MATH655 Final Presentation: Statistical analysis of the thermochemistry of polycyclic aromatic hydrocarbons (PAHs)

By Brendan Lucas

An Overview of the Proposal

- Does it make sense to interpret entropy as a measure of disorder in thermochemistry?
- Some theory:
  - Entropy: $S = k \log W$
  - Enthalpy: $H = E + PV$
  - Gibbs free energy: $G = H - TS$
  - Free entropy: $G/T = - k \log Z$
- Polycyclic aromatic hydrocarbons:
  - They are highly orderly and there is a lot of data about them.
  - Can we make more concrete statements about what thermodynamic correspondence to molecular structure than "ordered" or "disordered"?
  - The dataset approaches graphene in the limit.

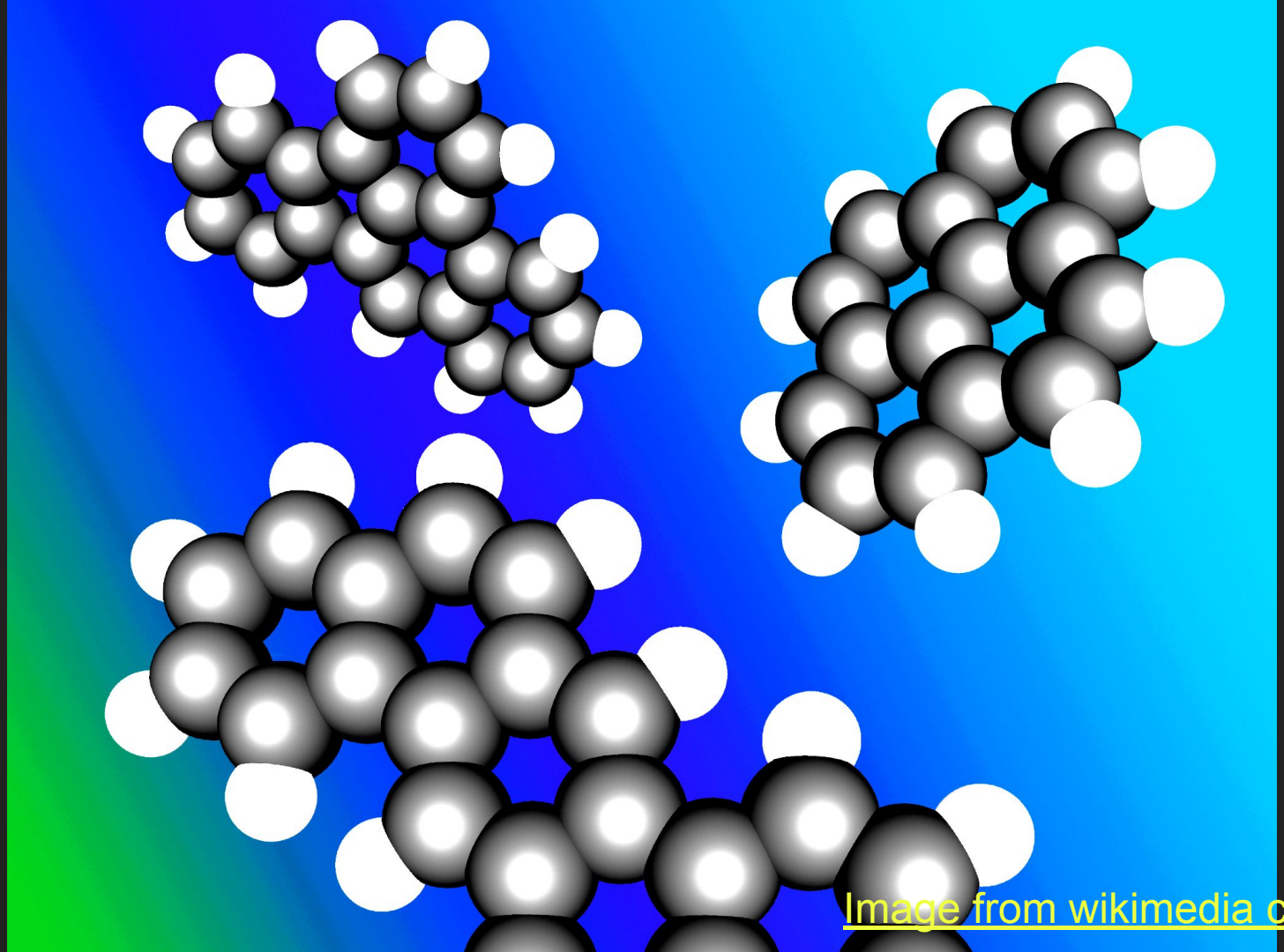


Image from wikimedia commons

TABLE 5: Thermodynamic Properties of PAH Molecules at 298 K[a]

species formula	C_p	S	ΔH	species formula	C_p	S	ΔH
A_1	83.5	269.9	14.4	A_2R5C_2H	188.6	410.1	29.3
A_1-	82.0	289.8	14.4	A_2R5C_2H*	186.9	416.4	29.3
A_1C_2H	117.6	329.8	19.9	$A_2R5(C_2H)_2$	223.3	454.6	35.3
A_1C_2H-2	116.0	341.4	19.9	$A_2R5C_2H_2$	195.2	429.1	30.6
A_1C_2H-3	116.0	340.7	19.8	$A_2R5C_2H_3$	196.7	424.5	30.8
A_1C_2H-4	115.9	334.7	19.8	A_3	185.9	391.7	28.3
$A_1C_2H_3$	120.8	345.7	20.4	A_3	185.5	398.5	28.3
$i-A_1C_2H_2$	124.7	350.0	21.5	A_3-	183.6	408.6	28.2
$n-A_1C_2H_2$	123.0	350.1	21.0	A_3C_2H	219.7	454.1	34.2
$A_1C_2H_3$*	121.1	357.1	20.9	$A_3C_2H_2$	222.4	455.9	34.4
$A_1(C_2H)_2$	152.0	374.7	25.6	$A_4(C_{16}H_{10})$	203.8	402.9	30.1
A_2	134.4	335.3	21.1	$A_4(C_{18}H_{12})$	237.7	448.2	35.6
A_2-1	132.8	352.3	21.0	$A_4(C_{18}H_{12})$	236.9	456.5	35.7
A_2-2	133.0	352.1	21.0	A_4-	201.9	419.8	30.0
A_2C_2HB	168.7	392.0	26.8	A_3R5	205.7	422.0	30.8
A_2C_2HB*	166.9	398.4	26.9	A_3R5-	203.9	427.2	30.7
A_2C_2HA	168.5	391.0	26.8	A_3R5C_2H	239.8	467.0	36.6
A_2C_2HA*	166.8	396.9	26.8	$A_3R5C_2H_2$	240.5	472.3	36.7
$A_2C_2H_3$	176.5	405.9	28.2	A_4R5	224.0	433.0	32.7
$A_2C_2H_2$	174.9	410.4	28.0	P_2	167.6	391.7	26.9
$A_2(C_2H)_2$	202.9	435.3	32.7	P_2-	165.0	401.1	26.8
A_2R5	154.7	359.9	23.5	A_5	255.6	469.2	38.0
A_2R5-	152.5	370.7	23.4	A_7	287.9	472.0	41.2

[a] Heat capacities and entropies are in J/mol/K; enthalpies are in kJ/mol.

