A Cautionary Tale About Properly Vetting Datasets for Supervised Machine Learning Predicting Metabolic Pathway Involvement

Hunter N.B. Moseley, PhD

University of Kentucky

http://bioinformatics.cesb.uky.edu/

https://github.com/MoseleyBioinformaticsLab

UK Superfund Research Center

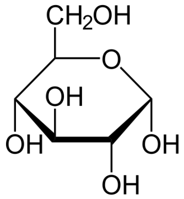
UK HealthCare MARKEY CANCER CENTER

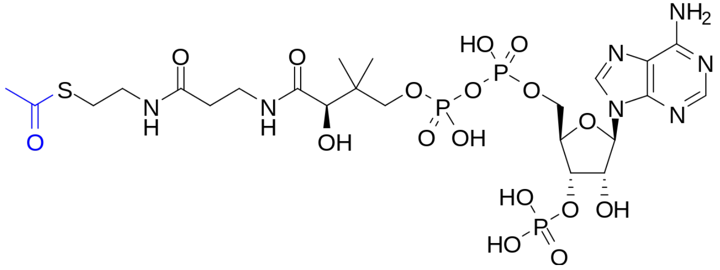
UK Institute for Biomedical Informatics

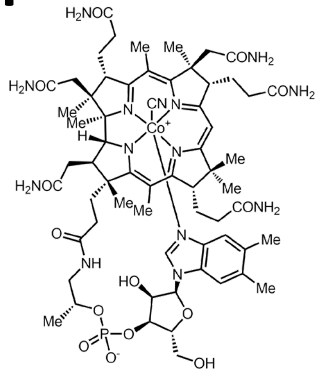
UK Center for Clinical and Translational Science

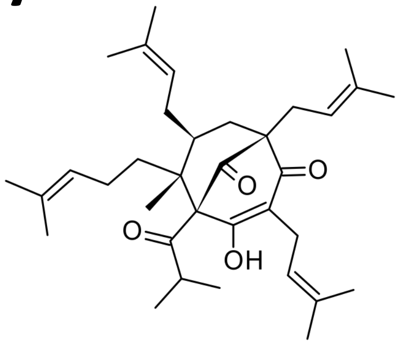
What is metabolomics and why is it hard to analyze?

- **Metabolomics is the systematic detection and characterization of small biomolecules generated from metabolism that are present in a biological sample.**
- **In comparison to other omics, the detected biomolecules are very chemically diverse and hard to comprehensively detect.**









- **Current metabolic databases are quite incomplete.**
- **Detection by any single analytical method (nuclear magnetic resonance spectroscopy or mass spectrometry) is grossly incomplete.**
  - **Systematic analysis of metabolites is limited by metabolite detection, database completeness, and availability of standards for identification.**

Given the difficulty, why use metabolomics?

Metabolomics provides a culminating molecular phenotype representing a final product of gene regulation and expression.

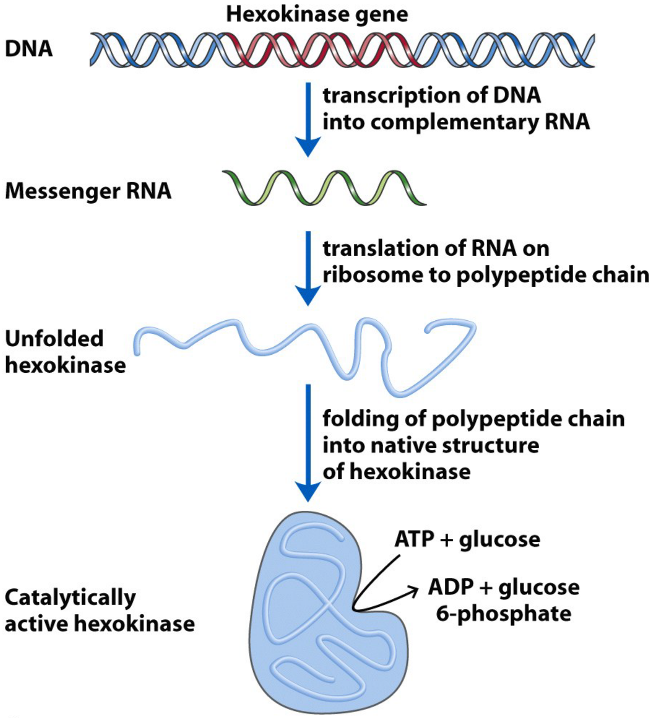

Hexokinase gene
DNA
transcription of DNA into complementary RNA
Messenger RNA
translation of RNA on ribosome to polypeptide chain
Unfolded hexokinase
folding of polypeptide chain into native structure of hexokinase
ATP + glucose
ADP + glucose 6-phosphate
Catalytically active hexokinase

