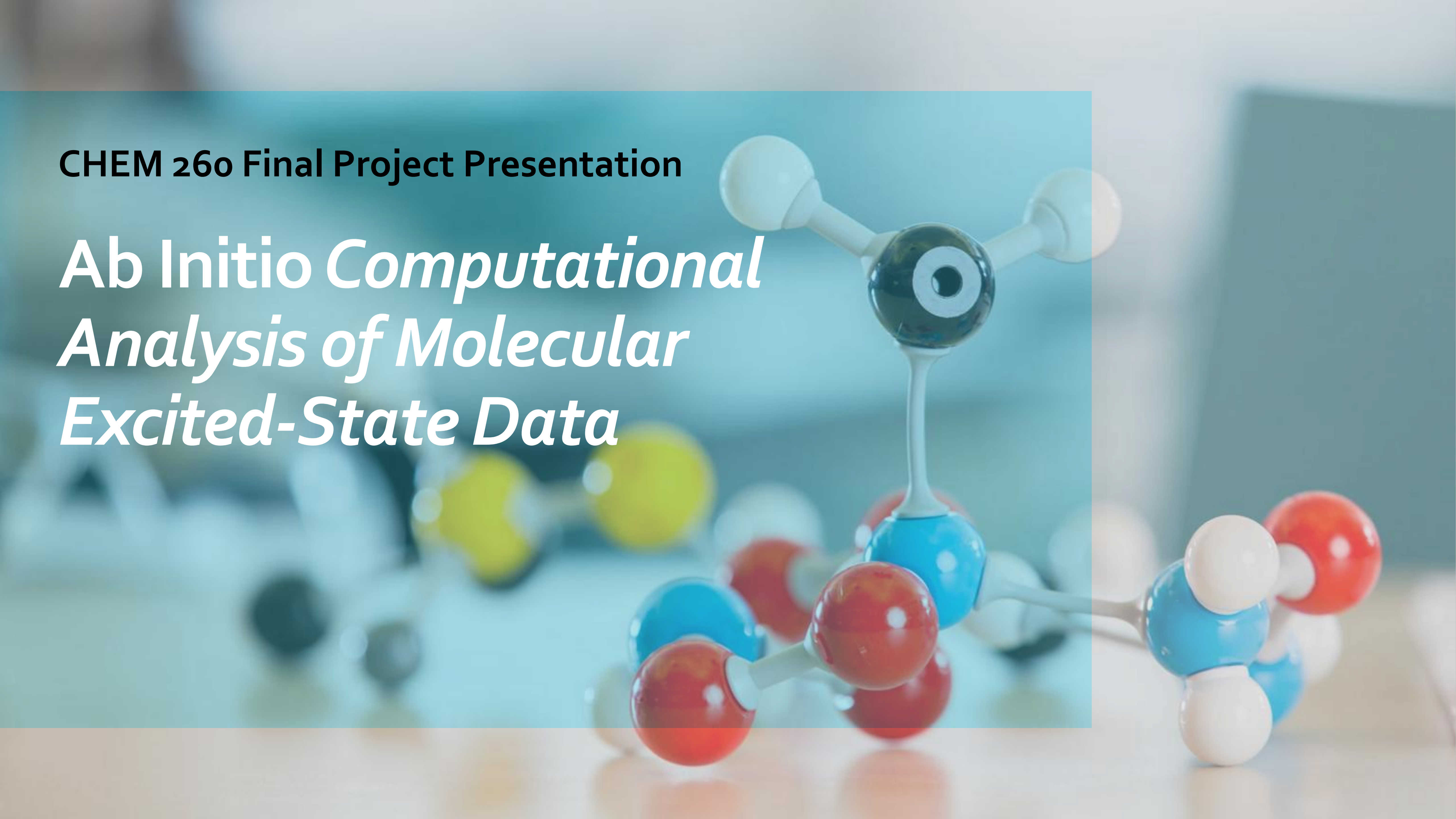
CHEM 260 Final Project Presentation

Ab Initio *Computational Analysis of Molecular Excited-State Data*



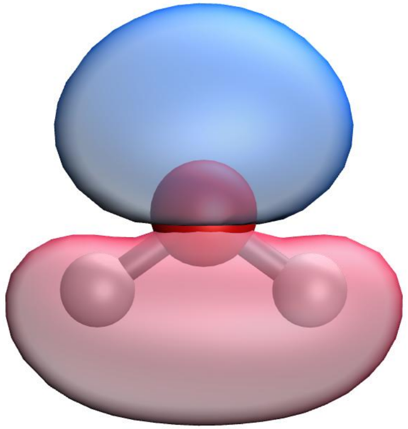
Presentation Outline

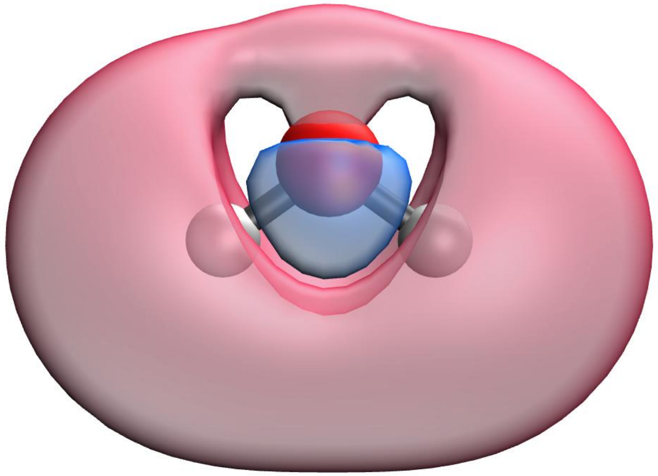
- Introduction and Background
- Project Strategy and Methods
- Calculation of Excitation Data
- Manual Analysis of Data
- Computational Analysis
- Application of Program
- Conclusion and Future Plans

Introduction and Background

Types of Electronic Excited States

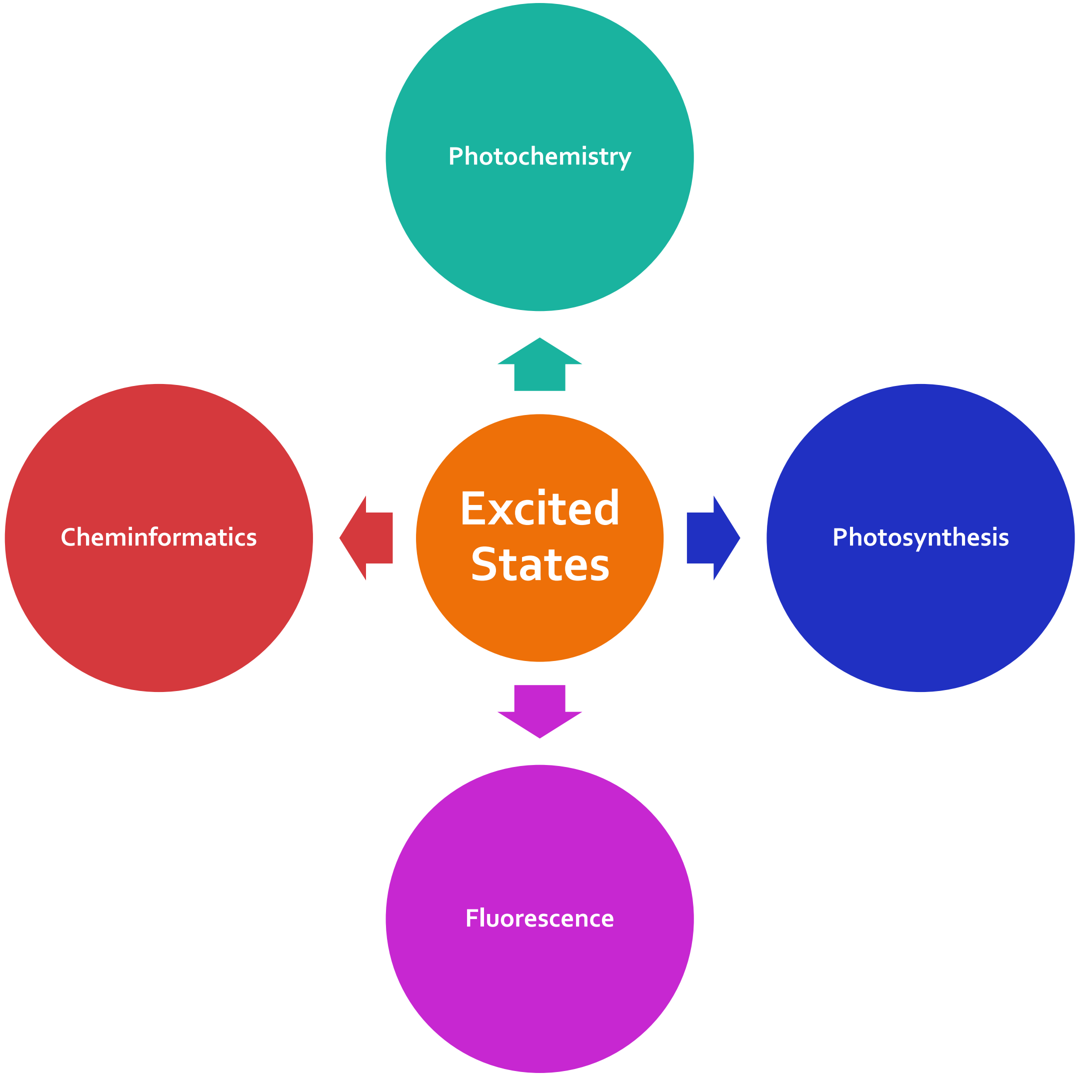
- **Valence States**
  - Lower Energy
  - Valence Orbitals
- **Rydberg States**
  - Higher Energy
  - Large, Diffuse Orbital
- **Core-Excited States**
  - Very High Energy
  - Valence or Rydberg





González, Escudero, Serrano-Andrés *ChemPhysChem* (**2012**) 28-51

Significance of Electronic Excited States



Excited States

Photochemistry

Photosynthesis

Fluorescence

Cheminformatics

Chemical Databases

Chemical/Biological

PubChem

Physical Data

NIST
National Institute of Standards and Technology

Drugs/Drug Targets

DRUGBANK

Solid State Data

OQMD
The Open Quantum Materials Database

3D Protein Structures

RCSB PDB
PROTEIN DATA BANK

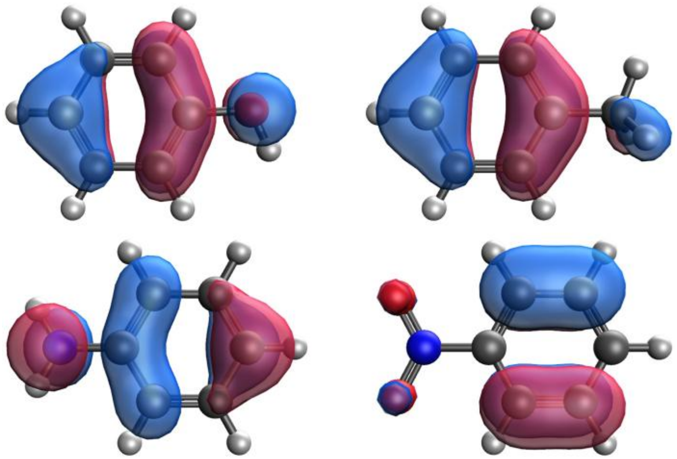
3D Crystal Structures

CCDC

Little to Nothing on Electronic Excited States!

