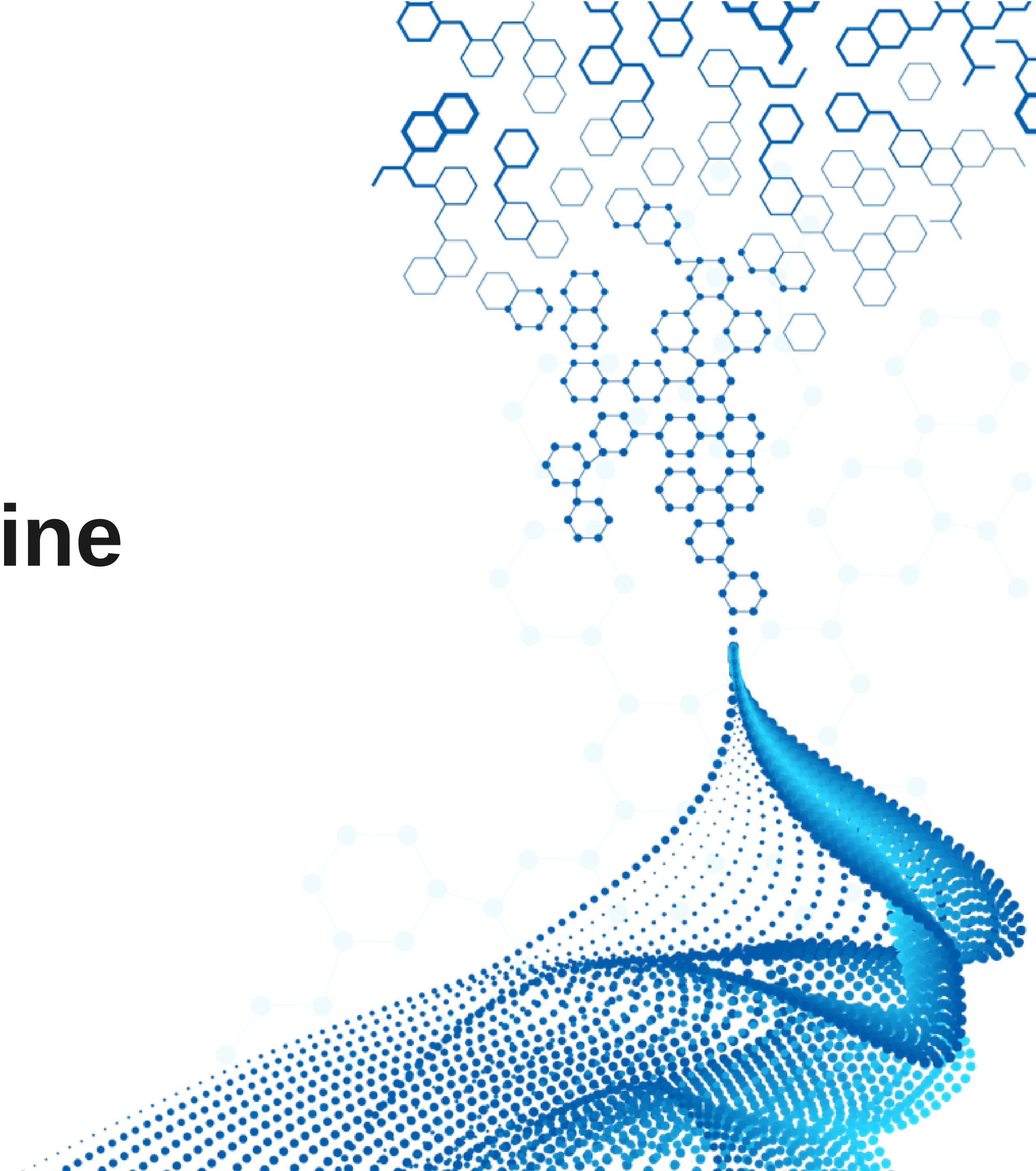
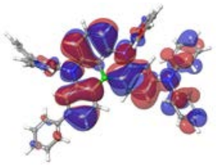
Data-driven materials innovation: where machine learning meets physics



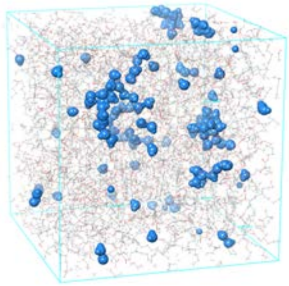
Anand Chandra
Product Manager, Materials Science Informatics
chandras@schrodinger.com

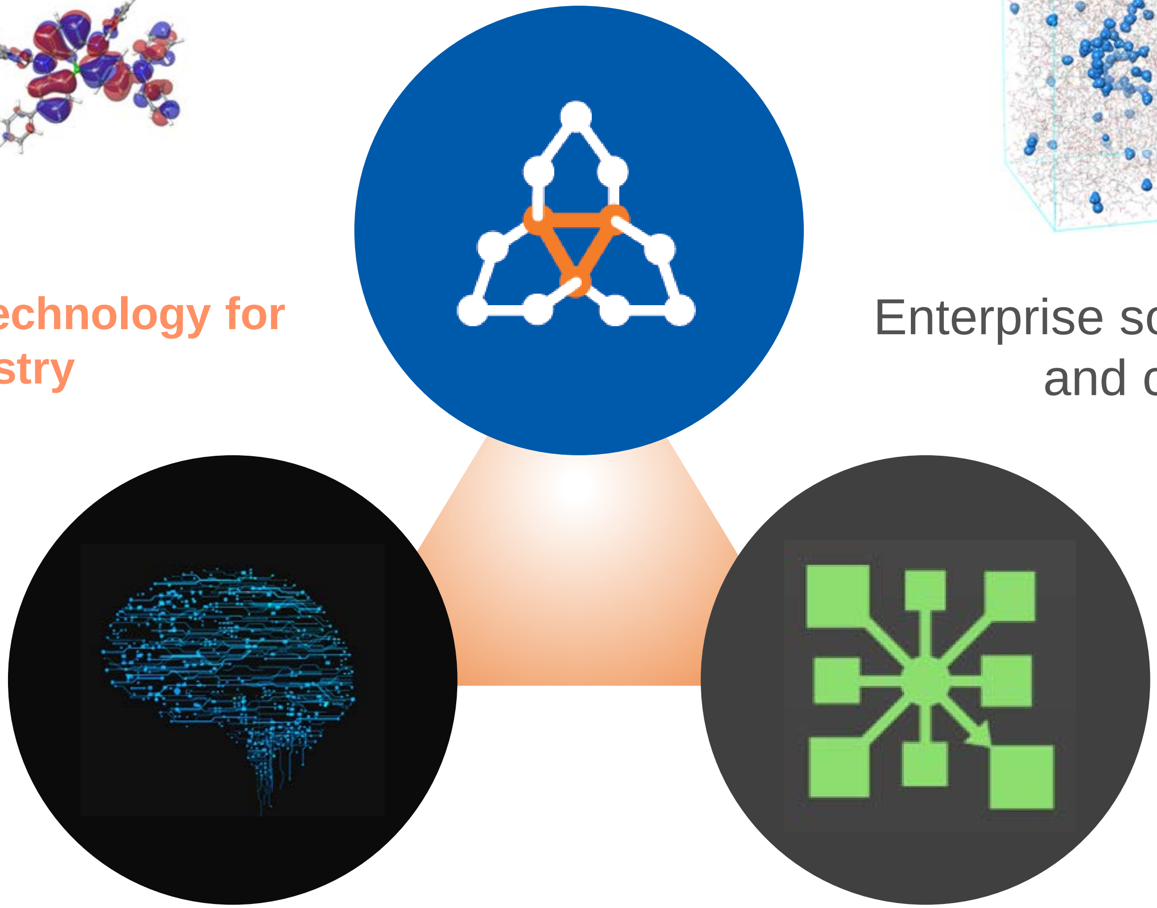
October 31st 2023

Machine Learning for Materials Design/Discovery at Schrödinger



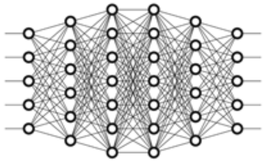
Expertise in physics-based simulation and domain knowledge



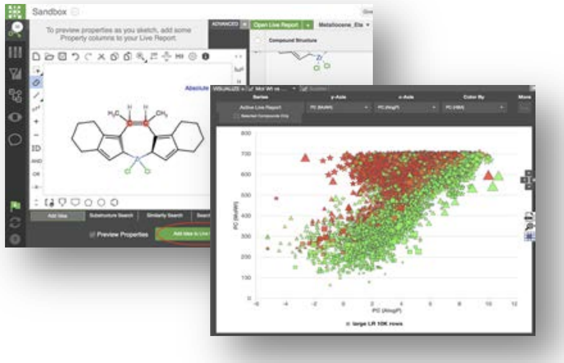


Latest machine learning technology for materials chemistry

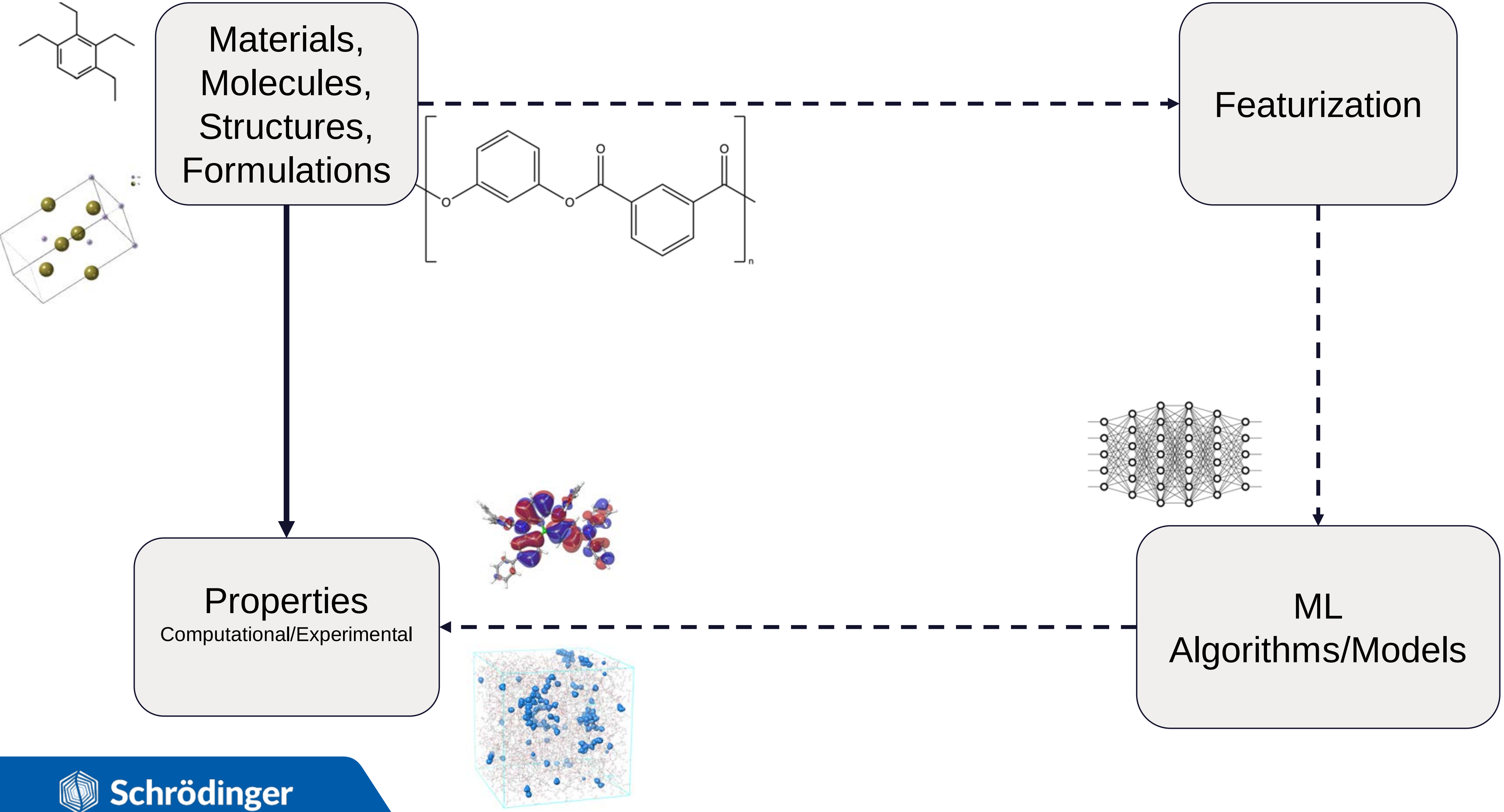
Enterprise solution for data management and collaborative ideation



```
OC(=O)COC1C=C(CN2C[C@@H](CC2)NC2=CC=CC3=CN=CC=C32)
COC1C=C2CC(COC2=CC=1)C1=NC2=CC=C(C=C2O1)C1C=NNC=1
NC1N=C(C=CN=1)C1=CC2N=C(OC=2C=C1)C1CC2=CC=CC=C2CO1
```



Supervised Learning in Materials Science



Materials, Molecules, Structures, Formulations

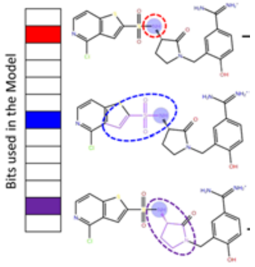
Featurization

ML Algorithms/Models

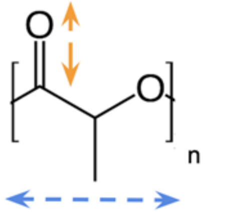
Properties
Computational/Experimental

Featurization in Diverse Materials Systems

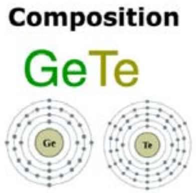
- Properly featurizing various chemical systems is key to building predictive machine learning models
- Small Molecules
  - Physiochemical, topographical descriptors
  - Binary fingerprints (RDKit, Canvas)
  - Graph-based convolution neural networks

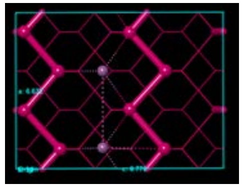


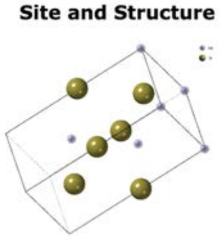
- Polymers
  - Taking into account connections between repeat units
  - RDKit fingerprints + customized descriptors



- Periodic Inorganic Solids
  - Element
  - Lattice structure
  - Oxidation state
  - Intercalation descriptors
  - 3D SOAP (with PCA)







- Formulations and Mixtures
  - Composition
  - Chemistry of the components
  - Experimental/Processing conditions

Schrödinger's Physics-based Simulation Provides Additional Power to Machine Learning

QM (Jaguar)
Catalysis (AutoRW)

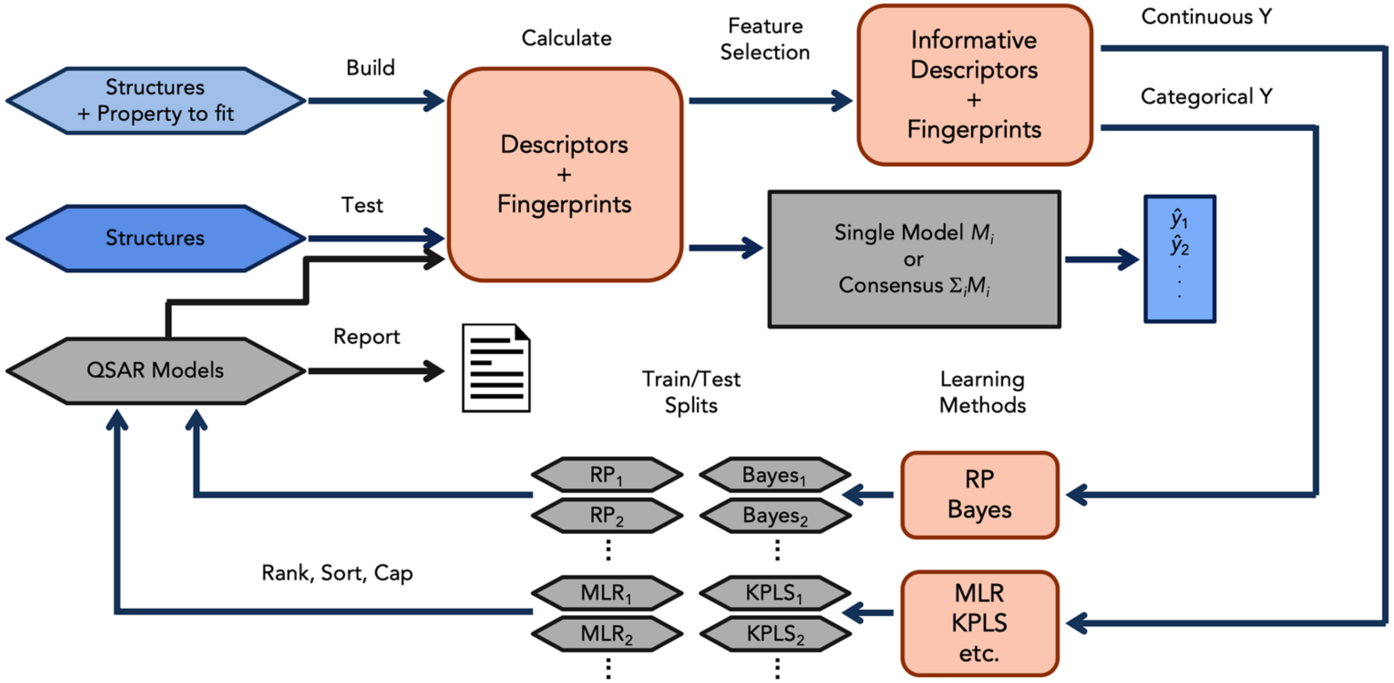
MD (Desmond)
MD + QM

Periodic QM (QE)

MD (Desmond)

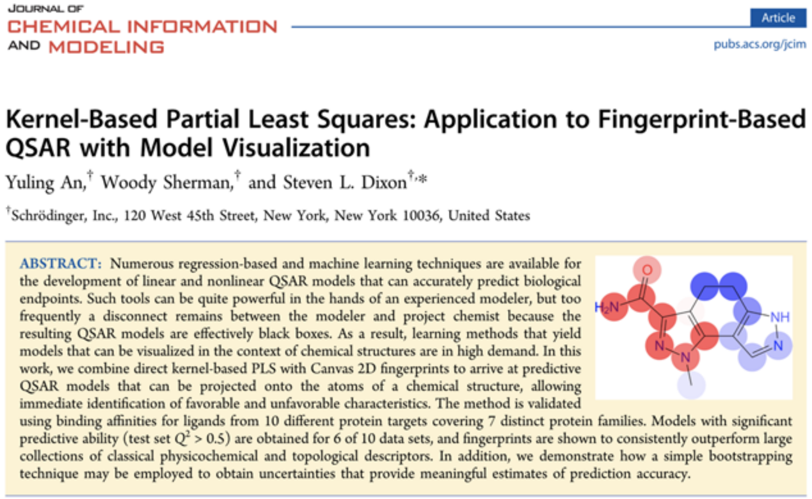
Automated Machine Learning and Visualization in Molecular Systems

- Supervised learning with 400+ built-in descriptors
- Integrated as automated HPC-supported workflow



Schrödinger’s automated model-building algorithm (AutoQSAR)

Machine Learning with Model visualization



JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

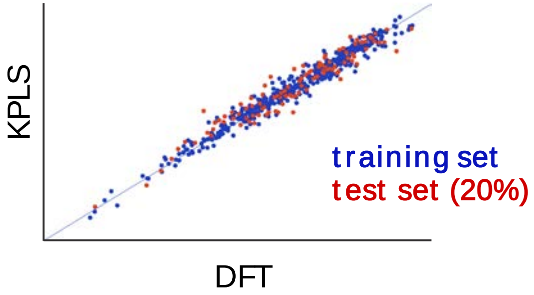
Kernel-Based Partial Least Squares: Application to Fingerprint-Based QSAR with Model Visualization

Yuling An,[†] Woody Sherman,[†] and Steven L. Dixon[†,*]

[†]Schrödinger, Inc., 120 West 45th Street, New York, New York 10036, United States

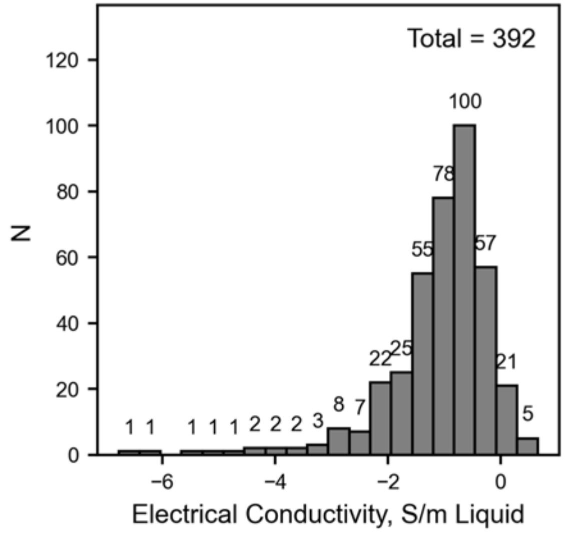
ABSTRACT: Numerous regression-based and machine learning techniques are available for the development of linear and nonlinear QSAR models that can accurately predict biological endpoints. Such tools can be quite powerful in the hands of an experienced modeler, but too frequently a disconnect remains between the modeler and project chemist because the resulting QSAR models are effectively black boxes. As a result, learning methods that yield models that can be visualized in the context of chemical structures are in high demand. In this work, we combine direct kernel-based PLS with Canvas 2D fingerprints to arrive at predictive QSAR models that can be projected onto the atoms of a chemical structure, allowing immediate identification of favorable and unfavorable characteristics. The method is validated using binding affinities for ligands from 10 different protein targets covering 7 distinct protein families. Models with significant predictive ability (test set $Q^2 > 0.5$) are obtained for 6 of 10 data sets, and fingerprints are shown to consistently outperform large collections of classical physicochemical and topological descriptors. In addition, we demonstrate how a simple bootstrapping technique may be employed to obtain uncertainties that provide meaningful estimates of prediction accuracy.

