Characterizing diesel oil composition using computational pentane-hexane data

Ralph Torres

1 Intro

This paper characterizes diesel oil's molecular composition by computationally generating and optimizing pentane-hexane conformers with varied orientations. Vibrational analysis yields infrared spectra which are compared to experimental terahertz time-domain spectroscopy data. The goal is to identify possible molecular structures of diesel oil based on spectral similarity acknowledging that pentane-hexane configurations are simplified models.

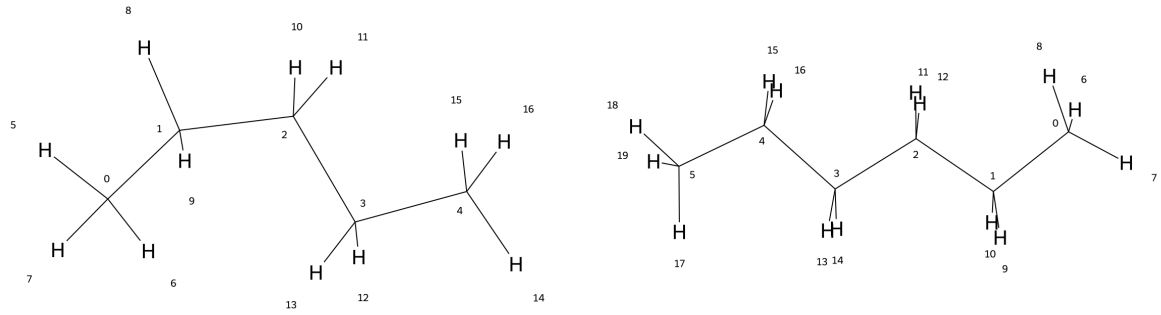

Figure 1: Computationally generated 3D conformers of pentane (left) and hexane (right).

2 Methods

This study employed a multi-step computation to investigate the structural and vibrational properties of pentane and hexane molecules in stacked configurations. The methodology encompassed conformer generation, rigorous geometry optimization using density functional theory (DFT), and subsequent vibrational analysis to simulate infrared (IR) and Raman spectra, particularly focusing on the terahertz (THz) region.

2.1 Initialize molecules and conformers

We started with the generation of conformers for pentane and hexane molecules. This was performed separately for each alkane using the `rdkit` cheminformatics toolkit. Following conformer generation, various molecular configurations were prepared such as individual molecules and pairs of identical molecules stacked on different molecules. The stacking arrangement was systematically varied by considering the number s of stacked molecules, the principal axis of stacking a $(\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2)$, and the intermolecular distance d between the adjacent molecules. An initial, less computationally demanding geometry optimization was performed on these configurations using the ETKDG method within `rdkit`. Representative molecular structures, corresponding to those depicted in Figure 2 and Figure 3, were visualized from these optimized conformers.

These generated conformers in xyz coordinates served as the input for the subsequent ab initio calculations. To facilitate these calculations, molecular objects compatible with `pyscf`, our chosen ab initio software package, were constructed. A key aspect of this study was the generation of two distinct datasets based on the choice of basis set. We imported the xyz

data from the conformers and then employed either the minimal `STO-3G` basis set or the more extensive `6-311++G(2d,2p)` basis set, which we will refer to as `sto3g` and `6311g`, respectively. This dual-basis set approach was chosen to compare the accuracy achievable with a minimal basis against a larger, more computationally expensive one, particularly for predicting properties relevant to non-covalent interactions.

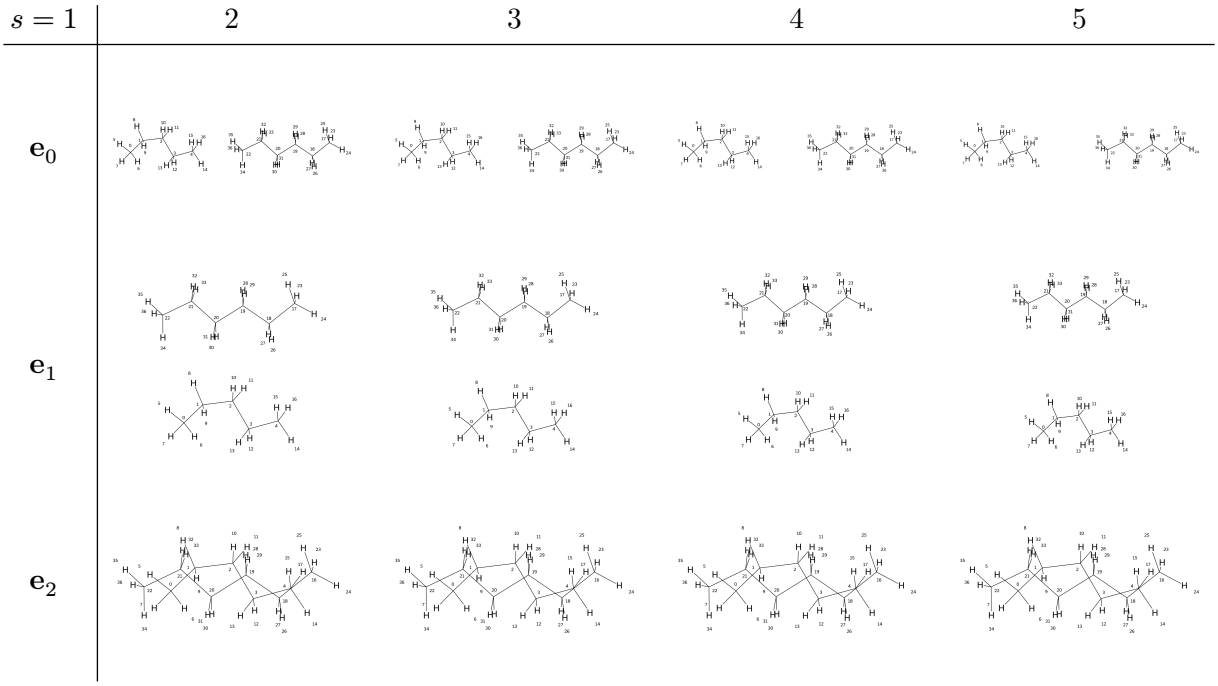

Figure 2: Pentane and hexane stacked $s = 1$ times along the axes $a = \{\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2\}$ with separation distances $d = \{2, 3, 4, 5\}$.

