**UNIVERSIDADE FEDERAL DE MINAS GERAIS**
**PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA**
**Instituto de Ciências Exatas**
**Programa de Pós-graduação em Física**

Arthur Patrocínio Pena

# RAMAN SPECTRA-BASED STRUCTURED CLASSIFICATORY ANALYSIS OF QUINOIDAL AND DERIVATIVE MOLECULAR SYSTEMS: an unsupervised machine learning approach

Belo Horizonte
2023

Arthur Patrocínio Pena

# RAMAN SPECTRA-BASED STRUCTURED CLASSIFICATORY ANALYSIS OF QUINOIDAL AND DERIVATIVE MOLECULAR SYSTEMS:

## an unsupervised machine learning approach

Thesis presented to the Physics post-graduation program from the *Instituto de Ciências Exatas da Universidade Federal de Minas Gerais* as a partial requisite to obtain the Ph. D. degree in Physics.

Supervisor: Ado Jorio de Vasconcelos

Belo Horizonte
2023

Dados Internacionais de Catalogação na Publicação (CIP)

P397r Pena, Arthur Patrocínio.
Raman spectra-based structured classificatory analysis of quinoidal and derivative molecular systems: an unsupervised machine learning approach / Arthur Patrocínio Pena. – 2023.
71 f. : il.

Orientador: Ado Jorio de Vasconcelos.
Tese (doutorado) – Universidade Federal de Minas Gerais, Departamento de Física.
Bibliografia: f. 49-61.

1. Espectroscopia de Raman. 2. Quinona. 3. Aprendizado do computador. I. Título. II. Vasconcelos, Ado Jorio de. III. Universidade Federal de Minas Gerais, Departamento de Física.

CDU – 543.42 (043)

Ficha catalográfica elaborada por Romário Martins Ribeiro – CRB6 3595
Biblioteca Professor Manoel Lopes de Siqueira – Departamento de Física - UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA

**FOLHA DE APROVAÇÃO**

A presente tese, intitulada **"Raman spectra-based structured classificatory analysis of Quinoidal and derivative molecular systems: a machine learning approach"** de autoria de **ARTHUR PATROCÍNIO PENA** submetida à Comissão Examinadora, abaixo-assinada, foi aprovada para obtenção do grau de **DOUTOR EM CIÊNCIAS**, em onze de agosto de 2023.

Belo Horizonte, 11 de agosto de 2023.

Prof. Ado Jorio de Vasconcelos
Orientador do estudante
Departamento de Física/UFMG

Prof. Omar Paranaíba
Departamento de Ciência da Computação/UFMG

Prof. Luiz Gustavo de Oliveira Lopes Cançado
Departamento de Física/UFMG

Prof. Antônio Gomes de Souza Filho
Departamento de Física/Universidade Federal do Ceará

Prof. Mario Sérgio de Carvalho Mazzoni
Departamento de Física/UFMG

Prof. Alexandre Magno Rodrigues Teixeira
Departamento de Física/Universidade Estadual do Ceará

Documento assinado eletronicamente por **Antonio Gomes de Souza Filho**, **Usuário Externo**, em 11/08/2023, às 14:56, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Mario Sergio de Carvalho Mazzoni**, **Membro**, em 11/08/2023, às 15:00, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ado Jorio de Vasconcelos**, **Coordenador(a)**, em 11/08/2023, às 15:49, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Alexandre Magno Rodrigues Teixeira**, **Usuário Externo**, em 11/08/2023, às 17:49, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Omar Paranaiba Vilela Neto**, **Professor do Magistério Superior**, em 14/08/2023, às 14:09, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Luiz Gustavo de Oliveira Lopes Cancado**, **Professor do Magistério Superior**, em 16/08/2023, às 17:20, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2541156** e o código CRC **A9D85315**.

**Referência:** Processo nº 23072.249088/2023-12 SEI nº 2541156

# Acknowledgements

Aos meus filhotes de quatro patas: Simba, Nala, Magali, Penadinho, Petúnia, Curica e Salém, por deixarem minha vida mais leve com seus pêlos e ronronados constantes.

Aos Deuses e entidades guias de minhas crenças. Eternas fontes de inspiração na minha vida.

Agradeço às instituições de fomento que apoiaram financeiramente o meu trabalho, direta, ou indiretamente, CAPES, CNPq e FAPEMIG.

Ao LCPNano por todo o apoio nas minhas medidas.

Ao Prof. Ado Jorio por todo o profissionalismo, ensinamentos, e toda a paciência que teve comigo em todo este tempo: Mestre, eu não retiro uma palavra de todos os elogios, agradecimentos, e admirações que já te direcionei, embora meu cansaço e inseguranças que tive com a vida nestes tempos, e mantive em meu íntimo, pudessem ter mostrado o contrário.

A todos os amigos que fiz nessa jornada acadêmica, e alguns que hoje se estenderam para o mercado de trabalho, em especial: Marcello Nery, Mário Fernandes, Clovis Güerim, Diego Ferreira (O time 'Data-Science Fisicurralados'), entre outros que eu possa ter esquecido.

A Talita Bergamaschi, companheira sem precedentes. Você não faz idéia do quanto eu aprendi e ainda aprendo com você.

A meus Irmãos Lucas Nézio e Fabiana Scalabrini, casal que tive a honra de ser o celebrante de seu casamento, por simplesmente tudo desde 2016. *"Virtus Junxit Mors non Separabit"*.

Aos Professores Helio e Eufrânio, e a Dra. Renata pelas valiosas contribuições para este trabalho.

# Resumo

Este trabalho traz um método de análise classificatória baseado nos espectros vibracionais Raman de 38 quinonas e estruturas relacionadas, ordenando e classificando espectralmente os compostos. Os sistemas moleculares são relevantes para processos químicos e biológicos, com aplicações em farmacologia, toxicologia e medicina. A estratégia classificatória usa uma combinação de análise de componentes principais com métodos de agrupamento $k$-means. Tanto as simulações teóricas como os dados experimentais são analisados, estabelecendo assim as suas características espectrais, relacionadas com as suas estruturas e propriedades químicas. O protocolo introduzido aqui deve ser amplamente aplicável em outros sistemas moleculares e de estado sólido, servindo de base para um protocolo de estudo de materiais fundamentado em espectroscopia Raman e aprendizado de máquina.

**Palavras-chave**: Espectroscopia Raman, Estrutura Vibracional, Quinonas, PCA, K-means, Aprendizado de Máquina.

# Abstract

This work brings a classificatory analysis method based on the vibrational Raman spectra of 38 quinones and related structures, spectrally ordering and classifying the compounds. The molecular systems are relevant for chemical and biological processes, with applications in pharmacology, toxicology and medicine. The classificatory strategy uses a combination of principal component analysis with $k$-Means clustering methods. Both theoretical simulations and experimental data are analysed, thus establishing their spectral characteristics, as related to their chemical structures and properties. The protocol introduced here should be broadly applicable in other molecular and solid state systems, providing a strucured protocol form materials study based in Raman spectroscopy and machine learning.

**Keywords**: Raman spectroscopy, Vibrational Structure, Quinones, PCA, K-means, Machine Learning..

# Contents

# 1 Introduction

Quinones are organic aromatic compounds that can be found in nature or synthesized. In nature, quinones can be found in chemical and biological processes, such as breath chain and photosynthesis [1–4]. Structurally, in the most simple form, quinones show two carbonyl residues, separated by vinyl groups within a ring (figure 1(a) left) or adjacent to each other (figure 1(a) right). Quinone compounds can sustain benzene (benzoquinone), naphthalene (naphthoquinone), anthracene (anthraquinone) ring structures, and similar [5, 6]. Quinones can also be used as a precursor for the synthesis of several derivative molecular systems, such as phenazines. Phenazines are organic, heterocyclic, nytrogenous aromatic compounds, also called as dibenzo[*b*,*e*]pyrazine [7]. Figure 1 (b) shows the most basic forms of a phenazine. The phenazines analysed in this work were synthesized from quinones [8]. It is possible to find these quinones and phenazines grouped with many other structures forming more complex molecules, as described in this work.

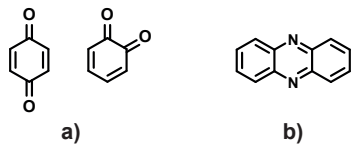



**Figure 1** – Most basic forms of **a)** quinone [*para*-benzoquinone(left) and *ortho*-benzoquinone(right)] and **b)** phenazine chemical structures.

In the last decades the study of the electronic [9] and chemical [10] properties of quinones has led to interesting results, especially in their applications in pharmacology, toxicology and medicine [1, 11, 12] with remarkably known antitumor [13–15], antimalarial [16,17], trypanocidal [18–20] and leishmanicidal [21] potential activity. Phenazines also have been widely explored in biology [7,22], where we can mention Barry et al. [23] investigations of its potential against tuberculosis disease and Cezairliyan et al. [24] identification of phenazines capable of killing nematodes. Most recently, Jardim et al. [8] reported on the synthesis of specific quinones and phenazines compounds for the development of new drugs against tuberculosis.

The vibrational modes of the p-benzoquinone molecule were firstly reported by Stammreich and Forneris, followed by the investigation of the polarization dependence of its Raman spectrum [25]. Durnick and Wait [26] published the investigation of the fundamental Raman active vibrations in phenazines using a He-Ne laser, along with some infrared active modes investigation. Stenman and Räsänen [27] investigated the symmetry as well as the Raman active modes of solid state 1,4-naphthoquinone. Delarmelina et al. [28] published a complete theoretical and experimental investigation of lapachol, $\alpha$- and $\beta$-lapachone Raman and infrared spectra. In addition, studies using time-resolved resonant

Raman spectroscopy [29], characterization via resonant Raman of quinones co-factors in solution [30, 31], in enzymatic catalysis [32], and surface-enhanced Raman spectroscopy (SERS) investigation [33] can be found in the literature.

As mentioned, Raman Spectroscopy provides detailed information about the composition and structure for molecules and other materials [34–36]. Its vibrational fingerprinting empowers researchers, enabling breakthroughs in fields from nanotechnology [37] to pharmaceuticals (as cited above), as well as the study of other optical phenomen [38, 39]. The invaluable insights gained through Raman Spectroscopy shape our understanding and drive innovation in material characterization. In this work we analyse the Raman spectra (both theoretical and experimental) of 38 quinones and derivative structures, some, to our knowledge, never characterized before using Raman spectroscopy. The relevance of comparing both simulated and experimental data in this analysis is that, when established that these data properly correlate, one can perform the analysis and predictions according with the information provided by the simulated data, avoiding the influence of experimental details.

Considering the complex vibrational structure, we make use Principal Component Analysis (PCA) and K-means Clustering [40–44] to analyse the data. These methods have been widely used in the last decades in material science, biology and chemistry to improve the extracting of information from data analysis in broader, automatic, fast, and efficient ways. The complexity of the data we analysed here is due to the number of analized compounts (38) and the number of vibrational Raman active modes, which goes up to 207 modes for the most complex analysed structure.

Therefore, here we bring an in depth study and the proposal of a classificatory analysis method using the combination of PCA with K-means clustering statistical learning methods, applied to the vibrational spectra of these 38 quinones and related structures from Raman spectroscopy. The analysis was initially performed to the simmulation data, which is free from experimental artefacts, and further compared to related experimental data, showing compatible results. Our contribution is, therefore, twofold: **(i)** we present new data and analysis related to these relevant organic aromatic compounds, the quinones e phenazines; **(ii)** we propose a methodology for Raman spectral analysis that might contribute for big data protocols such as the development of material´s genome initiative [45].

# 2 Theoretical background

## 2.1 Aspects of Raman spectroscopy

The phenomenon of light scattering can be divided in elastic or Rayleigh scattering, where light is scattered with the same energy of the incident light, and the inelastic scattering or Raman scattering, where a sample is excited by a beam of monochromatic light, and the interaction between the photons of that beam with the molecules' modes of vibration (or phonons for solid-state materials) of the sample causes the energy of the scattered light to be shifted, due to energy exchange between light and matter [46].

Figure 2 represents the transition of an electron from the fundamental state to a virtual excited state and its decaying process to the fundamental electronic state, after being excited by an incident photon from a light beam. Besides the electronic levels, there are the vibrational energy levels of the material. Figure 2 represents three possible outcomes of this process of excitation: **a)** represents the electron decaying back to its fundamental state, with no vibrational energy level variation in the material, and no shift in the energy of the scattered photon, which is the Rayleigh scattering; in **b)** the electron decays to the fundamental electronic state with a higher vibrational energy in the material (from $n = 0$ to $n = 1$), so that the scattered photon have less energy than before the interaction, which is called Stokes Raman scattering. **c)** shows the anti-Stokes Raman scattering, where the electron decays to the fundamental state, with a lower vibration energy level ($n = 1$ to $n = 0$) in the material, and the scattered photon has more energy after the interaction [36]. It is useful to mention that, for solid state, the vibrational levels are represented by energy bands, and the vibrational modes are usually named phonons.

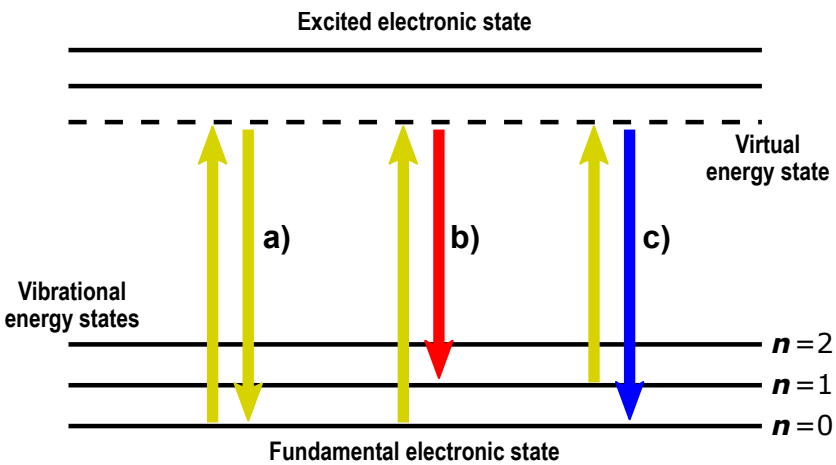



**Figure 2** – Representative diagram of light scattering. **a)** shows the elastic Rayleigh scattering, **b)** shows the Raman Stokes scattering and **c)** shows the Raman anti-Stokes scattering [36].

The classical approach that explains the Rayleigh and the Raman scattering phenomenon can be defined considering light as represented by the electric field $\vec{E}$ interacting with the material, and inducing a modulation in its diplole momentum $\vec{P}$ [46], as shown in the expression 2.1.

$$\vec{P} = \alpha \vec{E}, \tag{2.1}$$

where $\alpha$ is the electronic *polarizability*. Since an electromagnetic wave with frequency $\omega_0$ have its intensity ($E$) oscilating in time, $E = E_0 \cos(\omega_0 t)$ and we can write the induced polarization as:

$$\begin{aligned} E &= E_0 \cos(\omega_0 t), \\ P &= \alpha E_0 \cos(\omega_0 t). \end{aligned} \tag{2.2}$$

Within the material, the polarizability $\alpha$ usually depends on the generalized coordinate $Q$ of a vibrational mode

$$Q = Q_0 \cos(\omega_q t), \tag{2.3}$$

where $Q_0$ is the vibrational amplitude and $\omega_q$ is the molecule vibration frequency. For a small amplitude of vibration, we can assume that $\alpha$ is a linear function of $Q$. So we can expand it in a Taylor series such as

$$\alpha(Q) = \alpha_0 + \left(\frac{\partial \alpha}{\partial Q}\right)\bigg|_{Q=0} Q + O^2, \tag{2.4}$$

and the terms of second or higher order can be disregarded. Applying 2.4 in 2.2 it follows:

$$P = \alpha_0 E_0 \cos(\omega_0 t) + \left(\frac{\partial \alpha}{\partial Q}\right)\bigg|_{Q=0} Q_0 E_0 \cos(\omega_0 t) \cos(\omega_q t). \tag{2.5}$$

we can use the relation $2\cos(a)\cos(b) = \cos(a+b) + \cos(a-b)$, to obtain:

$$P = \alpha_0 E_0 \cos \omega_0 t + \frac{1}{2}\left(\frac{\partial \alpha}{\partial Q}\right)\bigg|_{Q=0} Q_0 E_0 \{\cos[(\omega_0 + \omega_q)t] + \cos[(\omega_0 - \omega_q)t]\}. \tag{2.6}$$

Here, the first term have the frequency of the elastic scattering (Rayleigh), and the other terms represent, respectively, the anti-Stokes, with resulting frequency $(\omega_0 + \omega_q)$, and the Stokes, with frequency $(\omega_0 - \omega_q)$. The Raman scattering occurs when $\frac{\partial \alpha}{\partial Q} \neq 0$.

The registered data of the Raman scattering is represented by the Raman spectrum, which is exemplified in figure 3, where we can notice the Rayleigh scattering in the center (the $0cm^{-1}$ Raman shift), which has to be blocked by a notch filter due to its intensity. The bands in the left and right hand of the figure represent the anti-Stokes and the Stokes energy bands, respectively. Each Stokes/anti-Stokes peaks represent a vibrational mode, related to a specific Raman shift.

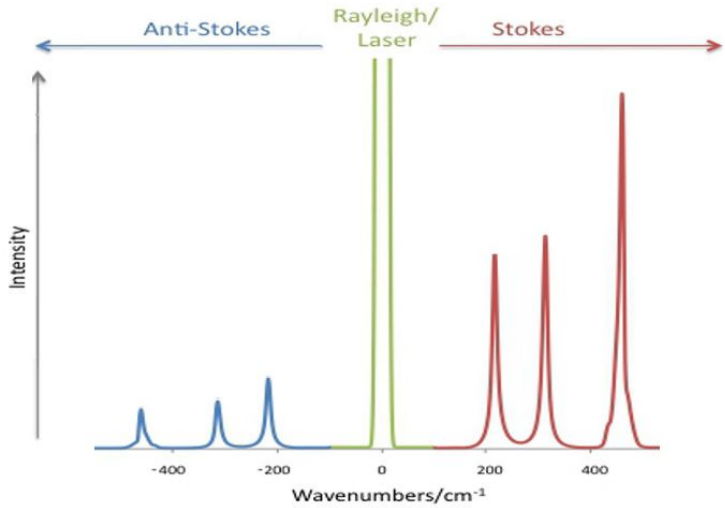



**Figure 3** – Representative model of Raman spectrum. At the center of the spectrum is located the energy band related to the Rayleigh scattering, and the left and right bands, respectively refers to the anti-Stokes and Stokes Raman scattering energy bands. This image was based in the referecne [47]

### 2.1.1 Density Funcional Theory (DFT) formalism for the simulations of Vibrational Spectra

In the present section we bring a brief discussion about the formalism behind these simulations, the Density Functional Theory (DFT). Since DFT formalism is not the main scope of this work, is not our intention to define the method itself, and we let some references along the text for further details and definitions as we bring the main aspects of DFT considered to obtain the simulational results for this work.

