Predicting Compound Amenability with Liquid Chromatography Mass Spectrometry to Improve Non-targeted Analysis

#246

Charles N. Lowe[1], Kristin K. Isaacs[1], Andrew McEachran[2], Christopher M. Grulke[1], Jon R. Sobus[1], Elin M. Ulrich[1], Ann Richard[1], Alex Chao[1], John Wambaugh[1], and Antony J. Williams[1]

[1]Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency (U.S. EPA), Research Triangle Park, North Carolina, USA; [2]Agilent Technologies, Inc., Santa Clara, CA, USA

ORCID: 0000-0001-9151-6157

OBJECTIVES

- Non-targeted analysis is a powerful tool for identifying chemicals within environmental samples, although only a fraction of these chemicals are detectable with a particular analytical method.
- EPA’s Non-Targeted Analysis Collaborative Trial (ENTACT) showed up to 40% of compounds (out of 1,269 tested) may be unamenable to LC-MS.
- We develop models capable of determining the amenability of chemical compounds in an LC-MS using electrospray ionization.

APPROACH

- LC-MS ESI+/- detected compounds from MassBank of North America (MoNA) as well as compounds detected/not-detected from ToxCast library.
- Sampling methods were applied to account for large disparity in amenable vs. unamenable compounds.
- Random forest models constructed to predict compound amenability based on PaDEL molecular descriptors.
- Models were validated using an external dataset and a simulated suspect screening.

MAIN RESULTS

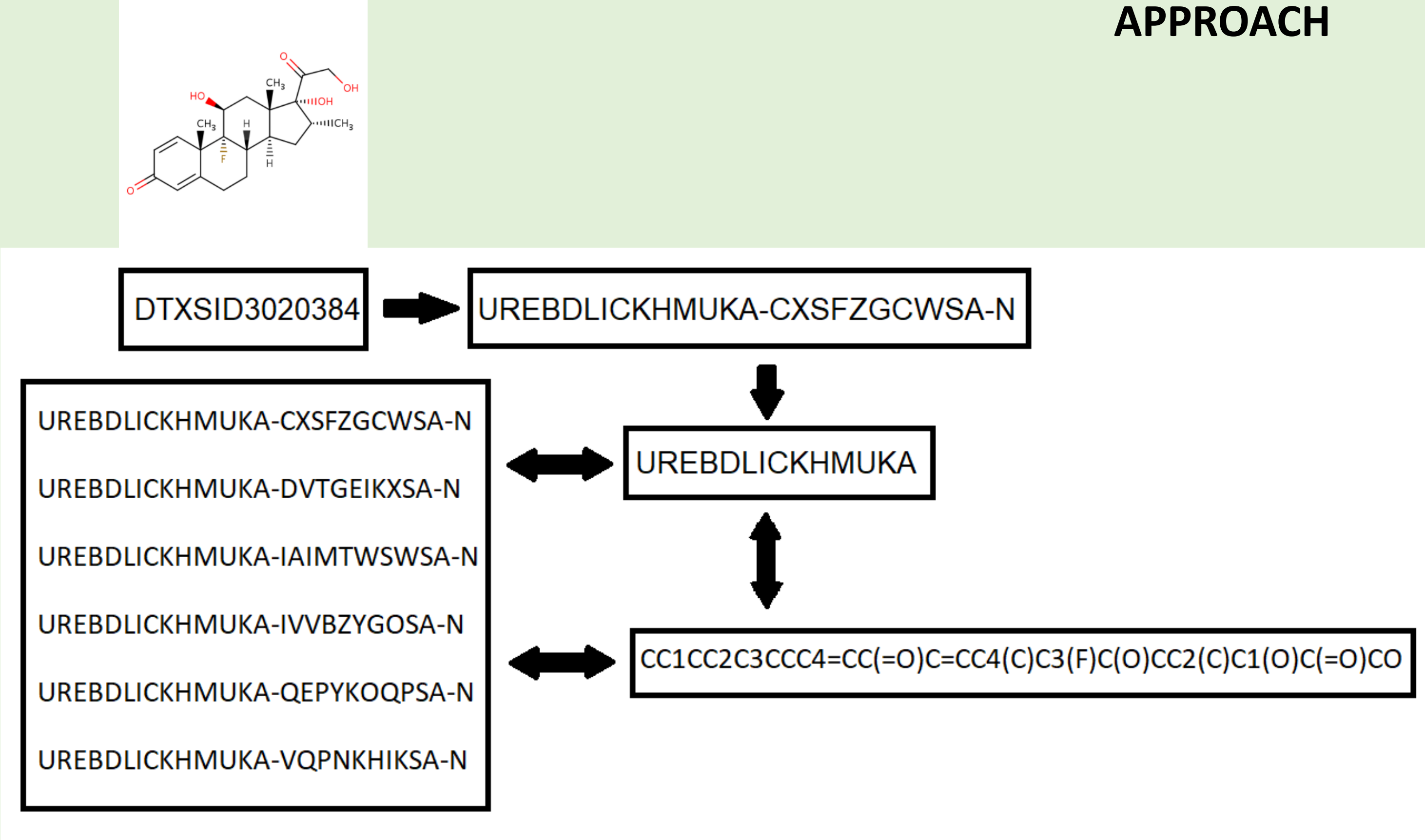
- Upsampling leads to over-fitted models, however downsampling leads to well-fitted models, albeit with a smaller descriptor space.
- Models showed better performance than those based on random chance.

