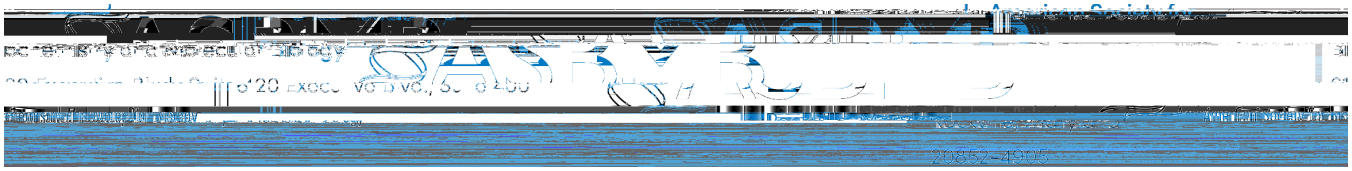


January 5, 2023

Kristine Willis, Ph.D.
Division of Cancer Biology
National Cancer Institute
9609 Medical Center Drive
Building 9609 MSC 9760
Bethesda, MD 20892

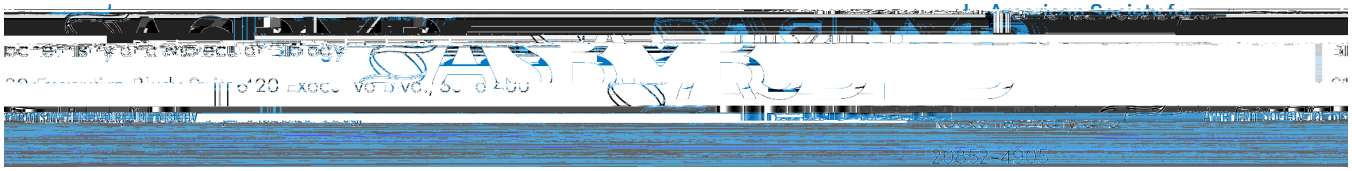
RE: National Cancer Institute s Request for Information on Soliciting Input on the Use and Reuse of Cancer Metabolomics Data

The American Society for Biochemistry and Molecular Biology is an international nonprofit scientific and educational organization that represents more than 10,000 students, researchers, educators and industry professionals. The ASBMB strongly advocates for strengthening the science, technology, engineering and mathematics workforce, supporting sustainable funding for the American research enterprise, and ensuring diversity, equity and inclusion in STEM.

The National Cancer Institute (NCI) published a request for information titled, Soliciting Input on the Use and Reuse of Cancer Metabolomics Data on Oct 19. The NCI aims to understand how to support privacy, reproducibility and harmonization in alignment with the new National Institutes of Health Data Management and Sharing Policy

The ASBMB shares the concern of many scientists that the new NIH data-management and -sharing policy has the potential to place significant burden on individual scientists, laboratories and core facilites that collect large, complex datasets. Because metabolomics is one such field producing highly diverse, complex datasets, the ASBMB recommends that the NIH issue more guidance on what level of data and information is required to be compliant. Importantly, these new clarifications also must remain sufficiently flexible to accommodate the diverse methods of collection and their individual technical limitations and/or assumptions.

Due to the high complexity and expert knowledge required to analyze and assess metabolomics data, the majority of the experimental data available in metabolomic repositories is not useful to the public or most scientists outside



the field; nor is it particularly easy to reuse by those within the field. This potentially high burden cost to investigators without much utility poses significant barriers for these scientists to comply with the new NIH data-management and -sharing policy.

Recommendation 2: NIH and NCI should continue to improve the deposition and retrieval process of repositories
The currently available software and tools for deposition and retrieval of metabolic data are cumbersome. In fact, metabolomics researchers are reluctant to extensively deposit their data into a repository due to the onerous effort required and lack of clear utility for the deposited data. To improve the deposition and retrieval process, we recommend that NIH and NCI ensure that repositories, such as the NIH Common Fund's National Metabolomics Data Repository, have the following attributes:

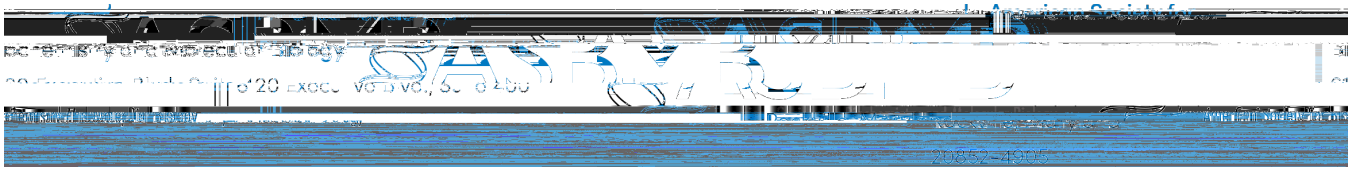
(1) streamlined to minimize the burden of deposition and protect scientists valuable time and effort
(2) updated to be compatible with stabile isotope tracer datasets
(3) regulated to require only the data and metadata necessary to comply with the policy in a format that supports sustainability in a rapidly evolving field (i.e., data on some file types from more than a decade ago are already inaccessible)
(4) structured to be sufficiently flexible so as to accommodate new technologies in the field and incorporate new functionalities with ease
(5) Embedded with thorough instructions on how to properly retrieve data to ensure they are correctly processed and analyzable by nonexperts

