for correspondence: filip.miljkovic@astrazeneca.com

First draft submitted: 4 October 2020; Accepted for publication: 5 October 2020; Published online: 19 October 2020

Keywords: artificial intelligence • computer-aided drug design • drug design

Small, colloiddally aggregating molecule Detective: accurate predictor of small, colloiddally aggregating molecules

Colloidal aggregation of small molecules remains the most common source of false positives in high-throughput screening campaigns. Due to the poor solubility, these small, colloiddally aggregating molecules (SCAMs) destabilize colloidal systems and undergo particle agglomeration which in turn causes artifactual protein inhibition. Despite the detailed understanding of this phenomenon, colloidal aggregation still threatens the early drug discovery efforts. Thus, appropriate computational methods are required to flag potential SCAMs and refine the compound selection for high-throughput screening procedures.

In this study, the authors have developed and validated multiple quantitative structure-interference relationship models using the data from several high-throughput screening campaigns [1]. These included an AmpC β -lactamase assay, a common counter-screen for aggregation and an additional cysteine protease cruzain assay, both run under different screening concentrations and assay conditions. Classification models distinguishing between aggregator and nonaggregator compounds were built using random forest and deep neural networks. Comparable accuracy was achieved by random forest and deep neural network models (67–77%) for both cross-validation and external withhold sets. Moreover, SCAM Detective increased the performance by ~53% in β -lactamase assay and ~46% in cruzain assay compared with previously published models. However, as the authors pointed out, the direct comparison with these models could not be made due to minor yet significant differences in modeling tasks. Nevertheless, carefully curated data sets and rigorously validated machine learning models that follow state-of-the-art practices make this research highly compelling.

To support the identification of potential SCAMs, the models were made publicly available as a web application [2]. Besides the predicted class for each compound, the models provide a confidence score, an applicability domain value and a molecular fragment contribution scheme supporting the prediction.

- Written by F Miljković

Automated identification of chemical series: classifying like a medicinal chemist

Identification of chemical series or groups of common scaffolds is an important step in the lead optimization process, which currently involves a significant amount of human input. Although there are multiple methods available to generate scaffolds and calculate similarities, efforts are needed to quantitatively compare human-defined scaffolds and chemical series to scaffolds/series generated by automated computational methods. In this study [3], the authors compared scaffolds generated from four different approaches, including a hierarchical agglomerative clustering approach UPGMA, a variant of the Butina clustering approach, a scaffold network-based approach by Bemis and Murcko and a BRICS fragmentation rules approach with human-defined series data from 30 internal projects at Novartis (Basel, Switzerland). Authors found that the UPGMA method performed better than the rest and suggested that it can be used retrospectively for project analysis.

newlands
press

As part of their study, the authors defined chemical series using each of the four approaches and assigned molecules to the series using a substructure match method. In addition, the article also discusses the maximum common substructure method as an alternative to series definition, which can be used in absence of human-defined scaffold/series datasets. To set up quality criteria, a universal specificity score is defined that mainly reflects the probability of random molecule getting assigned to the scaffold. In addition, an ambiguity score is calculated that provides information on multi-assignment of molecules concerning the intra project specificity. To link automated series with human-defined series, authors calculated F1 score ($F1 = PR/2 [P + R]$ where P = precision and R = recall) for given automated series against all human-defined series and linked the human-defined series with highest score to the respective automated series. Authors compared the automated series against 30 internal projects and the best performing method was further studied in detail against the second dataset of CDK2 antagonists.

Authors found that clustering approaches achieved number of series similar to the number of human-defined series. In general, the clustering methods (i.e., UPGMA and Butina) had better ambiguity scores compared with the fragmentation-based methods. Authors found that 75% of series from the UPGMA method linked to the human series and achieved a recall of 90% and a precision of above 70%. With similar results obtained from public datasets, the authors concluded that potential chemical series similar to human intuition can be identified using the proposed UPGMA approach without any bias. This method can be applied retrospectively and could provide a great learning in moving projects forward. Authors have provided the python class for automated classification algorithm and Jupiter notebook for data preprocessing (CDK2 dataset) for further development.

- Written by R Chaudhari

Financial & competing interests disclosure

F Miljković is an employee of AstraZeneca. 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. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

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

Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>

References

1. Alves VM, Capuzzi SJ, Braga RC *et al.* SCAM Detective: accurate predictor of small, colloiddally-aggregating molecules. *J. Chem. Inf. Model.* 60(8), 4056–4063 (2020).
2. SCAM Detective web application. (2020). <https://scamdetective.mml.unc.edu/>
3. Kruger F, Fechner N, Stiefl N. Automated identification of chemical series: classifying like a medicinal chemist. *J. Chem. Inf. Model.* 60(6), 2888–2902 (2020).