

SMILES-X: A tailored and effective molecular property inference and generation pipeline

LAMBARD Guillaume (ラムバール ギヨム)

LAMBARD.Guillaume@nims.go.jp

National Institute for Materials Science (NIMS)

Centre for Basic Research on Materials (CBRM)

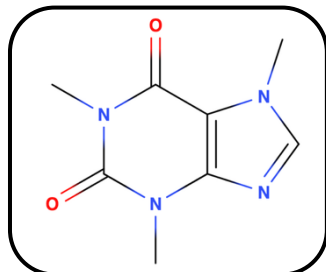
Data-driven Materials Design Group

SMILES-X context

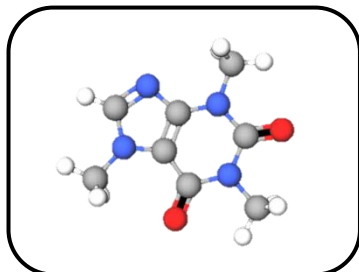
Representations

C[N]1C=NC2=C1C(=O)N(C)C(=O)N2C

1D SMILES



2D graph



3D mol

Target
Property

Features

1D, 2D, 3D fingerprints

and/or

1D, 2D, 3D physical descriptors

A.I. models

Linear regressions

Decision trees

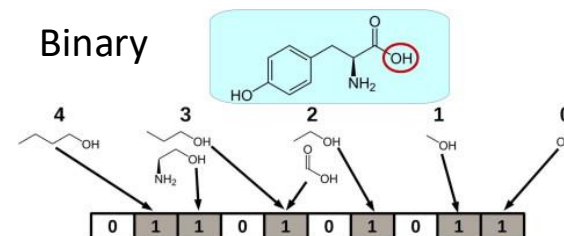
Neural networks

Inconveniences

- Small datasets (<<10,000 samples)
- Multiple representations

- Domain(task)-specific features
- Design new features? Yes, but hard
- Time-consuming to compute

Binary



ECFP

<https://doi.org/10.1016/j.ymeth.2014.08.005>

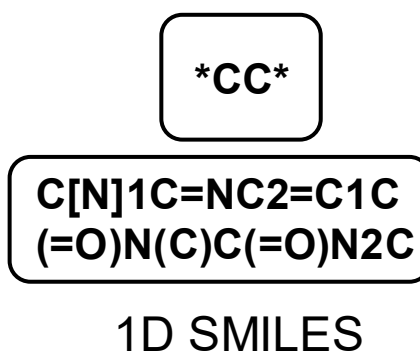
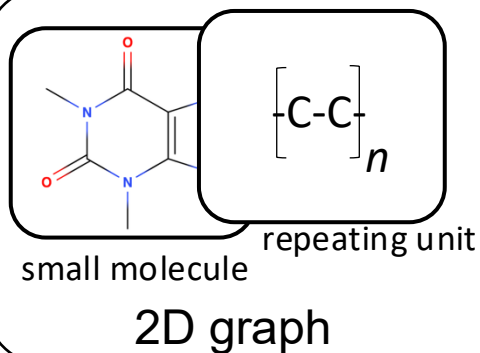
Physical descriptors: [Valence electrons, charge, molecular weight, number of aromatic rings, etc.]

- Try them all
Representations x Features space x Models
- No/little interpretation of outcomes
- Prediction uncertainties assessment

SMILES-X: Efficient physicochemical properties prediction for small molecules and homopolymers

From small (< 100) to big (>> 10,000 samples) datasets

Molecular representation



Targeted Property

Automated molecular description (i.e. featurization)

Automated “SMILES-to-Property” inference model design

“SMILES-to-property” high accuracy prediction + uncertainty

+

“SMILES-to-property” interpretation

SMILES-X

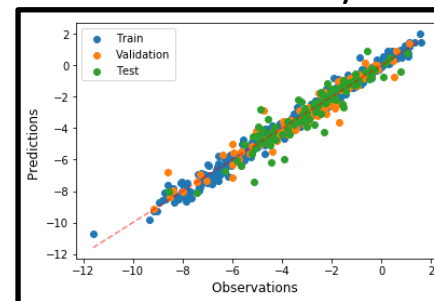
G. Lambard et al., Mach. Learn.: Sci. Technol., 1(2), 025004 (2020)

<https://github.com/Lambard-ML-Team/SMILES-X>

SMILES-X 2.x coming...

