Machine Learning for Design and Discovery of New Energetic Molecules

Daniel C. Elton, Ph.D.

6/4/18

UNIVERSITY OF MARYLAND 1856
THE DEPARTMENT OF MECHANICAL ENGINEERING
CECD
CENTER FOR ENGINEERING CONCEPTS DEVELOPMENT
energetics technology center
Office of Naval Research
USN
ONR
Science & Technology

THE DEPARTMENT OF MECHANICAL ENGINEERING

Overview of our grant

CECD
CENTER FOR ENGINEERING CONCEPTS DEVELOPMENT

Goal:

To accelerate the discovery of new energetic (explosive) molecules by using machine learning methods instead of traditional physics-based simulation for screening.

Thrust 1: Dataset construction

- Build dataset on energetic molecules and properties (finished 8/17)
- Identify design heuristics by searching literature and by interviewing subject matter experts

Thrust 2: Machine Learning

- Machine learning for property prediction (finished 1/18)
- Interpretation of machine learning models (finished 4/18)
- (extra) Generative models
- Perform quantum & thermochemical simulations to test newly generated molecules, use outputs from simulations as new training data.

My focus

Thrust 3: Natural Language Processing ("semantic discovery")

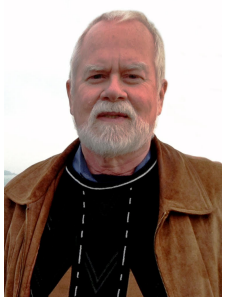
- Ingest pdfs and book scans to ASCII text, extract sentences
- Identify chemical names (named entity recognition)
- Test word embeddings (word2vec, GloVe, functional embedding)
- Use word embeddings to associate chemicals with properties and functionalities. Build a "recommender system" for molecules.

UMD-ETC machine learning team

Prof. Peter W. Chung
Computational materials science, energetic materials

Prof. Mark D. Fuge
Machine learning
data-driven design

Dr. Ruth M. Doherty
Energetic materials
chemistry, formulations



Dr. Bill Wilson
Energetic materials
properties, effects

Dr. Daniel Elton (Ph.D., Physics, ’16)
Machine learning
regression, generative models

Dr. Zois Boukouvalas (Ph.D., Math, ‘17)
Machine learning
dimensionality reduction

Mark S. Butrico (MechEng BS, ‘19)
Classification (moving to ARL this summer)

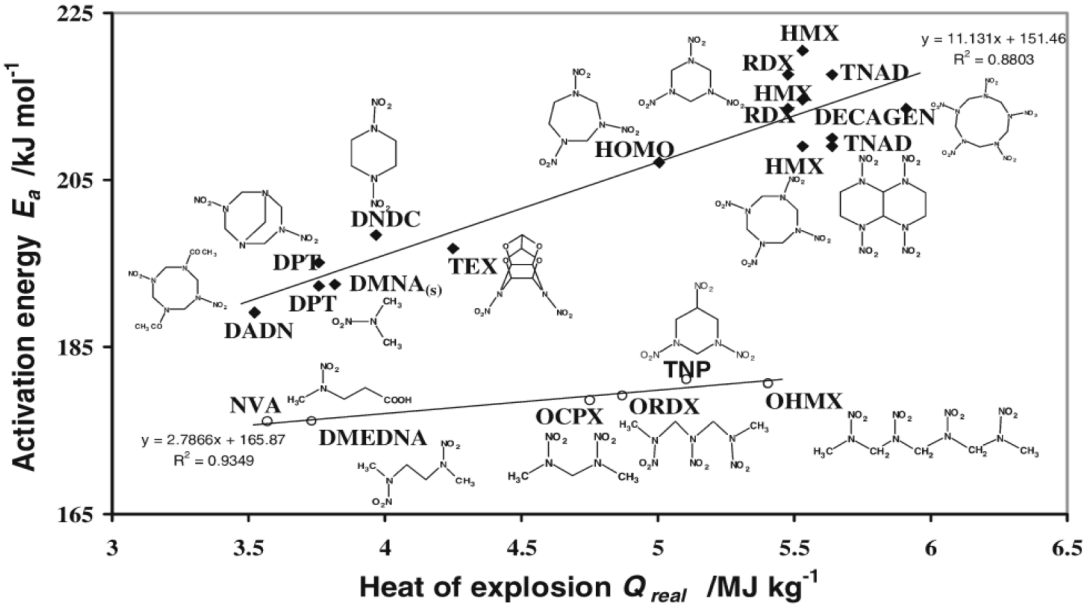
Dhruv Takhia (Telecomm. MS, ‘19)
NLP – word embeddings

Nischal Reddy (CS BS, ‘18)
NLP – word embeddings

Minh Truong (CS BS, ‘19)
NLP – named entity recognition

Austin Kim (CS BS, ‘21)
Classification

Previous modeling of energetic properties

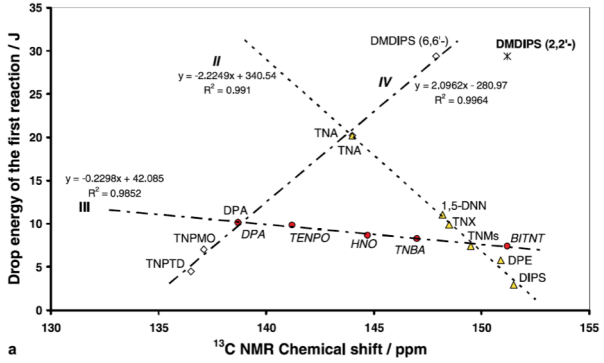


Linear regression only

Single hand-picked feature used

Each linear model is restricted to a single compound class

Sensitivity measurement



Compound classes differ significantly in structure-property relations

S. Zeman, in *High Energy Density Materials* (Ed: T. M. Klapötke), Springer-Verlag, Berlin 2007.

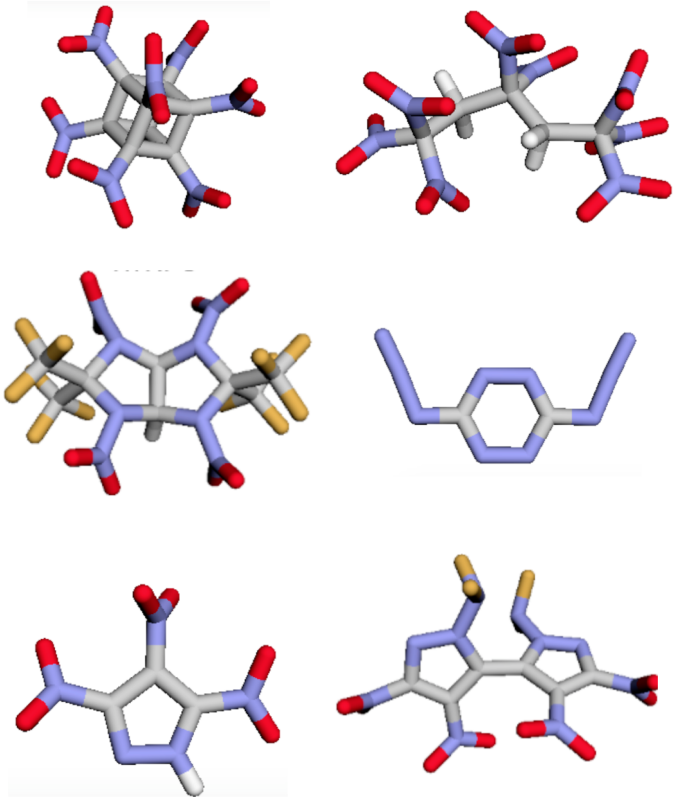
THE DEPARTMENT OF MECHANICAL ENGINEERING

