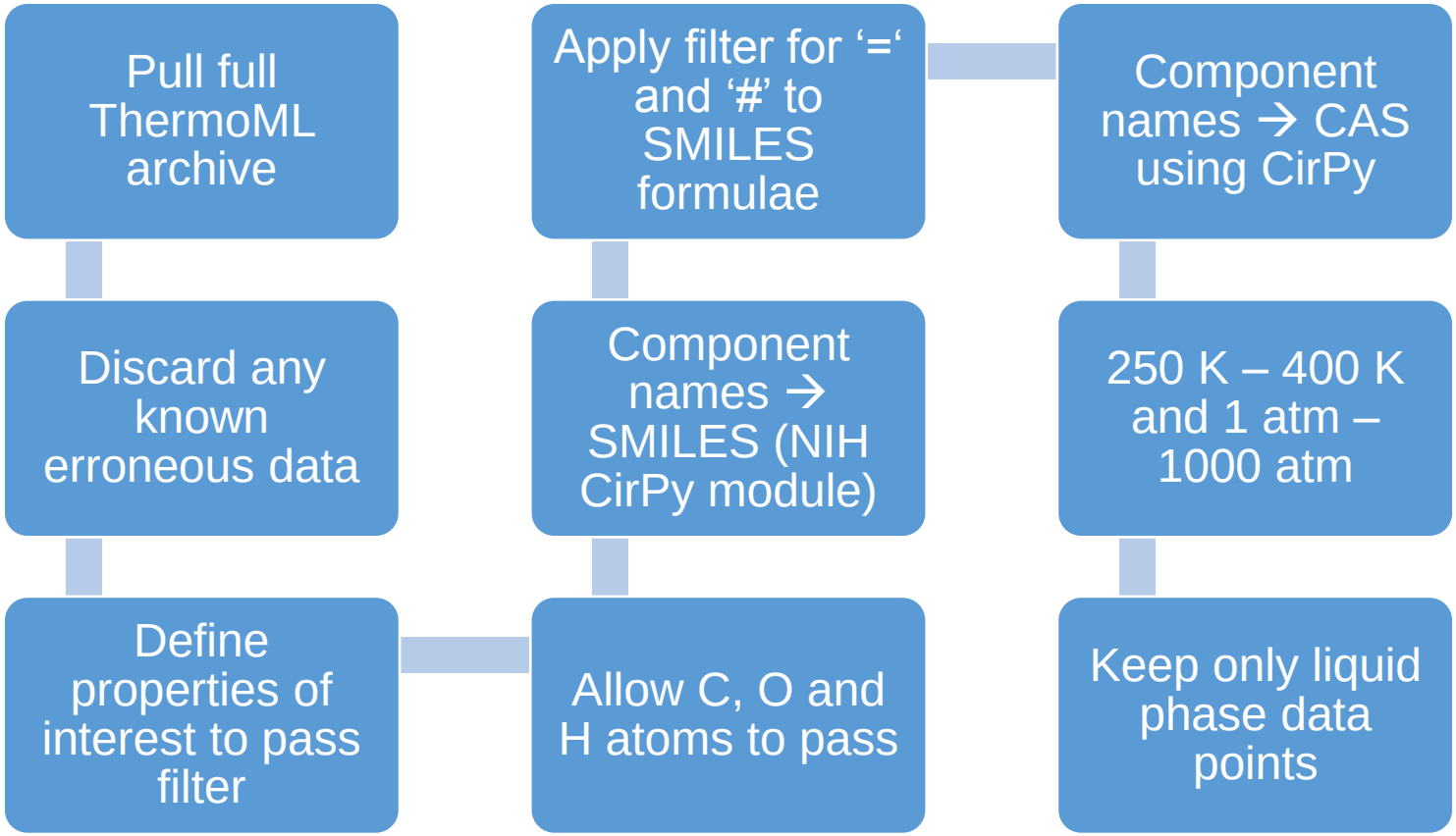
ThermoPyL can be used to sort and filter ThermoML data

- ThermoPyL
  - Chodera lab @MSKCC
  - Organizes data into Pandas dataframe
  - Each row of dataframe is single data point
- Very useful for extracting properties of interest
  - Pure solvent: ρ_{mass}, speed of sound, C_p, dielectric constant, V_{molar} and H_{molar}
  - Binary mix: ρ_{mass}, speed of sound, H_{excess}, dielectric constant, V_{excess}, $C_{p,excess}$, γ