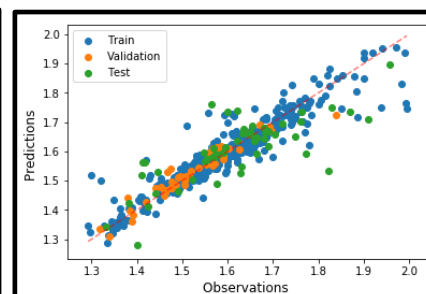
Properties Prediction

Water solubility



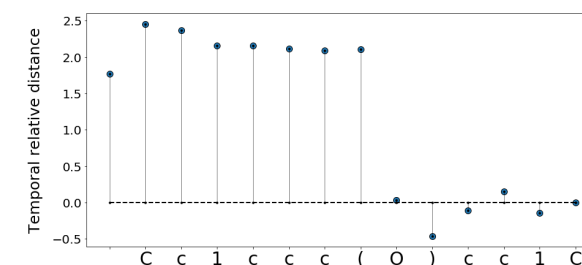
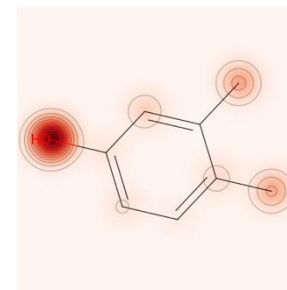
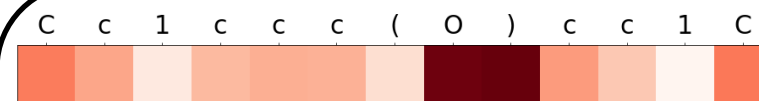
small molecules

Refractive index



homopolymers

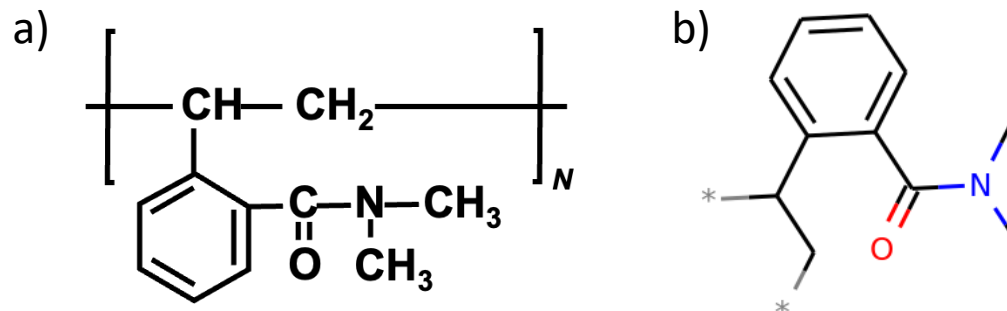
Interpretation



IUPAC name

poly[2-(dimethylcarbamoyl)styrene]