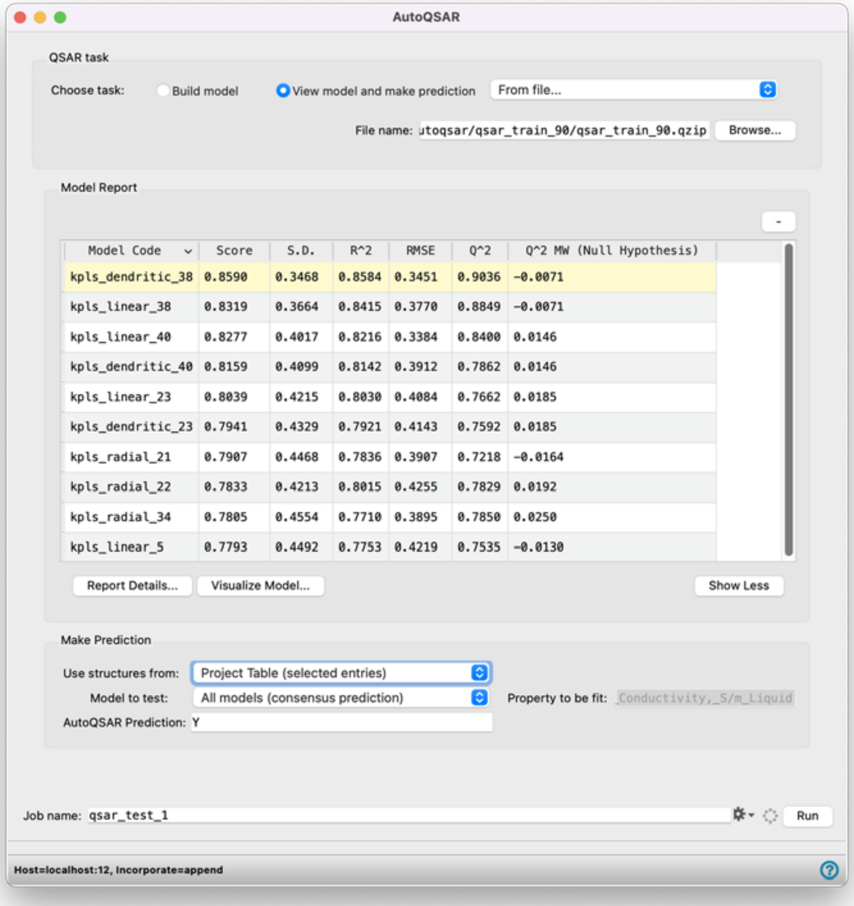
Automated cross-validation for model scoring and ranking



AutoQSAR for Ionic Liquids

- 392 ionic liquids from the NIST IL Thermo database
- Target Property → Electrical conductivity





QSAR task

Choose task: ○ Build model ● View model and make prediction From file...

File name: utoqsar/qsar_train_90/qsar_train_90.qzip

Model Report

Model Code	Score	S.D.	R^2	RMSE	Q^2	Q^2 MW (Null Hypothesis)
kpls_dendritic_38	0.8590	0.3468	0.8584	0.3451	0.9036	−0.0071
kpls_linear_38	0.8319	0.3664	0.8415	0.3770	0.8849	−0.0071
kpls_linear_40	0.8277	0.4017	0.8216	0.3384	0.8400	0.0146
kpls_dendritic_40	0.8159	0.4099	0.8142	0.3912	0.7862	0.0146
kpls_linear_23	0.8039	0.4215	0.8030	0.4084	0.7662	0.0185
kpls_dendritic_23	0.7941	0.4329	0.7921	0.4143	0.7592	0.0185
kpls_radial_21	0.7907	0.4468	0.7836	0.3907	0.7218	−0.0164
kpls_radial_22	0.7833	0.4213	0.8015	0.4255	0.7829	0.0192
kpls_radial_34	0.7805	0.4554	0.7710	0.3895	0.7850	0.0250
kpls_linear_5	0.7793	0.4492	0.7753	0.4219	0.7535	−0.0130

Make Prediction

Use structures from: Project Table (selected entries)

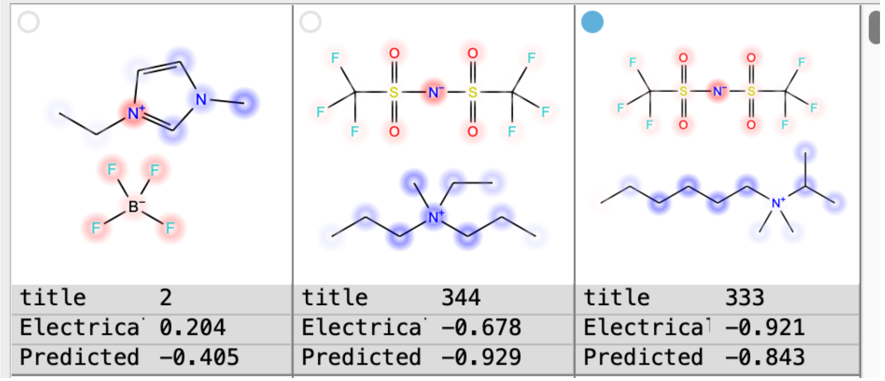
Model to test: All models (consensus prediction)

Property to be fit: Conductivity,_S/m_Liquid

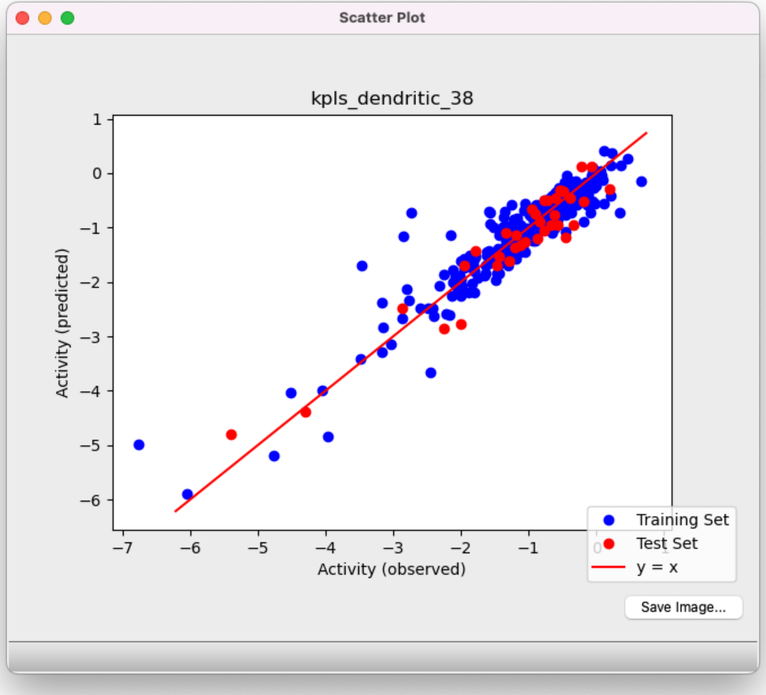
AutoQSAR Prediction: Y

Job name: qsar_test_1

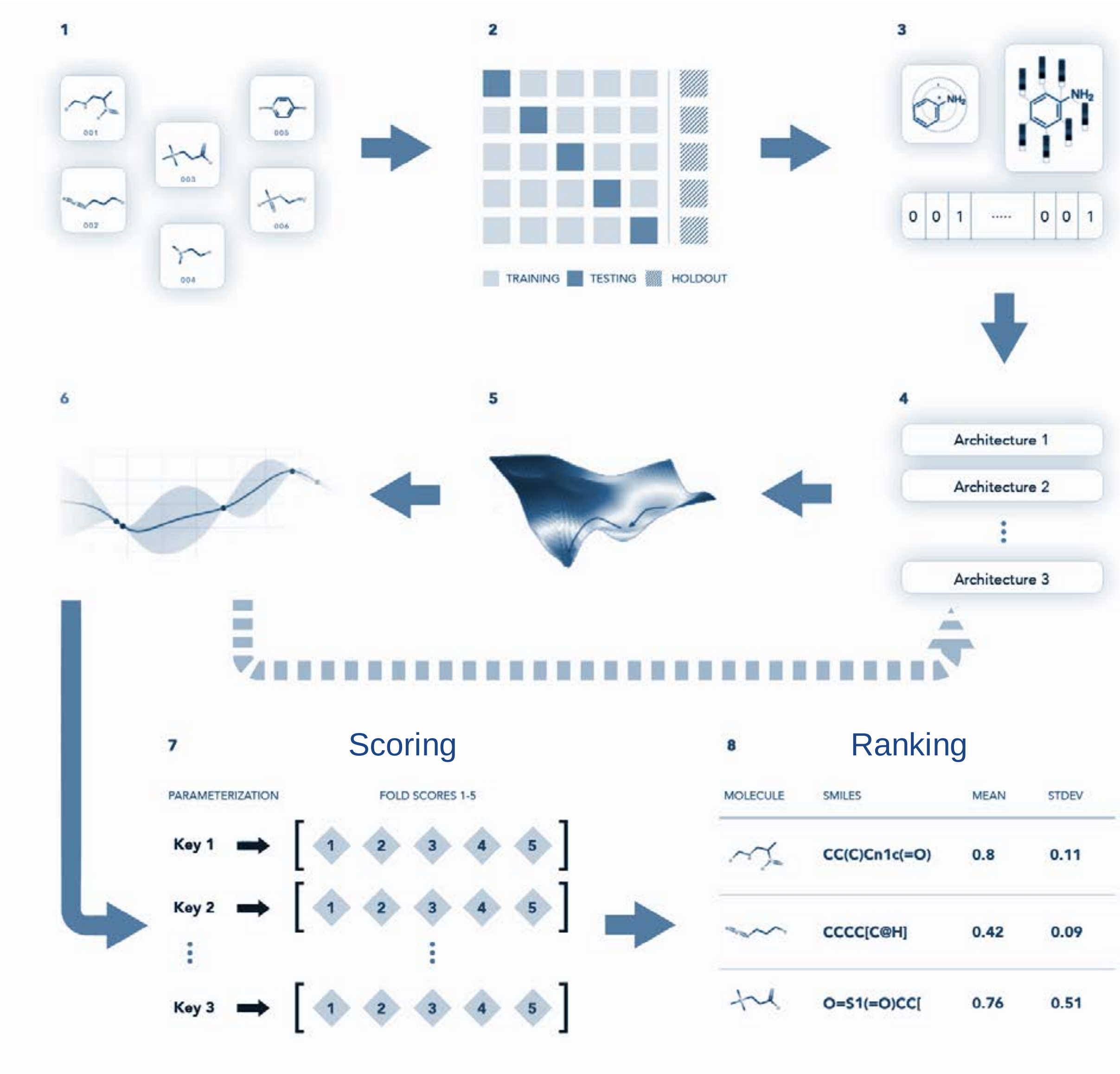
Host=localhost:12, Incorporate=append



title 2	title 344	title 333
Electrica 0.204	Electrica −0.678	Electrica −0.921
Predicted −0.405	Predicted −0.929	Predicted −0.843



DeepAutoQSAR: Automated Model Selection & Parameter Optimization



MOLECULE	SMILES	MEAN	STDEV
	CC(C)Cn1c(=O)	0.8	0.11
	CCCC[C@H]	0.42	0.09
	O=S1(=O)CC[	0.76	0.51

Data splits, featurization

Model architecture, descriptor and hyperparameter combinations explored and optimized via Bayesian Optimization

Models Sampled

- Dense Neural Network
- Random Forest Regressor
- XGBoost
- TorchGraphConv
- GCN
- GraphSAGE
- GIN
- TopK
- SAGPool
- EdgePool
- GlobalAttention
- Set2Set
- SortPool

Consensus Model

- Prediction = an average of the predictions for 5 best models
- Uncertainty = SD across the 5 predictions

For details on model training and performance see: Schrödinger LigandML Performance Whitepaper

Case Study - Redox Flow Batteries

- Design: Homobenzylic ethers (HBE) with oxidation potential in a pre-specified range.
- Oxidation potential of 1,400 HBEs calculated as the initial (training) dataset for machine learning

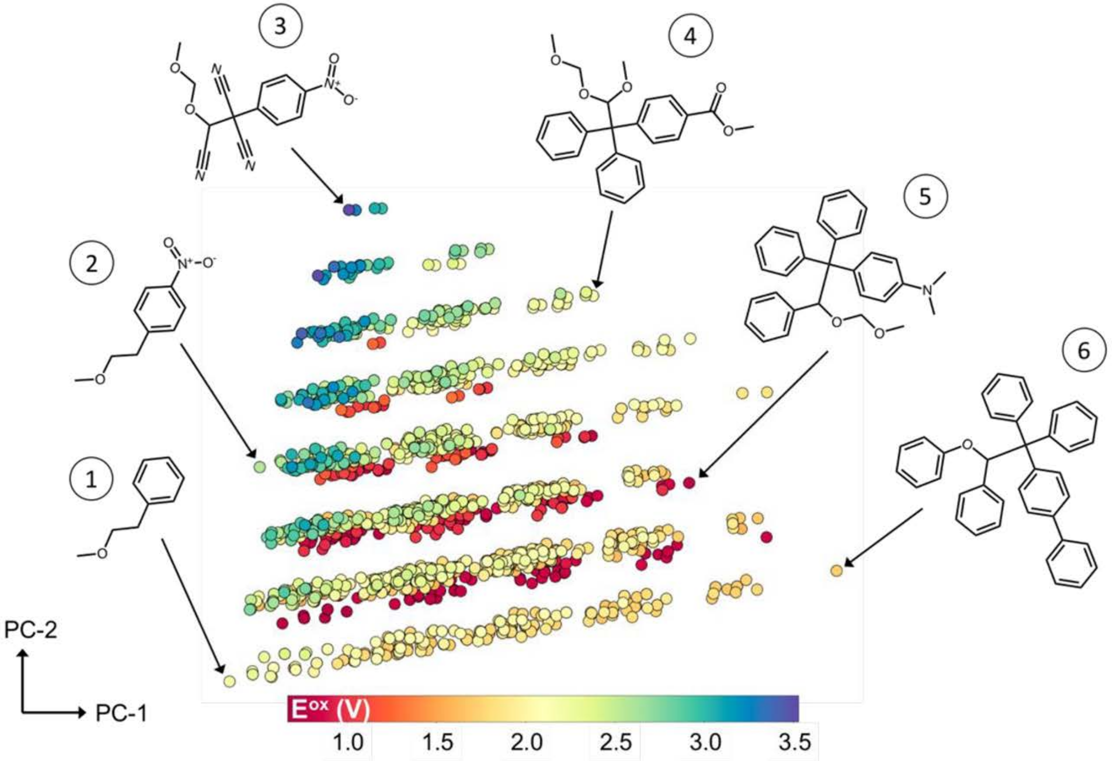
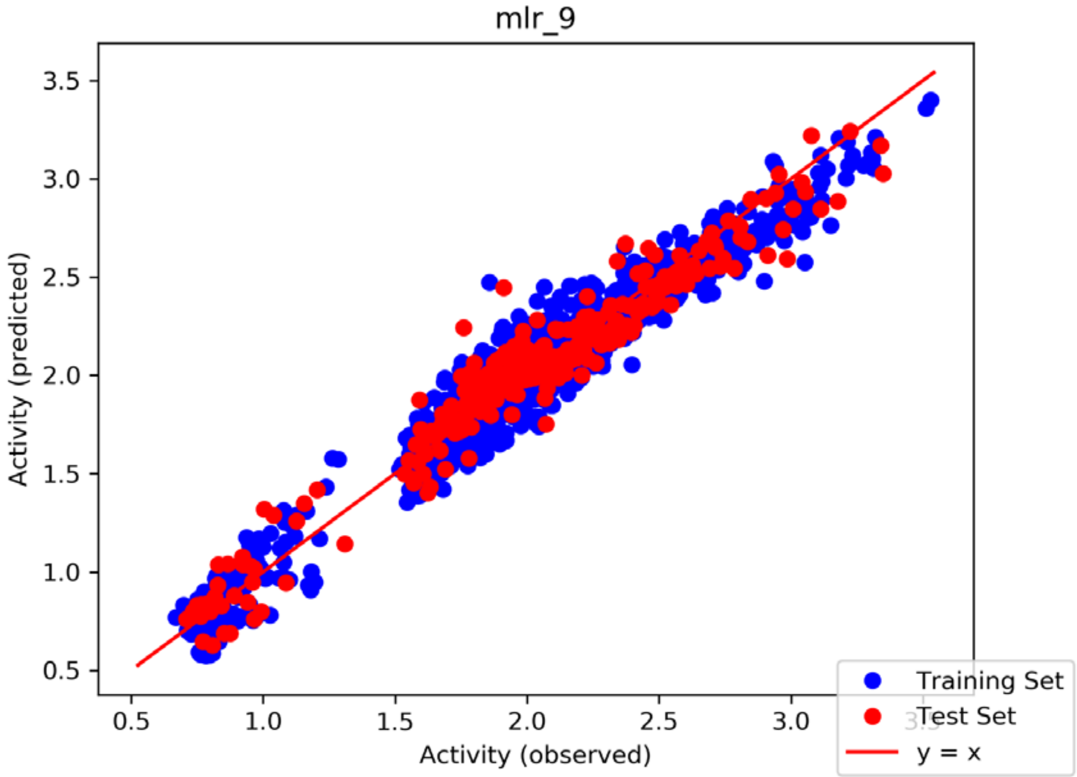


Figure 2 Graphical illustration of the chemical space of 1,400 HBEs and their computed oxidation potentials (E^{ox}). PC-1 and PC-2 represent principle component 1 and 2, respectively.

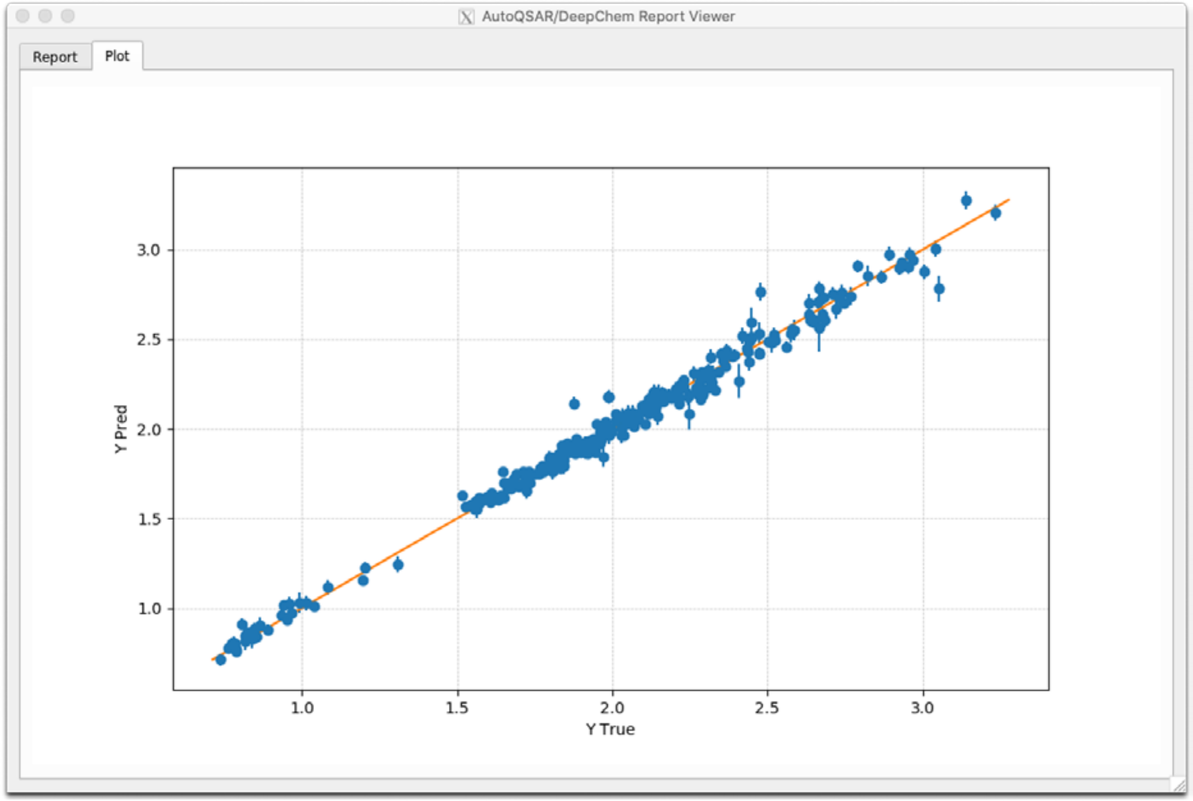
[1] Doan, Hieu A., Garvit Agarwal, Hai Qian, Michael J. Counihan, Joaquín Rodríguez-López, Jeffrey S. Moore, and Rajeev S. Assary. "**Quantum Chemistry-Informed Active Learning to Accelerate the Design and Discovery of Sustainable Energy Storage Materials**." *Chemistry of Materials* (2020).

