

AVT-340 Research Workshop on Preparation and Characterization of Energetic Materials

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

Carson Britt, CS Squared LLC, USA

10 February 2021





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

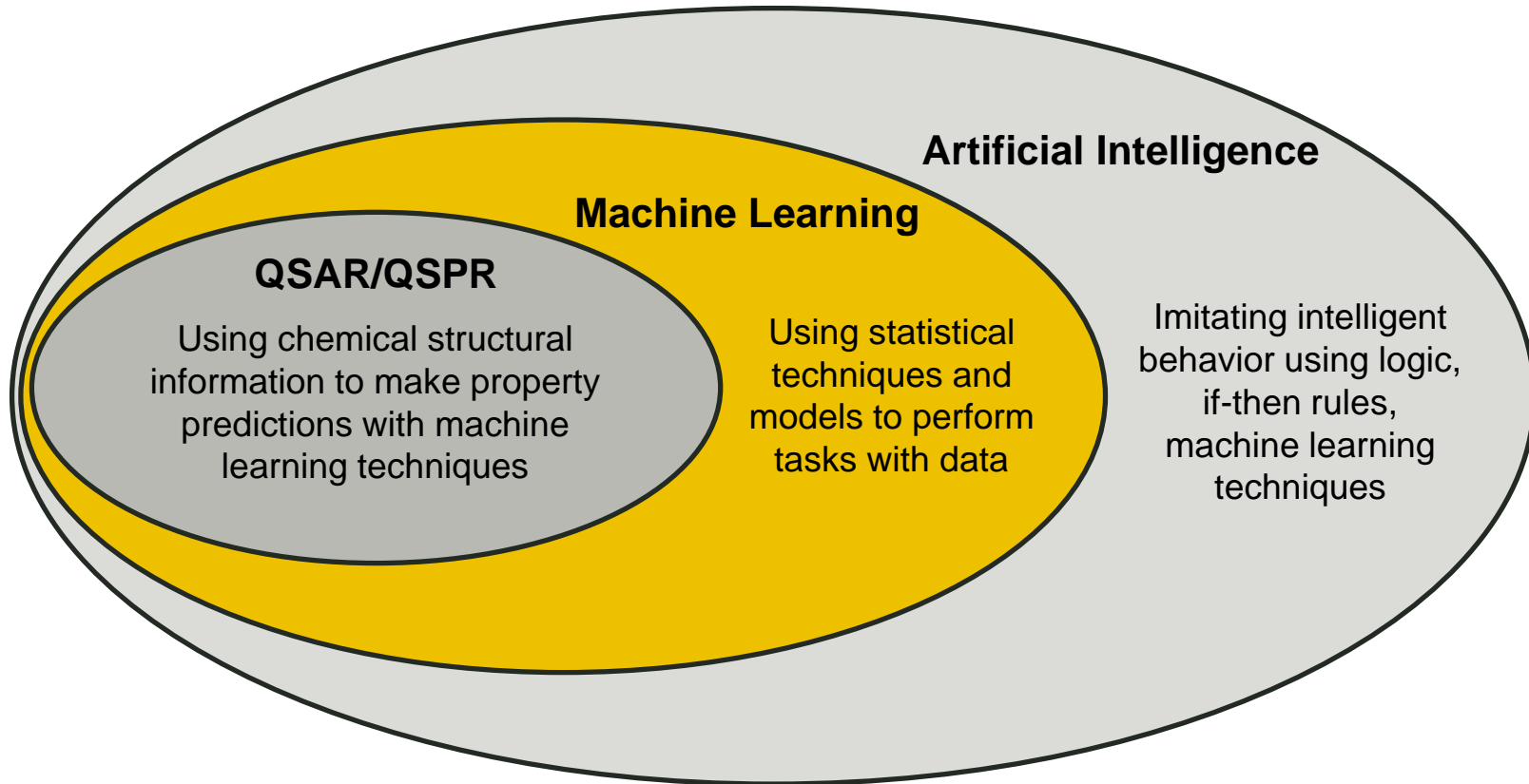


Quantitative Structure-Activity/Property Relationship Models for Secondary Explosive Compounds

