Summary of Presentations and the Panel Discussions

Disclaimer: These presentations reflect the views of the presenters and should not be construed to represent the agencies' views or policies. The findings and conclusions in these presentations are those of the authors. Mention of trade names or commercial products in the publications is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the agencies.

Chemical literature integration in PubChem

(https://doi.org/10.6084/m9.figshare.26940295.v1)

Jian Zhang, Evan Bolton

National Center for Biotechnology Information, National Library of Medicine, NIH

PubChem, an open chemistry database funded by the U.S. government, has been a valuable resource for the community for nearly two decades. It houses a vast amount of information, including chemical structures, properties, spectral data, chemical safety data, toxicity information, patents, relevant bioactivity data, and literature. PubChem actively integrates chemical literature data from a variety of sources, including esteemed publishers like Springer Nature, Thieme, Wiley, and Nature journals, as well as substance record contributors, annotation references, NLM-curated PubMed citations, and literature with co-occurrence references. This presentation will delve into the integration, consolidation, and retrieval of chemical literature data within PubChem.

The Environmental Protection Agency's (EPA) Substance Registry Services:

https://cdxapps.epa.gov/oms-substance-registry-services/search

Akshay Narang

Office of Mission Support, U.S. EPA

The Substance Registry Services (SRS) is the Environmental Protection Agency's (EPA's) authoritative resource for information about chemicals, biological organisms, and other substances tracked or regulated throughout the U.S. EPA. Different program offices may use different names for the same regulated substance. For example, there are eight different names for Lindane across EPA regulations. SRS was built to show the relations among the synonyms to help the tracking and reporting for both EPA and industry. SRS may be accessed at: EPA.gov/SRS.

Status of the IUPAC InChI project
Steve Heller

National Center for Biotechnology Information, National Library of Medicine, NIH

This presentation describes and updates the IUPAC chemical structure standard project – InChI. InChI is the International Chemical Identifier. InChI is an open source, freely available unique identifier for chemical compounds that allows the precise identification of any chemical compound. InChI's goal is to provide a sustainable standard that enables connections in chemistry for the advancement of science and medicine for the public benefit. The InChI algorithm was initially developed at NIST and then given over to IUPAC to maintain and expanding it capabilities. InChI is a freely available operational software program that supports the FAIR data principles, specifically in the F (Findability) and I (Interoperability) areas, which are critical to enable effective and efficient communication of scientific content. The technical aspects of the InChI project and its development and expansion to handle more areas of chemical structures, has gone very well over the past years. However, the financial support, and hence the sustainability of the project, has not kept up. The technical work has been contributed by many organizations around world. This is not the case for the increasingly complex testing and validation of new versions of InChI. Also, most organizations only think of InChI as "free" and there is very little interest in supporting its "sustainability" which has gotten to a critical point at this time. Without a considerable input of long-term funds, there will be fewer staff beginning in 2024 and the project will start shrinking or declining in quality or both. As this abstract was being prepared a new possible source to provide the needed sustainability is seriously looking into possible support and a status of this will be presented at this FDA meeting.

The National Institute for Allergy and Infectious Diseases' Anti-HIV/OI/TB Therapeutics Database: ChemDB
Margaret Rush, Louise Sumner, Mohamed Nasr

National Institute of Allergy and Infectious Diseases, NIH

Abstract was not provided.

Dissemination of model metadata via the WebTEST2.0 platform
Todd Martin

Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. EPA

WebTEST2.0 (Web-based Toxicity Estimation Software Tool, v2.0) is a web-based platform for real-time predictions using QSAR (quantitative structure activity relationship) models stored in a PostgreSQL database. For each model, WebTEST2.0 provides metadata via a QMRF (QSAR

Model Reporting Format) document and a summary spreadsheet. In addition, WebTEST2.0 provides a HTML report for each real-time prediction. The QMRF document gives a detailed description of the model and its associated dataset. The summary spreadsheet supplements the QMRF document by providing the original experimental data with source metadata, the training and test sets, model statistics, model predictions and plots, and molecular descriptor values. The statistics sheet includes an array of training set, test set, and applicability domain statistics, which facilitates the comparison of models. The HTML report provides the experimental and predicted values for the query chemical, an applicability domain analysis, and prediction results for similar chemicals from the training and test sets. In conclusion, WebTEST2.0 gives users a complete array of model metadata relevant for regulatory applications.