AutoQSAR vs DeepAutoQSAR Results

Traditional AutoQSAR
Test R^2 = 0.94



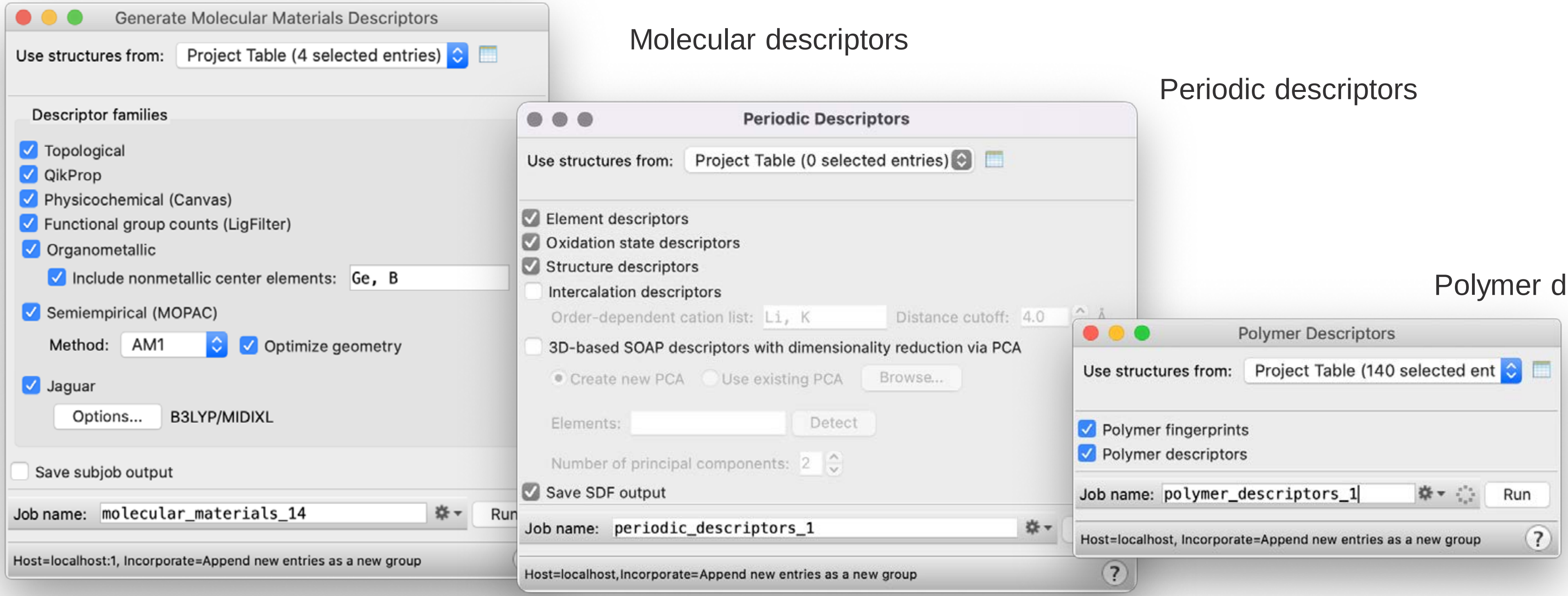
DeepAutoQSAR
Test R^2 = 0.98



- ML models were created for oxidation potential of 1,400 homobenzylic ethers for Redox Flow
- Both AutoQSAR and DeepAutoQSAR offer solid predictive capability.
- The deep-NN-based model (by DeepAutoQSAR) outperforms descriptor-based models for larger (>1000) training set.

Chemical Featurization using Physics

- 100+ additional physics-based descriptors by QM-bound properties, repeat-unit chemistry, and crystallinity
- Direct link to AutoQSAR and other workflows within the platform



Molecular descriptors

Generate Molecular Materials Descriptors

Use structures from: Project Table (4 selected entries)

Descriptor families

☑ Topological
☑ QikProp
☑ Physicochemical (Canvas)
☑ Functional group counts (LigFilter)
☑ Organometallic
☑ Include nonmetallic center elements: Ge, B
☑ Semiempirical (MOPAC)
Method: AM1 ☑ Optimize geometry
☑ Jaguar
Options... B3LYP/MIDIXL

□ Save subjob output

Job name: molecular_materials_14

Host=localhost:1, Incorporate=Append new entries as a new group

Periodic descriptors

Periodic Descriptors

Use structures from: Project Table (0 selected entries)

☑ Element descriptors
☑ Oxidation state descriptors
☑ Structure descriptors
□ Intercalation descriptors
Order-dependent cation list: Li, K Distance cutoff: 4.0
□ 3D-based SOAP descriptors with dimensionality reduction via PCA
● Create new PCA ○ Use existing PCA Browse...
Elements: Detect
Number of principal components: 2
☑ Save SDF output

Job name: periodic_descriptors_1

Host=localhost,Incorporate=Append new entries as a new group

Polymer descriptors

Polymer Descriptors

Use structures from: Project Table (140 selected ent

☑ Polymer fingerprints
☑ Polymer descriptors

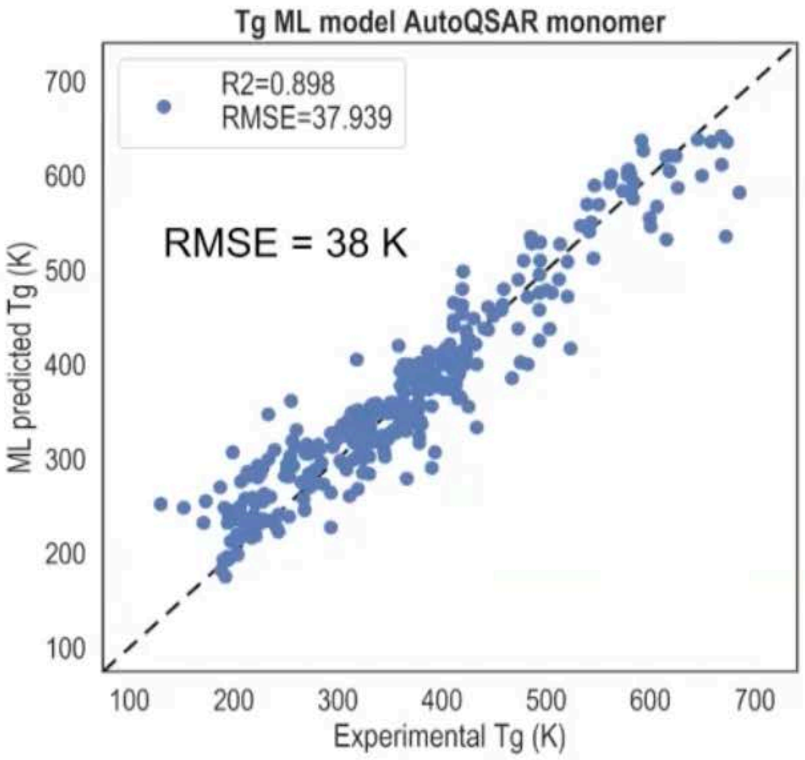
Job name: polymer_descriptors_1

Host=localhost, Incorporate=Append new entries as a new group

Customized Polymer Descriptors Outperform Simple Monomers

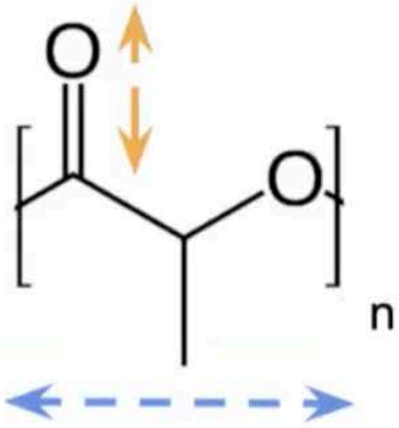
Polymer Descriptors

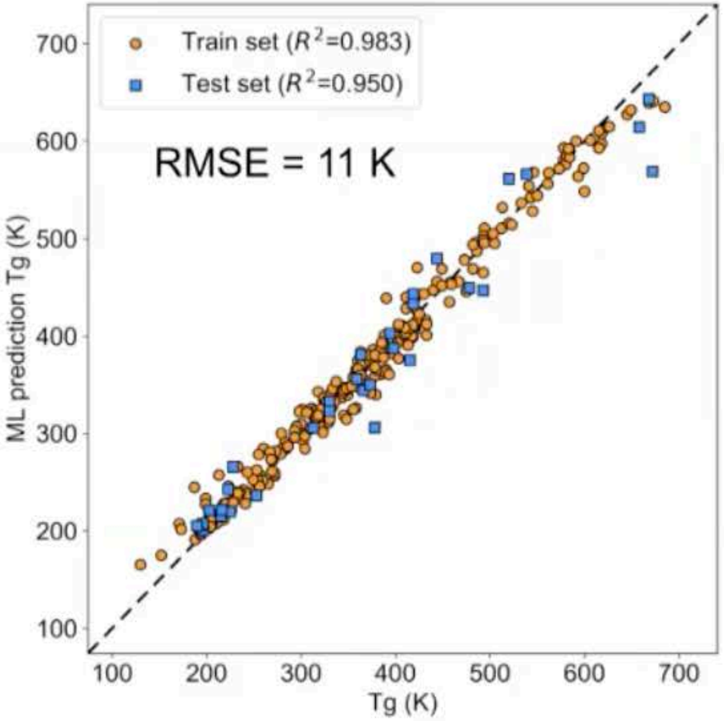
- Topological torsion fingerprints
- Number of rotatable bonds
- Number of ring atoms
- Fused ring atoms


Tg ML model AutoQSAR monomer

RMSE = 38 K

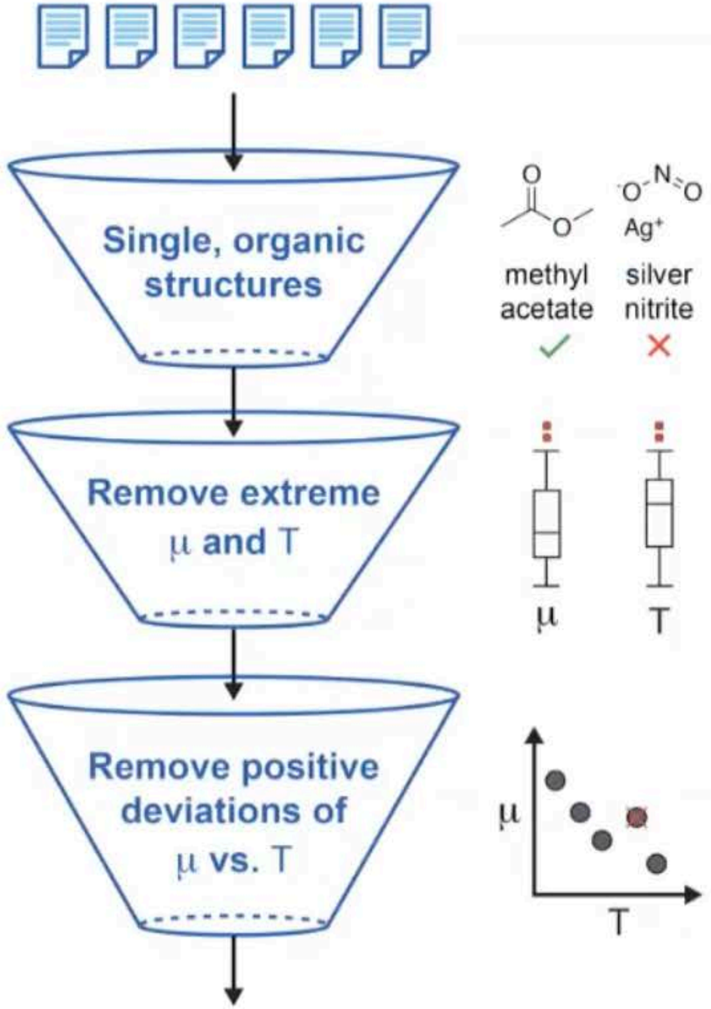
Data set of 315 polymers with Tg values




RMSE = 11 K

Viscosity Dataset for Machine Learning Models

Literature extraction of viscosity
~5,356 viscosities



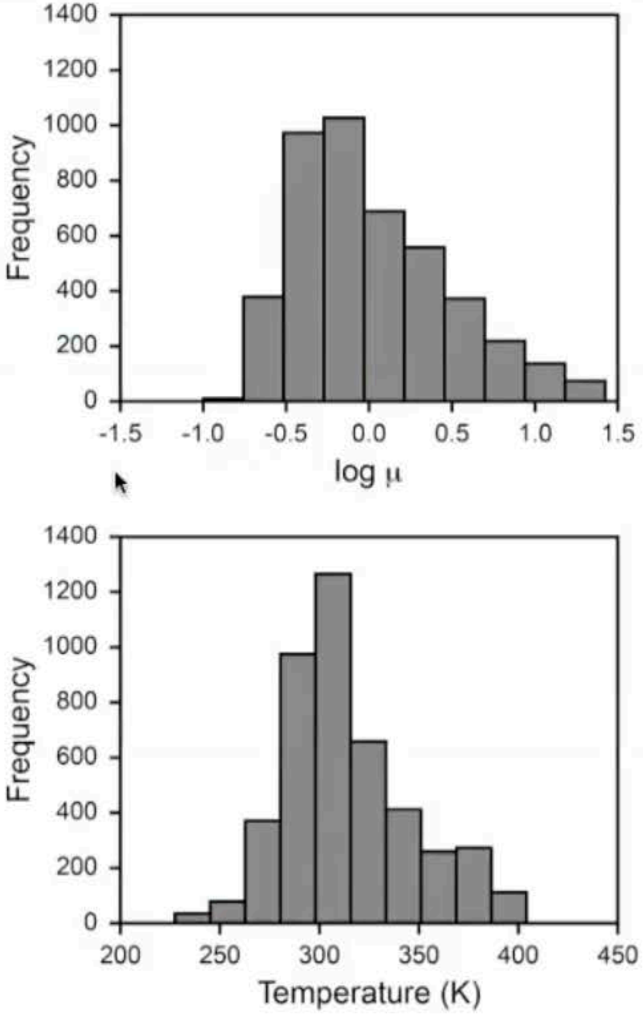
Single, organic structures

Remove extreme μ and T

Remove positive deviations of μ vs. T

4,400 viscosities

Distribution of viscosity and temperature

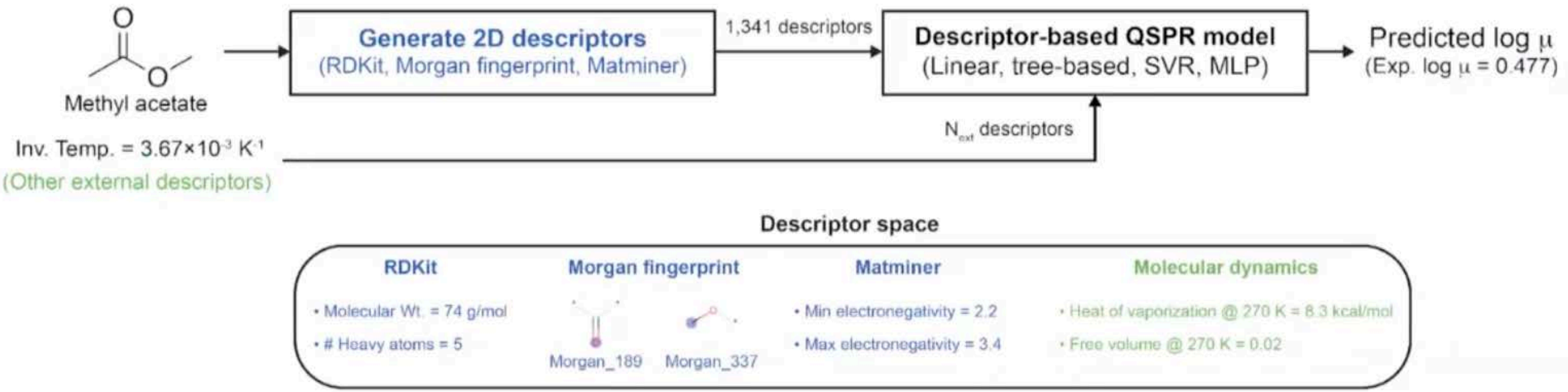


Dataset summary:

- **1,005** unique molecules
- Atomic elements of {H, C, N, O, F, Si, P, S, Cl, Br, and I}
- Viscosity is between **0.10** to **26.52 cP**
- Temperature is between **227 K** to **404 K**

Quantitative Structure-Property Relationships (QSPR)

Descriptor-Based



Methyl acetate

Generate 2D descriptors (RDKit, Morgan fingerprint, Matminer)

1,341 descriptors

Descriptor-based QSPR model (Linear, tree-based, SVR, MLP)

Predicted log μ (Exp. log μ = 0.477)

Inv. Temp. = 3.67×10^{-3} K^{-1} (Other external descriptors)

N_{ext} descriptors

Descriptor space

RDKit
- Molecular Wt. = 74 g/mol
- # Heavy atoms = 5

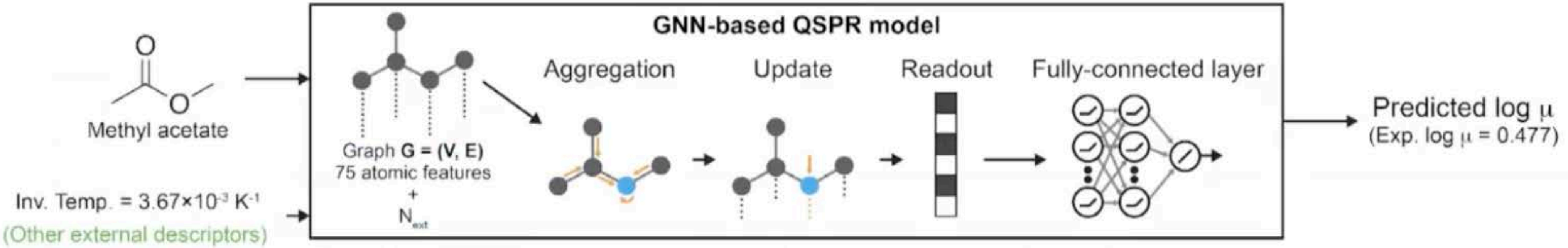
Morgan fingerprint

Morgan_189 Morgan_337

Matminer
- Min electronegativity = 2.2
- Max electronegativity = 3.4

Molecular dynamics
- Heat of vaporization @ 270 K = 8.3 kcal/mol
- Free volume @ 270 K = 0.02

Graph neural networks (GNN)-Based



Methyl acetate

Inv. Temp. = 3.67×10^{-3} K^{-1} (Other external descriptors)

GNN-based QSPR model

Graph **G = (V, E)** 75 atomic features + N_{ext}

Aggregation

Update

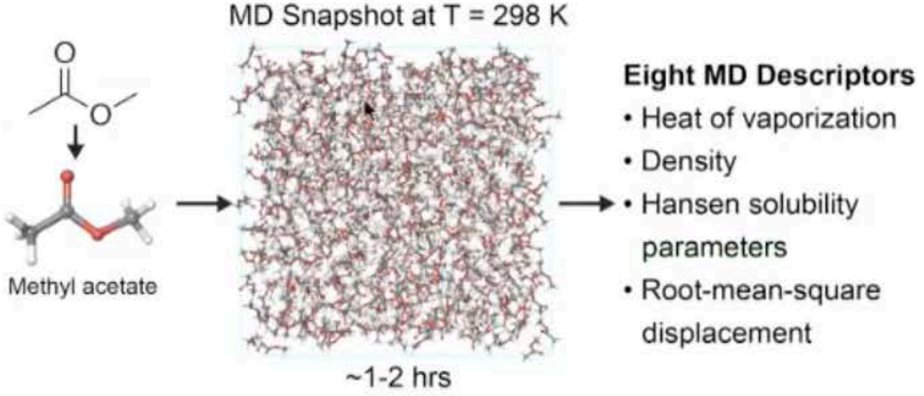
Readout

Fully-connected layer

Predicted log μ (Exp. log μ = 0.477)

GNNs: Graph convolution neural network (GCN), EdgePool, TopK, GraphSAGE, GIN, *etc.*

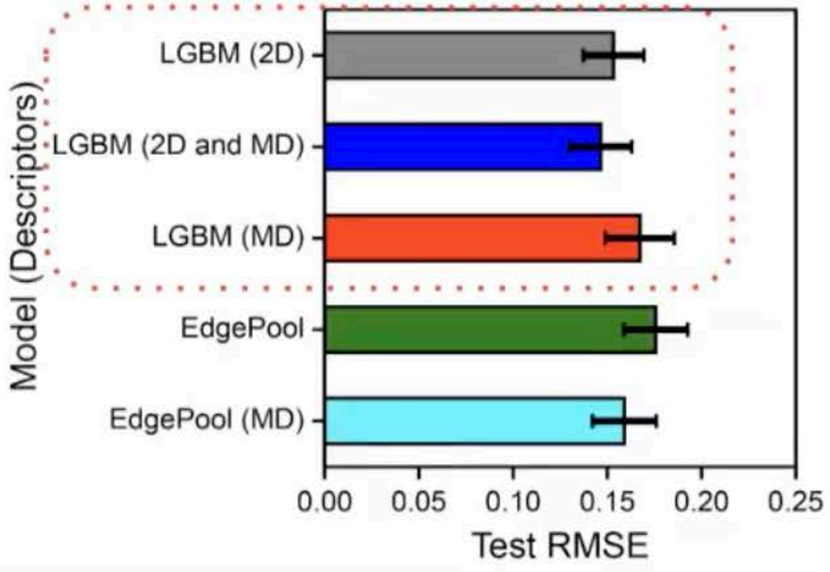
Impact of MD-Derived Simulation Descriptors