2D graphs



Successful depiction
of homopolymers
with SMILES

Canonical SMILES

CC()c1ccccc1C(=O)N(C)C

Non-canonical SMILES

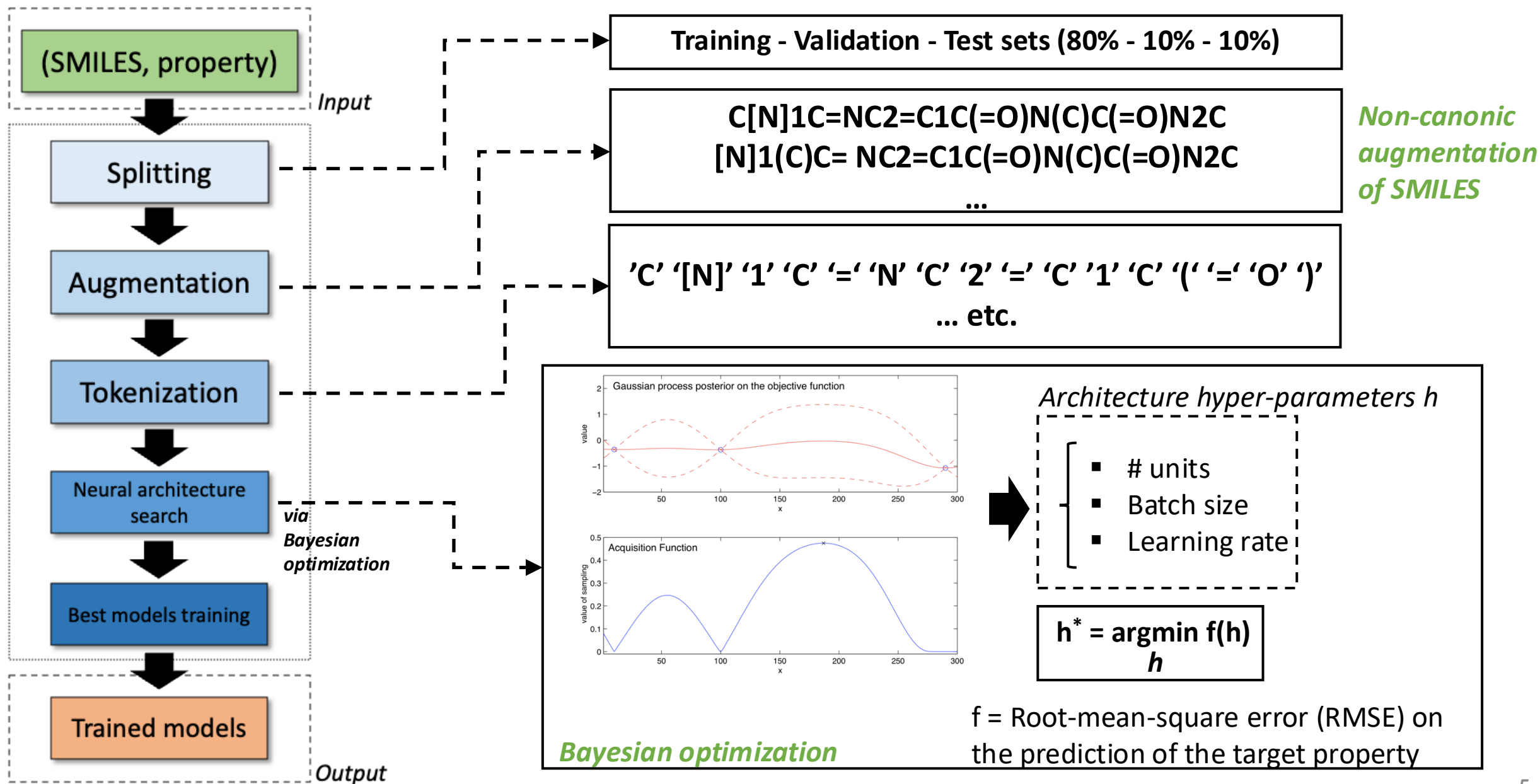
$\left[\begin{array}{l} C(C(*)c1ccccc1C(=O)N(C)C)* \\ C(*)c1ccccc1C(=O)N(C)C* \\ *C(c1ccccc1C(=O)N(C)C)* \\ c1(C(C*)*)ccccc1C(=O)N(C)C \\ c1cccc(C(=O)N(C)C)c1C(C*)* \\ c1ccc(C(=O)N(C)C)c(C(C*)*)c1 \\ c1cc(C(=O)N(C)C)c(C(C*)*)cc1 \\ c1c(C(=O)N(C)C)c(C(C*)*)ccc1 \\ c1(C(=O)N(C)C)c(C(C*)*)cccc1 \\ C(=O)(N(C)C)c1c(C(C*)*)cccc1 \\ O=C(N(C)C)c1c(C(C*)*)cccc1 \\ N(C)(C)c1c(C(C*)*)cccc1=O \\ CN(C)c1c(C(C*)*)cccc1=O \\ CN(C(c1c(C(C*)*)cccc1)=O)C \end{array} \right]$