The widely known Schrödinger's equation [48] is the base of quantum mechanics and can be represented as:

$$H\psi = E\psi. \tag{2.7}$$

This equation provides a description of the electronic structure of a molecule or a solid material sample [46, 48, 49]. In DFT, the total energy E is treated as a function of the electronic density $\rho$ as the basic variable. The objective of using the DFT is then to minimize the energy in relation to the electronic densities [50], as stated by Hohenberg and Kohn [51] in their two following theorems that served as base for the DFT formulations: 1. The external potential over the electrons is a functional of the electronic density; 2. The energy of the fundamental state is minimized if and only if, the electronic density is the exact density to the fundamental state. Based on these theorems, the electronic Hamiltonian for a systems of $M$ nuclei and $N$ electrons can be defined as:

$$\hat{H} = -\sum_{i}^{N} \frac{1}{2}\nabla_i^2 - \sum_{A}^{M}\sum_{i}^{N} \frac{Z_A}{|R_A - r_i|} + \sum_{i<j}^{N}\sum_{j}^{N} \frac{1}{|r_i - r_j|} \tag{2.8}$$

where $i$ and $j$ are indices that represent the electrons of the system, $A$ represents the atomic nuclei, $r_i$ and $R_A$ represent the positions of the electron $i$ and the atomic nuclei $A$, and $Z_A$ is the atomic number of the atom $A$ [52–54]. The Hamiltonian operator is defined by the kinectic energy operator (first term, represented by $\hat{T}$), the external potential operator (second term, $\hat{V}_{ext}$), which refers to the position and charges of the electrons, the electron-electron repulsion (third term, $\hat{V}_e$). We can rewrite the external potential like:

$$\nu(r_i) = \sum_{A}^{M} \frac{Z_A}{|R_A - r_i|} \quad \hat{V}_{ext} = \sum_{i}^{N} \nu(r_i), \tag{2.9}$$

where $\nu(r_i)$ is the nuclear attraction potential energy functional for an electron in a $r$ position and $\Psi_0$ is the solution function for the Hamiltonian at the funtamental state. The gound-state electronic density is then, defined as:

$$\rho_0(r) = \int \psi_0^*(r_1, \ldots, r_n) \sum_{i}^{N} \delta(r - r_i)\psi_0(r_1, \ldots, r_n) dr_1 \ldots dr_n, \tag{2.10}$$

so that we can write:

$$\begin{aligned} \langle \Psi_0 | \sum_{i}^{N} \nu(r_i) | \Psi_0 \rangle &= \int \psi_0^*(r_1, \ldots, r_n) \sum_{i}^{N} \nu(r_i)\psi_0(r_1, \ldots, r_n) dr_1 \ldots dr_n \\ &= \int \psi_0^*(r_1, \ldots, r_n) \sum_{i}^{N} \delta(r_p - r_i)\nu(r_p)\psi_0(r_1, \ldots, r_n) dr_1 \ldots dr_n dr_p \\ &= \int \rho_0(r)\nu(r) dr \end{aligned} \tag{2.11}$$

where $\psi_0(r_1, \ldots, r_n)$ is the solution of the Hamiltonian already mentioned above for the case of the fundamental state. Considering the separation of the external potential, the total energy of the system is, then:

$$\begin{aligned} E_0 = \langle \Psi_0 | \hat{H} | \Psi_0 \rangle &= \langle \Psi_0 | \hat{T} + \hat{V}_e + \hat{V}_{ext} | \Psi_0 \rangle \\ &= \langle \Psi_0 | \hat{T} + \hat{V}_e | \Psi_0 \rangle + \int \rho_0(r)\nu(r) dr. \end{aligned} \tag{2.12}$$

So, according to the Hohenberg and Kohn theorems, it is possible to calculate all properties of the system, without the necessicy of having determinated the wave function, by knowing the electronic density at the fundamental state. However, the theorems do not guide us to the calculation of $E_0$ from the density $\rho_0$, neither to calculate $\rho_0$ without determining the wave function. The Khon-shan formalism [55] offered a method to calculate $E_0$ and $\rho_0$

exactly if the used functionals are exact. In DFT, we still use approximated functionals to obtain approximated results.

In this work, as we shall see in section 3.2, we used the funcional $M06-2$x [56], which is classified as a meta-GGA (meta-Generalized Gradient Approximation) functional. Meta-GGA functionals derive from GGA functionals [57, 58], which are functionals that are derived from both, the electronic density and its gradient (how fast the density varies locally in the system). The meta-GGA functionals consider additionally the local kinectic energy density, allowing it to treat different chemical bonds more accuratelly than GGA functionals. The $M06-2x$ functional is used in the computational analysis of organic modecules, providing good results to thermodynamic properties, and so, it is expected the same for vibrational properties.

To optimize the structure, it was used the basis set $6-31+G(2d,p)$, which describe the using of 6 gaussians to describe the behavior of the core electrons, and 3 and 1 gaussians fo describe the valence electrons [59], plus the addition of diffuse functions (the "+" signal in the representation of the basis set) to enhance the accuracy in the calculation of the electrons behavior [60].

## 2.2 Quinones and derivate molecular systems

This section was based in the analogous section found in the work cited in the reference [61], since this present work was developed at the same period and with collaboration with the authors of the mentioned work.

The basic structure, origin and roles of quinones were already presented in the Introduction of this work, and, as already mentioned, their molecular structure can sustain different nuclei structures, which are illustrated in figure 4, respectivelly for the examples of a benzene ring (**a**)), naphthalene (**b**)), anthracene (**c**)) and phenanthrene (**d**)).

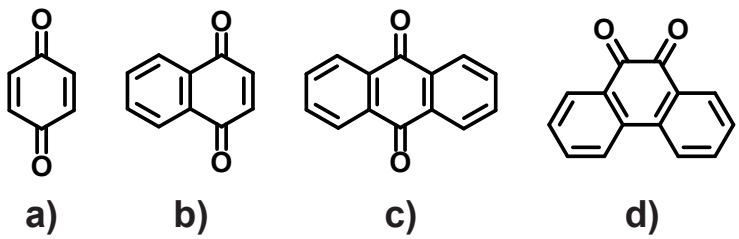



**Figure 4** – Schematic representation of some nuclei for quinones. **a)** Benzoquinone, **b)** Naphthoquinone, **c)** Anthraquinone and **d)** Phenanthroquinone.

Because of their different nuclei structures and properties, quinones can be used as a precursor for the synthesis of several derivative molecular systems, such as phenazines. Phenazines are organic, heterocyclic, nytrogenous aromatic compounds, also called as

dibenzo[*b,e*]pyrazine [7]. Figure 1 (b) shows the most basic forms of a phenazine. Since phenazines analysed in this work were synthesized from quinones [8], we keep our discussion centered in the properties of quinones, but in the introduction chapter, and table B we have some references for further details about the mentioned phenazines in our study. In the samples studied in this work, it is possible to find these quinones and phenazines grouped with many other structures forming more complex molecules, as described along the text.

Considering the wide role that quinones play in nature, we can begin mentioning as example the ubiquinone, which is a benzoquinoidal class molecule, also known as *Coenzyme Q10*, and which structure is illustrated in figure 5. They are present in all of the main tissues of the human body and it is also used as medication for, e. g., heart diseases [62], since it acts as an electron carrier in the mitochondrial breath electron transport chain.

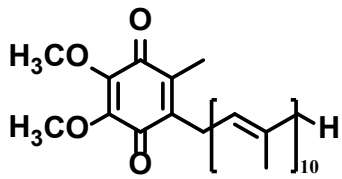



**Figure 5** – Chemical structure of the Ubiquinone, also known as *Coenzime Q10*.

Another molecule that we can bring in this section, which shows a naphthoquinoidal nucleus, is the vitamin K, and it exists in two versions: $K_1$, also called phylloquinone, and mostly found in plants, and $K_2$, also called menaquinone, and found synthesized by some kinds of bacteria [63]. These structures are presented below in figure 6. It is useful to mention that vitamin K is important for the biological activity of blood coagulation and bone metabolism.

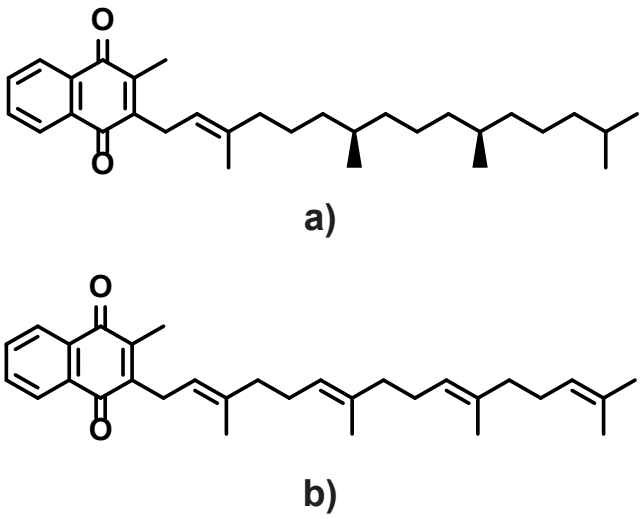



**Figure 6** – chemical sctructures of **a)** Vitamin K1 and **b)** Vitamin K2.

One last example that we can mention here, among a whole family of possible quinones either found in nature or sinthesized, is the Lapachol compound, which is shown in figure 7. Lapachol is a naphtoquinoidal molecule that can be extracted from trees that belong to the *Tabebuia* family (e.g. Brazilian *Ipê*). It is known as having high potential on biologial activity, being investigated in antitumor [64], anti-inflammatory [65] and antifungi [66] scientific research, among others [67]. Still in its biological activities, Lapachol can be used to obtain the $\beta$-lapachone molecule, which has been highly investigated due to its pharmacological activity in anti-tumor applications.

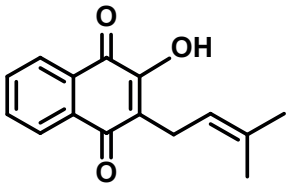



**Figure 7** – Chemical structure of the Lapachol molecule.

Lastly, quinones are highly reactive molecules, which chemical oxidative properties allow interaction with biological samples, acting in the electronic transference in bioreduction. In the last decades, the study of the electronic [9] and chemical [10] properties of quinones has led to interesting results, especially in their applications in pharmacology, toxicology and medicine [1, 11, 12] with remarkably known antitumor [13–15], antimalarial [16, 17], trypanocidal [18–20] and leishmanicidal [21] potential activity. Phenazines also have been widely explored in biology [7, 22], where we can mention Barry et al. [23] investigations of its potential against tuberculosis disease and Cezairliyan et al. [24] identification of phenazines capable of killing nematodes. Most recently, Jardim et al. [8] reported on the synthesis of specific quinones and phenazines compounds for the development of new drugs against tuberculosis.

## 2.3 Computational data processing

The Statistical Learning techniques can be divided in three different kinds of algorithms: supervised, unsupervised and reinforcement learning [68, 69]. These three kinds differ basically in how the statistical algorithm will be trained. For example, supervised learning algorithms deals usually with labeled data for training, with a predefined target variable of the dataset, unsupervised learning algorithms are oftenly found being used to process unlabeled data, and reinforcement learning algorithms work by interacting with the environment of the data analysis by means of errors or rewards: a chosen variable can vary according to the expected performance in the process of learning, guiding the model to the better accuracy. In this work, we use an unsupervised learning algorithm to classify data.

As mentioned previously, the machine will deal with data without any guidance, without, in principle, any necessary previous notion of classification or the dependency of a specific target variable. It is then put into an assignment of understanding patterns and behaviors of the data and then give the outcome. There are many examples of unsupervised learning algorithms available in the literature [69, 70], and as we discuss in the following sections, in this work we use clustering ($k$-eans) and dimensionality reduction algorithms (Principal Component Analysis) for Raman data classification.

### 2.3.1 *Principal Component Analysis* (PCA)

Given a problem in an initial $n$-dimensional space, one may find it necessary to represent a set of points as a best-fit regression into a specific, lower dimensional space. In order to make the processing (and/or the work) easier, keeping the most significant properties of the data set is maintained, the so-called Dimensionality Reduction [71–73] can be used, and we shall explore in this section the Principal Component Analysis (PCA) algorithm. The main Idea behind PCA is to convert a set of correlated variables into uncorrelated components (these are the principal components), such that these components are ordered by the value of each of their respective variance, as we shall see in the discussion bellow.

Consider that we have a set of $n$ points in a $p$-dimensional space, represented by a matrix $\mathbf{X}_{n\times p}$, the question here is how can we reduce this set into a $q$-dimensional space such that $q \leq p$, as we keep the main information of this set. The main objective of PCA is to define a projection of these points into the best fit regression lines and find the directions that maximizes the variances of the projected points into it [73, 74]. So, we can consider a vector $\vec{x}$ of $p$ random variables, as we are interested in the variances of these variables, and a vector $\vec{\alpha_1}$ of $p$ constants $\alpha_{11}, \alpha_{12}, ..., \alpha_{1p}$ that define a linear funcion

$$\vec{\alpha_1}^T \vec{x} = \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1p}x_p = \sum_{j=1}^{p} \alpha_{1j}x_j, \tag{2.13}$$

having maximum variance, in relation of the elements of $\vec{x}$. The next step is to define analogous linear functions $\vec{\alpha_2}^T\vec{x}, \vec{\alpha_3}^T\vec{x}, \cdots, \vec{\alpha_q}^T\vec{x}$, independent from each other, these being the "best fitting" linear regressions of the data set, for less than or each of the dimensions involved. It is hoped that most of the variation in $\vec{x}$ is accounted for by only a few of these $\vec{\alpha_q}^T\vec{x}$ functions, which is known as being the *Principal Components* of the data set.

A simple case, when $p = 2$, is illustrated in the figures 8 and 9. Figure 8 shows a set of data on two correlated variables $x_1$ and $x_2$, with the application of the functions that define the best fitting line. Figure 9 shows the result in terms of the transformation

in PC1 and PC2. In these, it is possible to notice that the variance in PC1 is higher than the variance in PC2.

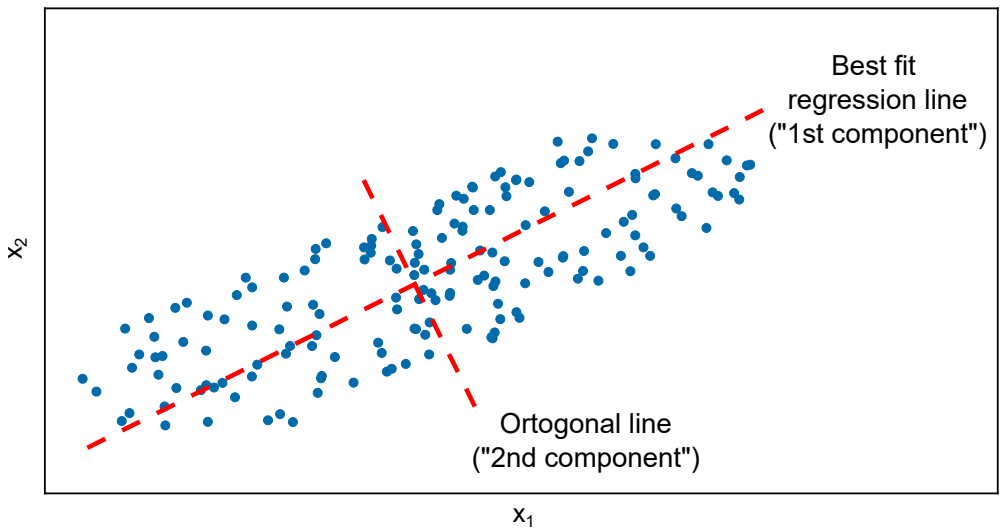



**Figure 8** – Randomly plotted points to illustrate how PCA works. The dashed lines is merely an example of regression which serves as a reference to define the samples space, from which we construct the two Principal Components of the dataset. From it we obtain the "direction of variance" of the whole data involved in the analysis.

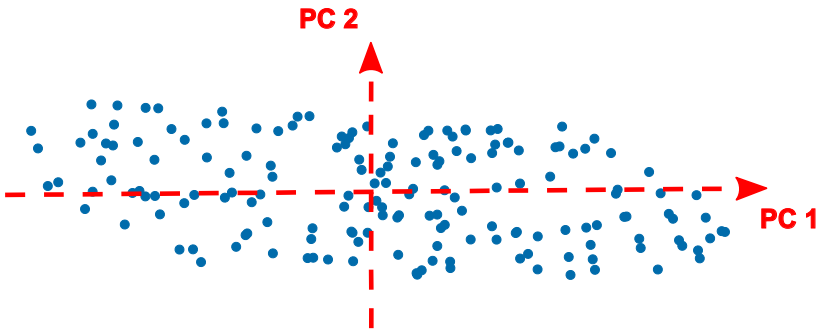



**Figure 9** – Representation of the points projected in the new "PC space" when $p = 2$, as it can be noticed from Figure 8.

The PCs can be found considering that $\vec{x}$ has a known covariance matrix $\mathbf{\Sigma}$. It is a $ij$th dimensional matrix which diagonal $(i = j)$ elements are the variance of $i$th element of $\vec{x}$, and the non-diagonal $(i \neq j)$ elements are the covariance between the $i$th and $j$th elements of $\vec{x}$. Defining the $q$th PC as $z_q = \vec{\alpha_q}^T \vec{x}$, where $\vec{\alpha_q}$ represents the eingenvectors of $\mathbf{\Sigma}$, which corresponds to the $q$th largest eigenvalue $\lambda_q$ and $\vec{\alpha_q}$ is chosen to have unit length $(\vec{\alpha_q}^T \vec{\alpha_q} = 1)$, so that

$$var(z_q) = var(\vec{\alpha_q}^T \vec{x}) = \lambda_q, \tag{2.14}$$

where $\lambda_1 > \lambda_2 > \cdots > \lambda_q$.

We shall define these relations in the next paragraphs, according with the references [73, 75, 76]. Considering the first PC, $\vec{\alpha_1}^T\vec{x}$, as the vector $\vec{\alpha_1}$ maximized the expression

$$var(\vec{\alpha_1}^T\vec{x}) = \vec{\alpha_1}^T\mathbf{\Sigma}\vec{\alpha_1}. \tag{2.15}$$

Here we need to consider such a constraint for normalization, so that we garantee the maximum of the expression to be achieved. We then consider $\vec{\alpha_1}^T\vec{\alpha_1} = 1$, and use the technique of Lagrange multipliers, in other words, we intent to maximize:

$$\vec{\alpha_1}^T\mathbf{\Sigma}\vec{\alpha_1} - \lambda(\vec{\alpha_1}^T\vec{\alpha_1} - 1), \tag{2.16}$$

here, for this case, $\lambda$ represents a Lagrange multiplier. If we apply a differentiation over $\vec{\alpha_1}$, we have

$$\begin{aligned} \mathbf{\Sigma}\vec{\alpha_1} - \lambda\vec{\alpha_1} &= 0; \\ (\mathbf{\Sigma} - \lambda\mathbf{I_p})\vec{\alpha_1} &= 0, \end{aligned} \tag{2.17}$$

where $\mathbf{I_p}$ represents a $p \times p$ identity matrix. We then have $\lambda$ being and eigenvalue of $\mathbf{\Sigma}$ and $\vec{\alpha_1}$ being the corresponding eigenvector. The quantity to be maximized is, then

$$\vec{\alpha_1}^T\mathbf{\Sigma}\vec{\alpha_1} = \vec{\alpha_1}^T\lambda\vec{\alpha_1} = \lambda\vec{\alpha_1}^T\vec{\alpha_1} = \lambda, \tag{2.18}$$

so that $\lambda$ must be as large as possible, and $\vec{\alpha_1}$ has to be the eigenvector corresponding to the largest eigenvalue of $\mathbf{\Sigma}$, and so

$$var(\vec{\alpha_1}^T\vec{x}) = \vec{\alpha_1}^T\mathbf{\Sigma}\vec{\alpha_1} = \lambda_1, \tag{2.19}$$

is the largest eigenvalue.

For the second PC, $\vec{\alpha_2}^T\vec{x}$ is searched for maximizing the expression $\vec{\alpha_2}^T\mathbf{\Sigma}\vec{\alpha_2}$, and can be obtained by considering the covariance between $\vec{\alpha_2}^T\vec{x}$ and $\vec{\alpha_1}^T\vec{x}$, which is zero, once they are uncorrelated. But we also have that

$$cov(\vec{\alpha_1}^T\vec{x}, \vec{\alpha_2}^T\vec{x}) = \vec{\alpha_1}^T\mathbf{\Sigma}\vec{\alpha_2} = \vec{\alpha_2}^T\mathbf{\Sigma}\vec{\alpha_1} = \vec{\alpha_2}^T\lambda_1\vec{\alpha_1} = \lambda_1\vec{\alpha_2}^T\vec{\alpha_1} = \lambda_1\vec{\alpha_1}^T\vec{\alpha_2} = 0. \tag{2.20}$$

Any of these equations could be used to explore the zero covariance between $\vec{\alpha_2}^T\vec{x}$ and $\vec{\alpha_1}^T\vec{x}$, but we shall explore the expression $\vec{\alpha_1}^T\vec{\alpha_2} = 0$ for simplicity. Remembering the constraint of normalization as used before for $\vec{\alpha_1}$, we have that the quantity to be maximized is

$$\vec{\alpha_2}^T\mathbf{\Sigma}\vec{\alpha_2} - \lambda(\vec{\alpha_2}^T\vec{\alpha_2} - 1) - \phi\vec{\alpha_2}^T\vec{\alpha_1} \tag{2.21}$$

where $\lambda$ and $\phi$ are Lagrange multipliers. Differentiating in respect to $\vec{\alpha_2}$ and multiplying, by the left, the resulting equation by $\vec{\alpha_1}^T$ yields

$$\begin{aligned} \mathbf{\Sigma}\vec{\alpha_2} - \lambda\vec{\alpha_2} - \phi\vec{\alpha_1} = 0; \\ \vec{\alpha_1}^T\mathbf{\Sigma}\vec{\alpha_2} - \lambda\vec{\alpha_1}^T\vec{\alpha_2} - \phi\vec{\alpha_1}^T\vec{\alpha_1} = 0, \end{aligned} \tag{2.22}$$

with the constraint $\vec{\alpha_1}^T\vec{\alpha_1} = 1$, and the first two terms being zero, yields $\phi = 0$. So again we have

$$\begin{aligned} \mathbf{\Sigma}\vec{\alpha_2} - \lambda\vec{\alpha_2} = 0; \\ (\mathbf{\Sigma} - \lambda\mathbf{I_p})\vec{\alpha_2} = 0, \end{aligned} \tag{2.23}$$

with $\lambda$ being and eigenvalue of $\mathbf{\Sigma}$ and $\vec{\alpha_2}$ being the corresponding eigenvector.