Fundamental machine learning & QSAR/QSPR concepts and terminology



OVERVIEW OF TERMS



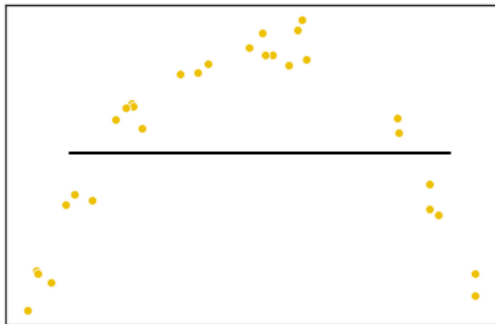
- **Usage of terms varies, but these are common definitions**
- **Subsets of machine learning – unsupervised/supervised learning**
 - Unsupervised learning – finding patterns or groupings within unlabeled data
 - Clustering unlabeled photos based on the content of each photo
 - Supervised learning – modeling using labeled data with a known “ground truth”
 - Predicting the selling price of homes based on data



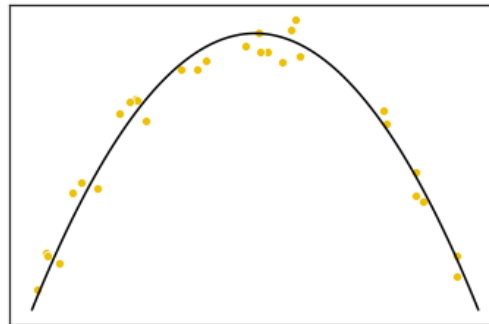
ERROR IN MACHINE LEARNING



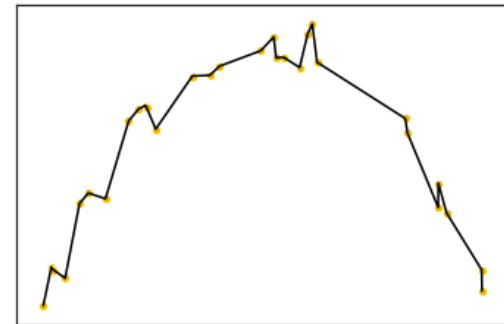
- **Modeling error – irreducible and reducible**
 - Irreducible error is error inherent in the “ground truth” input data
 - Irreducible because model alterations cannot resolve this (“garbage in, garbage out”)
 - Reducible error is error attributable to model
 - Bias – difference between model prediction and “ground truth” value
 - Variance – variability of model over perturbations of initial conditions, or “ground truth” data
- **Need to balance bias and variance to create a model that minimizes prediction error for given data while still generalizing well to new data**



