U.S. ARMY COMBAT CAPABILITIES DEVELOPMENT COMMAND – ARMAMENTS CENTER

Quantitative Structure-Activity/Property Relationship Models for Secondary Explosive Compounds

Carson Britt

Research Scientist I

CS Squared LLC

Approved For Public Release

BOTTOM LINE UP FRONT

- **Developed models to rapidly predict the properties of secondary explosive compounds from molecular structure alone**
  - Quantitative structure-activity/property relationships (QSAR/QSPRs) use regression or classification algorithms to relate structural features to properties
  - Density, melting temperature, vapor pressure, impact sensitivity, oral rat LD50 (LD50), 48 hour *Daphnia magna* LC50 (LC50DM), 96 hour fathead minnow LC50 (LC50FM), 40 hour *Tetrahymena pyriformis* IGC50 (IGC50), and bioaccumulation factor (BCF)

- **Statistically validated QSAR/QSPR's for accuracy, applicability domain**
  - Accuracy quantified by (repeated) nested five-fold cross validation
    - Root mean square error, mean absolute error, median error, Pearson correlation coefficient
  - Applicability domain by Tanimoto similarity to training set, predicted value
    - Gauge reliability of property predictions for new and unknown compounds

- **LiveDesign platform hosts QSAR/QSPR models, other tools**
  - Models available in easy to use, online interface

SCOPE AND AGENDA

- **Objective: using a dataset of "ground truth" data, establish correlations between the molecular structure and physical properties of known compounds. Use these correlations to predict the physical properties of new compounds from only their molecular structure.**

- **Requirements for a QSAR/QSPR model:**
  - Dataset of molecular structures and their properties
  - Routine to represent molecular structures in a machine understandable way
  - Algorithm to elucidate correlations between molecular representations and the corresponding physical properties
  - Criteria to evaluate the accuracy of correlations when applied to new compounds

- **Agenda:**
  - Background machine learning & QSAR/QSPR concepts and terminology
  - Technical aspects of QSAR/QSPR model development
  - Evaluation of accuracy and applicability of the developed models
  - Use of the models in new compound discovery workflow

Quantitative Structure-Activity/Property Relationship Models for Secondary Explosive Compounds

Fundamental machine learning & QSAR/QSPR concepts and terminology

TYPES OF MODELS

- **Minimizing error with supervised regression models**
  - Model selection – Linear, logistic, k-nearest neighbors, kernel ridge regression (KRR), support vector machines (SVM), decision trees, neural networks
  - Regularization – L1 regularization, L2 regularization, dropout
  - Model ensembling – combining multiple models to reduce bias and variance

- **For small structured datasets ensemble decision tree models are consistently among the best performing models**
  - Exhibit excellent predictivity
  - High level of interpretability
  - Relatively computationally inexpensive
  - Hyperparameters easily adjusted to needs of a given dataset

- **Gradient boosting algorithms**
  - CatBoost
  - LightGBM
  - XGBoost

GRADIENT BOOSTING

Input Dataset

Train decision tree

Predict on training data
Calculate residual prediction error
Use residual error as new training input

Train decision tree

Predict on training data
Calculate residual prediction error
Use residual error as new training input

Train decision tree

Predict on training data
Calculate residual prediction error
Use residual error as new training input

Sum outputs of each tree

Prediction

Quantitative Structure-Activity/Property Relationship Models for Secondary Explosive Compounds

Model Development

QSAR/QSPR MODEL DEVELOPMENT

- **A QSAR/QSPR model needs:**
  - Dataset of molecular structures and their physical properties
    - Molecules represented via simplified molecular-input line-entry system (SMILES)
  - Routine to represent molecular structures in a machine understandable way
    - RDKit – cheminformatics python library
  - Algorithm to elucidate correlations between molecular representations and the corresponding physical properties
    - XGBoost, dask, scikit-learn, hyperopt – machine learning python libraries
  - Criteria to evaluate the accuracy of correlations when applied to new compounds
    - scikit-learn, SciPy, NumPy, pandas – math and data visualization python libraries

DATASETS

Physical Property	# Compounds	Type Compounds	Sources
Density	15,435	Small molecules, drug-likes, energetics	ochem.eu, EMD
Melting temperature	3,171	Small molecules, drug-likes, energetics	DPG, EMD
Vapor pressure	3,268	Small molecules, pesticides, drug-likes, energetics	ochem.eu, EMD
Impact sensitivity	308	Energetics	Didier Mathieu
Oral rat LD50	7,294	Small molecules, pesticides, drug-likes	EPA T.E.S.T.
48 hour *Daphnia magna* LC50	353	Small molecules, pesticides, drug-likes	EPA T.E.S.T.
96 hour fathead minnow LC50	823	Small molecules, pesticides, drug-likes	EPA T.E.S.T.
40 hour *Tetrahymena pyriformis* IGC50	1,792	Small molecules, pesticides, drug-likes	EPA T.E.S.T.
Bioaccumulation factor	672	Small molecules, pesticides, drug-likes	EPA T.E.S.T.

