VOLUME 3, ISSUE 3

MOLECULAR DATA ORGANISER

Using AI, Deep Learning & Content Swamp

AUTOMATION IN DRUG & MOLECULAR DATA EXTRACTION

AutoML
DeepChem

AUTO ORGANISER OF DATA USING CONTENT SWAMP

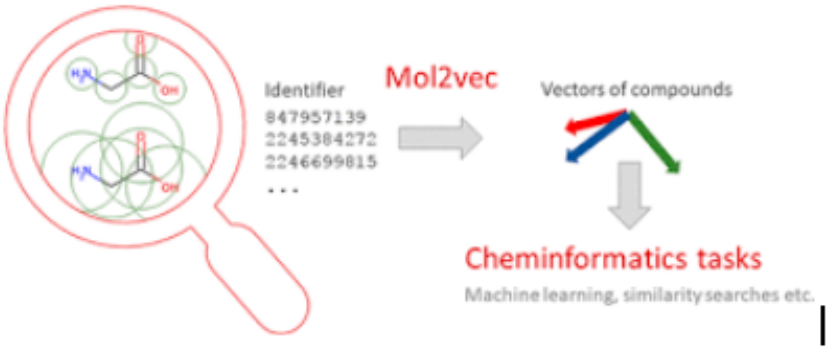
By iReadRx

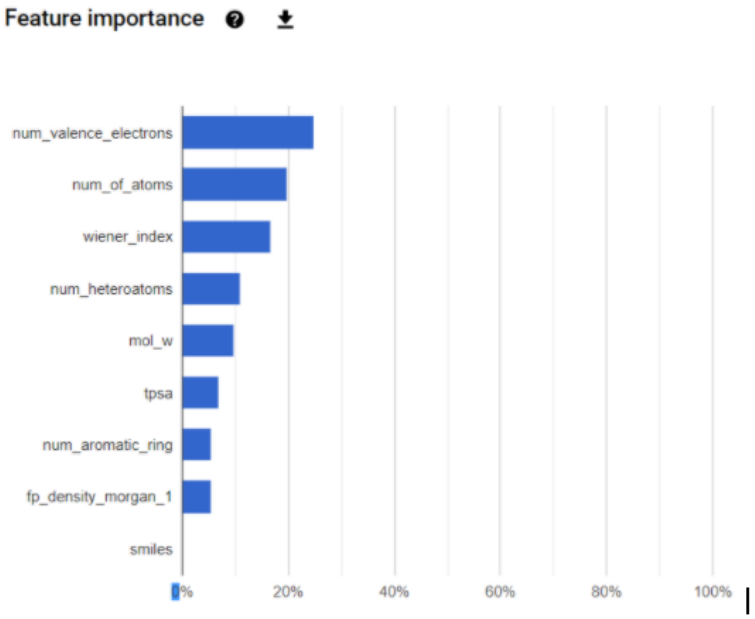
From scientific research papers and patent documents to data assets and journals, eighty percent of all enterprise pharma data is considered unstructured. This surge in digital documents and communications, otherwise known as Big Content, has completely shifted the paradigm of data collection and categorization because this information is the key to unlocking highly personalized experiences and extremely precise results to meet customer demands. That's where Content Lakes can come into play. A Content Lake is a specialized form of a Data Lake that stores language-based content and offers broad use access to original content as well as analytics, tabular data derivation and statistical processing.

iReadRx.ai

DATA EXTRACTION & RELEVANCY SCORE

Out of the various algorithms tried. Random Forest and AutoML are comparable. With the list of features. In both, data pre-processing is very fundamental. The compounds have multiple formats of representation such as IUPAC, SMILES, molecule and InChI etc. Even in each of these formats, the representation can differ and there is no standardisation. We have great sources of knowledge for this task. SureChemBL being one of them. Normalising the data and filling the data gaps are crucial in utilising all the data. Another important factor is **mol2vec**. Compounds can finally be encoded as vectors by summing the vectors of the individual substructures and, for instance, be fed into supervised machine learning approaches to predict compound properties. So, one of the important steps in building a solution for this problem is to normalise the data and convert them into molecule (image) structure.





Over the past few years, our team has worked with several clients to develop data lakes for storing enterprise-wide content. A data lake is a large storage repository that holds a vast amount of raw data in its native format until it is needed. Once the content is in the data lake, it can be searched, enriched, and used to generate insights to solve business problems and support diverse user needs.In a recent project for a pharmaceutical client, we tackled a different problem: ingesting over one petabyte (PB) of unstructured data into their data lake. To put this into perspective, according to Computer Weekly,

WWW.IREADRX.AI

Molecular Informatics
Cheminformatics
Deep Neural Networks