UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS QUÍMICAS
INSTITUTO DE QUÍMICA

ESTUDIO DE PROPIEDADES ESPECTROSCÓPICAS DE CROMÓFOROS EMPLEANDO MÉTODOS DE APRENDIZAJE AUTOMATIZADO QUE COMBINAN LA APROXIMACIÓN RELACIÓN ESTRUCTURA PROPIEDAD CON DESCRIPTORES TOPOLÓGICOS CUÁNTICOS.

TESIS

QUE PARA OPTAR POR EL GRADO DE

MAESTRO EN CIENCIAS

PRESENTA

Q. BERNARDO ARTURO SALCIDO SANTACRUZ

DR. JORGE PEÓN PERALTA
INSTITUTO DE QUÍMICA

CIUDAD UNIVERSITARIA, CIUDAD DE MÉXICO, JUNIO 2022

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS QUÍMICAS

ESTUDIO DE PROPIEDADES ESPECTROSCÓPICAS DE CROMÓFOROS EMPLEANDO MÉTODOS DE APRENDIZAJE AUTOMATIZADO QUE COMBINAN LA APROXIMACIÓN RELACIÓN ESTRUCTURA PROPIEDAD CON DESCRIPTORES TOPOLÓGICOS CUÁNTICOS

T E S I S
PARA OPTAR POR EL GRADO DE

MAESTRO EN CIENCIAS

P R E S E N T A

Q. BERNARDO ARTURO SALCIDO SANTACRUZ

Ciudad de México, junio, 2022

Jurado asignado

Presidente	Dr. Carlos Amador Bedolla
Vocal	Dr. J. Jesús Hernández Trujillo
Vocal	Dr. Juan Ignacio Rodríguez Hernández
Vocal	Dra. Karina Martínez Mayorga
Secretario	Dr. José Marco Antonio Franco Pérez

Bernardo Arturo Salcido Santacruz

Sustentante

Dr. Jorge Peón Peralta

Tutor

Dedico esta tesis a mi familia:

A mi esposa, Victoria Godínez López por compartir tantos momentos maravillosos junto a mí, por inspirarme a que nada es imposible, por motivarme a crecer personal y académicamente, pero especialmente por compartir conmigo y hacerme partícipe en sus sueños.

A mis padres, Alma y Arturo por apoyarme desde el principio y para siempre, brindándome todo su amor y confianza, por haberme enseñado todo lo necesario para poder forjar mi propio camino

A mis hermanos Jimena, Santiago y Héctor, por todos los momentos, buenos y malos, que compartimos juntos, por mostrarme la alegría de la vida.

Bernardo Arturo Salcido Santacruz

Agradecimientos

A la Universidad Nacional Autónoma de México, al Instituto de Química y al Programa de Maestría y Doctorado en Ciencias Químicas por brindarme las herramientas para mi desarrollo tanto profesional y académico como personal. Así mismo agradezco a la Fundación Alberto Y Dolores Andrade, por creer en mí y apoyarme durante todos mis estudios.

Al Consejo Nacional de Ciencia y Tecnología CONACyT por el proyecto Ciencia de Frontera 2019-51496, así como por el apoyo económico brindado a través la beca otorgada para la realización de mis estudios de maestría (CVU: 972500).

Al PAPIIT/DGAPA/UNAM por el proyecto IG200621

Al Dr. Jorge Peón Peralta por permitirme ser parte de su grupo de investigación, por brindarme las herramientas necesarias para crecer y desarrollarme en el área académica, por sus enseñanzas y especialmente por confiar en cada momento en mí, mis habilidades y capacidades.

A todos mis compañeros del laboratorio Jesús Durán, Leonardo Coello, Emmanuel García, Mariana Mejía, Óscar Guzmán, Andrea Cadena, Melissa Bravo, Francisco Reza y Mario Gutiérrez por brindarme siempre su apoyo, principalmente en los momentos frustrantes, pero especialmente por los buenos momentos que compartimos durante estos últimos 5 años.

Resumen

La demanda de nuevos cromóforos se ha incrementado en los últimos años debido a sus diversas aplicaciones, tales como terapia fotodinámica, microscopía de tiempos de vida y de superresolución. Conocer las propiedades espectroscópicas de estas moléculas previo a realizar su síntesis puede guiar de una mejor manera el proceso para el diseño de estos cromóforos. Usualmente, esto se lleva a cabo empleando las metodologías basadas en la teoría cuántica, sin embargo, estas son poco prácticas para cromóforos con muchos átomos. Debido a esto, se han incrementado los esfuerzos en el desarrollo de nuevas metodologías, las cuales hacen uso de la gran cantidad de información que hay acerca de estas propiedades, combinándola con modelos de aprendizaje automatizado. Es por ello, que el objetivo de este trabajo es encontrar una metodología basada en modelos de aprendizaje automatizado (figura 1) que permita predecir las propiedades espectroscópicas de cromóforos (la energía de absorción, la energía de emisión, el desplazamiento de Stokes y el tiempo de vida de fluorescencia) en diferentes medios de solvatación.

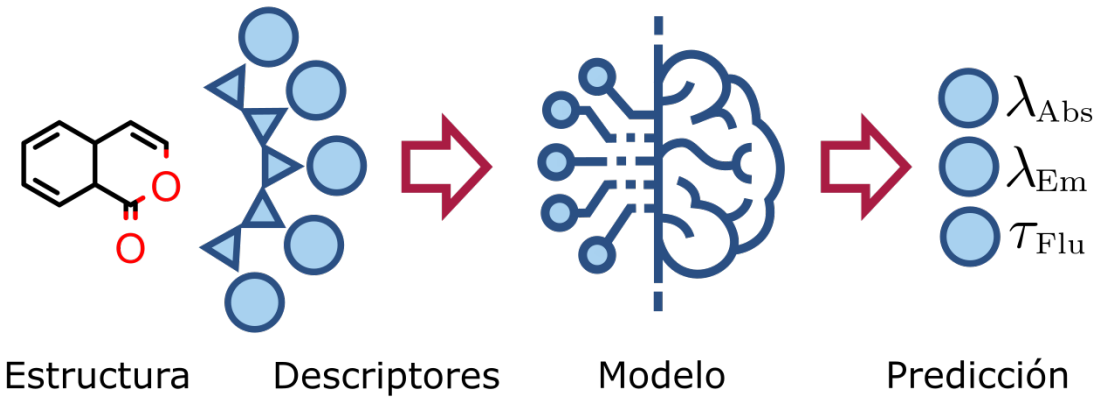


Figura 1. Esquema en donde se muestra el proceso realizado para el estudio de las propiedades espectroscópicas (energía de absorción, energía de emisión, desplazamiento de Stokes y tiempo de vida de fluorescencia). Para ello, se partió de la estructura de cada uno de los cromóforos estudiados, junto con la información espectroscópica para cada uno de ellos, incluyendo además el disolvente en el cual se realizaron las correspondientes mediciones. En una segunda etapa se calcularon una serie de propiedades (asociadas a los cromóforos y los disolventes) también llamados descriptores. Estos se categorizaron de acuerdo con su naturaleza física en descriptores estructurales, cuánticos y empíricos. Empleando los descriptores calculados en el paso anterior junto con los valores recabados de cada una de las propiedades espectroscópicas, se entrenó y evaluó a los correspondientes modelos de aprendizaje automatizado, los cuales mostraron ser capaces de predecir la mayoría de las propiedades estudiadas.

Con este fin, el presente trabajo proporciona una serie de modelos de aprendizaje automatizado basados en la relación estructura-propiedad y en la descripción de la topología de la densidad electrónica, los cuales son capaces de predecir la mayoría de las propiedades estudiadas, incluyendo el efecto del disolvente sobre dichas propiedades. Se obtuvieron modelos capaces de predecir la energía de absorción y emisión para un grupo de 500 cromóforos diferentes, con un error absoluto medio del orden de 0.1 eV y con un coeficiente de correlación de 0.95. Por otro lado, también se obtuvieron modelos capaces de predecir la energía de absorción, la energía de emisión y el desplazamiento de Stokes, para un grupo de 30 cumarinas. Para la energía de absorción y de emisión se obtuvo un error absoluto medio del orden de 0.06 eV. Mientras que para el desplazamiento de Stokes se obtuvo un error absoluto medio del orden de 0.04 eV, todos con un coeficiente de correlación de 0.95.

Además, este trabajo proporciona una herramienta práctica y eficiente para implementar métodos de aprendizaje automatizado enfocados en el estudio de propiedades moleculares que dependan tanto de la misma molécula como del disolvente en el cual ésta se encuentre. Esto se logra a través de una librería de Python ”Machine Learning Molecule”, la cual fue desarrollada como resultado de este trabajo. Ésta implementa una serie de funciones, clases y métodos que desempeñan tareas para la construcción de la base de datos, el cálculo de los descriptores estructurales y el entrenamiento de los modelos de aprendizaje automatizado, logrando simplificar en gran medida el trabajo requerido para la implementación de modelos de aprendizaje automatizado que tengan como objetivo la predicción de propiedades moleculares.