Knowledge sources and approaches to guide analysis and interpretation of metabolomics (and other omics) data
Ewy Mathe

National Center for Advancing Translational Sciences, NIH

Metabolomic and other omic data are being generated at an incredibly rapid pace that is not expected to slow down given large NIH, ARPA-H, and other programs that are developing and underway. Although clear processes for sharing these data are needed to enhance data reuse and our ability to build on existing knowledge, these processes are often unclear or not applied globally in the metabolomics and multi-omics field. In turn, building knowledge sources and associated tools that are comprehensive, up-to-date, accurate, and FAIR is challenging. This talk will discuss several NCATS DPI knowledge sources and tools (RaMP-DB, MetLinkR, and GSRS/Inxight), the gaps they fill, and challenges encountered in building them and using them for the analysis and interpretation of multi-omic and other translational research data. The balance between generalizability and specificity of resources, tools, and datasets will be discussed.

SEND Sanitizer: Generation of Synthetic SEND Data from Existing Studies (https://doi.org/10.6084/m9.figshare.26940295.v1)
Kevin Snyder

Center for Drug Evaluation and Research, U.S. FDA

Abstract was not provided.

Chemistry data delivery from the US-EPA Center for Computational Toxicology and Exposure to support environmental chemistry (https://doi.org/10.6084/m9.figshare.26940295.v1)
Antony Williams

Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. EPA

In recent years, the growth of scientific data and the increasing need for data-sharing and collaboration in the field of environmental chemistry has led to the creation of various software and databases that facilitate research and development into the safety and toxicity of chemicals. The U.S. EPA Center for Computational Toxicology and Exposure has been developing software and databases that serve the chemistry community for many years. This presentation will focus on several web-based software applications that have been developed at the U.S. EPA and made available to the community. Although the primary software application from the Center is the CompTox Chemicals Dashboard, almost a dozen proof-of-concept applications have been built serving various capabilities. The publicly accessible Cheminformatics Modules (https://www.epa.gov/chemical-research/cheminformatics)

provides access to six individual modules to allow for hazard comparison for sets of chemicals, structure-substructure-similarity searching, structure alerts and batch QSAR prediction of both physicochemical and toxicity endpoints. A number of other applications in development include a chemical transformations database and a database of analytical methods and open mass spectral data. Each of these depends on the underlying DSSTox chemicals database, a rich source of chemistry data for over 1.2 million chemical substances. We will provide an overview of all tools in development and the integrated nature of the applications based on the underlying chemistry data.

FDA GSRS: Systematically Describing Substances and Chemical Synthetic Schemes
Tyler Peryea

Office of Data, Analytics and Research, Office of Digital Transformation, Office of the Commissioner, U.S. FDA

Abstract was not provided.

Diving into the depths of decade-old analytical QC data associated with the Tox21 Project (https://doi.org/10.6084/m9.figshare.26940295.v1)
Antony Williams

Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. EPA

Over the past 15 years, analytical quality control data (i.e., NMR, LCMS and GCMS) has been measured for ~9000 chemicals to check identity, purity and stability of the chemicals that are screened through bioactivity assays. This presentation gives an overview of the multiple analytical analyses that were conducted and discusses the resulting AnalyticalQC application and observations resulting from the analysis of 10s of 1000s of spectra regarding amenability to methods.

ChemSTER: Visualizing Chemical Space for Non-Targeted Analysis
Nathaniel Charest

U.S. EPA

Abstract was not provided.

The Complexities of Chemical Information regarding Predicting Compound Amenability with Liquid Chromatography Mass Spectrometry
Charles N. Lowe, Nathaniel Charest, Antony J. Williams

Center for Computational Toxicology and Exposure, U.S. EPA

Abstract was not provided.

Developing an EPA database of open spectra to support non-targeted analysis
Gregory Janesch

U.S. EPA

Abstract was not provided.

Non-Targeted Analysis using LC/HR-MS for Food Safety Applications: Capabilities, Challenges, and Potential Opportunities (https://doi.org/10.6084/m9.figshare.26940295.v1)
Ann Knolhoff

Center for Food Safety and Applied Nutrition, U.S. FDA

Abstract was not provided.

EPA's Non-Targeted Analysis WebApp: Linking cheminformatics resources with NTA workflows
Alex Chao

U.S. EPA

Abstract was not provided.