Use by the SMILES-X software as
augmentation of canonical SMILES

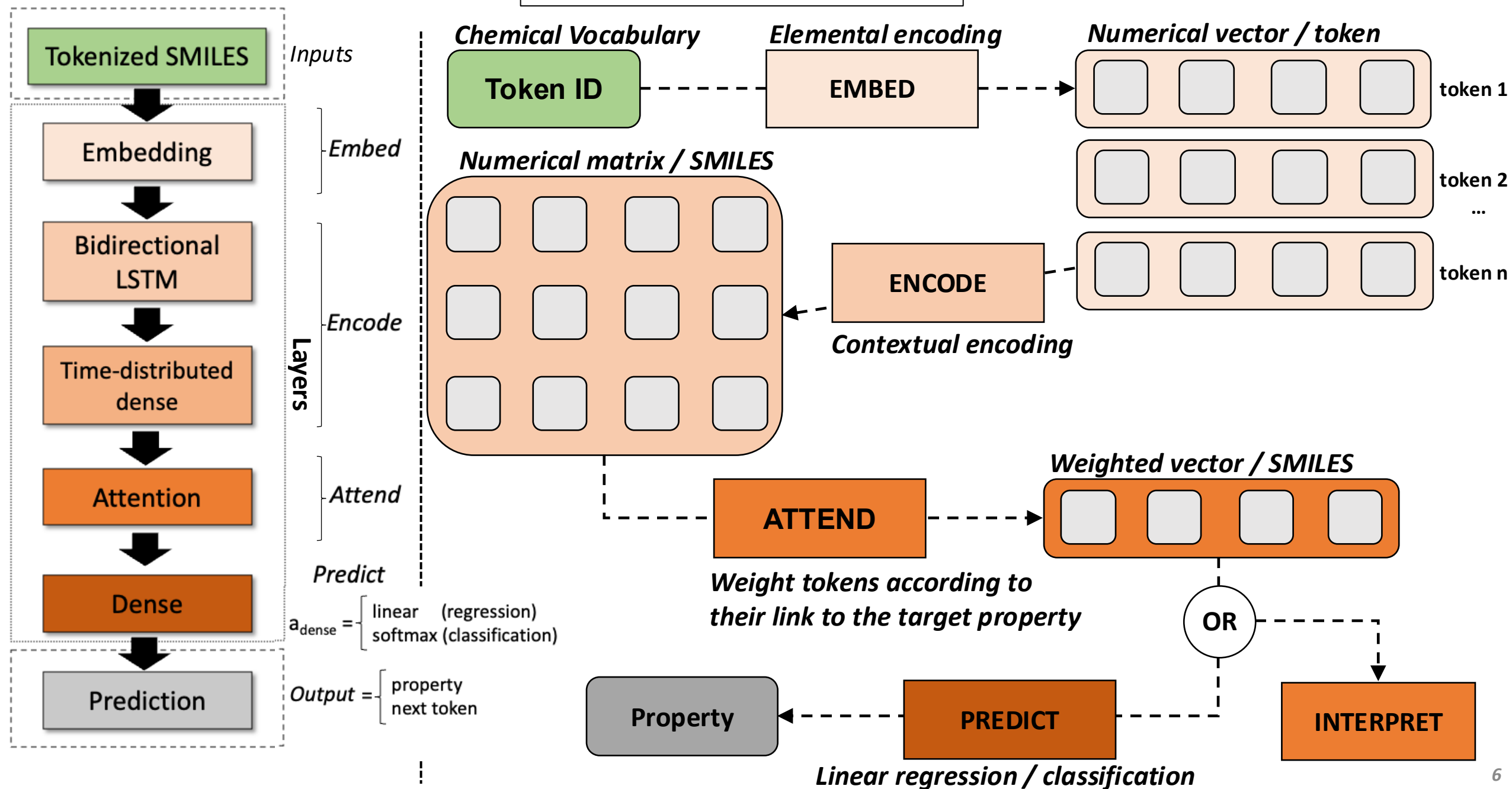
→ Common methodology for improving
convergence and performance of deep
learning models