Figure 1-31
Lehninger Principles of Biochemistry, Fifth Edition
© 2008 W. H. Freeman and Company

- **Allows a window into observing cellular and systemic metabolism.**
- **Changes in metabolism…**
  - **Reflect changes in cellular processes.**
  - **Typically occur on second and minute time scales.**
  - **Can be more easily achieved pharmacologically (via targeting enzymes).**
  - **Are a product of many disease processes.**
- **No model of a living system or process is complete without a metabolic component.**

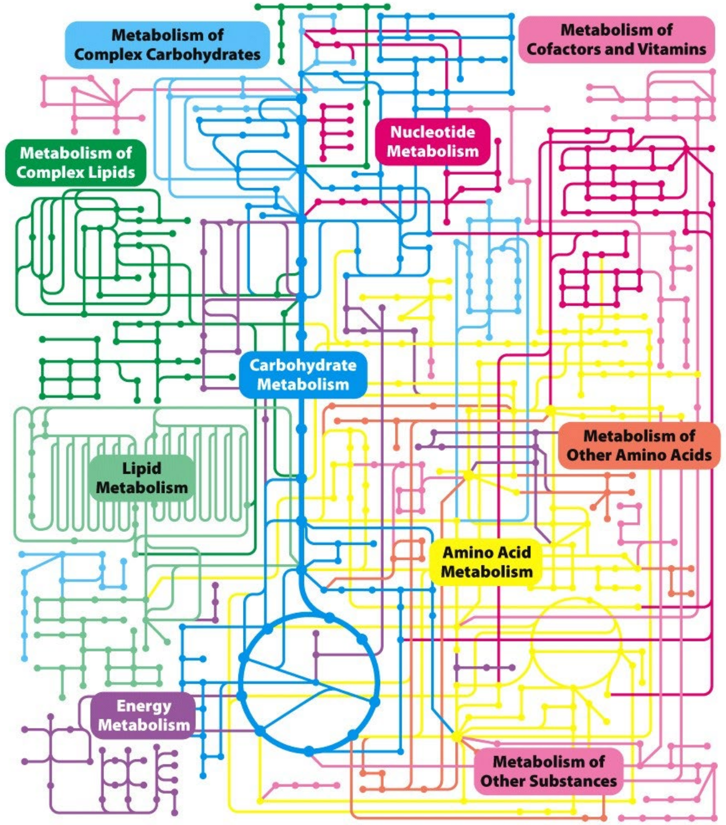
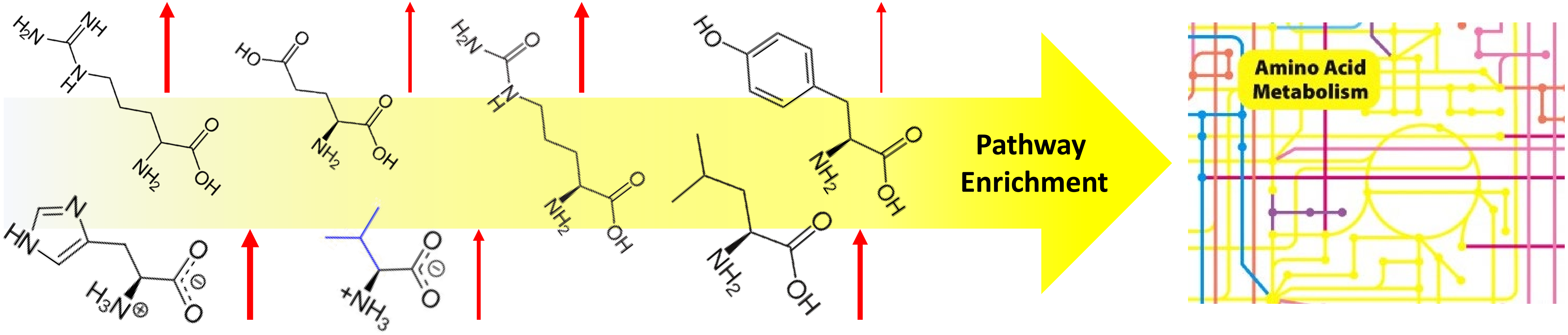