Our approach

CECD
CENTER FOR ENGINEERING CONCEPTS DEVELOPMENT

Goal:

- Predict properties over many compound classes with a single model
- Test many model types & latest featurizations developed for molecules (last 5 years)



Property data of Huang & Massa[1]

- **109 molecules, 10 compound classes**
- DFT & CHEETAH thermochemical code[2]
- **9 properties** – density, heat of formation, explosive energy, shock velocity, particle velocity, sound velocity, detonation pressure, detonation temperature, and TNT equivalent per cc

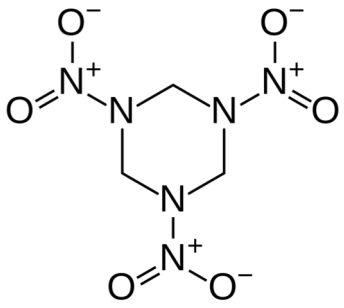
1.) Huang, L. & Massa, L. Int. J. Ener. Mat. Chem. Prop. **12** , 197–262 (2013)
2.) Huang, L., Massa, L. & Karle, J. Int. J. Ener. Mat. Chem. Prop. **10** , 33–44 (2011)

THE DEPARTMENT OF MECHANICAL ENGINEERING

Featurization

CECD
CENTER FOR ENGINEERING CONCEPTS DEVELOPMENT

2D Molecular graph



SMILES String

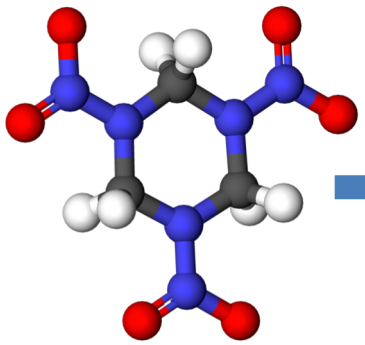
C1N(CN(CN1[N+](=O)[O-])[N+](=O)[O-])[N+](=O)[O-]

featurization method

Feature vector

$$\vec{x} = \begin{bmatrix} 0.34 \\ -0.13 \\ 0.59 \\ \vdots \\ 0.10 \end{bmatrix}$$

3D Molecular structure



(x,y,z) coodinates

H	-3.3804130	-1.1272367	0.5733036
N	0.9668296	-1.0737425	-0.8198227
C	0.0567293	0.8527195	0.3923156
N	-1.3751742	-1.0212243	-0.0570552
C	-1.2615018	0.2590713	0.5234135
C	-0.3068337	-1.6836331	-0.7169344
C	1.1394235	0.1874122	-0.2700900
N	0.5602627	2.0839095	0.8251589
O	-0.4926797	-2.8180554	-1.2094732

featurization method

Feature vector

$$\vec{x} = \begin{bmatrix} -0.26 \\ 0.33 \\ 0.73 \\ \vdots \\ 0.08 \end{bmatrix}$$

Featurization methods tested

- Oxygen balance
- Custom descriptor set
- Fingerprinting
- Sum over bonds
- Custom graph convolutional fingerprints
- Coulomb matrices
- Bag of bonds

Based on stoichiometry only

Based on 2D molecular graph

Based on 3D structure

Models explored

- Linear regression
- Ridge regression
- LASSO regression
- Bayesian ridge regression
- Elastic net regression
- Kernel ridge regression
- Support vector regression
- *k* nearest neighbors
- Gaussian process regression
- Random forests
- Gradient boosted trees
- Neural network

variants of linear regression

Decision tree based

Systematic comparison

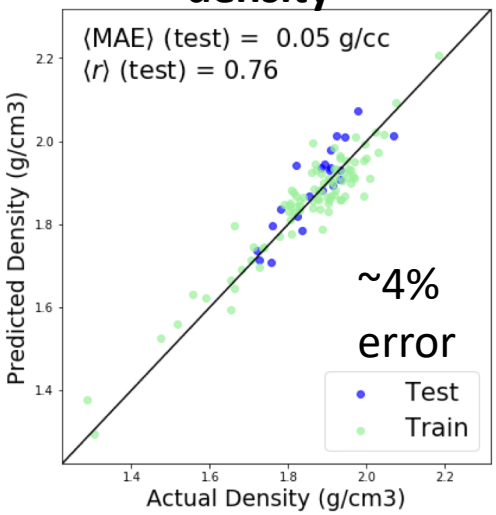
Models	Featurizations	Density ρ, $\frac{g}{cc}$	Heat of formation ΔH_f^s, $\frac{kJ}{mol}$	Explosive energy E_e, $\frac{kJ}{cc}$	Shock velocity V_s, $\frac{km}{s}$	Particle velocity V_p, $\frac{km}{s}$	Sound velocity V_{snd}, $\frac{km}{s}$	Detonation pressure P, GPa	Temperature T, K	TNT equiv. $\frac{TNT_{equiv}}{cc}$
KRR	Estate	0.06	231.61	0.41	0.43	0.10	0.35	3.65	387.56	0.14
	CDS	0.07	205.52	0.51	0.40	0.13	0.35	4.29	417.49	0.18
	SoB	0.06	78.20	0.37	0.30	**0.09**	0.27	3.17	333.02	**0.11**
	CM eigs	0.08	275.83	0.52	0.56	0.12	0.47	4.59	553.41	0.19
	Bag of Bonds	0.07	145.69	0.43	0.32	0.11	**0.24**	3.50	411.28	0.15
	Estate+CDS+SoB	**0.05**	**71.10**	**0.36**	0.36	**0.09**	0.26	3.12	370.32	0.13
Ridge	Estate	0.06	270.13	0.38	0.42	0.10	0.37	3.39	427.19	0.15
	CDS	0.06	198.10	0.50	0.47	0.12	0.39	4.30	431.77	0.17
	SoB	0.06	73.71	0.37	**0.30**	**0.09**	0.25	3.34	**309.98**	**0.11**
	CM eigs	0.08	326.18	0.57	0.65	0.14	0.51	5.18	569.68	0.20
	Bag of Bonds	0.07	168.93	0.45	0.32	0.12	**0.24**	3.47	476.96	0.18
	Estate+CDS+SoB	**0.05**	76.10	**0.36**	0.30	**0.09**	0.26	**3.01**	341.44	0.12
	OB_{1600}	0.08	331.29	0.59	0.59	0.14	0.51	4.61	503.17	0.20
⋮	⋮					⋮				

