United States Environmental Protection Agency

Predicting Chromatography-tandem Mass Spectrometry Amenability to Improve Non-targeted Analysis

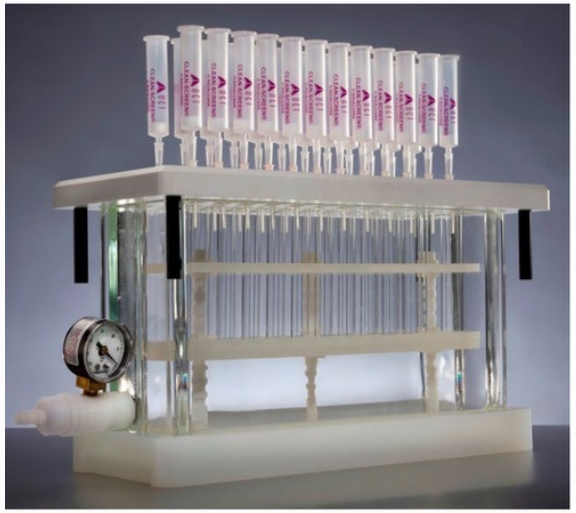
Charles N. Lowe[1], Kristin Isaacs[1],Andrew McEachran[2], Chris Grulke[1], Jon Sobus[1], Elin Ulrich[1], Alex Chao[1], and Antony J. Williams[1]

1. Center for Computational Toxicology and Exposure, U.S. Environmental Protection Agency, 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.

Complex samples, NTA, and the modeling problem

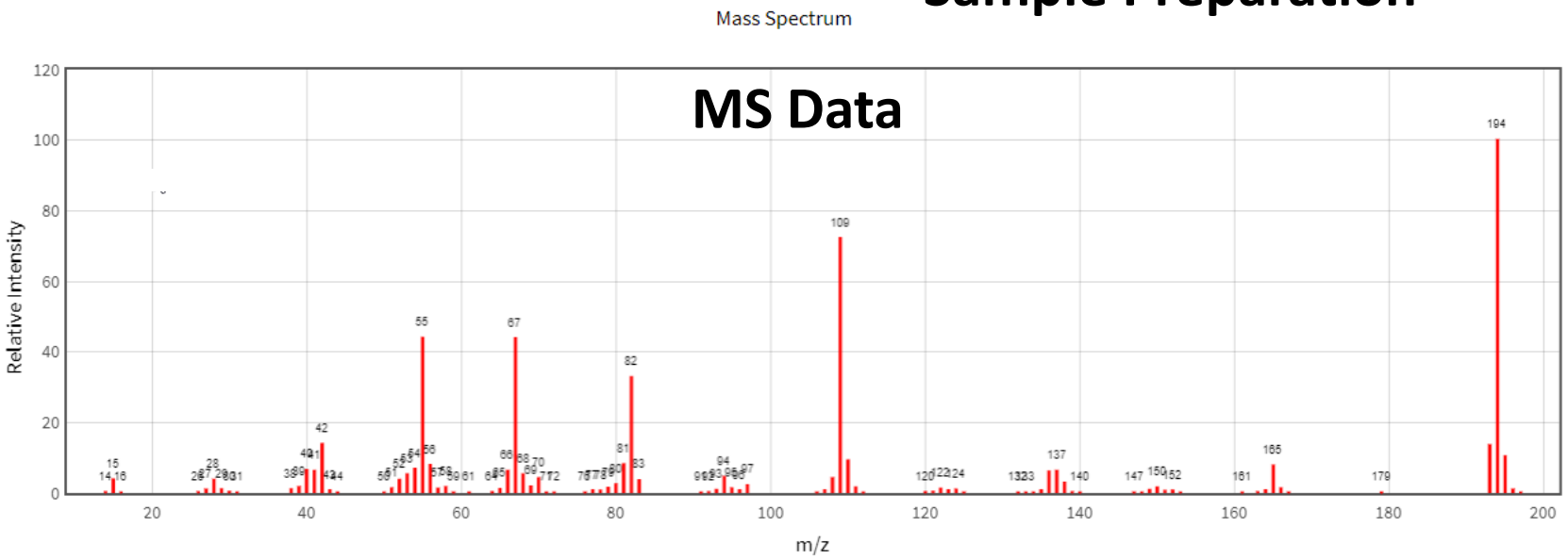
Media Sample



Extraction, Cleanup & Sample Preparation

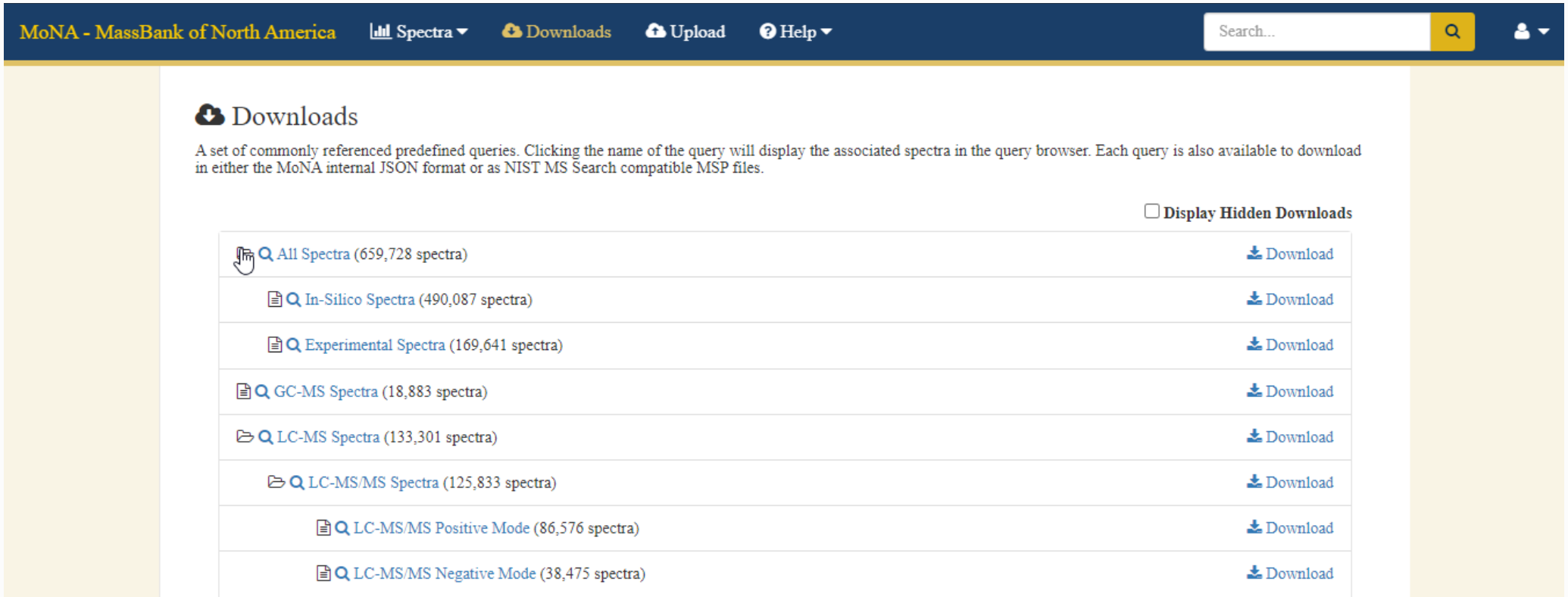
MS Analysis

???



https://images.app.goo.gl/ftRmhwxEtZs95uKv7

Curating a dataset for modeling

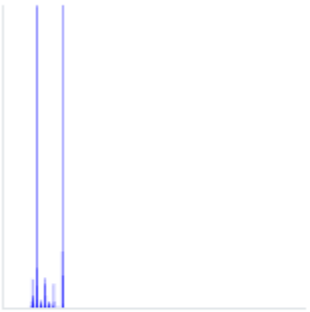


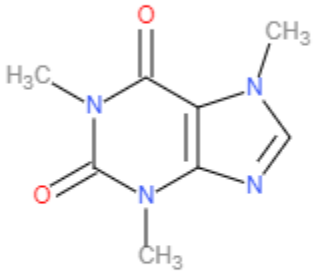
- 772 compounds in derivatized GCMS
- 7,199 compounds in non-derivatized GCMS
- **4,145 compounds in ESI+ LCMS**
- **2,981 compounds in ESI- LCMS**

United States Environmental Protection Agency

Caffeine

Score: ★★★★☆



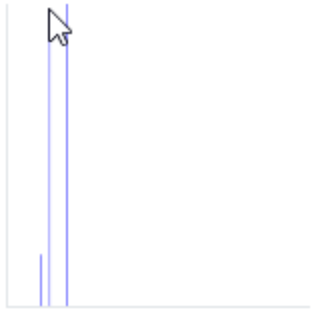


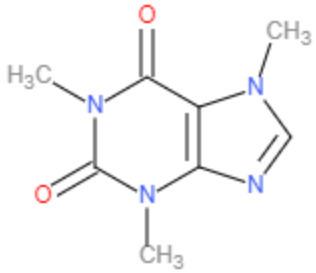
instrument	Pegasus III TOF-MS system...
instrument type	GC-EI-TOF
ms level	MS1
retention index	1880.2430
retention time	724.344 sec
ionization mode	positive
accession	OUF00133
date	2016.01.19 (Created 2010....
author	Tsujimoto Y, Tsugawa H, B...
license	CC BY-SA

Originally submitted to the MassBank High Quality Mass Spectral Database

Caffeine

Score: ★★★★☆





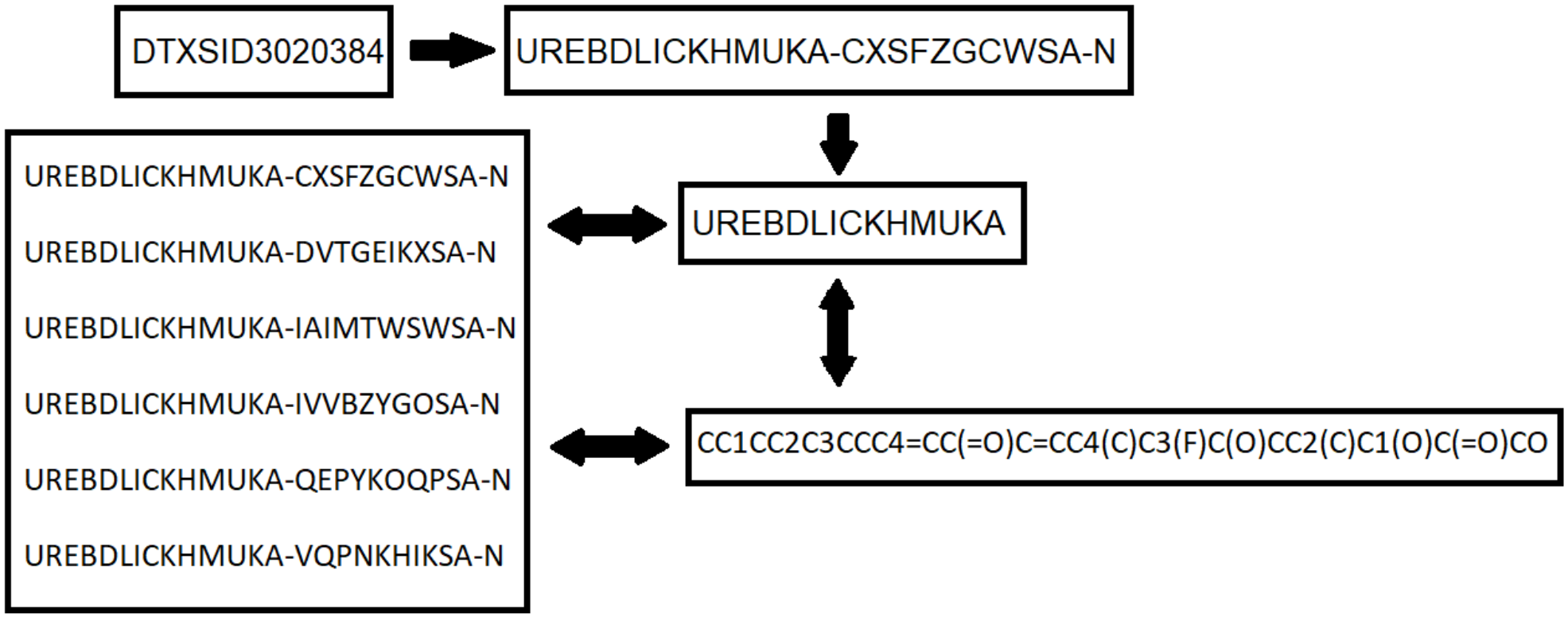
instrument type	QqQ
instrument	Micromass Quattromicro
collision energy	15eV
ionization	ESI
ionization mode	positive
ms level	MS2
precursor m/z	194.9000
precursor type	[M+H]+
accession	PM018511
publication	Alonso-Salces RM, Guillou...

Originally submitted to the RIKEN MS^n Spectral Database for Phytochemicals

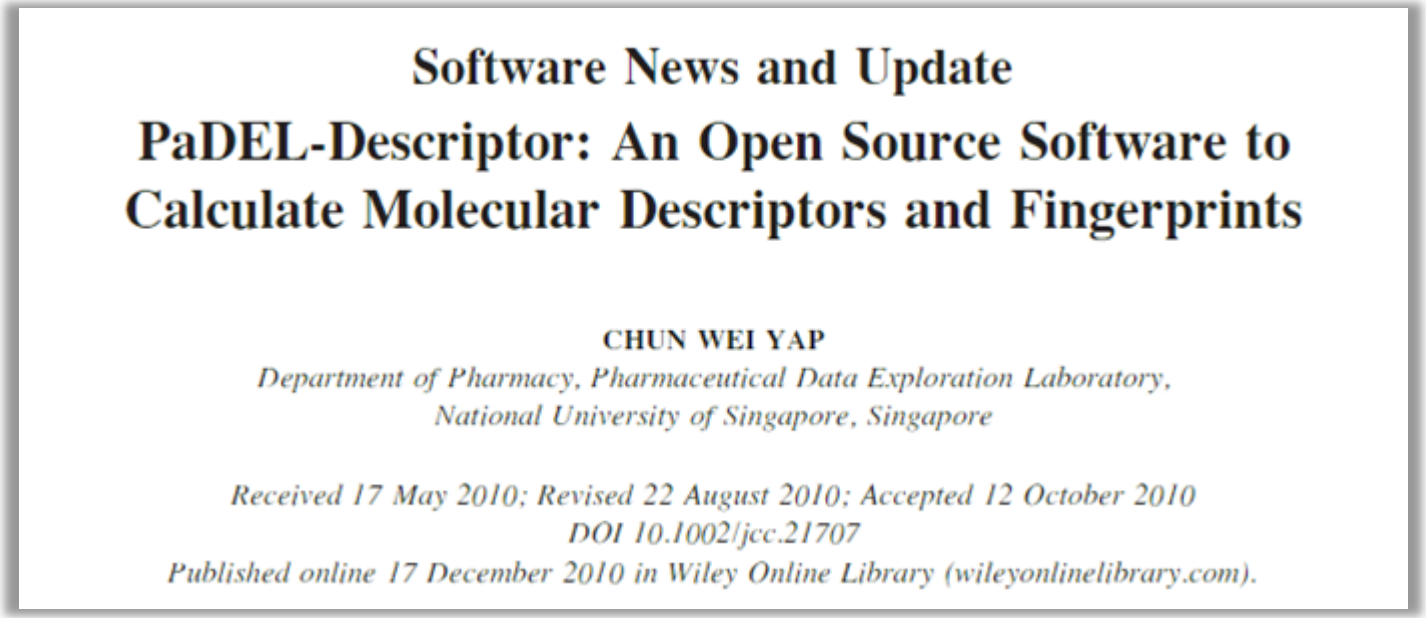
Curating a dataset for modeling

- Only amenable compounds identified in MoNA
  - No unamenable compound data
- ToxCast library LCMS curation
  - Spectra checked individually for quality
    - Provides unamenable compound data
- ESI+ LCMS
  - 403 amenable; 469 unamenable
- ESI- LCMS
  - 464 amenable; 415 unamenable
- Caveat: some of these unamenable compounds are amenable based on MoNA*

Curating a dataset for modeling



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
  - logP predictions, etc..

