SAND2016-9315PE

BIOCOMPOUNDML A GENERAL SCREENING TOOL FOR BIOLOGICAL COMPOUND PROPERTY PREDICTION USING MACHINE LEARNING

AUTHORS: LEANNE S. WHITMORE & COREY M. HUDSON

AFFILIATIONS: SANDIA NATIONAL LABORATORY (BIOSCIENCE) & JOINT BIOENERGY INSTITUTE (DECONSTRUCTION DIVISION)

EVALUATING CHEMICAL PROPERTIES

Technique	Strengths	Weaknesses
Experimental Validation	High Accuracy and Precision Gold Standard for property prediction	Expensive Time Consuming May require impossible to attain chemical quantities
Direct Computational Simulation (e.g., Quantum Mechanical Methods, Density Functional Theory)	Huge quantities of chemical information High Precision Trusted estimation quality	Huge computational requirements Long periods of simulation Potentially catastrophic accuracy failure
Machine Learning and Big Data Analytics	Measurable Accuracy and Precision Very Fast Leverages Chemical Property Correlations	Unclear underlying model (Black Box) Approximate measurement (unspecific) Extrapolation failure

SOFTWARE PURPOSE

- **BioCompoundML was developed to rapidly screen a very large number of biologically-producable compounds for chemical properties that are important in research and industrial settings**
- Any chemical to be seriously considered in manufacture will require experimental measurement
- But, if time-consuming and expensive measurements and estimations are the first stage, a large chunk of the chemical universe will not be considered
- Frequently, we've found that synthetic chemists and biologists have rather direct and binary criteria for evaluating chemical performance **(at least in the early stages)**

AN EXAMPLE: RESEARCH OCTANE NUMBER (RON)

- RON is a fundamental fuel property
- Measures the resistance of a spark ignition (SI) fuel to autoignition under compression
- Co-OPTIMA (a multilab DOE-funded project) is interested in evaluating a large number of potential Low-Greenhouse Gas produced chemicals for offsetting petroleum fuel in blendstocks (petroleum + chemical mix)
- One of the key components for any SI-added chemical is high RON

