Supporting Information for "Chemical descriptors for a large-scale study on drop-weight impact sensitivity of high explosives"

Frank W. Marrs,* Jack V. Davis, Alexandra C. Burch, Geoffrey W. Brown, Nicholas Lease, Patricia L. Huestis, Marc J. Cawkwell, and Virginia W. Manner*

Los Alamos National Laboratory, Los Alamos, New Mexico, USA, 87545

E-mail: fmarrs3@lanl.gov; vwmanner@lanl.gov

Correlations of predictors

We examined the Spearman correlation between log E_{50} and all numerical predictors. For predictors with statistically significant correlation with log E_{50}, we plotted the correlations between each pair of predictors and with log E_{50} in Figure S1. A summary of the correlations between each descriptor and log E_{50} are given in Table S1.

Hyperparameter tuning

To examine the effects of hyperparameters on the VSURF and random forest regression methods, we reran the first of the ten cross validations at a grid of hyperparmeters. Due to the computational demand of this procedure, we only reran the random forest method,

which forms the basis for both random forests and VSURF. We varied the size of the random pool of descriptors for each tree (`mtry` in the `randomForest` input) on the range $p[9^{-1}, 3^{-3/2}, 3^{-1}, 3^{-1/2}, 1]$, where p is the total number of decsriptors. This grid is centered on the default value for `mtry`, $p/3$, and equally spaced by multiplicative factors of $\sqrt{3}$. We also varied the minimal size of the terminal level of tree, `nodesize` in the `randomForest` input. The default value for `nodesize` is 5, and we explored values $[1, 2, 5, 10, 20]$. We kept the number of trees constant at a large but manageable value (2,000), which is consistent with recommendations in the literature.[1,2]

We reran the first of the ten cross validations at all combinations of `mtry` and `nodesize`, for both the means data, and the complete data set, including repeats. Then, we evaluated performance of random forest using $\overline{RMSE}$. For the complete data set, including repeats, the default settings of `mtry` and `nodesize` performed best. The trends of $\overline{RMSE}$ with `mtry` and `nodesize` are depicted in Figure S2, and values of $p/3\sqrt{3}$ and 1, respectively, performed best. However, we notice that the default values most consistently give low values of $\overline{RMSE}$.

Since tuning on the means data produced slightly lower values of the hyperparameters, we reran the remaining nine random cross validations at these settings, for both the random forest and VSURF methods. The resulting $\overline{RMSE}$ values, averaged across all ten random cross validations, was about 0.304 for both random forest and VSURF methods. This value of $\overline{RMSE}$ is consistent with the value of $\overline{RMSE}$ found when using the default tuning parameters of random forest and VSURF methods. Hence, we conclude that random forest and VSURF perform sufficiently well without changing the values of `mtry` and `nodesize` from the defaults.

Computation of approximate R^2

To compare fits of a our prediction method on the current data set to the performance of other methods on (potentially) other data sets, we wished to compute the coefficient of

determination (R^2). The R^2 value of predictor $\hat{y}$ for response y is typically defined as

$$R^2(\hat{y}) = 1 - \frac{rmse(\hat{y}, y)^2}{rmse(\bar{y}, y)^2}, \tag{S1}$$

where $rmse(a, b)$ represents the root mean square error between vectors a and b, and $\bar{y}$ represents a vector of the same length of y, with each entry occupied by the arithmetic mean of y (see Ref.[3] for example). In the main paper, we define a new root mean square error metric, $\overline{RMSE}$, which is meant to balance predictor performance between the complete and means data. However, $\overline{RMSE}$ represents the mean root mean square error on data sets with generally non-integer numbers of repeats (depending on the value of η). Hence, we define a pseudo R^2 that uses the $\overline{RMSE}$ metric. For a predictor $\hat{y}$ with metric $\overline{RMSE}(\hat{y}, y)$, we define the approximate R^2 as

$$\bar{R}^2(\hat{y}) = 1 - \frac{\overline{RMSE}(\hat{y}, y)^2}{\overline{RMSE}(\bar{y}, y)^2}, \tag{S2}$$

where $\overline{RMSE}(\bar{y}, y)$ is computed in the same way as $\overline{RMSE}(\hat{y}, y)$, but replacing the predictor $\hat{y}$ with a vector where every entry is the arithmetic mean of y. We report the out-of-sample $\bar{R}^2(\hat{y})$ for the selected prediction method in the main paper. We find $\overline{RMSE}(\bar{y}, y) \approx 0.502$.

References

(1) Scornet, E. Tuning Parameters in Random Forests. *ESAIM: Proceedings and Surveys* **2017**, *60*, 144–162.

