ISGC 2019
May 13th - 17th

N°692 / OC
TOPIC(s) : Industrial chemistry / Networking and education

FINDING THE NEEDLE IN THE HAYSTACK - A SOFTWARE-BASED APPROACH TO DATA-DRIVEN COMPOUND DESIGN

AUTHORS
Dora BARNA / CHEMAXON KFT, ZÁHON YUTCA 7, BUDAPEST

PURPOSE OF THE ABSTRACT

Discovery of new compounds and formulations is an iterative process of rational hypothesis generation and testing accomplished via the synthesis and screening of new molecules. The design cycle is an information-intensive process where an array of computational services are utilized by the project team to assist making decisions regarding synthesis candidates. The required information includes physico-chemical predictions, 3D properties and modeling results, rapid freedom-to-operate analysis, available biological data, potential screening compounds and building blocks.

A key factor in this process is collecting all relevant and up-to-date pieces of information from the disparate sources in public and proprietary domains. The number and size of databases holding chemical information have been increasing rapidly in the last couple of years. Searching these databases one-by-one and keeping track of the results is a huge task in itself. Crucial information may be missed if teams are unaware of the availability of new data sources.

The aim of the presentation is to introduce a novel search solution that helps chemists and biologists to find relevant data during the molecule design process. We have developed a single search interface to access chemical information collected and indexed from a variety of sources. The resulting database merges compound information from public sources including PubChem, ChEMBL, BindingDB, eMolecules, MolPort, SureChEMBL and MCULE, and consists of roughly 120 million unique structures. Furthermore, the Enamine REAL data set that contains around 470 million molecules is also available for chemical search from the same interface. The current collection can be extended with other public and commercial chemical databases or with custom in-house structure sources. With this universal, domain-agnostic approach, a single search can tell you the complete research history of our compounds, collect existing assay results, let you know who is working on similar molecules in your organisation, help stay outside the patented space or find building blocks for synthesis or screening compounds together with availability and price information.

This search solution can be combined with additional components for the central management and tracking of innovative chemical ideas. In order to effectively coordinate chemical hypotheses and compound series in projects where several teams are collaborating, this platform dynamically accesses changing sets of information for use in the evaluation, prioritization and triaging of the ideas.

Besides predictive models and various databases, querying the scientific literature also provides key information for the design of chemical structures. However, finding the truly relevant papers is not an easy task. Chemical and biological named entity recognition can be of assistance when looking for specific information in journal articles. A case study will be presented which demonstrates how the above solutions can be extended with a search tool that can identify both chemical entities and biological concepts in the literature, leading to a more rational prioritization of resources in the discovery of new chemical compounds.

FIGURES

FIGURE 1

FIGURE 2

KEYWORDS

large chemical data | domain agnostic chemical search | molecule design platform | extracting chemical data

BIBLIOGRAPHY