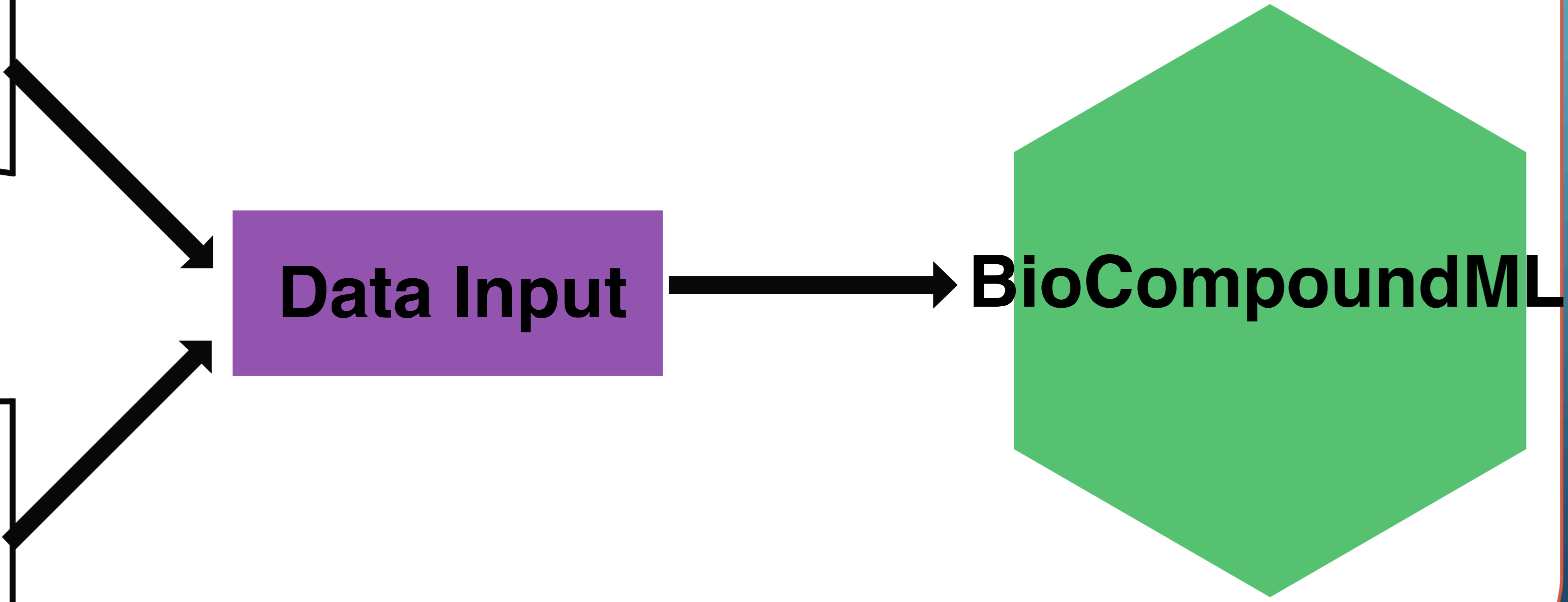
Input Requirements

Training Data

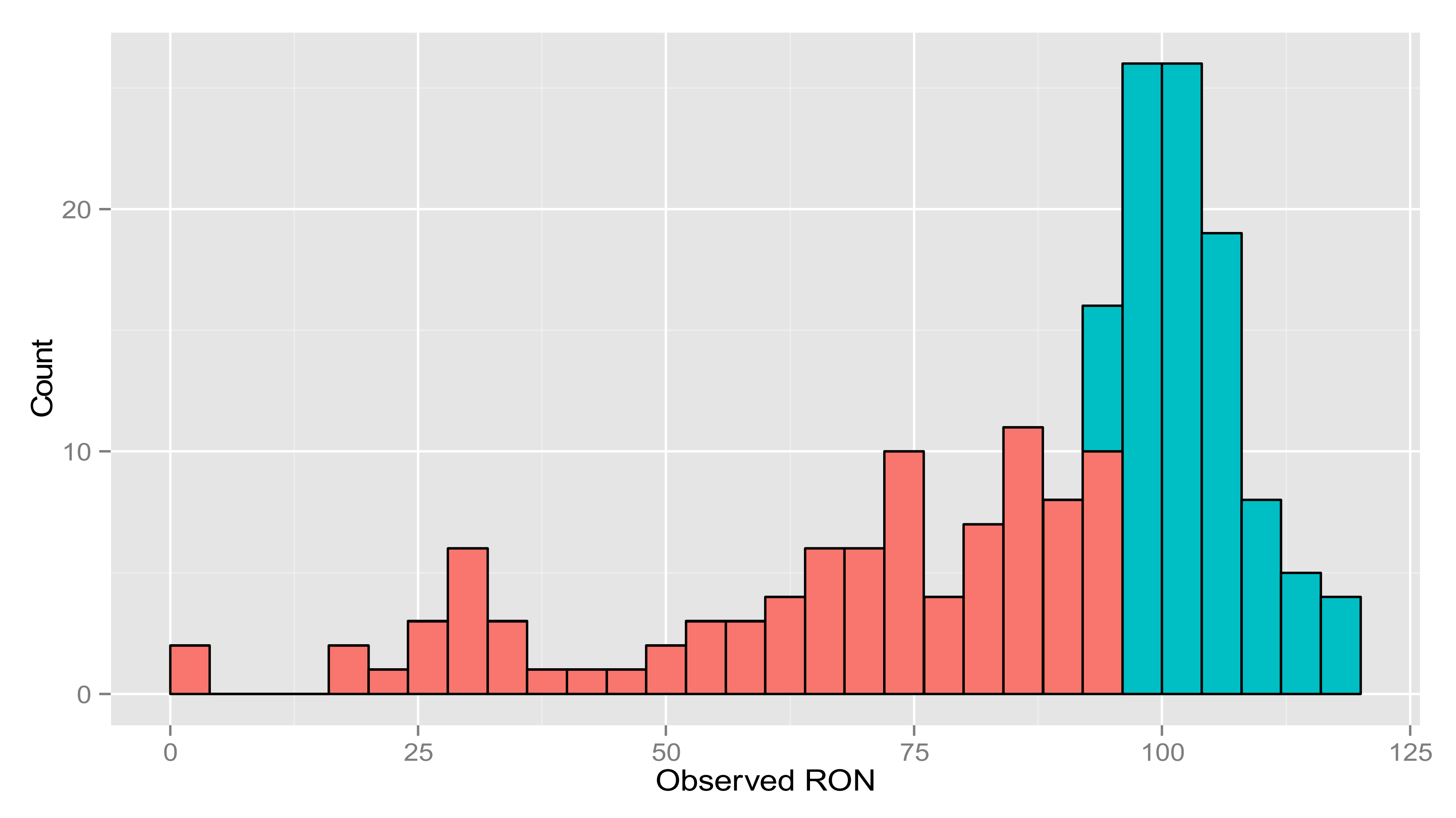
#Name	RON	PubChem
1-Butene	98.8	7844
1-Heptene	54.5	11610
1-Hexene	76.4	11597