(2) Probst, P.; Boulesteix, A.-L. To Tune or Not to Tune the Number of Trees in Random Forest. *J. Mach. Learn. Res.* **2017**, *18*, 6673–6690.

(3) Draper, N. R.; Smith, H. *Applied regression analysis*; John Wiley & Sons, 1998; Vol. 326.

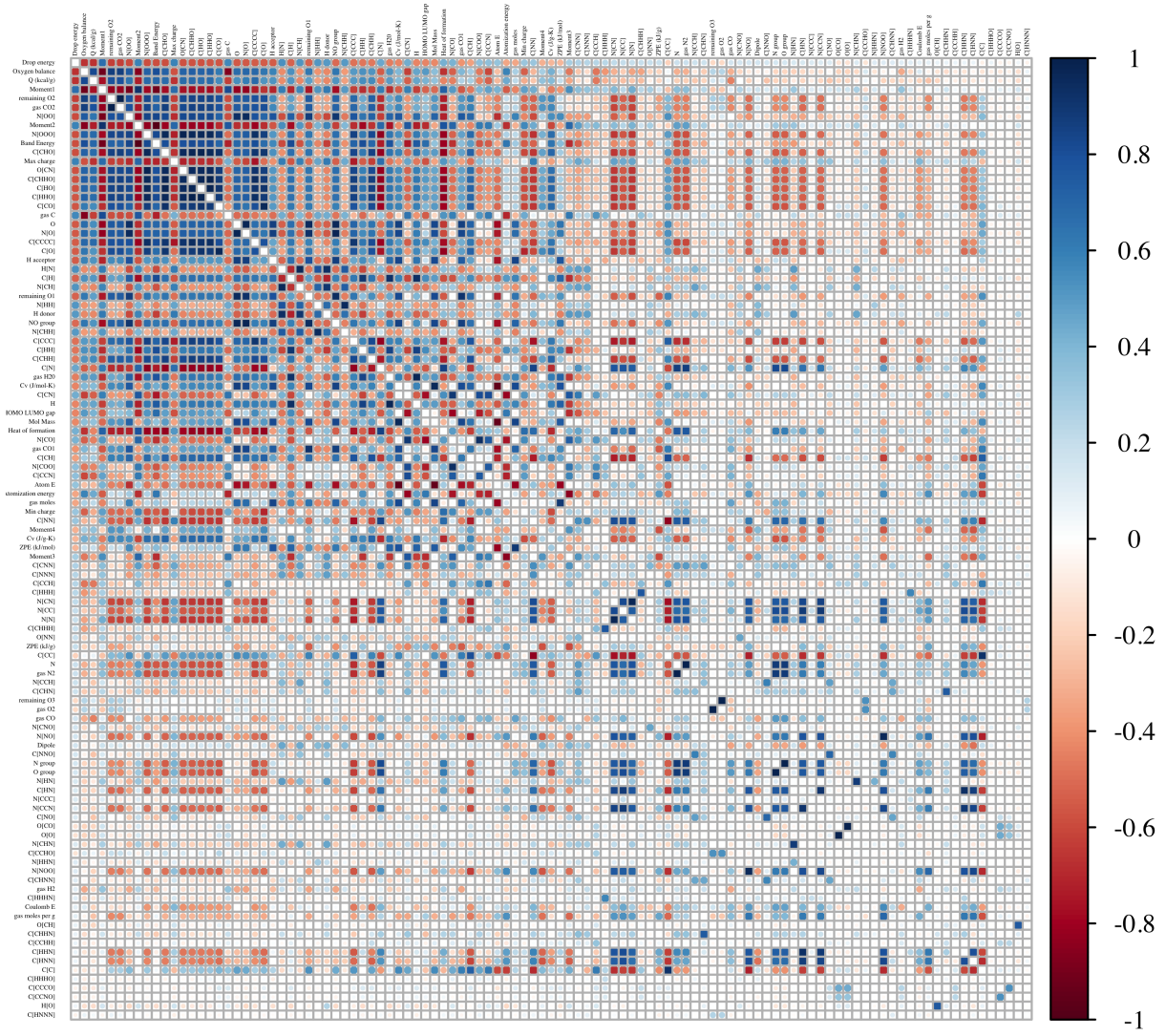

Figure S1: Spearman correlations of significantly correlated predictors with log E_{50}.

Table S1: Spearman correlation of predictors with drop energy, on the log scale.