MetaboPique: A High-Throughput Computational Workflow for Validating, Annotating, and Organizing Tandem MS/MS Spectra Derived from Biological Samples (https://doi.org/10.6084/m9.figshare.26940295.v1)
Tytus D. Mak, Kelly H. Telu, Meghan C. Burke, Concepcion A. Remoroza, Yamil Simón-Manso, Brian T. Cooper, and Stephen E. Stein

Mass Spectrometry Data Center, NIST

In recent years, metabolomics has emerged from its niche beginnings to become a critical platform for biological sciences. However, its continued rise is predicated on advancements in high-throughput identification and characterization of the metabolome, the compendium of all small molecules involved in metabolism for a given organism. Critical to this task is the creation, curation, and expansion of mass spectral libraries that are derived from analyzing pure chemical standards as well as collating spectra from metabolites *in situ* via biological sample analysis. The latter is especially important due to many metabolites not being commercially available for purchase, and difficult because of the uncertainties involved in collecting spectra from complex matrices, which necessitate tedious validation and annotation steps.

As such, we have developed MetaboPique, a novel computational workflow that automates the many steps necessary for creating spectral libraries from data acquired via LC-MS/MS analysis of biological samples. MetaboPique evaluates and annotates each MS2 spectrum in the dataset via comprehensive analysis of its MS1 precursor extracted ion chromatogram (XIC). Peak shape is examined via non-linear least-squares fitting, which enables peak tailing, fronting, and gaussian-ness to be used as filtering parameters. The chromatographic data is also exploited for adduct type and charge prediction of the precursor via a novel brute-force combinatoric approach for XIC analysis. MetaboPique is able to cluster and synergistically propagate annotations of MS2 spectra acquired across multiple samples, including technical and biological replicates that differ in preparation.

For demonstration purposes, untargeted LC-MS/MS data acquired during analysis of 4 technical replicate Chinese Hamster Ovary (CHO) cell 50% acetonitrile metabolite extract samples were collated, validated, and annotated via MetaboPique. A total of 51,714 ESI+ MS2 spectra were collected across all samples via data-dependent acquisition methods on a Thermo Orbitrap Fusion Lumos with an HCD normalized collision energy of 20. 14,517 MS2 spectra were associated with 9,660 well-behaved (i.e. gaussian) XICs and organized into 4,517 spectral clusters. Adduct type elucidation was conducted by extracting ancillary XICs for 156 potential adduct m/z values generated from pairwise combinations of 13 common ESI+ adducts. A total of 2,148 clusters (47.5%) were annotated with adduct information. In addition, a consensus spectrum consisting of peaks found in $\geq$50% of the constituent spectra was calculated for spectral clusters containing more than 1 MS2 spectrum, allowing for 415 clusters to be identified via spectral searching with the NIST23 Mass Spectral Library.

Taken as a whole, these annotated spectral clusters represent a nonredundant library of MS2 spectra acquired from analytes surveyed from CHO cells, many of which may be endogenous metabolites that cannot be characterized by any other means. This data can be further annotated via existing mass spectral libraries and used for interrogating and evaluating CHO cell derived metabolomics datasets.

Chemometrics and Machine Learning to Enable Applications of NMR in Biomanufacturing
Frank Delaglio

Institute for Bioscience and Biotechnology Research, NIST|UM

Nuclear magnetic resonance spectroscopy (NMR) is a powerful and diverse tool to characterize higher order structure of biologics (HOS), because NMR spectra are sensitive to molecular shape and intermolecular interactions as well as chemical structure, and NMR can reproducibly probe this information at atomic resolution. Furthermore, NMR has the advantage that it can be applied non-destructively to protein therapeutics as formulated, with little or no sample preparation. Intriguingly, since NMR can also be applied to quantify the small molecule mixtures comprising the metabolome, NMR has the potential to characterize cell growth metabolomics in a bioreactor with a time resolution of hours or minutes, to reveal potential downstream effects on HOS and yield. Exploiting NMR for these biomanufacturing needs leads to a series of computational challenges which we review, including metrics of spectral similarity, data handling for applications of principal component analysis (PCA), spectral analysis of mixtures, and identification of spectral features by machine learning.

