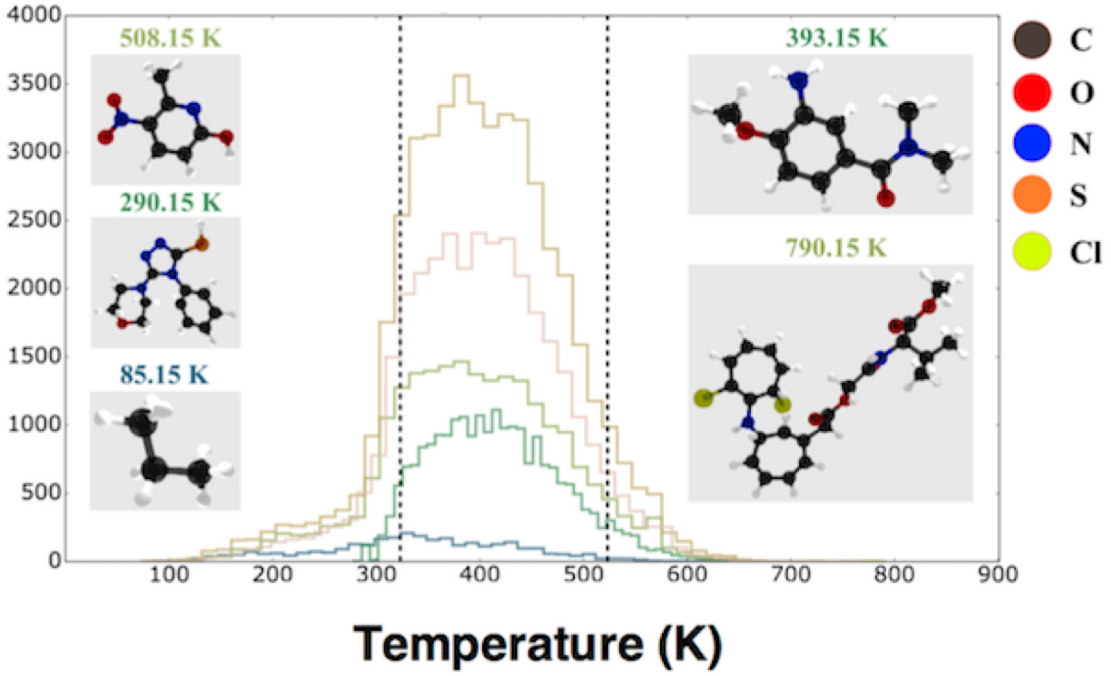
Argonne NATIONAL LABORATORY

A diversified machine learning strategy for predicting and understanding molecular melting points

GANESH SIVARAMAN
Postdoctoral Appointee
Argonne Leadership Computing Facility
EMAIL: gsivaraman@anl.gov

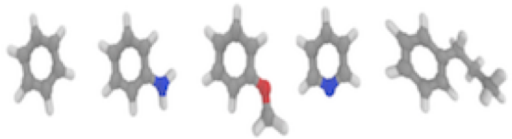
September 26 (2019)
Los Angeles, California

Experimental Melting Points



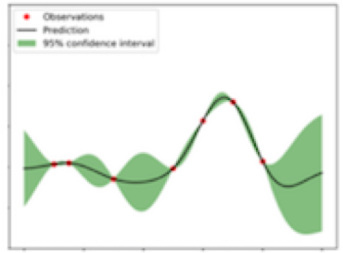
Machine Learning

Low-Bias Datasets



Chemical Insights

Predictive Models



Ganesh Sivaraman, Nicholas Jackson, Benjamin Sanchez-Lengeling, Álvaro Vázquez-Mayagoitia, Alán Aspuru-Guzik, Venkatram Vishwanath, and Juan de Pablo. “A diversified machine learning strategy for predicting and understanding molecular melting points” (2019) (In Review).

