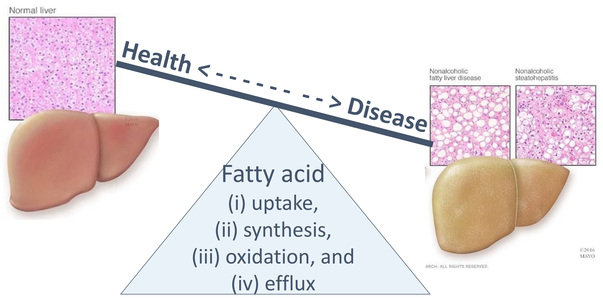
Development of a Curated Hepatic Steatosis (HS) Database & Quantitative Structure-Activity Relationship Modeling of HS Data

Nyssa N. Tucker[1*], Vinicius M. Alves[2], Eugene Muratov[2], Alexander Tropsha[2]

[1] Biological and Biomedical Sciences Program, UNC Chapel Hill , [2] Molecular Modelling Lab, UNC Chapel Hill

Introduction

- Hepatic steatosis, also known as non-alcoholic fatty liver disease, is characterized by abnormal fat accumulation in the liver.
- Disease impacts **one in three adults** and **one in ten children** in the US.
- Multifactorial causes include environment, diet, behavior, and genetics.



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

- HS can develop into adverse outcomes, including fibrosis, cirrhosis, cancer, and death.

We aimed to collect, curate, and integrate the largest chemogenomics HS dataset and use it to develop QSAR models of HS to enable the accurate identification of novel potential HS-causing agents.

Materials and methods

We performed extensive literature and web search, compiling data from:

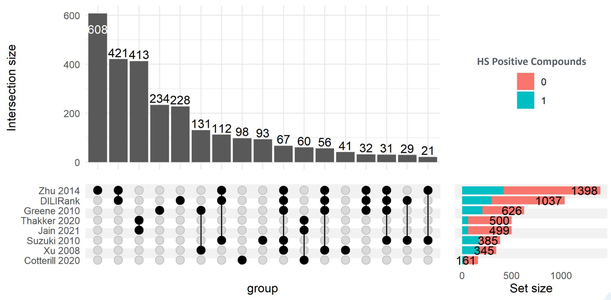
- Publications identified in PubMed
- Supplementary materials
- Publicly accessible electronic databases
- Private contributions

Data integration, curation, and analysis vis Knime, and visualization executed in R.

Results and Discussion

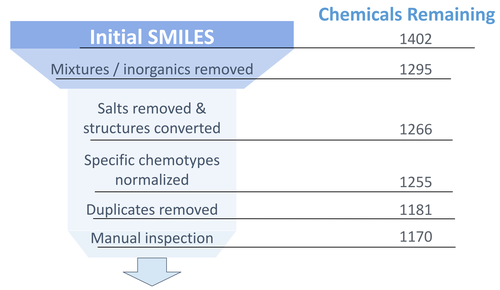
Data overview and curation

Data exploration



Chemical data overlap between top source datasets, visualized by count overlapping and entire set size color coded by HS positivity (0 = negative, 1 = positive). Visualized using R package: ComplexUpset.

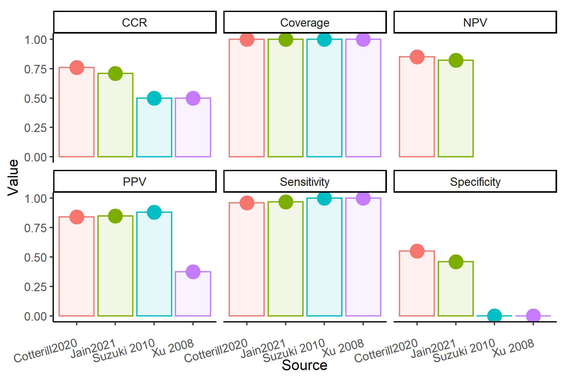
Key steps of chemical curation



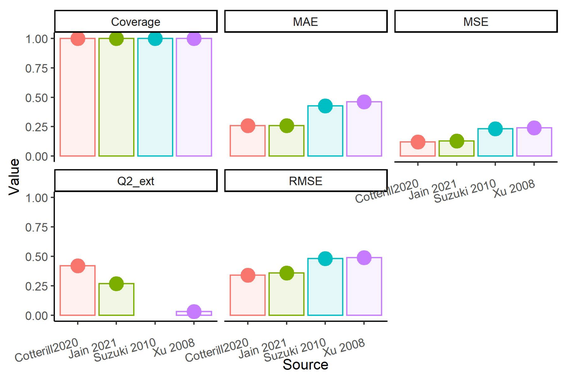
Summary of the chemical curation workflow, modified from [Fourches 2016]. Initial curation executed using subset of identified sources: Jain et al. 2021, Cotterill et al. 2020, Xu et al. 2008, and Suzuki et al. 2010.

Data modeling

Classification results



Regression results



Models developed using similarity balancing, RDKit and morgan fingerprints, and 5 fold partition random forest prediction.

Conclusions

- **Using public sources, developed the largest curated HS database incorporating 1170 unique compounds.**
- **Developed HS classification and regression QSAR models.**
- **Future studies include HS database enrichment and exploration of additional computational strategies to improve model accuracy.**

Future Directions

Data analysis

- HS database enrichment.
- Explore different data stratification strategies (e.g., by species).

Cheminformatics analysis and modeling

- Analyze SAR to identify chemical motifs related to HS.
- QSAR Modeling
  - Explore additional approaches to improve model accuracy.
  - Virtual screening of chemicals of interest
  - Model interpretation to identify statistically validated chemical moieties associated with HS.

Experimental validation

- Validate computational models using *in vitro* assays with EPA collaborator.

Key References

- Angrish et al. 2016 10.1093/toxsci/kfw018
- Fourches et al. 2016 10.1021/acs.jcim.6b00129
- Tropsha 2010 10.1002/minf.201000061
- Jain et al. 2021 10.1021/acs.chemrestox.0c00511
- Cottrill et al. 2020 10.1016/j.fct.2020.111494
- Suzuki t al. 2010 10.2165/11535340-000000000-00000
- Xu et al. 2008 10.1093/toxsci/kfn109

virtual poster

*Correspondence to: nyssa@unc.edu