THE DEPARTMENT OF MECHANICAL ENGINEERING

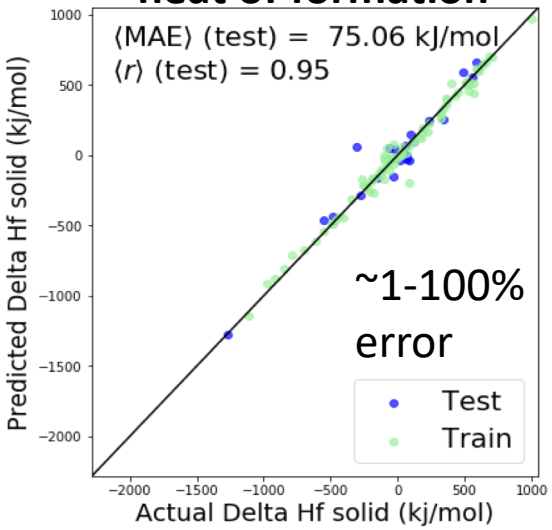
Model accuracies

CECD
CENTER FOR ENGINEERING CONCEPTS DEVELOPMENT

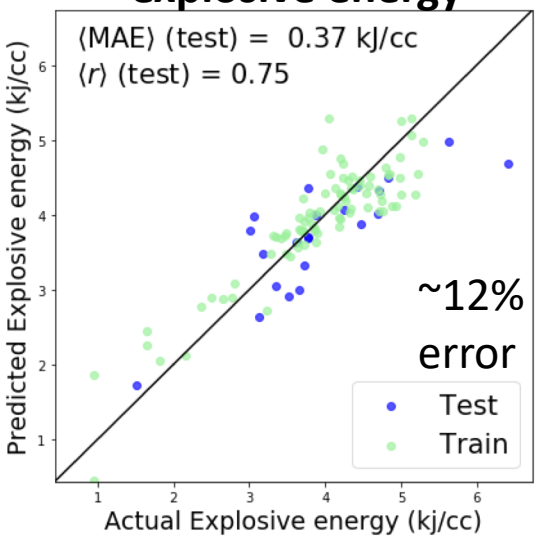
density



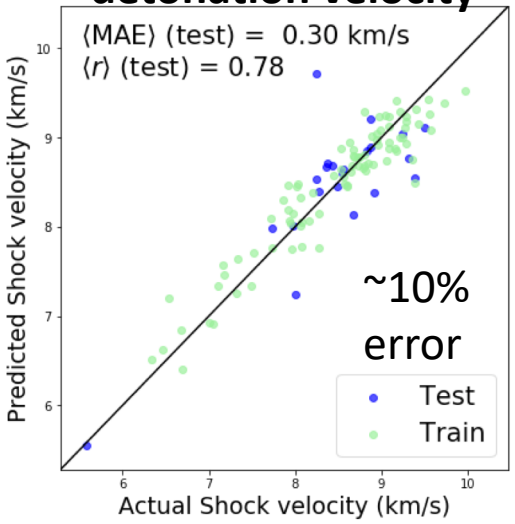
heat of formation



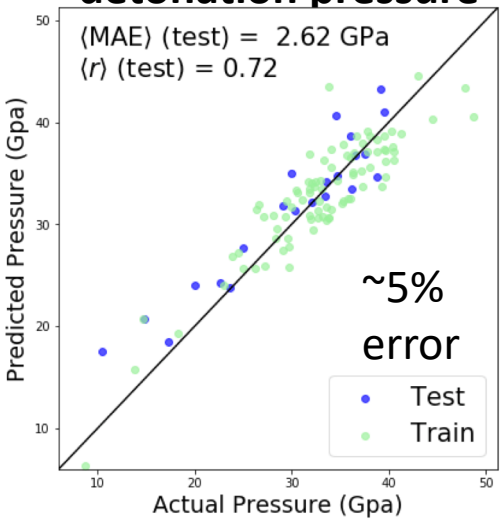
explosive energy



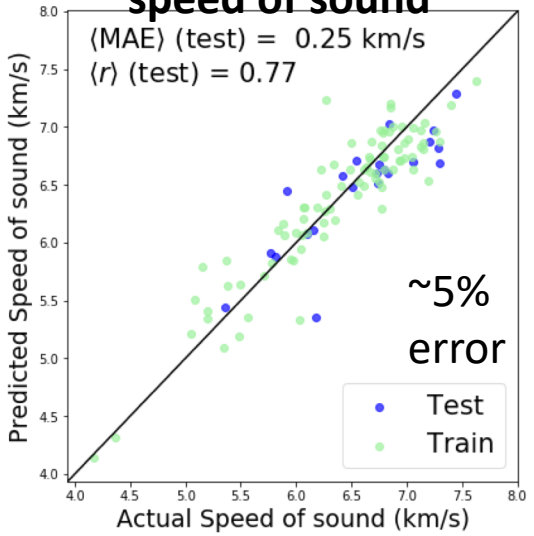
detonation velocity



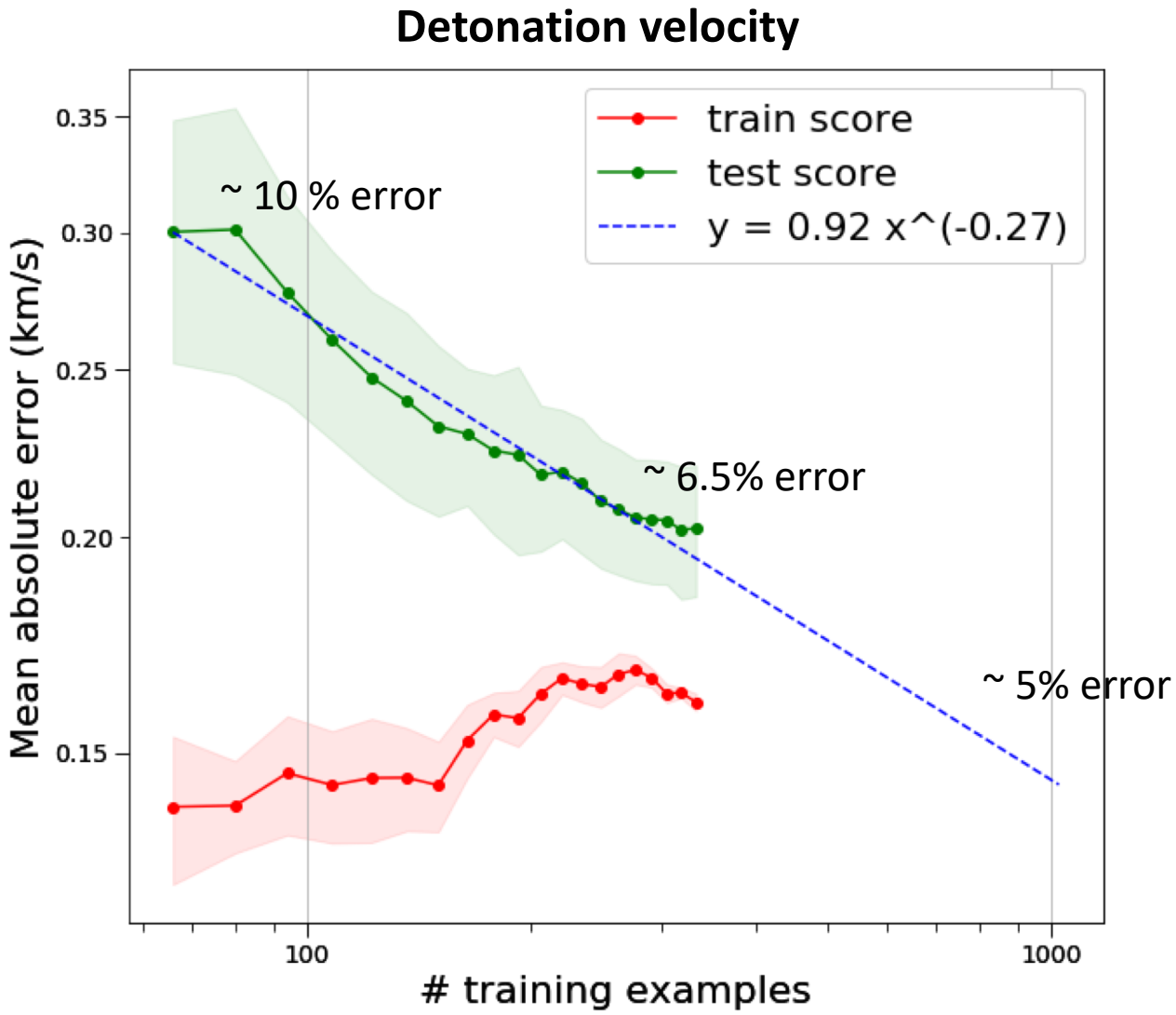
detonation pressure



speed of sound



Data dependence



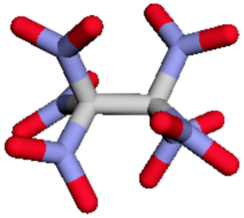
Cutting error in half requires increasing # of training molecules by 10 x