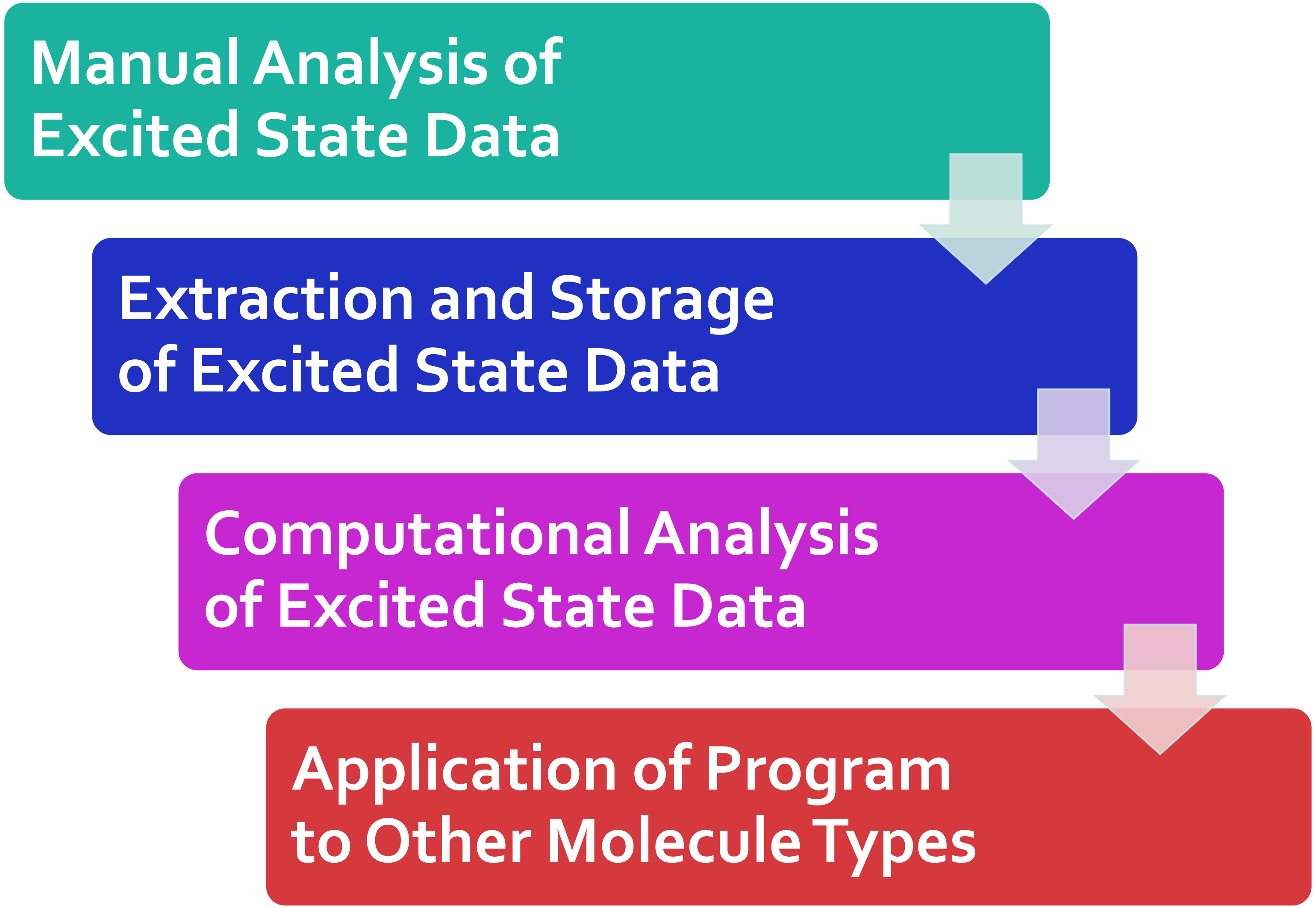
Hypothesis

- Molecular orbitals are used to visualize excited states
- Generating orbitals for multiple molecules is tedious
- **What about data?** – Energies, Oscillator Strengths
- Data is readily available in computation output file



Project Strategy and Methods

Four-Step Project Strategy



Manual Analysis of Excited State Data

Extraction and Storage of Excited State Data

Computational Analysis of Excited State Data

Application of Program to Other Molecule Types

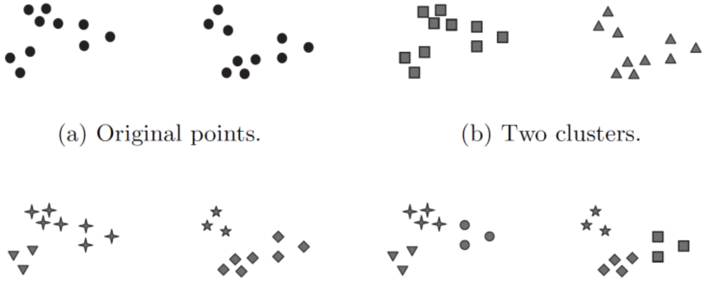
Hierarchical Clustering Analysis

Cluster Analysis

- Groups data objects using information that describes their relationships
- The greater the similarity within a group, the more distinct the clustering

Hierarchical Clustering

- ***Agglomerative*** – The closest pair of clusters or individual points merge at each step
- ***Divisive*** – One cluster is split until only single clusters of individual points remain

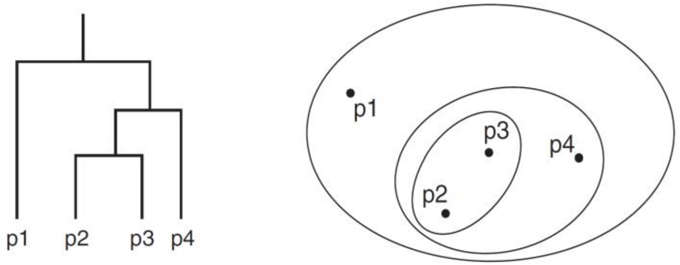


(a) Original points.

(b) Two clusters.

(c) Four clusters.

(d) Six clusters.



(a) Dendrogram.

(b) Nested cluster diagram.

Tan, Steinbach, Kumar *Introduction to Data Mining* (**2005**) 490-491, 515-516

Euclidean Distance

- Recall the Distance Formula from General Mathematics

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- Euclidean Distance = Extrapolation of Distance Formula

$$d = \sqrt{(a_B - a_A)^2 + (b_B - b_A)^2 + (c_B - c_A)^2 + \cdots}$$

Rokach, Maimon *Data Mining and Knowledge Discovery Handbook* (**2005**) 322-323

Methodology

Ground State Methods

- Density Functional Theory (B3LYP)
- Hartree-Fock Theory

Excited State Methods

- Configuration Interaction Singles (CIS)
- Tamm-Dancoff Approximation (TDA)
- Time-Dependent DFT (TDDFT)
- Gaussian Basis Set (6-31G)

Q-Chem Calculations

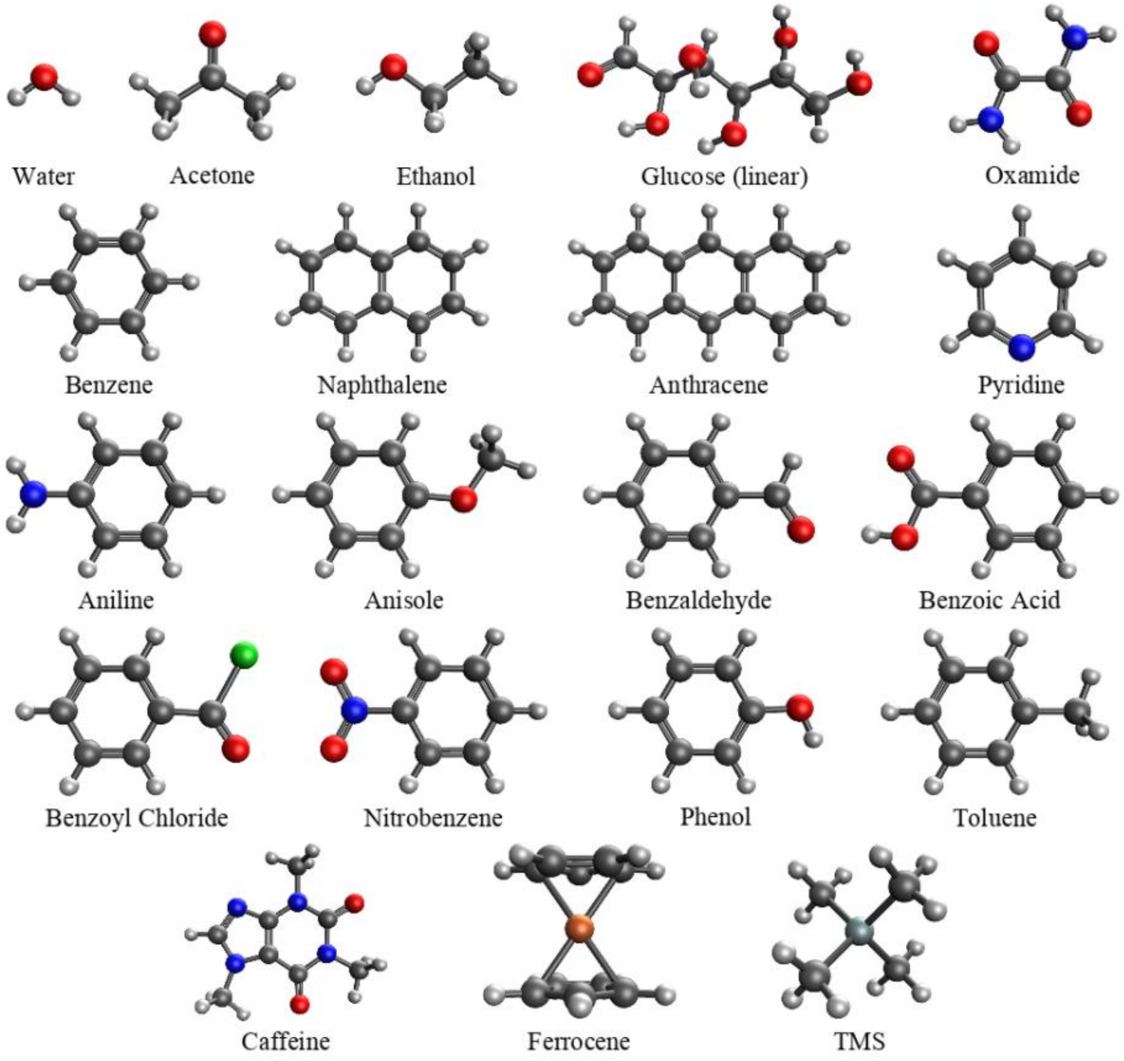
- Self-Consistent Field (SCF) Energy
- Optimized Molecular Geometry
- Infrared/Raman Frequencies
- Excitation Energies