High bias, low variance
Underfitting



Low bias, low variance
Good balance



Low bias, high variance
Overfitting



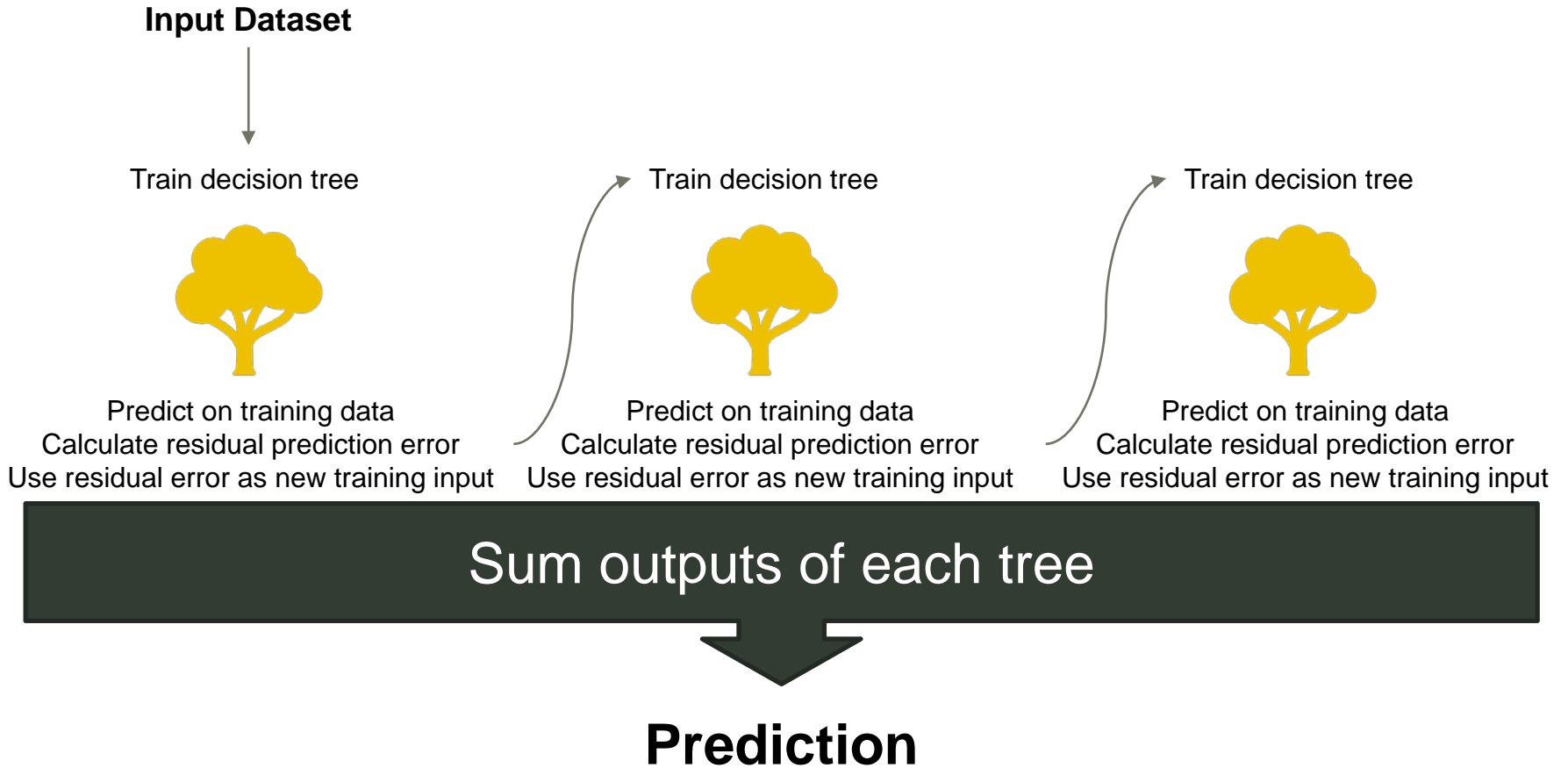
TYPES OF MODELS



- **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





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



Open-Source Cheminformatics
and Machine Learning





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 <i>Daphnia magna</i> 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 <i>Tetrahymena pyriformis</i> 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



