United States
Environmental Protection
Agency

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

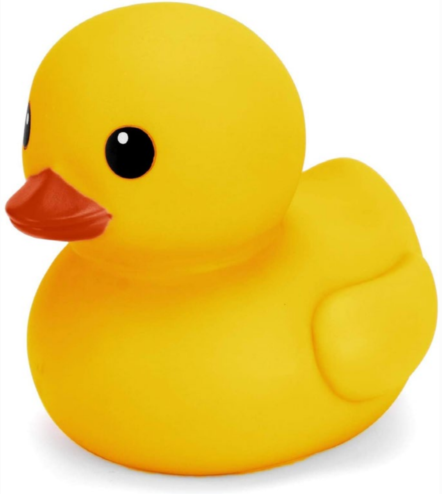
Charles N. Lowe[1], Kristin K. Isaacs[1], Andrew McEachran[2], Christopher M. Grulke[1], Jon R. Sobus[1], Elin M. Ulrich[1], Ann Richard[1], Alex Chao[1], John Wambaugh[1], and Antony J. Williams[1]

1. Center for Computational Toxicology and Exposure, U.S. EPA, Research Triangle Park, NC
2. Agilent Technologies, Inc., Santa Clara, CA

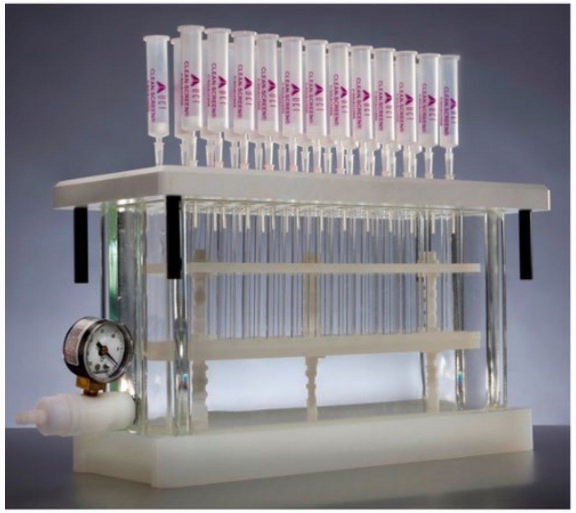
Disclaimer: The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

EPA United States Environmental Protection Agency

Complex samples, NTA, and the modeling problem



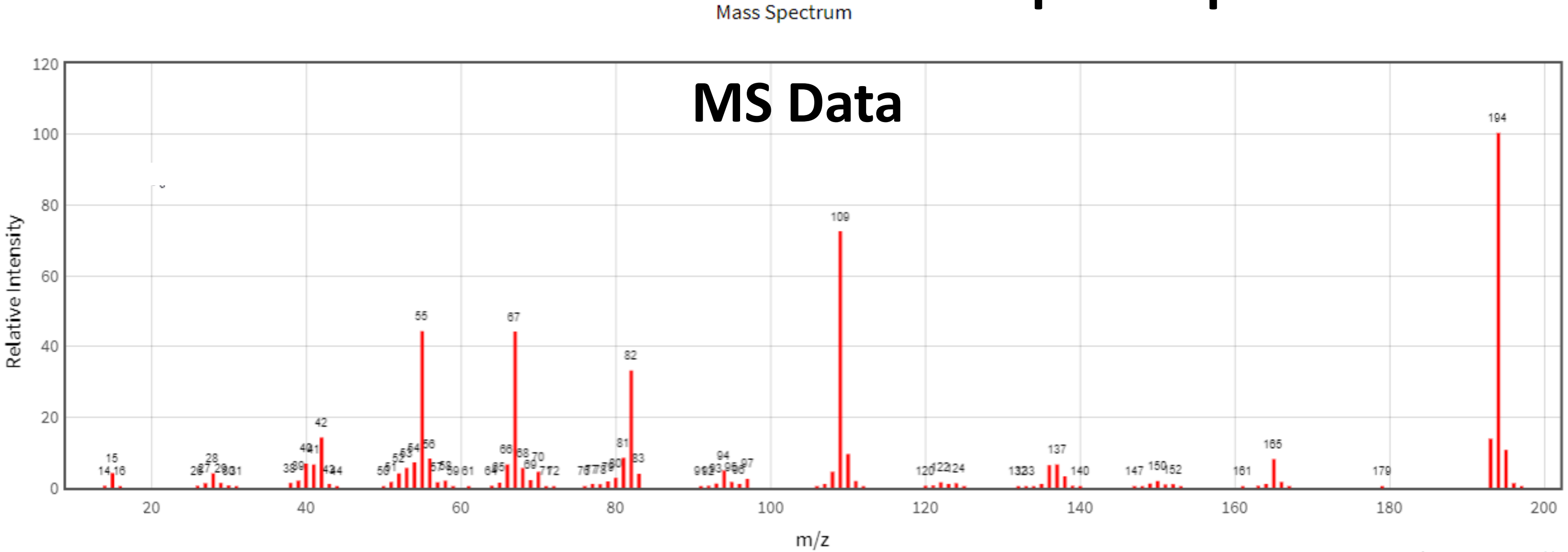
Media Sample



Extraction, Cleanup & Sample Preparation

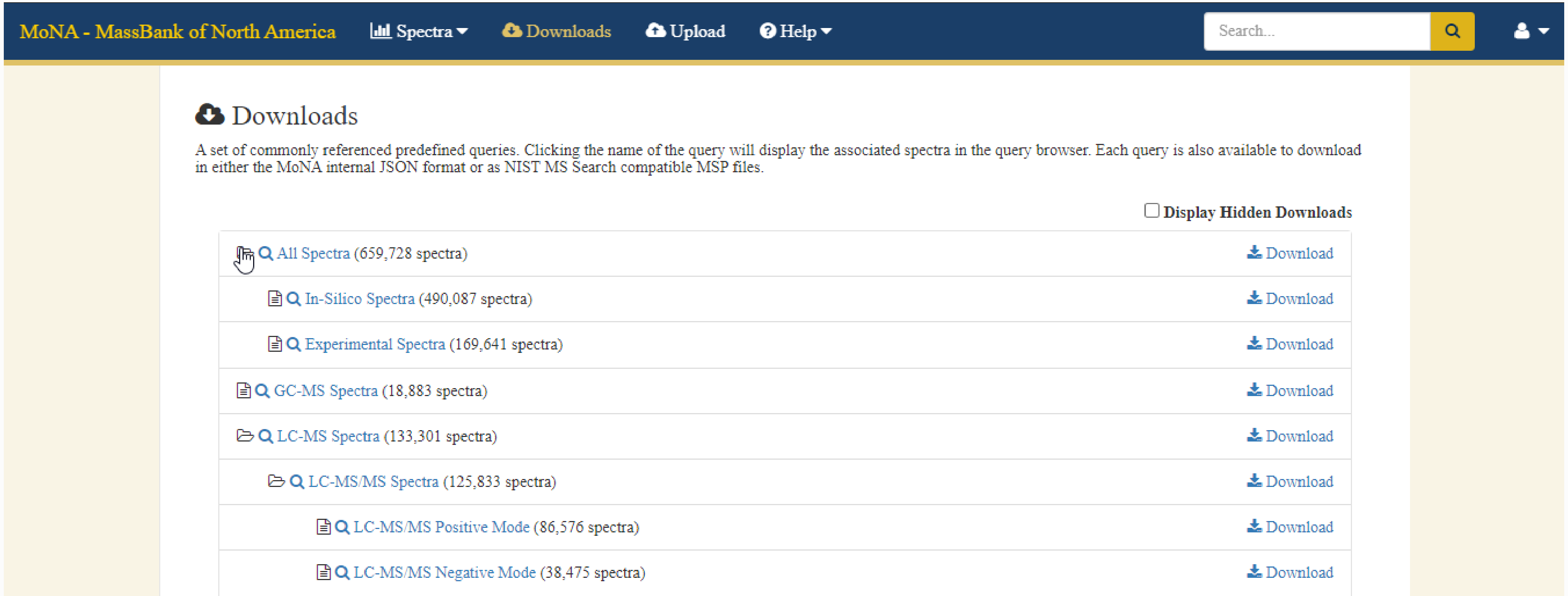
MS Analysis

???