DATASETS

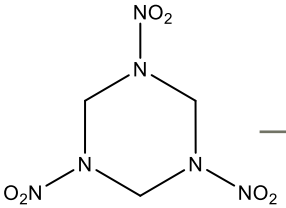
- **Sources of data**
  - Didier Mathieu impact sensitivity dataset (Didier Mathieu)
    - Mathieu, D., "Sensitivity of Energetic Materials: Theoretical Relationships to Detonation Performance and Molecular Structure“, Ind. Eng. Chem. Res., vol 56, no. 29, 2017, pp. 8191-8201
  - Jean-Claude Bradley double plus good melting point dataset (DPG)
    - Bradley, J-C., Lang, A., Williams, A.J., “Jean-Claude Bradley double plus good (highly curated and validated) melting point dataset”, 2019.
  - Energetic Materials Database (EMD)
    - Britt, C., and Hrudka, J., “Energetic Materials Database”, CS Squared LLC, 2019.
  - EPA Toxicology Estimation Software Tool (EPA T.E.S.T.)
    - Martin, T., “User’s Guide for T.E.S.T. (Toxicity Estimation Software Tool), U.S EPA/National Risk Management Research”, 2016.
  - Online Chemical Modeling Environment (ochem.eu)
    - Tetko, V. I., *et al*., “Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information”, J. Comput. Aided Mol. Des., vol. 25, no. 6, 2011, pp. 533-554
- **Datasets curated prior to use in model building**
  - No duplicates, salts, charged species, mixtures
  - No compounds containing elements besides C, H, N, O, B, P, S, F, Cl, Br, and I

NUMERICALLY REPRESENTING MOLECULES

- **Sample rows from melting point dataset:**

Compound	SMILES String	Melting Point (°C)
HMX	[O-][N+](=O)N1CN(CN(CN(C1)[N+](=O)[O-])[N+](=O)[O-])[N+](=O)[O-]	275
RDX	[O-][N+](=O)N1CN(CN(C1)[N+](=O)[O-])[N+](=O)[O-]	204
LLM-200	Nc1nonc1/N=[N+](\c1nonc1N)/O	182
TNT	[O-][N+](=O)c1cc([N+](=O)[O-])c(c(c1)[N+](=O)[O-])C	81
BODN	O=[N+](OCC1=NC(C2=NOC(CO[N+]([O-])=O)=N2)=NO1)[O-]	82

  - Convert SMILES to numeric format that can be used by XGBoost



CC1=C([N+]([O-])=O)C=C([N+]([O-])=O)C=C1[N+]([O-])=O

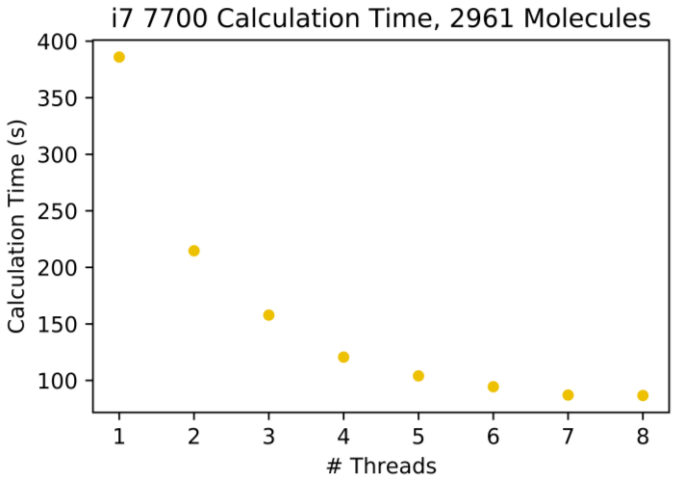
Atom counts
Functional group counts
Estate fingerprint
Etc.

- **Use RDKit to calculate 1447 molecular features**
  - Zero dimensional: Atom counts, atom ratios, oxygen balance, etc.
  - One dimensional: Bond counts, bond ratios, information indices, etc.
  - Two dimensional: Functional group fragments, fingerprints, etc.
  - Three dimensional: WHIM, geometric distances, inertial, etc.
- **Property and molecular structure now represented numerically:**

[275, 1.959, 0.476, 0.4, 0, … 12.453]

MOLECULAR FEATURE CALCULATION & FEATURE SELECTION

- **Calculation of molecular features not computationally expensive**
  - RDKit python library well optimized for an interpreted language
  - Calculation of molecular features is a trivially parallelized task