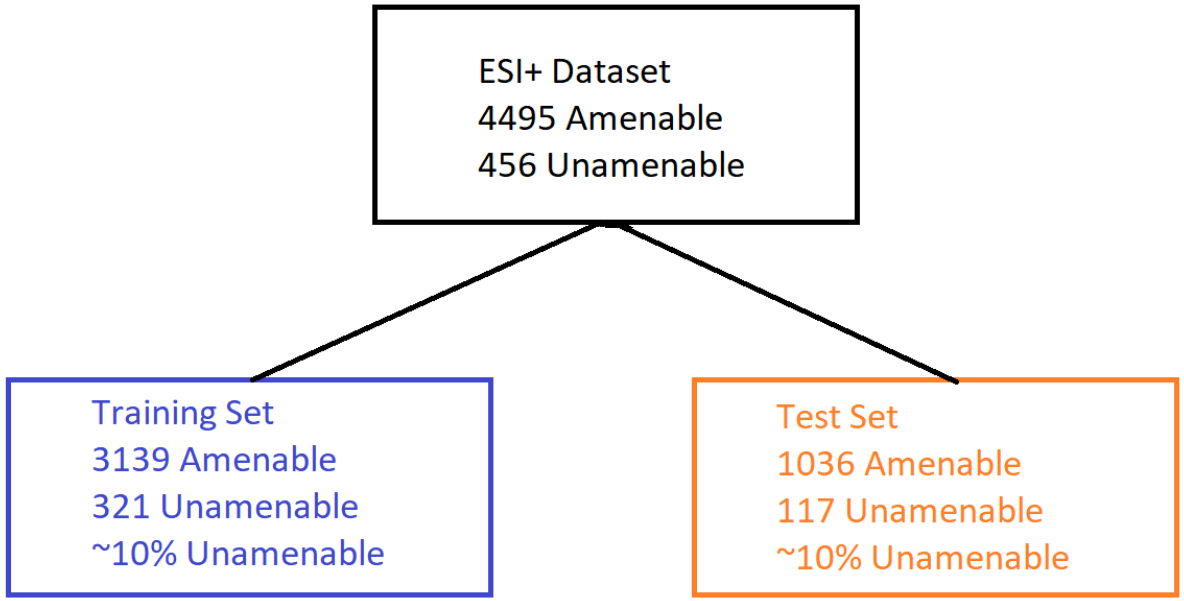
Cleaning and reduction of descriptor space

- 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 pair-wise 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 → 498 descriptors

Datasets suitable for modeling

- Models randomly divided into training and test sets
  - 75% of data for training, 25% for testing
  - Data stratified to maintain proportions in outcome variable
  - Different for each model
  - InChIKey skeleton as identifier



Machine learning approach

- Random forest models for two endpoints
  - ESI+ LCMS, ESI- LCMS
  - Balance training set with either upsampling or downsampling
  - Optimize hyperparameters via grid search
    - Number of decision trees
    - Number of random descriptors selected at each node
  - 5-fold cross validation
  - Y-randomization
    - Randomly scramble endpoint, descriptors left intact

Performance metrics

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

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

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

Office of Research and Development
Center for Computational Toxicology and Exposure

Machine learning approach

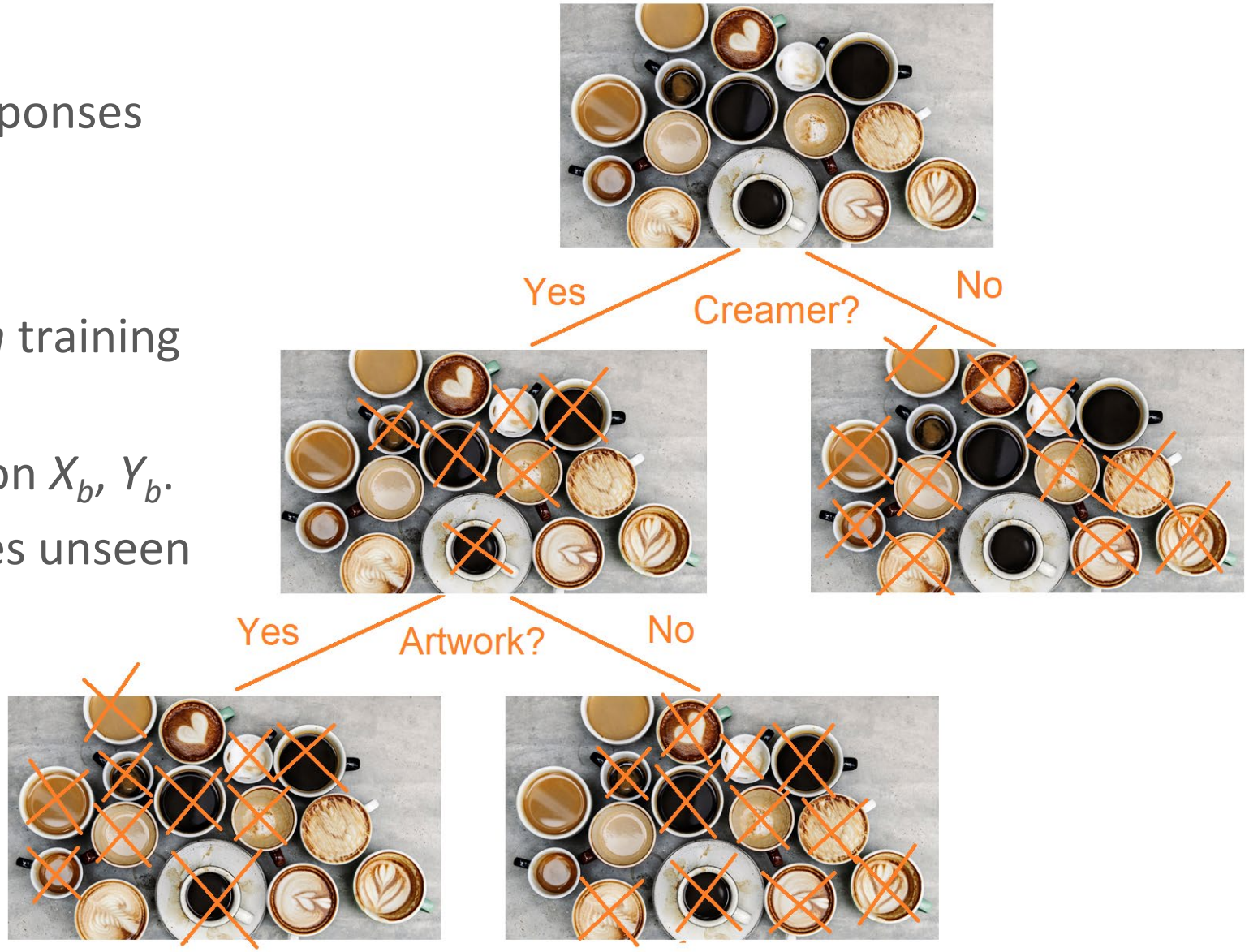
Random Forest Algorithm

Training set $X = x_1x_2...x_n$ with responses

$Y = y_1y_2...y_n$

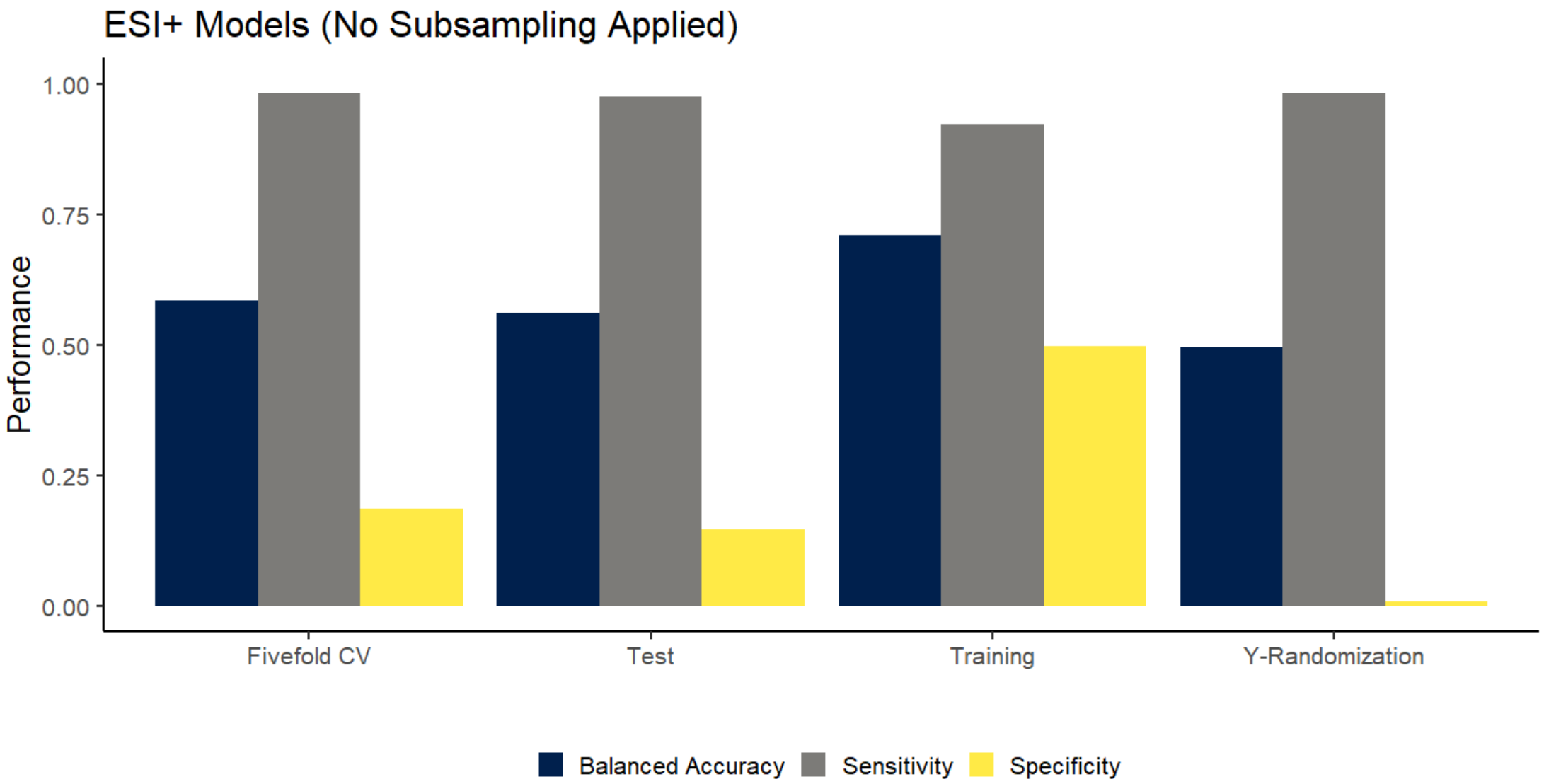
For $b = 1,...,B$

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.



Office of Research and Development
Center for Computational Toxicology and Exposure

Model performance

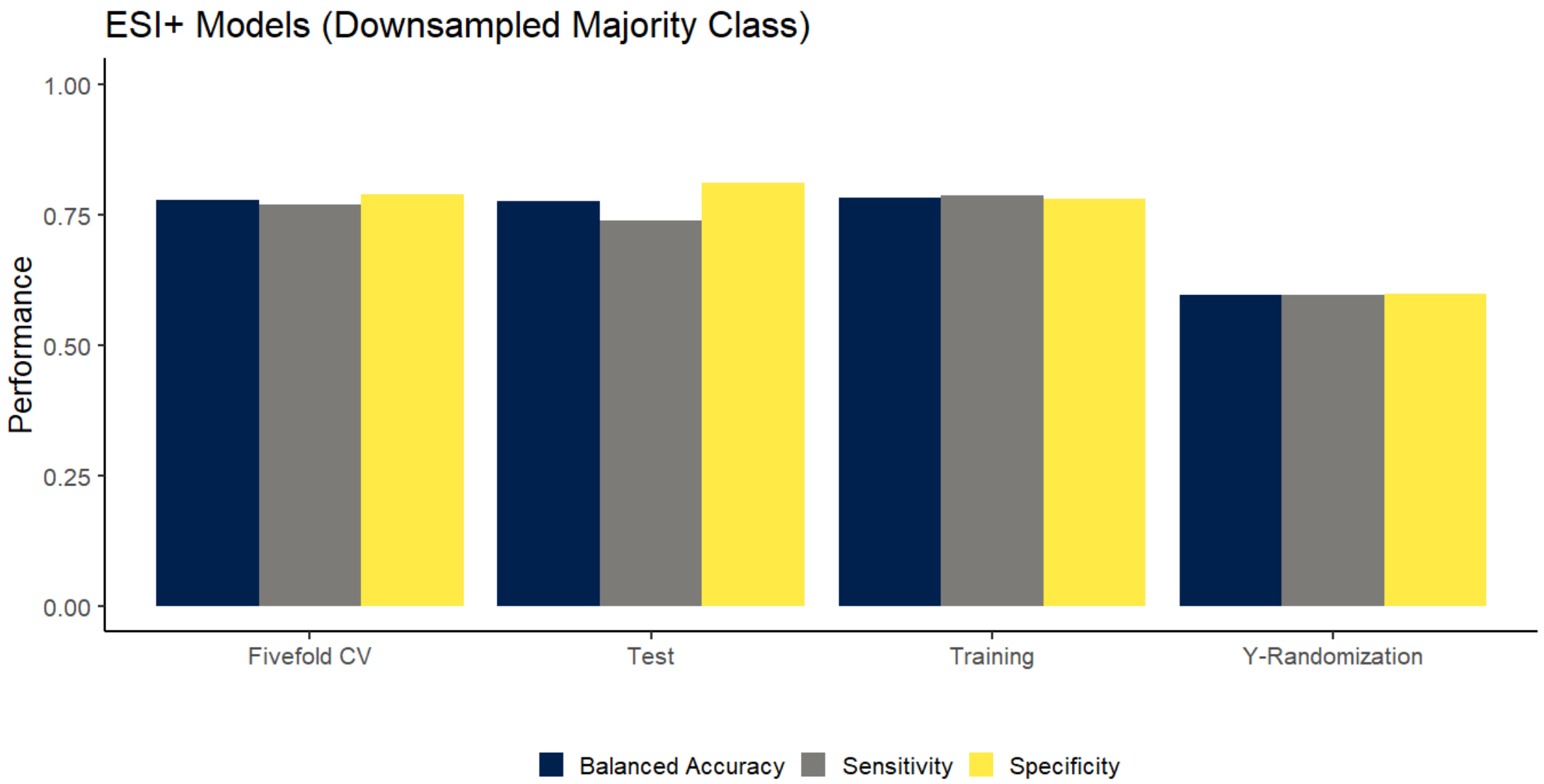


Office of Research and Development
Center for Computational Toxicology and Exposure

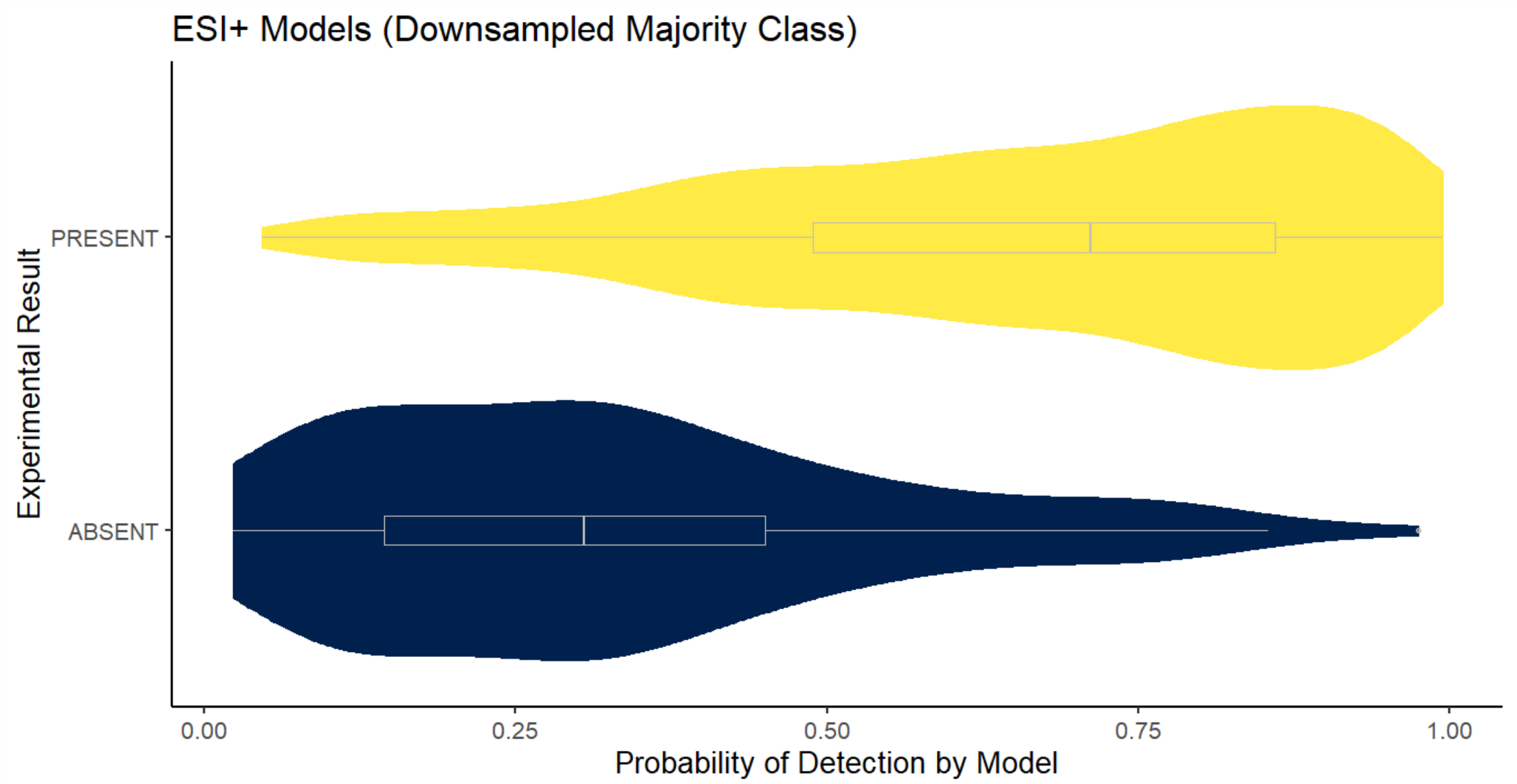
Model summary

Imbalanced training sets lead to bad models!

