Chemchart Enterprise: an Intelligent Platform for Chemical Data Management

July 2, 2019 by TJ Bozada

Chemical data management is an important process to a number of industries, especially those engaged in manufacturing or research and development. Unfortunately, chemical data is as unwieldy to manage as it is important. This is for a variety of reasons, but the biggest contributing factor is the sheer amount of data available to the public and managed by an enterprise. For decades, chemical data has been recorded in paper, and then excel sheets, and now databases. While efforts have been made to homogenize, or at least centralize this data, there is not one single solution for indexing and searching the wide variety of data types, and sources, managed by an individual enterprise. Chemchart Enterprise changes this paradigm by combining the flexibility of big data with the precision of machine learning, providing a single solution for managing the entire organization chemical space.

ToxTrack Inc, a developer of cheminformatics software, just released its new tool Chemchart Enterprise -- a machine learning (ML) based platform for managing chemical data. Chemchart Enterprise provides a single solution for managing the entire organizational chemical space, combining database management, document processing, and chemical exploration into an intuitive interface.

Built for expansion, Chemchart Enterprise's internal database can extend to new types of data, while facilitating semantic and structural queries, and providing tools to compare collections of chemicals. Chemchart Enterprise also has the ability to identify and extract chemical references in documents. The platform scans internal document collections (e.g. PDFs, spreadsheets, Word Docs, SDFs) and automatically indexes chemicals into the central database. In addition to organizing internal document repositories, Chemchart Enterprise can join internal data with external datasets like patents, regulatory, or news repositories.

A useful feature in Chemchart Enterprise is being able to get alerts whenever a chemical in the customer's supply chain lands on a new regulatory document, or finds a new use in a patent.

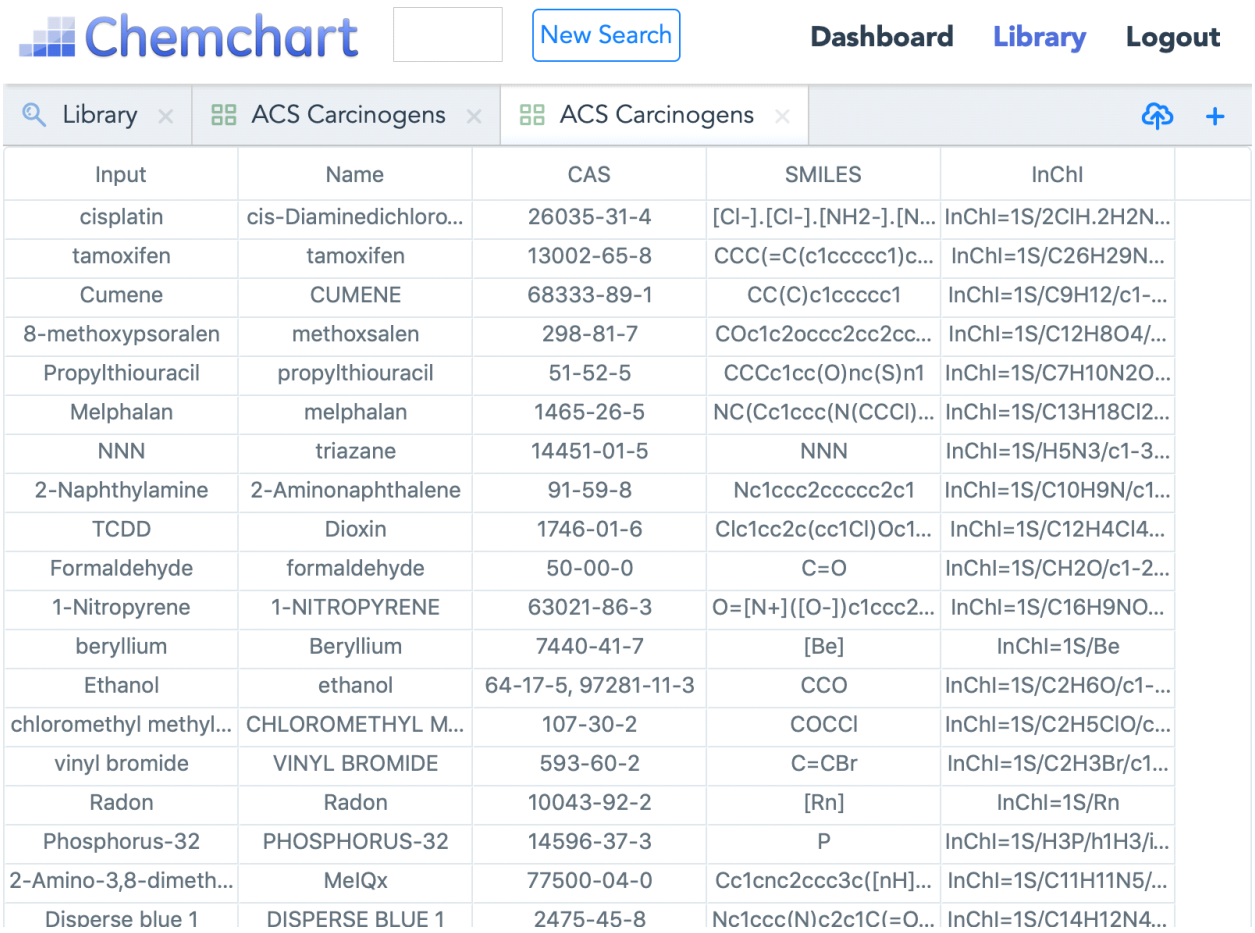
Data management options

The core of Chemchart Enterprise is a chemical database indexed by structure. Each structure is tied to a variety of meta-data, including identifiers, physicochemical properties, and toxicity data. The database

can also be expanded to include new sorts of data, whether it be the results of an assay or patent numbers associated with that chemical.