Theoretical limit with our current best featurization & model

Model interpretation

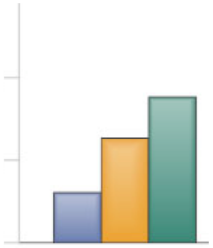
Forward modeling

Molecular structure



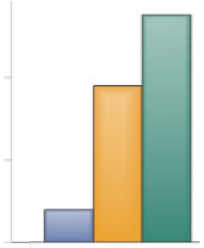
model

Properties



Inverse modeling

Properties



Interpretation of model

Molecular structure

?

Reasons for interpretation:

- To ensure the model is capturing known structure – property relationships
- To discover new structure – property relationships
- To discover latent features the model is using that may be useful for human designers

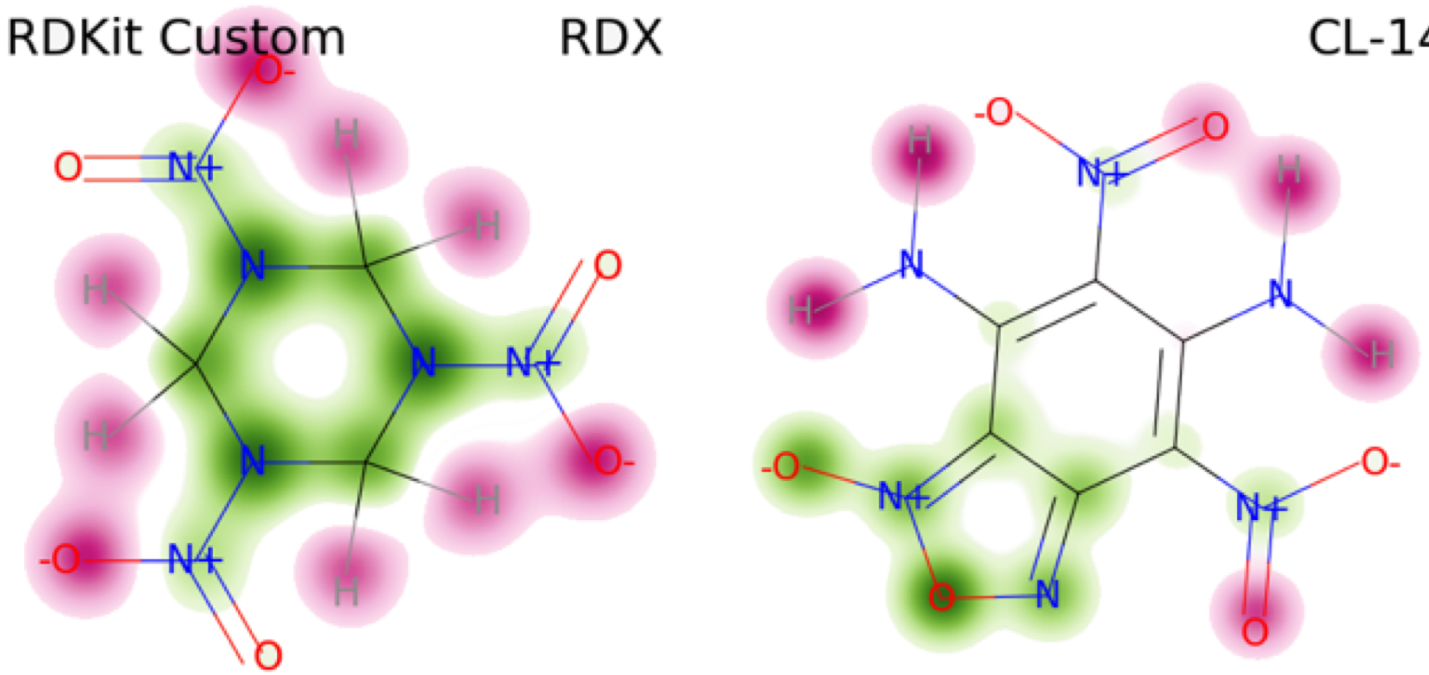
THE DEPARTMENT OF MECHANICAL ENGINEERING

One interpretation method : sensitivity analysis

CECD
CENTER FOR ENGINEERING CONCEPTS DEVELOPMENT

removal of a red atom causes the model to increase its predicted detonation velocity.
therefore red atoms are associated with *lower* detonation velocity according to the model

RDKit Custom RDX CL-14



Not all featurizations yield intuitive sensitivity analysis maps like this!!

See upcoming International Detonation Symposium paper for discussion & caveats

Feature ranking

Ranking of features associated with higher or lower **detonation velocity**

ranking	*Pearson correlation*		*Mutual information corr. coeff.*		*Maximal information criteria*		*LASSO coefficient size*		*LASSO stability selection*		*Random forest variance score*		*Random forest shuffling*	
1	C:C	-0.483	OB_{100}	0.353	C:C	0.342	C:C	-0.023	C:C	0.570	C:C	0.321	OB_{100}	1.422
2	aCa	-0.413	C:C	0.281	OB_{100}	0.307	C-H	-0.018	OB_{100}	0.545	OB_{100}	0.295	>NH-[+1]	1.013
3	OB_{100}	+0.393	C-H	0.227	aCa	0.300	N-O	+0.013	N-N	0.495	C-H	0.098	=0	1.009
4	N-N	+0.389	aCa	0.221	N-N	0.296	OB_{100}	+0.013	C-H	0.360	C-N	0.032	aaCa	1.007
5	n_{NNO_2}	+0.342	=0	0.189	>C<	0.267	n_{CO}	-0.011	aCa	0.225	C-O	0.028	C:N	1.006

Results align with chemical intuition here.

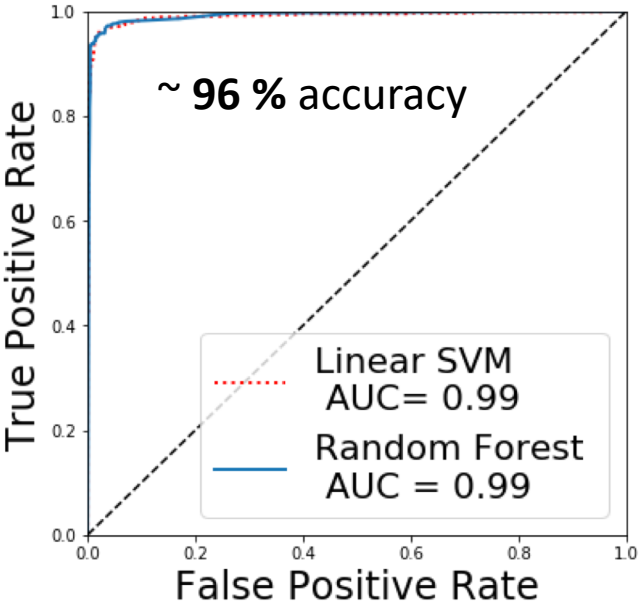
See upcoming *International Detonation Symposium* proceedings paper for discussion & caveats