See https://doi.org/10.1021/acs.est.1c01907 for more on rubber ducky analysis.

Curating a dataset for modeling



Downloads

A set of commonly referenced predefined queries. Clicking the name of the query will display the associated spectra in the query browser. Each query is also available to download in either the MoNA internal JSON format or as NIST MS Search compatible MSP files.

☐ Display Hidden Downloads

All Spectra (659,728 spectra)	Download
In-Silico Spectra (490,087 spectra)	Download
Experimental Spectra (169,641 spectra)	Download
GC-MS Spectra (18,883 spectra)	Download
LC-MS Spectra (133,301 spectra)	Download
LC-MS/MS Spectra (125,833 spectra)	Download
LC-MS/MS Positive Mode (86,576 spectra)	Download
LC-MS/MS Negative Mode (38,475 spectra)	Download

- **4,103 unique compounds detected in ESI+ LC-MS**
- **3,007 unique compounds detected in ESI- LC-MS**
- **1,542 unique compounds detected in both modes**

United States
Environmental Protection
Agency

Curating a dataset for modeling

- Only amenable compounds identified in MoNA
  - No unamenable compound data
- ToxCast library LC-MS/MS curation
  - Spectra checked individually for quality
    - Provides unamenable compound data

- **ESI+ LC-MS/MS**
  - **393 amenable; 456 unamenable**
- **ESI- LC-MS/MS**
  - **456 amenable; 402 unamenable**

Agilent

Curating a dataset for modeling

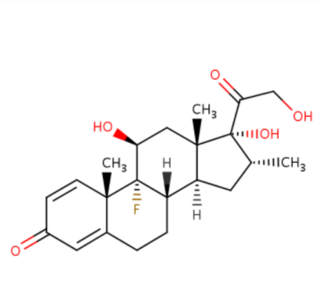
- Only amenable compounds identified in MoNA
  - No unamenable compound data
- ToxCast library LC-MS/MS curation
  - Spectra checked individually for quality
    - Provides unamenable compound data

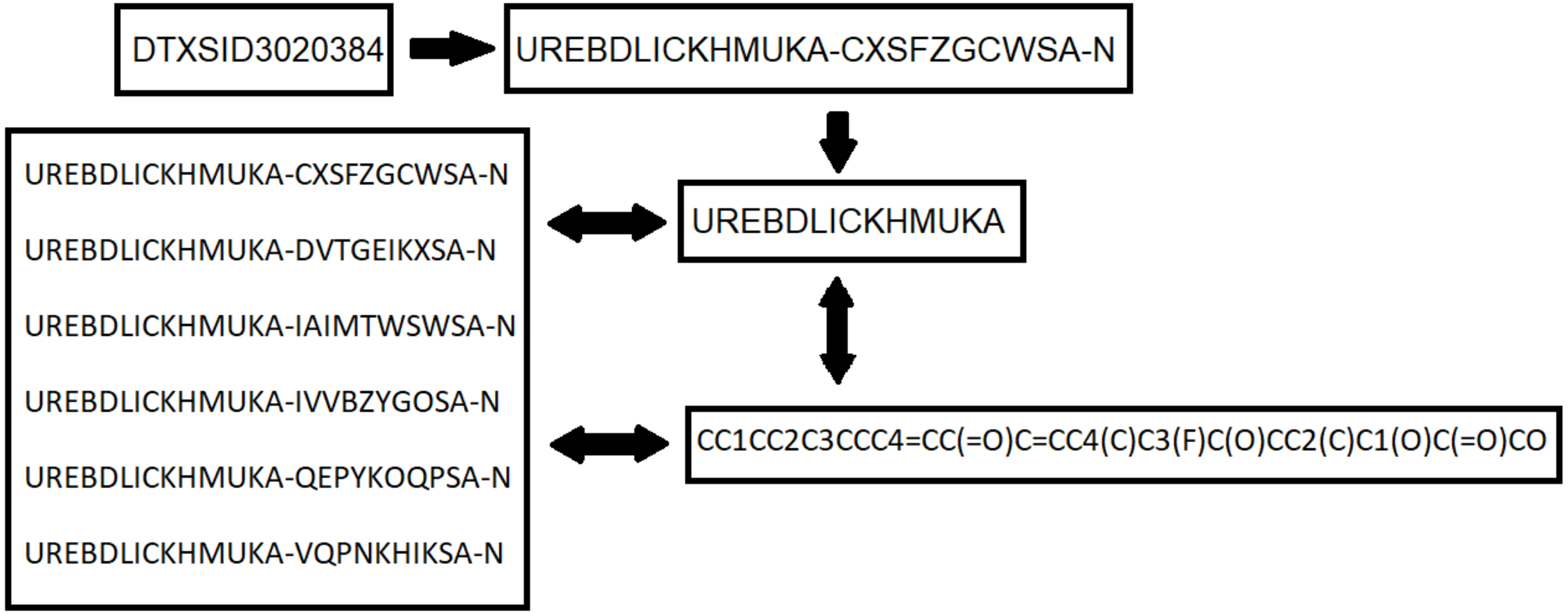
- **ESI+ LC-MS/MS**
  - **393 amenable; 456 unamenable**
- **ESI- LC-MS/MS**
  - **456 amenable; 402 unamenable**

Overall dataset

- **ESI+ LC-MS/MS**
  - **4,226 amenable; 387 unamenable**
- **ESI- LC-MS/MS**
  - **3,130 amenable; 360 unamenable**

Curating a dataset for modeling





DTXSID3020384

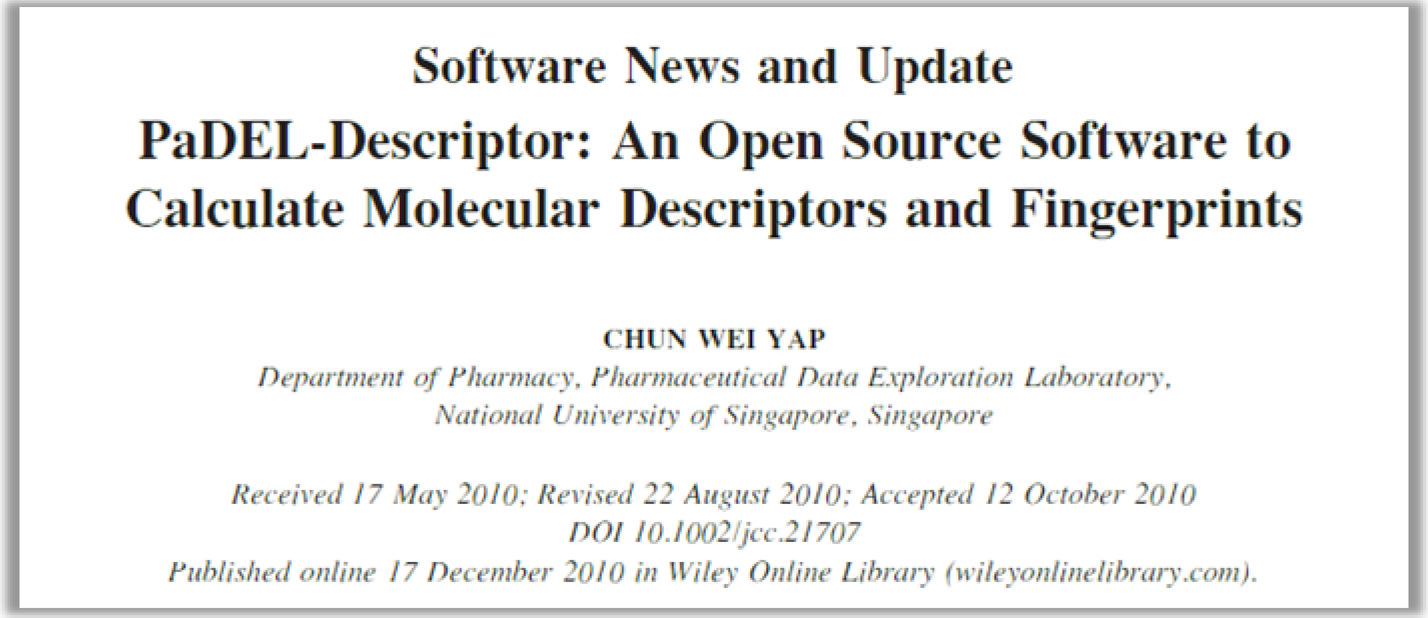
UREBDLICKHMUKA-CXSFZGCWSA-N

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

Describing molecular structures



Software News and Update

PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints

CHUN WEI YAP
*Department of Pharmacy, Pharmaceutical Data Exploration Laboratory,
National University of Singapore, Singapore*