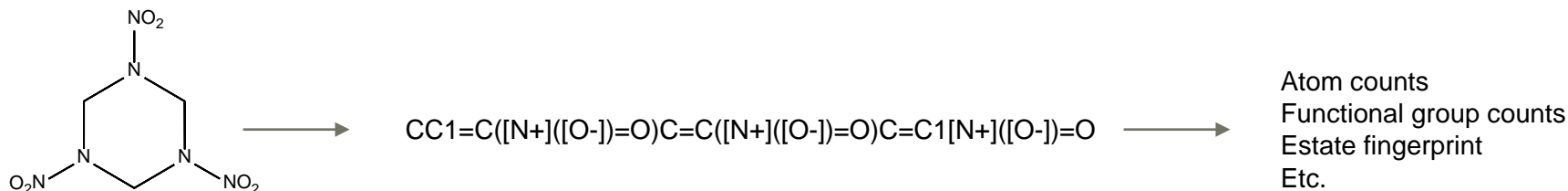
NUMERICALLY REPRESENTING MOLECULES



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

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

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



- **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



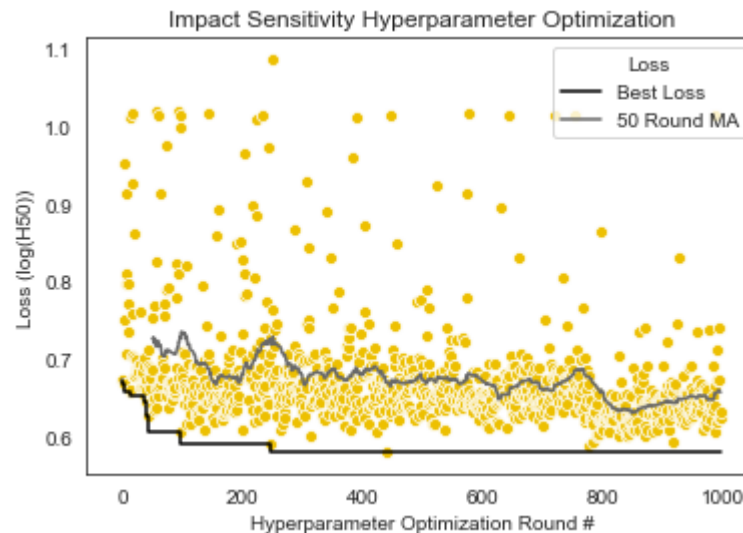
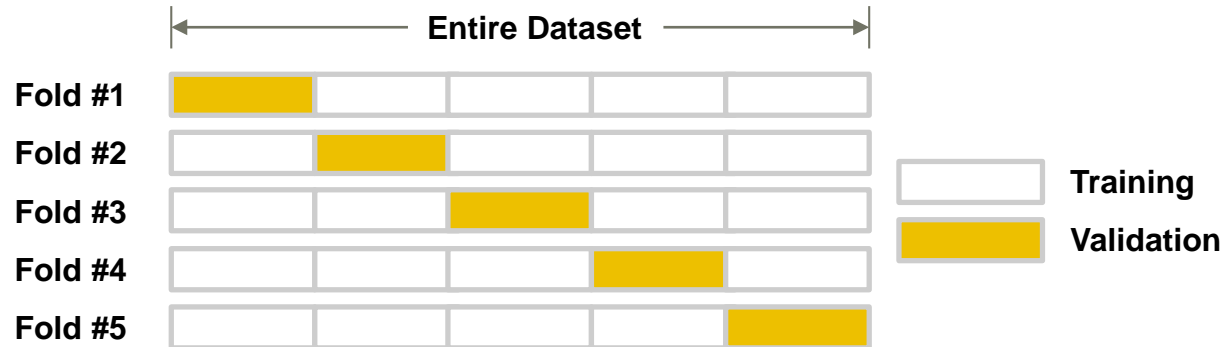
HYPERPARAMETER OPTIMIZATION



