Incorporating InChI into a polymeric database

Debra J. Audus

State and Future of the IUPAC InChI

August 16, 2017

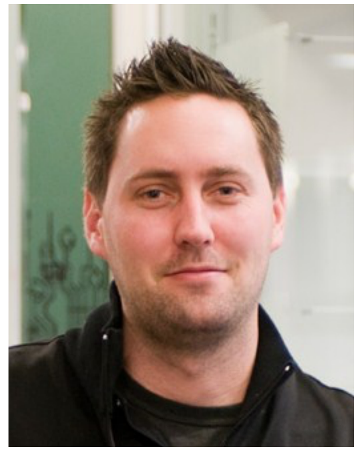
THE UNIVERSITY OF CHICAGO
CHiMaD
NIST National Institute of Standards and Technology U.S. Department of Commerce

Acknowledgements

THE UNIVERSITY OF CHICAGO

Computer Science

Roselyne Tchoua



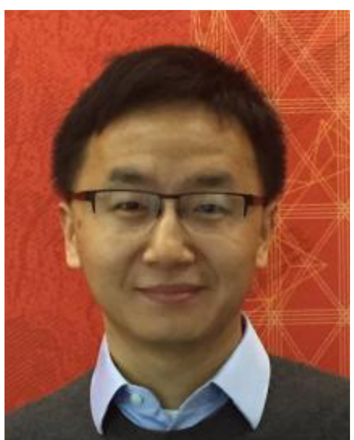
Kyle Chard

Logan Ward

Ian Foster

Stanford University

Chemical Engineering



Jian Qin

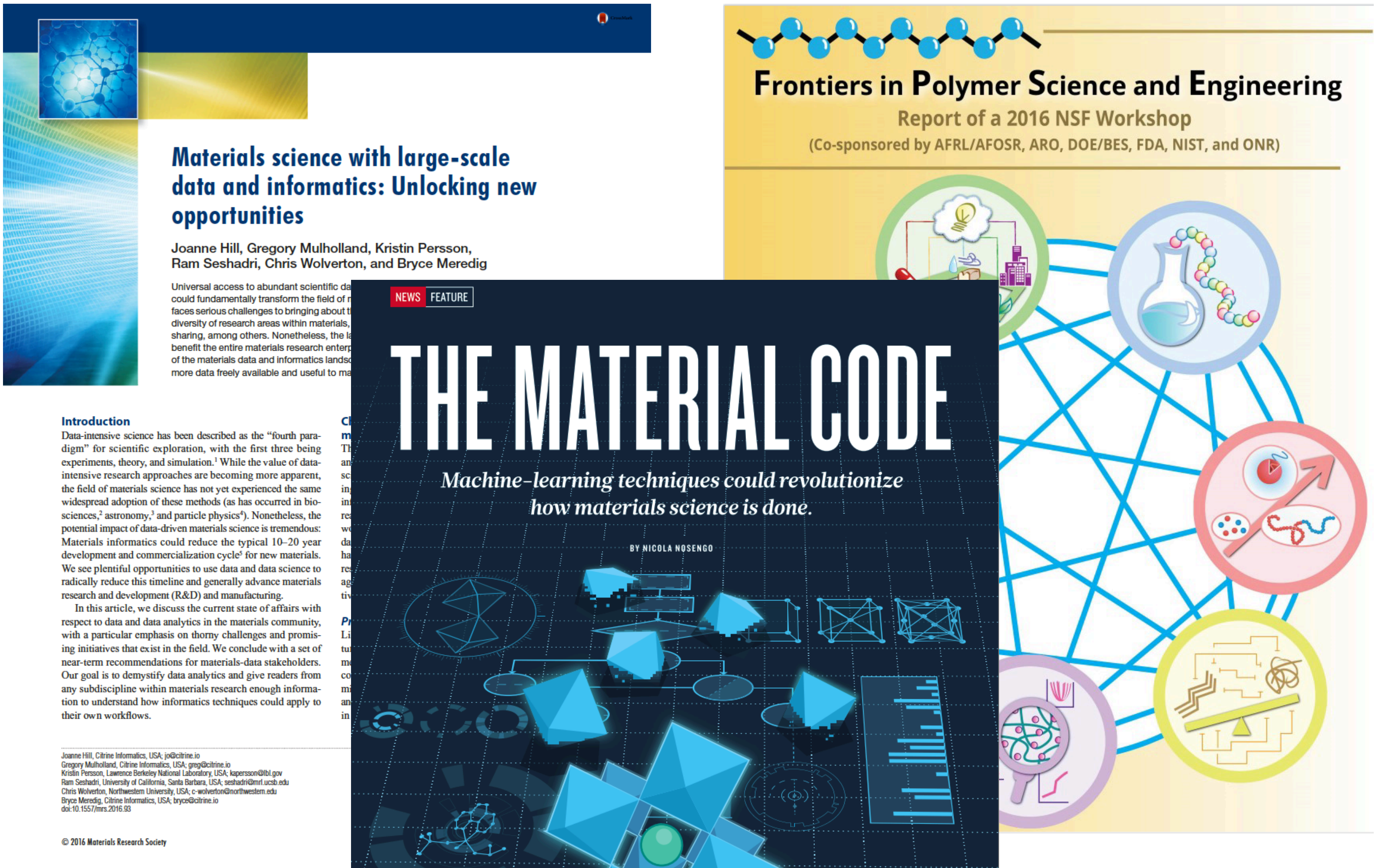
THE UNIVERSITY OF CHICAGO

Molecular Engineering

Joshua Lequieu

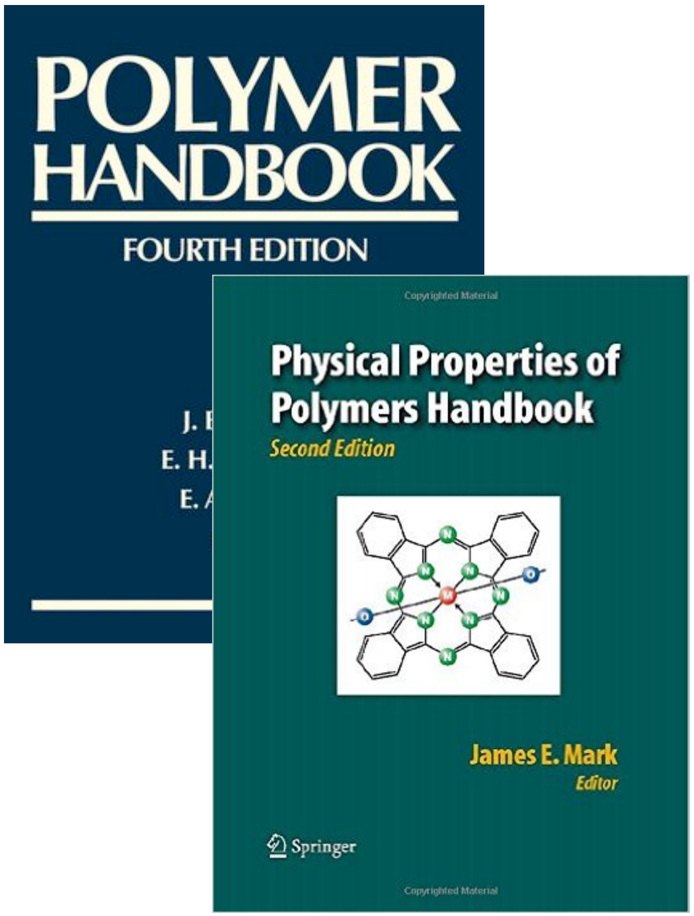
Juan de Pablo

Need for polymeric databases

