

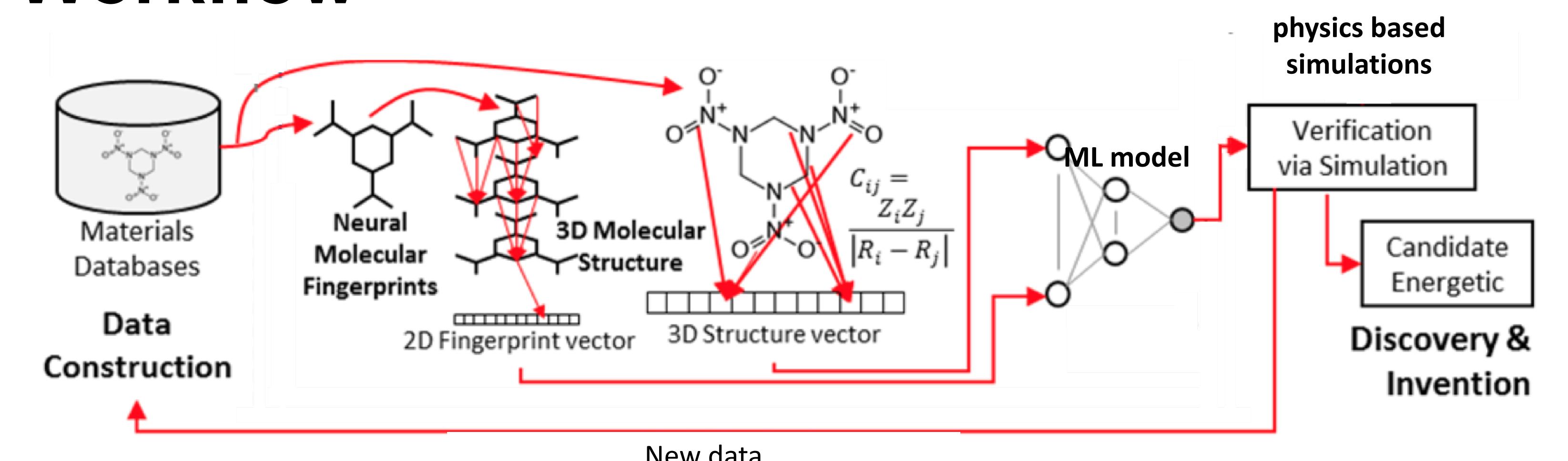
# Machine Learning for Design and Discovery of New Energetic Materials

Daniel C. Elton, Zois Boukouvalas, Mark S. Butrico, Mark D. Fuge, Peter W. Chung  
Department of Mechanical Engineering, University of Maryland, College Park

## Overview

The development of new energetic materials currently involves time consuming synthesis and characterization loops. Recently we presented a proof of concept that machine learning techniques can be used to accurately predict the properties of CNOHF energetic molecules from their molecular graphs (<https://arxiv.org/abs/1801.04900>). We worked with small datasets of 109 and ~400 molecules and studied several molecular featurization methods and a variety of models. We found that kernel ridge regression and sum over bonds generally performed best. We also trained a linear classifier which achieved a 95% accuracy (and ROC AUC of 0.96) when classifying "energetic" molecules and "non-energetic" molecules (here the non-energetic molecules were a collection of small organic molecules and pesticides). After demonstrating the utility of machine learning for prediction and classification, we next explored how machine learning interpretation methods such as sensitivity analysis and feature ranking techniques can illuminate structure-property relationships that may be useful for molecular design. Currently we are exploring how our machine learning models might be coupled with either a variational autoencoder (VAE) or generative adversarial network (GAN) deep learning architecture to rapidly explore chemical space and discover new molecules which are predicted to have user-specified properties.

