MassKit: A Flexible and High Performance API for Mass Spectrometry

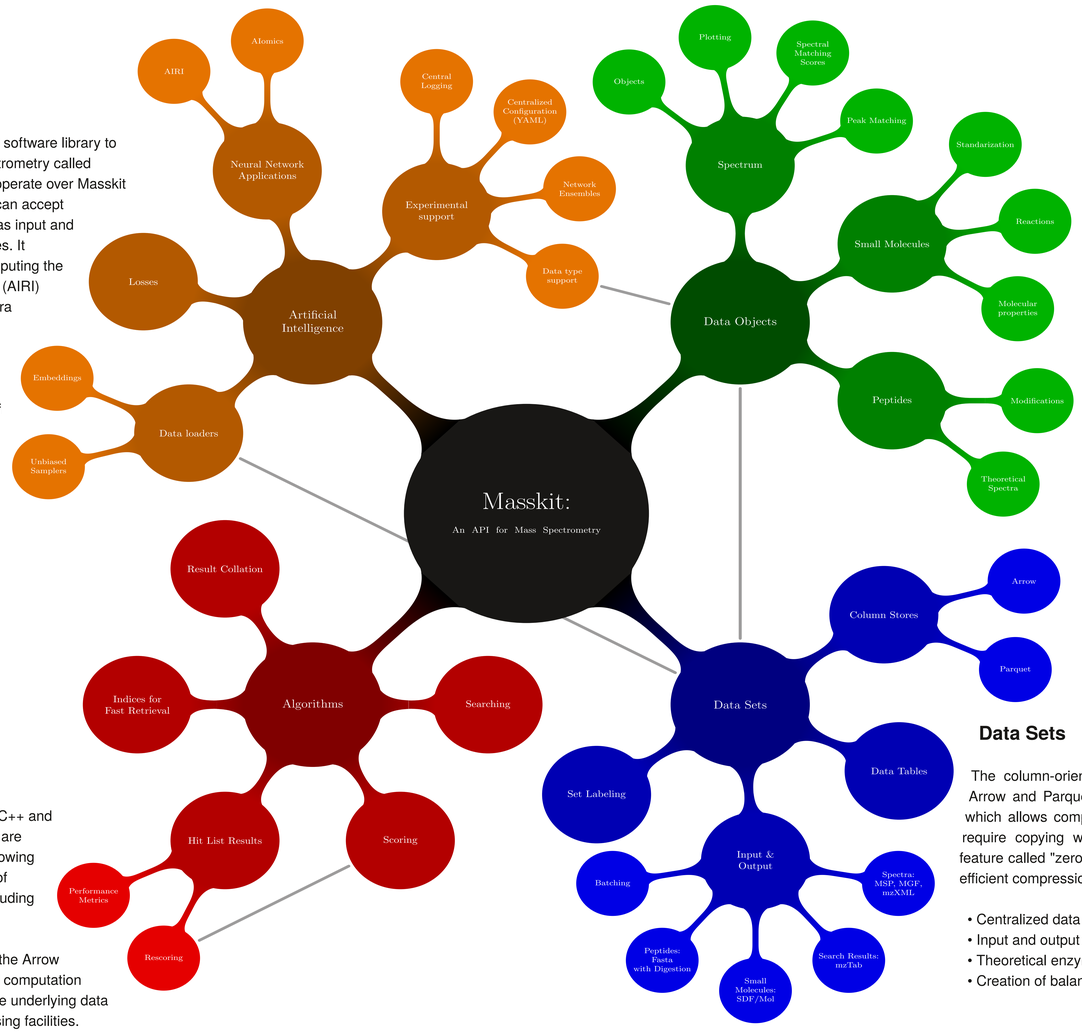
Douglas J. Slotta, Lewis Y. Geer. Mass Spectrometry Data Center, National Institute of Standards and Technology, Gaithersburg, MD, USA

The Masskit API

The number of possible molecules generated in biological and natural processes can be astronomical, leading experimentalists to cast a wide net when using mass spectrometry to identify these molecules. This approach leads to the brute force combinatorial generation of overly simplified theoretical spectra for matching, where the search space omits viable candidate molecules while admitting those that are not. However, the development of high quality spectral libraries, AI for backfilling those libraries, and increased knowledge of the chemistry and biology of experimentation hold the possibility of sensitive searching of large, well-defined chemical spaces. We have designed the MassKit software library using column-oriented data structures with the goal of allowing this type of flexible, high performance searching.

Artificial Intelligence

We use the functionality in the Masskit software library to create an AI framework for mass spectrometry called Masskit AI. This software library can operate over Masskit datasets to train neural networks that can accept spectra, small molecules, or peptides as input and output spectra and molecular properties. It includes published neural nets for computing the retention index of molecular structures (AIRI) and for computing tandem mass spectra from peptide sequences (Alomics). To accomplish this, Masskit AI contains code for losses, dataloaders, embeddings, and the training and prediction of singleton or ensembles of neural networks.



Data Objects

The foundation of the Masskit software library for mass spectrometry are the experimental data objects it operates on: objects for the manipulation of spectra, molecular structures, and peptide sequences. These data objects can then be organized into datasets and these datasets can then be labeled, indexed and searched.

