Summary of Presentations and the Panel Discussion

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.

The IUPAC InChI Chemical Identifier Standard Project (https://doi.org/10.6084/m9.figshare.21431931)
Stephen Heller
National Center for Biotechnology Information, National Library of Medicine, NIH

This presentation describes the IUPAC chemical structure standard - InChI- the International Chemical Identifier. InChI provides a sustainable standard that enables connections in chemistry for the advancement of science and medicine for the public benefit. 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

Open Access Chemical Information Integration in PubChem
(https://doi.org/10.6084/m9.figshare.21429483)
Jian Zhang
National Center for Biotechnology Information, National Library of Medicine, NIH

PubChem (https://pubchem.ncbi.nlm.nih.gov/) is an open chemistry database within the National Library of Medicine at the National Institutes of Health. PubChem has served the community since 2004. Today, PubChem provides information for more than 110 million unique compounds that link to nearly 300 million bioactivities, 42 million patents, and 34 million literatures. With such massive information, PubChem serves millions of users that yield more than 10 million page views per month. As an information hub, PubChem integrates chemical data from more than 860 data sources around the world. In this presentation, we will discuss the chemical data integration from open access databases including EPA, FDA, USDA, ATSDR, NIOSH, USGS, ECHA, WHO, FAO, ILO, EMA, and more. We will also discuss the challenges for data mapping and standardization. Lastly, we will present the data retrieving tools using PubChem's restful APIs and classification browser.

FDA Global Substance Registration
Tyler Peryea
Health Informatics Staff, Office of Data, Analytics and Research, Office of Digital Transformation, Office of the Commissioner, FDA

The Global Substance Registration System (GSRS) is an open-source substance database solution for indexing, defining, and exchanging substance information relevant to human health. The system supports defining and indexing substances from single atoms to complex gene and cell therapies using a variety of cheminformatics, bioinformatics and taxonomic tools and descriptors. The Food and Drug Administration (FDA) has partnered with the National Center for Advancing Translational Sciences (NCATS) as well as other international collaborators to maintain and distribute this software to support the ISO 11238 standards for substances. The FDA uses an internal database based on the GSRS software (FDA-GSRS) to issue Unique Ingredient Identifiers (UNIIs) which the agency uses to uniquely track active substances, metabolites, impurities, additives and excipients through the lab, clinical trials, and throughout the global marketplace. FDA's Global Substance Registration System (GSRS) enables efficient and accurate exchange of information on substances through their UNIIs, but much of the data feeding into the FDA's GSRS remains manually extracted. Embedded GSRS tools which use computer vision, OCR, and federated resolvers simplify the manual data registration process, but recent efforts have been on formatting machine-readable chemical data earlier in the regulatory process. FDA-GSRS has recently begun accepting SD files and GSRS JSON to facilitate faster and less error-prone data exchange from sponsors and collaborators. This presentation describes the background of how GSRS came to be, how the tools and datasets are used, and how these new forms of data exchange are beginning to simplify the registration process and improve the agency's ability to monitor the global supply chain.

Where to Find Regulated Substances at EPA: The Substance Registry Service (SRS)

Akshay Narang
Data Management Services Division, Office of Mission Support, US EPA

The Substance Registry Services (SRS) is EPA's authoritative resource for information about chemicals, biological organisms, and other substances tracked or regulated at U.S. EPA. Different program offices use different names for the same regulated substance. For example, there are eight different names for Lindane across EPA regulations. Therefore, SRS was built to show the relations among the synonyms to help the tracking and reporting for EPA and industry. This presentation also shows a demo on how to use the SRS online at EPA.gov/SRS.

U.S. Army Combat Capabilities Development Command Chemical Biological Center

Christopher Ellis
U.S. Army Combat Capabilities Development Command

In this presentation, the cheminformatics research efforts at the U.S. Army were highlighted. The approaches to profile the U.S. Army's chemicals of interest were discussed. These approaches included the predictions and calculations of the chemicals' physico-chemical properties, chemical structural similarity, and toxicity profiles. The future directions on the project were also mentioned.

Food and Food Packaging Chemicals at the FDA

Tammy Page
Center for Food Safety and Applied Nutrition, FDA

In this presentation, the cheminformatics information systems in the Office of Food Additive and Safety (OFAS) at the U.S. FDA were highlighted. The systems covered were the Chemical Evaluation and Risk Estimation System (CERES), the Scientific Terminology and Regulatory Information (STARI), and the Food Applications and Regulatory Management (FARM) systems. The data harmonization approach to integrate the chemical information between the three systems and their applications in supporting the safety assessment of the food contact substances reviewed in OFAS were summarized.