## Workflow



## Systematic comparison

		$P_{\text{E}}$	$\Delta H_f^{\circ}$	$E_a^{\circ}$	$T_{\text{d}}^{\circ}$	$V_{\text{m}}^{\circ}$	$T_{\text{NT}}^{\circ}$	$T_{\text{R}}^{\circ}$	$T_{\text{A}}^{\circ}$
KRR	Estate	0.10	240.02	0.63	0.41	4.05	500.19	0.18	
	CDS	0.08	198.81	0.50	0.44	0.11	3.07	462.63	0.17
	SoB	0.07	68.73	0.40	0.31	0.09	0.25	2.90	331.36 <b>0.11</b>
	CM eigs	0.09	288.41	0.67	0.67	0.18	0.61	5.67	600.08 <b>0.22</b>
	Bag of Bonds	<b>0.06</b>	166.64	0.47	0.33	0.11	0.29	3.78	478.93 <b>0.13</b>
Ridge	Estate	0.11	240.02	0.63	0.48	0.12	0.29	2.76	359.66 <b>0.13</b>
	CDS	0.07	193.19	0.43	0.39	0.11	0.33	3.23	438.27 <b>0.17</b>
	SoB	<b>0.06</b>	82.08	0.37	0.32	0.10	0.29	3.01	327.43 <b>0.11</b>
	CM eigs	0.06	355.12	0.79	0.60	0.16	0.55	5.82	590.69 <b>0.19</b>
	Bag of Bonds	<b>0.06</b>	163.76	0.68	0.52	0.11	0.31	3.37	472.93 <b>0.19</b>
	Estate-CDS-Sub	0.09	207.78	0.60	0.45	0.13	0.35	4.41	476.07 <b>0.13</b>
SVR	Estate	0.09	223.24	0.52	0.34	0.12	0.32	3.21	436.81 <b>0.18</b>
	CDS	0.07	130.78	0.40	0.31	0.10	0.28	2.97	331.27 <b>0.14</b>
	SoB	<b>0.06</b>	288.41	0.55	0.60	0.15	0.53	4.54	584.44 <b>0.21</b>
	CM eigs	0.06	189.24	0.55	0.62	0.15	0.56	3.75	590.88 <b>0.18</b>
	Bag of Bonds	<b>0.06</b>	189.24	0.55	0.62	0.15	0.56	3.75	593.18 <b>0.13</b>
RF	Estate	0.09	252.74	0.59	0.50	0.14	0.39	4.09	488.98 <b>0.19</b>
	CDS	0.07	241.67	0.46	0.36	0.11	0.29	3.34	435.77 <b>0.15</b>
	SoB	0.07	136.91	0.48	0.40	0.12	0.30	3.47	417.46 <b>0.15</b>
	CM eigs	0.09	286.89	0.67	0.62	0.15	0.51	5.52	512.22 <b>0.20</b>
	Bag of Bonds	<b>0.06</b>	223.24	0.52	0.34	0.12	0.32	3.21	436.81 <b>0.18</b>
	Estate-CDS-Sub	0.07	144.18	0.43	0.34	0.09	0.26	3.11	401.27 <b>0.15</b>
KNN	Estate	0.08	236.55	0.61	0.49	0.15	0.41	4.30	563.89 <b>0.20</b>
	CDS	0.07	242.99	0.55	0.39	0.13	0.33	3.56	478.50 <b>0.18</b>
	SoB	0.08	184.43	0.54	0.44	0.12	0.36	3.65	427.20 <b>0.17</b>
	CM eigs	0.08	238.05	0.53	0.40	0.11	0.32	3.51	527.57 <b>0.22</b>
	Bag of Bonds	0.08	238.05	0.53	0.40	0.11	0.32	3.58	515.25 <b>0.19</b>
	Estate-CDS-Sub	0.08	171.65	0.54	0.43	0.12	0.35	3.57	442.14 <b>0.17</b>
mean	n/a	0.11	309.75	0.69	0.65	0.15	0.55	4.88	629.20 <b>0.22</b>
		1.86	0.50	3.93	0.87	2.04	6.43	32.13	3568.65 <b>1.43</b>

We compared 8 model types and 8 featurizations. Hyperparameter optimization is performed in each case. Shown are the mean average errors in 5-fold cross validation.

