

# Applying machine learning techniques to predict the properties of energetic materials

Daniel C. Elton,<sup>1</sup> Zois Boukouvalas,<sup>1</sup> Mark S. Butrico,<sup>1</sup> Mark D. Fuge,<sup>1</sup> and Peter W. Chung<sup>1</sup>  
*Department of Mechanical Engineering, University of Maryland, College Park, 20742, United States*

(Dated: 24 May 2018)

## I. SUPPLEMENTARY METHODS: MACHINE LEARNING BASIC WORKFLOW & TERMINOLOGY

In the machine learning workflow, data is first collected, cleaned, and standardized. The target property values are put into a vector  $\mathbf{y}$ . Next, the raw data  $\mathbf{x}_i^{\text{raw}}$  is transformed into fixed-length feature vectors  $\mathbf{x}_i$  that are useful for machine learning, a step called featurization. The feature vectors become the rows in the design matrix  $\mathbf{X}$ . Machine learning algorithms take  $\mathbf{X}$  and  $\mathbf{y}$  as inputs and attempt to find a function  $f(\mathbf{x}) = y^{\text{pred}}$  such that a loss function is minimized. The machine learning model defines a space of functions through which the search is performed. The loss function we use is the squared error loss, defined as :

$$L_{\text{se}}(\{\mathbf{x}_i\}, \mathbf{y}, f(\mathbf{x})) = \sum_{i=1}^N (y_i - f(\mathbf{x}_i))^2 \quad (1)$$

Additionally, we add a regularization term to the loss function

$$L(\{\mathbf{x}_i\}, \mathbf{y}, f(\mathbf{x})) = L_{\text{se}} + \alpha \|\mathbf{w}\|^D \quad (2)$$

here  $\mathbf{w}$  is the “weight vector” containing the tuneable parameters in the model. In what follows we use  $D = 2$ , corresponding to an L2 norm, except for in LASSO regression, which uses  $D = 1$ . Regularization penalizes model functions of higher complexity. Regularization makes machine learning models less prone to overfitting (fitting noise in the data) and less sensitive to ill effects from outliers and strongly correlated features. The parameter  $\alpha$ , sometimes called the “regularization strength” sets the degree of regularization and is an example of a hyperparameter. Hyperparameters are set in advance before the machine learning model is trained.

## II. SUPPLEMENTARY METHODS: STACKING MODELS

We tried a few different variations on the idea of model stacking, a type of ensemble learning.<sup>1,2</sup> In model stacking the outputs of a set of trained models are used as features for another model which makes the final prediction. We experimented with a stacking strategy where a model for density prediction was trained first, and then the output of that model was used as an additional feature when training a model for explosive energy. The improvement in performance was minimal.

## III. SUPPLEMENTARY INFORMATION: TRAIN-TEST SPLITTING METHODS & GENERALIZATION PERFORMANCE

In an attempt to reduce the gap between the train and test errors (ie. generalization error or overfitting), we tried several popular train-test splitting strategies (table III). We compared standard shuffle split, stratification over the chemical groups and stratification over clusters. Clusters were determined using  $k$ -means clustering with  $k = 10$  (10 clusters). Plots of the clustering objective function (essentially a measure of average variance from the cluster centroids) vs the number of clusters used were analyzed. There was no clear “elbow point” indicating a natural number of clusters in the data, so we choose  $k = 10$  clusters. No significant differences were observed in the results, although in the case of the Huang & Massa data the gap between train and test  $r$  values decreased slightly. We also looked into using “scaffold split” which is implemented in the *deepchem* Python package. Scaffold split groups molecules by their chemical scaffold (backbone). We determined the method would not be very useful to implement as many of the scaffolds appeared only once in the data.

## IV. SUPPLEMENTARY INFORMATION: DESCRIPTIONS OF FINGERPRINTS

**Atom-Pairs** (Carhart et al. 1985)<sup>3</sup> - encodes the list of all atom pairs in the molecule. Each atom pair is a triplet of information consisting of (*atom description i*, *distance ij*, *atom description j*). *distance ij* is the distance between atom  $i$  and atom  $j$  as measured by the shortest path length along the molecular bond network, and each *atom description* is a specification of the atom type by atomic number, number of  $\pi$ -bonding electrons, and number of hydrogens. Each possible atom pair is encoded in a unique integer. The list of atom pairs becomes a list of integers which can then be transformed into a bit vector of desired length.

**Topological Torsions** (Ramaswamy et al., 1987)<sup>4</sup> - encodes the list of all topological torsions in the molecule. Each topological torsions is sequence of four bonded non-hydrogen atoms which is encoded in a quartet of information consisting of (*atom description i*, *atom description j*, *atom description k*, *atom description l*). Each *atom description* is an integer containing information on the atomic number and number of  $\pi$ -bonding electrons. As

name	MAE <sub>train</sub>	MAE <sub>test</sub>	MAPE <sub>test</sub>	RMSE <sub>test</sub>	$R^2_{train}$	$R^2_{test}$	$r_{train}$	$r_{test}$
shuffle	0.30 ± 0.01	0.35 ± 0.07	9.87	0.423	0.82	0.63	0.91	0.83
stratified over groups	0.30 ± 0.01	0.38 ± 0.09	11.16	0.454	0.82	0.71	0.91	0.86
stratified over clusters	0.30 ± 0.01	0.36 ± 0.06	9.90	0.435	0.83	0.62	0.91	0.85

TABLE I. Comparison of different splitting techniques for the Huang & Massa data, prediction of explosive energy, using kernel ridge regression and the combined (Estate+Custom Descriptor Set+Sum over Bonds) featurization. The groups were Pyrazole ( $N = 20$ ), Cubane ( $N = 12$ ), HMX ( $N = 6$ ), Linear ( $N = 18$ ), Butterfly ( $N = 10$ ), Ketone ( $N = 7$ ), TNT ( $N = 16$ ), RDX ( $N = 6$ ), Ring ( $N = 8$ ), CL20 ( $N = 6$ ).

name	MAE <sub>train</sub>	MAE <sub>test</sub>	MAPE <sub>test</sub>	RMSE <sub>test</sub>	$R^2_{train}$	$R^2_{test}$	$r_{train}$	$r_{test}$
shuffle	0.15 ± 0.00	0.17 ± 0.03	2.18	0.249	0.89	0.82	0.94	0.91
stratified over groups	0.15 ± 0.00	0.17 ± 0.03	2.20	0.254	0.89	0.81	0.94	0.91
stratified over clusters	0.15 ± 0.00	0.17 ± 0.03	2.16	0.250	0.89	0.82	0.94	0.91

TABLE II. Comparison of different splitting techniques for the sensitivity data, prediction of detonation velocity, using kernel ridge regression and the combined (Estate+Custom Descriptor Set+Sum over Bonds) featurization. The groups were designated Unstable ( $N = 21$ ), NAC ( $N = 86$ ), NNO2 ( $N = 77$ ), CNO2 ( $N = 66$ ), ONO2 ( $N = 13$ ), 5mArN ( $N = 45$ ).

with atom-pair, each resulting type of topological torsion in the molecule is encoded in a unique integer, and a list of integers is generated for further processing.