IMPACT

- Amenability models capable of predicting novel compound amenability with *good* accuracy.
- Can lead to significant time and resource cost by eliminating unnecessary testing.
- Amenability predictions for DSSTox database and web

Predicting Compound Amenability with Liquid Chromatography Mass Spectrometry to Improve Non-targeted Analysis

APPROACH



DTXSID3020384

UREBDLICKHMUKA-CXSFZGCWSA-N

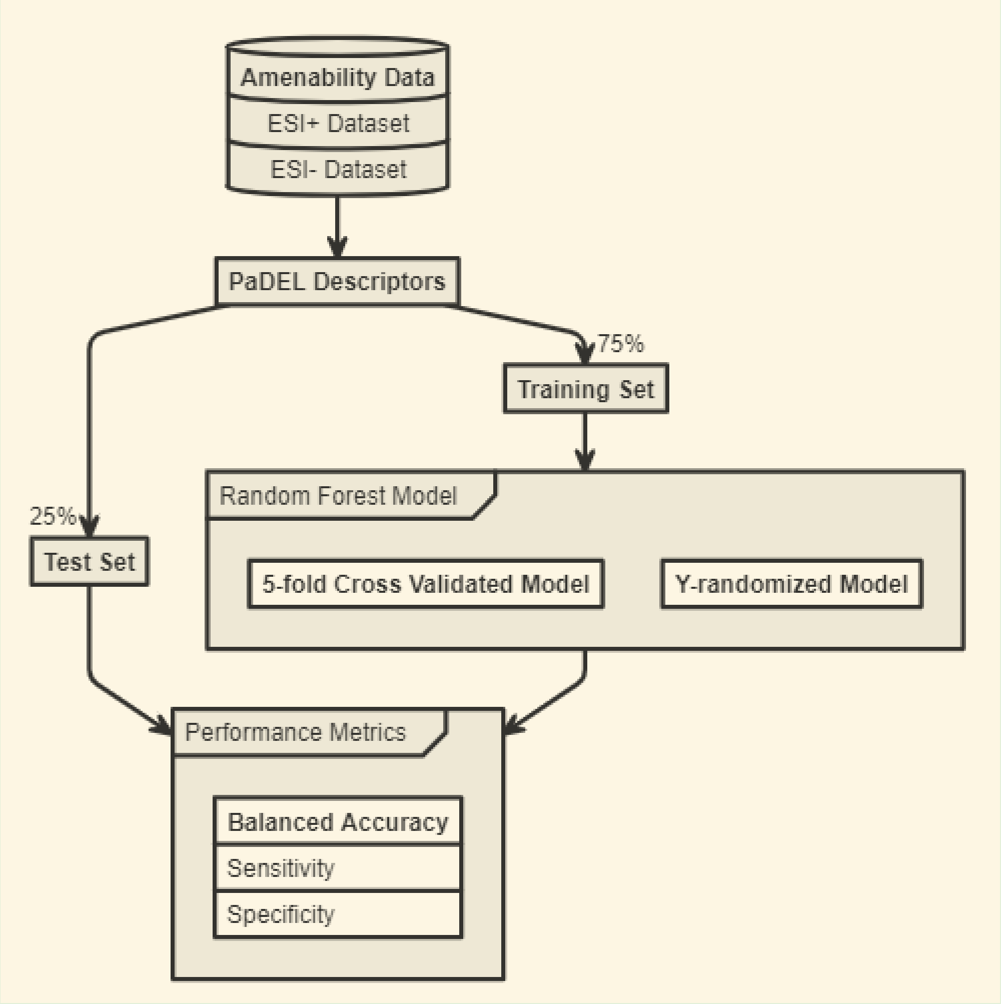
UREBDLICKHMUKA

CC1CC2C3CCC4=CC(=O)C=CC4(C)C3(F)C(O)CC2(C)C1(O)C(=O)CO

UREBDLICKHMUKA-CXSFZGCWSA-N
UREBDLICKHMUKA-DVTGEIKXSA-N
UREBDLICKHMUKA-IAIMTWSWSA-N
UREBDLICKHMUKA-IVVBZYGOSA-N
UREBDLICKHMUKA-QEPYKOQPSA-N
UREBDLICKHMUKA-VQPNKHIKSA-N

The first block of each InChIKey was chosen to represent each detected/not-detected compound*

**Mass spectrometers cannot determine stereo differences, which are captured in the second block of the InChIKey, but can distinguish*



Amenability Data
ESI+ Dataset
ESI- Dataset

PaDEL Descriptors

75%

Training Set

25%

Test Set

Random Forest Model

5-fold Cross Validated Model

Y-randomized Model

Performance Metrics

Balanced Accuracy
Sensitivity
Specificity

Simplified modeling workflow

Predicting Compound Amenability with Liquid Chromatography Mass Spectrometry to Improve Non-targeted Analysis

