Development of a curated hepatic steatosis database (HSDB) to enable Quantitative Structure-Activity Relationship (QSAR) modeling

N.N. Tucker[1], H.J. Martin[2], V.M. Alves[2], E. Muratov[2], A. Tropsha[2]

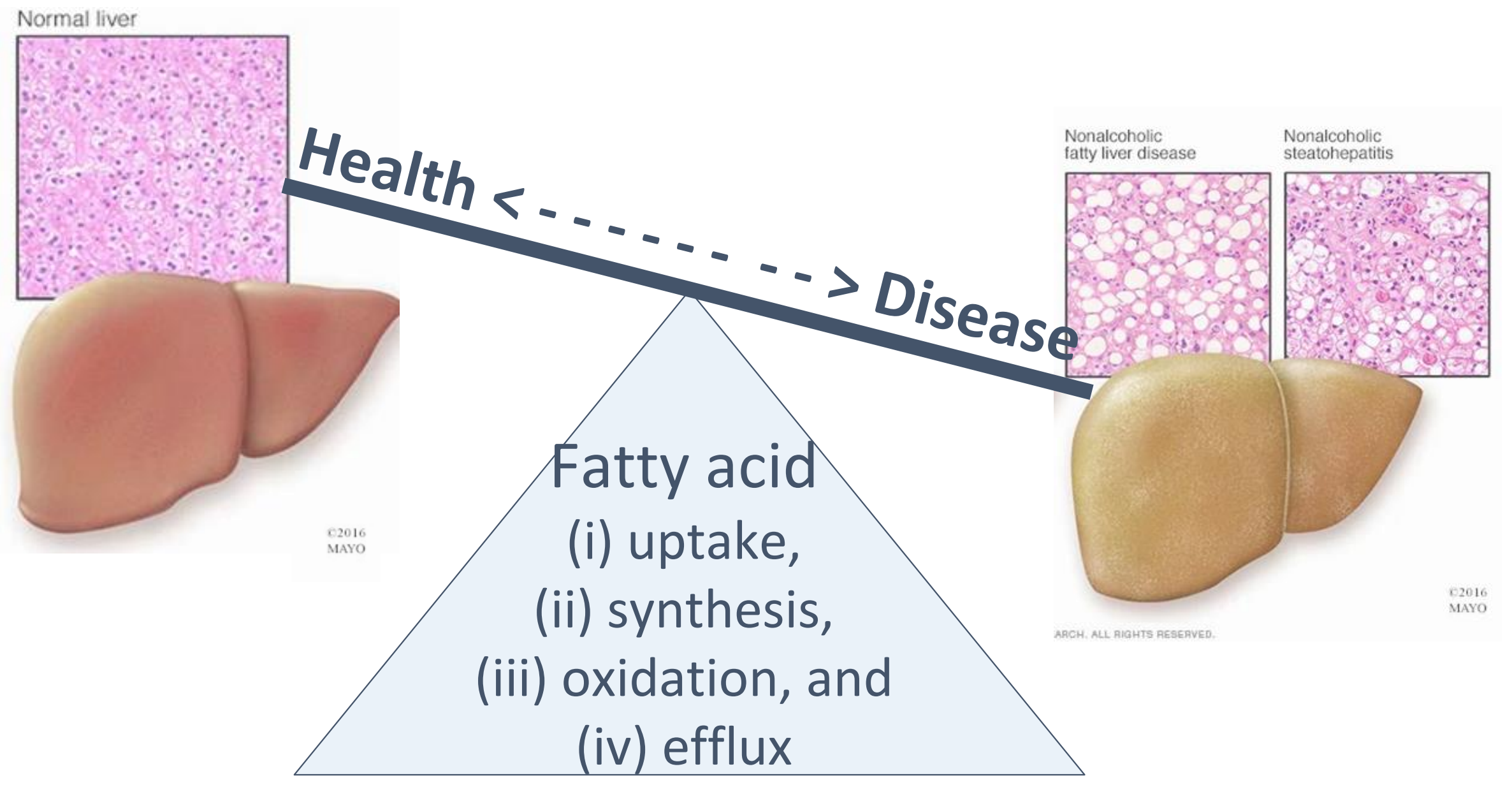
[1] Laboratory for Molecular Modelling, Curriculum of Toxicology and Environmental Medicine, University of North Carolina, Chapel Hill, NC 27599, United States;
[2] Laboratory for Molecular Modelling, Division of Chemical Biology and Medicinal Chemistry, Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC 27599, United States

