Machine Learning-Accelerated Molecular Design of Innovative Polymers: Shifting from Thomas Edison to Iron Man

Future Composites Symposium

November 13 - 14, 2024

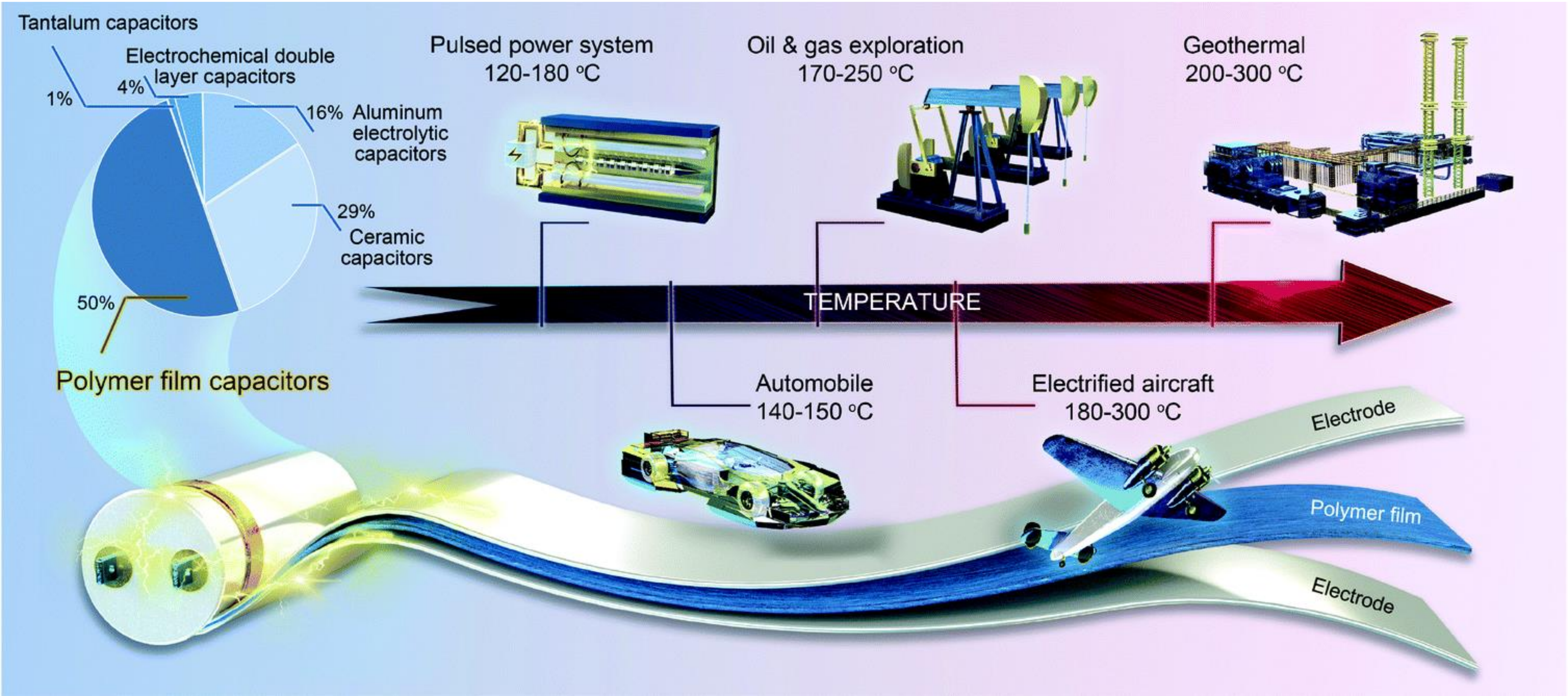
Ying Li, Associate Professor

Department of Mechanical Engineering,
University of Wisconsin-Madison
Email: yli2562@wisc.edu
Polymer Digital Engineering Lab: pdelab.engr.wisc.edu

Acknowledgment: Air Force's Young Investigator Research Program (FA9550-20-1-0183; Program Manager: Capt Derek Barbee), Air Force Research Laboratory/UES Inc. (FA8650-20-S-5008, PICASSO program), and the National Science Foundation (CMMI-2314424, CMMI-2316200, CMMI-2332276 and CAREER-2323108).

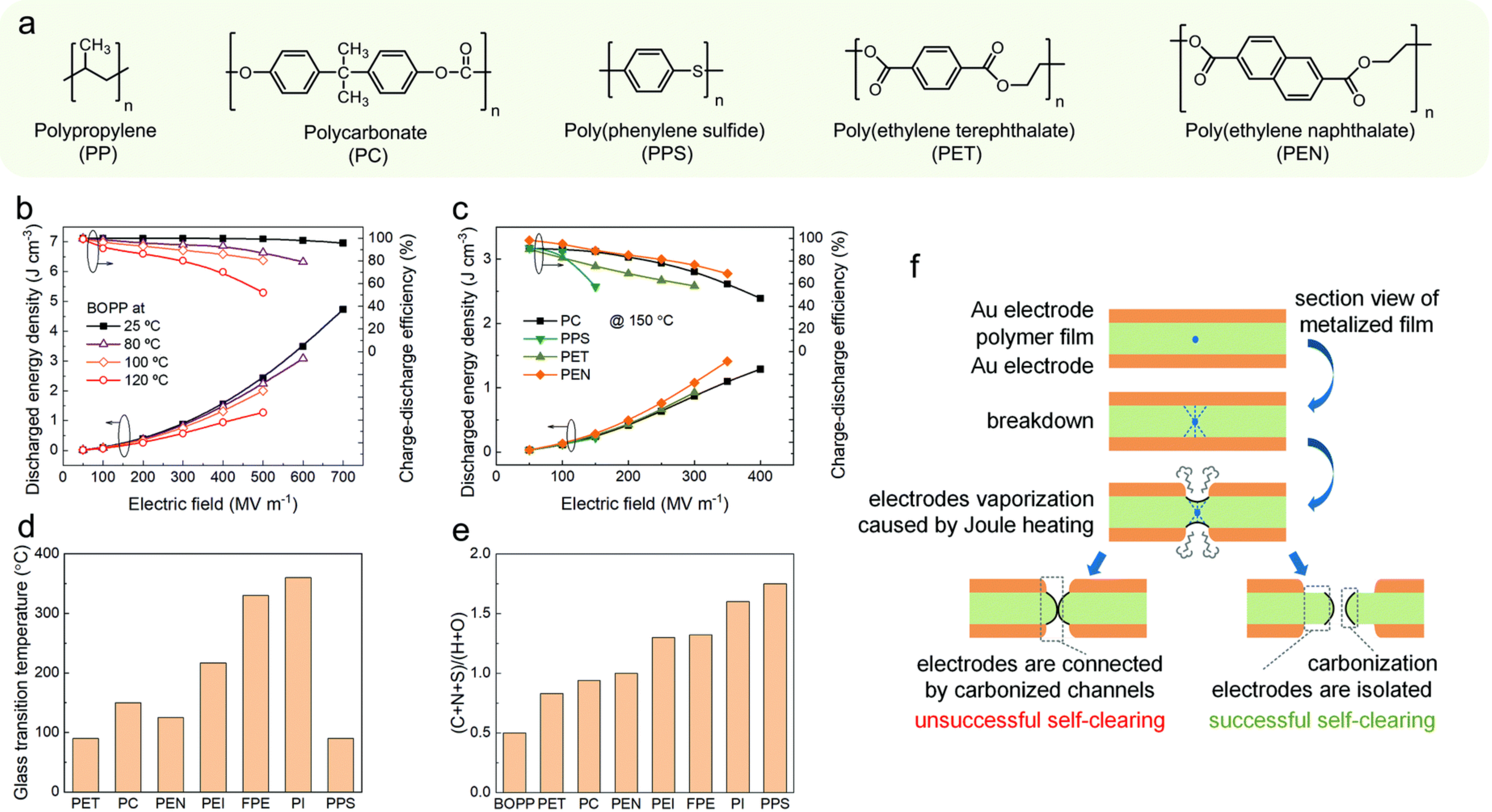
Co-Sponsors:

Dielectric polymers for capacitive energy storage

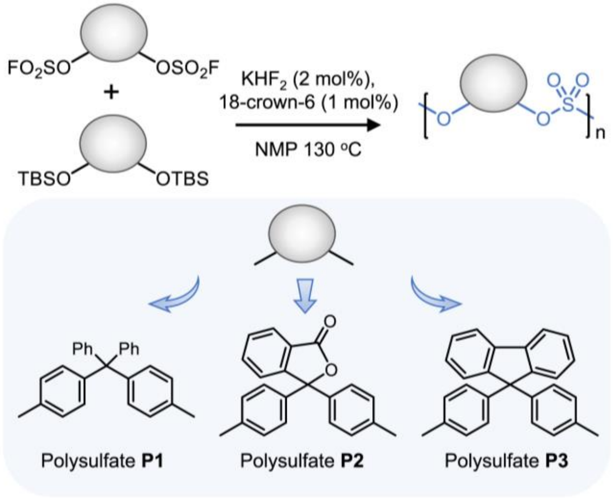


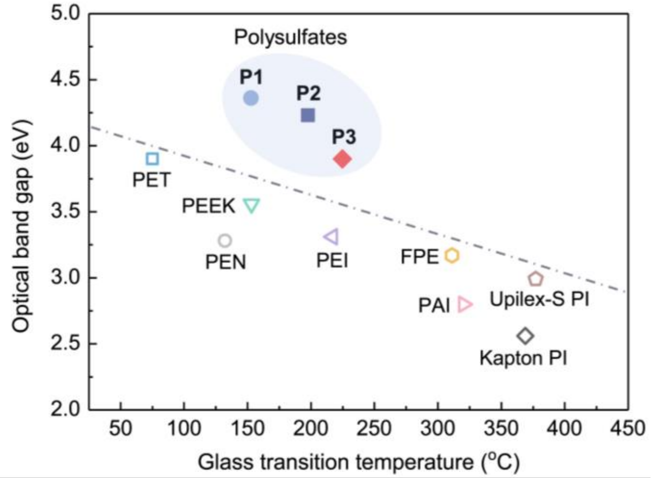
Li, He, Yao Zhou, Yang Liu, Li Li, Yi Liu, and Qing Wang. "Dielectric polymers for high-temperature capacitive energy storage." *Chemical Society Reviews* 50, no. 11 (2021): 6369-6400.

Dielectric polymers for capacitive energy storage

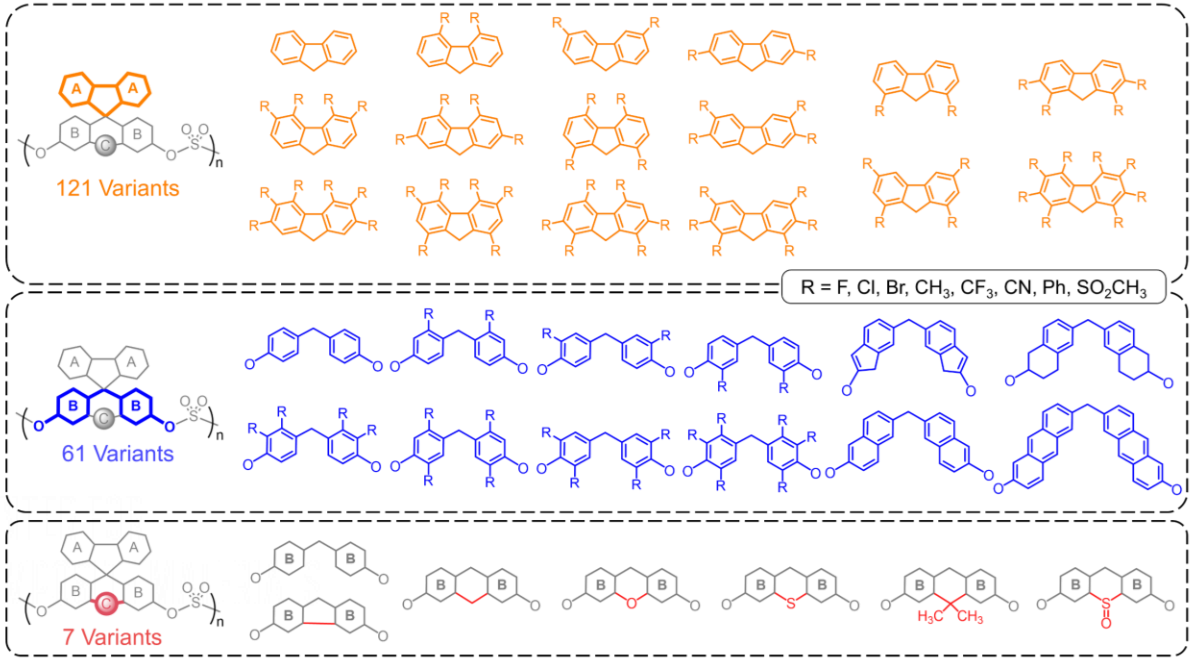


Building block assembly of heat-resistant polysulfates

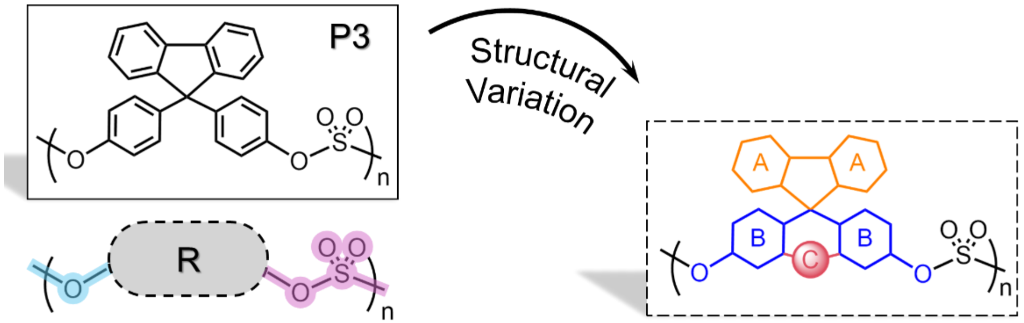




Li, He, et al. "High-performing polysulfate dielectrics for electrostatic energy storage under harsh conditions." Joule 7.1 (2023): 95-111..