Once again, we have, similarly, $\vec{\alpha_2}^T\mathbf{\Sigma}\vec{\alpha_2} = \lambda$, with $\lambda$ being as large as possible, but now assuming that $\mathbf{\Sigma}$ does not produce repeated eigenvalues, so it could violate the constraints of independence between the vectors $\vec{\alpha}_q$, $\lambda \neq \lambda_1$, so that $\lambda$ must be the second highest eigenvalue of $\mathbf{\Sigma}$ in this case.

The analogous can be demonstrated for $\lambda_3, \lambda_4, \cdots, \lambda_p$ and for the vectors of coefficients, and, by consequence, for the other $p$th PCs, remembering equation 2.14

Although the most challenging part of this procedure is to precisely interpret the Principal Components (PCs), we can say that in this work the values obtained in the PCA calculation will give us the coordinates in the samples space, guiding us to the relative distance between the analyzed samples, which gives us the notion of "how much different or similar" they can be among each other in function of their relative distances. We shall discuss further details in the section 3.2.2.

### 2.3.2 $k$-means clustering

In general, clustering algorithms intend to define, from the character of the dataset, the best division (by labeling) of groups of points. In the case of $k$-means clustering, it is made by the calculation of "cluster centroids" which are the arithmetic mean of the points that belongs to each cluster, with each point being closer to its own cluster centroid than to the centroid of any other cluster [77–79].

Mathematically, the $k$-means clustering can be defined based on the Sum of Squares(SSQ) criterion [78, 79], and can be described as follows: Given a set of $n$ data points $x_1, \cdots, x_n$ in the space $\mathbb{R}^p$ and a k-particioned set $\mathcal{C} = (C_1, \cdots, C_k)$. The discrete version of the SSQ criterion is defined as:

$$g_n(\mathcal{C}) := \sum_{i=1}^{k} \sum_{\ell \in C_i} \|x_\ell - \overline{x}_{C_i}\|^2 \to \min_{\mathcal{C}}, \tag{2.24}$$

with $\overline{x}_{C_i}$ representing the centroid of the points $x_\ell$ which belongs to $C_i$, and we look for a k-partition of the set $\mathcal{O}$ with minimum criterion value as in 2.24. We can use the equivalent form of 2.24 for two parameters,

$$g_n(\mathcal{C}, \mathcal{Z}) := \sum_{i=1}^{k} \sum_{k \in C_i} \|x_\ell - z_i\|^2 \to \min_{\mathcal{C},\mathcal{Z}}, \tag{2.25}$$

where the minimization problem is relationed to all the systems $\mathcal{Z} = (z_1, \cdots, z_n)$, which result come from the following theorem:

**Theorem 3.1:**

$(i)$ For any fixed k-partition $\mathcal{C}$, the criterion 2.24 is partially minimized in relation to $\mathcal{Z}$ by the sistem of class centroids $\mathcal{Z}^* = (\overline{x}_{C_1}, \cdots, \overline{x}_{C_k}) =: \mathcal{Z}(\mathcal{C})$:

$$g_n(\mathcal{C}, \mathcal{Z}) \geq g_n(\mathcal{C}, \mathcal{Z}^*) := \sum_{i=1}^{k} \sum_{k \in C_i} \|x_\ell - \overline{x}_{C_i}\|^2 = g_n(\mathcal{C}) \ \forall \ \mathcal{Z}, \tag{2.26}$$

$(ii)$ For any fixed prototype system $\mathcal{Z}$ the criterion $g_n(\mathcal{C})$ is partially minimized in relation to $\mathcal{C}$ by any minimum-distance partition $\mathcal{C}^* =: \mathcal{C}(\mathcal{Z})$ induced by $\mathcal{Z}$, i.e. with classes given by $C_i^* := \{\ell \in \mathcal{O} \,|\, d(x_\ell, z_i) = min_{j=1,\ldots,k} d(x_\ell, z_i)\}$, where $d(x, z) = \|x - z\|^2$ is the squared Eucliedean distances

$$g_n(\mathcal{C}, \mathcal{Z}) \geq g_n(\mathcal{C}^*, \mathcal{Z}) := \sum_{\ell=1}^{n} \min_{\mathcal{C},\mathcal{Z}} \{\|x - z\|^2\} \ \forall \ \mathcal{C}. \tag{2.27}$$

In simple words, the $k$-means method is set to find an optimum k-partition by iterating the partial minimization steps from the Theorem 3.1. It proceeds as shown bellow:
$t = 0$: Begin with an arbitrary prototype system $\mathcal{Z}^{(0)} = (z_1^{(0)}, \ldots, z_k^{(0)})$.
$t \to t + 1$:
$(i)$ Minimize the criterion $g_n(\mathcal{C}, \mathcal{Z}^{(t)})$ relationed to the k-partition $\mathcal{C}$, determining a minimum-distance partition $\mathcal{C}^{(t+1)} := \mathcal{C}(\mathcal{Z}(t))$. In other words, assign each register to the nearest group mean according to the measure of the square distance.
$(ii)$ Minimize the criterion $g_n(\mathcal{C}^{(t+1)}, \mathcal{Z})$ relationed to $\mathcal{Z}$, calculating the system of class centroids $\mathcal{Z}^{(t+1)} := \mathcal{Z}(\mathcal{C}(t + 1))$. This set the new mean of the group, based on the attribution of the registers.

The method converges when when the attribution of registers into groups does not change.

The $k$-means algorithm searches for a predefined number of clusters, and once the centroids are identified, the different clusters are separated by "mute coloured labels" for each group of points, with the colours being not related with any direct characteristics from the points themselves. This process can be seen on figure 10 where in the left side there is a random plot of points, which can be easily seen that there is something close to three different clusters. After the running of the algorithm, asking it to search three clusters of points, in the right side of the figure 10, it is possible to notice the labeling of

three clusters, marked by the coloured labels, as well as the cluster centroids, which are represented by the crossed red circles.

It is useful to say that there is not a unique way of choosing the number of clusters for the algorithm to find: it can depend on the context in which the dataset is being analyzed. Some can even make use of heuristic methods, like the so-called "*elbow method*", that uses the relation between the number of clusters and the behavior of the mean errors to find the best number of clusters your algorithm can be asked to calculate.

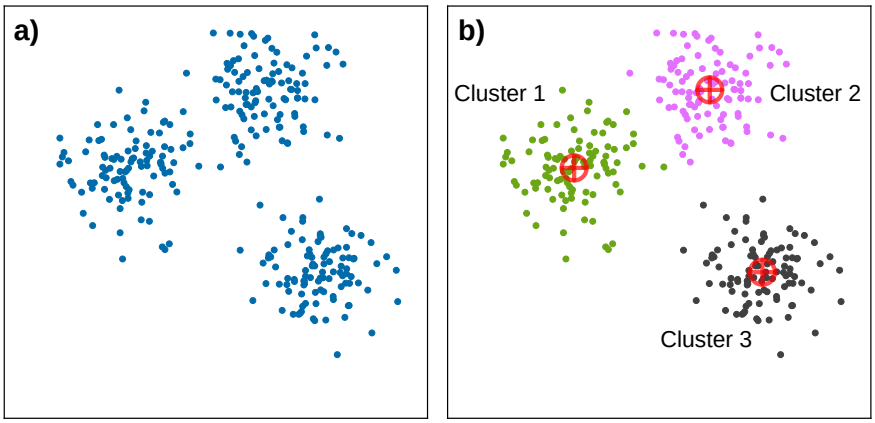



**Figure 10** – Example of how the clustering algorythm works. The clusters are defined by the different colors in the scatter plot. The colors are merelly "mute labels" and do not have anything to do with some kind of property of the analyzed data. The red crossed circles in the figure illustrate the calculated centroyd of the clusters, which serve as reference to compute the split of the points in well-defined clusters, as well as the involved errors of standart deviations, if necessary.

In this work we present the $k$-means clustering technique combined with the resulting plot of the PCA calculation our dataset gives us as result: We apply the PCA algorithm over the numerical Raman data of the simulations, followed by a three-dimensional plot with the three first PCs in order to obseve the distributions of the points which represent the samples, and apply a $k$-means cluster algorithms to investigate the grouping of the points according to their statistical interpretation of the algorithm. The ordering process, and the spectral reconstructions at the first principál component shall be discussed later in the Methodology section.

# 3 Methodology

## 3.1 Experimental details

### 3.1.1 Samples

The samples were obtained in collaboration with the *da Silva Júnior Group - Organic and Medicinal chemistry* [80] laboratory, at *Departamento de Química da Universidade Federal de Minas Gerais*, and the Apendix B brings the names of the compounds, chemical formulas, chemical structure representation for a single molecule and, most importantly, the references for how the 38 analyzed compounds (see Table B) were obtained.

The compounds were in a solid, microscopic, powder-like state, varying between crystalline and amorphous aspects (in some cases, both aspects could be found in the same sample) as shown in the figure 11 for the compound **(1)** (see Appendix B for compounds identification). In the middle image **b)** in the picture, when zooming in this captured region, it was possible to observe the formation of groups of small needle-like crystals, which Raman spectrum would vary according with the orientation of the sample.

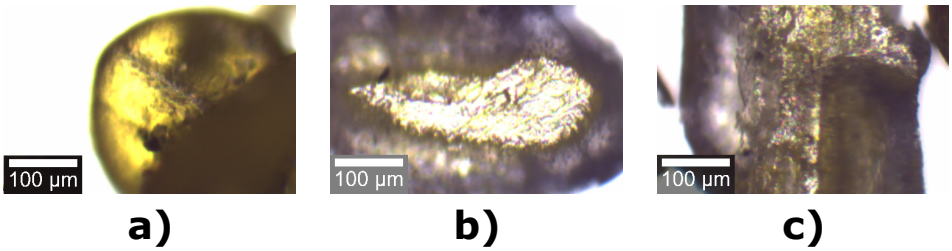



**Figure 11** – General picture of the morphologies found in compound **(1)**. It is possible to find some amorphous appearace in **a)** and **c)**, and **b)** shows the most crystalline aspect in the sample.

The studied compounds proved to be stable, but also sensitive to the laser power: for most of the samples, values such as 4.0 $mW$ in a 633 $nm$ laser wavelenght, were suficient to burn the region enlighted(more information in section 3.1.2). We also had to be careful when choosing the wavelenghts available in the apparatus, since for the 488 $nm$ and 532 $nm$ wavelenghts, small variations in the laser power could go from no suficient signal to a sample burning at the laser spot and neighborhood.

### 3.1.2 Raman Spectrocopy Measurements

In order to collect the Raman spectra of the samples, we used a WiTec Alpha 300 RA confocal Raman spectroscope, as shown in the figure 12.

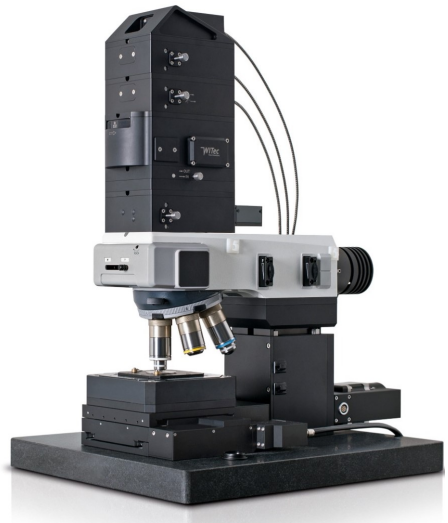


**Figure 12** – Central module microscope apparatus from the Witec Confocal Raman spectrometer, where the samples were measured.

The apparatus had available three possible laser lines: $457nm$, $523nm$ and $633mn$. For the measuring of the compounds presented in this work, we used the $633nm$ He-Ne line. The $633nm$ He-Ne laser sent in this spectroscope is linearly polarized, and both the laser-to-microscope and microscope-to-spectrometer coupling are made with optical fibers. The optics, including the gratings of the spectrometer, are polarization dependent, and the system configuration is chosen to maximize the system´s optical efficiency.

The backscattered Raman signals were collected by a 10 times/0.25 NA Zeiss EC Epiplan objective lens with accumulation time of 30 seconds, sent to a back-illuminated Charged-Coupled Device (CCD), located after a 600 g/mm, BLZ=500 nm grating. The laser power was adjusted to 4.0 mW as measured by at the sample location. In total, a set of 38 compounds were measured (see figure 16 in section 4.1), including quinones and derivative compounds. Since these molecules have aromatic rings in their structures, it was possible to observe a wide line of luminescence in the spectra of most of the compounds, generating a baseline in the Raman spectrum, which was removed in the data treatment with the Project FOUR 4.1 WiTec software.

## 3.2 Computational Methods Applications

### 3.2.1 Simulational data

This section explains the fundamental aspects of the vibrational simulations of the molecules, developed by Prof. Helio F. dos Santos, from *Núcleo de Estudos em Química Computacional (NEQC)* at Chemistry Department of the *Universidade Federal de Juíz de Fora (UFJF).*

The structure optimization and vibrational analysis were carried out in the gas phase. In general, the calculated molecules are rigid; however, for those with a flexible side chain, the conformation was defined by rotating the side chain in order to minimize steric contacts.

As a theoretical study, the first step is to optimize the molecular geometry of the studied systems, which in this case was made via the Density Functional Theory (DFT) method. In DFT, the energy of a molecular system is considered as a function of the electronic density in order to describe the many-body phenomena within a formalism of a single particle. The molecular geometries were optimized via DFT using the m062x functional and $6-31+G(2d,p)$ basis-set. The Raman spectra were calculated within the harmonic approximation considering a single molecule, in vacuum, for each compound. For the Raman intensities, was used the equation [81–84]:

$$I_i^R = C(\nu_0 - \nu_i)^4 \nu_i^{-1} B_i^{-1} S_i, \tag{3.1}$$

where $\nu_0$ is the laser excitation frequency, $\nu_i$ and $S_i$ the calculated frequency (in $cm^{-1}$) and Raman scattering activity (in Å$^4$ $amu^{-1}$) for each normal mode. The constant C was set to be $10^{-12}$ and $B_i$=1 [81, 82], this last one is a temperature factor that accounts for the contribution from excited vibrational modes. The calculations were performed using Gaussian 09® software, from which the output files containing the frequencies are visualized in the GaussView® software. Figure 13 illustrates the interface of the GausView software for the molecule of Benzoquinone.

Finally, in order to simulate the Raman spectra, a Lorentzian function was fitted to the calculated values of frequencies and intensities. Scaling factors were not used for frequencies in this first analysis, but it was considered in our methodological analysis to compare with the measured Raman spectra.

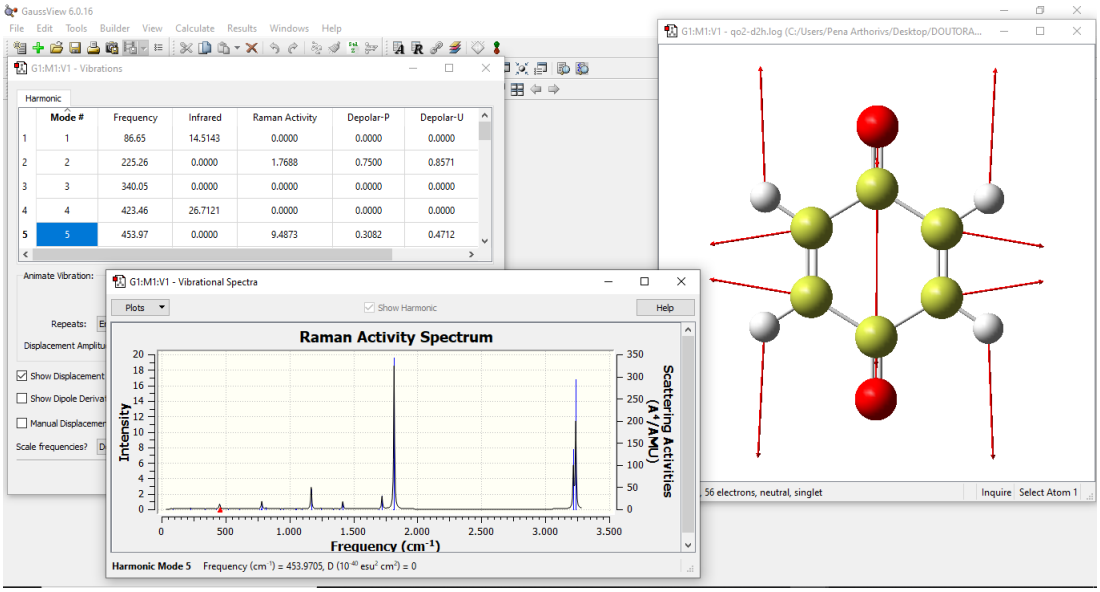



**Figure 13** – Interface of the GaussView Software, where is shown the output for the analysis of the benzoquinone molecule (compound **(1)**) with the Table of calculated vibrational modes, the simulated Raman spectrum and the 3-dimensional animation (with the displacement vectors) of the molecule for each selected mode from the table.

### 3.2.2 *Reduction of the dimensionality* (PCA)

Here the details of the reduction of data dimensionality using Principal Component Analysis (PCA) will be presented. In the process of dimensionality reduction, the number of dimensions (components) that our data set will have at the end of the process, will be equal to the minimum between the number of compounds (the rows of the input data frame) and the number of features (columns of the input data frame). For the simulated Raman spectra, the input data frame is a matrix of 37 rows by 4001 columns, as for the experimental data, in which the input data frame will have 37 rows and 977 columns. In both cases, the PCA Scores matrix, which contains the PCA components, will be a square matrix of $37x37$, as illustrated by figure 14 below for the case of the simulated data:

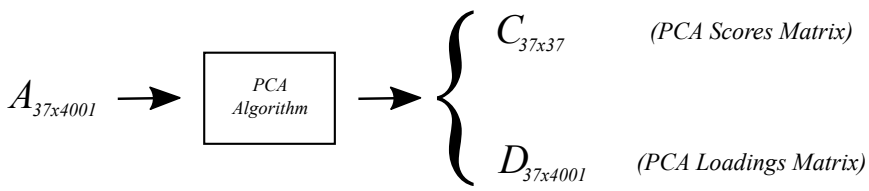



**Figure 14** – Schematic illustration of how the PCA works when applied in a data frame (matrix). The scores matrix represents the positions of the points of the data in the new coordinate system, and the loadings matrix brings the weights for each original variable when calculating the principal component.

Before running our data into PCA algorithm it is first necessary to scale the data in order to make all samples and features be in the same scaling criteria. The scaling

process we used standardizes the features by removing the mean and scaling them to unit variance, by using the StandardScaler library from scikit-learn. We then performed the PCA algorithm from which we selected the three first components, which were plotted it in a three dimensional diagram, and from which we could observe the varioational simmilarities between the compounds in space. The variance of these three components is illustrated below on figure 15. Here we can see that these three components correspond to 70.3% of the total variance.

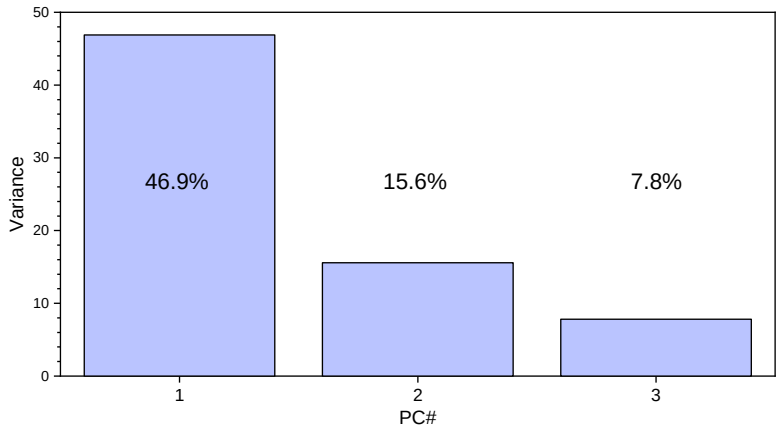



**Figure 15** – Explained variance of the three first PCs considered in our analysis. The first PC corresponds to almost a half of the total variance, and the three first correspond to an amount of 70.3% of the total variance.

The compound **(38)** had the most complex chemical structure of all the compounds (also in terms of its vibrational spectrum, simulated and experimental), so, the processes of scaling and and PCA calculation were being compromised due to the complexity of its data, so the best solution were to remove the compound **(38)** from the Scaling and PCA calculation processes. The treatment on this compound shall be discussed in the PCA reconstruction section.