Test Data

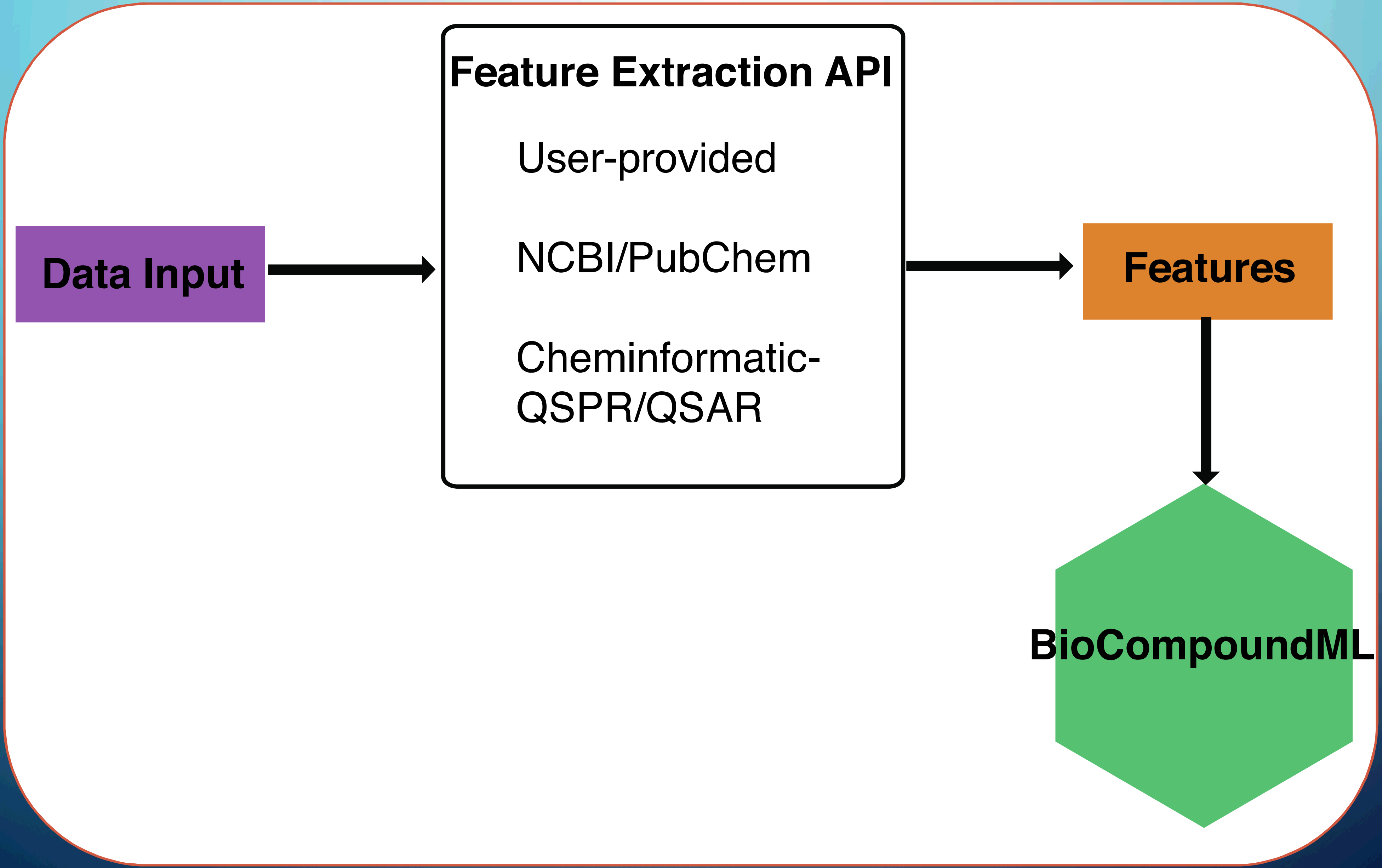
#Name	PubChem
isoamyl acetate	31276
myrcene	31253
eucalyptol	2758
3-carene	26049