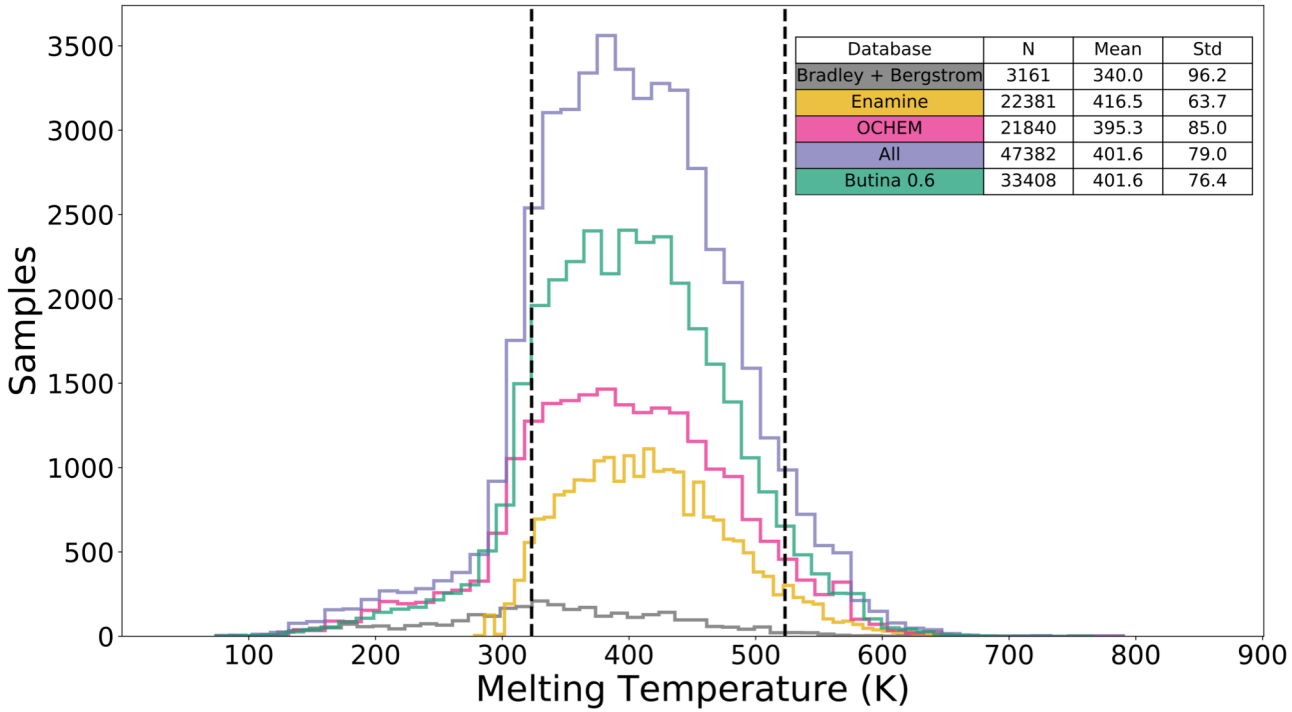
MOTIVATION

- The ability to predict multi-molecule processes, using only knowledge of single molecule structure, stands as a grand challenge for molecular modeling.

– **Molecular Melting points**

- Molecule's MP can be correlated with a number of industrially vital material properties. For example, solubilities of candidate drug-like molecules.