**Extended Connectivity Fingerprints (ECFPs)** (Rogers & Hahn, 2010)<sup>5</sup> This fingerprinting method was developed at Accelrys, Inc. in 2000 and is also known as Morgan circular fingerprinting, as it is an efficient version of the classic Morgan fingerprinting algorithm (Morgan, 1965).<sup>6</sup> ECFPs are one of the most popular graph-based fingerprints and encode information about graph substructures. The algorithm for generating an ECFP is as follows:

An initial assignment stage in which each atom has an integer identifier assigned to it. This initial list of integers is added to the fingerprint set. The iterative updating stage – each atom collects its own identifier and the identifiers of its immediately neighboring atoms into an array. A hash function is applied to reduce this array back into a new, single-integer identifier. This new list of integers is added to the fingerprint set. This process is iterated a specified number of times  $d$ , known as the “width” of the fingerprint. The duplicate identifier removal stage - multiple occurrences of the same substructure are reduced to a single representative in the final feature list.

In standard ECFPs, each initial atom identifier is a 32 bit ( $\approx$  ten digit) integer which contains the following information: 1. number of immediate neighbors who are heavy (non-hydrogen) atoms, 2. the valence minus the number of hydrogens, 3. the atomic number, 4. the atomic mass, 5. the atomic charge, 6. the number of attached hydrogens (both implicit and explicit), and 7. whether the atom is contained in at least one ring. What information is encoded in each atom identifier can be modified, and thus many variants of ECFPs exist.

**E-state fingerprints** (Hall & Kier, 1995)<sup>7</sup> The E-state fingerprint is based on electrotopological state (E-state) indices,<sup>8</sup> which encode information about associated functional group, graph topology, and the Kier-Hall

electronegativity of each atom. Mathematically, unlike in ECFPs, the E-state indices for each atom are graph invariants, which means their value does not change when the labeling of nodes is permuted. The E-state fingerprint differs from traditional fingerprints as it is fixed length, containing a vector with counts of 79 atom types. Thus, it is more of a descriptor set than a fingerprint. We found that only 31 of the atom types were relevant to the energetic materials we studied. The E-state fingerprint can also be calculated as a real-valued vector which sums the E-state indices for each atom, however we found the predictive performance was exactly the same as with the count vector version.

**Avalon fingerprints** (Gedeck, et al., 2006)<sup>9</sup> Avalon fingerprints hash together a variety of information on atom pairs, atom triplets, graph topology, atomic number, bonding types, and bonding environments.

**RDKit Fingerprint** (Landrum, 2006)<sup>10</sup> The RDKit fingerprint is a simple fingerprinting algorithm which enumerates all paths/subgraphs up to a certain size, hashes them into bit vectors, and then OR’s the bitvectors together into a fingerprint vector.

For two other RDKit fingerprints we tested, the extended reduced graph fingerprint (ErG),<sup>11</sup> and physio-chemical property fingerprints,<sup>12</sup> we were not able to vary the fingerprint length. (ErG is fixed length and physiochemical property fingerprints could only be used at the default length of 1024 due to constraints in RDKit). However both of these performed worse than the all others and are not shown on the plot. This is likely because both of these fingerprints were designed specifically with biological applications in mind.

**ErG fingerprints** (Stiefl, et al., 2006)<sup>11</sup> The extended reduced graph (ErG) fingerprint is generated by enumerating reduced graphs. Reduced graphs are a custom encoding scheme which retains chemically relevant information about graph substructures while throwing out superfluous information. ErG fingerprints contain information about graph topology, atomic charge, and H-bond donor and acceptor sites.

**Physiochemical property fingerprints** (Kearsley, et

al. 1997)<sup>12</sup> These are modification of atom-pair fingerprints were the atom types are determined by a triplet of information: binding property class, atomic *logP* contribution, and partial atomic charges.

<sup>1</sup>D. H. Wolpert, *Neural Networks* **5**, 241 (1992), ISSN 0893-6080.

<sup>2</sup>L. Breiman, *Machine Learning* **24**, 49 (1996), ISSN 1573-0565.

<sup>3</sup>R. E. Carhart, D. H. Smith, and R. Venkataraghavan, *Journal of Chemical Information and Computer Sciences* **25**, 64 (1985).

<sup>4</sup>R. Nilakantan, N. Bauman, J. S. Dixon, and R. Venkataraghavan, *Journal of Chemical Information and Computer Sciences* **27**, 82 (1987).

<sup>5</sup>D. Rogers and M. Hahn, *J. Chem. Info. Mod.* **50**, 742 (2010).

<sup>6</sup>H. L. Morgan, *Journal of Chemical Documentation* **5**, 107 (1965).

<sup>7</sup>L. H. Hall and L. B. Kier, *Journal of Chemical Information and Computer Sciences* **35**, 1039 (1995).

<sup>8</sup>L. B. Kier and L. H. Hall, *Pharmaceutical Research* **7**, 801 (1990).

<sup>9</sup>P. Gedeck, B. Rohde, and C. Bartels, *J. Chem. Info. Mod.* **46**, 1924 (2006).

<sup>10</sup>G. Landrum, *RDKit: Open-source cheminformatics*, <http://www.rdkit.org>.

<sup>11</sup>N. Stiefl, I. A. Watson, K. Baumann, and A. Zaliani, *J. Chem. Info. Mod.* **46**, 208 (2006).

<sup>12</sup>S. K. Kearsley, S. Sallamack, E. M. Fluder, J. D. Andose, R. T. Mosley, and R. P. Sheridan, *Journal of Chemical Information and Computer Sciences* **36**, 118 (1996).