Data Sets

The column-oriented storage substrate is based on the Apache Arrow and Parquet projects. Arrow is an in-memory column store which allows complex nested columnar data structures that do not require copying when being passed from process to process, a feature called "zero copy". Parquet is designed for data storage using efficient compression and encodings.

- Centralized data specifications using columnar stores.
- Input and output for many common mass spec data file formats.
- Theoretical enzymatic digestion of protein sequences.
- Creation of balanced training, validation and test sets.

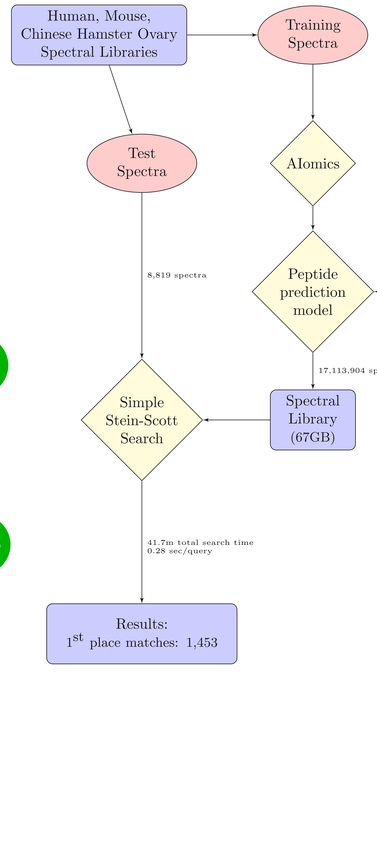
Algorithms

MassKit is open source and written in C++ and Python. The software libraries created are cross platform and cross language, allowing the use of these stores from a variety of operating systems and languages, including existing codebases.

Custom operators were created using the Arrow framework for searching, retrieval, and computation on spectra with zero copy access to the underlying data stores using their existing multiprocessing facilities.

Case Study: Search Algorithm Development

Searching massive AI generated spectral libraries.



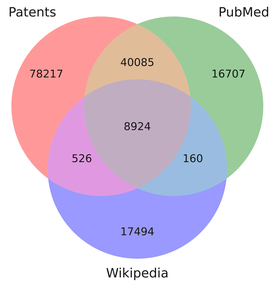
In our paper, *AIomics: Exploring More of the Proteome Using Mass Spectral Libraries Extended by Artificial Intelligence*[†], we discuss using Machine Learning to predict peptide spectra. The test and training sets for the model developed in that paper were derived from NIST high-resolution mass spectral libraries, created using Orbitrap HCD spectra from human, mouse, and Chinese hamster ovary cells (cell lines CHO-K1, CHO-S, CHO-DG44). The libraries include several modifications: carbamidomethyl cysteine, methionine oxidation, glutamine to pyro-glutamic acid, glutamic acid to pyro-glutamic acid, N-terminal acetylation, N-terminal pyro-carbamidomethyl cysteine, serine/threonine/tyrosine phosphorylation, and TMT6plex.

A script was used to generate a set of tryptic peptides with permutations of possible PTMs. These peptides were fed into the AIomics model to generate a spectral library. The library contains 17,113,904 spectra, stored in 6 parquet files totalling 67GB of storage. The 8,819 spectra with known peptides from the original CHO test set were searched against this predicted set.

[†] https://doi.org/10.1021/acs.jproteome.2c00807

Case Study: Experiments in Notoriety

We started an experiment to determine if a compound's notoriety could be useful when discriminating between closely scored matches. Data linking compounds to articles in PubMed, patents recorded at the US Patent Office, and Wikipedia entries were gathered from PubChem. For the Wikipedia entries we used the Wikimedia REST API to gather the average monthly pageviews for each entry since July 1, 2015.



In the combined NIST EI and Tandem libraries there are approximately 390K unique InChI keys, and 162,113 of them are linked to at least one of the three resources by PubChem. The Venn diagram shows the magnitude of the overlaps between the sets of compounds linked to each resource.

Table 1 shows the top ten compounds for each resource, as ranked by their respective counts. Obviously they do not entirely agree as to the order since they serve different audiences. The next step will be to determine which, if any, of these lists should be used, and how to integrate the results into spectral libraries.

PubMed		US Patent Office		Wikipedia	
Articles	Compound	Patents	Compound	Pageviews	Compound
233,597	Ethanol	680,997	D-sorbose	204,012	Fentanyl
193,424	Cholesterol	638,068	Hexitol	179,887	3,4-Methylenedioxymethamphetamine
190,208	Water	592,031	Mannitol	179,499	Cocaine
178,975	Oxygen	525,252	Benzoic Acid	163,931	Methamphetamine
178,554	Mannose, D-	509,079	DL-Arginine	163,424	Tramadol
127,007	Epicholesterol	464,982	p-Toluenesulfonic acid	152,967	Gabapentin
116,448	Adenosine-5'-triphosphate	461,797	Mannose, D-	148,085	Lysergide
109,502	Tyrosine	426,432	Elaidic Acid	145,589	Water
107,112	Dopamine	418,795	Histidine	142,300	Alprazolam
103,262	Serotonin	414,130	DL-Histidine	135,789	Diazepam

Table 1: The top ten compounds from each resource, ranked by their number of entries.

The authors declare no competing financial interest.