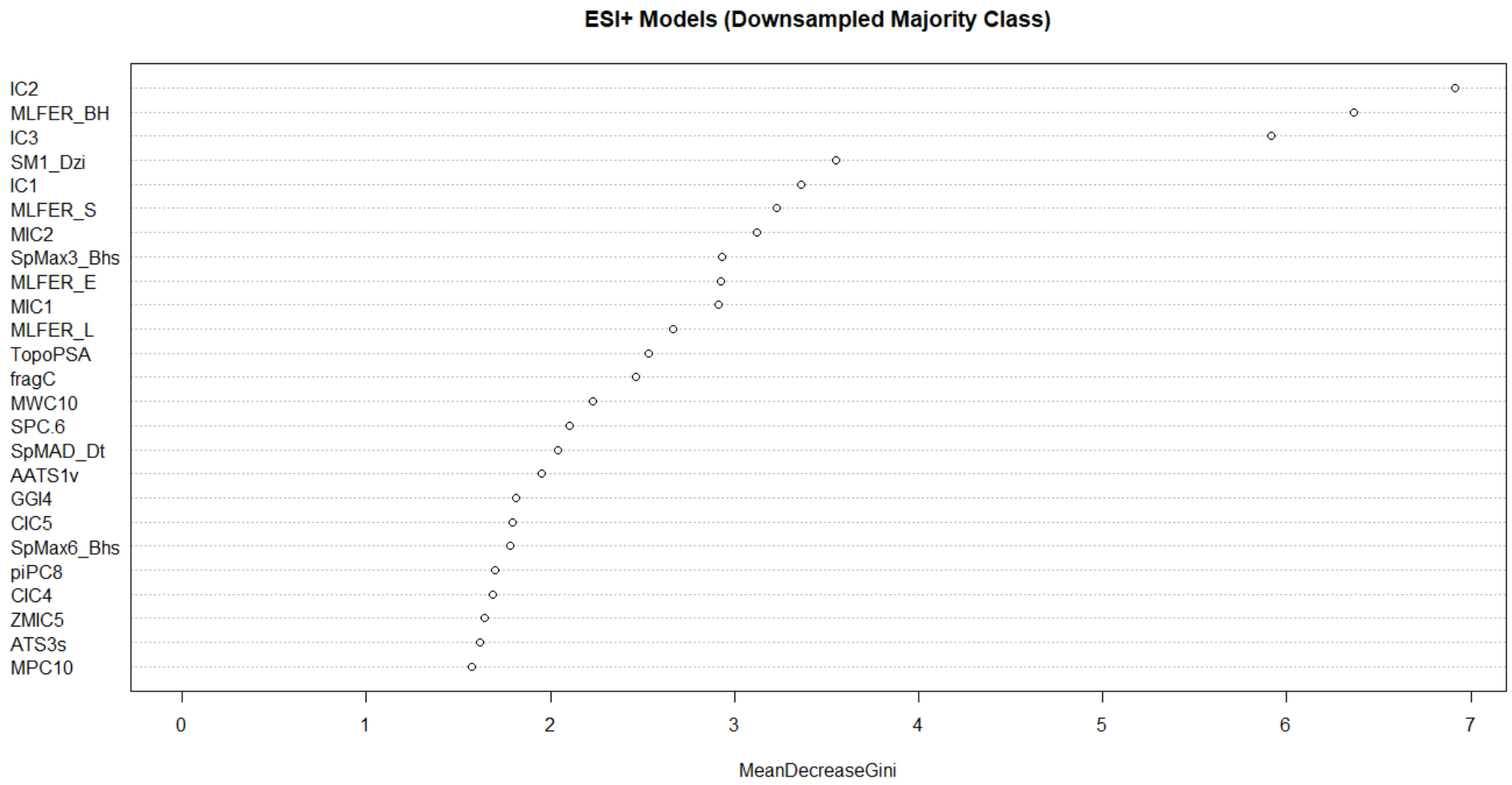
Model performance



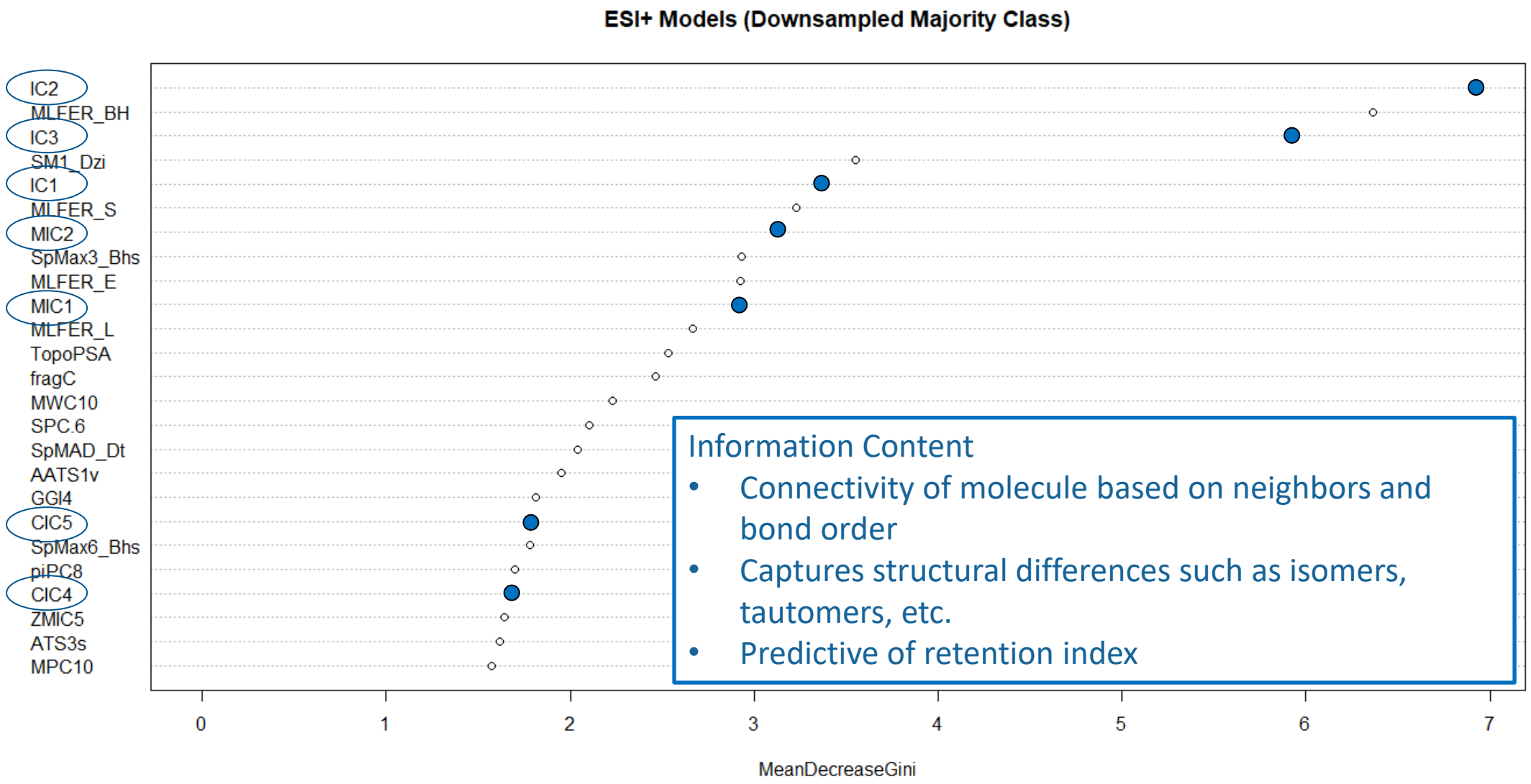
Model performance



Model performance



Mechanistic interpretation

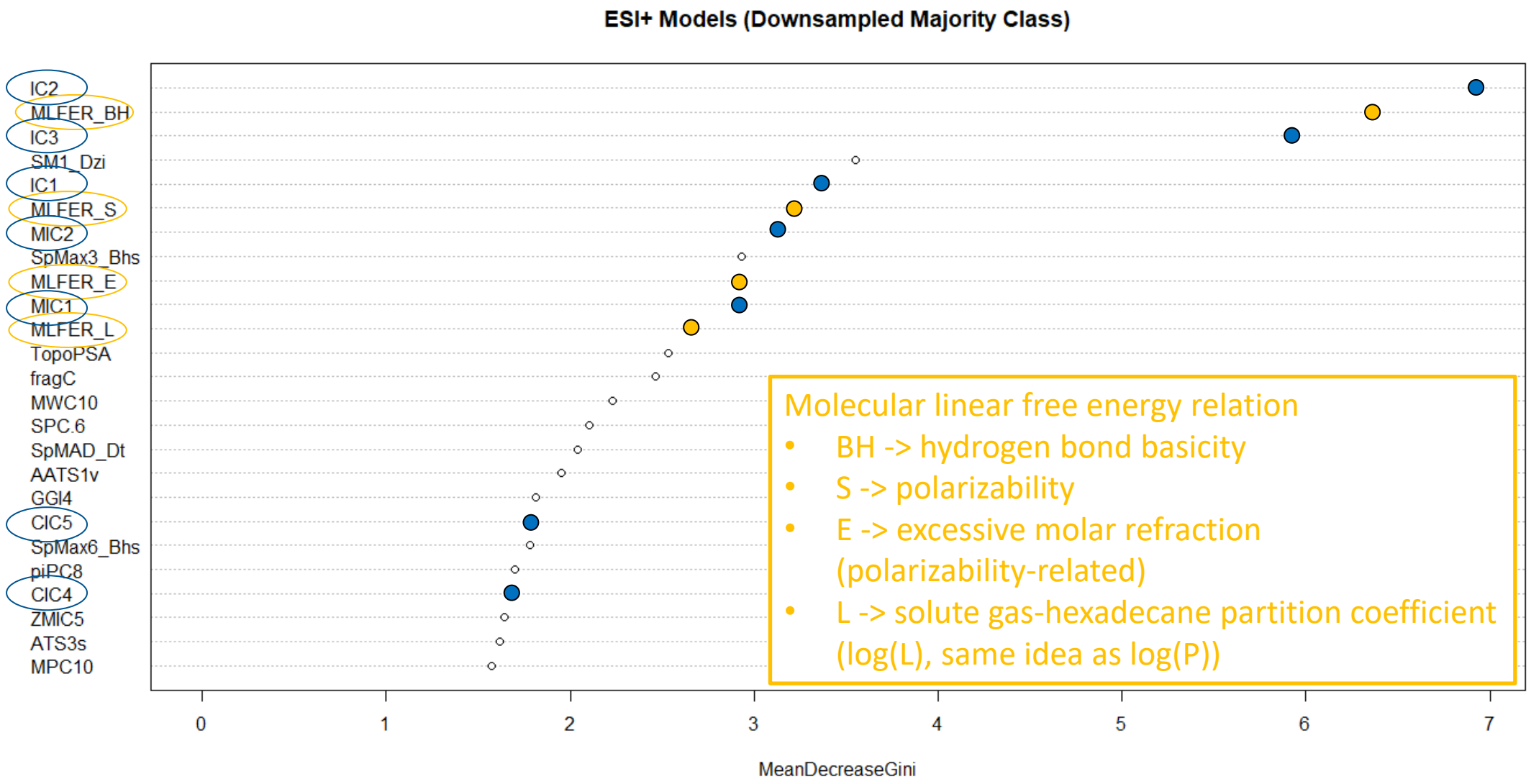


Information Content

- Connectivity of molecule based on neighbors and bond order
- Captures structural differences such as isomers, tautomers, etc.
- Predictive of retention index

Office of Research and Development
Center for Computational Toxicology and Exposure

Mechanistic interpretation

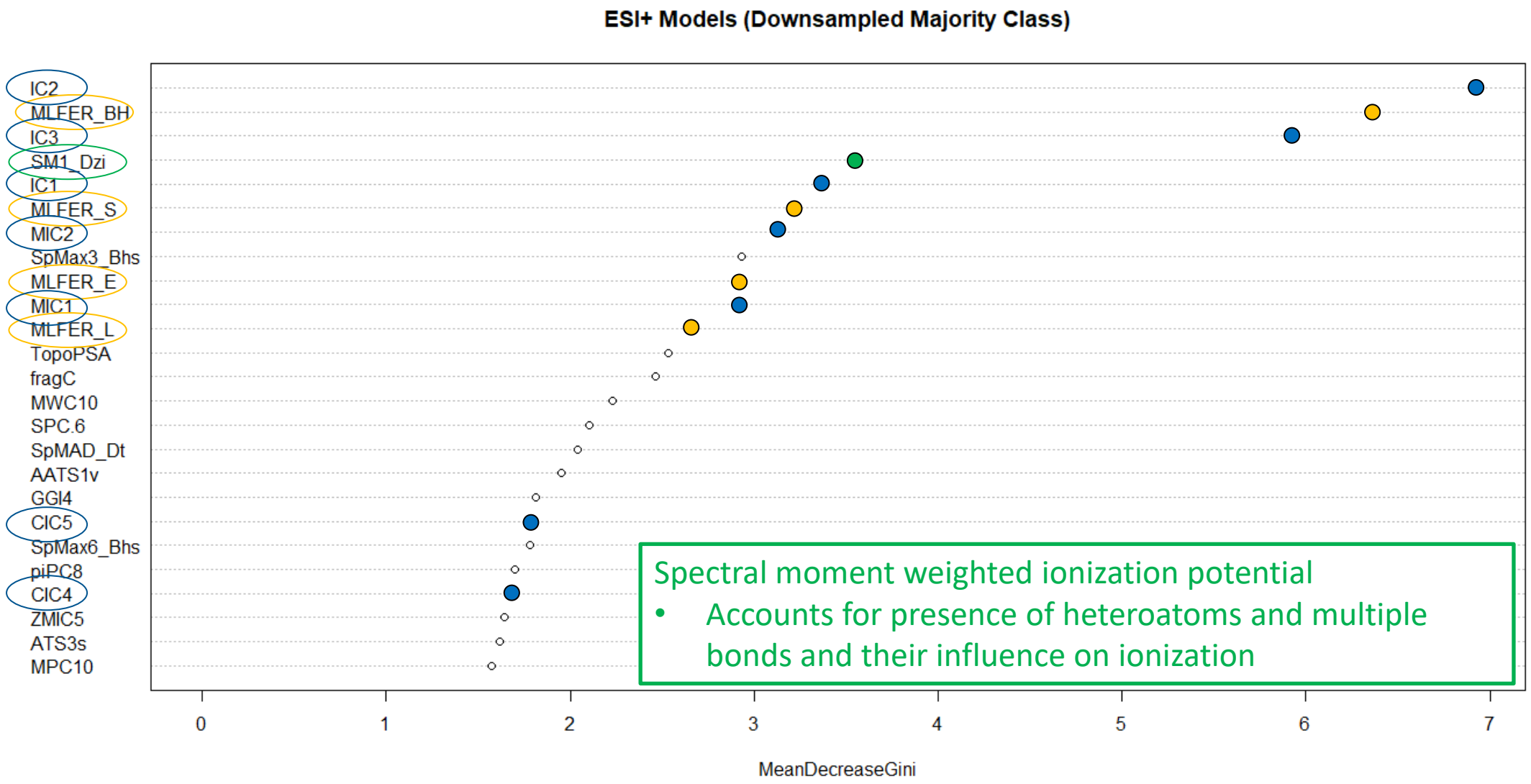


Molecular linear free energy relation

- BH -> hydrogen bond basicity
- S -> polarizability
- E -> excessive molar refraction (polarizability-related)
- L -> solute gas-hexadecane partition coefficient (log(L), same idea as log(P))

Office of Research and Development
Center for Computational Toxicology and Exposure