Índice

Nomenclatura x

Lista de figuras xv

Lista de tablas xvii

1. Introducción 1

1.1. Una nueva perspectiva para el estudio de las propiedades espectroscópicas 2

1.2. Sistemas de Estudio 5

1.3. Aprendizaje automatizado (Machine Learning) 8

1.3.1. Conceptos básicos 8

1.3.2. Implementación de los métodos 11

1.3.3. Modelos de aprendizaje automatizado basados en árboles de decisión 15

2. Objetivos e hipótesis 18

2.1. Objetivo general 18

2.2. Objetivos particulares 18

2.3. Hipótesis 19

3. Metodología 20

3.1. Construcción de las bases de datos 20

3.2. Descriptores 21

3.2.1. Descriptores estructurales 21

3.2.2. Descriptores cuánticos 23

3.2.3. Descriptores empíricos del disolvente 27

3.3. Aprendizaje automatizado 28

4. Resultados y discusiones 31

4.1. Resultados para el Sistema 1 31

4.1.1. Comparación entre diferentes tipos de cromóforos 31

4.1.2. Comparación entre diferentes tipos de descriptores 33

4.1.3. Resultados para cada una de las propiedades espectroscópicas 35
4.1.4. Importancia de las características 38
4.2. Resultados para el Sistema 2 41
4.2.1. Comparación entre diferentes tipos de descriptores 41
4.2.2. Resultados para cada una de las propiedades espectroscópicas 47
4.2.3. Importancia de las características 50
4.3. Mejores modelos para el Sistema 1 y para el Sistema 2 53

5. Conclusiones **54**
5.1. Conclusiones generales 54
5.2. Conclusiones particulares 55

Apéndice A **62**

Apéndice B **71**

Nomenclatura

Aprendizaje automatizado

$[X_j]$ Característica

$[Y_i]$ Elemento de salida

$(y_i, [X_i])$ Ejemplar, en donde y_i son los elementos de $[Y_i]$ y $[X_i]$ las filas en la matriz $[X_{i\,j}]$

$[X_{i\,j}]$ Vector de características

Conceptos matemáticos

r^2 Coeficiente de correlación

MAE Error absoluto medio

MSE Error cuadrático medio

RMSE Raíz cuadrada del error cuadrático medio $\sqrt{MSE}$

σ_{St} Desviación estándar

k Factor de cobertura

U Incertidumbre expandida

u Incertidumbre tipo A

Propiedades espectroscópicas

ε Coeficiente de extinción molar

λ_{abs} Longitud de onda de absorción

λ_{em} Longitud de onda de emisión

τ_{flu} Tiempo de vida de fluorescencia (Promedio)

Φ Rendimiento cuántico

ΔStokes Desplazamiento de Stokes

Otros Símbolos

c Velocidad de la luz en el vacio 2.99 x 10^{8} $\frac{\mathrm{m}}{\mathrm{s}}$

eV Unidad de energía equivalente a 1.602 x 10^{-19} J

h Constante de Plank equivalente a 4.135 x 10^{-15} eV s

Acrónimos / Abreviaturas

DFT Density functional theory “teoría de funcionales de la densidad”

HOMO Highest occupied molecular orbital “último orbital molecular ocupado”

LUMO Lowest unoccupied molecular orbital “primer orbital molecular desocupado”

MCSCF Multiconfiguration Self-Consistent Field “Métodos multiconfigracionales de campo auto consistente”

ND Dato no disponible

SMILES Simplified Molecular Input Line Entry System

TD-DFT Time dependent density functional theory “teoría de funcionales de la densidad dependiente del tiempo”

Lista de figuras

1. Esquema en donde se muestra el proceso realizado para el estudio de las propiedades espectroscópicas (energía de absorción, energía de emisión, desplazamiento de Stokes y tiempo de vida de fluorescencia). Para ello, se partió de la estructura de cada uno de los cromóforos estudiados, junto con la información espectroscópica para cada uno de ellos, incluyendo además el disolvente en el cual se realizaron las correspondientes mediciones. En una segunda etapa se calcularon una serie de propiedades (asociadas a los cromóforos y los disolventes) también llamados descriptores. Estos se categorizaron de acuerdo con su naturaleza física en descriptores estructurales, cuánticos y empíricos. Empleando los descriptores calculados en el paso anterior junto con los valores recabados de cada una de las propiedades espectroscópicas, se entrenó y evaluó a los correspondientes modelos de aprendizaje automatizado, los cuales mostraron ser capaces de predecir la mayoría de las propiedades estudiadas. VI

1.1. Esquema en donde se muestra el proceso para el cálculo de las propiedades moleculares. Las flechas negras muestran el esquema basándose únicamente en la teoría cuántica. Las flechas verdes y magentas muestran el esquema para el cálculo de las propiedades moleculares desarrollado para este proyecto, en el cual se sustituye la resolución de las ecuaciones de la mecánica cuántica por un algoritmo de aprendizaje automatizado. Aquí la flecha verde corresponde a emplear la aproximación relación estructura-propiedad (QSPR "Quantitative Structure–Property Relationship") mientras que la flecha magenta corresponde a emplear el enfoque de la topología de la densidad electrónica (QCT "Quantum Chemical Topology"). 3

1.2. Esquema en donde se muestra como se implementó el disolvente para este estudio. La flecha verde muestra el proceso para el cálculo de las propiedades del cromóforo, mientras que la flecha magenta muestra el proceso para el cálculo de las propiedades del disolvente. Aquí las flechas corresponden a emplear la aproximación relación estructura-propiedad (QSPR "Quantitative Structure–Property Relationship") y a emplear el enfoque de la topología de la densidad electrónica (QCT "Quantum Chemical Topology"). . . 4

1.3. Proceso de estandarización de una característica $[X_j]$ 11

1.4. Esquema en donde se muestra la distribución de los diferentes subconjuntos de datos (entrenamiento, validación y prueba) con sus respectivos porcentajes aproximados 12
1.5. Proceso por el cual se lleva a cabo la validación cruzada. En un primer paso se divide el conjunto de datos inicial en dos subconjuntos, uno de Entrenamiento-Validación y otro de Prueba). En un segundo paso (el cual se realiza iterativamente) se divide el conjunto de Entrenamiento-Validación en dos subconjuntos, uno de entrenamiento (verde) y otro de validación (rosa). Para cada uno de estos subconjuntos se obtiene el valor de las respectivas métricas (Coeficiente de correlación r^2, Error absoluto medio MAE y cuadrado del error medio MSE). Finalmente, se obtiene un promedio de las métricas obtenidas para cada iteración. 14
1.7. Esquema en donde se muestra como se construye el modelo de "GradientBoostingRegressor", el cual se construye a partir de una serie sucesiva de árboles de decisión, los cuales se crean corrigiendo los errores cometidos por el árbol previo. 17

3.1. Figura que muestra las tres subcategorías diferentes para los descriptores cuánticos: energías orbitales, descriptores electrónicos y descriptores vibracionales. 24
3.2. Figura en donde se muestran, de una manera esquemática, los dos diferentes tipos de descriptores que hacen uso de un análisis de átomos en moléculas 25
3.3. Figura en donde se muestran los grupos de la cumarina que se emplearon para el cálculo de los descriptores QTAIM por grupos. Los grupos 1,2 y 3 hacen referencia a los anillos aromáticos, mientras que los grupos 4, 5 y 6 al grupo carbonilo y a sus respectivos átomos adyacentes. 26
3.4. Esquema en donde se muestra el conjunto de datos después de añadir los descriptores, en donde la columna azul corresponde a la propiedad a estudiar (salida) y las columnas verdes y rosas a los descriptores empleados (entrada) tanto del cromóforo (verde) como del disolvente (rosa) 28
3.5. Figura en donde se muestra el proceso para la selección del valor óptimo para la reducción de dimensión. 29

4.1. Gráficas en donde se muestran los valores obtenidos del coeficiente de correlación r^2 y del error cuadrático medio MAE, para las diferentes pruebas realizadas con diferentes subconjuntos de cromóforos (Sistema 1). 33
4.2. Gráfica en donde se muestran los valores obtenidos del coeficiente de correlación r^2 y el error cuadrático medio MAE para las diferentes pruebas realizadas con diferentes subconjuntos de descriptores (Sistema 1). 34
4.3. Figura en donde se muestra la distribución de datos para las diferentes propiedades estudiadas (Sistema 1). Energía de absorción, energía de emisión, el desplazamiento de Stokes y el tiempo de vida de fluorescencia. 35

4.4. Gráficas de valor predicho contra el valor experimental para la energía de absorción y emisión (Sistema 1), en donde se emplearon todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente. 36
4.5. Gráficas de valor predicho contra el valor experimental para desplazamiento de Stokes y del tiempo de vida de fluorescencia (Sistema 1), en donde se emplearon todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente. 38
4.6. Figura en la que muestra la importancia relativa de los 34 descriptores más importantes para predecir la energía de absorción (Sistema 1). 39
4.7. Figura en la que se muestra la importancia relativa de los 34 descriptores más importantes para predecir la energía de emisión y el tiempo de vida de fluorescencia (Sistema 1). . . 40
4.8. Gráficas en donde se muestran los valores obtenidos del coeficiente de correlación r^2 y del error absoluto medio MAE para los diferentes subconjuntos de descriptores (estructurales, cuánticos y empiricos). 46
4.9. Gráficas en donde se muestran los valores obtenidos del coeficiente de correlación r^2 y del error absoluto medio MAE para los diferentes subconjuntos de descriptores (QTAIM por tipo de átomos y QTAIM por grupos). 46
4.10. Figura en donde se muestra la distribución de datos para las diferentes propiedades estudiadas (Sistema 2). Energía de absorción, energía de emisión y el desplazamiento de Stokes. 47
4.11. Gráficas de valor predicho contra el valor experimental para la energía de absorción y emisión (Sistema 2), en donde se emplearon todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente. 48
4.12. Gráficas de valor predicho contra el valor experimental para desplazamiento de Stokes (Sistema 2), en donde se emplearon todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente. 49
4.13. Figura que muestra la importancia relativa de los 38 descriptores más importantes para predecir la energía de absorción (Sistema 2). 50
4.14. Figura que muestra la importancia relativa de los 38 descriptores más importantes para predecir la energía de emisión y el desplazamiento de Stokes (Sistema 2). 52
4.15. Figura en donde se muestran los resultados de los mejores modelos para todas las propiedades y para ambos sistemas. 53

5.1. Gráficas de valor predicho contra el valor experimental para la energía de absorción y emisión (Sistema 1). En donde se emplearon todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente. 63

5.2. Gráficas de valor predicho contra el valor experimental para desplazamiento de Stokes y del tiempo de vida de fluorescencia (Sistema 1). En donde se emplearon todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente. 64
5.3. Figura en que se muestra la importancia relativa de los 34 descriptores más importantes para predecir la energía de absorción y emisión (Sistema 1). 65
5.4. Figura en que muestra la importancia relativa de los 34 descriptores más importantes para predecir el tiempo de vida de fluorescencia (Sistema 1). 66
5.5. Gráficas de valor predicho contra el valor experimental para la energía de absorción y emisión (Sistema 2). En donde se emplearon todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente. 68
5.6. Gráficas de valor predicho contra el valor experimental para desplazamiento de Stokes (Sistema 2). En donde se emplearon todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente. 68
5.7. Figura que muestra la importancia relativa de los 38 descriptores más importantes para predecir el desplazamiento de Stokes (Sistema 2). 69
5.8. Figura en que muestra la importancia relativa de los 38 descriptores más importantes para predecir la energía de absorción y emisión (Sistema 2). 70

