

Incorporating InChI into a polymeric database

Debra J. Audus

State and Future of the IUPAC InChI

August 16, 2017



Acknowledgements



Computer
Science



Roselyne Tchoua



Kyle Chard

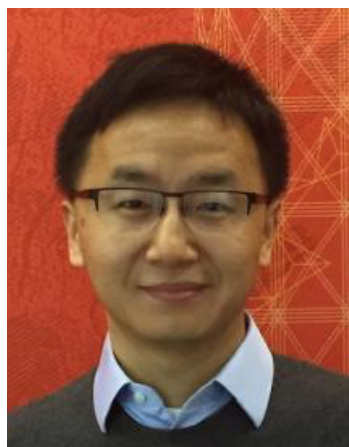


Logan Ward



Ian Foster

Stanford
University
Chemical
Engineering



Jian Qin



Molecular
Engineering

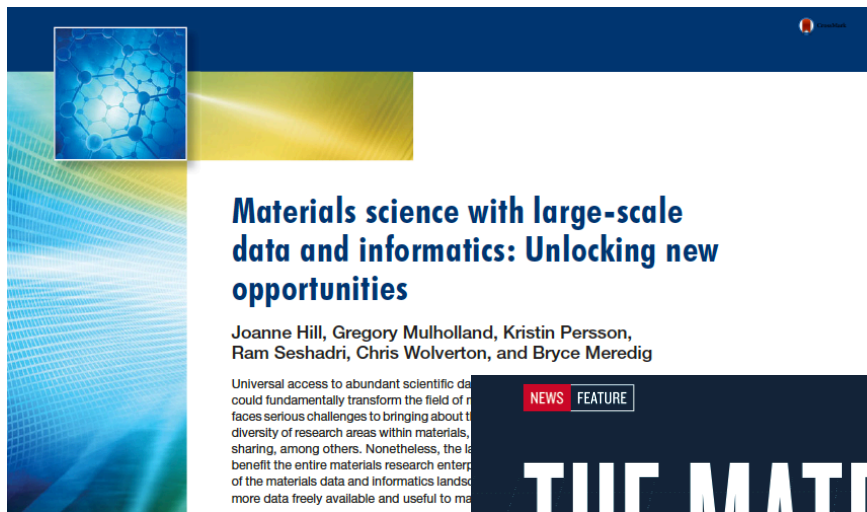


Joshua Lequieu



Juan de Pablo

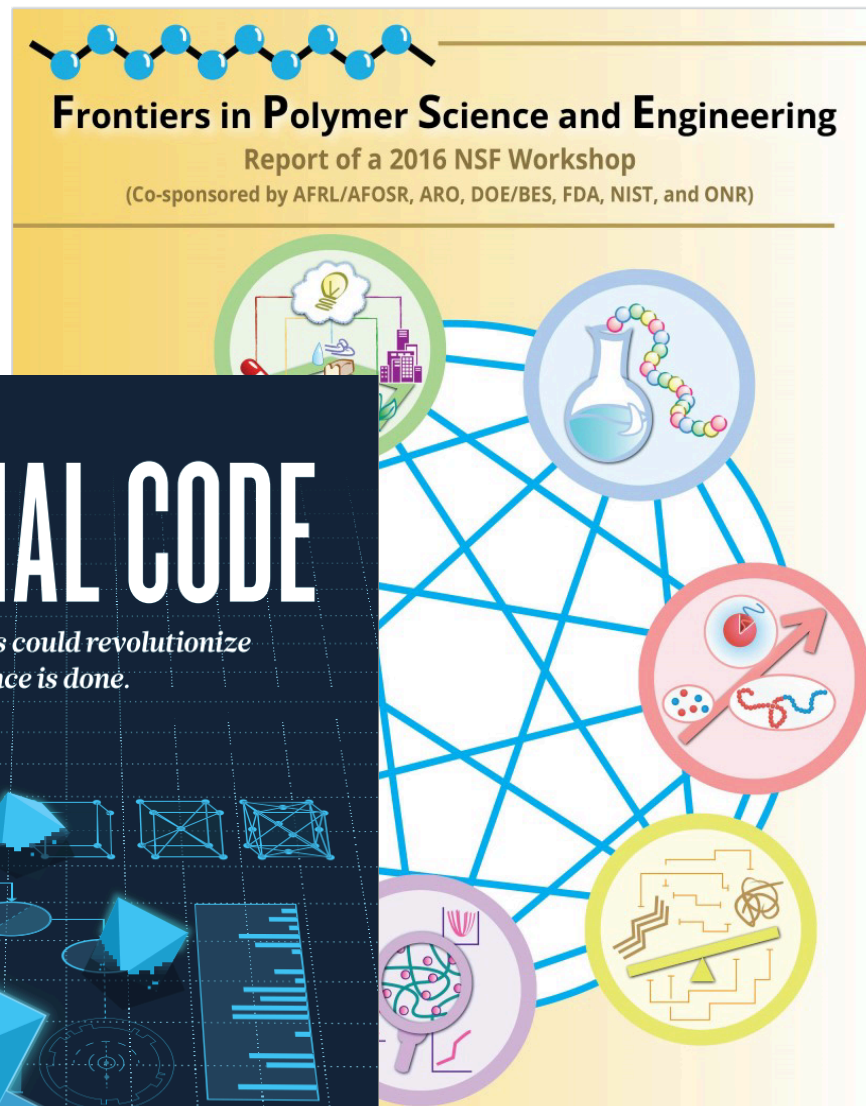
Need for polymeric databases



Materials science with large-scale data and informatics: Unlocking new opportunities

Joanne Hill, Gregory Mulholland, Kristin Persson, Ram Seshadri, Chris Wolverton, and Bryce Meredig

Universal access to abundant scientific data could fundamentally transform the field of materials science, but it also faces serious challenges to bringing about that transformation. The diversity of research areas within materials science, the need for data sharing, among others. Nonetheless, the benefits of data-driven materials science research benefit the entire materials research enterprise. The materials data and informatics landscape is becoming more data freely available and useful to materials scientists.