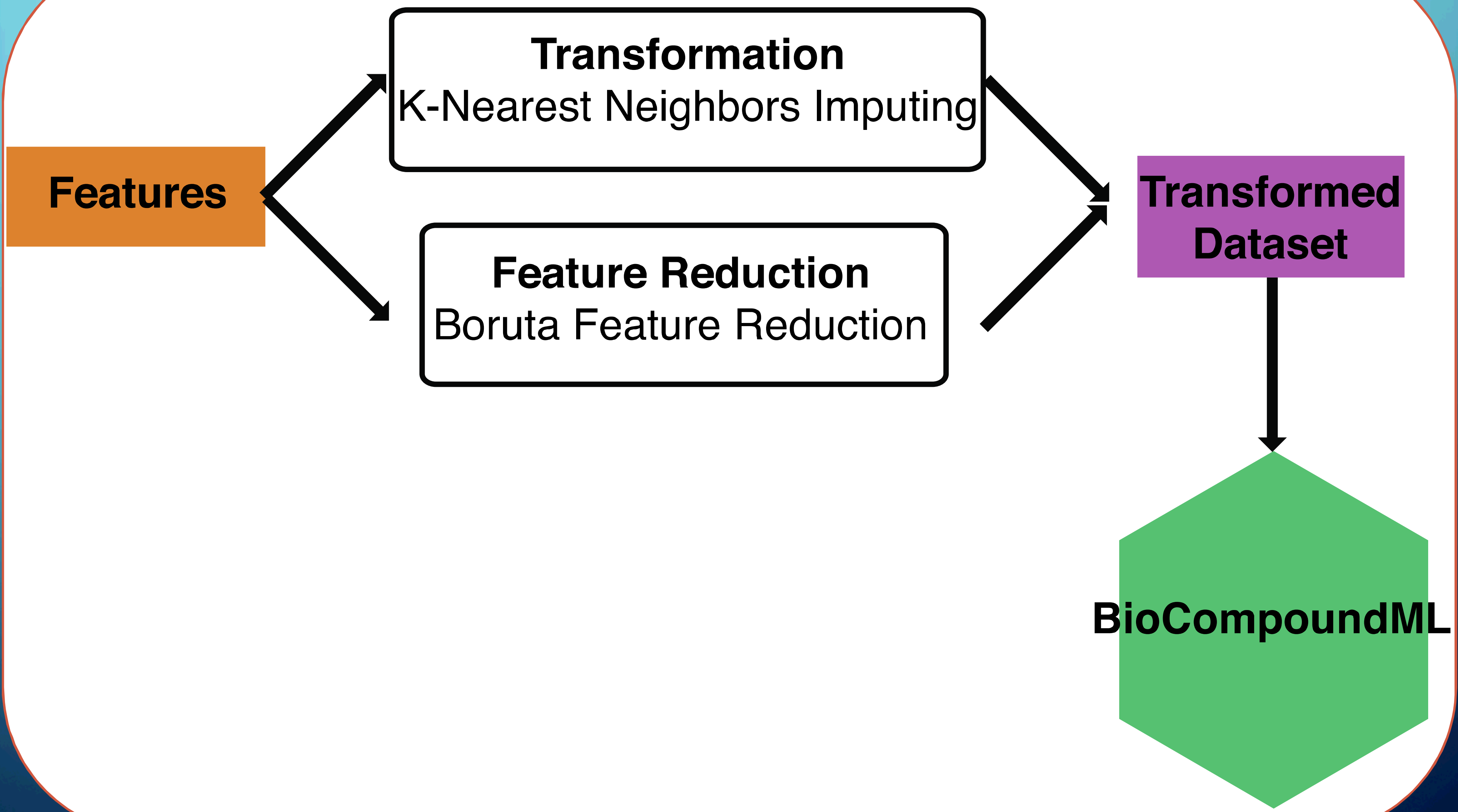
RON TRAINING DATA



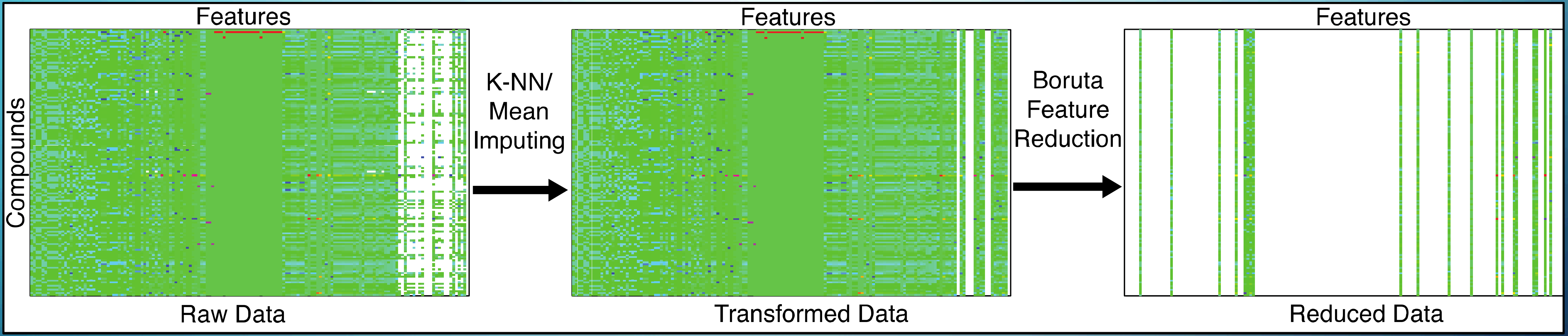
Feature Collection



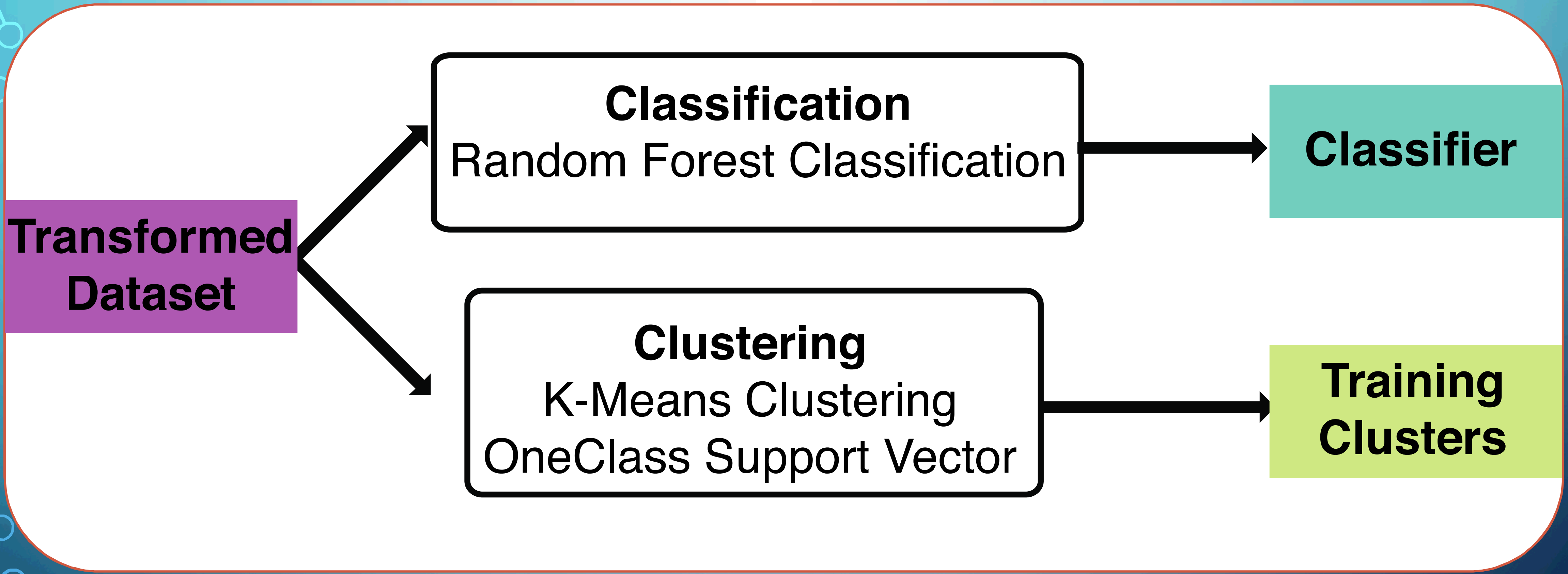
Data Reduction and Correction



