Incorporating InChI into a polymeric database

Debra J. Audus
State and Future of the IUPAC InChI
August 16, 2017
Acknowledgements

Roselyne Tchoua
Kyle Chard
Logan Ward
Ian Foster

Jian Qin
Joshua Lequieu
Juan de Pablo
Need for polymeric databases

Materials science with large-scale data and informatics: Unlocking new opportunities
Joanne Hill, Gregory Mulholland, Kristin Persson, Ram Seshadri, Chris Wolverton, and Bryce Mendig

Universal access to abundant scientific data could fundamentally transform the field of materials science by overcoming serious challenges to bringing about broad dissemination of research results within materials research communities. Nonetheless, this promise is constrained by a lack of effective database management and methods to share, compute, and integrate the large-scale materials data and informatics technologies.

Introduction

Data-intensive science has been described as the “fourth paradigm” for scientific exploration, with the first three being experiments, theory, and simulation. While the value of data-intensive research approaches are becoming more apparent, the field of materials science has not yet experienced the same widespread adoption of these methods (as has occurred in biosciences, astronomy, and particle physics). Nonetheless, the potential impact of data-driven materials science is tremendous. Materials informatics could reduce the typical 10-20 year development and commercialization cycle for new materials. We see plentiful opportunities to use data and data science to radically reduce this timeline and generally advance materials research and development (R&D) and manufacturing.

In this article, we discuss the current state of affairs with respect to data and data analytics in the materials community, with a particular emphasis on the potential of data-driven materials science.

The MATERIAL CODE

Machine-learning techniques could revolutionize how materials science is done.

BY NICOLA HOSEIN

© 2016 Materials Research Society
Existing resources

Paper-based

Web-based

limited accessibility of entire database and/or datasets that are too small
Polymer Property Predictor and Database

http://pppdb.uchicago.edu

The Center for Hierarchical Materials Design (CHIMAD) represents a Chicago-based consortium of the University of Chicago, Northwestern University, Northwestern-Argonne Institute for Science and Engineering (NASIE) that is a partnership between Argonne National Laboratory and Northwestern, and the Computational Institute that is a partnership between the University of Chicago and Argonne. It serves together with NIST as a national resource for the verified codes and curated databases that will enable proliferation of a materials-by-design strategy throughout US industrial partners. Numerous materials “use cases” of industrial relevance drive purposeful method and tool development, while aiding transfer to industry of both the new principles of computational materials design. Demonstrating a broad methodology for multicomponent, multiphase materials spanning metals and polymers for structural and multifunctional applications.

Search the Database

Polymer Name

Search

Browse the Database

Browse χ values
Browse Tg values
Flory Huggins $\chi$ parameter

Publications

376 articles from *Macromolecules*


Flory Huggins $\chi$ parameter

Automatically extract metadata (title, author, etc.)

Flory Huggins $\chi$ parameter

Undergrads review papers and enter $\chi$ into an online form

## Need for a polymer dictionary

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>poly(ethylene-alt-propylene)</td>
<td>polymer</td>
<td>PEP</td>
</tr>
<tr>
<td>protonated poly(ethylene-alt-propylene)</td>
<td>polymer</td>
<td>pPEP</td>
</tr>
<tr>
<td>Polybutadiene</td>
<td>polymer</td>
<td>PB</td>
</tr>
<tr>
<td>polybutadiene</td>
<td>polymer</td>
<td>PBD</td>
</tr>
<tr>
<td>poly(butyl methacrylate)</td>
<td>polymer</td>
<td>PbMA</td>
</tr>
<tr>
<td>Poly(n-butyl methacrylate)</td>
<td>polymer</td>
<td>PnBMA-115</td>
</tr>
<tr>
<td>poly(methacrylic acid)-b-poly(methyl methacrylate) (A)</td>
<td>polymer</td>
<td>PMAA-PMMA (A)</td>
</tr>
<tr>
<td>poly(methacrylic acid)-b-poly(methyl methacrylate) (C)</td>
<td>polymer</td>
<td>PMAA-PMMA (C)</td>
</tr>
<tr>
<td>styrene</td>
<td>polymer</td>
<td></td>
</tr>
</tbody>
</table>

- **prefixes**
- **capitalization**
- **ambiguous**
- **input errors**
The need for InChI

Multiple and trade names

- Registry Number: 9003-53-6
- Molecular Formula: (C8H8)x
- Chemical Name: Benzene, ethenyl-, homopolymer

Identify synonyms

poly(2,6-dimethyl-1,4-phenylene oxide)

poly(xylenyl ether)

