Constructing a database of asphaltenes: Quantum chemistry to contextualize single molecule experiments within ensemble properties

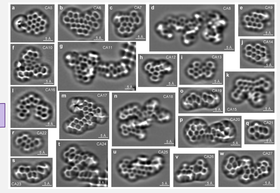
Gretel Stokes, Sydney Mazat, Benjamin Janesko*

Texas Christian University, Fort Worth, Texas

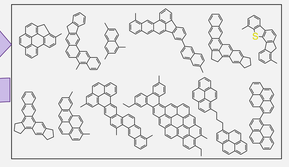
Introduction

Asphaltenes are the heaviest and most complex constituents of crude oil, with thousands of distinct species. Despite their structural diversity, which hampers complete understanding, modern studies have identified many asphaltene structures using atomic force microscopy. To address the challenge of storing and analyzing this information, we've created a database of 67 published asphaltene structures. Quantum chemistry calculations provide molecular properties such as weight, solubility, aromaticity, dipole moment, and HOMO-LUMO gap, which are stored in the database. Using this data, we generate graphs, like UV-visible absorbance spectra, to offer a comprehensive chemical description of asphaltene mixtures. Our computational predictions enhance understanding of individual asphaltene structures and their relationship to ensemble properties in crude oil.

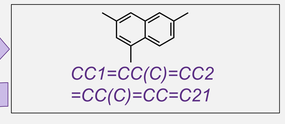
Constructing the Database



1. Single molecule imaging studies elucidate asphaltene chemical structures from three datasets: SCT28, CA12, and P27, where the number denotes the quantity of molecules in each dataset



2. The published molecules are converted to their SMILES strings



3. Descriptors are generated from SMILES strings

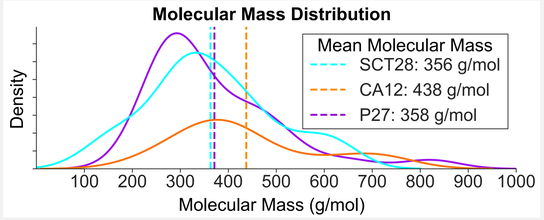
```
# create iupac name, molecular weight, and molecular formula list from smiles

iupac=[]
mw = []
formula = []
mol_withH = []
for smi in smile_list_SCT28:
    #iupac list
    compounds = pub.get_compounds(smi, namespace = 'smiles')
    match = compounds[0]
    iupac.append(match.iupac_name)
    #molecular weight list
    mol_noH = Chem.MolFromSmiles(smi)
    molwithH = Chem.AddHs(mol_noH)
    mol_withH.append(molwithH)
    mass = Descriptors.MolWt(molwithH)
    mw.append(mass)
    #molecular formula list
    form = CalcMolFormula(molwithH)
    formula.append(form)

# Generate lists with number of H, N, C, number of rings,
#  logP values, DBE, Z, and dipole moment columns for each molecule.

number_of_H = []
number_of_C = []
number_of_N = []
rings = []
logP = []
DBE_list =[]
Z_list = []
dipole_moment = []
HOMO_LUMO_gap_list = []

for mol in mol_withH:
```