Constructing the standard molar free entropy of formation

$$\Delta_f S^o_m(A_a B_b C_c) = S^o_m(A_a B_b C_c) - \frac{a}{\alpha} \times S^o_m(A_\alpha) + \frac{b}{\beta} \times S^o_m(B_\beta) + \frac{c}{\gamma} \times S^o_m(C_\gamma)$$

- (Greek letters denote lowest-enthalpy form of chemical elements). This formula is necessary to convert standard molar (absolute) entropy to standard molar entropy of formation to compute the Gibbs as shown:

$$\Delta_f \Phi^o_m = \frac{\Delta_f G^o_m}{T} = \frac{1}{T}\left(\Delta_f H^o_m - T\Delta_f S^o_m\right)$$

- How does free entropy, which is complementary to entropy, relate to the molecular structure of PAHs? How will that differ from entropy's relationship to these variables?
- "Standard" throughout here means 298K at ambient pressure; this is a condition of constant temperature and pressure, which means that the free entropy is related to the Gibbs free energy.
- Google AppsScript expands the original dataset to include these variables.

Methods: python, scikit-learn

- from sklearn.linear_model import LinearRegression
  - We will start with linear regression, treating standard molar entropy and standard molar free entropy of formation each separately as response variables for many predictor variables such as covalent bonds per molecule and molecular weight.
- from sklearn.linear_model import LassoCV, RidgeCV
  - Then, we will do regularized regression on both the l_1 and the l_2 norm to determine which of the molecular features are the most influential for predicting the entropy and free entropy of each compound.
- from sklearn.decomposition import PCA
  - Principal component analysis may be useful in determining which factors explain variance in the original thermochemical data.

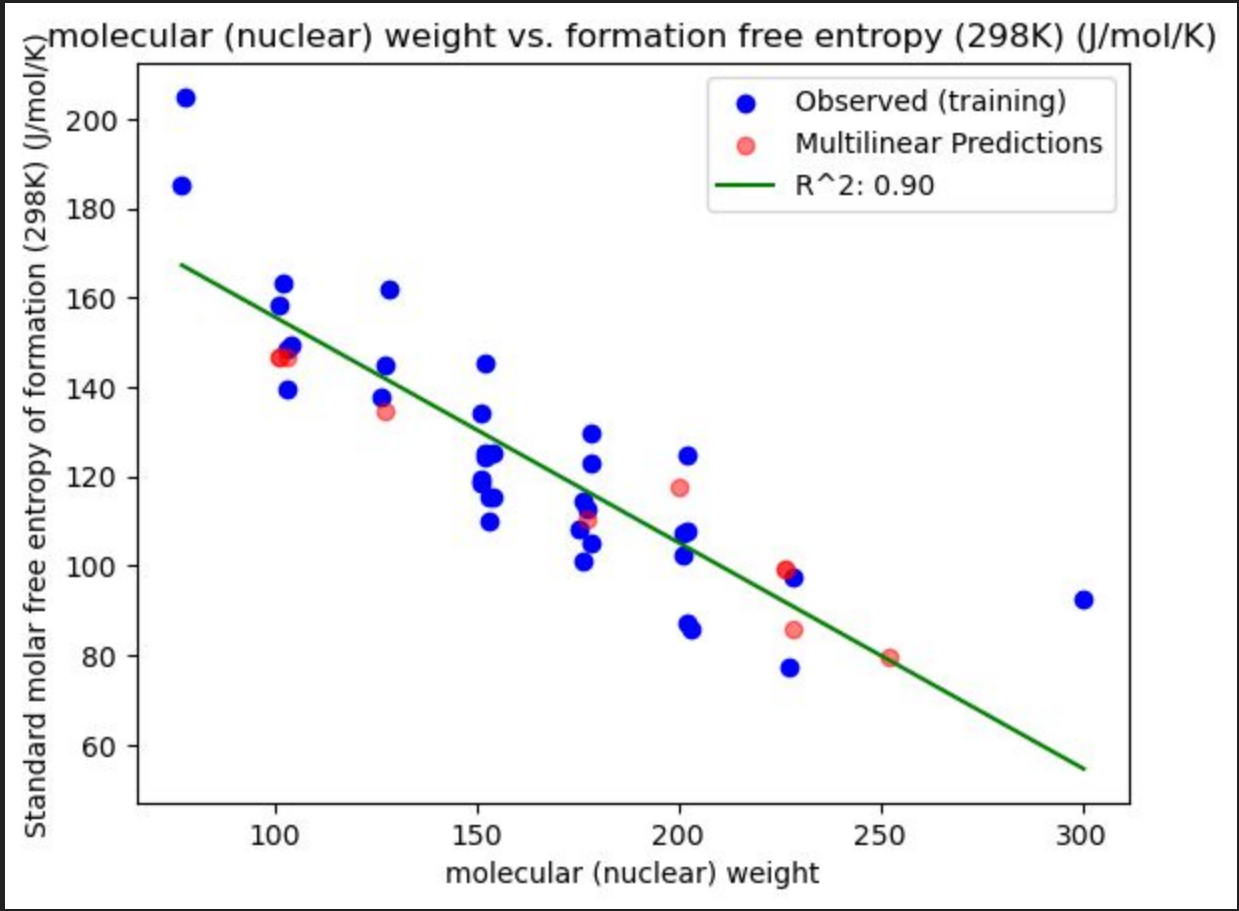
Our perspective on the data:

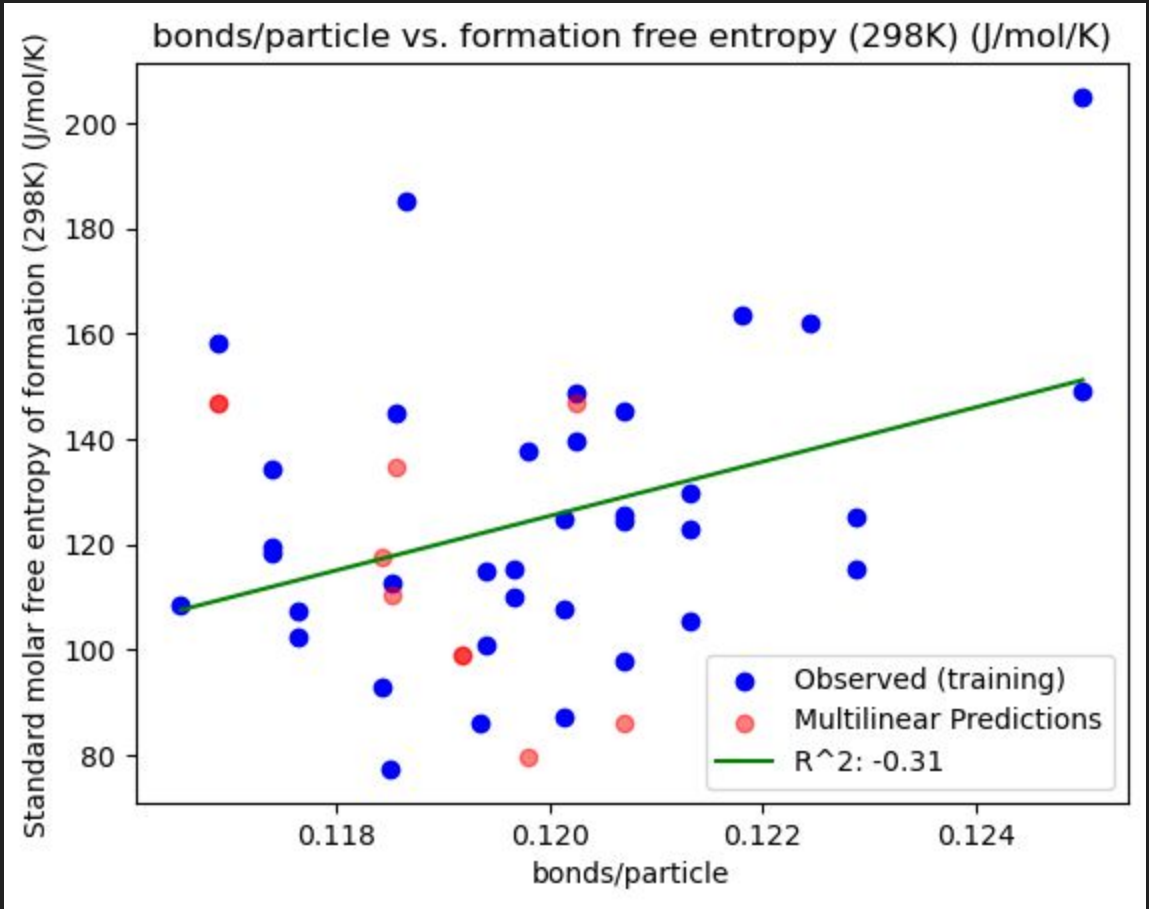
Predictors:

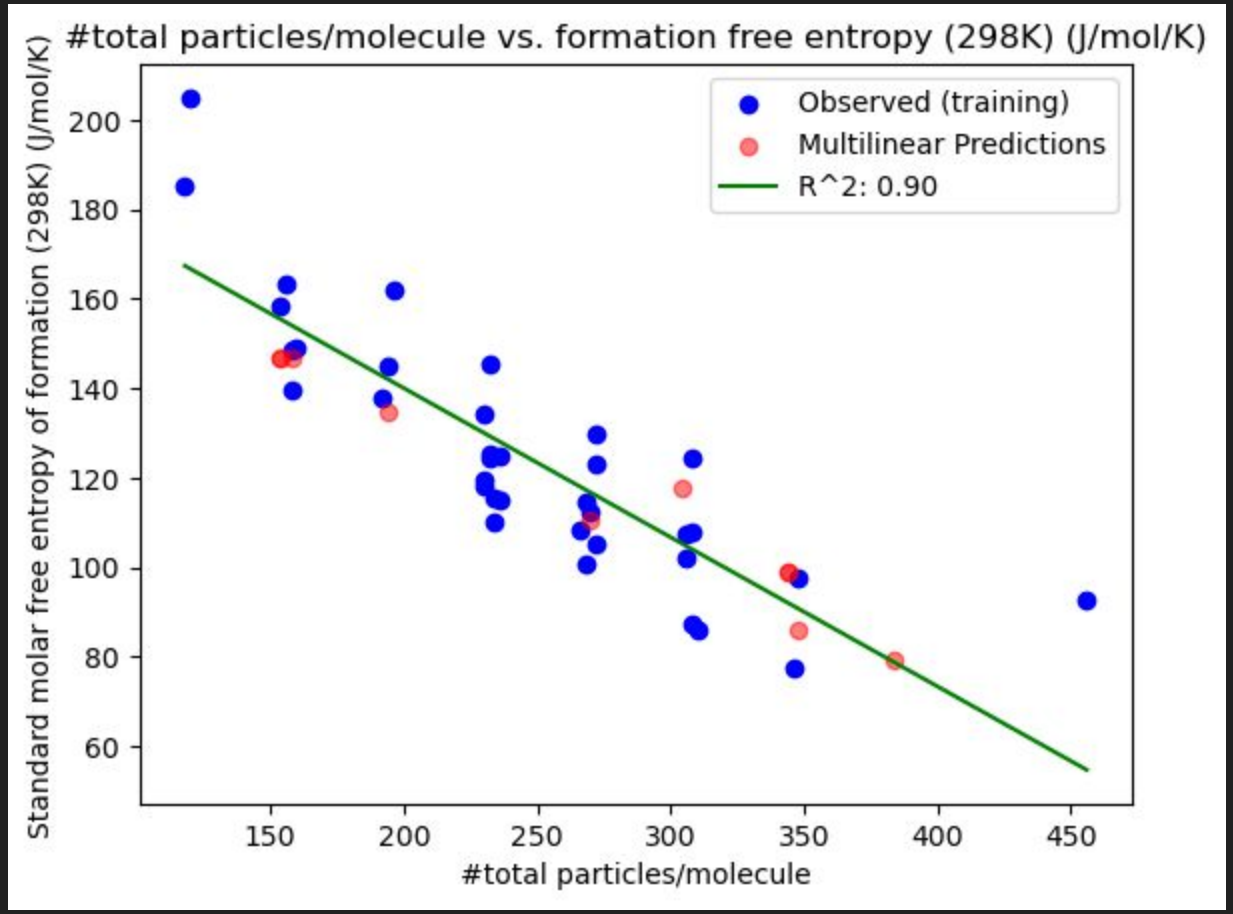
- # carbons / molecule
- # hydrogens / molecule
- # valence electrons / molecule
- # bonds / molecule
- # total subatomic particles / molecule
- # bonds / particle
- Molecular (nuclear) weight