Metabolism of Complex Carbohydrates
Metabolism of Cofactors and Vitamins
Nucleotide Metabolism
Metabolism of Complex Lipids
Carbohydrate Metabolism
Metabolism of Other Amino Acids
Lipid Metabolism
Amino Acid Metabolism
Energy Metabolism
Metabolism of Other Substances

Figure 15.2
Biochemistry, Seventh Edition
© 2012 W. H. Freeman and Company

Metabolome Mining is (Potentially) an Easier Approach.

- **"Metabolome mining is defined as the use of metabolite features, with chemical and other annotations, to derive metabolic information that is interpretable in a biological or biomedical context."**
  - https://www.mdpi.com/journal/metabolites/topical_collections/metabolome_mining
- **Identifying metabolites associated with specific metabolic pathways enables metabolic pathway enrichment analysis.**



But most metabolites detected in metabolomics experiments do not have metabolic pathway annotations!

Exploring Current State of the Art in Metabolic Pathway Involvement Prediction

Model / Feature Set	Accuracy (%)	Precision (%)	Recall (%)	F1
Hu et al. RF [1]	94.64	77.97	67.83	0.7254
Baranwal et al. GCN/RF [2]	97.58 ± .12	83.69 ± .78	83.63 ± .68	0.8366
Baranwal et al. GCN [2]	97.61 ± .12	91.61 ± .52	92.50 ± .44	0.9205
Yang et al. GAT [3]	97.50 ± .06	93.04 ± .28	93.22 ± .16	0.9313
Du et al. MLGL-MP [4]	98.64 ± 0.47	95.26 ± 2.25	94.21 ± 1.94	0.9473

Standard deviation of the model performance metrics across CV folds indicated by the ± symbol, if available from the publication.
RF – Random Forest; GCN – Graph Convolutional Network; GAT – Graph Attention Network;
MLGL-MP - Multi-Label Graph Learning framework enhanced by pathway interdependence for Metabolic Pathway prediction

[1] Hu L-L, Chen C, Huang T, Cai Y-D, Chou K-C. *PLoS ONE*. 2011 Dec 29;6(12):e29491.

[2] Baranwal M, Magner A, Elvati P, Saldinger J, Violi A, Hero AO. *Bioinformatics*. 2020 Apr 15;36(8):2547–53.

[3] Yang Z, Liu J, Wang Z, Wang Y, Feng J. *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE; 2020. p. 126–31.

[4] Du B-X, Zhao P-C, Zhu B, Yiu S-M, Nyamabo AK, Yu H, et al. *Bioinformatics*. 2022 Jun 24;38(Suppl 1):i325–32.

All of these methods used a Kyoto Encyclopedia of Gene and Genomes (KEGG) derived dataset with SMILES chemical structure representations (KEGG-SMILES dataset).

KEGG-SMILES Dataset(s) Used

Model / Feature Set	Data available	Code available	Dataset Size	Publication Date
Hu et al. RF [1]	No	No	3,137	December 2011
Baranwal et al. GCN/RF [2]	Yes	Yes	6,669*	April 2020
Baranwal et al. GCN [2]	Yes	Yes	6,669*	April 2020
Yang et al. GAT [3]	No	No	6,669*	December 2020
Du et al. MLGL-MP [4]	Yes	Yes	6,648*	June 2022

*Publications using the dataset originating with Baranwal et al.