virtual poster

*Email me at nyssa@unc.edu

Introduction

- Hepatic steatosis, also known as non-alcoholic fatty liver disease, is characterized by abnormal fat accumulation in the liver.

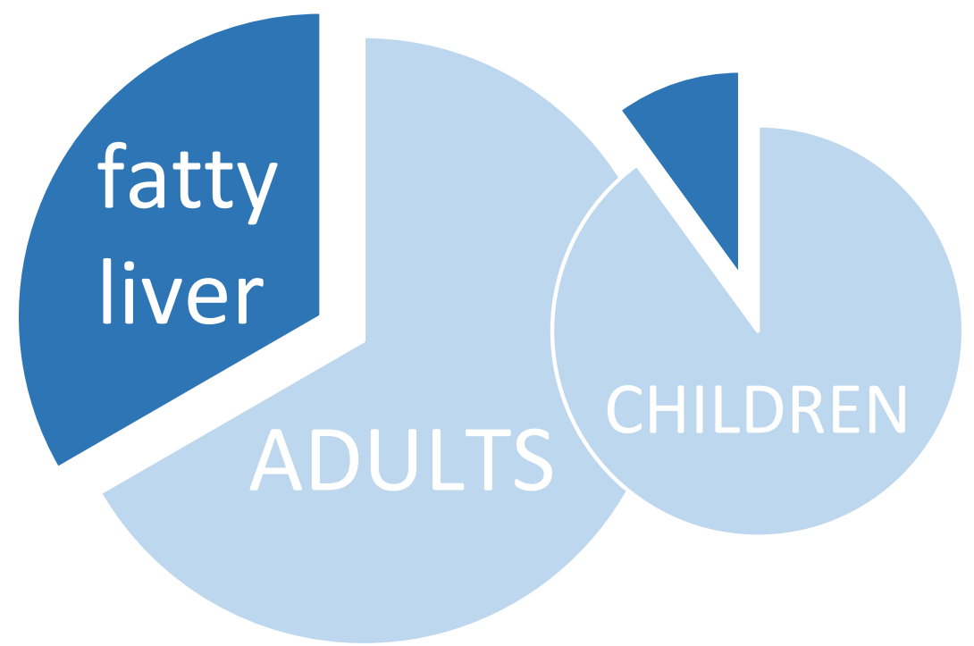

Normal liver
Nonalcoholic fatty liver disease
Nonalcoholic steatohepatitis
Health <- - - - - - - - -> Disease
©2016 MAYO

Fatty acid
(i) uptake,
(ii) synthesis,
(iii) oxidation, and
(iv) efflux

Four apical key events serve as the fulcrum potentiating additional disease outcomes of HS.

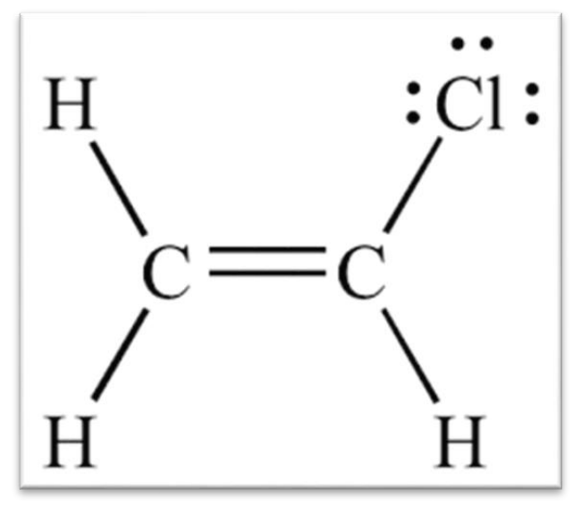
Angrish et al. 2016

- Disease impacts
  one in three adults
  one in ten children
  in the US.