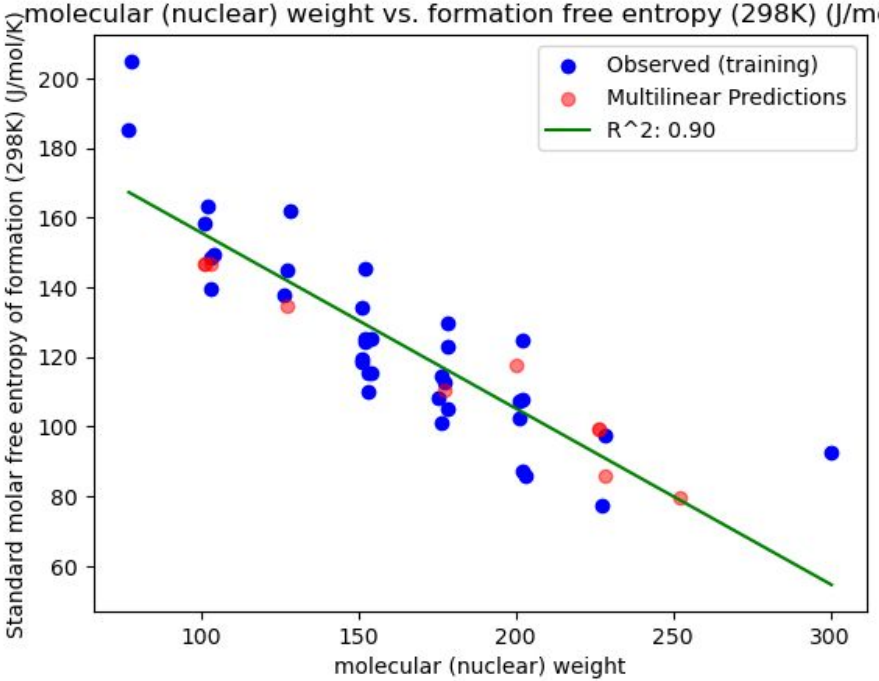
Responses:

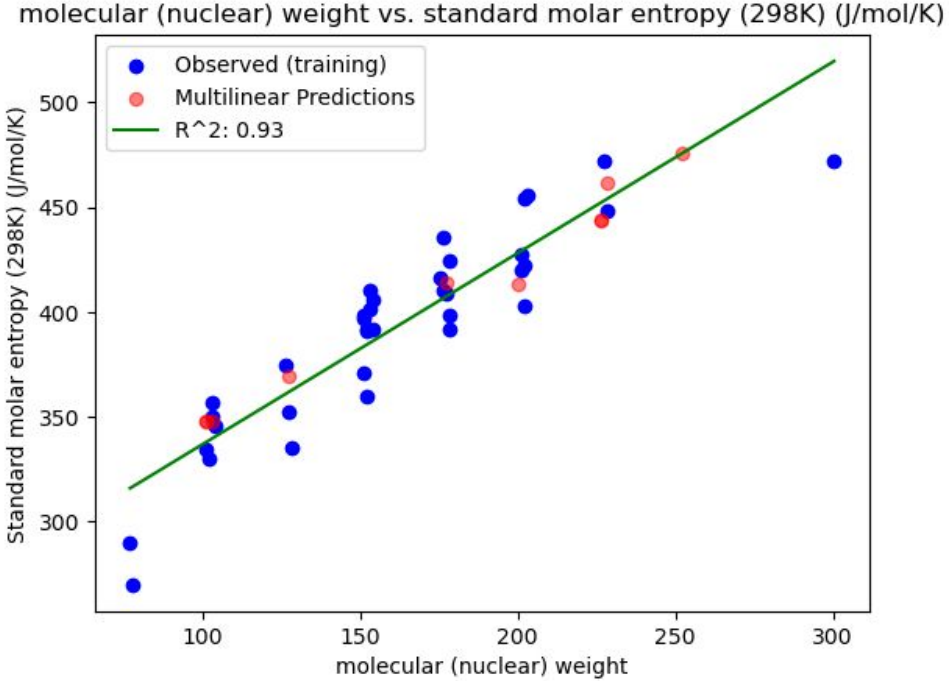
- standard molar entropy
- standard molar free entropy of formation

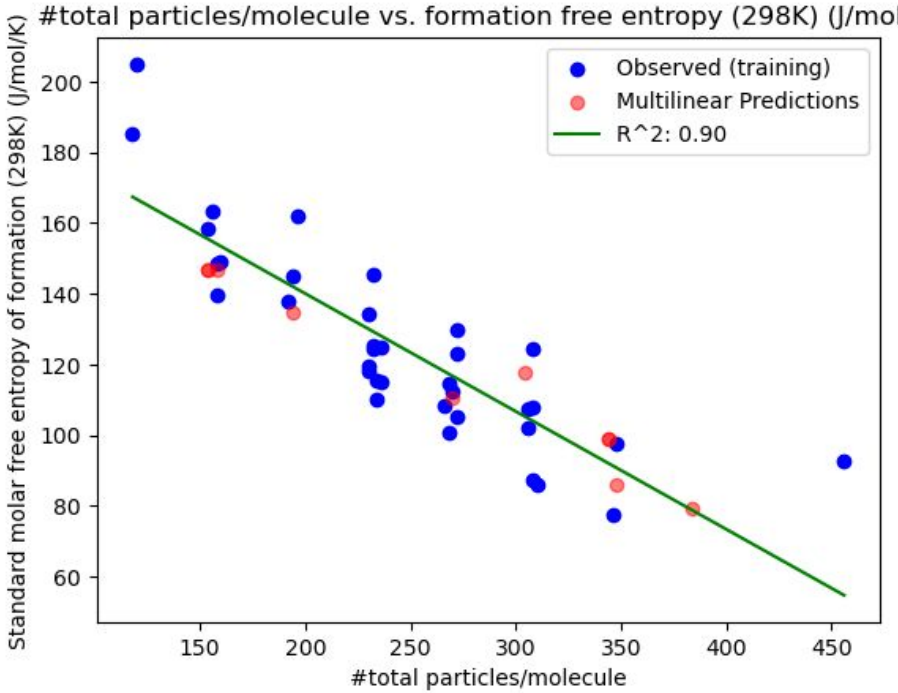

molecular (nuclear) weight vs. formation free entropy (298K) (J/mol/K)
Observed (training)
Multilinear Predictions
R^2: 0.90
Standard molar free entropy of formation (298K) (J/mol/K)
molecular (nuclear) weight
200
180
160
140
120
100
80
60
100
150
200
250
300