THE DEPARTMENT OF MECHANICAL ENGINEERING

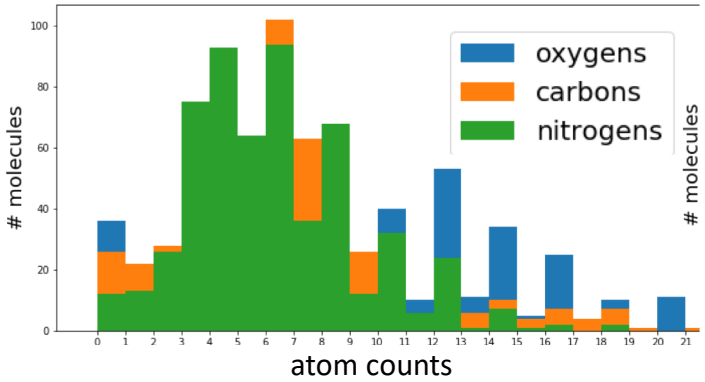
Classification

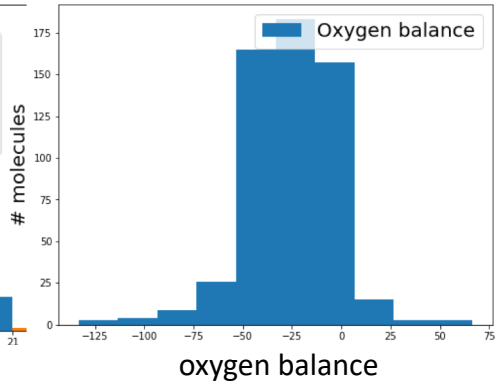
CECD CENTER FOR ENGINEERING CONCEPTS DEVELOPMENT

ROC curve

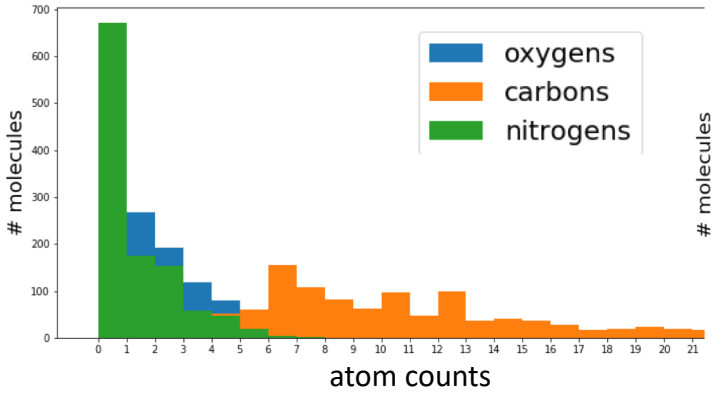


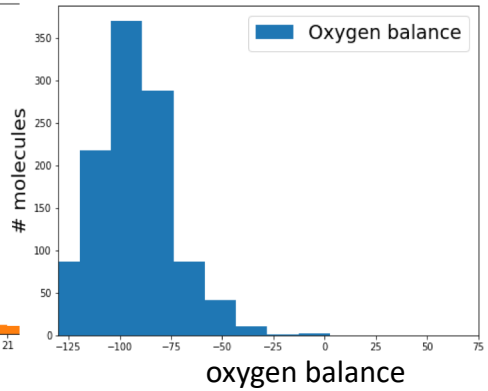
617 energetic molecules





1,128 non energetics - Delaney solubility dataset mostly pesticides and agrochemicals



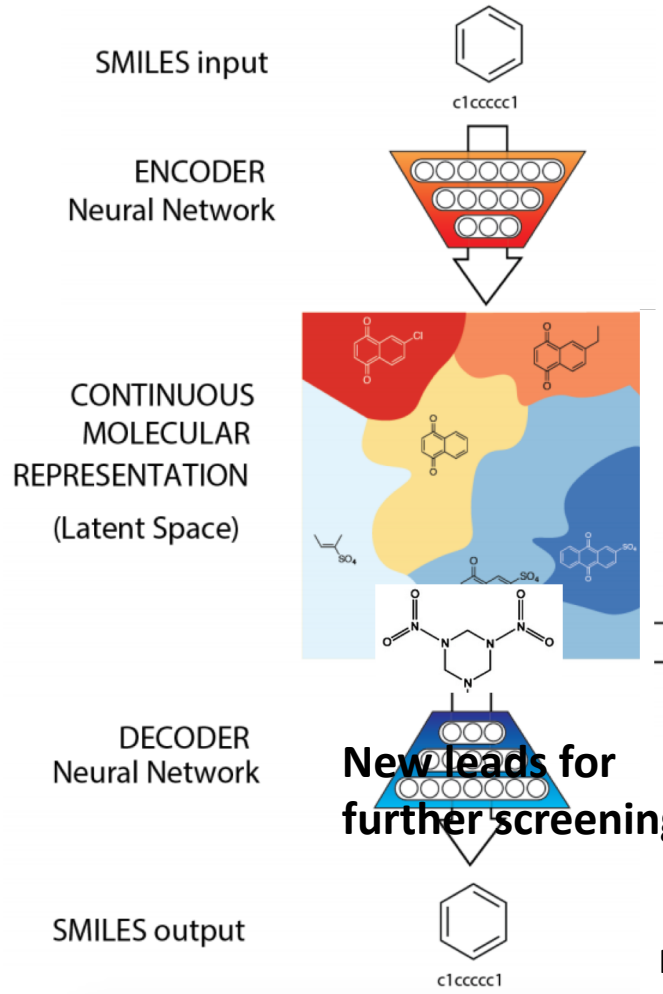


Random forest feature importances

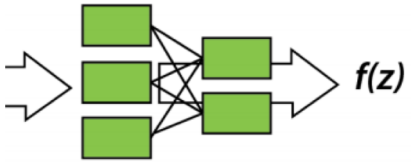
	importance	names
1	0.234166	N=O
2	0.164909	N-O
3	0.123097	=0
4	0.092792	OB_{100}
5	0.073828	n_NO

Generative method #1
Variational autoencoders

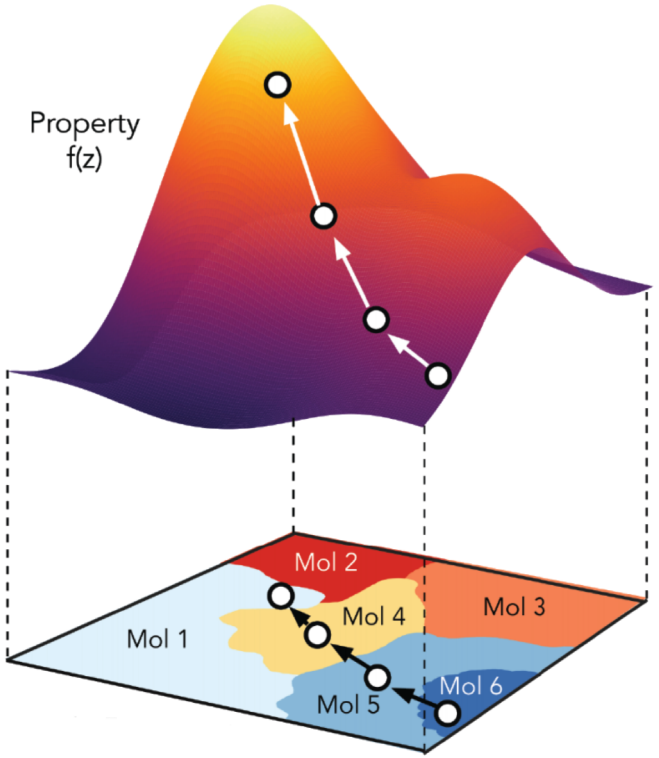
Stage 1: training autoencoder to learn low dimensional latent representation



attach property predicting model (or physics code)



New leads for further screening



Stage 2: optimization in latent space to find molecules which maximize target property

Figures from
Gómez-Bombarelli et al. *ACS Cent. Sci.,* (2018) 10.1021/acscentsci.7b00572
Aspuru-Guzik group, Harvard.