Mechanistic interpretation



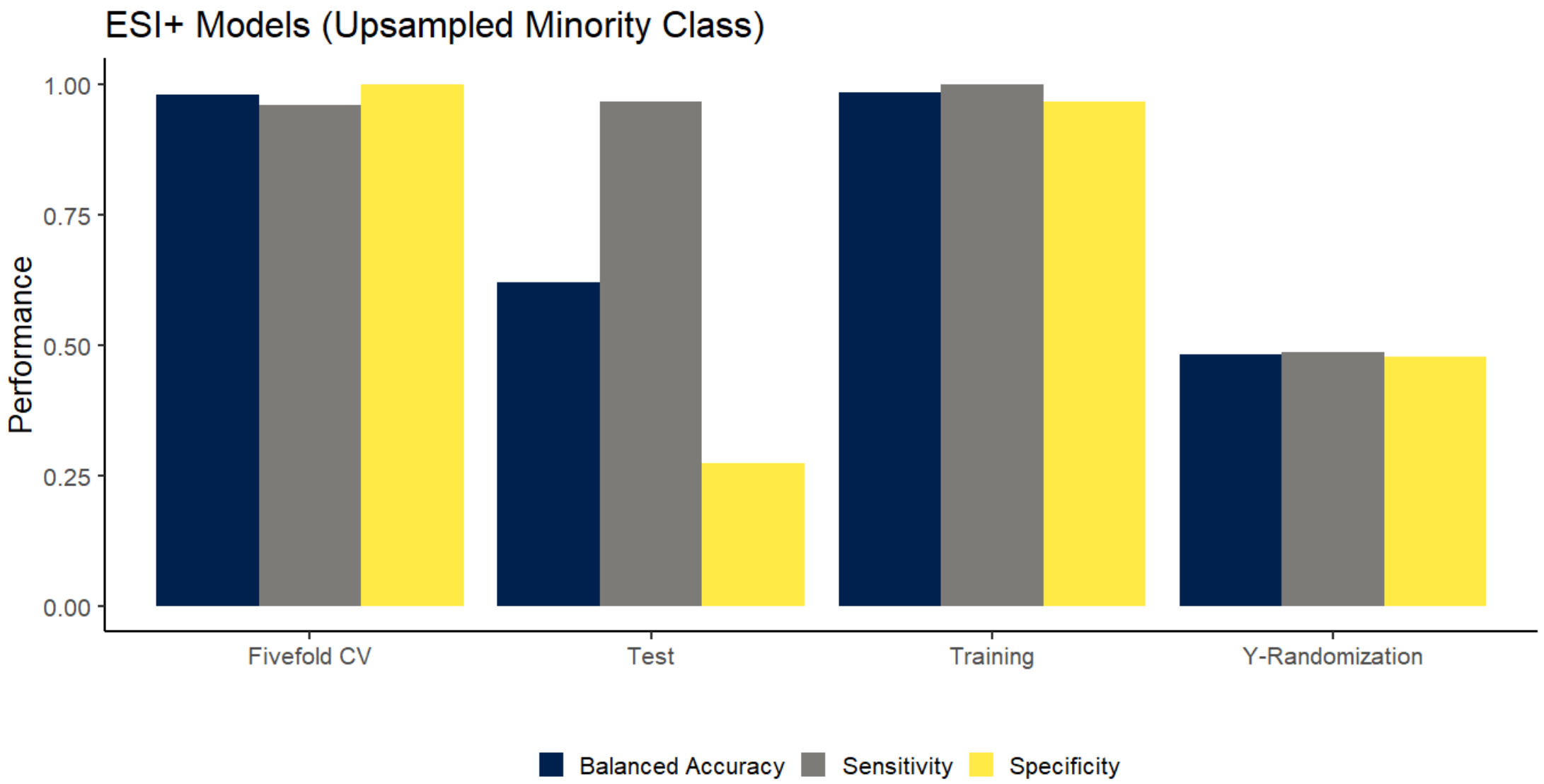
Spectral moment weighted ionization potential

- Accounts for presence of heteroatoms and multiple bonds and their influence on ionization

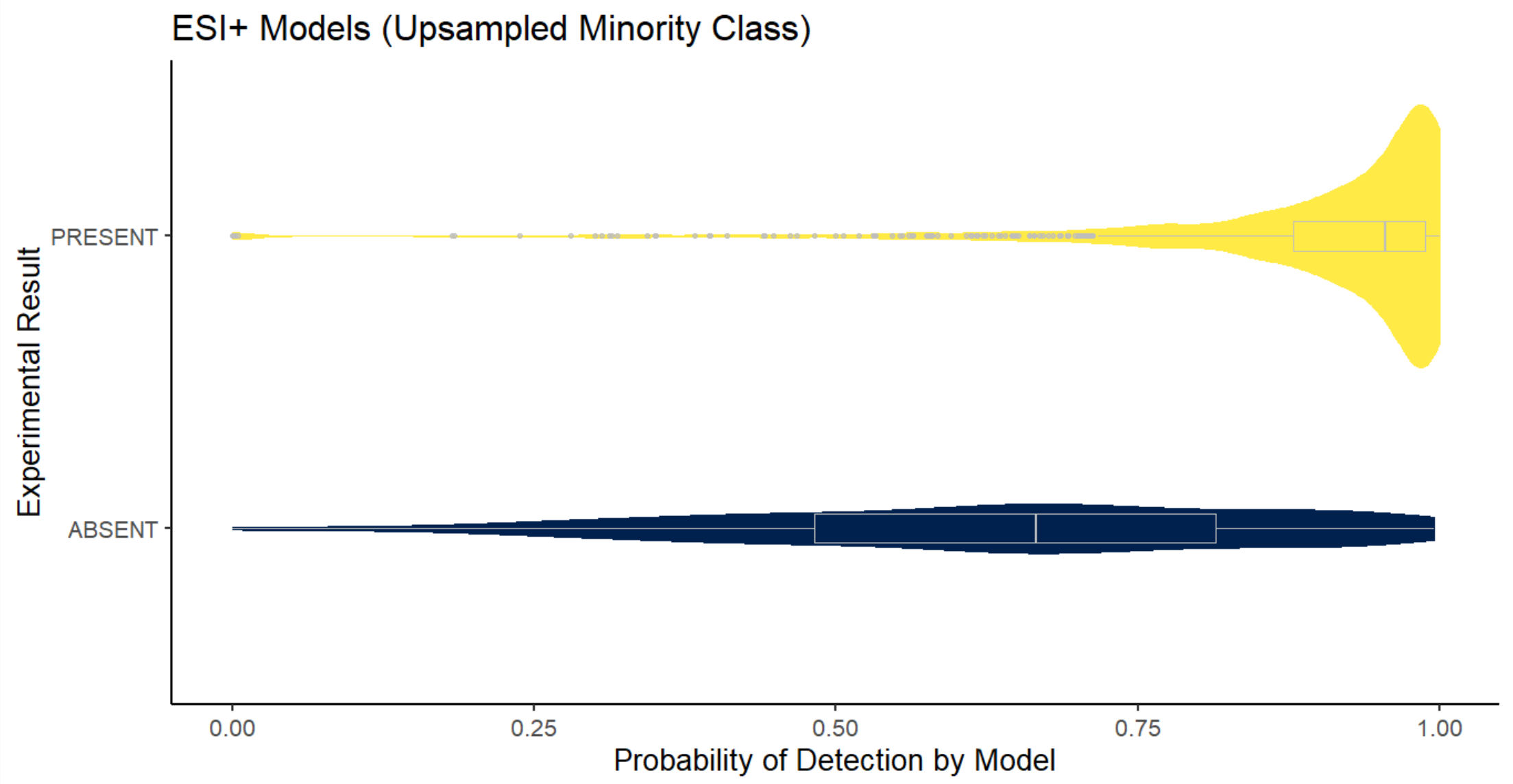
Model summary

- Downsampled model provides excellent predictions for both amenable and unamenable compounds
  - Caveat: reduces sample space of amenable compounds
    - May not accurately predict every amenable compound
- Preferred model for ranking candidates in a suspect-screening analysis

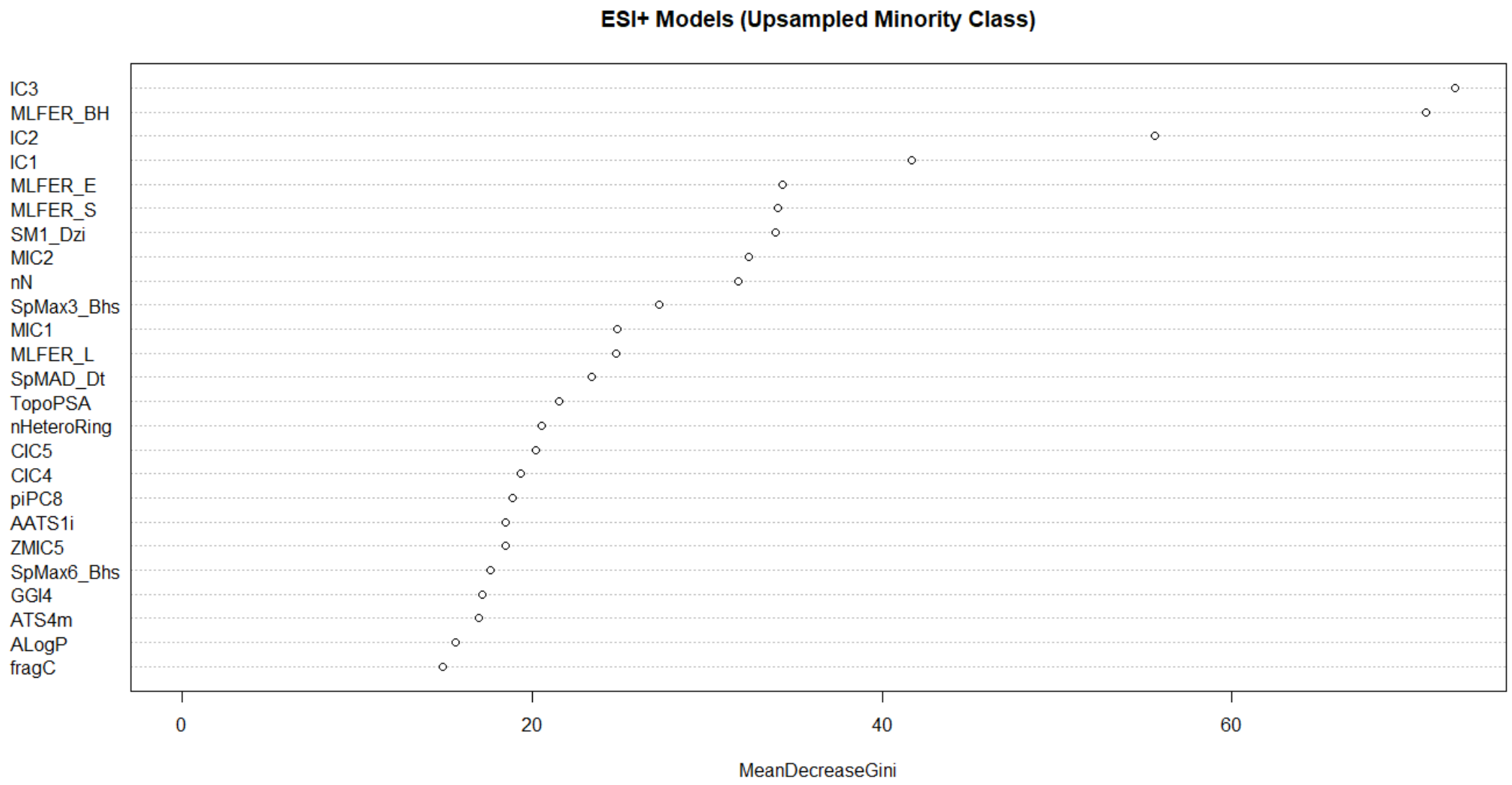
Model performance



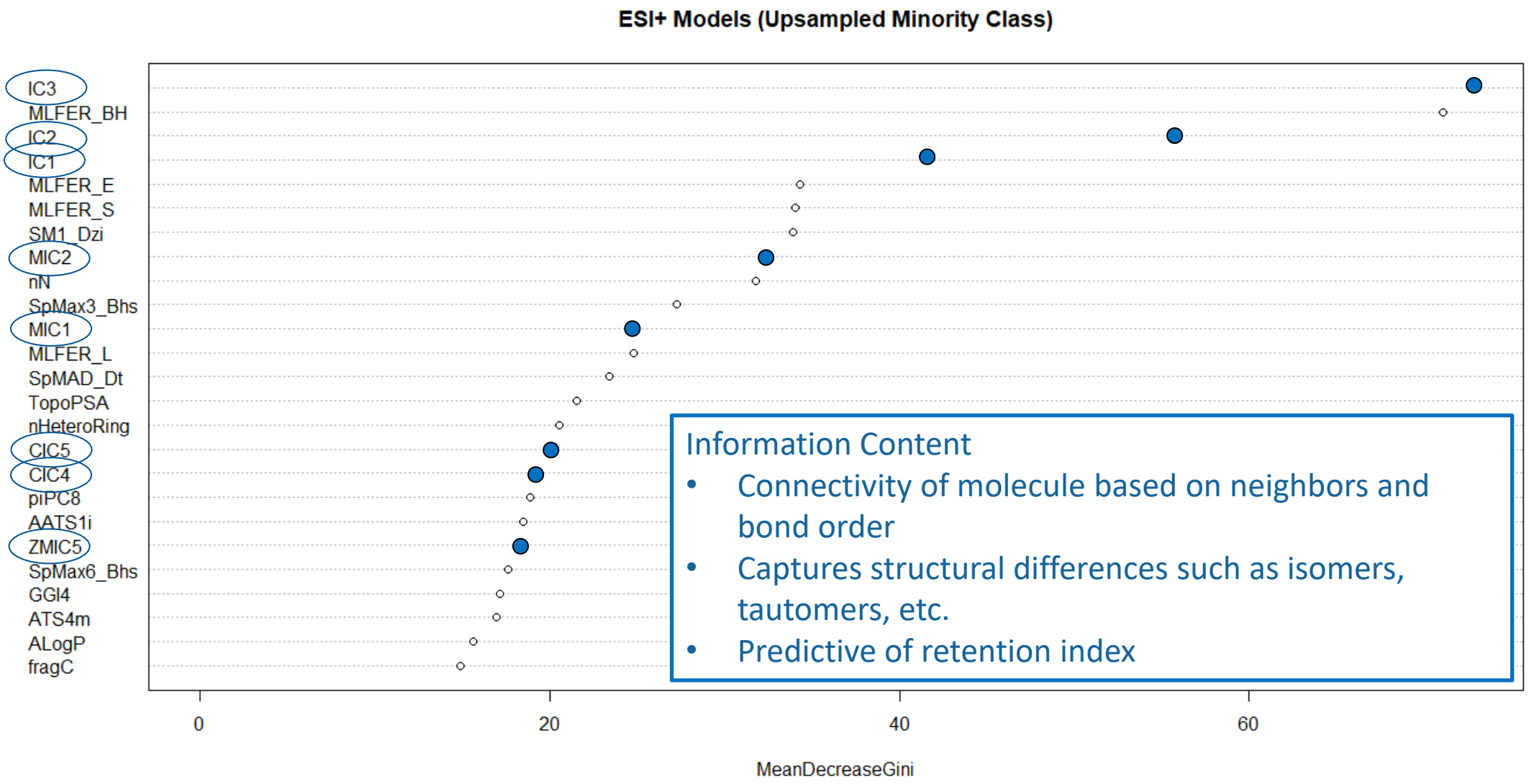
Model performance



Model performance



Mechanistic interpretation

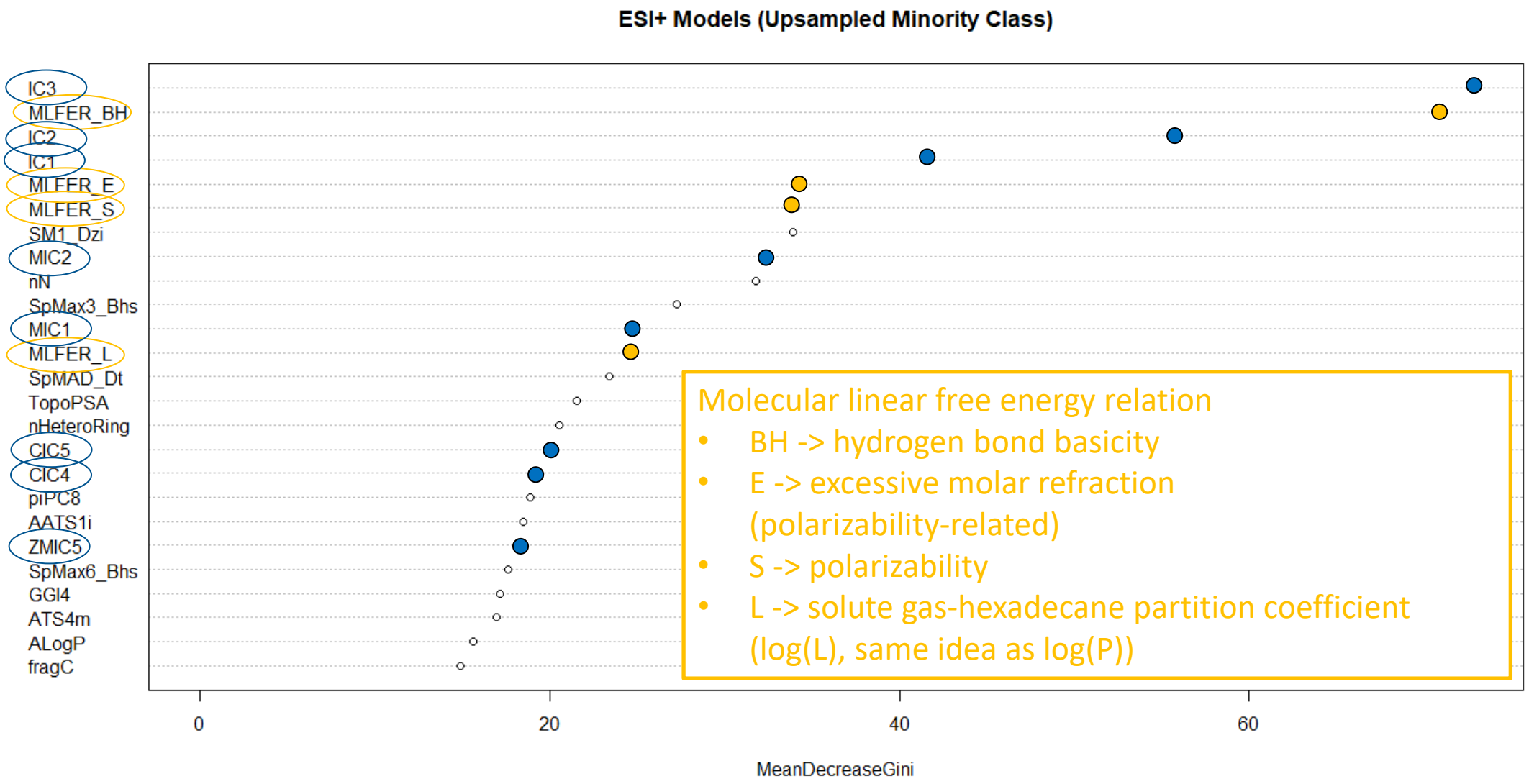


Information Content

- Connectivity of molecule based on neighbors and bond order
- Captures structural differences such as isomers, tautomers, etc.
- Predictive of retention index