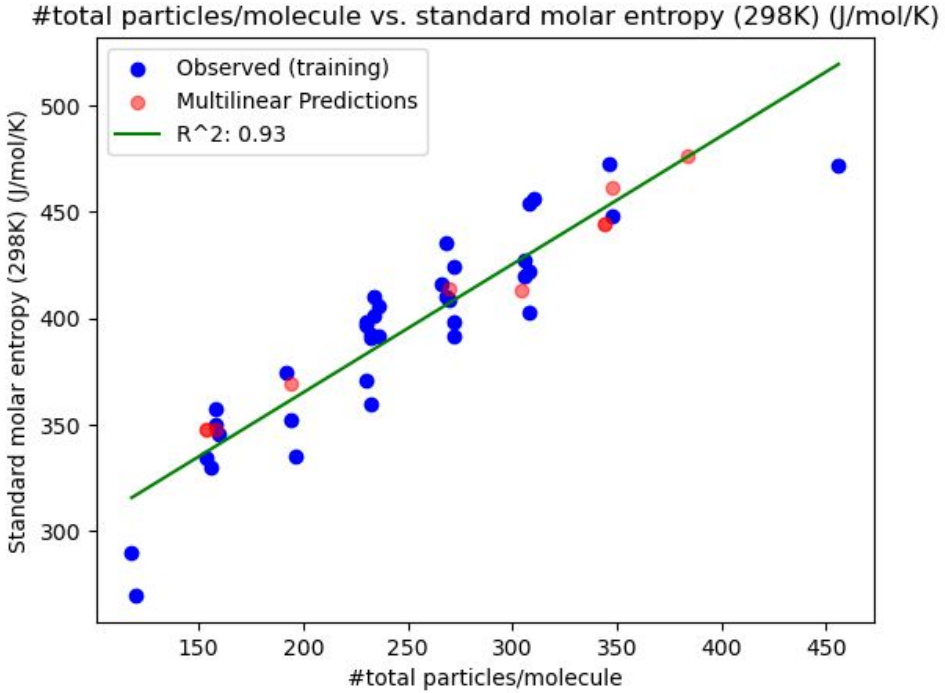

bonds/particle vs. formation free entropy (298K) (J/mol/K)
Standard molar free entropy of formation (298K) (J/mol/K)
200
180
160
140
120
100
80
0.118
0.120
0.122
0.124
bonds/particle
Observed (training)
Multilinear Predictions
R^2: -0.31

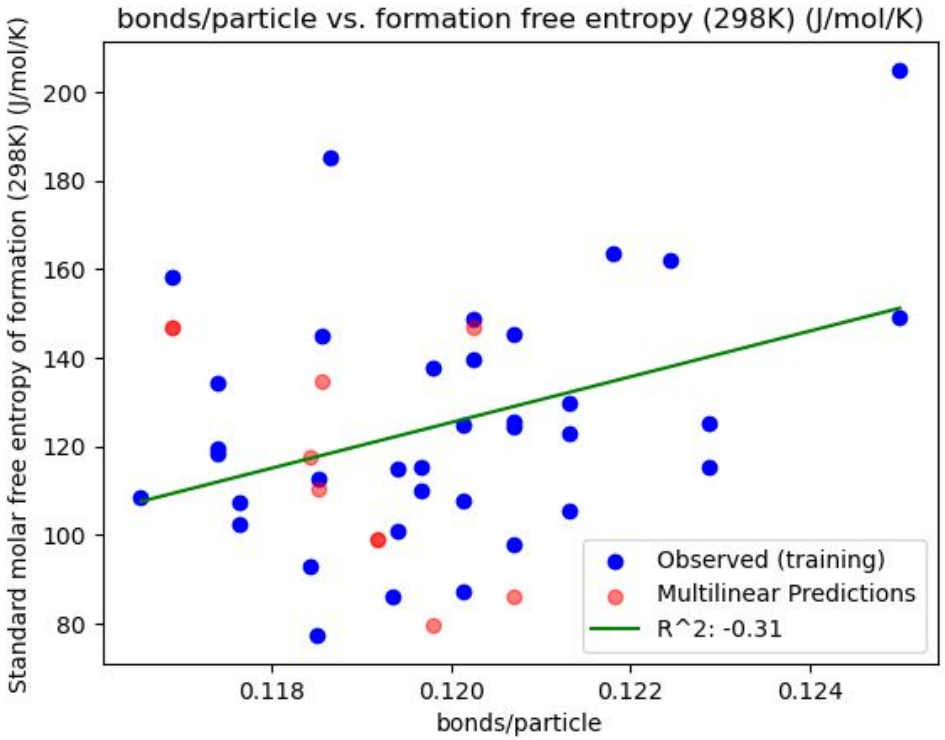

#total particles/molecule vs. formation free entropy (298K) (J/mol/K)
Observed (training)
Multilinear Predictions
R^2: 0.90
Standard molar free entropy of formation (298K) (J/mol/K)
#total particles/molecule
200
180
160
140
120
100
80
60
150
200
250
300
350
400
450

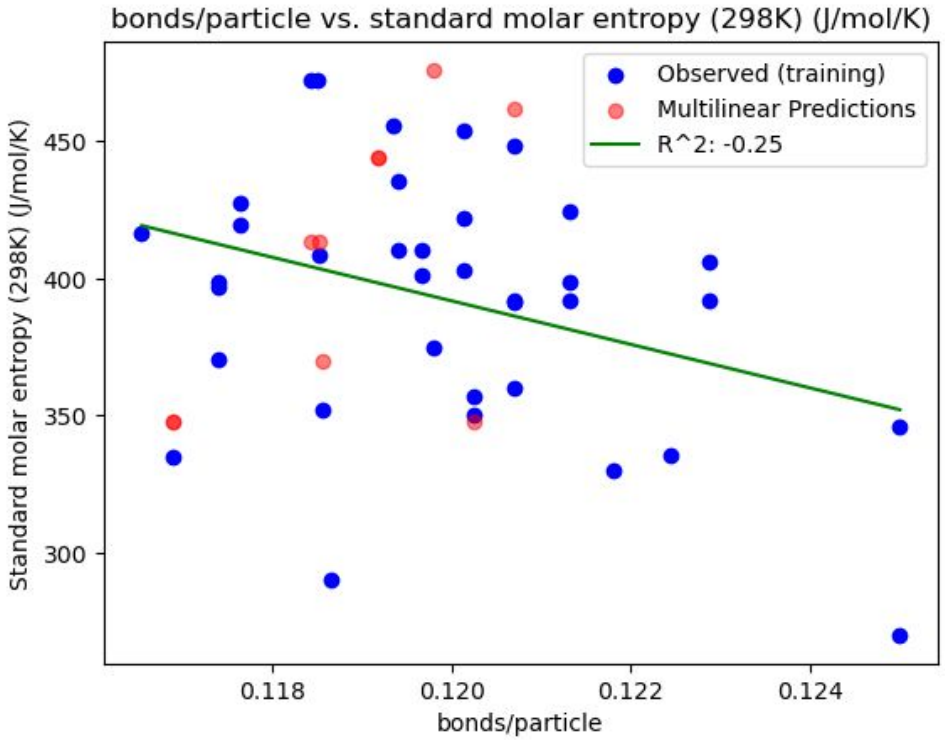

molecular (nuclear) weight vs. formation free entropy (298K) (J/mol/K)
Observed (training)
Multilinear Predictions
R^2: 0.90
Standard molar free entropy of formation (298K) (J/mol/K)
molecular (nuclear) weight
60
80
100
120
140
160
180
200
100
150
200
250
300


molecular (nuclear) weight vs. standard molar entropy (298K) (J/mol/K)
Observed (training)
Multilinear Predictions
R^2: 0.93
Standard molar entropy (298K) (J/mol/K)
molecular (nuclear) weight
300
350
400
450
500
100
150
200
250
300