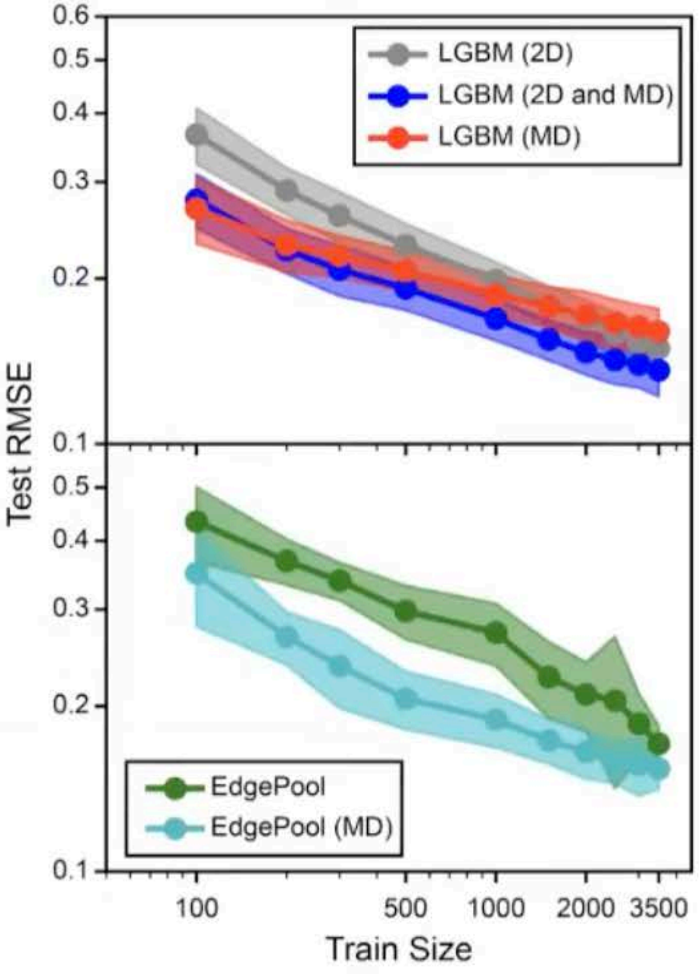
Eight MD Descriptors

- Heat of vaporization
- Density
- Hansen solubility parameters
- Root-mean-square displacement

Performance with MD Descriptors



Learning curve with and without MD Descriptors



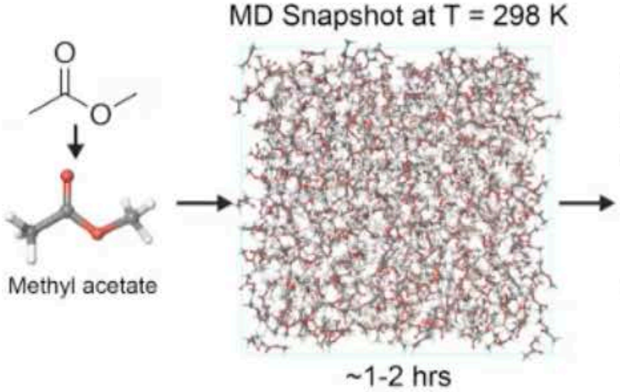
Main takeaways:

- Inclusion of MD Descriptors lowers test set RMSEs
- MD descriptors are most useful at small training size (<1,000)

Scheduled for 23-4

The technical features and projected timeline presented on this slide is for discussion purposes only. Such planned or potential capabilities are subject to change at any time.

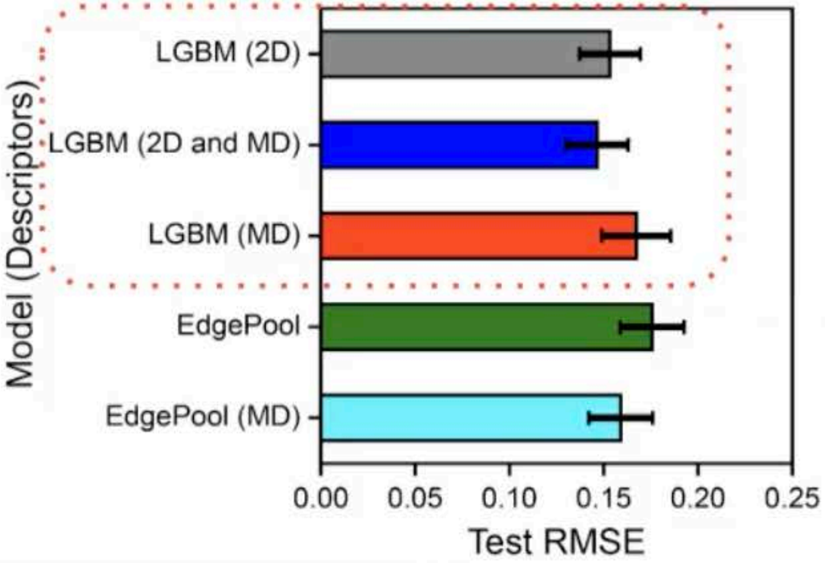
Impact of MD-Derived Simulation Descriptors



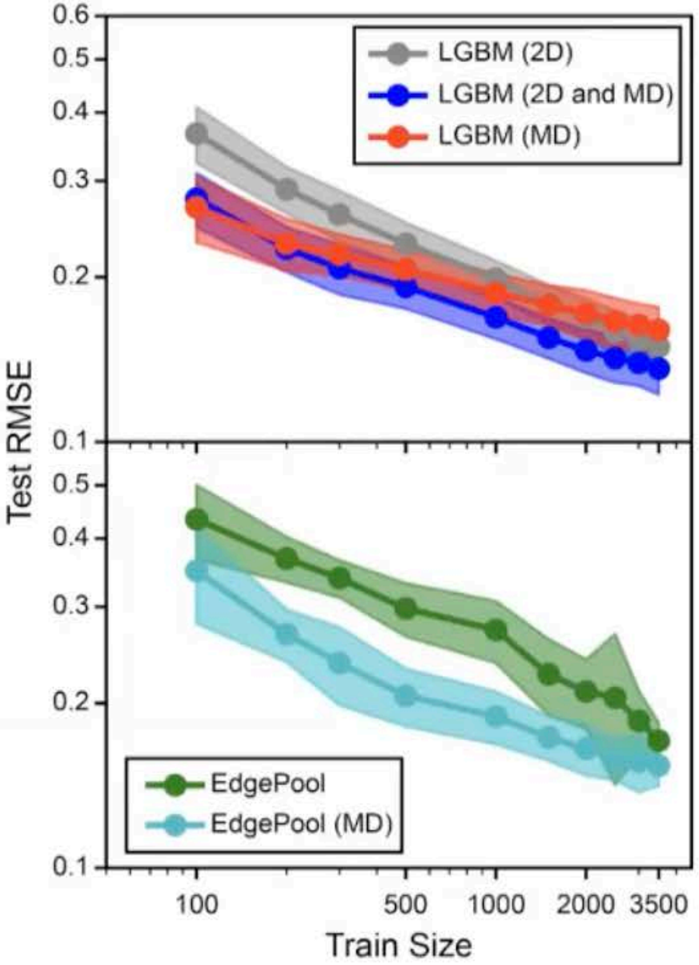
Eight MD Descriptors

- Heat of vaporization
- Density
- Hansen solubility parameters
- Root-mean-square displacement

Performance with MD Descriptors

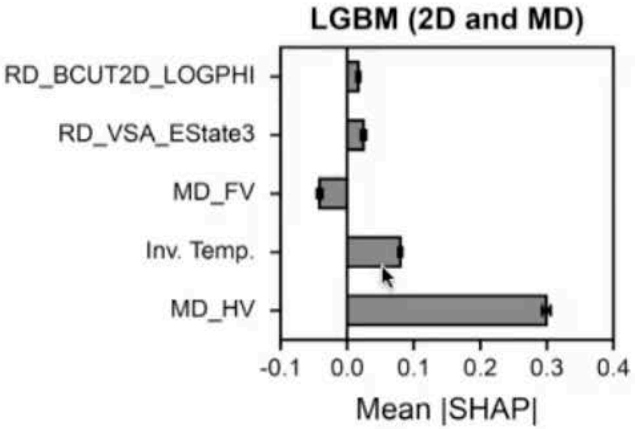


Learning curve with and without MD Descriptors



Main takeaways:

- Inclusion of MD Descriptors lowers test set RMSEs
- MD descriptors are most useful at small training size (<1,000)

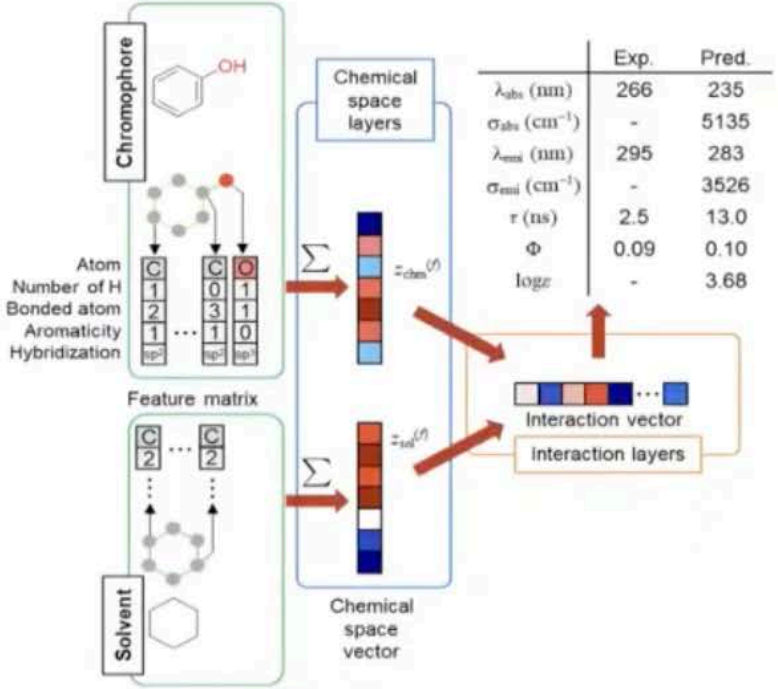


Which descriptors were most useful for viscosity?

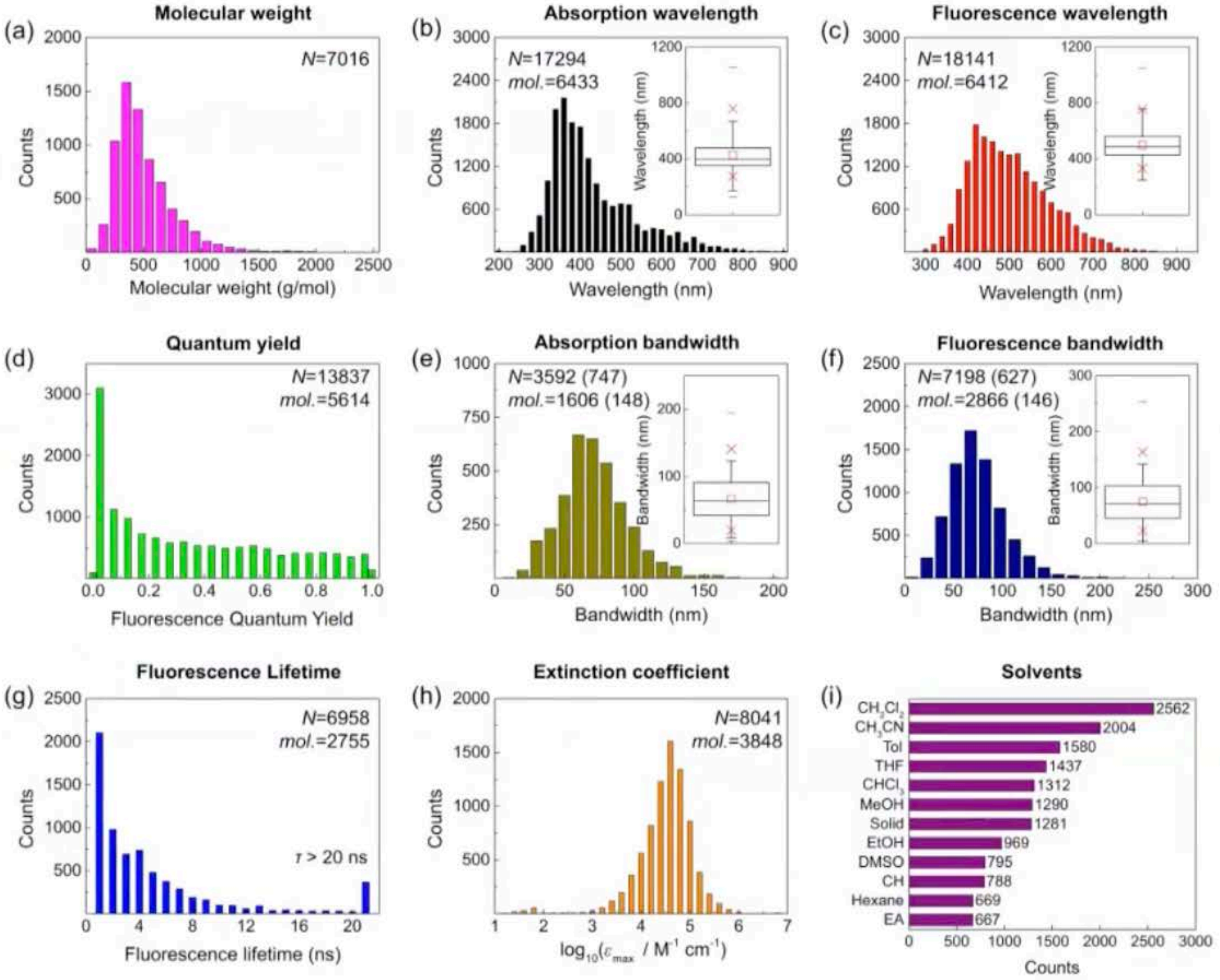
Machine Learning Optoelectronics Properties with DFT descriptors

Database of Optical Properties of Organic Compounds

- Experimental dataset of **20,236** combinations of **7,016** chromophores in **365** solvents



	Exp.	Pred.
λ_{abs} (nm)	266	235
σ_{abs} (cm^{-1})	-	5135
λ_{emi} (nm)	295	283
σ_{emi} (cm^{-1})	-	3526
τ (ns)	2.5	13.0
Φ	0.09	0.10
$\log\varepsilon$	-	3.68

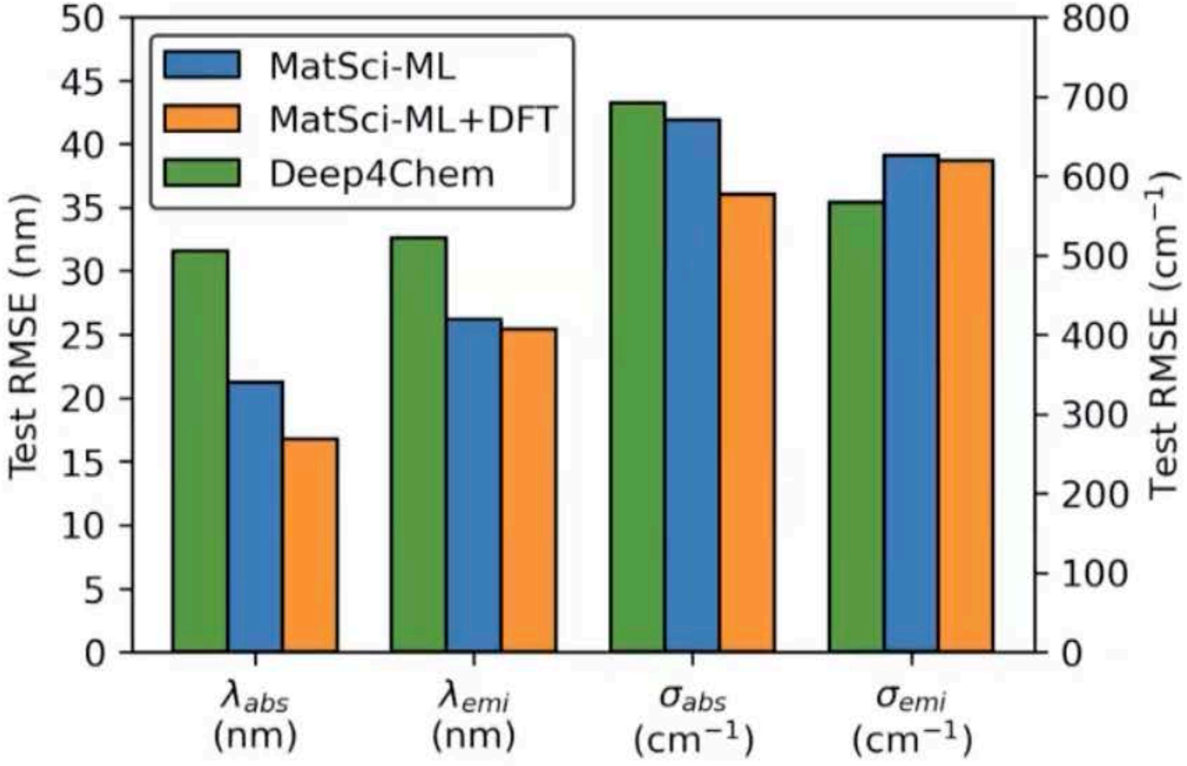


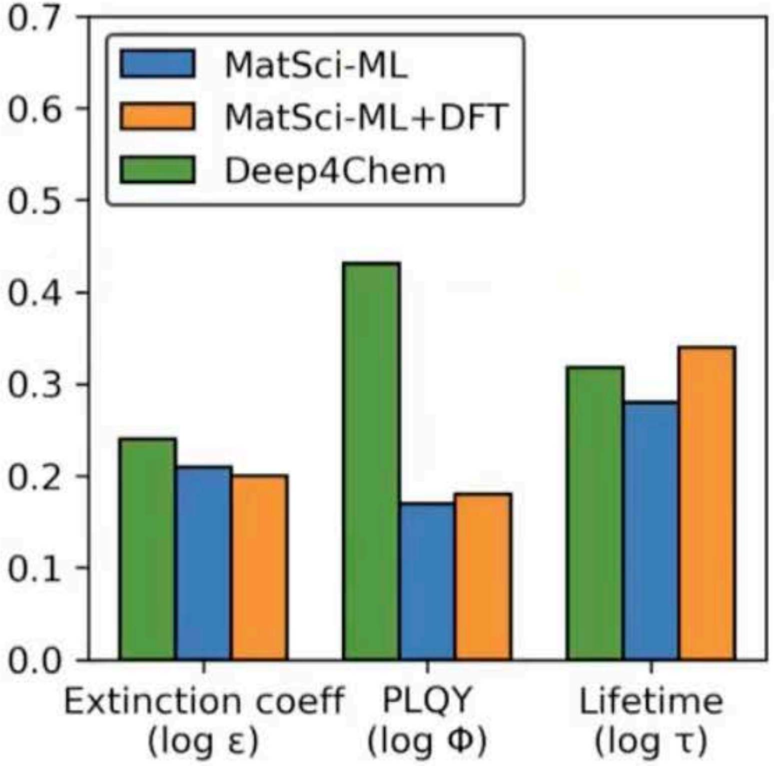
Schrödinger

Joung, Joonyoung F., et al. "Deep learning optical spectroscopy based on experimental database: potential applications to molecular design." *JACS Au* 1.4 (2021): 427-438.

Benchmark of DFT Descriptors

- Combining 2D and DFT descriptors leads to state-of-the art performance





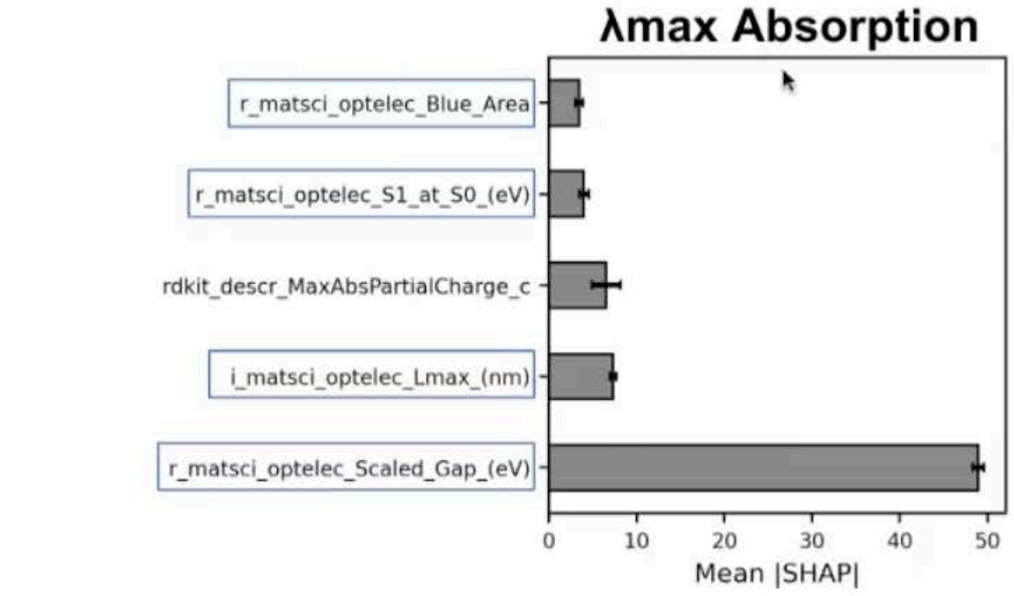
Scheduled for 23-4

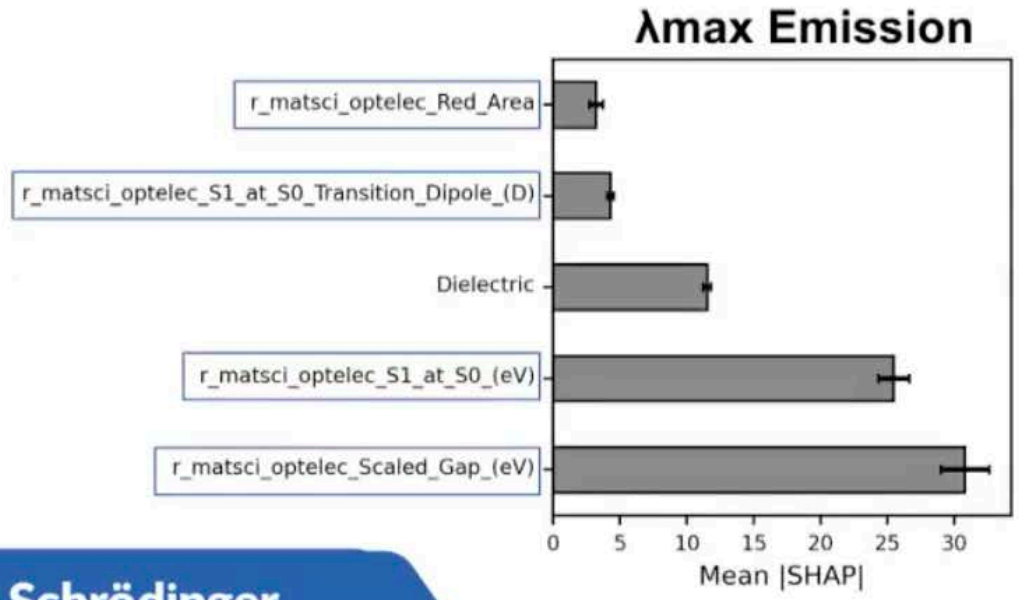
The technical features and projected timeline presented on this slide is for discussion purposes only. Such planned or potential capabilities are subject to change at any time.