Office of Research and Development
Center for Computational Toxicology and Exposure

Mechanistic interpretation

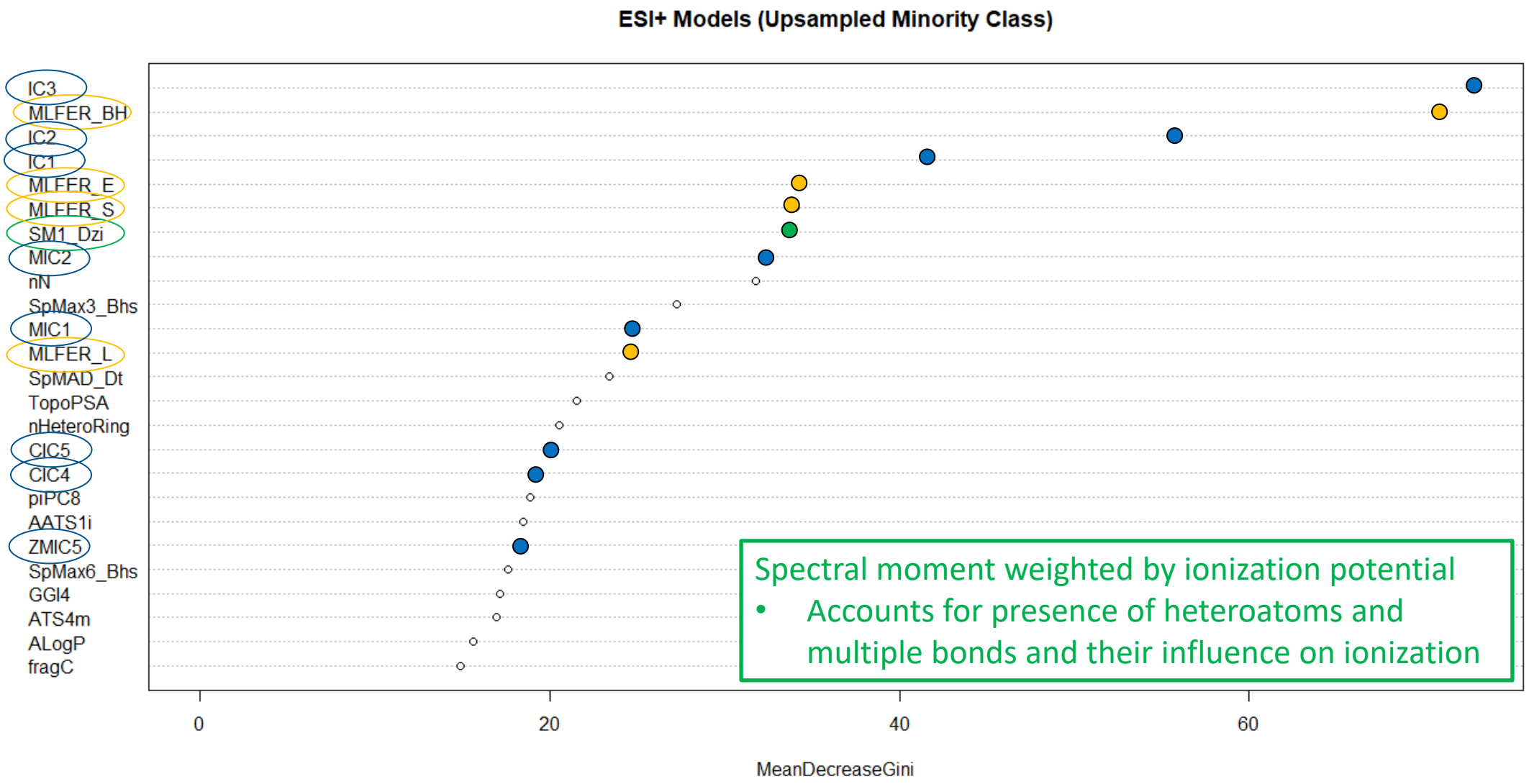


Molecular linear free energy relation

- BH -> hydrogen bond basicity
- E -> excessive molar refraction (polarizability-related)
- S -> polarizability
- L -> solute gas-hexadecane partition coefficient (log(L), same idea as log(P))

Office of Research and Development
Center for Computational Toxicology and Exposure

Mechanistic interpretation



Spectral moment weighted by ionization potential

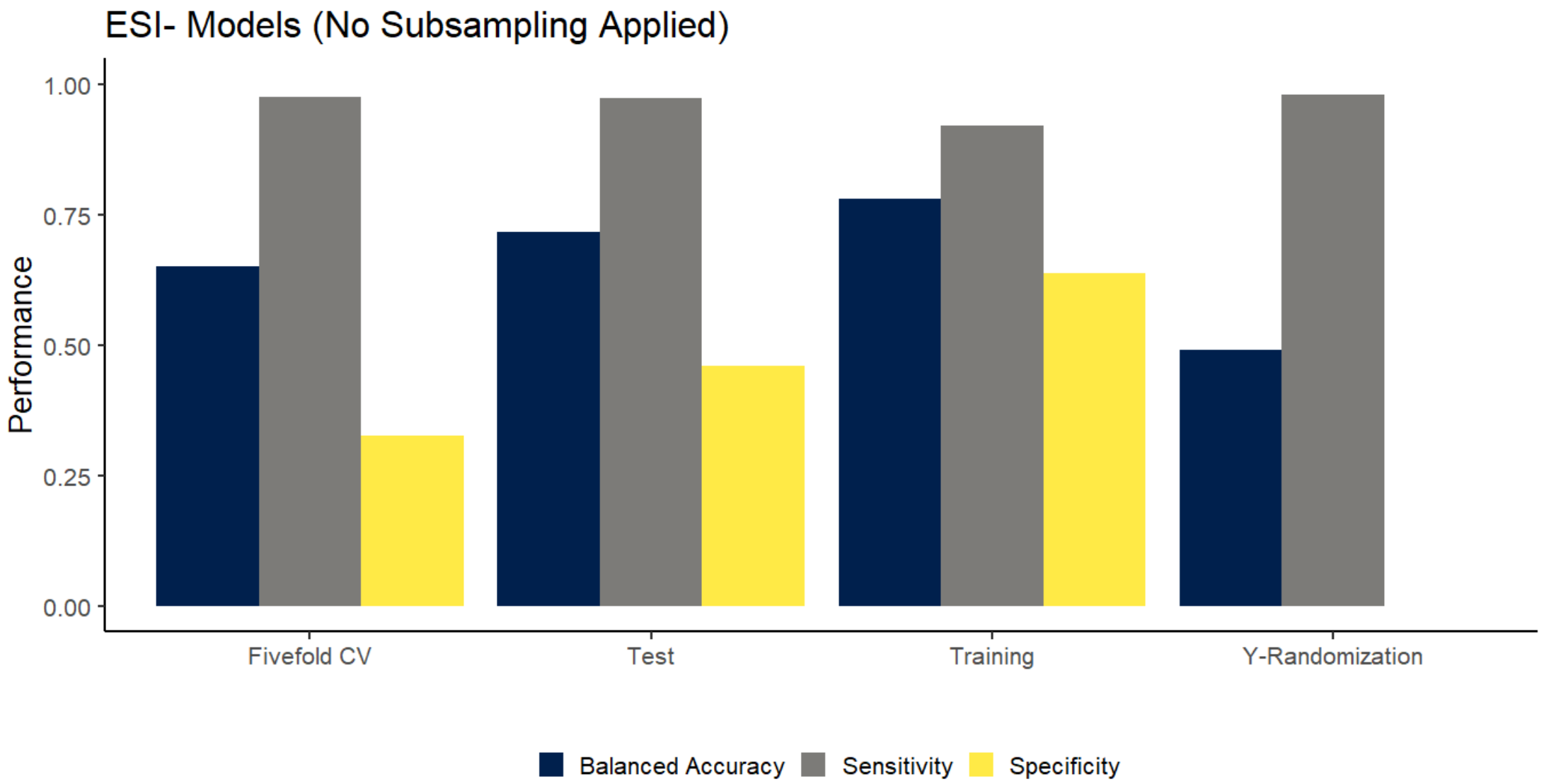
- Accounts for presence of heteroatoms and multiple bonds and their influence on ionization

Office of Research and Development
Center for Computational Toxicology and Exposure

Model summary

- Upsampled model provides excellent predictions for amenable compounds
  - Much larger sample space than downsampled model
  - Weak predictive power for unamenable compounds
  - Too optimistic for suspect-screening
- Preferred model for establishing which chemicals *may* be amenable to method
  - establishing a list of chemical standards

Model performance

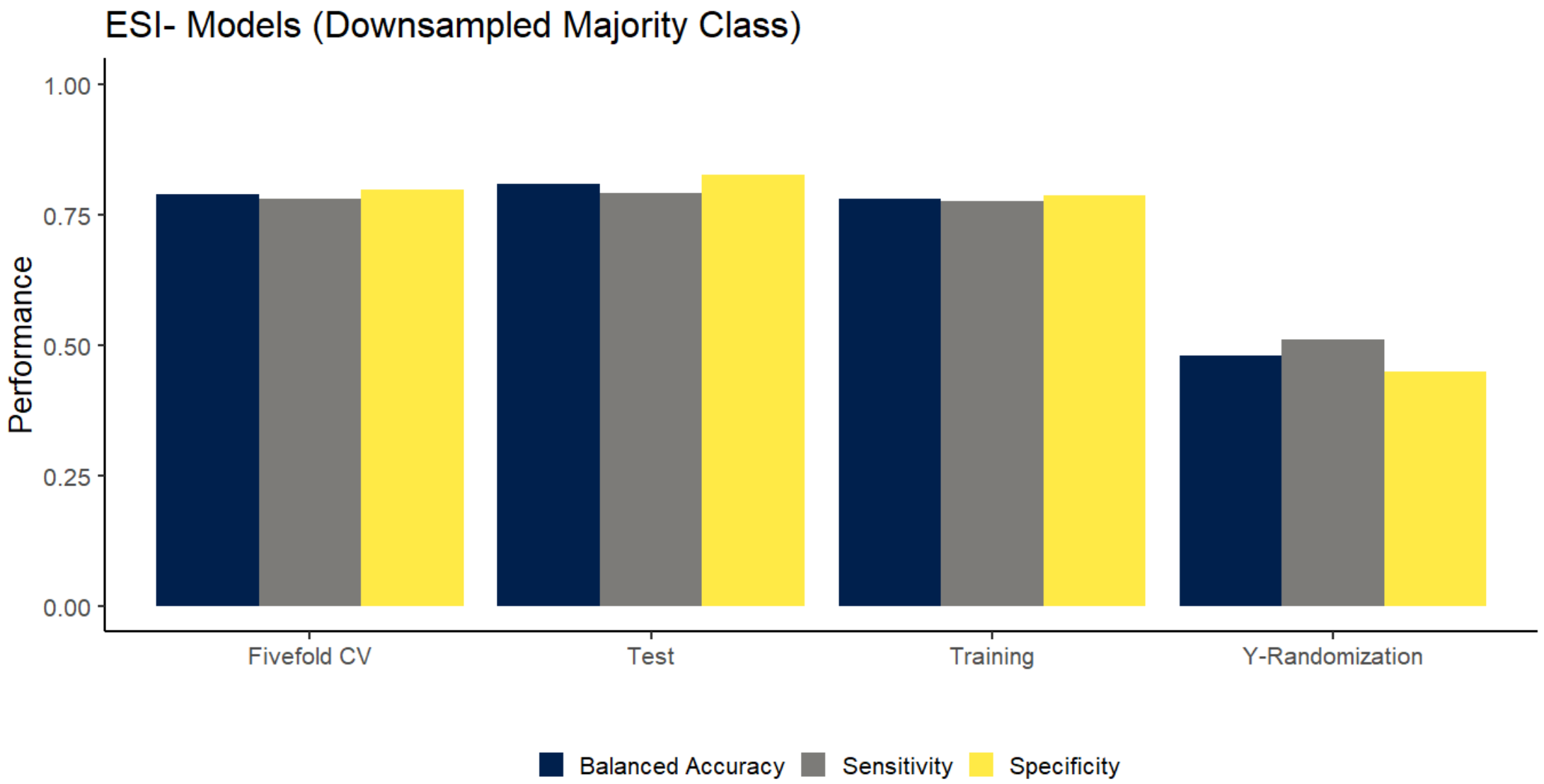


Office of Research and Development
Center for Computational Toxicology and Exposure

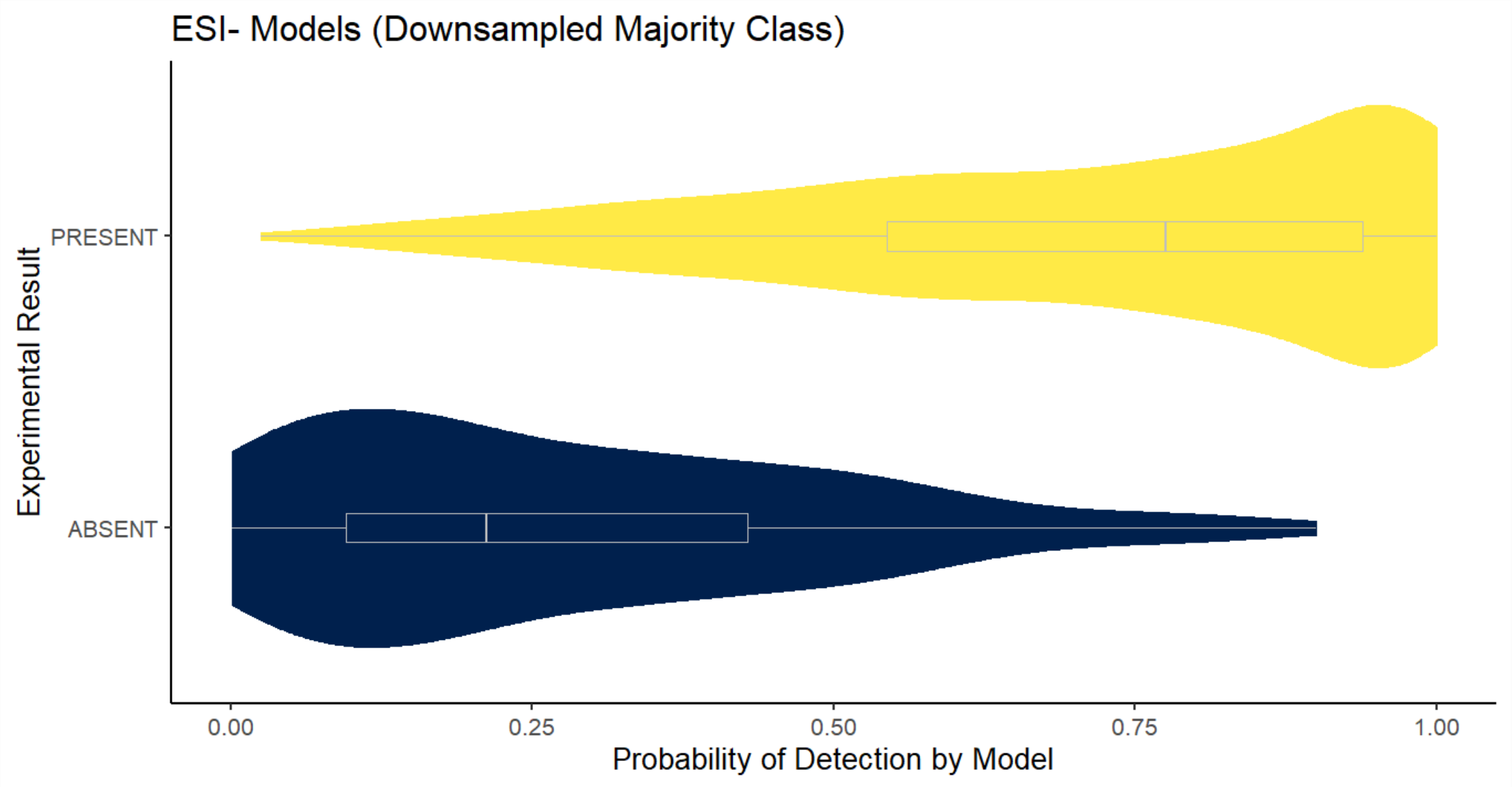
Model summary

Imbalanced training sets lead to bad models!

