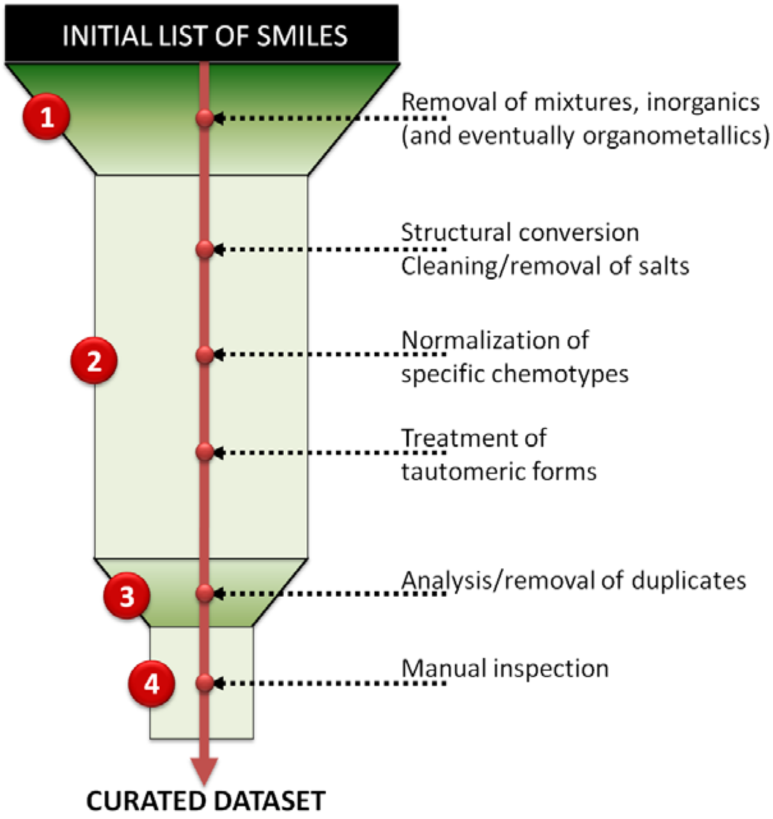
"Chemical Data Curation in Chemoinformatics"

Denis Fourches, *Laboratory for Molecular Modeling, Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA. Contact: fourches@email.unc.edu*

Molecular modelers and chemoinformaticians typically analyze experimental data generated by other scientists. Consequently, when it comes to the accuracy of experimental data, chemoinformaticians are always at the mercy of data providers who may inadvertently publish (partially) erroneous data. Thus, dataset curation is crucial for any chemoinformatics analysis such as similarity searching, clustering, QSAR modeling, virtual screening, etc. Despite the obvious importance of this preliminary step in the computational analysis of any dataset, there appears to be no commonly accepted guidance or set of procedures for chemical data curation.

The main objective of this training session is to emphasize the need for a standardized chemical data curation strategy that should be followed at the onset of any molecular modeling investigation. We will discuss several important steps for cleaning chemical records in a database including the removal of a fraction of the data that cannot be appropriately handled by conventional chemoinformatics techniques. Such steps include the removal of inorganic and organometallic compounds, counterions, salts and mixtures; ring aromatization; normalization of specific chemotypes; curation of tautomeric forms; and the deletion of duplicates.

To emphasize the importance of data curation as a mandatory step in data analysis, we will discuss several case studies where chemical curation of the original "raw" database enabled the successful modeling study (specifically, QSAR analysis) or resulted in a significant improvement of the model's prediction accuracy. We believe that good practices for curation of chemical records outlined in this training session will be of value to all scientists (especially graduate students) working in the fields of molecular modeling, chemoinformatics and QSAR studies.


INITIAL LIST OF SMILES
1
Removal of mixtures, inorganics (and eventually organometallics)
Structural conversion
Cleaning/removal of salts
2
Normalization of specific chemotypes
Treatment of tautomeric forms
3
Analysis/removal of duplicates
4
Manual inspection
CURATED DATASET

Training Session Program

- Brief overview of good practices for chemical structure curation.
- Removal of inorganics/organometallics using Chemaxon/JChem.
- Structural normalization using Chemaxon/Standardizer.
- Analysis and removal of duplicates using ISIDA/Duplicates.
- Hard cases (e.g., tautomers); the importance of manual inspection.

NB: We do not endorse any of the software packages mentioned in this work; however, we are naturally sensitive to the issue of software availability and did tend to select software that is freely available to academic investigators.