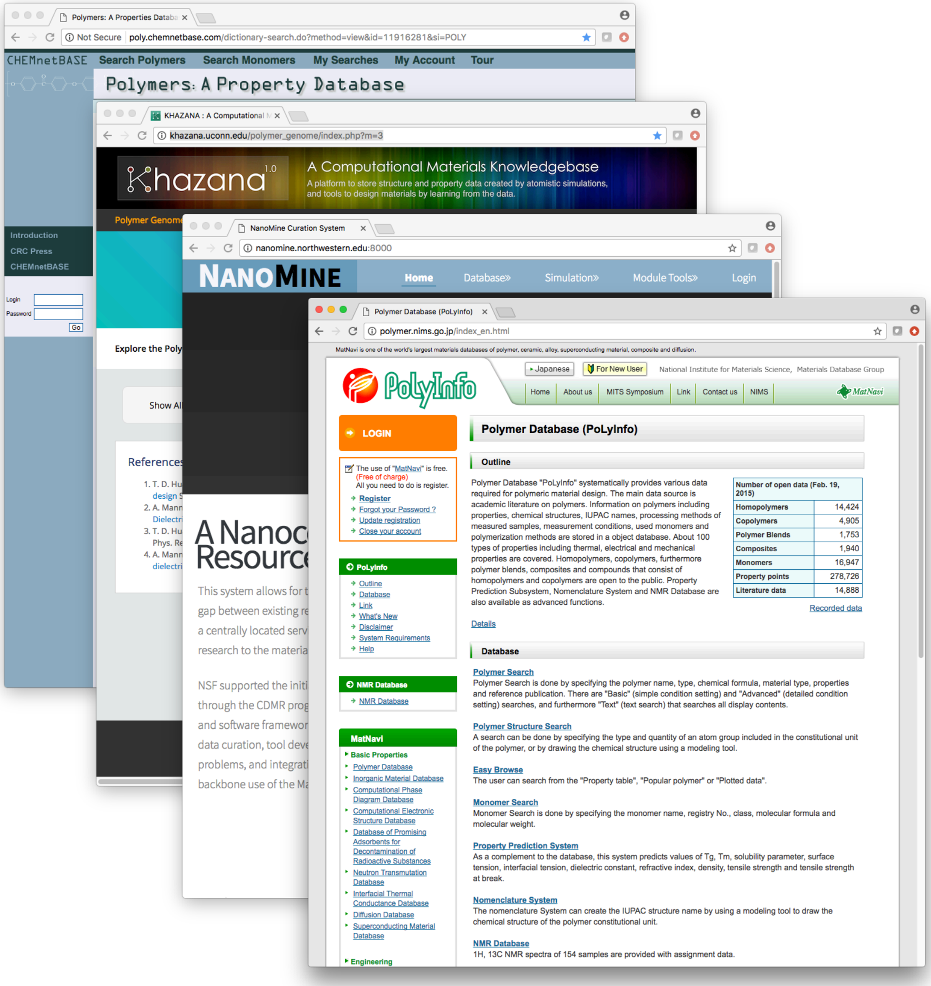


Existing resources

Paper-based



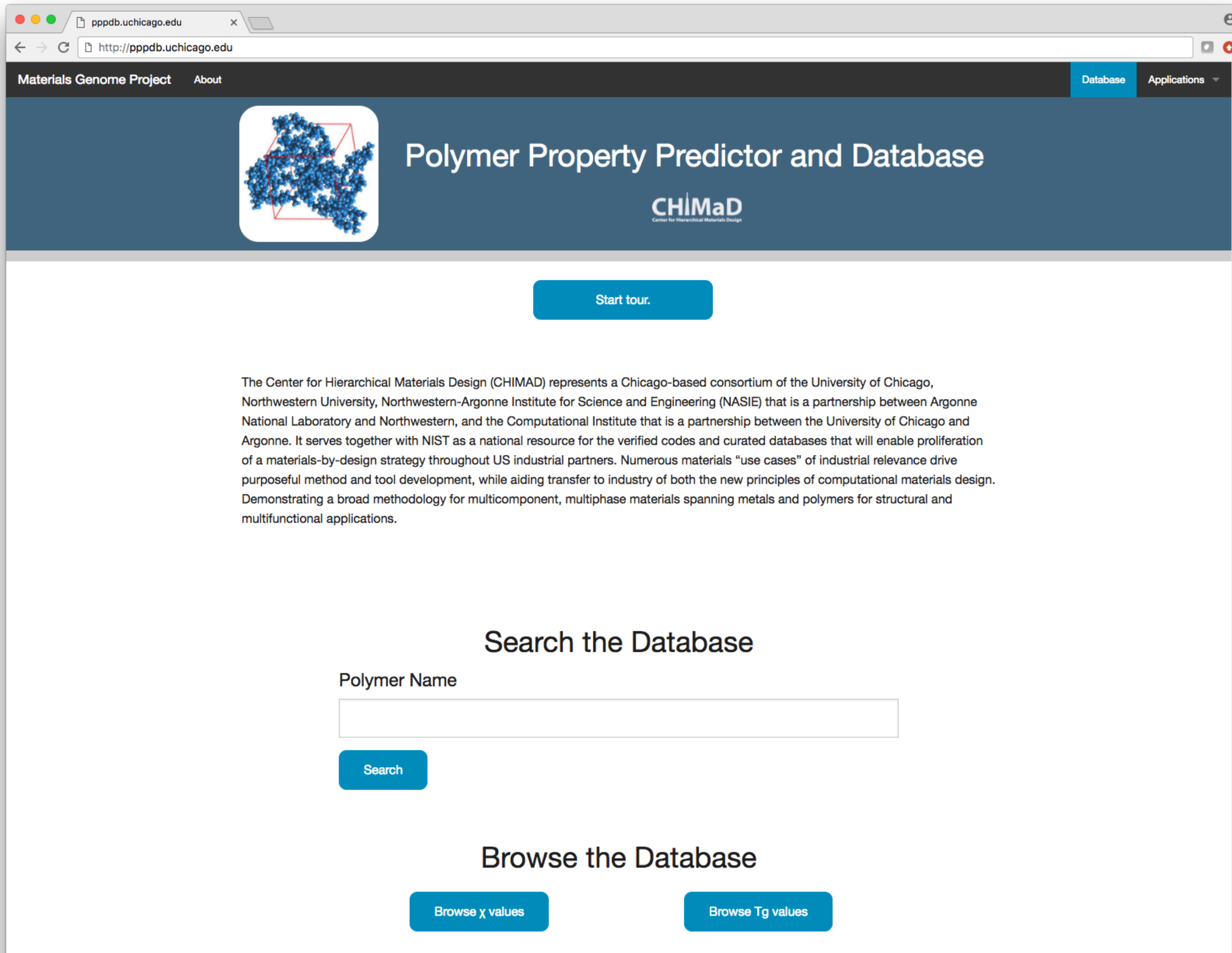
Web-based



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

Polymer Property Predictor and Database

http://pppdb.uchicago.edu



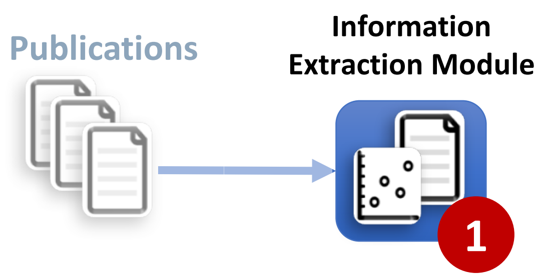
Flory Huggins χ parameter

Publications

376 articles from
Macromolecules

R.B. Tchoua, J. Qin, D.J. Audus, *et al. J. Chem. Educ.*, **93**, 1561-1568 (2016)
R.B. Tchoua, K. Chard, D.J. Audus, *et al. Procedia Comp. Sci.*, **80**, 386-397 (2016)

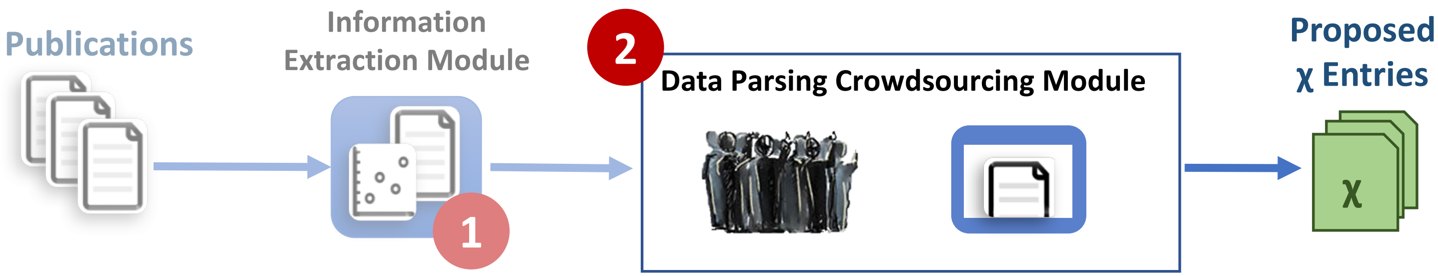
Flory Huggins χ parameter



Automatically extract
metadata (title, author, etc.)

R.B. Tchoua, J. Qin, D.J. Audus, *et al. J. Chem. Educ.*, **93**, 1561-1568 (2016)
R.B. Tchoua, K. Chard, D.J. Audus, *et al. Procedia Comp. Sci.*, **80**, 386-397 (2016)

Flory Huggins χ parameter



Undergrads review papers and enter χ into an online form

R.B. Tchoua, J. Qin, D.J. Audus, *et al. J. Chem. Educ.*, **93**, 1561-1568 (2016)
R.B. Tchoua, K. Chard, D.J. Audus, *et al. Procedia Comp. Sci.*, **80**, 386-397 (2016)

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 (rows 1–2)

capitalization (rows 3–5)

ambiguous (rows 6–7)

input errors (rows 8–10)

The need for InChI

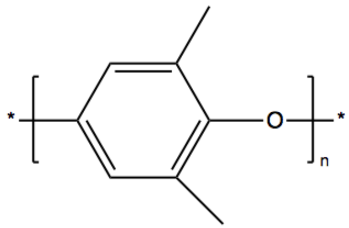
Multiple and trade names

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)