MatSci-ML Model: Neural Network

Features: 2D Descriptors of Chromophore (+DFT Features) + Dielectric constant of Solvent

Feature Importance Analysis





DFT Descriptors

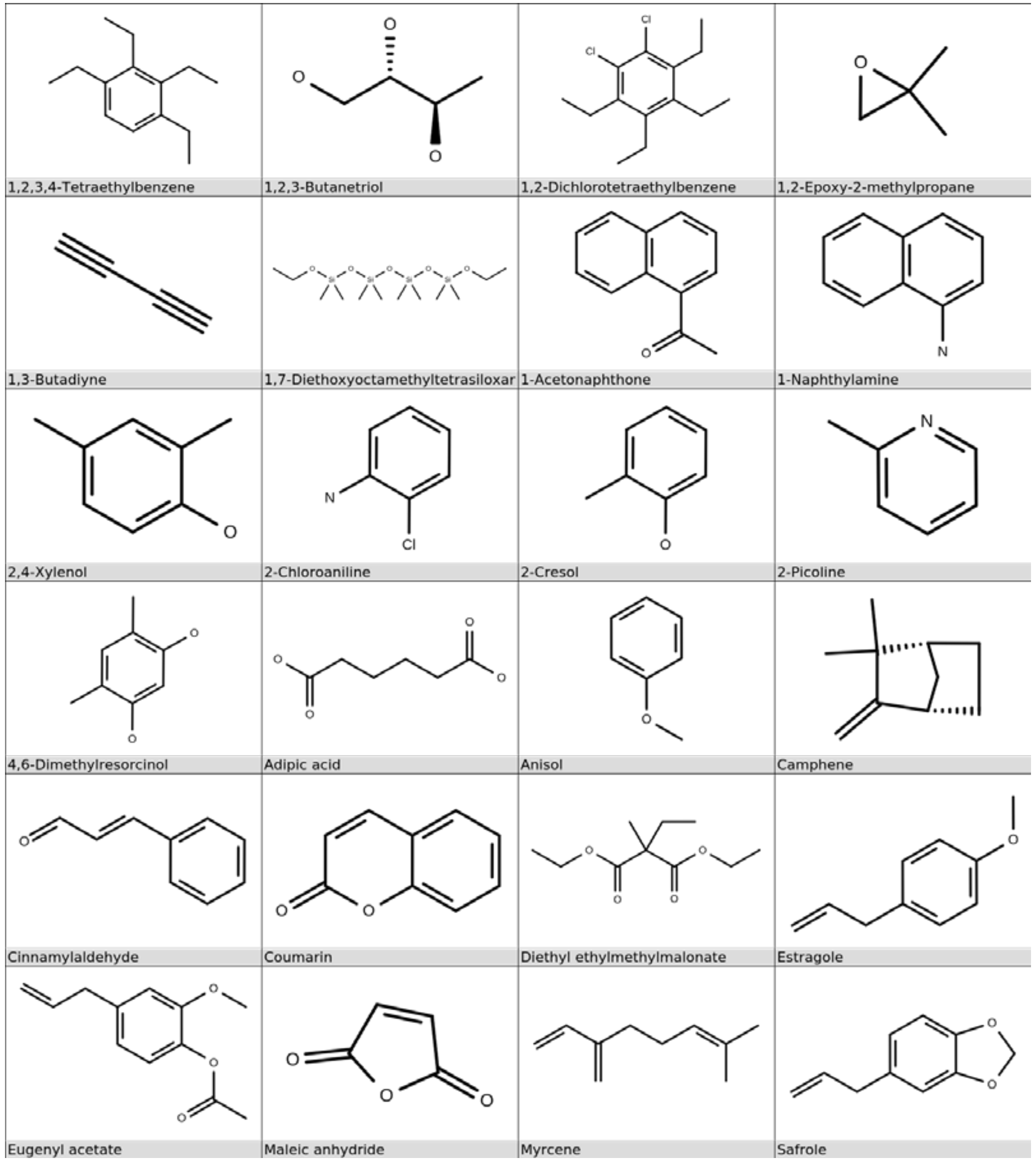
- **i_matsci_optelec_Lmax_(nm) ***
- r_j_Final_Energy
- r_j_Gas_Phase_Energy
- r_j_HOMO *
- r_j_LUMO *
- r_j_QM_Dipole_(debye)
- r_matsci_optelec_Blue_Area
- r_matsci_optelec_Dipole_(D)
- r_matsci_optelec_Green_Area
- r_matsci_optelec_Oxidation_Potential_(eV)
- r_matsci_optelec_Red_Area
- r_matsci_optelec_Reduction_Potential_(eV)
- **r_matsci_optelec_S1_at_S0_(eV) ***
- **r_matsci_optelec_S1_at_S0_Transition_Dipole_(D) ***
- r_matsci_optelec_S2_at_S0_(eV)
- r_matsci_optelec_S2_at_S0_Transition_Dipole_(D)
- r_matsci_optelec_S3_at_S0_(eV)
- r_matsci_optelec_S3_at_S0_Transition_Dipole_(D)
- **r_matsci_optelec_Scaled_Gap_(eV) ***
- r_matsci_optelec_Scaled_HOMO_(eV)
- r_matsci_optelec_Scaled_LUMO_(eV)

*** most impactful via feature importance**

Machine Learning for Volatility of Organic Molecules

Evaporation/Sublimation of Organic Molecules

- Training data:
  - 1,184 organic molecules containing C, O, Cl, N, Si, Br, S, F, P, I, B, As, Se
  - **12,169** experimental (*p*,*T*) datapoints
  - Pressure ranges from 1 Torr to 30 atm
- Generate 200 chemical descriptors and 1000 Morgan Fingerprints for each molecule from its 2D sketch
  - Examples of descriptors: molecular weight, solvent-accessible volume, max partial charge on atoms, electrotopological state descriptors …
- Log(*p*) was used as an additional descriptor and ML model was trained to predict 1/*T*



Benchmarking ML Algorithms

Top-performing machine learning algorithms:

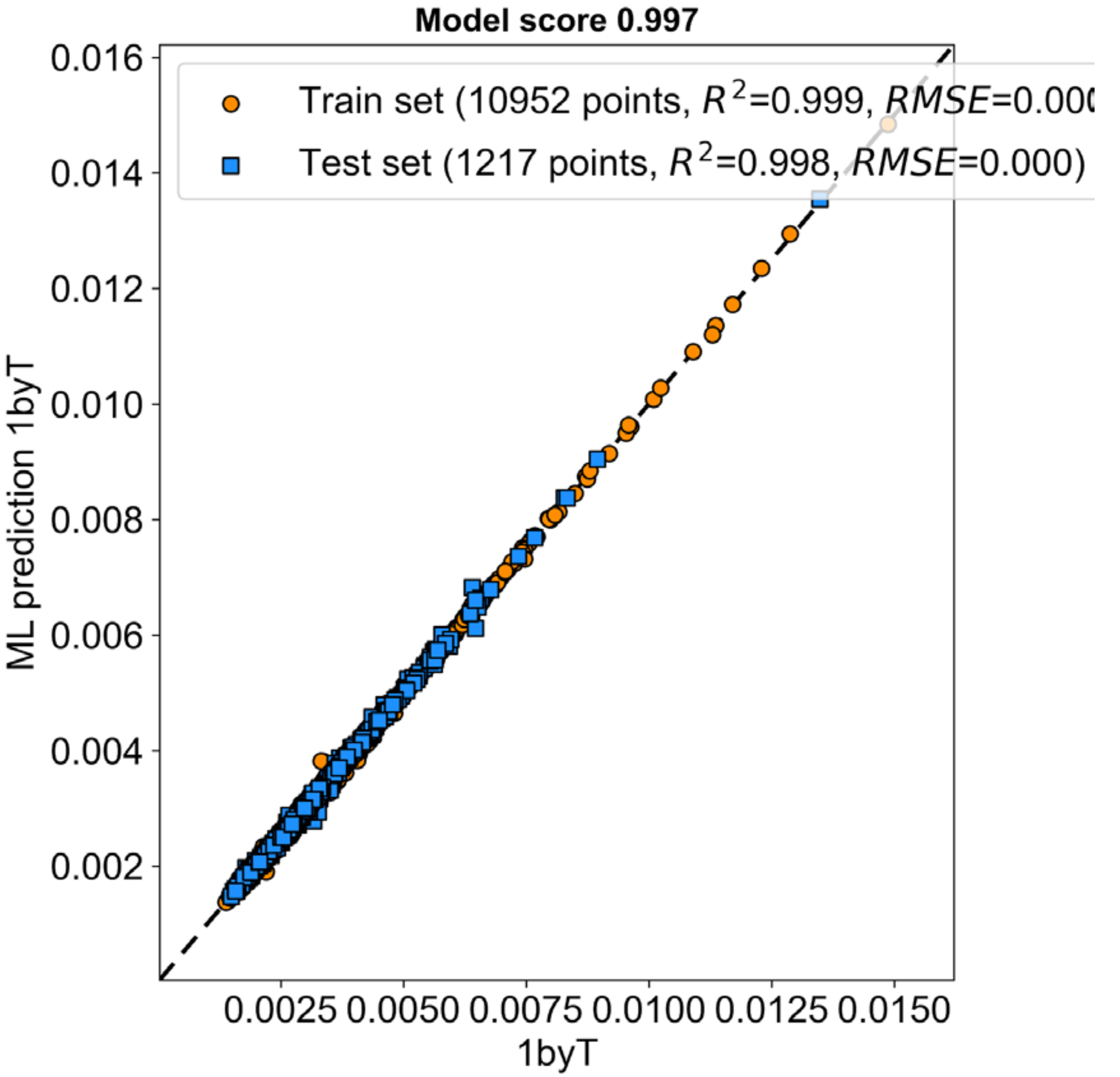
- Light Gradient Boosting Machine (LightGBM)
  - RMS error ±8°C
- Multi-Layer Perceptron (neural network)
  - RMS error ±2°C

Most literature QSPR models for boiling points of diverse organic molecules have errors ±18°C

“Quantitative structure-property relationships for prediction of boiling point, vapor pressure, and melting point”, J. C. Dearden, Environmental Toxicology and Chemistry, 22, 1696–1709 (2003).

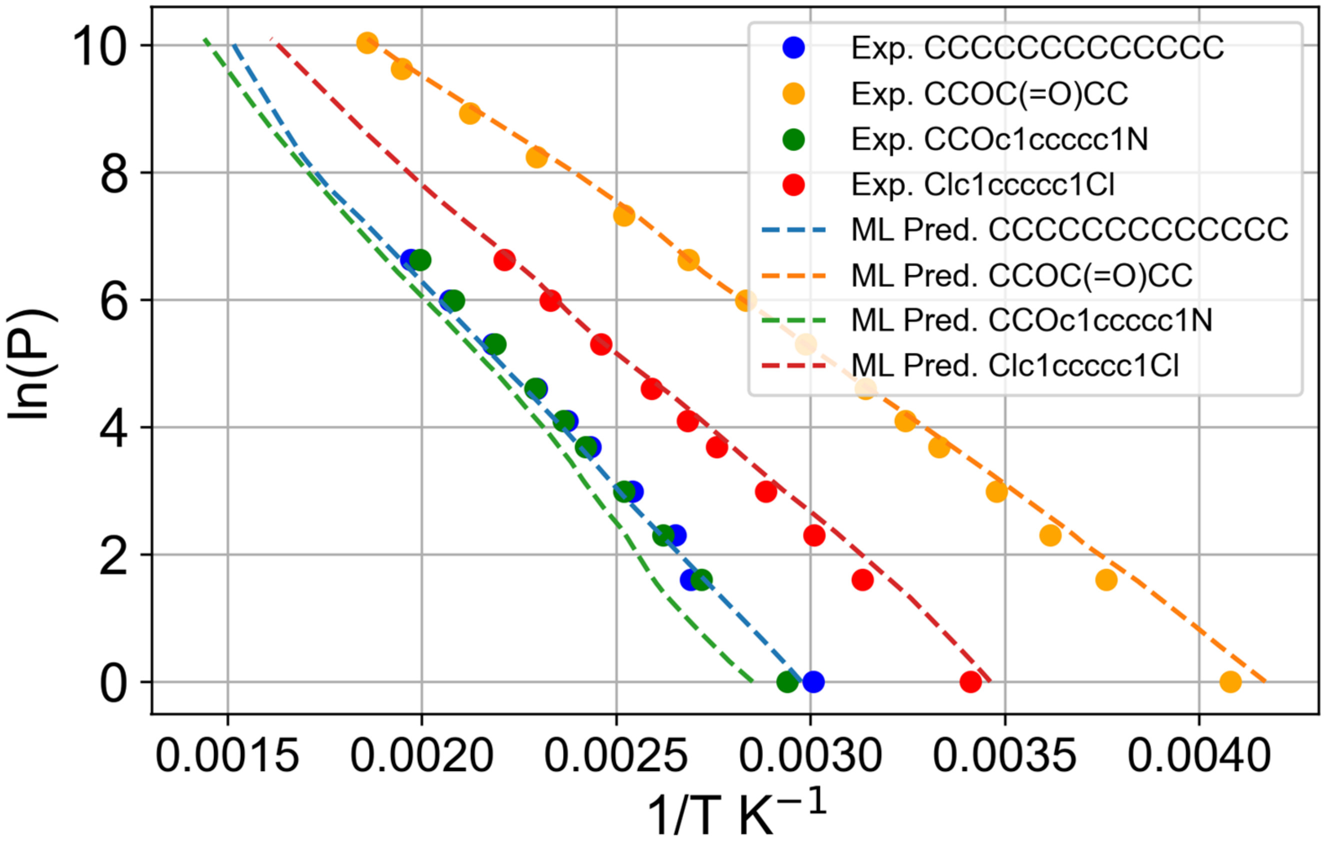
Best neural network gives RMS-error ±5°C and mean absolute error ±4°C

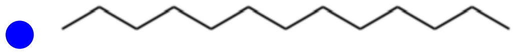
“Boiling point and critical temperature of a heterogeneous data set: QSAR with atom type electrotopological state indices using artificial neural networks”, L. H. Hall & C. T. Story, J. Chem. Inf. Comput. Sci. 36, 1004–1014 (1996).

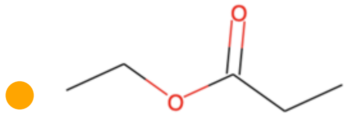


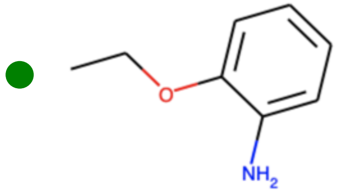
Prediction of Pressure-Temperature Relationships

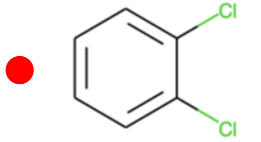
Performance of model on sample molecules *outside* training set











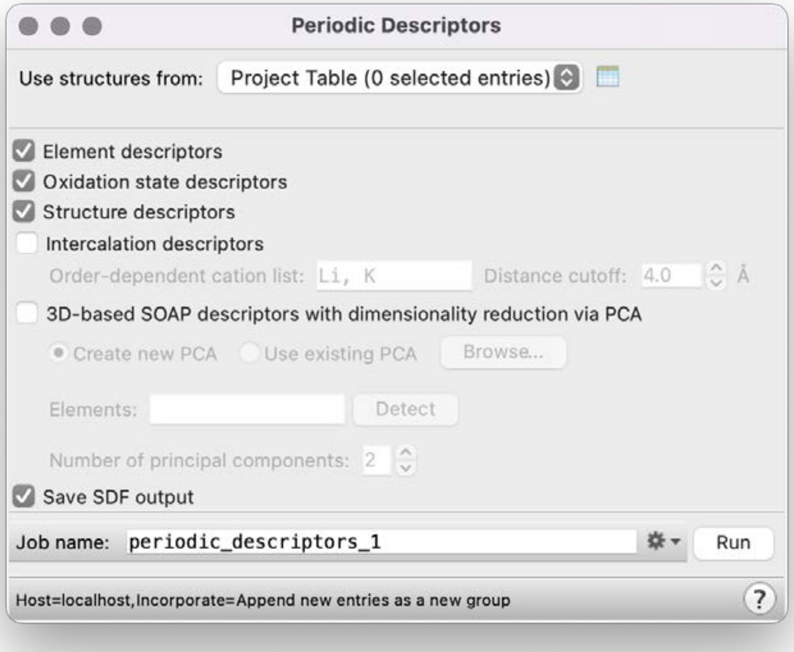
Applications of Volatility Machine Learning

- Atomic Layer Deposition / Chemical Vapor Deposition
- Thermal evaporation & jet-printing (Organic LED)
- Flavors & fragrances
- Equation of state for petroleum fluids
- Refrigerants
- Membrane separation/distillation
- Volatile Organic Compound Pollutants
- Explosion hazards

Machine Learning for Inorganic 3D Crystal Structures

Transparent Conducting Oxide Band Gap ML

- 3000 periodic structures containing indium, aluminum, gallium and oxygen
- These materials have applications in display devices and solar-cells
- Dataset was obtained from NOMAD 2018 Kaggle challenge on creating ML models for properties of transparent conducting oxides
- ML models for **Band Gap** were created using DeepAutoQSAR
- Composition (matminer) and 3D SOAP descriptors were used



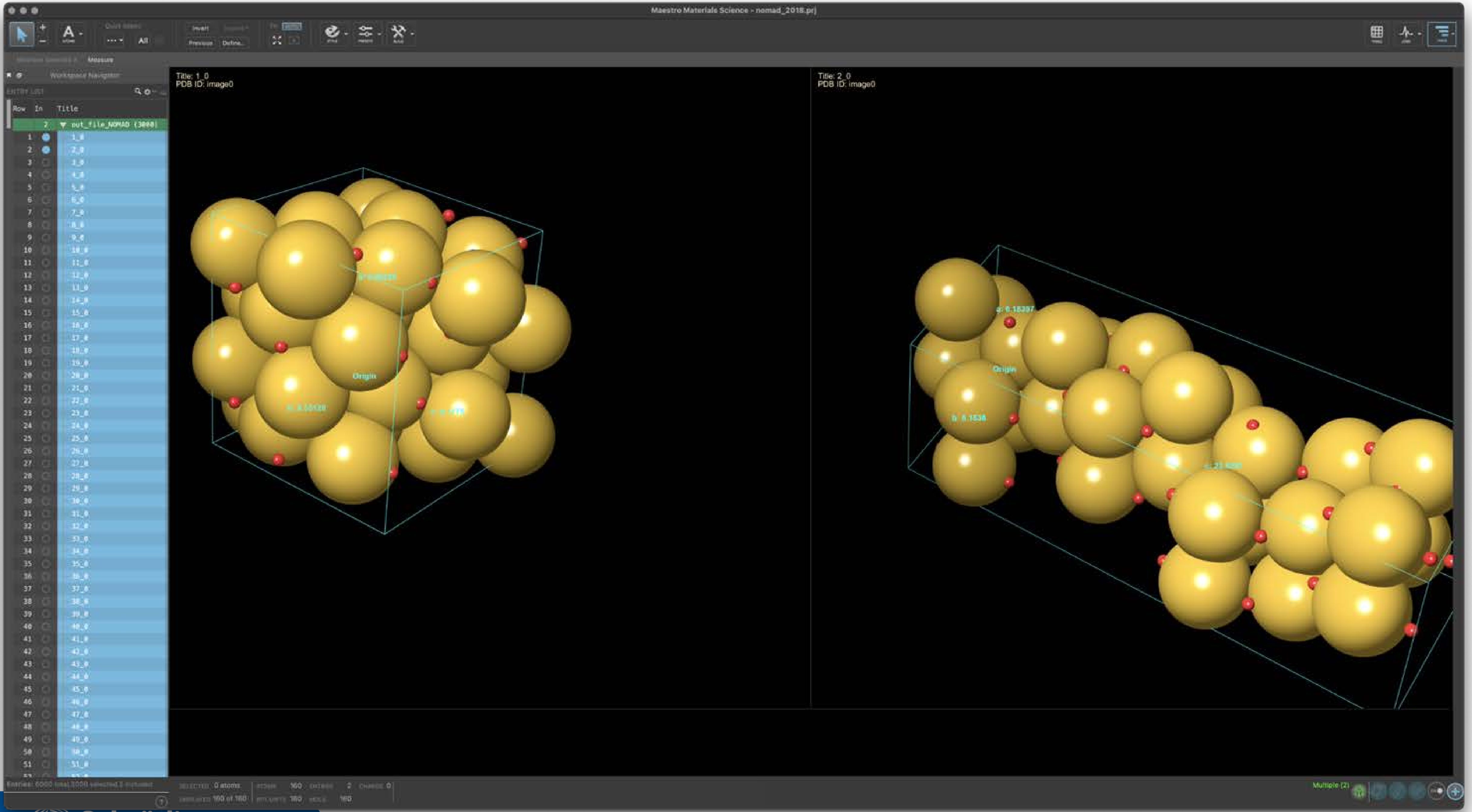
Periodic Descriptors

Use structures from: Project Table (0 selected entries)