- **Bayesian**

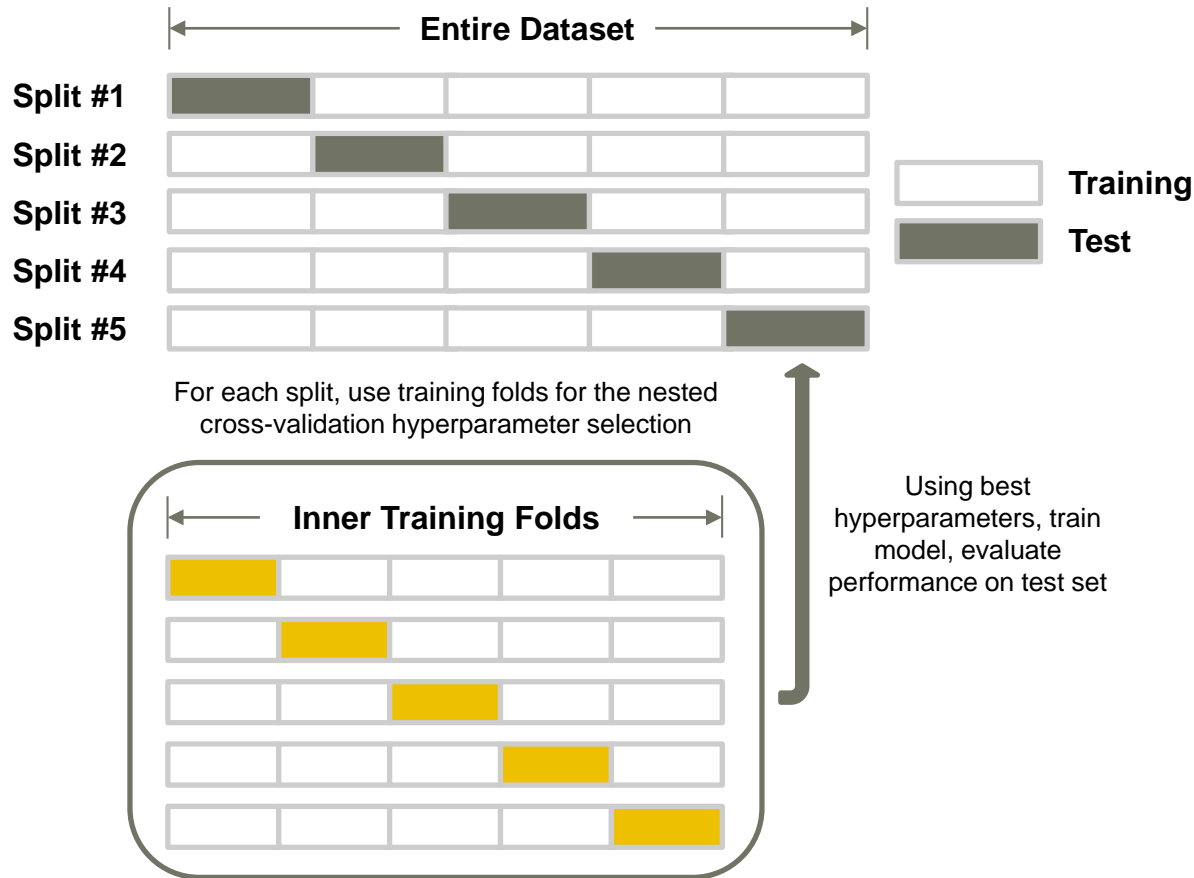
hyperparameter tuning

- 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





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



GPU TIME TO TRAIN



- **Cross validation can be time consuming**
 - 5 rounds CV * 5 outer folds * 4 inner folds * 1000 rounds per inner fold = 100,000 total XGBoost models for model qualification
- **Parallelize construction of XGBoost models on GPU's**
 - Benchmark with Covertypes dataset
 - Use dask to parallelize construction of inner folds across 4 x 2080ti's
 - Speedup of ~5x-10x depending on model

Dataset size (mb)	Dataset size (%)	i7 7700 avg. time, 10x runs (s)	2950x avg. time, 10x runs (s)	1 x 2080ti avg. time, 10x runs (s)
256	100	108.85	34.99	10.36
128	50	53.63	19.07	7.01
64	25	31.14	11.61	5.15
25.5	10	17.02	7.46	4.3
2.55	1	6.26	2.08	3.38



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

Model Evaluation



QSAR/QSPR STATISTICAL PROPERTIES



Property	MAE	RMSE	Median	R-squared
Density (g/cm ³)*	0.0214	0.0364	0.0119	0.988
Melting temperature (°C)*	25.5	34.68	19.34	0.873
Vapor pressure (Log10(mmHg))*	0.531	0.874	0.271	0.942
Impact sensitivity (Log10(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



REPEATED NESTED CROSS VALIDATION



- **Results show variation in CV results dependent on how data is split**
 - IGC50 (n=1792), 4.5% difference in RMSE
- **Results show significant variation dependent on 20% test set split**
 - IGC50 (n=1792), 27.4% difference in RMSE