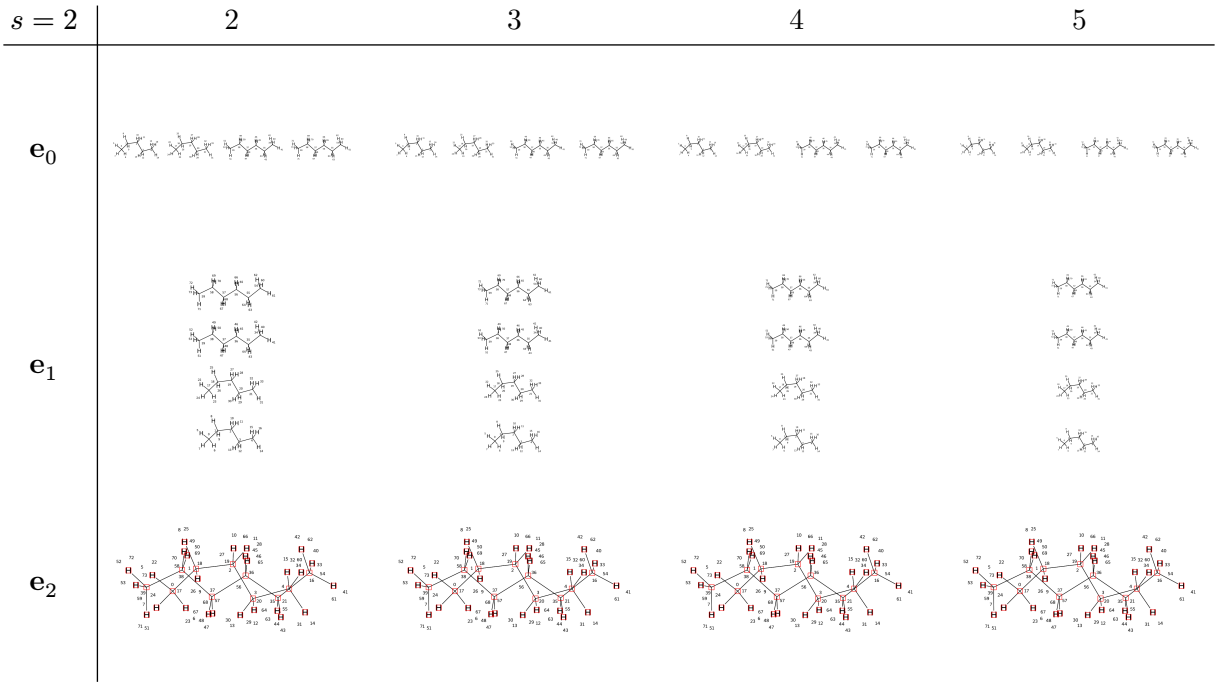

Figure 3: Pentane and hexane stacked $s = 2$ times along the axes $a = \{\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2\}$ with separation distances $d = \{2, 3, 4, 5\}$.

The selection of the `6311g` basis set was deliberate, given that the interactions in stacked alkanes are predominantly non-covalent van der Waals forces. For such interactions, basis sets require specific features.

1. Polarization functions. The `(2d,2p)` notation indicates the inclusion of d-type polarization functions on heavy non-hydrogen atoms and p-type polarization functions on hydrogen atoms. These are essential for accurately describing the distortions of electron clouds that occur during intermolecular interactions.
2. Diffusion functions. The `++` notation signifies the addition of diffusion functions on both heavy atoms and hydrogen atoms. Diffusion functions are crucial for describing the electron density far from the nucleus which is critical for modeling long-range interactions like van der Waals forces accurately.

While the `6-311++G(2d,2p)` basis set is relatively modest compared to those employed for high-accuracy quantum chemical calculations, its inclusion of both polarization and diffusion functions makes it suitable for capturing the essential physics of the systems under study. Future investigations with greater computational resources could explore even larger basis sets, such as correlation-consistent basis sets `aug-cc-pVDZ` or `aug-cc-pVTZ`, to further refine the results.

2.2 Optimize geometry

Following this, we performed a more rigorous geometry optimization for each configuration using the `geometric` optimizer interface with `pyscf`. The electronic structure calculations underpinning the optimization were carried out using DFT. Specifically, we employed the restricted Kohn-Sham (RKS) approach as the stacked alkane systems are expected to be closed-shell, that is all electrons are paired. RKS is generally the most computationally efficient and stable choice for such systems, assuming that alpha and beta spin electrons occupy the same set of spatial orbitals.

The choice of exchange-correlation (XC) functional was `B3LYP` augmented with the `D4` dispersion correction yielding `B3LYP-D4`. This selection is critical because alkane stacking is predominantly governed by van der Waals forces, specifically London dispersion forces. The `B3LYP` functional is a widely used hybrid functional, and the `D4` correction significantly enhances its ability to accurately model non-covalent interactions. While other functionals, such as the range-separated hybrid `ωB97X-D`, are also well-suited for non-covalent interactions and could be considered in future work, `B3LYP-D4` represents a robust general-purpose choice for this study.