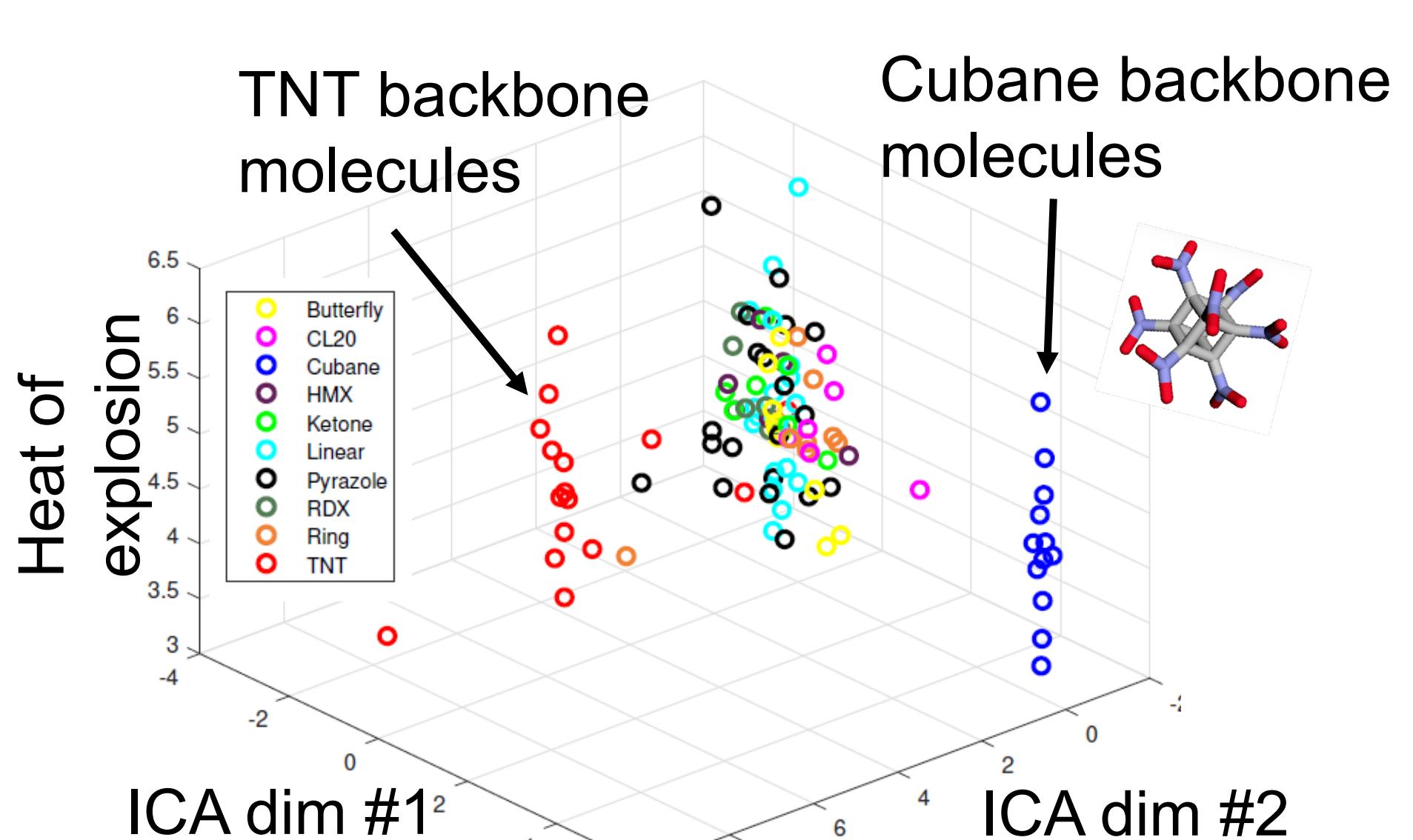
Kernel ridge regression and sum over bonds generally performed best.

There were slight gains from concatenating featurizations.