The Materials Project: A Community Data Resource for Accelerating New Materials Design

Anubhav Jain
Lawrence Berkeley National Laboratory

The Materials Project (www.materialsproject.org) is a free resource providing data and tools to help perform research and development of new materials. The database contains information on up to 150,000 materials, such as electronic structure, magnetic properties, and dielectric properties. The data could be accessed through the Materials Project API. The Materials Project has been widely used by the research community in designing new materials. To add value to the community of materials researchers, the materials project allows the research community to contribute their own data and help to disseminate the data. MPContribs provides a platform and advanced programming interface (API) to contribute computational as well as experimental data to the Materials Project. Matbench is an automated leaderboard for benchmarking state of the art machine learning algorithms predicting a diverse range of solid materials' properties.

USDA's AWIC: How to Conduct a Literature Search for Cheminformatics Information to Find Animal Use Alternatives (https://doi.org/10.6084/m9.figshare.21429474)

Jessie Carder
Animal Welfare Information Center, National Agricultural Library, Agricultural Research Service, USDA

USDA's Animal Welfare Information Center (AWIC) provides information and training on the Animal Welfare Act and the 3Rs principle, replacement, reduction, and refinement. One of AWIC's main activities is teaching how to search bibliographic databases for citations on the 3Rs. The workshop will teach the strategy for literature search by combining the descriptive keywords of 3Rs, toxicology, and cheminformaics. More information about the workshop and AWIC are available at the website: https://www.nal.usda.gov/programs/awic.

The National Institute for Allergy and Infectious Diseases' Anti HIV/OI/TB Therapeutics Database: ChemDB (https://doi.org/10.6084/m9.figshare.21430296)

Louise Sumner and Margaret Rush
National Institute of Allergy and Infectious Diseases, NIH

The ChemDB Anti HIV/OI/TB Therapeutics Database was established by the National Institute for Allergy and Infectious Diseases (NIAID) to track small molecules tested against HIV, TB, HBV and other opportunistic infections associated with HIV. The database currently includes approximately 410,000 public and proprietary compounds and associated biological testing data, derived from about 32,900 references. Public ChemDB data is accessible at https://chemdb.niaid.nih.gov. Approximately 151,000 public compounds have biological data associated with HIV, and 48,000 have biological data associated with Mycobacterium tuberculosis. The public portion of ChemDB is a unique resource for researchers around the world. Each compound in the database forms a record containing compound details, biological testing data, and literature citations. The database website enables researchers to search by chemical and structural details of a compound, as well as filtering for threshold biological activity values. ChemDB has been used by researchers for a wide range of cheminformatics analyses, such as structure-activity relationship studies, virtual drug discovery, machine learning/artificial intelligence training datasets, and examining research trends over time.

Per- and Polyfluoroalkyls: a “New” Class of Persistent Organic Pollutants

Eugene Demchuk
Office of Innovation and Analytics, Agency for Toxic Substances and Disease Registry

A gamut of per- and polyfluoroalkyl substance (PFAS) products is manufactured and ends up in the environment each year. The Organization for Economic Co-operation and Development (OECD) compiled nearly 5,000 PFAS structures. Examination of the Chemical Identifier Resolver (CIR) of the National Cancer Institute and ChemSpider revealed that this class of compounds is underrepresented in popular chemical informatics databases. Only 21% and 61% of OECD PFAS structures are covered by CIR and ChemSpider, respectively. An exigent need for better representation of PFAS in cheminformatics is stipulated by health concerns associated with exposure to them. According to the National Report on Human Exposure to Environmental Chemicals released by the U.S. Centers for Disease Control and Prevention, up to 98% of the U.S. population has been exposed to PFAS. Exposures to PFAS are associated with adverse health effects in animals and humans. The Agency for Toxic Substances and Disease Registry (ATSDR) has published minimal risk levels (MRLs) for four PFAS structures. MRLs for four others are under consideration. These are a tiny fraction of PFAS structures reported by the OECD. Therefore, a need for modeling or high-throughput screening (HTS) of the health effects of PFAS is imminent. Using 26 perfluorinated linear alkyls as a test case, the feasibility of computational modeling was investigated. It was found that top biological activity spectra of PFAS shown by CIR are in good agreement with the literature. Both emphasize developmental and endocrine effects as the most sensitive adverse outcomes of PFAS exposures. A comparison of literature data with the results of HTS screening points to similar health effects, along with a notion that PFAS structures with longer carbon chains are more toxic. Thus, the future development of a computational metric for modeling PFAS adverse health effects is likely. Disclaimer: The findings and conclusions in this presentation have not been formally disseminated by the Centers for Disease Control and Prevention/the Agency for Toxic Substances and Disease Registry and should not be construed to represent any agency determination or policy.