Python Libraries

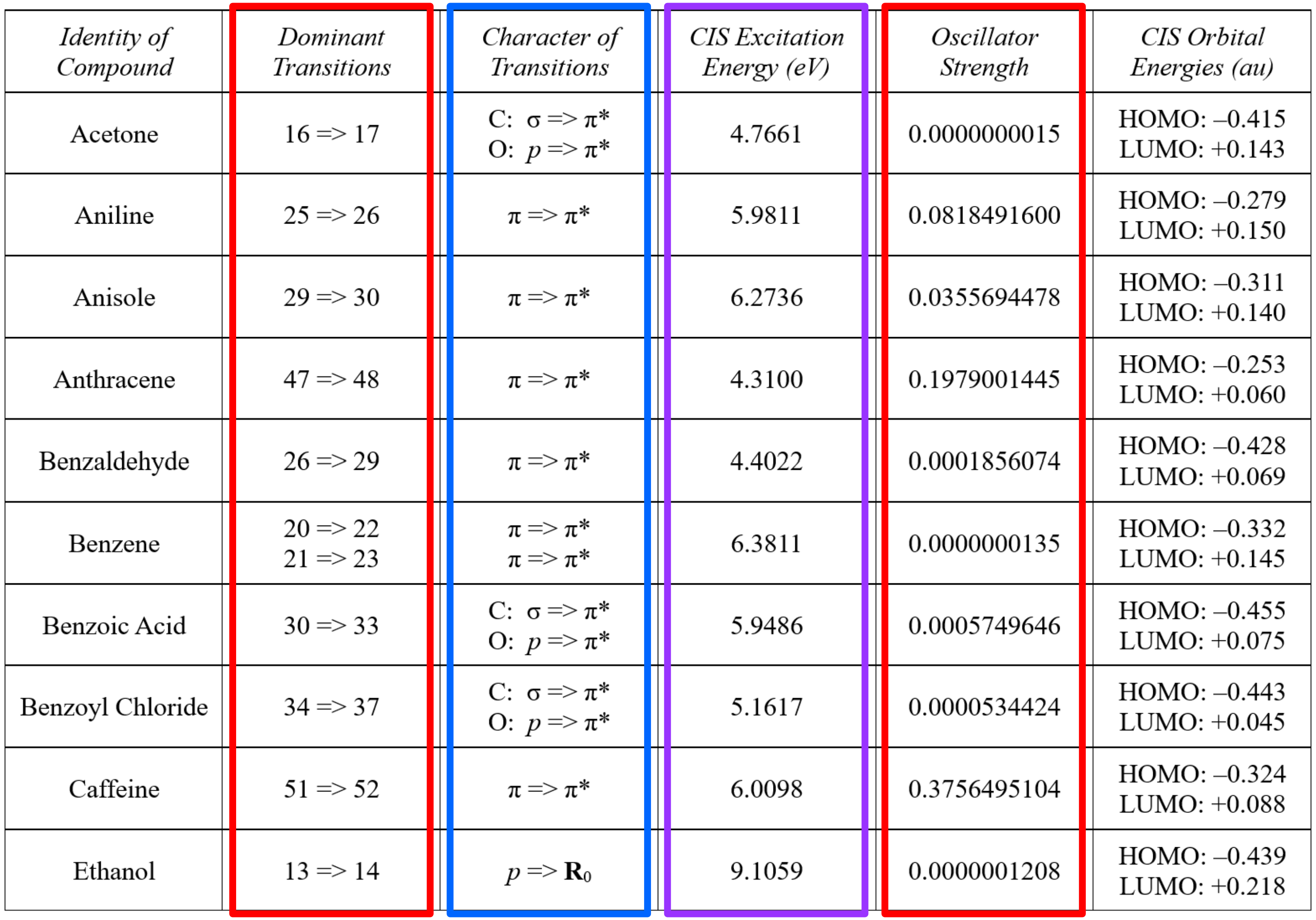
- from scipy.cluster.hierarchy import dendrogram, linkage
- from matplotlib import pyplot
- import pandas, math, sys

Calculation of Excitation Data

Training Set of Molecules



Generated Sample Data



Identity of Compound	*Dominant Transitions*	*Character of Transitions*	*CIS Excitation Energy (eV)*	*Oscillator Strength*	*CIS Orbital Energies (au)*
Acetone	16 => 17	C: σ => π* O: *p* => π*	4.7661	0.0000000015	HOMO: –0.415 LUMO: +0.143
Aniline	25 => 26	π => π*	5.9811	0.0818491600	HOMO: –0.279 LUMO: +0.150
Anisole	29 => 30	π => π*	6.2736	0.0355694478	HOMO: –0.311 LUMO: +0.140
Anthracene	47 => 48	π => π*	4.3100	0.1979001445	HOMO: –0.253 LUMO: +0.060
Benzaldehyde	26 => 29	π => π*	4.4022	0.0001856074	HOMO: –0.428 LUMO: +0.069
Benzene	20 => 22 21 => 23	π => π* π => π*	6.3811	0.0000000135	HOMO: –0.332 LUMO: +0.145
Benzoic Acid	30 => 33	C: σ => π* O: *p* => π*	5.9486	0.0005749646	HOMO: –0.455 LUMO: +0.075
Benzoyl Chloride	34 => 37	C: σ => π* O: *p* => π*	5.1617	0.0000534424	HOMO: –0.443 LUMO: +0.045
Caffeine	51 => 52	π => π*	6.0098	0.3756495104	HOMO: –0.324 LUMO: +0.088
Ethanol	13 => 14	*p* => $\mathbf{R}_0$	9.1059	0.0000001208	HOMO: –0.439 LUMO: +0.218

Blue = Manual Analysis, **Red** = Computational Analysis

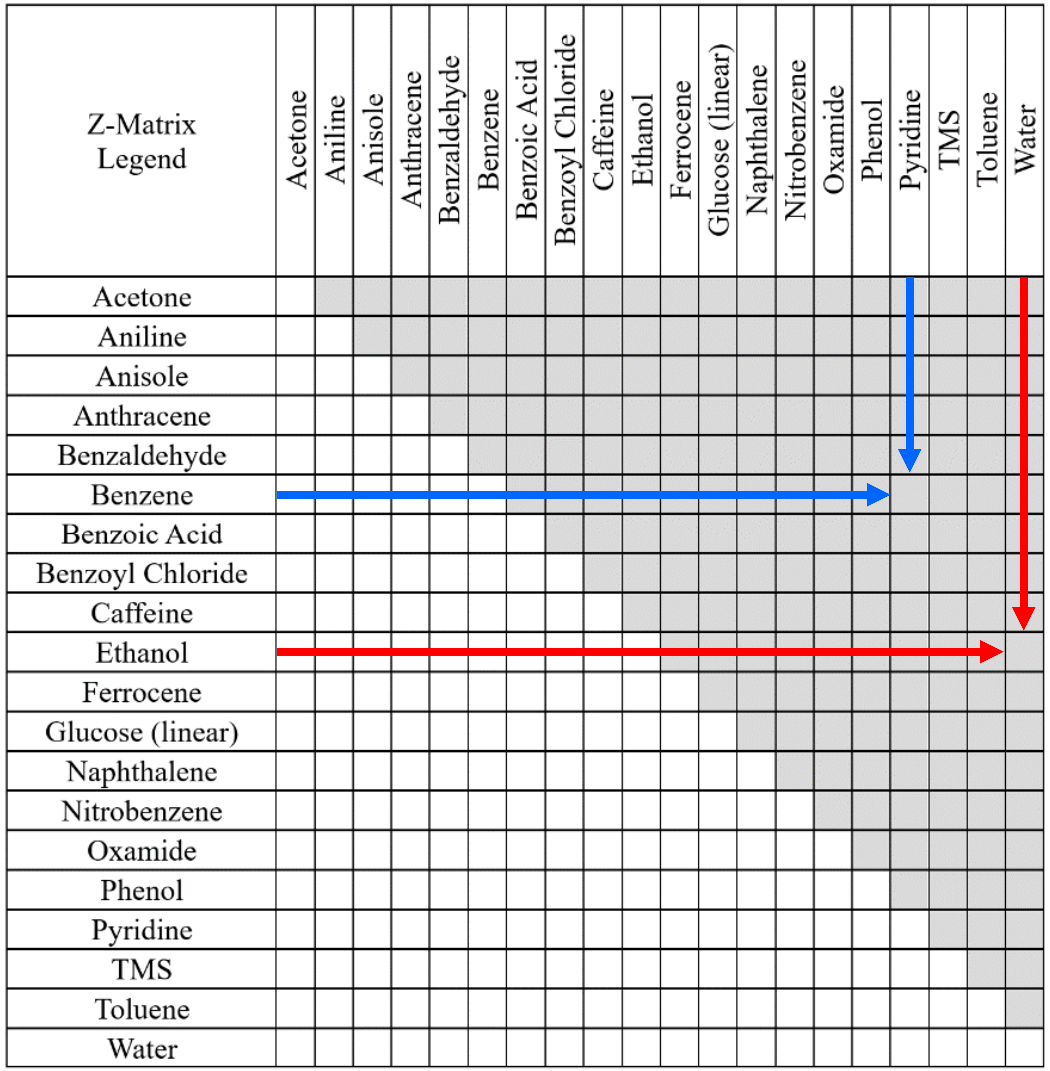
Manual Analysis of Excitation Data

Manual Analysis (MA) Dendrograms

- **Two Key Parameters**
    - (OS) **Orbital** and **Structure** Differences (Visual)
        - $\text{OS} = 1 - x$ where $x = \text{Level of Similarity}$
    - (E) **Excitation** Energy Differences (Numeric)
        - $\text{E} = 1 - \exp(-\Delta E / k_B T)$
    - ($\mathbb{Z}$) Weighed Combination of OS and E Values
        - $\mathbb{Z} = C_1(\text{OS}) + C_2(\text{E})$