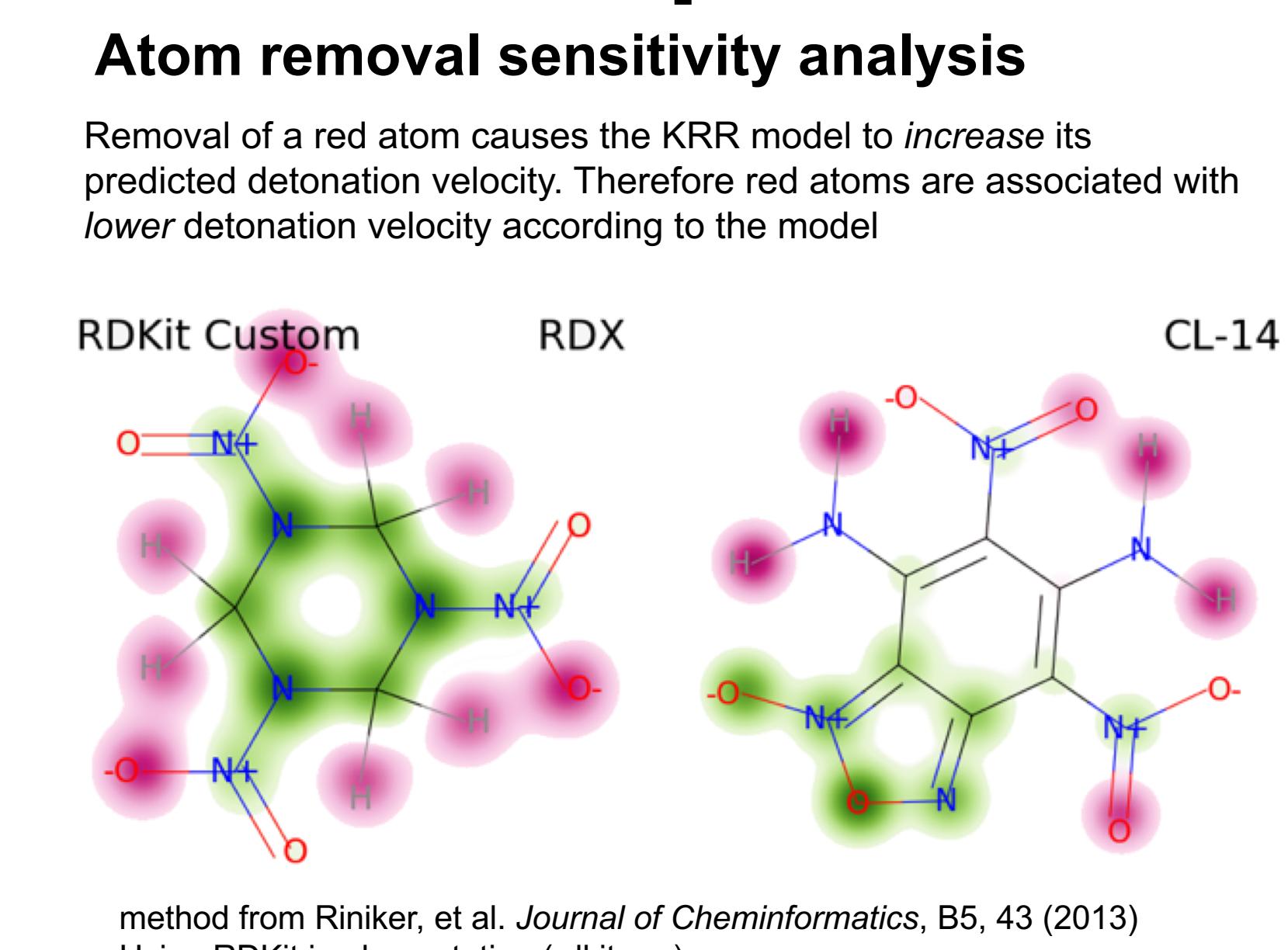
## 2D graph featurizations

- Custom descriptor set (hand picked features)
- Fingerprinting (test of 6 methods in RDKit)
- Custom "neural" graph convolutional fingerprints (Duvenaud et al., NIPS 2015)
- Sum over bonds (bond counting)