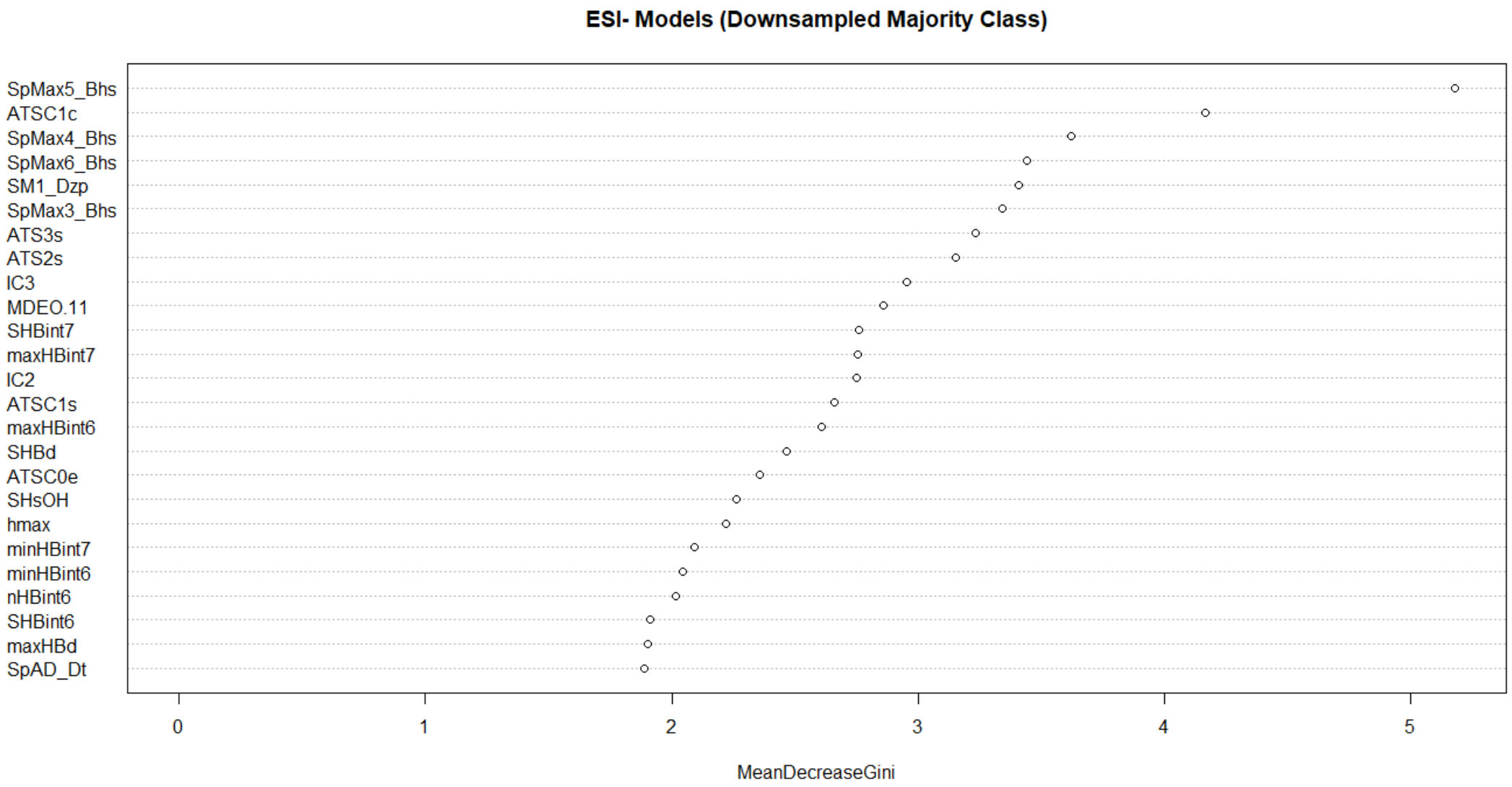
Model performance



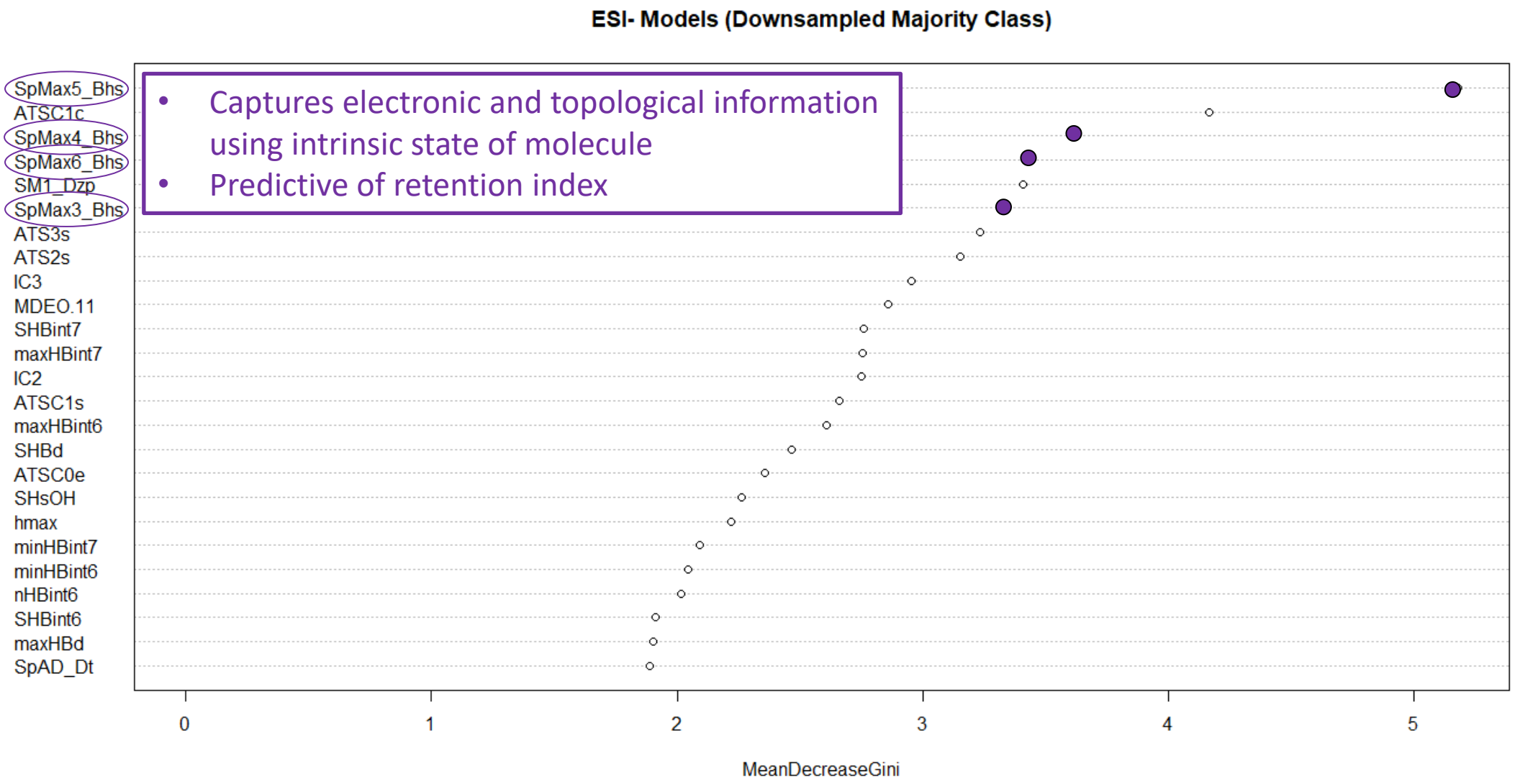
Model performance



Model performance



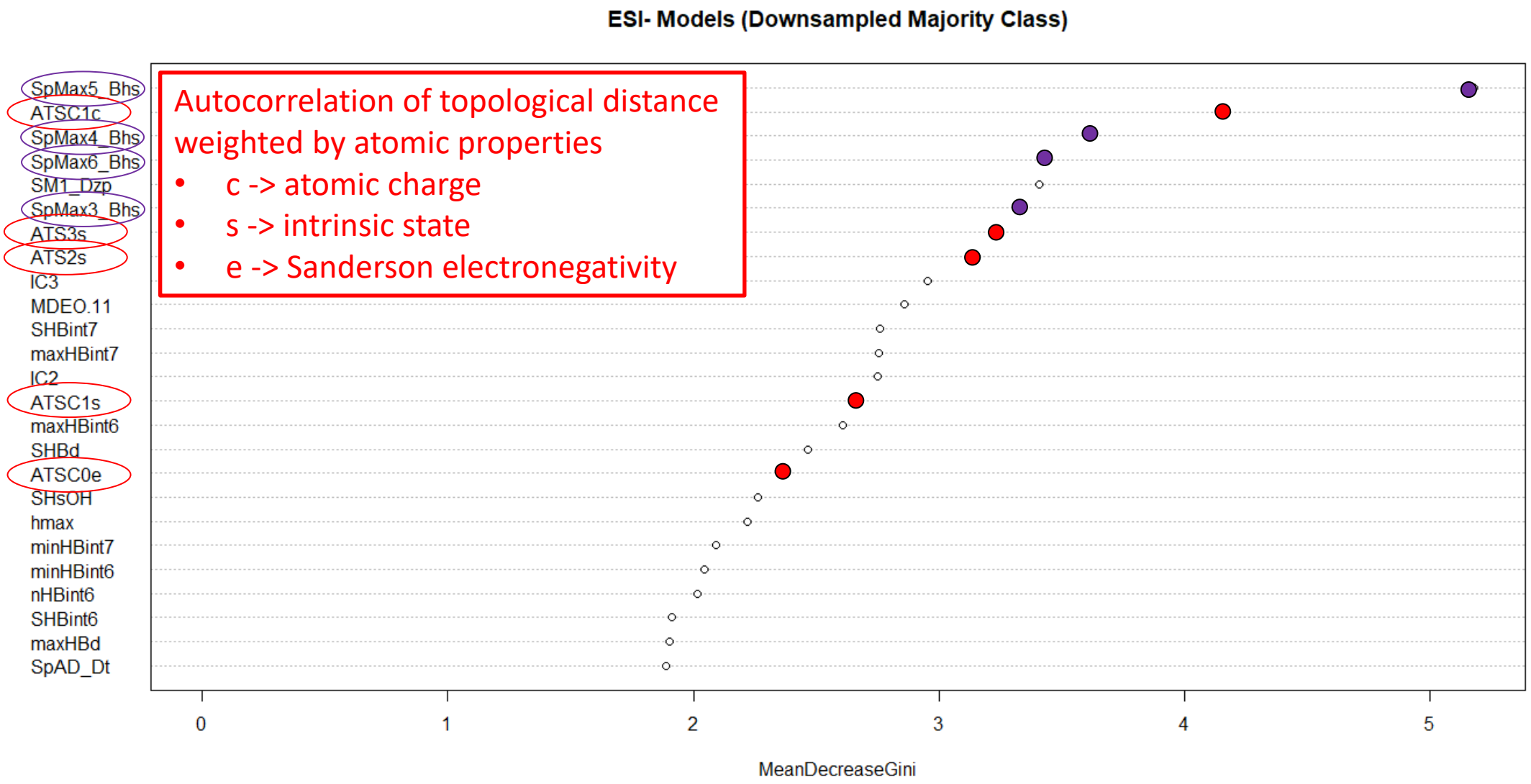
Mechanistic interpretation



- Captures electronic and topological information using intrinsic state of molecule
- Predictive of retention index

Office of Research and Development
Center for Computational Toxicology and Exposure

Mechanistic interpretation

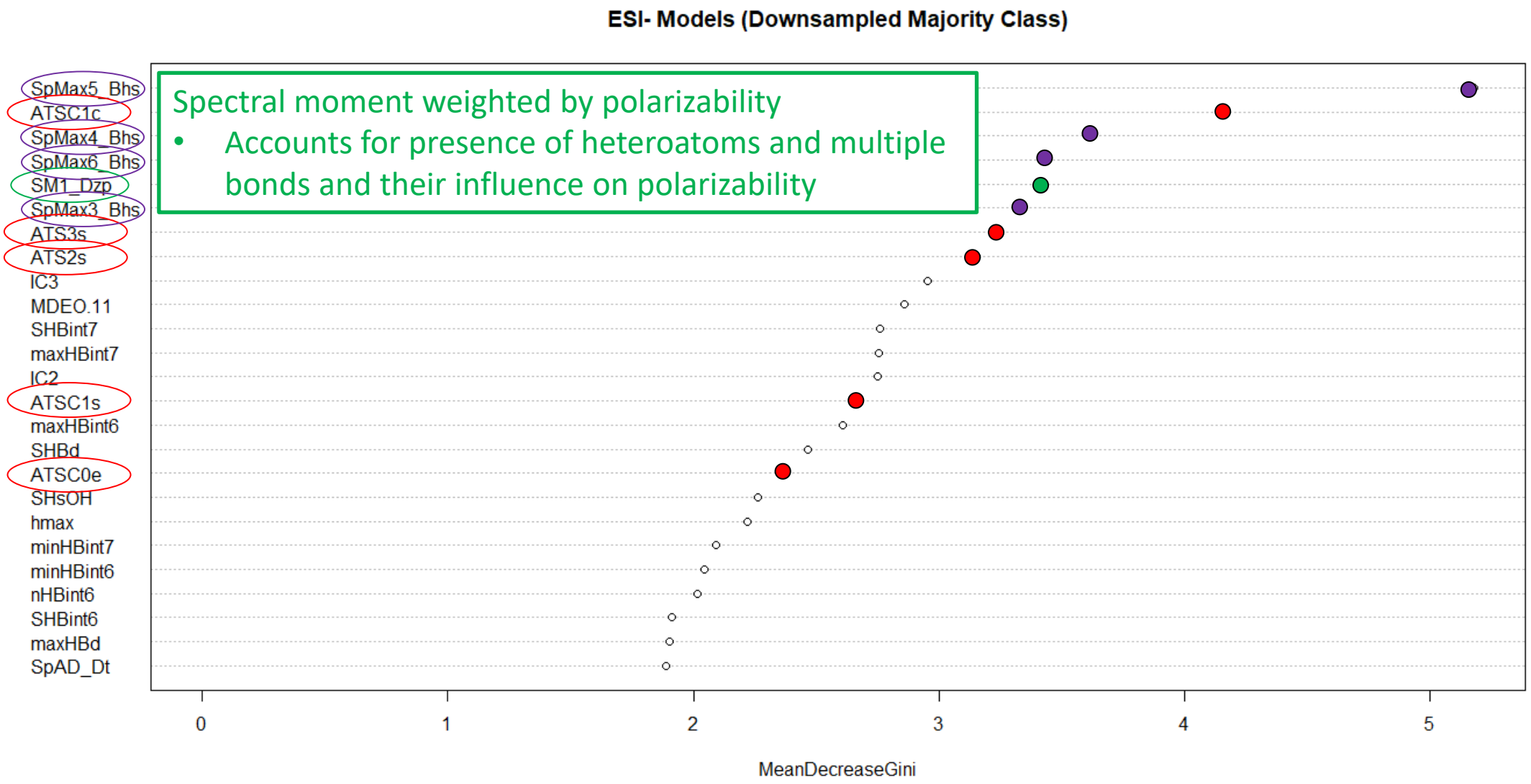


Autocorrelation of topological distance weighted by atomic properties

- c -> atomic charge
- s -> intrinsic state
- e -> Sanderson electronegativity

Office of Research and Development
Center for Computational Toxicology and Exposure