SMILES-X pipeline

Pipeline



SMILES-X architecture

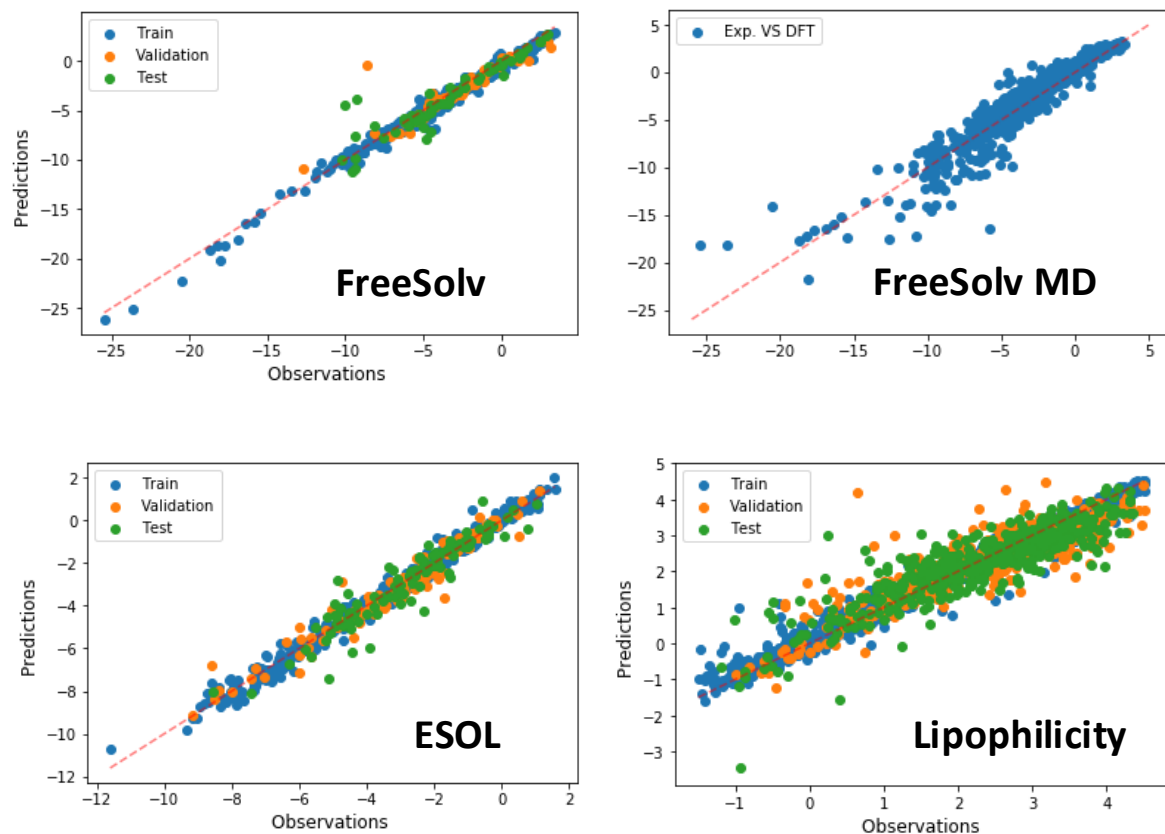


Some results

Physical Chemistry Datasets from <http://moleculenet.ai/datasets-1>

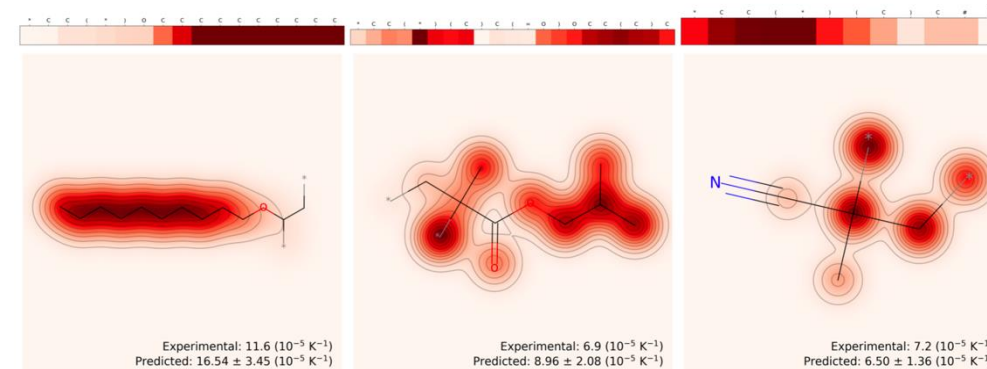
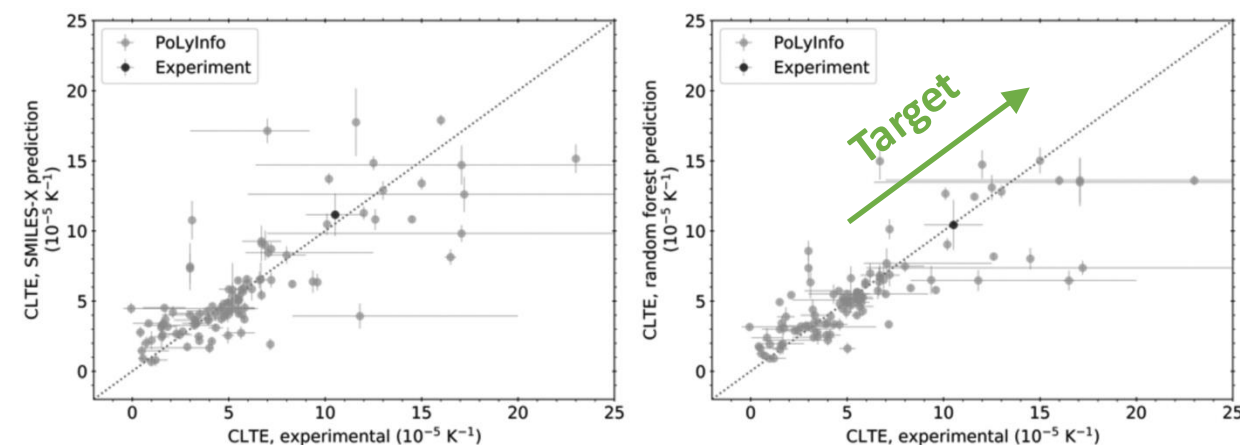
- * **ESOL**: Water solubility experimental data for common organic small molecules (log data in mols/litre) - **#compounds: 1128**
- * **FreeSolv**: Hydration free energy experimental and computational data for small molecules in water (kcal/mol) - **#compounds: 642**