In Kohn-Sham DFT (KS-DFT), as originally proposed by Kohn and Sham, the complex interacting electron system is mapped onto a fictitious system of non-interacting electrons that share the same ground-state electron density. This conceptual framework allows KS-DFT calculations to be computationally similar to Hartree-Fock (HF) theory, but with a modified effective potential. The total electronic energy in KS-DFT is expressed as

$$E = T_s + E_{\text{ext}} + E_J + E_{\text{XC}}$$

where T_s is the kinetic energy of the non-interacting reference system, E_{ext} is the energy due to the external potential from the nuclei, E_J is the classical coulomb repulsion or Hartree energy, and E_{XC} is the exchange-correlation energy. The E_{XC} term encapsulates all the many-body quantum mechanical effects and is approximated by a density functional.

The `geometric` optimizer iteratively refines the molecular geometry. In each step, it calculates the total energy and the forces, which are gradients, on each atom using the specified DFT

method, which is RKS with `B3LYP-D4` functional and the chosen basis set, for the current geometry. These forces are then used by the optimizer to predict atomic displacements that will lead to a lower energy structure. The atomic coordinates within the `pyscf` molecule object are updated, and the process is repeated until the geometry converges to a local minimum on the potential energy surface, or a predefined maximum number of optimization steps is reached. The mean-field (mf) object representing the RKS solution holds all necessary information including energy and gradients required by the optimizer.

2.3 Perform vibrational analysis

Subsequent to successful geometry optimization, we conducted vibrational analyses to predict IR and Raman spectra. An RKS object was again created using the `dft` module, specifying the `B3LYP-D4` exchange-correlation functional. To enhance computational efficiency, density fitting also known as resolution of identity (RI) was applied. This technique approximates the computationally expensive two-electron integrals using an auxiliary basis set, significantly reducing the cost of DFT calculations. For these demanding calculations, we leveraged GPU acceleration using an available NVIDIA A100 instance, which can substantially speed up the self-consistent field (SCF) iterations and subsequent Hessian computation.

After the SCF procedure converged, yielding the electronic energy, electron density, and molecular orbitals, a Hessian object associated with the mf object was created. The Hessian is crucial for characterizing stationary points on the potential energy surface; a true local minimum exhibits a positive definite Hessian where all eigenvalues are positive. Furthermore, the eigenvalues and eigenvectors of the mass-weighted Hessian matrix directly relate to the vibrational frequencies and normal modes of the molecule, respectively.

The vibrational frequencies and IR intensities were then calculated. This involved diagonalizing the mass-weighted Hessian matrix. The square roots of the resulting eigenvalues yield the vibrational frequencies. IR intensities were computed based on the changes in the molecular dipole moment during each normal mode of vibration, indicating how strongly each mode will absorb infrared radiation.

To facilitate comparison with experimental spectra, the discrete set of calculated IR frequencies (in wavenumbers, cm^{-1}) and their corresponding intensities were transformed into a continuous, broadened spectrum. This broadening was achieved by convolving each discrete vibrational peak with a Lorentzian lineshape function. The wavenumbers were also converted to frequencies in THz using the conversion factor

$$\frac{1}{\mathrm{cm}} \times 2.998 \times 10^8 \frac{\mathrm{m}}{\mathrm{s}} \times \frac{1\ \mathrm{THz}}{10^{12}\ \mathrm{Hz}} = 10^2 \times 2.998 \times 10^8 \times 10^{-12}\ \mathrm{THz} = 2.998 \times 10^{-2}\ \mathrm{THz}.$$

Intensities were subsequently normalized for presentation. The resulting spectral data files were named systematically using the format `ir-{stack}{axis}{distance}`, where `stack` refers to the number of times a component molecule is stacked onto itself before being paired with the stack of another component molecule, `axis` denotes the geometric stacking axis, and `distance` indicates the approximate intermolecular separation defined relative to C-H bond lengths.

2.4 Compare with experimental data

The simulated THz spectra obtained from these computational procedures, particularly for hexane and pentane, were then qualitatively and quantitatively compared with experimental terahertz time-domain spectroscopy (THz-TDS) data for components of diesel fuel, as reported by Ponceca et al. in their study on Kuwaiti diesel oil. This comparison aims to validate the

computational methodology and provide insights into the molecular origins of spectral features observed in complex hydrocarbon mixtures.

3 Results

This section presents a series of plots comparing experimental THz spectra of diesel components with simulated spectra of pentane and hexane conformers.