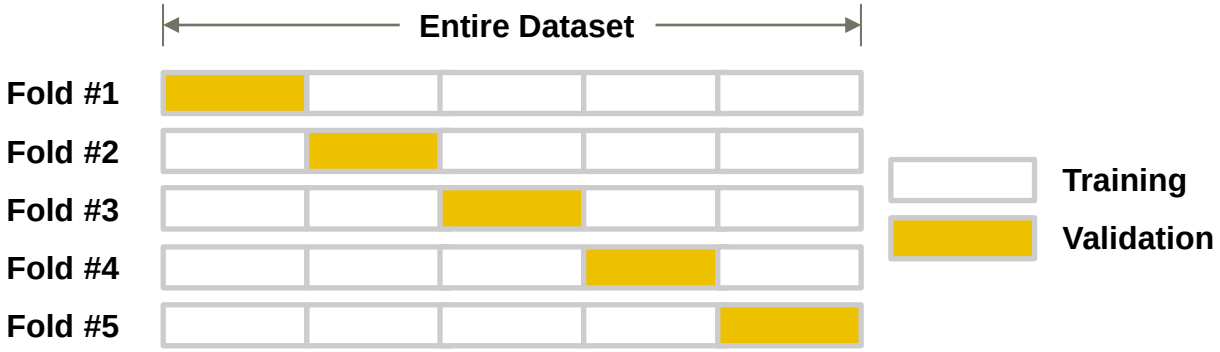
Number Threads	Calculation time: Intel i7 7700 (s)	Calculation time: AMD 2950x (s)
1	386	496
2	215	254
4	121	140
8	87	73
16	N/A	48
32	N/A	42

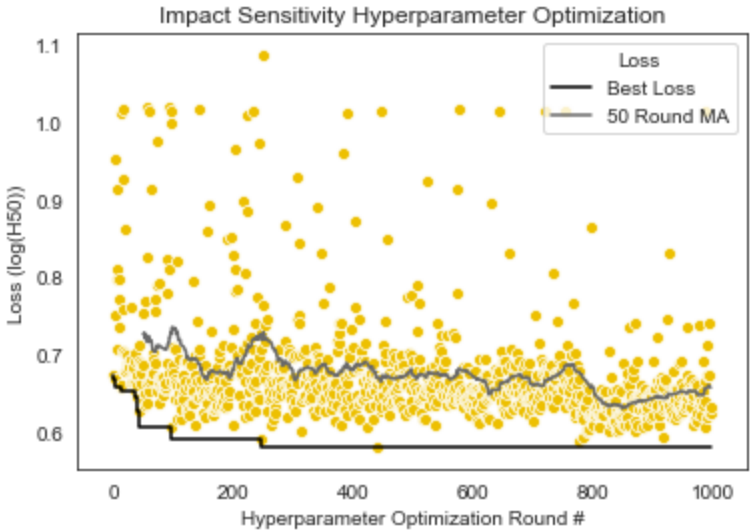
- **Feature selection using Boruta method**
  - Feature selection can improve model accuracy, interpretability, decrease wall-time
  - Duplicate each feature, y-scramble the rows creating “shadow features”
  - Use shadow features and original features to train a random forest model
  - Rank feature importance within random forest model, reject original features that perform worse than specified percentage of shadow features
  - Features that perform better than random noise are retained

Kursa, B. M., Jankowski, A., Rudnicki, W. R., “Boruta – A system for feature selection” Fundamenta Informaticae, vol. 101, 2010, pp. 271–285, doi: 10.3233/FI-2010-288

HYPERPARAMETER OPTIMIZATION

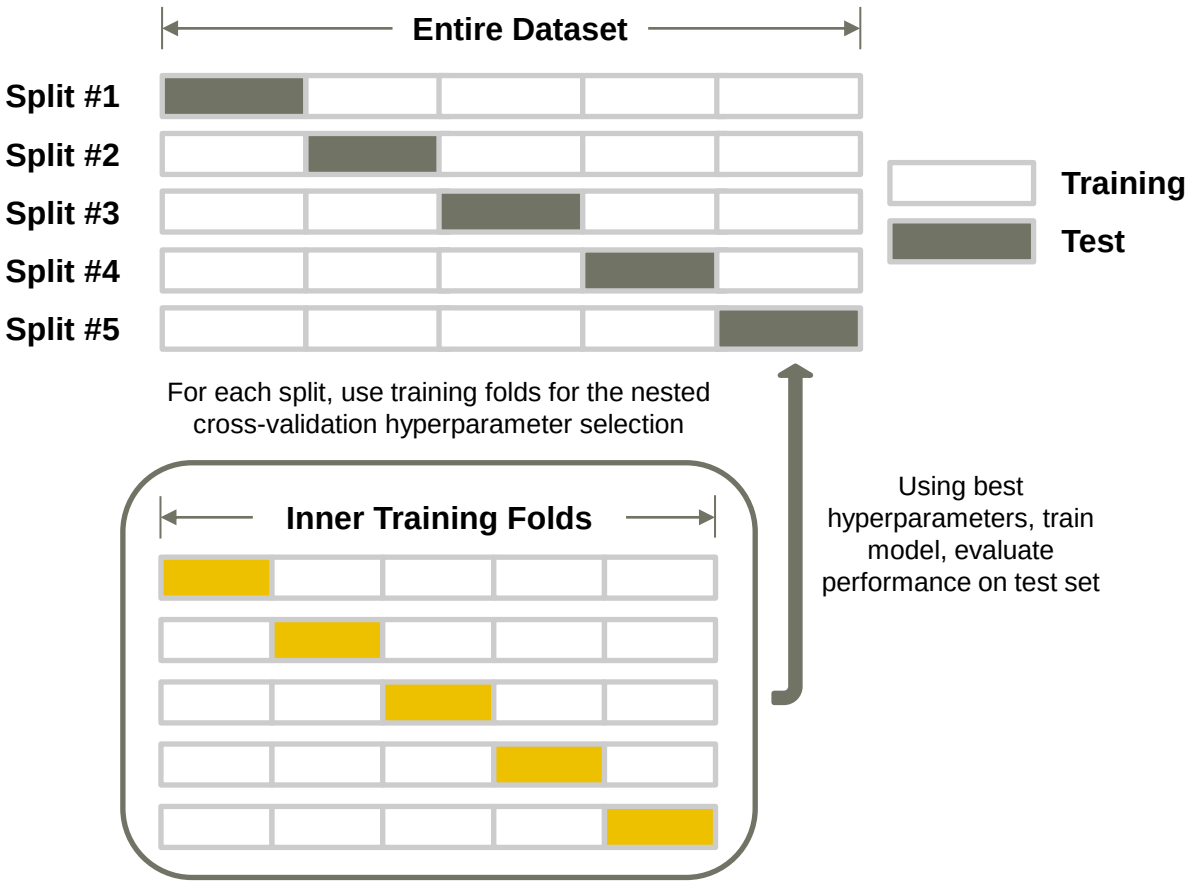
- **Bayesian hyperparameter tuning**
  - Hyperparameters are specifications for how a model is constructed, i.e., number of boosting rounds, max depth, etc.
  - Generate distribution of possible hyperparameters
  - Test a set of hyperparameters with cross-validation, use test results to inform the selection of the next set of hyperparameters to try
  - Converge toward optimal set of hyperparameters