#total particles/molecule vs. formation free entropy (298K) (J/mol/K)
Observed (training)
Multilinear Predictions
R^2: 0.90
Standard molar free entropy of formation (298K) (J/mol/K)
#total particles/molecule
200
180
160
140
120
100
80
60
150
200
250
300
350
400
450


#total particles/molecule vs. standard molar entropy (298K) (J/mol/K)
Observed (training)
Multilinear Predictions
R^2: 0.93
Standard molar entropy (298K) (J/mol/K)
#total particles/molecule
500
450
400
350
300
150
200
250
300
350
400
450


bonds/particle vs. formation free entropy (298K) (J/mol/K)
Standard molar free entropy of formation (298K) (J/mol/K)
200
180
160
140
120
100
80
0.118
0.120
0.122
0.124
bonds/particle
Observed (training)
Multilinear Predictions
R^2: -0.31


bonds/particle vs. standard molar entropy (298K) (J/mol/K)
Standard molar entropy (298K) (J/mol/K)
450
400
350
300
0.118
0.120
0.122
0.124
bonds/particle
Observed (training)
Multilinear Predictions
R^2: -0.25

Why is R^2 negative for the bond/particles ratio as a predictor?

- The definition of the coefficient of determination (denoted R^2) in scikit-learn is (1 - u/v) where u is the total sum of squares (y_true - y_true.mean()) ** 2).sum() and v is the residual sum of squares ((y_pred - y_true) ** 2).sum().

- The thermodynamic properties of polycyclic aromatic hydrocarbons do not vary by any noticeable trend with the molar thermodynamic properties, sticking around a mean of ~0.12.

- This means that the total sum of squares (y_true - y_true.mean()) ** 2) is very small.

- Having a very small v in the denominator means that u/v is greater than 1.

- That is why the coefficient of determination is negative, despite the presence of a "square" in the notation.

- Thermochemically, this negative R^2 result means that the bonds/particle ratio is constant for PAHs despite variation in molar thermodynamic properties, and that a linear regression model is not useful for predicting this essential lack of change.

Lasso regression (free entropy)

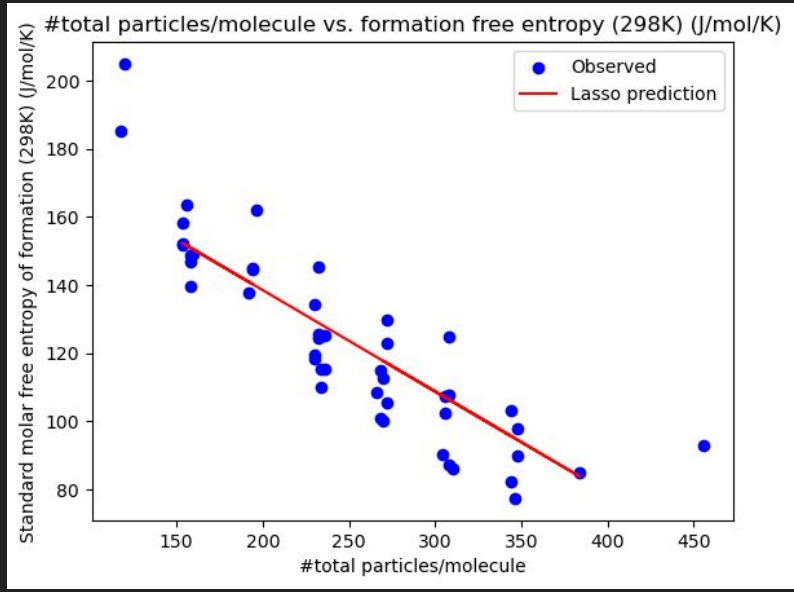
```
#repeating the previous regression with lasso regularization.
from sklearn.linear_model import LassoCV
model_2_lasso_molar_free_entropy = LassoCV(cv=10)
model_2_lasso_molar_free_entropy.fit(X_train, y_train)

y_pred = model_2_lasso_molar_free_entropy.predict(X_test)

summarize(model_2_lasso_molar_free_entropy)
```

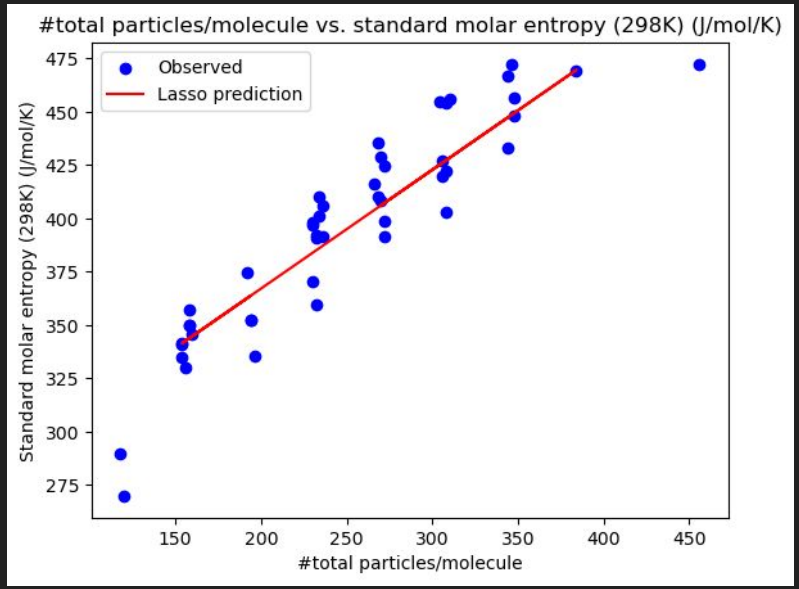
```
Intercept: 197.97101492481193
Coefficients:
  #carbons/molecule: -0.0
  #hydrogens/molecule: -0.0
  # valence electrons / molecule: -0.0
  #bonds / molecule: -0.0
  #total particles/molecule: -0.2974585835620723
  bonds/particle: 0.0
  molecular (nuclear) weight: -0.0
Mean Squared Error: 89.92096690775008
R^2 Score: 0.8906619250232303
```

Note: It is interesting that the lasso selected exclusively for total particles / molecule