Frontiers in Polymer Science and Engineering

Report of a 2016 NSF Workshop
(Co-sponsored by AFRL/AFOSR, ARO, DOE/BES, FDA, NIST, and ONR)

NEWS FEATURE

THE MATERIAL CODE

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

BY NICOLA NOSENGO

© 2016 Materials Research Society

Introduction

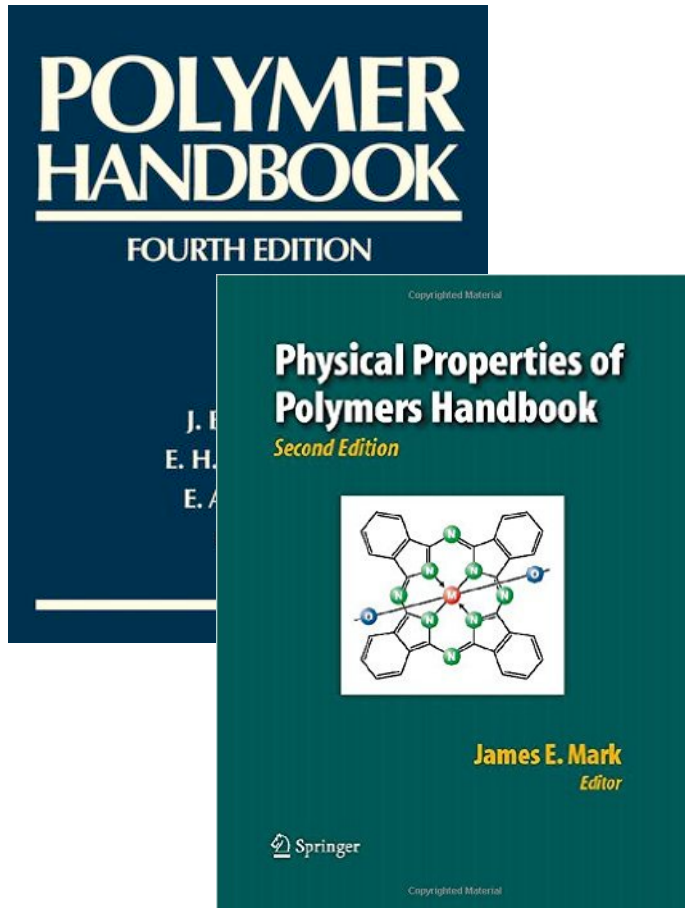
Data-intensive science has been described as the “fourth paradigm” for scientific exploration, with the first three being experiments, theory, and simulation.¹ While the value of data-intensive research approaches are becoming more apparent, the field of materials science has not yet experienced the same widespread adoption of these methods (as has occurred in bio-sciences,² astronomy,³ and particle physics⁴). Nonetheless, the potential impact of data-driven materials science is tremendous: Materials informatics could reduce the typical 10–20 year development and commercialization cycle⁵ 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 thorny challenges and promising initiatives that exist in the field. We conclude with a set of near-term recommendations for materials-data stakeholders. Our goal is to demystify data analytics and give readers from any subdiscipline within materials research enough information to understand how informatics techniques could apply to their own workflows.

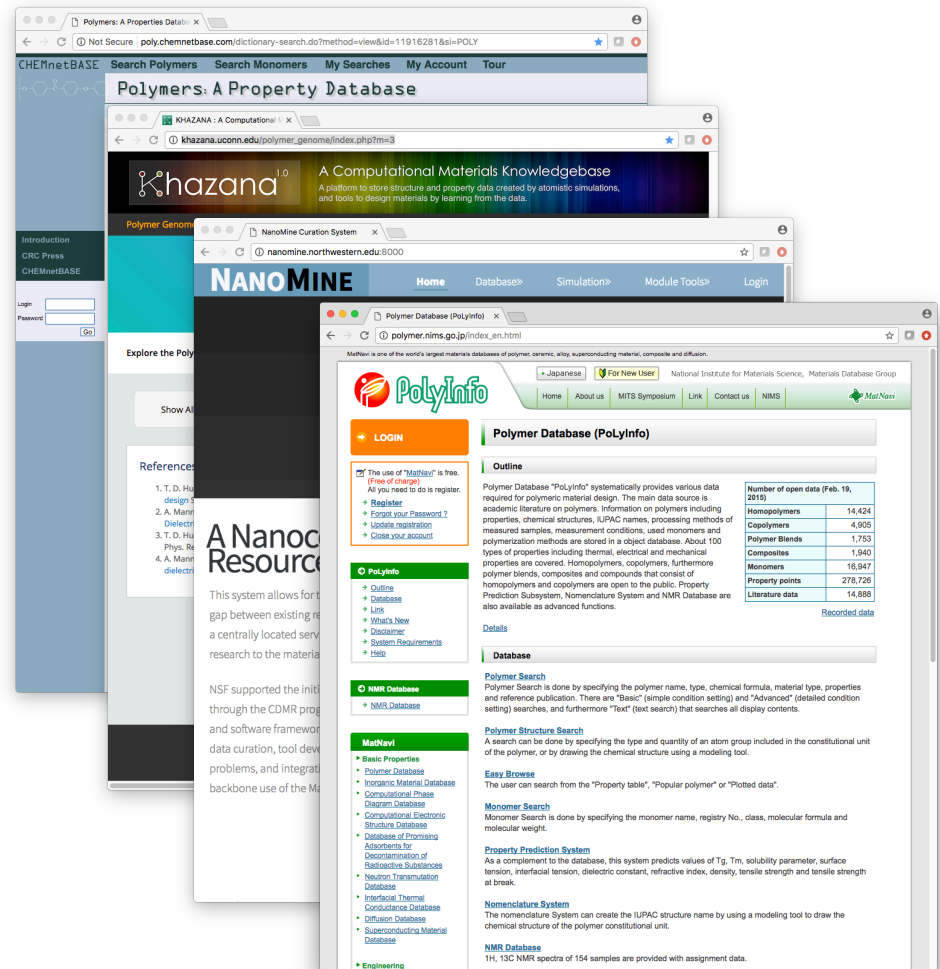
Joanne Hill, Citrine Informatics, USA, jh@citrine.io
Gregory Mulholland, Citrine Informatics, USA, greg@citrine.io
Kristin Persson, Lawrence Berkeley National Laboratory, USA, kpersson@lbl.gov
Ram Seshadri, University of California, Santa Barbara, USA, seshadri@mt.ucsb.edu
Chris Wolverton, Northwestern University, USA, c-wolverton@northwestern.edu
Bryce Meredig, Citrine Informatics, USA, bryce@citrine.io
doi:10.1557/mrs.2016.93