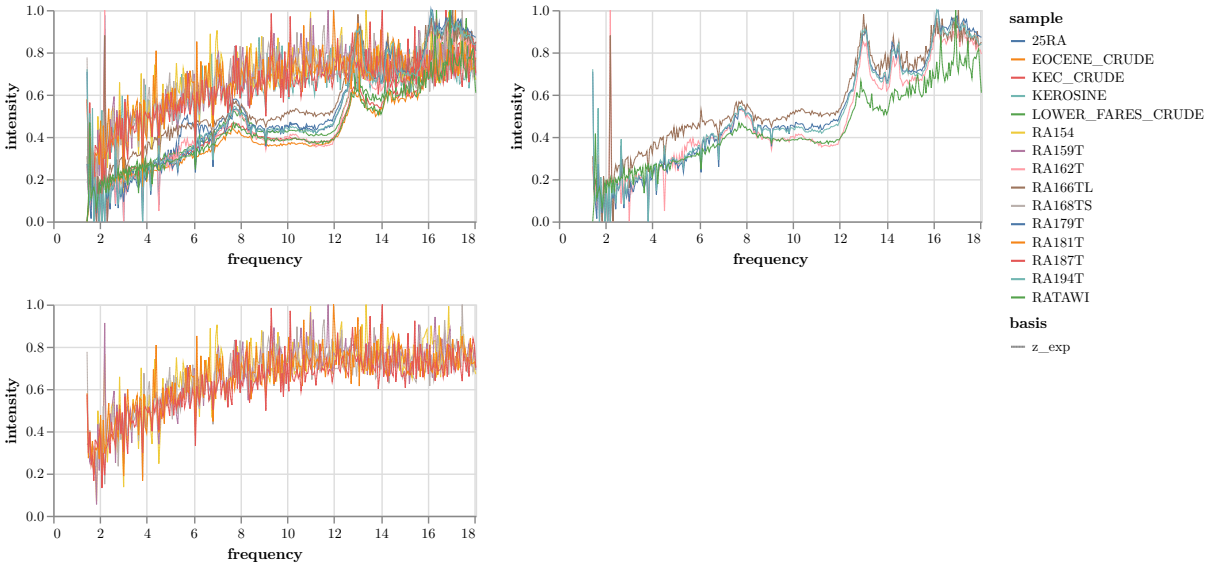

Figure 4: Experimental THz spectra for various RA series diesel oil samples.

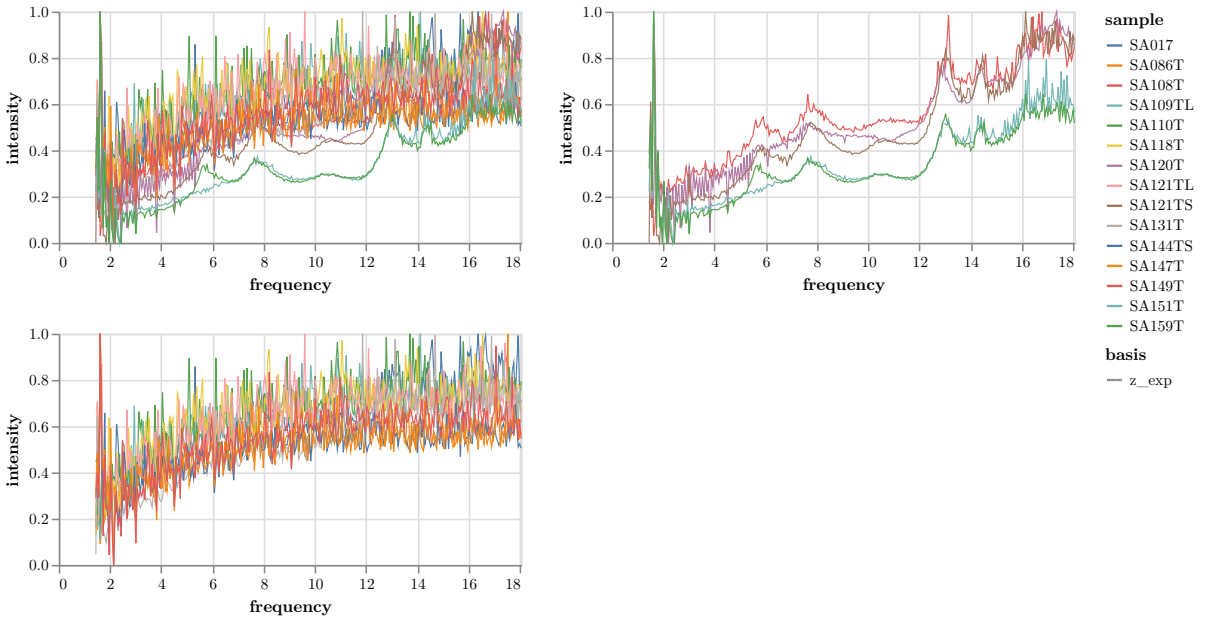

Figure 5: Experimental THz spectra for various SA series diesel oil samples.

The experimental THz spectra for various diesel oil samples, labeled as RA and SA series, consistently exhibit broad absorption features across the plotted frequency range at approximately 0-18 THz. While the overall shape is a broad continuum, there are discernible variations in intensity and the subtle contours of these features among the different experimental samples. Some samples in Figure 4 show more pronounced shoulders or peaks in the 12-16 THz region compared to others. Similar variability is seen in Figure 5. This suggests that while the general THz response of diesel components is characterized by broad absorption, the specific

composition or concentration of different molecular species within each sample leads to these observed variations.

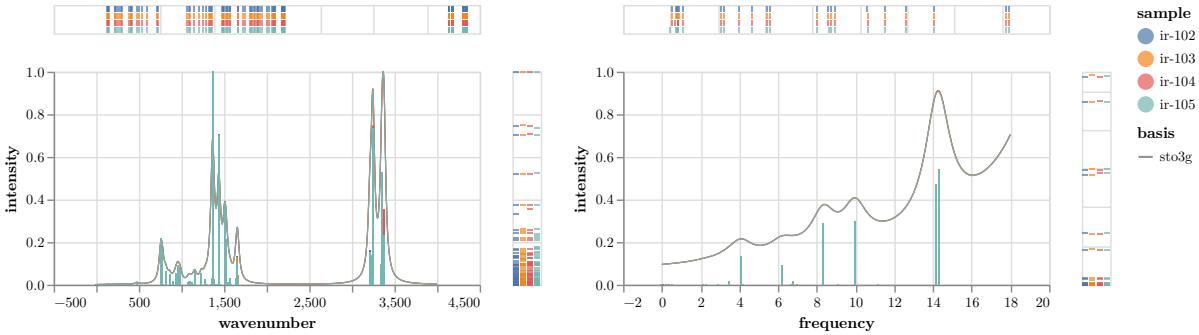

Figure 6: Simulated THz spectra for the `STO-3G` basis set showing intensity vs. wavenumber (left) and intensity vs. frequency in THz (right).