☑ Element descriptors
☑ Oxidation state descriptors
☑ Structure descriptors
□ Intercalation descriptors
Order-dependent cation list: Li, K Distance cutoff: 4.0 Å
□ 3D-based SOAP descriptors with dimensionality reduction via PCA
● Create new PCA ○ Use existing PCA Browse...
Elements: Detect
Number of principal components: 2
☑ Save SDF output

Job name: periodic_descriptors_1 Run

Host=localhost,Incorporate=Append new entries as a new group



Maestro Materials Science - nomad_2018.prj

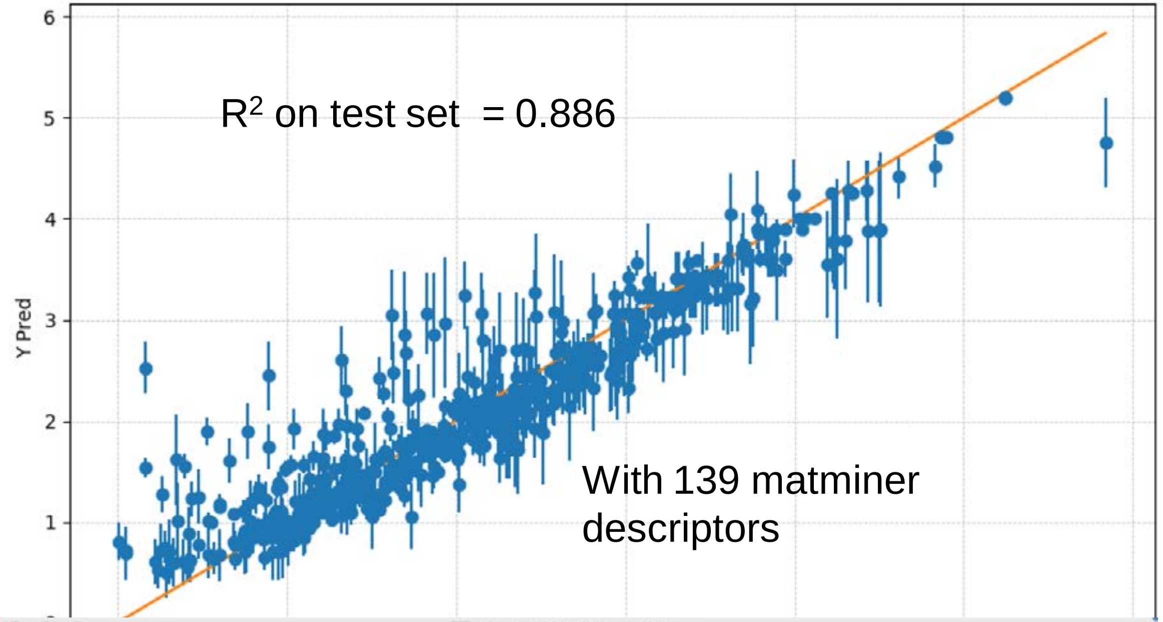
Row	In	Title
2		out_file_NOMAD (3000)
1	●	1_0
2	●	2_0
3	○	3_0
4	○	4_0
5	○	5_0
6	○	6_0
7	○	7_0
8	○	8_0
9	○	9_0
10	○	10_0
11	○	11_0
12	○	12_0
13	○	13_0
14	○	14_0
15	○	15_0
16	○	16_0
17	○	17_0
18	○	18_0
19	○	19_0
20	○	20_0
21	○	21_0
22	○	22_0
23	○	23_0
24	○	24_0
25	○	25_0
26	○	26_0
27	○	27_0
28	○	28_0
29	○	29_0
30	○	30_0
31	○	31_0
32	○	32_0
33	○	33_0
34	○	34_0
35	○	35_0
36	○	36_0
37	○	37_0
38	○	38_0
39	○	39_0
40	○	40_0
41	○	41_0
42	○	42_0
43	○	43_0
44	○	44_0
45	○	45_0
46	○	46_0
47	○	47_0
48	○	48_0
49	○	49_0
50	○	50_0
51	○	51_0

Title: 1_0
PDB ID: image0

Title: 2_0
PDB ID: image0

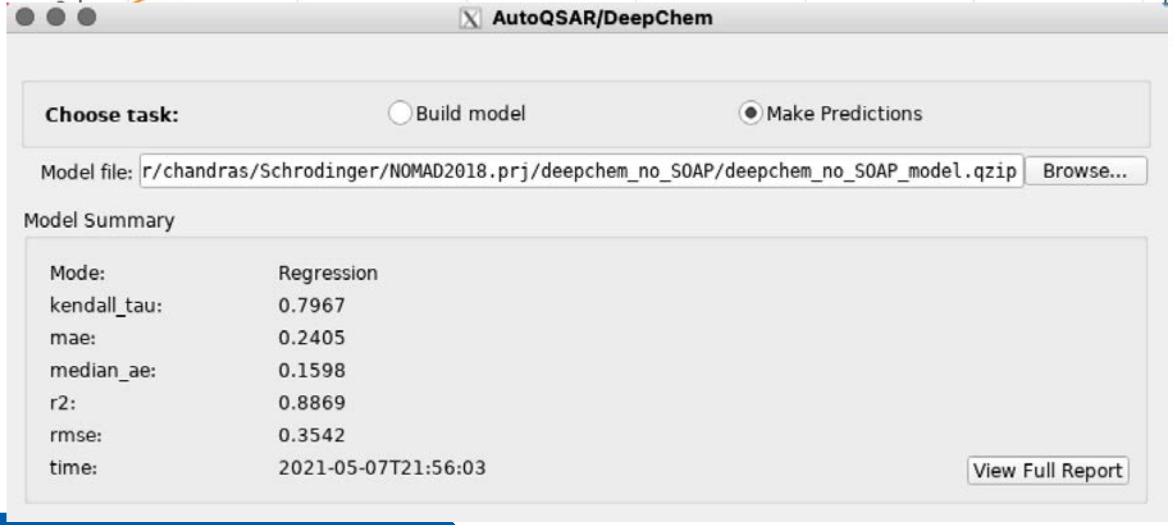
SELECTED 0 atoms
ATOMS 160 ENTRIES 2 CHARGE 0
DISPLAYED 160 of 160
MOLECULES 160

DeepAutoQSAR Results



R^2 on test set = 0.886

With 139 matminer descriptors

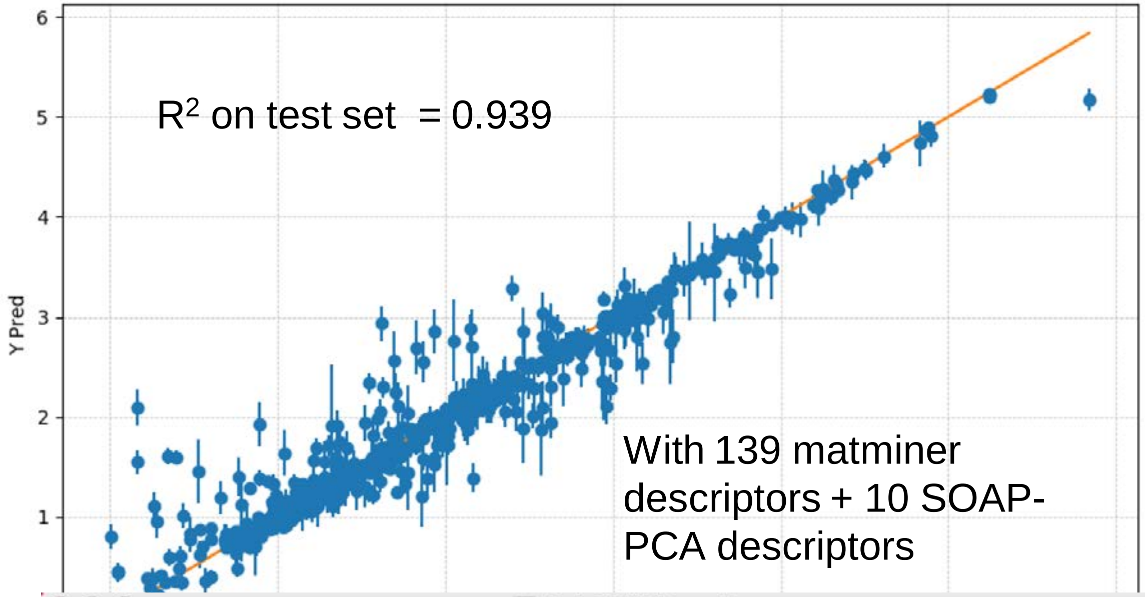


Choose task: ○ Build model ● Make Predictions

Model file: r/chandras/Schrodinger/NOMAD2018.prj/deepchem_no_SOAP/deepchem_no_SOAP_model.qzip

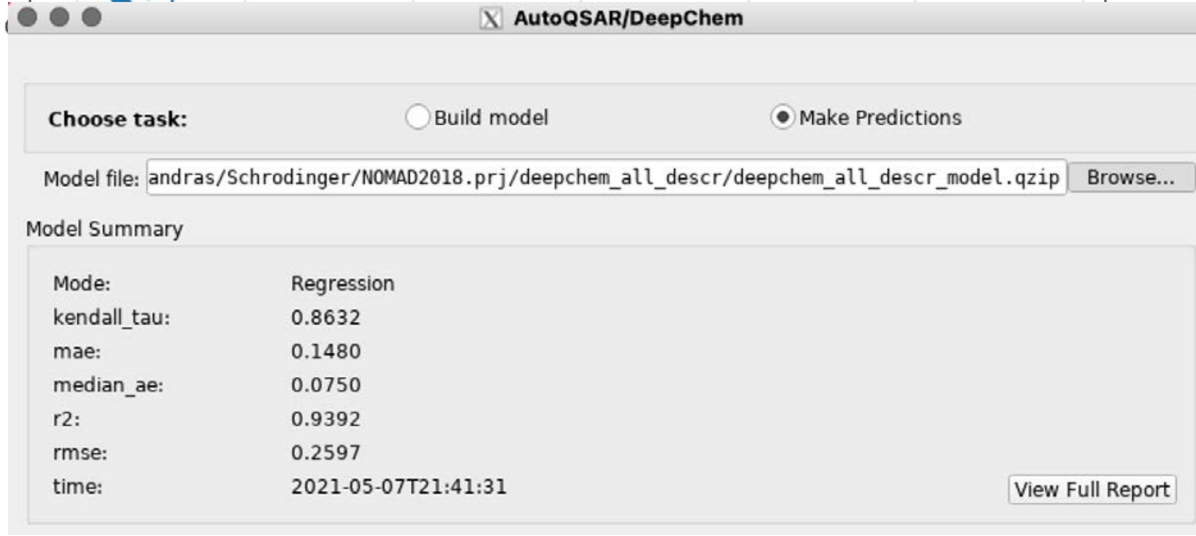
Model Summary

Mode:	Regression
kendall_tau:	0.7967
mae:	0.2405
median_ae:	0.1598
r2:	0.8869
rmse:	0.3542
time:	2021-05-07T21:56:03



R^2 on test set = 0.939

With 139 matminer descriptors + 10 SOAP-PCA descriptors



Choose task: ○ Build model ● Make Predictions

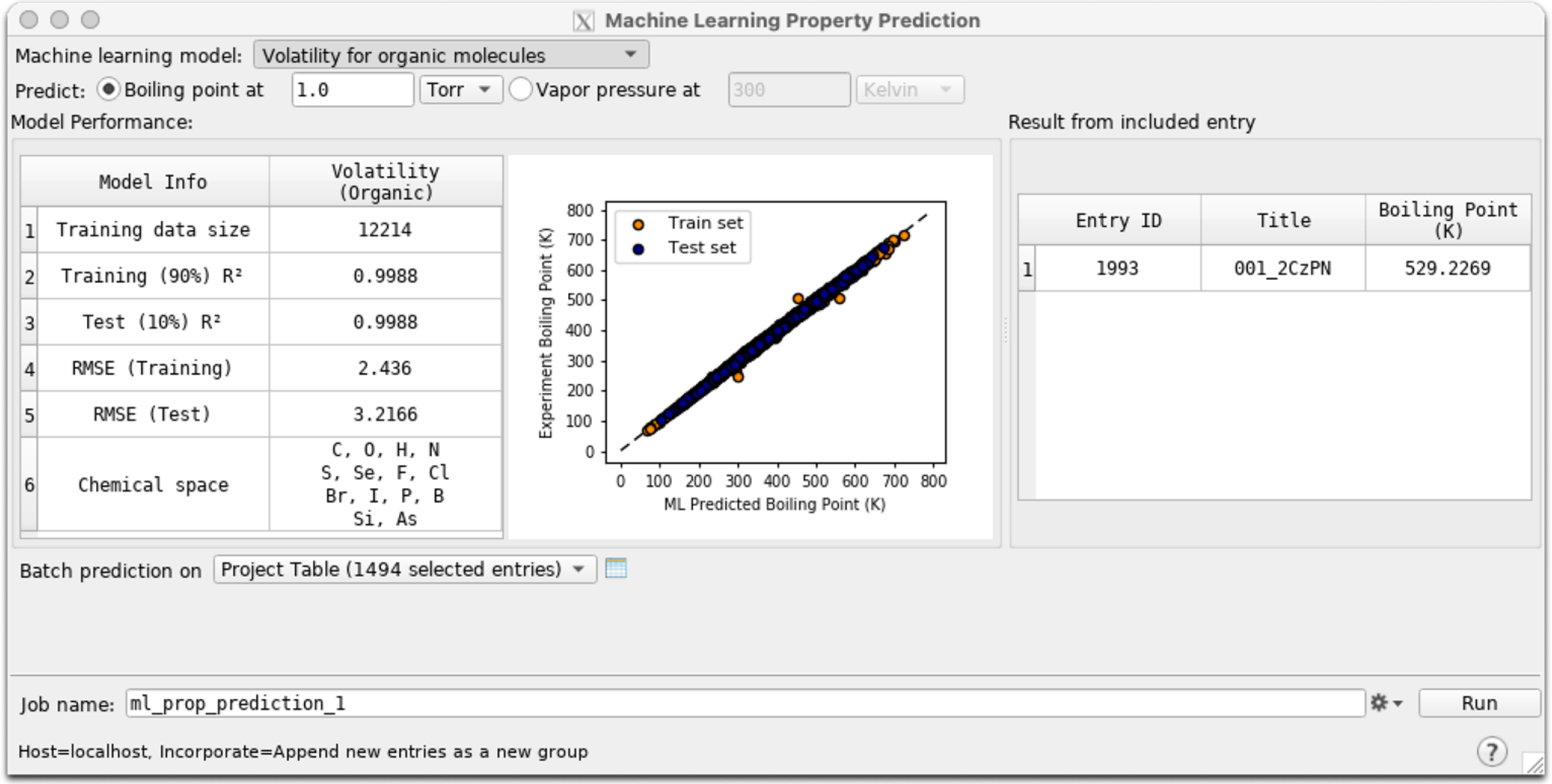
Model file: andras/Schrodinger/NOMAD2018.prj/deepchem_all_descr/deepchem_all_descr_model.qzip

Model Summary

Mode:	Regression
kendall_tau:	0.8632
mae:	0.1480
median_ae:	0.0750
r2:	0.9392
rmse:	0.2597
time:	2021-05-07T21:41:31

Machine Learning Property Prediction Panel

- The following properties/models are currently available
  - Volatility of organic molecules (Both Boiling Point and Vapor pressure)
  - Volatility of organometallic molecules
  - Polymer Tg
  - Frequency dependent Df
  - Frequency dependent Dk
  - Density
  - Viscosity



Machine Learning Property Prediction

Machine learning model: Volatility for organic molecules

Predict: ● Boiling point at 1.0 Torr ○ Vapor pressure at 300 Kelvin

Model Performance:

	Model Info	Volatility (Organic)
1	Training data size	12214
2	Training (90%) R²	0.9988
3	Test (10%) R²	0.9988
4	RMSE (Training)	2.436
5	RMSE (Test)	3.2166
6	Chemical space	C, O, H, N S, Se, F, Cl Br, I, P, B Si, As

Result from included entry

	Entry ID	Title	Boiling Point (K)
1	1993	001_2CzPN	529.2269

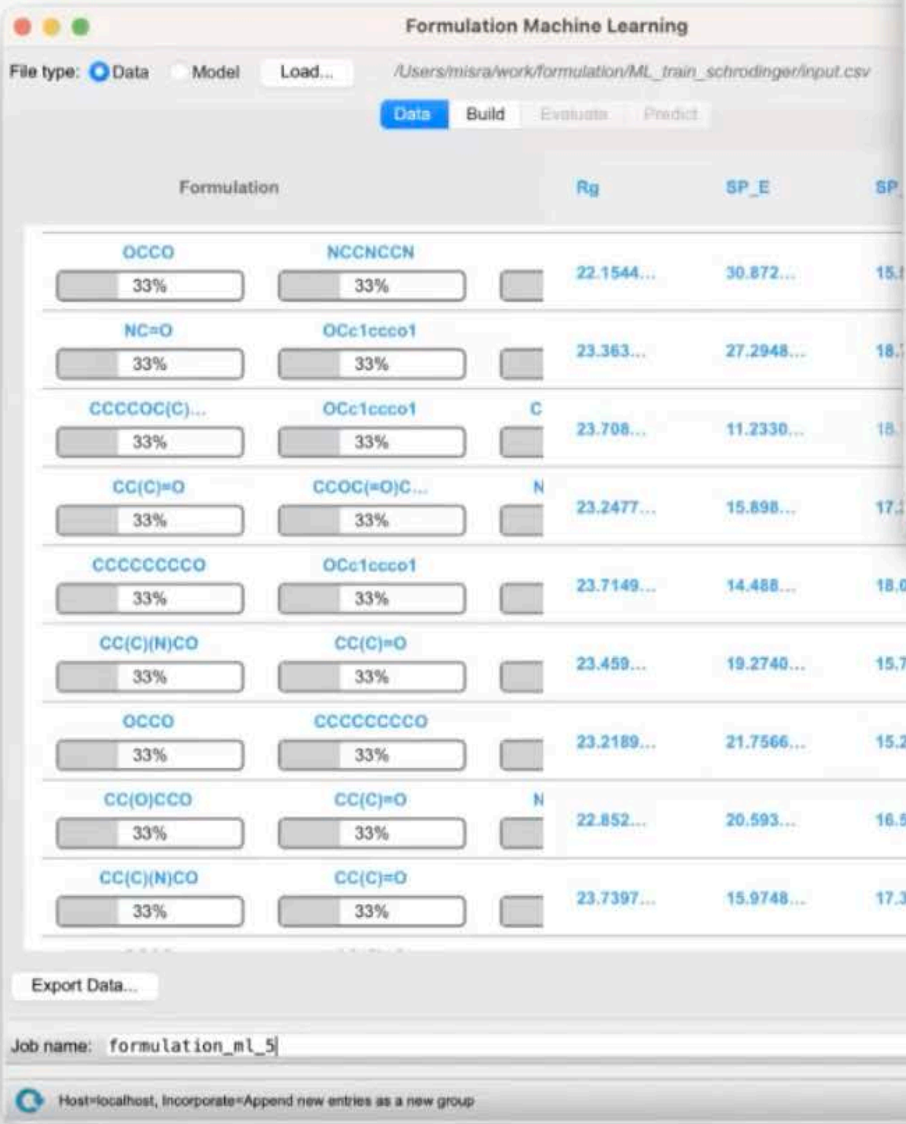
Batch prediction on Project Table (1494 selected entries)

Job name: ml_prop_prediction_1

Host=localhost, Incorporate=Append new entries as a new group

ML for Formulations

- Create ML models for mixtures and formulations containing multiple molecules
- Identify most import features and descriptors
- Design/optimize new formulations with novel compositions and chemistries



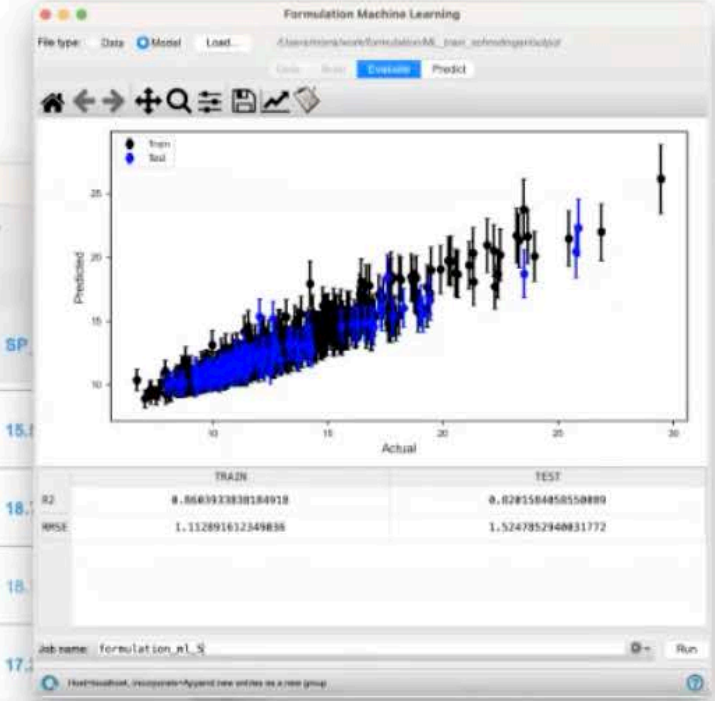
Formulation Machine Learning

File type: ● Data ○ Model Load... /Users/misra/work/formulation/ML_train_schrodinger/input.csv

Formulation			Rg	SP_E	SP	
OCCO 33%	NCCNCCN 33%		22.1544...	30.872...	15.	
NC=O 33%	OCc1ccco1 33%		23.363...	27.2948...	18.	
CCCCOC(C)... 33%	OCc1ccco1 33%	C	23.708...	11.2330...	18.	
CC(C)=O 33%	CCOC(=O)C... 33%	N	23.2477...	15.898...	17.	
CCCCCCCCO 33%	OCc1ccco1 33%		23.7149...	14.488...	18.0358...	23.1
CC(C)(N)CO 33%	CC(C)=O 33%		23.459...	19.2740...	15.7063...	24.8
OCCO 33%	CCCCCCCCO 33%		23.2189...	21.7566...	15.2313...	26.5
CC(O)CCO 33%	CC(C)=O 33%	N	22.852...	20.593...	16.5877...	26.4
CC(C)(N)CO 33%	CC(C)=O 33%		23.7397...	15.9748...	17.3363...	23.5

Job name: formulation_ml_5

Host=localhost, Incorporate=Append new entries as a new group



Formulation Machine Learning

File type: ○ Data ● Model Load...