Existing resources

Paper-based



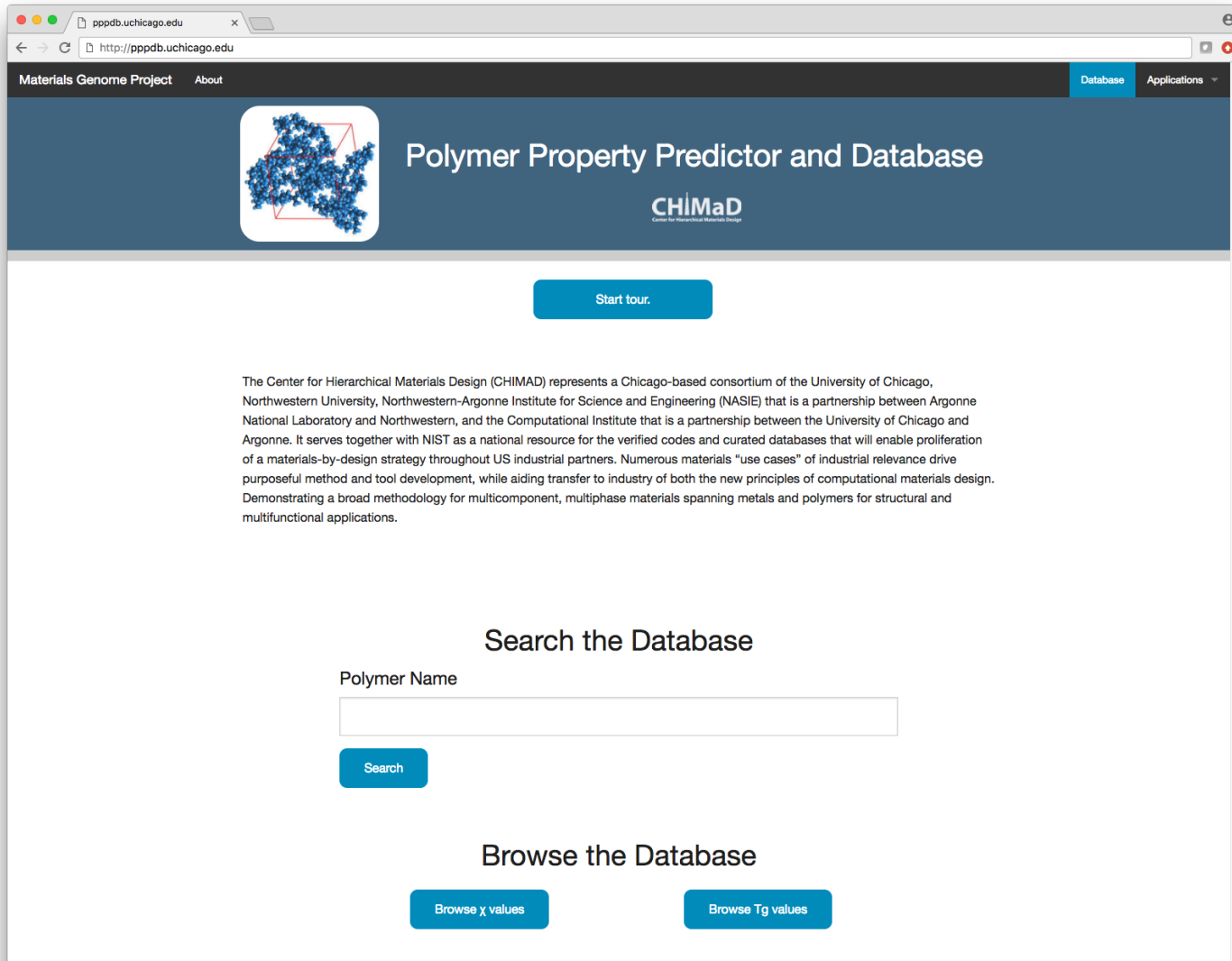
Web-based



limited accessibility of entire database and/or datasets that are too small

Polymer Property Predictor and Database

http://pppdb.uchicago.edu



The screenshot shows a web browser window with the URL `http://pppdb.uchicago.edu`. The page features a dark blue header with a navigation menu containing "Materials Genome Project", "About", "Database", and "Applications". A central banner displays a blue molecular structure icon, the title "Polymer Property Predictor and Database", and the CHIMaD logo (Center for Hierarchical Materials Design). Below the banner is a "Start tour." button. A paragraph of text describes the CHIMAD consortium and its mission. A "Search the Database" section includes a text input field for "Polymer Name" and a "Search" button. At the bottom, a "Browse the Database" section contains two buttons: "Browse χ values" and "Browse T_g values".

Materials Genome Project About Database Applications

Polymer Property Predictor and Database

CHIMaD
Center for Hierarchical Materials Design

[Start tour.](#)

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 \$\chi\$ values](#) [Browse \$T_g\$ values](#)

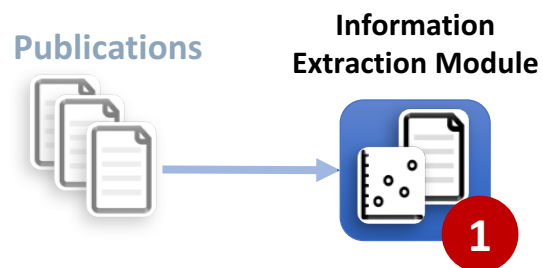
Flory Huggins χ parameter

Publications



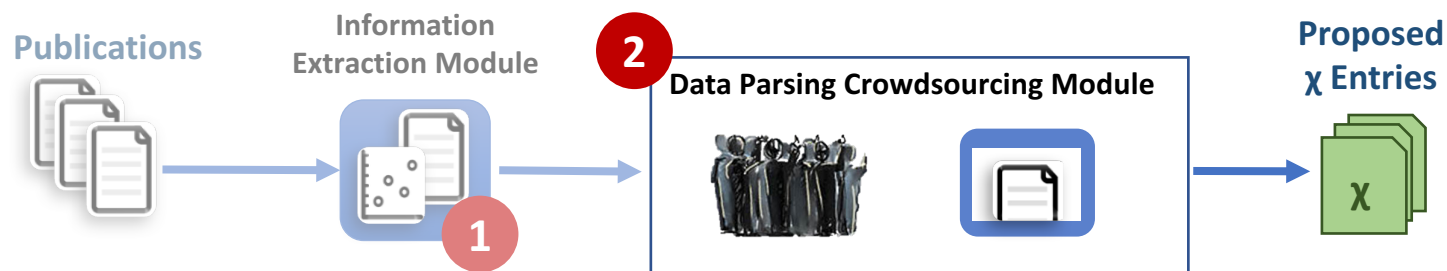
376 articles from
Macromolecules

Flory Huggins χ parameter



Automatically extract
metadata (title, author, etc.)

Flory Huggins χ parameter



Undergrads review papers and enter χ into an online form

Need for a polymer dictionary

Name	Type	Abbreviation
poly(ethylene-alt-propylene)	polymer	PEP
protonated poly(ethylene-alt-propylene)	polymer	pPEP
Polybutadiene	polymer	
polybutadiene	polymer	PB
polybutadiene	polymer	PBD
poly(butyl methacrylate)	polymer	PbMA
Poly(n-butyl methacrylate)	polymer	PnBMA-115
poly(methacrylic acid)-b-poly(methyl methacrylate) (A)	polymer	PMAA-PMMA (A)
poly(methacrylic acid)-b-poly(methyl methacrylate) (C)	polymer	PMAA-PMMA (C)
styrene	polymer	

prefixes

capitalization

ambiguous

input errors


The need for InChI

Multiple and trade names

Registry Number: 9003-53-6

Molecular Formula: (C₈H₈)_x

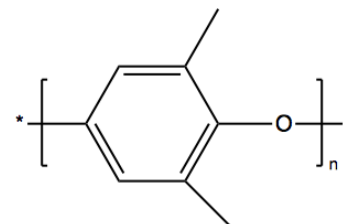
Chemical Name: Benzene, ethenyl-, homopolymer

[Hide Synonyms](#) 

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

1800+
for polystyrene

Identify synonyms




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

Breadth of CAS

Registry Number: 9004-34-6

Molecular Formula: W₉₉


Chemical Name: Cellulose

[Show Synonyms](#) 