Property	MAE	RMSE	Median	R-Squared (95% conf. int.)
IGC50 (-Log10(mol/L)) best CV	0.3175	0.4452	0.2223	0.820 (0.807, 0.832)
IGC50 (-Log10(mol/L)) worst CV	0.3266	0.4654	0.2259	0.803 (0.789, 0.816)

Property	MAE	RMSE	Median	R-Squared (95% conf. int.)
IGC50 (-Log10(mol/L)) best fold	0.303	0.4127	0.2284	0.846 (0.819, 0.869)
IGC50 (-Log10(mol/L)) worst fold	0.3729	0.5259	0.2613	0.746 (0.705, 0.782)



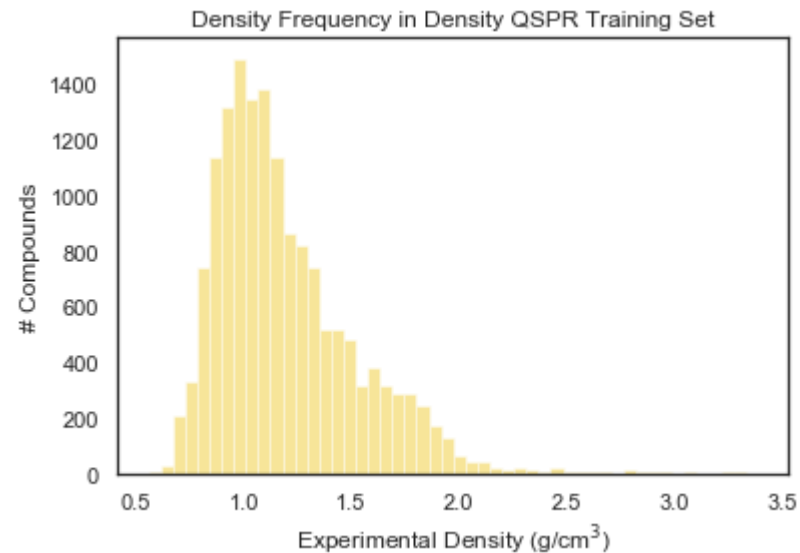
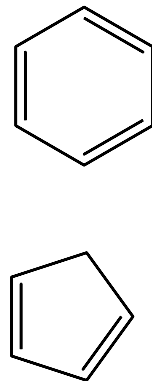
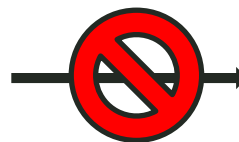
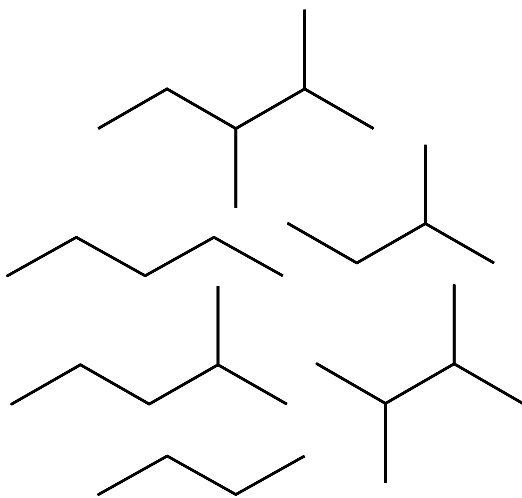
APPLICABILITY DOMAIN