EPA’s DSSTox Database: The Strategic Role and Requirements of Chemical Curation (http://doi.org/10.6084/m9.figshare.21429381)

Ann M Richard
Office of Research and Development, Center for Computational Toxicology and Exposure, US EPA

The expansion of chemical-bioassay data in the public domain is a boon to science; however, inaccuracies in linkages of CAS registry number (CASRN) to structure, and unreliability of names and synonyms assigned to a particular structure are well known. EPA’s DSSTox project, through strategic use of expert manual curation and strict enforcement of database rules, has earned a reputation for providing quality chemical-structure annotations to the environmental toxicology community. The DSSTox database, supported by modern cheminformatics tools, has grown to contain more than 1.2 million unique chemical records, and provides the underpinning for EPA’s CompTox Chemicals Dashboard that uses chemical structure to link diverse data streams and resources in support of EPA’s chemical research and regulatory programs. A history of DSSTox’s expansion from 25K to over 1million substances, while retaining a focus on quality chemical associations, is presented, with a particular focus on the critical role of expert manual curation, how and when this resource is implemented and applied, and the impact of expert curation on real Agency problems. Examples include: curation of more than 10,000 per- and poly-fluorinated alkyl substances (PFAS) structures and over 1000 Markush-type structures to represent ill-defined PFAS substances; curation of ToxCast and Tox21 sample libraries supporting EPA’s high-throughput testing programs; and creation of reference libraries in support of non-targeted screening objectives. Whereas other large public chemical databases also employ automated chemical curation approaches to address quality issues, the DSSTox project owes its reputation for quality to the necessary and strategic use of expert manual curation.

NCI CADD Group Cheminformatics Resources: From Web Tools to Billion-Molecule Databases (https://doi.org/10.6084/m9.figshare.21430311)

Marc Nicklaus
Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, NIH

The CADD Group of the NCI, NIH, has a nearly 25 year history of making chemoinformatics resources freely available via a public web server, now accessible under https://cactus.nci.nih.gov. The presentation recounts the beginning of these web tools, the "NCI Database Browser," which was the first public Web GUI for a large small-molecule database with advanced capabilities such as full substructure search. Other early tools offer the capability of generating 2D structures of chemicals, conversion of SMILES into other representations, and optical structure recognition. The most widely used tool is the Chemical Identifier Resolver, which “resolves” structure identifiers or representations, i.e., converts one structure identifier/representation into another. Several services and downloadable data sets are related to the analysis and handling of tautomerism. One of the newest tools is the Synthetically Accessible Virtual Inventory (SAVI), an ultra-large database of 1.75 billion easily synthesizable screening samples, with a proposed synthetic route associated with each SAVI product. This is one of the currently nine downloadable data sets comprising nearly 2 billion structures in total.

Development of Cheminformatics Systems for Facilitating Application of Computational Chemistry in Regulatory Sciences (http://doi.org/10.6084/m9.figshare.21429489)

Jie Liu
Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, FDA

Cheminformatics systems (databases, tools, and algorithms) play an important role in regulating the chemicals in FDA-regulated products. Division of bioinformatics and biostatistics at NCTR, US FDA has developed multiple cheminformatics systems which are publicly available (http://www.fda.gov/ScienceResearch/BioinformaticsTools/default.htm). This presentation introduced several cheminformatics systems, including chemical databases EDKB (Endocrine Disruptor Knowledge Base), EADB (Estrogenic Activity Database (EADB), MAAR (Molecules with Androgenic Activity Resource), and OAK (Opioid Agonists/antagonists Knowledgebase), software tool Mold2 for molecular descriptors calculation, and machine learning algorithm Decision Forest. Application cases will be shared to demonstrate the usefulness of the developed cheminformatics systems in risk assessment of chemicals.