1800+ for polystyrene

Broadness of CAS

- Registry Number: 9004-34-6
  - Molecular Formula: W99
  - Chemical Name: Cellulose

- Registry Number: 9000-11-7
  - Molecular Formula: C2H4C3.xUnspecified
  - Chemical Name: Cellulose, carboxymethyl ether

- Registry Number: 9004-32-4
  - Molecular Formula: C2H4C3.xNa.xW99
  - Chemical Name: Cellulose, carboxymethyl ether, sodium salt
# Need for a polymer dictionary

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>poly(ethylene-alt-propylene)</td>
<td>polymer</td>
<td>PEP</td>
</tr>
<tr>
<td>protonated poly(ethylene-alt-propylene)</td>
<td>polymer</td>
<td>pPEP</td>
</tr>
<tr>
<td>Polybutadiene</td>
<td>polymer</td>
<td>PB</td>
</tr>
<tr>
<td>polybutadiene</td>
<td>polymer</td>
<td>PBD</td>
</tr>
<tr>
<td>poly(butyl methacrylate)</td>
<td>polymer</td>
<td>PbMA</td>
</tr>
<tr>
<td>Poly(n-butyl methacrylate)</td>
<td>polymer</td>
<td>PnBMA-115</td>
</tr>
<tr>
<td>poly(methacrylic acid)-b-poly(methyl methacrylate) (A)</td>
<td>polymer</td>
<td>PMAA-PMMA (A)</td>
</tr>
<tr>
<td>poly(methacrylic acid)-b-poly(methyl methacrylate) (C)</td>
<td>polymer</td>
<td>PMAA-PMMA (C)</td>
</tr>
<tr>
<td>styrene</td>
<td>polymer</td>
<td></td>
</tr>
</tbody>
</table>

- **prefixes**
- **capitalization**
- **ambiguous**
- **input errors**

Need something like PubChem for polymers

![PubChem](https://example.com/pubchem.png)
Flory Huggins $\chi$ parameter

1. Information Extraction Module
2. Data Parsing Crowdsourcing Module
3. Curated Polymer Dictionary Module

Developing the polymer dictionary

Name
poly(2-vinylpyridine)

Abbreviation
P2VP

Structure (saved as .mol file)