Bergestra, J., Yamins, D., Cox, D. D., "Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures" ICML'13: Proceedings of the 30th International Conference on International Conference on Machine Learning, vol. 28, 2013, pp. 115-123

NESTED K-FOLD CROSS VALIDATION



- **Algorithm:**
  - Divide data into k folds
  - Withhold each fold once as a test fold, use remaining folds for feature selection, hyperparameter selection, and training
  - Divide training folds again into k equal inner training folds, select features
  - Hyperparameters selected with cross validation
  - Use hyperparameters to train model with all inner training folds
  - Evaluate model on the withheld test set
  - Repeat for each split
- **Repeat algorithm**

Quantitative Structure-Activity/Property Relationship Models for Secondary Explosive Compounds

Model Evaluation

QSAR/QSPR STATISTICAL PROPERTIES

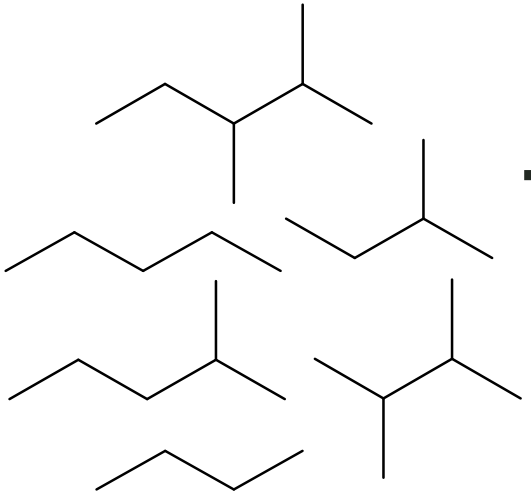
Property	MAE	RMSE	Median	Q-squared
Density (g/cm³)*	0.0214	0.0364	0.0119	0.988
Melting temperature (°C)*	25.5	34.68	19.34	0.873
Vapor pressure (Log10(mmg))*	0.531	0.874	0.271	0.942
Impact sensitivity (ln(H50))	0.458	0.616	0.335	0.629
LD50 (-Log10(mol/kg))*	0.420	0.573	0.317	0.638
LC50DM (-Log10(mol/L))	0.795	1.065	0.594	0.618
FC50FM (-Log10(mol/L))	0.593	0.827	0.410	0.692
IGC50 (-Log10(mol/L))	0.321	0.455	0.223	0.812
BCF (Log10)	0.466	0.625	0.342	0.791