### 3.2.3 Choosing and finding the K Clusters

As discussed in the section 2.3.2, the way one can choose the number of clusters for the algorithm to find will depend on the context in where the problem is. In our case we do have compounds with structural similarities and differences that can be noticed by eye. It was possible to estimate that we had between 6 and 8 groups with different aspects, combined with the interpretation of the resulting PCA plot (if analyzed separately before the running of the k-means algorithm). Then, we tested the K-means algorithm for 6, 7 and 8 clusters, and we found more suitable to keep a total of 7 clusters for the algorithm to find, according with the chemical structural aspects of the samples and the dispersion of

the points in the PCA plot. The algorithm was applied the dataframe containing the three selected PCA coordinates. Since the k-means follows an euclidean method of distancing for the calculation of the centroids , and considering that PCA respects the (already mentioned) variance hierarchy among its components, we multiplied the each considered PC (which were represented by the columns of the matrix) by its respective variance before applying the k-means algorithm, considering the 37 samples scaled for the PCA calculations. The result of the clustering would reflect into a new, numeric column in the PC matrix which we called the "labels", from which, each number would correspond into a color code in the PCA plot. We shall see the final coloured plot in the results discussion section 4.

### 3.2.4 Spectral Ordering

One of our intentions in this work was to develop a method to analyse the compounds also in terms of their chemical structure complexities, and since PCA and K-means showed a good behavior in terms of the grouping in three-dimensional space, we decided to investigate how the samples would order from the most simple structure to the most complex. When running the PCA with all 38 spectra, we find compound **(38)** to be too far away from all the others, as already mentioned, compromising the metrics and variational calculations. In order to investigate the behavior of the other 37 compounds, which where closer to each other, in relation to the distance of the compound **(38)**, we considered compound **(38)** as the most different and executed PCA again excluding compound **(38)** and considering it the last in the ordering process. The spectral ordering for the other 37 samples were then calculated by an Euclidean-based metric calculation, considering as the three-dimensional coordinates, the three first principal components in the PCA, weighted by their respective variance, as follows:

$$d = \sqrt{PC1_{var}(x_{37} - x_j)^2 + PC2_{var}(y_{37} - y_j)^2 + PC3_{var}(z_{37} - z_j)^2}, \quad j = 36, 35, 34, ..., 1, \tag{3.2}$$

where x stands for the PC1 axis, y for PC2 axis, z for the PC3 axis, and j stands for each of the samples in decrescent order, from 1 to 36, all calculated with respect to the most distant sample in this case, which is sample 37.

### 3.2.5 Spectral reconstructions at the first principal component

Once obtained the disposal of the samples spectra points in the PCA space, we decided to investigate their relative positioning in terms of their variation from a spectrum to another. To do this, we use the data referent to the 3 PCs we considered in our PCA plot, by selecting three first columns from the scores matrix (represented by $C'_{37\times3}$), multiplying them by the three first rows of the Loadings matrix (represented by $D'_{3\times4001}$), and applying an inverse transformation of scaling in the resulting matrix, which we call $Rc_{37\times4001}$, as shown bellow:

$$C'^{i}_{37\times1} \cdot D'^{i}_{1\times4001} = Rc^{i}_{37\times4001}, \qquad i = 1, 2, 3. \tag{3.3}$$

We have as a result three matrices of 37 rows by 4001 columns, representing the reconstruction of the Raman Spectra at each of the three Principal Components, each row representing one of the 37 considered samples (remember that we disregarded compound **(38)**), and each columns being a Raman spectrum point. Plotting each of these rows shall give us the variational behavior of the spectra in relation to each of the three Principal components, and show which band of the Raman spectra most contributed to the sample to be in that position at the three-dimensional PCA space.

# 4 Results and discussion

## 4.1 Experimentally measured Raman spectra of the 38 compounds

Figure 16 shows their experimental Raman spectra in the region between 40 and 1800 $cm^{-1}$. Spectra **(1")** and **(5")** relates to, respectively, the amorphous character of compound **(1)**, and the spectrum compound **(5)** with the light polarized to the largest crystal axis (90° rotation from spectrum **(5)**). The compounds are ordered based on the degree of complexity of their Raman spectra, according to *Principal Component Analysis (PCA)*, as discussed here.

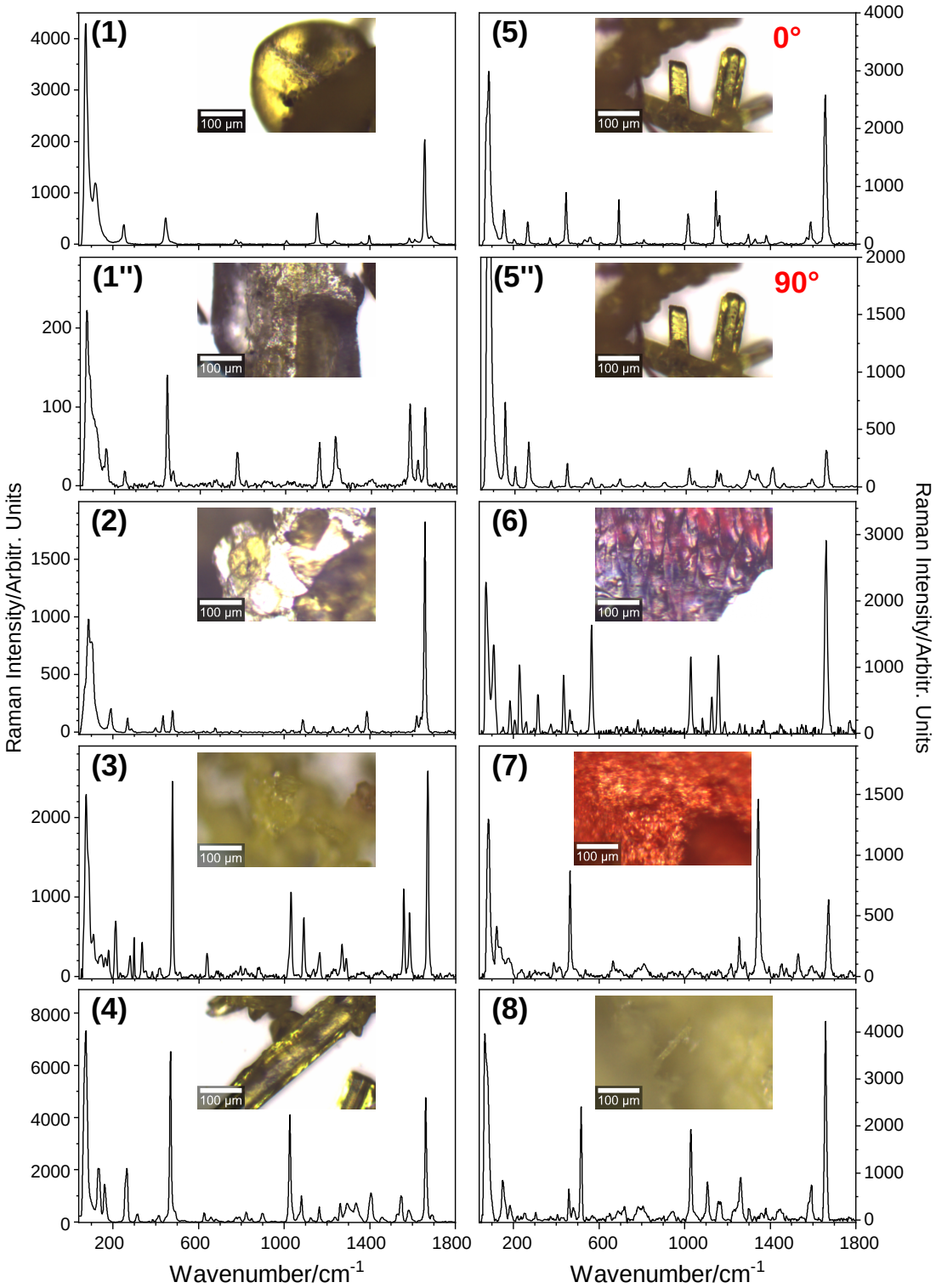



**Figure 16** – Raman spectra of the compounds (see Table 1 in appendix B for names) in the spectral region between 40 and 1800 $cm^{-1}$. Also shown are the compounds' photo-image obtained through a microscope (10x objective). **(1")** and **(5")** show the spectra of, respectively, the amorphous character of compound **(1)**, and the spectrum of 90° rotation of compound **(5)** with respect to the larger crystal axes.

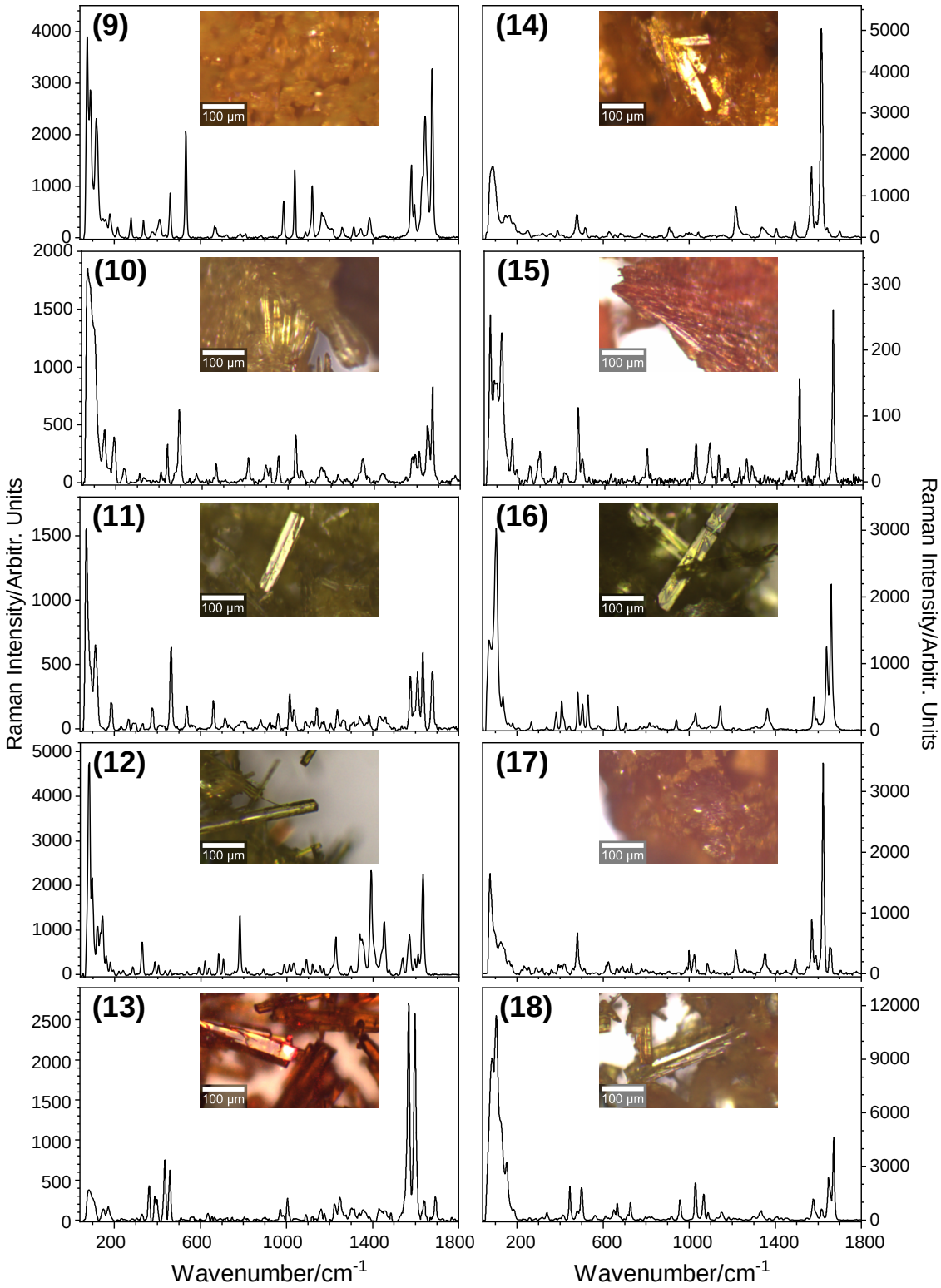



**Figure 16** (cont.)

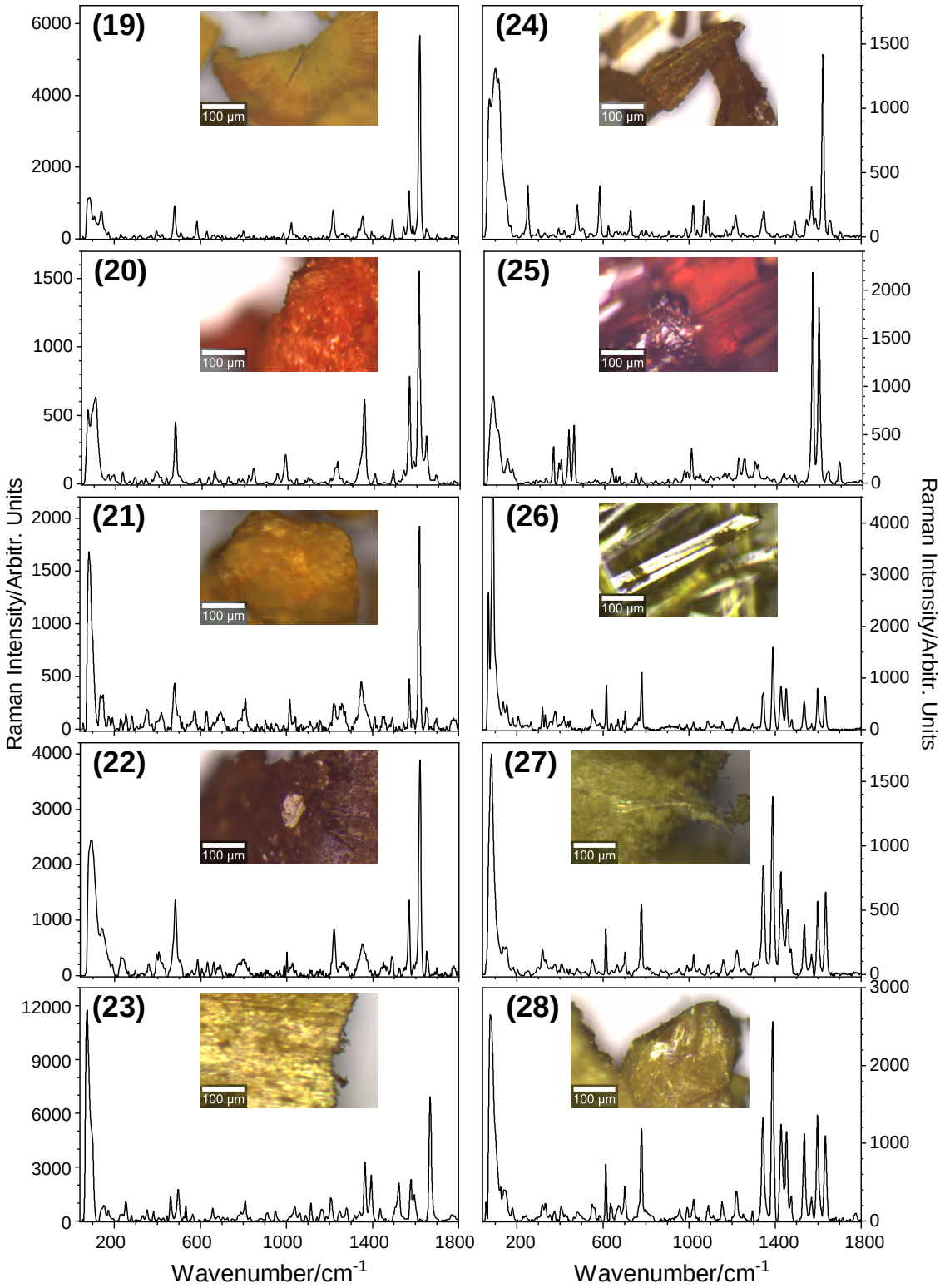



**Figure 16** (cont.)

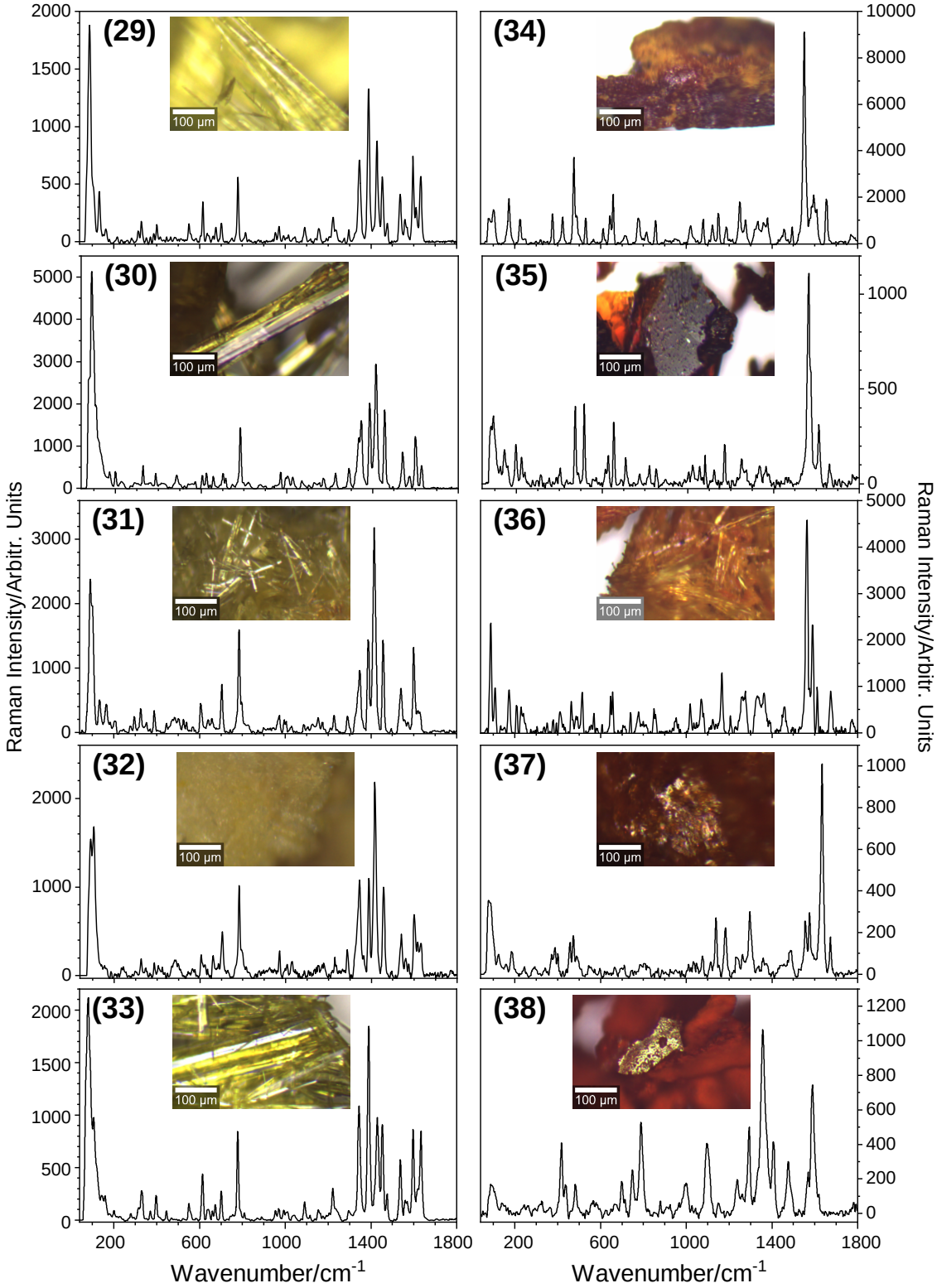



**Figure 16** (cont.)

## 4.2 Comparison Between the simulated and measured spectra

In order to check the accuracy of the simulational results of the vibrational spectra, we made a detailed study comparing the simulated vibrational spectra with the obtained experimentaly. The experimental data onde has influence from other optical phenomena like luminescence, creating a baseline in the spectrum. Futhermore, samples instabilities cause loss of signal, ethalon fringes appear due to the grating of the spectrometer, etc, influencing the signal quality. Figure 17 shows the comparison between the measured and simulated Raman spectra for the compounds **(1)**, **(5)**, **(4)** and **(24)**.