Registry Number: 9000-11-7

Molecular Formula: C₂H₄O₃.xUnspecified


Chemical Name: Cellulose, carboxymethyl ether

[Show Synonyms](#) 

Registry Number: 9004-32-4

Molecular Formula: C₂H₄O₃.xNa.xW₉₉

Chemical Name: Cellulose, carboxymethyl ether, sodium salt

[Show Synonyms](#) 

Need for a polymer dictionary

Name	Type	Abbreviation
poly(ethylene-alt-propylene)	polymer	PEP
protonated poly(ethylene-alt-propylene)	polymer	pPEP
Polybutadiene	polymer	
polybutadiene	polymer	PB
polybutadiene	polymer	PBD
poly(butyl methacrylate)	polymer	PbMA
Poly(n-butyl methacrylate)	polymer	PnBMA-115
poly(methacrylic acid)-b-poly(methyl methacrylate) (A)	polymer	PMAA-PMMA (A)
poly(methacrylic acid)-b-poly(methyl methacrylate) (C)	polymer	PMAA-PMMA (C)
styrene	polymer	

prefixes

capitalization

ambiguous

input errors

Need something like PubChem for polymers



BioAssay

Compound

Substance

polystyrene

Go

Limits
Advanced

STYRENE



STRUCTURE



VENDORS



PHARMACOLOGY



LITERATURE



PATENTS



BIOACTIVITIES

PubChem CID: 7501

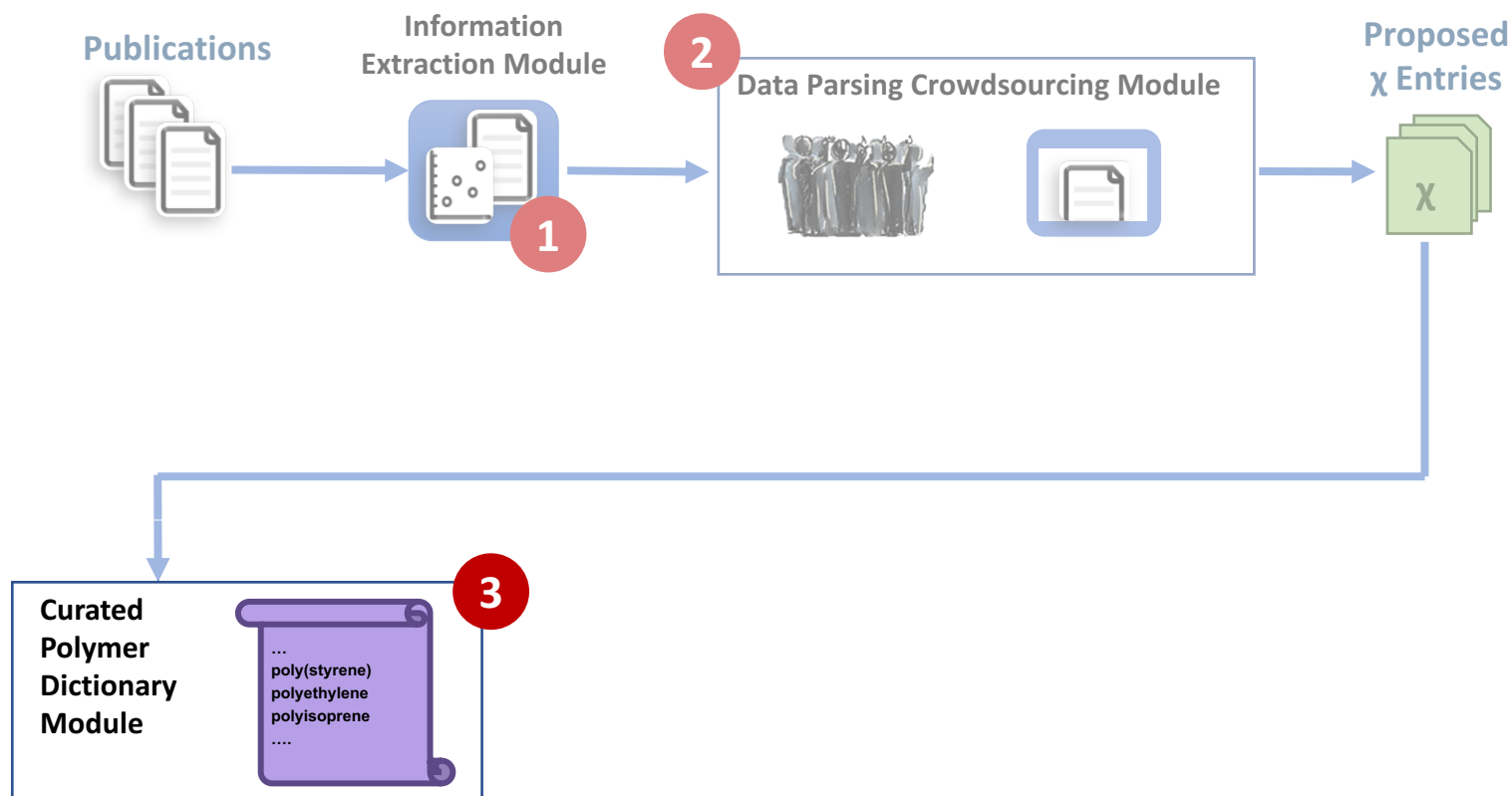
Chemical Names: STYRENE; Ethenylbenzene; Phenylethylene; Vinylbenzene; 100-42-5; Styrol [More...](#)

Molecular Formula: C_8H_8 or $C_6H_5CHCH_2$

Molecular Weight: 104.152 g/mol

InChI Key: PPBRXRYQALVLMV-UHFFFAOYSA-N

Flory Huggins χ parameter



Developing the polymer dictionary

Name

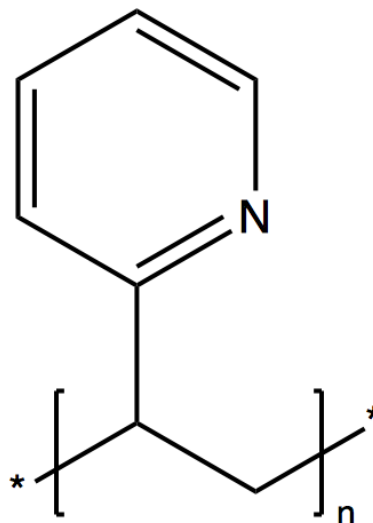
poly(2-vinylpyridine)

Abbreviation

P2VP

Structure

(saved as .mol file)