Beyond the Top Hit: Interactive Visual Interpretation of Hybrid Similarity Search Hit Lists (https://doi.org/10.6084/m9.figshare.26940295.v1)
Brian T. Cooper

Mass Spectrometry Data Center, NIST

The sheer diversity of chemical structures in complex biological samples guarantees that mass spectral libraries will never be complete. But current libraries can be expected to contain "cognates" of most unknown metabolites—compounds with a structural difference confined to a single region of the molecule that does not substantially alter its fragmentation behavior. The NIST “hybrid” similarity search expands the scope of library searching by returning a list of library compounds likely to share structural features with an unknown metabolite. The big remaining challenge is to use the structural information in hybrid search hit lists to confidently identify authentic unknowns. This is exceedingly difficult to automate, so we developed an interactive graphical user interface to help researchers suggest and evaluate possible unknown structures. Hybrid searches are performed by calling NIST’s MSPepSearch command-line search tool and parsing its text output. The GUI displays the entire hit list and displays chemical structures and associated metadata for the first 12 hits, optionally skipping structures for potential isomers of the unknown (mass difference within tolerance of zero). The user may visually select two or more hits to include in a maximum common substructure (MCS) calculation, after which the MCS is overlaid on those structures that contain it. The RDKit cheminformatics library is used to draw structures, perform the MCS calculations, and determine the mass and molecular formula differences between the completed-valence MCS and each hit that contains it. Two versions of the GUI were developed: one for high-resolution tandem MS (which allows the user to specify the likely unknown precursor ion type), and one for unit-mass-resolution electron ionization spectra (which allows the user to specify the unknown molecular weight if it cannot be accurately determined from its spectrum). Relevant examples are presented for both types of searches.

NMR metabolomics from design to metabolite ID
Goncalo Gouveia

NIST

Presentation was withdrawn.

Disparately Acquired LC-MS Alignment with Applications in Metabolomics
(https://doi.org/10.6084/m9.figshare.26940295.v1)
Hani Habra

Mass Spectrometry Data Center, NIST

Metabolite profiling studies are typically performed under replicated LC-MS settings for all experimental samples. Changes to important parameters, such as chromatography and instrumentation, have major consequences for metabolomics analyses. Specialized procedures are required for spectral alignment of signals corresponding to identical compounds and normalization of spectral abundances acquired under non-identical experimental conditions. Here we present applications of disparate LC-MS alignment and data analysis in the metabolomics field, using the *metabCombiner* software package. First, plasma metabolomics data was collected from two sets of maternal subjects and their infants three years apart using distinct instruments and LC-MS procedures. We briefly outline a procedure for merging and harmonizing these study subsets into a unified data matrix. Bioinformatics analysis on the combined sample set reveals large-scale metabolic changes associated with gestation. Second, four institutions performed untargeted lipidomics using in-house procedures as part of the inter-laboratory Unknown Lipids consortium. We demonstrate the alignment of these data sets; determine overlapping known and unknown compounds; and compare annotation rates among the laboratories in both ionization modes. Finally, we describe the Publicly Available Metabolomics Data Alignment (PAMDA) project, an ongoing effort to assemble lists of commonly detected urinary compounds detected in multiple studies deposited in public data repositories. In addition, a new workflow integrating tandem MS comparisons will be discussed. Together, these applications of disparate LC-MS alignment demonstrate opportunities to improve metabolomics data inter-operability, increase compound annotation rates, and augment sample sizes for improved statistical power.

Confident Identification of Extractables in Medical Devices: The need for a standardized approach for chemical characterization to facilitate toxicological risk assessment
Eric Miller

Center for Devices and Radiological Health, U.S. FDA

Abstract was not provided.

Using AI to Improve Predictive Models for Chemistry
(https://doi.org/10.6084/m9.figshare.26940295.v1)
Eric Stahlberg

Cancer Data Science Initiatives, Frederick National Laboratory for Cancer Research

The Accelerating Therapeutics for Opportunities in Medicine project has been a collaboration with academia, DOE laboratories, and industry involvement to develop an open platform for AI model-driven drug development with active learning. The platform is available on Github, has been installed at multiple sites, and supports both predictive AI models as well as generative AI models. The platform also has the capacity to integrate and support biomedical digital twin and virtual human models in the overall treatment optimization process. The presentation also covered the open resources for sharing models and data through the Predictive Oncology Model and Data Clearinghouse (modac.cancer.gov). The platform will be increasingly available in the cloud and within government resources such as precisionFDA. The depth and breadth of models and datasets will also grow through collaborations.