Data Leakage Problem in Baranwal KEGG-SMILES Dataset

Label ID	Pathway Category	Number Of Compounds In Dataset (Original)	Fraction Of Dataset (Original)	Percentage Of Duplicates	Number Of Compounds In Dataset (De-duplicated)	Fraction Of Dataset (De-duplicated)
0	Carbohydrate metabolism	1126	0.169	67.05	371	0.075
1	Energy metabolism	750	0.113	72.80	204	0.041
2	Lipid metabolism	1066	0.16	38.93	651	0.132
3	Nucleotide metabolism	342	0.051	49.12	174	0.035
4	Amino acid metabolism	1440	0.217	54.37	657	0.133
5	Metabolism of other amino acids	597	0.09	59.80	240	0.049
6	Glycan biosynthesis and metabolism	325	0.049	64.00	117	0.024
7	Metabolism of cofactors and vitamins	948	0.143	44.83	523	0.106
8	Metabolism of terpenoids and polyketides	1483	0.223	35.13	962	0.195
9	Biosynthesis of other secondary metabolites	1906	0.287	35.78	1224	0.248
10	Xenobiotics biodegradation and metabolism	1452	0.218	32.58	979	0.199
N/A	Total Dataset	6,648	N/A	25.86	4,929	N/A

Over 25% of the dataset are complete duplicates! This creates a catastrophic data leakage problem for training!

The Good, the Bad, and the Ugly!

The Bad

- A catastrophic data leakage was created within the Baranwal KEGG-SMILES dataset.

The Ugly

- This dataset affected at least 3 publications in highly reputable journals and conferences, since none of the authors properly vetted the dataset.

The Good (Silver Lining)

- Baranwal et al and Du et al followed many best practices for scientific reproducibility in computational research, enabling the detection of this catastrophically-flawed dataset and highly flawed results.
- These analyses are available in the following preprint and are under review:
  - Erik D. Huckvale and Hunter N.B. Moseley. "A cautionary tale about properly vetting datasets used in supervised learning predicting metabolic pathway involvement" bioRxiv 2023.10.03.560711 (2023).
- These findings prompted us to create a new benchmark dataset for metabolic pathway involvement prediction, which is also under review:
  - Erik D. Huckvale, Christian D. Powell, Huan Jin, and Hunter N.B. Moseley. "Benchmark dataset for training machine learning models to predict the pathway involvement of metabolites" bioRxiv 2023.10.03.560715.

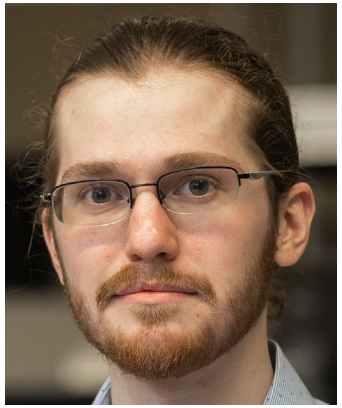
Acknowledgements

- Mr. Erik Huckvale did the majority of the KEGG-SMILES dataset evaluation and creation of a new benchmark dataset.
- Mr. Christian Powell started our metabolic pathway prediction project.
- Dr. Huan Jin provided atom coloring methodology used in the metabolic pathway prediction based on prior work done by Dr. Joshua Mitchell.
- Dr. Robert Flight provided feedback on the dataset evaluations.
- This work was supported by funding from:
  - NSF 2020026 (PI Moseley)
  - NIH NIEHS P42 ES007380 (UKSRC, PD Pennell)