Quantum Chemical and Quantitative Structure-Activity Relationship Studies to Reduce Exposure to Mycotoxins in Food and Feed

Michael Appell
Mycotoxin Prevention and Applied Microbiology Research Unit, National Center for Agricultural Utilization Research, Agricultural Research Service, USDA

Cheminformatics has helped find solutions to many agricultural and food chemistry problems, including pest management, food design, quality assurance of agricultural commodities, agro-based biomaterial development, and food safety. Contamination of commodities by harmful pests, microbes, and mycotoxin producing fungi threaten our safe and nutritious food and feed supply. Historically, phytochemicals are sources for lead biochemicals that exhibit desirable biological activities to benefit human health and agriculture. Popular plant-based foods are potential sources for safer lead biochemicals with favorable properties to address food safety problems. Quantum chemical and cheminformatics methods were applied to discover food components with predicted antimicrobial activities against harmful micro-organisms that occasionally contaminate agricultural commodities and pose health risks. Predictive models using resources developed in-house and provided by the U.S. FDA and U.S. EPA enabled the rapid and cost-effective analysis of the absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties. The models were built using genetic function algorithms and artificial intelligence/machine learning technology. Quantitative structure-activity relationship (QSAR) modeling identified two lead components of popular foods and allowed rapid assessment of other antimicrobial candidates. Models were trained on experimental studies and validated using test sets of chemicals not used to train the models. The cheminformatics resources provided by governmental organizations identified lead compounds that exhibited broad antifungal activities in experimental fungal growth assays. This robust methodology also developed validated predictive models for mycotoxin toxicity, phytotoxicity, cytotoxicity, and properties associated with false-positive detection results. The U.S. governmental cheminformatics resources advanced a high-priority program that directly and positively affects food safety.

In Silico Evaluation of Biokinetics and Toxicity of Phytochemicals
(https://doi.org/10.6084/m9.figshare.21430335)

Yitong Liu

Division of Toxicology, Office of Applied Research and Safety Assessment, Center for Food Safety and Applied Nutrition, FDA

The use of botanicals as traditional medicines, herbal supplements, and natural health products continues to grow globally. Along with the broad use, toxicities have been associated with certain botanicals, such as fo-ti and ephedra. Limited data are available on their toxic mechanisms, including toxic constituents and their biokinetics in the human body. Botanicals contain complex mixtures of large numbers of active phytochemicals and structural information are often the only data available. Since the traditional animal testing is resource intensive and potentially lack of human relevance, new approach methodologies including in silico modeling are valuable in conducting initial screening and filling data gaps. Here, quantitative structure-activity relationship (QSAR) and physiologically based pharmacokinetic (PBPK) modeling were used in three case studies to predict the biokinetics and hepatotoxicity of phytochemicals. The first study adopted a tiered QSAR approach that incorporated absorption and metabolism prediction into evaluating the liver toxicity of 255 phytochemicals in 20 botanicals. Both the absorbed phytochemicals and their metabolites were predicted for causing potential human liver toxicities. The second study investigated the concordance between in silico QSAR prediction and in vitro testing for inhibiting the activities of liver metabolizing enzymes (cytochromes P450, CYP). Between 75% and 90% of the 22 anthraquinones in aloe and fo-ti were correctly predicted for CYP1A2, CYP2C9, CYP2C19 and CYP2D6. The third study used PBPK modeling to extrapolate in vitro hepatotoxicity of 16 anthraquinones to in vivo human toxicity. A toxic dose was predicted and compared with human daily intake levels of anthraquinones in botanicals to make safety assessment. These in silico results agreed with existing literature reports on the biokinetics and toxicity of phytochemicals and could be used to guide further testing.

Chemometric Trends for Persistent Anthropogenic Toxins Found in Seabirds of the Northern Pacific
Nathan Mahynski
NIST

The NIST Seabird Tissue Archival and Monitoring Project (STAMP) has collected egg contents for more than twenty years to create a geospatial and temporal record of conditions throughout areas of the northern Pacific Ocean. These have been collected and analyzed to monitor ubiquitous contaminants and other analytes as these species are consuming the same food as humans. In some areas, eggs are used as part of subsistence diet, serving a role in nutrition for indigenous peoples. Contaminant profiles in eggs are different across species, however, eggs are often not easily identifiable at the species level unless the bird is observed sitting on the nest; this represents a large point of uncertainty for wildlife managers and researchers alike. To address this issue, we employed machine learning techniques to develop a chemometric classification scheme for seabirds represented in the STAMP collection. We focused on 487 samples from 5 species originating in Alaska and Hawaii. To date, these samples are covered by more than 50,000 individual data points representing seven contaminant classes, collated into a curated chemometric database linked to data describing sample origins. Specifically, we developed models to identify a bird's genus, species, and geographic origin using only chemometric data. Our results suggest chemometric data, commonly generated as part of environmental monitoring efforts, provides sufficient information to enable identification of the genus, species, and geographic origin of tissue samples, with some confidence, when manual identification is not possible.