Lista de tablas

1.1. Esquema de la base de datos para el Sistema 1, en donde se tienen registrados los valores de las diferentes propiedades espectroscópicas estudiadas para cada pareja específica del cromóforo junto con su disolvente. Los guiones corresponden a valores registrados de cada propiedad, mientras que "ND" (no disponible) hace referencia a los valores no reportados. 6
1.2. Esquema de la base de datos para el Sistema 2, en donde se tienen registrados los valores de las diferentes propiedades espectroscópicas estudiadas para cada pareja específica de cromóforo junto con su disolvente. Los guiones corresponden a valores registrados de cada propiedad, mientras que "ND" (no disponible) hace referencia a los valores no reportados. 7

4.1. Tabla en donde se muestra una comparación del desempeño de los modelos para los diferentes subconjuntos de cromóforos (Sistema 1). Para el entrenamiento se empleó el modelo de "GradientBoostingRegressor" junto con todos los tipos de descriptores calculados (estructurales, cuánticos y empíricos). 32
4.2. Tabla en donde se muestra una comparación del desempeño de los modelos para los diferentes subconjuntos de descriptores (Sistema 1). 34
4.3. Tabla en donde se muestra una comparación del desempeño del modelo "GradientBoostingRegressor" para predecir la energía de absorción (E_{abs}) empleando diferentes descriptores tanto para el cromóforo como para el disolvente. 42
4.4. Tabla en donde se muestra una comparación del desempeño del modelo "GradientBoostingRegressor" para predecir la energía de emisión (E_{em}) empleando diferentes descriptores tanto para el cromóforo como para el disolvente. 43
4.5. Tabla en donde se muestra una comparación del desempeño del modelo "GradientBoostingRegressor" para predecir el desplazamiento de Stokes empleando diferentes descriptores tanto para el cromóforo como para el disolvente. 44

5.1. Tabla en donde se muestra una comparación del desempeño de los modelos para los diferentes subconjuntos de cromóforos (Sistema 1). Para el entrenamiento se empleó el modelo de "RandomForestRegressor" junto con todos los tipos de descriptores calculados (estructurales, cuánticos y empíricos). 62
5.2. Tabla en donde se muestra una comparación del desempeño de los modelos para los diferentes subconjuntos de descriptores (Sistema 1). 63
5.3. Tabla en donde se muestra una comparación del desempeño del modelo "RandomForestRegressor" para predecir la energía de absorción (E_{abs}) empleando diferentes descriptores tanto para el cromóforo como para el disolvente. 66
5.4. Tabla en donde se muestra una comparación del desempeño del modelo "RandomForestRegressor" para predecir la energía de emisión (E_{em}) empleando diferentes descriptores tanto para el cromóforo como para el disolvente. 67
5.5. Tabla en donde se muestra una comparación del desempeño del modelo "RandomForestRegressor" para predecir el desplazamiento de Stokes empleando diferentes descriptores tanto para el cromóforo como para el disolvente. 67

Introducción

La demanda de nuevos cromóforos con propiedades espectroscópicas específicas ha aumentado debido a la necesidad de mejores formas de caracterizar procesos físicos y químicos, a través de señales ópticas. Esto incluye el desarrollo de nuevos pares de moléculas capaces de transferirse energía unas a otras (pares donador-aceptor), los cuales son ampliamente usados en el estudio de procesos biológicos [1] tales como el plegamiento y los cambios conformacionales en proteínas [2]. Además, estos esquemas permiten el desarrollo de nuevos sistemas moleculares antena-efector [3], los cuales son usados en fotofarmacología [4], [5] y [6] y en microscopía de superresolución [7] y [8]. A su vez, estudios han mostrado qué nuevos tipos de cromóforos son útiles en tareas como la sensibilización en tecnologías de celdas solares [9] y en esquemas fotocatalíticos [6]. Para todos estos estudios, se requiere de nuevas moléculas con propiedades fisicoquímicas y espectroscópicas específicas, tales como la longitud de onda de absorción, de emisión, y el tiempo de vida del estado fluorescente.

Al mismo tiempo, las técnicas de aprendizaje automatizado son cada vez más utilizadas en química por su gran potencial para resolver problemas complejos. Estos estudios tienen como objetivo principal predecir propiedades moleculares empleando técnicas basadas en la relación estructura-propiedad, las cuales se basan en la hipótesis de que la estructura de una molécula es lo que la define y, por tanto, a sus propiedades. Empleando estas técnicas es posible conocer las propiedades de cada molécula conociendo únicamente su estructura. Estudios previos que han hecho uso de estas metodologías lograron, con éxito, construir modelos capaces de predecir diferentes propiedades fisicoquímicas entre las cuales destacan la energía libre de solvatación [10], las constantes de acidez [11] y de estabilidad, así como estados de espín [12] [13] y [14]. Sin embargo, sólo hay unos pocos ejemplos que utilizan ese tipo de metodologías para predecir propiedades espectroscópicas asociadas a las transiciones electrónicas (longitud de onda de absorción y emisión) [15], [16], [17], [18] y [19].

Por otro lado, el enfoque de la topología de la densidad electrónica ha mostrado ser útil para la descripción de los estados excitados. Estos métodos permiten predecir y correlacionar propiedades estáticas y dinámicas, tanto para el estado base como para estados excitados. Por ejemplo, se ha mostrado que ciertas tendencias en la reactividad química pueden ser explicadas a partir de estudios en la topología de la densidad [20]. Además, este enfoque ha podido explicar una gran cantidad de fenómenos de interés químico tales como la evolución en las superficies de energía potencial de los estados electrónicos [21], el carácter de transferencia de carga en los estados excitados [22] y las interacciones específicas como los enlaces de hidrógeno en el estado base y el emisivo [23] y [24].

También, estudios que emplean el enfoque de la topología de la densidad electrónica han tenido éxito al predecir constantes de disociación [25], constantes de Hamett [26] y reemplazos bioisostéricos [27]. Lo anterior muestra el potencial de este enfoque tanto para describir como para predecir fenómenos físicos y químicos. Por último, el estudio de la topología de la densidad electrónica promete ser útil para la descripción de la distribución de la densidad de electrones tras la excitación electrónica, la cual es crucial para determinar la interacción de la molécula excitada con su entorno de solvatación, definiendo así la posición exacta de las bandas de emisión en los cromóforos en función de su entorno local [28] y [29].

1.1. Una nueva perspectiva para el estudio de las propiedades espectroscópicas

Como se vio en la sección anterior, el desarrollo de nuevos cromóforos es un campo de oportunidad muy grande y con aplicaciones en diferentes áreas. Sin embargo, algunas de las limitaciones de este tipo de proyectos son el tiempo y el esfuerzo necesarios para diseñar y sintetizar nuevas moléculas. De modo que es deseable contar con metodologías fiables y prácticas para predecir las propiedades espectroscópicas antes de embarcarse en costosos esfuerzos de síntesis.

En química computacional existen dos metodologías ampliamente usadas para el cálculo de propiedades moleculares, en particular propiedades espectroscópicas. El primer tipo de metodología emplea a la función de onda de diferentes configuraciones electrónicas. Esto se realiza para tener una correcta descripción de los estados electrónicos de la molécula, la cual incluye, de manera acertada, el efecto de la correlación electrónica [30]. A este tipo de metodologías se les conoce como métodos multiconfiguracionales o “Multiconfiguration Self-Consistent Field” (MCSCF) en inglés. Sin embargo, este tipo de métodos son computacionalmente muy costosos, haciéndolos poco prácticos o incluso inviables para moléculas con muchos electrones. Este hecho provoca que, en ocasiones, realizar un estudio de este nivel puede llegar a ser igual o más laborioso que el mismo proceso de síntesis.

Por otro lado, el remplazar a la función de onda por la densidad electrónica e implementar funcionales de ésta (teoría de funcionales de la densidad o “Density Functional Theory” (DFT) en inglés), ha mostrado ser una alternativa computacionalmente menos costosa para el cálculo de propiedades moleculares [31], [32], [33] y [34]. Sin embargo, el análogo dependiente del tiempo de esta teoría (teoría de funcionales de la densidad dependiente del tiempo o “Time-dependent density functional theory” en inglés) es bastante menos precisa que su contraparte independiente del tiempo, la cual presenta errores típicos de 0.2 eV (esto ya realizando un proceso de calibración posterior al cálculo) sobre la energía de la transición electrónica del estado base al primer estado excitado singlete [35], [36], [37] y [38].

Lo ya mencionado nos dejan pocas alternativas. Una de ellas es emplear métodos confiables, pero poco prácticos para una gran cantidad de moléculas o moléculas con muchos electrones (MCSCF). Otra opción es emplear métodos relativamente asequibles para moléculas no muy grandes (la mayoría de los cromóforos, excluyendo a las cianinas) pero que llegan a ser poco confiables, como lo es la teoría de funcionales de la densidad dependiente del tiempo (TD-DFT) o los métodos semiempíricos [39] y [40], los cuales son una alternativa computacionalmente menos costosa para el cálculo de propiedades moleculares. A la problemática anterior se le puede añadir qué propiedades espectroscópicas como el tiempo de vida de fluorescencia no se pueden predecir mediante estos métodos, ya que ésta no sólo depende de las transiciones electrónicas, sino que esta propiedad se describe a través de una competencia cinética entres diferentes estados electrónicos y vibracionales de la molécula [41].

El formalismo de la mecánica cuántica nos proporciona una manera de calcular el valor de cualquier propiedad molecular (observable físico) a través de conocer la función de onda electrónica. Para llevar a cabo esto, primero se resuelve la ecuación de Schrödinger bajo la aproximación de Born-Oppenheimer, la cual nos permite obtener la función de onda electrónica. Posteriormente se resuelve la integral de valor esperado empleando el correspondiente operador hermítico asociado a la propiedad deseada (figura 1.1).