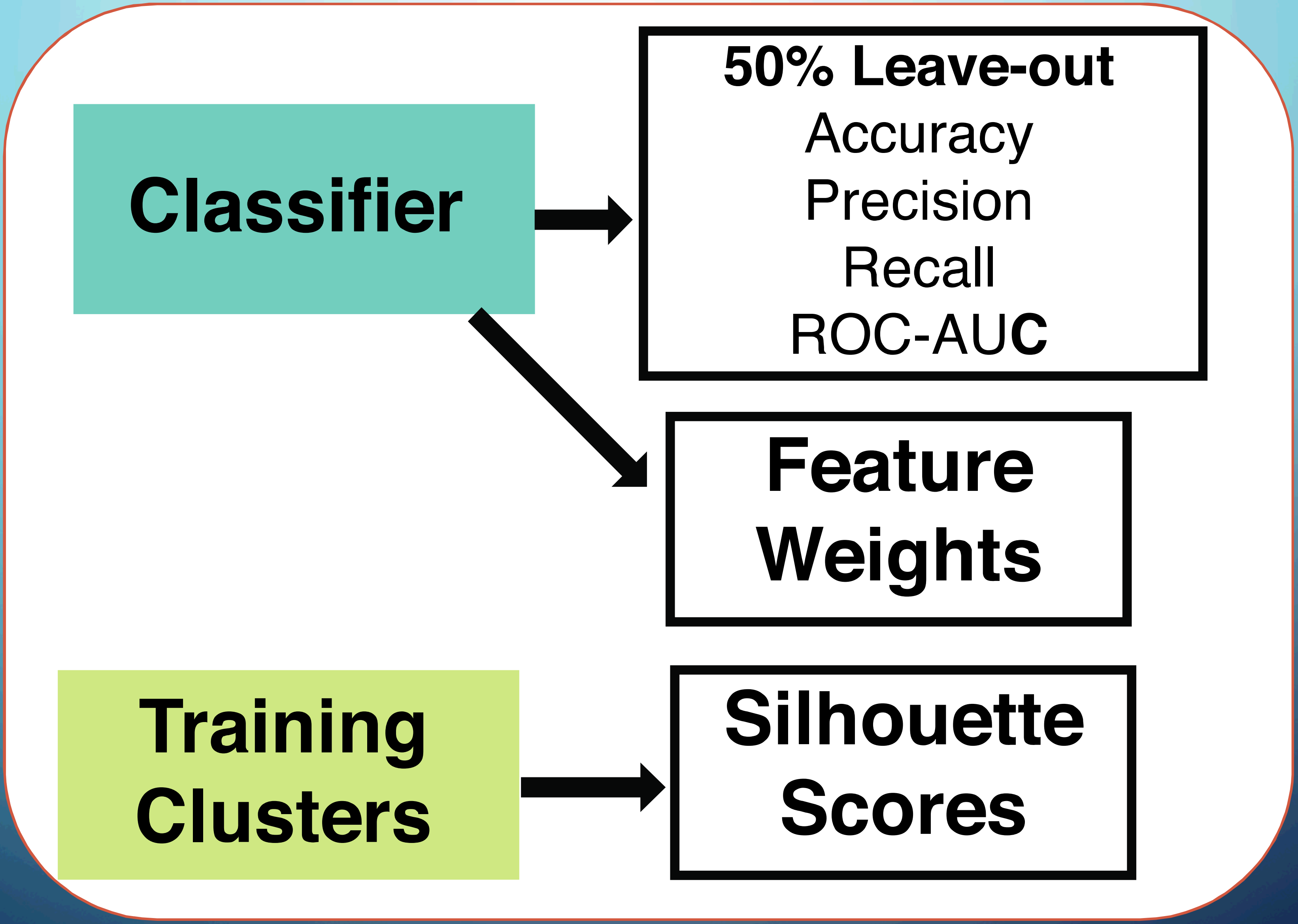
SCALE OF DATA REDUCTION – INITIAL VS. FINAL FEATURE SET



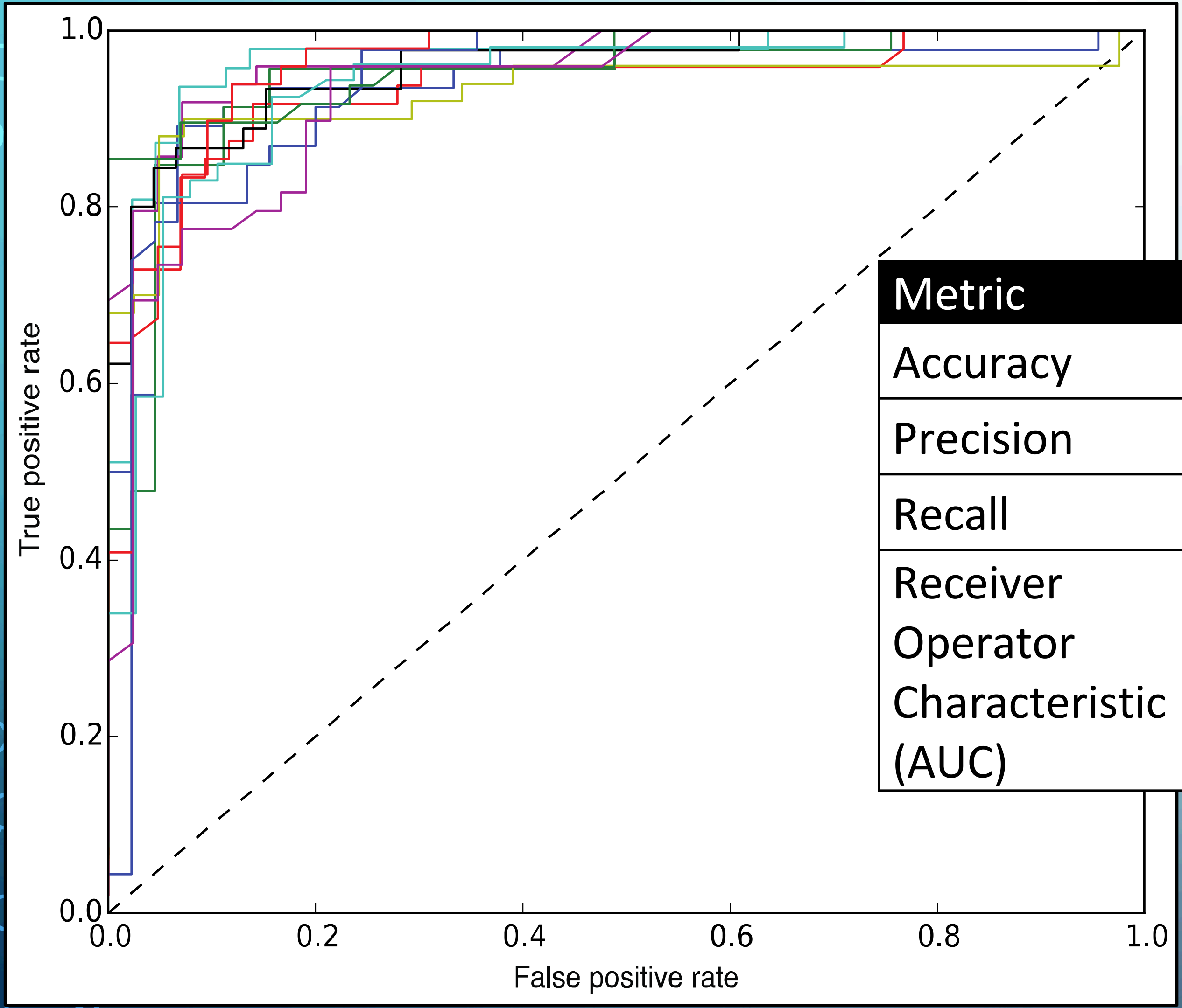
Classification and Clustering



Assessment


Classifier
50% Leave-out
Accuracy
Precision
Recall
ROC-AUC
Feature Weights
Training Clusters
Silhouette Scores

RON MODEL ASSESSMENT



Metric	Mean value	Std. dev
Accuracy	0.88	0.06
Precision	0.88	0.11
Recall	0.88	0.11
Receiver Operator Characteristic (AUC)	0.88	0.06