Daniel C. Elton

THE DEPARTMENT OF MECHANICAL ENGINEERING

Molecular Generation with Deep Learning Architectures

CECD CENTER FOR ENGINEERING CONCEPTS DEVELOPMENT

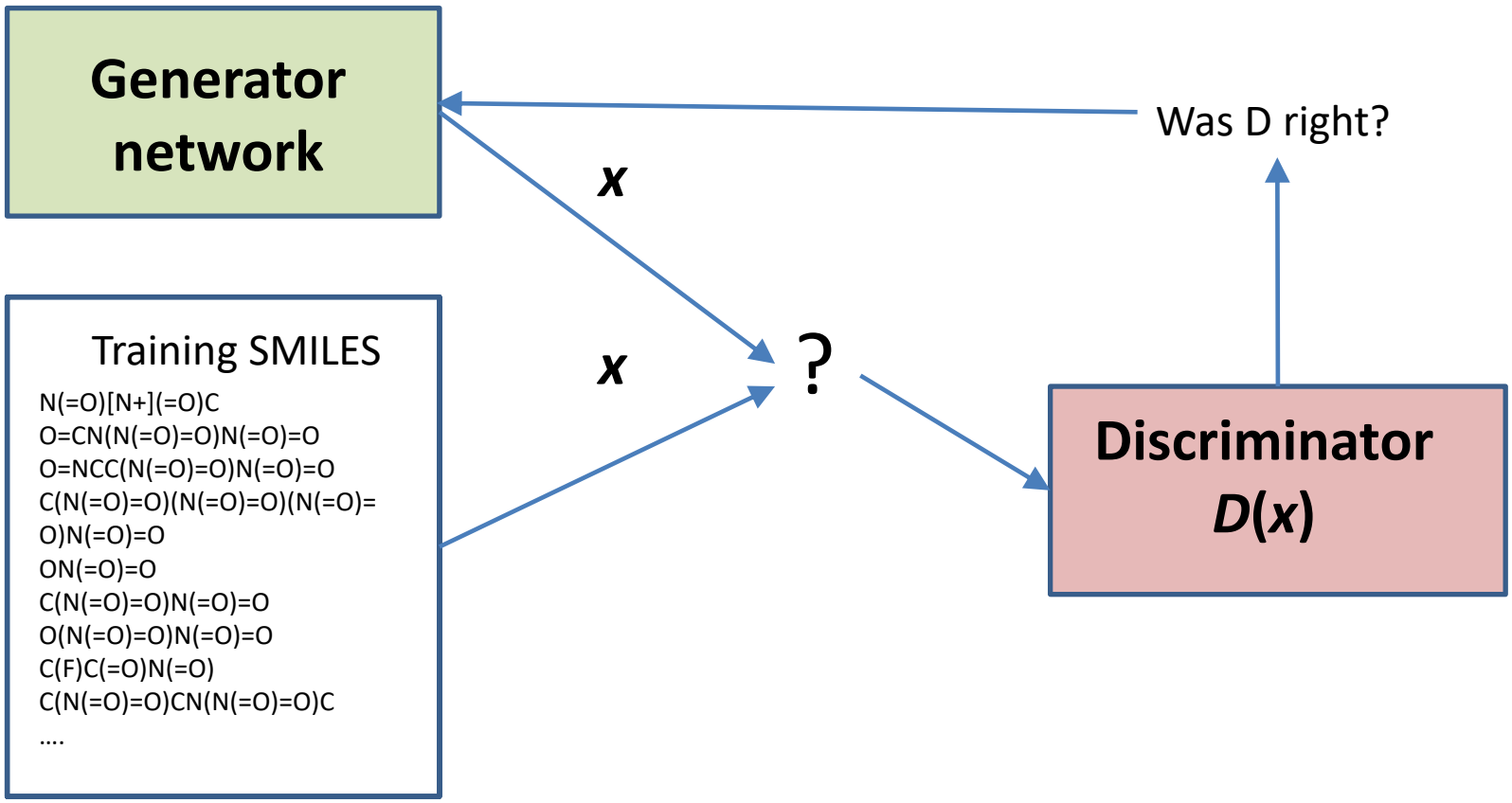
Very hot area - new papers coming out every month

- **Variational Autoencoders (VAEs)**
  - **Molecular autoencoder,** Gomez-Bombarelli, Aspuru-Guzik, et al. *ACS Central Science,* 2018
  - **Deep adversarial autoencoder,** Kadurin, et al. *Oncotarget*, **8**(7) p. 10883, (2017)
    - novel hybrid architecture – VAE but has GAN discriminator working on the latent space variables to enforce drug-likeness and drug binding properties

- **Generative Adverarial Networks (GANs)**
  - **ORGANIC** Sanchez-Lengeling, Aspuru-Guzik, et al. 2018, chemrxiv:5309668
    - based on **ORGAN,** Guimaraes, Aspuru-Guzik, et al. 2017, arxiv:1705.10843v3
      - based on **SeqGAN**, Yu, et al. 2017, AAAI-17 proceedings
        - uses reinforcement learning with policy gradients (Sutton et al. 1999)

- **Reccurent Neural Networks (RNNs)**
  - Olivecrona, et al. 2017 (Astrazenca)
  - Neil, Segler, et al. ICLR 2018 (BenevolantAI)
  - Bjerrum and Threlfall, arXiv:1705.04612 (Wildcard Pharmaceutical Consulting)