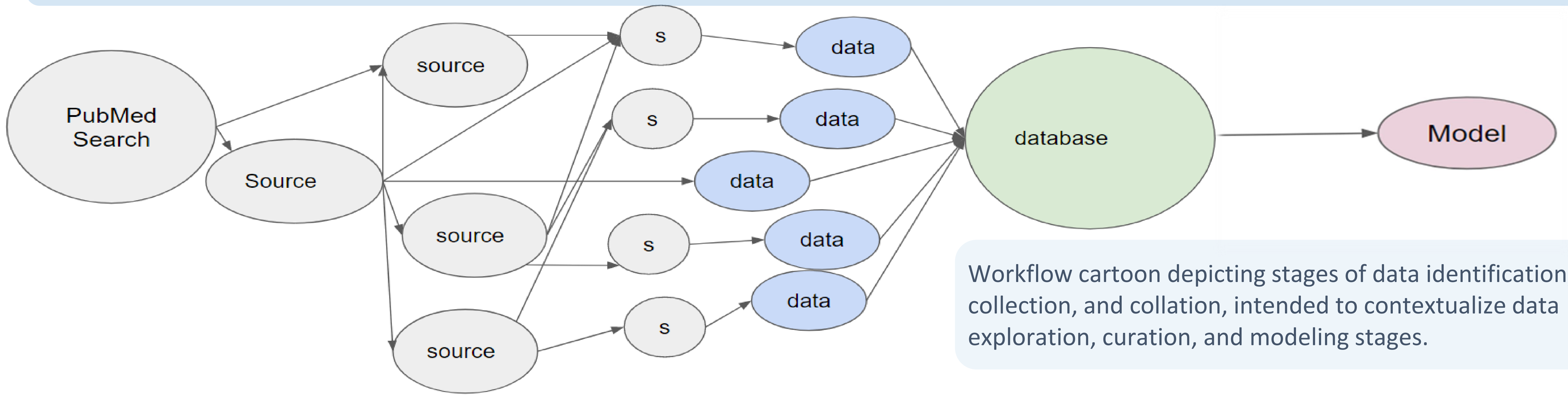
- Multifactorial causes / susceptibility sources:

Genetics, chemical exposure, psychosocial factors



- Hepatic steatosis can progress into an additional adverse outcome including fibrosis, cirrhosis, cancer, and death.

Materials and Method



Workflow cartoon depicting stages of data identification, collection, and collation, intended to contextualize data exploration, curation, and modeling stages.