Recommendation 3: NCI should factor in the experimental challenges of metabolomics data collection as they move toward the goal of reuse
Experimental variations contribute to a lot of uncertainty in reusing metabolic data. The individual instrumentation, chromatography columns and sample preparation methods that are utilized will produce unique spectra and must be standardized within an experiment. It is extraordinarily difficult to standardize across the whole metabolomics field. For example, standardization committees coordinated through NIH (e.g., mQACC) currently are addressing standards for LC-MS only. Additionally, the metabolic content from cell extracts can (1) vary based on extraction method (which varies significantly across the field and biases the number of recoverable metabolites) and (2) rapidly change during sample preparation, potentially skewing the data.

Recommendation 4: General pathway software tools need improvement
Another area of concern for metabolomics researchers is the use of metabolic pathway analysis software. These tools can be an excellent starting point but are highly reductive and can lead to significant misinterpretations. Because metabolism varies by tissue type and organism, the results of general pathway software can be highly inaccurate. It must be clearly communicated to users that these tools generate hypothetical outputs that must be validated and not taken as evidence.

Recommendation 5: NIH must establish clear nomenclature for metabolites
The lack of clarity regarding chemical and metabolite nomenclature is another barrier to the use and reuse of metabolomics data. There still remains some debate in the field as to what is considered a metabolite. For example, is a protein or nucleic acid a metabolite? Futhermore, there are considerations around exogenous versus endogenous and interorganismal transformations. The NIH should provide a clear definition for what it considers a metabolite.

Additionally, there are several different standardized formats used to identify and distinguish one chemical from another, e.g., InChIKey, SMILES, PubChem, ChemSpider, CHEBI and several others. The lack of consistency in chemical names can create confusion and difficulty in communicating and reusing data. The ASBMB encourages more standardization of chemical naming in a manner that works for metabolomics as well as across scientific fields. In metabolomics, InChIKey and SMILES were reported as the current front-runners, but they are not fully



compatible with tracer studies. As progress is made on this barrier, the NCI and NIH should prioritize interoperability between format types.

The data and metadata most necessary to reproduce results reported by metabolomics studies

Recommendation 6: Require metadata for metabolomics data deposition
Metadata are the information necessary to understand the context of experimental data, such as experimental design, sample preparation and equipment details. Without this information, metabolomics data loses significant value and reduces confidence and reproducibility. Alarmingly, certain metabolomics repositories do not require deposition of this information. In contrast, amending raw data with metadata has the potential to be an inordinately burdensome process. The ASBMB recommends requiring a reasonable degree of metadata that is standardized in format and interoperable with international standards.

Due to the complexity of some datasets, especially multiomics ones, the ASBMB recognizes that reporting the metadata in a consistent and retrievable manner will be quite challenging and continued engagement from stakeholders will be critical.

Recommendation 7: The NIH and NCI should carefully balance the necessity of data and metadata for validity with the utility and the burden cost to researchers.
The minimum metadata required to reproduce a metabolomics experiment would include information on the instrument platform and settings, chromatography columns used and mobile phase program (if used), and experimental design (including sample treatment, drugs/inhibitors, tracer description, solute preparation, media information and internal standards, when applicable).

The data that are most necessary for reproducing a metabolomic study would of course vary depending on what type of data, e.g., mass spectrometry data or nuclear magnetic resonance (NMR) spectroscopy data. In the case of mass spectrometry, the essential data would include both MS1 and MS2 spectra, indication of data-dependent or data-independent collection, the area under the curve, its transpose and all identifiers and/or annotations. For NMR, the information provided should include the observed spectra and/or the chemical shift data, the peak areas, coupling patterns for all observed nuclei, and concentration of the internal reference standard, which can be used to identify and quantify a metabolite.

The researchers to whom the ASBMB spoke were divided on whether or not to include unknown peaks in metabolomics datasets. The chemocentric approach providing data on the only metabolites that can be identified at a certain level of confidence would be the most practical while also making the data easier to reuse and view with respect to the biological significance. Alternatively, the umbrella approach of depositing all spectra, regardless of confidence in identity, has potential utility in the future as technology and computational methods become more sophisticated and new breakthroughs occur. The ASBMB recommends that NIH thoroughly engage the community of stakeholders and experts in metabolomics on this topic to thoughtfully determine the best course of action.

Considerations for selecting and using software and informatics tools for metabolomics

The researchers ASBMB consulted did not overwhelmingly consider any software to be the standard, in fact, they shared that many labs, core facilities and industrial providers have proprietary software.