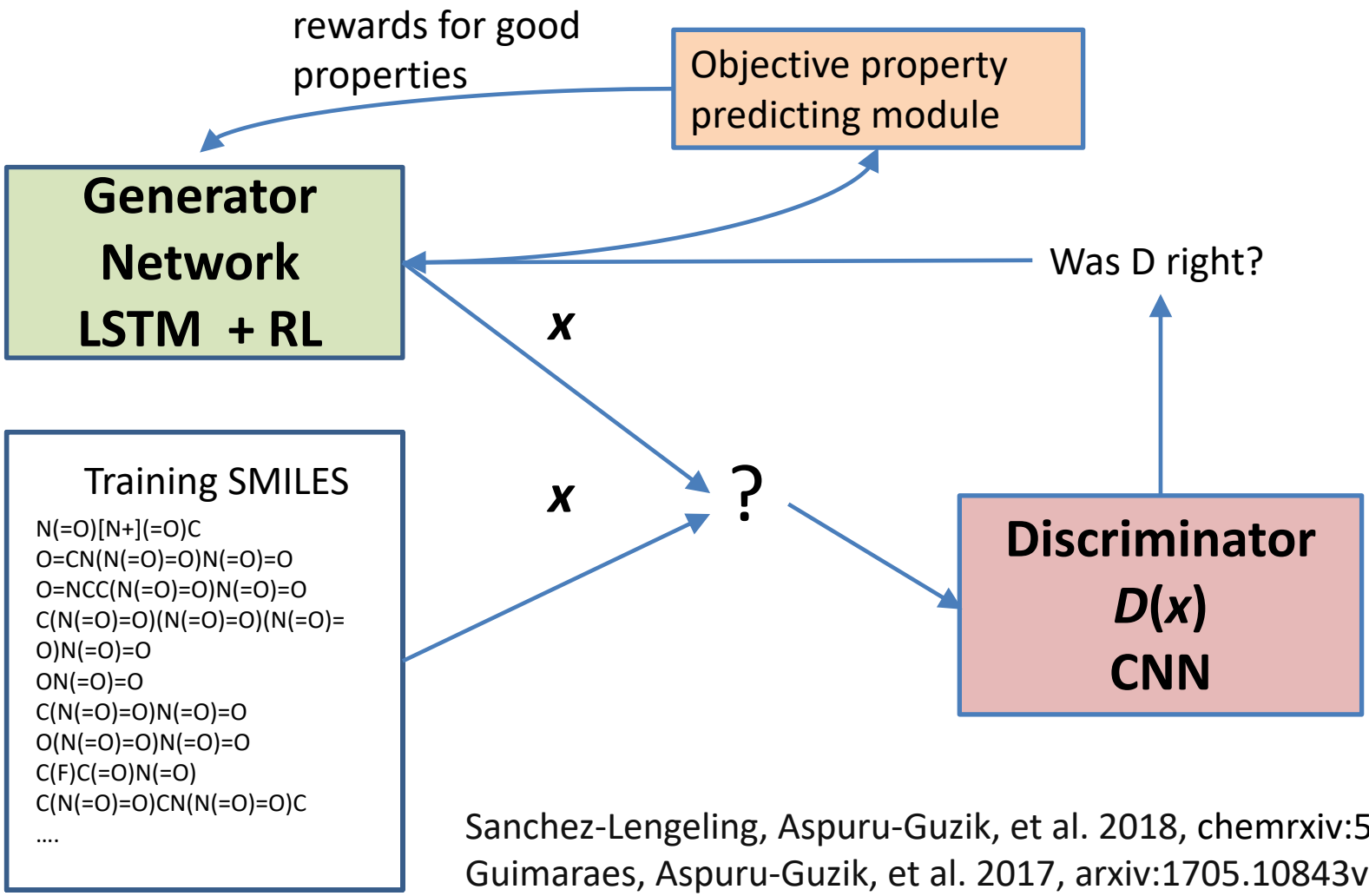
Generative method #2
Generative Adversarial Networks



THE DEPARTMENT OF MECHANICAL ENGINEERING

ORGAN
Objective Reinforced GAN

CECD
CENTER FOR ENGINEERING CONCEPTS DEVELOPMENT



Sanchez-Lengeling, Aspuru-Guzik, et al. 2018, chemrxiv:5309668
Guimaraes, Aspuru-Guzik, et al. 2017, arxiv:1705.10843v3

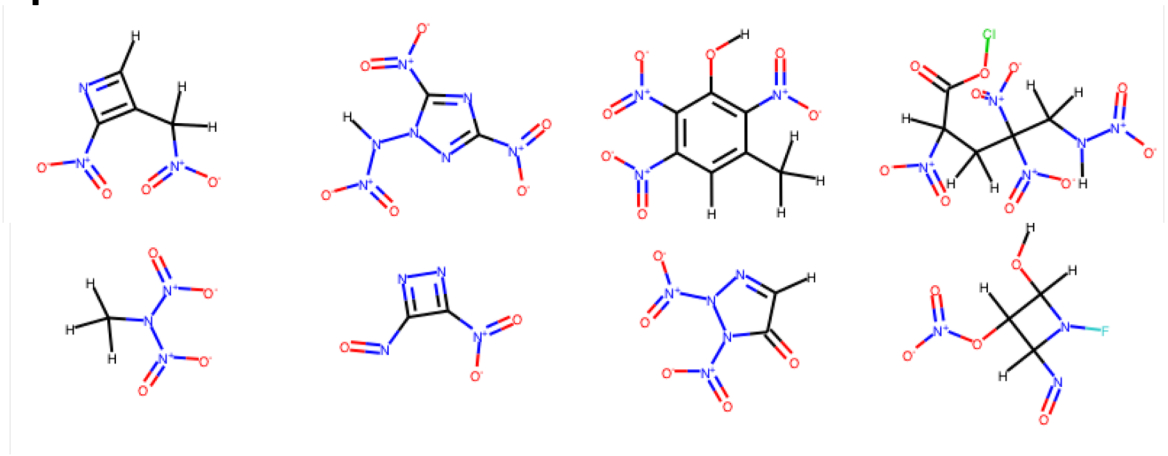
THE DEPARTMENT OF MECHANICAL ENGINEERING

ORGANIC GAN
Initial results

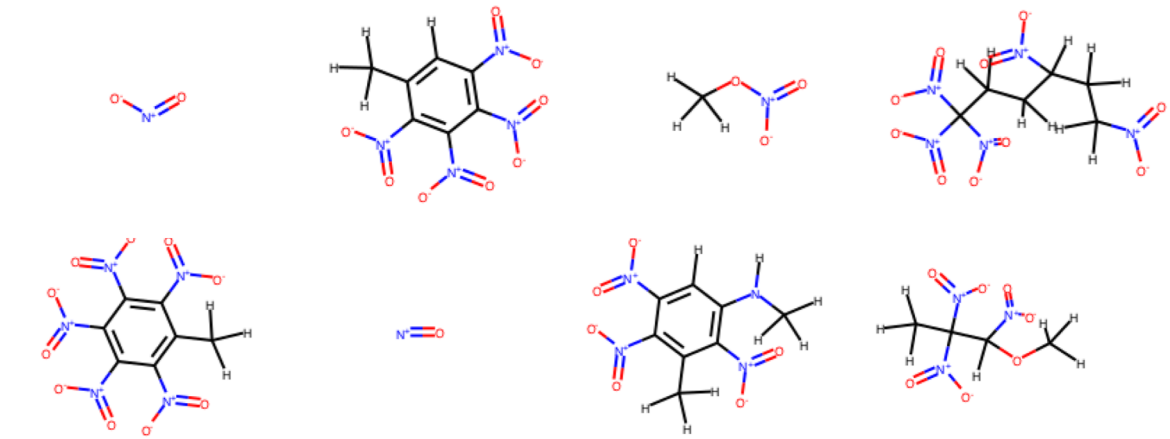
CECD
CENTER FOR ENGINEERING CONCEPTS DEVELOPMENT

Trained on 580 organic (CNOHFCl) energetic molecules

Epoch 1



Epoch 48

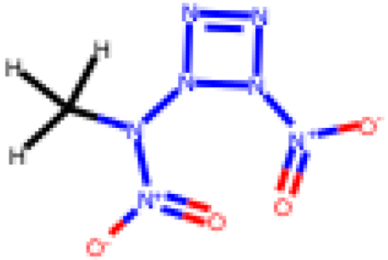


Batch rewards for variety (avg. in group distances) and diversity (avg. distance to random sample of training data)