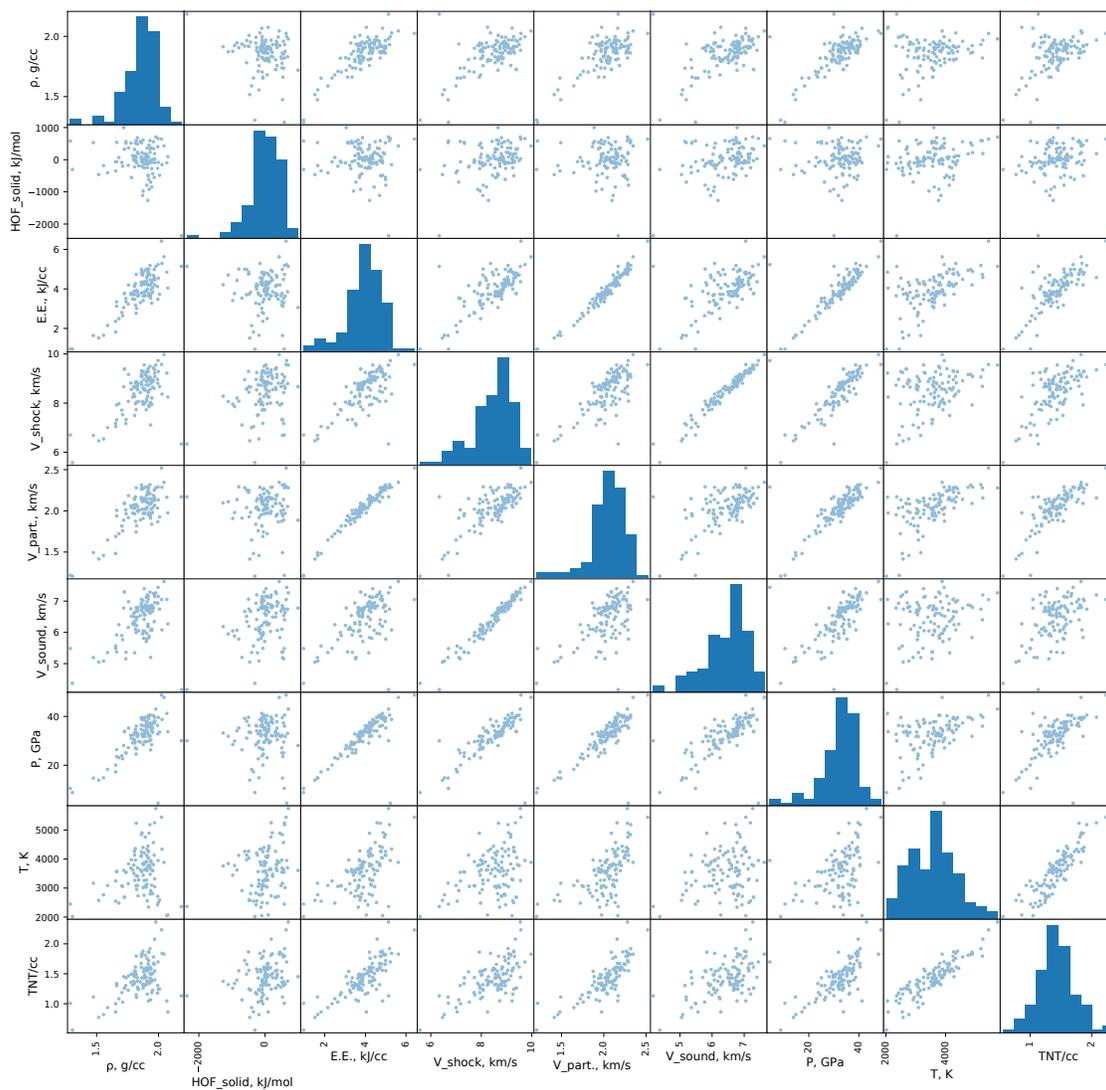


FIG. 1. Correlation plot visualization of the Huang & Massa property data.

name	$MAE_{train}$	$MAE_{test}$	$MAPE_{test}$	$RMSE_{test}$	$R^2_{train}$	$R^2_{test}$	$r_{train}$	$r_{test}$
Kernel Ridge	$0.32 \pm 0.02$	$0.35 \pm 0.06$	10.48	0.464	0.80	0.74	0.80	0.76
Bayesian Ridge	$0.31 \pm 0.01$	$0.36 \pm 0.05$	11.16	0.480	0.80	0.68	0.80	0.70
Elastic Net	$0.31 \pm 0.01$	$0.37 \pm 0.05$	11.13	0.485	0.80	0.68	0.80	0.72
Lasso	$0.32 \pm 0.01$	$0.38 \pm 0.04$	11.13	0.489	0.79	0.71	0.79	0.72
Ridge	$0.32 \pm 0.01$	$0.39 \pm 0.05$	11.44	0.500	0.79	0.66	0.80	0.73
Linear	$0.30 \pm 0.01$	$0.39 \pm 0.02$	11.52	0.510	0.81	0.65	0.81	0.68
Random Forest	$0.19 \pm 0.01$	$0.48 \pm 0.06$	14.48	0.631	0.92	0.49	0.94	0.53
Support Vector	$0.18 \pm 0.01$	$0.48 \pm 0.10$	17.33	0.671	0.89	0.45	0.93	0.50
Gradient Boosted Trees	$0.04 \pm 0.01$	$0.49 \pm 0.04$	14.45	0.628	0.99	0.43	1.00	0.49
k-Nearest Neighbors	$0.35 \pm 0.03$	$0.49 \pm 0.09$	15.67	0.668	0.78	0.44	0.79	0.52
Gaussian Process	$0.00 \pm 0.00$	$0.63 \pm 0.15$	23.45	0.889	1.00	0.04	1.00	0.15
Neural Network	$0.76 \pm 0.02$	$0.96 \pm 0.10$	28.03	1.250	-0.04	-0.92	0.29	0.19
Take the mean	$0.68 \pm 0.02$	$0.69 \pm 0.08$	24.91	0.935	0.00	-0.05	0.00	0.00

TABLE III. A complete comparison of all the models tested for explosive energy prediction with the sum over bonds featurization and 5-fold cross validation. One train-test split ( $n_{train}/n_{test} = 5$ ) is shown for case, along with the mean average error and Pearson correlation  $r$  averaged over the 5 train-test splits. Gaussian process regression can improved with by implementing custom kernels (our result uses the default RBF kernel). We found neural network models could be found with more more extensive hyperparameter optimization, but still were worse than Kernel Ridge.