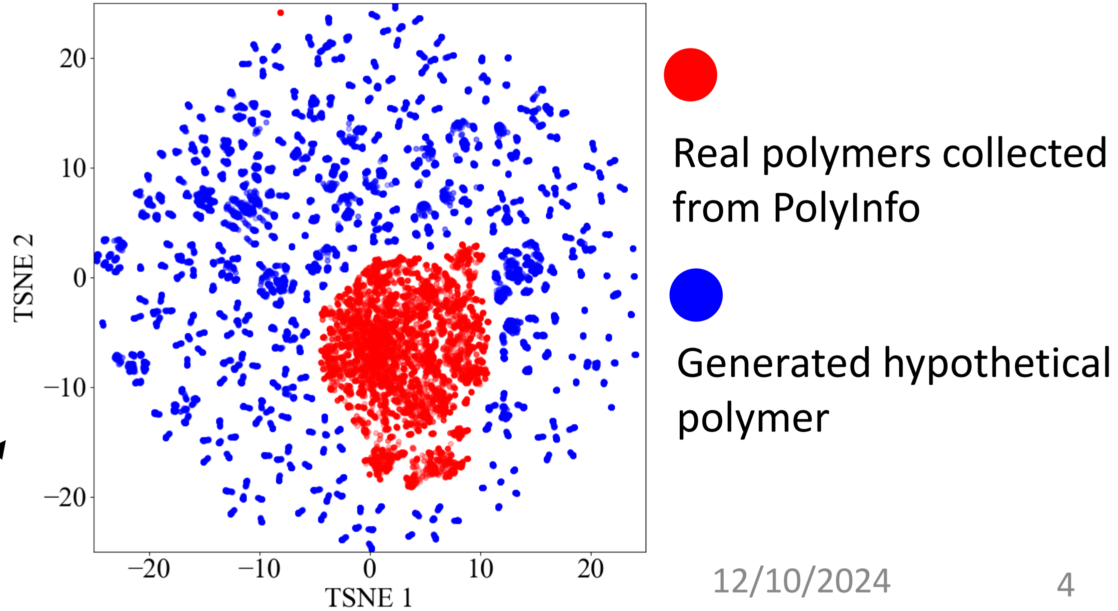
Inspired



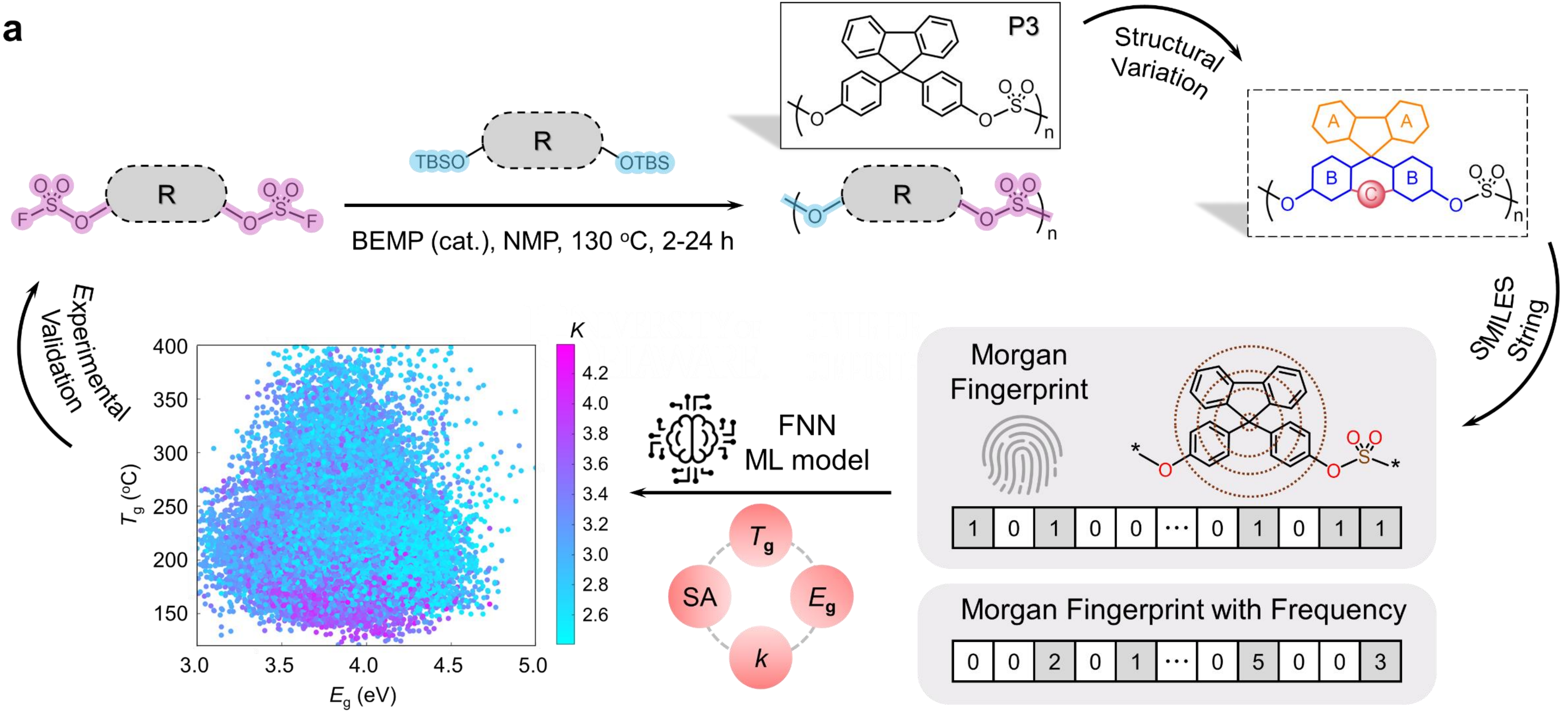
Generation

49,731 hypothetical polymer structures

Chemical space

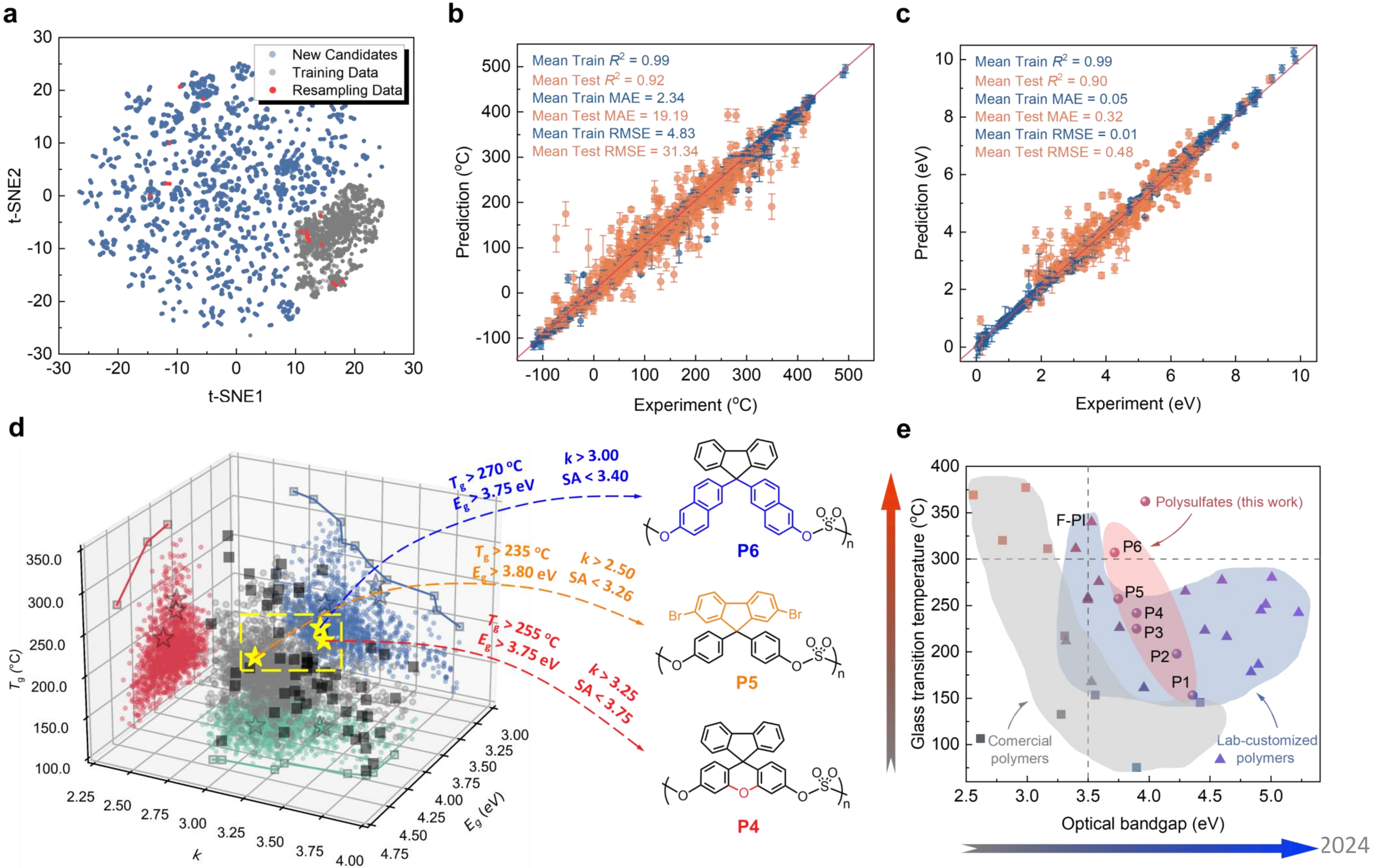


Discovery workflow integrating both ML and Exp

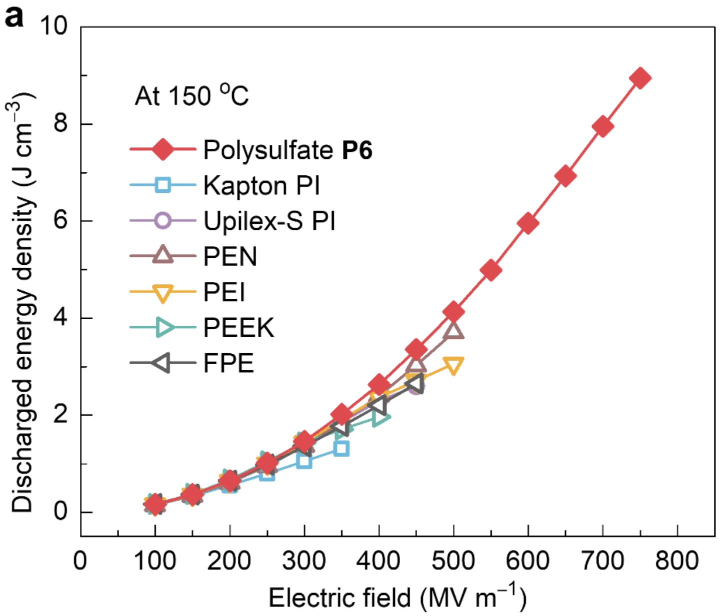


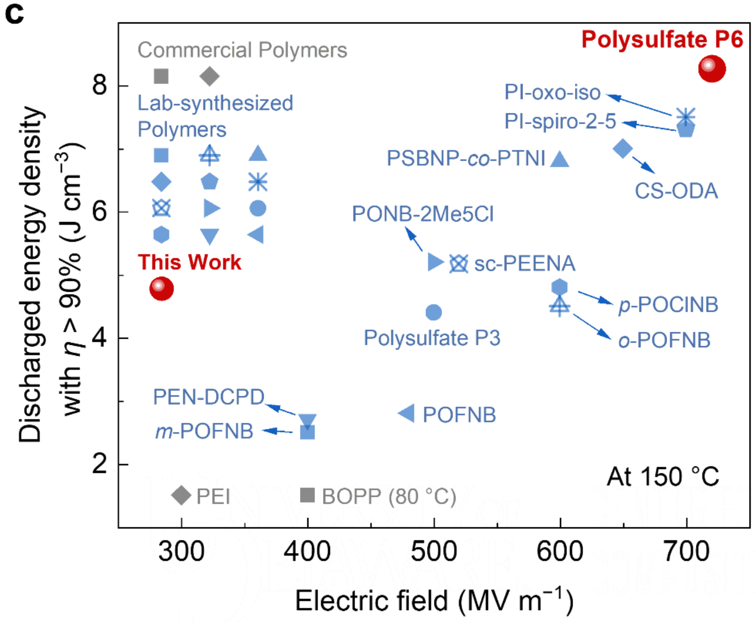
He Li, Hongbo Zheng, Tianle Yue, Zongliang Xie, Shaopeng Yu, Ji Zhou, Topprasad Kapri, Yunfei Wang, Zhiqiang Cao, Haoyu Zhao, Aidar Kemelbay, Jinlong He, Ge Zhang, Priscilla Pieters, Eric Dailing, John Cappiello, Miquel Salmeron, Xiaodan Gu, Ting Xu, Peng Wu, Ying Li†, Karl Sharpless, and Yi Liu. (2024) ***Nature Energy***. Accepted.

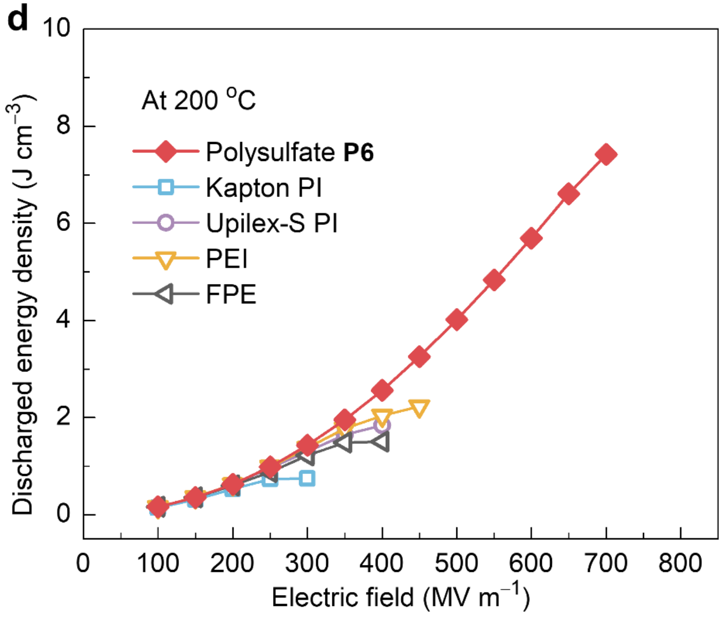
Predictions on thermal and electronic parameters

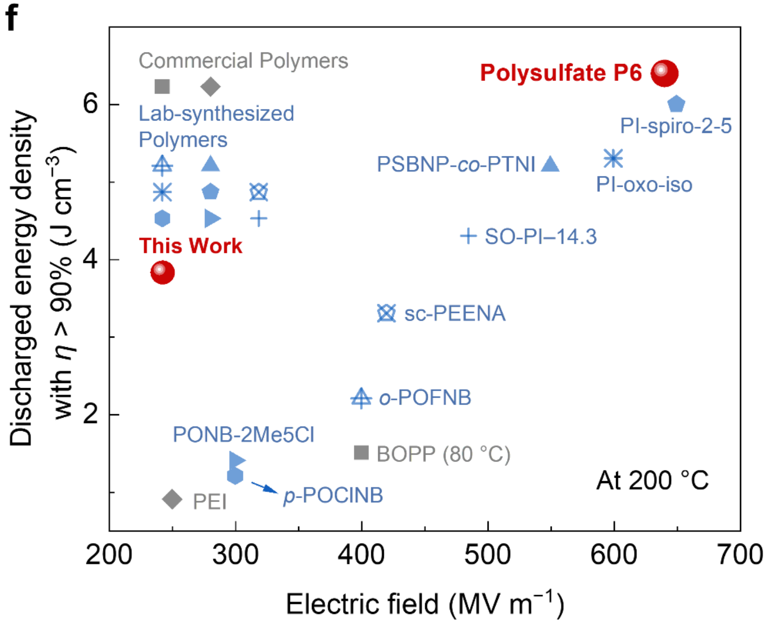


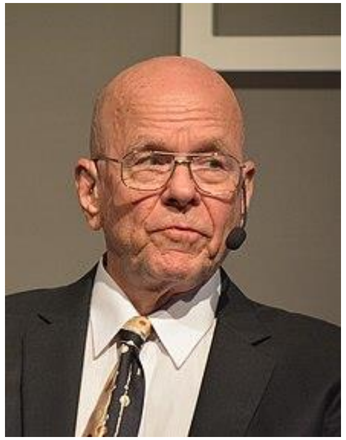
Electrostatic energy storage and reliability