- **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



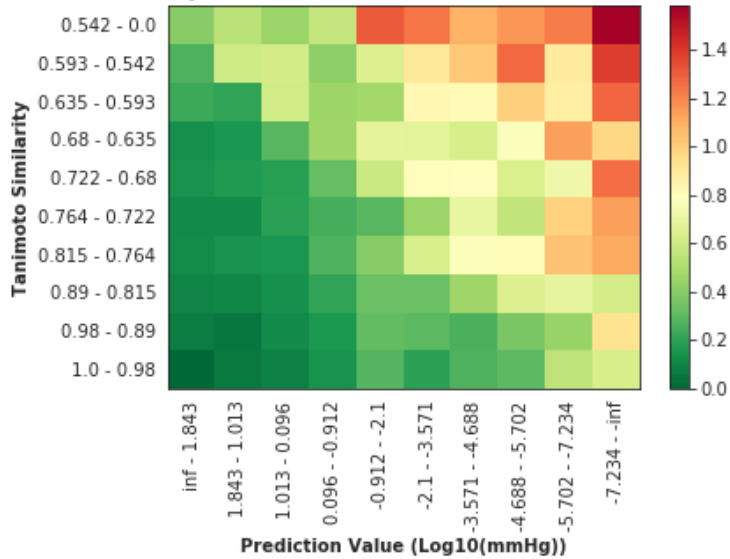
- **Applicability domain by chemical similarity**
 - External prediction for each dataset molecule in process of cross validation
 - Quantifiable measure of similarity to the training set
 - Use Morgan fingerprints to generate Tanimoto similarities
 - Average Tanimoto similarity to the five most similar compounds in the training set
- **Applicability domain by predicted property value**
 - External prediction for each training set molecule in process of cross validation
- **Combine similarity and predicted property**
 - Create a grid with the value of predicted property on one axis, and similarity to the molecules in the training set on the other axis, fill cells with cross validation data
 - Find mean absolute, median, root mean square error of each cell
 - Determine appropriate cell for new compounds to estimate prediction error