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

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	

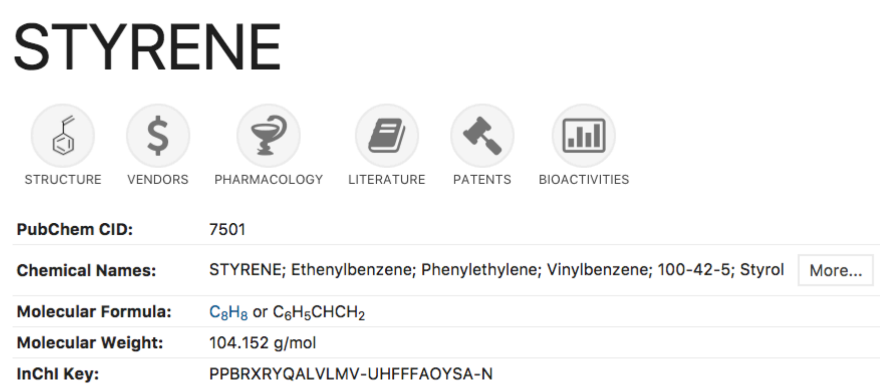
prefixes

capitalization

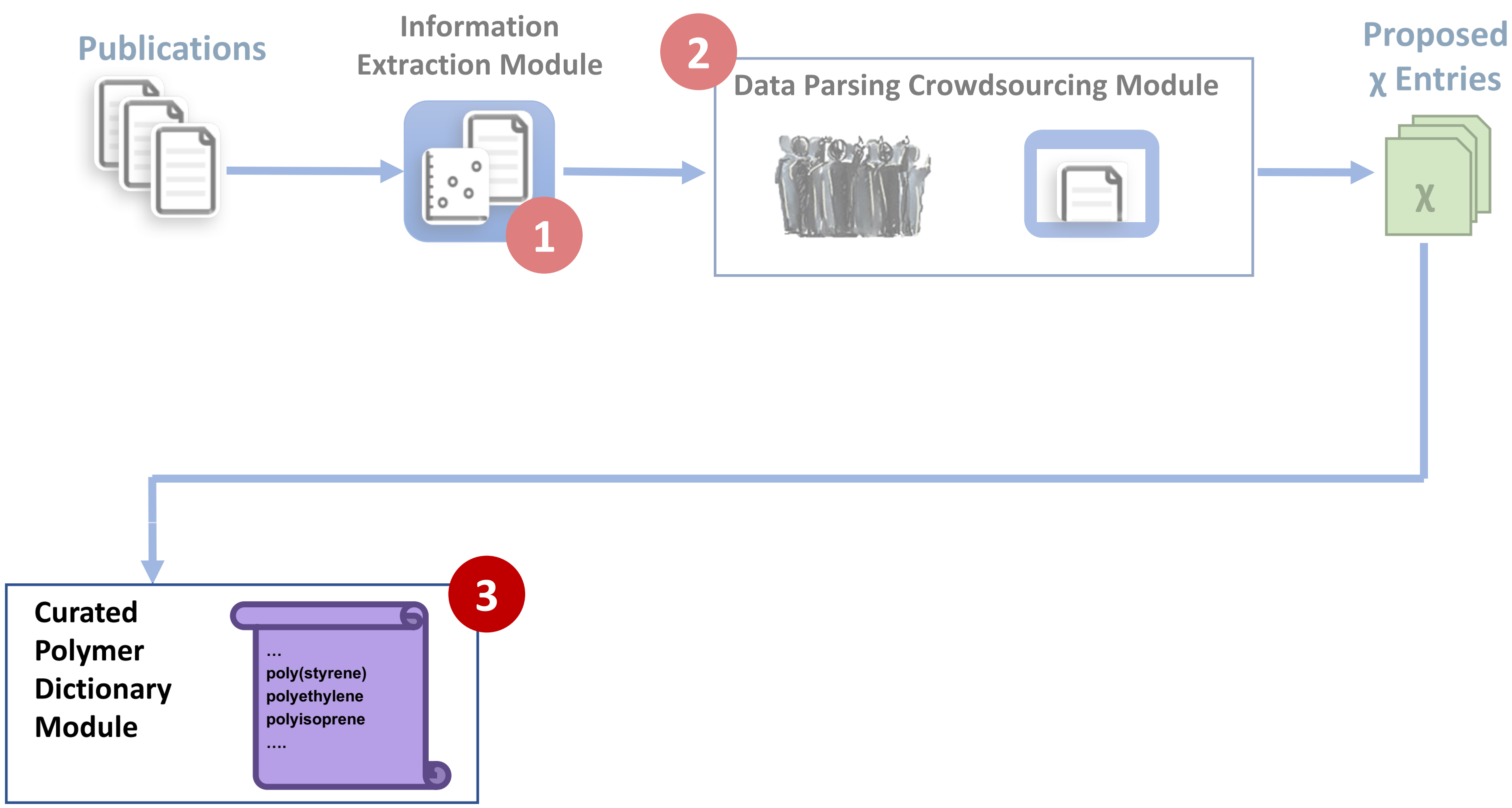
ambiguous

input errors

Need something like PubChem for polymers



Flory Huggins χ parameter



R.B. Tchoua, J. Qin, D.J. Audus, *et al. J. Chem. Educ.*, **93**, 1561-1568 (2016)
R.B. Tchoua, K. Chard, D.J. Audus, *et al. Procedia Comp. Sci.*, **80**, 386-397 (2016)

Developing the polymer dictionary

Name

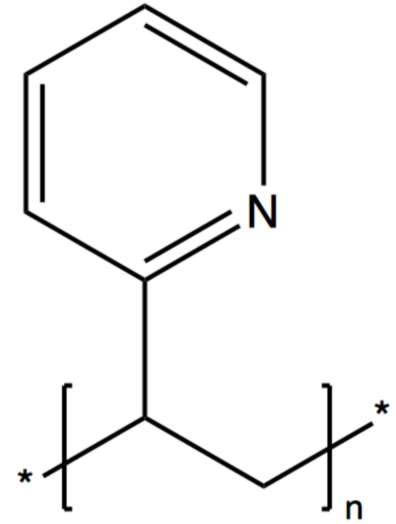
poly(2-vinylpyridine)

Abbreviation

P2VP

Structure

(saved as .mol file)



InChIKey and InChI

KGIGUEBEKRSTEW-BBVYVPKKBA-N

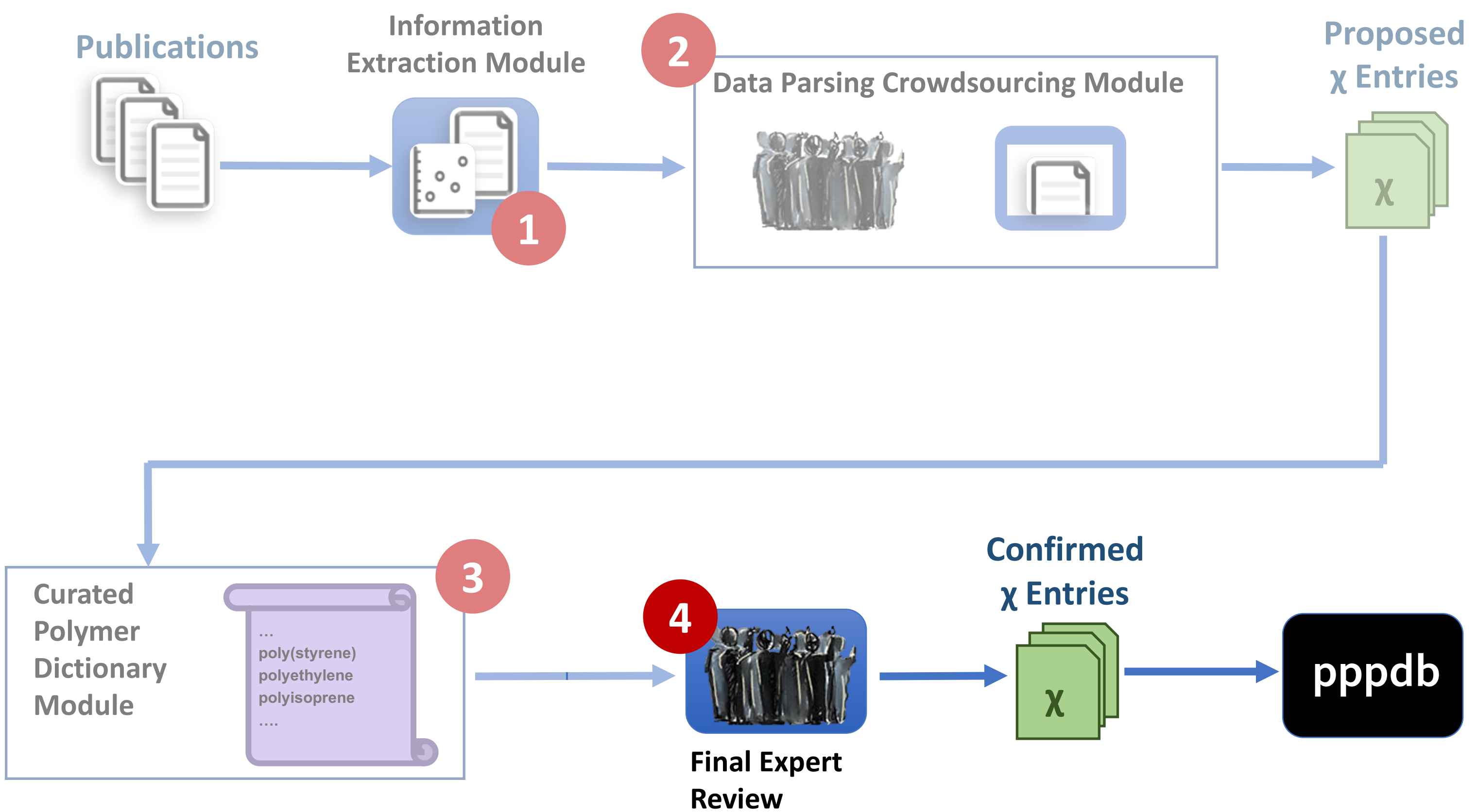
1B/C7H7N/c1-2-7-5-3-4-6-8-7/h2-6H,1H2/z101-1-8(1.2)