- A first approximation considered OS and E equally
- Data contained using Microsoft Excel spreadsheets
- Python script created dendrograms from spreadsheets

Manual Analysis (MA) Dendrograms



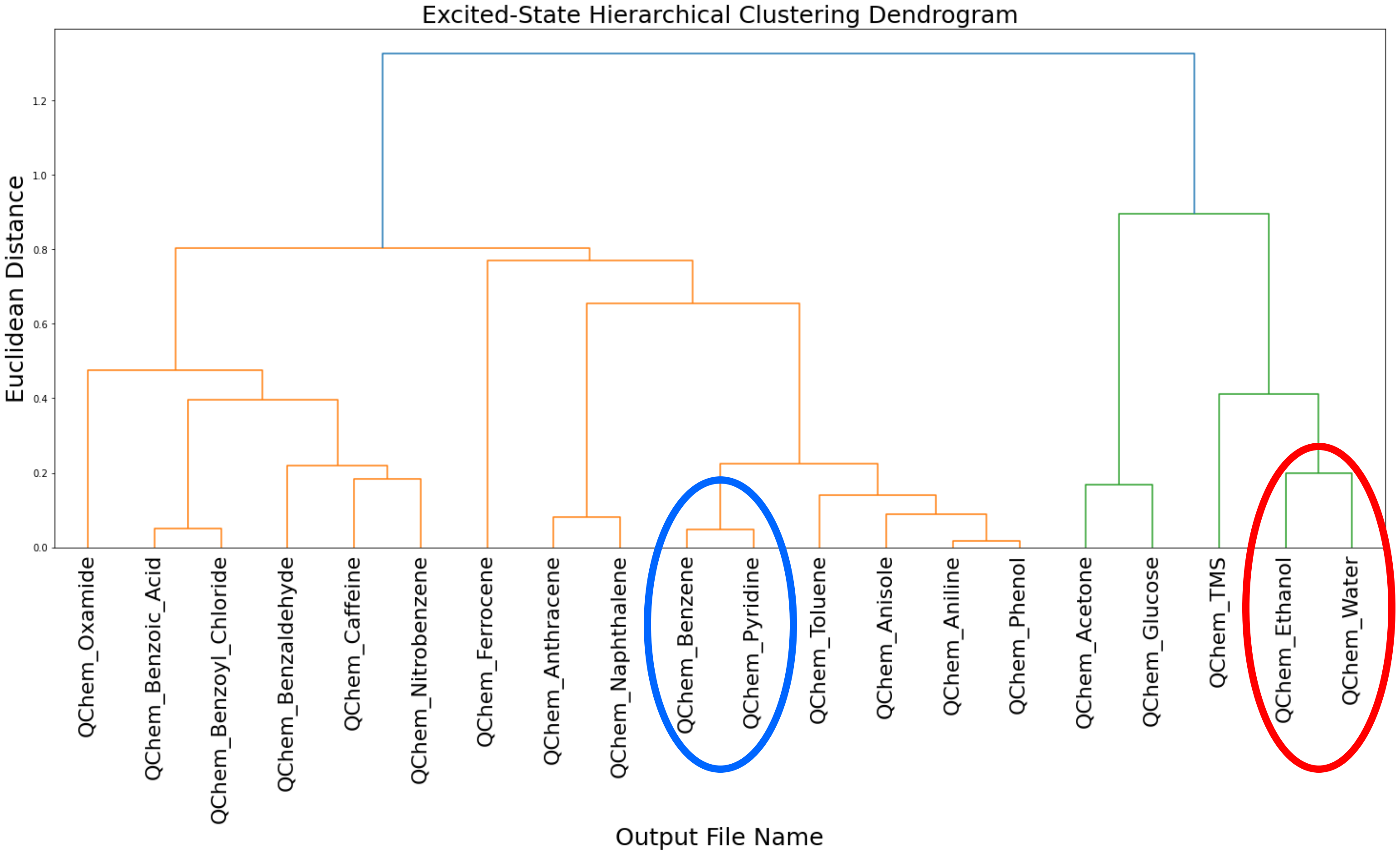
Z-Matrix Legend	Acetone	Aniline	Anisole	Anthracene	Benzaldehyde	Benzene	Benzoic Acid	Benzoyl Chloride	Caffeine	Ethanol	Ferrocene	Glucose (linear)	Naphthalene	Nitrobenzene	Oxamide	Phenol	Pyridine	TMS	Toluene	Water
Acetone		▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Aniline			▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Anisole				▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Anthracene					▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Benzaldehyde						▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Benzene							▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Benzoic Acid								▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Benzoyl Chloride									▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Caffeine										▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Ethanol											▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Ferrocene												▨	▨	▨	▨	▨	▨	▨	▨	▨
Glucose (linear)													▨	▨	▨	▨	▨	▨	▨	▨
Naphthalene														▨	▨	▨	▨	▨	▨	▨
Nitrobenzene															▨	▨	▨	▨	▨	▨
Oxamide																▨	▨	▨	▨	▨
Phenol																	▨	▨	▨	▨
Pyridine																		▨	▨	▨
TMS																			▨	▨
Toluene																				▨
Water																				