*Received 17 May 2010; Revised 22 August 2010; Accepted 12 October 2010
DOI 10.1002/jcc.21707
Published online 17 December 2010 in Wiley Online Library (wileyonlinelibrary.com).*

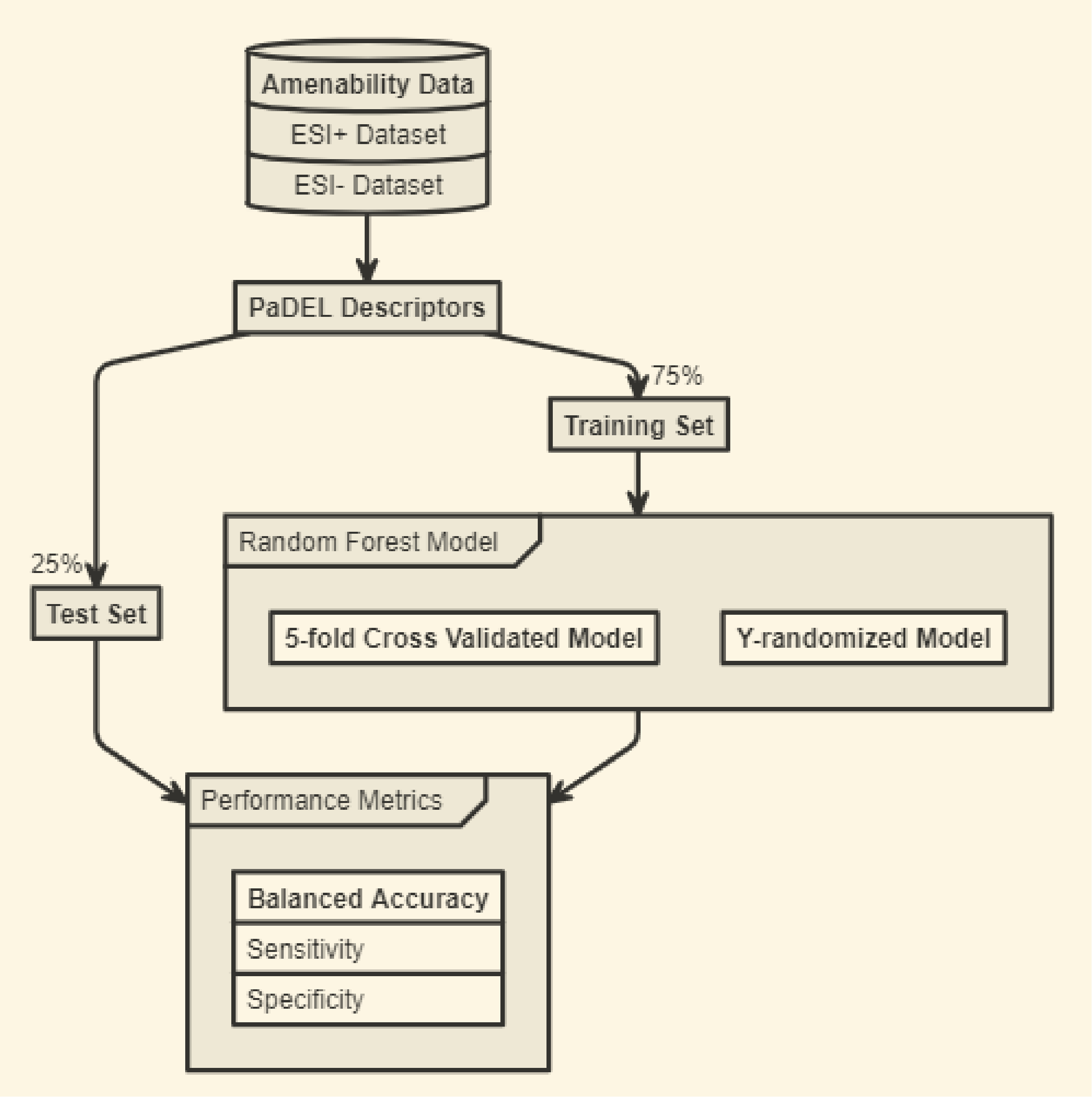
- 1,444 1D & 2D Molecular descriptors from QSAR-ready SMILES. Examples include…
  - Electrotopological states weighted by atomic properties
  - Molecular linear free energy relationships weighted by atomic properties
  - Atom, bond, & ring counts
  - LogKow (logP) predictions, etc..

Cleaning and reduction of descriptor space

- Dimension reduction will do two things:
  - improve interpretability of models
  - make model calculations faster
- Remove chemicals missing descriptors*
- Remove any constant descriptors (variance(x) = 0)
- Remove near-constant descriptors (sd(x) < 0.25)
  - 0.25 gives a good balance between reduction and retention
- Calculate pairwise correlations between remaining descriptors
  - Eliminate based on a cutoff = 0.96 correlation
    - descriptor showing largest pair correlation with other descriptors was excluded

1,444 descriptors → 451 descriptors

Machine learning approach



Performance Metrics

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$\text{balanced accuracy} = \frac{\text{sensitivity} + \text{specificity}}{2}$$