* Results from only one round of external CV

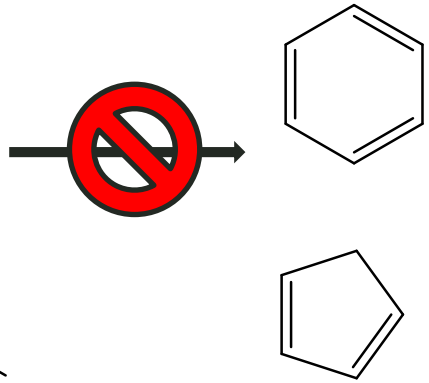
APPLICABILITY DOMAIN

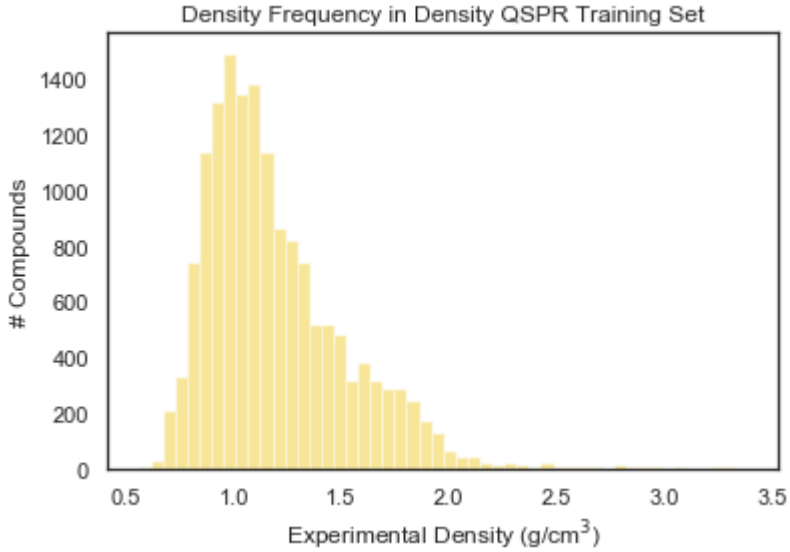
- **Physical properties of some molecules will be better predicted than others, quantify confidence in prediction for specific molecule**
  - Applicability domain is the chemical space where models will give good predictions
  - Predictions for molecules that are similar to training set molecules will be better
  - Numeric value of predicted property can also be used to gauge prediction accuracy