CIS Method Dendrogram

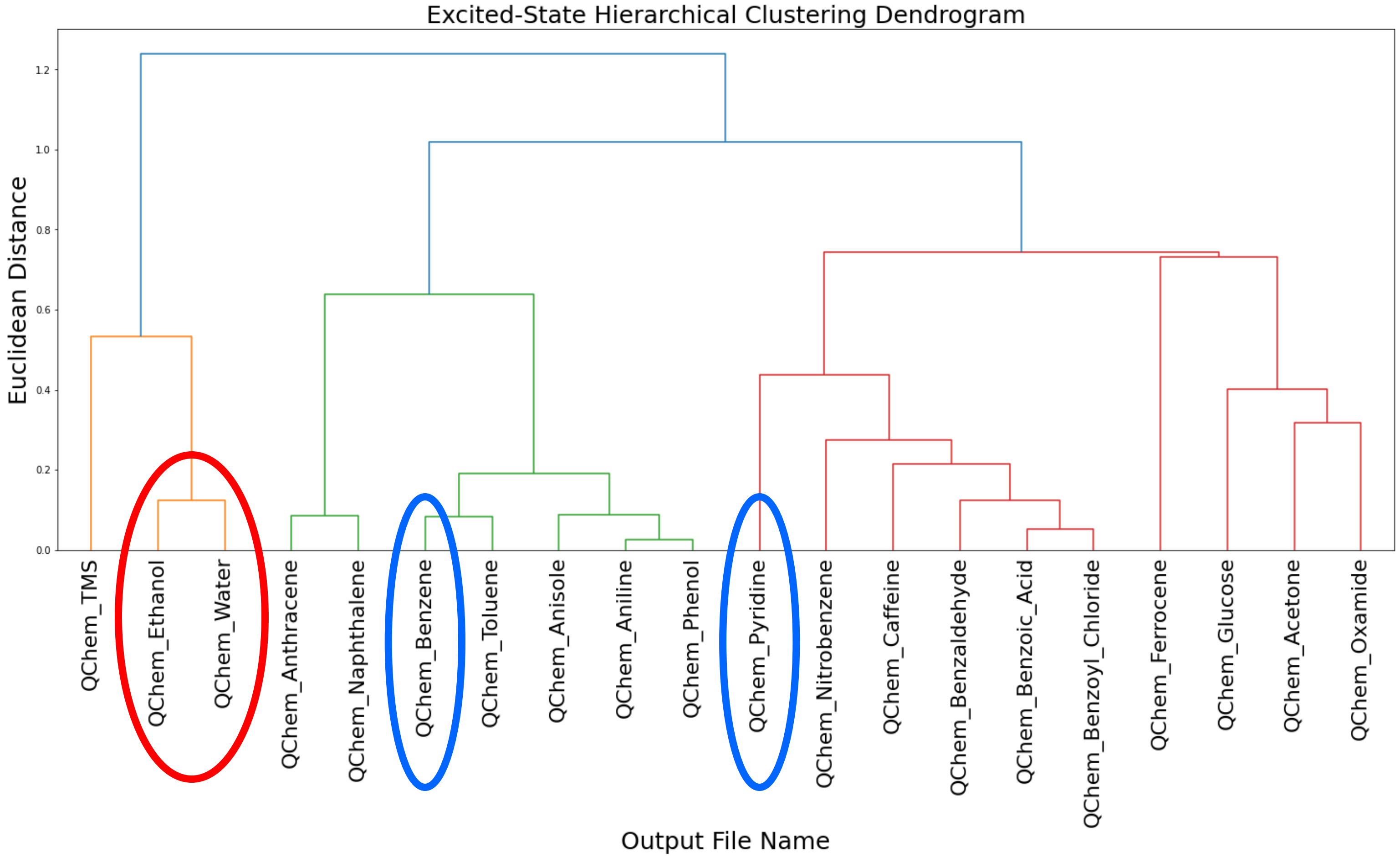
Dendrogram Parameters

Factor Combination:
Average of OS + E

State Combination:
60% Ex State 1
30% Ex State 2
10% Ex State 3



TDA Method Dendrogram



Dendrogram Parameters

Factor Combination:
Average of OS + E

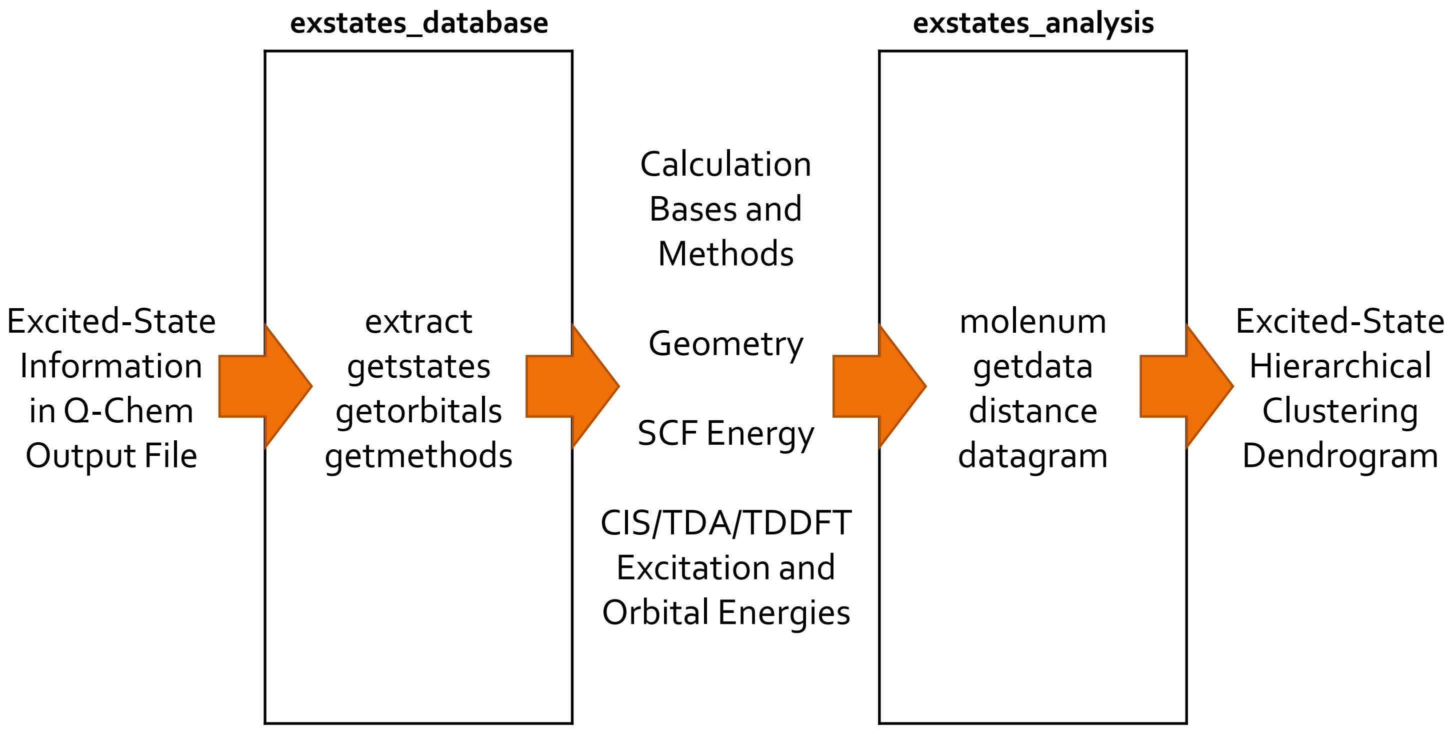
State Combination:
60% Ex State 1
30% Ex State 2
10% Ex State 3

Reasoning for Final Parameters

- First Three Excited States (60/30/10)
  - Requires more energy to reach higher states
- 50/50 Weighing of Data Parameters
  - Not highly sensitive to changes
- TDA is More Stable and Accurate
  - Theory, Parameter Variation

Computational Analysis of Data

Computation Input/Output Diagram



Excited-State Information in Q-Chem Output File

exstates_database

extract
getstates
getorbitals
getmethods

Calculation Bases and Methods

Geometry

SCF Energy

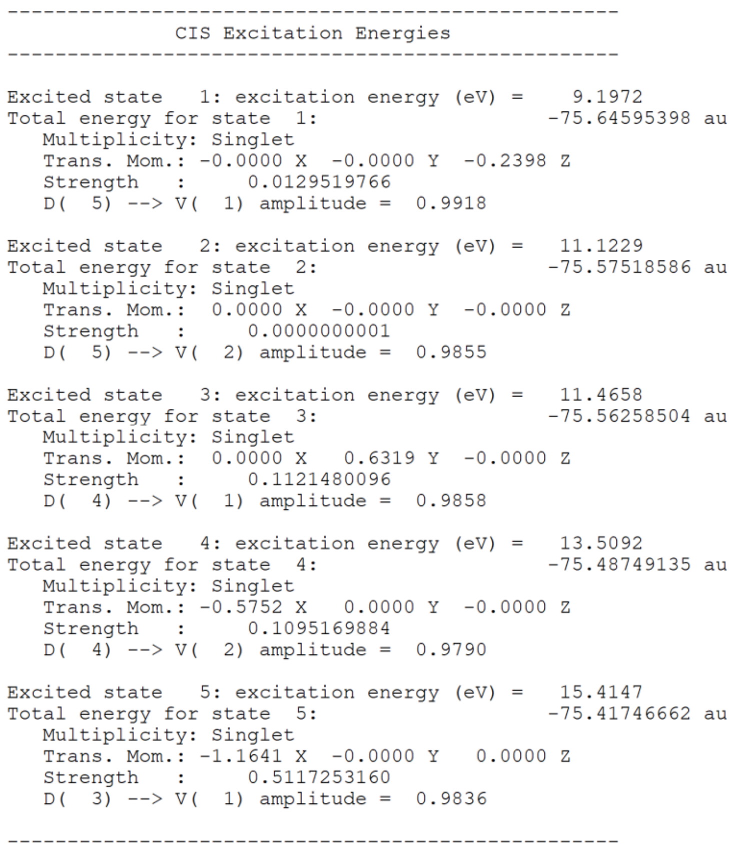
CIS/TDA/TDDFT Excitation and Orbital Energies