	TRAIN	TEST
R2	0.8603933830184918	0.8201584058550089
RMSE	1.112891612349836	1.5247852940031772

Job name: formulation_ml_5

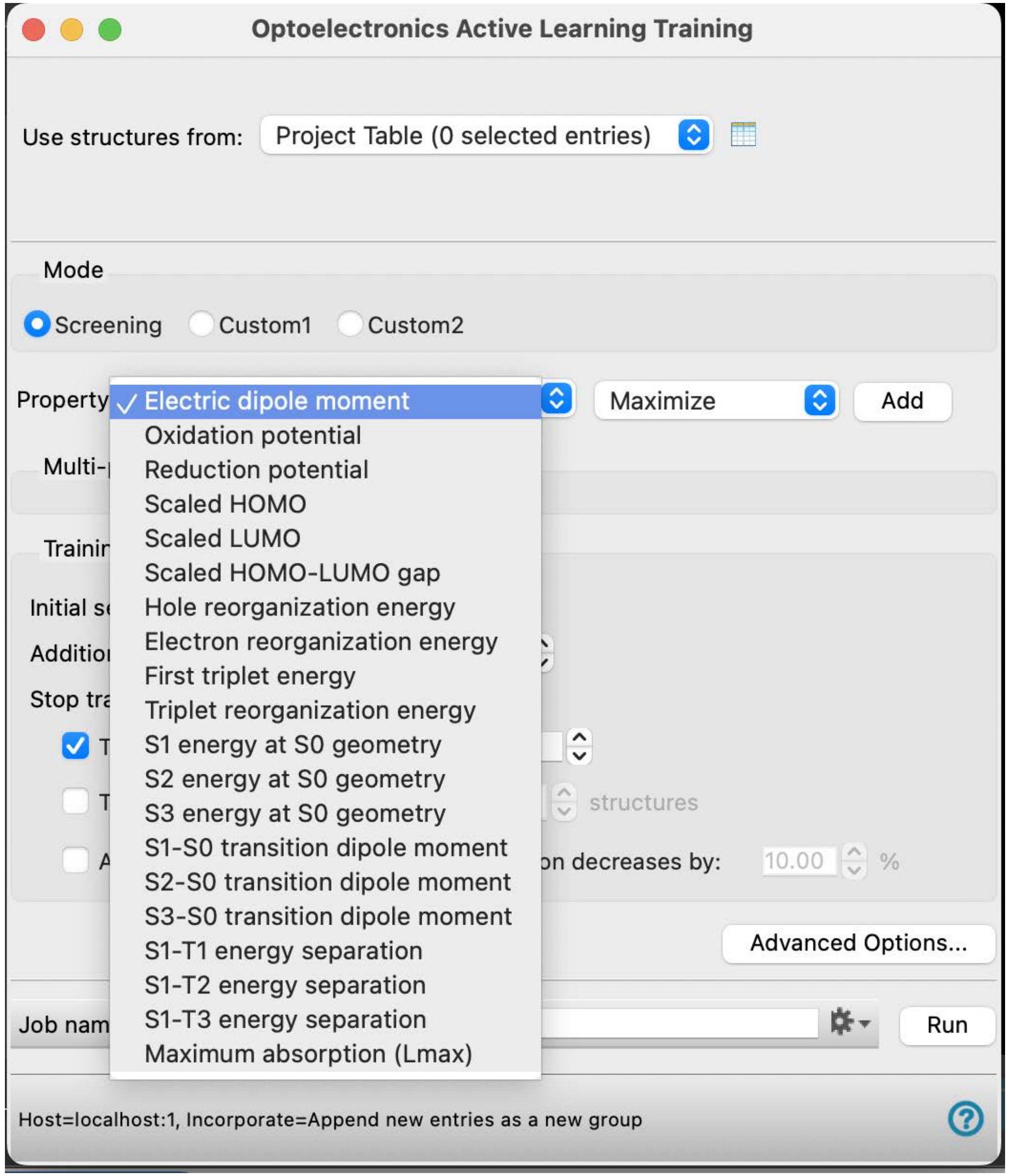
Host=localhost, Incorporate=Append new entries as a new group

Scheduled for 24-1

The technical features and projected timeline presented on this slide is for discussion purposes only. Such planned or potential capabilities are subject to change at any time.

Active Learning and Genetic Optimization

Active Learning OptoElectronics Multi-Parameter Optimization (MPO)



Optoelectronics Active Learning Training

Use structures from: Project Table (0 selected entries)

Mode

● Screening ○ Custom1 ○ Custom2

Property ✓ Electric dipole moment | Maximize | Add

- Oxidation potential
- Reduction potential
- Scaled HOMO
- Scaled LUMO
- Scaled HOMO-LUMO gap
- Hole reorganization energy
- Electron reorganization energy
- First triplet energy
- Triplet reorganization energy
- S1 energy at S0 geometry
- S2 energy at S0 geometry
- S3 energy at S0 geometry
- S1-S0 transition dipole moment
- S2-S0 transition dipole moment
- S3-S0 transition dipole moment
- S1-T1 energy separation
- S1-T2 energy separation
- S1-T3 energy separation
- Maximum absorption (Lmax)

Multi-|

Trainir

Initial s

Additior

Stop tra

☑ T

□ T structures

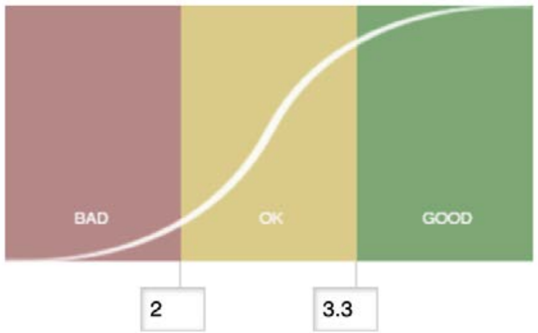
□ A on decreases by: 10.00 %

Advanced Options...

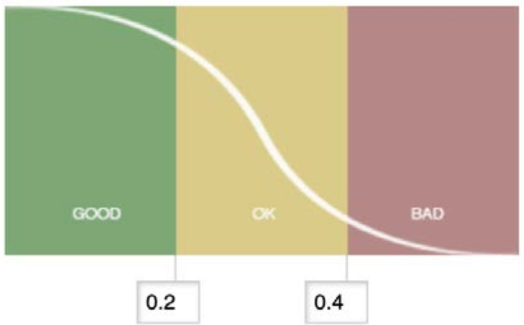
Job nam Run

Host=localhost:1, Incorporate=Append new entries as a new group

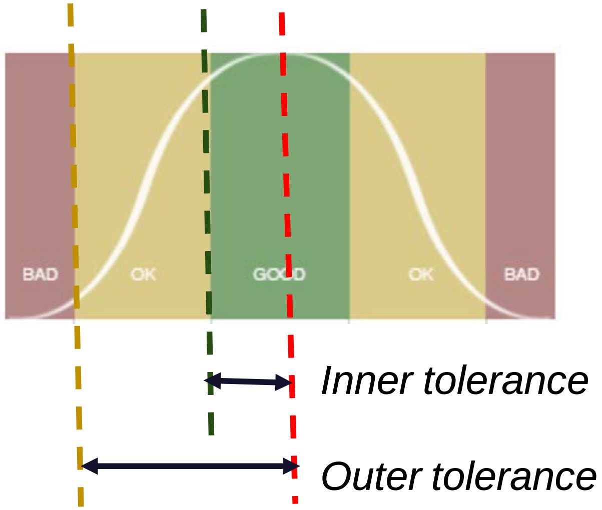
Maximize



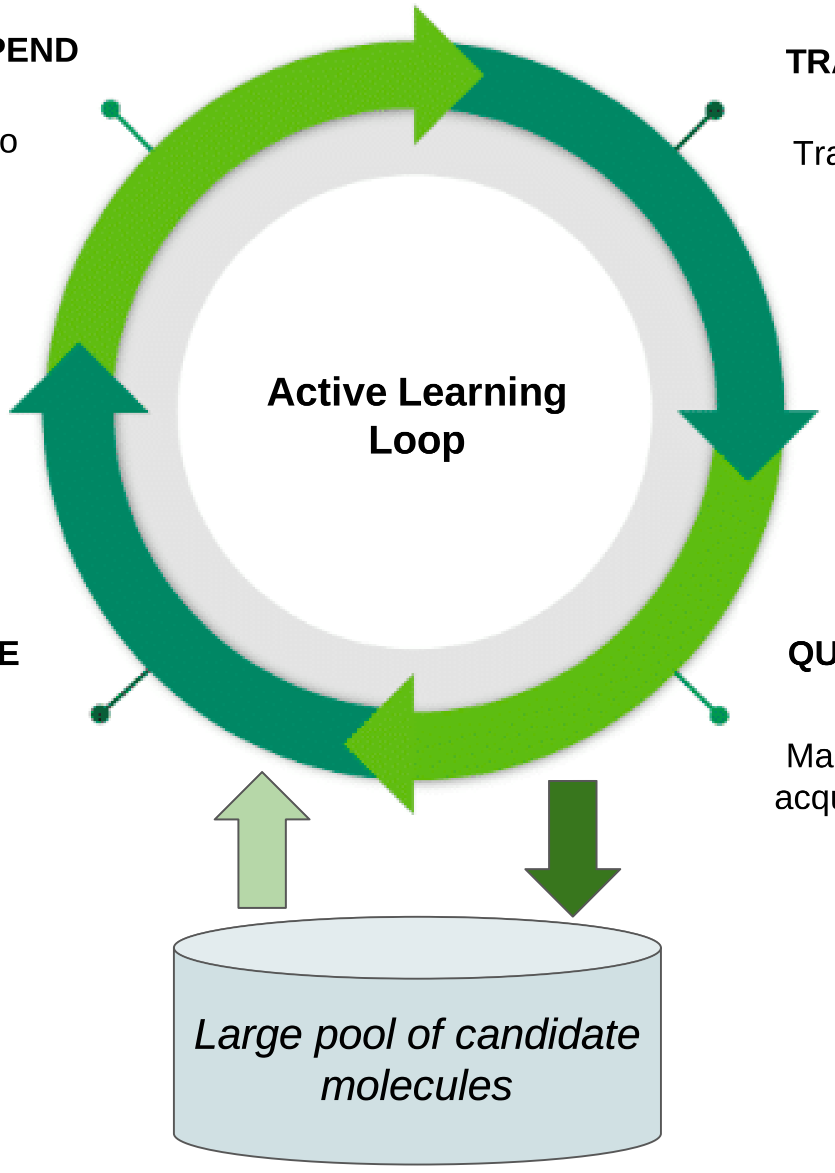
Minimize



Target Value



Active Learning Workflow for OptoElectronics



APPEND

Add new molecules to Training Set

TRAIN

Train ML models on new Training Set

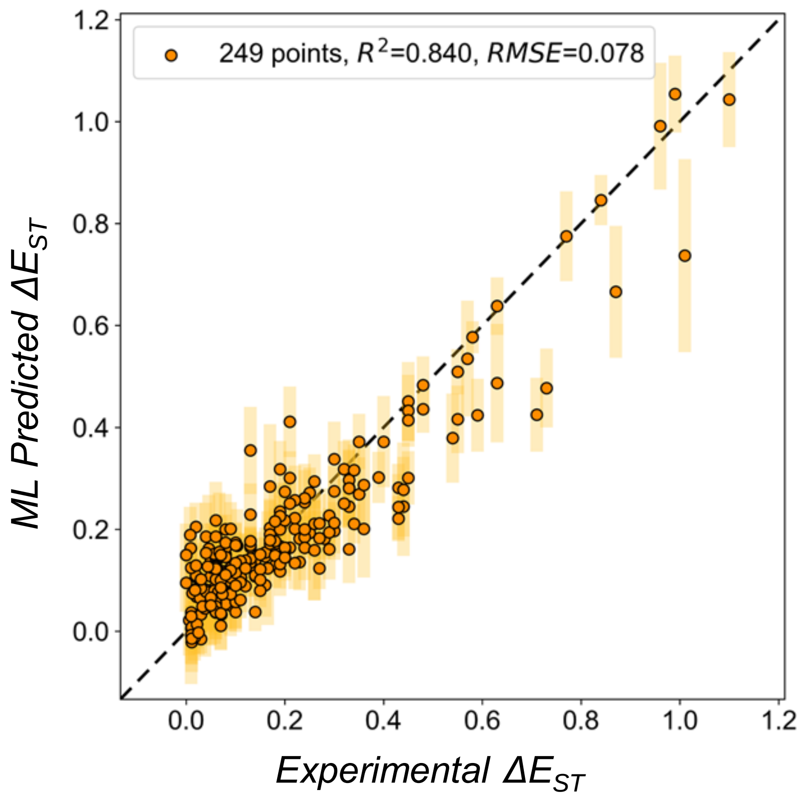
QUERY

Make MPO predictions and use acquisition function to select new molecules

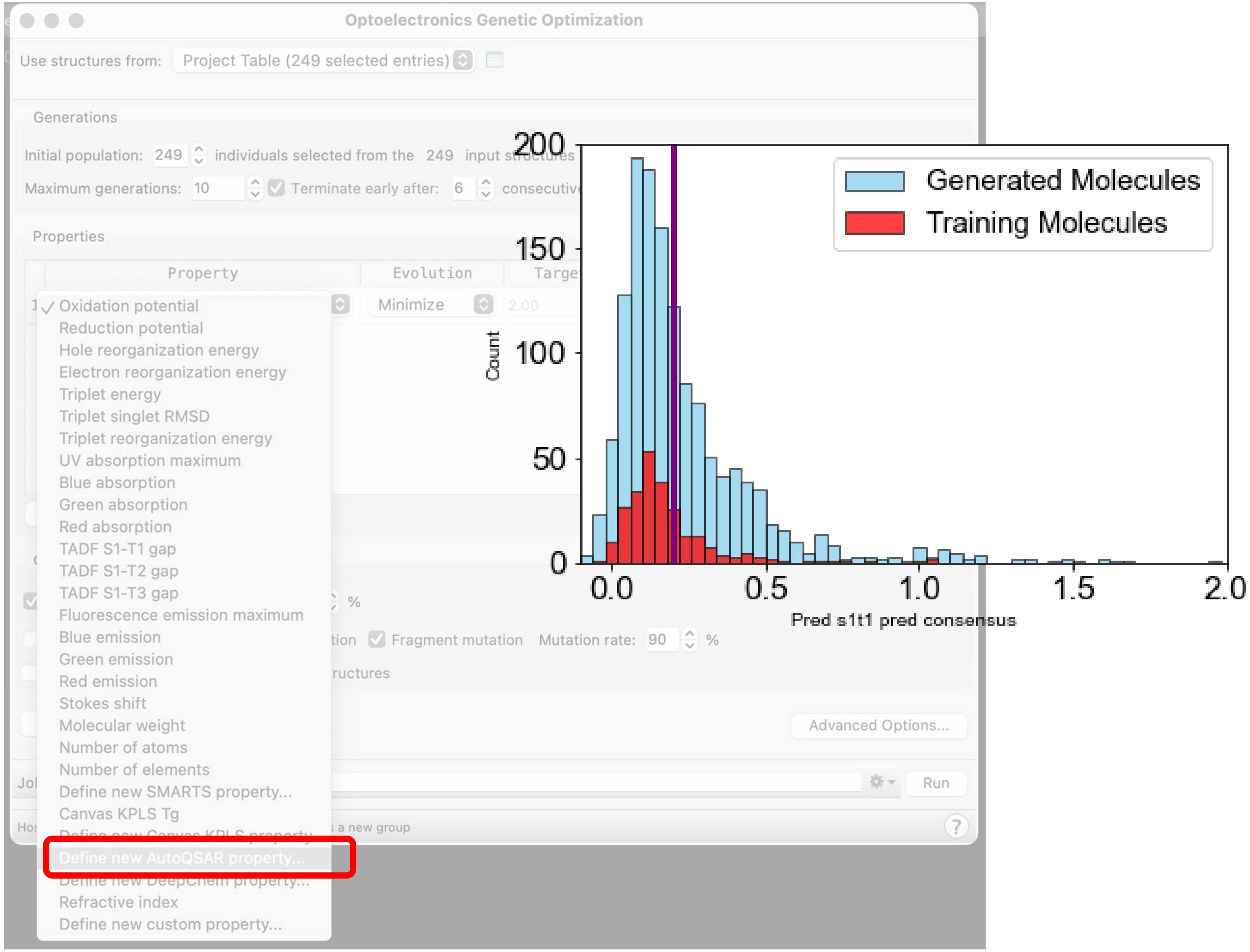
COMPUTE

Calculate DFT properties of selected molecules

Optoelectronic Genetic Optimization



AutoQSAR consensus prediction for experimental ΔE_{ST} dataset



Optoelectronics Genetic Optimization

Use structures from: Project Table (249 selected entries)