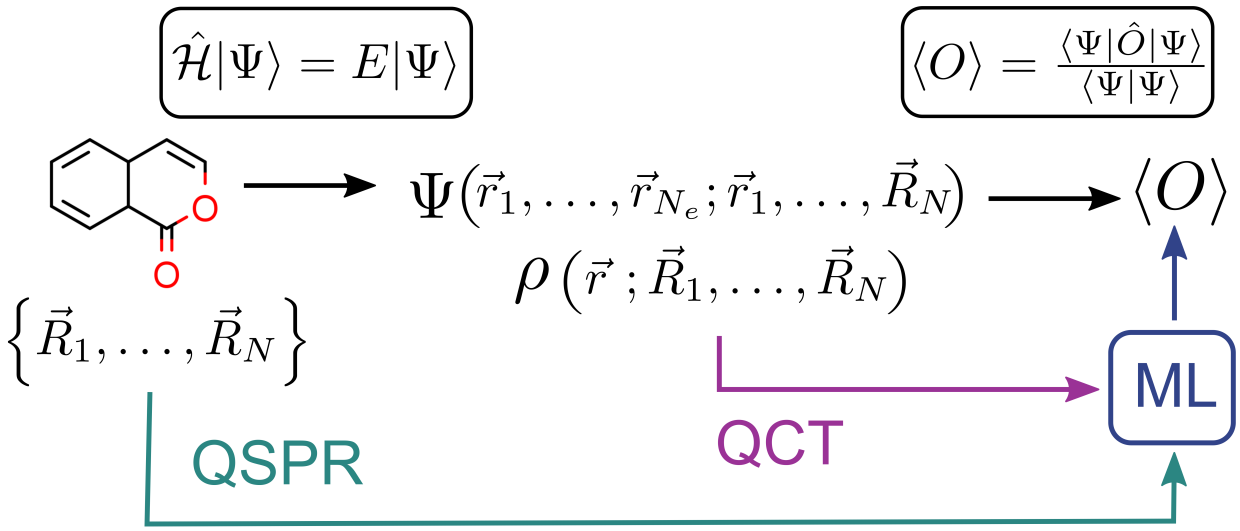


Figura 1.1. Esquema en donde se muestra el proceso para el cálculo de las propiedades moleculares. Las flechas negras muestran el esquema basándose únicamente en la teoría cuántica. Las flechas verdes y magentas muestran el esquema para el cálculo de las propiedades moleculares desarrollado para este proyecto, en el cual se sustituye la resolución de las ecuaciones de la mecánica cuántica por un algoritmo de aprendizaje automatizado. Aquí la flecha verde corresponde a emplear la aproximación relación estructura-propiedad (QSPR "Quantitative Structure–Property Relationship") mientras que la flecha magenta corresponde a emplear el enfoque de la topología de la densidad electrónica (QCT "Quantum Chemical Topology").

Es importante hacer notar que debido a la aproximación de Born-Oppenheimer (la cual nos permite hacer una separación en una función de onda nuclear y otra electrónica) la función de onda electrónica adquiere una dependencia paramétrica con respecto a las coordenadas nucleares $\left\{\vec{R}_1,\ldots,\vec{R}_N\right\}$ (figura 1.1). Es por este último hecho que toma "fuerza" la hipótesis de la relación estructura propiedad ("Quantitative Structure–Property Relationship" (QSPR) en inglés). Esta hipótesis postula que se puede conocer cualquier propiedad molecular únicamente conociendo la estructura de dicha molécula.

Para este estudio se propuso trabajar sobre esta hipótesis y desarrollar un modelo que sustituya el resolver la ecuación de Schrödinger para encontrar la función de onda electrónica y la obtención del valor esperado, por un modelo de aprendizaje automatizado (flecha verde de la figura 1.1). Para realizar esto, es necesario decodificar la estructura de la molécula (cromóforo y disolvente) en valores numéricos que el algoritmo de aprendizaje automatizado pueda interpretar. Estos valores son conocidos como "descriptores", los cuales describen propiedades asociadas a la molécula. El cómo se lleva a cabo esta decodificación se describirá más adelante en la sección 3.2.

Una propuesta similar a la aproximación QSPR, es únicamente sustituir la obtención del observable por un modelo de aprendizaje automatizado (flecha magenta de la figura 1.1). Es decir, obtener la función de onda empleando la ecuación de Schrödinger y únicamente remplazar la obtención del observable por un modelo de aprendizaje automatizado. Para ello, se propone calcular dos tipos de descriptores (propiedades que decodifiquen la información de la función de onda). El primer tipo de descriptores se obtiene directamente de la función de onda, mientras que el segundo tipo de descriptores se calculan a partir de un análisis posterior de la topología de la densidad electrónica, para lo cual se empleó la teoría cuántica de átomos en moléculas [42] ("Quantum theory of atoms in molecules"(QTAIM) en inglés). Este tipo de metodología puede llegar a ser especialmente útil para los casos en donde no se tenga una expresión para calcular el observable, como es el caso del tiempo de vida de fluorescencia.

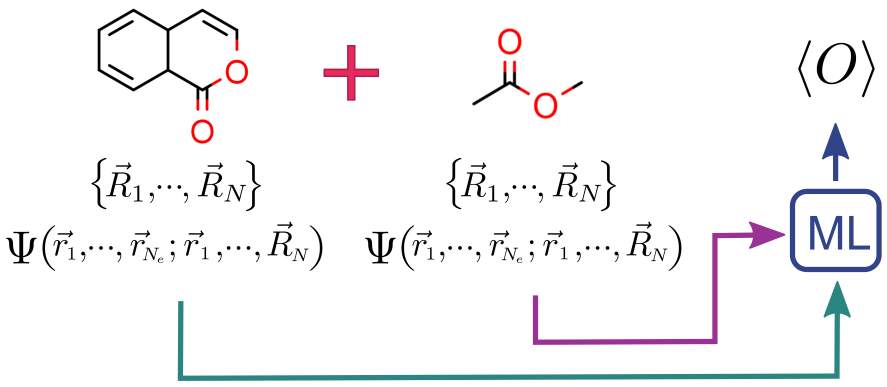


Figura 1.2. Esquema en donde se muestra como se implementó el disolvente para este estudio. La flecha verde muestra el proceso para el cálculo de las propiedades del cromóforo, mientras que la flecha magenta muestra el proceso para el cálculo de las propiedades del disolvente. Aquí las flechas corresponden a emplear la aproximación relación estructura-propiedad (QSPR "Quantitative Structure–Property Relationship") y a emplear el enfoque de la topología de la densidad electrónica (QCT "Quantum Chemical Topology").

Por último, y debido a que las propiedades espectroscópicas de los cromóforos no solamente dependen del cromóforo en sí, sino que también dependen del disolvente en el cual estos se encuentren, se decidió incluir al modelo de aprendizaje tanto la estructura del cromóforo como la del disolvente (figura 1.2)

Como se mencionó en la sección anterior, se han realizado estudios que tienen como objetivo principal emplear este tipo de metodologías para predecir propiedades espectroscópicas [43], [44] y [45]. Sin embargo, estos estudios carecen de un análisis detallado, además de una interpretación de las características o variables más importantes o cruciales que hacen que el modelo tenga éxito. Debido a esto, el presente trabajo proporciona un análisis sistemático de cuáles características (descriptores) son importantes para la predicción de las correspondientes propiedades espectroscópicas (energía de absorción, energía de emisión, desplazamiento de Stokes y el tiempo de vida de fluorescencia), así como de su importancia. En particular, se categorizaron a los diferentes descriptores de acuerdo con su naturaleza física, siendo los descriptores estructurales, cuánticos y empíricos, las categorías principales. En la sección 3.2 se ahondará más sobre en qué consiste cada categoría y cómo se calcularon cada uno de estos descriptores.

1.2. Sistemas de Estudio

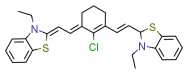
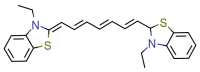
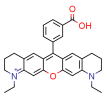
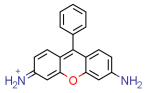
La idea de remplazar toda la teoría cuántica por un modelo de aprendizaje automatizado suena muy tentadora. Sin embargo, esto nos pone frente una gran cantidad de formas diferentes de abordar este problema, pudiendo implementar una gran cantidad de modelos de aprendizaje automatizado, así como una gran cantidad de variables para el entrenamiento de dichos modelos. Lo anterior convierte a este trabajo en un proceso de prueba y error, en el cual, se requiere probar diferentes modelos junto con diferentes variables de entrada para los mismos.

Esta situación nos pone frente al problema de encontrar qué variables son las más relevantes para la correcta descripción de nuestro sistema de estudio. En este trabajo estamos interesados en propiedades espectroscópicas relacionadas con las transiciones electrónicas. Estas propiedades dependen fuertemente de la estructura de la molécula. Por ejemplo, las cumarinas (las cuales tienen una estructura de dos anillos conjugados con un grupo éster) tienden a mostrar un máximo de absorción en longitudes de onda de entre 300 a 500 nm, mientras que, moléculas como las cianinas (las cuales se caracterizan por tener dos centros con nitrógeno unidos entre sí mediante una cadena polimetínica) presentan máximos de absorción en longitudes de onda más grandes (700 a 900 nm).

Además, la presencia de grupos funcionales tanto electrodonadores como electroatractores modifican la distribución de densidad electrónica alrededor del centro del cromóforo o “núcleo” [46], [47] (conjunto de átomos dentro de un cromóforo mayormente responsable de las propiedades espectroscópicas y que se mantienen presentes para un tipo de cromóforo en particular). Estos hechos sugieren que la variabilidad en la estructura de los cromóforos es un factor crucial para la descripción del sistema. Por esto, el modelo debería de incluir a los diferentes tipos de cromóforos, así como los efectos de los sustituyentes sobre cada uno de ellos.

Por último, pero no menos importante, un factor que modifica el valor de estas propiedades es el entorno de solvatación. Éste modifica la distribución de la densidad electrónica (estabilizándola) en diferente medida tanto si la molécula se encuentra en el estado basal, como si la molécula se encuentra en algún estado excitado, siendo diferente la magnitud de estas interacciones para ambos casos. Esto modifica directamente los máximos de absorción y emisión en diferente medida, haciendo que el implementar descriptores que contemplen el efecto del disolvente sea particularmente importante para la predicción de estas propiedades.

Tabla 1.1. Esquema de la base de datos para el Sistema 1, en donde se tienen registrados los valores de las diferentes propiedades espectroscópicas estudiadas para cada pareja específica del cromóforo junto con su disolvente. Los guiones corresponden a valores registrados de cada propiedad, mientras que "ND" (no disponible) hace referencia a los valores no reportados.

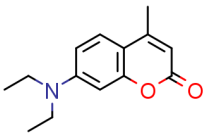
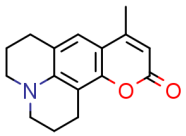
Estructura	Tipo de molécula	Disolvente	λ_{abs}	λ_{em}	ΔStokes	τ_{flu}
	Cianina	EtOH	—	—	—	—
		MeOH	—	—	—	—
		DMSO	—	—	—	ND
	Cianina	EtOH	—	—	—	—
		MeOH	—	—	—	—
		DMF	—	ND	ND	—
	Bodipy	THF	—	—	—	—
		MeOH	—	—	—	—
		CH_3CN	—	—	—	ND
	Rodamina	EtOH	—	—	—	—
		MeOH	—	—	—	—
		DMF	—	—	—	ND
	Rodamina	MeOH	—	—	—	—
		CH_3CN	—	ND	ND	—
		DMF	—	ND	ND	—

Tomando en cuenta tanto los efectos estructurales del cromóforo, así como las modificaciones a éste, ocasionadas por el disolvente, se propuso realizar dos sistemas de estudio; uno general que incluya predominantemente los parámetros asociados a la estructura del cromóforo (Sistema 1) y otro más específico que considere los efectos de los sustituyentes y el efecto del entorno de solvatación (Sistema 2).