InChIKey and InChI
KGIGUEBEKRSWTEW-BBVYVPKKBAN

1B/C7H7N/c1-2-7-5-3-4-6-8-7/h2-6H,1H2/z101-1-8(1.2)
## The polymer dictionary

<table>
<thead>
<tr>
<th>Name</th>
<th>Abbreviation</th>
<th>InChI</th>
<th>InChI Key</th>
<th>Thesaurus Names</th>
<th>Thesaurus Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>poly(2-vinylpyridine)</td>
<td>PVCVP</td>
<td>18/C7H7N/c1-2-7-5-3-4-6-1-2/h2-6H,1H2/t1H2-1-17-18-5-3</td>
<td>KG1GJ6BEBKXSTEN-BBVYYPKXBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(2,2-bis(trifluoromethyl)-4,5-difluoro-1,3-dioxole)</td>
<td>PVCFBO2</td>
<td>18/C5F8O2/c6-1-2(7)15-3(14-1,4(8,9)10)11-12</td>
<td>YSYR6SKCBP96-XLALVCSBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(2,6-dimethyl-1,4-phenylene oxide)</td>
<td>PXE</td>
<td>18/C8H10O/c-1-8-3-7-6-2(8)9-7-3H2,1H2/t1H2-2-1-3/2-1-3</td>
<td>GVLQZVRE1MQB9-NW4WQDBA-N</td>
<td>poly(xylene ether)</td>
<td></td>
</tr>
<tr>
<td>poly(3-(2-ethylhexyloxy)thiophene)</td>
<td>P3HT</td>
<td>18/C18H15OSS/c1-3-5-6-9(4-2)-7-8-9-11-12(10)13-UCZKVR2JTHAH-HZUOHMBBA-N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(3-butylnaphthylene)</td>
<td>P3BN</td>
<td>18/C8H14/c-1-8-3-6-7-8-9-10-11-12-13-OMPLFPCOMNIQ-DRAQJUJBA-N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(3-dodecylthiophene)</td>
<td>P3DT</td>
<td>18/C16H26S/c6(2)-3-5-7-8-9-10-12-14-13-OMPLFPCOMNIQ-DRAQJUJBA-N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(3-hexylthiophene)</td>
<td>P3HT</td>
<td>18/C18H15/c1-3-4-5-6-7-8-9-10(8)11-9-7H2,1-2/WACCBCCHELGMWS-RHPJ3LWABA-N</td>
<td>poly(3-hexylthiophene-2,5-diy)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(3-methylthiophene)</td>
<td>PMT</td>
<td>18/C5H5S/c1-3-4-5-6-7H2,1H2/t1H2-1-16(4)-17(2)-18(3)-19/1-7</td>
<td>IZAAQJHPQLQI-KVYRQBDAFBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(3-octylthiophene)</td>
<td>P3OT</td>
<td>18/C18H17S/c1-3-4-5-6-7-8-9-11-12(10)13-IXMDQQBTSNN8-HZUOHMBBA-N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(4-acetoxyacrylate)</td>
<td>PAA</td>
<td>18/C8H8O4/c1-3-4-5-6-7/h2-6H,1H2/t2H2-1-3-4-5-7-6</td>
<td>JAMKSXLYLVPNC-XEPF3TWABA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(4-hydroxystyrene)</td>
<td>PHS</td>
<td>18/C8H8O2/c1-2-3-5-8-9-6-4-7-2-6H,9H2/t2H2-1-3-5-8-6-4-7-2-9</td>
<td>FUSYQGDSNSXORN-KYRQYF8KBBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(4-tert-butylstyrene)</td>
<td>PCSBS</td>
<td>18/C12H16/c1-5-10-8-6-11(9-7-8-10-12)-2(3)4-9-5/1-6QEDQ9M7LJUIC-KX2QDQPOBA-N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(4-vinylbenzyltrimethylammonium chloride)</td>
<td>PVBTMAc</td>
<td>18/C12H9N.CCl/c1-5-11-6-8-12(9-7-11-16)-13-17-12/1-7/VXN5QRAIZONH1-IPOBBXVBA-N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(4-vinylpyridine)</td>
<td>P4VP</td>
<td>18/C7H7N/c2-7-3-5-6-8-4-7-2-6H,1H2/t1H2-2-3-4-5-6-8-7-1</td>
<td>KFDVPUPPSTJTHAH-HZUOHMBBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(6-methyl-1,2,4-naphthalenetricarboxylate)</td>
<td>PMLC</td>
<td>18/C9H8O4/c1-2-3-5-6-7-8-9(8)-4-10-11(10)12-13(12)</td>
<td>3NCXQG4PQXTK-YXATYHDBBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(acrylic acid)</td>
<td>PAA</td>
<td>18/C3H4O2/c1-2-3(4)h5/h2-3(4)h5,1H2/t1H2-1-3-4-5-6-7-9</td>
<td>NIXONLZIIQK2J-ZJLSCBCBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(2,5-diacylamidophenyl isocyanate)</td>
<td>PDA</td>
<td>18/C8H8N2O2/c1-2-3-5-6-7/h2-6H,1H2/t2H2-1-3-5-6-7</td>
<td>1L3DQQJHOXQ4-RWAKDQJBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(2,5-diacylamidophenyl isocyanate)</td>
<td>PDA</td>
<td>18/C8H8N2O2/c1-2-3-5-6-7/h2-6H,1H2/t2H2-1-3-5-6-7</td>
<td>1L3DQQJHOXQ4-RWAKDQJBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(2,5-diacylamidophenyl isocyanate)</td>
<td>PDA</td>
<td>18/C8H8N2O2/c1-2-3-5-6-7/h2-6H,1H2/t2H2-1-3-5-6-7</td>
<td>1L3DQQJHOXQ4-RWAKDQJBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(2,5-diacylamidophenyl isocyanate)</td>
<td>PDA</td>
<td>18/C8H8N2O2/c1-2-3-5-6-7/h2-6H,1H2/t2H2-1-3-5-6-7</td>
<td>1L3DQQJHOXQ4-RWAKDQJBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(2,5-diacylamidophenyl isocyanate)</td>
<td>PDA</td>
<td>18/C8H8N2O2/c1-2-3-5-6-7/h2-6H,1H2/t2H2-1-3-5-6-7</td>
<td>1L3DQQJHOXQ4-RWAKDQJBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(2,5-diacylamidophenyl isocyanate)</td>
<td>PDA</td>
<td>18/C8H8N2O2/c1-2-3-5-6-7/h2-6H,1H2/t2H2-1-3-5-6-7</td>
<td>1L3DQQJHOXQ4-RWAKDQJBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(2,5-diacylamidophenyl isocyanate)</td>
<td>PDA</td>
<td>18/C8H8N2O2/c1-2-3-5-6-7/h2-6H,1H2/t2H2-1-3-5-6-7</td>
<td>1L3DQQJHOXQ4-RWAKDQJBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(2,5-diacylamidophenyl isocyanate)</td>
<td>PDA</td>
<td>18/C8H8N2O2/c1-2-3-5-6-7/h2-6H,1H2/t2H2-1-3-5-6-7</td>
<td>1L3DQQJHOXQ4-RWAKDQJBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(2,5-diacylamidophenyl isocyanate)</td>
<td>PDA</td>
<td>18/C8H8N2O2/c1-2-3-5-6-7/h2-6H,1H2/t2H2-1-3-5-6-7</td>
<td>1L3DQQJHOXQ4-RWAKDQJBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(2,5-diacylamidophenyl isocyanate)</td>
<td>PDA</td>
<td>18/C8H8N2O2/c1-2-3-5-6-7/h2-6H,1H2/t2H2-1-3-5-6-7</td>
<td>1L3DQQJHOXQ4-RWAKDQJBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(2,5-diacylamidophenyl isocyanate)</td>
<td>PDA</td>
<td>18/C8H8N2O2/c1-2-3-5-6-7/h2-6H,1H2/t2H2-1-3-5-6-7</td>
<td>1L3DQQJHOXQ4-RWAKDQJBA-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poly(2,5-diacylamidophenyl isocyanate)</td>
<td>PDA</td>
<td>18/C8H8N2O2/c1-2-3-5-6-7/h2-6H,1H2/t2H2-1-3-5-6-7</td>
<td>1L3DQQJHOXQ4-RWAKDQJBA-N</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Flory Huggins $\chi$ parameter

