UChicago Research Data Inventory

Cesar Lema
cl4393@nyu.edu

Summer 2020

Contents

1 Directory of online data sources

A directory of online data source we are currently looking to get data data from.

1. repo Tapir database
2. url Polymer Gas Separation Membrane Database (CSIRO project)
3. url 1 url 2 Crow's Polymer property database
4. url MoleculeNet (sub module of DeepChem)
5. rdkit url RDKit rdkit.Chem.Descriptors module
6. Additional rdkit module: rdkit.Chem.rdMolDescriptors, rdkit.Chem.rdPartialCharges, rdkit.Chem.EState, rdkit.Chem.ChemUtils.DescriptorUtilities, rdkit.Chem, GraphDescriptors, MolSurf, Lipinski, Fragments, Crippen, Descriptors3D, rdkit.ML.Descriptors.MoleculeDescriptors

Journal papers to look into for data

1. arXiv "Designing exceptional gas-separation polymer membranes using machine learning"
2. article "Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology" data is available part of rdkit
3. book An introduction to cheminformatics
4. book rdkit documentation book
5. book Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes:Descriptors from Molecular Geometry
6. about finger prints

2 Data available in Sources

This section gives a description of data available in the online sources listed in the previous section.

2.1 Tapir database

The data is associated with the Tapir database github repo and consists of a collection of polymer thermo-physical property values for different polymers. Note additional polymer descriptors can be added (collected from RDKit) using the Repo or the polymerDataManager class.

Number of polymers	660
Number of properties	14

Remark: Raw data is saved at directory ”/Users/cesarlema/Developer/uchicago reu/code/Uchicago-Research/polymer data/data” or github repo url

Available thermo-physical properties in the data are

1. polymer_name
2. smiles
3. molar_volume
4. density
5. solubility_parameter
6. molar_cohesive_energy
7. glass_transition_temperature
8. molar_heat_capacity
9. entanglement_molecular_weight
10. refraction_index
11. thermal_expansion_coefficient
12. repeat_unit_weight
13. waals_volume
14. inchi

Available polymer descriptors are the same as RDKit available descriptors. See RDKit section for more details.

2.1.1 Structure of data:

The data is structured as a 665x14 csv file.

2.2 Polymer Gas Separation Membrane Database (CSIRO project)

The data from the Polymer Gas Separation Membrane Database is a collection of polymer membrane gas permeability values for specific gas and polymer combinations.

Number of polymers	≈ 1470
Number of gases	15
Total number of permeability values	

Remark: Raw data is saved at directory "/Users/cesarlema/Developer/uchicago reu/code/Uchicago-Research/polymer data/data" or github repo url

Polymer membrane gas separation data is documented from published articles for the 15 gasses below:

1. He
2. H2
3. O2
4. N2
5. CO2
6. CH4
7. C2H4
8. C2H6
9. C3H6
10. C3H8
11. C4H8
12. n-C4H10
13. CF4
14. C2F6
15. C3F8

2.2.1 Structure of data:

The data is structured as a 1502x21 csv file. The 1502 rows consists of mainly different polymers. The first few rows are data used to plot Robeson limit upper bounds. The columns consist of gas species the permeability values are for and additional information.

A description of each column is listed below (from left to right in the table).

1. Category: Polymer type
2. Brief Description:
3. Extended Description: Polymer

4. Data: He (Barrer)
5. Data: H2 (Barrer)
6. Data: O2 (Barrer)
7. Data: N2 (Barrer)
8. Data: CO2 (Barrer)
9. Data: CH4 (Barrer)
10. Data: C2H4 (Barrer)
11. Data: C2H6 (Barrer)
12. Data: C3H6 (Barrer)
13. Data: C3H8 (Barrer)
14. Data: C4H8 (Barrer)
15. Data: n-C4H10 (Barrer)
16. Data: CF4 (Barrer)
17. Data: C2F6 (Barrer)
18. Data: C3F8 (Barrer)
19. In Reference Data Location:
20. Reference Name:
21. Reference URL:

2.3 Crow's Polymer property database

The Crow's Polymer property database consists of articles on polymer physics (with sparse references to data sets for the topic of the respective article) and pages with identifiers and thermo-physical properties for a specific polymer.

Available data and thermo-physical properties for each polymer include:

1. Names and identifiers of polymers
2. Identifiers of monomer(s)
3. Thermo-Physical Properties: Experimental / Literature Data
4. Thermo-Physical Properties: Calculated Data

2.3.1 Structure of data:

The thermo-physical Properties of specific polymers are in the encyclopedia style website. The polymer type and name are used as indices and specific pages are dedicated to each that contain the data.