El Sistema 1 está enfocado en predecir los efectos asociados a la estructura del cromóforo. Para esto se construyó una base de datos (tabla 1.1) con un aproximado de 500 cromóforos, de entre las cuales se encuentran: cianinas, cumarinas, bodipys, rodaminas y hemicianinas. Además, se recabó la información espectroscópica para cada cromóforo aproximadamente en 3 disolventes diferentes.

De este modo, el Sistema 1 logra incluir una gran variabilidad en la estructura del cromóforo, sin dejar de lado el efecto del disolvente. Las propiedades incluidas en el Sistema 1 son las siguientes: longitud de onda del máximo de absorción (λ_{abs}), longitud de onda del máximo de emisión (λ_{em}) y tiempo de vida de fluorescencia (τ_{flu}). Por otro lado, el Sistema 2 está enfocado principalmente en describir los efectos en las propiedades espectroscópicas causados por el entorno de solvatación y los grupos funcionales presentes en los cromóforos con un mismo "núcleo" (mismo tipo de cromóforo). Para esto se construyó un base de datos (tabla 1.2) con 30 moléculas, únicamente cumarinas, las cuales tiene una gran similitud estructural entre sí (tienen el mismo "núcleo"). Se seleccionaron cumarinas sobre los otros tipos de cromóforos, ya que éstas presentan un efecto solvatocrómico importante y sistemático [48], [49], [50], lo cual puede ayudar al modelo a identificar y predecir dicho efecto. Además, se registraron las propiedades espectroscópicas para cada molécula en al menos 10 disolventes diferentes. Lo anterior asegura incluir el efecto del disolvente sobre las propiedades espectroscópicas, en donde las propiedades para el Sistema 2 son solamente la longitud de onda del máximo de absorción (λ_{abs}) y longitud de onda del máximo de emisión (λ_{em}).

Tabla 1.2. Esquema de la base de datos para el Sistema 2, en donde se tienen registrados los valores de las diferentes propiedades espectroscópicas estudiadas para cada pareja específica de cromóforo junto con su disolvente. Los guiones corresponden a valores registrados de cada propiedad, mientras que "ND" (no disponible) hace referencia a los valores no reportados.

Estructura	Tipo de molécula	Disolvente	λ_{abs}	λ_{em}	ΔStokes
	Cumarina	EtOH	—	—	—
		MeOH	—	—	—
		CH_3CN	—	—	—
		DMSO	—	ND	ND
		$CHCl_3$	—	—	—
		THF	—	—	—
		DMF	—	—	—
	Cumarina	EtOH	—	—	—
		MeOH	—	—	—
		CH_3CN	—	ND	ND
		DMSO	—	—	—
		THF	—	—	—
		DMF	—	—	—
		$CHCl_3$	—	ND	ND

1.3. Aprendizaje automatizado (Machine Learning)

Gran parte de este estudio consta de evaluar el desempeño de los diferentes modelos de aprendizaje automatizado empleando diferentes subconjuntos de datos (diferentes propiedades y descriptores). Por lo ya mencionado le dedicaremos una sección completa a abordar este tema. En esta sección se abordarán desde los conceptos más básicos sobre el aprendizaje automatizado, pasando por los métodos para el entrenamiento y evaluación de los modelos de aprendizaje automatizado, finalizando con una breve descripción del funcionamiento de los modelos implementados para este trabajo.

El concepto de aprendizaje automatizado, como ya se mencionó, es ampliamente usado en diferentes áreas del conocimiento. Esto incluye áreas ajenas a las ciencias naturales. Debido a esto, existen diferentes tipos de aprendizaje automatizado, los cuales difieren mucho uno del otro, tanto en los modelos que implementan como en las metodologías para aproximarse al resultado final, así como las métricas para evaluar el desempeño del modelo. Debido a esto y para no alejarse del objetivo de este estudio, esta sección se limitará a desglosar únicamente el procedimiento empleado para el presente estudio, el cual, pertenece a la categoría de aprendizaje supervisado empleando modelos de regresión.

1.3.1. Conceptos básicos

El objetivo del aprendizaje automatizado es abordar un problema de predicción de algún fenómeno haciendo uso de información previamente recabada sobre dicho fenómeno. Este tipo de metodología es especialmente útil cuando no se conocen con precisión las variables ni sus relaciones involucradas en la descripción del fenómeno. Estos métodos también son apropiados cuando el resolver las ecuaciones para dicho fenómeno conlleva un costo computacional muy elevado, siendo esta última la razón principal de emplear aprendizaje automatizado para este estudio.

Principalmente, los fenómenos químicos, físicos, biológicos e incluso económicos, pueden ser descritos mediante una función $f(x)$ que correlaciona una o más variables independientes x con un variable dependiente y.

$$y = f(x) \tag{1.1}$$

Por ejemplo, la desintegración radioactiva es un fenómeno que puede ser descrito mediante un decaimiento exponencial, en donde la función que lo describe es la siguiente:

$$y = y_0 \, e^{-\frac{x}{\tau}} \tag{1.2}$$

donde x es el tiempo, $y(x)$ es la cantidad de componente radioactivo en función del tiempo, y_0 es la cantidad inicial de componente radioactivo y τ es el tiempo de vida medio (el tiempo que tarda en descomponerse la mitad del componente radioactivo).

En este caso, la función tiene una forma analítica y un número de variables suficientemente pequeño, lo cual permite que el fenómeno se pueda entender de manera intuitiva. Sin embargo, existen casos en donde el número de variables es mucho mayor o la forma de la función es desconocida. Es para este tipo de casos en donde se emplea la metodología de aprendizaje automatizado, la cual, a través de identificar patrones puede lograr descifrar la correlación que existe entre las variables independientes (variables de entradas) con la de variables dependientes (variables de salida).

Para plantear un esquema de aprendizaje, se requiere proporcionar al algoritmo automatizado una serie de variables de entrada junto con su correspondiente valor de salida. Esto para que el algoritmo pueda correlacionar los valores de entrada con los de salida. A este conjunto de un valor de salida junto con sus correspondientes valores de entrada se le conoce como ejemplar:

$$(y_i, [X_i]) = (y_i, [x_{i,1}, x_{i,2}, \ldots, x_{i,n}]) \tag{1.3}$$

Para que el algoritmo prediga apropiadamente, se necesita proporcionarle a éste el suficiente número de ejemplares, los cuales deben ser representativos del fenómeno que se desea estudiar. De este modo, el algoritmo de aprendizaje automatizado queda representado como la función F(), la cual correlaciona las variables de entrada y salida para cada uno de los ejemplares. Cabe recalcar que la forma de esta función depende completamente del algoritmo de aprendizaje automatizado empleado.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = F \begin{bmatrix} x_{1\,1} & x_{1\,2} & \ldots & x_{1\,m} \\ x_{2\,1} & x_{2\,2} & \ldots & x_{2\,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n\,1} & x_{n\,2} & \ldots & x_{n\,m} \end{bmatrix} \quad ; \, [y_i] = F[x_{ij}] \tag{1.4}$$

donde, n es el número de ejemplares y m el número de variables independientes diferentes. En la notación de aprendizaje automatizado, a la matriz de entradas $[X_{i\,j}]$ se le conoce como vector de características, en donde cada columna $[X_j]$ corresponde a una característica diferente.

Es importante mencionar que no existe un único algoritmo de aprendizaje automatizado. La forma en la que cada uno de los diferentes modelos funciona es completamente diferente. Sin embargo, la mayoría de los modelos funcionan mediante un proceso de minimización, en el cual se ajusta el modelo para minimizar una función de pérdida, la cual usualmente es el error cuadrático medio. Por ejemplo, una regresión lineal realiza el proceso de mínimos cuadrados, en el cual se construye una recta en donde se busca minimizar el error cuadrático medio. Por otro lado, los árboles de decisión realizan un proceso de minimización muy similar. Esto se abordará más a detalle en secciones posteriores.

Notación

- Vector de características $[X_{i\,j}]$

$$[X_{i\,j}] = \begin{bmatrix} x_{1\,1} & x_{1\,2} & \dots & x_{1\,m} \\ x_{2\,1} & x_{2\,2} & \dots & x_{2\,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n\,1} & x_{n\,2} & \dots & x_{n\,m} \end{bmatrix} \tag{1.5}$$

  donde i itera sobre cada elemento (filas) y j itera sobre las diferentes características $[X_j]$ (columnas).

- Característica $[X_j]$

$$[X_j] = \begin{bmatrix} x_{1\,j} \\ x_{2\,j} \\ \vdots \\ x_{n\,j} \end{bmatrix} \tag{1.6}$$

- Salida $[Y_i]$

$$[Y_i] = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{1.7}$$

- Ejemplar $(y_i, [X_i])$. Para este caso particular, un ejemplar corresponde a un cromóforo en un disolvente específico.

$$(y_i, [x_{i,1}, x_{i,2}, \dots, x_{m,i}]) \tag{1.8}$$

1.3.2. Implementación de los métodos

La implementación de los métodos de aprendizaje automatizado se realiza en 4 pasos. El primer paso es realizar un preprocesado al vector de características (entradas), esto para asegurarse de que el modelo pueda entrenarse sin ninguna clase de error. El segundo paso es entrenar al modelo empleando tanto las variables de entrada como las de salida. Por último, se realiza la evaluación del modelo, lo cual se lleva a cabo en dos etapas: la validación y la prueba.

Preprocesado

La etapa del preprocesado consta de un conjunto de pasos que modifican al vector de características. Un primer paso es imputar los datos faltantes. Existen casos en donde no hay datos disponibles de alguna característica. Por ejemplo, si una molécula no tiene un triple enlace carbono-carbono, el valor de la distancia del enlace carbono-carbono será imposible de calcular. Por lo que es necesario reemplazar los datos faltantes por valores que se aproximen al comportamiento de los demás datos. Para ello existe una gran variedad de algoritmos y métodos. También es necesario estandarizar los valores de cada característica, esto para evitar que el modelo presente errores asociados a valores muy grandes o alejados del cero (figura 1.3). Para ello se modifican los valores de cada característica empleando la ecuación 1.9.