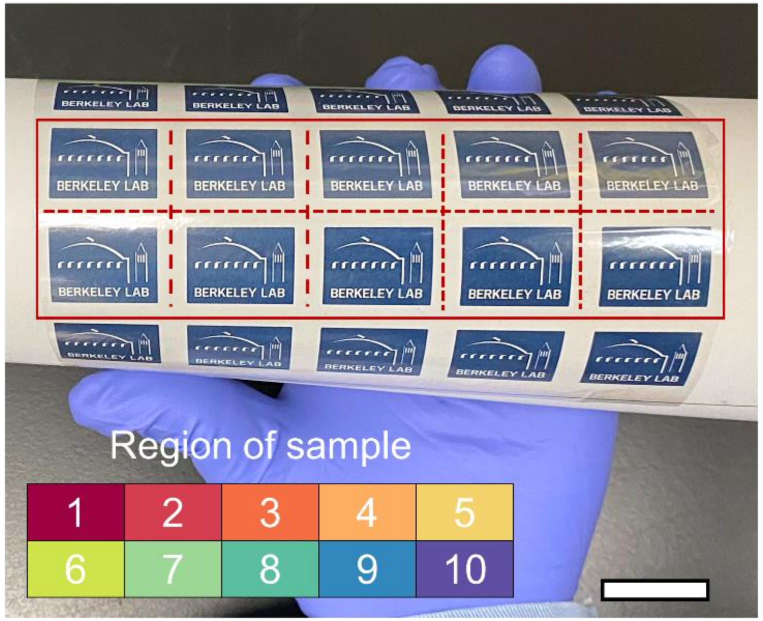






Karl Barry Sharpless
Scripps Research

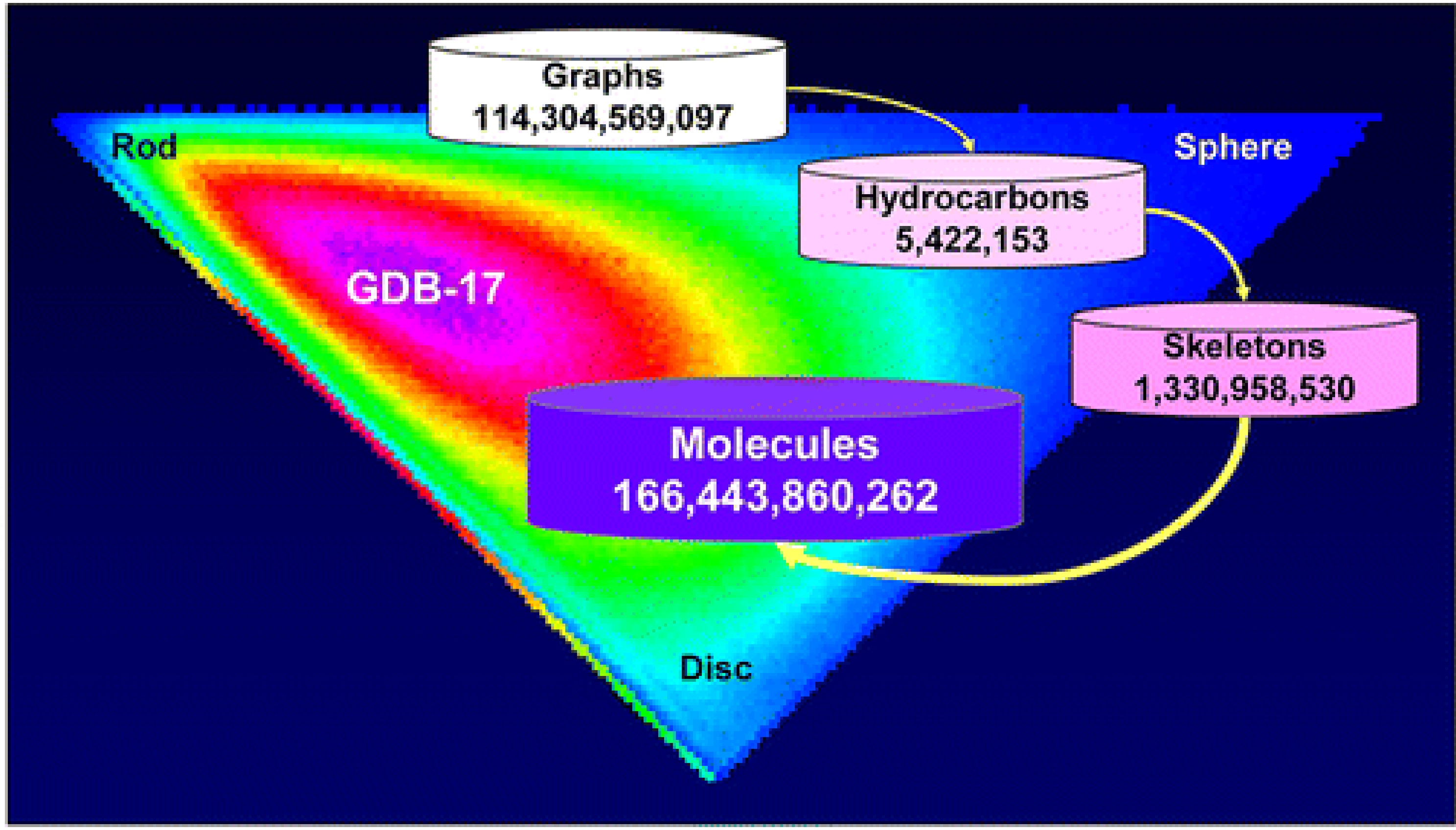
Yi Liu, Lawrence Berkeley
National Lab



Roll of the polysulfate P6 film

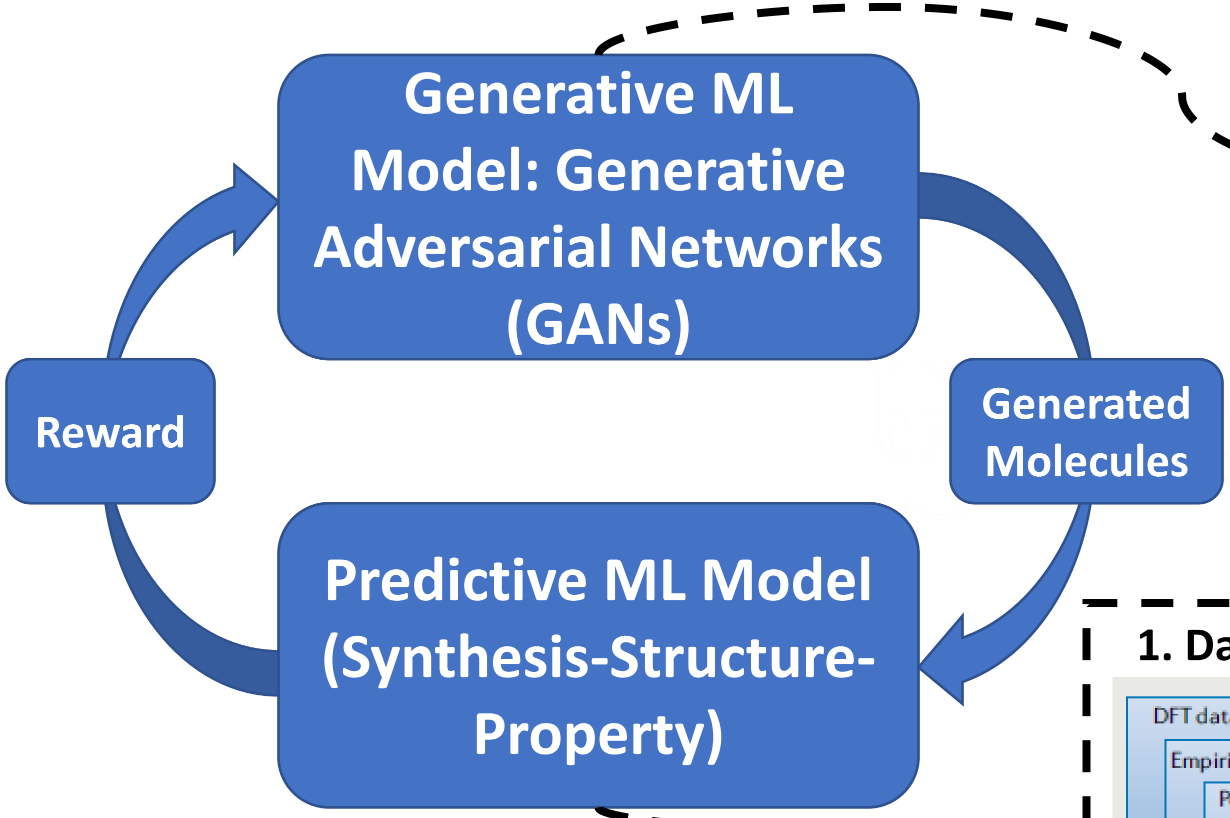
Scientific challenge in molecular design of polymers

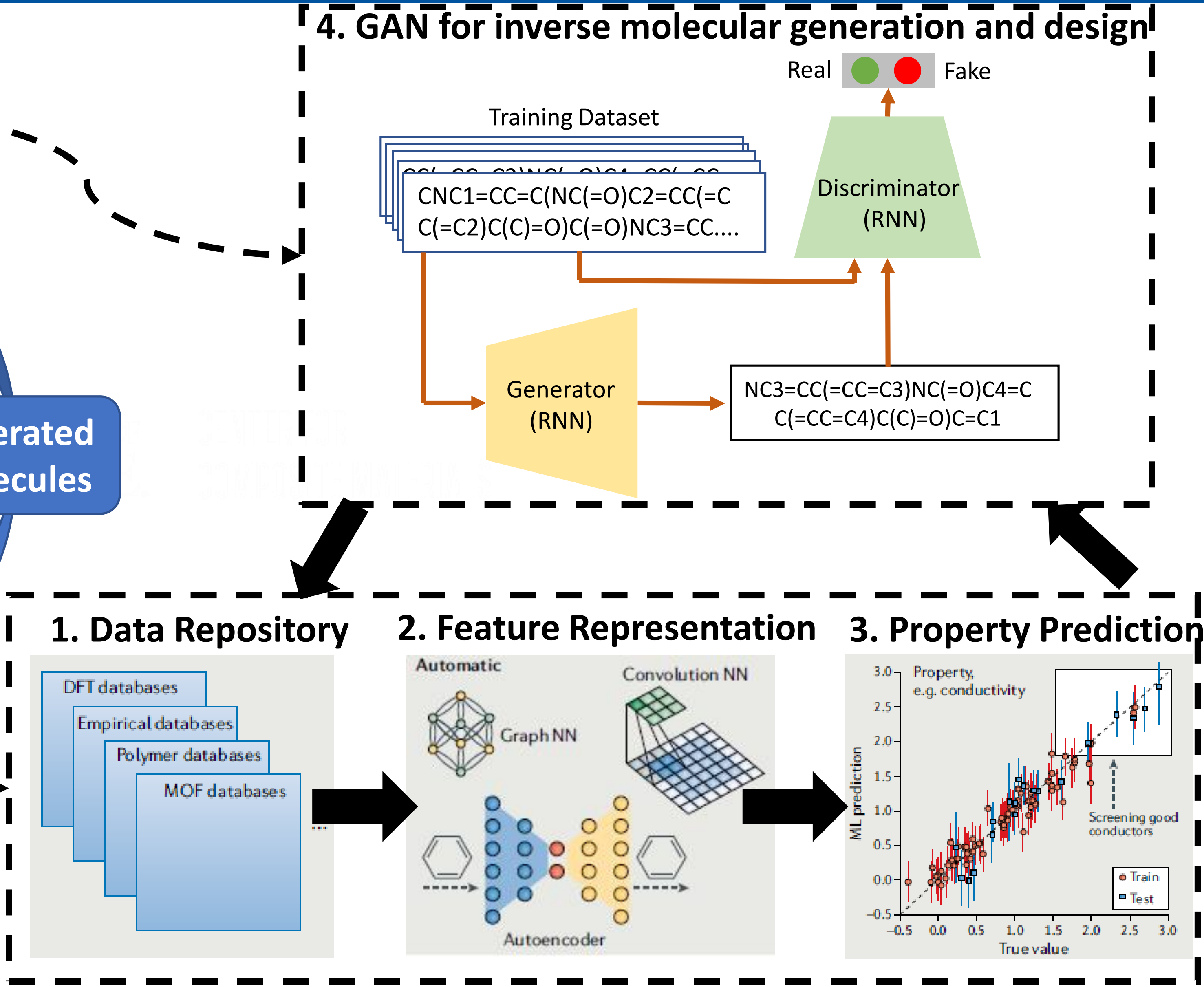
LEGO building blocks



For example, GDB-17 database enumerates small organic molecules up to <u>17 atoms of C, N, O, S, and halogens</u> following all possible chemical structures, resulting in **>166.4 billion** molecule designs. [J. Chem. Inf. Model. 2012, 52, 11, 2864-2875]

ML-directed Molecular Design of Polymer





Chen, Guang, Zhiqiang Shen, Akshay Iyer, Umar Farooq Ghumman, Shan Tang, Jinbo Bi, Wei Chen, and **Ying Li**. "Machine-Learning-Assisted De Novo Design of Organic Molecules and Polymers: Opportunities and Challenges." *Polymers* 12, no. 1 (2020): 163.

Step 1 Data Repository and Chemical Space

Dataset	Number of Polymers	T_g (°C)	Source
Dataset-1	6923	-118~495	Real polymers from PoLyInfo
Dataset-2	5690	Unknown	Real polymers from PoLyInfo

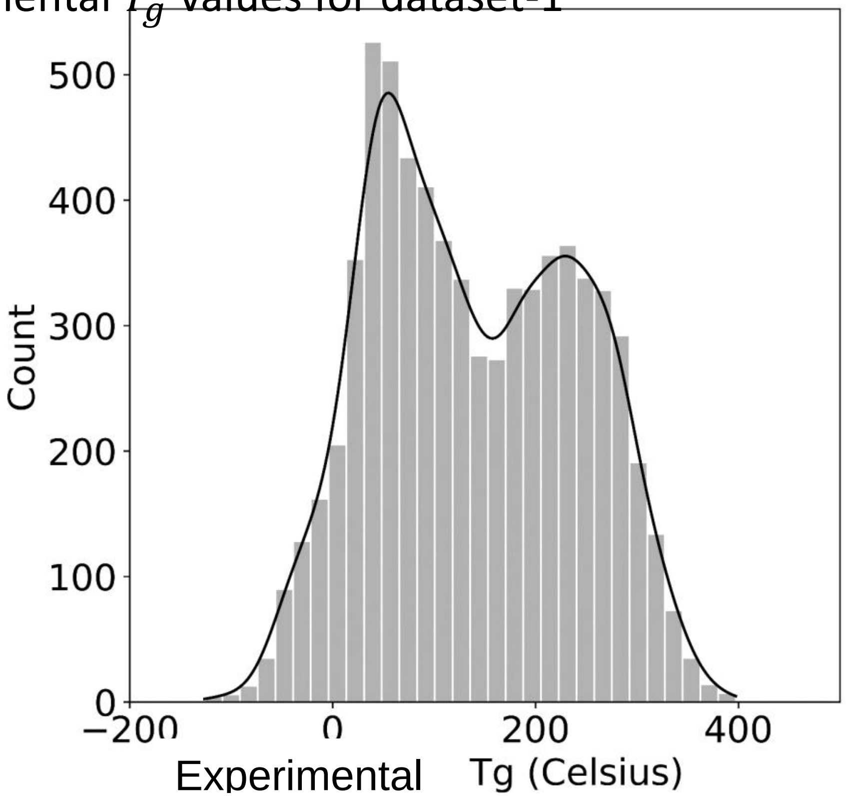
Principal Component Analysis (PCA)

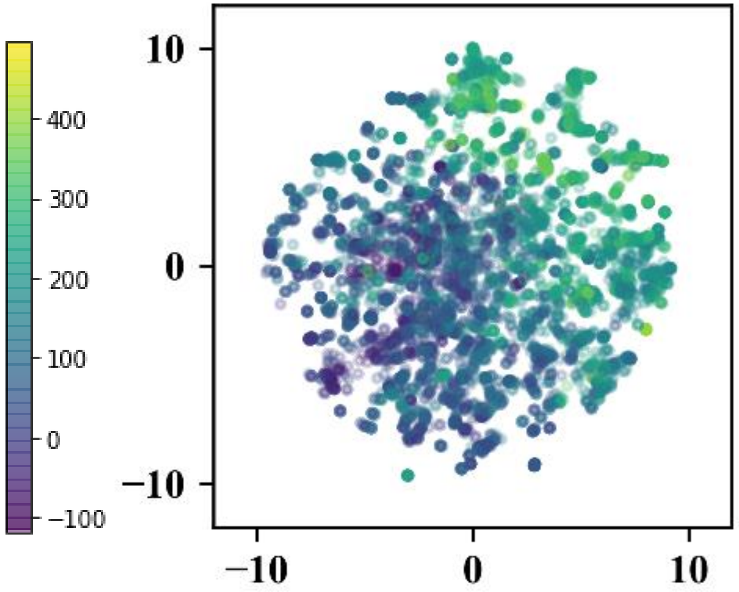
t-distributed stochastic neighbor embedding (t-SNE)

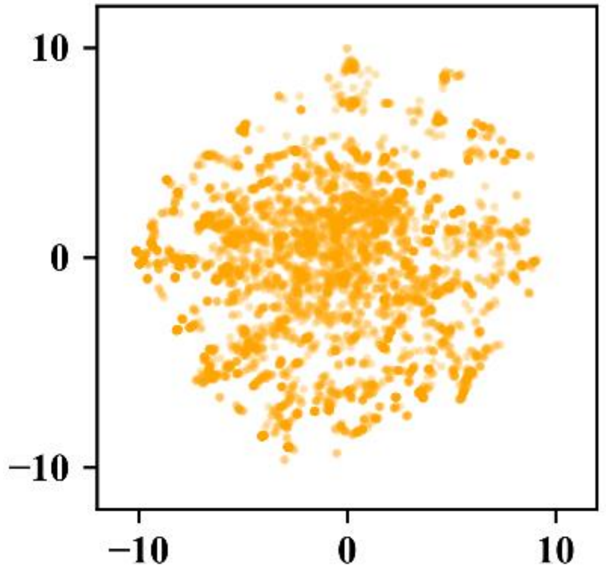
Dataset-1
6,923 polymers (w/ T_g)

Dataset-2
5,690 polymers (w/o T_g)