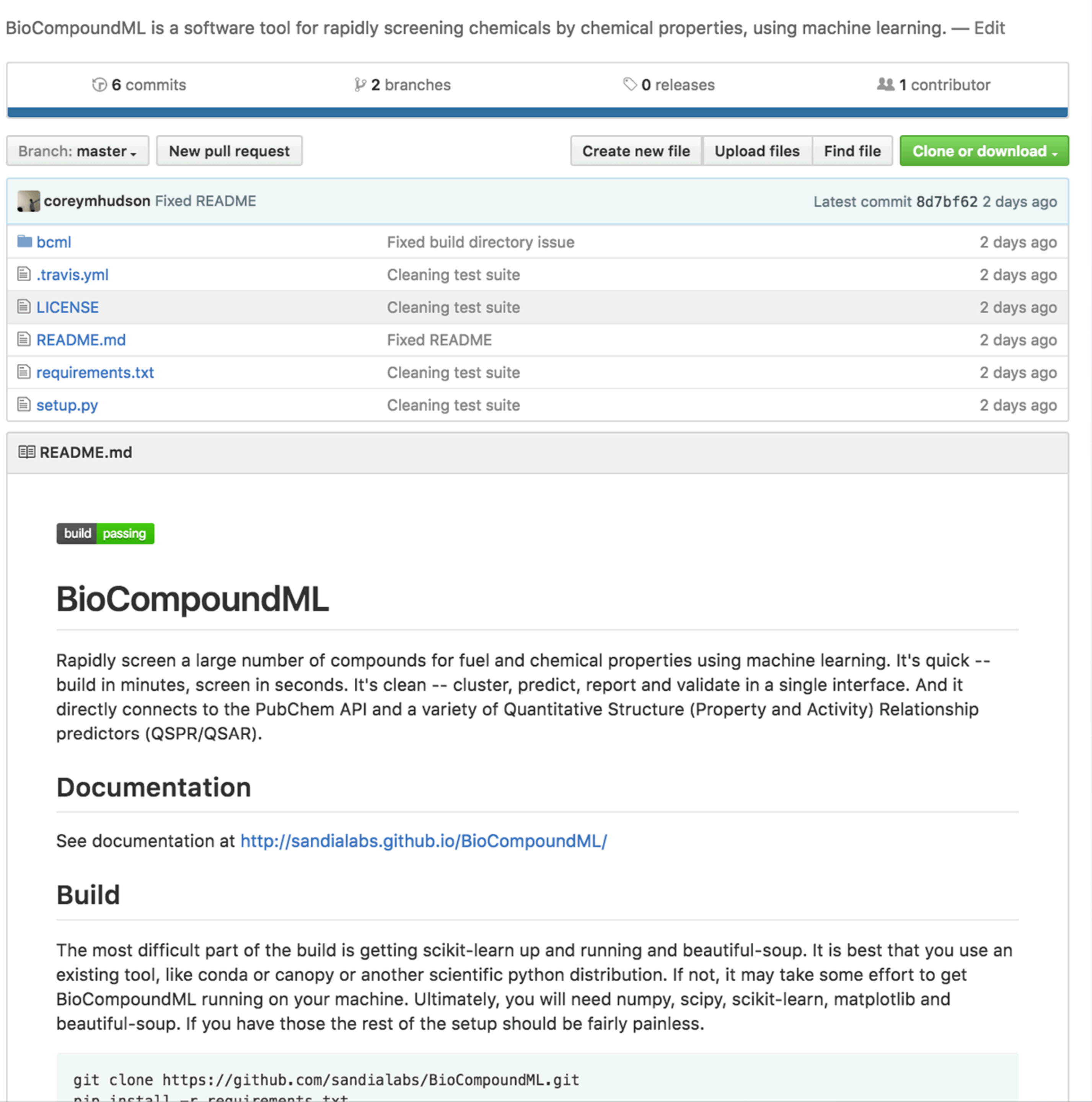
TESTING RON MODEL – EXPERIMENTAL VALIDATION

CAS-No	Compound	Measured RON	Prediction	Probability in High RON Class	Accurate
106-21-8	3,7-Dimethyl-1-Octanol	64.9	Not High RON	0.707	Yes
13466-78-9	3-Carene	68.9	Not High RON	0.754	Yes
13877-91-3	Ocimene	72.9	Not High RON	0.463	Yes
78-69-3	3,7-Dimethyl-3-octanol	76.3	Not High RON	0.76	Yes
123-35-3	Myrcene	82.5	Not High RON	0.799	Yes
80-56-8	a-Pinene	83.3	Not High RON	0.63	Yes
5989-27-5	(R)-(+)-Limonene	87.6	Not High RON	0.695	Yes
78-70-6	Linalool	96.7	Unclear	0.869	Marginal
470-82-6	Eucalyptol	99.2	High RON	0.916	Yes
142-62-1	Butyl Acetate	100.7	High RON	0.99	Yes
123-92-2	Isoamyl Acetate	101	High RON	0.967	Yes
93-58-3	Methyl-Benzoate	101.1	High RON	0.998	Yes
115-18-4	2-methyl-3-buten-2-ol	103.5	High RON	0.967	Yes
110-19-0	Isobutyl Acetate	108.7	High RON	0.977	Yes
67-64-1	Acetone	111.3	High RON	0.908	Yes
209-117-3	Isopropyl Acetate	>120	High RON	0.971	Yes

TOP 20 METACYC COMPOUNDS

Compound	Probability High RON	CAS	PubChem	Formula
butyl acetate	0.99	123-86-4	31272	$C6H12O2$
1,4-benzoquinone	0.98	106-51-4	4650	$C6H4O2$
fumarate	0.98	110-17-8	5460307	$C4H2O4$
ethanol	0.97	64-17-5	702	$C2H6O$
diacetyl	0.97	431-03-8	650	$C4H6O2$
1-O-methylsalicylate	0.97	119-36-8	4133	$C8H8O3$
2-methylbutanol	0.97	137-32-6	8723	$C5H12O$
anisole	0.97	100-66-3	7519	$C7H8O$
ethyl acetate	0.97	141-78-6	8857	$C4H8O2$
2-methylbut-3-en-2-ol	0.97	115-18-4	8257	$C5H10O$
methylglyoxal	0.97	78-98-8	880	$C3H4O2$
patulin	0.96	149-29-1	4696	$C7H6O4$
3-methylbutanol	0.96	123-51-3	31260	$C5H12O$
cyclopentanone	0.96	120-92-3	8452	$C5H8O$
acetoin	0.96	513-86-0	179	$C4H8O2$
1,3-propanediol	0.96	504-63-2	10442	$C3H8O2$
(-)-camphor	0.96	464-48-2	444294	$C10H16O$
(R)-mevalonate	0.96	150-97-0	5288798	$C6H11O4$
2-butyne-1,4-diol	0.96	110-65-6	8066	$C4H6O2$
2-methylphenol	0.96	95-48-7	335	$C7H8O$