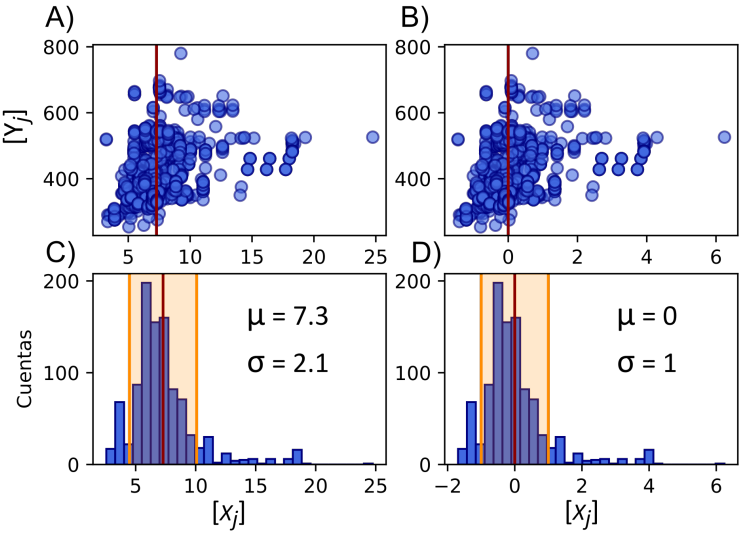


Figura 1.3. Proceso de estandarización de una característica $[X_j]$, en donde A) y B) muestran la gráfica de la característica $[X_j]$ contra el valor de salida $[Y_i]$ y las gráficas C) y D) muestran el histograma de dicha característica. Las gráficas A) y C) representan a la característica antes de la estandarización mientras que las gráficas B) y D) a la característica después del proceso de estandarización. Como se puede apreciar, después de la estandarización, el valor de la media μ es igual 0 y el valor de la desviación σ es uno. Para este ejemplo en particular la salida $[Y_i]$ es la longitud de onda del máximo de absorción en nanómetros y la característica $[X_j]$ es un descriptor estructural.

$$x_{ij}^{St} = \frac{x_{ij} - \widehat{x_j}}{\sigma_{x_j}} \tag{1.9}$$

Otro paso muy importante, es la reducción de dimensión. Durante esta etapa se seleccionan las características (columnas en el vector de características) más relevantes, reduciendo así la dimensión del vector de características. Este proceso se puede realizar empleando diferentes algoritmos. Para este estudio se empleó un método basado en una regresión F, la cual es una prueba estadística que nos dice qué variables de entrada (características) correlacionan más con las variables de salida.

Por último, y debido a que los modelos de aprendizaje automatizado no están basados en alguna teoría rigurosa, es necesario asegurarse que el modelo funcione correctamente. Para ello se entrena al modelo con un conjunto de datos (entrenamiento) y se evalúa al modelo con otros conjuntos completamente diferentes (validación y prueba). Esto garantiza que el modelo sea capaz de desempeñarse correctamente con un conjunto de datos con el que no se entrenó.

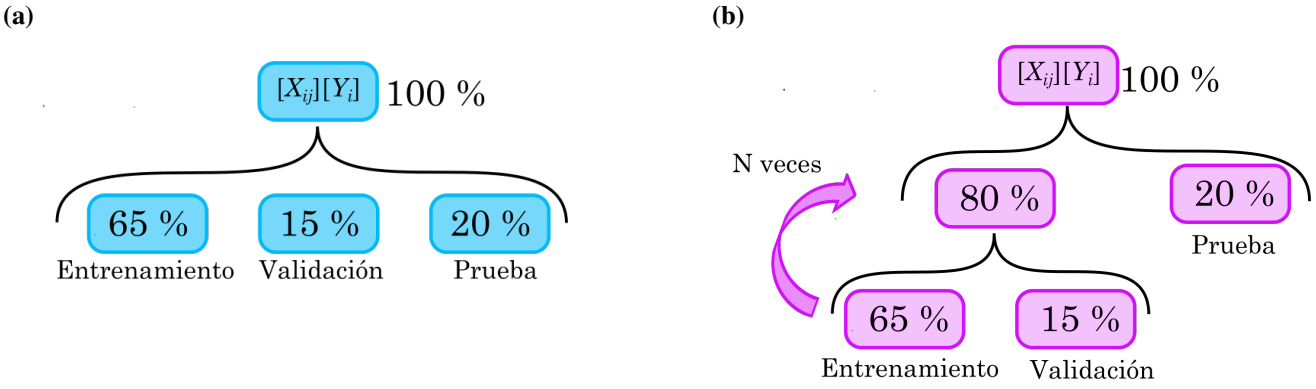


Figura 1.4. Esquema donde se muestra la distribución de los diferentes subconjuntos de datos (entrenamiento, validación y prueba) con sus respectivos porcentajes aproximados. La sub-figura **(a)** muestra el esquema de distribución para una validación simple, mientras que la sub-figura **(b)** muestra el esquema donde se considera una validación cruzada, en donde la división para obtener el conjunto de entrenamiento y validación se realiza una vez por cada iteración en el algoritmo de validación cruzada.

Algo importante a puntualizar es que la división del conjunto de validación y de prueba, depende del método de validación que se va a emplear. Para una validación simple se hace una división en tres grupos, uno para cada etapa: entrenamiento, validación y prueba. Mientras que para la validación cruzada sólo se separa en dos grupos, uno de entrenamiento y otro de prueba, esto debido a que el mismo algoritmo de validación generará varias subdivisiones de entrenamiento y validación sobre el grupo de entrenamiento inicial (Figura 1.4).

Entrenamiento

Como ya se mencionó en la sección 1.3.1, cada algoritmo propone una forma diferente de aproximarse a la función de predicción. Esto lleva a que la selección del modelo sea un paso crucial para el éxito o fracaso del modelo. Por ejemplo, para un conjunto de datos dado, un modelo puede predecir con buena precisión, mientras que otro modelo puede que no lo haga con la misma precisión o que incluso no encuentre correlación alguna. En este trabajo se emplearon mayoritariamente modelos basados en árboles de decisión como lo son el modelo de un bosque aleatorio o "RandomForestRegressor" en inglés y el modelo de "GradientBoostingRegressor" en inglés.

Validación

El proceso de validación consta de hacer una serie de iteraciones en donde se evalúa el modelo (empleando el subconjunto de validación) para, posteriormente, realizar ajustes sobre los hiperparámetros hasta encontrar el algoritmo óptimo que realice la mejor predicción posible. Los hiperparámetros de un modelo son parámetros ajustables del algoritmo o modelo. Por ejemplo, para un árbol de decisión, algunos hiperparámetros son el número de hojas o la profundidad del árbol. Además, existe un tipo de validación conocido como validación cruzada, la cual hace N validaciones internas con diferentes selecciones de los subconjuntos de entrenamiento y el de validación (Figura 1.5).

Este tipo de validación tiene la ventaja de que nos muestra qué tan robusto es nuestro modelo con respecto a la selección de un conjunto de datos en particular. Por ejemplo, puede ser el caso de que a la hora de dividir nuestro conjunto de datos entre entrenamiento y validación, el subconjunto de datos que se escogió para el entrenamiento sea tal que el modelo correlacionó con un $r^2 = 0.9$. En dicho caso uno podría pensar que ya se tiene el modelo ideal. Sin embargo, a la hora de volver a hacer la misma prueba, pero con una división del conjunto de entrenamiento y validación diferentes, el resultado puede llegar a ser muy diferente, por ejemplo, una correlación de $r^2 = 0.8$. Esto evidencia que los resultados del modelo dependen fuertemente de qué subconjunto de datos sean seleccionados para el entrenamiento y la validación, haciendo al modelo poco robusto.

Para realizar la validación, se divide el subconjunto de entrenamiento en un subconjunto de entrenamiento y otro de validación. Se selecciona un número k de datos con los cuales se evaluará el modelo (validación). Posteriormente, se realizará el entrenamiento y la evaluación, obteniendo los valores de las métricas para esta iteración. En una segunda iteración, se selecciona un nuevo conjunto de k elementos del conjunto original como conjunto de validación y se vuelven a realizar las respectivas evaluaciones. Estos pasos se repiten N/k veces donde N es el número de datos del subconjunto inicial (figura 1.5).

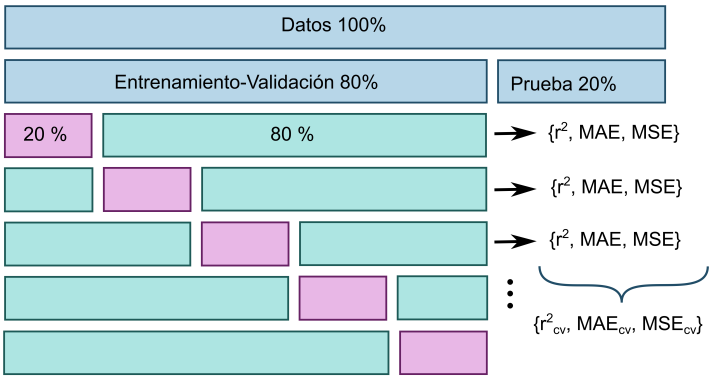


Figura 1.5. Proceso por el cual se lleva a cabo la validación cruzada. En un primer paso se divide el conjunto de datos inicial en dos subconjuntos, uno de Entrenamiento-Validación y otro de Prueba). En un segundo paso (el cual se realiza iterativamente) se divide el conjunto de Entrenamiento-Validación en dos subconjuntos, uno de entrenamiento (verde) y otro de validación (rosa). Para cada uno de estos subconjuntos se obtiene el valor de las respectivas métricas (Coeficiente de correlación r^2, Error absoluto medio MAE y cuadrado del error medio MSE). Finalmente, se obtiene un promedio de las métricas obtenidas para cada iteración.

Por último, con los valores de las métricas para cada iteración se obtiene un valor promedio para cada métrica y su valor de desviación e incertidumbre correspondiente. Por ejemplo, para una validación iterando 5 veces se obtuvieron los siguientes valores de coeficiente de correlación: 0.9, 0.9, 0.8, 0.8, 0.8. Por lo tanto, el valor de coeficiente de correlación a reportar empleando la validación cruzada sería el promedio de los resultados en cada iteración junto con su correspondiente valor de incertidumbre dada ($u = \sigma_{St}/\sqrt{N}$), donde σ_{St} es la desviación estándar. El valor de incertidumbre es multiplicado por el factor de cobertura, $k = 2$ obteniendo así la incertidumbre expandida $U = k\,u$. El factor de cobertura es un factor estadístico que representa el nivel de confianza (o cobertura) de la medición, un factor de cobertura $k = 2$ (bajo una distribución normal) corresponde a una confiabilidad del 95 % es decir que al realizar una medición se tiene un 95 % de probabilidad de que el valor real de la medición esté en el intervalo especificado por la incertidumbre ([medición -U, medición +U]).