Simulations using the minimal `sto3g` basis set produce spectra with several relatively sharp, distinct peaks. The left plot in Figure 6 shows peaks in wavenumbers at around 700, 1400, and 3000 cm^{-1}. The right plot with frequency in THz shows features at around 4, 6, 8, and 10 Thz, and a dominant peak around 14 THz.

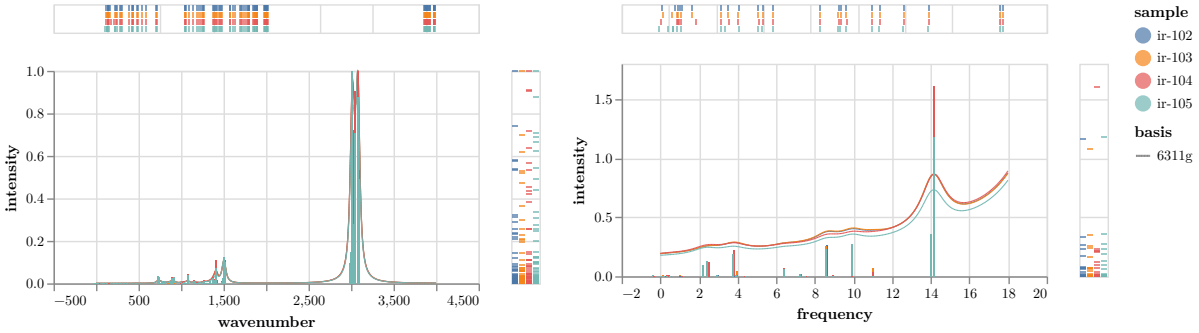

Figure 7: Simulated THz spectra for the `6-311++G(2d,2p)` basis set showing intensity vs. wavenumber (left) and intensity vs. frequency in THz (right).

The more extensive `6311g` basis set also yields spectra with sharp peaks as shown in Figure 7. In the wavenumber plot, prominent peaks are seen near 1500 cm^{-1} and a very strong set of peaks around 3000 cm^{-1}. The corresponding THz plot shows activity below 2 THz, distinct peaks around 4 and 8-10 THz, and a very intense, sharp peak at 14 THz.

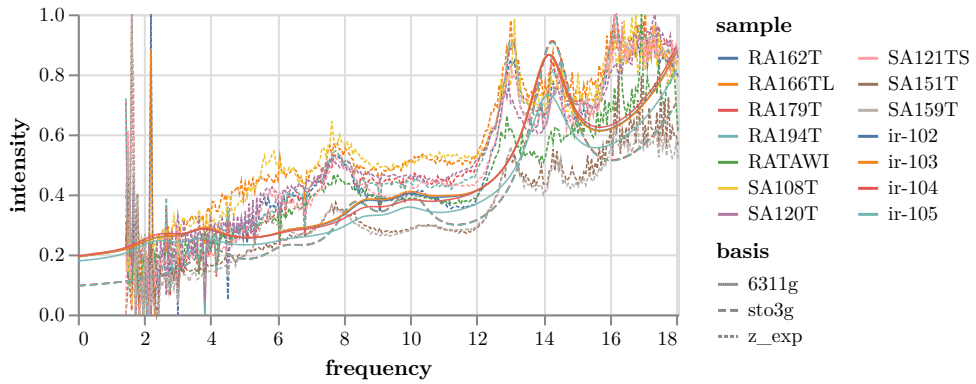

Figure 8: Comparison of experimental THz spectra with simulated spectra from `STO-3G` and `6-311++G(2d,2p)` basis sets for select configurations.

Figure 8 overlays experimental data with simulated spectra from both `sto3g` and `6311g` basis sets for select configurations from ir-102 to ir-105. Recall that spectral data were named systematically using the scheme `ir-{stack}{axis}{distance}`. The `6311g` simulations generally show higher intensity peaks, particularly the strong feature around 14 THz. Both basis sets predict activity in similar regions such as at low THz, 4-10 THz, and 14 THz. Note that we had to scale the spectra upon conversion.

Visually, the `6311g` simulations appear to align somewhat better with the regions where experimental data shows absorption, especially in capturing the higher frequency activity around 12-16 THz, although the experimental features are much broader. This is expected courtesy of this basis set's inclusion of polarization and diffusion functions. So we prefer this moving forward in the analysis as it better models the intermolecular interactions dominant in these systems, likely leading to more physically realistic though still simplified spectra.

The next figures explore the effect of stacking axis $(\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2)$ and the number of stacked molecules ($s = 1$ in Figure 9, $s = 2$ in Figure 10) on the simulated THz spectra.