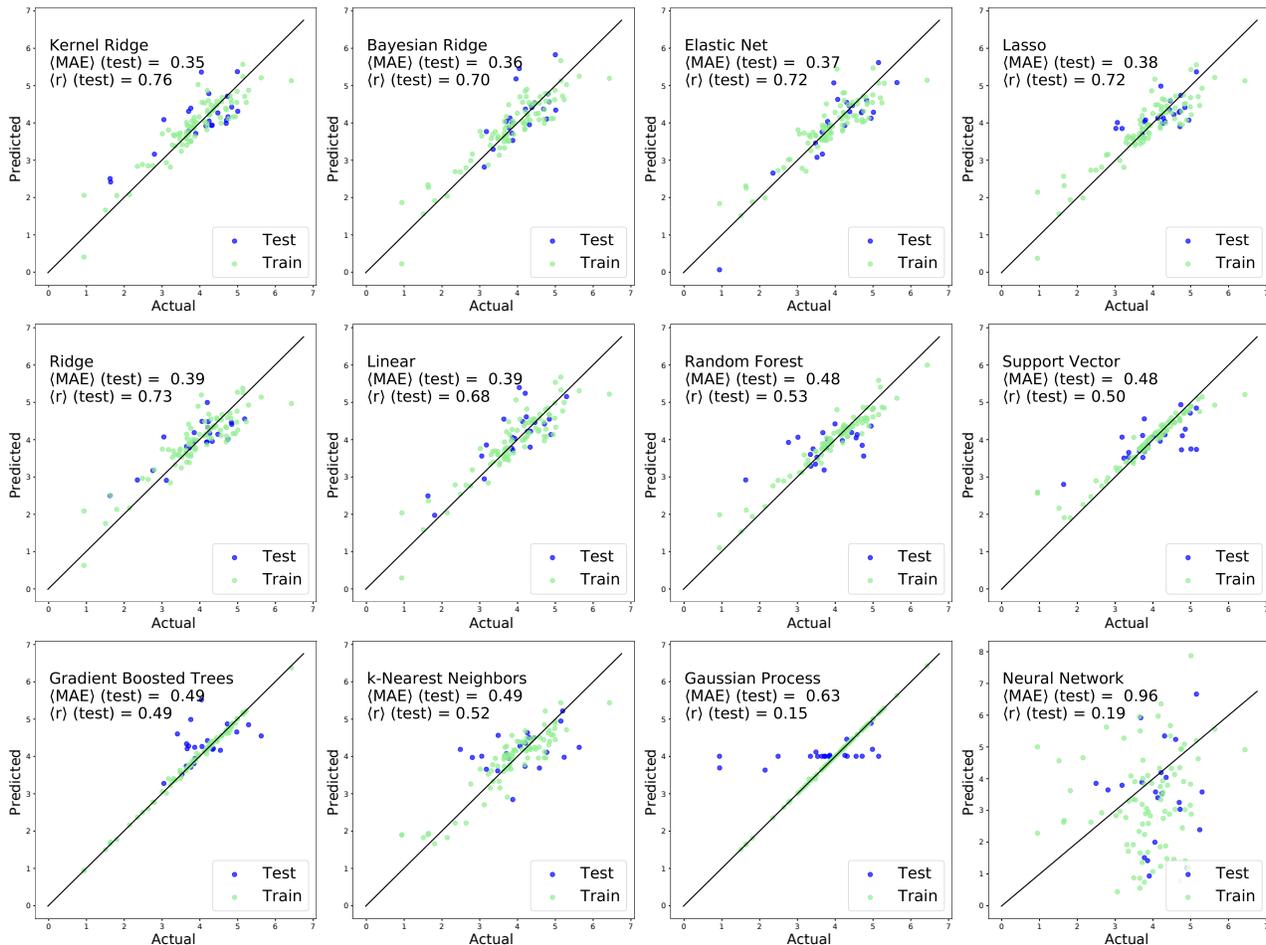


FIG. 2. A complete comparison of all the models tested for explosive energy prediction with the sum over bonds featurization and 5-fold cross validation.