Para el ejemplo anterior, el resultado se reportaría de la siguiente forma: r^2(Validación)$= 0.840 \pm 0.049$. Este proceso se realiza para cada una de las diferentes métricas, obteniendo así los valores de las correspondientes métricas: coeficiente de correlación r^2(Validación), error absoluto medio *MAE* (Validación) y error cuadrático medio *MSE* (Validación).

Prueba

En la etapa de prueba se evalúa el desempeño del modelo final (después de haber realizado la validación). Para ellos se emplean una serie de métricas de evaluación, las cual nos proporcionan diferente información acerca del desempeño de los modelos. Las métricas más usuales para regresión y las que se trabajaron durante el desarrollo de este proyecto son las siguientes:

- **Coeficiente de correlación r^2**

- **Error cuadrático medio (MSE)**

$$MSE = \frac{1}{n}\sum_{i}^{n}(\hat{y} - y_i)^2 \tag{1.10}$$

  Donde y_i es cada elemento de $[Y_i]$, donde $\hat{y}$ es el valor predicho por el modelo.

- **Raíz cuadrada del error cuadrático medio (RMSE)**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(\hat{y} - y_i)^2} \tag{1.11}$$

- **Error absoluto medio (MAE)**

$$MAE = \frac{1}{n}\sum_{i}^{n}|(\hat{y} - y_i)| \tag{1.12}$$

- **Error relativo a la media (% Error)**

$$\%\text{Error} = \frac{\text{MAE}}{\mu}\text{x}100\,\% \tag{1.13}$$

  donde μ es el valor medio de la distribución (de cada propiedad) y MAE el error absoluto medio obtenido por el modelo.

El coeficiente de correlación r^2 se refiere al grado de correlación entre las características $[X_{ij}]$ con la salida $[Y_i]$, mientras que el error absoluto medio (MAE) y la raíz cuadrada del error cuadrático medio (RMSE) se refiere al grado de desviación de los valores predichos con respecto a los valores experimentales. Por último, el error relativo a la media (% Error) se calcula con el objetivo de visualizar porcentualmente el error cometido por el modelo. Empleando estas métricas se tienen casi todos los elementos para poder evaluar el desempeño del modelo. Sin embargo, es necesario realizar las gráficas de valor predicho contra valor real, ya que estas gráficas nos pueden indicar si el modelo está sobre ajustando o cometiendo algún otro tipo de error que no se puede saber a priori únicamente con los valores de las métricas de la validación.

1.3.3. Modelos de aprendizaje automatizado basados en árboles de decisión

Uno de los objetivos principales de este proyecto es proporcionar herramientas para el diseño de nuevos cromóforos que presenten propiedades espectroscópicas específicas. Para ello se requiere de conocer como las propiedades físicas y estructurales de una molécula modifican sus propiedades espectroscópicas. Es por ello que se decidió trabajar con modelos basados en árboles de decisión, ya que estos nos proporcionan una métrica (importancia relativa de características) que nos dice qué características (descriptores) son las más relevantes para la predicción de las propiedades estudiadas.

Un árbol de decisión es una función construida en secciones (ecuación 1.14), en donde para un intervalo dado en la coordenada "x" le corresponde un valor en la coordenada "y" (figura 1.6-a). Esta forma de entender a los árboles de decisión es muy sencilla. Sin embargo, tiene sus limitaciones. Los árboles que dependan de más de dos características son difíciles de representar mediante una ecuación e imposible de visualizar, ya que esto correspondería a una gráfica en más de 3 dimensiones.

$$y = F(x) = \begin{cases} 3, & 0 < x < 5 \\ 5, & 5 < x < 10 \\ 7, & 10 < x < 15 \\ 3, & x > 15 \end{cases} \tag{1.14}$$

Otra forma de entender a un árbol de decisión es mediante una estructura ramificada (figura 1.6-b), de ahí el nombre de árbol, en donde partimos de un nodo raíz (Q) el cual contiene a todos los datos q_i. Posteriormente, cada nodo se divide en dos intervalos (nodos hijos Q_{left} y Q_{right}) dividido por una separación $\theta(t_m)$ que depende del valor de la variable en donde se realizará la división t_m. Teniendo como resultado que los valores que sean mayores al valor de t_m se distribuyan en el nodo de la derecha ($q_i < t_m$), mientras que los datos que sean iguales o menores se colocan en el nodo de la izquierda ($q_i \geq t_m$).

Esta forma de representar a un árbol hace que sea más fácil entenderlo, debido a que ya no es necesario representar una dimensión para cada característica $[X_j]$. De esta forma se implementa una dependencia de la división con respecto a cada característica. Expresando a la división de la siguiente forma $\theta(t_m, [X_j])$, en donde ésta depende de en qué característica se va a realizar la división $[X_j]$ y el valor en el cual se realiza la división t_m. Un hecho a recalcar es que la división se puede hacer sobre cualquier característica, no importando el orden o si es una característica repetida. Los nodos que ya no tienen ninguna división (finales) son conocidos como nodos finales u hojas.

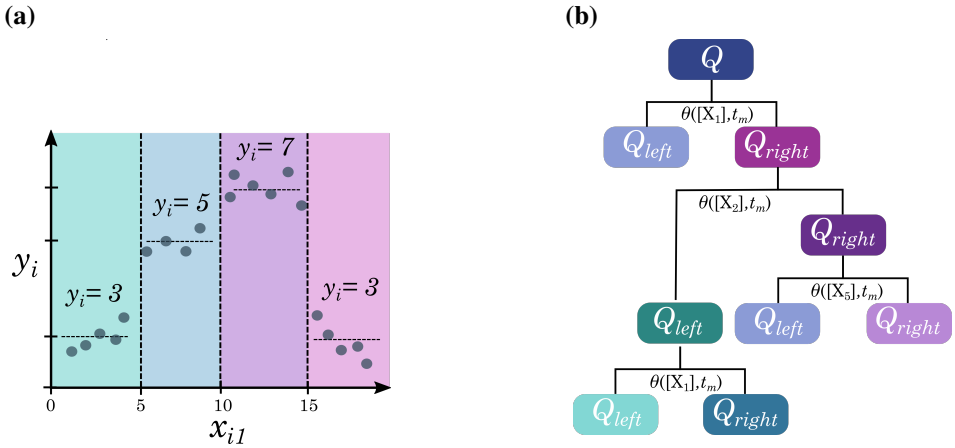


Figura 1.6. (a) Gráfica en donde se muestran los intervalos de la función (ecuación 1.14) junto con los datos que ajustan a ésta. **(b)** Esquema de un árbol de decisión. Este está construido a partir de un nodo raíz, el cual se divide sucesivamente en nodos hijos, finalizando con las hojas, los cuales ya no se dividen. La división $\theta(t_m, [X_j])$ depende de la correspondiente característica $[X_j]$ y en el valor t_m.

Para este proyecto se emplearon dos tipos de modelos, los cuales se construyen a partir de ensamblar varios árboles de decisión. Esto con el objetivo de reducir el sobreajuste que presentan los árboles de decisión de manera individual. El primer tipo de modelo (bosque aleatorio o "RandomForestRegressor") como su nombre lo dice, crea un bosque de árboles (conjunto de árboles de decisión), en donde la forma de construir árboles estructuralmente diferentes unos de otros es a través de hacer un sub muestreo del conjunto de datos original "Bootstrap". Ese se construye tomando, uno a uno, diferentes elementos del conjunto original, obteniendo un conjunto con el mismo número de elementos que el original pero con ciertos datos repetidos "aleatoriamente". Finalmente, el error (determinado por las métricas de evaluación) asociado al bosque completo (conjunto de árboles) corresponde al promedio de los errores asociados a cada árbol.

Para el segundo tipo de modelo ("GradientBoostingRegressor") se construye una serie sucesiva de árboles de decisión, los cuales se crean corrigiendo los errores cometidos por el árbol anterior. Esto se logra a través de construir árboles sucesivamente, los cuales envés de predecir la salida $[Y_i]$ predigan los residuos $[r_i] = [y_i] - [y_i^{\text{Predicho}}]$ obtenidos por modelo en la iteración anterior. De modo que el modelo para cada iteración (cada árbol en la serie sucesiva) sé construye sumando el valor que predice cada árbol hasta ese punto construido, ponderándolos con una variable conocida como taza de aprendizaje. De esta forma se logra que este modelo converja en un único árbol con un menor error que todos los anteriores 1.7).

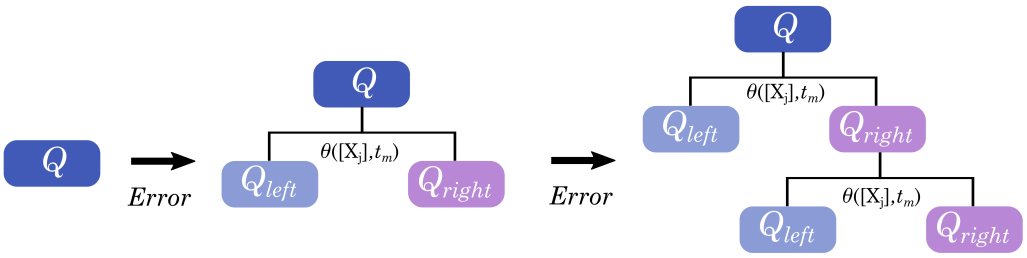


Figura 1.7. Esquema en donde se muestra como se construye el modelo de "GradientBoostingRegressor", el cual se construye a partir de una serie sucesiva de árboles de decisión, los cuales se crean corrigiendo los errores cometidos por el árbol previo.

Objetivos e hipótesis

2.1. Objetivo general

Construir modelos de aprendizaje automatizado capaces de predecir las propiedades espectroscópicas de ciertos grupos de cromóforos en diferentes disolventes. En particular, la energía de absorción, la energía de emisión, el desplazamiento de Stokes y el tiempo de vida de fluorescencia.