Machine learning approach

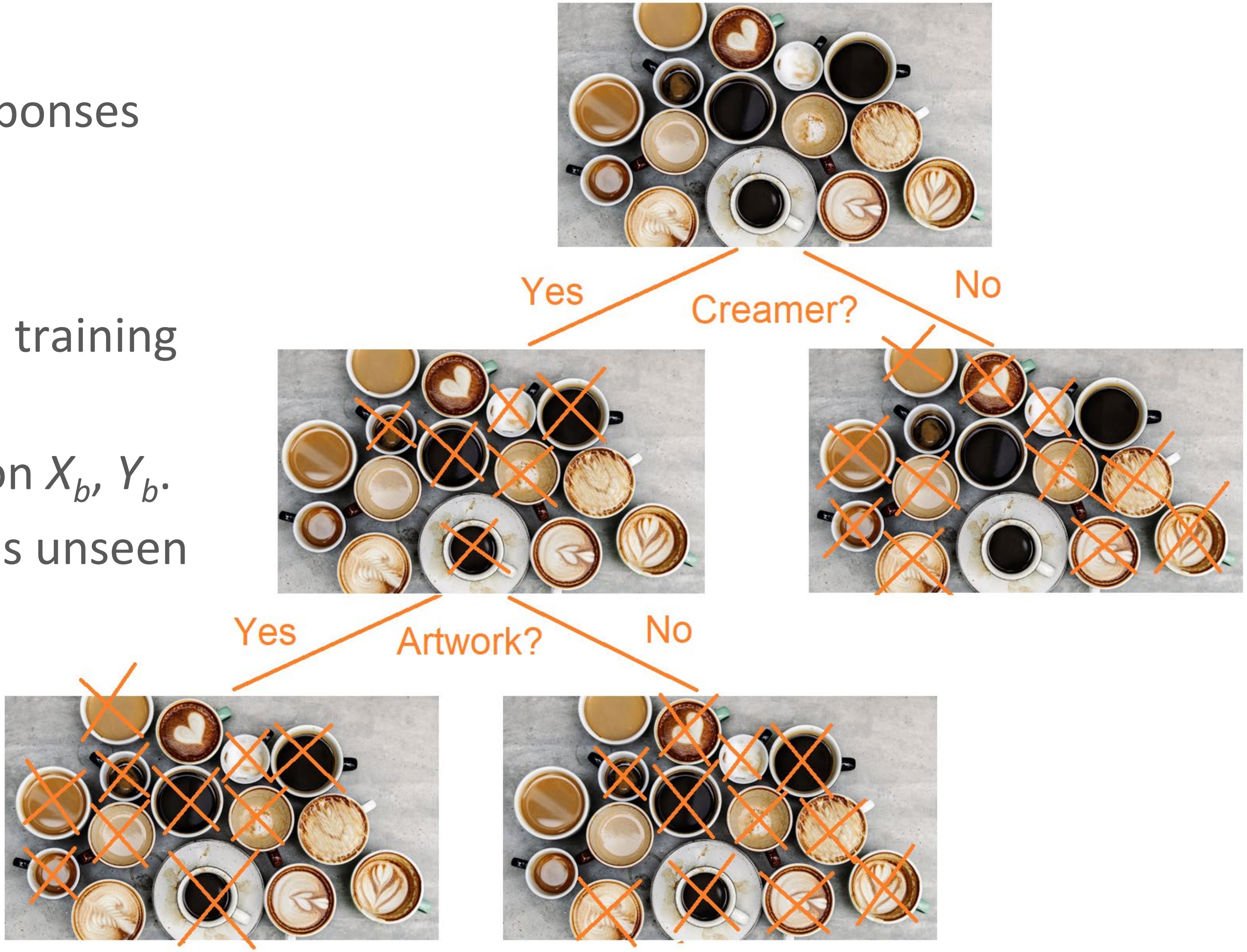
Random Forest Algorithm

Training set *X* = $x_1x_2…x_n$ with responses

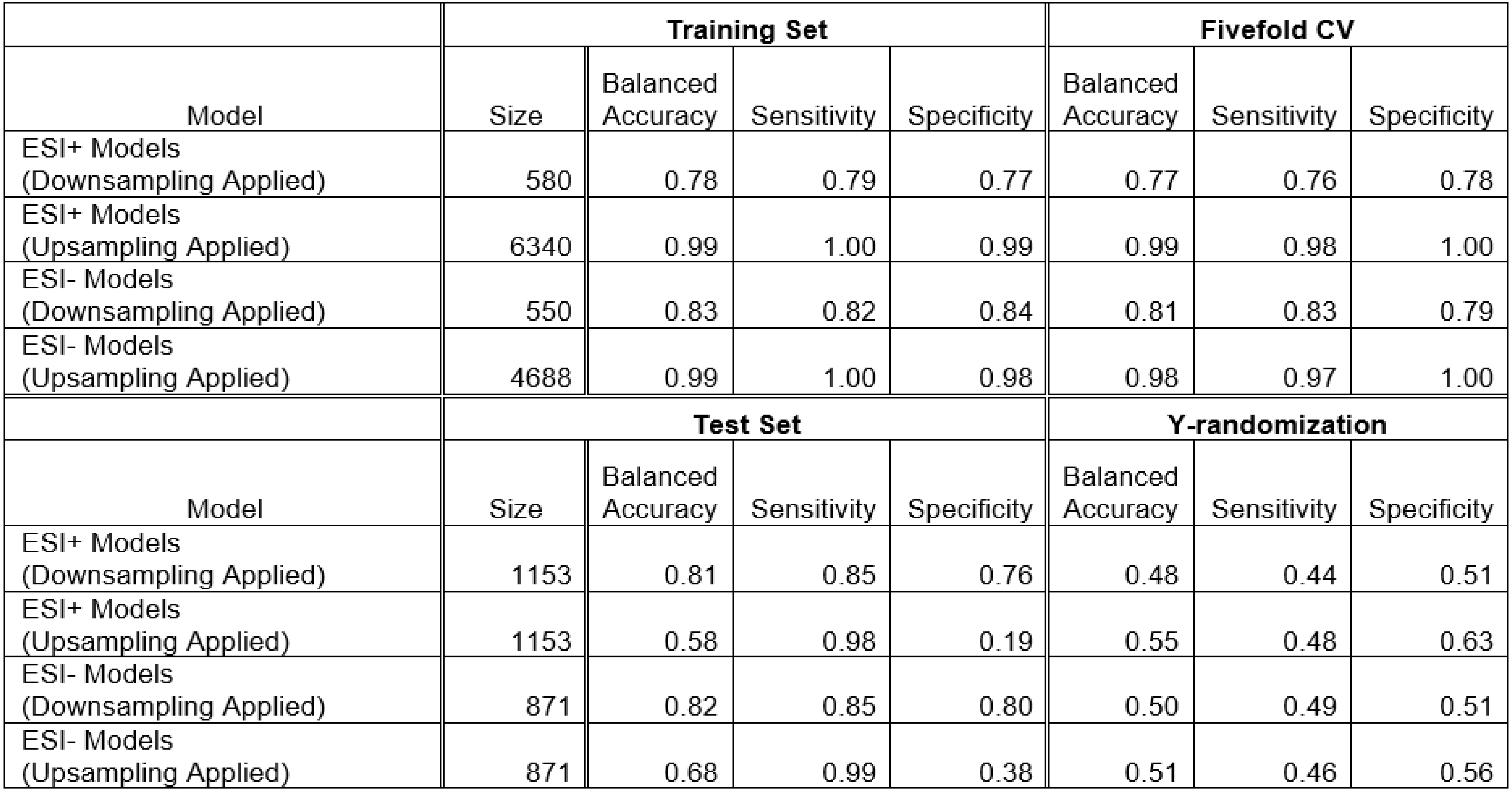
$Y = y_1y_2…y_n$

For number of trees, *b = 1,…,B*

1. Sample, with replacement, *n* training examples from *X*, *Y*; X_b, Y_b.
2. Train a classification tree f_b on X_b, Y_b.
3. The majority of all f_b classifies unseen endpoints.