Training set

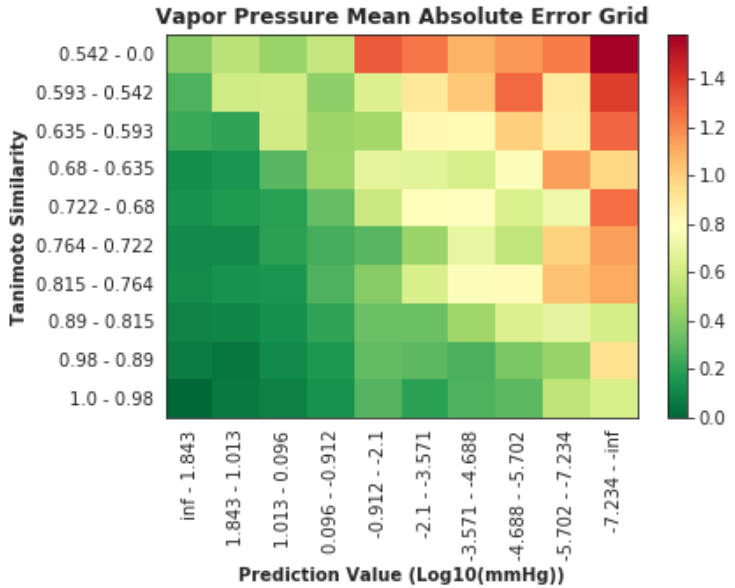


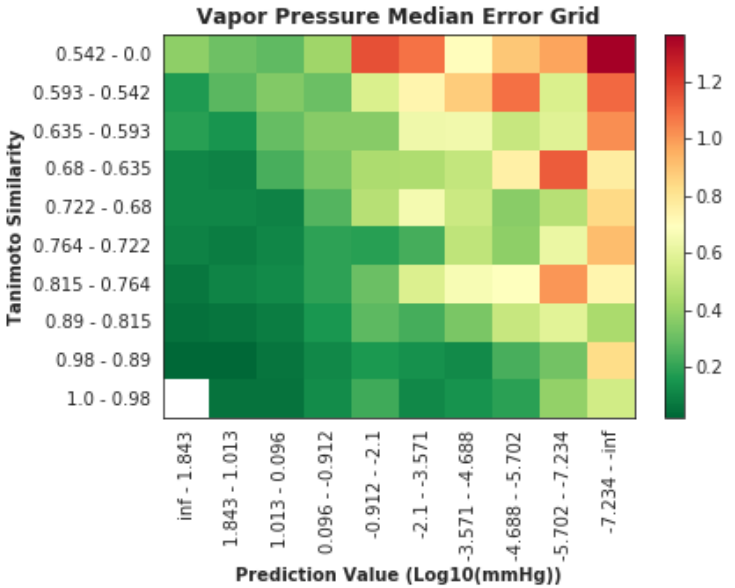
Test set

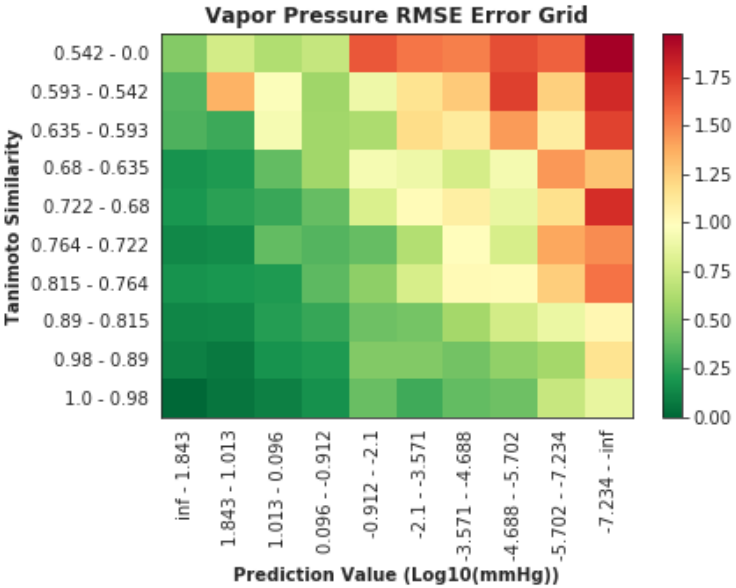


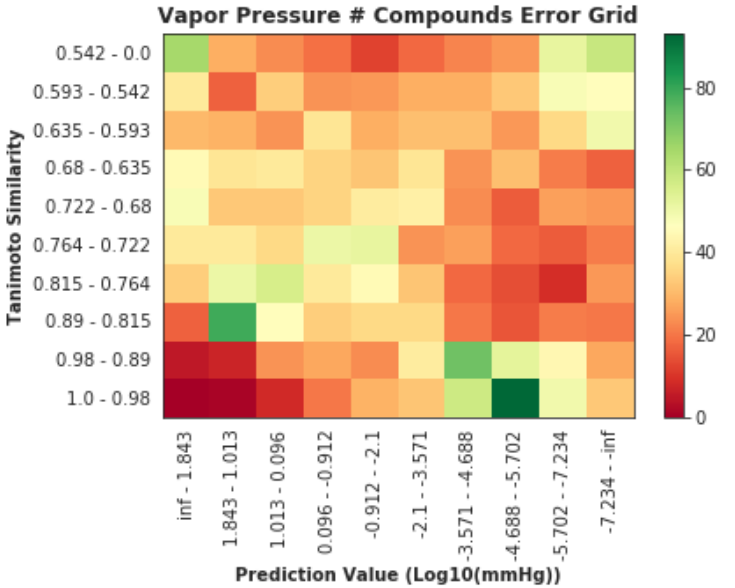


APPLICABILITY DOMAIN







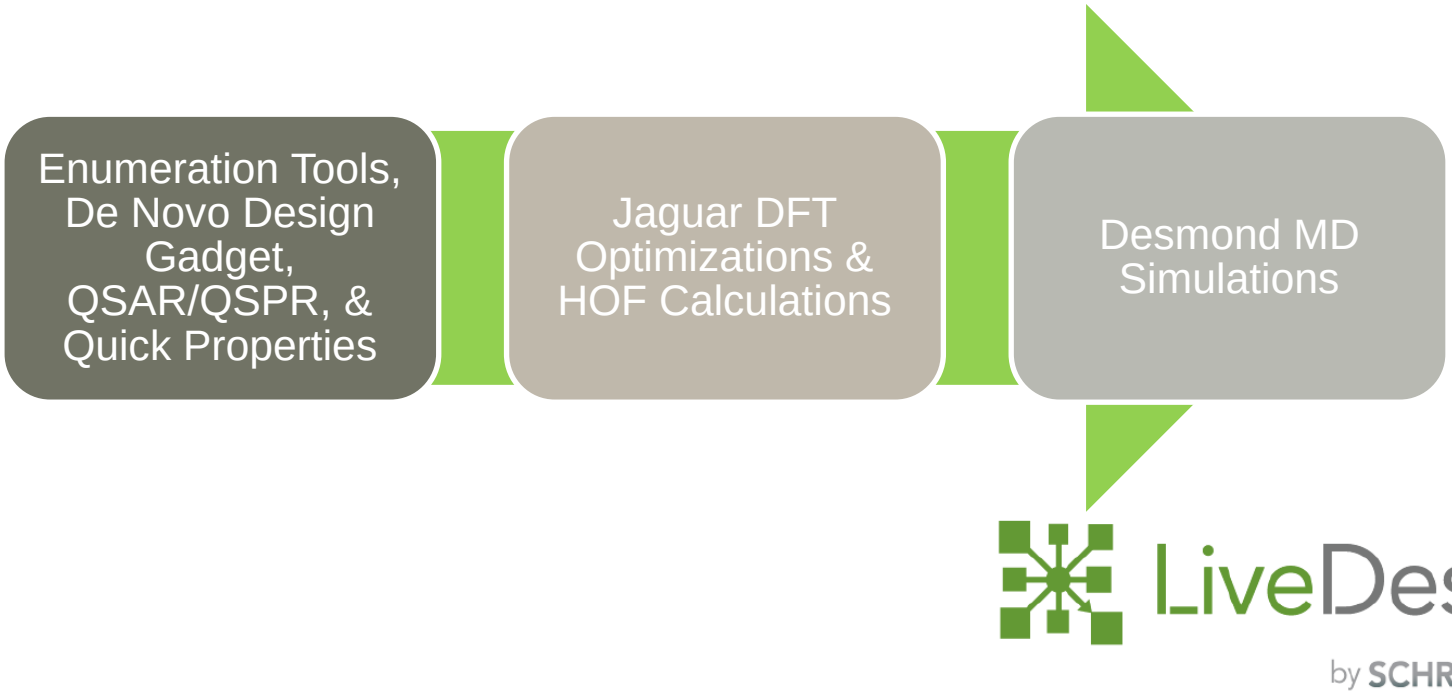


Quantitative Structure-Activity/Property Relationship Models for Secondary Explosive Compounds