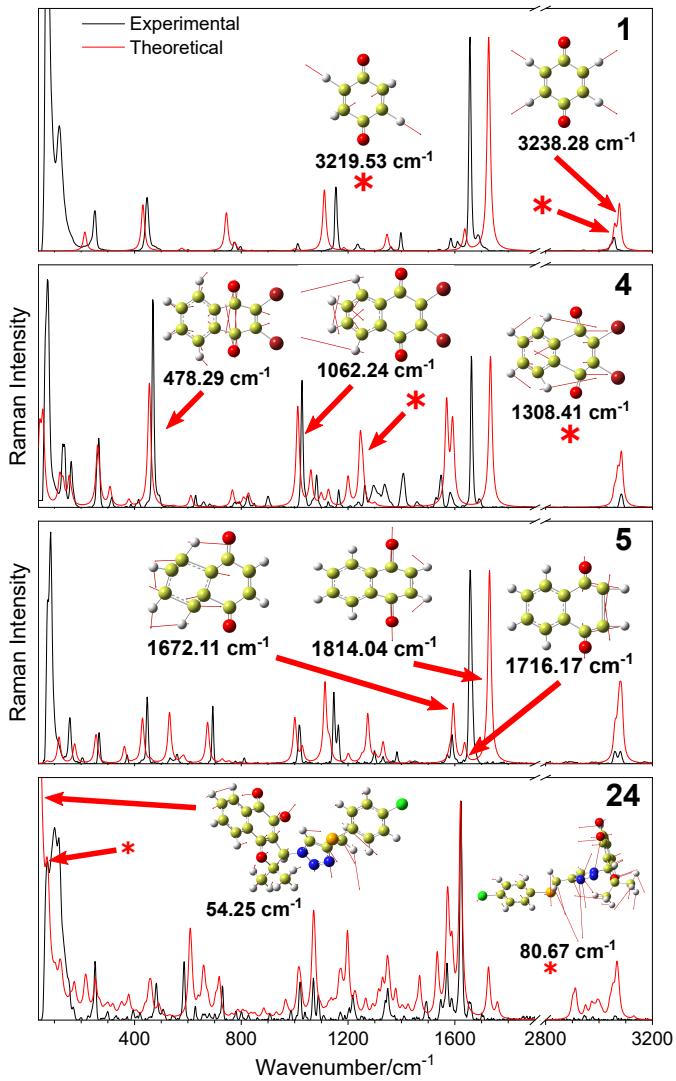



**Figure 17** – Experimental (black line) and calculated (red line) Raman spectra for the molecules **(1)**, **(5)**, **(4)** and **(24)** (see table B in the appendix B). Some specific vibrational modes are highlighted for some spectral regions to illustrate the type of vibration for different frequency ranges.

By comparing the predicted high frequency region (>1000 $cm^{-1}$) profile with experimental, we see that the calculated frequencies are overestimated due to the use of

harmonic approximation, making the calculated spectra to look wider than the measured data. A multiplication factor of 0.95 [85] was applied to the frequency scale in order to make the highest frequency bands (above 3000 $cm^{-1}$) graphically aligned, such that the simulated and experimental spectra could fit each other reasonably. In the region below 100 $cm^{-1}$ (Fig. 17) strong peaks are experimentally observed. Some normal modes are also calculated in this region, assigned to out-of-plane vibration of the entire molecule, as shown in Fig. 17 for the molecule **(24)**, for the vibrational modes 54.25 and 80.67 $cm^{-1}$, respectively. These modes have very small Raman scattering activity, but high Raman intensity due the low frequencies (see Eq. 1). The analysis of these vibrational modes represented in Fig. 17 and assignment must be done with care. Some molecules from the set studied here (1,4-benzoquinone, naftoquinone, lausone, among others) showed intense bands in this low frequency region, which is not predicted theoretically, because while theory considers only a single molecule (as already discussed) in vacuum, the real compounds are in a solid phase, some with a well defined crystalline character.

In Fig. 17 it is possible to notice that the more complex the chemical structures of the compounds are, the more complex is the measured and calculated Raman spectra, and it is possible to notice that the experimental data does not show all the modes activated by the laser, due to the mentioned phenomena at the begining of this section. As the number of scatterers increases, more susceptible to luminescence phenomena the samples are, such that, when subtracting the baseline of the experimental spectra, the peaks of some regions shall be lost. This loss os information is better noticeable when looking for the vibrational modes in the region arround 3000 $cm^{-1}$, when comparing samples **(1)** and **(24)** experimental spectra, one can notice that for the former, it is easy to see the peak in 3056 $cm^{-1}$, as for the latter, almost none of the peaks arround 3000 $cm^{-1}$ can be seen.

## 4.3 Discussing the Principal Components

### 4.3.1 Ordering of the Samples Through the PCA Scores

To order the spectra according to spectral complexity, we applied PCA for processing the similarity among the 37 simulated Raman spectra data, and the K-means clustering classification method in order to partition the clusters observed in the PCA. The first three principal components (PC1, PC2 and PC3) accounting for 70.25% of the total spectral variance(PC1: $46,88\%$; PC2: $15,55\%$; PC3: $7,82\%$)(see figure 15), are shown in figure 18, each point representing one of the 37 spectra, collored accoding to the K-means clustering labeling output. From these images it is already possible to notice the formation of small clusters of points, even without considering the collored labels. Compound (38) was not

considered in this analysis due to a signicantly larger distance from the others, interfering in the understanding of the plot by grouping too closely all the other 37 data points. (38) appears further away along the same PC-direction as compound (37). We defined then 7 clusters (or 8, including sample**(38)**) to better describe the similarities and differences among the samples.

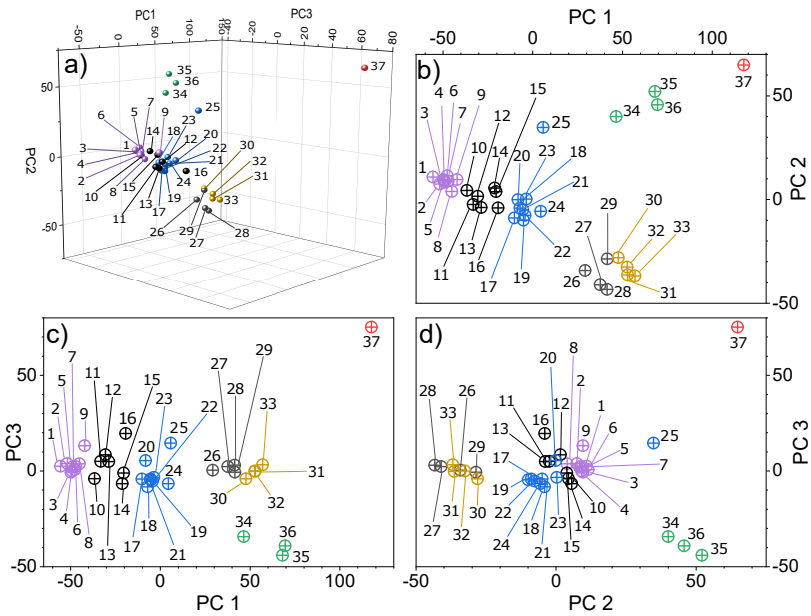



**Figure 18** – PCA scores plots relative to the theoretical spectra of compounds **(1)** to **(37)**. a) Three-dimensional (3D) scatter plot of the three first Principal Components (PCs) (70.3% of the total variance). The 2D plots are shown in b), c) and d) to give a better notion about the relative distances between the compounds. The distances between points were calculated as a weighted norm relative to the most isolated (in this case, **(37)**).

As we mentioned in the section 3.2, the PCA algorithm had as input, for the simulated data a 38x3800 dimensional matrix, where 38 is the number of compounds and 3800 is the number of points in one spectrum, one point per $cm^{-1}$. For the experimental data, similarly we utilized a 38x977 dimensional matrix, where 977 is the number of experimental spectral Raman data, one point per 2.1 $cm^{-1}$ on average within the 40-1800 $cm^{-1}$ spectral range. In the process of comparing both data sets, we checked that the difference between both pitches did not interfere in the PCA output data.

Figure 19 shows the Raman spectra of the 38 compounds, both (a) the simulated and (b) the experimental data in a heat map (see figure 16 for each experimental data separately). From this figure is possible to observe the general behavior of the Raman Peaks that define the spectrum of each sample, and how the spectra complexity evolves as does the molecule complexity. The spectra are ordered, from bottom to top, according to increased spectral complexity, as defined by the PCA ordering trained with the simulated data. Figure 19(b) shows the same ordering of (a) applied in the experimental data by hand to show the behavior in experimental Raman spectra. The ordedring of experimental

data that results from the application of the PCA model and the ordering in 3.2 and the comparison with the simulated data will be shown in figure 20.

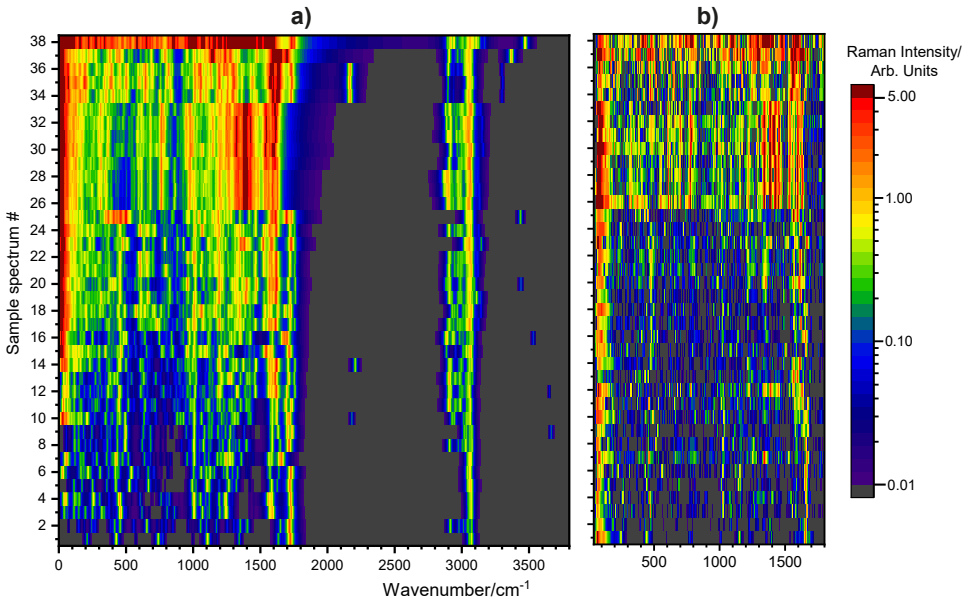



**Figure 19** – Heat scale plot for the Raman spectra of the 38 compounds. Each horizontal line corresponds to one Raman Spectrum. **a)** Simulated Raman spectra in the region between 0 and 3800 $cm^{-1}$. **b)** Experimental Raman spectra in the region between 40 and 1800 $cm^{-1}$. In **b)**, the region above 1800 $cm^{-1}$ was removed due to the presence of Etalon fringes.

Figure 20(a) plots the PCA compounds ordering of the simulated versus the experimental data, showing that the simulated data is a considerably consistent representation of the experimental data, so that analyses and predictions can be made here according to the information provided by the simulated data. The relevance of polarization configuration dependence is shown in figure 20(b), where the polarization scattering geometry of samples 1, 4, 5 and 24, which are samples with macroscopic crystalline aspect, were modified (see caption). Figure 20(c) shows the plot of the PCA-based theoretical spectral ordering on the X axis, and on the Y axis the respective number of atoms N, and will be moreee discussed in the section 4.3.3.

### 4.3.2 Spectral reconstructions at the first principal component

Figure 21 shows the reconstructions of the Raman spectra of some samples in PC1, using the methodology applied by Campos, *et al* [41]. From **a)** to **f)** three examples are displayed, representing the center and the two extremes of each cluster partition. For each partition, the Raman spectra of the selected samples (above), and their reconstructions in PC1 plot (below) are shown. These composition plots give the weights of the Raman modes that mostly contributed for the PC1 variance (and, consequently, the distancing

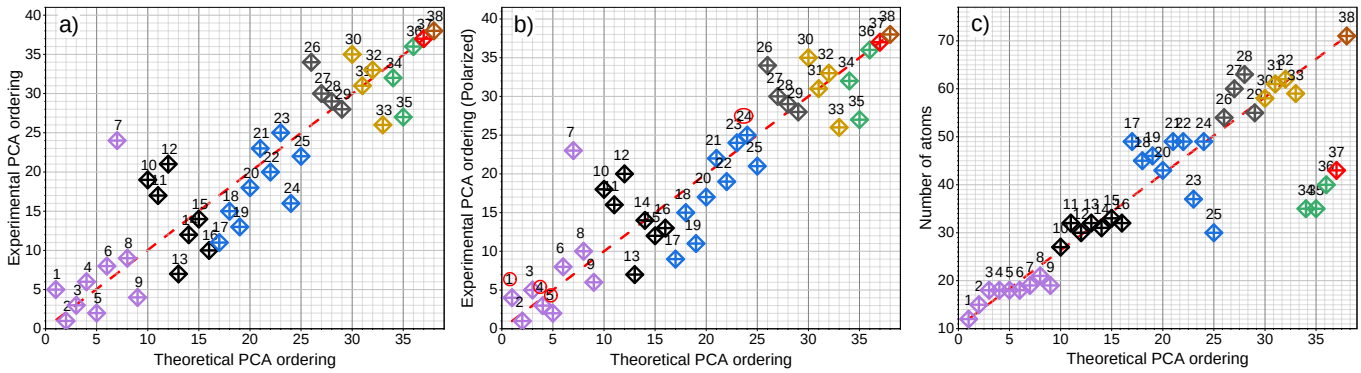



**Figure 20** – a) Plot of the PCA compounds ordering of the simulated versus experimental data. Compound numbers on top of each data point and cluster colors indicating the K-means partitioning are based on the simulated data analysis. The red dashed line represents a figurative perfect match between theoretical and experimental orderings. b) Plot of the PCA compounds ordering of the simulated versus experimental data for 90° rotation of some of the samples (circled numbers). c) Plotting the PCA-based theoretical spectral ordering versus the respective number of atoms N.

between the points in figure 18) individually for each sample.

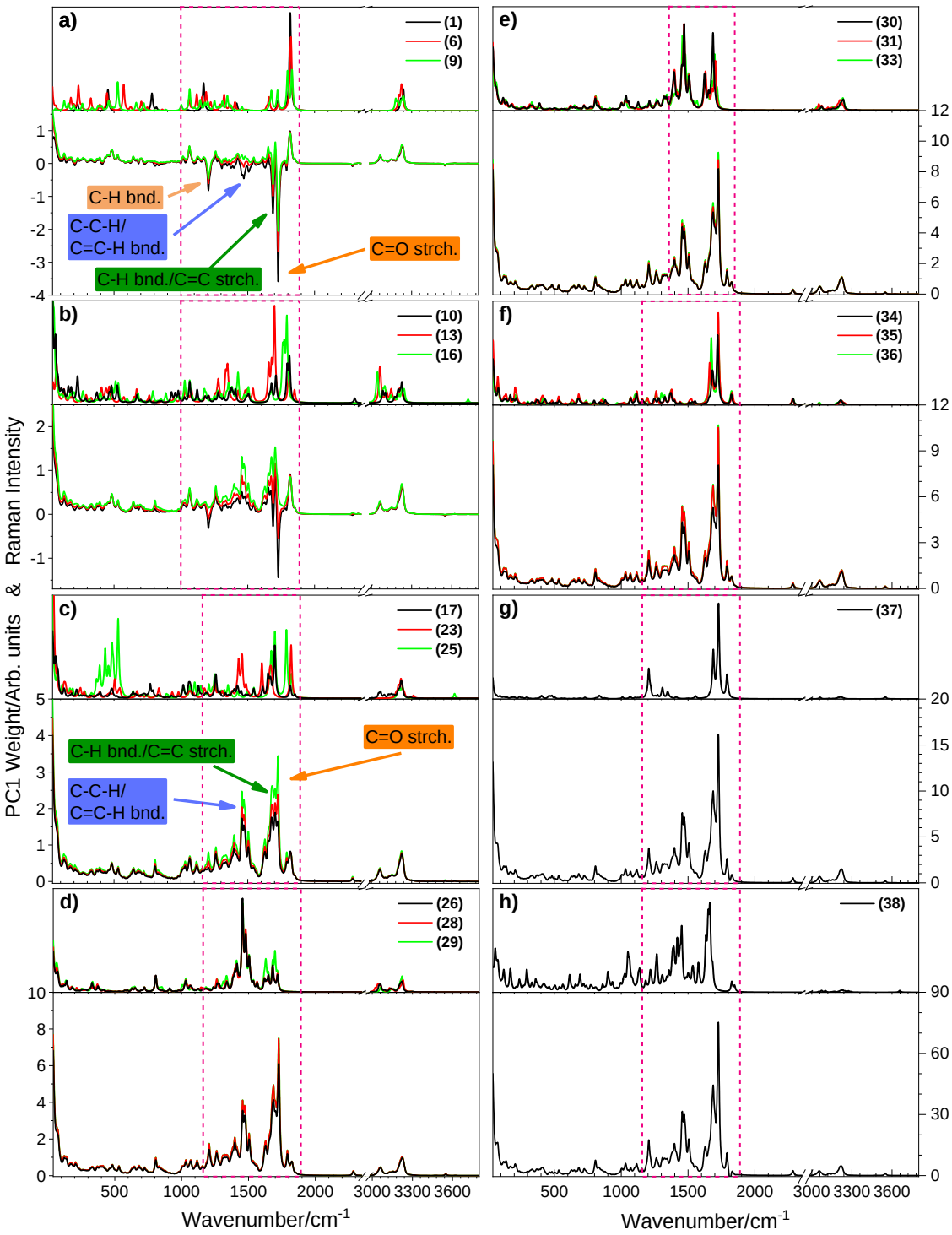



**Figure 21** – **a)** to **h)**: Raman spectra (top) and Raman spectra reconstructions in PC1 (bottom) of selected samples. Each curve stands for one sample, as displayed in the legends. At the bottom plot of **a)** and **c)** the main vibrational modes with larger variance are indicated ("bnd." stands for bending, and "strch." means stretching). Partition **h)** (bottom) shows the prediction of the spectral composition to compound **(38)** using the PCA parameterized to the other 37.

Between the dashed red lines in each plot are the most characteristic modes of quinoidal compounds (top) [30–32] and the analogue PCA composition regions with the

most expressive variations (bottom) between the samples: mostly, the C-H bendings, C=C and/or C=O stretchings, as well as the association between those vibrational modes. In figure 22, we bring some visual examples of the vibrational modes in the range between 1700 $cm^{-1}$ and 1900 $cm^{-1}$ for the samples **(1)** (1723.52 $cm^{-1}$), **(25)** (1702.63 $cm^{-1}$), **(35)** (1725.56 $cm^{-1}$) and **(38)** (1826.70 $cm^{-1}$). The main vibrational modes associated with the regions of higher variation are labeled in a) and c) in figure 21. From figure 21 it is possible to realize that the general variance of the molecular vibrations within one cluster partition is similar, changing most significantly from one partition to another.

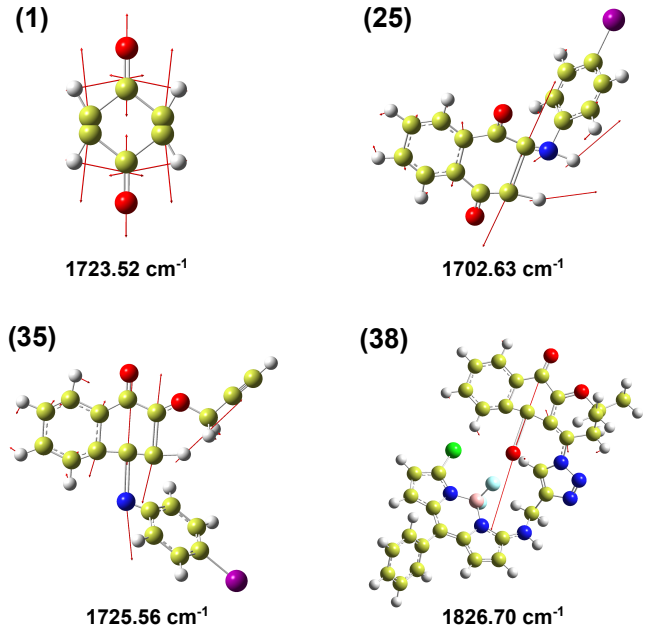



**Figure 22** – Visual examples of the vibrational modes in the range between 1700 $cm^{-1}$ and 1900 $cm^{-1}$ for the samples **(1)**, **(25)**, **(35)** and **(38)**. The relative vibrational modes are labeled for each respective sample.

### 4.3.3 Ordering and clustering interpretation

Figure 23 shows the molecular structures for the 38 quinoidal and derivative molecular systems, ordered according to the spectra-based PCA. The more complex the chemical structure is, the more complex will be the Raman spectrum (compare figures 19 and 23).