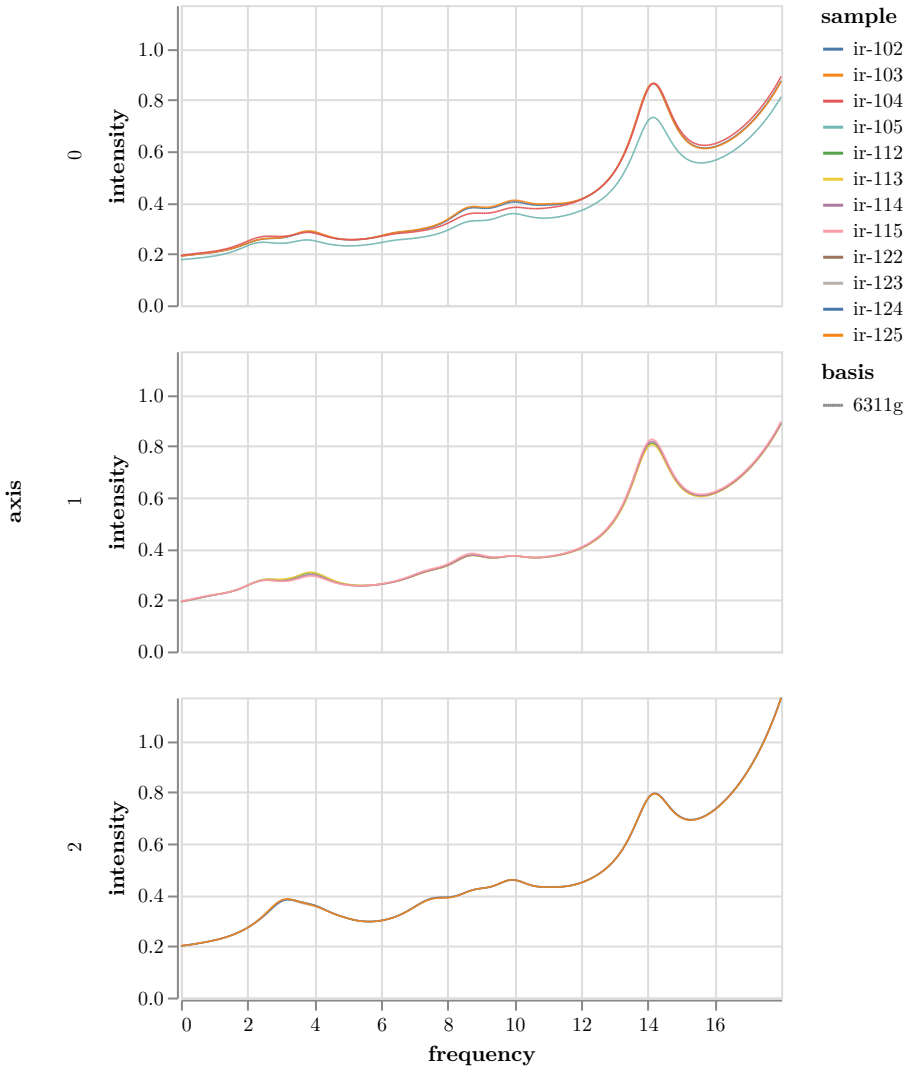

Figure 9: Spectral features from simulated THz data from the `6-311++G(2d,2p)` basis set for $s = 1$ stacks on various axis $(\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2)$.

For x-axis $(\mathbf{e}_0)$, the spectra are quite similar, showing a broad rise to a dominant peak around 14 THz, with minor features around 2-4 THz and 8-10 THz, as shown in Figure 9. For y-axis

($\mathbf{e}_1$), the spectra also show a dominant 14 THz peak, but the features in the 2-10 THz region are slightly more pronounced and varied among the different samples. For z-axis ($\mathbf{e}_2$), the 14 THz peak persists. The 2-10 THz region shows more distinct features, particularly a noticeable peak or shoulder developing around 3-4 THz and another around 8 THz, with more variability between samples.