Experimental T_g values for dataset-1







Polymers without T_g (Dataset-2) → Data generation → Glass transition temperature T_g

via Laborious computations or experiments

Feature engineering → Instant property prediction

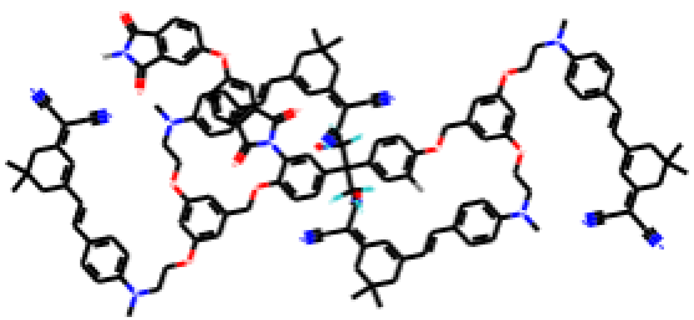
via **Machine Learning** on Dataset-1

Lei Tao, Guang Chen, **Ying Li**, 2021, "Machine Learning Discovery of High-Temperature Polymers", Cell/Patterns

Step 2 Feature Engineering and Representation

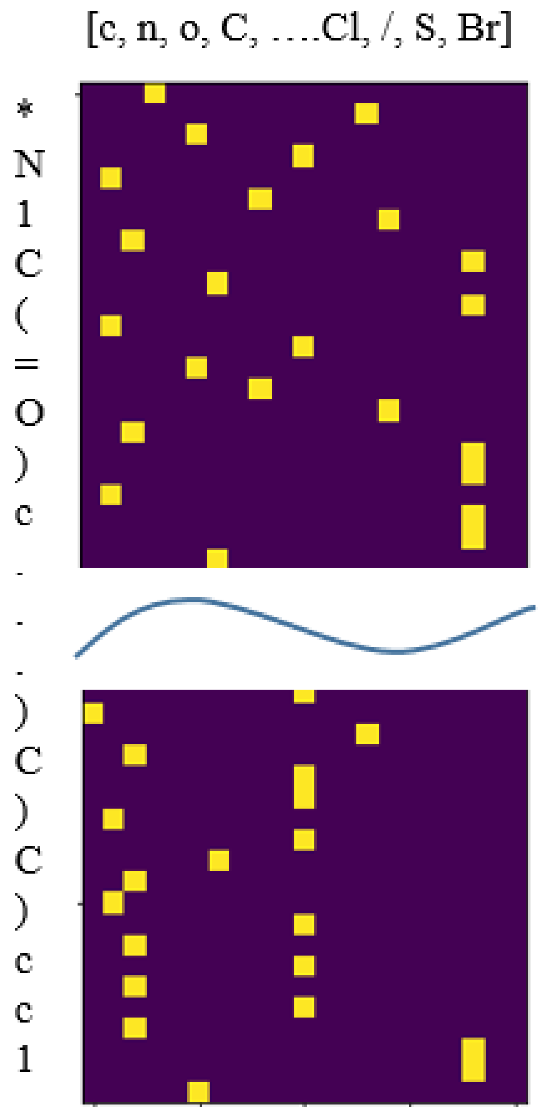
[632.88, 7.911, 80.3268, 54.5686, 89.8954,1.004085, 0.682107, 1.123692, 0.04862, 80 ... 0.5, 0.333333, 0.5, 0.8, 1.0, 0.333333, 0.0, 0.495238, 0.166667, 0.5]

Descriptor



N1C(=O)c2c(C1=O)cc(cc2)Oc1cc2c(C(=O)N(C2=O)c2c(ccc(c2)C(c2cc(c(cc2)OCc2cc(cc(c2)OCCN(c2ccc(cc2)C=CC2=CC(=C(C#N)C#N)CC(C2)(C)C)C)OCCN(c2ccc(cc2)C=CC2=CC(=C(C#N)C#N)CC(C2)(C)C)C))(C(F)(F)F)C(F)(F)F)OCc2cc(cc(c2)OCCN(c2ccc(cc2)C=CC2=CC(=C(C#N)C#N)CC(C2)(C)C)C)OCCN(c2ccc(cc2)C=CC2=CC(=C(C#N)C#N)CC(C2)(C)C)C)cc1

Image



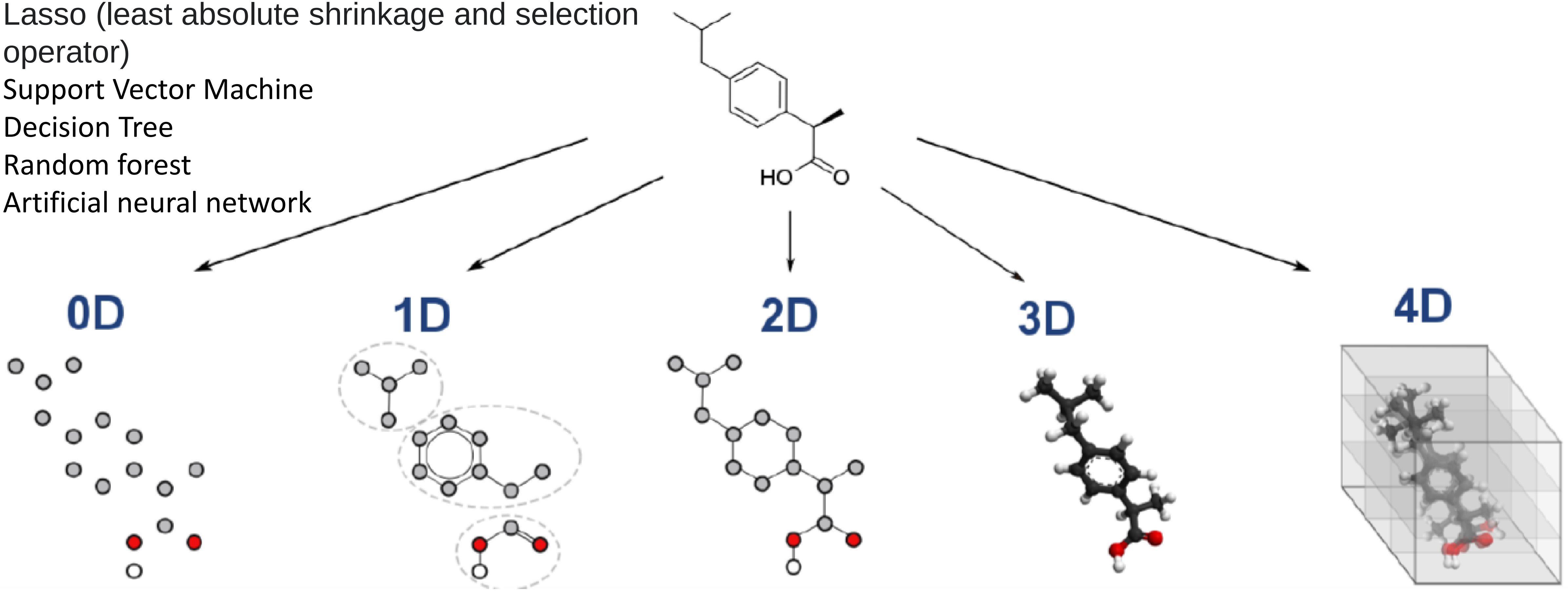
Fingerprint

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0..... 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0]

Three types of feature representation calculated based on the polymer's SMILES (simplified molecular-input line-entry system) notation for ML models: molecular descriptor, Morgan fingerprint, and image.

Molecular Descriptors and Structure-Activity Relationship

- Lasso (least absolute shrinkage and selection
- operator)
- Support Vector Machine
- Decision Tree
- Random forest
- Artificial neural network



- Atom counts
- Molecular weight
- Atomic properties

- Fragment counts
- Fragment presence

- Topo-structural
- Topo-chemical

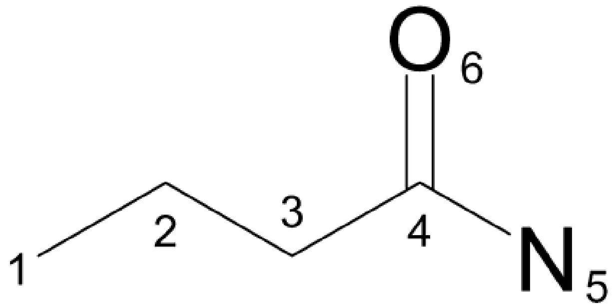
- Geometrical
- Atomic coordinates

- Grid-based
- Ensemble-based

Todeschini, Roberto, and Viviana Consonni. *Handbook of molecular descriptors*. John Wiley & Sons, 2008.

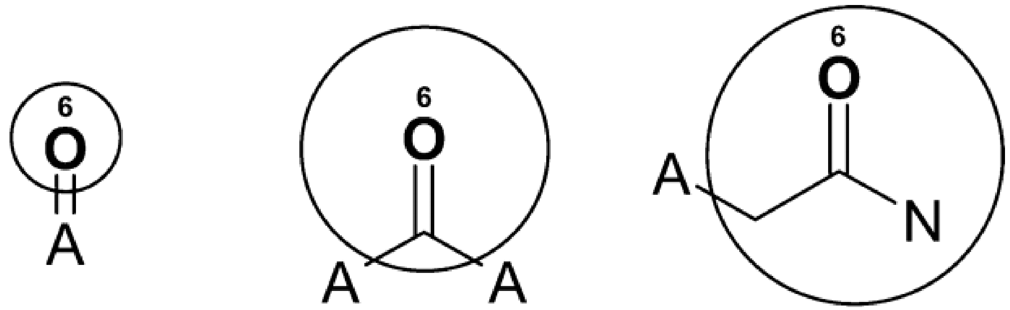
Extended Connectivity FingerPrinting (ECFP)

1. Assign each atom with an identifier



1: 734603939
2: 1559650422
3: 1559650422
4: -1100000244
5: 1572579716
6: -1074141656

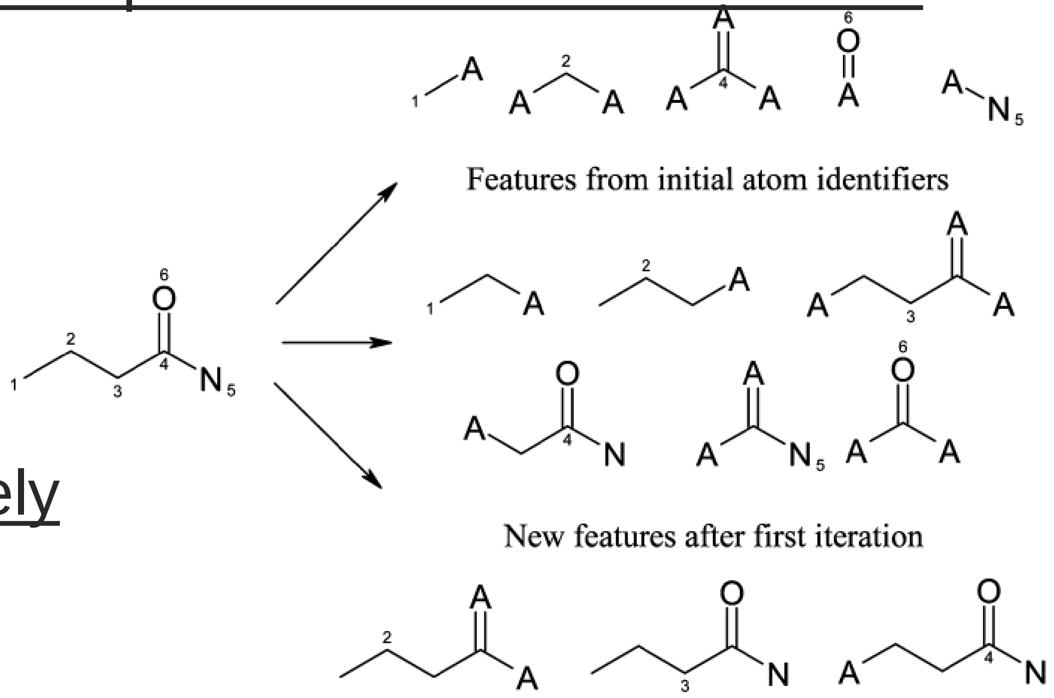
2. Update the identifiers of each atom, iteratively



Iteration 0 Iteration 1 Iteration 2

Rogers, David, and Mathew Hahn. "Extended-connectivity fingerprints." *Journal of Chemical Information and Modeling* 50, no. 5 (2010): 742-754.

3. Duplicate substructure removal

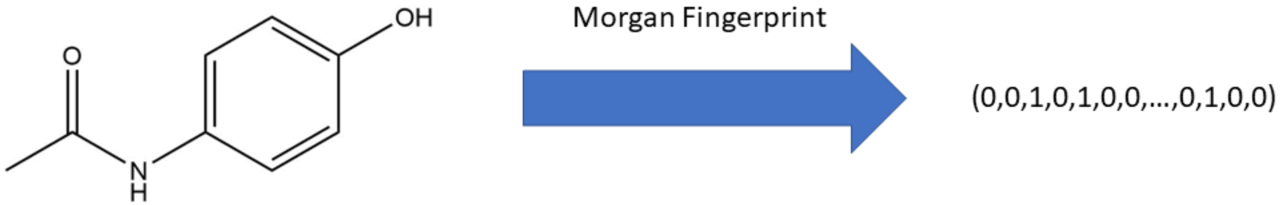


Features from initial atom identifiers

New features after first iteration

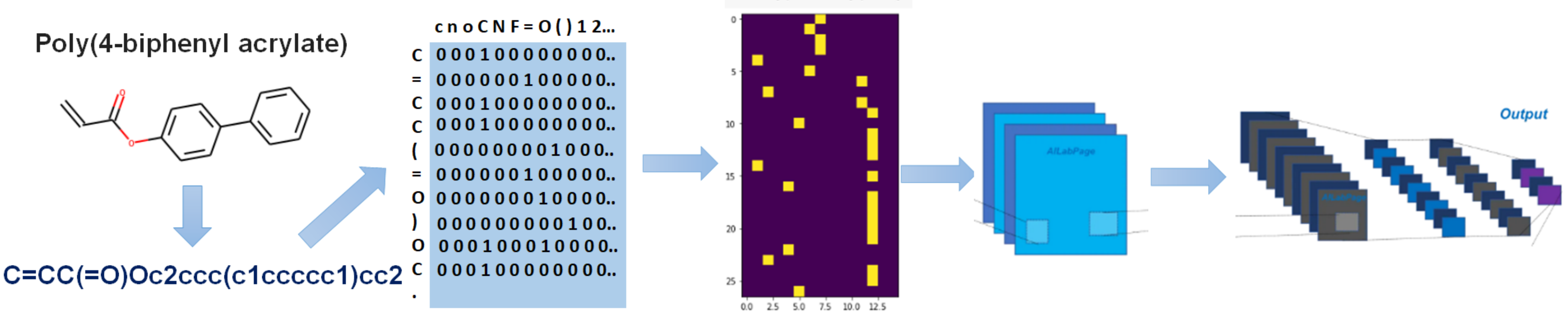
New features after second iteration
(additional iterations discover no new features)

4. Fold list of identifiers into a 2048-bit vector



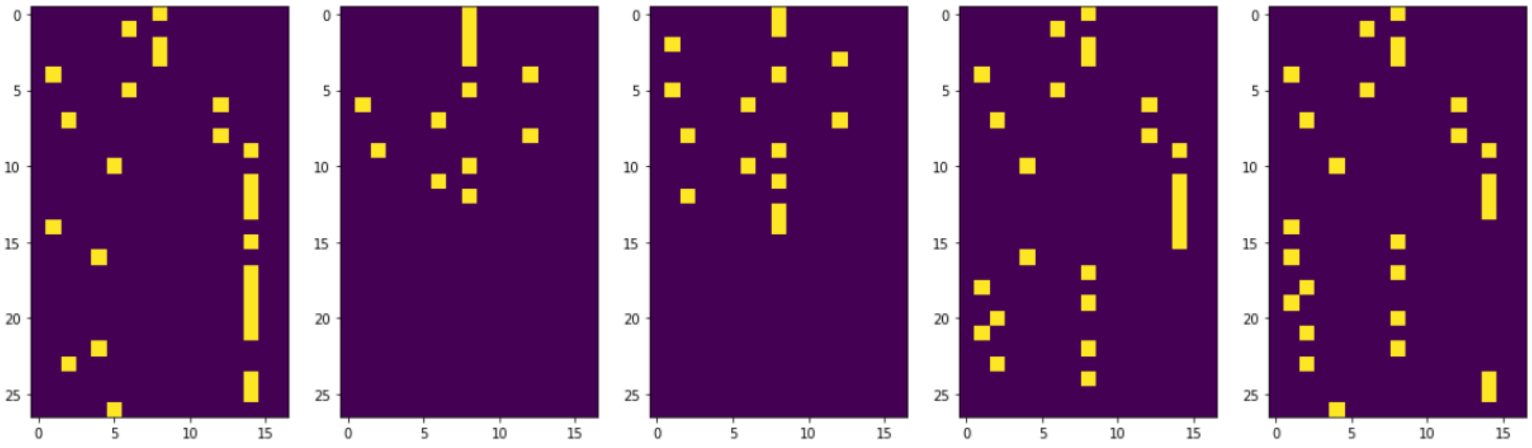
paracetamol

Using Image to represent the polymers



SMILES Code | One-Hot Encoding | Encoded Image | Convolution | Pooling and Fully Connected Layers