Additional resources are sparsely linked in the polymer articles.

Remark: In the "Barrier properties of polymers" and "Polymer solubility and solubility parameter" polymer physics articles *moisture vapor and oxygene transmission rates* data for some polymers are listed and *Solubility Parameters for Homopolymers* is linked in each article respectively.

2.4 MoleculeNet

MoleculeNet is a benchmark specially designed for testing machine learning methods of molecular properties. The work curates a number of dataset collections. The datasets are integrated as parts of the open source DeepChem package(MIT license).

Quantum Mechanical datasets include:

1. QM7:
   is a subset of GDB-13 (a database of nearly 1 billion stable and synthetically accessible organic molecules) containing up to 7 heavy atoms C, N, O, and S.
2. QM8:
   the dataset used in a study on modeling quantum mechanical calculations of electronic spectra and excited state energy of small molecules. Multiple methods, including time-dependent density functional theories (TDDFT) and second-order approximate coupled-cluster (CC2), are applied to a collection of molecules that include up to eight heavy atoms (also a subset of the GDB-17 database).
3. QM9:
   a comprehensive dataset that provides geometric, energetic, electronic and thermodynamic properties for a subset of GDB-17 database, comprising 134 thousand stable organic molecules with up to 9 heavy atoms.

Remark: Info on the GDB-17 database:
"To better define the unknown chemical space, we have enumerated 166.4 billion molecules of up to 17 atoms of C, N, O, S, and halogens forming the chemical universe database GDB-17, covering a size range containing many drugs and typical for lead compounds. GDB-17 contains millions of isomers of known drugs, including analogs with high shape similarity to the parent drug. "

2.4.1 Structure of data:

MolculeNet consists of multiple datasets for different levels of physics. They are available to download here. The data is clean and ready to use.

2.5 rdkit.Chem.Descriptors Module

This RDKit module can give various molecule descriptors.

Available descriptors implemented as methods of the descriptors module are given below (Note these descriptors were collected from the attributes of the RDKit.Chem.Descriptor module and are the names of its methods that compute descriptors with most taking in a mol instance and outputing the corresponding value):

1. MaxEStateIndex
2. MinEStateIndex
3. MaxAbsEStateIndex
4. MinAbsEStateIndex
5. qed
6. MolWt
7. HeavyAtomMolWt
8. ExactMolWt
9. NumValenceElectrons
10. NumRadicalElectrons
11. MaxPartialCharge
12. MinPartialCharge
13. MaxAbsPartialCharge
14. MinAbsPartialCharge
15. FpDensityMorgan1
16. FpDensityMorgan2
17. FpDensityMorgan3
18. BalabanJ
19. BertzCT
20. Chi0
21. Chi0n
22. Chi0v
23. Chi1
24. Chi1n
25. Chi1v
26. Chi2n
27. Chi2v
28. Chi3n

29. Chi3v
30. Chi4n
31. Chi4v
32. HallKierAlpha
33. Ipc
34. Kappa1
35. Kappa2
36. Kappa3
37. LabuteASA
38. PEOE_VSA1
39. PEOE_VSA10
40. PEOE_VSA11
41. PEOE_VSA12
42. PEOE_VSA13
43. PEOE_VSA14
44. PEOE_VSA2
45. PEOE_VSA3
46. PEOE_VSA4
47. PEOE_VSA5
48. PEOE_VSA6
49. PEOE_VSA7
50. PEOE_VSA8
51. PEOE_VSA9
52. SMR_VSA1
53. SMR_VSA10
54. SMR_VSA2
55. SMR_VSA3
56. SMR_VSA4
57. SMR_VSA5
58. SMR_VSA6
59. SMR_VSA7
60. SMR_VSA8
61. SMR_VSA9
62. SlogP_VSA1