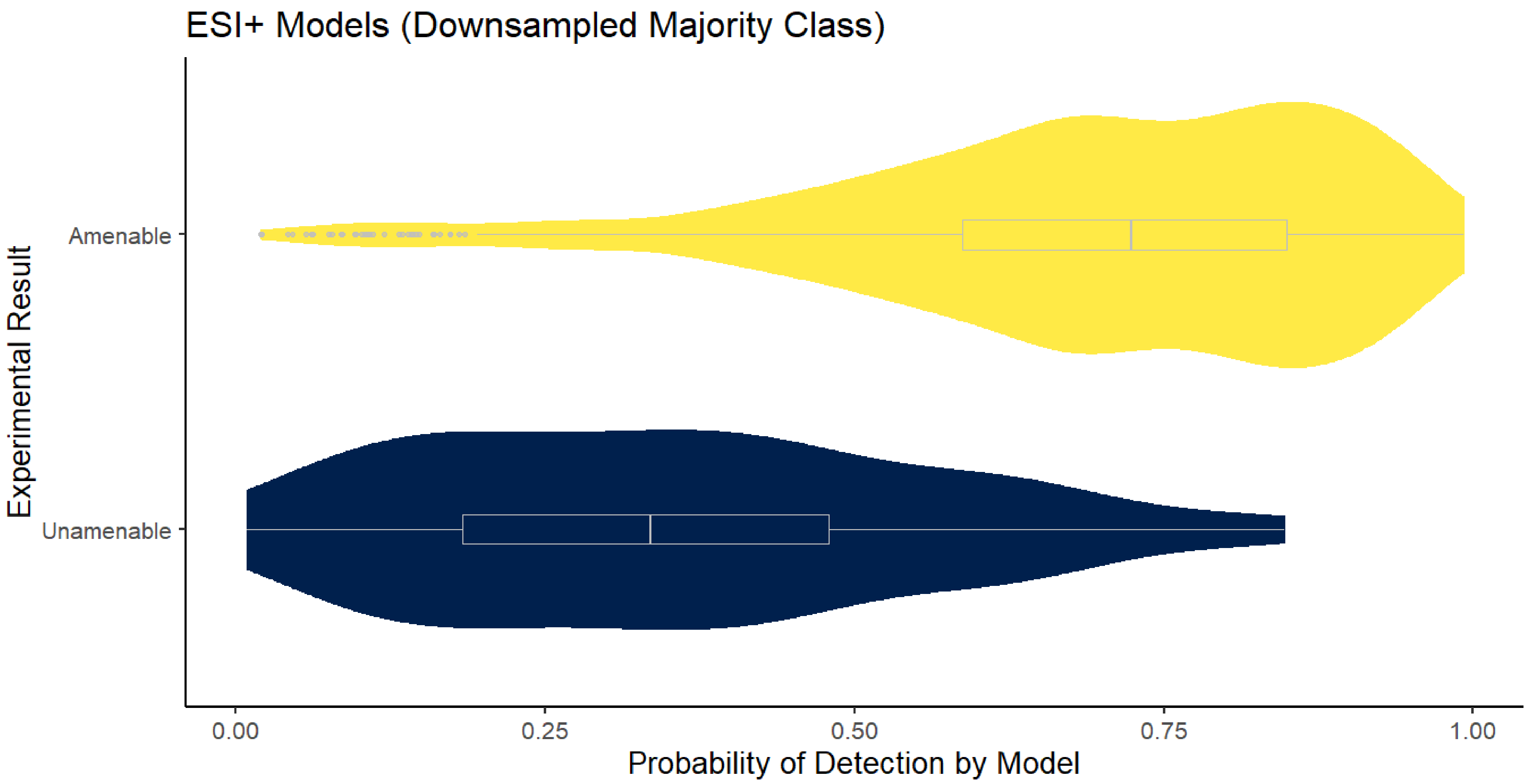
Model performance



	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

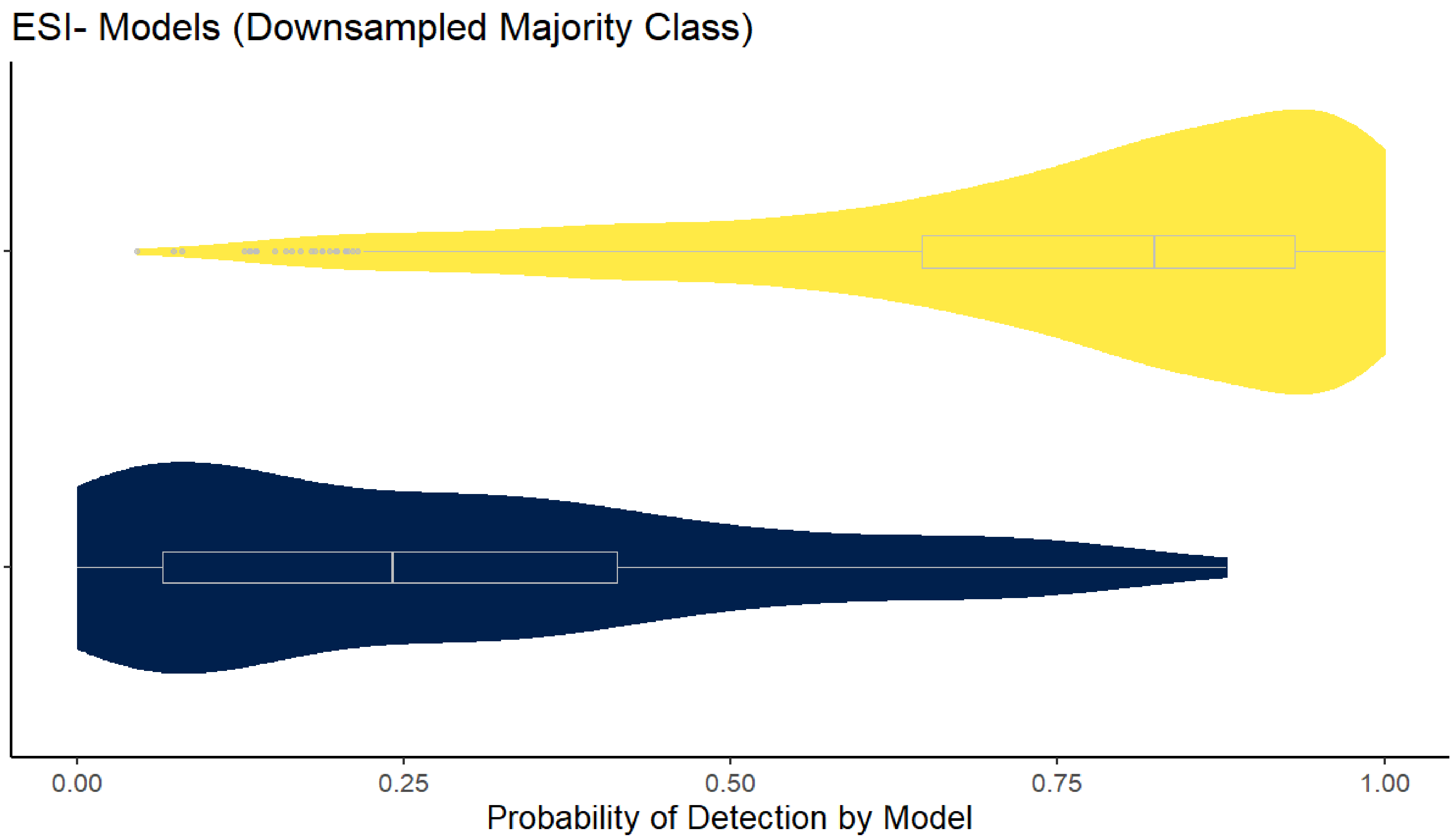
	Test Set				Y-randomization		
Model	Size	Balanced Accuracy	Sensitivity	Specificity	Balanced Accuracy	Sensitivity	Specificity
ESI+ Models (Downsampling Applied)	1153	0.81	0.85	0.76	0.48	0.44	0.51
ESI+ Models (Upsampling Applied)	1153	0.58	0.98	0.19	0.55	0.48	0.63
ESI- Models (Downsampling Applied)	871	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