4. Properties are stored in an SQL database

smiles_string	IUPAC_name	molar_mass	molecular...	number_of_H	number_of_C	number_of_N	number_of_rin...	LogP	DBE	Z	dipole_moment	HOMO_LUMO_Gap
CC1=C(C=CC2=C3C=...	18-methyl-41-thiaund...	547	C41H22S	< 22	41	0	11	12.36...	31	-60	0.308	0.4
CC1=C(C2=C3=C(C=C4C...	5-methylhexacyclo[11....	290	C23H14	< 14	23	0	6	8.636...	17	-32	0.037	1.33
CC1=C2C(C)(C=CC3=C...	None	387	C28H18S	< 18	28	0	7	8.72194	20	-38	0.258	1.41
CC1=CC2=C(C=C1)C3...	2-methylhpheny1ene	242	C19H14	< 14	19	0	4	5.454...	13	-24	0.041	3.02
CC1=CC2=CC=C3C4C...	None	432	C33H21N	< 21	33	1	9	8.215...	24	-45	0.045	1.45
C13=CC(C)(C)C2=C3=C4C(...	heptacyclo[11.11.1.14,...	328	C26H14	< 14	26	0	7	7.481...	20	-38	0.003	1.36
CC1=CC=CC2=C1C(C3...	1,7-dimethyl-9H-fluore...	194	C15H14	< 14	15	0	3	3.874...	9	-16	0.052	3.36
O=C1C2=CC=CC3=C2...	None	318	C24H14O	< 14	24	0	6	6.152...	18	-34	0.524	1.29
CC1=CC2=C3C(C4=C...	None	329	C25H13N	< 15	25	1	7	7.117...	19	-35	0.109	1.89
CC(C1=C2C3=CC4=C(...	None	356	C28H20	< 20	28	0	6	8.069...	19	-36	0.046	1.42
CC1=CC2=C(CC3=...	None	437	C32H20S	< 20	32	0	8	9.875...	23	-44	0.3	1.15
CC1=C2C3=C(C=C1)C...	None	278	C22H14	< 14	22	0	6	5.949...	16	-30	0.039	2.23
CC1=CC=CC=C2C3=C...	None	330	C26H18	< 18	26	0	6	7.025...	18	-34	0.032	1.96
CC1=CC(C)=CC2=CC(...	1,3,6-trimethylnaphtha...	170	C13H14	< 14	13	0	2	3.765...	7	-12	0.011	2.84
CC1=C2C(C=C(C=C3C...	None	457	C36H24	< 24	36	0	8	10.23...	25	-48	0.013	1.85
C12=CC3=CC(C)(C)C4=C...	None	344	C27H20	< 20	27	0	7	8.886...	18	-34	0.661	1.98
CC1=CC=CC2=C1SC3...	None	300	C21H16S	< 16	21	0	5	6.242...	14	-26	0.307	1.92
CC1=CC=CC2=C1C(C3...	None	627	C47H30S	< 30	47	0	12	13.80...	33	-64	0.166	1.04
CC1=C(C)C(=CC2)C=C...	1,6,7-trimethylnaphtha...	170	C13H14	< 14	13	0	2	3.765...	7	-12	0.057	2.81
CC1=CC2=C3C(C4=C(...	None	621	C47H24S	< 24	47	0	14	14.29...	36	-70	0.335	0.9
C(C(C=C1)=CC2=C1C13...	None	383	C30H22	< 22	30	0	7	8.136...	20	-38	0.055	2.01
CC1=C(C2C3=CC3=C4...	None	310	C24H22	< 22	24	0	5	6.478...	14	-26	0.019	2.06
CC1C2=CC3=C(C(C)=C...	None	589	C45H32O	< 32	45	0	11	11.08...	30	-58	0.477	0.99
CC1=C(C)C(=C(C)C=C1	1,2,4-trimethylbenzene	120	C9H12	< 12	9	0	1	2.611...	4	-6	0.022	4.28

Exploiting the Database

Mass Spectra: Asphaltenes exhibit molecular weights ranging from several hundred to several thousand g/mol. A plot that estimates the molecular mass distribution for each of the three datasets was generated, with masses recorded in the database being on the lower end of the spectrum.

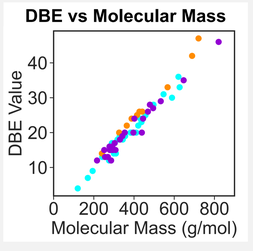


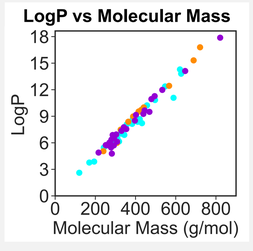
Double bond equivalents (DBE): a measure of unsaturation in organic molecules calculated by:

$$DBE = N_C - \frac{N_H}{2} + \frac{N_N}{2} + 1$$

For asphaltenes, DBE offers insight into their structural complexity and reactivity due to the presence of fused aromatic rings and aliphatic chains.

LogP: the logarithm of a molecule's octanol:water partition coefficient, which indicates relative solubility. Plotting logP against asphaltene molecular mass shows larger molecules are more hydrophobic, consistent with previous research. Predicted logP values span a wide range, signifying substantial variation in water solubility among asphaltenes.

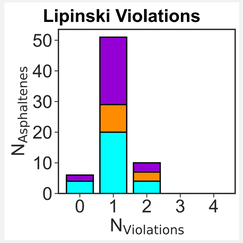




Druglikeness: Lipinski's Rule of Five is widely used in pharmaceuticals to gauge oral availability. For a molecule to be considered orally available, it shouldn't violate more than one of these criteria:

1. No more than five hydrogen bond donors
2. No more than ten hydrogen bond acceptors
3. Molecular mass less than 500 daltons
4. LogP not greater than five

The number of violations of each asphaltenes was computed. Most had one or fewer violations, suggesting they might be suitable drug candidates for oral delivery.



UV-Vis Spectral Modeling

The UV-Visible absorbance spectrum of the asphaltenes was modeled using computed values of the HOMO-LUMO gap and transition dipole moment.