Indexing by structure gives Chemchart Enterprise the foundation for substantial flexibility. For Chemchart Enterprise, flexibility means predictive technologies. Predictive models can be incorporated into this software, using the underlying database as training data to predict for data gaps.

Recognizing that research is the collective work of many researchers, Chemchart Enterprise is a multi-user application with both easily approachable and technical user interfaces. Non-technical users can access an innovative, spreadsheet-like web app to view and contribute to the database. This makes accessing and updating data as simple as using Excel. For users with more technical applications, like building machine learning models, there are standardized application programming interfaces (APIs) with read/write access to the database.



Input	Name	CAS	SMILES	InChI
cisplatin	cis-Diaminedichloro...	26035-31-4	[Cl-].[Cl-].[NH2-].[N...	InChI=1S/2ClH.2H2N...
tamoxifen	tamoxifen	13002-65-8	CCC(=C(c1ccccc1)c...	InChI=1S/C26H29N...
Cumene	CUMENE	68333-89-1	CC(C)c1ccccc1	InChI=1S/C9H12/c1-...
8-methoxypsoralen	methoxsalen	298-81-7	COc1c2occc2cc2cc...	InChI=1S/C12H8O4/...
Propylthiouracil	propylthiouracil	51-52-5	CCCc1cc(O)nc(S)n1	InChI=1S/C7H10N2O...
Melphalan	melphalan	1465-26-5	NC(Cc1ccc(N(CCCl)...	InChI=1S/C13H18Cl2...
NNN	triazane	14451-01-5	NNN	InChI=1S/H5N3/c1-3...
2-Naphthylamine	2-Aminonaphthalene	91-59-8	Nc1ccc2ccccc2c1	InChI=1S/C10H9N/c1...
TCDD	Dioxin	1746-01-6	Clc1cc2c(cc1Cl)Oc1...	InChI=1S/C12H4Cl4...
Formaldehyde	formaldehyde	50-00-0	C=O	InChI=1S/CH2O/c1-2...
1-Nitropyrene	1-NITROPYRENE	63021-86-3	O=[N+]([O-])c1ccc2...	InChI=1S/C16H9NO...
beryllium	Beryllium	7440-41-7	[Be]	InChI=1S/Be
Ethanol	ethanol	64-17-5, 97281-11-3	CCO	InChI=1S/C2H6O/c1-...
chloromethyl methyl...	CHLOROMETHYL M...	107-30-2	COCCl	InChI=1S/C2H5ClO/c...
vinyl bromide	VINYL BROMIDE	593-60-2	C=CBr	InChI=1S/C2H3Br/c1...
Radon	Radon	10043-92-2	[Rn]	InChI=1S/Rn
Phosphorus-32	PHOSPHORUS-32	14596-37-3	P	InChI=1S/H3P/h1H3/i...
2-Amino-3,8-dimeth...	MeIQx	77500-04-0	Cc1cnc2ccc3c([nH]...	InChI=1S/C11H11N5/...
Disperse blue 1	DISPERSE BLUE 1	2475-45-8	Nc1ccc(N)c2c1C(=O...	InChI=1S/C14H12N4...

Data navigation options

Chemchart Enterprise repurposes the functionality of the publicly available Chemchart.com – ToxTrack's original chemical search engine. This search engine combines structural similarity and natural language processing (NLP) algorithms. On top of the familiar structure and sub-structure searches, users are able to search by percent similarity – that is, the chemicals that are a certain percent similar to the queried structure. This ability is powerful when combined with Chemchart Enterprise's other search modality: natural language.

Chemchart Enterprise, having been trained on tens of millions of journals articles, titles, and abstracts, utilizes natural language models to facilitate semantic queries. Simply put, the models have learned associations between words, as well as the ability to identify chemicals in text. When users enter a keyword, Chemchart Enterprise returns chemicals associated with this keyword. As the models first learned associations between words, they are automatically able to process 'new' keywords and match to novel chemicals introduced into the database.

Document extraction features

One of the greatest obstacles in enterprise chemical data management is the sheer amount of inaccessible data – specifically data that exists in unstructured documents. Using the same natural language models, Chemchart Enterprise is able to identify chemicals in text. Each chemical reference is mapped to a structure within the central database and the location of each reference is logged --creating a supporting document index. Users can search for documents or chemicals using all of Chemchart Enterprise's search functionalities. Chemchart Enterprise can also extract and log references from external data sources, including online sources.