- * **Lipophilicity**: Octanol/water distribution coefficient (logD at pH 7.4) experimental data - **#compounds: 4200**

G. Lambard et al., *Mach. Learn.: Sci. Technol.*, 1(2), 025004 (2020)



* **Prediction of the coefficient of linear thermal expansion (CLTE) for amorphous homopolymers (10^{-5} K^{-1}) - #compounds: 106**

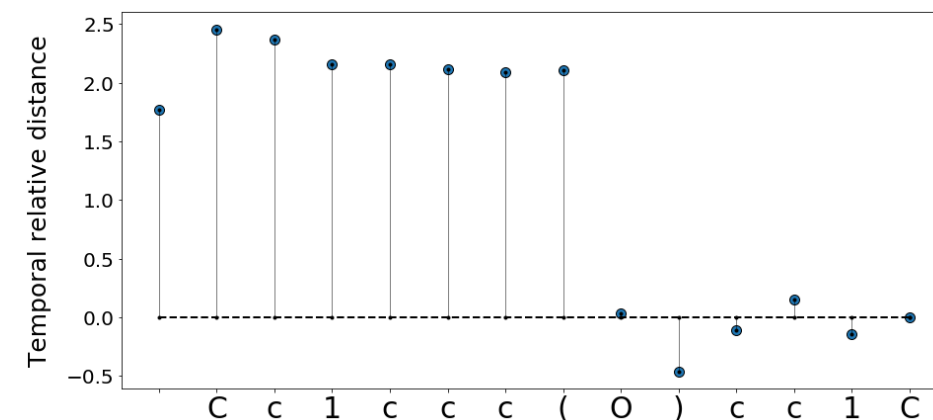
E. Gracheva, G. Lambard, S. Samitsu, K. Sodeyama, A. Nakata, *STAM: Methods*, 1:1, 213-224 (2021)