InChIKey and InChI

KGIGUEBEKRSTEW-BBVYVPKBA-N

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

The polymer dictionary

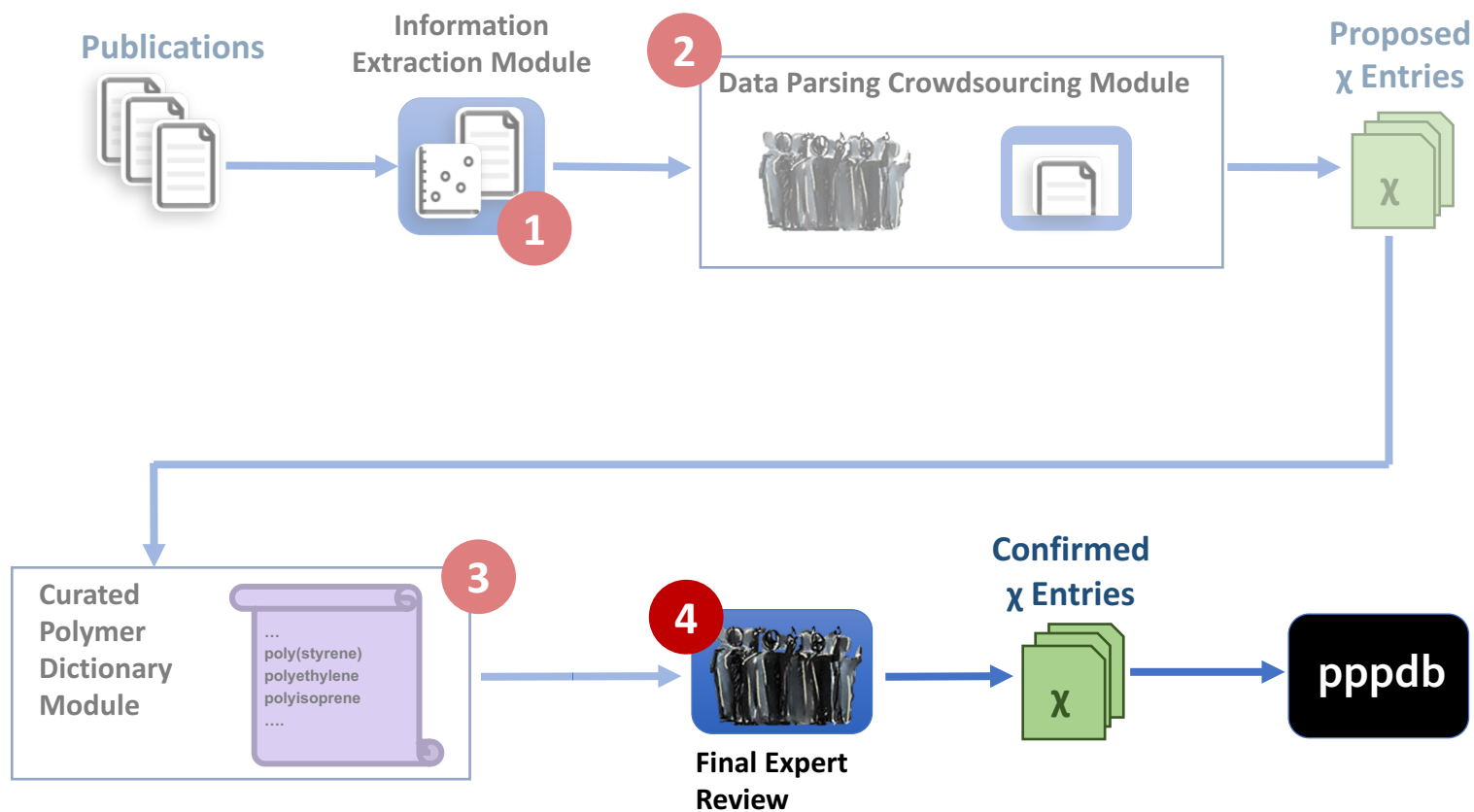
88 entries

3 without InChI

associated .mol files

Name	Abbreviation	InChI	InChI Key	Thesaurus Names	Thesaurus Abbreviation
poly(2-vinylpyridine)	P2VP	1B/C7H7N/c1-2-7-5-3-4-6-8-7/h2-6H,1H2/z101-1-	KGIGUEBEKRSTEW-BBVYVPKKBAN		
poly(2,2-bis(trifluoromethyl)-4,5-difluoro-1,3-dioxole)		1B/C5F8O2/c6-1-2(7)15-3(14-1,4(8,9)10)5(11,12)	YSYRISKCBOPJRG-XLALVCCSBA-N		
poly(2,6-dimethyl-1,4-phenylene oxide)	PXE	1B/C8H8O/c1-5-3-7-4-6(2)8(5)9-7/h3-4H,1-2H3/z-	GVLZQVREHWQBJN-NBWNIIOTOBAN	poly(xylylenyl ether)	
poly(3-(2-ethylhexyl)thiophene)	P3EHT	1B/C12H18S/c1-3-5-6-9(4-2)7-10-8-11-12(10)13-	UCZXRGRJTHIALN-HUZCROMBBAN		
poly(3-butylthiophene)	P3BT	1B/C8H10S/c1-2-3-4-6-5-7-8(6)9-7/h5H,2-4H2,1H-	MFKYCRZHEWXER-WCYUVGCXBAN		
poly(3-dodecylthiophene)	P3DDT	1B/C16H26S/c1-2-3-4-5-6-7-8-9-10-11-12-14-13-	OWMLCFMOUNBSQ-ORBAQIJBUBAN		
poly(3-hexylthiophene)	P3HT	1B/C10H14S/c1-2-3-4-5-6-8-7-9-10(8)11-9/h7H,2-	VWCCBCCELGMSR-RHPJWNLABAN	poly(3-hexylthiophene-2,5-diyl)	
poly(3-methylthiophene)	P3MT	1B/C5H4S/c1-3-2-4-5(3)6-4/h2H,1H3/z101-1-6(4,-	JZAQUQHOPLBQPI-XVBQSDAFBAN		
poly(3-octylthiophene)	P3OT	1B/C12H18S/c1-2-3-4-5-6-7-8-10-9-11-12(10)13-	IXMZQXBZTSSNAG-HUZCROMBBAN		
poly(4-acetoxystyrene)	P4AS	1B/C10H10O2/c1-3-9-4-6-10(7-5-9)12-8(2)11/h3-	JAMNSIXSLVPLNC-YUFRJTRMBAN		
poly(4-hydroxystyrene)	PHS	1B/C8H8O/c1-2-7-3-5-8(9)6-4-7/h2-6,9H,1H2/z10-	FUGYGGDSWSUORM-RRFBGHKIBAN		
poly(4-tert-butylstyrene)	PtBS	1B/C12H16/c1-5-10-6-8-11(9-7-10)12(2,3)4/h5-9H	QEDJMOONZLUMC-KJZCOQPOBAN		
poly(4-vinylbenzyltrimethylammonium chloride)	PVBTAmC	1B/C12H18N.ClH/c1-5-11-6-8-12(9-7-11)10-13(2,-	TVXNKQRAZONMHJ-IPOBBXJVBBAN		
poly(4-vinylpyridine)	P4VP	1B/C7H7N/c1-2-7-3-5-8-6-4-7/h2-6H,1H2/z101-1-	KFDVVPJUYSEJTH-BBVYVPKKBAN		
poly(6-methyl-Δ ⁵ -caprolactone)	PMCL	1B/C7H12O2/c1-6-4-2-3-5-7(8)9-6/h6H,2-5H2,1H3-	WZRNGGFHDMOCEA-PBJUEKBZBAN		
poly(acrylic acid)	PAA	1B/C3H4O2/c1-2-3(4)5/h2H,1H2,(H,4,5)/z101-1-5	NIXOWILDQLNWCW-YZJLBCBGBAN		
poly(allylamine hydrochloride)	PAH	1B/C3H7N.ClH/c1-2-3-4;/h2H,1,3-4H2;1H/z101-1-	MLGWTRRHANFCC-XEQKNLBJBAN		
poly(benzyl methacrylate)	PbnMA	1B/C11H12O2/c1-9(2)11(12)13-8-10-6-4-3-5-7-10	AQJQJEFVRHODZFN-PQSYNHGEBAN		
poly(butyl acrylate)	PBA	1B/C7H12O2/c1-3-5-6-9-7(8)4-2/h4H,2-3,5-6H2,1H	CQEYYJKESMYFG-YSNYPLFEBAN	poly(n-butyl acrylate)	