MELTING POINT DATASETS



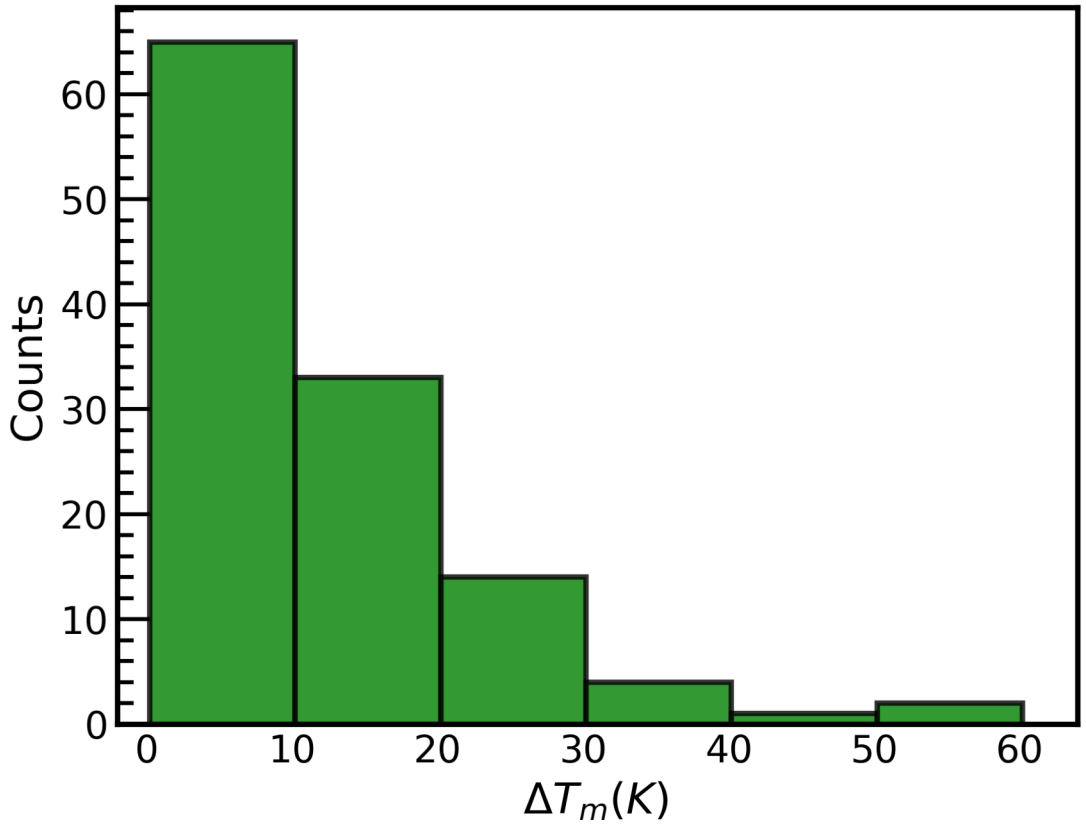
Database	N	Mean	Std
Bradley + Bergstrom	3161	340.0	96.2
Enamine	22381	416.5	63.7
OCHEM	21840	395.3	85.0
All	47382	401.6	79.0
Butina 0.6	33408	401.6	76.4

- An integrated dataset of 47k molecules with experimental meting points of organic molecules, and augmented with 3D molecular structures and quantum chemical properties.

Tetko, Igor V., *et al.* "How accurately can we predict the melting points of drug-like compounds?." *Journal of chemical information and modeling* 54.12 (2014)

DISTRIBUTION OF MP INTERVALS

119 literature molecules exhibiting multiple crystal polymorphs.