DISTRIBUTED ON GITHUB WITH BSD-OPEN SOURCE LICENSE



POTENTIAL USERS

- Larger synthetic chemistry and biology communities
- Academic researchers seeking chemical classification
- Industry seeking new target compounds
- JBEI and National Lab researchers looking to classify and rank chemicals for biological production across a gradient of important features

LIMITATIONS IN REACHING USERS

- Difficulty of use
  - Command line driven
  - Requires modern understanding of Python and its dependencies
  - Huge parameter set
  - Difficult install
- Collection and curation of training data
  - No central repository for measured training data
- Exposure
  - Reach the actual audience of interest
- Underlying machine learning paradigm not explained through visuals

HOW THE GRASSROOTS SOFTWARE TOOL COMPETITION COULD HELP

- Movement of the software from research to production quality
- Provide web-based presentation of visuals (both results and machine learning tutorials)
- Associate the tool with an existing suite tools for synthetic biologists
- Expose JBEI and National Laboratory researchers to tool

ACKNOWLEDGEMENTS

This research was conducted as part of the Co-Optimization of Fuels & Engines (Co-Optima) project sponsored by the U.S. Department of Energy (DOE) Office of Energy Efficiency and Renewable Energy (EERE), Bioenergy Technologies and Vehicle Technologies Offices. Co-Optima is a collaborative project of multiple National Laboratories initiated to simultaneously accelerate the introduction of affordable, scalable, and sustainable biofuels and high-efficiency, low-emission vehicle engines.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. This work was part of the DOE Joint BioEnergy Institute (http://www.jbei.org) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02- 05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy.

THANK YOU!