1. Publications
2. Information Extraction Module
3. Curated Polymer Dictionary Module
4. Final Expert Review

Proposed $\chi$ Entries
Confirmed $\chi$ Entries

Final review and push to database

Glass transition temperature

Publications

6,090 articles from Macromolecules

Glass transition temperature

Tries to find compound-\(T_g\) pairs automatically

Glass transition temperature

Automatically create a dictionary of polymers (only names) using “P” and “poly”
## NLP Polymer Dictionary

<table>
<thead>
<tr>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polystyrene</td>
</tr>
<tr>
<td>poly(styrene)</td>
</tr>
<tr>
<td>polystyrene</td>
</tr>
<tr>
<td>polystyrenes</td>
</tr>
<tr>
<td>PS</td>
</tr>
<tr>
<td>PSS</td>
</tr>
<tr>
<td>polyimides</td>
</tr>
<tr>
<td>polyolefin</td>
</tr>
<tr>
<td>copolymer 10</td>
</tr>
<tr>
<td>poly(2,4’-BF-a)</td>
</tr>
<tr>
<td>macroporous poly(N-isopropylacrylamide)gel</td>
</tr>
</tbody>
</table>

- various forms
- not plural of PS
- family names
- labels not names
- prefixes/suffixes

12,814 polymers in the dictionary

Work in progress to clean up errors above and adding InChI
Glass transition temperature

Many other steps to final product!

The need for InChI

Multiple and trade names

Registry Number: 9003-53-6
Molecular Formula: (C8H8)x
Chemical Name: Benzene, ethenyl-, homopolymer
Hide Synonyms

- MS 555
- Fostarene 20D9
- JSR-BK 2500
- Styron 680
- BSB-S 40
- Daicel Styrol 20

1800+ for polystyrene

poly(2,6-dimethyl-1,4-phenylene oxide)
poly(xylenyl ether)

Identify synonyms

Broadness of CAS

Registry Number: 9004-34-6
Molecular Formula: W99
Chemical Name: Cellulose
Show Synonyms

Registry Number: 9000-11-7
Molecular Formula: C2H4C3.xUnspecified
Chemical Name: Cellulose, carboxymethyl ether
Show Synonyms

Registry Number: 9004-32-4
Molecular Formula: C2H4C3.xNa.xW99
Chemical Name: Cellulose, carboxymethyl ether, sodium salt
Show Synonyms

Input/output for machine learning

Natural Language Processing Module

1B/C8H8O/c1-5-3-7-46(2)8(5)9-7/h3-4H,12H3/z101-1-9(7,9,8,9)
Limitations of current InChI

Organometallic

Branching / crosslinks

Markush
Conclusions and outlook

http://pppdb.uchicago.edu

\[ \chi^{263} \quad T_g^{258} \]

Future work

• Add .mol files and InChI to pppdb
• Cleaning up NLP polymer dictionary

Advances still need for InChI

• Organometallics
• Branching / cross-links
• Markush
Flory Huggins $\chi$ parameter

1. Publications
2. Information Extraction Module
3. Curated Polymer Dictionary Module
4. Final Expert Review

Proposed $\chi$ Entries
Confirmed $\chi$ Entries

Glass transition temperature

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