QSAR Modeling for Drug-Induced Liver Injury
Minjun Chen
Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, FDA

Drug-induced liver injury (DILI) is a leading cause of drugs failing during the drug development phase. Identifying the drug candidates associated with DILI potential in human in the discovery stage can improve the drug pipeline. Many efforts have focused on developing preclinical models to enhance prediction accuracy for a drug's DILI potential in human, including in silico models designed to prioritize drug candidates for further experiments. We developed a quantitative structure-activity relationship (QSAR) model using in-house developed decision forest algorithm coupled with the Molds2 chemical descriptors. 197 drugs were used to develop model, and the model was subsequently challenged by the left of drugs serving as an external validation set with an overall prediction accuracy of 68.9%. The performance of the model was further assessed by the use of 2 additional independent validation sets. The presented QSAR model, used together with other models developed in our group, could be useful in drug discovery to assess drug candidates for their DILI potential in humans.

The FDA/CDER Computational Toxicology Consultation Service Chemical Dictionary and Consult Database
Jahan Cooper
Center for Drug Evaluation and Research, FDA

For over 15 years the Computational Toxicology and Consultation Service (CTCS) within FDA/CDER has provided (Q)SAR toxicity predictions for CDER reviewers to aid in regulatory decision-making for drug impurities within new and generic drug applications. In order to more easily track and access the wealth of data CTCS generates each year, an internal chemical database—the CTCS Chemical Dictionary—was created to house chemical structures, consultation reports, and related toxicology data with a structurally searchable interface. Today, the Chemical Dictionary contains over 35,000 chemicals including proprietary and non-proprietary drug substances, drug impurities, extractables and leachables, degradation products, metabolites, and excipients. An alphanumerical code is assigned for each compound registered into the database. These unique identifiers reduce redundant entries, facilitate structure-based searches within the database, and allow linking of records or endpoint values to a structure. The database is made available to our internal collaborators through a web-client so they may find past regulatory decisions or other associated data for compounds of interest without the need for a specialized software installation.

ATOM: Integrated AI Modeling to Accelerate Cancer Drug Discovery
(https://doi.org/10.6084/m9.figshare.21431922)
Eric Stahlberg
Frederick National Laboratory for Cancer Research

The use of AI, high performance computing and growing levels of reliable data is creating a paradigm shift in the approach to drug discovery and development. The ATOM (Accelerating Therapeutics for Opportunities in Medicine) collaborative effort, a multi-agency public-private partnership involving Frederick National Laboratory for Cancer Research, several DOE labs, and supported by the Department of Energy and the National Cancer Institute is leading the way forward to deliver an open ecosystem to dramatically accelerate the identification and successful delivery of new molecules to improve health. Traditionally, the sequential laboratory-based process of discovery is being replaced with increasingly automated, predictive model-driven approaches, which enable simultaneous optimization of molecular properties and characteristics. Using a challenge project approach to drive innovation and development of the ecosystem, the ATOM team has successfully developed a model creation workflow, AMPL, to standardize model development, executed several projects to optimize compounds against multiple design criteria, developed approaches to identify novel drug candidates, and delivered models to the community for general use. The ATOM Research Alliance serves as the new hub for growing the collaboration into the future (www.atomscience.org).