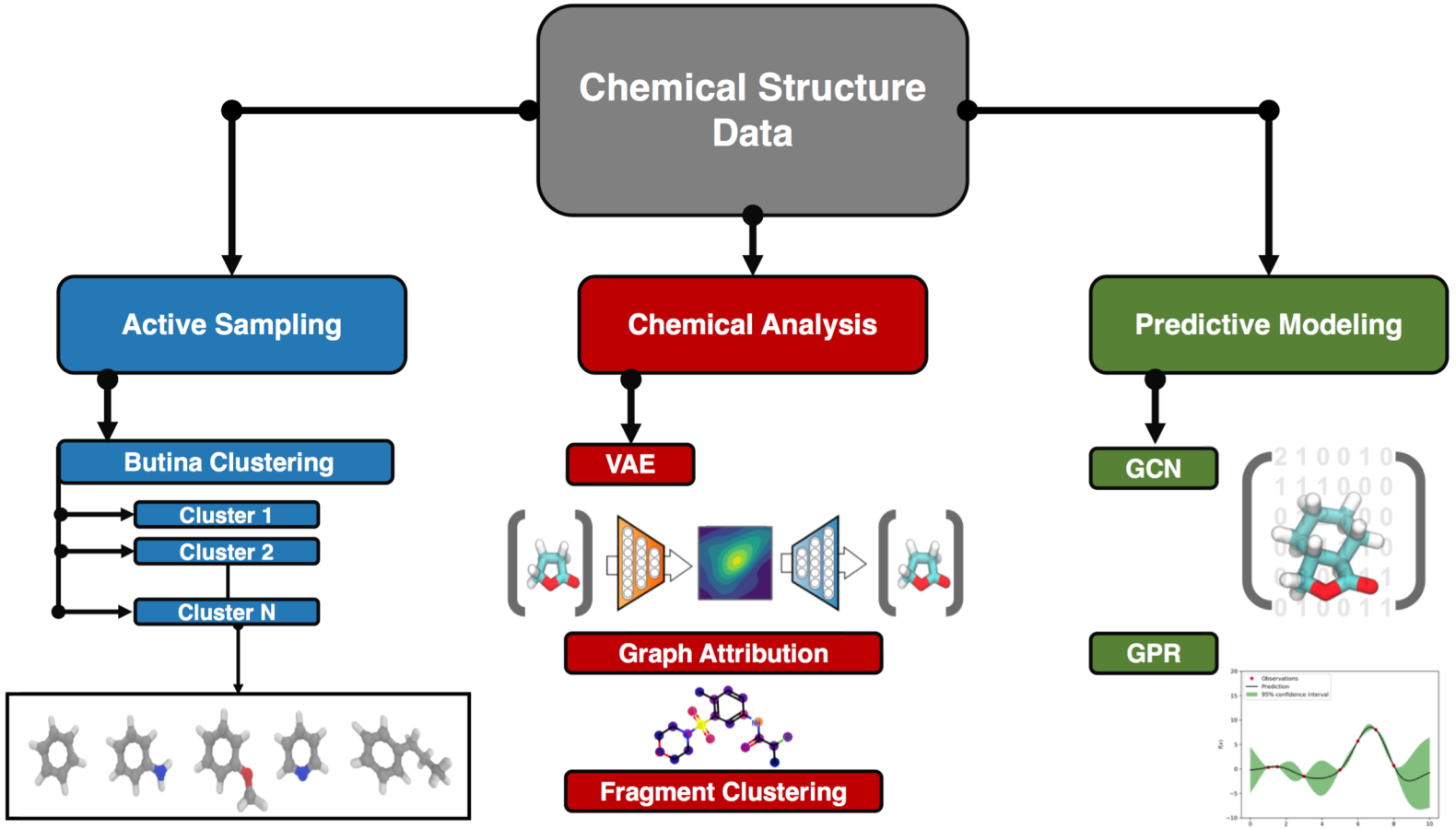


Acknowledgment
Prof. Lian Yu, University of Wisconsin–Madison

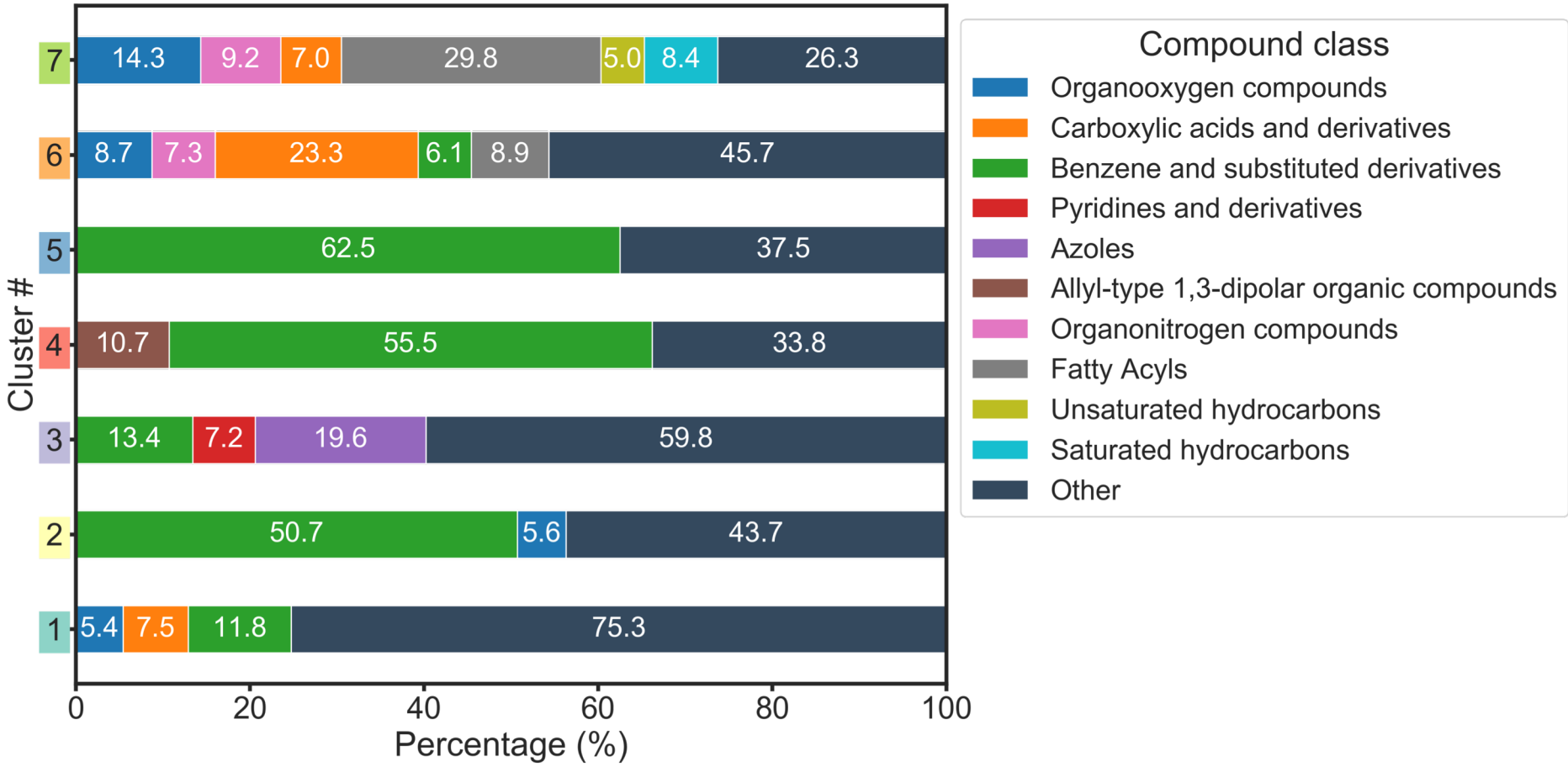
- ML strategies have been widely employed for the prediction of MPs, with models routinely achieving prediction errors of 35-50K[1].

A DATA-DRIVEN APPROACH UTILIZING MACHINE LEARNING

Predict and understand the melting points (MP) of molecules.