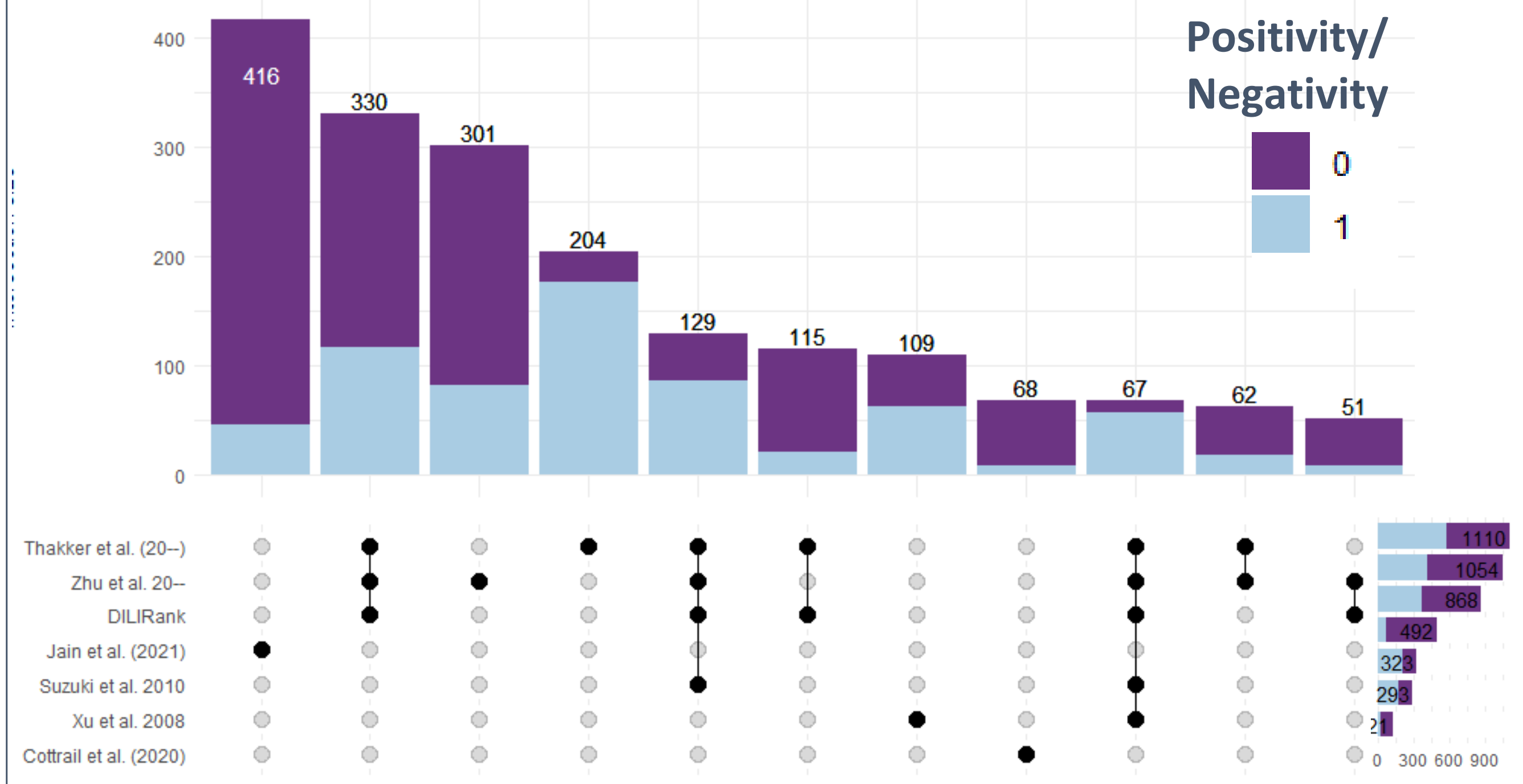
a. Data compiled via literature and web search using PubMed, supplementary materials, publicly accessible electronic databases, & private contributions.

b. Data integration, curation, analysis, and visualization executed in R and KNIME.