		$\rho$ , $\frac{g}{cc}$	$\Delta H_f^s$ , $\frac{kJ}{mol}$	$E_c$ , $\frac{kJ}{cc}$	$V_s$ , $\frac{km}{s}$	$V_p$ , $\frac{km}{s}$	$V_{snd}$ , $\frac{km}{s}$	$P$ , GPa	$T$ , K	TNT equiv cc
KRR	Estate	0.10 <sup>0.11</sup> <sub>0.09</sub>	261.02 <sup>229.53</sup> <sub>229.51</sub>	0.63 <sup>0.67</sup> <sub>0.59</sub>	0.48 <sup>0.54</sup> <sub>0.43</sub>	0.13 <sup>0.14</sup> <sub>0.12</sub>	0.41 <sup>0.45</sup> <sub>0.37</sub>	4.95 <sup>5.32</sup> <sub>4.59</sub>	500.19 <sup>542.56</sup> <sub>457.82</sub>	0.18 <sup>0.20</sup> <sub>0.16</sub>
	CDS	0.08 <sup>0.09</sup> <sub>0.07</sub>	198.81 <sup>211.87</sup> <sub>185.76</sub>	0.50 <sup>0.55</sup> <sub>0.46</sub>	0.44 <sup>0.49</sup> <sub>0.39</sub>	0.11 <sup>0.12</sup> <sub>0.10</sub>	0.37 <sup>0.41</sup> <sub>0.34</sub>	3.07 <sup>3.38</sup> <sub>2.76</sub>	462.63 <sup>513.95</sup> <sub>411.32</sub>	0.17 <sup>0.19</sup> <sub>0.16</sub>
	CM eigs	0.09 <sup>0.11</sup> <sub>0.08</sub>	288.41 <sup>315.10</sup> <sub>261.72</sub>	0.67 <sup>0.75</sup> <sub>0.59</sub>	0.67 <sup>0.74</sup> <sub>0.61</sub>	0.18 <sup>0.20</sup> <sub>0.16</sub>	0.61 <sup>0.71</sup> <sub>0.52</sub>	5.67 <sup>6.10</sup> <sub>5.23</sub>	600.08 <sup>658.08</sup> <sub>542.09</sub>	0.22 <sup>0.24</sup> <sub>0.20</sub>
	Bag of Bonds	0.06 <sup>0.07</sup> <sub>0.06</sub>	166.66 <sup>189.10</sup> <sub>144.23</sub>	0.47 <sup>0.51</sup> <sub>0.44</sub>	0.33 <sup>0.35</sup> <sub>0.30</sub>	0.11 <sup>0.12</sup> <sub>0.10</sub>	0.29 <sup>0.33</sup> <sub>0.25</sub>	3.38 <sup>3.66</sup> <sub>3.09</sub>	478.93 <sup>522.32</sup> <sub>435.54</sub>	0.18 <sup>0.20</sup> <sub>0.17</sub>
	Estate+CDS+SoB	<b>0.06</b> <sup>0.07</sup> <sub>0.06</sub>	<b>71.40</b> <sup>79.69</sup> <sub>63.12</sub>	<b>0.36</b> <sup>0.39</sup> <sub>0.33</sub>	<b>0.32</b> <sup>0.35</sup> <sub>0.30</sub>	<b>0.10</b> <sup>0.10</sup> <sub>0.09</sub>	<b>0.29</b> <sup>0.28</sup> <sub>0.28</sub>	<b>2.76</b> <sup>2.98</sup> <sub>2.55</sub>	<b>359.66</b> <sup>323.31</sup> <sub>323.31</sub>	<b>0.13</b> <sup>0.15</sup> <sub>0.12</sub>
SVR	OB <sub>1600</sub>	0.07 <sup>0.08</sup> <sub>0.07</sub>	368.32 <sup>400.49</sup> <sub>336.15</sub>	0.48 <sup>0.52</sup> <sub>0.45</sub>	0.53 <sup>0.59</sup> <sub>0.48</sub>	0.18 <sup>0.28</sup> <sub>0.07</sub>	0.50 <sup>0.58</sup> <sub>0.41</sub>	3.43 <sup>3.91</sup> <sub>2.96</sub>	704.57 <sup>924.34</sup> <sub>484.80</sub>	0.22 <sup>0.24</sup> <sub>0.20</sub>
	Estate	0.09 <sup>0.10</sup> <sub>0.09</sub>	207.78 <sup>234.27</sup> <sub>181.30</sub>	0.60 <sup>0.64</sup> <sub>0.57</sub>	0.45 <sup>0.50</sup> <sub>0.41</sub>	0.13 <sup>0.14</sup> <sub>0.11</sub>	0.35 <sup>0.37</sup> <sub>0.33</sub>	4.41 <sup>4.81</sup> <sub>4.00</sub>	476.06 <sup>518.97</sup> <sub>433.15</sub>	0.17 <sup>0.18</sup> <sub>0.16</sub>
	CDS	0.07 <sup>0.08</sup> <sub>0.07</sub>	223.24 <sup>257.71</sup> <sub>188.77</sub>	0.52 <sup>0.56</sup> <sub>0.49</sub>	0.34 <sup>0.37</sup> <sub>0.31</sub>	0.12 <sup>0.13</sup> <sub>0.11</sub>	0.32 <sup>0.35</sup> <sub>0.29</sub>	3.21 <sup>3.42</sup> <sub>3.01</sub>	436.81 <sup>475.64</sup> <sub>397.99</sub>	0.18 <sup>0.19</sup> <sub>0.16</sub>
	SoB	0.06 <sup>0.07</sup> <sub>0.06</sub>	130.78 <sup>152.42</sup> <sub>109.13</sub>	0.40 <sup>0.43</sup> <sub>0.37</sub>	<b>0.31</b> <sup>0.33</sup> <sub>0.29</sub>	0.10 <sup>0.11</sup> <sub>0.09</sub>	0.28 <sup>0.30</sup> <sub>0.26</sub>	2.97 <sup>3.16</sup> <sub>2.78</sub>	331.27 <sup>364.66</sup> <sub>297.88</sub>	0.14 <sup>0.16</sup> <sub>0.13</sub>
	CM eigs	0.08 <sup>0.09</sup> <sub>0.07</sub>	288.41 <sup>326.75</sup> <sub>250.08</sub>	0.55 <sup>0.58</sup> <sub>0.52</sub>	0.60 <sup>0.65</sup> <sub>0.55</sub>	0.15 <sup>0.16</sup> <sub>0.14</sub>	0.53 <sup>0.58</sup> <sub>0.48</sub>	4.54 <sup>4.94</sup> <sub>4.13</sub>	584.44 <sup>624.58</sup> <sub>544.29</sub>	0.21 <sup>0.23</sup> <sub>0.20</sub>
Ridge	Bag of Bonds	0.07 <sup>0.08</sup> <sub>0.07</sub>	159.24 <sup>188.03</sup> <sub>130.45</sub>	0.47 <sup>0.50</sup> <sub>0.44</sub>	0.35 <sup>0.37</sup> <sub>0.32</sub>	0.12 <sup>0.12</sup> <sub>0.11</sub>	0.28 <sup>0.31</sup> <sub>0.26</sub>	3.34 <sup>3.53</sup> <sub>3.15</sub>	385.59 <sup>419.20</sup> <sub>351.97</sub>	0.18 <sup>0.19</sup> <sub>0.17</sub>
	Estate+CDS+SoB	0.06 <sup>0.06</sup> <sub>0.05</sub>	129.89 <sup>148.84</sup> <sub>110.93</sub>	0.37 <sup>0.40</sup> <sub>0.35</sub>	0.34 <sup>0.38</sup> <sub>0.31</sub>	0.10 <sup>0.10</sup> <sub>0.09</sub>	0.28 <sup>0.31</sup> <sub>0.26</sub>	<b>2.73</b> <sup>2.91</sup> <sub>2.55</sub>	353.18 <sup>379.95</sup> <sub>326.41</sub>	0.13 <sup>0.15</sup> <sub>0.12</sub>
	OB <sub>1600</sub>	0.07 <sup>0.08</sup> <sub>0.06</sub>	337.94 <sup>373.47</sup> <sub>302.41</sub>	0.52 <sup>0.56</sup> <sub>0.47</sub>	0.44 <sup>0.47</sup> <sub>0.41</sub>	0.12 <sup>0.13</sup> <sub>0.11</sub>	0.41 <sup>0.43</sup> <sub>0.38</sub>	3.78 <sup>4.02</sup> <sub>3.54</sub>	546.42 <sup>590.40</sup> <sub>502.44</sub>	0.19 <sup>0.20</sup> <sub>0.18</sub>
	Estate	0.09 <sup>0.10</sup> <sub>0.09</sub>	269.11 <sup>302.51</sup> <sub>235.70</sub>	0.58 <sup>0.63</sup> <sub>0.53</sub>	0.57 <sup>0.60</sup> <sub>0.53</sub>	0.14 <sup>0.15</sup> <sub>0.14</sub>	0.45 <sup>0.48</sup> <sub>0.42</sub>	4.71 <sup>5.01</sup> <sub>4.41</sub>	491.21 <sup>534.08</sup> <sub>448.34</sub>	0.19 <sup>0.20</sup> <sub>0.17</sub>
	CDS	0.07 <sup>0.07</sup> <sub>0.06</sub>	193.19 <sup>207.17</sup> <sub>179.21</sub>	0.43 <sup>0.47</sup> <sub>0.40</sub>	0.39 <sup>0.43</sup> <sub>0.35</sub>	0.11 <sup>0.12</sup> <sub>0.10</sub>	0.33 <sup>0.35</sup> <sub>0.31</sub>	3.23 <sup>3.56</sup> <sub>2.91</sub>	438.27 <sup>467.92</sup> <sub>408.61</sub>	0.17 <sup>0.18</sup> <sub>0.15</sub>
RF	SoB	<b>0.06</b> <sup>0.06</sup> <sub>0.05</sub>	<b>82.00</b> <sup>88.76</sup> <sub>75.24</sub>	<b>0.37</b> <sup>0.40</sup> <sub>0.34</sub>	<b>0.32</b> <sup>0.34</sup> <sub>0.30</sub>	0.10 <sup>0.11</sup> <sub>0.09</sub>	0.29 <sup>0.31</sup> <sub>0.27</sub>	3.01 <sup>3.20</sup> <sub>2.82</sub>	<b>327.43</b> <sup>358.22</sup> <sub>296.65</sub>	<b>0.11</b> <sup>0.12</sup> <sub>0.11</sub>
	CM eigs	0.09 <sup>0.10</sup> <sub>0.08</sub>	355.12 <sup>396.42</sup> <sub>313.81</sub>	0.79 <sup>0.92</sup> <sub>0.65</sub>	0.60 <sup>0.64</sup> <sub>0.57</sub>	0.16 <sup>0.19</sup> <sub>0.13</sub>	0.55 <sup>0.59</sup> <sub>0.51</sub>	5.82 <sup>6.37</sup> <sub>5.26</sub>	590.69 <sup>635.69</sup> <sub>545.69</sub>	0.19 <sup>0.20</sup> <sub>0.18</sub>