poly(butyl methacrylate)	PbMA	1B/C8H14O2/c1-4-5-6-10-8(9)7(2)3/h2,4-6H2,1,3-	SOGAXMICEFXMKE-WWNZXXXVBAN	poly(n-butyl methacrylate)	
poly(butylene oxide)	PBO	1B/C4H8O/c1-2-4-5-3-1/h1-4H2/z101-1-5(1,2,1,3)	WYURNTSHIVDZCO-PYASYNHBBAN		
poly(cyclohexylethylene)	PCHE	1B/C8H14/c1-2-8-6-4-3-5-7-8/h2,8H,1,3-7H2/z10-	LDLDYFCDDKENPD-BBVYVPKKBAN		
poly(diethylhexyloxy-p-phenylenevinylene)		1B/C24H38O2/c1-5-9-11-19(7-3)17-25-23-15-22-14	QSFVDVZTDIATNP-FRAFILGZBAN		
poly(diisooamyl itaconate)		1B/C15H26O4/c1-4-6-8-10-18-14(16)12-13(3)15(1)	NJCKCJUVRQPTKF-IXYATHDYBAN		
poly(ethyl ethylene)	PEE	1B/C4H8/c1-3-4-2/h3H,1,4H2,2H3/z101-1-4(1.3)	VXNZUUAJNFGPBY-XMMBCNSCBAN		
poly(ethyl methacrylate)	PEMA	1B/C6H10O2/c1-4-8-6(7)5(2)3/h2,4H2,1,3H3/z101-	SUPCQIBBMFXVTL-ANYLYJCWBAN		
poly(ethylene oxide)	PEO	1B/C2H4O/c1-2-3-1/h1-2H2/z101-1-3(1,2,1,3,2,3)	IAYPIBMASNFSP-L-GCGQHNKHBAN	poly(ethylene glycol)	PEG
poly(ethylene-alt-propylene)	PEP	1B/C5H10/c1-5-3-2-4-5/h5H,2-4H2,1H3/z101-1-5(1	BDJAEZIRGNQCBZ-HZCQOQKVBAN	poly(ethylene-propylene); polymethylbutylene	
poly(ethylene-r-butylene)	PEB	1B/C4H8.C2H4/c1-3-4-2;/1-2/h3H,1,4H2,2H3;1-2H2	WXCCUWHSJWOTRV-LWCKEHRHBAN		
poly(ferrocenylmethylsilyl silane)					
poly(hexafluoroisopropylidene diphthalic anhydride-alt-2)	6FDA-TMPD	1B/C29H18F6N2O4/c1-11-12(2)22-14(4)13(3)21(11)	UWVNHDDZPHOGRQ-AGENGVHNBAN		
poly(hexene oxide)	PHO	1B/C2H3BrO/c3-2-1-4-2/h2H,1H2/z101-1-4(1,2,1,-	XOOVDXMMNOFROX-UAKBPNCWBAN		
poly(hydroxyethyl acrylate)		1B/C6H10O3/c1-5(2)6(8)9-4-3-7/h7H,1,3-4H2,2H3-	WOBHKFSMXKNTIM-MWVPMDSBBAN		
poly(ε-caprolactone)	PCL	1B/C6H10O2/c7-6-4-2-1-3-5-8-6/h1-5H2/z101-1-8	PAPBSGBWRJIAAV-CMRMDLKMBAN		
poly(lactic acid)	PLA	1B/C3H4O2/c1-2-3(4)5-2/h2H,1H3/z101-1-5(2,3,2-	YCHXNMPJFFFEIJG-MGVHKUGVBAN	poly(lactide); poly(lactic acid)	
poly(methacrylic acid)	PMAA	1B/C4H6O2/c1-3(2)4(5)6/h1H2,2H3,(H,5,6)/z101-	CERQOIWHTDAKMF-IXKXVMLBBAN		
poly(methyl acrylate)	PMA	1B/C4H6O2/c1-3-4(5)6-2/h3H,1H2,2H3/z101-1-6(1-	BAPJBEWLBFGYME-IXKXVMLBBAN		
poly(methyl methacrylate)	PMMA	1B/C5H8O2/c1-4(2)5(6)7-3/h1H2,2-3H3/z101-1-7(1-	VVQNEPGJFQJSBK-KCAOANCSBAN		
poly(n-hexyl methacrylate)	PHMA	1B/C10H18O2/c1-4-5-6-7-8-12-10(11)9(2)3/h2,4-	LNCPIVMCTVXXOY-XJFATUNXBAN		
poly(N-isopropylacrylamide)	PNIPAM	1B/C6H11NO/c1-4-6(8)7-5(2)3/h4-5H,1H2,2-3H3,(1	QNILTEGFHQSXKF-YWFYHYGJBAN		
poly(n-pentyl methacrylate)	PnPMA	1B/C9H16O2/c1-4-5-6-7-11-9(10)8(2)3/h2,4-7H2,	GYDSPAVLTMAXHT-VVXONTJNBAN		