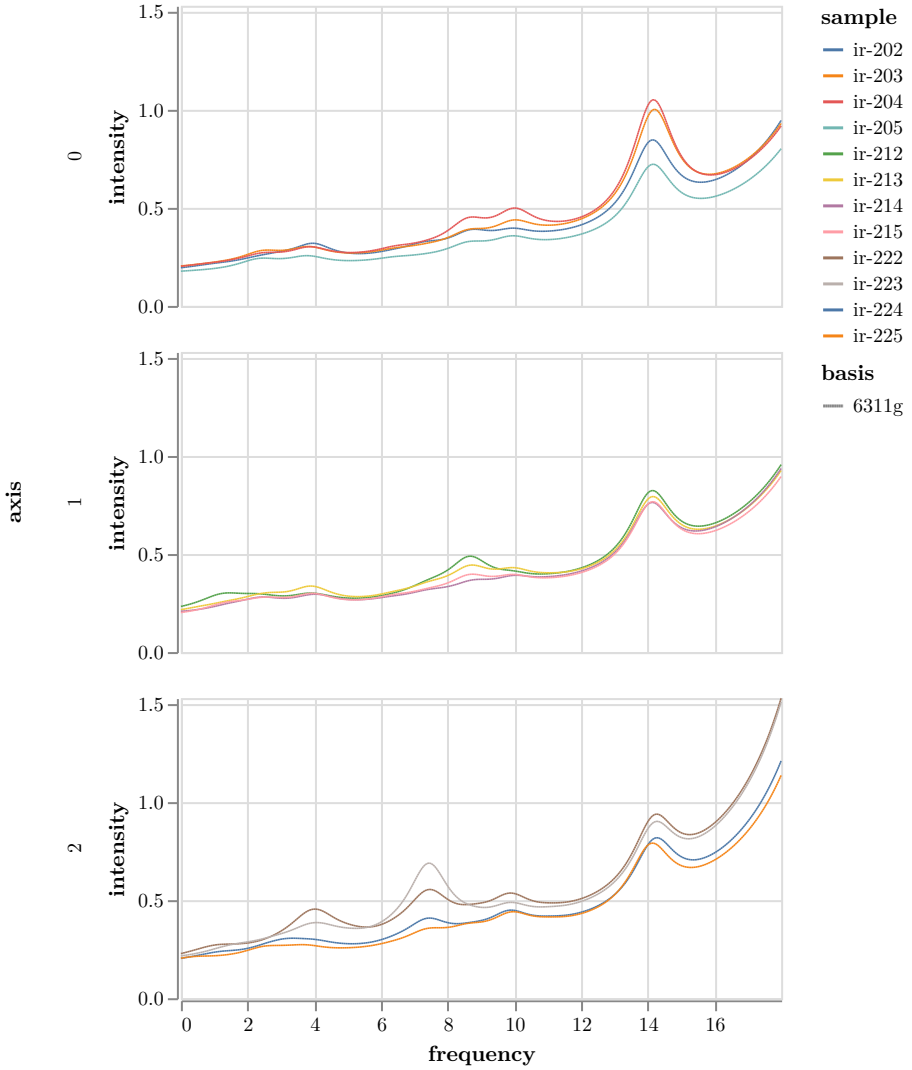

Figure 10: Spectral features from simulated THz data from the `6-311++G(2d,2p)` basis set for $s = 2$ stacks on various axis ($\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2$).

In Figure 10, trends are similar to that of $s = 1$ stacks. The x-axis ($\mathbf{e}_0$) shows the most consistent spectra among its samples. The y-axis ($\mathbf{e}_1$) shows slightly more variation in the mid-frequency range. The z-axis ($\mathbf{e}_2$) again shows the most complex features in the 2-10 THz range, with some configurations exhibiting sharper and more intense peaks around 7-9 THz compared to that of $s = 1$ stacks.

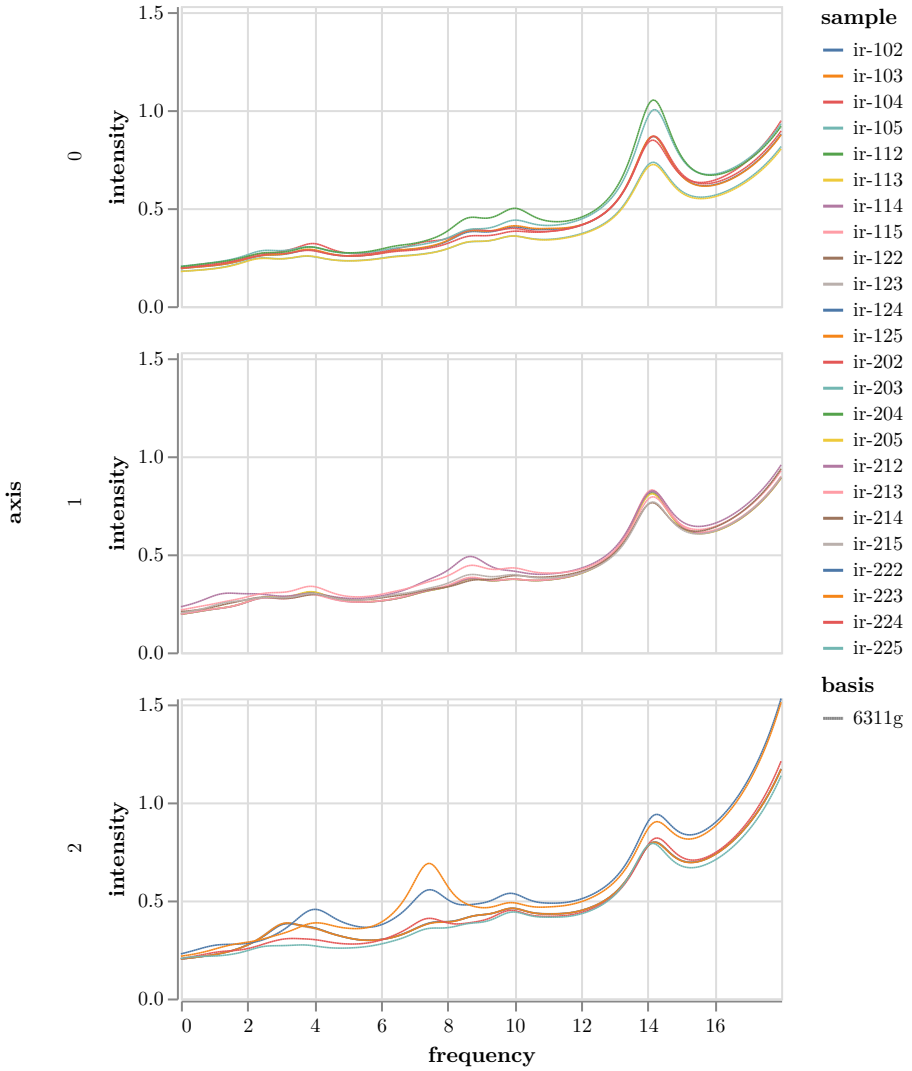

Figure 11: Spectral features from simulated THz data from the `6-311++G(2d,2p)` basis set for both $s = 1$ and $s = 2$ stacks on various axis $(\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2)$.

Combining both $s = 1$ and $s = 2$ stacking in Figure 11, the dominant 14 THz feature is present across all configurations and stack numbers. Stacking along $\mathbf{e}_2$ axis consistently produces more structured spectra in the 2-10 THz region compared to $\mathbf{e}_0$and $\mathbf{e}_1$. Increasing the stack number from $s = 1$ to $s = 2$ appears to slightly enhance or sharpen some features, particularly for $\mathbf{e}_2$ axis configurations in the 6-10 THz region.

Overall, we observe that the simulated THz spectra are sensitive to the specific 3D arrangement (stacking axis, intermolecular distance) and, to a lesser extent, the number of stacked units. The consistent appearance of features in certain frequency bands (2-5, 6-10, and 14 THz) across many configurations, which are also present in the experimental THz-TDS data of diesel oil, suggests these could be characteristic vibrational modes of interacting pentane-hexane system. This supports the paper's premise that these simpler alkane configurations can serve as foundational models for understanding the more complex molecular interactions and vibrational signatures in diesel fuel.