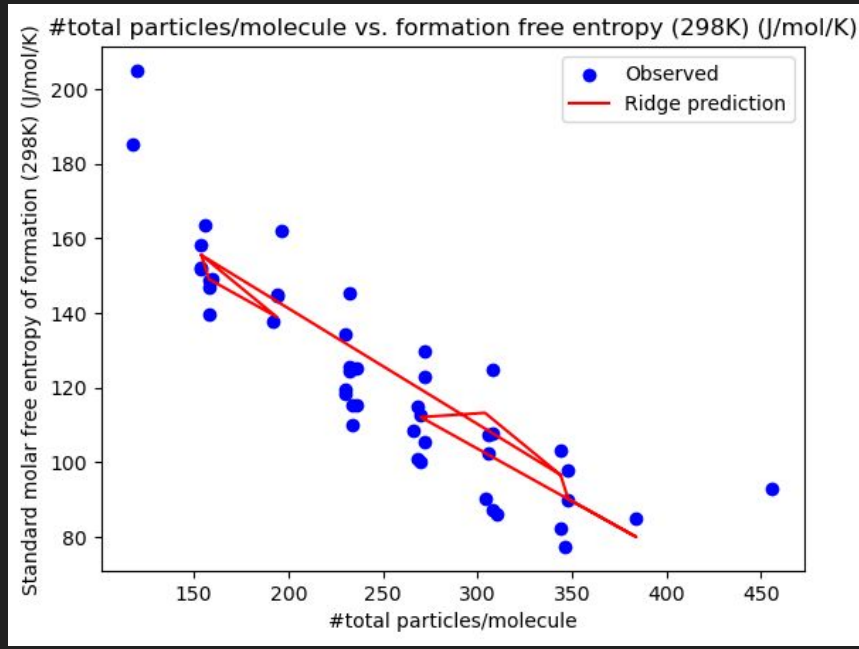
Lasso regression (entropy)

```
Intercept: 255.7939650466574
Coefficients:
  #carbons/molecule: 0.0
  #hydrogens/molecule: 0.0
  # valence electrons / molecule: 0.0
  #bonds / molecule: 0.0
  #total particles/molecule: 0.5567170211343394
  bonds/particle: -0.0
  molecular (nuclear) weight: 0.0
Mean Squared Error: 221.70566451838317
R^2 Score: 0.9214391635506651
```



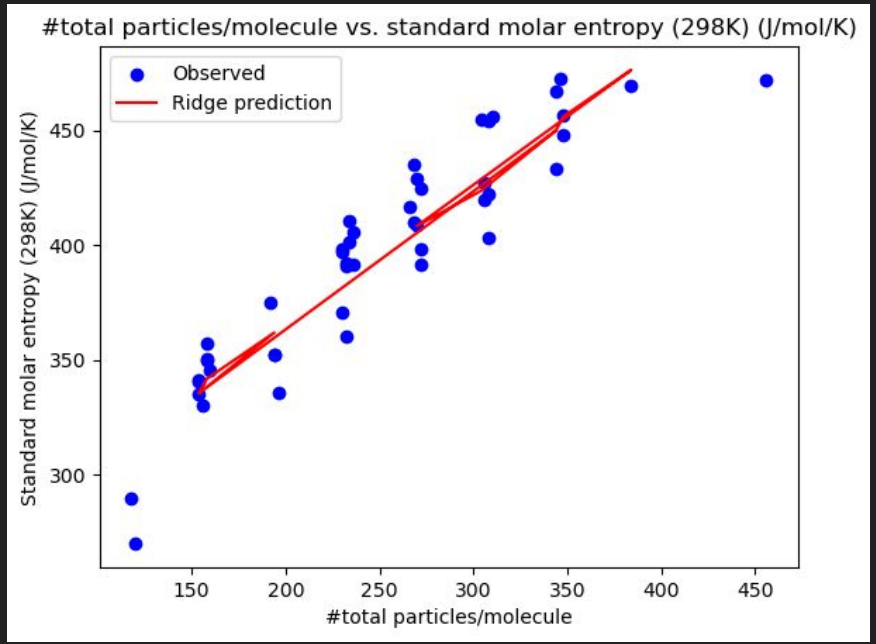
Ridge regression (free entropy)

```
Intercept: 220.34813807451587
Coefficients:
  #carbons/molecule: 0.4746590740764867
  #hydrogens/molecule: -5.137263608526174
  # valence electrons / molecule: -3.2386273122260048
  #bonds / molecule: 16.19751885364877
  #total particles/molecule: -1.730663883702601
  bonds/particle: 0.17359422040207853
  molecular (nuclear) weight: 0.5120192351254899
Mean Squared Error: 99.88277644492965
R^2 Score: 0.8785490094759826
```



Ridge regression (entropy)

```
Intercept: 237.43009494589643
Coefficients:
  #carbons/molecule: -0.19448041961286958
  #hydrogens/molecule: 2.1439739186976037
  # valence electrons / molecule: 1.36605224024702
  #bonds / molecule: -3.6521876787476533
  #total particles/molecule: 0.7873002843654628
  bonds/particle: -0.03049884570808079
  molecular (nuclear) weight: -0.17029618166974794
Mean Squared Error: 224.85472778162574
R^2 Score: 0.9203233010194591
```



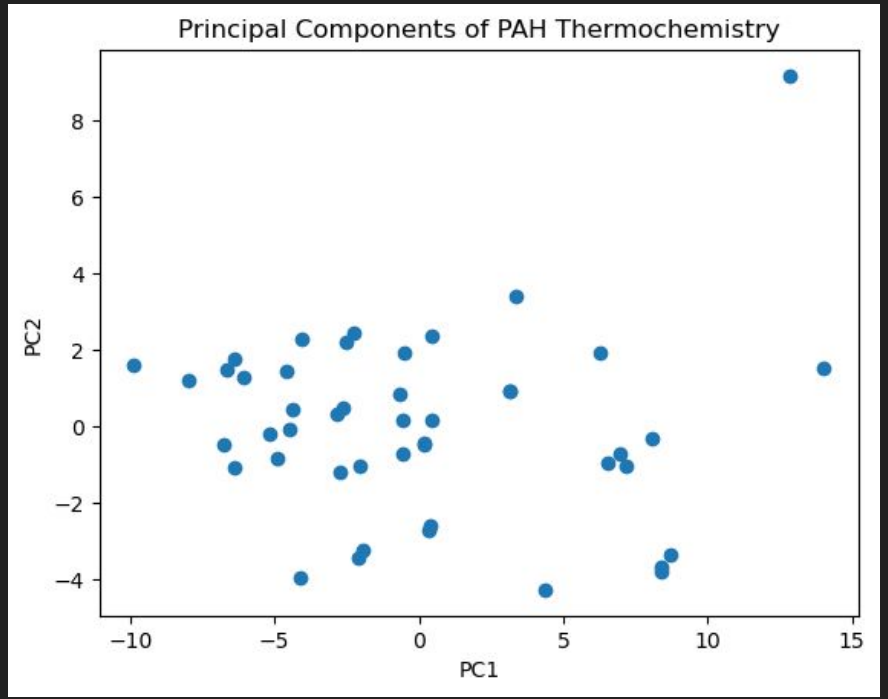
Comments on Regression

- It is interesting to note that, for both standard molar entropy and standard molar free entropy of formation, the lasso automatically selects the total number of subatomic particles per molecule as the sole predictor.
- As the ratio of subatomic particles per molecule in a polycyclic aromatic hydrocarbon increases, the standard molar entropy tends to increase and the standard molar free entropy of formation tends to decrease.
- This trend might challenge the conventional notion that entropy is a measure of disorder, because we find increasing standard molar entropies for the increasing size of organic molecules which are highly ordered (PAHs).
- The trend emerges because each individual particle carries a proportional amount of thermodynamic entropy, decreasing molar free energy due to the negative entropy term in $G = H - TS$.