	Bag of Bonds	0.06 <sup>0.07</sup> <sub>0.06</sub>	163.76 <sup>185.79</sup> <sub>141.74</sub>	0.48 <sup>0.51</sup> <sub>0.45</sub>	0.32 <sup>0.35</sup> <sub>0.29</sub>	0.11 <sup>0.12</sup> <sub>0.10</sub>	0.31 <sup>0.33</sup> <sub>0.29</sub>	3.37 <sup>3.67</sup> <sub>3.07</sub>	472.99 <sup>512.03</sup> <sub>433.83</sub>	0.19 <sup>0.21</sup> <sub>0.18</sub>
	Estate+CDS+SoB	0.06 <sup>0.06</sup> <sub>0.06</sub>	77.31 <sup>84.13</sup> <sub>70.48</sub>	0.39 <sup>0.41</sup> <sub>0.36</sub>	0.32 <sup>0.33</sup> <sub>0.31</sub>	0.10 <sup>0.11</sup> <sub>0.09</sub>	0.28 <sup>0.31</sup> <sub>0.25</sub>	2.78 <sup>2.95</sup> <sub>2.60</sub>	383.07 <sup>415.60</sup> <sub>350.54</sub>	0.13 <sup>0.14</sup> <sub>0.12</sub>
	OB <sub>1600</sub>	0.06 <sup>0.07</sup> <sub>0.06</sub>	362.89 <sup>396.80</sup> <sub>328.98</sub>	0.49 <sup>0.54</sup> <sub>0.45</sub>	0.51 <sup>0.54</sup> <sub>0.47</sub>	0.11 <sup>0.12</sup> <sub>0.10</sub>	0.46 <sup>0.49</sup> <sub>0.43</sub>	3.39 <sup>3.65</sup> <sub>3.12</sub>	545.27 <sup>575.00</sup> <sub>515.54</sub>	0.19 <sup>0.20</sup> <sub>0.18</sub>
kNN	Estate	0.09 <sup>0.10</sup> <sub>0.09</sub>	252.74 <sup>291.84</sup> <sub>213.65</sub>	0.59 <sup>0.64</sup> <sub>0.54</sub>	0.50 <sup>0.54</sup> <sub>0.46</sub>	0.14 <sup>0.15</sup> <sub>0.13</sub>	0.39 <sup>0.42</sup> <sub>0.36</sub>	4.09 <sup>4.38</sup> <sub>3.80</sub>	488.98 <sup>533.33</sup> <sub>443.63</sub>	0.19 <sup>0.21</sup> <sub>0.18</sub>
	CDS	0.07 <sup>0.08</sup> <sub>0.07</sub>	241.67 <sup>265.41</sup> <sub>217.93</sub>	0.46 <sup>0.50</sup> <sub>0.43</sub>	0.36 <sup>0.39</sup> <sub>0.33</sub>	0.11 <sup>0.12</sup> <sub>0.10</sub>	0.29 <sup>0.32</sup> <sub>0.27</sub>	3.34 <sup>3.57</sup> <sub>3.11</sub>	435.77 <sup>475.37</sup> <sub>396.17</sub>	0.16 <sup>0.18</sup> <sub>0.15</sub>
	SoB	0.07 <sup>0.08</sup> <sub>0.07</sub>	136.91 <sup>157.28</sup> <sub>116.55</sub>	0.48 <sup>0.51</sup> <sub>0.45</sub>	0.40 <sup>0.44</sup> <sub>0.37</sub>	0.12 <sup>0.13</sup> <sub>0.10</sub>	0.30 <sup>0.33</sup> <sub>0.28</sub>	3.47 <sup>3.74</sup> <sub>3.21</sub>	417.46 <sup>456.54</sup> <sub>378.39</sub>	0.15 <sup>0.16</sup> <sub>0.13</sub>
	CM eigs	0.09 <sup>0.10</sup> <sub>0.08</sub>	286.89 <sup>311.55</sup> <sub>262.23</sub>	0.67 <sup>0.72</sup> <sub>0.61</sub>	0.62 <sup>0.66</sup> <sub>0.58</sub>	0.15 <sup>0.17</sup> <sub>0.14</sub>	0.51 <sup>0.56</sup> <sub>0.47</sub>	5.52 <sup>6.02</sup> <sub>5.02</sub>	512.22 <sup>550.43</sup> <sub>474.00</sub>	0.20 <sup>0.22</sup> <sub>0.18</sub>
	Bag of Bonds	0.07 <sup>0.07</sup> <sub>0.06</sub>	172.41 <sup>194.19</sup> <sub>150.63</sub>	0.46 <sup>0.50</sup> <sub>0.43</sub>	0.36 <sup>0.39</sup> <sub>0.33</sub>	0.10 <sup>0.11</sup> <sub>0.10</sub>	0.29 <sup>0.32</sup> <sub>0.25</sub>	3.10 <sup>3.29</sup> <sub>2.92</sub>	418.35 <sup>449.87</sup> <sub>386.83</sub>	0.16 <sup>0.18</sup> <sub>0.15</sub>
kNN	Estate+CDS+SoB	0.07 <sup>0.08</sup> <sub>0.07</sub>	144.18 <sup>165.03</sup> <sub>123.34</sub>	0.43 <sup>0.47</sup> <sub>0.39</sub>	0.34 <sup>0.36</sup> <sub>0.31</sub>	<b>0.09</b> <sup>0.10</sup> <sub>0.08</sub>	<b>0.26</b> <sup>0.29</sup> <sub>0.24</sub>	3.11 <sup>3.39</sup> <sub>2.82</sub>	401.27 <sup>441.62</sup> <sub>361.01</sub>	0.15 <sup>0.16</sup> <sub>0.14</sub>
	OB <sub>1600</sub>	0.08 <sup>0.08</sup> <sub>0.07</sub>	443.56 <sup>473.61</sup> <sub>413.51</sub>	0.48 <sup>0.53</sup> <sub>0.43</sub>	0.48 <sup>0.51</sup> <sub>0.46</sub>	0.12 <sup>0.13</sup> <sub>0.11</sub>	0.45 <sup>0.48</sup> <sub>0.42</sub>	3.58 <sup>3.95</sup> <sub>3.21</sub>	611.27 <sup>654.07</sup> <sub>568.47</sub>	0.21 <sup>0.23</sup> <sub>0.19</sub>
	Estate	0.08 <sup>0.09</sup> <sub>0.08</sub>	236.55 <sup>266.19</sup> <sub>206.91</sub>	0.61 <sup>0.64</sup> <sub>0.57</sub>	0.49 <sup>0.52</sup> <sub>0.46</sub>	0.15 <sup>0.16</sup> <sub>0.14</sub>	0.41 <sup>0.43</sup> <sub>0.38</sub>	4.30 <sup>4.62</sup> <sub>3.99</sub>	563.89 <sup>609.25</sup> <sub>518.53</sub>	0.20 <sup>0.21</sup> <sub>0.18</sub>
	CDS	0.07 <sup>0.08</sup> <sub>0.07</sub>	242.99 <sup>271.36</sup> <sub>214.62</sub>	0.55 <sup>0.59</sup> <sub>0.51</sub>	0.39 <sup>0.42</sup> <sub>0.36</sub>	0.13 <sup>0.14</sup> <sub>0.12</sub>	0.33 <sup>0.35</sup> <sub>0.30</sub>	3.56 <sup>3.75</sup> <sub>3.38</sub>	478.50 <sup>508.42</sup> <sub>448.57</sub>	0.18 <sup>0.19</sup> <sub>0.17</sub>
	SoB	0.08 <sup>0.08</sup> <sub>0.07</sub>	184.43 <sup>209.57</sup> <sub>159.29</sub>	0.54 <sup>0.58</sup> <sub>0.50</sub>	0.44 <sup>0.48</sup> <sub>0.39</sub>	0.12 <sup>0.14</sup> <sub>0.11</sub>	0.36 <sup>0.38</sup> <sub>0.33</sub>	3.63 <sup>3.93</sup> <sub>3.38</sub>	427.20 <sup>461.44</sup> <sub>392.96</sub>	0.17 <sup>0.19</sup> <sub>0.16</sub>
kNN	CM eigs	0.10 <sup>0.11</sup> <sub>0.09</sub>	343.48 <sup>368.70</sup> <sub>318.27</sub>	0.62 <sup>0.67</sup> <sub>0.56</sub>	0.67 <sup>0.72</sup> <sub>0.62</sub>	0.15 <sup>0.16</sup> <sub>0.13</sub>	0.51 <sup>0.54</sup> <sub>0.49</sub>	5.52 <sup>5.84</sup> <sub>5.21</sub>	570.59 <sup>604.71</sup> <sub>536.39</sub>	0.22 <sup>0.24</sup> <sub>0.21</sub>
	Bag of Bonds	0.08 <sup>0.09</sup> <sub>0.07</sub>	238.05 <sup>258.83</sup> <sub>217.27</sub>	0.53 <sup>0.57</sup> <sub>0.49</sub>	0.40 <sup>0.44</sup> <sub>0.37</sub>	0.11 <sup>0.12</sup> <sub>0.10</sub>	0.32 <sup>0.35</sup> <sub>0.29</sub>	3.58 <sup>3.88</sup> <sub>3.29</sub>	515.29 <sup>546.78</sup> <sub>483.71</sub>	0.19 <sup>0.20</sup> <sub>0.18</sub>
	Estate+CDS+SoB	0.08 <sup>0.08</sup> <sub>0.07</sub>	171.65 <sup>195.35</sup> <sub>147.94</sub>	0.54 <sup>0.59</sup> <sub>0.49</sub>	0.43 <sup>0.46</sup> <sub>0.40</sub>	0.12 <sup>0.13</sup> <sub>0.11</sub>	0.35 <sup>0.37</sup> <sub>0.33</sub>	3.57 <sup>3.76</sup> <sub>3.37</sub>	442.14 <sup>476.56</sup> <sub>407.73</sub>	0.17 <sup>0.18</sup> <sub>0.16</sub>
	OB <sub>1600</sub>	0.08 <sup>0.08</sup> <sub>0.07</sub>	333.95 <sup>359.50</sup> <sub>308.41</sub>	0.49 <sup>0.53</sup> <sub>0.44</sub>	0.50 <sup>0.53</sup> <sub>0.46</sub>	0.11 <sup>0.12</sup> <sub>0.10</sub>	0.45 <sup>0.49</sup> <sub>0.42</sub>	3.45 <sup>3.73</sup> <sub>3.17</sub>	515.62 <sup>554.44</sup> <sub>476.80</sub>	0.19 <sup>0.21</sup> <sub>0.17</sub>