Results

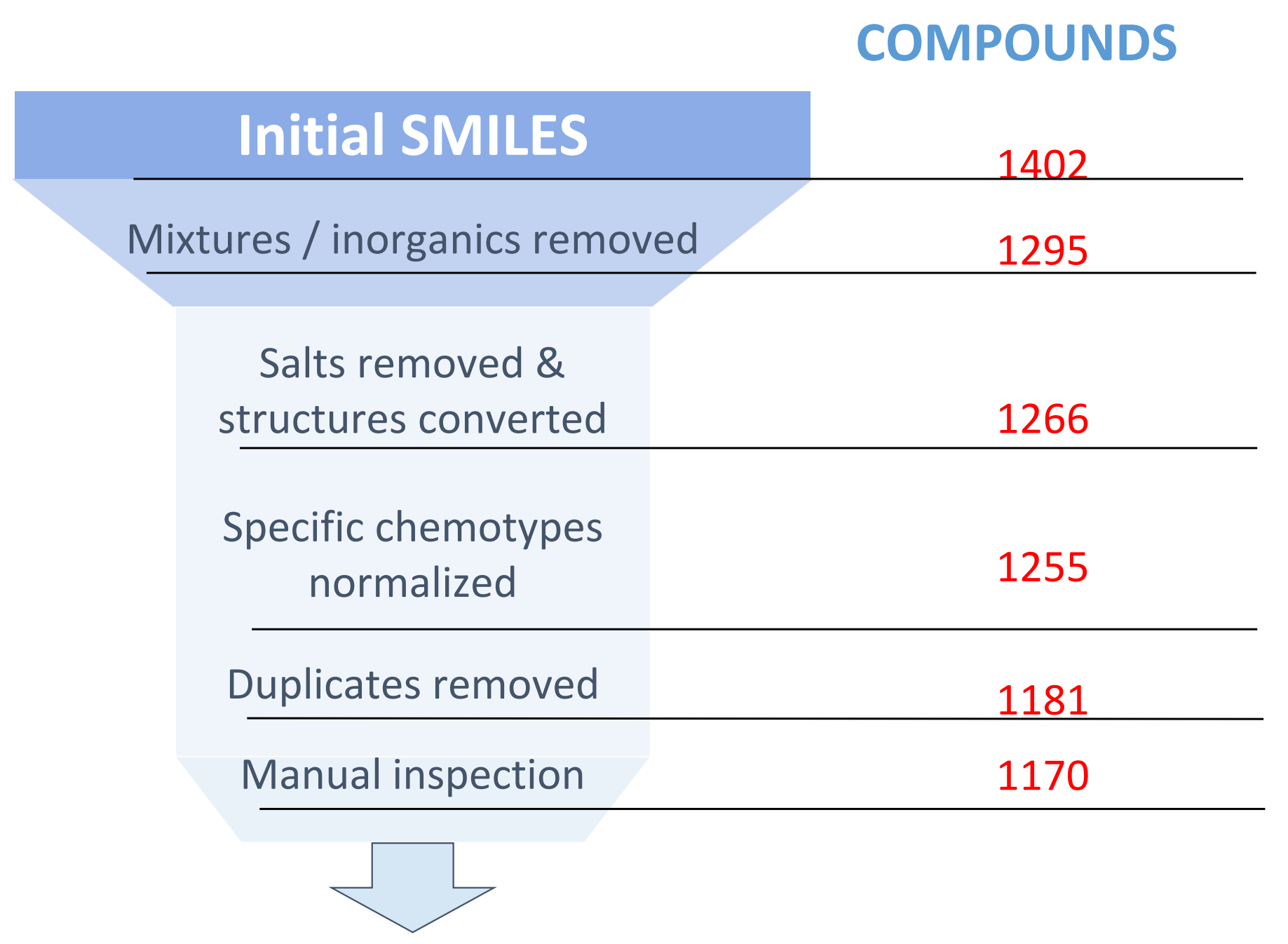
Data overview and curation

Data exploration



Chemical data overlap between top source datasets, visualized by count overlapping and entire set size, color coded by consensus steatosis positivity/negativity.