One important aspect defining the complexity of the Raman spectra is the number of atoms N, which defines the number of vibrational modes as 3N-6. This aspect is explored in figure 20(c), where we plotted the PCA-based theoretical spectral ordering on the X axis, and on the Y axis the respective number of atoms N. The data points follow roughly the diagonal (dashed line), indicating the relevance of N (or the equivalent 3N-6) on defining

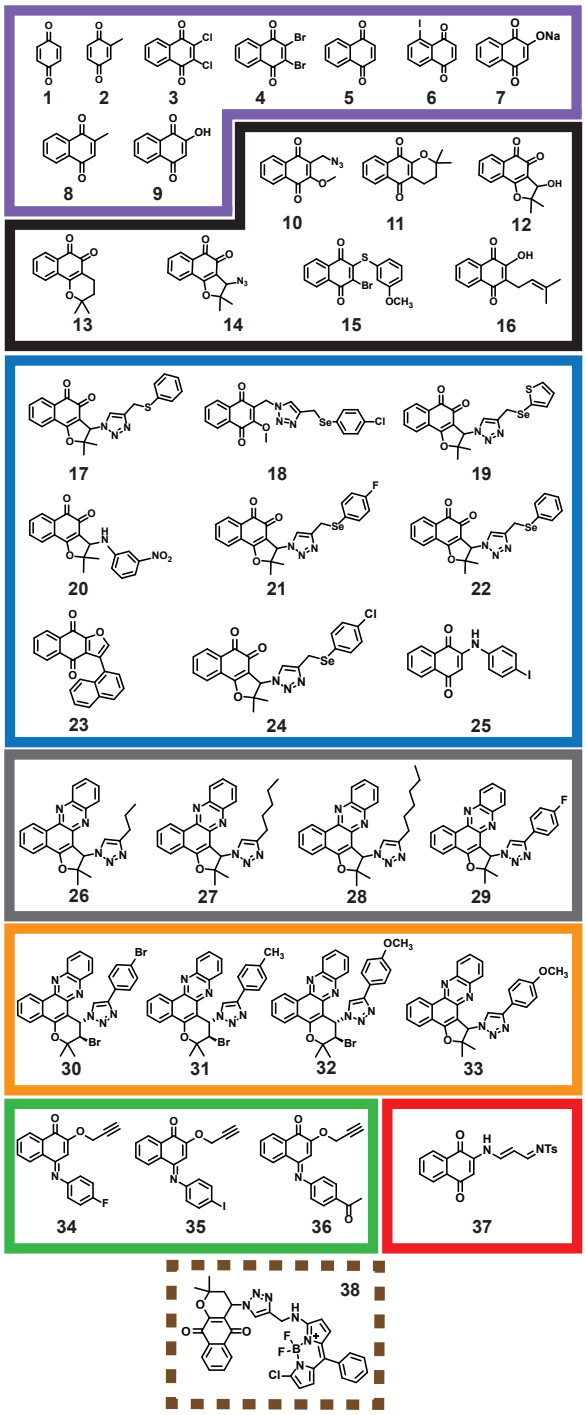



**Figure 23** – Schematic organization of how the molecules grouped together in according to the PCA relative distances. The grouping boxes follow the same color-code used in figure 3. The dashed brown box refers to the compound number **(38)**, disregarded in figure 18.

the spectral complexity, as expected, and is supported by figure 19(a). However, the data spread from the dashed line shows that the spectral PCA depends not only on the number of vibrational modes, but also on their specific Raman cross sections and frequencies, which depend on the type of elements and their location in the molecular structure. For example, the spectral ordering within the cluster of spectra from 1 to 9, or the clustering

of larger molecules, such as the ones related to spectra from 25 to 37, cannot be explained only by N.

The first cluster (purple box) is composed by the simplest structures, namely, *p*-benzoquinones and *p*-naphthoquinones, with single atoms or small substitutions (for instance Cl, Br, I, OH, ONa or $CH_3$) bonded to the main benzo- or naphthoquinone structure. In the second cluster (black box) are found the first *o*-quinones of the whole set of samples (**(12)**, **(13)**, **(14)**), and the molecules have substitutions larger than the first cluster, with aromatic ring substituents or a long open chain, like for sample **(16)**. The third (blue box) cluster shows the set of quinones with longer and more complex pattern of substitutions, being mainly characterized by the presence of sequential aromatic substituents or by the presence of nitrogen atoms in the substitutions. Notice that the samples being "*ortho*-quinone" or "*para*-quinone" do not represent a determinant factor for the ordering/classification considering their vibrational characteristics.

The fourth (gray box) and the fifth (yellow box) clusters are characterized by phenazines with more complex substituents. Open chains of aliphatic compounds (Alkanes) or aromatic sequences are found. These two clusters are very close to each other in the PCA scores (see figure 18). Compound **(29)**, for example, which contains triazole ring and substituted phenyl as all compounds in group 5, falls into group 4 according to the K-means analysis. From the mathematical point of view, the ordering is dictated by PC1, which has the highest variance (notice the PC1 ordering of samples **(28)** and **(29)**, for example). From the physical-chemistry point of view, the fifth cluster is characterized mainly by the presence of a bromine atom in the aromatic chain substituents and by an aromatic ring bonded in the triazole, and these structural aspects should be responsible for the actually obtained clustering. The sixth (green box) cluster is characterized by the group of alkynes substituents, with the complexity defined by the size of the structure that ends the bond of the aromatic ring, the last one being a carbonyl bonded in the aromatic ring. Sample **(37)**, characterized by the tosyl substituent, illustrated in figure 24, is by itself the seventh (red box) cluster. The bodipy substituent characterizes the eighth (brown box) cluster, with sample **(38)** (see figure 22) being the most complex structure, with the larger bonded structure, relatively to the other molecules.

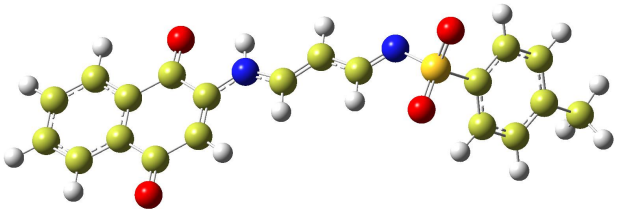


**Figure 24** – Three-dimensional representation of the sample **(37)**. The Tosyl substituent, represented by a TS in the structure represents a more complex structure containing a sulfur atom bonded by two oxygem atoms and to a benzene ring, ending the structure with a $CH_3$ bond.

# 5 Conclusions

In this study, 38 samples of different quinoidal compounds and derivative molecular systems were measured via back scattering confocal Raman spectroscopy and simulated via DFT and molecular dynamics under harmonic approximation.

Our algorithm was able to compute the ordering of the Raman spectra (and so the structures) based on the variace in the regions related mostly to the C-H bendings, C-C and C=X stretching (X = C, O or N) vibrational modes, with the higher weight relative to the C-H bending and C=X stretching from the quinoidal or phenazinic nuclei structures (C-H and C=X modes) and sustituents (C-H modes) (exception for the cases of the sixth-cluster (green box) samples **(34)**, **(35)** and **(36)**, where there was the presence of an alkyne (C≡C), which was not present in any other sample). The obtained ordering was found to be relative not only to the size (number of atoms) of the chemical structure, but also to how the aromatic substitutions are bonded to the main structure. The analysis of the first principal compontent (PC1) shows that the spectral distribution in the PC1 weights are similar within a same K-means partition, changing significantly when compared to the spectral composition distribution among clusters.

Therefore, we demonstrate that PCA and K-means clustering Raman-based analysis can be utilized to structurally order and classify molecular systems. Interestingly, we found in the literature information that indicates a link between the clusters divisions and biological/pharmacological aspects of some of the samples, like the antifungi activity for the samples **(3), (4)** and **(5)** [86] from the purple cluster in figs. 18 and 23 and HIV-1 inhibition activity for the samples **(10), (11)** and **(15)** [87] (black cluster in figs. 18 and 23), indicating that the method utilized here might be a way of grouping and/or selecting similar compounds not only by its physical/spectroscopic characteristics, but also biological/pharmacological applications.

Finally, the method discussed here should not be applicable only to molecules, but also to other amorphous or crystalline solids. In this sense, it is important to stress that with the advance of lasers and detectors, Raman spectroscopy is gaining importance very rapidly (see Figure 25) [88]. Furthermore, the development of theoretical techniques has triggered new and large amount of theoretical Raman data, within the materials genome initiative [45]. For example, Taghizadeh *et al.* [89] created the "Computational 2D Materials Database (C2DB)" based on calculated Raman spectra of 733 different two-dimensional systems. In this perspective, the method introduced here might be very helpful for the analysis of greater amounts of vibrational and spectral data in physical chemistry, useful in the concept of accelerated discovery of novel materials with specific functionalities.

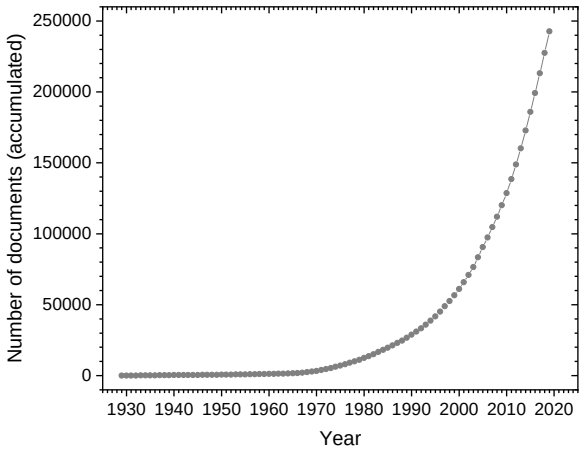



**Figure 25** – Accumulative number of Raman papers in the literature. The data are built based on the Scopus database using the following search expressions in the "keyword, title, or abstract" fields (date of search, September 17, 2020): RAMAN: "Raman spectr*" OR "Raman microsc*" OR "Raman scat*".

# Bibliography


[1] Bolton, Judy L, Michael A Trush, Trevor M Penning, Glenn Dryhurst e Terrence J Monks: *Role of quinones in toxicology.* Chem. Res. in Tox., 13(3):135–160, 2000. Citado 2 vezes nas páginas 10 e 18.

[2] Ghosheh, Omar A, Abdulghani A Houdi e Peter A Crooks: *High performance liquid chromatographic analysis of the pharmacologically active quinones and related compounds in the oil of the black seed (Nigella sativa L.).* Journal of Pharmaceutical and Biomedical Analysis, 19(5):757–762, 1999. Citado na página 10.

[3] Castro, Frederico Augusto Vieira, Diana Mariani, Anita Dolly Panek, Elis Cristina Araújo Eleutherio e Marcos Dias Pereira: *Cytotoxicity mechanism of two naphthoquinones (menadione and plumbagin) in Saccharomyces cerevisiae.* PloS One, 3(12):e3999, 2008. Citado na página 10.

[4] Sousa, Eliane Teixeira, Wilson A Lopes e Jailson B de Andrade: *Fontes, formação, reatividade e determinação de quinonas na atmosfera.* Química Nova, 39(4):486–495, 2016. Citado na página 10.

[5] Dantas-Pereira, Luíza, Edézio F Cunha-Junior, Valter V Andrade-Neto, John F Bower, Guilherme AM Jardim, Eufrânio N da Silva Júnior, Eduardo C Torres-Santos e Rubem FS Menna-Barreto: *Naphthoquinones and Derivatives for Chemotherapy: Perspectives and Limitations of their Anti-trypanosomatids Activities.* Current Pharmaceutical Design, 27:1807–1824, 2021. Citado na página 10.

[6] Lucas, Nanci C de, Aurélio BB Ferreira e José Carlos Netto-Ferreira: *Fotoquímica de naftoquinonas.* Revista Virtual de Química, 7(1):403–463, 2015. Citado na página 10.

[7] Chaudhary, Ankita e Jitender M Khurana: *Synthetic routes for phenazines: an overview.* Research on Chemical Intermediates, 44(2):1045–1083, 2018. Citado 3 vezes nas páginas 10, 17 e 18.

[8] Jardim, Guilherme AM, Eduardo HG Cruz, Wagner O Valença, Jarbas M Resende, Bernardo L Rodrigues, Daniela F Ramos, Ronaldo N Oliveira, Pedro EA Silva e Eufrânio N da Silva Júnior: *On the search for potential antimycobacterial drugs: synthesis of naphthoquinoidal, phenazinic and 1, 2, 3-triazolic compounds and evaluation against Mycobacterium tuberculosis.* Journal of the Brazilian Chemical Society, 26(5):1013–1027, 2015. Citado 6 vezes nas páginas 10, 17, 18, 68, 69 e 70.

[9] Sidman, Jerome W: *Electronic States of p-Benzoquinone1.* Journal of the American Chemical Society, 78(11):2363–2367, 1956. Citado 2 vezes nas páginas 10 e 18.

[10] Eckert, Timothy S e Thomas C Bruice: *Chemical properties of phenanthrolinequinones and the mechanism of amine oxidation by o-quinones of medium redox potentials.* Journal of the American Chemical Society, 105(13):4431–4441, 1983. Citado 2 vezes nas páginas 10 e 18.

[11] Silva, Milton N da, Vítor F Ferreira e Maria Cecília BV Souza: *Um panorama atual da química e da farmacologia de naftoquinonas, com ênfase na beta-lapachona e derivados.* Química Nova, 26(3):407–416, 2003. Citado 2 vezes nas páginas 10 e 18.

[12] Rötig, Agnès, Julie Mollet, Marlene Rio e Arnold Munnich: *Infantile and pediatric quinone deficiency diseases.* Mitochondrion, 7:S112–S121, 2007. Citado 2 vezes nas páginas 10 e 18.

[13] Ribeiro, Magno R, Pablo P Souza, L DM Ferreira, Sharlene L Pereira, Ingrid da S Martins, Rosangela de A Epifanio, Leticia V Costa-Lotufo, Paula C Jimenez, Claudia Pessoa, Manoel O de Moraes *et al.*: *Natural Furano Naphtoquinones from Lapachol: Hydroxyiso-$\beta$-Lapachone, Stenocarpoquinone-B and Avicequinone-C.* Letters in Organic Chemistry, 8(5):347–351, 2011. Citado 2 vezes nas páginas 10 e 18.

[14] Costa-Lotufo, Leticia V, Raquel C Montenegro, Ana Paula NN Alves, Socorro Vanesca F Madeira, Cláudia Pessoa, Maria Elisabete A Moraes e Manoel Odorico Moraes: *A contribuição dos produtos naturais como fonte de novos fármacos anticâncer: estudos no Laboratório Nacional de Oncologia Experimental da Universidade Federal do Ceará.* Revista Virtual de Química, 2(1):47–58, 2010. Citado 2 vezes nas páginas 10 e 18.

[15] Lu, Jin Jian, Jiao Lin Bao, Guo Sheng Wu, Wen Shan Xu, Ming Qing Huang, Xiu Ping Chen e Yi Tao Wang: *Quinones derived from plant secondary metabolites as anti-cancer agents.* Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Cancer Agents), 13(3):456–463, 2013. Citado 2 vezes nas páginas 10 e 18.

[16] Wan, Yieh Ping, Thomas H Porter e Karl Folkers: *Antimalarial quinones for prophylaxis based on a rationale of inhibition of electron transfer in Plasmodium.* Proceedings of the National Academy of Sciences, 71(3):952–956, 1974. Citado 2 vezes nas páginas 10 e 18.

[17] Carr, Gavin, Emily R Derbyshire, Eric Caldera, Cameron R Currie e Jon Clardy: *Antibiotic and antimalarial quinones from fungus-growing ant-associated Pseudonocardia sp.* Journal of natural products, 75(10):1806–1809, 2012. Citado 2 vezes nas páginas 10 e 18.

[18] Pinto, AV, RFS Menna-Barreto, SL De Castro *et al.*: *Naphthoquinones isolated from Tabebuia: a review about the synthesis of heterocyclic derivatives, screening against Trypanosoma cruzi and correlation structure-trypanocidal activity.* Phytomedicines, páginas 109–127, 2007. Citado 2 vezes nas páginas 10 e 18.

[19] Silva Júnior, Eufrânio N, Maria Aline BF Moura, Antonio V Pinto, Maria do Carmo FR Pinto, Maria Cecília BV Souza, Ana J Araújo, Claudia Pessoa, Letícia V Costa-Lotufo, Raquel C Montenegro, Manoel Odorico de Moraes *et al.*: *Cytotoxic, trypanocidal activities and physicochemical parameters of nor-$^2$-lapachone-based 1, 2, 3-triazoles.* Journal of the Brazilian Chemical Society, 20(4):635–643, 2009. Citado 2 vezes nas páginas 10 e 18.

[20] Silva Junior, Eufranio N, Isadora MM Melo, Emilay BT Diogo, Verenice A Costa, José D Souza Filho, Wagner O Valença, Celso A Camara, Ronaldo N de Oliveira, Alexandre S de Araujo, Flávio S Emery *et al.*: *On the search for potential anti-Trypanosoma cruzi drugs: Synthesis and biological evaluation of 2-hydroxy-3-methylamino and 1, 2, 3-triazolic naphthoquinoidal compounds obtained by click chemistry reactions.* European Journal of Medicinal Chemistry, 52:304–312, 2012. Citado 2 vezes nas páginas 10 e 18.

[21] Guimarães, Tiago T, FR Maria do Carmo, Juliane S Lanza, Maria N Melo, L Rubens, Isadora MM de Melo, Emilay BT Diogo, Vitor F Ferreira, Celso A Camara, Wagner O Valença *et al.*: *Potent naphthoquinones against antimony-sensitive and-resistant Leishmania parasites: Synthesis of novel $\alpha$-and nor-$\alpha$-lapachone-based 1, 2, 3-triazoles by copper-catalyzed azide–alkyne cycloaddition.* European Journal of Medicinal Chemistry, 63:523–530, 2013. Citado 2 vezes nas páginas 10 e 18.

[22] Laursen, Jane Buus e John Nielsen: *Phenazine natural products: biosynthesis, synthetic analogues, and biological activity.* Chemical Reviews, 104(3):1663–1686, 2004. Citado 2 vezes nas páginas 10 e 18.

[23] Barry, VINCENT C, JG Belton, Michael L Conalty, Joan M Den-steny, Deirdre W Edward, JF O'sulli-van, Dermot Twomey, Frank Winder *et al.*: *A new series of phenazines (rimino-compounds) with high antituberculosis activity.* Nature, 179:1013–15, 1957. Citado 2 vezes nas páginas 10 e 18.

[24] Cezairliyan, Brent, Nawaporn Vinayavekhin, Daniel Grenfell-Lee, Grace J Yuen, Alan Saghatelian e Frederick M Ausubel: *Identification of Pseudomonas aeruginosa phenazines that kill Caenorhabditis elegans.* PLoS Pathog, 9(1):e1003101, 2013. Citado 2 vezes nas páginas 10 e 18.

[25] Stammreich, H e Th Teixeira Sans: *Molecular vibrations of quinones. IV. Raman spectra of p-benzoquinone and its centrosymmetrically substituted isotopic derivatives*

*and assignment of observed frequencies.* The Journal of Chemical Physics, 42(3):920–931, 1965. Citado na página 10.

[26] Durnick, Thomas J e Samuel C Wait Jr: *Vibrational spectra and assignments for phenazine.* Journal of Molecular Spectroscopy, 42(2):211–226, 1972. Citado na página 10.

[27] Stenman, Folke e Jaakko Räsänen: *On the vibrational spectrum of 1, 4-naphthoquinone.* Spectrochimica Acta Part A: Molecular Spectroscopy, 29(2):405–410, 1973. Citado na página 10.

[28] Delarmelina, Maicon, Glaucio B Ferreira, Vitor F Ferreira e Jose W de M Carneiro: *Vibrational spectroscopy of lapachol, $\alpha$-and $\beta$-lapachone: Theoretical and experimental elucidation of the Raman and infrared spectra.* Vibrational Spectroscopy, 86:311–323, 2016. Citado na página 10.

[29] Sahoo, Sangram Keshari, Siva Umapathy e Anthony W Parker: *Time-resolved resonance Raman spectroscopy: Exploring reactive intermediates.* Applied Spectroscopy, 65(10):1087–1115, 2011. Citado na página 11.

[30] Wang, Sophie X, Nobuhumi Nakamura, Minae Mure, Judith P Klinman e Joann Sanders-Loehr: *Characterization of the native lysine tyrosylquinone cofactor in lysyl oxidase by Raman spectroscopy.* Journal of Biological Chemistry, 272(46):28841–28844, 1997. Citado 2 vezes nas páginas 11 e 42.

[31] Backes, Gabriele, Victor L Davidson, Fienke Huitema, Johannis A Duine e Joann Sanders-Loehr: *Characterization of the tryptophan-derived quinone cofactor of methylamine dehydrogenase by resonance Raman spectroscopy.* Biochemistry, 30(38):9201–9210, 1991. Citado 2 vezes nas páginas 11 e 42.

