NTP
National Toxicology Program

Data Storage, Analysis, and Interpretation

Co-Chairs:

Dr. Pauline Gee

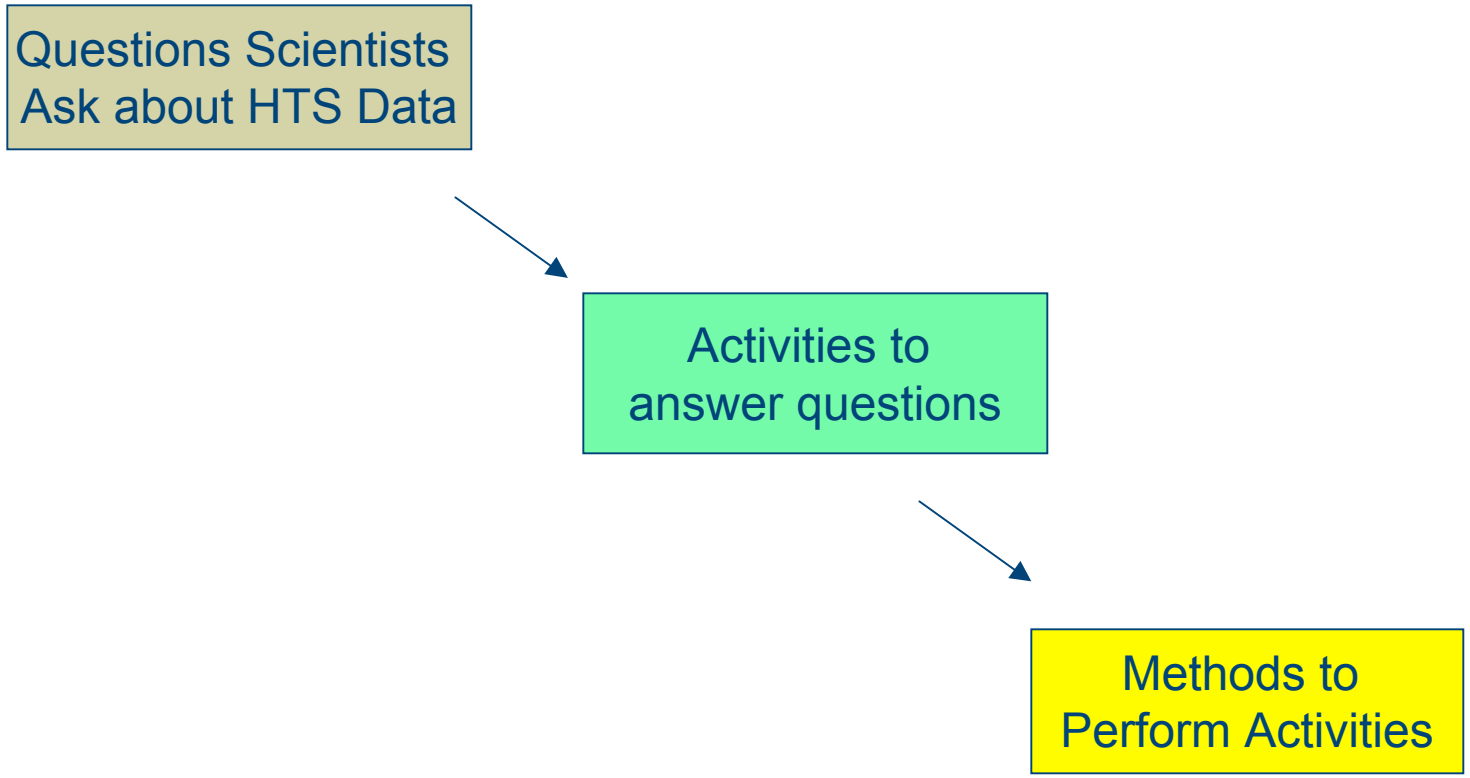
Dr. Alexander Tropsha

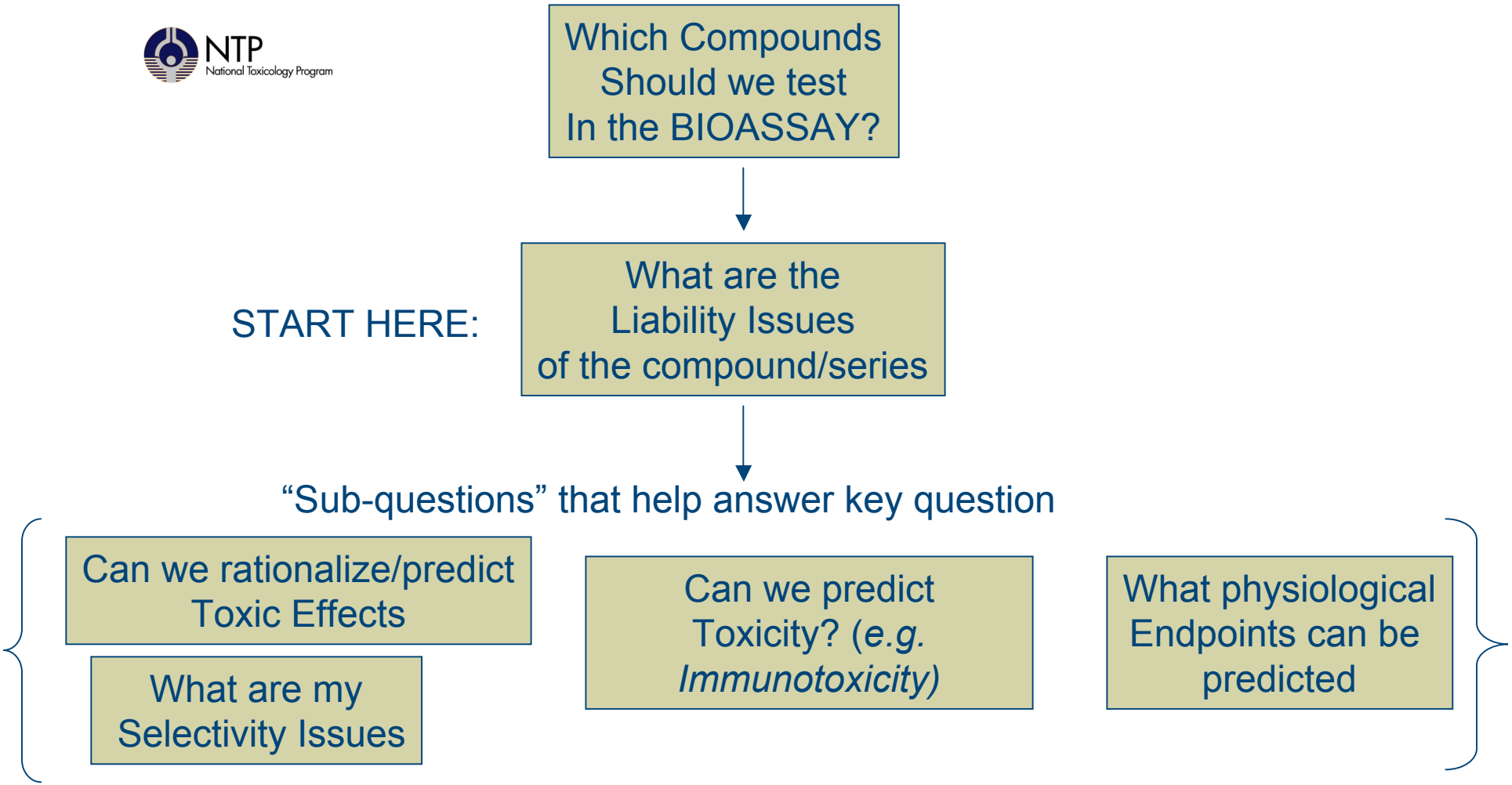
NTP
National Toxicology Program

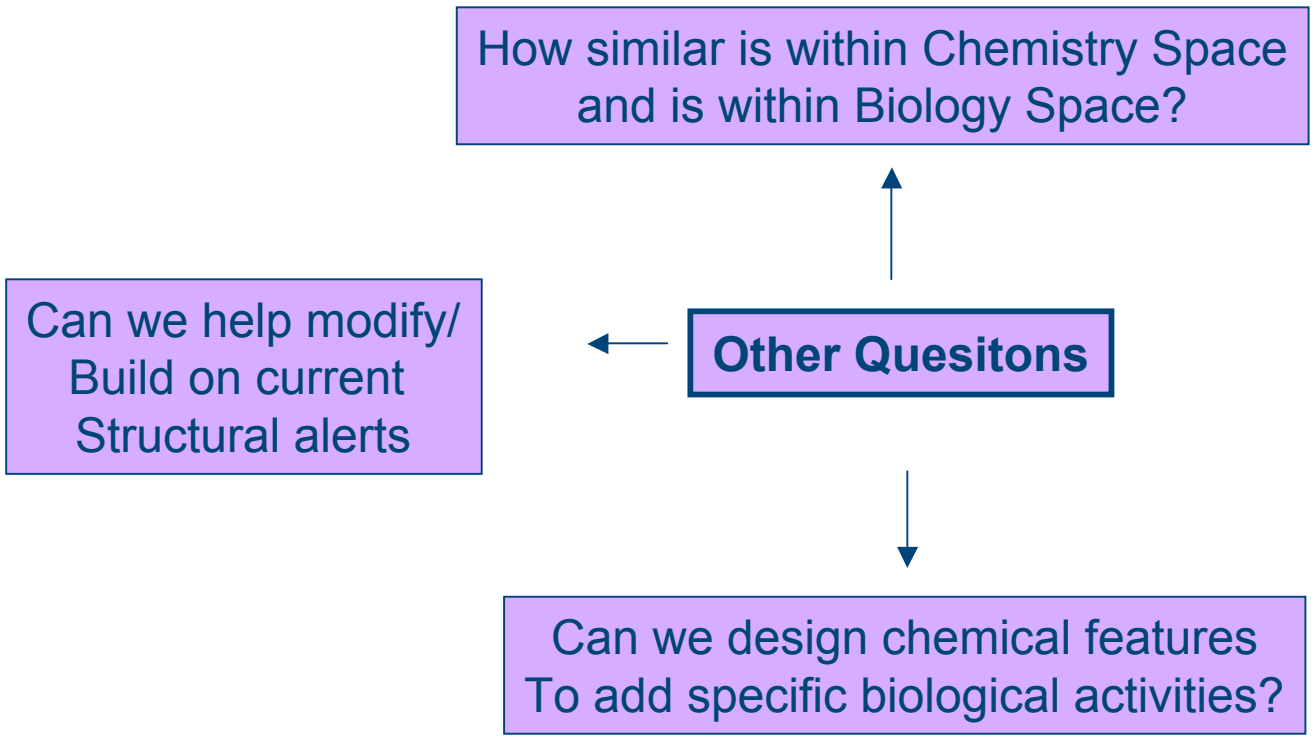
Interpretations to Support Decisions

- The ultimate goal is to provide enough critical scientific support to make public health decisions

- Prioritize compounds for *in vivo* testing / Bioassay
- Implies that information coming from HTS will be valuable
- Key Questions/queries of a database:


Questions Scientists Ask about HTS Data
Activities to answer questions
Methods to Perform Activities


NTP
National Toxicology Program
Which Compounds Should we test In the BIOASSAY?
START HERE:
What are the Liability Issues of the compound/series
"Sub-questions" that help answer key question
Can we rationalize/predict Toxic Effects
What are my Selectivity Issues
Can we predict Toxicity? (e.g. Immunotoxicity)
What physiological Endpoints can be predicted


How similar is within Chemistry Space
and is within Biology Space?
Can we help modify/
Build on current
Structural alerts
Other Quesitons
Can we design chemical features
To add specific biological activities?