The DOE/NCI IMPROVE Framework for AI Model Improvement
Tom Brettin

Argonne National Laboratory

The IMPROVE project aims to improve deep learning models for predicting drug response in tumors. It leverages relevant DOE assets like exascale class computing and data infrastructure, and large-scale simulation capabilities, and involves developing protocols for generating new data and comparing deep learning models to improve predictive models of drug response. The project also seeks to generate well-curated, clinically relevant, and standardized training and testing datasets, and engage the community for expert opinions and collaborations on developing a model evaluation framework and generating benchmark data. The project has the potential to generate new hypothesis and identify previous hidden cancer types and treatment targets.

SAVI - Going Back to Beyond AI
Marc Nicklaus

National Cancer Institute, NIH

We describe the motivation for, components of, properties and usage of, as well as future plans for, the NCI CADD Group's Synthetically Accessible Virtual Inventory (SAVI). SAVI in its 2020 version ("SAVI-2020") is a computer database of 1.75 billion easily synthesizable small molecules useful for drug development. For each of these structures, commercially available

building blocks and the proposed synthetic route are provided in the database. For those 170 or so SAVI-2020 products that were submitted to actual synthesis, the synthetic success rate has been >90%. We outline the possibility of AI approaches for generation and use of future SAVI versions.

Machine learning models for predicting multi-generation reproductive toxicity of chemicals
Jie Liu

Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. FDA

Reproductive toxicity is one of the prominent endpoints in the risk assessment of environmental and industrial chemicals. Due to the complexity of the reproductive system, traditional reproductive toxicity testing in animals--especially guideline multigeneration reproductive toxicity studies--takes a long time and is expensive. We curated rat multigeneration reproductive toxicity testing data and developed predictive models using seven machine learning algorithms (decision tree, decision forest, random forest, k-nearest neighbors, support vector machine, linear discriminant analysis, and logistic regression). The performances of individual and consensus models were evaluated using 500 iterations of 5-fold cross-validations and the external validation data set. The balanced accuracy of the models ranged from 58% to 65% in the 5-fold cross-validations and 45% - 61% in the external validations. We demonstrate the importance of using consensus models for harnessing the benefits of multiple machine learning models. While we continue to build upon the models to better characterize weak toxicants, there is current usefulness in saving resources by being able to screen out strong reproductive toxicants before investing in vivo testing. Our results suggest that machine learning may be a promising alternative approach to evaluate the potential reproductive toxicity of chemicals.

Development, Evaluation, and Application of QSUR Models
(https://doi.org/10.6084/m9.figshare.26940295.v1)
Katherine Phillips

Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. EPA

Functional use (also called technical function) is the role a chemical is intended to play when it is added to a product or process. Function useful is informative to both formulation chemists and health assessors. A chemical's structure informs what functional uses a chemical can have in various products or processes. Despite technical function being so informative, this information (at least publicly available information) for most chemicals in the Toxics Substance Control Act (TSCA) public, active inventory is limited.

To fill these gaps in publicly available functional use information, quantitative structure-use relationship (QSUR) models were developed in 2017. QSURs use a chemical's structure to predict functional uses that a chemical could likely serve; prediction values range from 0 (unlikely to serve a function) to 1 (likely to serve a function). QSURs have been used to parameterize high-throughput exposure models, propose functional alternatives, and bolster non-targeted analysis tentative identifications. However, the original models suffered from a relatively small, consumer product-focused training set and were not found to be predictive outside of the domain of consumer product chemicals. To improve the domain of applicability of the original QSURs, a larger, more diverse training set was constructed. A data curation tool that has been developed at U.S. Environmental Protection Agency (EPA) was used to allow for tracking data provenance and quality assurance of the data being collected and curated for use in developing new QSUR models. In addition, the Organisation for Economic Co-operation and Development's (OECD) harmonized technical function categories were used to standardize functional uses across different data sources.

The new QSUR models trained with this larger, standardized training set were then evaluated with EPA's Chemical Data Reporting (CDR) data set collected for the 2020 reporting cycle. This data set reports chemicals used in industrial and commercial settings with annotation to OECD technical functions. Comparing both the original and newly re-trained models showed an overall improvement in the domain of applicability of the new QSUR models as well as improved numbers of true and false positive (i.e., correct predictions) across the CDR 2020 data set. This talk was an example of how models can be used to fill data gaps, and how improved collection, curation, and storage of appropriate data can be used to refine and improve those models.