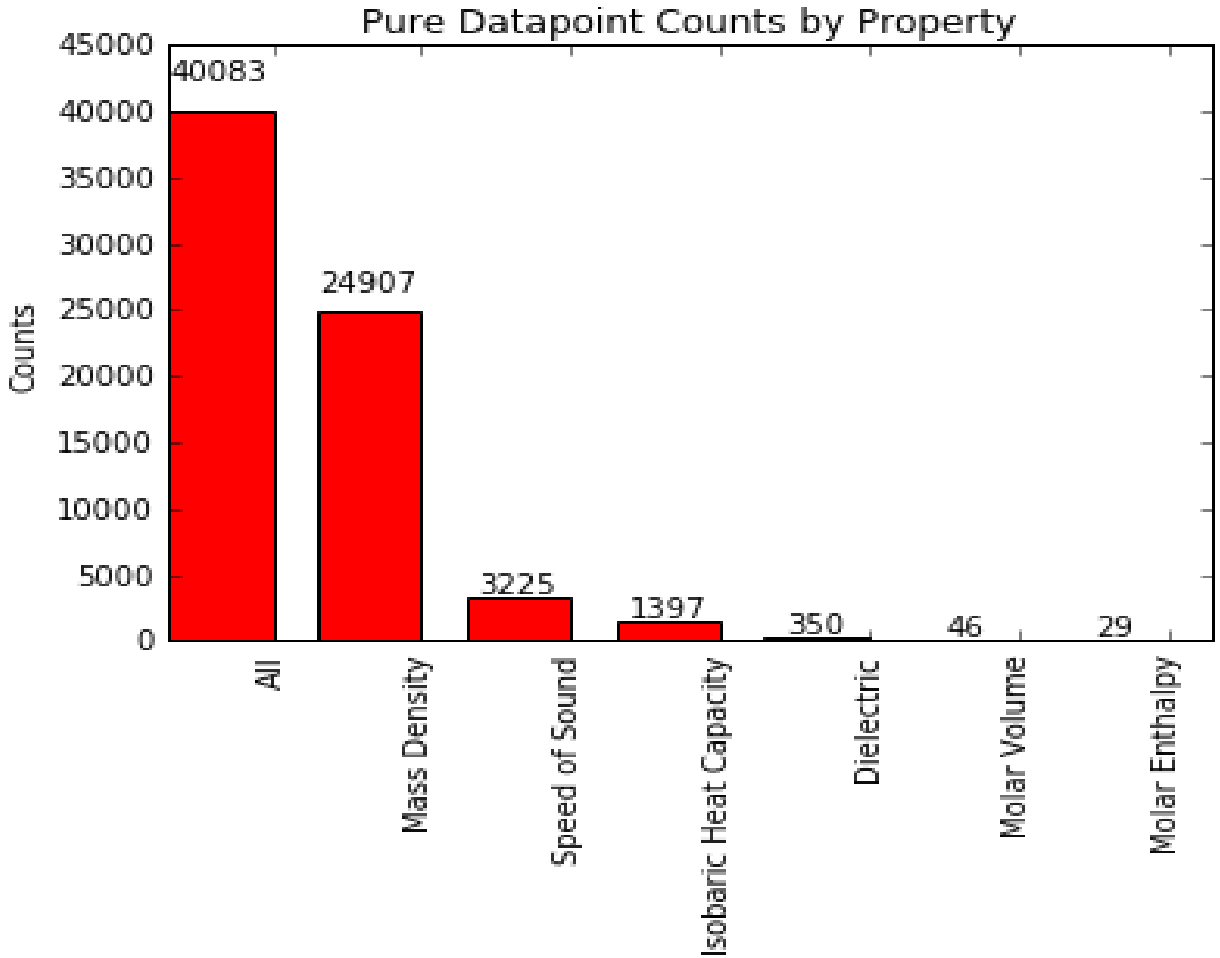
Specific search criterion are applied to narrow the data set