- Exploring the chemical space is hard
 - >> **10^{60} possibilities**
 - for small molecules, or homopolymers repeated unit
 - More for copolymers (x structural dependency)
 - More for polymer blends (x mixing ratios)
- Combinatorial puzzle with limited hardware/software, time, cost
 - **Can't rely on experiments alone**
 - **Can't rely on computational chemistry alone** (e.g. DFT, TD-DFT, MD, etc.)
- **Materials \rightarrow properties**: likelihood $p(o|s)$ is estimated
 - But limited to joint space of states s (tokens) and observables o (properties) presently known
- **Properties \rightarrow Materials**: posterior $p(s|o)$ can be estimated through **Bayesian inversion** of the likelihood $p(o|s)$
 - **No** need of Generative Adversarial Networks (**GANs**), or reinforcement learning (**RL**) here
 - **Bayesian principle**: $p(s|o) \propto p(o|s) \cdot p(s)$ (neglecting evidence, $p(o) = \sum_s p(o|s) \cdot p(s)$), with $p(s)$ the prior over states s
 - A molecular structure $p(s) = p(s_0) \cdot \prod_t p(s_t | s_{t-1}, \dots, s_0)$



<https://doi.org/10.1038/432823a>



SMILES-X: AI-assisted generation of small molecules or homopolymers

Tokens = any SMILES characters in a given dataset e.g., PolyInfo

At time t

- S_t : structure to be extended
- s_t : possible token to be added
- \tilde{o}_t : distance to target observation

C()C(=O)

Structure S_t

+

'C'
'('
)'
'F'
'O'
.
.
.
.
.
.
.
[Si]

Vector of possible
next tokens

Prior

$$p(s_{t+1})$$

SMILES-X on tokens
enumeration

X

Bayesian inversion

Likelihood

$$p(\tilde{o}_{t+1} | s_{t+1})$$

SMILES-X on %Biodeg.
predictions

~

Posterior

$$p(s_{t+1} | \tilde{o}_{t+1})$$

Most likely
token s_{t+1} to
be added

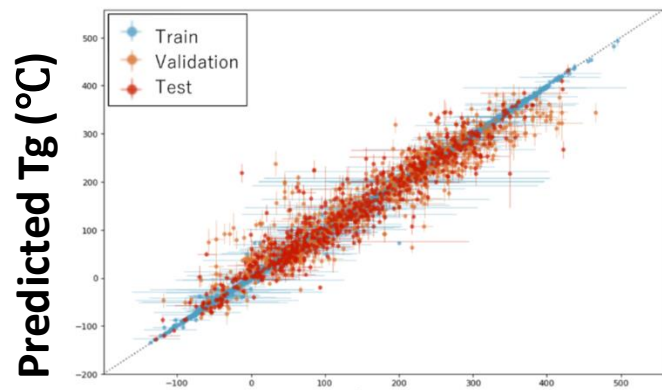
C()C(=O)O

Best structure S_{t+1} at time t+1 closest
to the target

If not, continue extending S_{t+1} until the target is reached

AI-assisted generation of high Tg homopolymers

SMILES-X training on PoLyInfo data

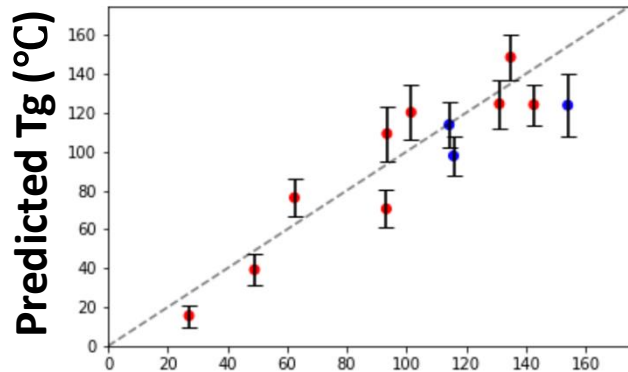


MAE ~ 20.9 °C
RMSE ~ 32 °C
 $R^2 \sim 0.91$

Observed Tg (°C)

Even though PoLyInfo data comes from various sources, SMILES-X performs very well on unseen lab data