AI Methods for Chemical Datasets (https://doi.org/10.6084/m9.figshare.26940295.v1)
Samir Lababidi

Office of Data, Analytics and Research, Office of Digital Transformation, Office of the Commissioner, U.S. FDA

There has been a tremendous interest in using artificial intelligence (AI) across all medical applications. Recently FDA released its “Artificial Intelligence and Medical Products: How CBER, CDER, CDRH, and OCP are Working Together,” which outlines how FDA’s medical product centers are working together to protect public health while fostering responsible innovation in AI used in medical products and their development. To explore AI/ML methods in Cheminformatics, this talk provided a general overview of AI and machine learning (ML) concepts. A comparison between supervised and unsupervised approaches was presented and some of the basic principles for developing algorithms and classifiers in ML were highlighted. As an application to Cheminformatics, a dataset was presented where an ML algorithm was developed to predict whether a substance is active or inactive by training on a dataset of chemical descriptors with Multilevel Neighborhoods of Atoms. The ML algorithm resulted in a very high specificity of 92.9% with a total sensitivity + specificity larger than 100%. The

application presented some important concepts and lessons learned for future development of ML algorithms in Cheminformatics.

Machine Learning with Chemical Data and Molecular Representations
Neeraj Kumar

Pacific Northwest National Laboratory

Abstract was not provided.

Data and Computing Resources for Building Predictive Safety Models
Felice Lightstone

Lawrence Livermore National Laboratory

Abstract was not provided.

Scalable Workflow for On-Demand Lead Optimization - A Pilot
(https://doi.org/10.6084/m9.figshare.26940295.v1)
[1]Mike Mikailov, [2]Yulia Borodina, [1]Fu-Jyh Luo, [1]Kenny Cha

[1]Office of Science and Engineering Laboratories Center for Devices and Radiological Health, U.S. FDA

[2]Office of Data, Analytics and Research, Office of Digital Transformation, Office of the Commissioner, U.S. FDA

The recent explosion of chemical libraries at the National Cancer Institute (NCI), reaching beyond a billion molecules, has enabled large-scale simulations for virtual screening (VS). VS is a simulation technique used in drug discovery to search libraries of molecules to identify structures that are likely to bind to a drug target. The FDA CDRH High-Performance Computing (HPC) team is working with the NCI to apply scaling techniques and take advantage of advances in computational power to make this mission-critical task feasible and accomplishable in a timely manner. A scalable workflow on the HPC clusters has been created for on-demand lead optimization. This workflow reduces the dimensionality of the libraries and enriches the dataset with more structurally diverse, lead-like substances. This workflow reduced the computation time from more than a month to less than 10 minutes, enabling quick and efficient search through large libraries of molecules to find potential drug candidates. This new method may help supplement the traditional techniques in accelerating the drug discovery process.

Report from the Day-1 panel discussion

Panelists:
Antony Williams (EPA/ORD)
Ewy Mathe (NIH/NCATS)
Jian (Jeff) Zhang (NIH/NCBI)
Kevin Snyder (FDA/CDER)
Margaret (Meg) Rush (Gryphon Scientific at NIH)
Steve Heller (NIH)
Tyler Peryea (FDA/OC)

Moderator: Evan Bolton (NIH/NCBI)

The panelists were invited to present their personal opinions about several pain-points described below. Here are the highlights of the questions followed by proposed solutions:

We have a number of different issues in trying to make data FAIR. FAIR is basically the ability of making information machine readable in some forms. For many people, FAIR is not findable, accessible, or reusable. It really means AI-ready and you can start to do things like machine learning and data analysis. But the process of taking human readable information to make it machine readable is long, difficult, and arduous. Which issues do you encounter in making data FAIR at your organizations? What works well and what does not? Any pain points?

- Make data machine readable by building data standards, ontologies, controlled vocabularies, etc.
- Come to a common data standard in the research community
- Develop open-source tools or web form to help data standardization
- Both incentives and resources are important in making data FAIR

We are all working towards making data available to our end users in one format or another. We are digesting various kinds of information. How are you approaching making your data more machine readable and usable for end users?

- Use API with standardization of programmatic interfaces
- Make data also available outside the API and in a variety of formats for downloading
- Implement standards and ontologies in research communities
- Add dropdown with controlled vocabulary in excel files to standardize the data input
- Use topic clustering to help harmonize similar data sets from different sources
- Develop international standards across the organization and practical prototypes in parallel
- Make more convenient ways -- creating graphical interfaces, providing more accessibility and findability

The data is very complex and there are always all types of issues when consuming the data. Also, there is a huge diversity in how we are working with machine readable data, including our customers. As agencies, we are approaching different technologies and software tools. How do you consume machine readable data? (What technologies? What workflows? Cloud vs non-cloud? Graph DB vs relational DB vs spreadsheet?)