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-BBVYVPKKBA-N		
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-NBWNIOTOBA-N	poly(xylenyl ether)	
poly(3-(2-ethylhexyl)thiophene)	P3EHT	1B/C12H18S/c1-3-5-6-9(4-2)7-10-8-11-12(10)13-	UCZXRGRJTIHALN-HUZCROMBBA-N		
poly(3-butylthiophene)	P3BT	1B/C8H10S/c1-2-3-4-6-5-7-8(6)9-7/h5H,2-4H2,1H	MFKYCRZHETWXER-WCYUVGCXBA-N		
poly(3-dodecylthiophene)	P3DDT	1B/C16H26S/c1-2-3-4-5-6-7-8-9-10-11-12-14-13-	OWMLCFMCOUNBSQ-ORBAQIJUBA-N		
poly(3-hexylthiophene)	P3HT	1B/C10H14S/c1-2-3-4-5-6-8-7-9-10(8)11-9/h7H,2	VWCCCBCELGDMSR-RHPJWNLABA-N	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-XVBQSDAFBA-N		
poly(3-octylthiophene)	P3OT	1B/C12H18S/c1-2-3-4-5-6-7-8-10-9-11-12(10)13-	IXMZQXBZTSSNAG-HUZCROMBBA-N		
poly(4-acetoxystyrene)	P4AS	1B/C10H10O2/c1-3-9-4-6-10(7-5-9)12-8(2)11/h3-	JAMNSIXSLVPNLC-YUFRJTRMBA-N		
poly(4-hydroxystyrene)	PHS	1B/C8H8O/c1-2-7-3-5-8(9)6-4-7/h2-6,9H,1H2/z10	FUGYGGDSWSUORM-RRFBGHKIBA-N		
poly(4-tert-butylstyrene)	PtBS	1B/C12H16/c1-5-10-6-8-11(9-7-10)12(2,3)4/h5-9	QEDJMOONZLUIMC-KJZCOQPOBA-N		
poly(4-vinylbenzyltrimethylammonium chloride)	PVBTAmC	1B/C12H18N.ClH/c1-5-11-6-8-12(9-7-11)10-13(2,	TVXNKQRAZONMHJ-IPOBBXJVBA-M		
poly(4-vinylpyridine)	P4VP	1B/C7H7N/c1-2-7-3-5-8-6-4-7/h2-6H,1H2/z101-1-	KFDVPJUYSDEJTH-BBVYVPKKBA-N		
poly(6-methyl-ÃŽÂµ-caprolactone)	PMCL	1B/C7H12O2/c1-6-4-2-3-5-7(8)9-6/h6H,2-5H2,1H3	WZRNGGFHDMOCEA-PBJUEKBZBA-N		
poly(acrylic acid)	PAA	1B/C3H4O2/c1-2-3(4)5/h2H,1H2,(H,4,5)/z101-1-5	NIXOWILDQLNWCW-YZJLBCBGBA-N		
poly(allylamine hydrochloride)	PAH	1B/C3H7N.ClH/c1-2-3-4;/h2H,1,3-4H2;1H/z101-1-	MLGWTHRHHANFCC-XEQKNLBJBA-N		
poly(benzyl methacrylate)	PbnMA	1B/C11H12O2/c1-9(2)11(12)13-8-10-6-4-3-5-7-10	AOJOEFVRHOZDFN-PQSYHNEGBA-N		
poly(butyl acrylate)	PBA	1B/C7H12O2/c1-3-5-6-9-7(8)4-2/h4H,2-3,5-6H2,1	CQEYYJKEWSMYFG-YSNYPLFEBA-N	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-WVNZXXKVBA-N	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-PYASYJNHBA-N		
poly(cyclohexylethylene)	PCHE	1B/C8H14/c1-2-8-6-4-3-5-7-8/h2,8H,1,3-7H2/z10	LDLDYFCCDKENPD-BBVYVPKKBA-N		
poly(diethylhexyloxy-p-phenylenevinylene)		1B/C24H38O2/c1-5-9-11-19(7-3)17-25-23-15-22-1	QSFDVDZTDIATNP-FRAFILGZBA-N		
poly(diisoamyl itaconate)		1B/C15H26O4/c1-4-6-8-10-18-14(16)12-13(3)15(1	NJCKCUVRQPTKTF-IXYATHDYBA-N		
poly(ethyl ethylene)	PEE	1B/C4H8/c1-3-4-2/h3H,1,4H2,2H3/z101-1-4(1.3)	VXNZUUAINFGPBY-XMMBCNCSBA-N		
poly(ethyl methacrylate)	PEMA	1B/C6H10O2/c1-4-8-6(7)5(2)3/h2,4H2,1,3H3/z101	SUPCQIBBMFXVTL-ANYLYJCWBA-N		
poly(ethylene oxide)	PEO	1B/C2H4O/c1-2-3-1/h1-2H2/z101-1-3(1,2,1,3,2,3	IAYPIBMASNFSPL-GCGQHNKHBA-N	poly(ethylene glycol)	PEG
poly(ethylene-alt-propylene)	PEP	1B/C5H10/c1-5-3-2-4-5/h5H,2-4H2,1H3/z101-1-5(	BDJAEZRIGNCQBZ-HZQCOQKVBA-N	poly(ethylene-propylene); polymethylbutylene	
poly(ethylene-r-butylene)	PEB	1B/C4H8.C2H4/c1-3-4-2;1-2/h3H,1,4H2,2H3;1-2H2	WXCZUWHSJWOTRV-LWCKEHRHBA-N		
poly(ferrocenylmethylethylsilane)					
poly(hexafluoroisopropylidene diphthalic anhydride-alt-2	6FDA-TMPD	1B/C29H18F6N2O4/c1-11-12(2)22-14(4)13(3)21(11	UWVNHDDZPHOGRQ-AGENGVHNBA-N		
poly(hexene oxide)	PHO	1B/C2H3BrO/c3-2-1-4-2/h2H,1H2/z101-1-4(1,2,1,	XOOVDXMNWOFROX-UAKBPNCWBA-N		
poly(hydroxyethyl acrylate)		1B/C6H10O3/c1-5(2)6(8)9-4-3-7/h7H,1,3-4H2,2H3	WOBHKFSMXKNTIM-MWVPMSDBBA-N		
poly(Îµ-caprolactone)	PCL	1B/C6H10O2/c7-6-4-2-1-3-5-8-6/h1-5H2/z101-1-8	PAPBSGBWRJIAAV-CMRMDLKMBA-N		
poly(lactic acid)	PLA	1B/C3H4O2/c1-2-3(4)5-2/h2H,1H3/z101-1-5(2,3,2	YCHXNMPJFFEIJG-MGVHKUGVBA-N	poly(lactide); poly(lactic acid)	
poly(methacrylic acid)	PMAA	1B/C4H6O2/c1-3(2)4(5)6/h1H2,2H3,(H,5,6)/z101-	CERQOIWHTDAKMF-IKXXVMLBBA-N		
poly(methyl acrylate)	PMA	1B/C4H6O2/c1-3-4(5)6-2/h3H,1H2,2H3/z101-1-6(1	BAPJBEWLBFYGME-IKXXVMLBBA-N		
poly(methyl methacrylate)	PMMA	1B/C5H8O2/c1-4(2)5(6)7-3/h1H2,2-3H3/z101-1-7(	VVQNEPGJFQJSBK-KCAOANCDBA-N		
poly(n-hexyl methacrylate)	PHMA	1B/C10H18O2/c1-4-5-6-7-8-12-10(11)9(2)3/h2,4-	LNCPIMCVTKXXOY-XJFATUNXBA-N		
poly(N-isopropylacrylamide)	PNIPAM	1B/C6H11NO/c1-4-6(8)7-5(2)3/h4-5H,1H2,2-3H3,(	QNILTEGFHQSKFF-YWFYHYGJBA-N		
poly(n-pentyl methacrylate)	PnPMA	1B/C9H16O2/c1-4-5-6-7-11-9(10)8(2)3/h2,4-7H2,	GYDSPAVLTMAXHT-YVXONTJNBA-N		