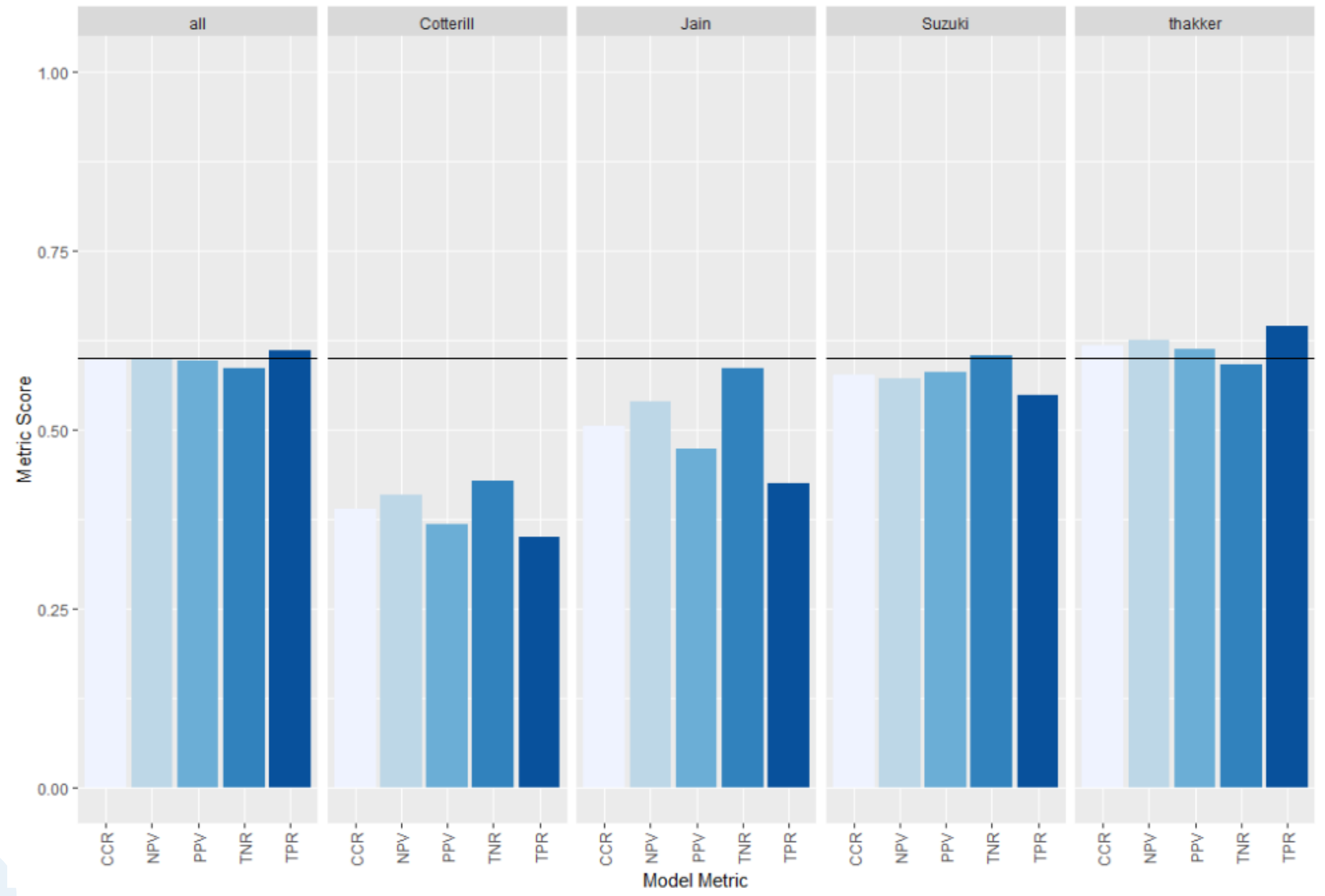
Key steps of curation


COMPOUNDS
Initial SMILES
1402
Mixtures / inorganics removed
1295
Salts removed & structures converted
1266
Specific chemotypes normalized
1255
Duplicates removed
1181
Manual inspection
1170

Compounds used for model development →

Summary of the chemical curation workflow, modified from [Fourches 2016]. Initial curation executed using subset of identified sources.

Data modeling



Models developed using similarity balancing, Random Forest with RDKit/Morgan fingerprints, and 5-fold external validation.

Conclusions

Here represents an attempt to collect, curate, and integrate the largest annotated liver steatosis dataset and use it to develop QSAR models to enable the accurate identification of novel potential steatosis causing agents.

Using public sources, developed the largest curated hepatic steatosis database incorporating 1170 unique compounds.

Future Directions

Data collection

○ Augment with known steatosis treatments

○ Annotate sources with assay variables, i.e. organism, endpoint, modality

QSAR Modeling

○ Silo models per data source hierarchy

○ Build individual event models

○ Model interpretation to identify statistically validated chemical moieties associated with HS.

Key References

- Angrish et al. 2016 10.1093/toxsci/kfw018
- Fourches et al. 2016 10.1021/acs.jcim.6b00129

" The views expressed in this presentation are mine and not official policy stances by NIH. "

" AT and ENM are co-founders of Predictive, LLC, which develops computational methodologies and software for toxicity prediction. "