HOMO-LUMO gap (E_{HL}):

- Energy in eV is converted to corresponding wavelength
- Determines wavelength where absorption is the highest

Transition dipole moment (μ):

- Determines intensity of absorption peak.

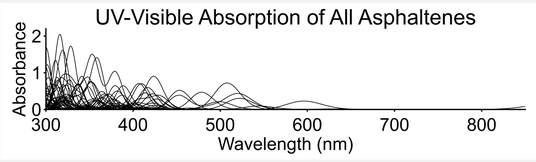
Thus, the absorption, A, is modeled by a Gaussian broadening function:

$$A(E) = \mu e^{-(E-E_{HL})^2/\sigma^2}$$

Where:

- E is the photon energy over which UV-visible spectra are plotted
- σ is a broadening factor, chosen to be 0.05, which determines how sharp the peaks appear.

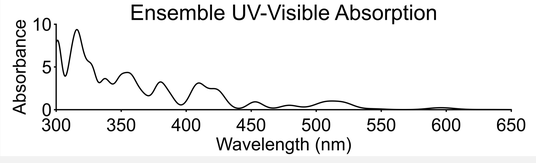
The following graph gives the plotted absorbance of each of the 67 asphaltene molecules:



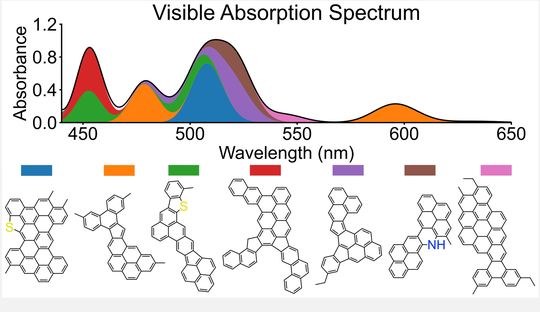
Summing the absorption of all asphaltenes and plotting against photon energy gives the following ensemble spectrum:

$$\sum_{n=1}^{67} A_n$$

Where A_n is the absorbance of each asphaltene



Most asphaltenes absorb in the 300 nm – 400 nm range. Focusing on the visible spectrum, we illustrate the spectral absorption of the seven most significant asphaltene structures.



Conclusions

Our study illuminates the complexity of asphaltenes in crude oil. Through computational analysis, we've cataloged 67 asphaltene structures in a SQL database and elucidated their predicted molecular properties. Our findings reveal insights into properties of asphaltenes such as molecular weight distribution, unsaturation, solubility, and drug-likeness, with many adhering to Lipinski's Rule of Five. By modeling UV-visible absorbance spectra based on molecular characteristics, we enhance understanding of asphaltene behavior. Our interdisciplinary approach integrates experimental data and computational modeling, offering valuable tools for further research in petroleum science.

Future Directions

Asphaltene mixtures are structurally complex, containing thousands of distinct structures. The upcoming phase involves employing machine learning techniques to generate predictive structures based on the properties of real asphaltenes stored in the database, advancing our understanding of crude oil composition and behavior.

AFM imaging elucidates the real structures of asphaltene molecules → Real structures are inputted into a workflow to predict properties → Machine learning is implemented to generate predictive structures

Acknowledgements

This work is supported by the American Chemical Society Petroleum Research Fund.

References

1. Schuler, B.; Meyer, G.; Peña, D.; Mullins, O. C.; Gross, L. Unraveling the Molecular Structures of Asphaltenes by Atomic Force Microscopy. *Journal of the American Chemical Society* **2015**, *137* (31), 9870-9876. DOI: 10.1021/jacs.5b04056.
2. Chen, P.; Metz, J. N.; Mennito, A. S.; Merchant, S.; Smith, S. E.; Siskin, M.; Rucker, S. P.; Dankworth, D. C.; Kushnerick, J. D.; Yao, N.; et al. Petroleum pitch: Exploring a 50-year structure puzzle with real-space molecular imaging. *Carbon* **2020**, *161*, 456-465. DOI: https://doi.org/10.1016/j.carbon.2020.01.062.
3. Zhang, Y.; Schuler, B.; Fatayer, S.; Gross, L.; Harper, M. R.; Kushnerick, J. D. Understanding the Effects of Sample Preparation on the Chemical Structures of Petroleum Imaged with Noncontact Atomic Force Microscopy. *Industrial & Engineering Chemistry Research* **2018**, *57* (46), 15935-15941. DOI: 10.1021/acs.iecr.8b03962.

Contact Information

gretel.jordan@tcu.edu
b.janesko@tcu.edu