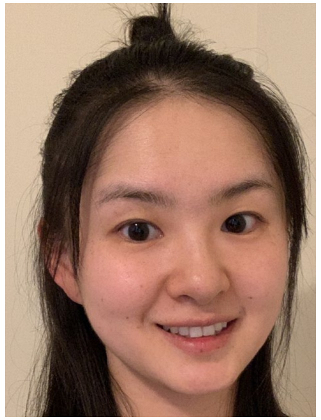
Hunter Moseley

Robert Flight

Joshua Mitchell

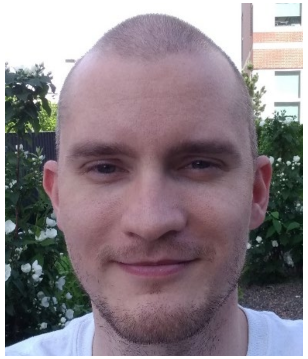


Christian Powell



Huan Jin

Andrew Smelter

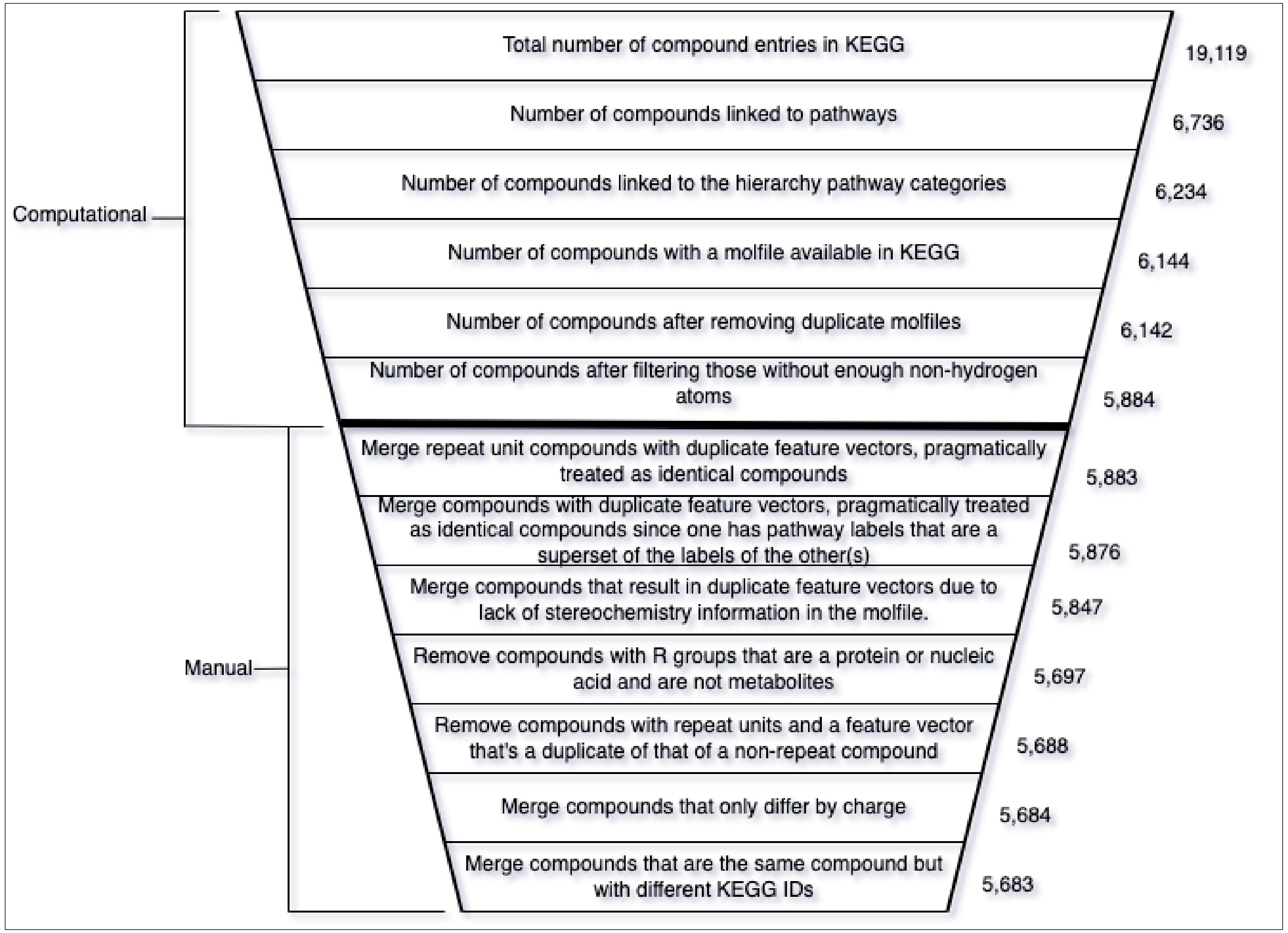


Travis Thompson

Erik Huckvale

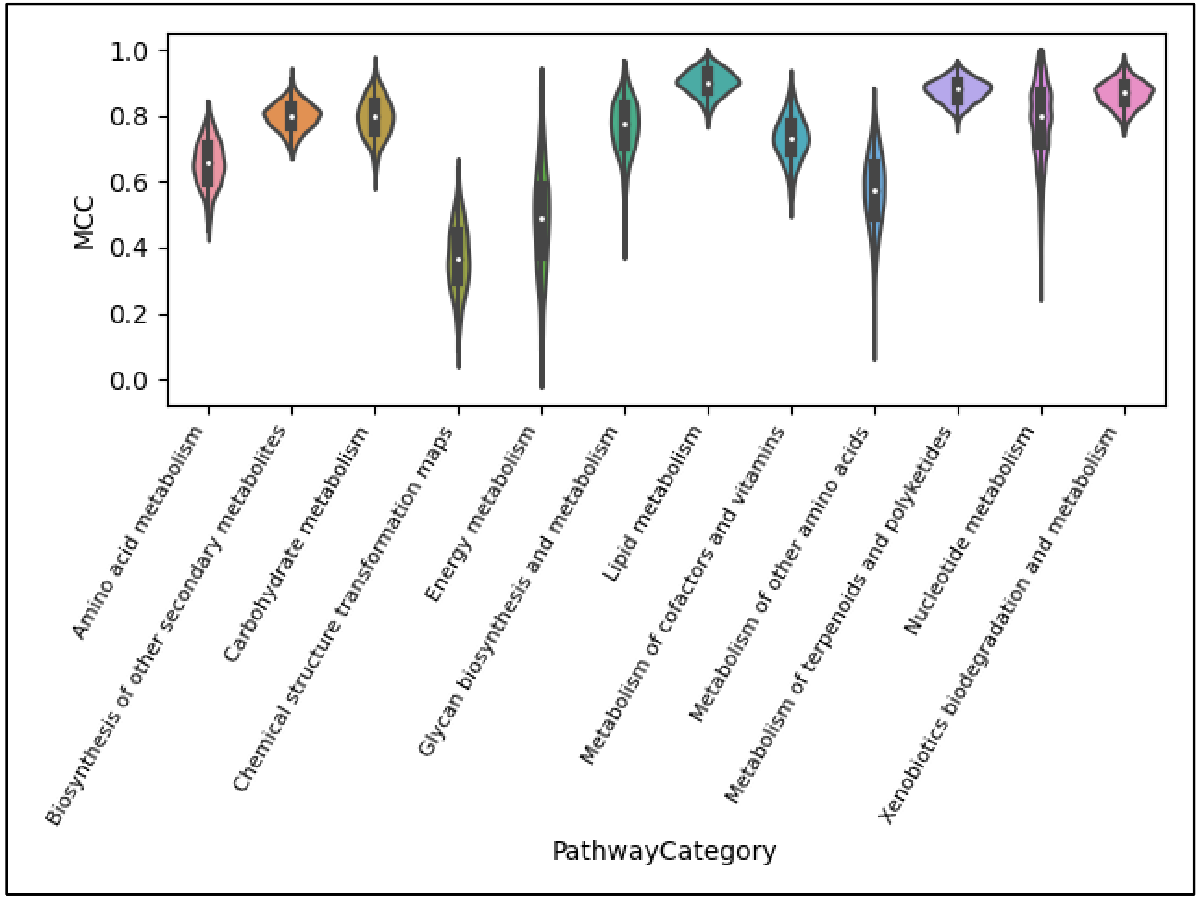
New Benchmark Dataset for Metabolic Pathway Involvement Prediction

Dataset Creation Workflow



Computational	Total number of compound entries in KEGG	19,119
	Number of compounds linked to pathways	6,736
	Number of compounds linked to the hierarchy pathway categories	6,234
	Number of compounds with a molfile available in KEGG	6,144
	Number of compounds after removing duplicate molfiles	6,142
	Number of compounds after filtering those without enough non-hydrogen atoms	5,884
Manual	Merge repeat unit compounds with duplicate feature vectors, pragmatically treated as identical compounds	5,883
	Merge compounds with duplicate feature vectors, pragmatically treated as identical compounds since one has pathway labels that are a superset of the labels of the other(s)	5,876
	Merge compounds that result in duplicate feature vectors due to lack of stereochemistry information in the molfile.	5,847
	Remove compounds with R groups that are a protein or nucleic acid and are not metabolites	5,697
	Remove compounds with repeat units and a feature vector that's a duplicate of that of a non-repeat compound	5,688
	Merge compounds that only differ by charge	5,684
	Merge compounds that are the same compound but with different KEGG IDs	5,683

XGBoost Performance Evaluation



Erik D. Huckvale, Christian D. Powell, Huan Jin, and Hunter N.B. Moseley. "Benchmark dataset for training machine learning models to predict the pathway involvement of metabolites" bioRxiv 2023.10.03.560715.