Principal Component Analysis (PCA) (BIPLOT THIS!)

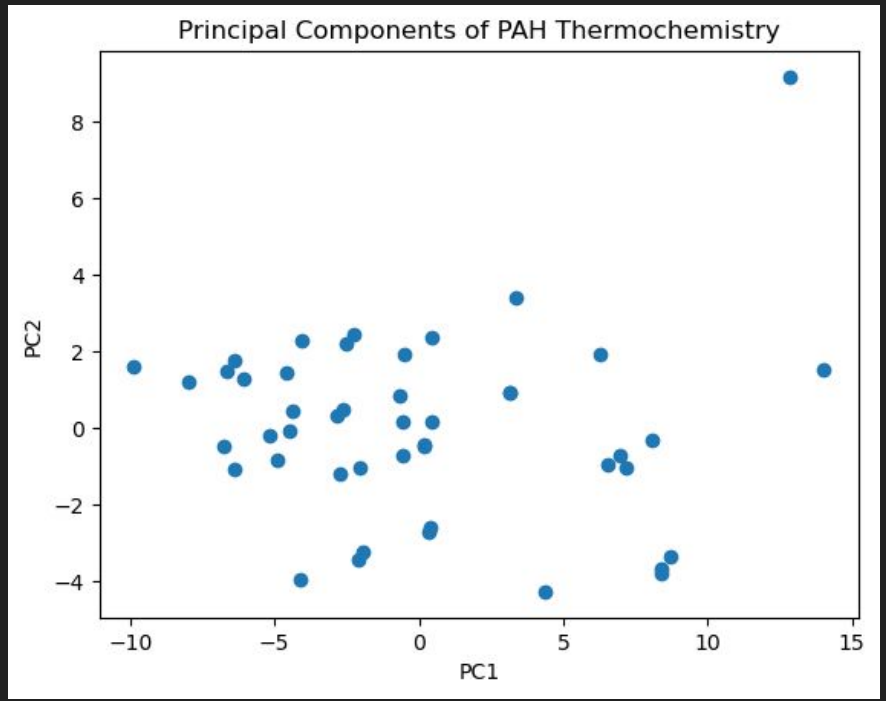
```
Variance explained by each component: [0.76225045 0.13782869 0.05802059 0.02003238 0.01141601]
Loadings in the first principle component:

#carbons/molecule: -0.17216072726105885
#hydrogens/molecule: -0.15666430631730832
# valence electrons / molecule: -0.1728732839384492
#bonds / molecule: -0.17266950566245895
#total particles/molecule: -0.1725668603749473
bonds/particle: 0.026524501360732648
molecular (nuclear) weight: -0.1724829024987645
entropy S (298K) (J/mol/K): -0.17016120373361862
entropy S (298K) (J/g/K): 0.17366905525363488
formation entropy Sf (298K) (J/mol/K): -0.17016120373361868
formation entropy Sf (298K) (J/g/K): -0.1684511840657803
formation enthalpy (298K) (kJ/mol): -0.17093074829576305
formation enthalpy (298K) (kJ/g): 0.13930011625986796
formation Gibbs (298K) (kJ/mol) (dG = dH − TdS): 0.16610196485434392
formation Gibbs (298K) (kJ/g) (dG = dH − TdS): 0.17262016247682443
formation free entropy (298K) (J/mol/K): 0.16610196485383438
formation free entropy (298K) (J/g/K): 0.1726201624720837
Standard molar entropy / mol of bonds (J / mol / K): 0.1724311820508677
standard molar entropy of formation per mol of bonds (298K) (J/mol/K): -0.16933791905831502
standard molar enthalpy of formation / mol of bonds (298K) (kJ/mol): 0.13470810696987515
standard molar free entropy of formation / mol of bonds (298K) (J/mol/K): 0.17359512561857557
standard molar entropy / mol of valence electrons (298K) (J/mol/K): 0.1732934089826309
Standard molar molar entropy of formation / mol of valence electrons (298K) (J/mol/K): -0.16914010284687647
Standard molar free entropy of formation per mol of valence electrons (298K) (J/mol/K): 0.1733087770406729
Standard molar enthalpy of formation / mol of valence electrons (298K) (kJ/mol): 0.1312600305979607
standard molar entropy / carbons (298K) (J/mol/K): 0.17347458833361182
standard molar entropy of formation / mol of carbon atoms (J/mol/K): -0.16799700590151806
standard molar enthalpy of formation / mol of carbon atoms (298K) (kJ/mol): 0.14228628771913318
Standard molar free entropy of formation / mol of carbon atoms (298K) (J/mol/K): 0.17212975159977198
standard molar entropy / mol of particles (298K) (J/mol/K): 0.17367057986071757
standard molar entropy of formation / mol of particles (298K) (J/mol/K): -0.16858011540115125
standard molar enthalpy of formation / mol of particles (298K) (kJ/mol): 0.1381680274420011
Standard molar free entropy of formation / mol of particles (298K) (J/mol/K): 0.1727548328677048
standard molar entropy / molar of hydrogen atoms (J/mol/K) (298K): 0.10542925000274063
standard molar entropy of formation / mol of hydrogen atoms (J/mol/K) (298K): -0.17184817463831967
standard molar enthalpy of formation / mol of hydrogen atoms: -0.017288156563410227
standard molar free entropy of formation / mol of hydrogen atoms: 0.1705411553010496
standard molar entropy / mol / nuclear weight (J / g / K): 0.04597748376457837
standard molar entropy of formation / mol / nuclear weight  (J / g / K): -0.1573694336855763
standard molar enthalpy of formation / mol / nuclear weight (kJ / g / K): -0.0436137997450919
standard molar free entropy of formation / mol / nuclear weight (J / g / K): 0.1375847546790233
```



Comments on PCA

- The variance as explained by the first two principal components in PCA of the PAH thermochemistry dataframe does not add all the way to one.
- For the first several components (of which we have computed the first five), there is at least some non-noise signal.
- The dominant signal is of the first principal component, representing the trend between particle count and standard molar entropy.
- The variance of the thermodynamic and molecular properties of PAHs cannot be simply reduced to one determining factor.
- The loadings in the first principle component shows that the signal in the data set is very much a combination of all the features, except with low loading coming from the standard molar entropy over nuclear weight and standard molar enthalpy of formation over nuclear weight
    - (both of these variables have <0.1 loading in the first principal component).-



?