NTP
National Toxicology Program

Issues

Effect of Missing Data

What is the effect of assay noise

How do results track for duplicates

Data Pre-processing Prior to analysis

Missing Data imputation techniques

Effects of assay selection

Accuracies and Errors In Data

What is the optimal Screening panel makeup?

Accuracy and errors In Predictions

NTP
National Toxicology Program

NTP HTS Priorities in Context of MLI HTS

This discussion is focused on the:

1. storage
2. analysis and
3 interpretation

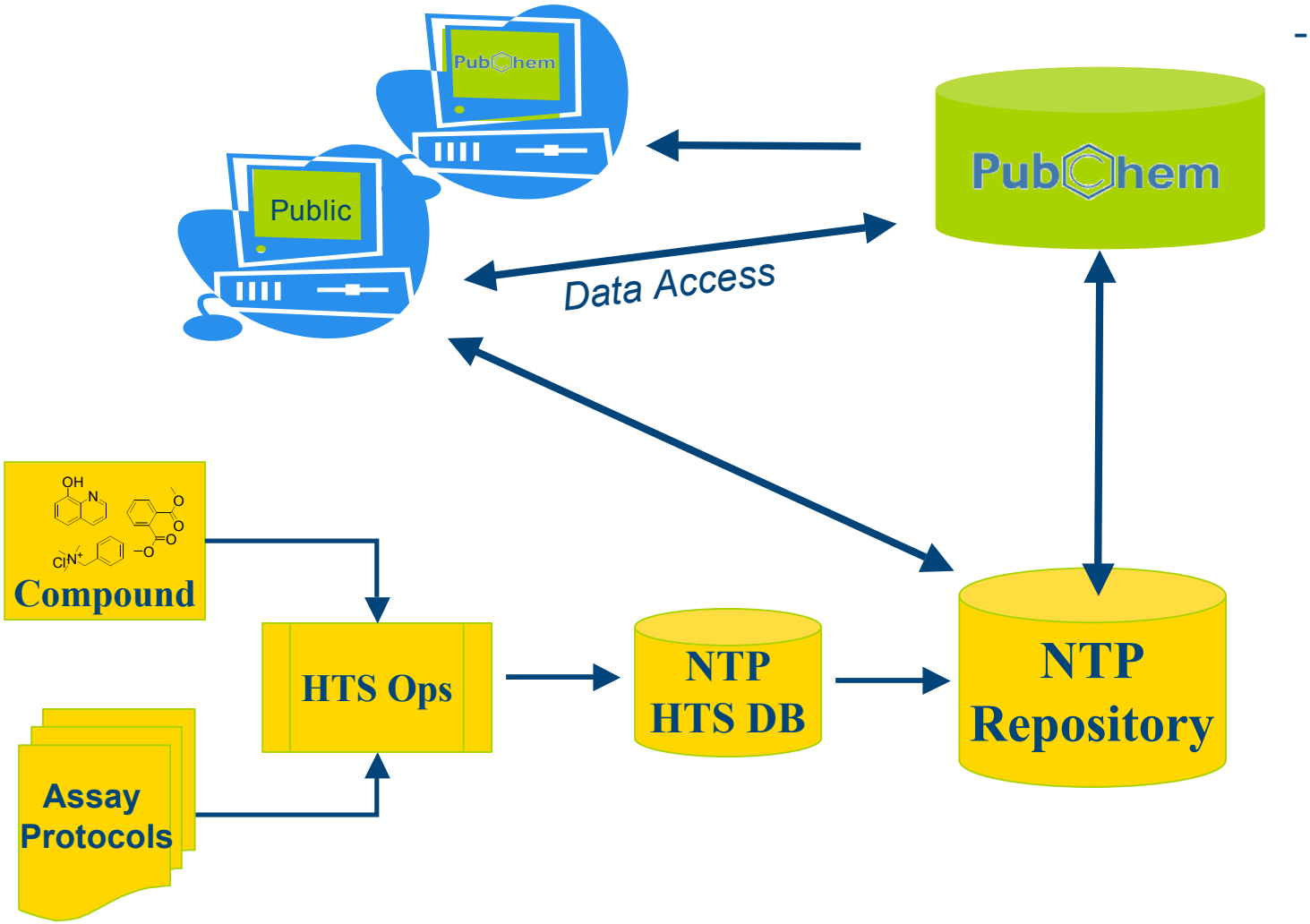
- of HTS data for the first 1,408 of perhaps 100,000 compounds submitted by the NTP
- to 1 of about 10 screening centers (~20 assays per center)
- set up as part of the Molecular Libraries and Imaging (MLI)
- to accelerate medical discovery to improve health in the NIH Roadmap Initiative Prioritize compounds for *in vivo* Study/Bioassay