Flory Huggins χ parameter



Final review and push to database

R.B. Tchoua, J. Qin, D.J. Audus, *et al. J. Chem. Educ.*, **93**, 1561-1568 (2016)
R.B. Tchoua, K. Chard, D.J. Audus, *et al. Procedia Comp. Sci.*, **80**, 386-397 (2016)

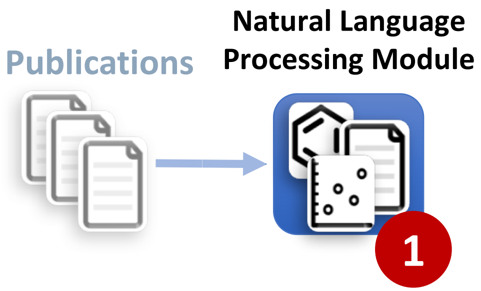
Glass transition temperature

Publications

6,090 articles from *Macromolecules*

R.B. Tchoua, K. Chard, D.J. Audus, *et al.* submitted to 13th IEEE International Conference on e-Science

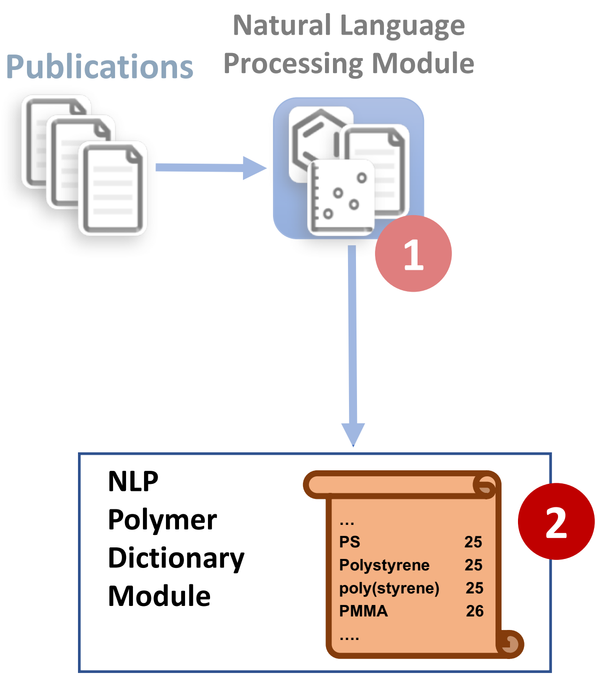
Glass transition temperature



Tries to find compound-T_g pairs automatically

R.B. Tchoua, K. Chard, D.J. Audus, *et al.* submitted to 13th IEEE International Conference on e-Science

Glass transition temperature



Automatically create a dictionary of polymers (only names) using "P" and "poly"

R.B. Tchoua, K. Chard, D.J. Audus, *et al.* submitted to 13th IEEE International Conference on e-Science