Mechanistic interpretation



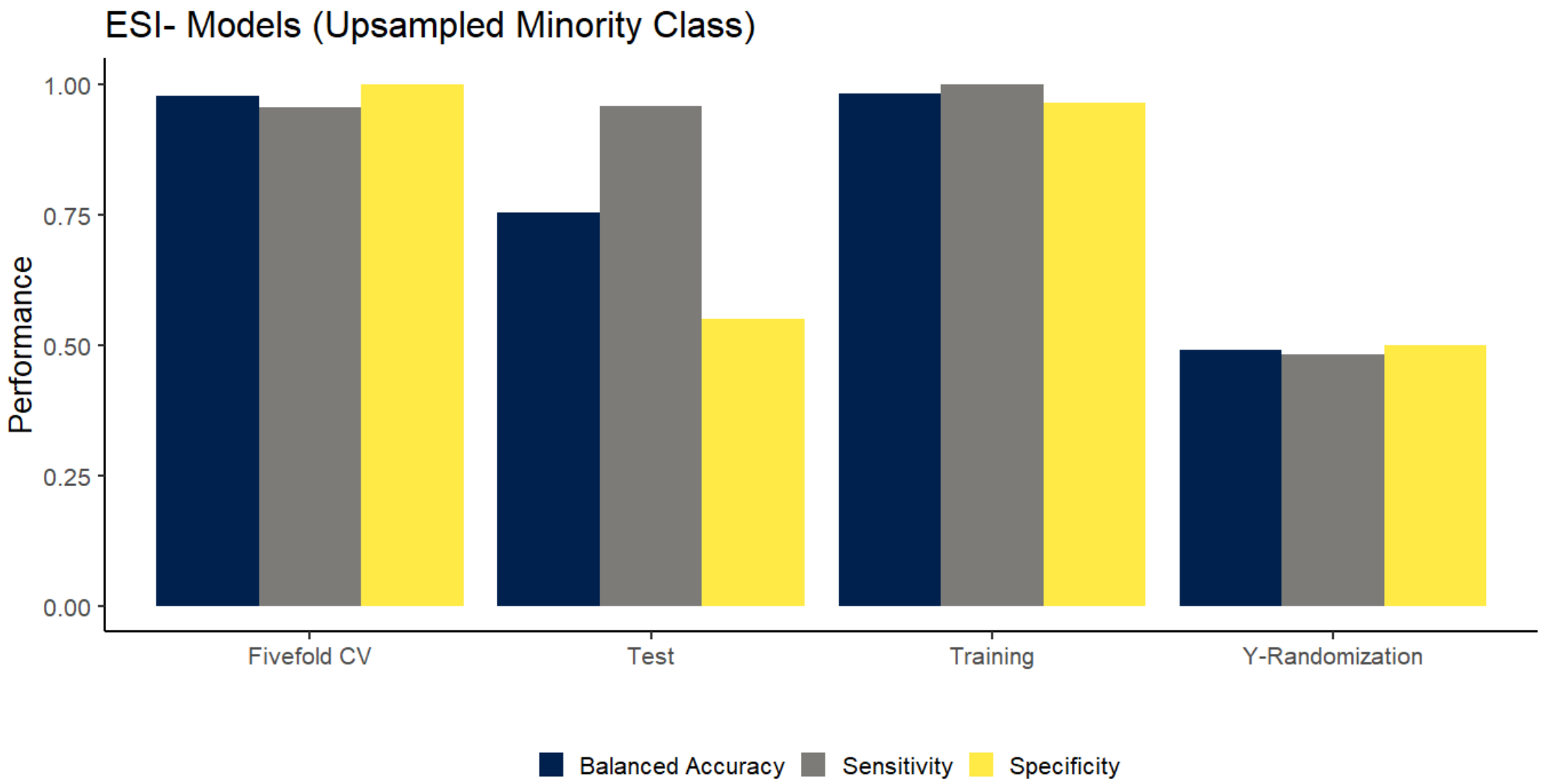
Spectral moment weighted by polarizability

- Accounts for presence of heteroatoms and multiple bonds and their influence on polarizability

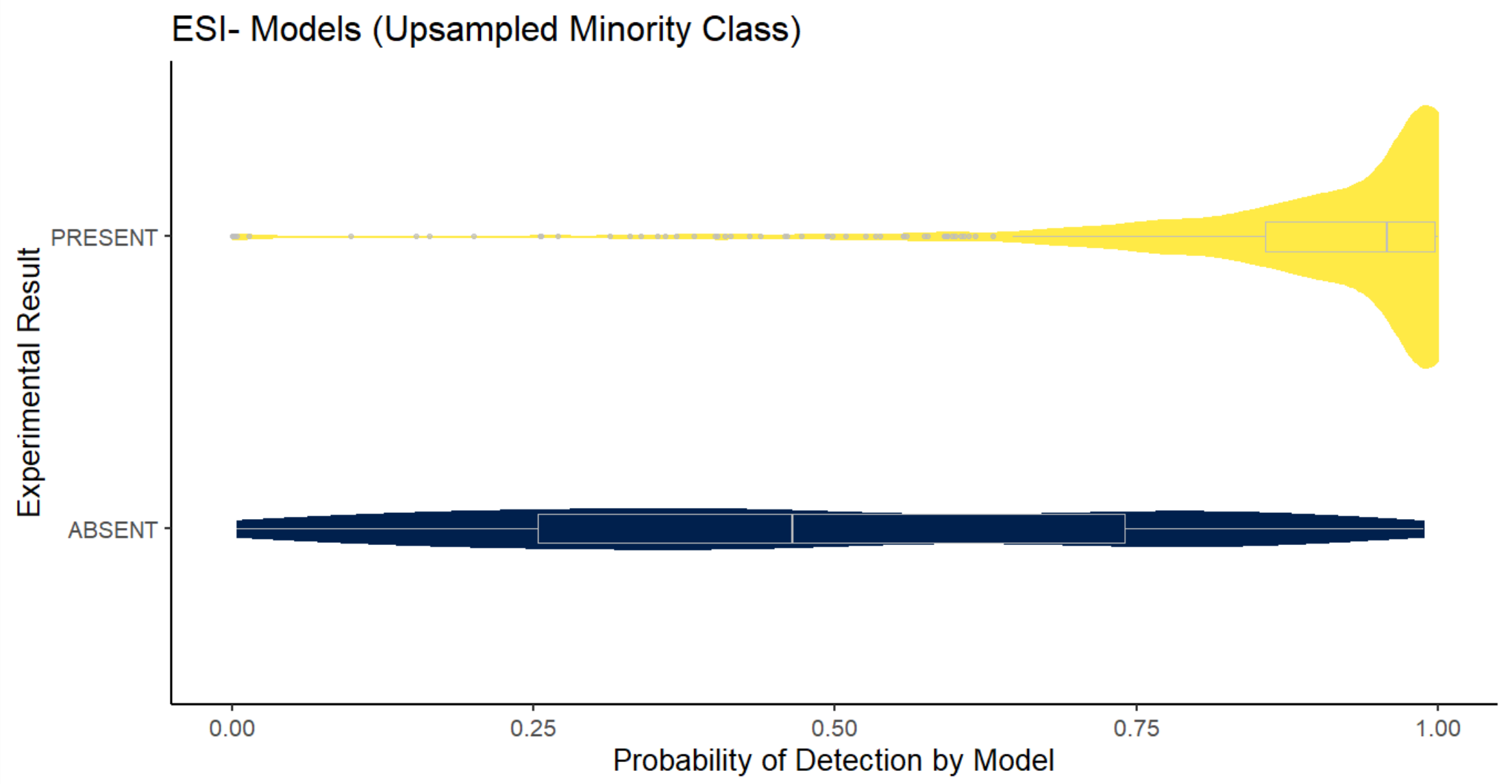
Model summary

- Downsampled model provides excellent predictions for both amenable and unamenable compounds
  - Caveat: reduces sample space of amenable compounds
    - May not accurately predict every amenable compound
- Preferred model for ranking candidates in a suspect-screening analysis