Jaccard / Tanimoto distance

$$f(A,B) = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B}$$

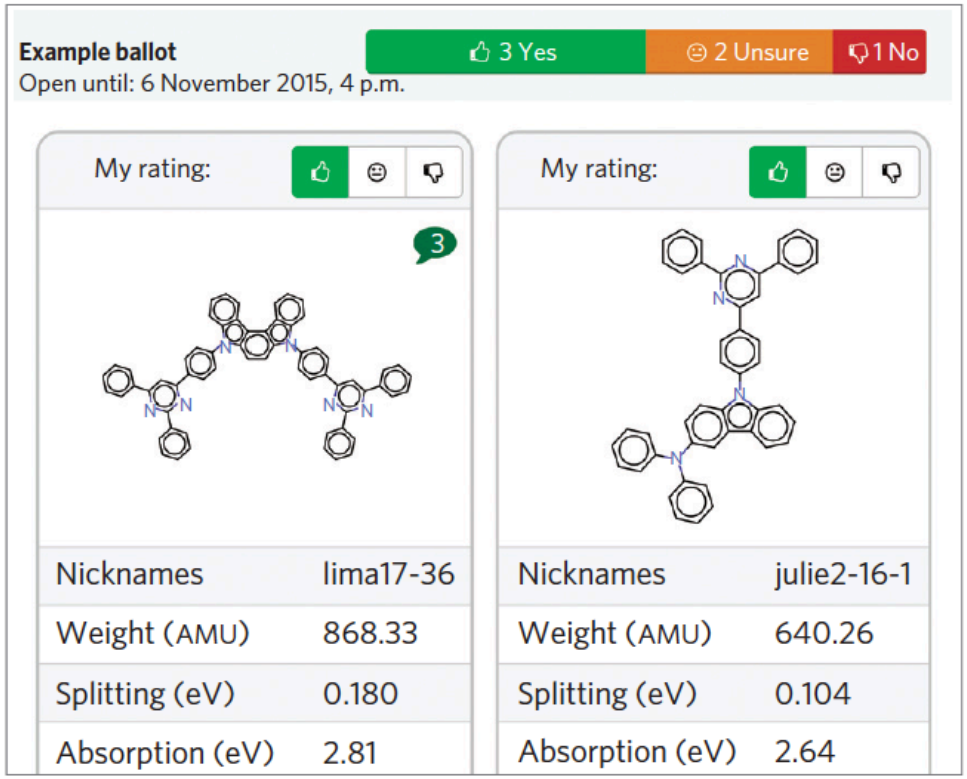
Not all of the ones generated are synthesizable!



Daniel C. Elton

"Molecular Tinder"

Allows subject matter experts to be part of screening pipeline and train deep learning models



R. Gómez-Bombarelli et al., Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach, *Nat. Mater.* 15, p. 1120-1127, 2016.

THE DEPARTMENT OF MECHANICAL ENGINEERING

Summary

CECD
CENTER FOR ENGINEERING CONCEPTS DEVELOPMENT

- **We presented a proof of principle that machine learning for property prediction and classification is feasible starting with small dataset of energetic materials**
- **We showed how how feature ranking can illuminate structure-property relations**
- **We showed initial efforts towards molecular generation and how it might be incorporated into a screening pipeline**

Acknowledgments

- Office of Naval Research (N00014-17-1-2108)
- Energetics Technology Center (2044-001)
- Center for Engineering Concepts Development (CECD)

Our paper:

D. C. Elton, Z. Boukouvalas, M. S. Butrico, M. D. Fuge, and P. W. Chung, "Applying machine learning techniques to predict the properties of energetic materials" accepted in *Scientific Reports,* 2018 (arXiv:1801.04900)

Upcoming paper in the 2018 Detonation Symposium (collaboration with ARL)

B. C. Barnes, D. C. Elton, Z. Boukouvalas, D. E. Taylor, W. D. Mattson, M. D. Fuge, and P. W. Chung, "Machine Learning and Discovery for Energetic Materials", in 16th International Detonation Symposium, Cambridge MD, USA, July 2018.

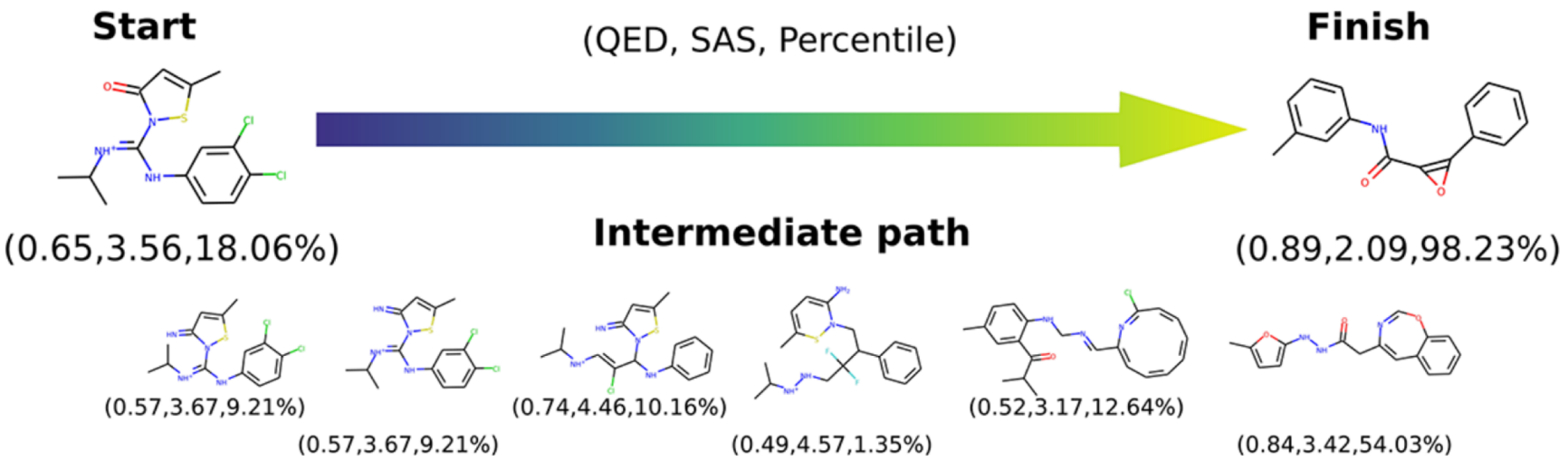
Open Source Python *Molecular Machine Learning Toolkit*

https://github.com/delton137/mmltoolkit

Molecular autoencoder
Prior work at Harvard

Optimization of

- **Quantitative Estimation of Drug-likeness**
- **Synthetic accessibility score**



Results from Gómez-Bombarelli et al. *ACS Cent. Sci.,* (2018) 10.1021/acscentsci.7b00572
Aspuru-Guzik group, Harvard.

THE DEPARTMENT OF MECHANICAL ENGINEERING

My background

CECD
CENTER FOR ENGINEERING CONCEPTS DEVELOPMENT

D. C. Elton, M. Fritz, and M.-V Fernández-Serra. "Using a monomer potential energy surface to perform approximate path integral molecular dynamics simulation of ab-initio water at near-zero added cost" (arXiv:1803.05740), 2018

D.C. Elton "The microscopic origin of the Debye relaxation in liquid water and fitting the high frequency excess response" *Phys. Chem. Chem. Phys.*, **19**, 18739 (2017)

D.C. Elton and M.-V Fernández-Serra. "The hydrogen bond network of water supports propagating optical phonon-like modes" *Nat. Comm.* **7**, 10913 (2016)

D.C. Elton and M.-V Fernández-Serra. "Polar nanoregions in water – a study of the dielectric properties of TIP4P/2005,TIP4P/2005f and TTM3F" *J. Chem. Phys.*, **140**, 124504 (2014)