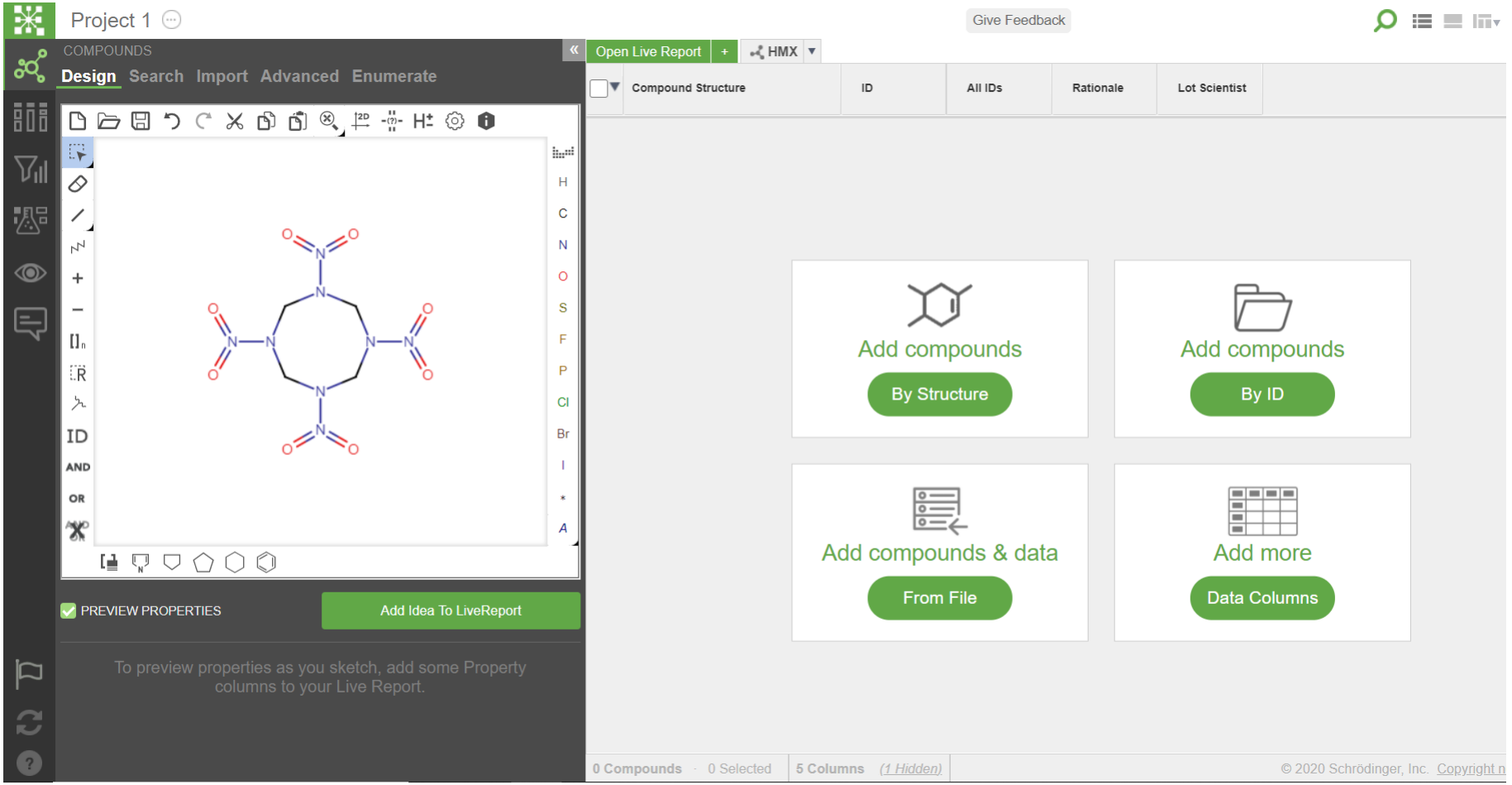
Implementation in LiveDesign

LIVEDESIGN: WEB-BASED COLLABORATION

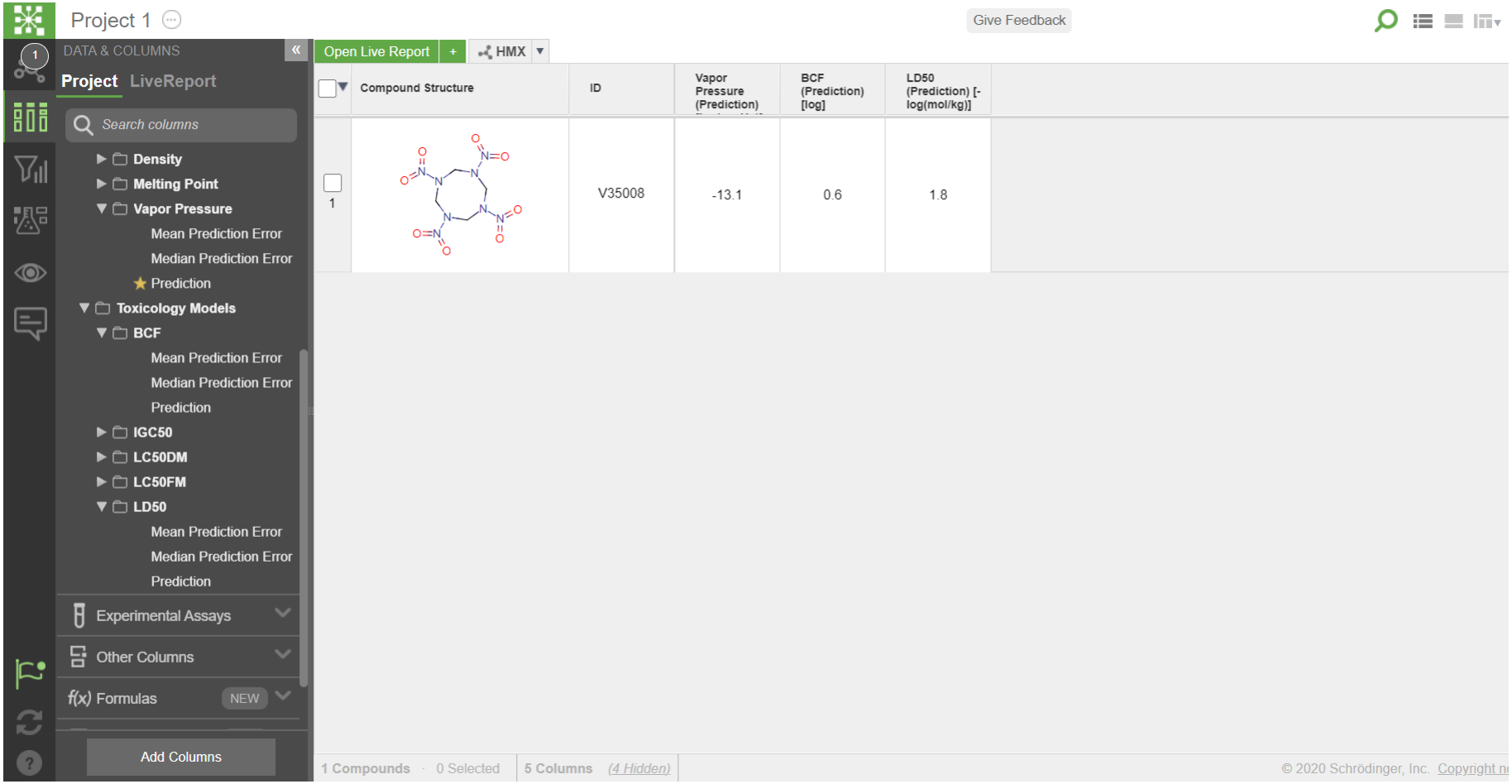
- **LiveDesign by Schrödinger**
  - An online informatics platform that allows teams to collaborate, design, and experiment in one centralized location
  - Enables teams of computational and synthetic chemists and engineers to work together and share results on one platform
  - Add data, visualize structures, run calculations with a single click
  - Keeps data and files in one place that can be easily searched and accessed



LIVEDESIGN: ADDING COMPOUNDS



LIVEDESIGN: MAKING PREDICTIONS



RECAP

- **Developed models to predict the properties of secondary explosive compounds from molecular structure alone**
  - Fast molecular descriptor calculation, feature selection with Boruta method
  - Bayesian hyperparameter optimization
  - Models trained in parallel using GPU's
  - Models developed for density, melting temperature, vapor pressure, impact sensitivity, IGC50, LD50, LC50DM, LC50FM, and BCF
- **Statistically validated QSAR/QSPR's for accuracy, applicability domain**
  - Accuracy quantified by (repeated) nested five-fold cross validation
    - Root mean square error, mean absolute error, median error, Pearson correlation coefficient
  - Applicability domain by Tanimoto similarity to training set, predicted value
    - Gauge reliability of property predictions for new and unknown compounds
- **LiveDesign platform hosts QSAR/QSPR models, other tools**
  - Models available in easy to use, online interface
- **Questions?**
  - carson.britt@csssquaredllc.com

ACKNOWLEDGEMENTS

- **This work was funded by the following Prime Contract/Subcontract:**
  - Prime Contract No. W911NF-15-D-0014 Task/Delivery Order No. W911NF18F0061 from the US Army Contacting Command
  - Leidos Subcontract No. P010220377
- **The presenter would like to acknowledge the support and guidance of the following individuals:**
  - Mr. Anthony Di Stasio, Deputy Director, DPA Title III Program, Industrial Policy, OUSD (A&S)
  - Dr. Kimberly Spangler, Program Manager, Joint DoD/DOE Munitions Program OUSD (R&E)
  - Dr. Paritosh Dave, Senior Program Manager, Leidos