Generations

Initial population: 249 individuals selected from the 249 input structures

Maximum generations: 10 ☑ Terminate early after: 6 consecutiv

Properties

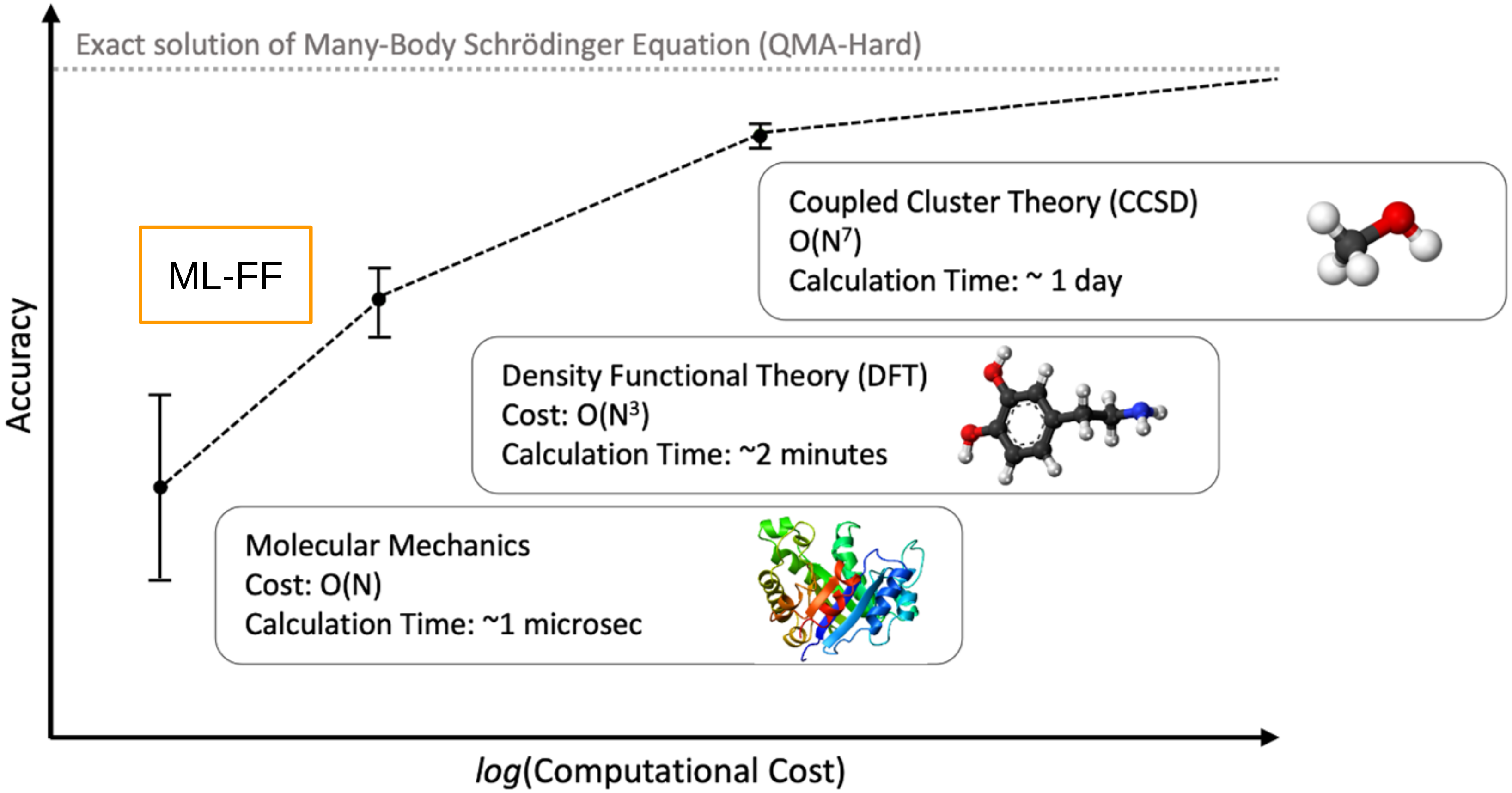
	Property	Evolution	Targe
1	Oxidation potential	Minimize	2.00

tion ☑ Fragment mutation Mutation rate: 90 %

ructures

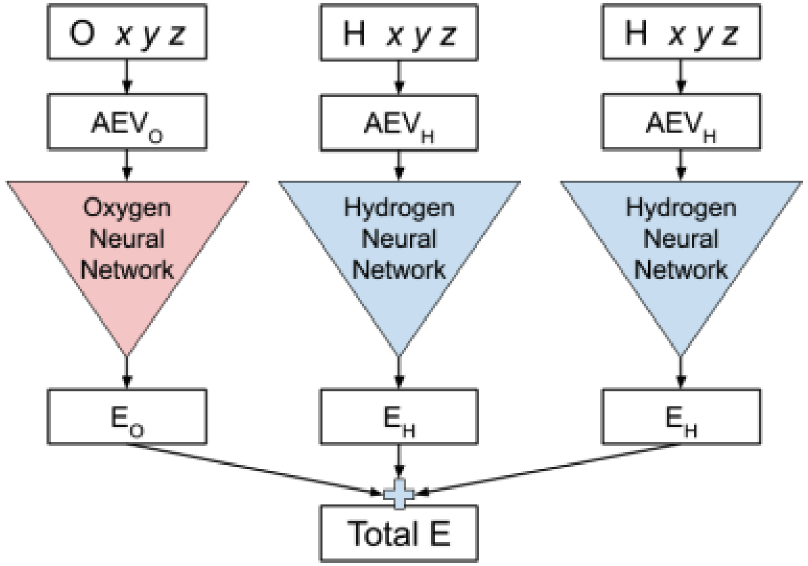
Machine Learning Forcefields

Neural Network Potentials (NNPs)



Achieving QM accuracy at the cost of classical forcefields is an exciting prospect for neural network potentials to accelerate design of next-generation materials

Our First NN Model: Schrödinger-ANI (SANI)



O xyz → AEV_O → Oxygen Neural Network → E_O

H xyz → AEV_H → Hydrogen Neural Network → E_H

H xyz → AEV_H → Hydrogen Neural Network → E_H

Total E

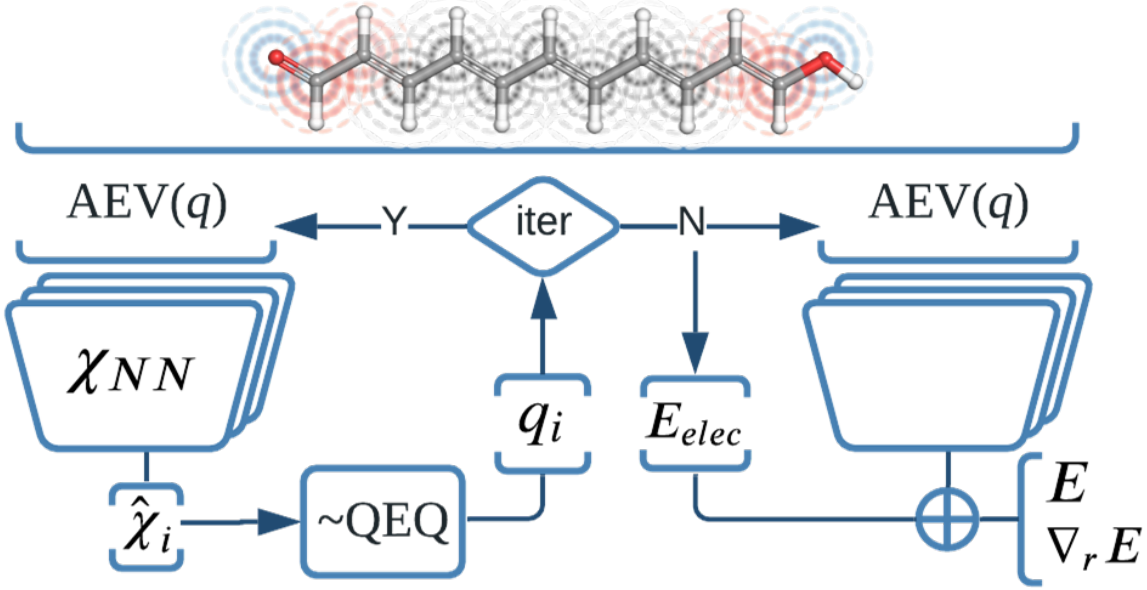
AEV_H	r_{HH}	r_{HC}	r_{HN}	r_{HO}	θ_{HHH}	θ_{HHC}	θ_{HHN}	θ_{HHO}	θ_{CHH}	...
AEV_C	r_{CH}	r_{CC}	r_{CN}	r_{CO}	θ_{HCH}	θ_{HCC}	θ_{HCN}	θ_{HCO}	θ_{CCH}	...
AEV_N	r_{NH}	r_{NC}	r_{NN}	r_{NO}	θ_{HNH}	θ_{HNC}	θ_{HNN}	θ_{HNO}	θ_{CNH}	...
AEV_O	r_{OH}	r_{OC}	r_{ON}	r_{OO}	θ_{HOH}	θ_{HOC}	θ_{HON}	θ_{HOO}	θ_{COH}	...

Stevenson et. al., arXiv, 1912.05079 (2019)

- SANI is extension of ANI[1]-family of NN potential[2]
- Supports 8 elements covering 94% druglike molecules in ChEMBL[3]
- Inputs are cartesian coordinates and element type for each atom
- Each element has a separate NN learning mapping from features to energies
- Trained to DFT energies (wB97X/6-31G(d))
- Limitations: Neglect long-range effects, no information about charge state, charge distribution

[1] *Smith et. al.,Chem. Sci., 8, 3192-3203 (2017)*
[2] *Behler et. al., Phys. Rev. Lett., 98, 146401 (2007)*
[3] *Gaulton et. al., Nucleic Acids Res., 45, D945-D954 (2016)*

QRNN: Charge-Recursive Neural Network

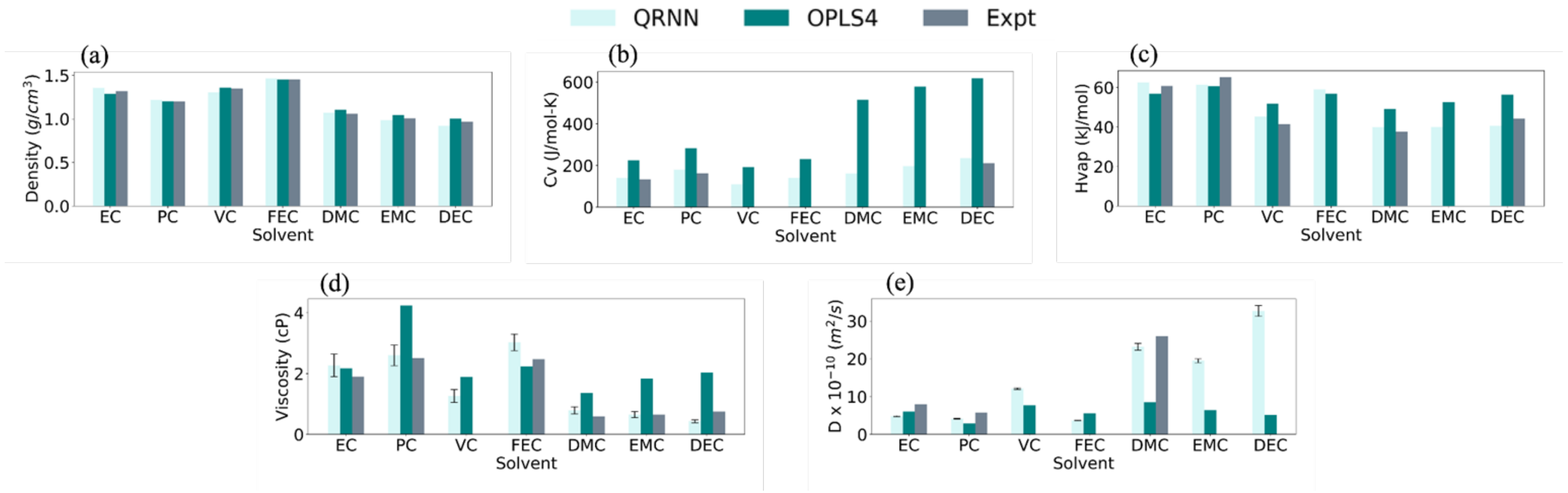


Jacobson et. al., J. Chem. Theory Comput. 18, 2354-2366 (2022)

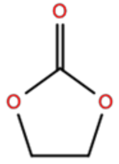
Predicts energy, atomic charges, dipole and atomic forces

- Extension of SANI to provide support for ionic systems
- Involves recursive charge correction
- Predicted atomic charges added as features to the NN
- Include charge dependent AEVs to learn radial charge distribution
- Empirical dispersion correction and coulomb interaction using the predicted charges added to energy

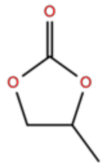
Bulk Properties of Liquid Electrolytes



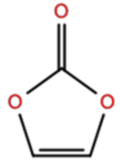
EC



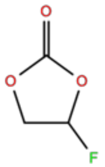
PC



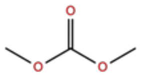
VC



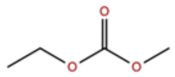
FEC



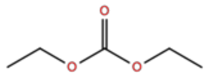
DMC



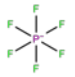
EMC



DEC



PF6-



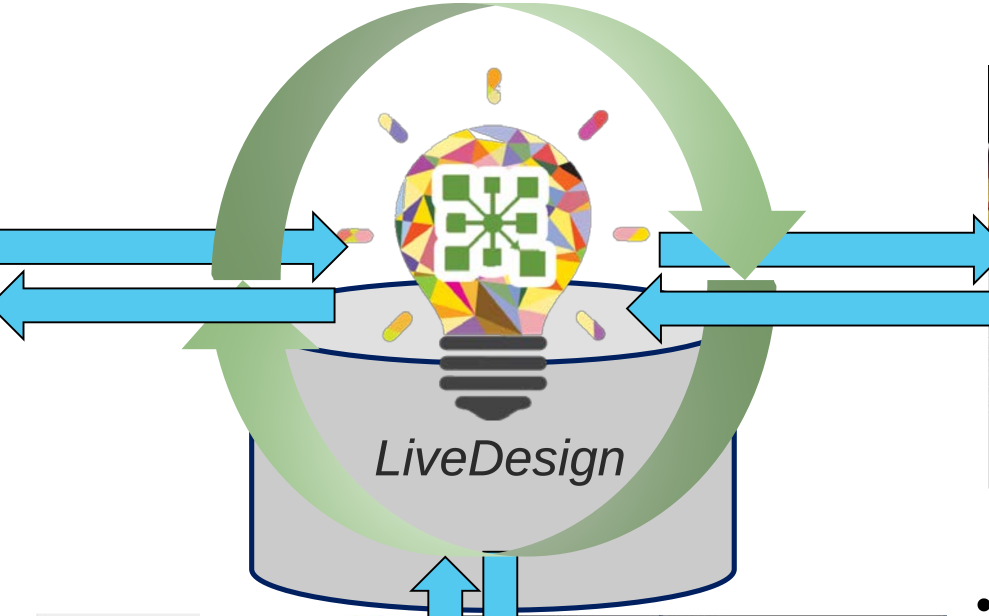
- ML-FF computed bulk properties are compared to experiments and OPLS4
- Excellent property predictions for all electrolyte components
- Significant improvement in prediction of diffusivity and viscosity compared to OPLS4

The technical features and projected timeline presented on this slide is for discussion purposes only. Such planned or potential capabilities are subject to change at any time.

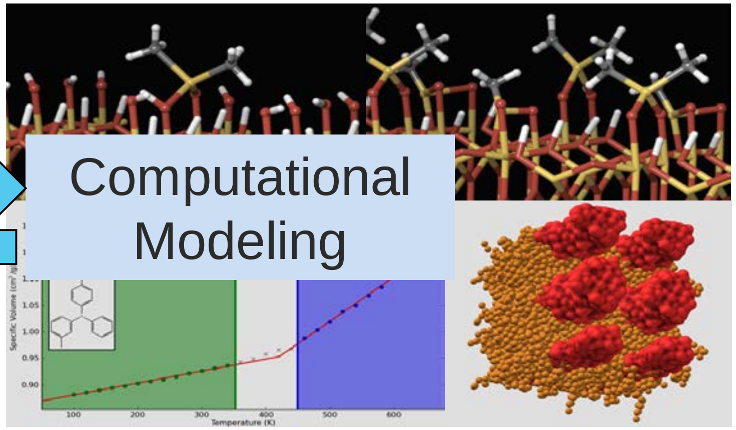
Enterprise Informatics

Schrodinger’s Informatics Platform - LiveDesign®

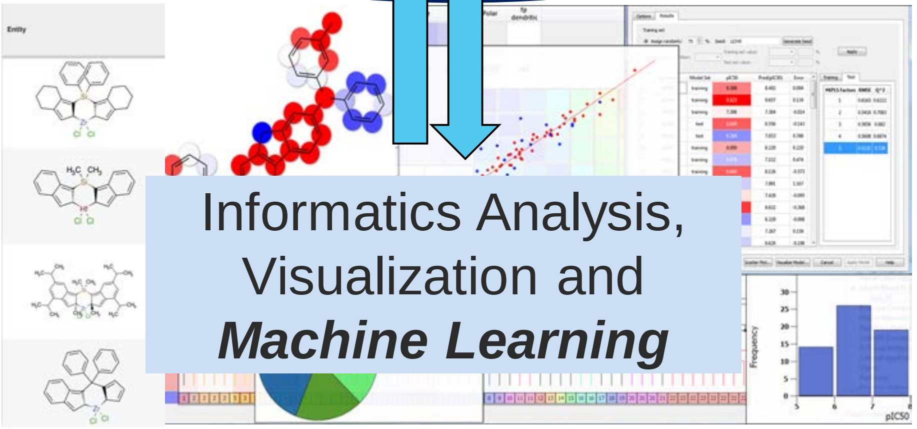
Experimental data



LiveDesign



Computational Modeling



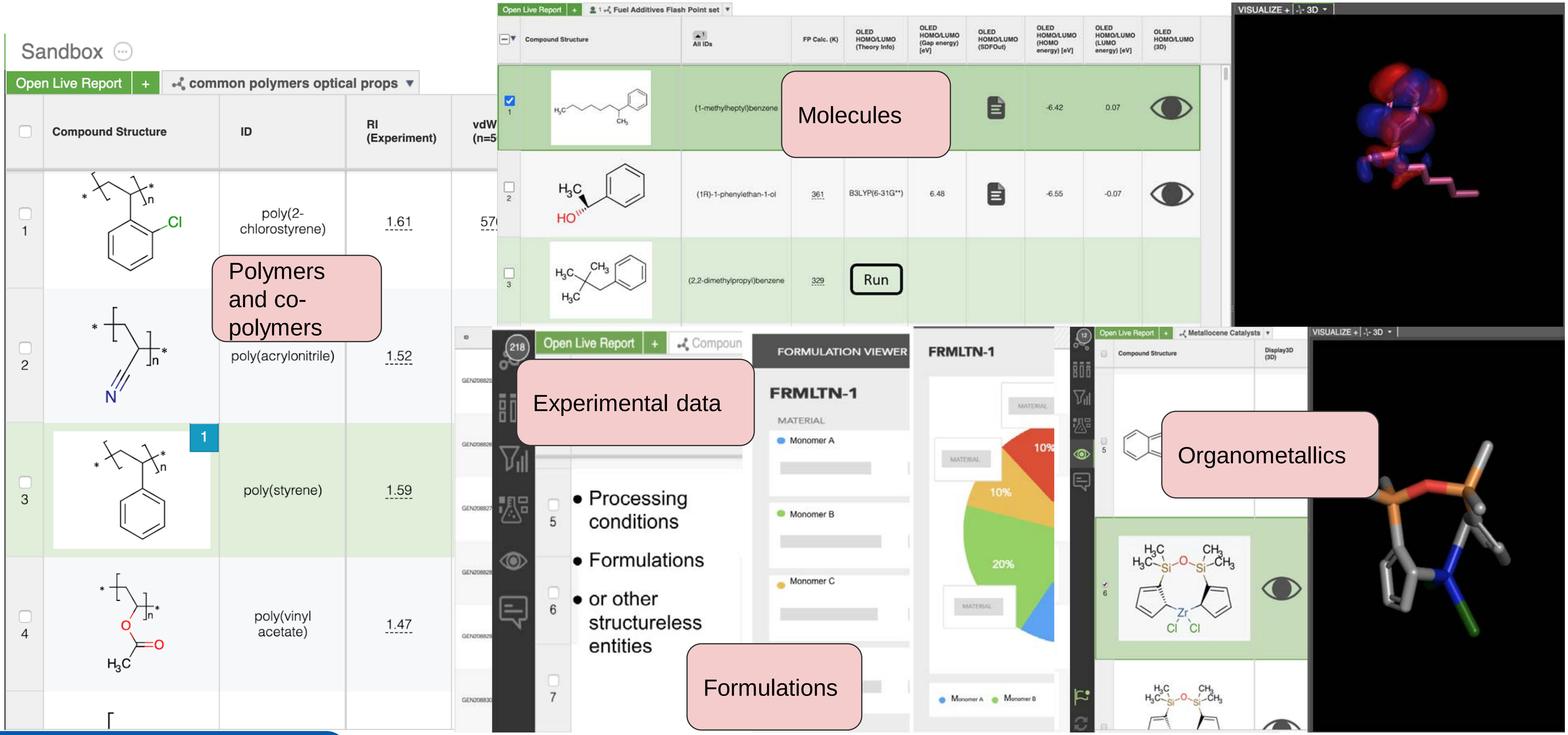
Informatics Analysis, Visualization and ***Machine Learning***

- Web-based: Instantly LIVE to all users
- Scalable: Performant for global sized organizations
- Informatics: Visualization and Analysis
- Central Platform

- Easy agnostic access to expert computational tools
  - Machine learning
  - Advanced QM properties
- Execute modeling jobs, analyze results alongside all other data

Schrödinger’s core values of modeling supported DESIGN

Suitable for Diverse Materials and Data Types



Sandbox

Open Live Report + common polymers optical props

	Compound Structure	ID	RI (Experiment)	vdW (n=5
□ 1		poly(2-chlorostyrene)	1.61	57
□ 2		poly(acrylonitrile)	1.52	
□ 3		poly(styrene)	1.59	
□ 4		poly(vinyl acetate)	1.47	

Polymers and co-polymers

Open Live Report + 1 Fuel Additives Flash Point set

	Compound Structure	All IDs	FP Calc. (K)	OLED HOMO/LUMO (Theory Info)	OLED HOMO/LUMO (Gap energy) [eV]	OLED HOMO/LUMO (SDFOut)	OLED HOMO/LUMO (HOMO energy) [eV]	OLED HOMO/LUMO (LUMO energy) [eV]	OLED HOMO/LUMO (3D)
☑ 1		(1-methylheptyl)benzene					-6.42	0.07	
□ 2		(1R)-1-phenylethan-1-ol	361	B3LYP(6-31G**)	6.48		-6.55	-0.07	
□ 3		(2,2-dimethylpropyl)benzene	329	Run					

Molecules

VISUALIZE + 3D

Open Live Report + Compoun

Experimental data

- Processing conditions
- Formulations
- or other structureless entities

FORMULATION VIEWER

FRMLTN-1

MATERIAL

Monomer A

Monomer B

Monomer C

Formulations

Open Live Report + Metallocene Catalysts

Compound Structure

Display3D (3D)

Organometallics

VISUALIZE + 3D

Summary

- Customized featurization based on chemical domain knowledge is critical in developing machine learning models
- High throughput physics-based modeling (QM, MD) provides various advantages in enhancing machine learning technology
- Our machine learning technology has been successfully applied to a wide range of materials systems and can be easily adapted to experiment design (i.e experimental design)
- Web-based materials informatics platform (LiveDesign) enables data digitization with advanced data analysis/visualization and machine learning technology
- ML Forcefields offer MD simulations with DFT-level accuracy
- Schrödinger’s technology is to ***empower*** users and increase efficiency and productivity

Thank you

Schrödinger

Feel free to reach out to me at:
chandras@schrodinger.com