	Nomenclature Name	Molecular Structure	Tg	Class_of_Polymer
0	Poly(4-biphenyl acrylate)	C=CC(=O)Oc2ccc(c1ccccc1)cc2	383.0	
1	Poly(butyl acrylate)	CCCCOC(=O)C=C	219.0	
2	Poly(sec-butyl acrylate)	CC(OC(=O)C=C)CC	250.0	
3	Poly(2-tertbutylphenyl acrylate)	C=CC(=O)Oc1ccccc1C(C)(C)C	345.0	
4	Poly(4-tertbutylphenyl acrylate)	C=CC(=O)Oc1ccc(C(C)(C)C)cc1	344.0	



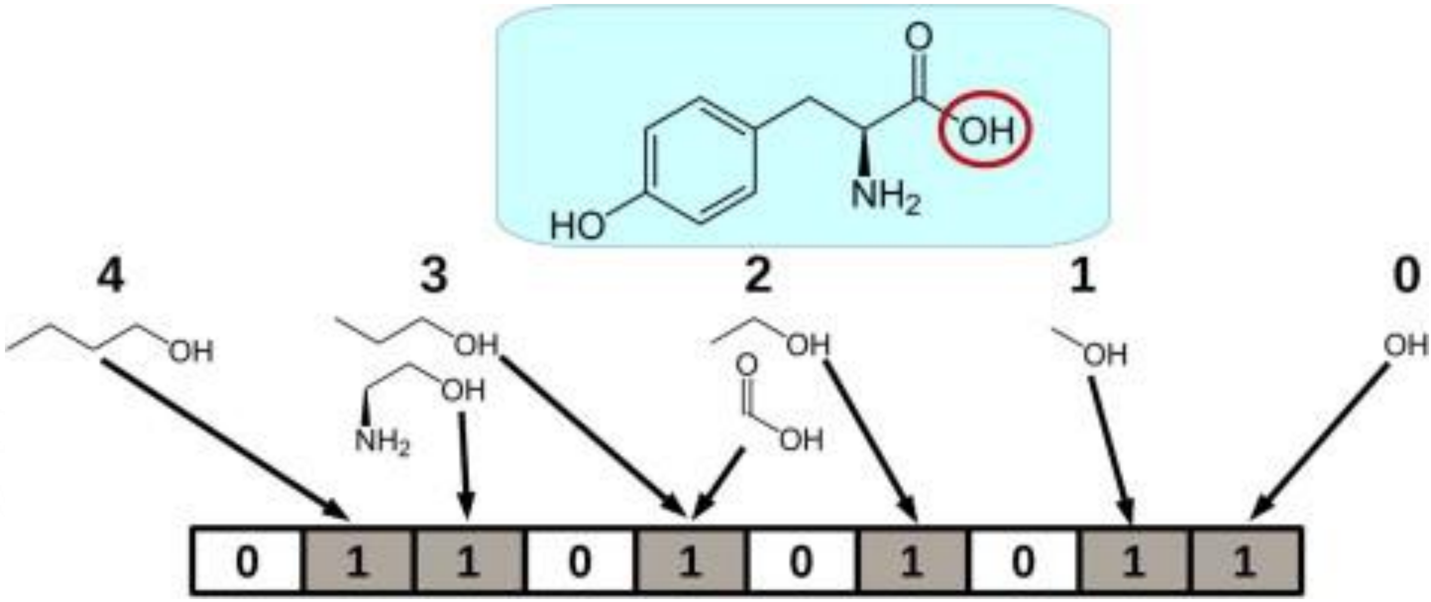
Feature Engineering Examples

Comparison Between Feature Representations

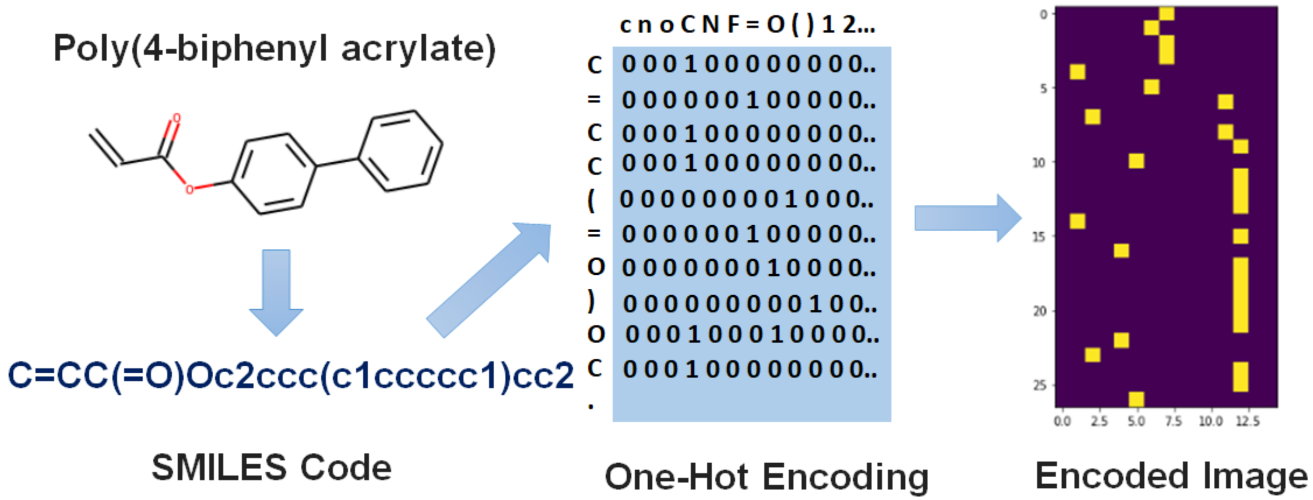
Molecular Descriptors (~5000): Very time-consuming

N	Group
1	Constitutional descriptors
2	Topological descriptors
3	Walk and path counts
4	Connectivity indices
5	Information indices
6	2D autocorrelations
7	Edge adjacency indices
8	BCUT descriptors
9	Topological charge indices
10	Eigenvalue-based indices
11	Randic molecular profiles
12	Geometrical descriptors
13	RDF descriptors
14	3D-MoRSE descriptors
15	WHIM descriptors
16	GETAWAY descriptors
17	Functional group counts
18	Atom-centred fragments
19	Charge descriptors
20	Molecular properties

Fingerprints (hash 45,000 distinct substructures into 2048 bits)



Images (a sparse matrix, 21×310): Very fast



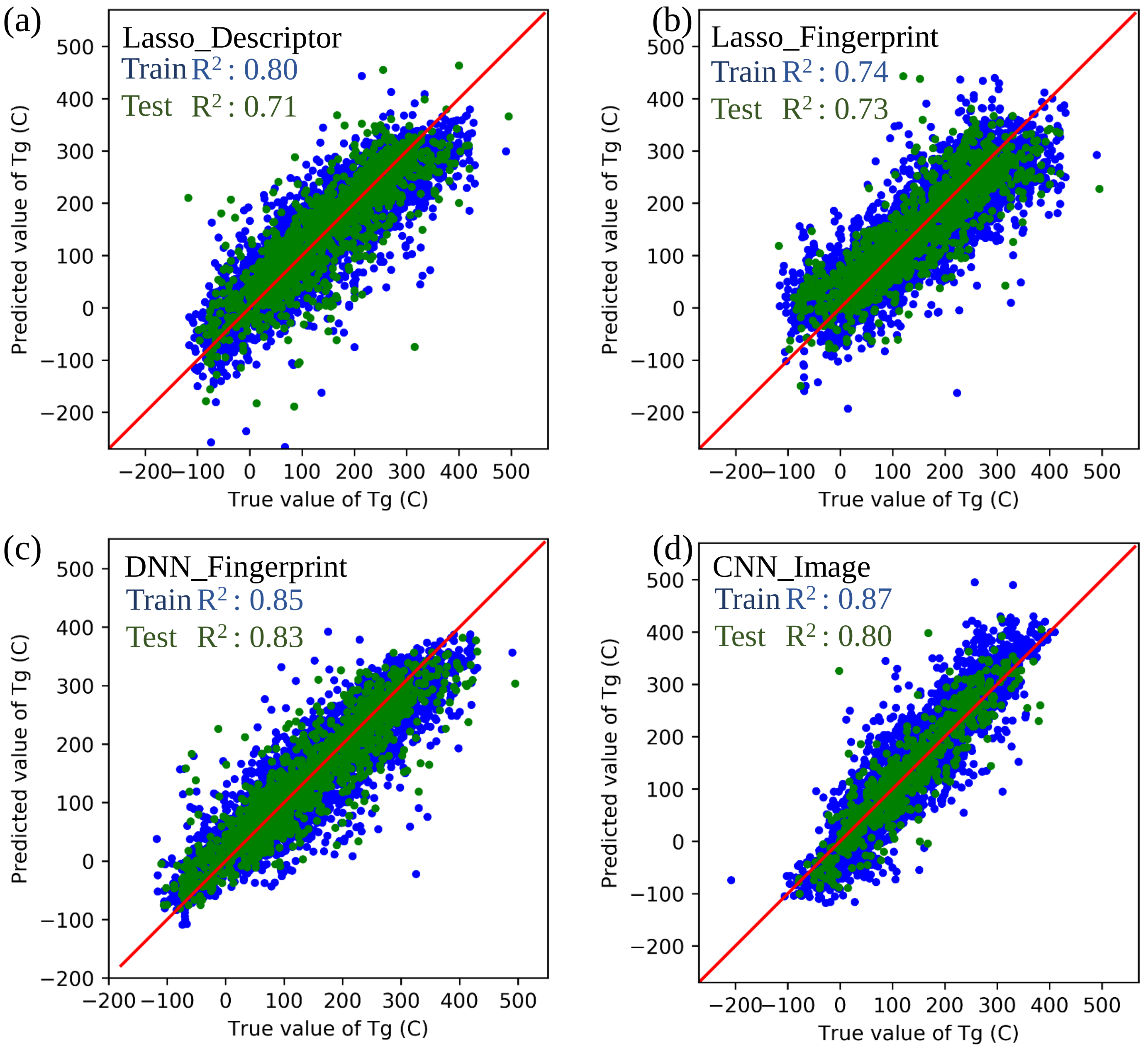
Step 3 Property Prediction (Predictive ML Model)

<u>Three different feature representations</u>

- Molecular descriptors
- Morgan fingerprints
- Images

<u>Three different ML models</u>

- Least absolute shrinkage and selection operator (Lasso) regression
- Deep neural network (DNN)
- Convolutional neural network (CNN)

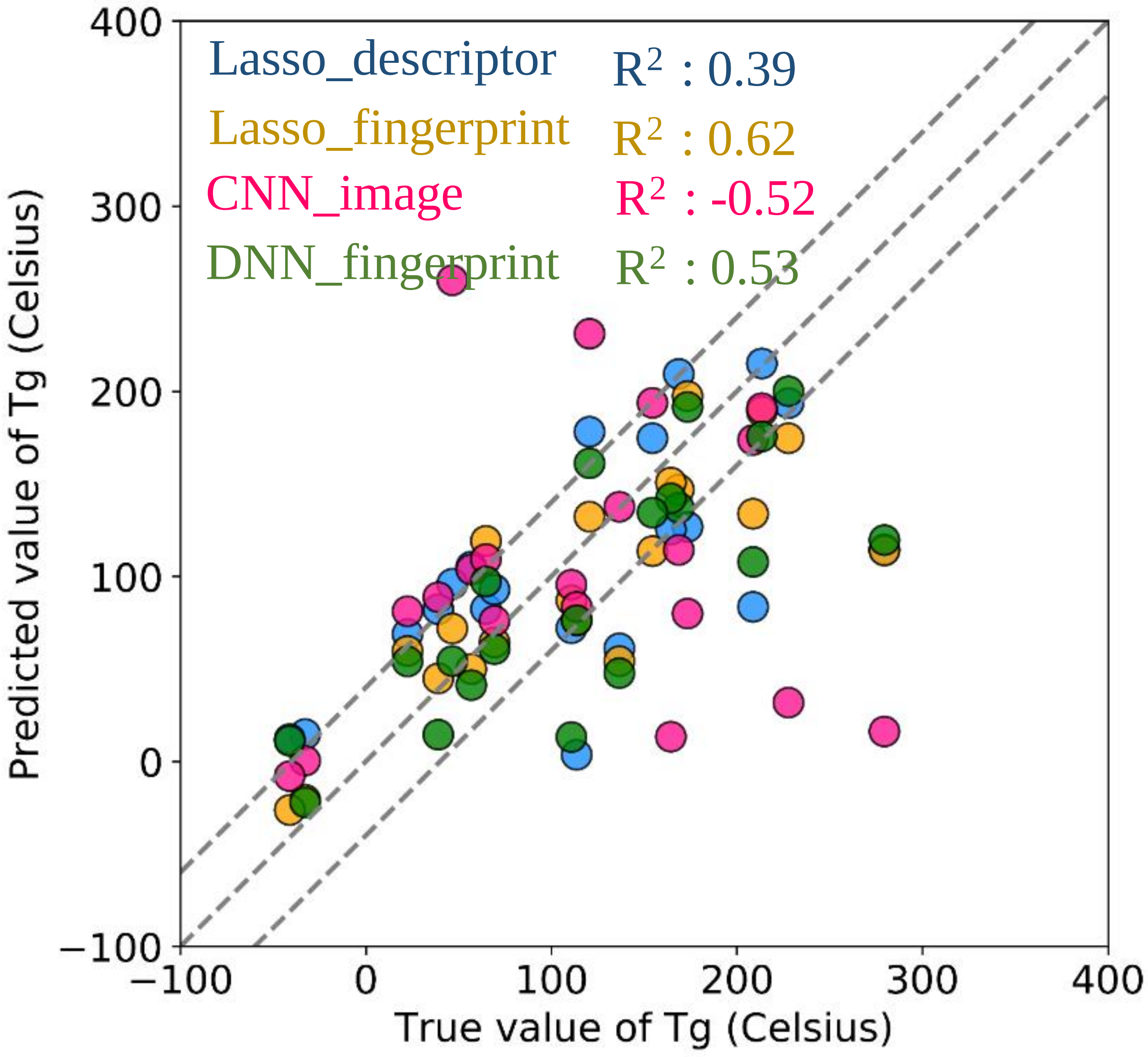


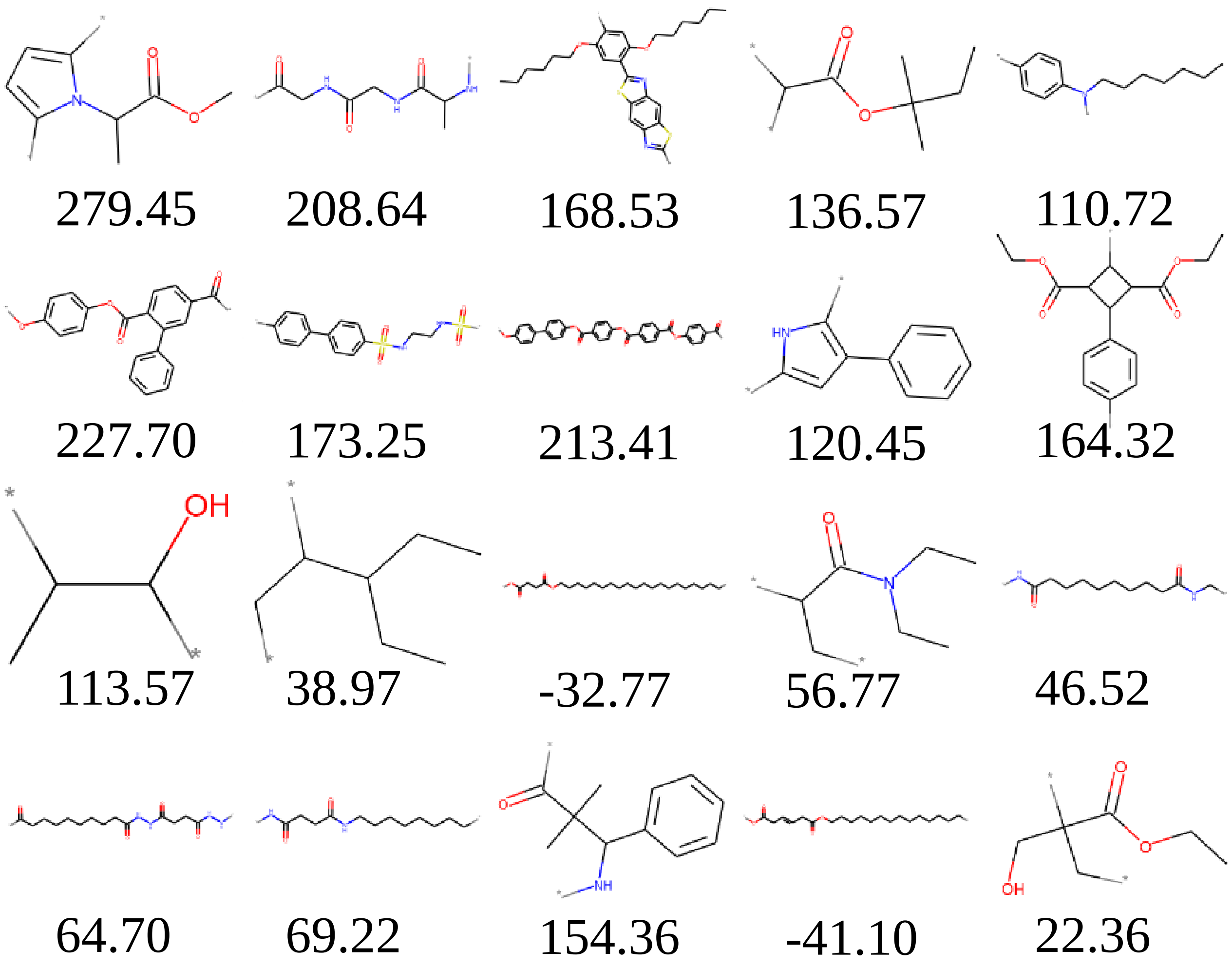
<u>Four predictive ML models trained on dataset-1</u>

Name	ML Model	Features	R^2 (train/test)
Lasso_Descriptor	Lasso regression model	3579 descriptors	0.80/0.71
Lasso_Fingerprint	Lasso regression model	2048 fingerprints	0.74/0.73
DNN_Fingerprint	Deep neural network	2048 fingerprints	0.85/0.83
CNN_Image	Convolutional neural network	310×21 binary images	0.87/0.80

Which ML model is trustworthy on dataset-2?