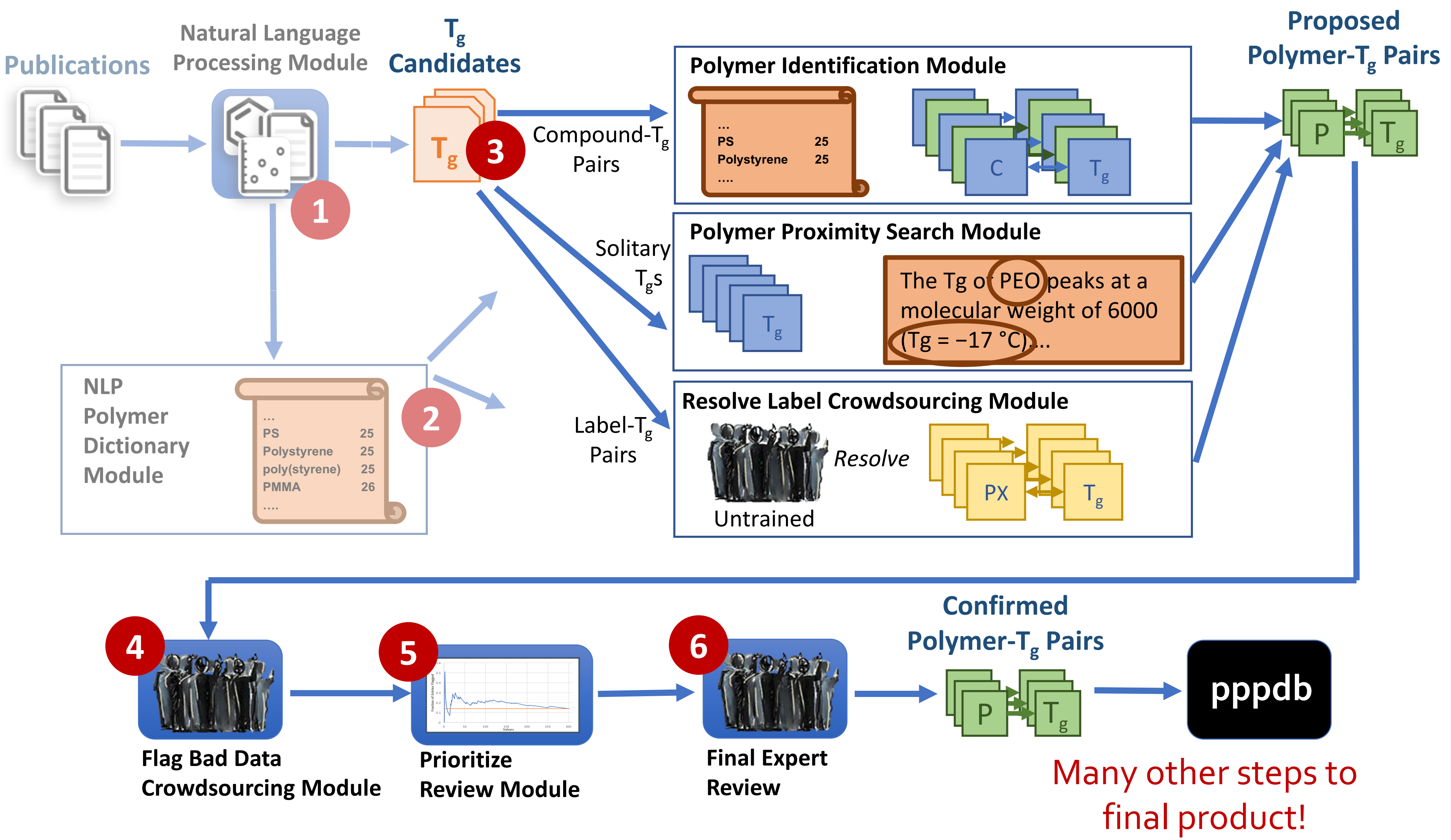
NLP Polymer Dictionary

Name	
Polystyrene	various forms
poly(styrene)	
polystyrene	
polystyrenes	
PS	
PSS	not plural of PS
polyimides	family names
polyolefin	
copolymer 10	labels not names
poly(2,4’-BF-a)	
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



R.B. Tchoua, K. Chard, D.J. Audus, *et al.* submitted to 13th IEEE International Conference on e-Science

The need for InChI

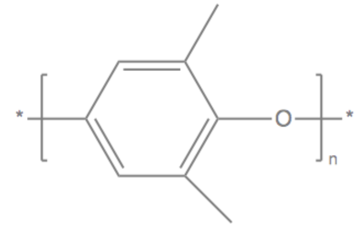
Multiple and trade names

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)

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

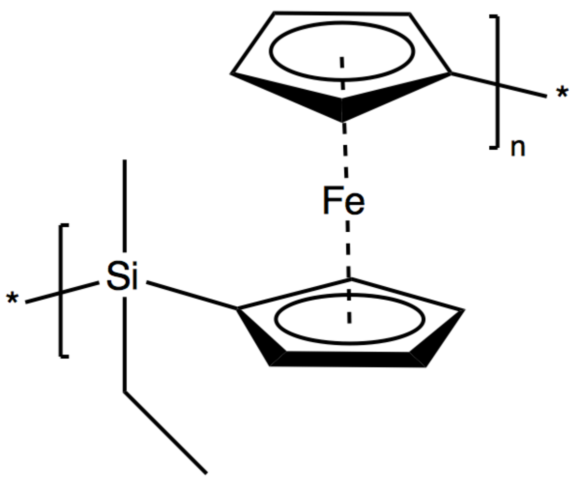
Input/output for machine learning

Natural Language Processing Module

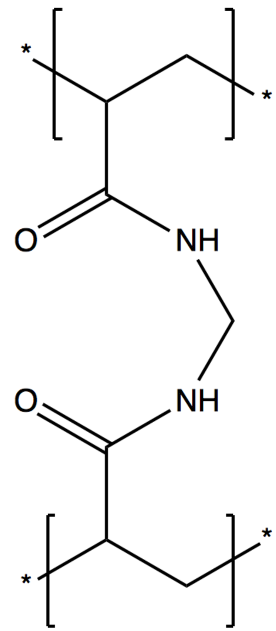
1B/C8H8O/c1-5-3-7-46(2)8(5)9-7/
h3-4H,12H3/z101-1-9(7,9,8,9)