Flory Huggins χ parameter



Final review and push to database

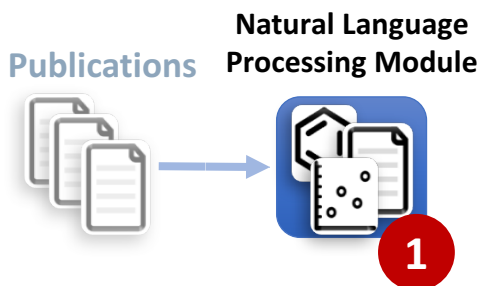
Glass transition temperature

Publications



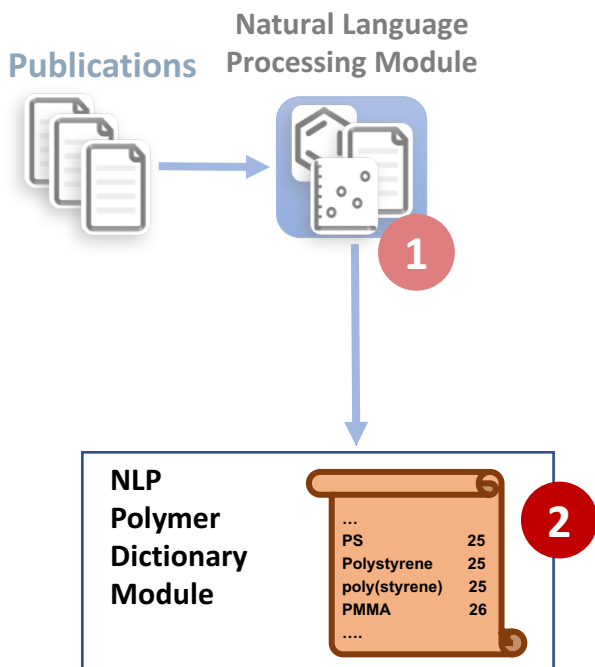
6,090 articles from
Macromolecules

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"

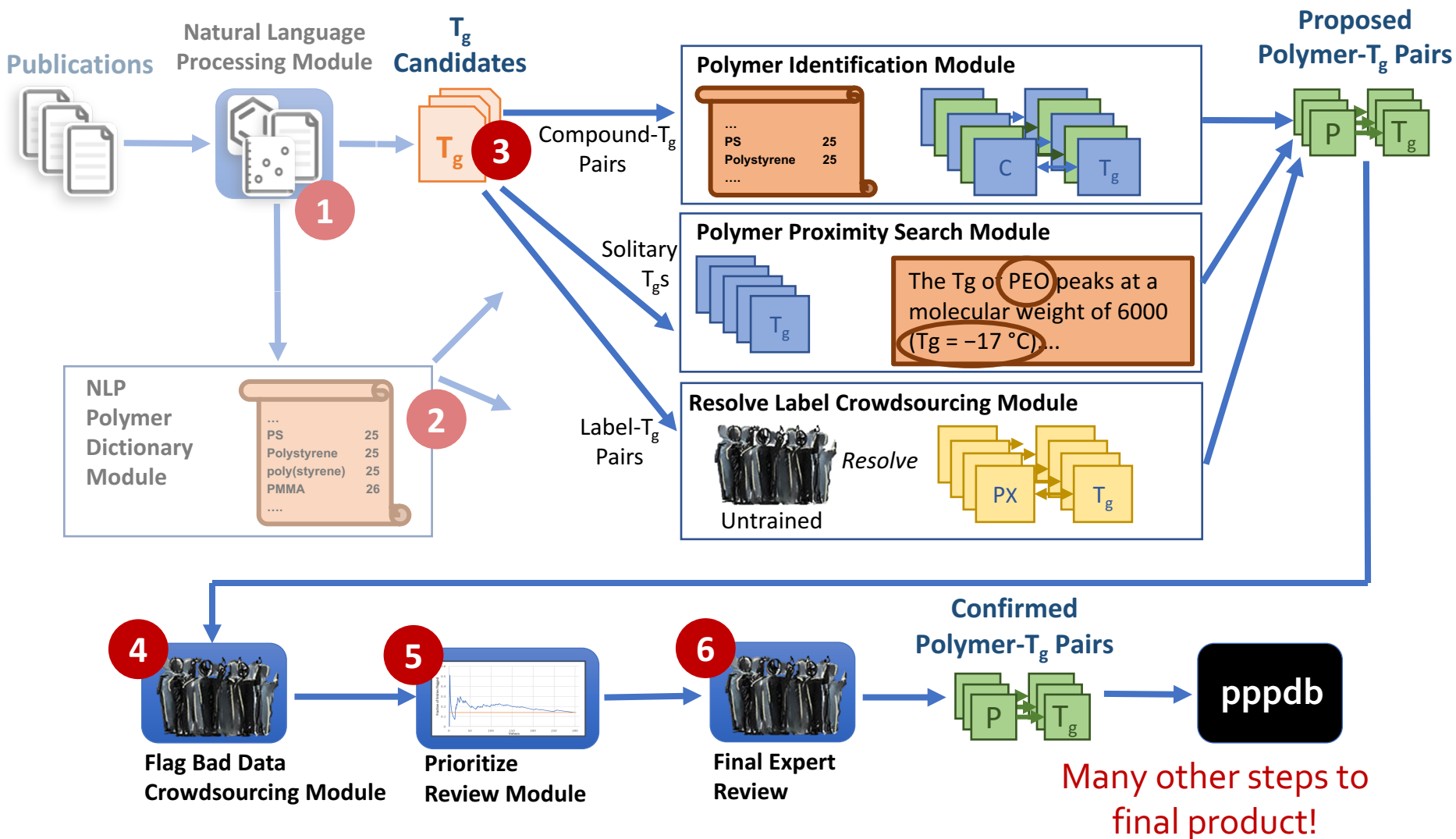
NLP Polymer Dictionary

Name		
Polystyrene	}	
poly(styrene)		various forms
polystyrene		
polystyrenes		
PS		
PSS	} not plural of PS	
polyimides	}	
polyolefin		family names
copolymer 10	}	
poly(2,4'-BF-a)		labels not names
macroporous poly(N-isopropylacrylamide)gel	prefixes/suffixes	

12,814 polymers in the dictionary

Work in progress to clean up errors above and adding InChI

Glass transition temperature




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

Broadness of CAS

Registry Number: **9004-34-6**

Molecular Formula: W₉₉


Chemical Name: Cellulose

Show Synonyms 

Registry Number: **9000-11-7**

Molecular Formula: C₂H₄O₃.xUnspecified

Chemical Name: Cellulose, carboxymethyl ether

Show Synonyms 

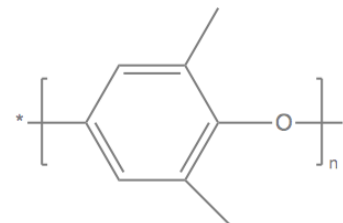
Registry Number: **9004-32-4**

Molecular Formula: C₂H₄O₃.xNa.xW₉₉

Chemical Name: Cellulose, carboxymethyl ether, sodium salt

Show Synonyms 

Identify synonyms



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

Input/output for machine learning

Natural Language
Processing Module

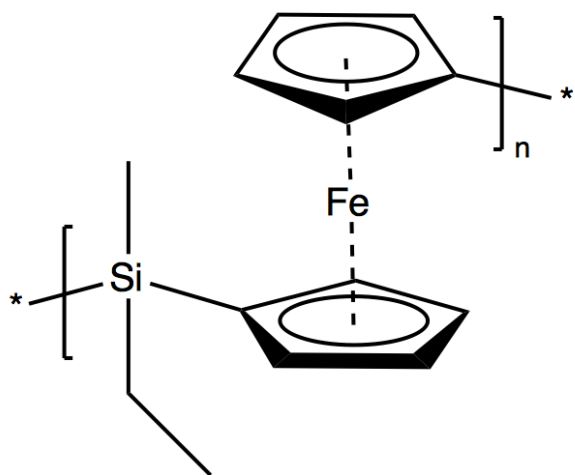


1

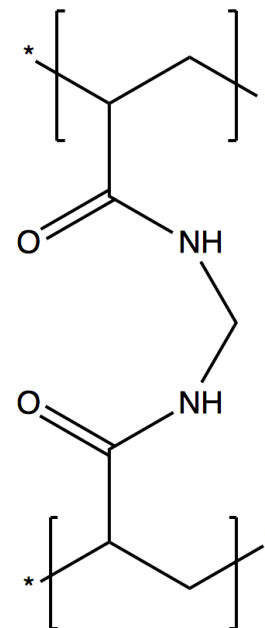
1B/C₈H₈O/c1-5-3-7-46(2)8(5)9-7/
h3-4H,12H3/z101-1-9(7,9,8,9)