Comparison between the MD simulated T_g and the ML predicted T_g on 20 polymers randomly selected from dataset-2.

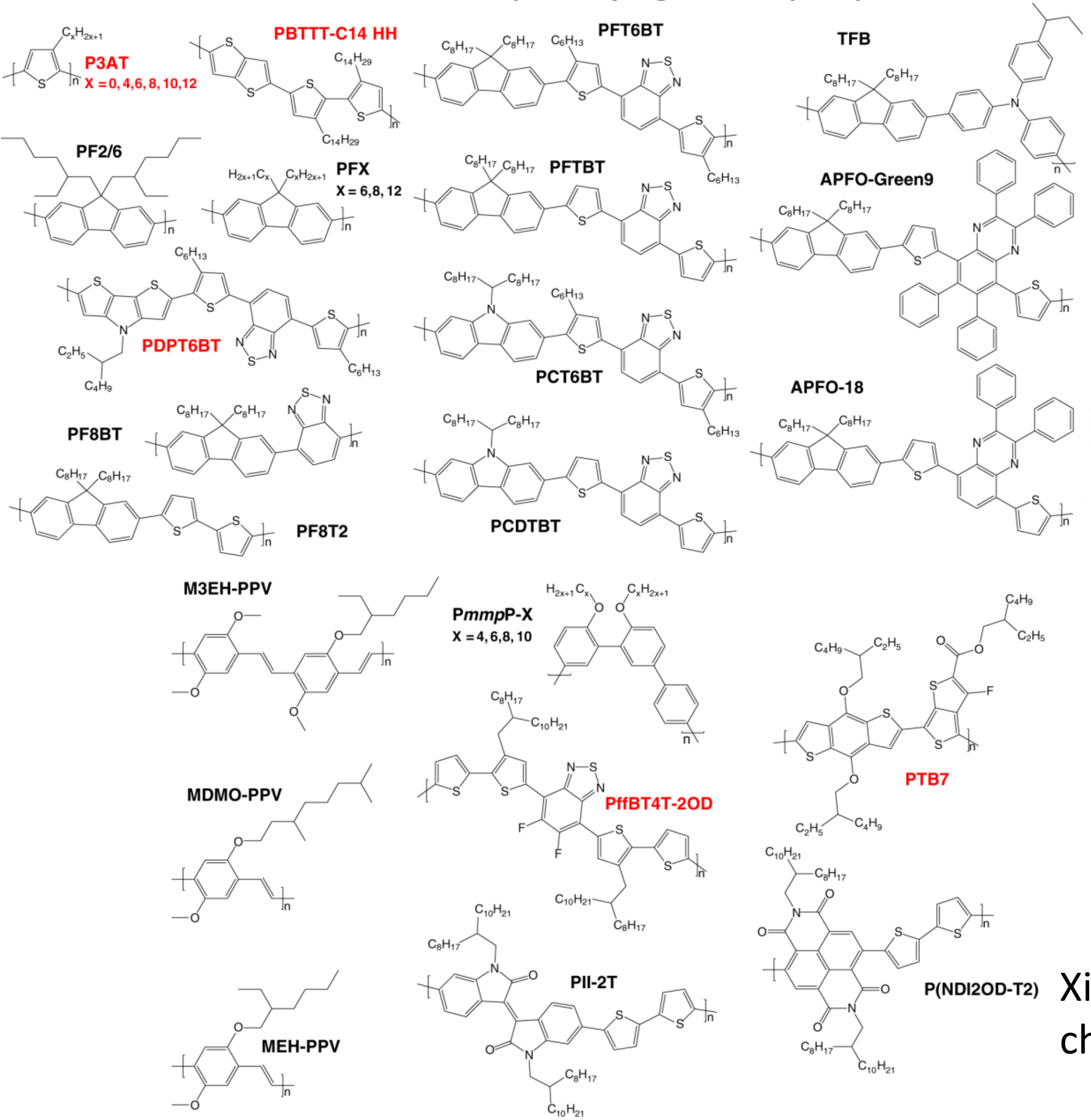


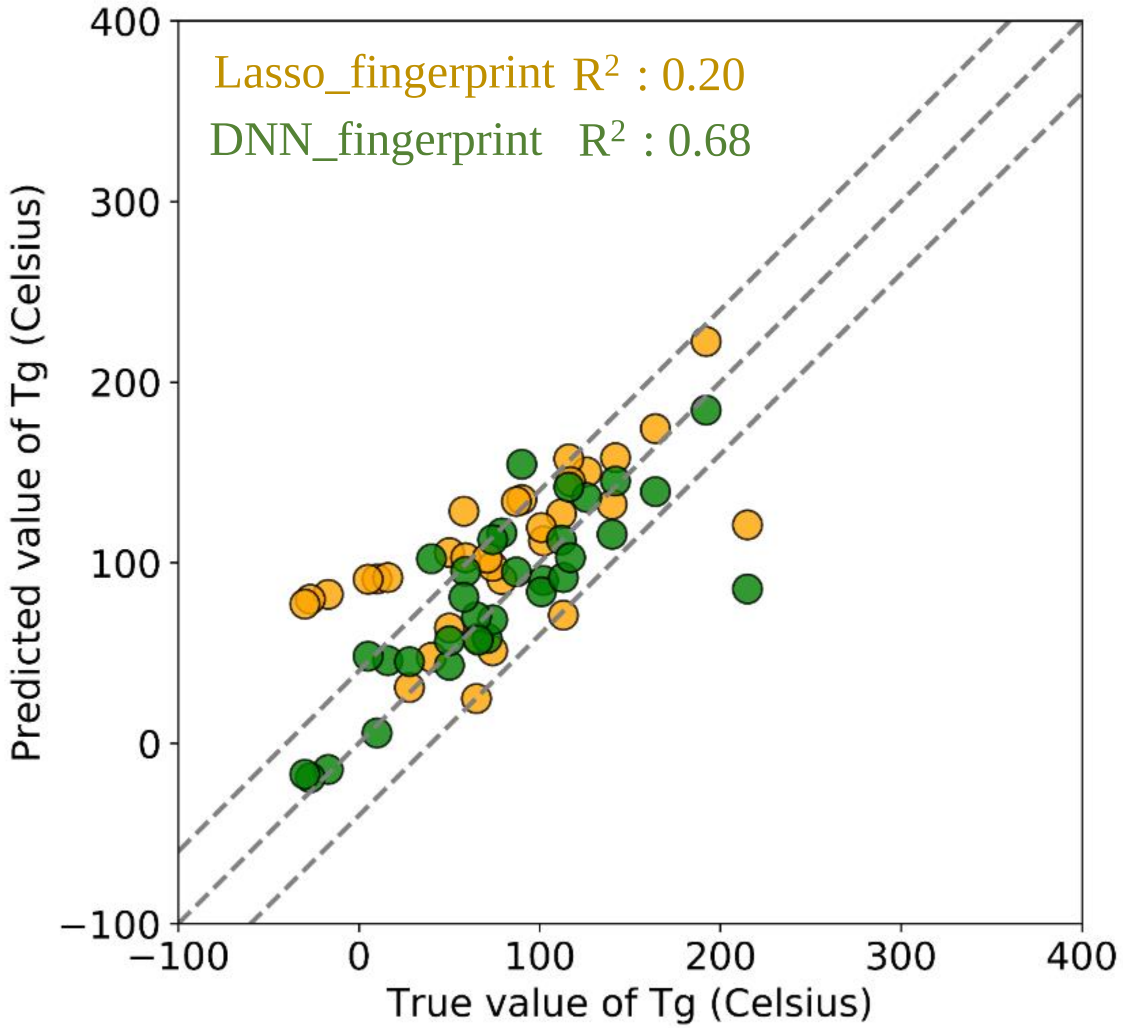


Lei Tao, Guang Chen, **Ying Li**, 2021, “Machine Learning Discovery of High-Temperature Polymers”, Cell/Patterns

Which ML model is trustworthy on experiments?

32 semiflexible (mostly conjugated) polymers that differ drastically in aromatic backbone and alkyl side chain chem





Xie, R., Weisen, A.R., Lee, Y. et al. Glass transition temperature from the chemical structure of conjugated polymers. Nat Commun 11, 893 (2020)

[Ralph H. Colby & Enrique D. Gomez @ Penn State University]

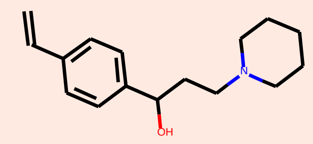
Benchmarking ML Models for Polymer Informatics

Q: Which machine learning (ML) model is trustworthy for polymer property (Tg) prediction?

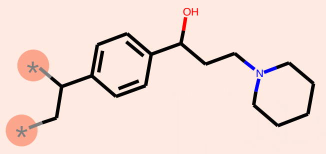
L Tao, V Varshney, Y Li
Journal of Chemical Information and Modeling
61 (11), 5395-5413, 2021

Structure

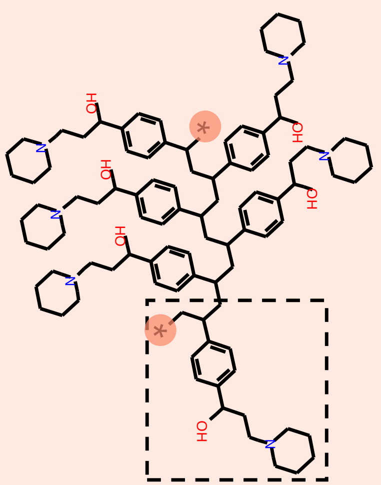
Monomer



Repeat unit



Repeat units polymerized

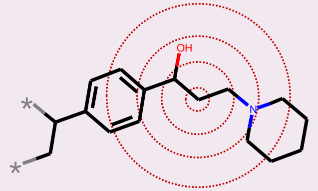


Feature

Smiles

C(C)c1ccc(cc1)C(CCN1CCCCC1)O

Morgan fingerprint



1	0	1	0	0	···	0	0	1	0	1	0

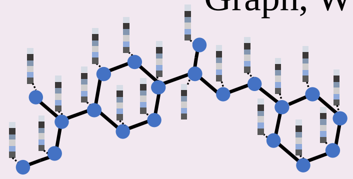
Morgan fingerprint with frequency

0	0	2	0	1	···	0	6	0	0	0	3

Descriptors

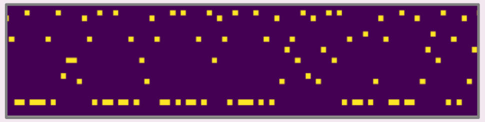
0.02	134	73	0.05	-5.1	···	53	46	6	177	0.6	0.8

