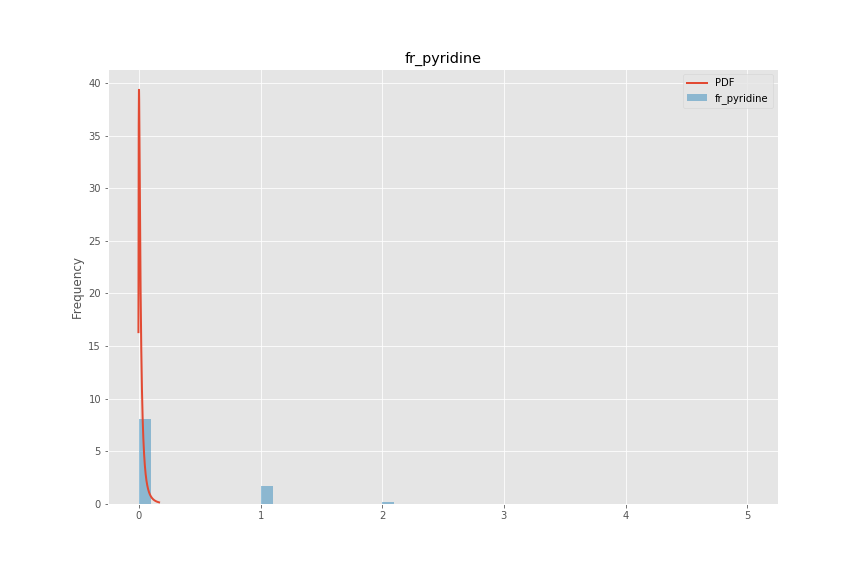
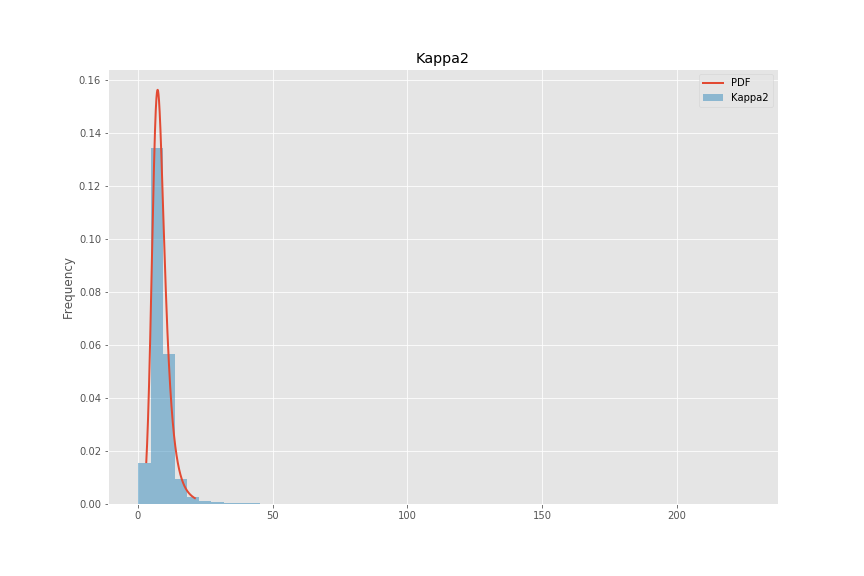
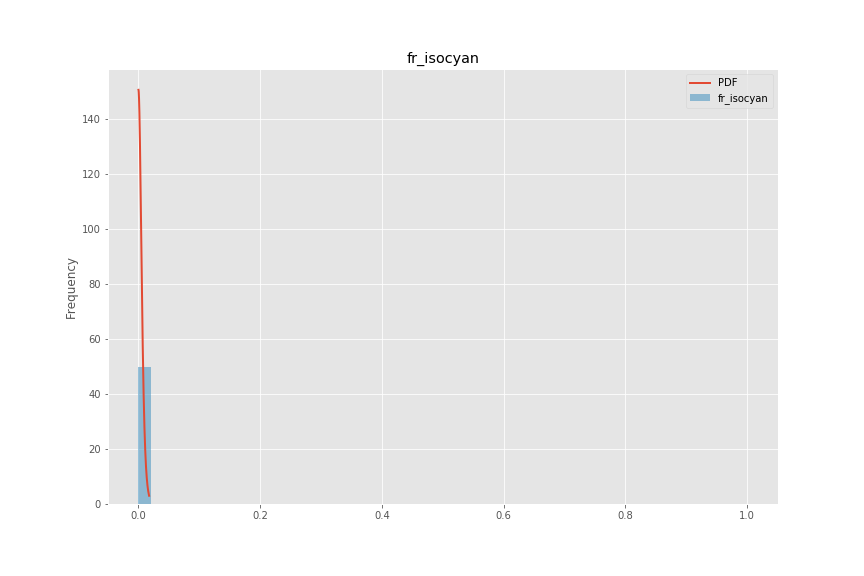
Methods