[32] Moenne-Loccoz, Pierre, Nobuhumi Nakamura, Vincent Steinebach, Johannis A Duine, Minae Mure, Judith P Klinman e Joann Sanders-Loehr: *Characterization of the topa quinone cofactor in amine oxidase from Escherichia coli by resonance Raman spectroscopy.* Biochemistry, 34(21):7020–7026, 1995. Citado 2 vezes nas páginas 11 e 42.

[33] Umadevi, M, A Ramasubbu, P Vanelle e V Ramakrishnan: *Spectral investigations on 2-methyl-1, 4-naphthoquinone: solvent effects, host–guest interactions and SERS.* Journal of Raman Spectroscopy, 34(2):112–120, 2003. Citado na página 11.

[34] Dieringer, Jon A, Adam D McFarland, Nilam C Shah, Douglas A Stuart, Alyson V Whitney, Chanda R Yonzon, Matthew A Young, Xiaoyu Zhang e Richard P Van Duyne: *Introductory lecture surface enhanced Raman spectroscopy: new materials,*

*concepts, characterization tools, and applications.* Faraday discussions, 132:9–26, 2006. Citado na página 11.

[35] Knight, Diane S e William B White: *Characterization of diamond films by Raman spectroscopy.* Journal of Materials Research, 4(2):385–393, 1989. Citado na página 11.

[36] Jorio, Ado, M Dresselhaus, Riichiro Saito e GF Dresselhaus: *Raman Spectroscopy in Graphene Related Systems.* Spectroscopy, 2013. Citado 2 vezes nas páginas 11 e 12.

[37] Yılmaz, Deniz, Beyza Nur Günaydın e Meral Yüce: *Nanotechnology in food and water security: On-site detection of agricultural pollutants through surface-enhanced Raman spectroscopy.* Emergent Materials, 5(1):105–132, 2022. Citado na página 11.

[38] Gadelha, Andreij C, Douglas AA Ohlberg, Cassiano Rabelo, Eliel GS Neto, Thiago L Vasconcelos, João L Campos, Jessica S Lemos, Vinícius Ornelas, Daniel Miranda, Rafael Nadas *et al.*: *Localization of lattice dynamics in low-angle twisted bilayer graphene.* Nature, 590(7846):405–409, 2021. Citado na página 11.

[39] Saraiva, André, Filomeno S de Aguiar Júnior, Reinaldo de Melo e Souza, Arthur Patrocínio Pena, Carlos H Monken, Marcelo F Santos, Belita Koiller e Ado Jorio: *Photonic counterparts of cooper pairs.* Physical review letters, 119(19):193603, 2017. Citado na página 11.

[40] Bonnier, Franck e HJ Byrne: *Understanding the molecular information contained in principal component analysis of vibrational spectra of biological systems.* Analyst, 137(2):322–332, 2012. Citado na página 11.

[41] Campos, João Luiz Elias, Hudson Miranda, Cassiano Rabelo, Emil Sandoz-Rosado, Sugandha Pandey, Juha Riikonen, Abraham G Cano-Marquez e Ado Jorio: *Applications of Raman spectroscopy in graphene-related materials and the development of parameterized PCA for large-scale data analysis.* Journal of Raman Spectroscopy, 49(1):54–65, 2018. Citado 2 vezes nas páginas 11 e 40.

[42] Ye, Zhengmao, Yongmao Ye, Habib Mohamadian, Pradeep Bhattacharya e Kai Kang: *Fuzzy filtering and fuzzy K-means clustering on biomedical sample characterization.* Em *Proceedings of 2005 IEEE Conference on Control Applications, 2005. CCA 2005.*, páginas 90–95. IEEE, 2005. Citado na página 11.

[43] Holliday, John D, Sarah L Rodgers, Peter Willett, Min You Chen, Mahdi Mahfouf, Kevin Lawson e Graham Mullier: *Clustering files of chemical structures using the fuzzy k-means clustering method.* Journal of Chemical Information and Computer Sciences, 44(3):894–902, 2004. Citado na página 11.

[44] Junlin, Li e Fu Hongguang: *Molecular dynamics-like data clustering approach.* Pattern Recognition, 44(8):1721–1737, 2011. Citado na página 11.

[45] de Pablo, Juan J., Barbara Jones, Cora Lind Kovacs, Vidvuds Ozolins e Arthur P. Ramirez: *The Materials Genome Initiative, the interplay of experiment, theory and computation.* Current Opinion in Solid State and Materials Science, 18(2):99–117, 2014, ISSN 1359-0286. `https://www.sciencedirect.com/science/article/pii/S1359028614000060`. Citado 2 vezes nas páginas 11 e 47.

[46] Sala, Oswaldo: *Fundamentos da espectroscopia Raman e no infravermelho.* Unesp, 1996. Citado 3 vezes nas páginas 12, 13 e 14.

[47] Fonseca, Lucas Lafetá Prates da: *Propriedades ópticas não-lineares de terceira ordem em materiais bidimensionais.* 2017. Citado na página 14.

[48] Zettili, Nouredine: *Quantum mechanics: concepts and applications*, 2003. Citado na página 14.

[49] Ballentine, Leslie E: *Quantum mechanics: a modern development.* World Scientific Publishing Company, 2014. Citado na página 14.

[50] Morgon, Nelson H e Kaline Rabelo Coutinho: *Métodos de química teórica e modelagem molecular.* 2007. Citado na página 14.

[51] Hohenberg, Pierre e Walter Kohn: *Inhomogeneous electron gas.* Physical review, 136(3B):B864, 1964. Citado na página 14.

[52] Levine, Ira N, Daryle H Busch e Harrison Shull: *Quantum chemistry*, volume 6. Pearson Prentice Hall Upper Saddle River, NJ, 2009. Citado na página 15.

[53] Morgon, Nelson H e Kaline Rabelo Coutinho: *Métodos de química teórica e modelagem molecular.* 2007. Citado na página 15.

[54] Cramer, Christopher J: *Essentials of computational chemistry: theories and models.* John Wiley & Sons, 2013. Citado na página 15.

[55] Kohn, Walter e Lu Jeu Sham: *Self-consistent equations including exchange and correlation effects.* Physical review, 140(4A):A1133, 1965. Citado na página 15.

[56] Zhao, Yan e Donald G Truhlar: *The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals.* Theoretical chemistry accounts, 120:215–241, 2008. Citado na página 16.

[57] Simón, Luis e Jonathan M Goodman: *How reliable are DFT transition structures? Comparison of GGA, hybrid-meta-GGA and meta-GGA functionals.* Organic & biomolecular chemistry, 9(3):689–700, 2011. Citado na página 16.

[58] Zhao, Yan e Donald G Truhlar: *A density functional that accounts for medium-range correlation energies in organic chemistry.* Organic letters, 8(25):5753–5755, 2006. Citado na página 16.

[59] Hehre, Warren J, Robert Ditchfield e John A Pople: *Self—consistent molecular orbital methods. XII. Further extensions of Gaussian—type basis sets for use in molecular orbital studies of organic molecules.* The Journal of Chemical Physics, 56(5):2257–2261, 1972. Citado na página 16.

[60] Clark, Timothy, Jayaraman Chandrasekhar, Günther W Spitznagel e Paul Von Ragué Schleyer: *Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+ G basis set for first-row elements, Li–F.* Journal of Computational Chemistry, 4(3):294–301, 1983. Citado na página 16.

[61] Almeida, Renata Gomes de: *Aplicação de métodos sintéticos via catálise com sais de cobre(I) e nanotubos de carbono decorados com nanopartículas de rutênio para síntese de novos compostos antitumorais.* Tese de Doutoramento, Universidade Federal de Minas Gerais, 2022. Citado na página 16.

[62] Overvad, K, B Diamant, L Holm, G Hølmer, SA Mortensen e S Stender: *Coenzyme Q10 in health and disease.* European Journal of Clinical Nutrition, 53(10):764–770, 1999. Citado na página 17.

[63] Weber, Peter: *Vitamin K and bone health.* Nutrition, 17(10):880–887, 2001. Citado na página 17.

[64] Araújo, Evani L, João Rui B Alencar e Pedro J Rolim Neto: *Lapachol: segurança e eficácia na terapêutica.* Revista Brasileira de Farmacognosia, 12:57–59, 2002. Citado na página 18.

[65] Müller, Klaus, Andreas Sellmer e Wolfgang Wiegrebe: *Potential antipsoriatic agents: lapacho compounds as potent inhibitors of HaCaT cell growth.* Journal of Natural Products, 62(8):1134–1136, 1999. Citado na página 18.

[66] Eyong, Kenneth Oben, Gabriel Ngosong Folefoc, Victor Kuete, Veronique Penlap Beng, Karsten Krohn, Hidayat Hussain, Augustin Ephram Nkengfack, Michael Saeftel, Salem Ramadan Sarite e Achim Hoerauf: *Newbouldiaquinone A: A naphthoquinone–anthraquinone ether coupled pigment, as a potential antimicrobial and antimalarial agent from Newbouldia laevis.* Phytochemistry, 67(6):605–609, 2006. Citado na página 18.

[67] Fernandes, MC, EN Da Silva, AV Pinto, SL De Castro e RFS Menna-Barreto: *A novel triazolic naphthofuranquinone induces autophagy in reservosomes and impairment of mitosis in Trypanosoma cruzi*. Parasitology, 139(1):26–36, 2012. Citado na página 18.

[68] Mitchell, Tom M e Machine Learning: *Mcgraw-hill science*. Engineering/Math, 1:27, 1997. Citado na página 18.

[69] James, Gareth, Daniela Witten, Trevor Hastie e Robert Tibshirani: *An introduction to statistical learning*, volume 112. Springer, 2013. Citado 2 vezes nas páginas 18 e 19.

[70] Geron, Aurelien: *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. OReilly Media, Inc., 2022. Citado na página 19.

[71] James, Gareth, Daniela Witten, Trevor Hastie e Robert Tibshirani: *An introduction to statistical learning*, volume 112. Springer, 2013. Citado na página 19.

[72] Hastie, Trevor, Robert Tibshirani e Jerome Friedman: *The elements of statistical learning*. Springer Series in Statistics. Springer, 1ª edição, julho 2003. Citado na página 19.

[73] Jolliffe, Ian T: *Principal component analysis for special types of data*. Springer, 2002. Citado 2 vezes nas páginas 19 e 21.

[74] Berrar, Daniel P, Werner Dubitzky, Martin Granzow *et al.*: *A practical approach to microarray data analysis*. Springer, 2003. Citado na página 19.

[75] Wen, Shaokai: *Application of principal component analysis to decision support system*. Tese de Doutoramento, Oklahoma State University, 1997. Citado na página 21.

[76] Alani, Ahmed Sami Abdulghafour: *Principal component analysis in statistics*. Tese de Doutoramento, Eastern Mediterranean University (EMU)-Doğu Akdeniz Üniversitesi (DAÜ), 2014. Citado na página 21.

[77] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot e E. Duchesnay: *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12:2825–2830, 2011. Citado 2 vezes nas páginas 22 e 62.

[78] Bruce, Peter e Andrew Bruce: *Practical statistics for data scientists*. O'Reilly Media, junho 2017. Citado na página 22.

[79] Bock, Hans Hermann: *Clustering Methods: A History of k-Means Algorithms.* Em *Selected Contributions in Data Analysis and Classification*, páginas 161–172. Springer Berlin Heidelberg, 2007. https://doi.org/10.1007/978-3-540-73560-1_15. Citado na página 22.

[80] *Eufrânio Da Silva Júnior Research Group.* https://www.eufraniolab.com. Accessed: 209-03-15. Citado na página 25.

[81] Polavarapu, Prasad L: *Ab initio vibrational Raman and Raman optical activity spectra.* Journal of Physical Chemistry, 94(21):8106–8112, 1990. Citado na página 27.

[82] Michalska, Danuta e Rafał Wysokiński: *The prediction of Raman spectra of platinum(II) anticancer drugs by density functional theory.* Chemical Physics Letters, 403(1):211–217, 2005, ISSN 0009-2614. https://www.sciencedirect.com/science/article/pii/S0009261404020664. Citado na página 27.

[83] Wysokiński, Rafał, Katarzyna Hernik, Roman Szostak e Danuta Michalska: *Electronic structure and vibrational spectra of cis-diammine(orotato)platinum(II), a potential cisplatin analogue: DFT and experimental study.* Chemical Physics, 333(1):37–48, 2007, ISSN 0301-0104. https://www.sciencedirect.com/science/article/pii/S0301010407000055. Citado na página 27.

[84] Santos, Hélio F dos, Wagner B de Almeida, Amélia MG Do Val e Afonso C Guimarães: *Espectro infravermelho e análise conformacional do composto 3-fenil-2-oxo-1, 2, 3-oxatiazolidina.* Química Nova, 22(5):732–736, 1999. Citado na página 27.

[85] Ünal, Yener, Wassim Nassif, Burak Can Özaydin e Koray Sayin: *Scale factor database for the vibration frequencies calculated in M06-2X, one of the DFT methods.* Vibrational Spectroscopy, 112:103189, 2021. Citado na página 38.

[86] Tandon, Vishnu K., Rakeshwar B. Chhor, Ravindra V. Singh, Sanjay Rai e Dharmendra B. Yadav: *Design, synthesis and evaluation of novel 1,4-naphthoquinone derivatives as antifungal and anticancer agents.* Bioorganic & Medicinal Chemistry Letters, 14(5):1079–1083, 2004, ISSN 0960-894X. https://www.sciencedirect.com/science/article/pii/S0960894X04000277. Citado na página 47.

[87] Ilina, TV, EA Semenova, TR Pronyaeva, AG Pokrovskii, IV Nechepurenko, EE Shults, OI Andreeva, SN Kochetkov e GA Tolstikov: *Inhibition of HIV-1 reverse transcriptase by aryl-substituted naphto-and anthraquinones.* 382(1):56–59, 2002. Citado na página 47.

[88] Costa, Márcia Dias Diniz *et al.*: *Desenvolvimento de instrumentação científica em nanotecnologia: inferências para nanoespectroscopia Raman a partir de Technology*

*Roadmapping retrospectivo adaptado de duas tecnologias consolidadas.* 2021. Citado na página 47.

[89] Taghizadeh, Alireza, Ulrik Leffers, Thomas G Pedersen e Kristian S Thygesen: *A library of ab initio Raman spectra for automated identification of 2D materials.* Nature Communications, 11(1):1–10, 2020. Citado na página 47.

[90] Lin, Yangming, Zigeng Liu, Yiming Niu, Bingsen Zhang, Qing Lu, Shuchang Wu, Gabriele Centi, Siglinda Perathoner, Saskia Heumann, Linhui Yu e Dang Sheng Su: *Highly Efficient Metal-Free Nitrogen-Doped Nanocarbons with Unexpected Active Sites for Aerobic Catalytic Reactions.* ACS Nano, 13(12):13995–14004, 2019. https://doi.org/10.1021/acsnano.9b05856, PMID: 31765120. Citado na página 65.

[91] Jardim, Guilherme A.M., Thaissa L. Silva, Marilia O.F. Goulart, Carlos A. de Simone, Juliana M.C. Barbosa, Kelly Salomão, Solange L. de Castro, John F. Bower e Eufrânio N. da Silva Júnior: *Rhodium-catalyzed C-H bond activation for the synthesis of quinonoid compounds: Significant Anti-Trypanosoma cruzi activities and electrochemical studies of functionalized quinones.* European Journal of Medicinal Chemistry, 136:406–419, 2017, ISSN 0223-5234. https://www.sciencedirect.com/science/article/pii/S0223523417303720. Citado na página 65.

[92] de Carvalho, Renato L., Guilherme A. M. Jardim, Augusto C. C. Santos, Maria H. Araujo, Willian X. C. Oliveira, Ana Cristina S. Bombaça, Rubem F. S. Menna-Barreto, Elumalai Gopi, Edmond Gravel, Eric Doris e Eufrânio N. da Silva Júnior: *Combination of Aryl Diselenides/Hydrogen Peroxide and Carbon-Nanotube/Rhodium Nanohybrids for Naphthol Oxidation: An Efficient Route towards Trypanocidal Quinones.* Chemistry – A European Journal, 24(57):15227–15235, 2018. https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/chem.201802773. Citado na página 65.

[93] Jardim, Guilherme A. M., Eufrânio N. da Silva Júnior e John F. Bower: *Overcoming naphthoquinone deactivation: rhodium-catalyzed C-5 selective C–H iodination as a gateway to functionalized derivatives.* Chem. Sci., 7:3780–3784, 2016. http://dx.doi.org/10.1039/C6SC00302H. Citado na página 65.

[94] Silva, Raphael S.F., Elaine M. Costa, Úrsula L.T. Trindade, Daniel V. Teixeira, Maria de Carmo F.R. Pinto, Gustavo L. Santos, Valeria R.S. Malta, Carlos Alberto De Simone, Antonio Ventura Pinto e Solange L. de Castro: *Synthesis of naphthofuranquinones with activity against Trypanosoma cruzi.* European Journal of Medicinal Chemistry, 41(4):526–530, 2006, ISSN 0223-5234. https://www.sciencedirect.com/science/article/pii/S0223523406000304. Citado na página 65.

[95] Jardim, Guilherme A. M., Daisy J. B. Lima, Wagner O. Valença, Daisy J. B. Lima, Bruno C. Cavalcanti, Claudia Pessoa, Jamal Rafique, Antonio L. Braga, Claus Jacob, Eufrânio N. Da Silva Júnior e Eduardo H. G. Da Cruz: *Synthesis of Selenium-Quinone Hybrid Compounds with Potential Antitumor Activity via Rh-Catalyzed C-H Bond Activation and Click Reactions.* Molecules, 23(1), 2018, ISSN 1420-3049. https://www.mdpi.com/1420-3049/23/1/83. Citado 2 vezes nas páginas 66 e 67.

[96] Salas, Cristian, Ricardo A. Tapia, Karina Ciudad, Verónica Armstrong, Myriam Orellana, Ulrike Kemmerling, Jorge Ferreira, Juan Diego Maya e Antonio Morello: *Trypanosoma cruzi: Activities of lapachol and α- and β-lapachone derivatives against epimastigote and trypomastigote forms.* Bioorganic & Medicinal Chemistry, 16(2):668–674, 2008, ISSN 0968-0896. https://www.sciencedirect.com/science/article/pii/S0968089607009066. Citado na página 66.

[97] Silva Júnior, Eufrânio N. da, Clara F. de Deus, Bruno C. Cavalcanti, Cláudia Pessoa, Letícia V. Costa-Lotufo, Raquel C. Montenegro, Manoel O. de Moraes, Maria do Carmo F. R. Pinto, Carlos A. de Simone, Vitor F. Ferreira, Marilia O. F. Goulart, Carlos Kleber Z. Andrade e Antônio V. Pinto: *3-Arylamino and 3-Alkoxy-nor-β-lapachone Derivatives: Synthesis and Cytotoxicity against Cancer Cell Lines.* Journal of Medicinal Chemistry, 53(1):504–508, 2010. https://doi.org/10.1021/jm900865m, PMID: 19947600. Citado na página 66.

[98] da Silva, Eufrânio N., Rubem F.S. Menna-Barreto, Maria do Carmo F.R. Pinto, Raphael S.F. Silva, Daniel V. Teixeira, Maria Cecília B.V. de Souza, Carlos Alberto De Simone, Solange L. De Castro, Vitor F. Ferreira e Antônio V. Pinto: *Naphthoquinoidal [1,2,3]-triazole, a new structural moiety active against Trypanosoma cruzi.* European Journal of Medicinal Chemistry, 43(8):1774–1780, 2008, ISSN 0223-5234. https://www.sciencedirect.com/science/article/pii/S0223523407003984. Citado na página 66.

[99] Miyaki, Komei e Nisaburo Ikeda: *antibacterial Properties of 2 and 2,3-Substituted 1, 4-Naphthoquinones, - Dithiophenyl-1, 4-n2phthoquinone and of Mercapto-1,4-naphthoquinone III. Dimorphism of 2, 3 Infrared Spectra Derivertives.* YAKUGAKU ZASSHI, 73(9):964–968, 1953. Citado na página 66.

[100] Jardim, Guilherme A. M., Wallace J. Reis, Matheus F. Ribeiro, Flaviano M. Ottoni, Ricardo J. Alves, Thaissa L. Silva, Marilia O. F. Goulart, Antonio L. Braga, Rubem F. S. Menna-Barreto, Kelly Salomão, Solange L. de Castro e Eufrânio N. da Silva Júnior: *On the investigation of hybrid quinones: synthesis, electrochemical studies and evaluation of trypanocidal activity.* RSC Adv., 5:78047–78060, 2015. http://dx.doi.org/10.1039/C5RA16213K. Citado 2 vezes nas páginas 67 e 68.