Limitations of current InChI

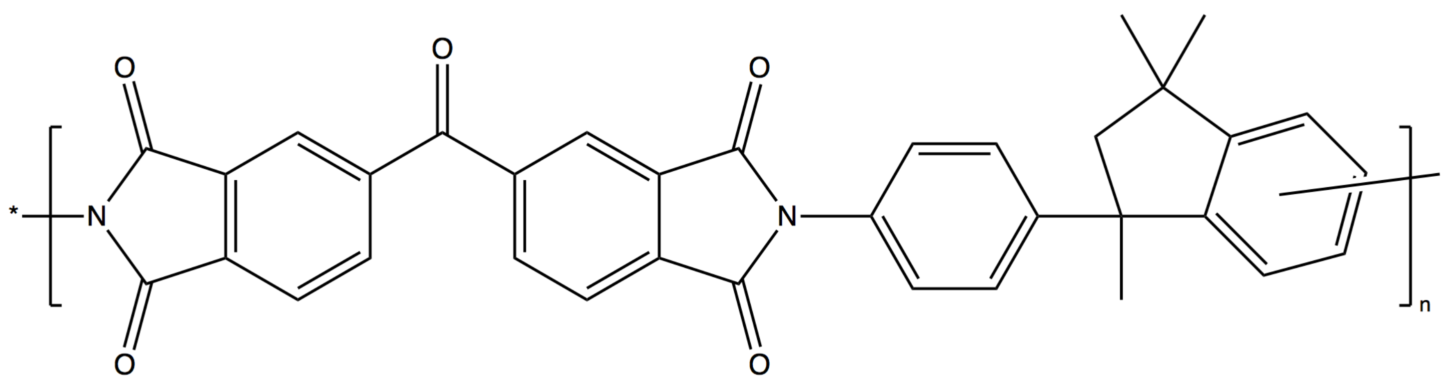
Organometallic



Branching / crosslinks



Markush



Conclusions and outlook

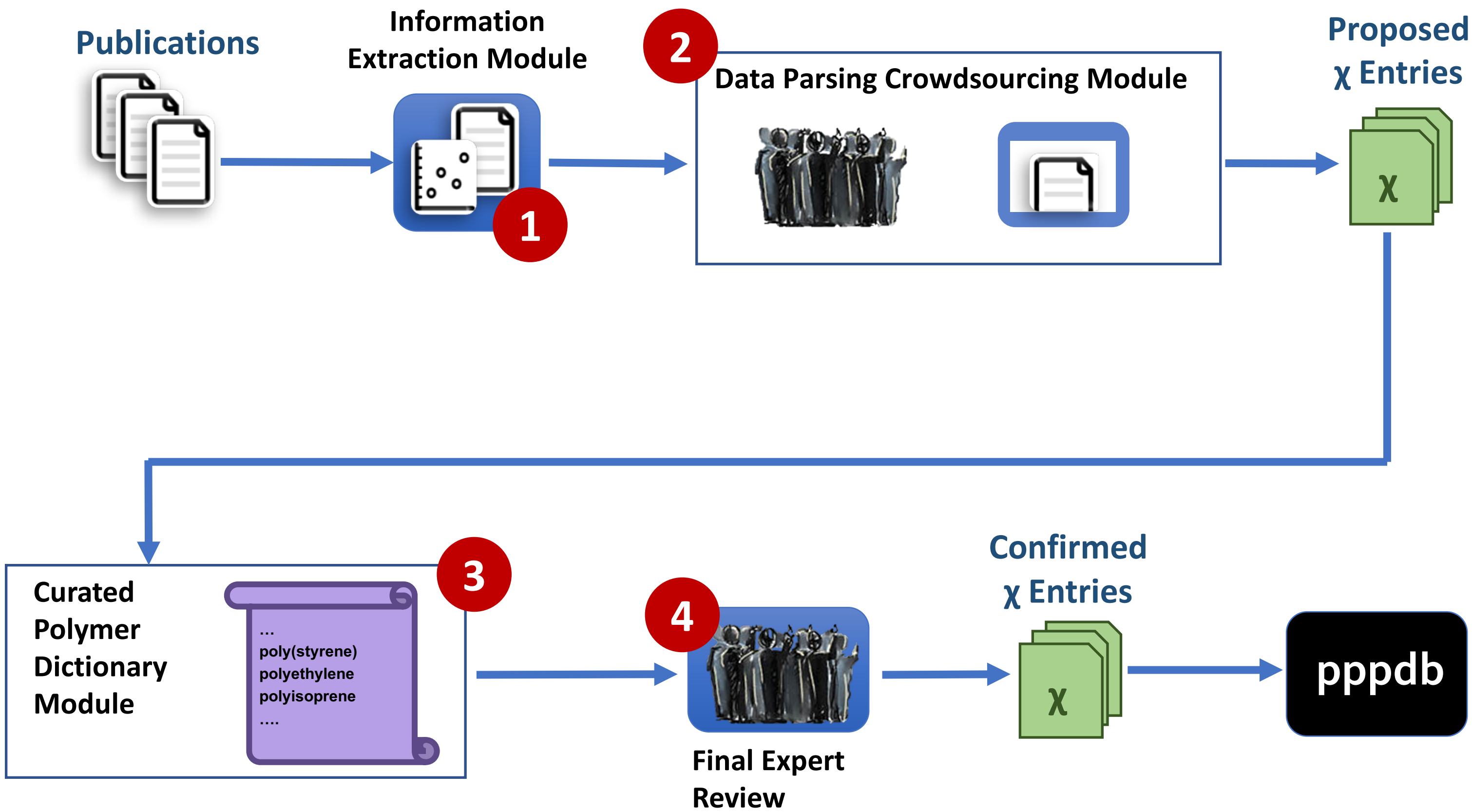
`http://pppdb.uchicago.edu`

263 χ 258 T_g

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

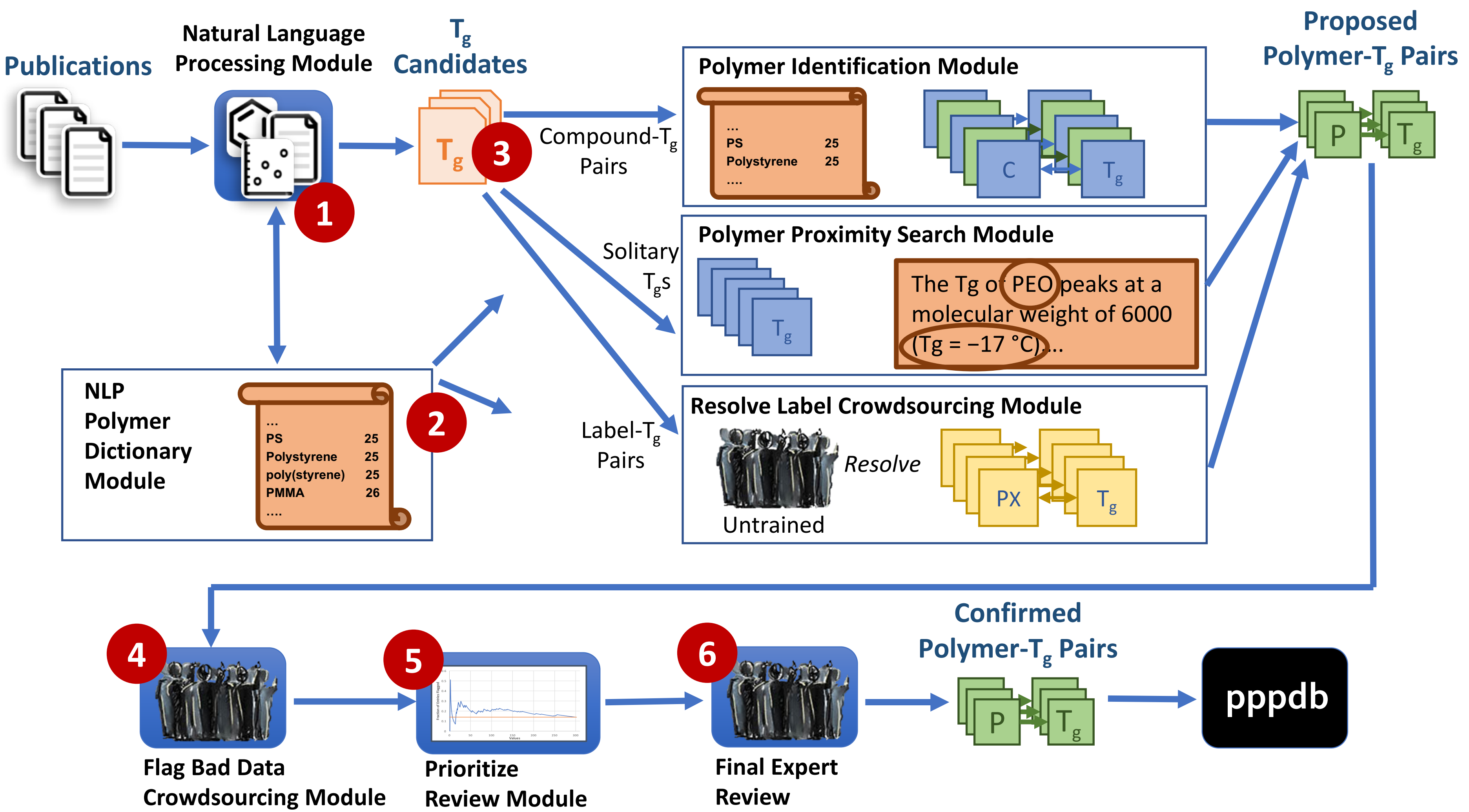
Advances still need for InChI
- Organometallics
- Branching / cross-links
- Markush

Flory Huggins χ parameter



R.B. Tchoua, J. Qin, D.J. Audus, *et al. J. Chem. Educ.*, **93**, 1561-1568 (2016)
R.B. Tchoua, K. Chard, D.J. Audus, *et al. Procedia Comp. Sci.*, **80**, 386-397 (2016)

Glass transition temperature



R.B. Tchoua, K. Chard, D.J. Audus, *et al.* submitted to 13th IEEE International Conference on e-Science