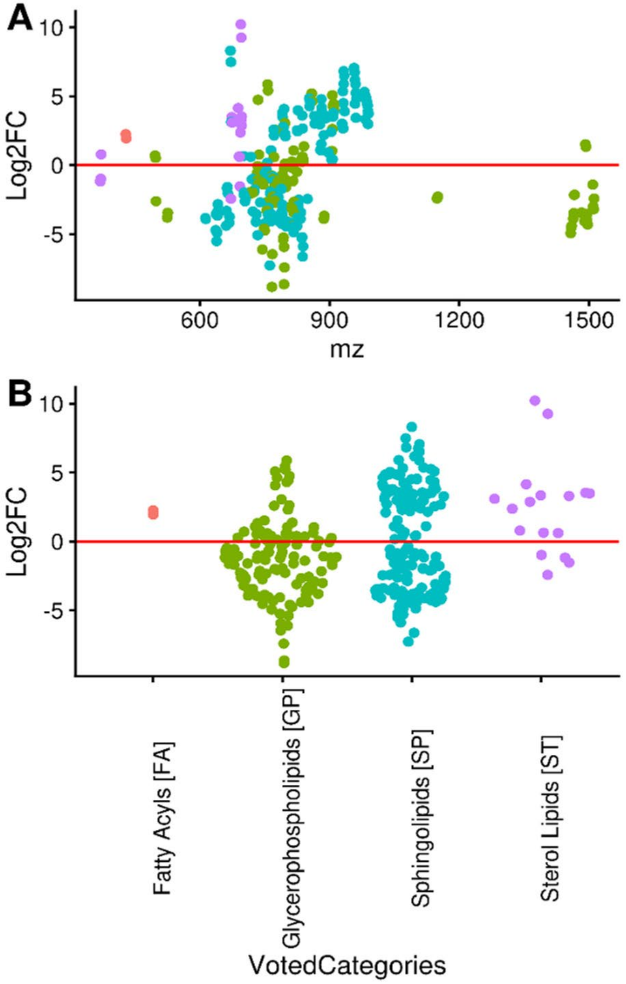
Untargeted lipidomics of non-small cell lung carcinoma demonstrates differentially abundant lipid classes in cancer vs non-cancer tissue

Joshua M. Mitchell, Robert M. Flight, and Hunter N.B. Moseley. *Metabolites* 11, 740 (2021).

- Most untargeted approach to metabolomics which derives molecular formula from Fourier transform mass spectra using SMIRFE (US patent 10,607,723 B2).
- Resulting molecular formulas were classified into lipid categories and classes using a hierarchical set of Random Forest binary classifiers.
- High abundances of sterol esters were observed in NSCLC tissue, suggesting altered SCD1 or ACAT1 activity.
- Low abundances of cardiolipins were observed, suggesting altered human cardiolipin synthase 1 or lysocardiolipin acyltransferase activity which is known to confer apoptotic resistance.

Category	Total	More-Abundant Features			Less-Abundant Features		
		Expected	Observed	p-adjust	Expected	Observed	p-adjust
Fatty Acyls [FA]	12	2.989	2	1	3.947	0	1
Glycerophospholipids [GP]	205	51.055	37	1	67.424	88	0.00503
Prenol Lipids [PR]	5	1.245	0	1	1.644	0	1
Sphingolipids [SP]	281	69.983	79	0.09861	92.420	81	1
Sphingolipids [SP] – Low M/Z	33	8.219	3	1	10.854	16	0.141
Sphingolipids [SP] – High M/Z	248	61.764	76	0.00967	81.567	65	1
Sterol Lipids [ST]	23	5.728	13	0.00643	7.084	3	1

Log2 Fold Changes of Consistent Assigned Metabolites


A
Log2FC
mz
B
Log2FC
Fatty Acyls [FA]
Glycerophospholipids [GP]
Sphingolipids [SP]
Sterol Lipids [ST]
VotedCategories