Graph, Weave



Connection matrix
+
Atom feature vector
+
Bond feature vector

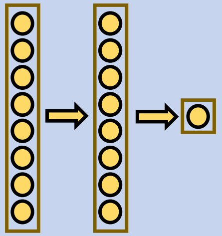
Image



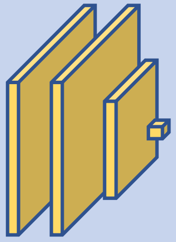
One-hot encoding matrix

Model

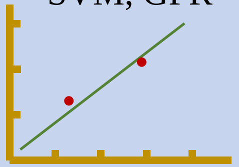
FFNN, RNN



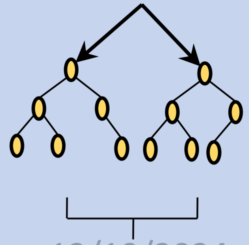
CNN, GCN



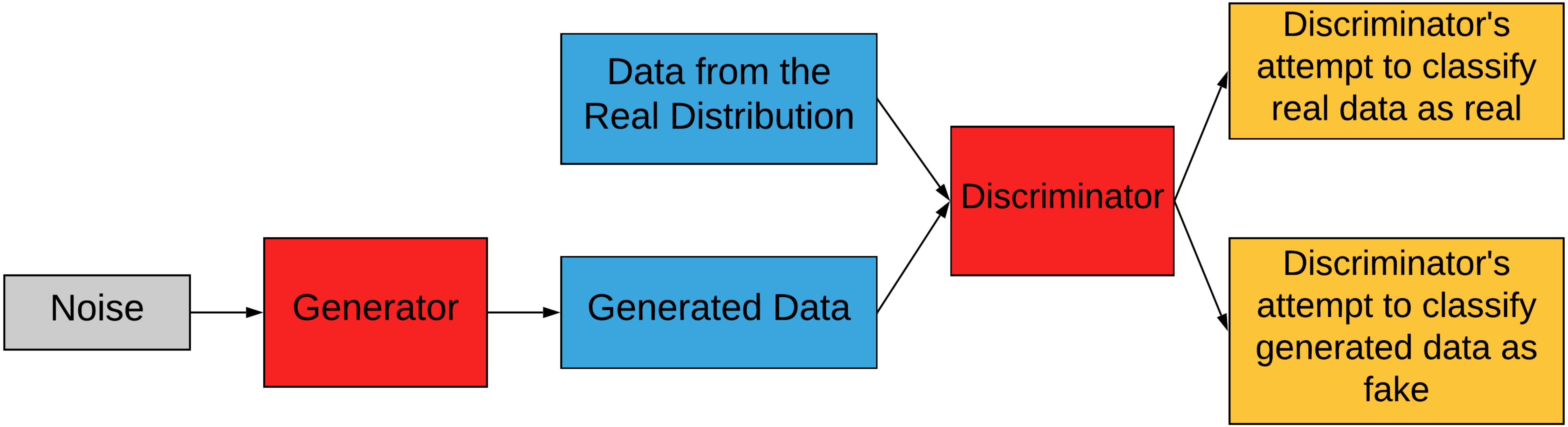
Linear Regression, SVM, GPR



Random Forest



Step 4 Inverse Molecular Generation and Design

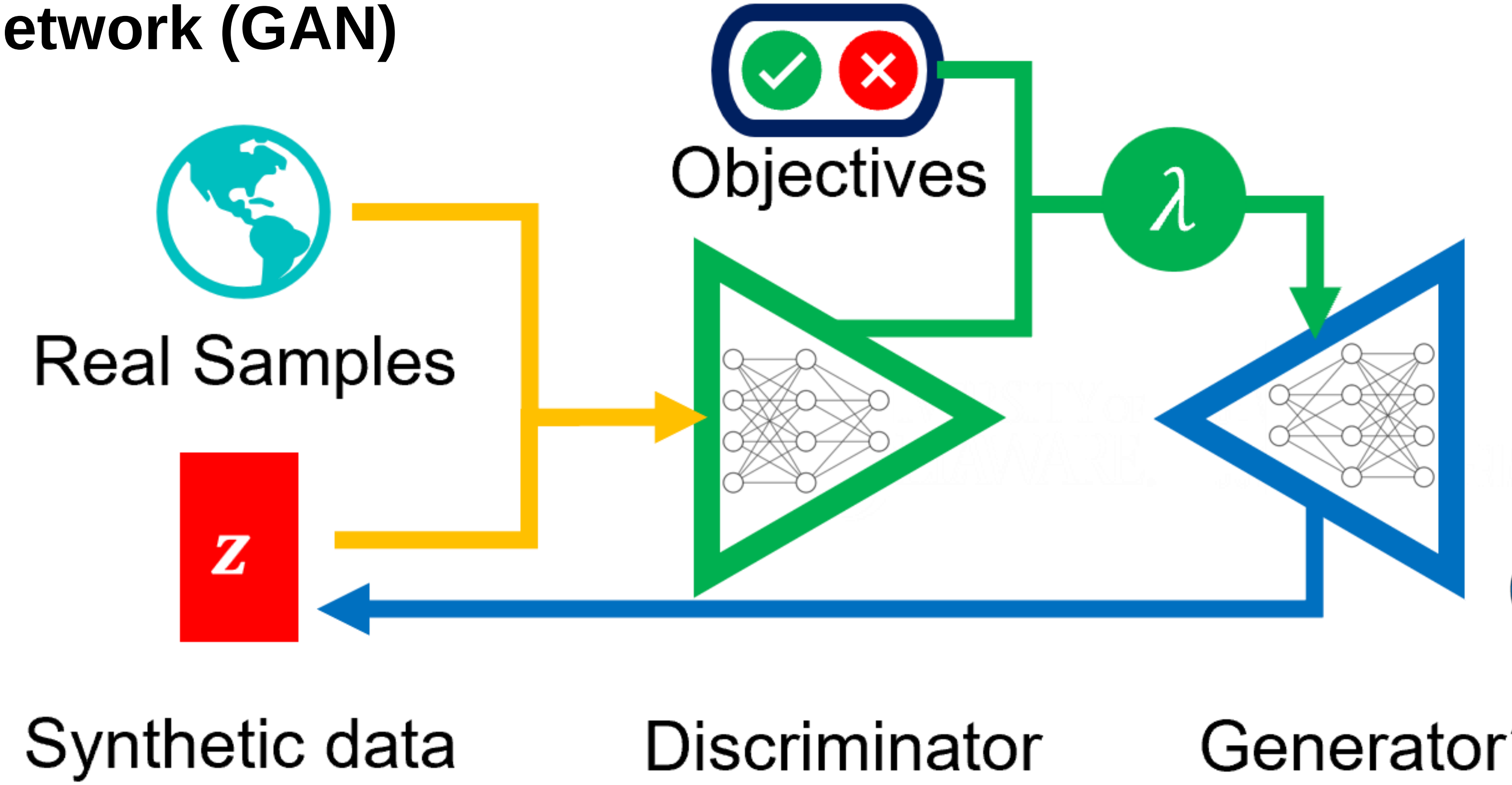


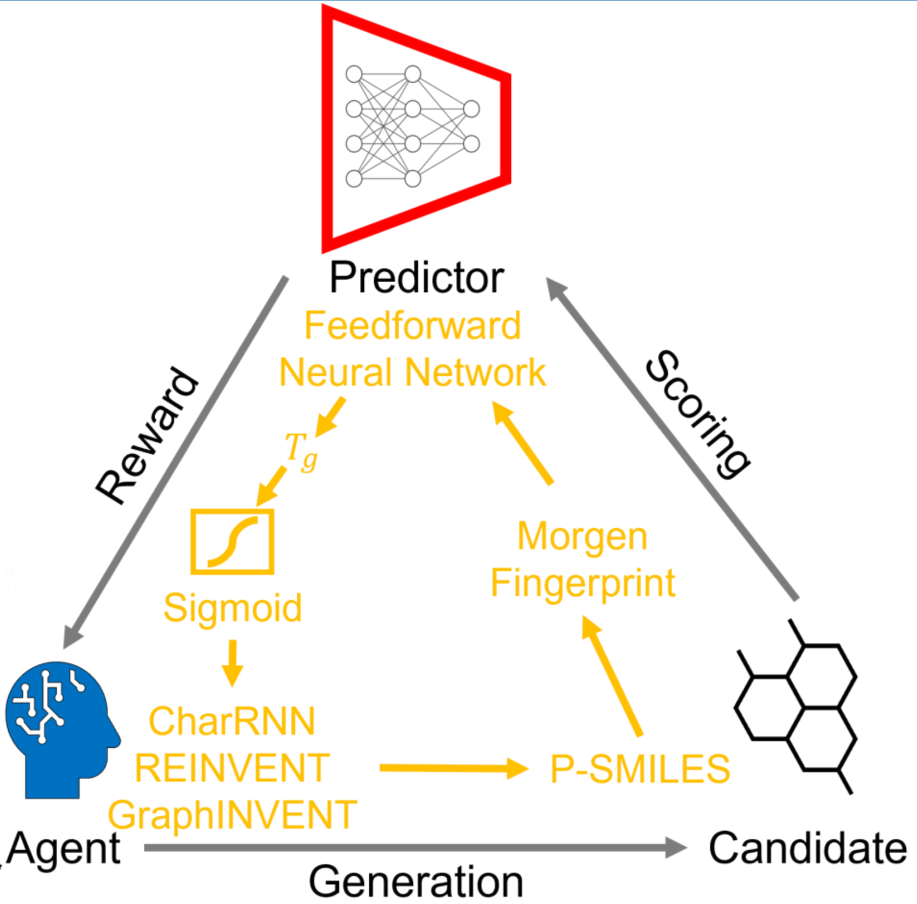
As training progresses, the generator gets closer to producing output that can fool the discriminator:

Finally, if generator training goes well, the discriminator gets worse at telling the difference between real and fake. It starts to classify fake data as real, and its accuracy decreases.

Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Yoshua (2014). Generative Adversarial Nets. Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014). pp. 2672–2680.

Generative Adversarial Network (GAN)

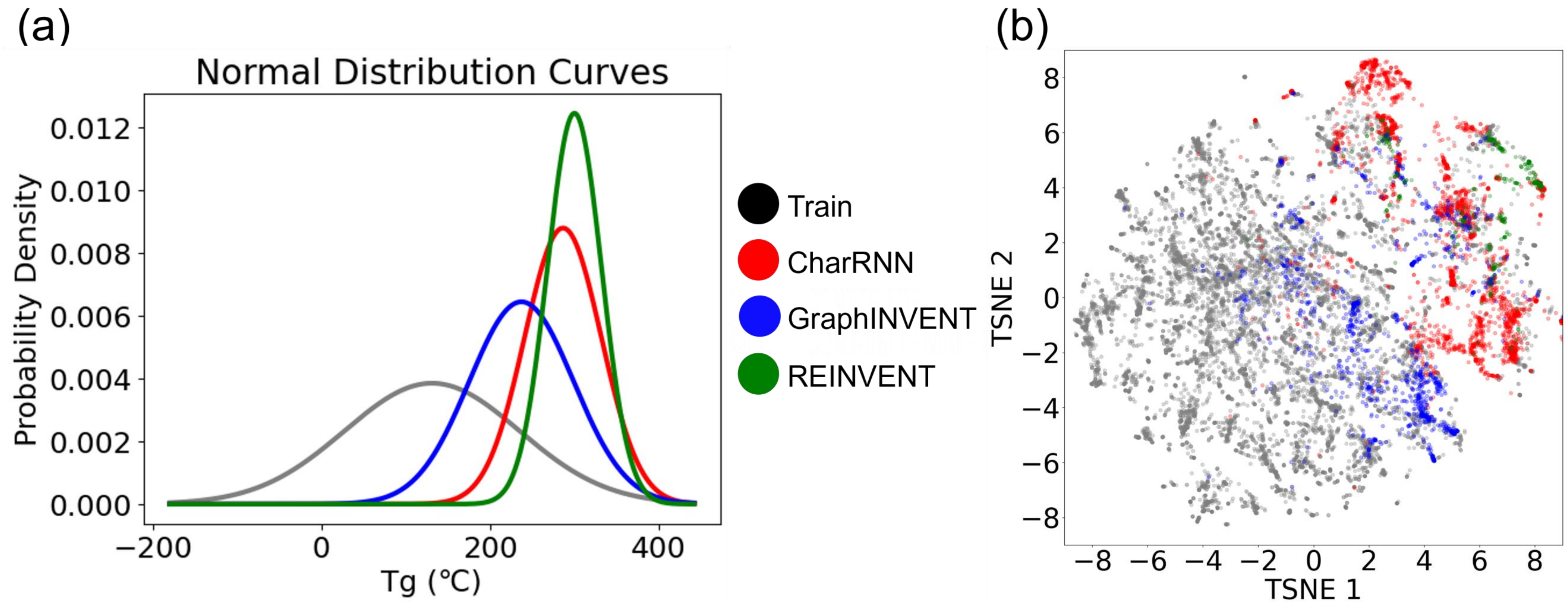




Sanchez-Lengeling, Benjamin, Carlos Outeiral, Gabriel L. Guimaraes, and Alan Aspuru-Guzik. "Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC)." *ChemRxiv* 2017 (2017).

Yue, Tianle, Lei Tao, Vikas Varshney, and Ying Li. "Benchmarking Study of Deep Generative Models for Inverse Polymer Design." *Digital Discovery* (2024).

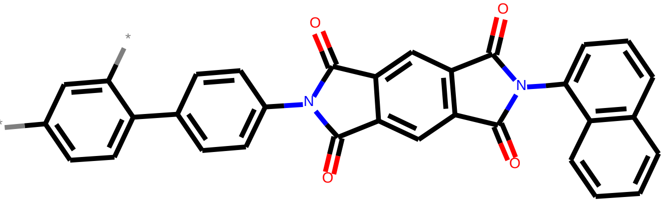
Reinforcement Learning for High Tg Polymers



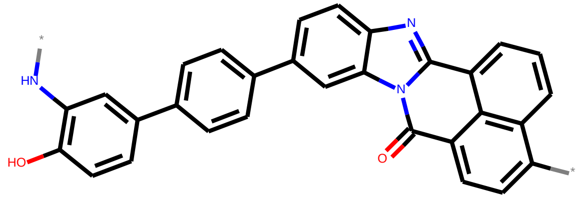
(a) Normalized probability density distribution of predicted Tg values and (b) chemical space distribution of the hypothetical valid unique polymers generated by CharRNN (red), GraphINVENT (blue), REINVENT (green), and the real polymers.