"Collections" feature

One way to view Chemchart Enterprise's semantic search is as an ontology, whereby users are able to search by any combination of keywords and classify the result as a unit. In Chemchart Enterprise, these units are called *Collections*. *Collections* allow users to quickly compare and analyze different chemistries

Collections can be created in a variety of ways. First and foremost, *Collections* can be created by manually entering identifiers (name, CAS, SMILES, or InChi) into the simple spreadsheet like web-app. *Collections* can also be created by saving the results of any search query. Finally, *Collections* are

automatically created when Chemchart Enterprise scans internal or external sources.

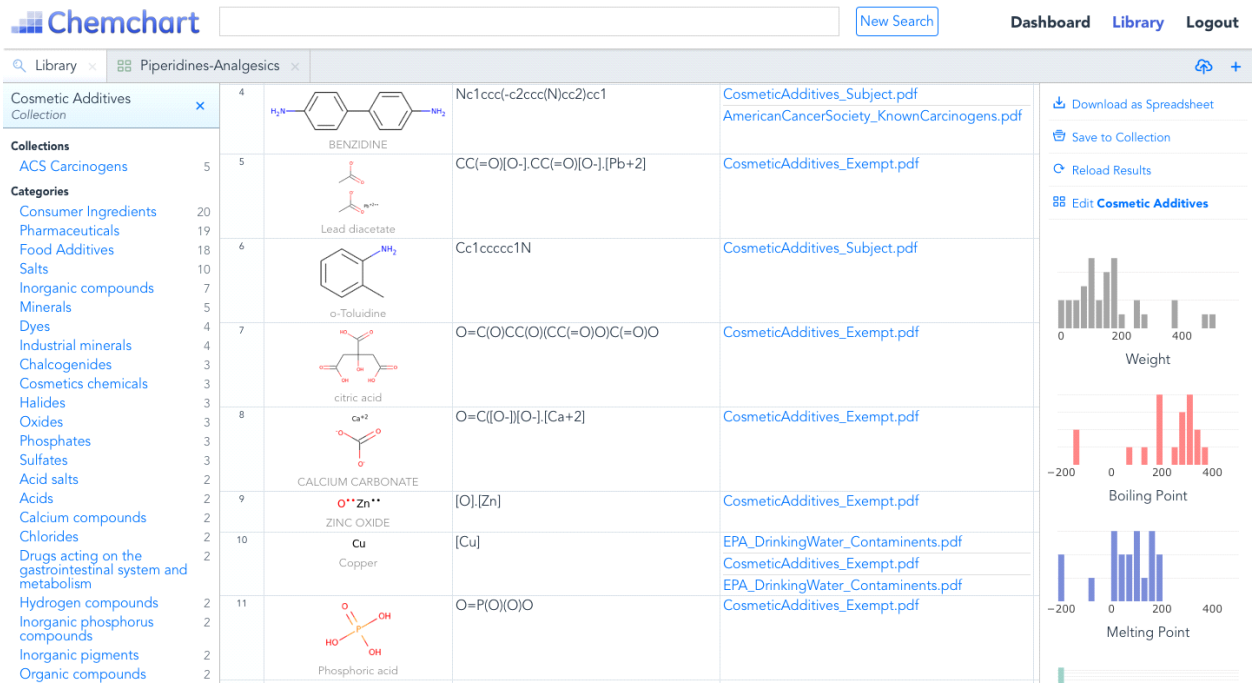
Patent and regulatory monitoring

Organizations spend a lot of time and money assuring their company's compliance with a variety of regulations. Chemchart Enterprise can provide automatic alerts to changes in the regulatory landscape based on specific portfolios. First, Chemchart Enterprise learns the organizational chemical space by processing ingredient and component lists. Once the portfolio is known, Chemchart Enterprise can monitor external data sources for chemicals of interest. Common data sources include patents and regulatory guidelines.

Chemical exploration

When exploring a new chemical space, it can be difficult to balance all the different criteria. This can be especially true for the industry's more rapidly changing landscape. Chemchart Enterprise optimizes the exploration process through the tracking of these many factors. Imagine a cosmetic company that is evaluating potential alternatives to a specific ingredient. They begin the exploration by searching for chemicals that are structurally similar to the original ingredient. Depending on the similarity threshold, the results may be quite numerous and so they need to refine the search.

As this cosmetic company understands the value of perception, they also want to remove any chemicals associated with pesticides, as well as any expected carcinogens or water contaminants. The company is able to remove unwanted chemicals by excluding the relevant semantic searches. Next, there are a number of patents filed by competitors with chemicals that cannot be used. These documents are uploaded into Chemchart Enterprise, the patented chemicals identified, and relevant chemicals are then excluded from the search result. This process can be repeated with federally regulated chemicals. In these ways, companies can combine multiple data sources and search queries to facilitate chemical exploration.



Chemchart Enterprise is an intuitive and flexible solution to chemical data management. Providing a centralized database, Chemchart Enterprise consolidates internal and external data source into an easily navigable interface. The combination of big data and machine learning means a robust solution capable of quickly extending into new chemistries. The result is the unique ability to identify, monitor, analyze distinct chemical spaces in one centralized platform.