Feature Normalization

Features used in conjunction with the graph convolution are computed with the RDKit toolkit[X]. The neural network architecture requires that the features are appropriately scaled to prevent features with large ranges dominating smaller ranged features as well as preventing issues where features in the training set are not drawn from the same sample distribution as the testing sets. prevent these issues, a large sample of molecules was used to fit cumulative density function (CDF) to all features. CDF were used as opposed to simpler scaling algorithms mainly because CDFs have the useful property that each value has the same meaning: the percentage of the population observed below the raw feature value. Min-max scaling can be easily biased with outliers and Z-score scaling assumes a normal distribution which is most often not the case for chemical features, especially if they are based on counts.

The CDFs were fit to a sampling of 100k compounds from the Novartis internal catalog using the distributions available in the scikit learn package[X], a sampling of which can be seen in Figure X. scikit learn was used primarily due to the simplicity of fitting and the final application, however more complicated techniques can be used in the future to fit to empirical CDFs such as finding the best fit general logistic function[X] which has been shown to be successful for other biological datasets. The sample distributions and fits are available in the supplementary material. No review was taken to remove odd distributions; for example, azides are hazardous and rarely used outside of a few specific reactions. This is reflected in the fr\_azide distribution. As such, since the sample data was primarily used for chemical screening against biological targets, the distribution used here may not accurately reflect the distribution of reagents used for chemical synthesis.

Future work should include detecting ill-fitting distributions, such as for fr\_pyridine as seen in Figure X, and using alternative methods, such as empirical distributions.



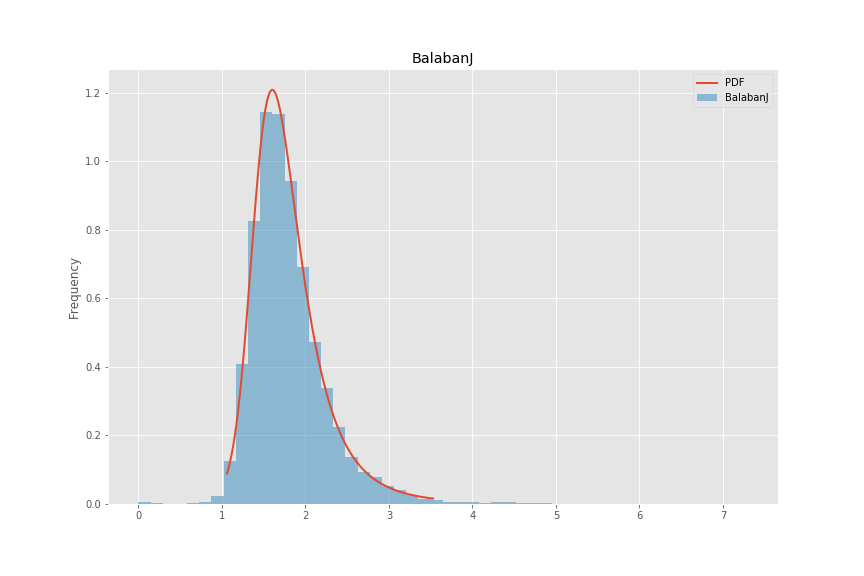


Figure 1 Four example distributions fit to a random sample of 100,000 of compounds used for biological screening in Novartis. Note that some distributions for discrete calculations, such as fr\_pyridine are not fit especially well. This is an active area for improvement.

List of descriptors used:

1. BalabanJ
2. BertzCT
3. Chi0
4. Chi0n
5. Chi0v
6. Chi1
7. Chi1n
8. Chi1v
9. Chi2n
10. Chi2v
11. Chi3n
12. Chi3v
13. Chi4n
14. Chi4v
15. EState\_VSA1
16. EState\_VSA10
17. EState\_VSA11
18. EState\_VSA2
19. EState\_VSA3
20. EState\_VSA4
21. EState\_VSA5
22. EState\_VSA6
23. EState\_VSA7
24. EState\_VSA8
25. EState\_VSA9
26. ExactMolWt
27. FpDensityMorgan1
28. FpDensityMorgan2
29. FpDensityMorgan3
30. FractionCSP3
31. HallKierAlpha
32. HeavyAtomCount
33. HeavyAtomMolWt
34. Ipc
35. Kappa1
36. Kappa2
37. Kappa3
38. LabuteASA
39. MaxAbsEStateIndex
40. MaxAbsPartialCharge
41. MaxEStateIndex
42. MaxPartialCharge
43. MinAbsEStateIndex
44. MinAbsPartialCharge
45. MinEStateIndex
46. MinPartialCharge
47. MolLogP
48. MolMR
49. MolWt
50. NHOHCount
51. NOCount
52. NumAliphaticCarbocycles
53. NumAliphaticHeterocycles
54. NumAliphaticRings
55. NumAromaticCarbocycles
56. NumAromaticHeterocycles
57. NumAromaticRings
58. NumHAcceptors
59. NumHDonors
60. NumHeteroatoms
61. NumRadicalElectrons
62. NumRotatableBonds
63. NumSaturatedCarbocycles
64. NumSaturatedHeterocycles
65. NumSaturatedRings
66. NumValenceElectrons
67. PEOE\_VSA1
68. PEOE\_VSA10
69. PEOE\_VSA11
70. PEOE\_VSA12
71. PEOE\_VSA13
72. PEOE\_VSA14
73. PEOE\_VSA2
74. PEOE\_VSA3
75. PEOE\_VSA4
76. PEOE\_VSA5
77. PEOE\_VSA6
78. PEOE\_VSA7
79. PEOE\_VSA8
80. PEOE\_VSA9
81. RingCount
82. SMR\_VSA1
83. SMR\_VSA10
84. SMR\_VSA2
85. SMR\_VSA3
86. SMR\_VSA4
87. SMR\_VSA5
88. SMR\_VSA6
89. SMR\_VSA7
90. SMR\_VSA8
91. SMR\_VSA9
92. SlogP\_VSA1
93. SlogP\_VSA10
94. SlogP\_VSA11
95. SlogP\_VSA12
96. SlogP\_VSA2
97. SlogP\_VSA3
98. SlogP\_VSA4
99. SlogP\_VSA5
100. SlogP\_VSA6
101. SlogP\_VSA7
102. SlogP\_VSA8
103. SlogP\_VSA9
104. TPSA
105. VSA\_EState1
106. VSA\_EState10
107. VSA\_EState2
108. VSA\_EState3
109. VSA\_EState4
110. VSA\_EState5
111. VSA\_EState6
112. VSA\_EState7
113. VSA\_EState8
114. VSA\_EState9
115. fr\_Al\_COO
116. fr\_Al\_OH
117. fr\_Al\_OH\_noTert
118. fr\_ArN
119. fr\_Ar\_COO
120. fr\_Ar\_N
121. fr\_Ar\_NH
122. fr\_Ar\_OH
123. fr\_COO
124. fr\_COO2
125. fr\_C\_O
126. fr\_C\_O\_noCOO
127. fr\_C\_S
128. fr\_HOCCN
129. fr\_Imine
130. fr\_NH0
131. fr\_NH1
132. fr\_NH2
133. fr\_N\_O
134. fr\_Ndealkylation1
135. fr\_Ndealkylation2
136. fr\_Nhpyrrole
137. fr\_SH
138. fr\_aldehyde
139. fr\_alkyl\_carbamate
140. fr\_alkyl\_halide
141. fr\_allylic\_oxid
142. fr\_amide
143. fr\_amidine
144. fr\_aniline
145. fr\_aryl\_methyl
146. fr\_azide
147. fr\_azo
148. fr\_barbitur
149. fr\_benzene
150. fr\_benzodiazepine
151. fr\_bicyclic
152. fr\_diazo
153. fr\_dihydropyridine
154. fr\_epoxide
155. fr\_ester
156. fr\_ether
157. fr\_furan
158. fr\_guanido
159. fr\_halogen
160. fr\_hdrzine
161. fr\_hdrzone
162. fr\_imidazole
163. fr\_imide
164. fr\_isocyan
165. fr\_isothiocyan
166. fr\_ketone
167. fr\_ketone\_Topliss
168. fr\_lactam
169. fr\_lactone
170. fr\_methoxy
171. fr\_morpholine
172. fr\_nitrile
173. fr\_nitro
174. fr\_nitro\_arom
175. fr\_nitro\_arom\_nonortho
176. fr\_nitroso
177. fr\_oxazole
178. fr\_oxime
179. fr\_para\_hydroxylation
180. fr\_phenol
181. fr\_phenol\_noOrthoHbond
182. fr\_phos\_acid
183. fr\_phos\_ester
184. fr\_piperdine
185. fr\_piperzine
186. fr\_priamide
187. fr\_prisulfonamd
188. fr\_pyridine
189. fr\_quatN
190. fr\_sulfide
191. fr\_sulfonamd
192. fr\_sulfone
193. fr\_term\_acetylene
194. fr\_tetrazole
195. fr\_thiazole
196. fr\_thiocyan
197. fr\_thiophene
198. fr\_unbrch\_alkane
199. fr\_urea
200. qed

[X] RDKit 2018.09.01

[X] scikit learn version …