SMILES-X testing on lab data

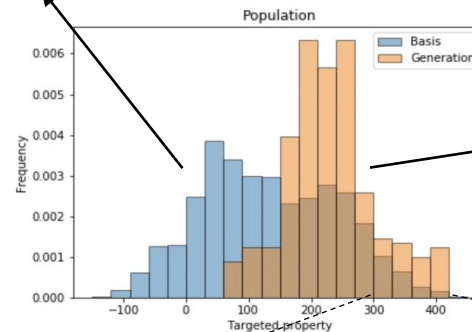


● Found in training set
● Seen for the first time by SMILES-X

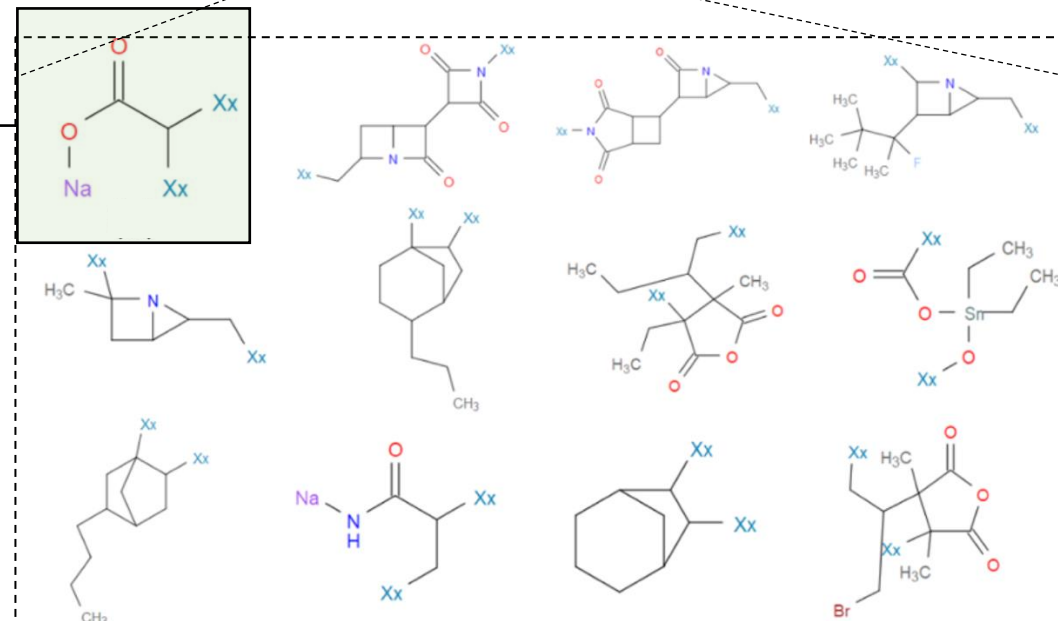
Observed Tg (°C)

PoLyInfo data distribution

SMILES-X + SMILES-Neo generation



Generated data distribution



Polymethylene with predicted Tg = 340.2 ± 115.2 °C and Tg > 300 °C in the laboratory

Material Synthesis 2025 (MatSyn25) Data Set for 2D Materials

Chengbo Li, Ying Wang, Qianying Wang, Zhizhi Tan*, Haiqing Jia, Yi Liu, Li Qian*, Nian Ran*, Jianjun Liu*, and Zhixiong Zhang



Loading Institutional Login
Options...

Other Access Options

Supporting Information (1)

Abstract

Two-dimensional (2D) materials have shown broad application prospects in fields such as energy, environment, and aerospace owing to their unique electrical, mechanical, thermal, and other properties. With the development of artificial intelligence (AI), the discovery and design of novel 2D materials have been significantly accelerated. However, due to the lack of basic theories of material synthesis, identifying reliable synthesis processes for theoretically designed materials is a challenge. The emergence of large language model offers approaches for the reliability prediction of material synthesis processes. However, its development is limited by the lack of publicly available data sets of material synthesis processes. To address this, we present the Material Synthesis 2025 (MatSyn25), a large-scale open data set of 2D material synthesis processes. MatSyn25 contains 163,240 pieces of synthesis process information extracted from 85,160 high-quality research articles, each including basic material information and detailed synthesis process steps. Based on MatSyn25, we developed MatSyn AI, which specializes in material synthesis, and provided an interactive web platform that enables multifaceted exploration of the data set (<https://matsynai.stpaper.cn/>). MatSyn25 is publicly available, allowing the research community to build upon our work and further advance AI-assisted materials science.