TABLE IV. Average mean absolute errors in the test sets for different combinations of target property, model and featurization. Hyperparameter optimization and 20-fold shuffle split with a train-test ratio of 4:1 was used. 95% confidence intervals for the mean errors were calculated using the standard statistical method with the corresponding  $t$  value.

		$\rho, \frac{\text{g}}{\text{cc}}$	$\Delta H_f^s, \frac{\text{kJ}}{\text{mol}}$	$E_c, \frac{\text{kJ}}{\text{cc}}$	$V_s, \frac{\text{km}}{\text{s}}$	$V_p, \frac{\text{km}}{\text{s}}$	$V_{\text{snd}}, \frac{\text{km}}{\text{s}}$	$P$ (GPa)	$T$ (K)	$\frac{\text{TNT}_{\text{equiv}}}{\text{cc}}$	avg
KR	Estate	0.69	0.39	0.72	0.53	0.58	0.56	0.55	0.51	0.60	0.57
	CDS	0.57	0.63	0.56	0.57	0.44	0.50	0.45	0.50	0.42	0.52
	SoB	0.70	0.95	0.72	0.70	0.67	0.73	0.67	0.61	0.70	<b>0.72</b>
	CM eigs	0.40	0.33	0.52	0.25	0.51	0.20	0.33	0.21	0.34	0.34
	Bag of Bonds	0.61	0.78	0.72	0.73	0.56	0.75	0.61	0.46	0.49	0.63
	Estate+CDS+SoB	0.74	0.94	0.78	0.70	0.65	0.71	0.65	0.50	0.65	0.70
SVR	Estate	0.55	0.57	0.56	0.70	0.70	0.61	0.60	0.52	0.53	0.59
	CDS	0.39	0.06	0.38	0.51	0.37	0.48	0.46	0.34	0.31	0.37
	SoB	0.38	0.61	0.60	0.62	0.59	0.58	0.57	0.58	0.57	0.57
	CM eigs	0.02	0.05	0.03	0.05	0.03	0.05	0.03	0.06	0.11	0.05
	Bag of Bonds	0.04	0.12	0.04	0.14	0.10	0.15	0.10	0.14	0.10	0.10
	Estate+CDS+SoB	0.32	0.58	0.46	0.56	0.43	0.54	0.47	0.56	0.58	0.50
Ridge	Estate	0.69	0.28	0.69	0.56	0.66	0.44	0.61	0.45	0.51	0.54
	CDS	0.64	0.63	0.60	0.40	0.46	0.45	0.40	0.50	0.49	0.51
	SoB	0.71	0.93	0.69	0.78	0.63	0.74	0.63	0.68	0.71	<b>0.72</b>
	CM eigs	0.46	0.20	0.43	0.13	0.32	0.12	0.30	0.19	0.22	0.26
	Bag of Bonds	0.47	0.74	0.63	0.73	0.53	0.75	0.57	0.36	0.39	0.58
	Estate+CDS+SoB	0.70	0.94	0.79	0.76	0.69	0.69	0.65	0.60	0.70	<b>0.72</b>
Lasso	Estate	0.58	0.34	0.69	0.50	0.61	0.49	0.44	0.39	0.41	0.49
	CDS	0.54	0.68	0.55	0.41	0.48	0.41	0.43	0.53	0.48	0.50
	SoB	0.62	0.94	0.73	0.77	0.69	0.78	0.60	0.63	0.63	0.71
	CM eigs	0.41	0.26	0.38	0.12	0.31	0.13	0.21	0.24	0.23	0.25
	Bag of Bonds	0.43	0.66	0.50	0.75	0.55	0.75	0.45	0.33	0.39	0.53
	Estate+CDS+SoB	0.71	0.95	0.73	0.75	0.69	0.78	0.62	0.54	0.67	<b>0.72</b>
BR	Estate	0.66	0.38	0.70	0.54	0.64	0.51	0.59	0.36	0.48	0.54
	CDS	0.62	0.67	0.57	0.47	0.47	0.39	0.40	0.46	0.42	0.50
	SoB	0.67	0.93	0.73	0.73	0.75	0.77	0.59	0.55	0.70	0.71
	CM eigs	0.42	0.18	0.37	0.10	0.34	0.26	0.18	0.21	0.21	0.25
	Bag of Bonds	0.55	0.78	0.55	0.72	0.48	0.72	0.57	0.27	0.41	0.56
	Estate+CDS+SoB	0.77	0.95	0.74	0.77	0.66	0.75	0.65	0.59	0.60	<b>0.72</b>
GBoost	Estate	0.43	0.39	0.52	0.43	0.53	0.54	0.39	0.37	0.35	0.44
	CDS	0.49	0.47	0.43	0.48	0.46	0.56	0.39	0.49	0.43	0.47
	SoB	0.60	0.78	0.62	0.61	0.60	0.56	0.45	0.60	0.40	0.58
	CM eigs	0.23	0.42	0.13	0.15	0.07	0.02	0.10	0.27	0.27	0.18
	Bag of Bonds	0.66	0.75	0.58	0.73	0.56	0.70	0.49	0.55	0.56	0.62
	Estate+CDS+SoB	0.64	0.81	0.61	0.64	0.60	0.63	0.43	0.59	0.52	0.61
RF	Estate	0.47	0.41	0.39	0.50	0.62	0.56	0.42	0.35	0.41	0.46
	CDS	0.56	0.44	0.42	0.44	0.41	0.46	0.41	0.44	0.42	0.45
	SoB	0.56	0.79	0.55	0.63	0.55	0.60	0.41	0.52	0.41	0.56
	CM eigs	0.22	0.32	0.22	0.18	0.23	0.10	0.19	0.30	0.29	0.23
	Bag of Bonds	0.68	0.72	0.63	0.67	0.62	0.71	0.47	0.51	0.46	0.61
	Estate+CDS+SoB	0.63	0.79	0.64	0.62	0.56	0.68	0.53	0.52	0.49	0.60
kNN	Estate	0.42	0.43	0.46	0.52	0.44	0.47	0.49	0.36	0.39	0.44
	CDS	0.47	0.06	0.42	0.28	0.38	0.28	0.38	0.42	0.41	0.35
	SoB	0.49	0.65	0.46	0.62	0.46	0.58	0.46	0.48	0.53	0.53
	CM eigs	0.20	0.26	0.22	0.03	0.16	0.13	0.07	0.14	0.18	0.16
	Bag of Bonds	0.41	0.34	0.53	0.70	0.56	0.65	0.43	0.32	0.36	0.48
	Estate+CDS+SoB	0.54	0.50	0.55	0.58	0.44	0.54	0.58	0.40	0.45	0.51