Model performance

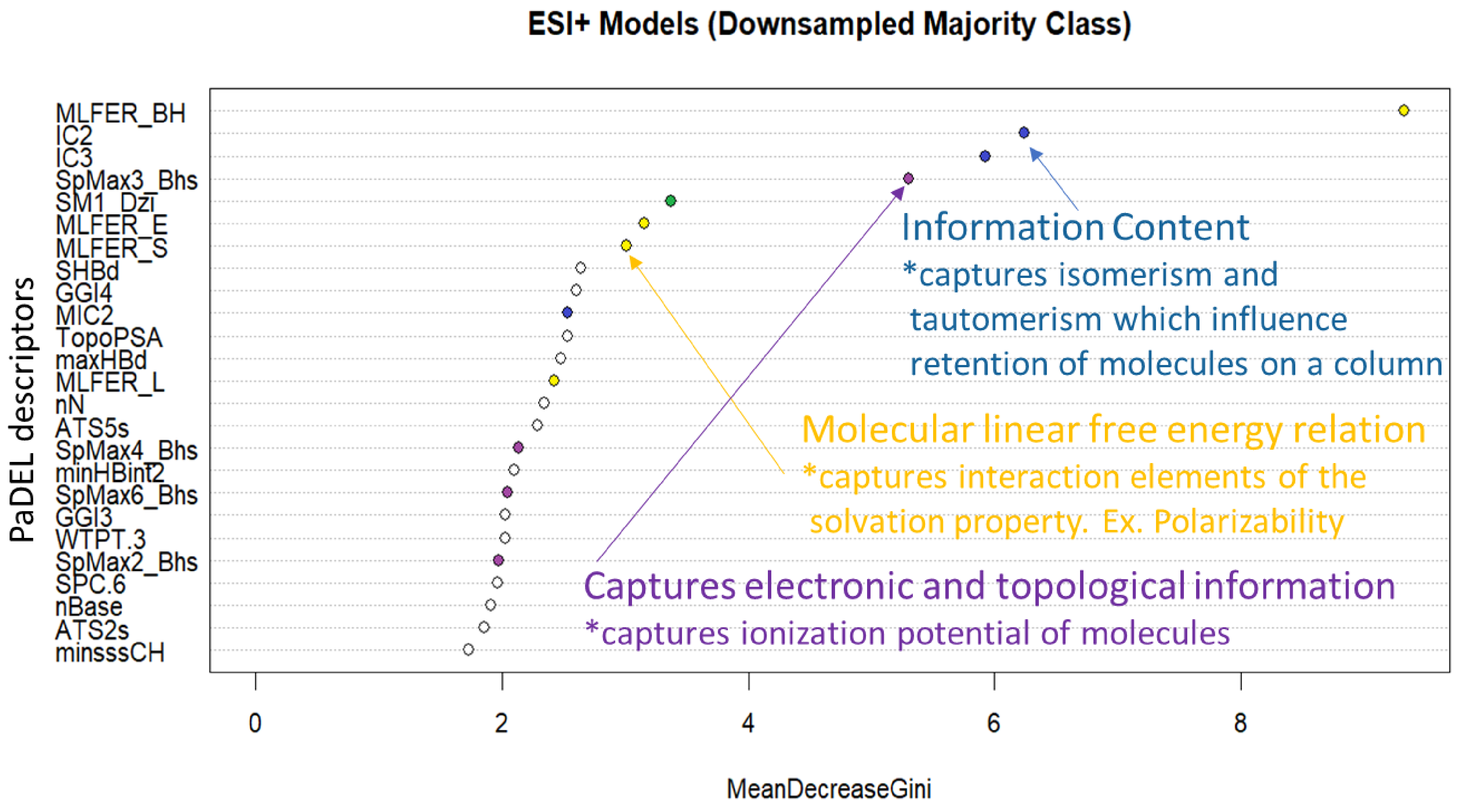


United States Environmental Protection Agency

Model performance



Mechanistic Interpretation



Information Content

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

Molecular linear free energy relation

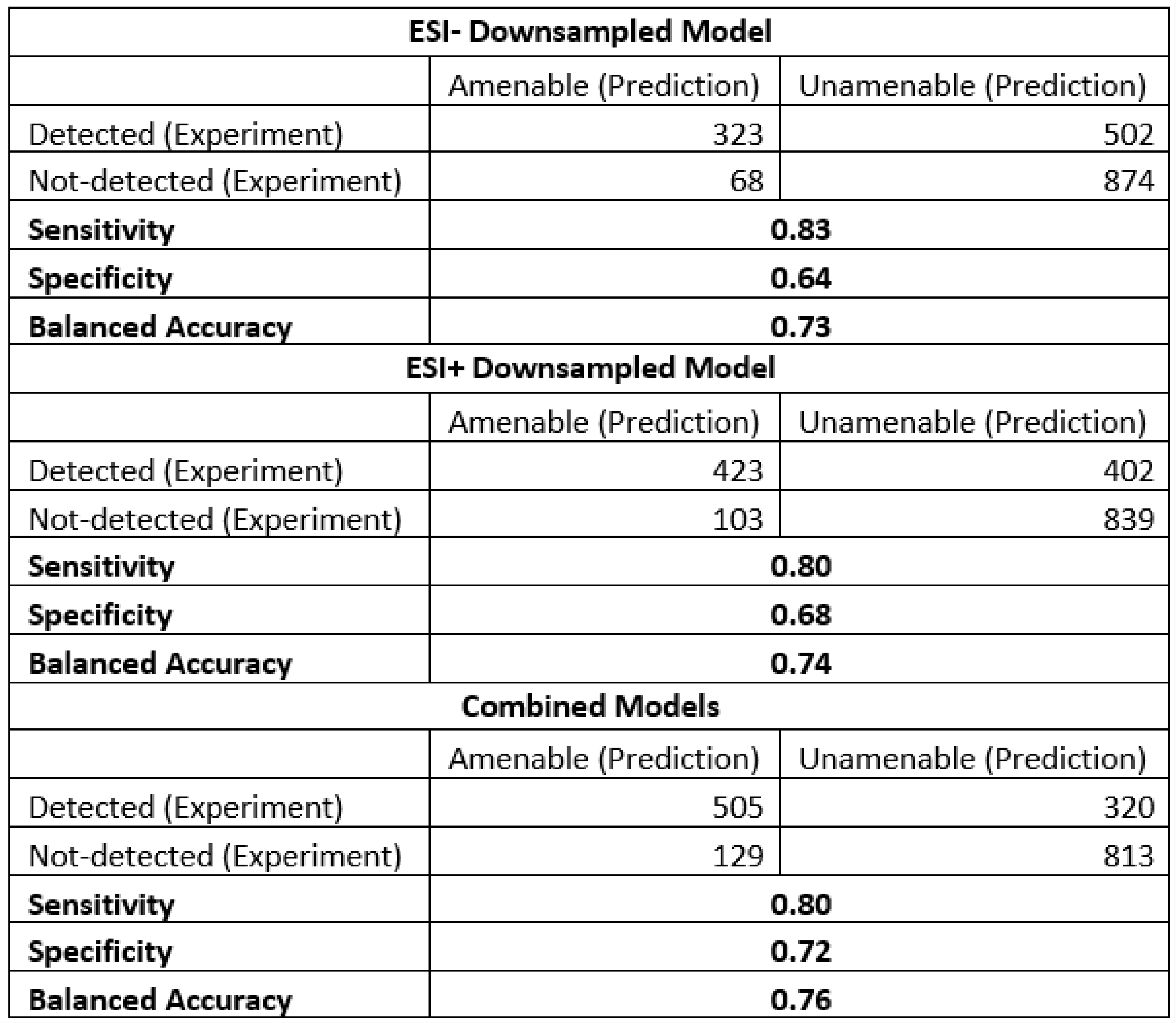
*captures interaction elements of the solvation property. Ex. Polarizability

Captures electronic and topological information

*captures ionization potential of molecules

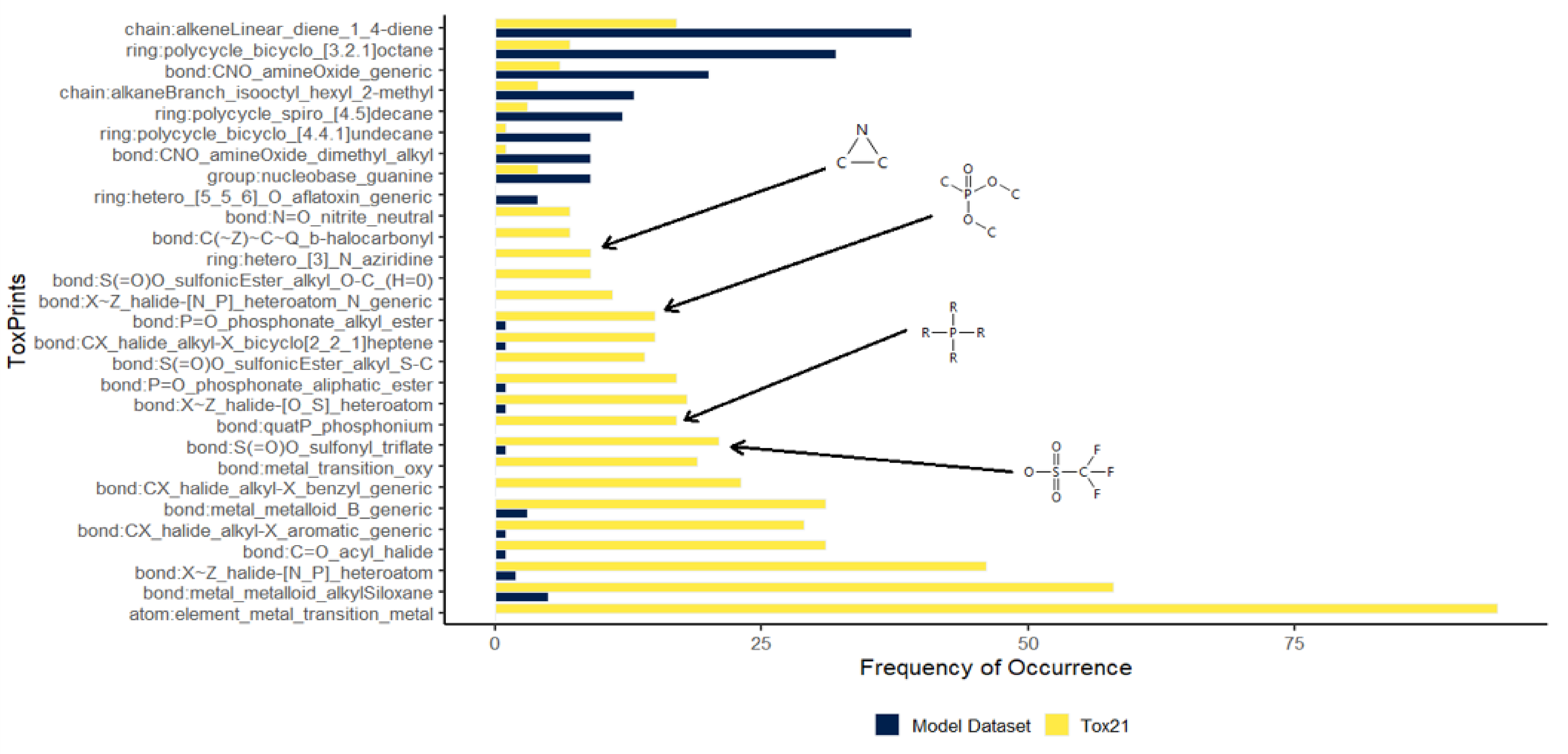
External Validation

- 1,768 chemical compounds analyzed for LC-MS amenability as part of ToxCast program
  - All compounds exclusive of modeling dataset