MAIN RESULTS

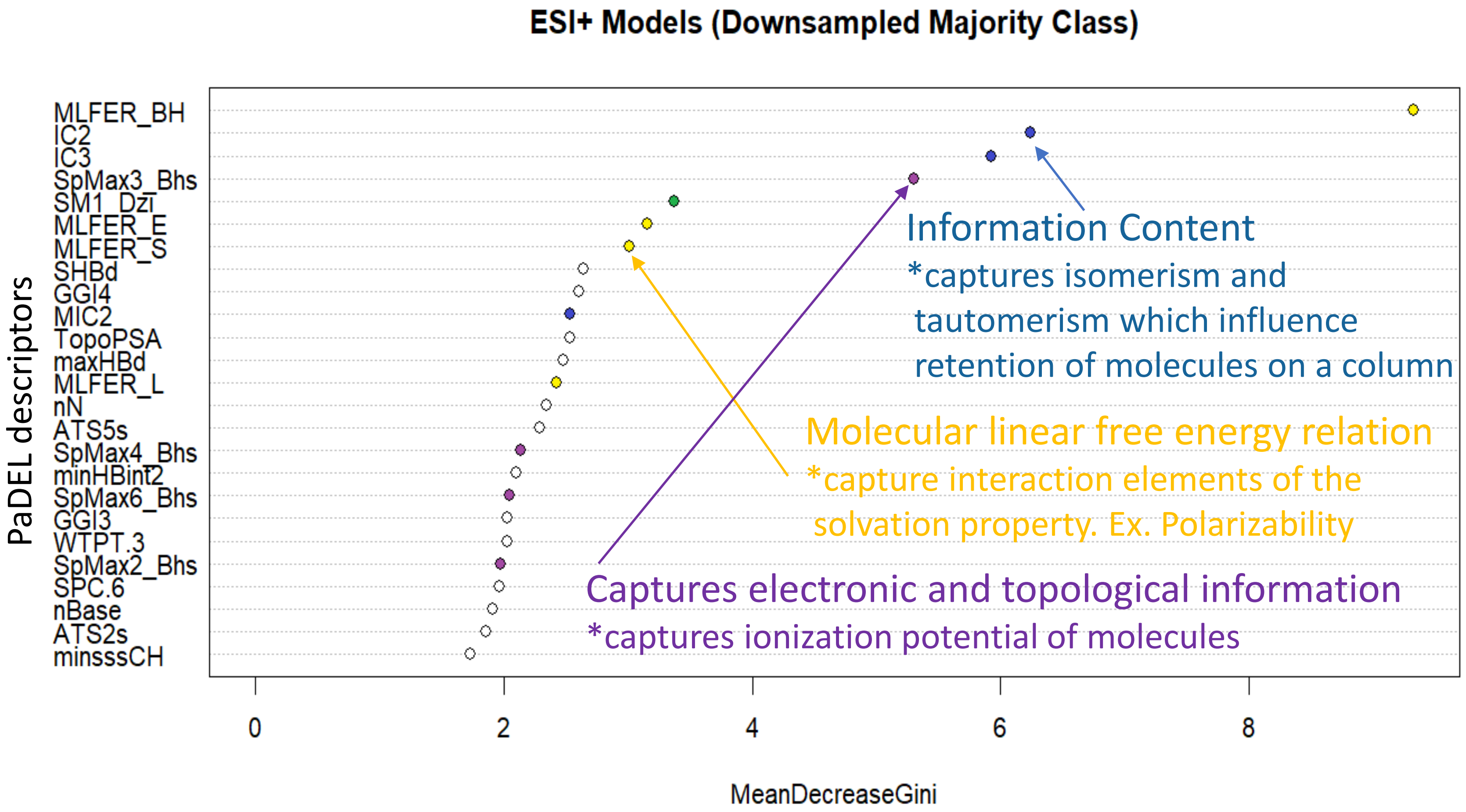
Model performances for fitting, training, and cross-validated sets for the test and Y-randomized test sets.

	Training Set				Fivefold CV		
Model	Size	Balanced Accuracy	Sensitivity	Specificity	Balanced Accuracy	Sensitivity	Specificity
ESI+ Models (Downsampling Applied)	580	0.78	0.79	0.77	0.77	0.76	0.78
ESI+ Models (Upsampling Applied)	6340	0.99	1.00	0.99	0.99	0.98	1.00
ESI- Models (Downsampling Applied)	550	0.83	0.82	0.84	0.81	0.83	0.79
ESI- Models (Upsampling Applied)	4688	0.99	1.00	0.98	0.98	0.97	1.00
	Test Set				Y-randomization		
Model	Size	Balanced Accuracy	Sensitivity	Specificity	Balanced Accuracy	Sensitivity	Specificity
ESI+ Models (Downsampling Applied)	1153	0.81	0.85	0.76	0.48	0.44	0.51
ESI+ Models (Upsampling Applied)	1153	0.58	0.98	0.19	0.55	0.48	0.63
ESI- Models (Downsampling Applied)	871	0.82	0.85	0.80	0.50	0.49	0.51
ESI- Models (Upsampling Applied)	871	0.68	0.99	0.38	0.51	0.46	0.56

- The upsampled models initially appear superior, however the results for the test set show poor accuracy – a result of overfitting.
- The downsampled models show similar accuracy for both training and test data in the ~0.8 range, much better than those built based on random chance in the Y-randomization models.

