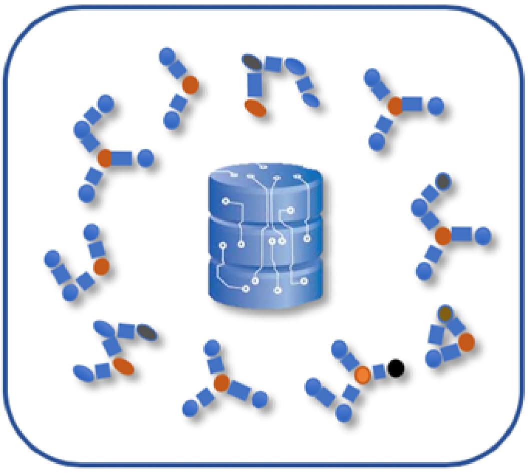
ConstruQt â€“ The Beginnings of the Chemical Data Revolution

Feb. 4, 2020 by Peter Jarowski

Chemical Data Has Problems

The state of data access, quality and dissemination in Chemistry is extremely poor - so poor that it is blocking advances in machine learning (ML) and artificial intelligence (AI), and also impeding research and development in traditional methods. The recent surge in AI skepticism is a direct consequence of years of over-hype and promises based on precarious data. Over-the-top expectation were offered without enough consideration for the data quality and volume required to train fancy algorithms. The old adage "^&$% in, ^&$% out" holds true (we can say 'crap' right?). This opinion is in line with recent statements by the CEO of Novartis, for example, who runs the second largest pharmaceutical company in the world, lamenting the difficulty in accessing quality datasets to make AI effective.



ChemAlive is building a platform to solve this problem by offering advanced computational tools in exchange for academic raw and live chemical data. ConstruQt is our first freemium web application module showing how this interaction works. It provides quantum mechanical molecular energetic and structural analysis designed for library scale (big data) research and simply asks that researchers deposit their chemical structures upon submission.

Chemical Data is Damn Valuable

Let’s set the context. “Data is the new gold” is a statement of increasing significance. Scientific data as an industry is at least 20 billion USD p.a. based on the size of the peer reviewed scientific publication market. About 5 billion of that is related to Chemistry. In fact, chemical data is the most valuable scientific data based on highest frequency of journal article piracy – in other words, it is data (recipes) worth stealing. Chemistry and Electronic Engineering (EE) account together for over 50% of all patents filed. The monetary value of patents in Chemistry, based on patent sale price, far exceeds EE. Companies distributing published chemical data like Scifinder, Reaxys are only aggregators of the hard-earned, high-value data of, mostly, academic researchers, but they also act as gatekeepers to organized chemical information, with quite an upside for themselves. Public initiatives like Pubchem and Chemspider do not have the same pizazz as their business counterparts, but are widely used to retrieve basic chemical information. Government initiatives like NIST can be important tools as well. However, there is a problem. None of these platforms (commercial or public) are built for big data analytics. Their business models or data architectures do not allow a non-sparse connection between large curated datasets of molecular properties connected directly to molecular structure freely searchable.

A Walk through the Chemical Data Freak Show

In 2016, we set out to build algorithms for ConstruQt, and discovered the extent of the data problem in chemistry. An early attempt to find acidity constants was particularly disappointing. The largest list (20,000 or so) is maintained in secret by a prominent software vendor, in collaboration with a prominent pharmaceutical company. Their prerogative. Public datasets of organic acids and bases are all over the place. We ultimately had to assemble our own list of about 5,000 values by hand. There is a company that makes its daily bread on selling a list of comparable size to data hungry medicinal chemists.

The problem was even worse when we looked for conformational energies to train and validate our conformational sub-routines. Thank you Crystallographic Open Database (COD) for at least a sizable set of known XRD conformers (we can have a separate discussion about the value of this data for solution phase molecular shape). True energetic information is non-existent (I hope to be proven wrong) and NIST was helpful in securing at least a few data points.

In general, good datasets linked to chemical structure in a usable format are rare. Public initiatives to solve this do not work. Recently, we wanted a list of organic dyes and their data. A simple google search revealed a Chemspider dataset of 150 molecules. This could be downloaded in SMILES format (good start) after registration. Was the spectral data uploaded to these entries and connected to each SMILES? Of course not. The reason is the heart of the problem – there are no researchers in this beautiful world with the extra time to manually upload spectra to Chemspider with their only motivation a desire to be a good chemical Samaritan. “Ain’t nobody got time for that!” Such initiatives are designed by academic minds without consideration for what really motivates people.

Building a Proper Chemical Open Data Platform

To get something, you need to give something. At ChemAlive we give state-of-the-art chemical predictive analytics based on quantum chemistry. With ConstruQt, we offer freemium tier quantum chemical calculations fully automated for all chemists connected to powerful computational infrastructure. The module performs automatic conformational, tautomeric and stereoisomeric enumeration of the input molecules (copy paste them as SMILES or draw them) and allows the user to choose to compute the entire molecular space or only a part of it using the fast semi-empirical PM6 method. Future integrations will allow DFT for highly engaged or paying users (that costs more money). The results are fully interactive allowing navigation of the structure with energy graphs and 3D structural display.

In exchange for this service, we want your meta data. That is our price. We are trying to monetize chemical information to eat and live. We are not Google or Facebook. We do not have ads, nor are we trying to influence you in any way. We want to connect chemical structure to intent. Who are you? What are you trying to accomplish? What kind of chemistry are you doing with what molecules? We are asking no more than what the publishing house is asking, except we will not block you from accessing your own data fi you can’t afford the subscription. In this way, we can accelerate industrial chemical research with high-value tagged datasets of relevant (fresh) molecules to inform their commercial applications. We believe in helping the chemical industry design better, more efficient, molecules, with a lower environmental and climate impact. We are the good guys (for the moment).

ConstruQt can process hundreds of molecules per minute matching your current high-throughput requirements and enhancing the utility in screening applications. It can scale to thousands of CPUs in minutes and uses inexpensive spot computing. ConstruQt will deploy calculations based only on SMILES list input and return quantum chemical data in an SD file ready for integration with standard cheminformatics and drug discovery platforms or viewable on our platform.

Please try it here by simply registering here: app.chemalive.com

Read here for more information: https://www.chemalive.com/construqt-api/

For larger libraries and advanced quantum mechanics methods for increased accuracy or advanced properties (spectra, charges, reactions), please get in touch at info@chemalive.com.

Happy quantum chemistry!

The ChemAlive Team

- ChemAlive SA