NTP
National Toxicology Program

Storage of HTS Data

- The HTS data for the NTP sets of compounds will be acquired for each assay at the screening centers and stored in PubChem formats
- To provide the most accurate use and to maximize the value for modeling of the data it should be stored in their:
  - Original and raw form
  - Normalized and summarized data
  - Appropriate annotation including skeleton protocol
- ALL of the data should be stored in a publicly accessible database
- A major pitfall in storage of data that is not easy to access and retrieve that resources

NTP Recommended Data Flow



NTP
National Toxicology Program

Analysis of HTS Data

- The data for each plate will be normalized according to the specific protocol dictated by the assay at the screening centers.
- Similarly the statistical analysis normally used to handle plate-to-plate differences will be used for the NTP set taken into account
  - The dose response curves
  - The doses at the higher ranges
  - The higher number of replicates at each dose
- The outliers identified and eliminated using Assay specific QA/QC protocols?
- Yes, use the NIH analysis pipeline that is compatible with the data storage formats in PubChem
- REFERENCE: Gunter *et al.*, 2004

NTP
National Toxicology Program

Reduction of HTS Data

- Guidelines used to reduce the primary set of data:
  - ANOVA (various statistical cutoffs appropriate to the assay)
  - For example, use 3 sigma cutoffs (Bill Janzen)
  - For Dose Response Data, select 1 or 2 summary parameters

NTP
National Toxicology Program

Linkage of HTS Data with Bioassay Data

- Reconciling the different responses when storing and analyzing the NTP data set in one database with the rest of the MLI data:

Advantages	Disadvantages
Easier to administer and maintain	Separate db support
	Track Version numbers
Likely faster retrieval	Query multiple time
Ability to relate to the larger data	Cannot Customization and may be Easy to analyze

- The HTS data on NTP compounds must be linked to the NTP databases to obtain maximal value in prioritizing NTP compounds for further testing. This is a first step to building an engine to predict well enough to rank compounds for toxicity testing criteria in a bioassay.

NTP
National Toxicology Program

Linkage of HTS Data with NTP Data

- **What would be required to link the HTS set of data with the BIOASSAY data (all within the boundaries of the NTP compounds)**
- **"Essential" Steps in handshaking include:**
  - **Translation of CAS numbers to standard machine readable structures such that the bioassay data can be retrieved at the same level as the HTS data**
  - ***In vivo* data needs to be organized in a uniform ontology framework of pathology (and other domains) consistent across all studies and needs to machine readable.**

NTP
National Toxicology Program

What Components are Goals for Prediction?

- Given the complexity of the types of data likely to result from the NTP HTS initiative, do you see common themes (or can you suggest approaches) in how we can analyze these data to address NTP priority setting and prediction?
- Given the complexity and exquisite detail gathered in an NTP study for a single given compound (or stressor):
- Once:
  - the critical components that contribute to a judgment call are identified –
- THEN:
  - Appropriate statistical analyses for these different data types can be determined (novel statistical research needed)
  - Identify relationships amongst the critical components
- within the complexities of the internal data that describes the toxicities?

NTP Database Site Map