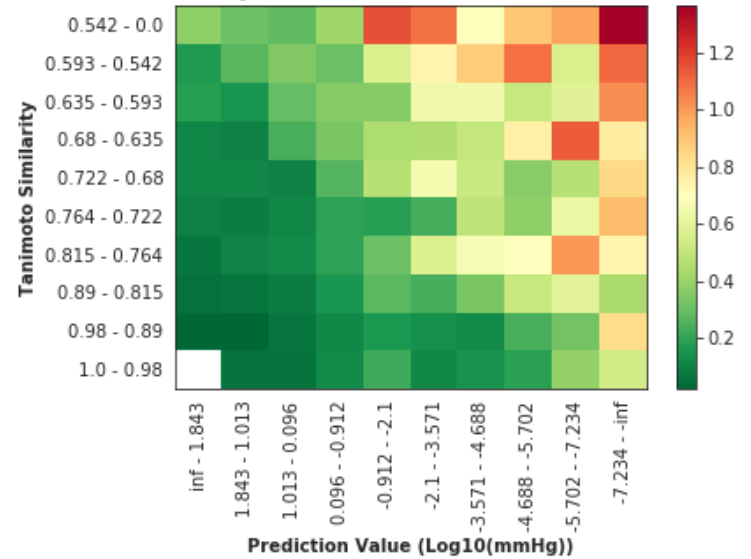
APPLICABILITY DOMAIN



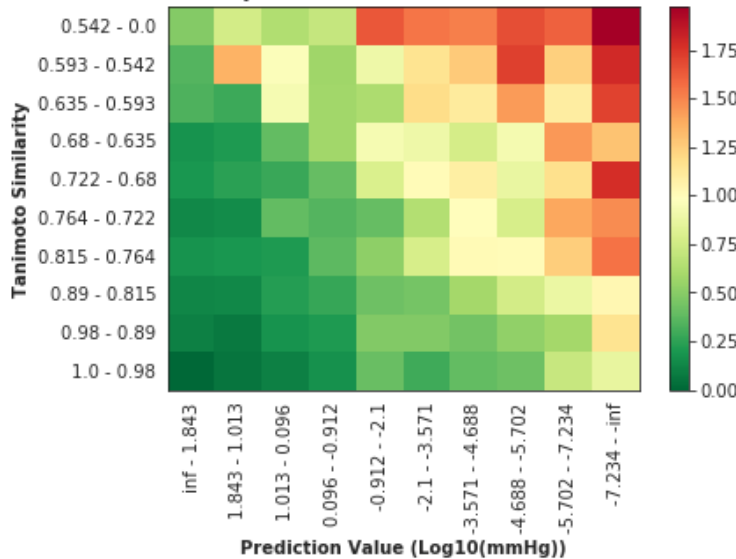
Vapor Pressure Mean Absolute Error Grid



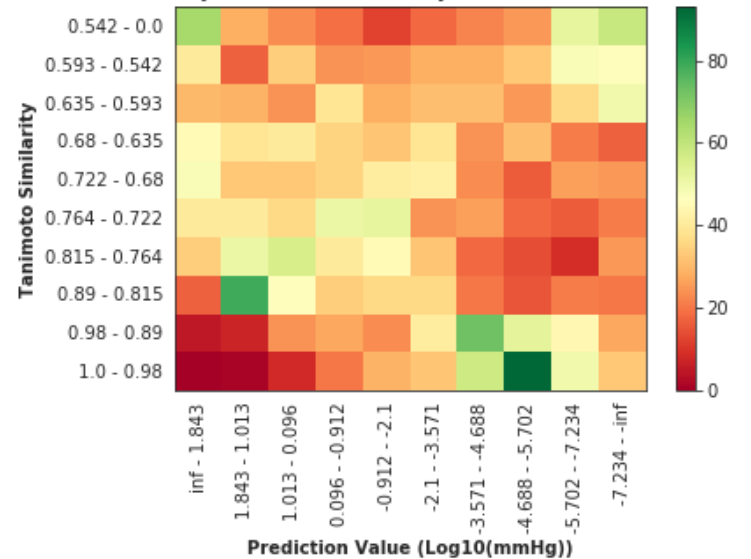
Vapor Pressure Median Error Grid



Vapor Pressure RMSE Error Grid



Vapor Pressure # Compounds Error Grid





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™

by SCHRÖDINGER.



LIVEDESIGN: CAPABILITIES



- **Cheminformatics workflows**

- Quick Properties: properties calculated in real time, examples: molecular weight, CO₂ oxygen balance, CO oxygen balance, and nitrogen content
- QSAR/QSPR Models: density, melting temperature, vapor pressure, impact sensitivity, IGC50, LD50, LC50DM, LC50FM, and BCF

- **DFT molecular workflows**

- Gas phase heat of formation, bond dissociation energy, and thermochemical property (internal energy, entropy, enthalpy) calculations

- **Molecular dynamics workflows**

- Crystal Workflow: density, heat of sublimation, melting point
- Amorphous Workflow: density, heat of sublimation, glass transition temperature



LIVEDESIGN: ADDING COMPOUNDS



Project 1 ☰ Give Feedback

COMPOUNDS Open Live Report HMX

Design Search Import Advanced Enumerate

Compound Structure ID All IDs Rationale Lot Scientist

Add compounds **By Structure**

Add compounds **By ID**

Add compounds & data **From File**

Add more **Data Columns**

PREVIEW PROPERTIES Add Idea To LiveReport

To preview properties as you sketch, add some Property columns to your Live Report.

0 Compounds - 0 Selected 5 Columns (1 Hidden) © 2020 Schrödinger, Inc. Copyright



LIVEDESIGN: MAKING PREDICTIONS



Project 1 ⋮ Give Feedback

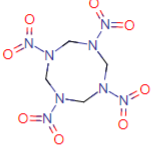
DATA & COLUMNS Open Live Report + HMX ▼

Project LiveReport

Search columns

- Density
- Melting Point
- Vapor Pressure
 - Mean Prediction Error
 - Median Prediction Error
 - Prediction
- Toxicology Models
 - BCF
 - Mean Prediction Error
 - Median Prediction Error
 - Prediction
 - IGC50
 - LC50DM
 - LC50FM
 - LD50
 - Mean Prediction Error
 - Median Prediction Error
 - Prediction
- Experimental Assays ▼
- Other Columns ▼
- f(x) Formulas NEW ▼

Add Columns

<input type="checkbox"/>	Compound Structure	ID	Vapor Pressure (Prediction)	BCF (Prediction) [log]	LD50 (Prediction) [-log(mol/kg)]
<input type="checkbox"/>		V35008	-13.1	0.6	1.8

1 Compounds · 0 Selected · 5 Columns (4 Hidden) © 2020 Schrödinger, Inc. Copyright



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