Contact Information:

Name of principal investigator: [redacted]
E [redacted]:
Phone: [redacted]

Affiliation:Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Prague, Czech Republic

---

Background information:
Please attach a short Biography of the Principle Investigator (s) and references to relevant publications

PI's CV attached in a separate file.

Relevant publication:
1. Brungs C, Schmid R, Heuckeroth S, Mazumdar A, Drexler M, Šácha P, et al. Efficient generation of open multi-stage fragmentation mass spectral libraries. *ChemRxiv 2024*; https://doi.org/10.26434/chemrxiv-2024-l1tqh-v2

Project name:

MERLIN – Mass spEctRal Library Network: Generation of the biggest open multi-stage fragmentation mass spectral libraries

Biological rationale (target):
If proposing a target based approach please describe the Target mechanism of interest
Not applicable.
Clarification: We don't target any molecule or (bio)chemical pathway.

Biological rationale (phenotype):
If proposing a Phenotypic approach, please describe the Primary cellular assay system and its relationship to the disease or cellular mechanism/pathway of interest
Not applicable.
Clarification: We are not proposing a phenotypic approach.

Novelty and impact (limitations):
of proposed project
Our goal is to collect comprehensive reference datasets of multi-stage tandem mass spectra of small molecules. However, in practice some molecules are difficult to detect with our mass spectrometry instrumentation. This can be caused by their molecular weight being out of target range, by inability to ionize the compounds using electrospray ionization, or by instability of the compounds. Based on our preliminary results (*ChemRxiv 2024*; https://doi.org/10.26434/chemrxiv-2024-l1tqh-v2) we estimate that about 90% of screened molecules can be successfully measured.

Novelty and impact (previous):
Untargeted analysis based on high-resolution mass spectrometry is a key tool in clinical metabolomics, natural product discovery, and exposomics. The major bottleneck of untargeted mass spectrometry analysis limiting its full exploitation is the identification of detected compounds. Spectral library matching is a confident way for compound annotation.

Novelty and impact (reason):

The number of compounds covered by current open MS repositories is fairly small. The most common open databases are MassBankEU, MassBank NorthAmerica, and GNPS. All of them are including less than unique 30,000 structures. Commercial libraries have marginally more compounds (NIST and mzCloud), but these data cannot be downloaded and used for machine learning training or other tool development. These tools are needed for confident annotation in untargeted metabolomics workflows. Nowadays, training data are limited to the compounds included in public databases, resulting in a poor coverage of the known chemical space.

We have developed a high-throughput method for acquiring MSn trees and an automated workflow for generating open MSn libraries (described in the above-mentioned publication). This workflow will be used for the continuous extension of the open MSn library, a unique resource for the community, that can be subsequently used for the research of a variety of (bio)chemical processes. The novelty/peerless of the library generated by our team lies in its free availability to the researchers and by largest publicly available collection of high-quality labeled MS/MS spectra.

Technical feasibility*:

The technical solution of this project is described in the manuscript entitled „Efficient generation of open multi-stage fragmentation mass spectral libraries" deposited in ChemRxiv. under doi:10.26434/chemrxiv-2024-l1tqh-v2.

Technical feasibility (pilot):

For the first generation of our MsnLib we acquired data for 30,000 unique compounds in 23 days. Since access to compound libraries is limited and costly, Openscreen provides a great resource with adding new compounds to the mass spectral library and increasing the chemical space coverage with new structures.

Technical feasibility (proposed format):

Assay-ready plates with compounds will be delivered by CZ-OS to be measured using the data acquisition pipeline established at IOCB (*ChemRxiv 2024*; https://doi.org/10.26434/chemrxiv-2024-l1tqh-v2). The compounds final concentration 20 uM will be obtain by dilution with MeOH/H$_2$0 or MeOH/ACN/H$_2$O.

Technical feasibility (secondary):

Please describe briefly the Secondary and Selectivity assays (needed to validate any Hits emerging from a Primary screen) and their status of development

Not applicable.

Plans for post screening stages*:

Data will be shared with the MS community for free and also uploaded to ECBD database. All generated spectral libraries will be published in MGF and json formats in public access data repositories (e.g., Zenodo). The collected raw MS spectra will be published in the open mzML format in a public mass spectrometry data repository (e.g., MassIVE).

Funding status*:

If funding is not in place, please identify potential funding opportunities and the format, timescale and anticipated likelihood of success when seeking funding

Funding is available from the ERC Consolidator Grant of the PI to cover the ECBL access fee and the costs of consumables for the open-access collaboration.

Confirmation of willingness to release screening results available into the public domain once a reasonable time to allow secure IP has passed*:

CZ-OPENSCREEN uses public funding to provide Scientists access to a high-quality library of compounds to allow identification of chemical probes. As part of its public service mission, data generated in CZ-OPENSCREEN projects will be deposited in our database. Data deposition is expected to occur no more than 36 months following completion of a project. (The data will be also deposited in the European Chemical Biology Database after project owners have had a chance to secure IP on the results.)

We are supporting the FAIR science strategy and uploading all generated data on public repositories.

YES