- Standards and workflows would be useful in consuming the data
- Develop a software to use agency-wide or in small groups
- Spreadsheet is still widely used. Machine learning and AI would also help to consume the data

What are your needs for machine readable data? (What ontologies? What data structure? Where does data need to be?) Are there areas that we need to work together on to harmonize cheminformatics and related data in a fashion that would make more sense to make the data more accessible and usable?

- Resource Description Framework (RDF) is helpful for data sharing and exchanging
- Need easy understanding and use data structure and ontologies
- Make the concept of ontology more accessible

As a data provider, are you able to share the information you want? (e.g., do you have internal restrictions preventing you from sharing data?)

- Make the inaccessible data public by separating out the confidential information
- Lots of efforts are spent to absolutely make sure that the data we want to share are shareable
- Data use agreements would help on data sharing
- Define the level of the data for public

Report from the Day-2 panel discussion

Panelists:
Ann Knolhoff (FDA)
Antony Williams (EPA)
Tytus Mak (NIST)
Frank Delaglio (NIST)

Moderator: Evan Bolton (NIH)

The topic of the Day-2 panel discussion focused on the application of cheminformatics in analytical chemistry, particularly in the areas of mass spectral and NMR data analysis. The

panelists addressed multiple issues including the reproducibility of spectral data in both large biomolecules and small molecules studied by NMR and mass spectrometry (MS). The panelists commented that in general chemical shifts of common nuclei in proteins and organic small molecules in NMR spectra are highly reproducible (with many examples where they are not: exotic nuclei, nuclei in exchange, solvent and temperature dependency), However, line shapes are not always reproducible because they can change significantly with the experimental conditions and instruments used. Comparing spectral data can be challenging, and the approach must be planned accordingly with consideration given to the acquisition of the data and the assembly of a reference database. For example, in NMR the data from multiple instruments on biomolecules (such as monoclonal antibodies) can be compared using the chemical shift data. For small molecules searchable databases containing chemical shifts are available including commercial solutions (e.g., ACD/Labs) and free online databases such as NMRShiftDB (https://nmrshiftdb.nmr.uni-koeln.de/). Comparing NMR spectral data for small molecules can also be challenging in NMR, particularly for tautomers as their spectra can easily be influenced by the experimental conditions (e.g., sample concentration, pH, temperature).

In MS, chromatographic reproducibility for LC-MS platforms is difficult to achieve even within the same lab. Furthermore, complex matrix analysis often results in "chimeric" MS/MS spectra that makes it difficult to compare the same analyte from different biological samples. For complex biomolecules, a common practice in both NMR and MS is establishing the measurement variance, using the collection of the spectra that represent the variances. Potential applications of cheminformatics to alleviate some of the mentioned pain points in evaluating/comparing the spectral data was also discussed. However, the panelists believed that cheminformatics alone would not be able to solve the problems but rather a combination of the computational approaches and experimental data would be needed.

Another topic that was discussed at the session was information sharing. All panelists agreed that data format and metadata describing the data were two of the major challenges in sharing information. This includes results from different labs using different software platforms (either proprietary or open source) to generate data. Solutions proposed by the panelists included the use of virtual machines that contain scripts that automatically process the data generated by the software. However, this topic remains a challenge though it was also agreed that there is an increasing abundance of open file formats standardizing vendor formats (e.g., JCAMP or NMRML for NMR and mzML for mass spectrometry.

Report from the Day 3 panel discussion

Panelists:
Eric Stahlberg (FNLCR)
Huixiao Hong (FDA/NCTR)
Tyler Peryea (FDA/OC)
Marc Nicklaus (NIH/NCI)
Margaret Rush (Gryphon Scientific)
Mike Mikailov (FDA/CDRH)

Moderator: Samir Lababidi (FDA/OC)

The panelists were invited to present their personal opinions and comment on several questions related to AI technology in Chemo-informatics. Here are the questions and the highlights of their comments and proposed solutions:

What is needed to establish community reference datasets for evaluating model performance? Can these be public, or do they need to be private?