Molecular Simulations of New High-Tg Polymers

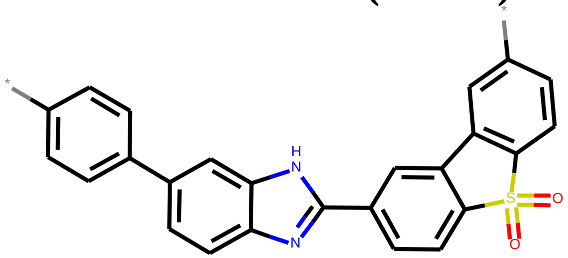
Enumerations



686.77 (658.42) K

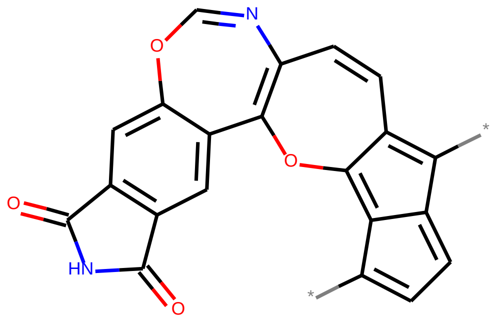


661.90 (641.99) K

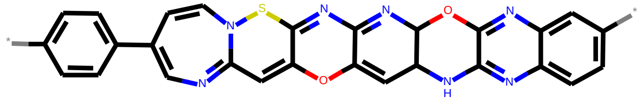


649.04 (632.67) K

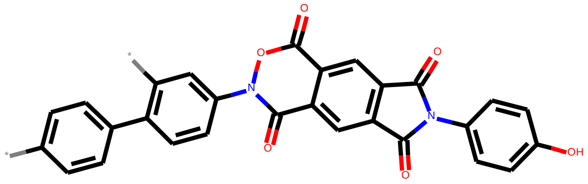
Inverse design by GAN



737.76 (673.45) K

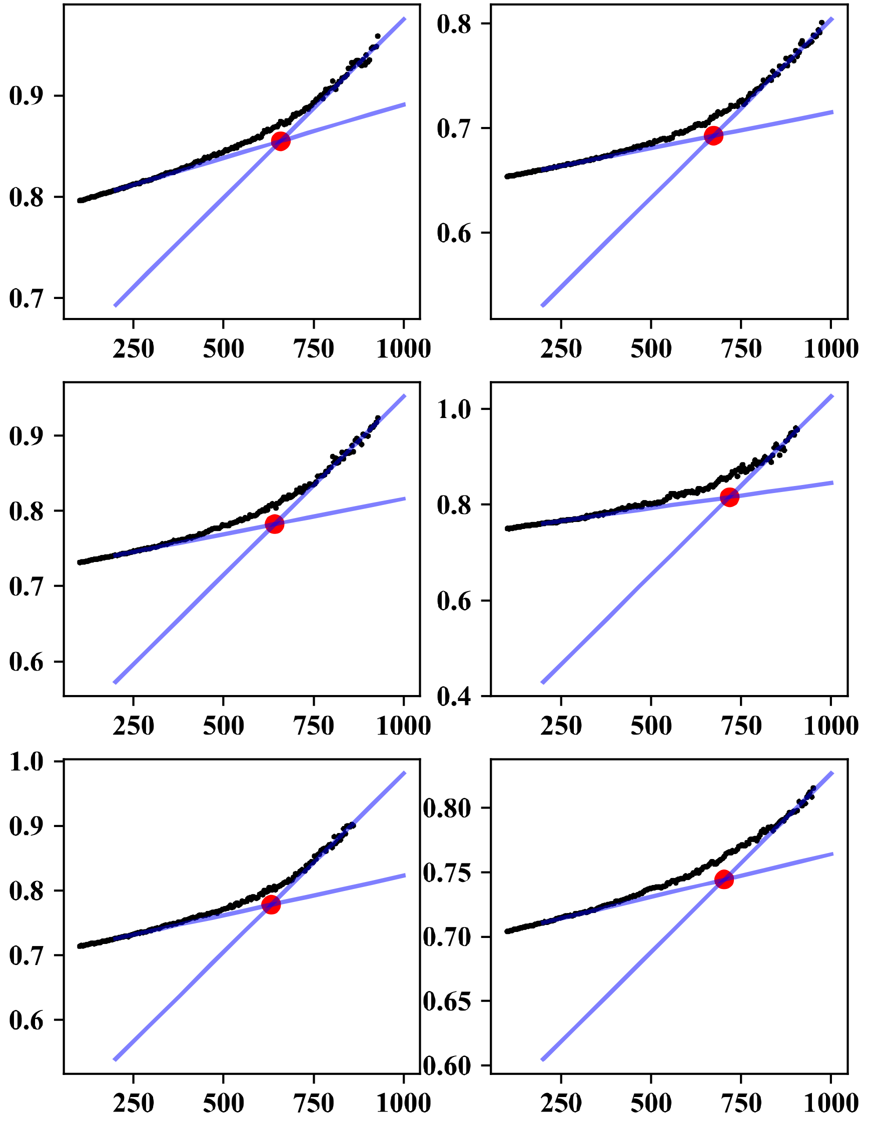


708.59 (717.64) K

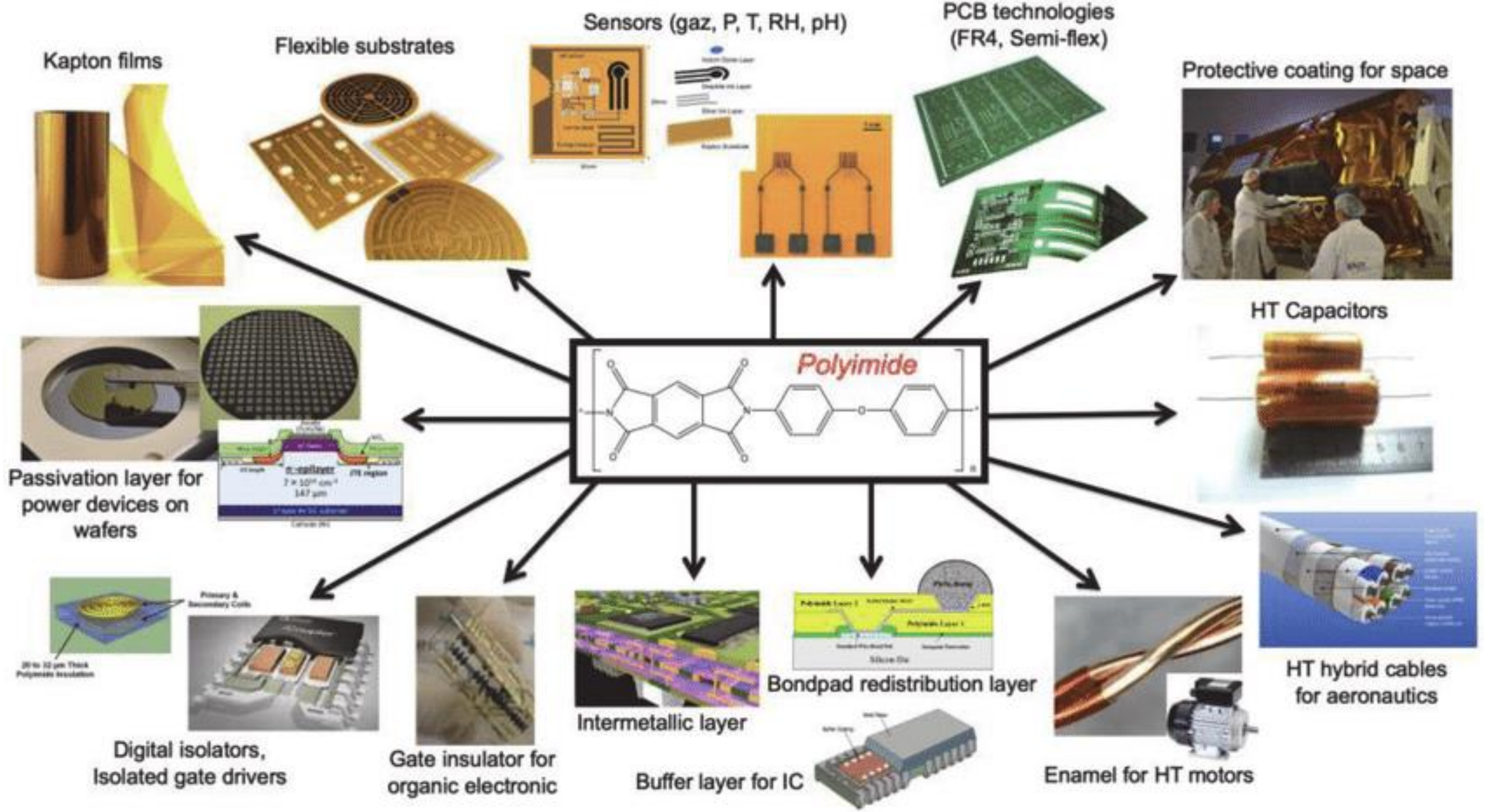


707.83 (702.67) K

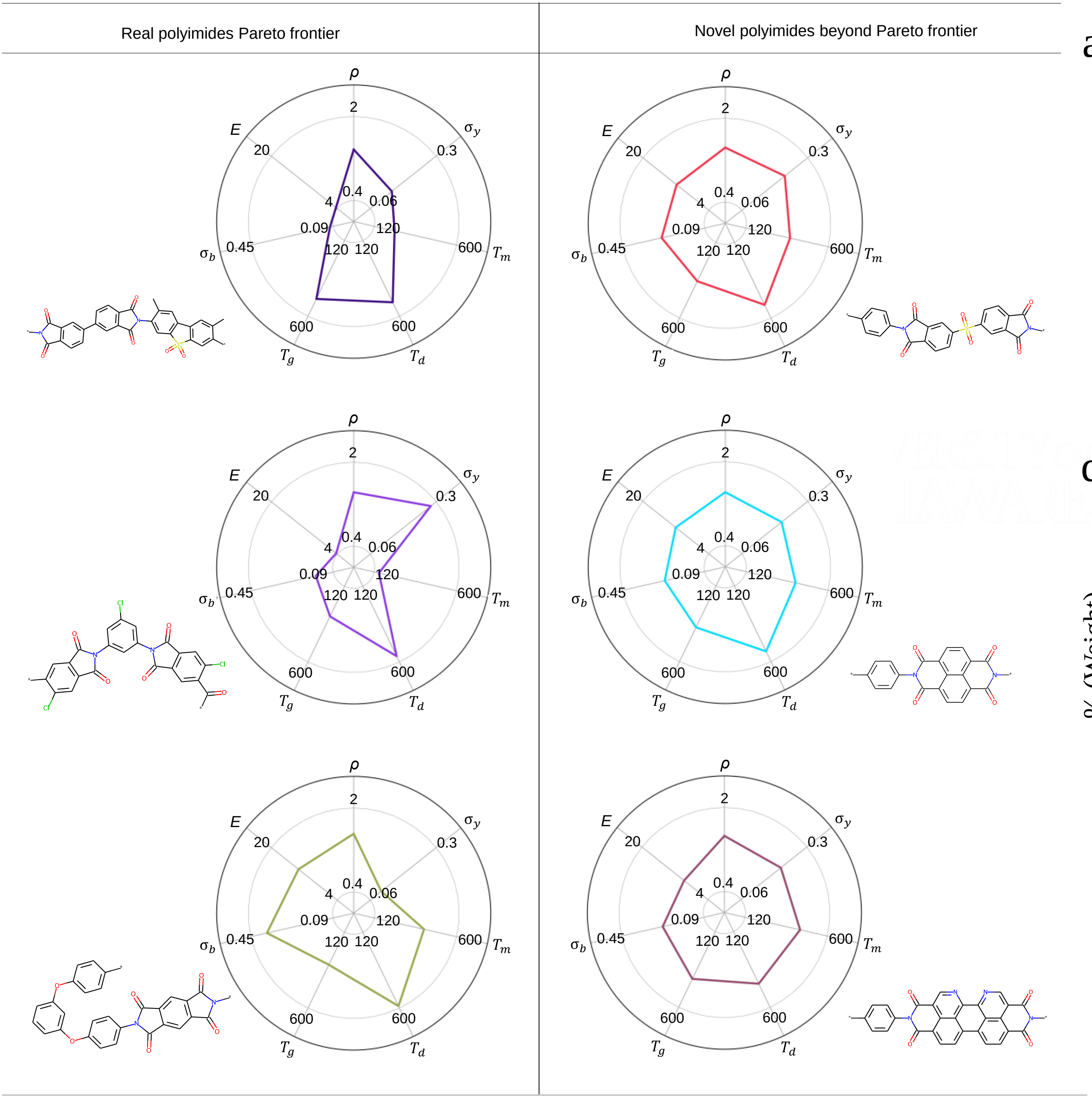
MD simulations

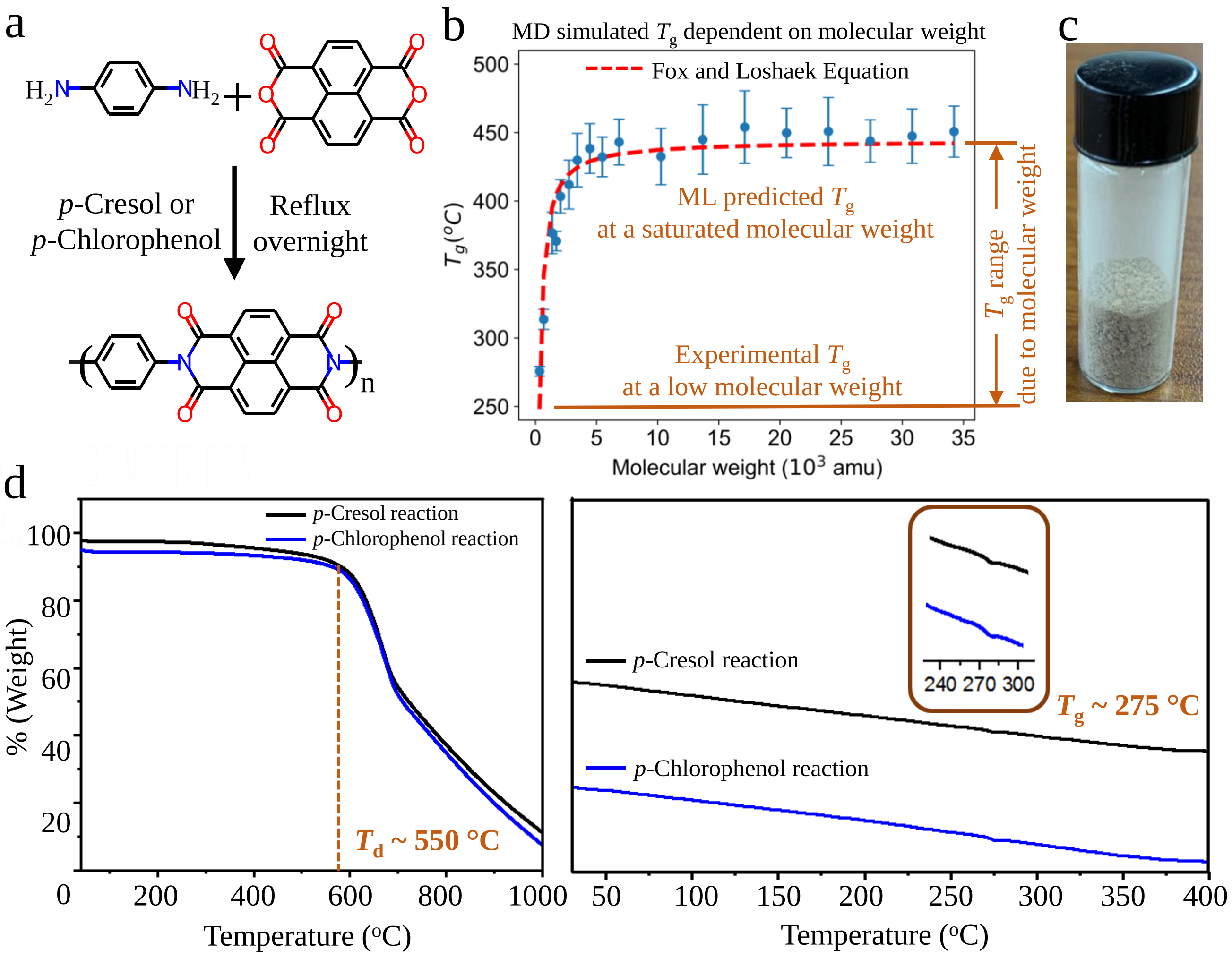


Application Example: Polyimide, "Golden Plastic"



ML-assisted discovery of novel polyimides





"Discovery of Multi-Functional Polyimides through High-Throughput Screening using Explainable Machine Learning" Lei Tao, Jinlong He, Nuwayo Eric Munyaneza, Vikas Varshney, Wei Chen, Guoliang Liu, Ying Li, Chemical Engineering Journal, 2023, 465, 142949.

Polyimide Explorer

polyimide.herokuapp.com

FlashBack Express 5 Recorder

Polyimide Explorer

This tool explores 8 million hypothetical polyimides for three thermal/mechanical properties including Young's Modulus E, Tensile Yield Strength σy and Glass Transition Temperature Tg. All hypothetical are obtained via the computational polycondensation of compounds that are commercially available. The best-performing 77,000 hypothetical polyimide are included for visualization. More details/data can be found in Tao, Lei, Jinlong He, Vikas Varshney, Wei Chen, and Ying Li. 'Machine Learning Discovery of Multi-Functional Polyimides'

Operation 1. ***Hover*** over a polyimide in the 3D plot to see its structure, and imide groups contained are highlighted

Operation 2. ***Click*** a polyimide in the graph to see the info of its structure/property and reacting compounds at the bottom of the page.

Operation 3. ***Select*** the interested ranges of the three properties to filter desired hypothetical polyimides.

Operation 4. ***Predict*** the properties of polyimides based on a SMILES input.

Range of E: [0.00, 18.00] (GPa)

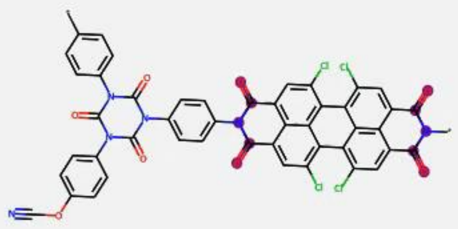
0 6 12 18

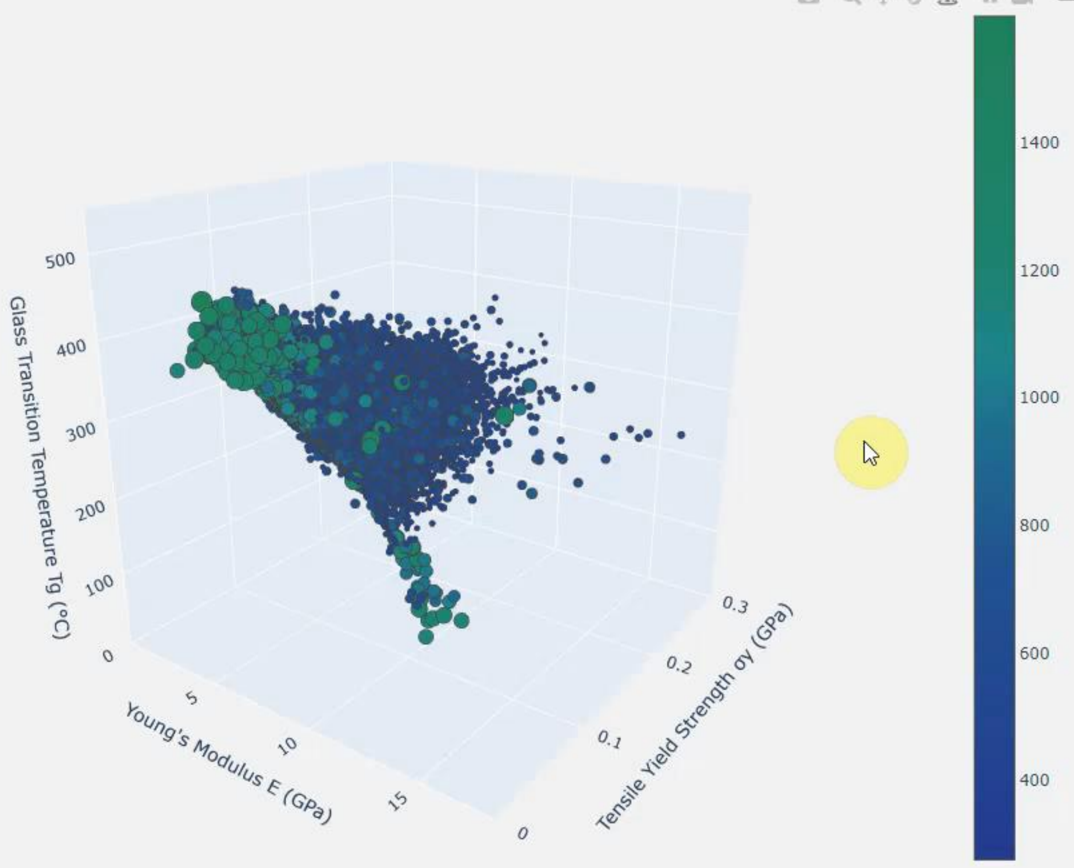
Range of σy: [0.00, 0.30] (GPa)

0 0.1 0.2 0.3

Range of Tg: [0.00, 550.00] (°C)

0 100 200 300 400 500 550





Input box for SMILES

PREDICT

Enter a polyimide SMILES and press predict for property prediction

6:20 PM
7/29/2022

Take-Home Message

- ML is a powerful method for the prediction and rapid screening of innovative polymers, particularly with growing large sets of experimental and computational data for polymeric materials.
- By establishing an *inverse* mapping from property to polymer's synthesis (polyGAN), we can overcome the limitations of the property-prediction (or *forward* problem-based) approaches that screen polymers from a predetermined dataset and suffer from selection bias.
- My personal suggestion: stop with trial-and-error (Edisonian) and embrace

Machine learning & Optimization (AlphaGo, AlphaGo Zero, AlphaFold, AlphaCode, AlphaTensor, AlphaGeometry, …)

All of us can be Iron Man in the future with our J.A.R.V.I.S. (ML models).

This presentation does not contain any proprietary, confidential, or otherwise restricted information

Thank You !