	Spearman correlation
Oxygen balance	-0.668
Q (kcal/g)	-0.668
Moment1	0.664
remaining O2	-0.623
gas CO2	-0.622
N[OO]	-0.596
Moment2	0.589
N[OOO]	-0.586
C[CHO]	-0.583
Band Energy	-0.582
O[CN]	-0.577
Max charge	0.577
C[CHHO]	-0.575
C[HO]	-0.571
C[CO]	-0.568
C[HHO]	-0.567
gas C	0.560
O	-0.556
N[O]	-0.550
C[CCCC]	-0.548
C[O]	-0.545
H acceptor	-0.543
H[N]	0.528
C[H]	-0.527
N[CH]	0.522
remaining O1	-0.518
N[HH]	0.507
NO group	-0.499
H donor	0.498
N[CHH]	0.493
C[CCC]	-0.485
C[HH]	-0.473
C[CHH]	-0.463
gas H20	-0.452
C[N]	0.451
Cv (J/mol-K)	-0.439
C[CN]	0.438
H	-0.436
HOMO LUMO gap	-0.432
Mol Mass	-0.422
Heat of formation	0.419
N[CO]	0.415
gas CO1	-0.389
N[COO]	0.358
C[CH]	-0.356
C[CCN]	0.354
Atom E	0.340
Atomization energy	-0.338
gas moles	-0.335
Min charge	0.326
C[NN]	0.319
Moment4	-0.301
Cv (J/g-K)	-0.287
ZPE (kJ/mol)	-0.284
Moment3	0.280
C[CNN]	0.263
C[CCH]	0.250
C[HHH]	0.245
C[NNN]	0.243
N[CN]	0.216
N[CC]	0.207
C[CHHH]	0.204
N[N]	0.196
O[NN]	0.195
ZPE (kJ/g)	0.193
C[CC]	-0.189
N	0.176
gas N2	0.176
N[CCH]	0.175
C[CHN]	0.169
gas CO	0.165
remaining O3	-0.162
gas O2	-0.162
N[CNO]	0.157
N[NO]	0.156
Dipole	0.150
C[NNO]	0.147
N group	0.139
O group	0.139
N[HN]	0.136
C[HN]	0.129
N[CCC]	0.127
N[CCN]	0.123
C[NO]	0.122
O[CO]	-0.121
O[O]	-0.121
C[CCHO]	-0.119
N[CHN]	0.118
C[CHNN]	0.117
N[NOO]	0.113
N[HHN]	0.111
C[HHHN]	0.107
gas H2	0.105
Coulomb E	0.100
gas moles per g	0.093
O[CH]	0.073
C[CHHN]	0.069
C[CCHH]	0.068
C[HHN]	0.067
C[HNN]	0.067
C[HHHO]	0.064
C[C]	-0.061
C[CCCO]	-0.060
C[CCNO]	-0.059
H[O]	0.055

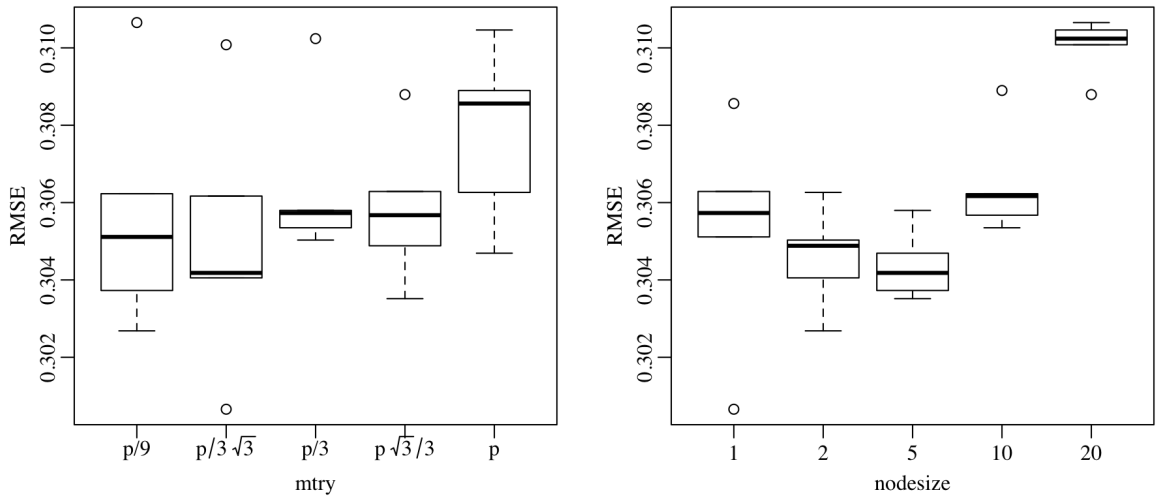

Figure S2: $\overline{RMSE}$ in first cross validation of random forest method, varying tuning parameters `mtry` and `nodesize`.