exstates_analysis

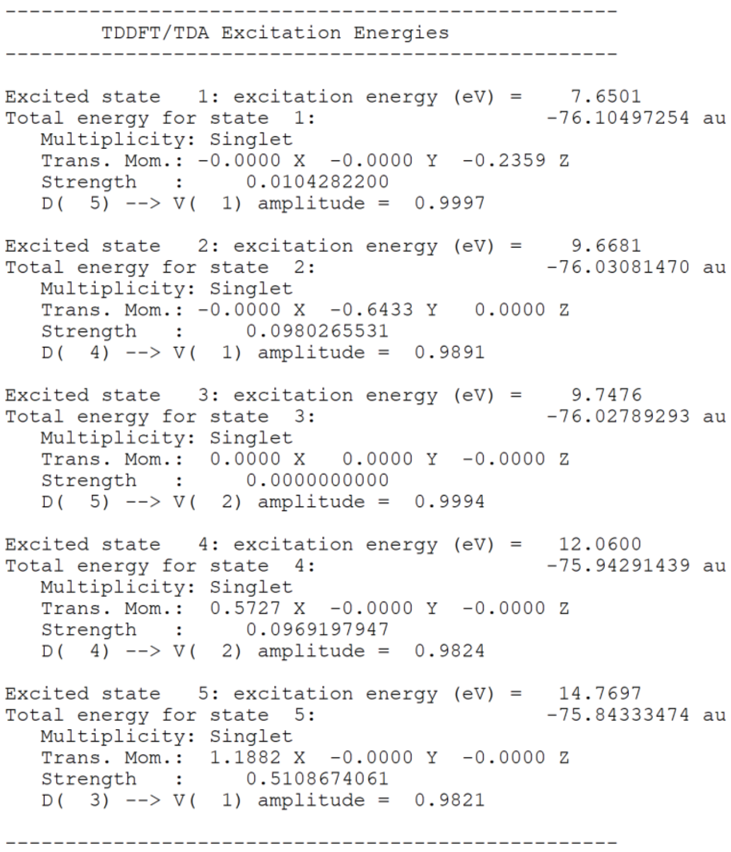
molenum
getdata
distance
datagram

Excited-State Hierarchical Clustering Dendrogram

Compression of Excitation Information



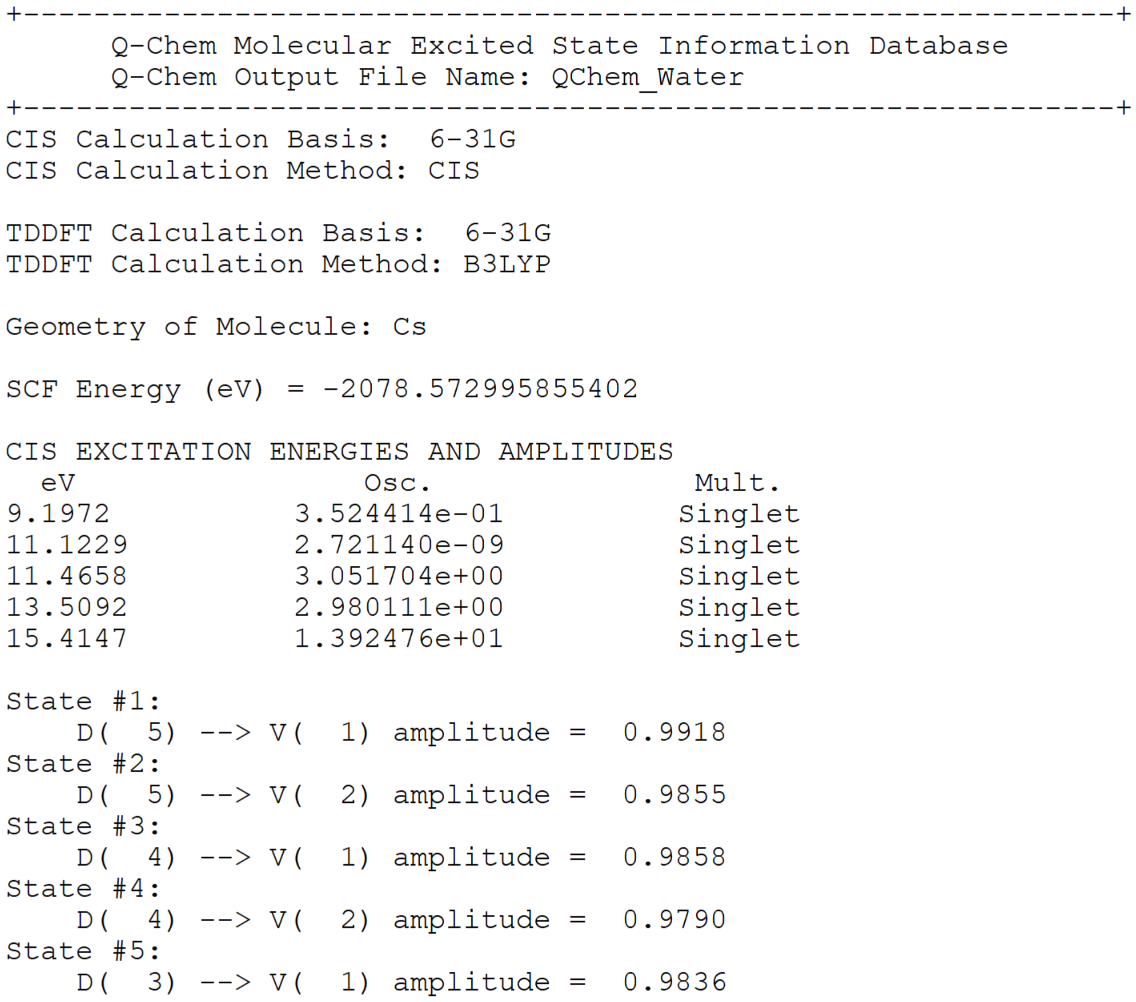
```
---------------------------------------------------
              CIS Excitation Energies
---------------------------------------------------

Excited state   1: excitation energy (eV) =    9.1972
Total energy for state  1:                  -75.64595398 au
   Multiplicity: Singlet
   Trans. Mom.: -0.0000 X  -0.0000 Y  -0.2398 Z
   Strength   :     0.0129519766
   D(  5) --> V(  1) amplitude =  0.9918

Excited state   2: excitation energy (eV) =   11.1229
Total energy for state  2:                  -75.57518586 au
   Multiplicity: Singlet
   Trans. Mom.:  0.0000 X  -0.0000 Y  -0.0000 Z
   Strength   :     0.0000000001
   D(  5) --> V(  2) amplitude =  0.9855

Excited state   3: excitation energy (eV) =   11.4658
Total energy for state  3:                  -75.56258504 au
   Multiplicity: Singlet
   Trans. Mom.:  0.0000 X   0.6319 Y  -0.0000 Z
   Strength   :     0.1121480096
   D(  4) --> V(  1) amplitude =  0.9858

Excited state   4: excitation energy (eV) =   13.5092
Total energy for state  4:                  -75.48749135 au
   Multiplicity: Singlet
   Trans. Mom.: -0.5752 X   0.0000 Y  -0.0000 Z
   Strength   :     0.1095169884
   D(  4) --> V(  2) amplitude =  0.9790

Excited state   5: excitation energy (eV) =   15.4147
Total energy for state  5:                  -75.41746662 au
   Multiplicity: Singlet
   Trans. Mom.: -1.1641 X  -0.0000 Y   0.0000 Z
   Strength   :     0.5117253160
   D(  3) --> V(  1) amplitude =  0.9836

---------------------------------------------------
```



```
---------------------------------------------------
        TDDFT/TDA Excitation Energies
---------------------------------------------------

Excited state   1: excitation energy (eV) =    7.6501
Total energy for state  1:                  -76.10497254 au
   Multiplicity: Singlet
   Trans. Mom.: -0.0000 X  -0.0000 Y  -0.2359 Z
   Strength   :     0.0104282200
   D(  5) --> V(  1) amplitude =  0.9997

Excited state   2: excitation energy (eV) =    9.6681
Total energy for state  2:                  -76.03081470 au
   Multiplicity: Singlet
   Trans. Mom.: -0.0000 X  -0.6433 Y   0.0000 Z
   Strength   :     0.0980265531
   D(  4) --> V(  1) amplitude =  0.9891

Excited state   3: excitation energy (eV) =    9.7476
Total energy for state  3:                  -76.02789293 au
   Multiplicity: Singlet
   Trans. Mom.:  0.0000 X   0.0000 Y  -0.0000 Z
   Strength   :     0.0000000000
   D(  5) --> V(  2) amplitude =  0.9994

Excited state   4: excitation energy (eV) =   12.0600
Total energy for state  4:                  -75.94291439 au
   Multiplicity: Singlet
   Trans. Mom.:  0.5727 X  -0.0000 Y  -0.0000 Z
   Strength   :     0.0969197947
   D(  4) --> V(  2) amplitude =  0.9824

Excited state   5: excitation energy (eV) =   14.7697
Total energy for state  5:                  -75.84333474 au
   Multiplicity: Singlet
   Trans. Mom.:  1.1882 X  -0.0000 Y  -0.0000 Z
   Strength   :     0.5108674061
   D(  3) --> V(  1) amplitude =  0.9821

---------------------------------------------------
```

Length of Output File = ~1400 Lines

Compression of Excitation Information



```
+--------------------------------------------------------------+
      Q-Chem Molecular Excited State Information Database
      Q-Chem Output File Name: QChem_Water
+--------------------------------------------------------------+
CIS Calculation Basis:  6-31G
CIS Calculation Method: CIS

TDDFT Calculation Basis:  6-31G
TDDFT Calculation Method: B3LYP

Geometry of Molecule: Cs

SCF Energy (eV) = -2078.572995855402

CIS EXCITATION ENERGIES AND AMPLITUDES
  eV                  Osc.                  Mult.
9.1972          3.524414e-01          Singlet
11.1229         2.721140e-09          Singlet
11.4658         3.051704e+00          Singlet
13.5092         2.980111e+00          Singlet
15.4147         1.392476e+01          Singlet

State #1:
    D(  5) --> V(  1) amplitude =  0.9918
State #2:
    D(  5) --> V(  2) amplitude =  0.9855
State #3:
    D(  4) --> V(  1) amplitude =  0.9858
State #4:
    D(  4) --> V(  2) amplitude =  0.9790
State #5:
    D(  3) --> V(  1) amplitude =  0.9836
```

Length of Database Entry = 93 Lines

Parameter Optimization Formulas

S–Matrix Formula:

$$\|\mathbb{S}\| = \sqrt{(c_1 s_1 + c_2 s_2 + c_3 s_3)^2}$$

Description of Variables:

- c_x = State Coefficient
- s_x = Excited State

For this project:

- Constant State Coefficients
  - $c_1 = 0.60$
  - $c_2 = 0.30$
  - $c_3 = 0.10$
- Based on MA Dendrograms

Parameter Optimization Formulas

Parameters:

$$s_x^2 = a_1^2 e_x^2 + a_2^2 o_x^2 + a_3^2 n_x^2$$

$$e_x = 1 - e^{-\frac{|e_A - e_B|}{k_B T}}$$

$$o_x = |o_A - o_B|$$

$$n_x = \left|\frac{1}{n_A} - \frac{1}{n_B}\right|$$

Description of Variables:

- s_x = Excited State
- a_x = Element Coefficient
- e_x = Excitation Energy
- o_x = Oscillator Strength
- n_x = Number of Transitions

Parameter Optimization Formulas

Euclidean Distance:

$$d = \sqrt{\sum (\mathbb{Z} - \mathbb{S})^2}$$

Description of Variables:

- $\mathbb{Z}$ = Z-Matrix Cell Value
- $\mathbb{S}$ = S-Matrix Cell Value

- Differences between Each Respective Molecular Pair
- Measured the Accuracy of Experiment Dendrograms
- Best Dendrogram Tracked during Program Execution

Optimization Results (TDA)

Parameter Distribution:

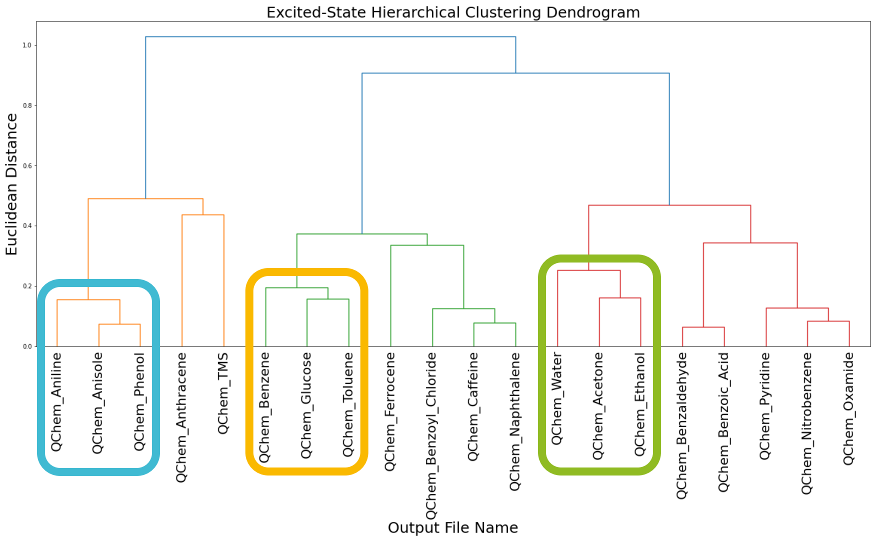
$c_1 = 0.60 \quad a_1 = 0.62$

$c_2 = 0.30 \quad a_2 = 0.55$

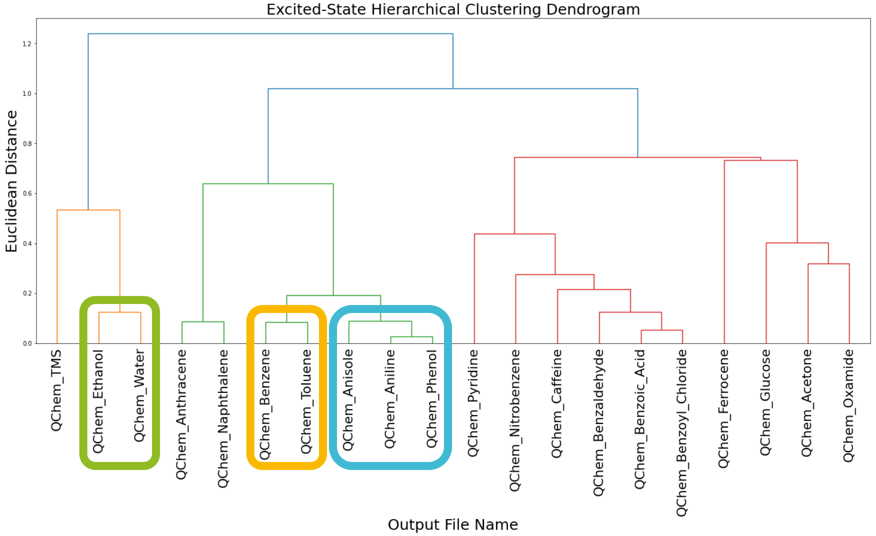
$c_3 = 0.10 \quad a_3 = 0.57$

$$a_1^2 + a_2^2 + a_3^2 = 1$$

Optimized Parameters Dendrogram



Manual Analysis Dendrogram



Application of Current Program

Complete Training Set Dendrogram

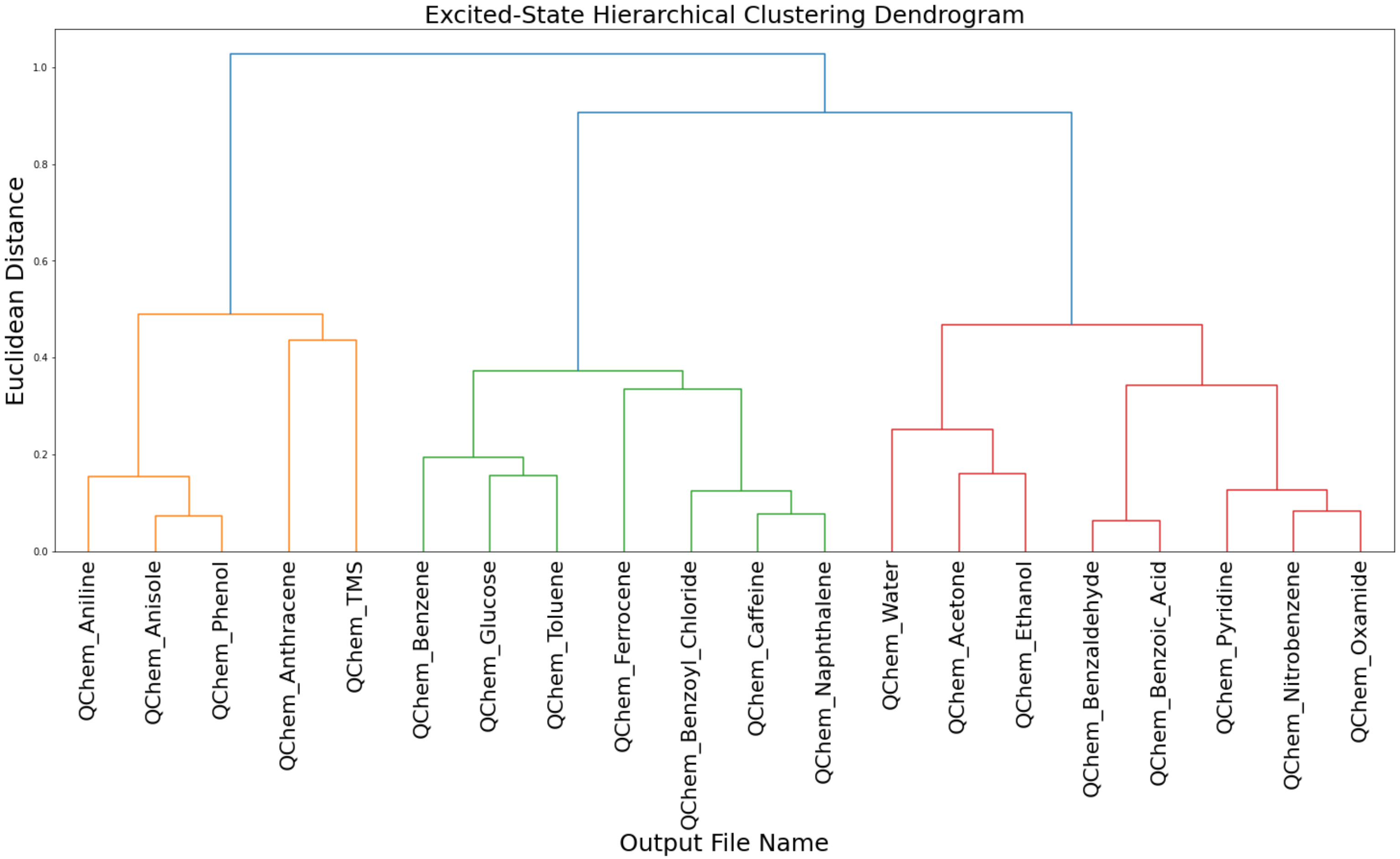
Parameter Distribution:

$c_1 = 0.60 \quad a_1 = 0.62$

$c_2 = 0.30 \quad a_2 = 0.55$

$c_3 = 0.10 \quad a_3 = 0.57$

$$a_1^2 + a_2^2 + a_3^2 = 1$$



Aromatic Training Set Dendrogram

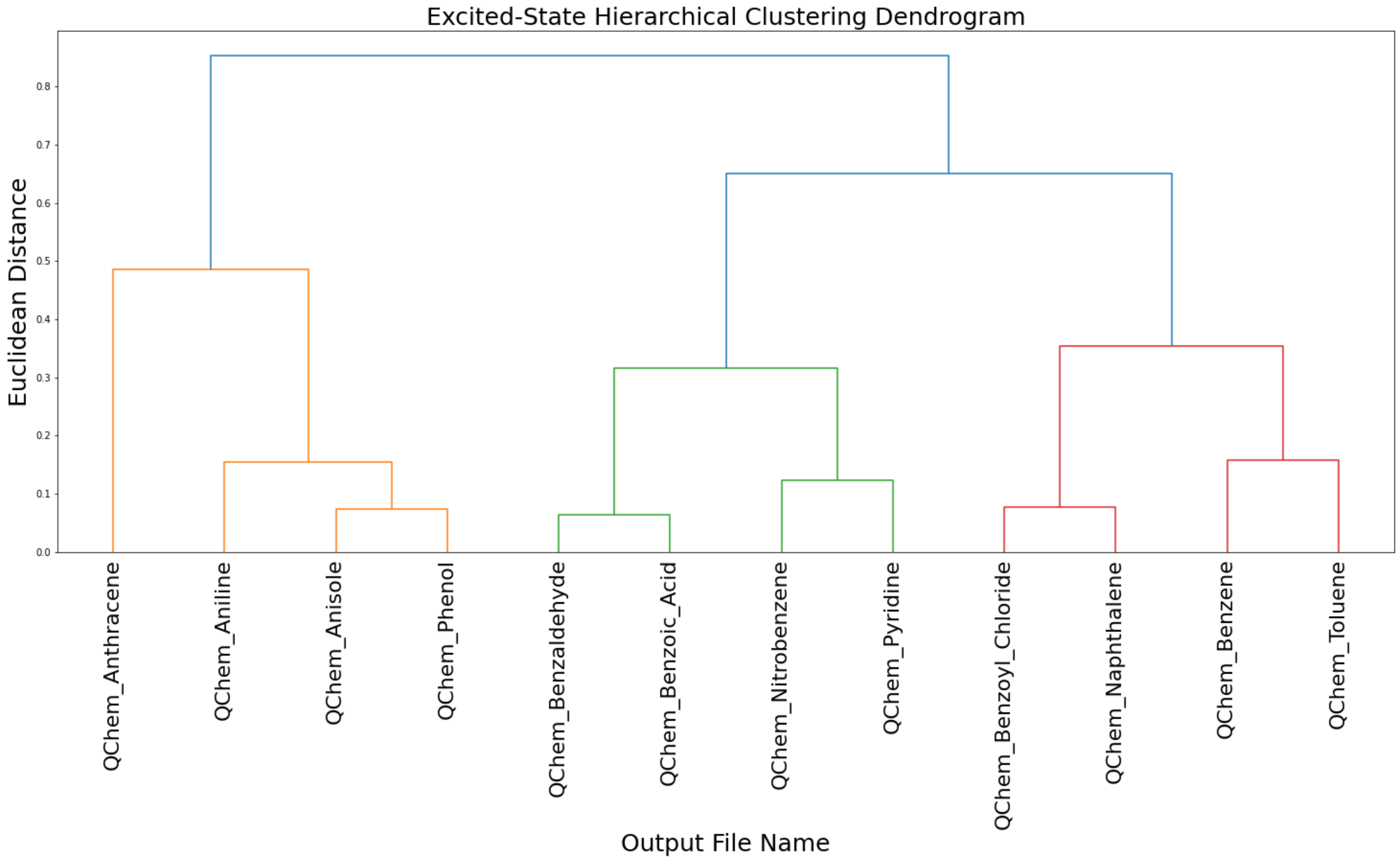
Parameter Distribution:

$c_1 = 0.60 \quad a_1 = 0.62$

$c_2 = 0.30 \quad a_2 = 0.55$

$c_3 = 0.10 \quad a_3 = 0.57$

$$a_1^2 + a_2^2 + a_3^2 = 1$$



Phenylureas Testing Set Dendrogram

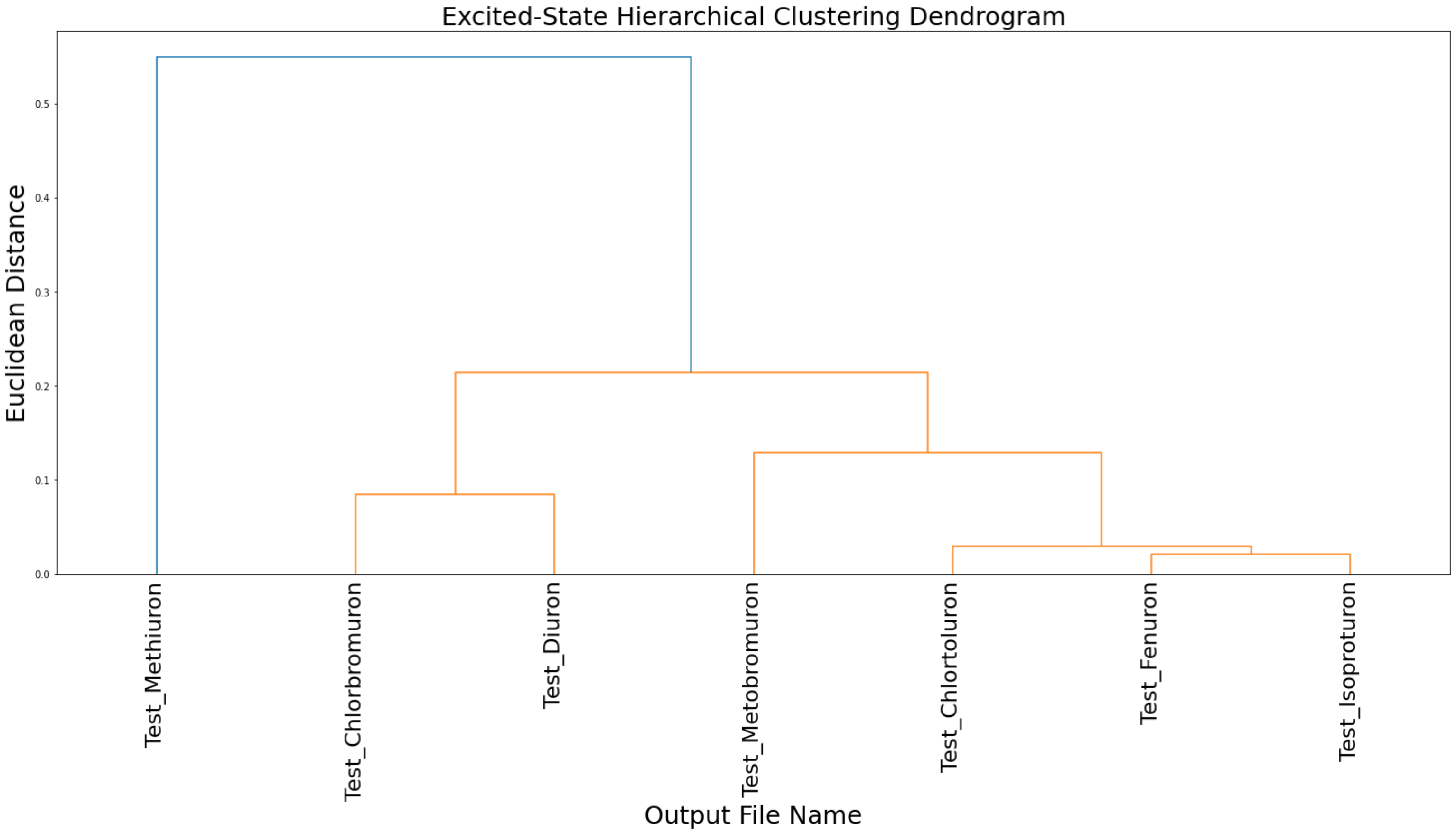
Parameter Distribution:

$c_1 = 0.60 \quad a_1 = 0.62$

$c_2 = 0.30 \quad a_2 = 0.55$

$c_3 = 0.10 \quad a_3 = 0.57$

$$a_1^2 + a_2^2 + a_3^2 = 1$$



Amines Testing Set Dendrogram

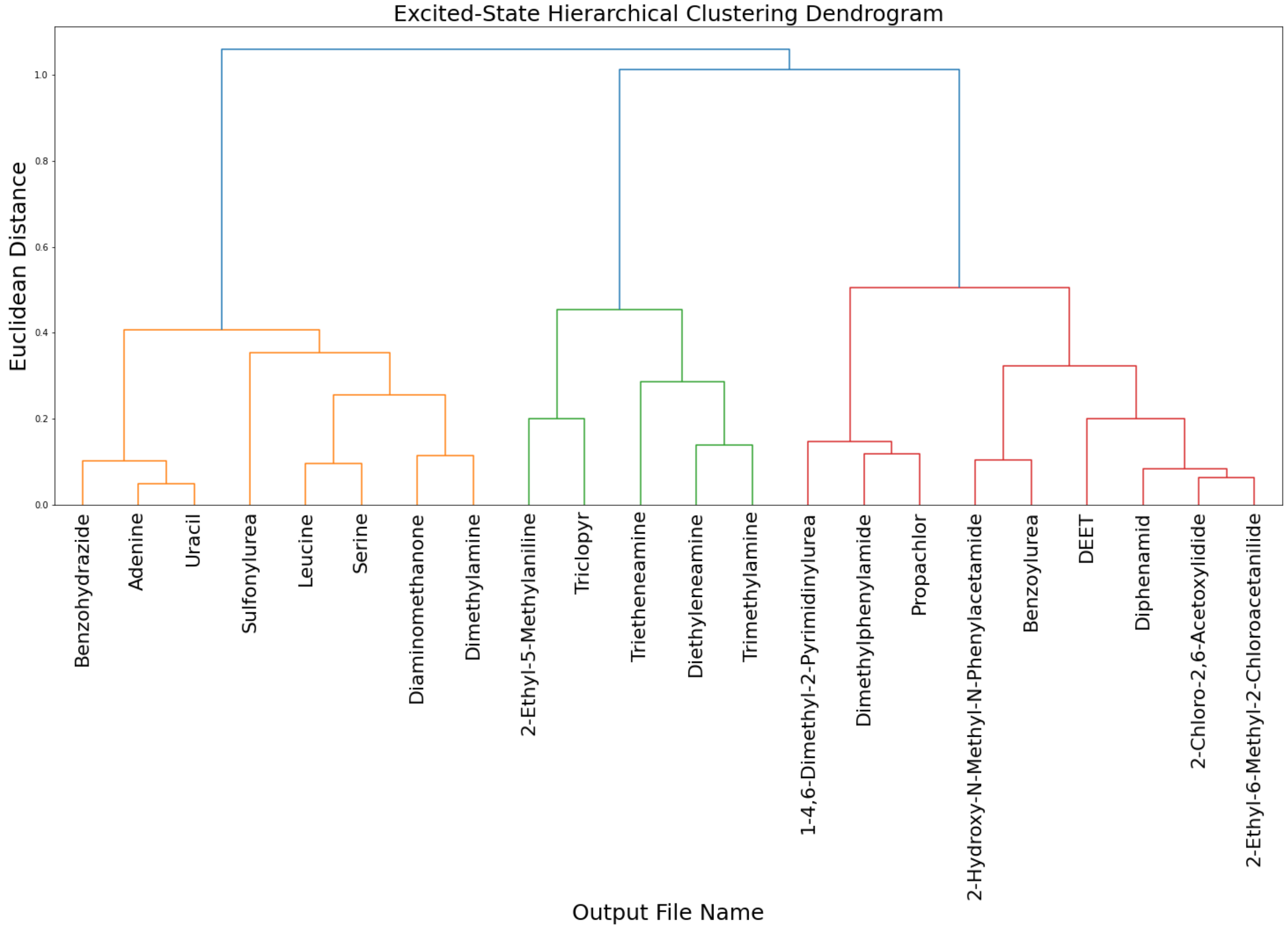
Parameter Distribution:

$c_1 = 0.60 \quad a_1 = 0.62$

$c_2 = 0.30 \quad a_2 = 0.55$

$c_3 = 0.10 \quad a_3 = 0.57$

$$a_1^2 + a_2^2 + a_3^2 = 1$$

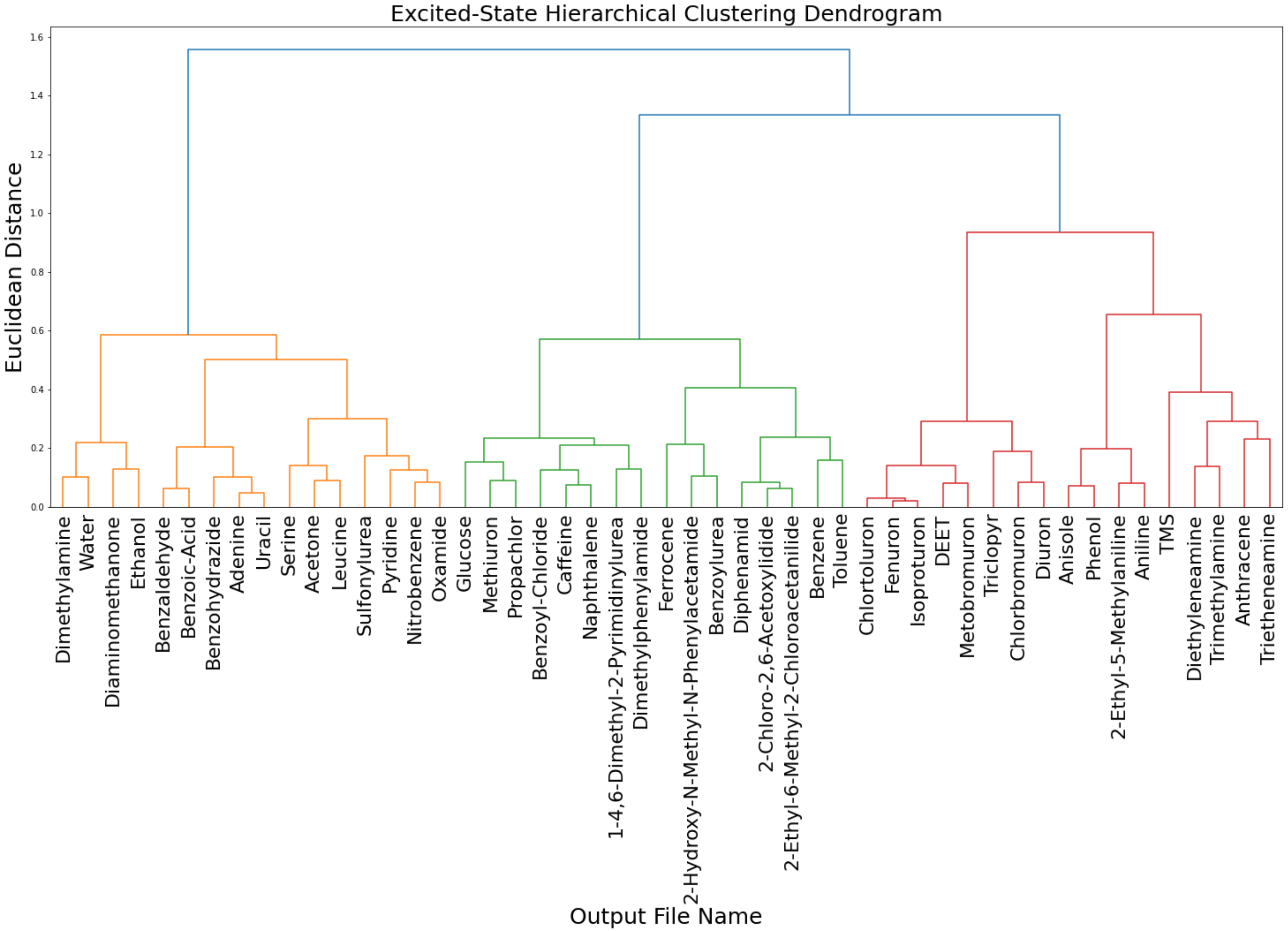


All Molecules Dendrogram

Parameter Distribution:

$$c_1 = 0.60 \quad a_1 = 0.62$$
$$c_2 = 0.30 \quad a_2 = 0.55$$
$$c_3 = 0.10 \quad a_3 = 0.57$$

$$a_1^2 + a_2^2 + a_3^2 = 1$$



Conclusion and Future Plans

Conclusion and Future Plans

- Manual clustering based on molecular orbitals and excitation energies used to generate reference data
- Automated clustering model developed that avoids visualization of molecular orbitals
  - Clustering optimized within model
  - Some clustering similarities vs. manual analysis
  - Additional parameters such as orbital energies should be incorporated to improve model
- Additional testing of molecules, basis sets, and methods

Thank You For Listening

Acknowledgments

Any Questions?