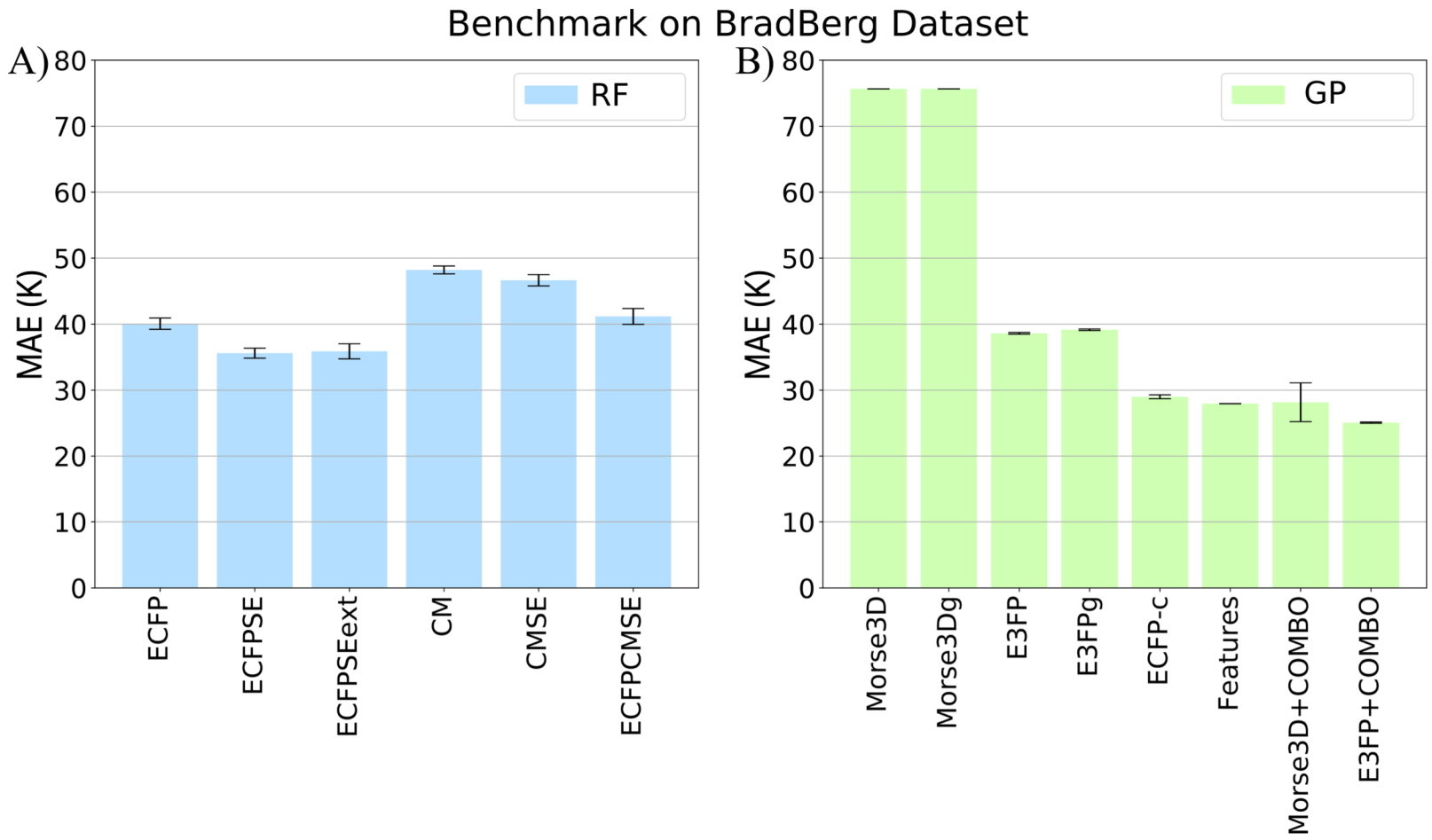


CHEMICAL CLASSIFICATION ANALYSIS



REGRESSSION BENCHMARK

Gold Standard ‘Bradley + Bergstrom’data set



REGRESSION RESULTS

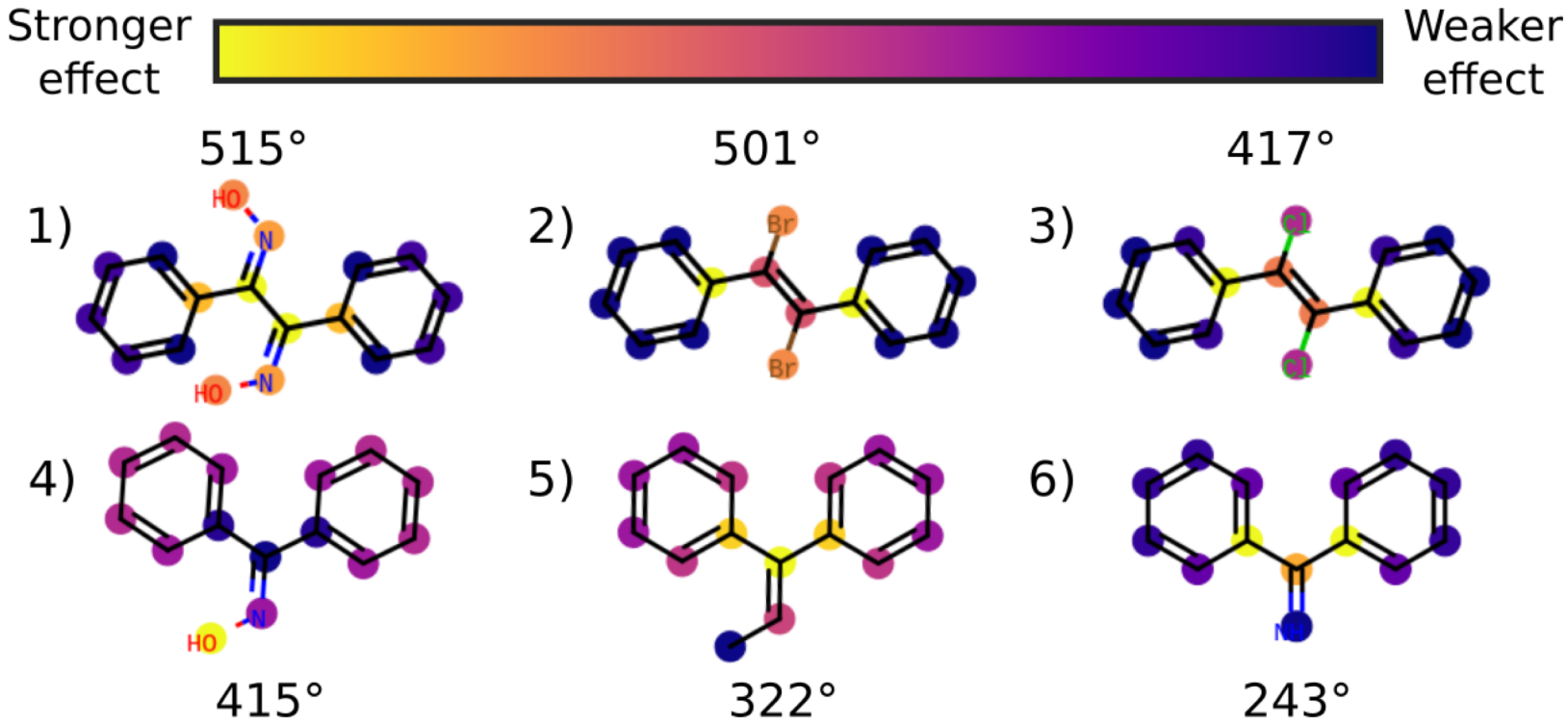
Table 1: MP Regression Results for Experimental Data Sets. † Model using only systems with melting temperatures in the drug-like region [323.15, 523.15]K

Method	OCHEM MAE (K)/R^2	Enamine MAE (K)/R^2	BradBerg MAE (K)/R^2	All MAE/R^2
GPR	30.03(0.01)/0.77	28.60(0.00)/0.64	25.06(0.03)/0.88	28.85(0.01)/0.78
GPR†	26.34(0.05)/0.60	25.65(0.02)/0.59	24.64(0.15)/0.64	25.80(0.03)/0.61
RF	37.56(0.07)/0.66	32.01(0.09)/0.56	35.60(0.75)/0.76	34.62(0.13)/0.66
GCN	31.59(0.83)/0.75	29.45(0.55)/0.62	28.51(0.80)/0.84	29.41(0.26)/0.75

MP Regression Results for Butina 0.6 Clustering Data Sets

Method	MAE (K)	R^2
GPR	28.24 (0.02)	0.75
RF	32.31 (0.28)	0.69
GCN	29.26 (0.27)	0.74

GRAPH ATTRIBUTION OF MP ON SIMILAR MOLECULES.



SUMMARY & OUTLOOK

- The inclusion of 3D structural and quantum-chemically derived features in improving MP prediction accuracy by a modest ~1 - 2K improvement relative to graph-based methods when using GPR, obtaining predicted MAE in the range of 25-29 K.

- GCN has the potential to approaches the target MAE~ 20K

Thank you!
QUESTIONS?

This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.