63. SlogP_VSA10

64. SlogP_VSA11

65. SlogP_VSA12

66. SlogP_VSA2

67. SlogP_VSA3

68. SlogP_VSA4

69. SlogP_VSA5

70. SlogP_VSA6

71. SlogP_VSA7

72. SlogP_VSA8

73. SlogP_VSA9

74. TPSA

75. EState_VSA1

76. EState_VSA10

77. EState_VSA11

78. EState_VSA2

79. EState_VSA3

80. EState_VSA4

81. EState_VSA5

82. EState_VSA6

83. EState_VSA7

84. EState_VSA8

85. EState_VSA9

86. VSA_EState1

87. VSA_EState10

88. VSA_EState2

89. VSA_EState3

90. VSA_EState4

91. VSA_EState5

92. VSA_EState6

93. VSA_EState7

94. VSA_EState8

95. VSA_EState9

96. FractionCSP3

97. HeavyAtomCount
98. NHOHCount
99. NOCount
100. NumAliphaticCarbocycles
101. NumAliphaticHeterocycles
102. NumAliphaticRings
103. NumAromaticCarbocycles
104. NumAromaticHeterocycles
105. NumAromaticRings
106. NumHAcceptors
107. NumHDonors
108. NumHeteroatoms
109. NumRotatableBonds
110. NumSaturatedCarbocycles
111. NumSaturatedHeterocycles
112. NumSaturatedRings
113. RingCount
114. MolLogP
115. MolMR
116. fr_Al_COO
117. fr_Al_OH
118. fr_Al_OH_noTert
119. fr_ArN
120. fr_Ar_COO
121. fr_Ar_N
122. fr_Ar_NH
123. fr_Ar_OH
124. fr_COO
125. fr_COO2
126. fr_C_O
127. fr_C_O_noCOO
128. fr_C_S
129. fr_HOCCN
130. fr_Imine

131. fr_NH0
132. fr_NH1
133. fr_NH2
134. fr_N_O
135. fr_Ndealkylation1
136. fr_Ndealkylation2
137. fr_Nhpyrrole
138. fr_SH
139. fr_aldehyde
140. fr_alkyl_carbamate
141. fr_alkyl_halide
142. fr_allylic_oxid
143. fr_amide
144. fr_amidine
145. fr_aniline
146. fr_aryl_methyl
147. fr_azide
148. fr_azo
149. fr_barbitur
150. fr_benzene
151. fr_benzodiazepine
152. fr_bicyclic
153. fr_diazo
154. fr_dihydropyridine
155. fr_epoxide
156. fr_ester
157. fr_ether
158. fr_furan
159. fr_guanido
160. fr_halogen
161. fr_hdrzine
162. fr_hdrzone
163. fr_imidazole
164. fr_imide

165. fr_isocyan
166. fr_isothiocyan
167. fr_ketone
168. fr_ketone_Topliss
169. fr_lactam
170. fr_lactone
171. fr_methoxy
172. fr_morpholine
173. fr_nitrile
174. fr_nitro
175. fr_nitro_arom
176. fr_nitro_arom_nonortho
177. fr_nitroso
178. fr_oxazole
179. fr_oxime
180. fr_para_hydroxylation
181. fr_phenol
182. fr_phenol_noOrthoHbond
183. fr_phos_acid
184. fr_phos_ester
185. fr_piperdine
186. fr_piperzine
187. fr_priamide
188. fr_prisulfonamd
189. fr_pyridine
190. fr_quatN
191. fr_sulfide
192. fr_sulfonamd
193. fr_sulfone
194. fr_term_acetylene
195. fr_tetrazole
196. fr_thiazole
197. fr_thiocyan
198. fr_thiophene
199. fr_unbrch_alkane
200. fr_urea

2.5.1 Structure of data:

The Descriptors module implements these descriptors as methods with the name of the methods listed above. The method takes in an rd kit mol instance and returns its corresponding value.

2.6 Additional RDKit modules

The rdkit Descriptors module is the centralized location for all chemical descriptors however other modules contain additional descriptors.

Some Additional submodules of rdkit that can calculate descriptors

1. rdkit.Chem.rdMolDescriptors:
   Module containing functions to compute molecular descriptors. ”Low level” and used by many of the other modules to compute descriptor values.
2. rdkit.Chem.rdPartialCharges
   Module containing functions to set [compute] partial charges - currently Gasteiger Charges
3. rdkit.Chem.AllChem:
   Import all RDKit chemistry modules, contains many more helpful functions.
4. rdkit.Chem.GraphDescriptors: (Most are in descriptors module)
   Calculation of topological/topochemical descriptors.
5. rdkit.Chem.MolSurf:
   Exposes functionality for MOE-like approximate molecular surface area descriptors
6. rdkit.Chem.Lipinski:
   Calculation of Lipinski parameters for molecules
7. rdkit.Chem.Crippen:
   Atom-based calculation of LogP and MR using Crippen’s approach
8. rdkit.Chem.Descriptors3D:
   Descriptors derived from a molecule’s 3D structure
9. rdkit.Chem.EState:
   A module for Kier and Hall’s EState Descriptors. Defined by the article “Molecular Structure Description: The Electrotopological State”
10. rdkit.ML.Descriptors.MoleculeDescriptors:
    High level interface for descriptors from rdkit.Chem.Descriptors. Consists of a class to calculate descriptors from methods available in the rdkit.Chem.Descriptors module.