TABLE V. Pearson correlation values in the test set for different combinations of target property, model and featurization. Hyperparameter optimization and 5-fold cross validation was used here.

		$\rho, \frac{\text{g}}{\text{cc}}$	$\Delta H_f^s, \frac{\text{kJ}}{\text{mol}}$	$E_c, \frac{\text{kJ}}{\text{cc}}$	$V_s, \frac{\text{km}}{\text{s}}$	$V_p, \frac{\text{km}}{\text{s}}$	$V_{\text{snd}}, \frac{\text{km}}{\text{s}}$	$P, \text{GPa}$	$T, \text{K}$	$\frac{\text{TNT}_{\text{equiv}}}{\text{cc}}$
KR	CDS	3.97	108.13	16.22	5.38	6.95	5.99	13.57	12.69	12.95
	SoB	2.97	59.83	11.11	3.60	4.96	4.04	11.07	13.13	10.27
SVR	CDS	3.55	107.53	15.24	4.55	6.35	5.22	12.09	11.97	14.88
	SoB	3.31	67.36	11.21	4.17	5.21	4.55	9.56	10.68	8.33
Ridge	CDS	3.85	127.68	16.80	6.13	7.13	6.80	14.87	13.10	13.44
	SoB	3.24	70.76	11.98	3.60	4.85	4.41	10.25	10.60	8.52
mean	CDS	5.67	106.41	24.78	8.17	8.46	8.91	21.56	18.40	17.60
	SoB	5.70	113.62	24.55	7.98	8.62	8.87	21.45	18.58	17.60

TABLE VI. Mean absolute percentage errors for some of the models and two of the best featurization. Note that these numbers are often inflated by the presence of small target values, and may not accurately reflect model accuracy.

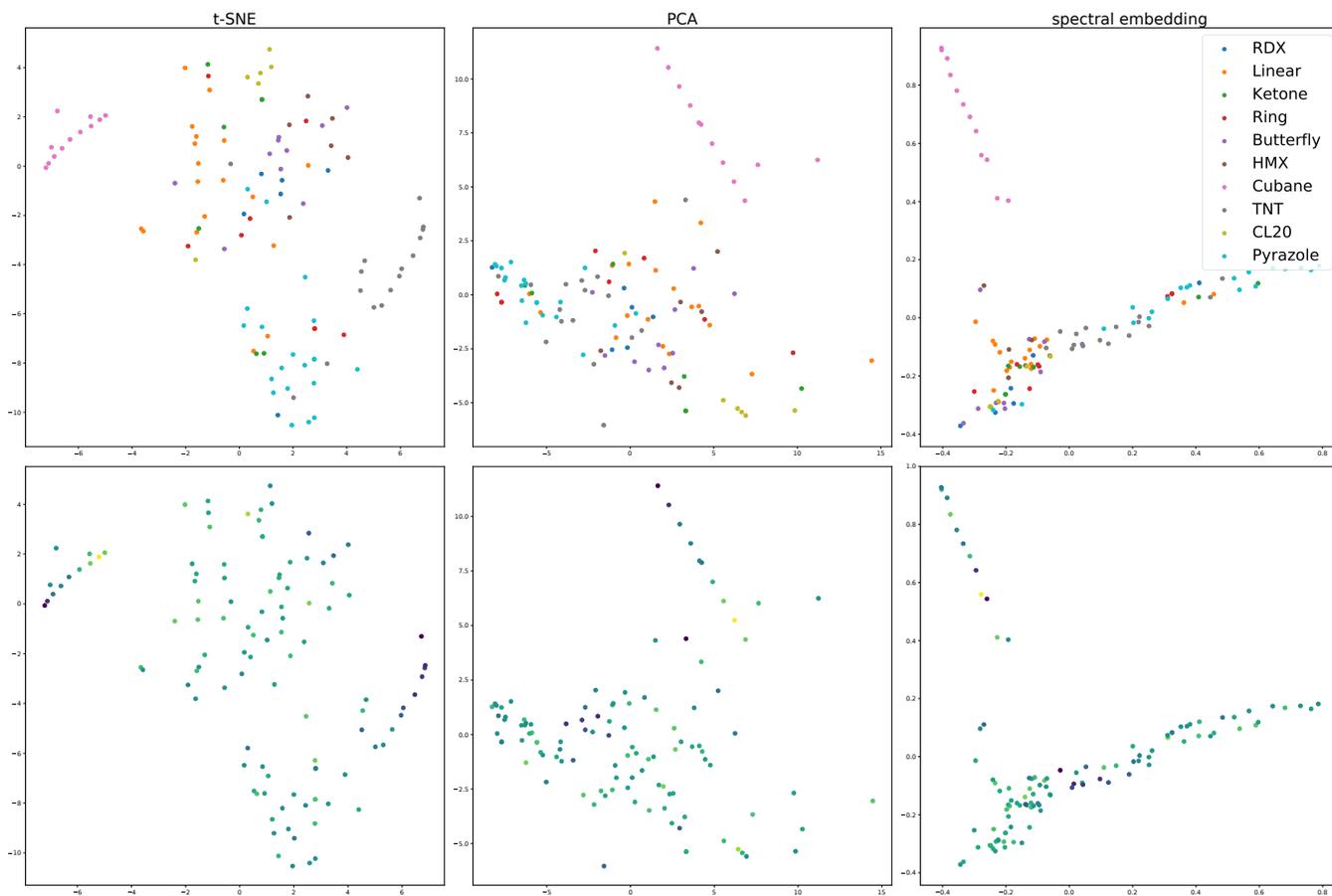


FIG. 3. Two dimensional embeddings of the Huang & Massa data using the sum over bonds featurization and three different embedding techniques - t-SNE, PCA, & spectral embedding. The upper figures show the different chemical groups while the bottom figures show the explosive energy (dark = low, light = high).