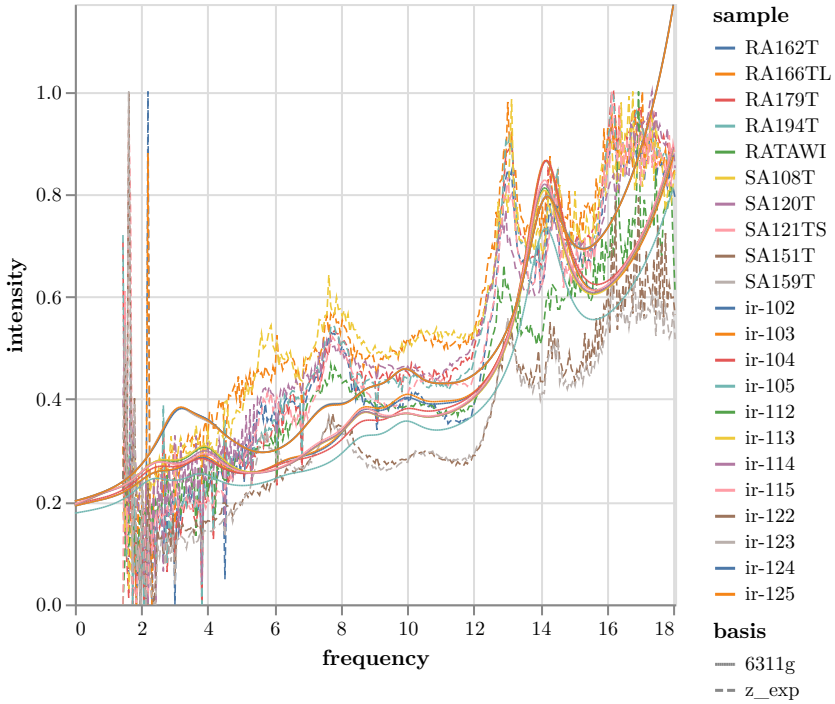

Figure 12: Experimental THz spectra with simulated spectra from the `6-311++G(2d,2p)` basis set for $s = 1$ stacks.

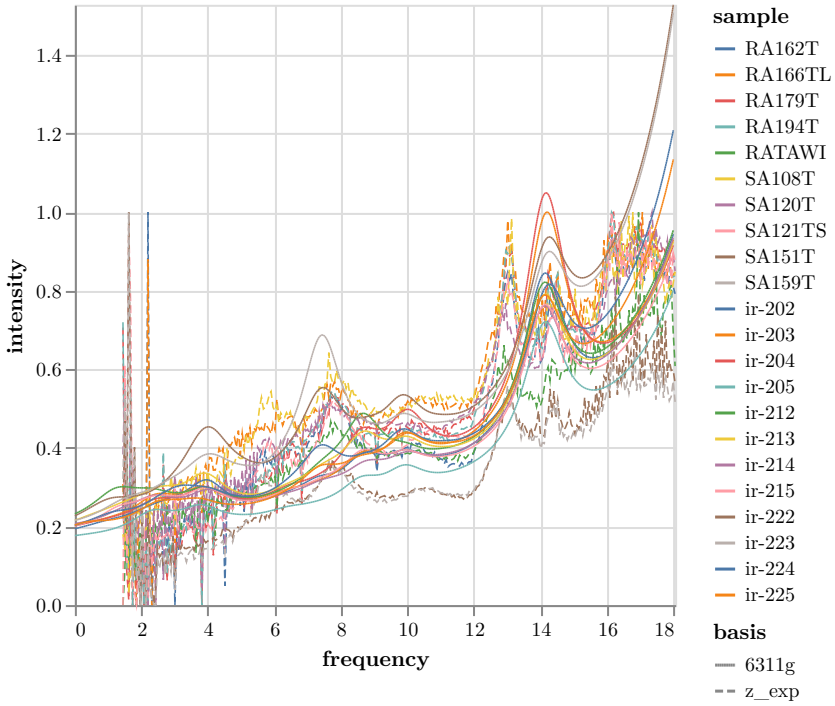

Figure 13: Experimental THz spectra with simulated spectra from the `6-311++G(2d,2p)` basis set for $s = 2$ stacks.

4 Conclusion

The computational THz spectra of stacked pentane-hexane conformers, particularly those generated using the `6-311++G(2d,2p)` basis set, show absorption features in frequency regions that are also present in the experimental THz-TDS data of diesel oil components. Despite the

simplified nature of the models, the consistent prediction of THz activity in specific bands across various stacking orientations and distances suggests these may represent characteristic intermolecular vibrational modes or collective motions of short-chain alkanes in condensed or aggregated phases. These simulated features, while sharper than the broad experimental bands, can guide the interpretation of the experimental diesel spectra.

Future work could involve simulating mixtures or larger, more representative alkane structures to achieve even closer agreement and potentially deconvolve the contributions of different molecular motifs to the overall diesel THz spectrum. The current study successfully demonstrates that computational modeling of simplified systems provides valuable insights into the types of molecular vibrations that contribute to the THz spectrum of complex fuels like diesel.

5 References

- Ponseca et al's Kuwaiti diesel oil data
- https://cccbdb.nist.gov/vibnotesx.asp
- https://github.com/leeping/geometric
- https://github.com/pyscf/dispersion
- https://github.com/pyscf/gpu4pyscf
- https://github.com/pyscf/properties
- https://pyscf.org
- https://rdkit.org