Predicting Compound Amenability with Liquid Chromatography Mass Spectrometry to Improve Non-targeted Analysis

MAIN RESULTS



Information Content

*captures isomerism and tautomerism which influence retention of molecules on a column

Molecular linear free energy relation

*capture interaction elements of the solvation property. Ex. Polarizability

Captures electronic and topological information

*captures ionization potential of molecules

A plot of variable importance based on the mean decrease in the Gini

Predicting Compound Amenability with Liquid Chromatography Mass Spectrometry to Improve Non-targeted Analysis

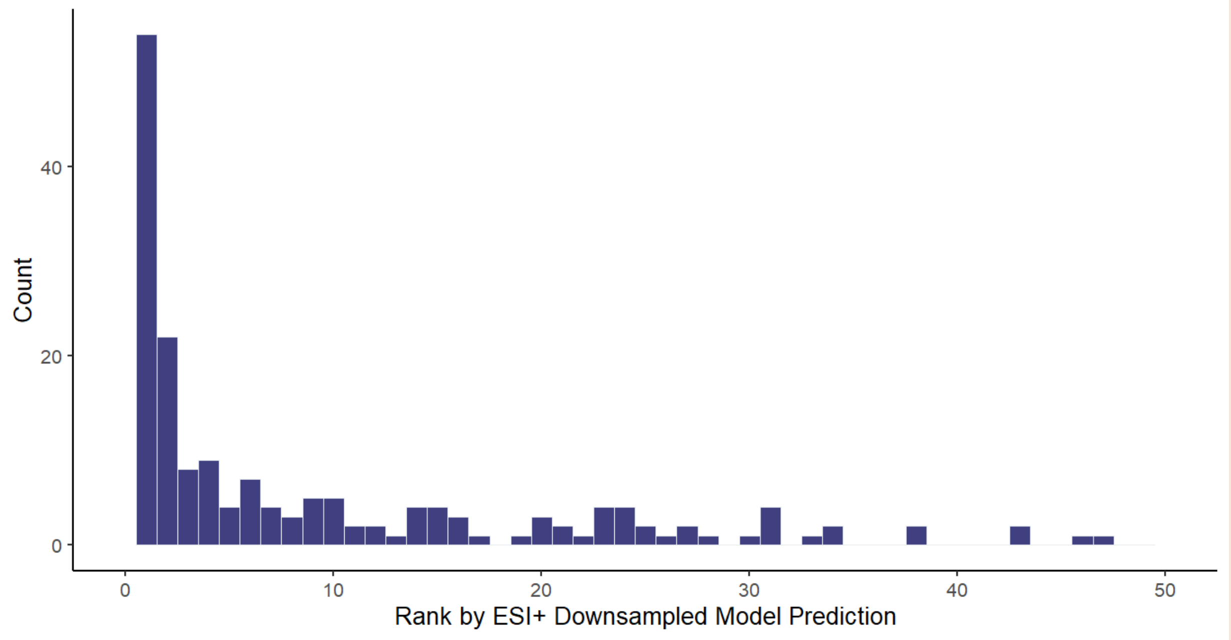
MAIN RESULTS

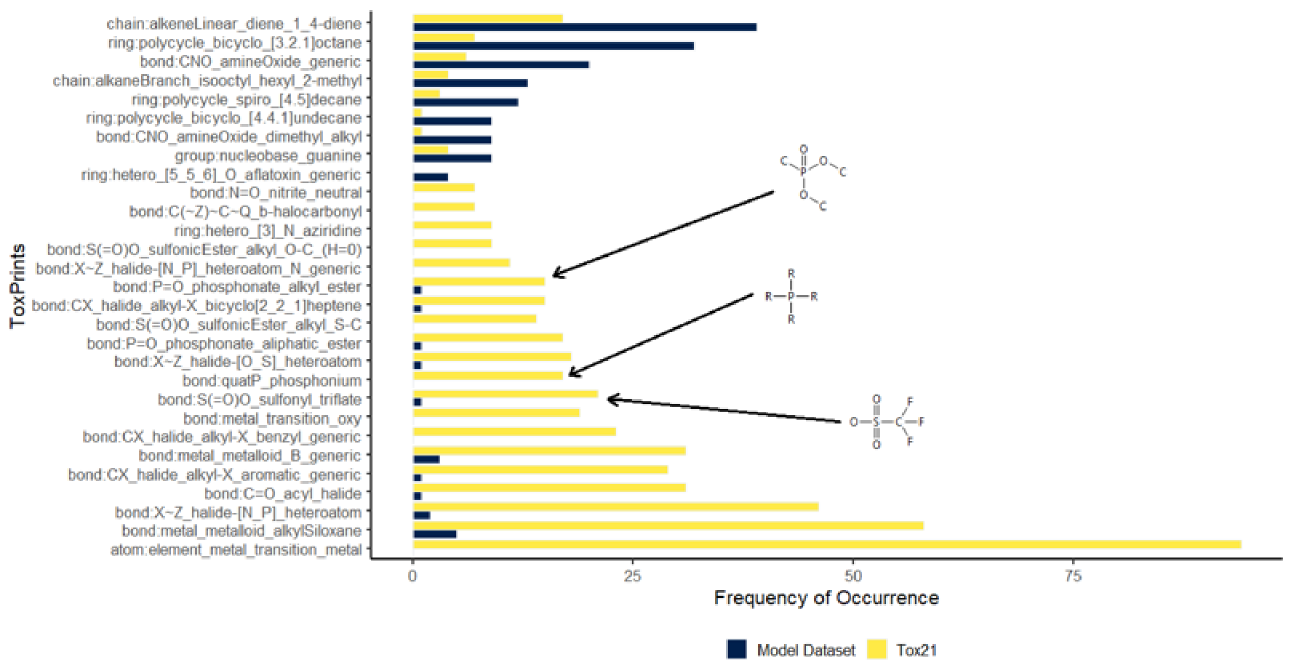
Model performances for ESI+ and ESI- downsampled model predictions compared to external validation data. While accuracy is lower (still acceptable) than observed for the test set taken from the modeling dataset, this dataset strictly excluded any chemicals from the modeling dataset and is thus biased toward unseen parts of chemical space.

ESI- Downsampled Model		
	Amenable (Prediction)	Unamenable (Prediction)
Detected (Experiment)	323	502
Not-detected (Experiment)	68	874
Sensitivity	**0.83**	
Specificity	**0.64**	
Balanced Accuracy	**0.73**	
ESI+ Downsampled Model		
	Amenable (Prediction)	Unamenable (Prediction)
Detected (Experiment)	423	402
Not-detected (Experiment)	103	839
Sensitivity	**0.8**	
Specificity	**0.68**	
Balanced Accuracy	**0.74**	
Combined Models		
	Amenable (Prediction)	Unamenable (Prediction)
Detected (Experiment)	505	320
Not-detected (Experiment)	129	813
Sensitivity	**0.8**	
Specificity	**0.72**	

Predicting Compound Amenability with Liquid Chromatography Mass Spectrometry to Improve Non-targeted Analysis

MAIN RESULTS





Frequency counts of candidate compounds found to be a match for a suspect screening compound ordered by prediction rank value (with 1 being the highest confidence rank, and 50 the lowest) based on ESI+ LC-MS amenability predictions.

This result shows that ranking candidates based on the confidence measure from the random forest model can provide reasonable candidates for conformation by chemical standard.

A plot of prevalent chemotypes in the Tox21 dataset and the model dataset used in this work, selected based on the absolute difference of frequency of occurrence between datasets.

This plot provides insight into the portion of chemical space not currently represented in the models (some of which will never be represented). Future work will attempt to address these deficiencies to improve usefulness for analysis of environmental