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	**0.80**	
Specificity	**0.68**	
Balanced Accuracy	**0.74**	
Combined Models		
	Amenable (Prediction)	Unamenable (Prediction)
Detected (Experiment)	505	320
Not-detected (Experiment)	129	813
Sensitivity	**0.80**	
Specificity	**0.72**	
Balanced Accuracy	**0.76**	

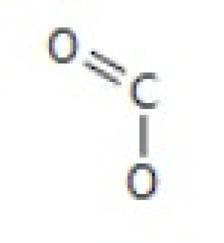
Model Applicability to ToxCast



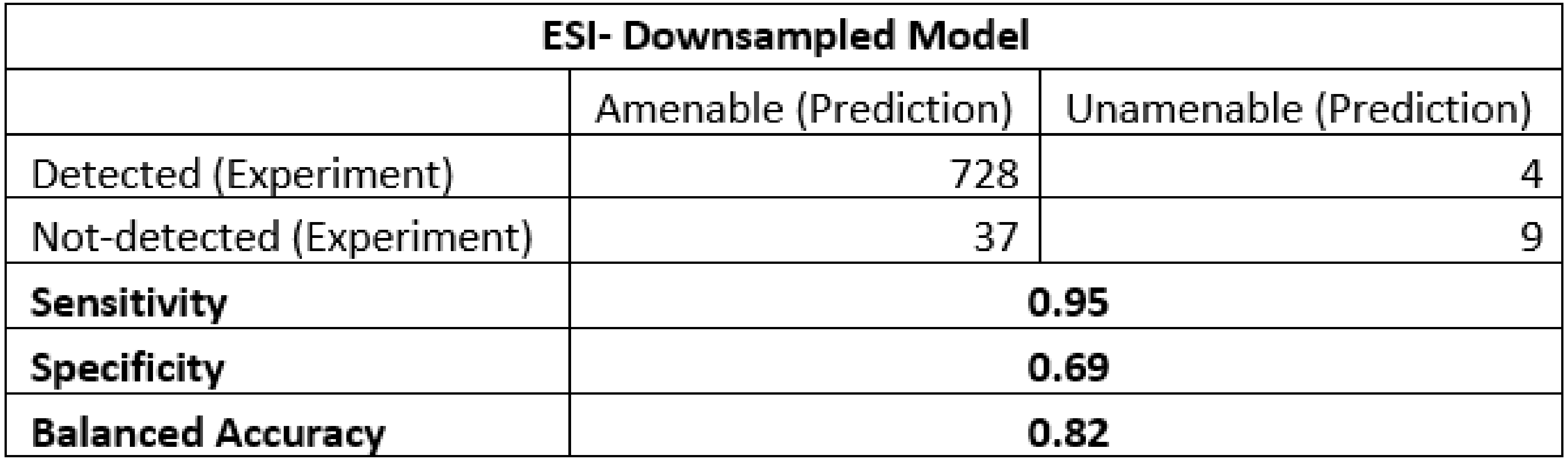
Comparison of prevalent ToxPrint chemotypes in amenability dataset against the ToxCast dataset

Model Comparison with Expert Intuition

- A small molecule containing a carboxylic acid functional group *should* be amenable to ESI- LC-MS



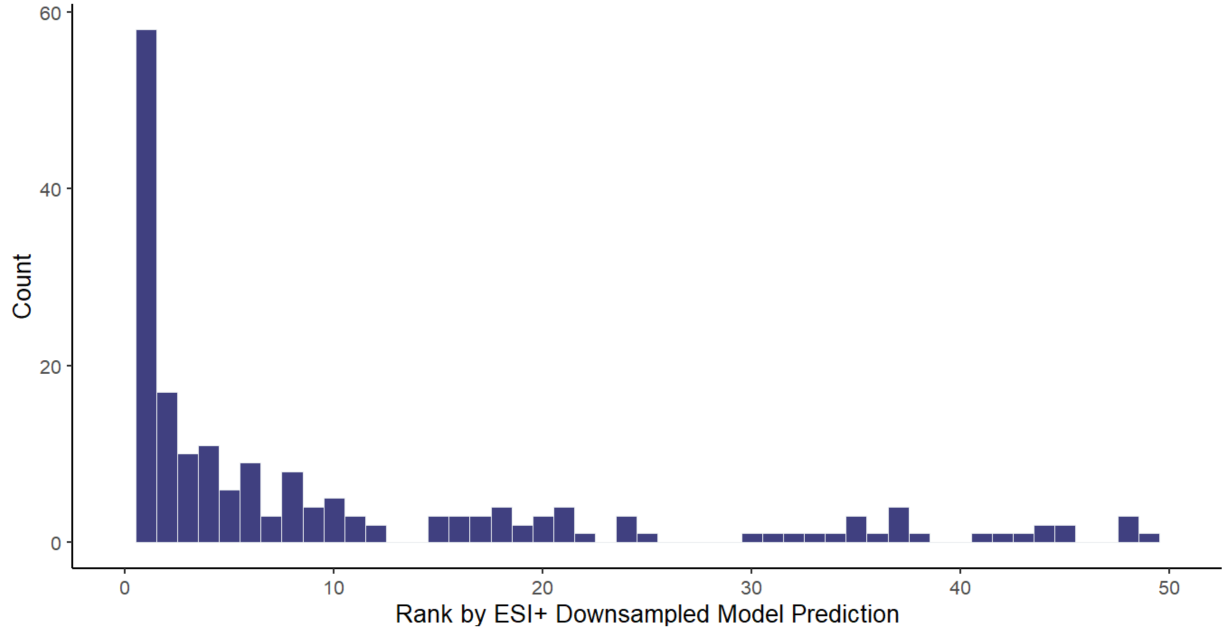
- 773 compounds contained the ToxPrint “bond:C(=O)O_carboxylicAcid_generic” in amenability dataset