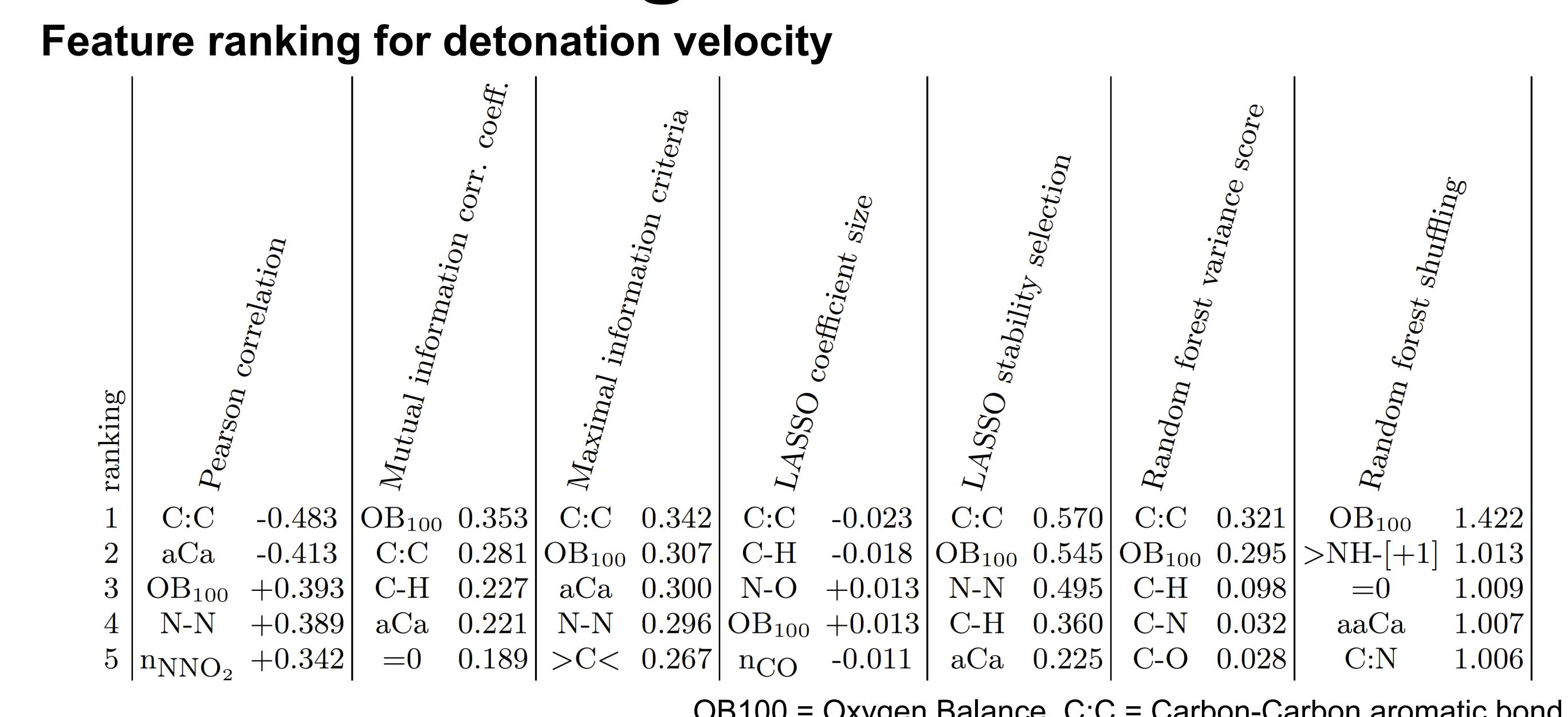
## ICA clusters



## Model interpretation



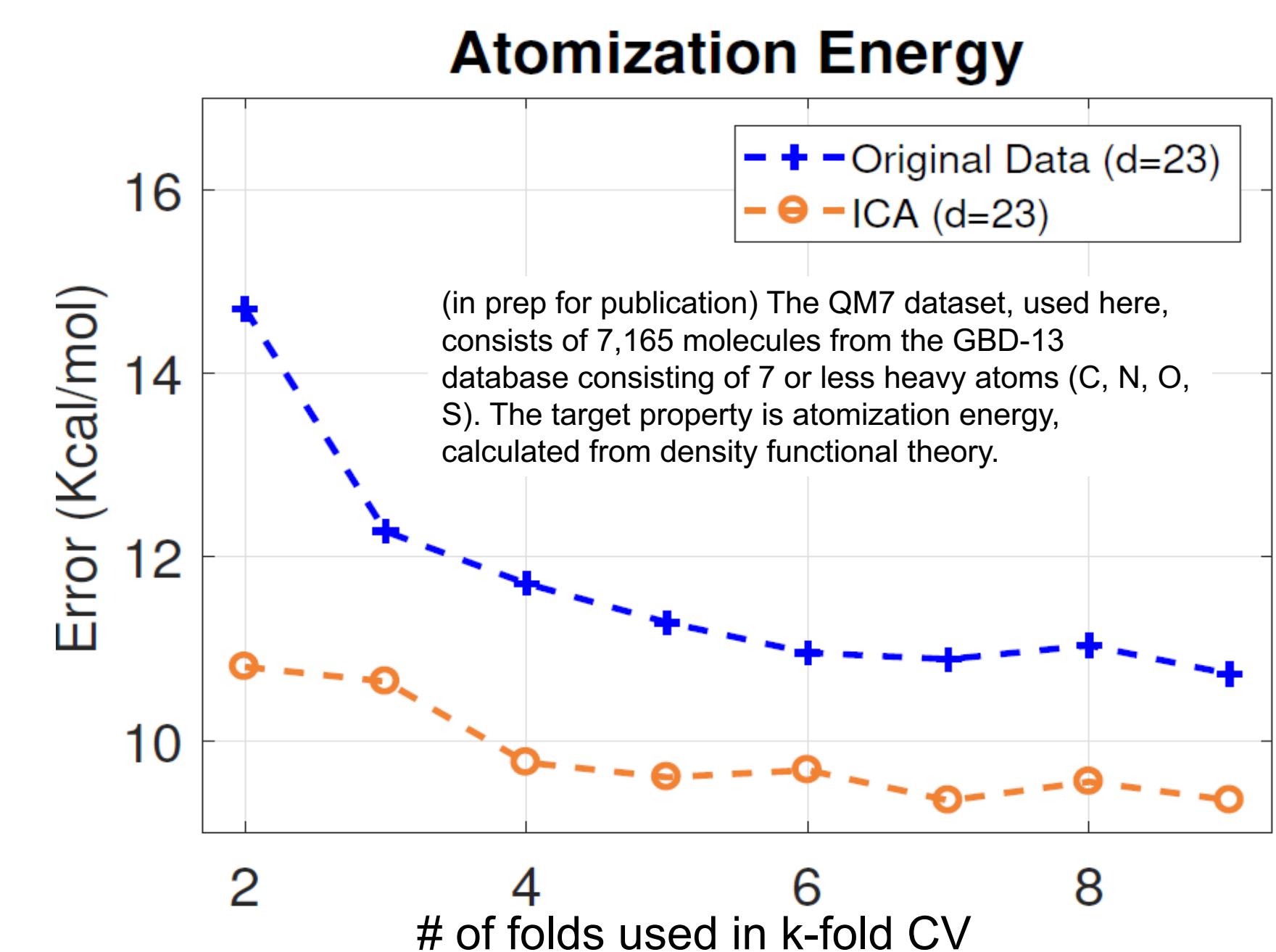
## Feature ranking methods



## 3D structure featurizations

- Coulomb matrix eigenvalues (Rupp et al., 2013)
- Bag of bonds (Hansen et al., 2015)
- Summed bag of bonds

## ICA performance boost



## Molecular generation

- We are exploring several methods to generate SMILES strings, which encode molecules:
- Generative Adversarial Networks (GANs)
    - Goodfellow, et al., 2014 arXiv:1406.2661
  - Variational Autoencoders (VAEs)
    - Gomez-Bombarelli, Aspuru-Guzik, et al. ACS Central Science, 2018
  - Recurrent Neural Networks (RNNs)
- With an objective/evaluation function in hand, one can do Bayesian or gradient ascent optimization with a VAE. With a GAN or RNN, one can learn how to maximize the evaluation function by using reinforcement learning applied to the SMILES sequence generation problem.

## Acknowledgements, Papers & Code

Contacts: delton@umd.edu, pchung15@umd.edu

Thanks to Drs. Ruth Doherty, Bill Wilson, and Brian Barnes for helpful discussions.

### Papers:

- D. C. Elton, Z. Boukouvalas, M. S. Butrico, M. D. Fuge, and P. W. Chung, "Applying machine learning techniques to predict the properties of energetic materials" ([arXiv:1801.04900](https://arxiv.org/abs/1801.04900)), 2018
- B. C. Barnes, D. C. Elton, Z. Boukouvalas, D. E. Taylor, W. D. Mattson, M. D. Fuge, and P. W. Chung, "Machine Learning and Discovery for Energetic Materials", in 16th International Detonation Symposium, Cambridge MD, USA, July 2018.

### Code:

Open Source Python Molecular Machine Learning Toolkit:

<https://github.com/delton137/mml toolkit>

Open Source Python Jupyter Notebooks to reproduce our results:

<https://github.com/delton137/Machine-Learning-Energetic-Molecules-Notebooks>