- We need a set of data that is trustworthy to serve as a truth dataset. It does not have to be complete, but it should represent a yardstick to compare what is measured.
- There has not been proposed such a reference dataset previously. If such a reference should exist, it must be publicly available.
- Nevertheless, there are challenges to develop a reference dataset that is universal. One problem is that the dataset depends on the assay used to produce the chemicals and this assay can vary, e.g., with different labs. Another problem is that the reference dataset depends on the specific endpoint under investigation and on the contest of use.
- This kind of reference dataset operates at the intersection of multiple disciplines and so it would require a diversity of expertise to develop it. It is also needed to quantify common controls as part of the process.
- Additionally, the reference dataset would have to be somehow different depending on whether it is for toxicological purposes or for a drug design. A dataset focusing on toxicology should be open whereas it is not necessary the case for drug design.

Given the excess of breadth and the size of chemical space, how do we characterize the standardized domain of applicability for models?

- An important question in model development is how to assess whether a compound is close to the training set and whether the training set is appropriate for this assessment. Can we agree on an approach on how to determine the distance between a compound and a training set to assess its ability to characterize the predictions sought of by the model?

- There are billions or several billions of molecules out there, but we can have only a small number of building blocks of molecules with little overlap. Therefore, it may not be necessary to look at billions of molecules; instead, we can use the building blocks.
- There are two different parts for this question, one about the science and the chemical space and the other one is how to communicate the model to the computer. It seems that the computer part is more approachable, where we can build standards and check if a model makes sense. On the computer side, the CDRH's High Performance Computing (HPC) is quite suitable to perform such computational models.
- There are different ways to define the domain of applicability for models. However, the choice of the metric is the important factor in the model development.
- In Chemo-informatics, the range of applications spans from prediction of 3-dimensional structure to how small molecules are used. Through this continuum there are different requirements we need in developing models.
- In the process of building these models, it is worth looking at how we transcribe the chemical space on the computer and explore some of its nuances. In this respect, we note here that we could find a cancer cure using mouse models, but with no applicability in human.
- Finally, in the models we develop we should consider the situation where molecules have features that do not exist in the training set. Some of the reasons this may happen include non-generalizability of the training set and improper choice of the training set.

How do you see the use of AI/ML models can improve Chemo-informatics as compared to what had been done so far?

- Although there may be some skepticism in the use of AI/ML in Chemo-informatics, it is doing a good job and the future seems to be bright with such technologies such as Large Language Models (LLMs) in ChatGPT. The community is evolving in understanding products and results we get using LLMs, especially those in chemical space.
- However, these models should be verified, and their assumptions should be investigated further to confirm the results. Particularly, it's dangerous to rely 100% on the results we get using LLMs. For instance, there is a danger in self-learning by machines based on made-up materials without further verification of these products.

What are the opportunities to broadening the use of AI/ML in Chemo-informatics?

- A good use of AI/ML is to help us be in-line versus out of line. Particularly, it can make sure things are consistent and can effectively integrate different types of data to make it more operable.
- One powerful use of AI/ML is to bring unstructured data into structured format. Along the same lines, AI can take care of simple repetitive tasks and improve the data format and its use.
- Another advantage of the use of AI/ML is to integrate complex data from different sources, which would, otherwise, require different people with different background and expertise. In fact, ChatGPT can help in this direction, but one should not completely

depend on it at this time. Nevertheless, ChatGPT is great for writing computer programs.

How do you see the research in Chemo-informatics is shaping up in the near future?

- With increased use of automation, we are getting better precision and reduced variability. This will provide a growing amount of reliable data while at the same time capturing the information and the context that the information is generated in, which will help us establish the ground truth. This does not seem to be far off and we are already seeing early adoption of that.
- Another direction that may take place is to move into quantum chemistry (at the level of molecules), which may change Chemo-informatics as we see it now or perhaps get a hybrid of this and what we have so far. AI can help in this direction as well. There is some work going now in the area of quantum mechanics on compounds. In fact, with more adoption of AI/ML approaches, AI can be trained on quantum chemistry and help drive the best rule of thumb in addition to discovering new things that weren't possible before.
- We would like to show better understanding of chemistry in life to biologists, with applications to precision medicine, which, for instance, can help our understanding of new drugs and how they are metabolized in the body, so we could get more effective drugs and less toxicity!
- Another thought is that Chemo-informatics may take on the same effect as the bioinformatics has been on biologists, which is helping them to better move the field forward.
- At the intersection of biology and Chemo-informatics, we could elucidate the interaction between chemical space and the disease space through, e.g., protein-protein interaction, which may lead to more biomarker discovery.