Limitations of current InChI

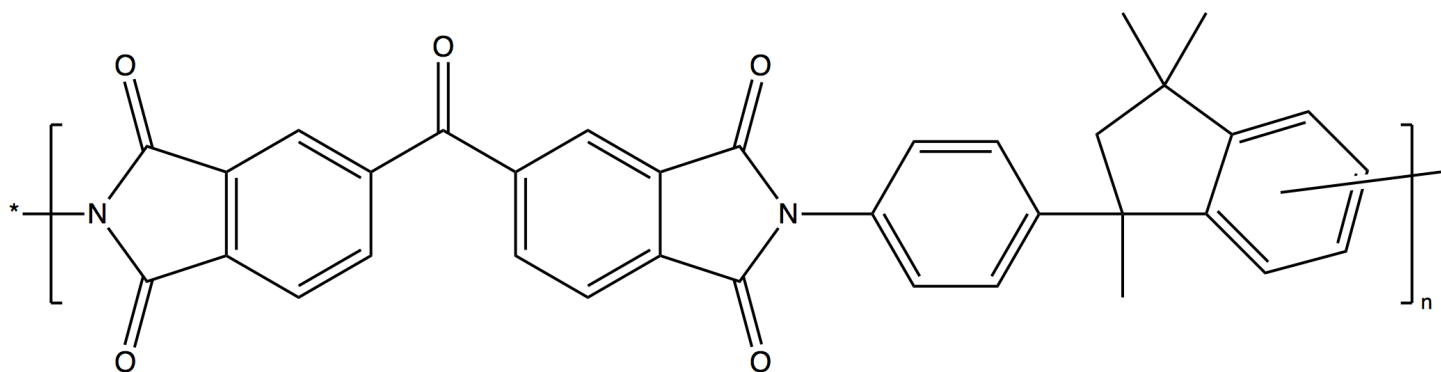
Organometallic



Branching / crosslinks



Markush



Conclusions and outlook

<http://pppdb.uchicago.edu>

263 χ

258 T_g

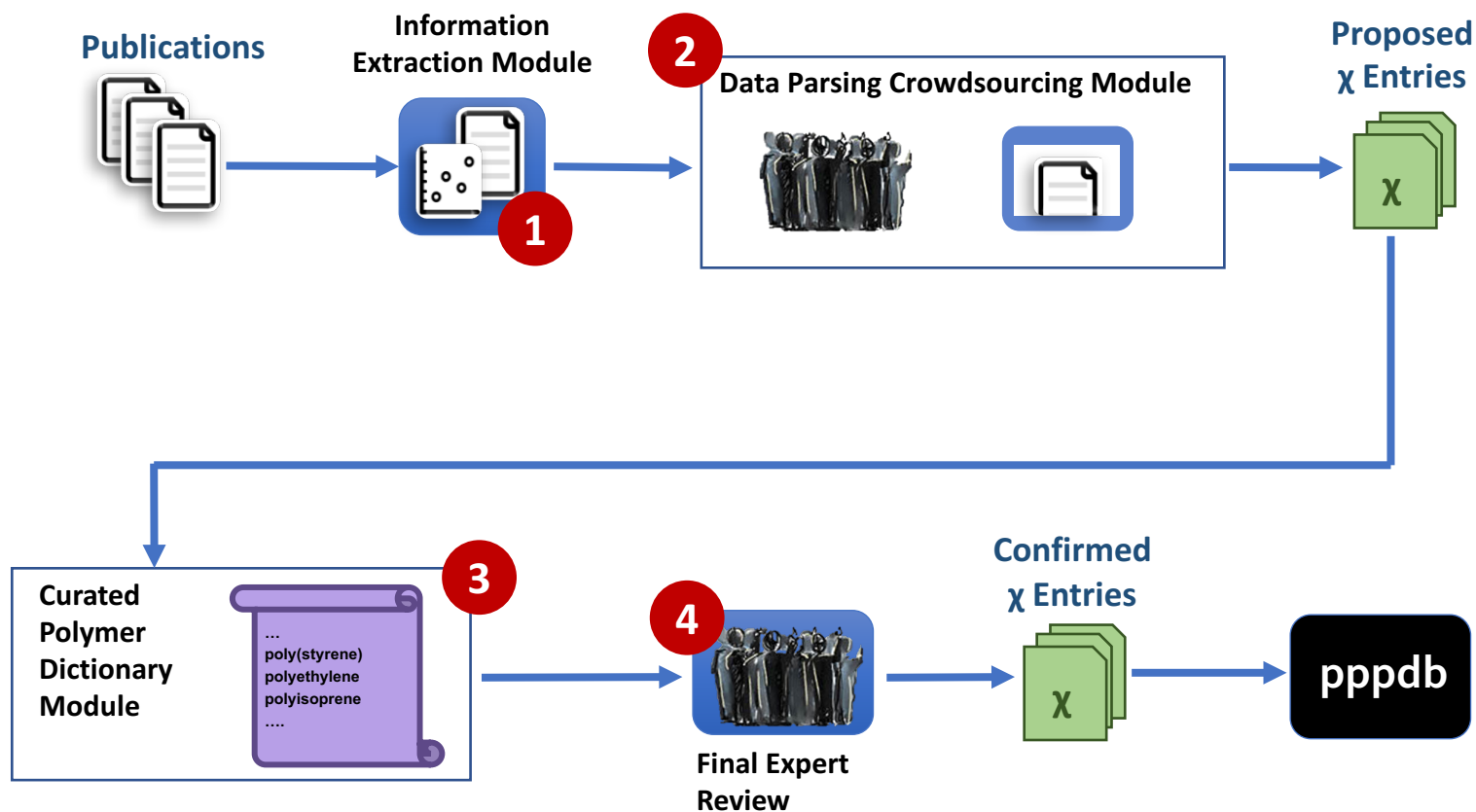
Future work

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

Advances still need for InChI

- Organometallics
- Branching / cross-links
- Markush

Flory Huggins χ parameter



Glass transition temperature