[101] Bahia, Samara Ben B. B., Wallace J. Reis, Guilherme A. M. Jardim, Francielly T. Souto, Carlos A. de Simone, Claudia C. Gatto, Rubem F. S. Menna-Barreto, Solange L. de Castro, Bruno C. Cavalcanti, Claudia Pessoa, Maria H. Araujo e Eufrânio N. da Silva Júnior: *Molecular hybridization as a powerful tool towards multitarget quinoidal systems: synthesis, trypanocidal and antitumor activities of naphthoquinone-based 5-iodo-1,4-disubstituted-, 1,4- and 1,5-disubstituted-1,2,3-triazoles.* Med. Chem. Commun., 7:1555–1563, 2016. http://dx.doi.org/10.1039/C6MD00216A. Citado 2 vezes nas páginas 67 e 68.

[102] da Cruz, Eduardo H.G., Molly A. Silvers, Guilherme A.M. Jardim, Jarbas M. Resende, Bruno C. Cavalcanti, Igor S. Bomfim, Claudia Pessoa, Carlos A. de Simone, Giancarlo V. Botteselle, Antonio L. Braga, Divya K. Nair, Irishi N.N. Namboothiri, David A. Boothman e Eufrânio N. da Silva Júnior: *Synthesis and antitumor activity of selenium-containing quinone-based triazoles possessing two redox centres, and their mechanistic insights.* European Journal of Medicinal Chemistry, 122:1–16, 2016, ISSN 0223-5234. https://www.sciencedirect.com/science/article/pii/S0223523416304974. Citado 2 vezes nas páginas 67 e 68.

[103] da Silva Júnior, Eufrânio N., Maria Cecília B.V. de Souza, Antônio V. Pinto, Maria do Carmo F.R. Pinto, Marilia O.F. Goulart, Francisco W.A. Barros, Claudia Pessoa, Letícia V. Costa-Lotufo, Raquel C. Montenegro, Manoel O. de Moraes e Vitor F. Ferreira: *Synthesis and potent antitumor activity of new arylamino derivatives of nor-β-lapachone and nor-α-lapachone.* Bioorganic & Medicinal Chemistry, 15(22):7035–7041, 2007, ISSN 0968-0896. https://www.sciencedirect.com/science/article/pii/S0968089607006608. Citado na página 67.

[104] Baiju, Thekke V., Renata G. Almeida, Sudheesh T. Sivanandan, Carlos A. de Simone, Lucas M. Brito, Bruno C. Cavalcanti, Claudia Pessoa, Irishi N.N. Namboothiri e Eufrânio N. da Silva Júnior: *Quinonoid compounds via reactions of lawsone and 2-aminonaphthoquinone with α-bromonitroalkenes and nitroallylic acetates: Structural diversity by C-ring modification and cytotoxic evaluation against cancer cells.* European Journal of Medicinal Chemistry, 151:686–704, 2018, ISSN 0223-5234. https://www.sciencedirect.com/science/article/pii/S0223523418303222. Citado na página 68.

[105] Valença, Wagner O., Thekke V. Baiju, Fernanda G. Brito, Maria H. Araujo, Claudia Pessoa, Bruno C. Cavalcanti, Carlos A. de Simone, Claus Jacob, Irishi N. N. Namboothiri e Eufrânio N. da Silva Júnior: *Synthesis of Quinone-Based N-Sulfonyl-1,2,3-triazoles: Chemical Reactivity of Rh(II) Azavinyl Carbenes and Antitumor Activity.* ChemistrySelect, 2(16):4301–4308, 2017. https://chemistry-europe.

onlinelibrary.wiley.com/doi/abs/10.1002/slct.201700885. Citado 2 vezes nas páginas 68 e 70.

[106] Almeida, Renata G., Wagner O. Valença, Luísa G. Rosa, Carlos A. de Simone, Solange L. de Castro, Juliana M. C. Barbosa, Daniel P. Pinheiro, Carlos R. K. Paier, Guilherme G. C. de Carvalho, Claudia Pessoa, Marilia O. F. Goulart, Ammar Kharma e Eufrânio N. da Silva Júnior: *Synthesis of quinone imine and sulphur-containing compounds with antitumor and trypanocidal activities: redox and biological implications*. RSC Med. Chem., 11:1145–1160, 2020. http://dx.doi.org/10.1039/D0MD00072H. Citado na página 70.

[107] Gontijo, Talita B., Rossimiriam P. de Freitas, Flavio S. Emery, Leandro F. Pedrosa, José B. Vieira Neto, Bruno C. Cavalcanti, Claudia Pessoa, Aaron King, Fabio de Moliner, Marc Vendrell e Eufrânio N. da Silva Júnior: *On the synthesis of quinone-based BODIPY hybrids: New insights on antitumor activity and mechanism of action in cancer cells*. Bioorganic & Medicinal Chemistry Letters, 27(18):4446–4456, 2017, ISSN 0960-894X. https://www.sciencedirect.com/science/article/pii/S0960894X17307941. Citado na página 71.

# APPENDIX A – PCA and K-means clustering Application

In this work, we apply de PCA method in our data set using the Python Library *Scikit-Learn* [77] through the following importing command:

```
from sklearn.decomposition import PCA,
```

and by naming an instance *PCA()* from the PCA library, fitting and transforming your target data:

```
pca = PCA()
pca.fit(target_data)
pca.transform(target_data).
```

The scikit-learn PCA uses Singular Value Decomposition to reduce de dimensionality of the target data set. In simple words, it is to say that it transforms the data as the following equation:

$$X = USV^T \tag{A.1}$$

where $\mathbf{U}$ is an $m \times n$ matrix, $\mathbf{S}$ is an $n \times n$ diagonal matrix, and $\mathbf{V^T}$ is also an $n \times n$ matrix. The Matrix $\mathbf{SV^T}$ is usually called the *loadings* matrix, and the matrix $\mathbf{U}$ is called the *scores* matrix of the decomposed data set. One can access the score and the loadings matrices, respectively, by the commands:

```
loadings_matrix = pca.transform(target_data)
scores_matrix = pca.components_
```

where you can see the dispersion of your data points in a 3-dimensional space by plotting the 3 first columns of the scores matrix.

The k-means clustering method is also part of the Scikit-Learn library and can be imported through the following command:

```
from sklearn.cluster import KMeans,
```

as in mentioned in the PCA algorithm, we need also to name an instance, where we need to set the number of clusters que algorithm have to search for:

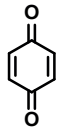
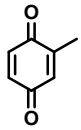
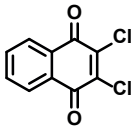
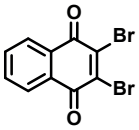
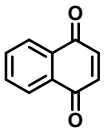
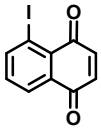
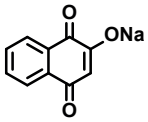
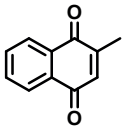
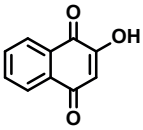
```
kmeans = KMeans(n_clusters=K)
kmeans.fit(target_data).
```

in our work, it was sufficient to only "$.fit$" the target data in order to produce a numpy array with the labels for our clustered data. One can call the labels array by the command:

```
target_data_labels = kmeans.labels_
```

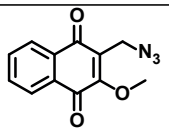
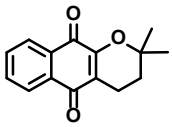
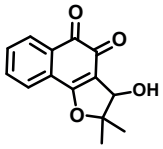
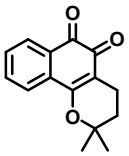
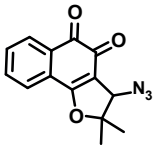
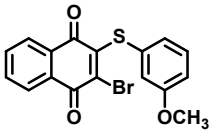
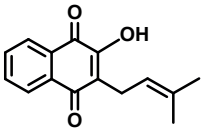
# APPENDIX B – Table of the studied molecules

**Table 1** – Sample numbers, names, chemical structures, chemical formulas and reference methods of how the compounds were obtained.

| Compound | Name | Chemical Formula | Obtained through |
|---|---|---|---|
| **1** |  1,4-Benzoquinone | $C_6H_4O_2$ | Commercially |
| **2** |  2-methylcyclohexa-2,5-diene-1,4-dione | $C_7H_6O_2$ | Ref. 90 |
| **3** |  2,3-dichloronaphthalene-1,4-dione | $C_10H_4Cl_2O_2$ | Ref. 91 |
| **4** |  2,3-dibromonaphthalene-1,4-dione | $C_{10}H_4Br_2O_2$ | Ref. 91 |
| **5** |  naphthalene-1,4-dione | $C_{11}H_8O_2$ | Ref. 92 |
| **6** |  5-iodonaphthalene-1,4-dione | $C_{10}H_5IO_2$ | Ref. 93 |
| **7** |  sodium 1,4-dioxo-1,4-dihydronaphthalen-2-olate | $C_{10}H_5NaO_3$ | Ref. 94 |
| **8** |  2-methylnaphthalene-1,4-dione | $C_{11}H_8O_2$ | Commercially |
| **9** |  2-hydroxynaphthalene-1,4-dione | $C_{10}H_6O_3$ | Commercially |

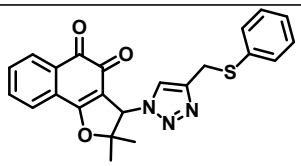
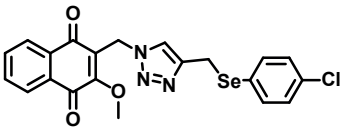
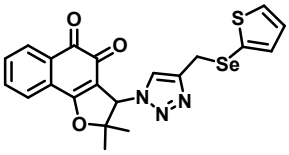
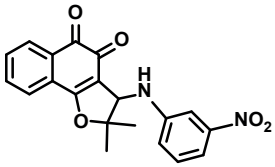
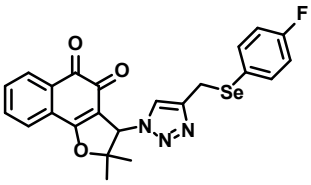
Continued on next page

**Table 1 – continued from previous page**

| Compound | Name | Chemical Formula | Obtained through |
|---|---|---|---|
| **10** |  2-(azidomethyl)-3-methoxynaphthalene-1,4-dione | $C_{12}H_9N_3O_3$ | Ref. 95 |
| **11** |  2,2-dimethyl-3,4-dihydro-2*H*-benzo[*g*]chromene-5,10-dione | $C_{15}H_{14}O_3$ | Ref. 96 |
| **12** |  3-hydroxy-2,2-dimethyl-2,3-dihydronaphtho[1,2-*b*]furan-4,5-dione | $C_{14}H_{12}O_4$ | Ref. 97 |
| **13** |  2,2-dimethyl-3,4-dihydro-2*H*-benzo[*h*]chromene-5,6-dione | $C_{15}H_{14}O_3$ | Ref. 96 |
| **14** |  3-azido-2,2-dimethyl-2,3-dihydronaphtho[1,2-*b*]furan-4,5-dione | $C_{14}H_{11}N_3O_3$ | Ref. 98 |
| **15** |  2-bromo-3-((3-methoxyphenyl)thio)-naphthalene-1,4-dione | $C_{17}H_{11}BrO_3S$ | Ref. 99 |
| **16** |  2-hydroxy-3-(3-methylbut-2-en-1-yl)naphthalene-1,4-dione | $C_{15}H_{14}O_3$ | Commercially |

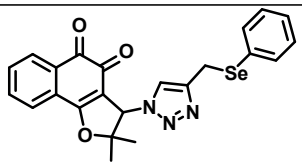
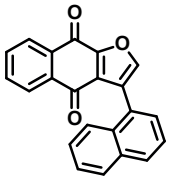
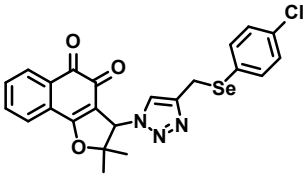
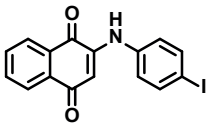
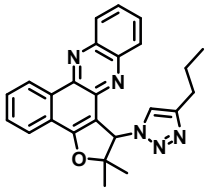
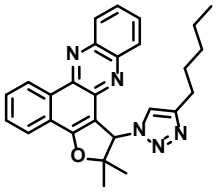
Continued on next page

**Table 1 – continued from previous page**

| Compound | Name | Chemical Formula | Obtained through |
|---|---|---|---|
| **17** |  2,2-dimethyl-3-(4-((phenylthio)methyl)-1*H*-1,2,3-triazol-1-yl)-2,3-dihydronaphtho[1,2-*b*]furan-4,5-dione | $C_{23}H_{19}N_3O_3S$ | Refs. 100–102 |
| **18** |  2-((4-(((4-chlorophenyl)selanyl)methyl)-1*H*-1,2,3-triazol-1-yl)methyl)-3-methoxynaphthalene-1,4-dione | $C_{21}H_{16}ClN_3O_3Se$ | Ref. 95 |
| **19** |  2,2-dimethyl-3-(4-((thiophen-2-ylselanyl)methyl)-1*H*-1,2,3-triazol-1- yl)-2,3-dihydronaphtho[1,2-*b*]furan-4,5-dione | $C_{21}H_{17}N_3O_3SSe$ | Refs. 100–102 |
| **20** |  2,2-dimethyl-3-((3-nitrophenyl)amino)-2,3-dihydronaphtho[1,2-*b*]furan-4,5-dione | $C_{20}H_{16}N_2O_5$ | Ref. 103 |
| **21** |  3-(4-(((4-fluorophenyl)selanyl)methyl)-1*H*-1,2,3-triazol-1-yl)-2,2-dimethyl-2,3-dihydronaphtho[1,2-*b*]furan-4,5-dione | $C_{23}H_{18}FN_3O_3Se$ | Refs. 100–102 |

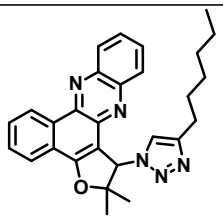
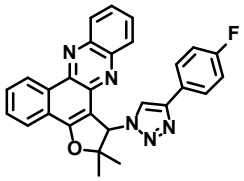
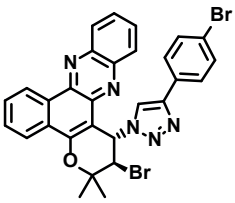
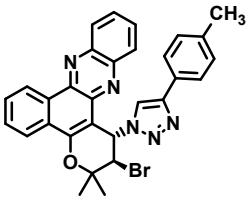
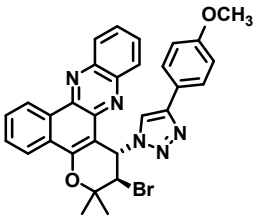
Continued on next page

**Table 1 – continued from previous page**

| Compound | Name | Chemical Formula | Obtained through |
|---|---|---|---|
| **22** |  2,2-dimethyl-3-(4-((phenylselanyl)methyl)-1*H*-1,2,3-triazol-1-yl)-2,3-dihydronaphtho[1,2-*b*]furan-4,5-dione | $C_{23}H_{19}N_3O_3Se$ | Refs. 100–102 |
| **23** |  3-(naphthalen-1-yl)naphtho[2,3-*b*]furan-4,9-dione | $C_{22}H_{12}O_3$ | Ref. 104 |
| **24** |  3-(4-(((4-chlorophenyl)selanyl)methyl)-1*H*-1,2,3-triazol-1-yl)-2,2-dimethyl-2,3-dihydronaphtho[1,2-*b*]furan-4,5-dione | $C_{23}H_{18}ClN_3O_3Se$ | Refs. 100–102 |
| **25** |  2-((4-iodophenyl)amino)naphthalene-1,4-dione | $C_{16}H_{10}INO_2$ | Ref. 105 |
| **26** |  2,2-dimethyl-1-(4-propyl-1*H*-1,2,3-triazol-1-yl)-1,2-dihydrobenzo[*a*]furo[2,3-*c*]phenazine | $C_{25}H_{23}N_5O$ | Ref. 8 |
| **27** |  2,2-dimethyl-1-(4-pentyl-1*H*-1,2,3-triazol-1-yl)-1,2-dihydrobenzo[*a*]furo[2,3-*c*]phenazine | $C_{27}H_{27}N_5O$ | Ref. 8 |

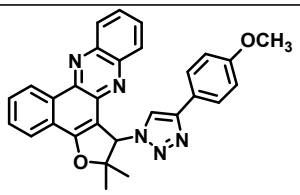
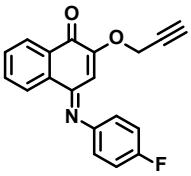
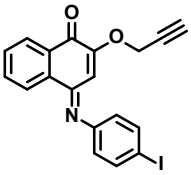
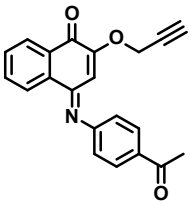
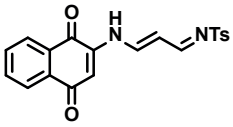
Continued on next page

**Table 1 – continued from previous page**

| Compound | Name | Chemical Formula | Obtained through |
|---|---|---|---|
| **28** |  1-(4-hexyl-1*H*-1,2,3-triazol-1-yl)-2,2-dimethyl-1,2-dihydrobenzo[*a*]furo[2,3-*c*]phenazine | $C_{28}H_{29}N_5O$ | Ref. 8 |
| **29** |  1-(4-(4-fluorophenyl)-1*H*-1,2,3-triazol-1-yl)-2,2-dimethyl-1,2-dihydrobenzo[*a*]furo[2,3-*c*]phenazine | $C_{28}H_{20}FN_5O$ | Ref. 8 |
| **30** |  (1S,2R)-2-bromo-1-(4-(4-bromophenyl)-1*H*-1,2,3-triazol-1-yl)-3,3-dimethyl-2,3-dihydro-1*H*-benzo[*a*]pyrano[2,3-*c*]phenazine | $C_{29}H_{21}Br_2N_5O$ | Ref. 8 |
| **31** |  (1S,2R)-2-bromo-3,3-dimethyl-1-(4-(p-tolyl)-1*H*-1,2,3-triazol-1-yl)-2,3-dihydro-1*H*-benzo[*a*]pyrano[2,3-*c*]phenazine | $C_{30}H_{24}BrN_5O$ | Ref. 8 |
| **32** |  (1S,2R)-2-bromo-1-(4-(4-methoxyphenyl)-1*H*-1,2,3-triazol-1-yl)-3,3-dimethyl-2,3-dihydro-1*H*-benzo[*a*]pyrano[2,3-*c*]phenazine | $C_{30}H_{24}BrN_5O_2$ | Ref. 8 |

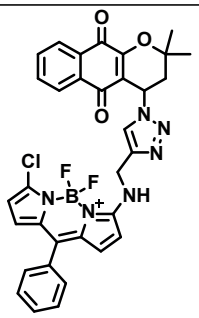
Continued on next page

**Table 1 – continued from previous page**

| Compound | Name | Chemical Formula | Obtained through |
|---|---|---|---|
| **33** |  <br>1-(4-(4-methoxyphenyl)-1*H*-1,2,3-triazol-1-yl)-2,2-dimethyl-1,2-dihydrobenzo[*a*]furo[2,3-*c*]phenazine | $C_29H_23N_5O_2$ | Ref. 8 |
| **34** |  <br>(*E*)-4-((4-fluorophenyl)imino)-2-(prop-2-yn-1-yloxy)naphthalen-1(4*H*)-one | $C_{19}H_{12}FN_5O_2$ | Ref. 106 |
| **35** |  <br>(*E*)-4-((4-iodophenyl)imino)-2-(prop-2-yn-1-yloxy)naphthalen-1(4*H*)-one | $C_{19}H_{12}INO_2$ | Ref. 106 |
| **36** |  <br>(*E*)-4-((4-acetylphenyl)imino)-2-(prop-2-yn-1-yloxy)naphthalen-1(4*H*)-one | $C_{21}H_{15}NO_3$ | Ref. 106 |
| **37** |  <br>*N*-((1*E*,2*E*)-3-((1,4-dioxo-1,4-dihydronaphthalen-2-yl)amino)allylidene)-4-methylbenzenesulfonamide | $C_{20}H_{16}N_2O_4S$ | Ref. 105 |

Continued on next page

**Table 1 – continued from previous page**

| Compound | Name | Chemical Formula | Obtained through |
|---|---|---|---|
| **38** |  <br>4-(4-(((7-chloro-5,5-difluoro-10-phenyl-5*H*-4l$^4$,5l$^4$-dipyrrolo[1,2-*c*:2',1'-*f*][1,3,2]diazaborinin-3-yl)amino)methyl)-1*H*-1,2,3-triazol-1-yl)-2,2-dimethyl-3,4-dihydro-2*H*-benzo[*g*]chromene-5,10-dione | $C_{33}H_{26}BClF_2N_6O_3$ | Ref. 107 |