Separation by property allows for further analyses

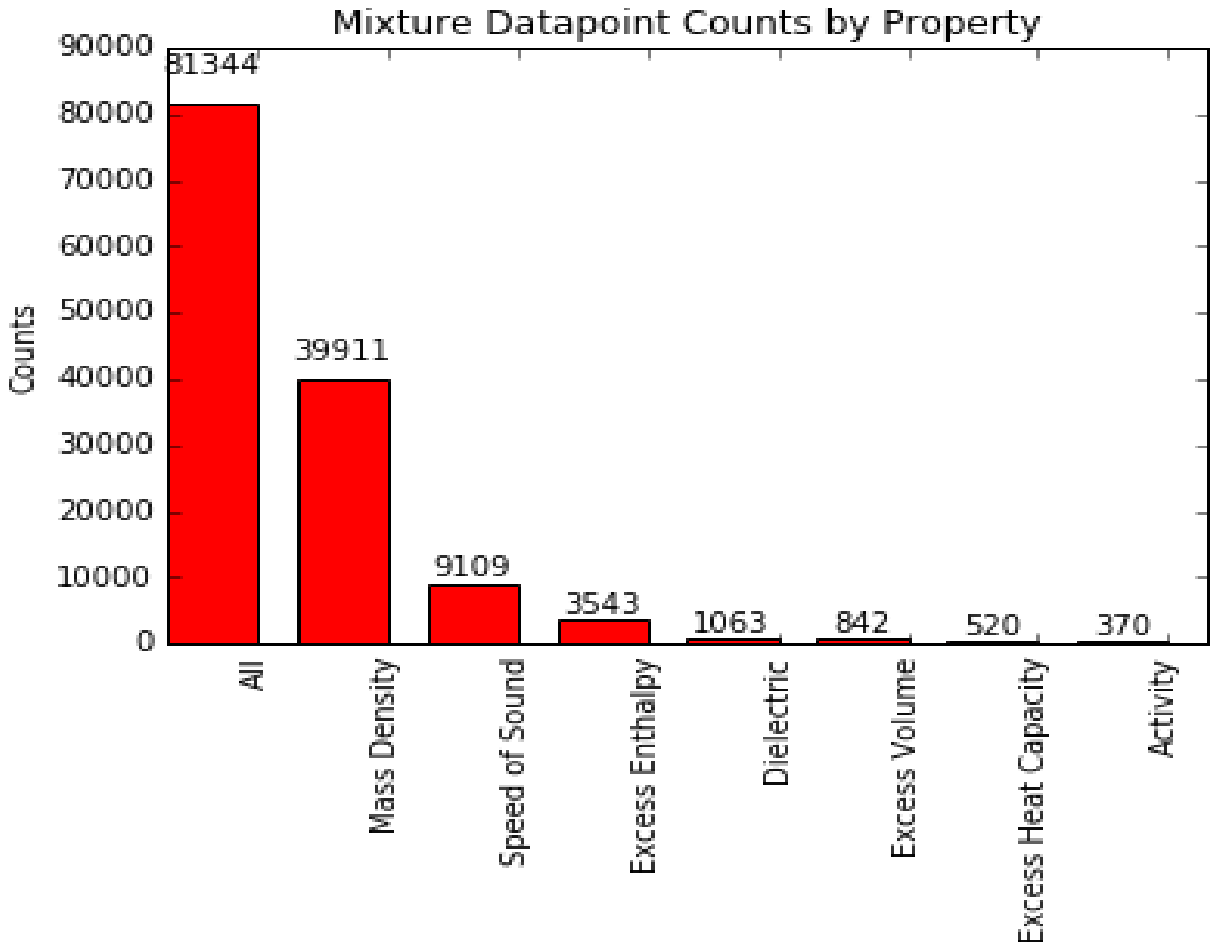
- The final large dataframes are separated into subframes by property of interest
  - Data that has no associated uncertainty is removed
  - Only applied to the subframes
- Counts by component and journal article are created for all dataframes
- Everything then saved as text .csv

Pure datacounts illustrate the "clustering" problem



Plot shows the extremely lopsided distribution of pure data

Binary datacounts illustrate the same issue



Plot shows the same lopsided distribution of binary data

Initial set of molecules chosen based on pure density data

Matrix of binary mixture data for a small number of components

	Water	Heptane	Octane	Decane	Methanol	Ethanol	1-propanol	1-butanol	Benzene	Toluene
Water		0	0	0	1117	1050	713	0	0	0
Heptane	0		0	750	0	1664	378	261	0	151
Octane	0	0		750	0	720	77	112	1080	111
Decane	0	750	750		0	720	17	126	0	10
Methanol	1117	0	0	0		469	105	105	248	24
Ethanol	1050	1664	720	720	469		74	0	0	805
1-propanol	713	378	77	17	105	74		0	0	986
1-butanol	0	261	112	126	105	0	0		189	214
Benzene	0	0	1080	0	248	0	0	189		25
Toluene	0	151	111	10	24	805	986	214	25	