A Hybrid In-Silico Approach for Identification of Novel Inhibitors of SARS-Cov-2
(https://doi.org/10.6084/m9.figshare.21430332)
Sankalp Jain
National Center for Advancing Translational Sciences, NIH

The National Center for Advancing Translational Sciences (NCATS) has been actively generating SARS-CoV-2 high-throughput screening data and disseminates it through the OpenData Portal (https://opendata.ncats.nih.gov/covid19). Here, we provide a hybrid approach that utilizes NCATS screening data from the SARS-CoV-2 cytopathic effect reduction assay to build predictive models, using both machine learning and pharmacophore-based modeling. Optimized models were used to perform two iterative rounds of virtual screening to predict small molecules active against SARS-CoV-2. Experimental testing with live virus provided 100 (~16% of predicted hits) active compounds (Efficacy > 30%, IC50 ≤ 15 µM). Systematic clustering analysis of active compounds revealed three promising chemotypes which have not been previously identified as inhibitors of SARS-CoV-2 infection. Further analysis identified allosteric binders to host receptor angiotensin-converting enzyme 2, which were able to inhibit the entry of pseudoparticles bearing spike protein of wild type SARS-CoV-2 as well as South African B.1.351 and UK B.1.1.7 variants.

Application of (Quantitative) Structure-Activity Relationships to Pharmaceuticals at FDA/CDER (https://doi.org/10.6084/m9.figshare.21430383)

Naomi Kruhlak
Center for Drug Evaluation and Research, FDA

This presentation describes the regulatory use of (quantitative) structure-activity relationship, or (Q)SAR modeling in the safety assessment of impurities and contaminants in drug products. A general overview of the (Q)SAR modeling methodology is provided, as well as information on the interpretation of model predictions to support regulatory decision-making. Key factors that have led to the regulatory acceptance of (Q)SAR modeling for pharmaceuticals are discussed, and the importance of chemical databasing to support a comprehensive (Q)SAR analysis workflow is highlighted. Lastly, success stories are presented on the value of collaborative research in advancing the state-of-the-science in this area.

Bayesian Methods for Enhanced Model Fitting and Uncertainty Analysis of Simulation Data

Jacob I. Monroe
NIST

Molecular simulations have become important tools in the prediction of properties of fluids and materials. A great strength of these techniques lies in their ability to provide insight into the molecular and atomistic details, which are often missing from experiments, that contribute to structural and thermodynamic properties. Calculation of properties as a function of state conditions, such as temperature and density, is a common task that we show is beneficially accelerated through Gaussian Process Regression (GPR). In contrast to many applications of Gaussian Processes to data fitting, uncertainty estimates and derivative information from molecular simulation results may be leveraged as inputs to a GPR. We highlight methods and best practices for incorporating this information into GPR models, finding that the accuracy of uncertainty estimates plays a key role in determining the behavior of the fit. With these lessons in hand, we demonstrate how active learning based on GPR models can efficiently direct additional simulations to achieve high certainty in estimates of properties over prespecified ranges of state conditions.

Scaling Cheminformatics Applications on FDA CDRH HPC Clusters (https://doi.org/10.6084/m9.figshare.21430326)

Mike Mikailov
Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, FDA

The recent explosion of chemical libraries of National Cancer Institute (NCI) beyond a billion molecules led to large scale simulations for Virtual Screening (VS). VS is a simulation technique used in drug discovery to search libraries of molecules to identify structures, likely to bind to a drug target. Over 950 years is needed for processing the billion-sized libraries. The FDA CDRH High-Performance Computing team is working with NCI to apply scaling techniques and use more powerful hardware resources (i. e., GPUs) make this mission critical task feasible and accomplishable in timely manner. Using the scaling techniques described in this presentation, would potentially reduce the 950 years of VS time to 70 days on 5,000 CPU cores on the FDA CDRH HPC clusters.

Encoding Chemical Information in The Age of AI

Raul Cachau
National Institute of Allergy and Infectious Diseases, NIH

There are many molecular representations, such as InChI, SMILES, topological descriptors, 3D descriptors, Molfiles, etc., which have their applications for indexing, discovery and optimization of drug-like molecules. Are they optimal for working with modern technologies, such as convolutional neural networks (CNN)? The presentation discusses requirements for CNN inputs in general and invites the cheminformatics community to suggest chemical encodings which fulfill the requirements. It emphasizes that encodings producing best CNN results may be computationally expensive and challenges the scientific community to produce reusable data.

