globus labs

The Department of Computer Science

THE UNIVERSITY OF CHICAGO

Hybrid Human-Machine Information Extraction Decision Engine

Roselyne Tchoua[1,2], Kyle Chard[2], Debra Audus[3], Logan Ward[2], Juan de Pablo[4], Ian Foster[1,2]
[1]Department of Computer Science, [2]The Computation Institute, [3]The National Institute of Standards and Technology
[4]The Institute for Molecular Engineering, {University of Chicago and Argonne National Laboratory}

DESIGN GOALS

- *Create a human and machine accessible database of materials properties extracted from journal publications*
- *Combine automatic software and crowdsourcing to optimize the efficiency and quality of the extracted information*
- *Learn from the collaboration between software and crowds to improve next generation software*

Background:

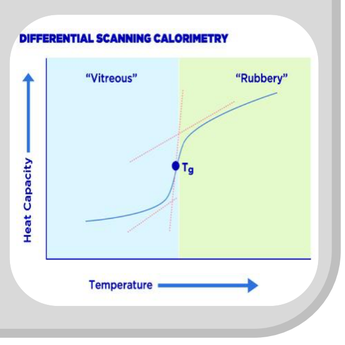
- 28,000 Journals, 1.8 million publications per year
- Thousands of properties locked within publications
- Properties are unstructured and not machine accessible

Flory-Huggins Interaction (χ) Parameter

- Measures the interaction of polymer chains with solvent molecules and polymer-polymer interactions

Glass Transition Temperature (T_g)

- T_g is the temperature region where the polymer transitions from a hard, glassy material to a soft, rubbery material

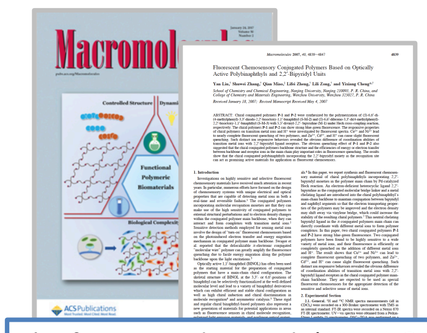


ChemDataExtractor (CDE)*

- Parses and extracts information from journal articles and patents
- Resolves chemical names, abbreviations and identifiers
- Extracts experimental properties and spectroscopy
  - **We added new rule to extract T_g**

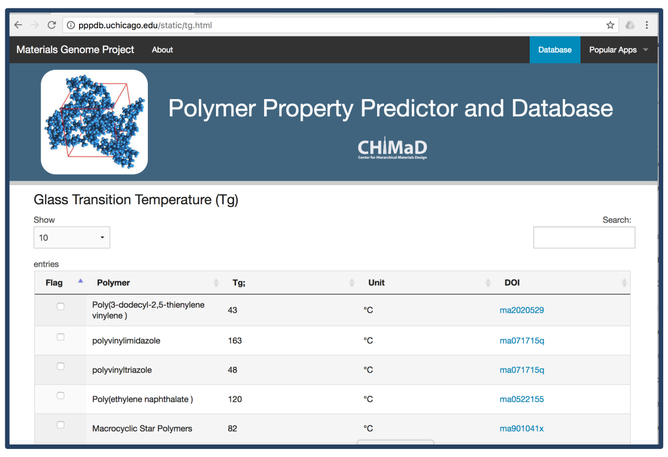
* Ref: Swain, Cole. J Chem. Inf. Model. (2016) 1894

Goal: Create a Database of Polymer Properties



The four-armed compd. (ANTH-OXA6t-OC12) with the dodecyloxy surface group is a high glass transition temp. (Tg: 211°) material and exhibits good solubility.

ANTH-OXA6t-OC12 —has_tg_of→ 211°



p³DB: http://pppdb.uchicago.edu/

Results are neither perfect nor complete…

Solution: build a pipeline of automatic information extraction using a combination of rule-based, machine learning and crowdsourcing approaches.

(1) Extract polymers: run CDE on thousands of publications

Categories of results

- T_g values for identified polymers (21%) ✓
- T_g values for unresolved identifiers (~3%) ✗
- T_g values for non-polymers (28%) ✗
- Tg values for missing compounds (30%) ✗
- Multiple Tg values for compounds (18%) ✗

(2) Classify polymers: apply heuristics to find synonyms/acronyms

…	
PS	25
Polystyrene	25
poly(styrene)	25
PMMA	26
….	

9555 compounds – 12257 synonyms

(3) Learn acronyms: learn to associate labels & author-defined acronyms

… glass transition temperature (T_g) of PVDF is not observed, though pure PVDF exhibit T_g at −46 °C as evident from loss modulus peak.

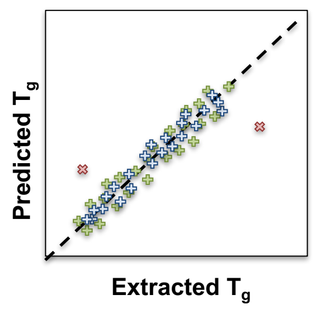
PVDF → poly(vinylidene fluoride)

(4) Proximity search: link properties to compounds

The Tg of PEO peaks at a molecular weight of 6000 (Tg = −17 °C)…

Resolves ~50% of missing compounds

(5) Validate values: apply machine learning models



Machine | Humans

Domain Experts

Trained crowds

Untrained crowds

RESULTS

Potential T_g data extracted:

- 1442 glass transition records (single and multiple temperatures)
- 1884 glass transition temperatures for more 800 compounds

Processed T_g data :

- 6090 documents processed
- 300 T_gs extracted
- Dictionary of ~ 9000 polymers
- ~300 resolved polymer-T_g pairs