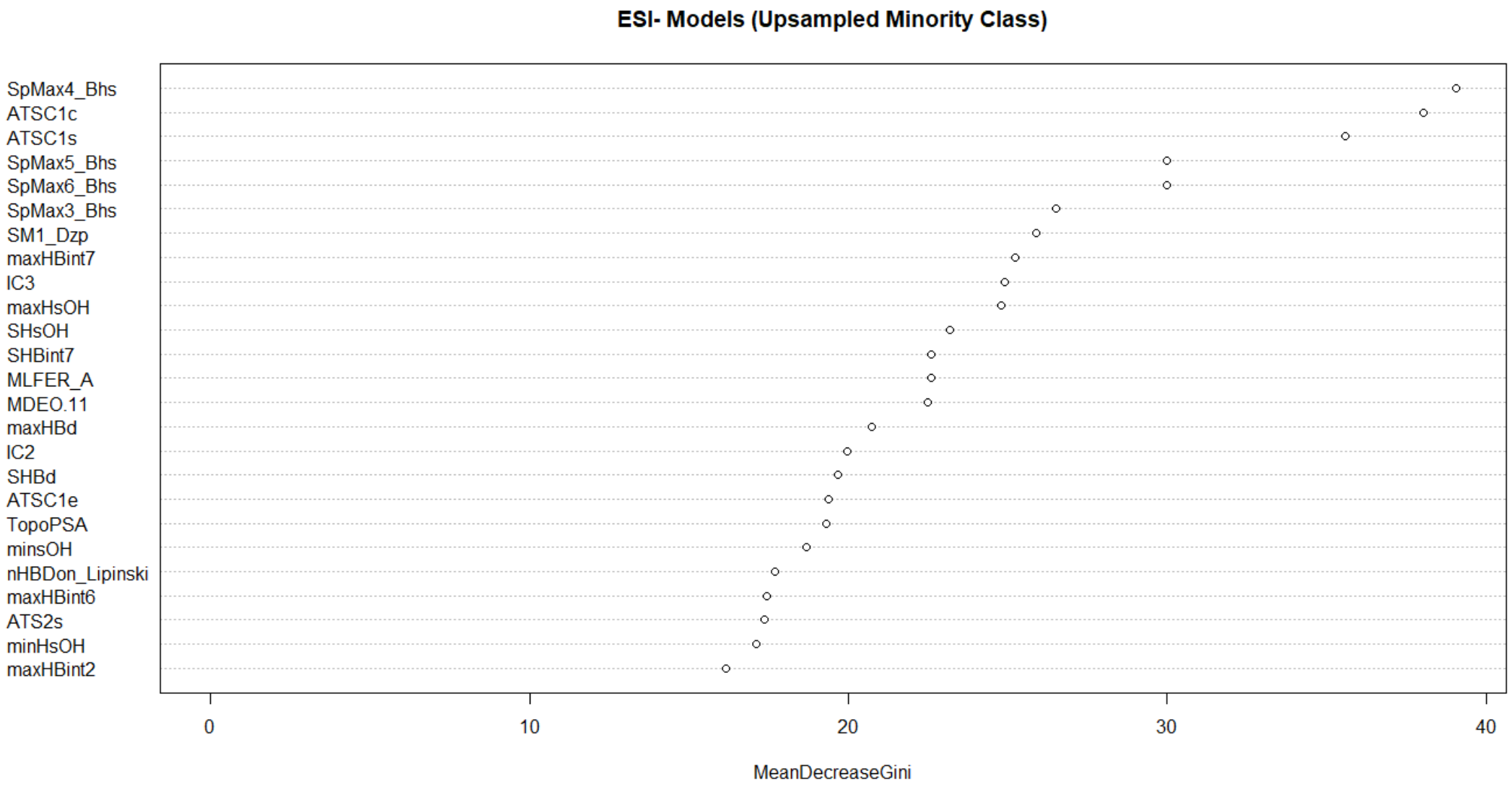
Model performance



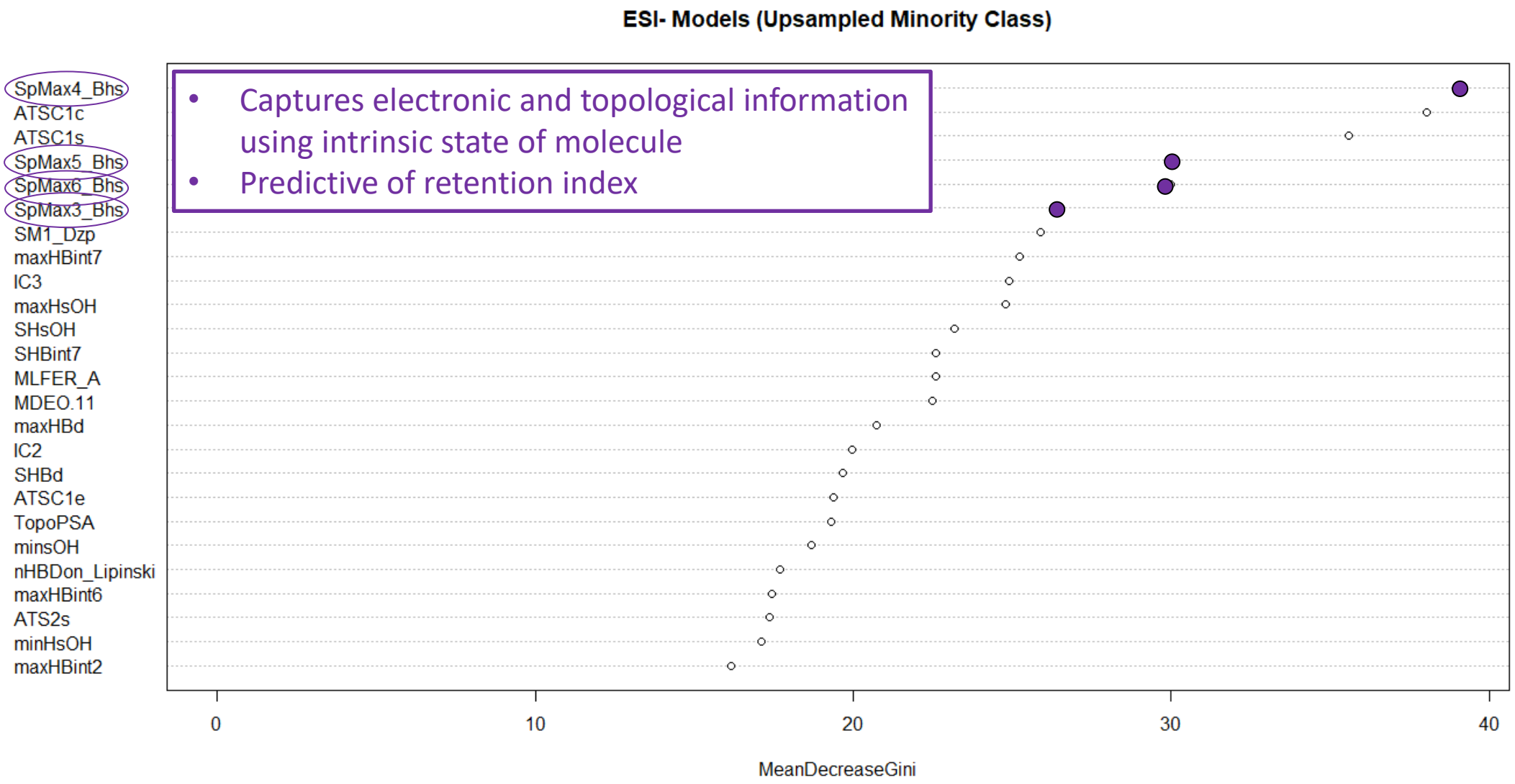
Model performance



Model performance



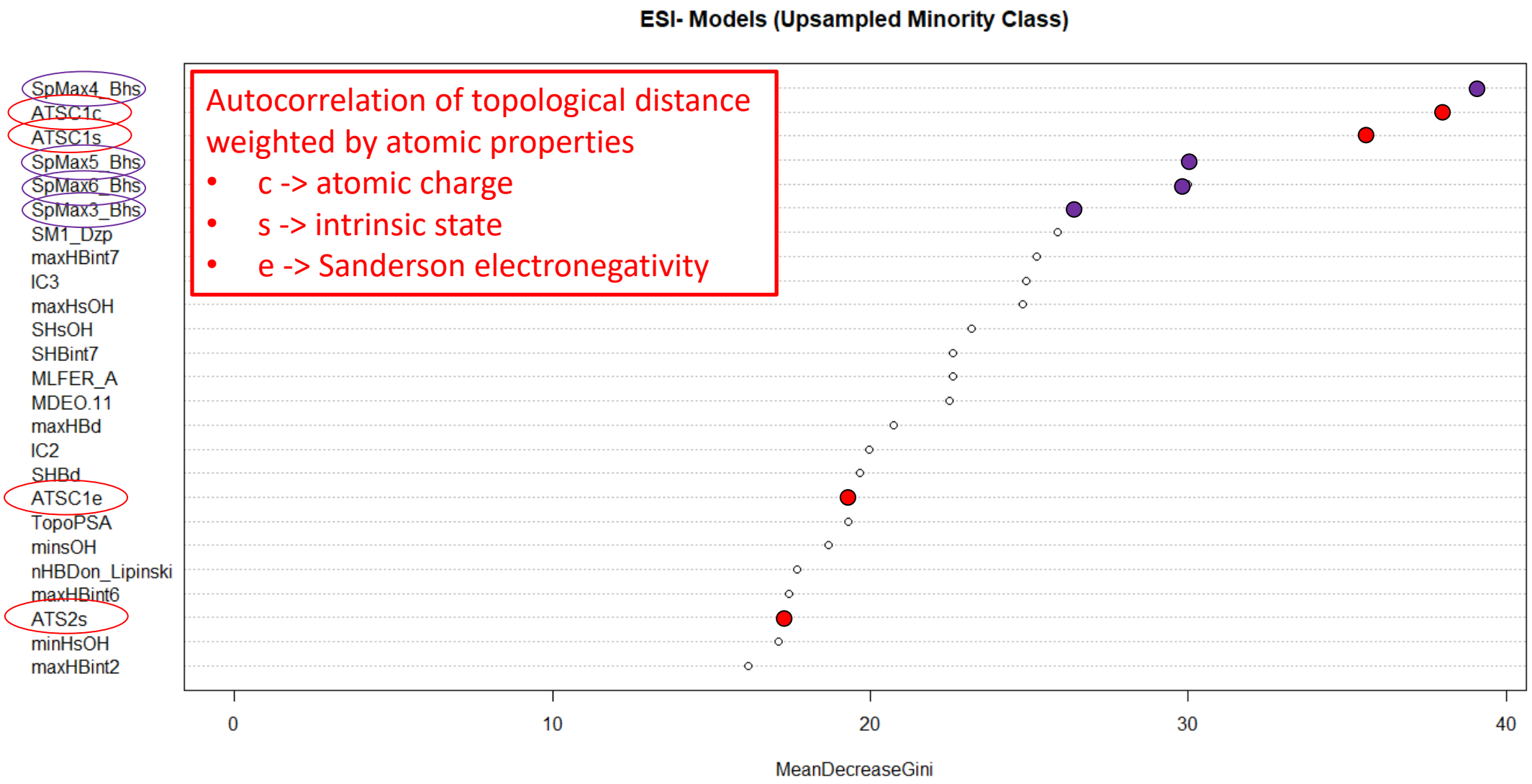
Mechanistic interpretation



- Captures electronic and topological information using intrinsic state of molecule
- Predictive of retention index

Office of Research and Development
Center for Computational Toxicology and Exposure

Mechanistic interpretation

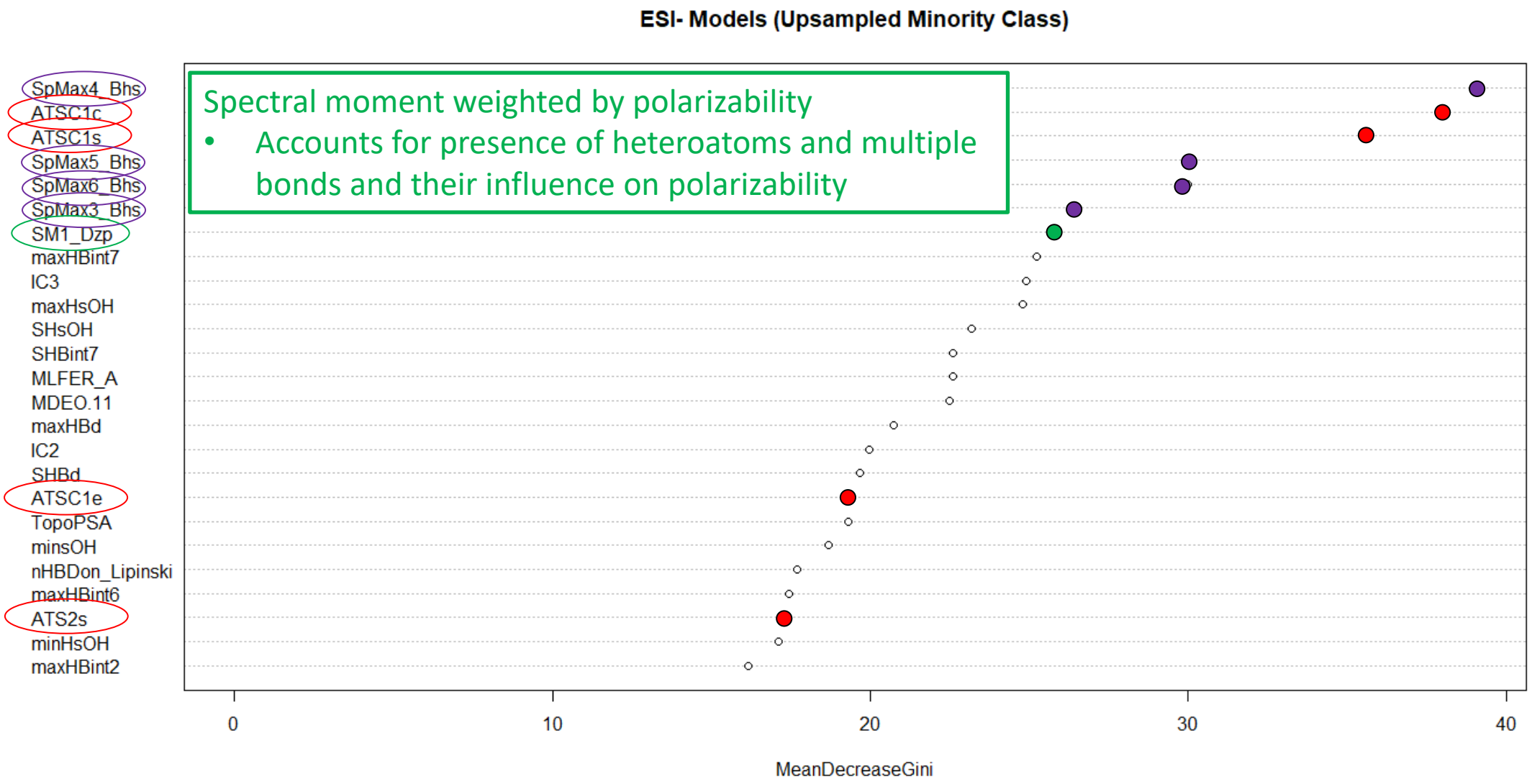


Autocorrelation of topological distance weighted by atomic properties

- c -> atomic charge
- s -> intrinsic state
- e -> Sanderson electronegativity

Office of Research and Development
Center for Computational Toxicology and Exposure

Mechanistic interpretation



Spectral moment weighted by polarizability

- Accounts for presence of heteroatoms and multiple bonds and their influence on polarizability

Office of Research and Development
Center for Computational Toxicology and Exposure

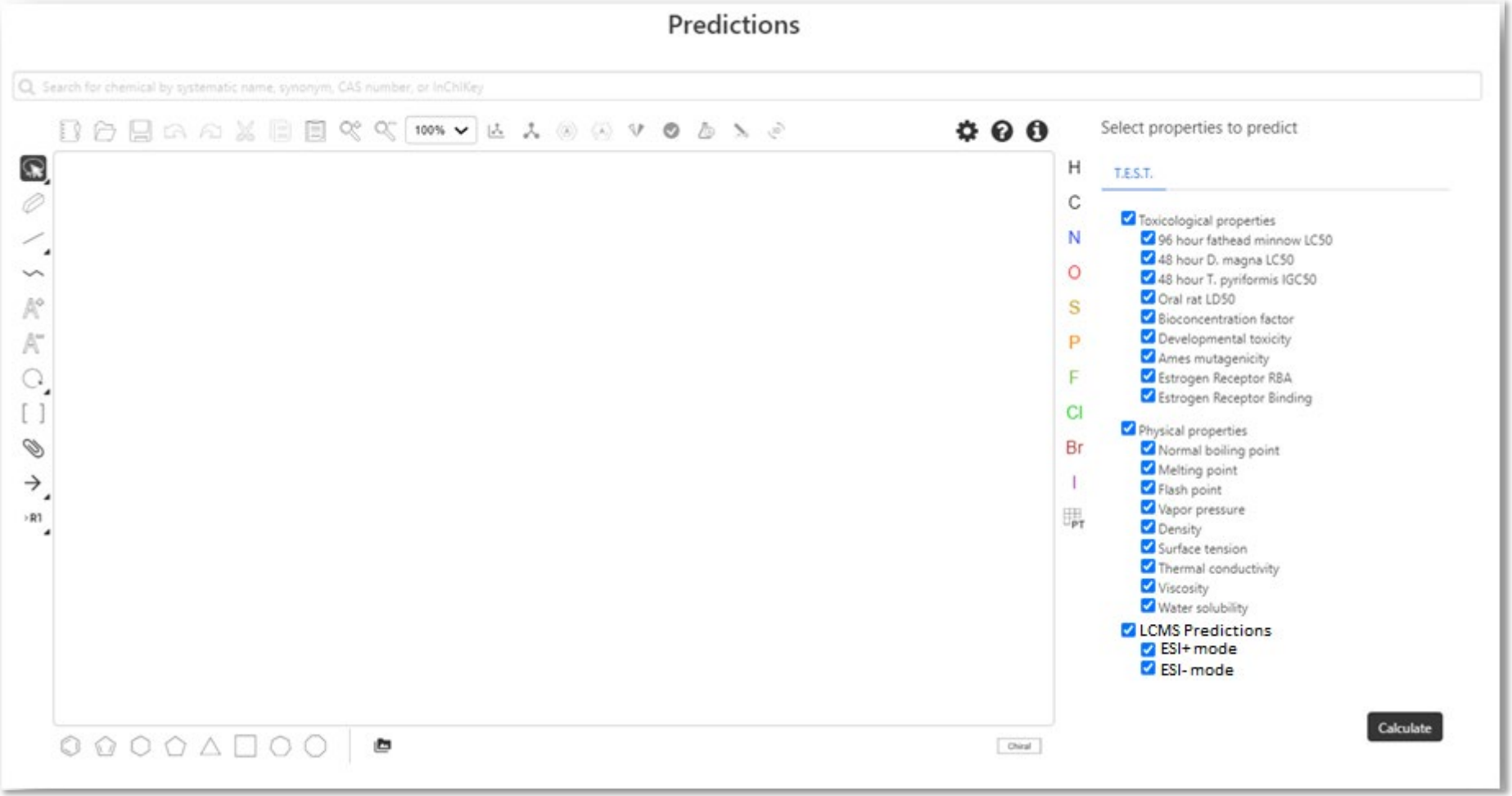
Model summary

- Upsampled model provides excellent predictions for amenable compounds
  - Much larger sample space than downsampled model
  - Weak predictive power for unamenable compounds
  - Too optimistic for suspect-screening
- Preferred model for establishing which chemicals *may* be amenable to method
  - establishing a list of chemical standards

Current & future work

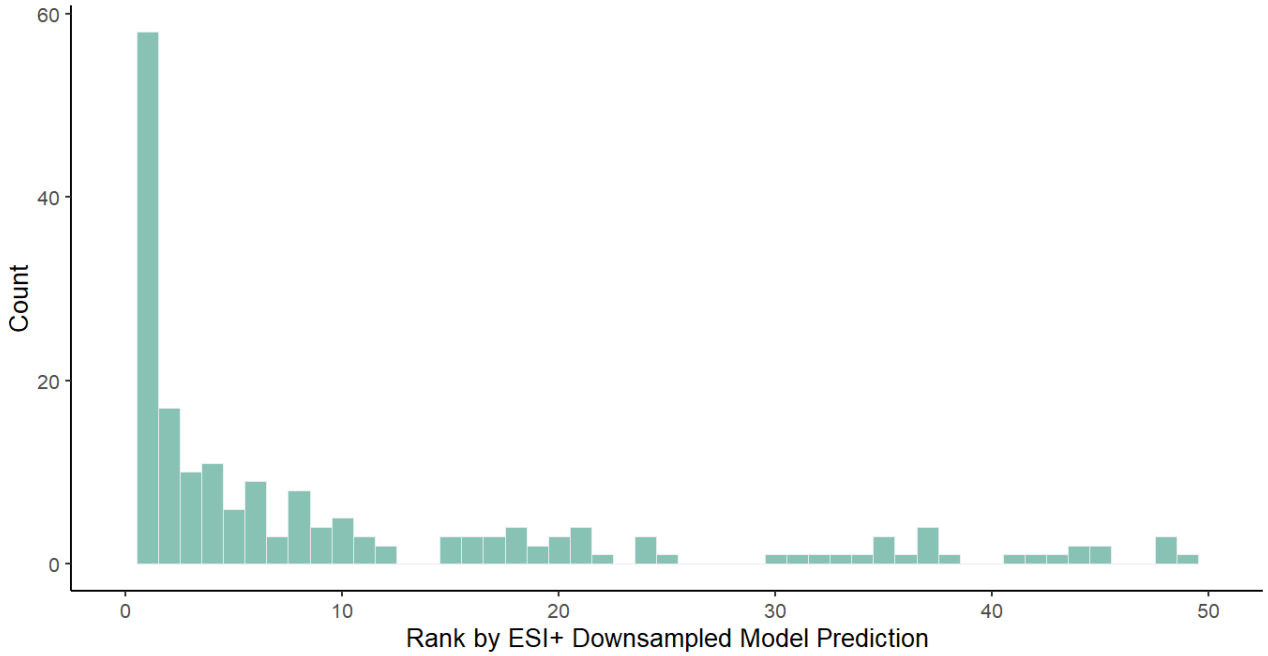
- Currently wrapping up manuscript
- Comparison of model results to Analytical QC data for ToxCast library
  - Good examples – no signal in LCMS ESI+, ESI- or in GCMS BUT present and high purity by NMR
- Compare model results to ENTACT results
    - Model predictions vs. independent labs, consensus of labs
- Future plans
  - Ensemble of upsampled and downsampled models?
  - Predictions for entirety DSSTox
  - Application for on-the-fly predictions based on a drawn structure

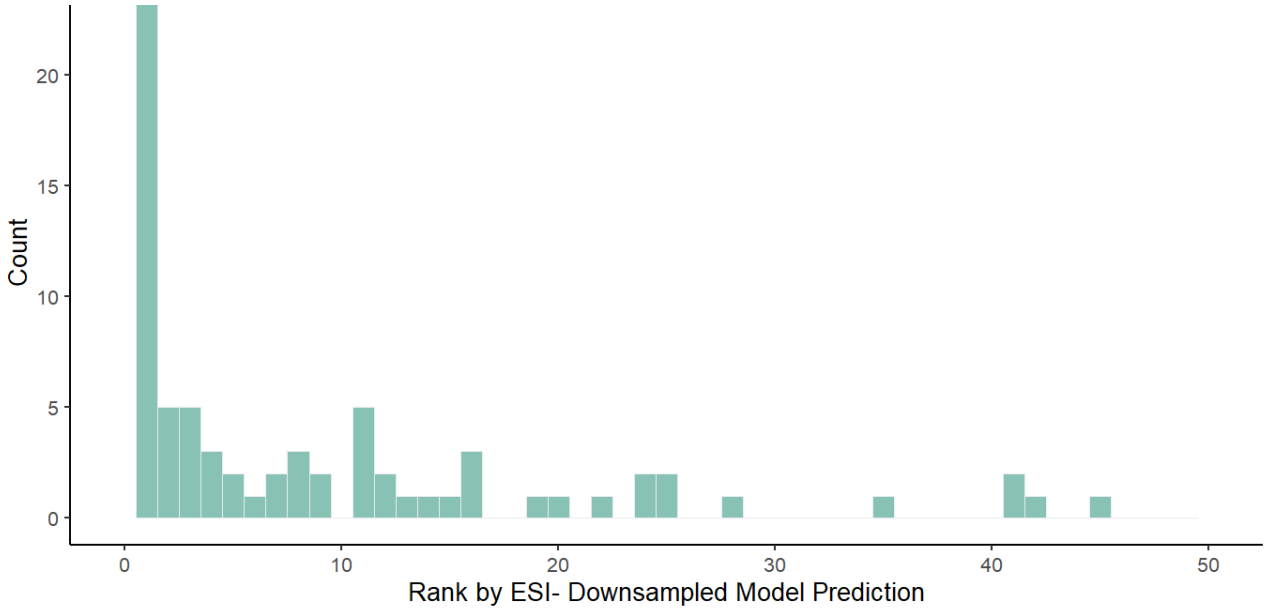
CompTox Chemicals Dashboard mockup - Predictions



Suspect-screening application

- List of ENTACT compounds identified in ESI+ & ESI- LCMS
  - 214 in ESI+
  - 105 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





Contributing researchers

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!