Delivering Chemical-Associated Data Via EPA Web Applications
(http://doi.org/10.6084/m9.figshare.21429414)

Antony Williams
Office of Research and Development, Center for Computational Toxicology and Exposure, EPA

As part of its mission the Center for Computational Toxicology and Exposure (CCTE) delivers access to chemicals related data via online Dashboards. The CompTox Chemicals Dashboard (available at https://comptox.epa.gov/dashboard) provides access to >900,000 chemicals and associated data including experimental and predicted property data, in vivo hazard data, in vitro bioactivity data, exposure data, and various other data types. The application provides a set of flexible searches allowing for search, visualization and downloads of the data to the desktop for further interrogation. This presentation will provide an overview of the Dashboard and other proof-of-concept applications. For example, the proof-of-concept cheminformatics modules (https://www.epa.gov/chemical-research/cheminformatics) has a module which allows profiling of chemicals based on toxicity types (https://doi.org/10.1007/s10098-019-01795-w). This presentation will provide an overview of the CompTox Chemicals Dashboard and introduce proof-of-concept modules in development.

Overcoming Scarcity in The Scientific Literature

Andrei Kazakov
Thermodynamics Research Center, Applied Chemicals and Materials Division, NIST

NIST Thermodynamics Research Center (TRC) supplies industry and academia with critically evaluated property data to inform manufacturing, public health safety, and research. Over 75 years TRC has been collecting, evaluating and curating essentially all thermodynamic and transport properties of organic and common inorganic compounds. It has been recognized that collection of data from the scientific literature typically requires a big effort of data interpretation, clean-up and structurization. In order to improve the quality of data at origin, a cooperation has been created between NIST and the scientific publishers to utilize the data communication format ThermoML (https://trc.nist.gov/ThermoML) for capturing thermodynamic study data from scientific articles before their acceptance for publication. This practice implemented by five journals significantly decreased the data reporting problems, such as substance identification, errors in describing experiments, errors in numerical properties and units of measure, typos, etc. The data obtained from the journal cooperation has been disseminated with compliance to FAIR principles through https://data.nist.gov. The challenges encountered by TRC included a lack of the end user’s preparedness to receive a complex structured data (“give me an excel spreadsheet” attitude), high pollution of data at its origin, no trend of the data quality improvement with an increase of the publication year or a size of a dataset.

Reaction Informatics Innovations in ASPIRE
Gergely Zahoranszky-Kohalmi
National Center for Advancing Translational Sciences, NIH

A Specialized Platform for Innovative Research Exploration (ASPIRE) (https://ncats.nih.gov/aspire) is a project developed by NCATS to transform chemistry from an individualized craft to a modern, information-based science. ASPIRE is designed to bring novel, safe and effective treatments to more patients more quickly at lower cost. The users access all functionalities through ASPIRE integrated Computational Platform (AICP), including Synthesis Planning Core, Reaction Knowledgebase, and Computer Aided Reaction Registration. For the future work, source code of API and UI will be released and more functionalities will be integrated to ASPIRE.

NIST Mass Spectral Libraries (https://doi.org/10.6084/m9.figshare.21431937)
Lewis Geer
Mass Spectrometry Data Center, NIST

The NIST mass spectral libraries are widely adopted data standards used for the identification of unknown chemical compounds. The NIST MSDC compiles these libraries from experimental data and uses cheminformatics and other approaches to develop algorithms used in the identification process.

Structures in ChemIDplus: Continuous Improvement in Small Steps
(https://doi.org/10.6084/m9.figshare.21430392)
Mitch Miller
National Library of Medicine, NIH

The ChemIDplus system has been providing searching of structures, names, locators and properties of selected chemicals related to human health since 1999. Our structure database contains over 350,000 records, entered over more than 20 years by several dozen contributors. We periodically evaluate tools to clean and standardize the structures. This talk covers a couple of different approaches that involve the use of SGroups – an extension to the well-established molfile standard that provides the ability to layer high-level information on top of the atoms and bonds. SGroups that convey no information were replaced with SGroups that add meaning or were removed.
Note: the ChemIDplus system will be retired at the of 2022. Please see https://www.nlm.nih.gov/pubs/techbull/ja22/ja22_pubchem.html for more information.