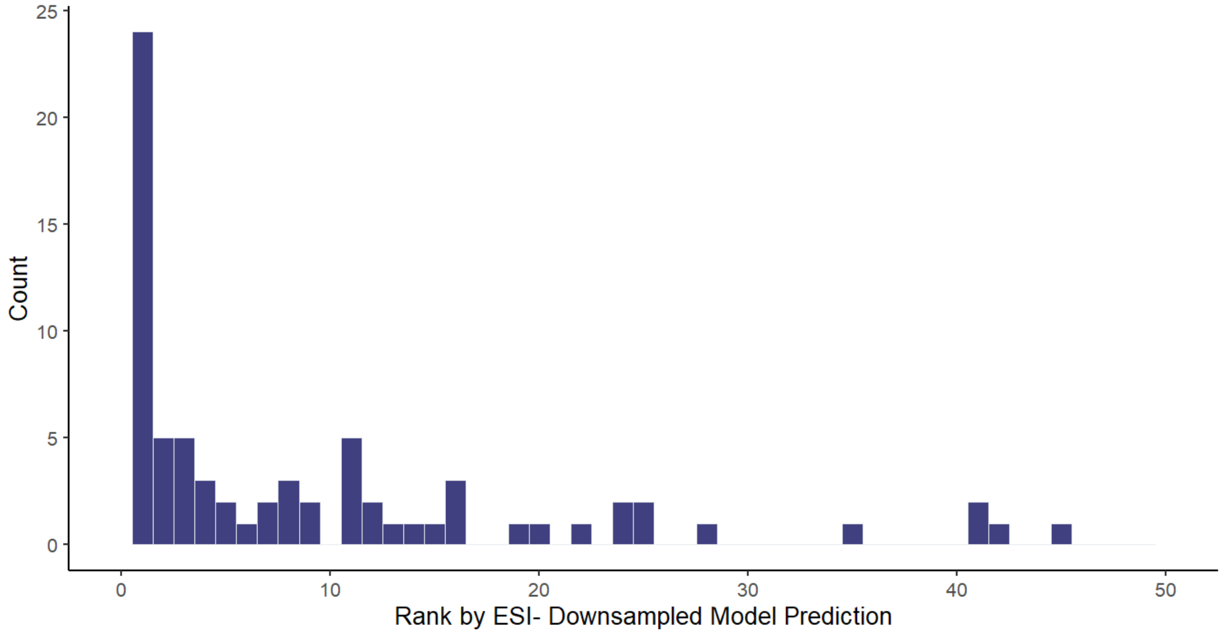


ESI- Downsampled Model		
	Amenable (Prediction)	Unamenable (Prediction)
Detected (Experiment)	728	4
Not-detected (Experiment)	37	9
Sensitivity	**0.95**	
Specificity	**0.69**	
Balanced Accuracy	**0.82**	

Suspect-screening application

- List of ENTACT compounds identified in ESI+ & ESI- LC-MS
  - 228 in ESI+
  - 108 in ESI-
- Retrieved candidates for each molecular formula via Dashboard
  - 13,325 candidates for ESI+
  - 7,079 candidates for ESI-
- Generated amenability predictions for candidate structures
- Rank ordered candidates by amenability probability

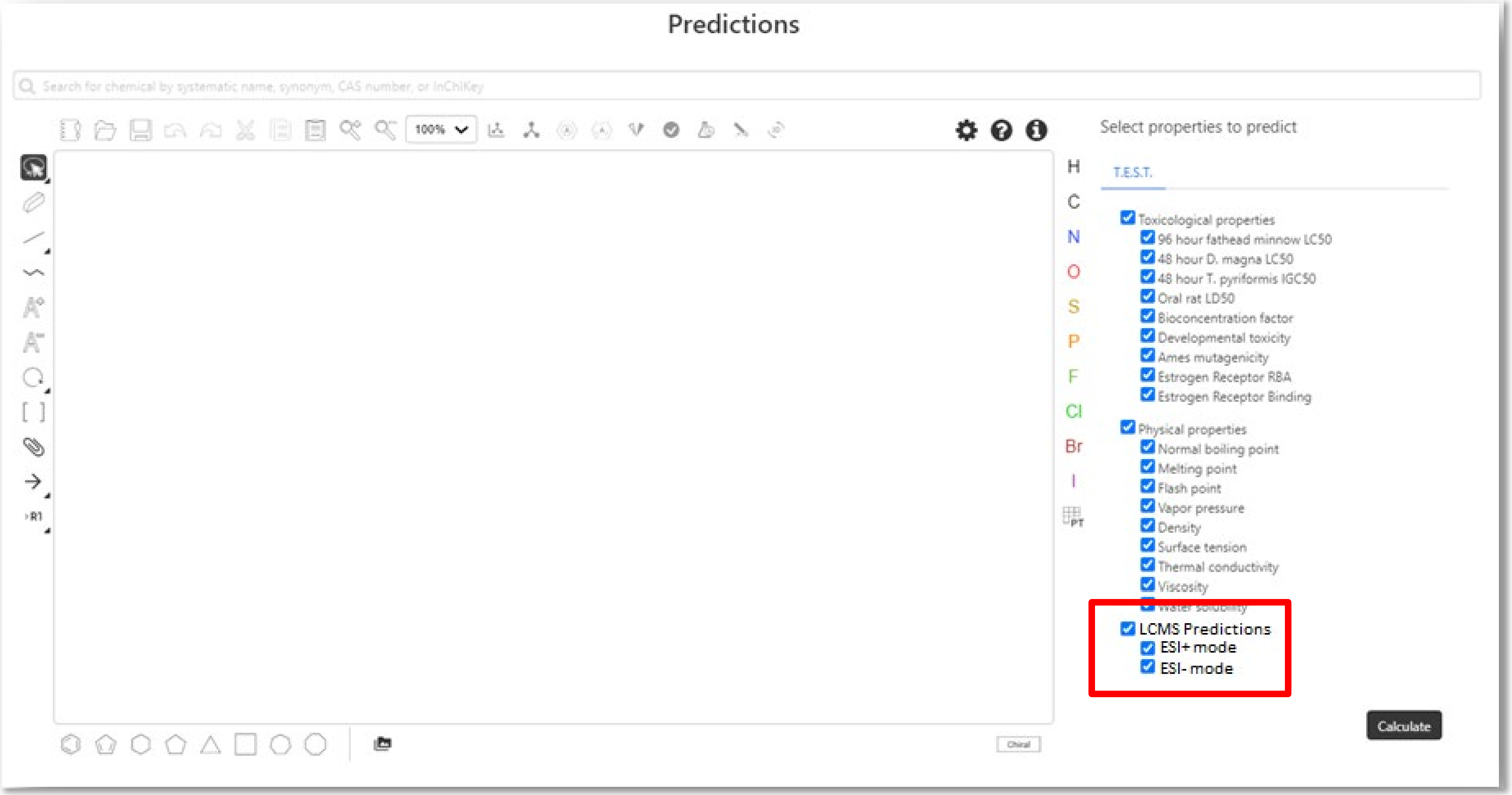




Current & future work

- Manuscript is currently undergoing peer review
- Comparison of model results to Analytical QC data for ToxCast library
  - Good examples – no signal in LC-MS ESI+, ESI- or in GC-MS BUT present and high purity by NMR
- Working with collaborators to gather additional data, particularly unamenable compounds
  - Additional collaborators would be appreciated!
- Future plans
  - Predictions for entire DSSTox database
  - Application for on-the-fly predictions based on a drawn structure

CompTox Chemicals Dashboard mockup - Predictions



Predictions

Search for chemical by systematic name, synonym, CAS number, or InChIKey
100%
H
C
N
O
S
P
F
Cl
Br
I
PT
Chiral
T.E.S.T.
Calculate

Select properties to predict

- ☑ Toxicological properties
  - ☑ 96 hour fathead minnow LC50
  - ☑ 48 hour D. magna LC50
  - ☑ 48 hour T. pyriformis IGC50
  - ☑ Oral rat LD50
  - ☑ Bioconcentration factor
  - ☑ Developmental toxicity
  - ☑ Ames mutagenicity
  - ☑ Estrogen Receptor RBA
  - ☑ Estrogen Receptor Binding
- ☑ Physical properties
  - ☑ Normal boiling point
  - ☑ Melting point
  - ☑ Flash point
  - ☑ Vapor pressure
  - ☑ Density
  - ☑ Surface tension
  - ☑ Thermal conductivity
  - ☑ Viscosity
  - ☑ Water solubility
- ☑ LCMS Predictions
  - ☑ ESI+ mode
  - ☑ ESI- mode

Contributing researchers

Credit: the Research Triangle Foundation

EPA ORD
Hussein Al-Ghoul*
Alex Chao*
Louis Groff*
Jarod Grossman*
Kristin Isaacs
Sarah Laughlin*
Hannah Liberatore
James McCord
Kelsey Miller
Jeff Minucci
Seth Newton
Katherine Phillips
Allison Phillips*
Tom Purucker
Randolph Singh*
Jon Sobus
Mark Strynar
Elin Ulrich
Nelson Yeung*

EPA ORD (cont.)
Kathie Dionisio
Chris Grulke
Kamel Mansouri*
Andrew McEachran*
Ann Richard
Adam Swank
John Wambaugh
Antony Williams

Agilent
Jarod Grossman
Andrew McEachran

GDIT
Ilya Balabin
Tom Transue
Tommy Cathey

* = ORISE/ORAU

Thank you for Listening!