PrecisionFDA – A Platform Where Government Cheminformaticians Can Collaborate with Academia and Industry (https://doi.org/10.6084/m9.figshare.21430365)
Yulia Borodina
Health Informatics Staff, Office of Data, Analytics and Research, Office of Digital Transformation, Office of the Commissioner, FDA

PrecisionFDA is a secure, cloud-based, high-performance platform implemented by FDA for collaboration in the areas of biological, clinical and chemical informatics. It allows users to upload, download, and share files, create executable packages (apps) that can run scripts on files, share tools with other users, combine apps into multi-stage workflows, and participate in organized challenges sponsored by community groups. It has been designed for a diverse community of experts working with genomic datasets in order to advance precision medicine. It has expanded to assist the broader community of data scientists for analysis of clinical data and to host data-driven competitions that push the limits of tools and algorithms for large-scale data processing. The latest application of this platform provides analysis of chemical data sets, as well as benchmarking computational approaches to cheminformatics, drug discovery and toxicology. In fall 2022 precisionFDA will be hosting its first cheminformatics challenge "Crowdsourced evaluation of InChI based tautomer identification"

Report from the panel discussion

Panelists:
Andrey Kazakov (NIST)
Ann Richard (EPA/ORD)
Antony Williams (EPA/ORD)
Evan Bolton (NIH/NCBI)
Huixiao Hong (FDA/NCTR)
Marc Nicklaus (NIH/NCI)
Raul Cachau (NIH/NIAID)
Tammy Page (FDA/CFSAN)
Tyler Peryea (FDA/OC)

Moderator: Iwona Weidlich (FDA/CDER)

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

The US Governmental organizations have collectively accumulated millions of chemical structures and their properties in chemical databases, and they have developed numerous computational approaches based on chemical data. Yet, cheminformatics is not sufficiently recognized by governmental organizations as an important field. Chemists working in the Government are often not aware of cheminformatics data resources, standards, and good practices, and don't use them in their everyday work. What can be done to improve the awareness of existence and use of these resources?

- Educating chemists with regard to cheminformatics approaches in the form of regular scientific meetings, workshops, webinars, etc.
- Bringing cheminformatics to end users in more convenient ways -- creating graphical interfaces, providing more accessibility and findability on the Internet
- Providing methods that can be integrated with chemical synthesis robotics
- Developing cheminformatics approaches that provide immediate benefits to the end users

Most organizations represented at the meeting (such as NIH, EPA, FDA) maintain several chemical databases varying in scope and data life cycle. Big efforts are being made to extract chemical data from literature and non-structured data submissions, mapping data between databases, and data curation. Regulatory agencies tend to stay isolated from the rest of the cheminformatics community and between themselves because of the need to protect intellectual property of businesses whose data they regulate. This leads to the creation of silos of information and duplication of efforts. What can be done to reduce and de-duplicate the effort?

- Addressing specific challenges of regulatory organizations dealing with confidential data by exchanging skill sets instead of exchanging datasets
- Separating out confidential business information, so that the vast majority of the data could be made public
- Publishing information in a machine-readable format
- Facilitating integration of resources through promulgating cheminformatics data standards
- Using AI for replacing human judgement for curation of data wherever possible

There are a number of chemical databases that maintain unique chemical identifiers. Examples are government databases (PubChem, GSRS, EPA DSSTox/Dashboard) and the commercial database highly utilized by the Government (CAS registry). However, each entity defines chemical substances with different degrees of granularity, therefore using the identifiers for mapping chemical records cannot guarantee a lossless exchange. Machine-readability of chemical data does not necessarily guarantee a lossless exchange either, because different systems use different machine-readable formats. The problem becomes bigger when we look at physicochemical properties and chemical processes. Government organizations and programs may curate data with methodologies developed in-house, which can lead to modification of the original data in different ways depending on the task of the program performing the curation and the curation method. How can we reduce data discrepancy between sources and make knowledge extraction from the data more efficient?

- Utilizing CAS Registry as the source of truth for CAS Registry Numbers and not relying on CAS numbers republished by other databases
- Making data not just machine-readable but also machine-processible by using data standards, ontologies, controlled vocabularies, etc.
- Exchanging information about standards and practices in Government forums such as this one, and also via publications
- Replacing old data formats/encapsulators (e.g., SDF) with modern data standards more appropriate for use with modern AI applications

Cheminformatics computational resources have historically been developed by the private sector. There are many computational firms that develop and support commercial cheminformatics solutions. Computational resources developed by Government cheminformaticians are usually publicly available but often not sustainable for prolonged use and maintenance. What are the pros and cons of in-house and collaborative development vs. using commercial software?

- Maintaining one's own software is beneficial because it can be adjusted more easily to meet new requirements, however this is time-consuming. Commercial providers have a lot of resources; creating similar capabilities in-house would require a lot of effort.
- For collaborative development with companies, the Government should not agree to let the commercial entity take data out of the public domain in order to commercialize them, not even in exchange for free access to software for the Government institution