2.2. Objetivos particulares

- La construcción de las dos bases de datos correspondientes a los dos sistemas de estudio. La primera base de datos con información espectroscópica de una variedad de diferentes tipos de cromóforos (en varios disolventes). Y la segunda base de datos con información espectroscópica únicamente con sólo un tipo de cromóforo (cumarinas), pero con una mayor cantidad de disolventes diferentes por cada cromóforo.
- Realizar el cálculo de descriptores estructurales, cuánticos y empíricos de los diferentes cromóforos y disolventes presentes en las bases de datos, los cuales van a servir como entradas para el entrenamiento de los algoritmos de aprendizaje automatizado.
- Construcción de una librería en Python (“ML_molecule”) que contenga todas las herramientas necesarias (métodos y funciones) para poder construir, entender y evaluar los modelos de aprendizaje automatizado. Para ello se requiere de seleccionar los métodos de aprendizaje automatizado que se van a emplear. Incluyendo tanto los modelos de aprendizaje automatizado, así como otros algoritmos necesarios previos al entrenamiento (método de reducción de dimensión).
- Entrenar y evaluar el desempeño de los modelos construidos para la predicción de las correspondientes propiedades espectroscópicas.

- Realizar un análisis sistemático del desempeño (en la predicción) de cada tipo de descriptor, para lo cual se evaluará a los modelos de aprendizaje empleando diferentes subconjuntos de descriptores, incluyendo un análisis de la importancia relativa de estos descriptores para el modelo. Esto nos brindará una forma de identificar qué propiedades físicas y estructurales pueden ser útiles para la modificación y diseño para nuevos cromóforos.

2.3. Hipótesis

Implementar propiedades (descriptores) cuánticas como características en modelos basados en árboles de decisión, prometen proporcionar tanto una ventaja cuantitativa en la predicción de las propiedades estudiadas, así como una ventaja cualitativa, proporcionando herramientas de diseño para la síntesis de nuevos cromóforos, esto debido a que estos tipos de descriptores se relacionan directamente con las propiedades electrónicas de los cromóforos.

Metodología

3.1. Construcción de las bases de datos

Como ya se mencionó en la sección 1.2, se trabajaron dos sistemas de estudio diferentes, para lo cual se requirió de la construcción de una base de datos para cada uno de los sistemas. Para el Sistema 1 se recabó información espectroscópica de un total de 475 cromóforos diferentes, incluyendo Cianinas, Bodypis, Cumarinas y Rodaminas. Para cada uno de estos cromóforo se buscó información de sus propiedades espectroscópicas, incluyendo: el máximo de la longitud de onda de absorción λ_{abs} y emisión λ_{em}, junto con el tiempo de vida de fluorescencia τ_{flu}. A partir de estas propiedades se calcularon las energías de absorción E_{abs} y emisión E_{em} empleando la ecuación 3.1. Los valores se reportaron en unidades de energía "electrón volt" [eV].

$$E = \frac{hc}{\lambda} \tag{3.1}$$

donde h es la constante de Plank y c la velocidad de la luz en el vacío.

Además, se calculó el valor del desplazamiento de Stokes (ΔStokes) haciendo la resta entre la energía de absorción y la energía de emisión. Para el tiempo de vida de fluorescencia, algunas fuentes reportaban más de un componente del tiempo de vida, estas componentes pueden ser ocasionadas por la presencia de diferentes isómeros o conformeros, así como a mecanismos cinéticos más complejos. Debido a esto, se decidió tomar en cuenta el tiempo de vida promedio (ponderado) obtenido a partir de la ecuación 3.2.

$$\tau_{\text{flu}}(\text{Promedio}) = \sum_i \alpha_i \tau_i \tag{3.2}$$

donde $\tau_{\text{flu}}(\text{Promedio})$ es el tiempo de vida promedio, τ_i es cada una de las componentes del tiempo de vida y α_i los pesos de cada uno de los correspondientes tiempos de vida sobre la función de decaimiento total. Los valores de tiempo de vida de fluorescencia son reportados en nanosegundos [ns].

Para cada cromóforo se recabó información espectroscópica en diferentes disolventes (81 disolventes diferentes). De modo que al final se obtuvieron un total de 1430 datos (parejas cromóforo – disolvente). Además, la información de la estructura de cada molécula, tanto cromóforos como disolventes, se capturó en formato SMILES ("Simplified Molecular Input Line Entry System") a partir de la estructura proporcionada en cada uno de los artículos de donde se recabó la información espectroscópica.

Para la base de datos del Sistema 2 se incluyeron un total de 30 cromóforos diferentes, únicamente cumarinas. Para cada molécula se buscó información de sus propiedades espectroscópicas, incluyendo: el máximo de la longitud de onda de absorción λ_{abs} y emisión λ_{em}, a partir de estas propiedades se calcularon las energías de absorción E_{abs} y emisión E_{em} correspondientes, reportándolas en unidades de energía "electrón volt" [eV]. Posteriormente, se calculó el valor del desplazamiento de Stokes (ΔStokes), haciendo la resta entre la energía de absorción y la energía de emisión.

Por último, para cada cumarina se encontró información espectroscópica en diferentes disolventes, con un promedio de 10 disolventes por cada molécula, teniendo al final un total de 371 datos (parejas de cromóforo – disolvente) y un total de 59 disolventes diferentes. Los conjuntos de datos completos tanto para el sistema 1 y 2, junto con sus correspondientes referencias, se puede encontrar en [51] (https://github.com/BernaASS/Chromophore-database).

3.2. Descriptores

Los descriptores los podemos clasificar, basándose en su naturaleza, en tres grupos principales. El primero corresponde a los descriptores estructurales, el segundo a los cuánticos y el tercero a los empíricos. Se realizó el cálculo de los descriptores estructurales y cuánticos tanto para el cromóforo como para el disolvente, mientras que los descriptores empíricos solo se calcularon para el disolvente.

A lo largo de esta sección y en secciones posteriores se mencionarán algunas funciones específicas, las cuales desempeñan tareas tanto para la construcción de la base de datos, el cálculo de los descriptores y el entrenamiento de los modelos de aprendizaje automatizado. Todas ellas son funciones implementadas en la librería de Python "Machine Learning Molecule" [52]. Esta librería fue desarrollada como resultado de este trabajo y tiene como objetivo principal simplificar la tarea de implementar y desarrollar proyectos que requieran de metodologías de aprendizaje automatizado en moléculas. Esto lo logra al compactar todos los pasos requeridos para ello en unas pocas clases y métodos que funcionan de una manera simple e intuitiva.

3.2.1. Descriptores estructurales

Como se vio en la sección 1.3, los algoritmos de aprendizaje automatizado requieren de una entrada consistente en tamaño $[X_{ij}]$, donde para cada fila i (la cual corresponde a una pareja cromóforo - disolvente) se tenga el mismo número de características $[X_j]$. Debido a esto, no se puede introducir la posición de cada uno de los átomos (x, y, z) directamente al algoritmo de aprendizaje, ya que cada molécula tiene diferentes tipos y números de átomos. Con esto, se requiere decodificar la posición y la información de la conectividad de una molécula en valores que no dependan del número de átomos para proporcionar una entrada consistente en tamaño al algoritmo de aprendizaje automatizado.

A este tipo de descriptores se les conoce como descriptores estructurales, los cuales describen alguna propiedad o característica de la molécula (no necesariamente física), la cual se relaciona directamente con la estructura de la misma. Estos descriptores pueden ser tan simples como únicamente el número de átomos de oxígeno o átomos de carbono, el número de anillos aromáticos o bien funciones más complejas que ponderen la posición de todos los átomos con respecto a propiedades atómicas. Existen varios programas que calculan este tipo de descriptores. Ejemplo de estos son el programa CODESSA [53] o la librería de Python MORDRED [54].

Debido a que este trabajo está construido sobre un entorno de programación de Python 3, el programa más cómodo de emplear es MORDRED, ya que ésta es una librería de Python la cual permite que se calcule el valor de los descriptores en la interfaz de Python ("on the fly"). De otro modo, con los demás programas es necesario obtener los descriptores en el programa correspondiente y posteriormente importarlos empleando la librería de Pyhton "Machine Learning Molecule". Sin embargo, y debido a que emplear un único programa para calcular estos descriptores podría ocasionar algún tipo de error sistemático, se decidió también emplear el programa CODESSA.

Previo a realizar el cálculo de los descriptores estructurales, es necesario obtener una estructura optimizada de las moléculas. Para ello se utiliza la siguiente metodología.

- Se parte de la estructura del cromóforo y del disolvente en formato SMILES presente en la base de datos.
- Mediante una librería de Python llamada "RD-kit" [55] se interpretó la estructura SMILES instanciando una clase (especial de "RD-kit") la cual contiene la información de la estructura 2D y la conectividad de la molécula.
- Esta librería también implementa un método de optimización mediante mecánica molecular, el cual se emplea para obtener una estructura 3D optimizada de la molécula. Para ello se empleó el método ETKDG [56] de optimización.

Cada uno de estos pasos se realiza automáticamente para todos los cromóforos y disolventes, empleando la función ("Structure_generator"). Posteriormente, se exportaron las estructuras 2D y 3D de las moléculas, esto en dos archivos diferentes (formato .sdf) uno para la estructura 2D y otro para la estructura 3D. Este tipo de archivos es ampliamente usado en el manejo de bases de datos de moléculas y se construyen a partir de concatenar la estructura de cada una de las moléculas en formato MOL en un único archivo (formato .sdf), el cual contiene la estructura de todas las moléculas.

Los descriptores incluidos por la librería MORDRED se calculan a partir de la base de datos de moléculas, empleando la función ("Descriptor_generator"). Esta exporta una tabla junto con todos los descriptores calculados (columnas) para todas las moléculas (filas). Se exportaron dos archivos (en formato.csv), uno para los descriptores del cromóforo y otro para los descriptores del disolvente.

Para el cálculo de los descriptores empleando el programa CODESSA se sigue una metodología similar. El programa requiere de importar la base de datos de moléculas y como resultado exporta un archivo (en formato.csv) con una tabla que incluye todos los descriptores calculados (columnas) para todas las moléculas (filas). Una breve lista y descripción de los descriptores implementados se encuentra en el apéndice B.

3.2.2. Descriptores cuánticos

De igual manera que en el caso de los descriptores estructurales, no se puede introducir directamente la función de onda al algoritmo de aprendizaje automatizado, por lo que se requiere decodificar la información de la función de onda en valores numéricos que a su vez no dependan del número de átomos en la molécula. A este tipo de variables que describen a la molécula, con información obtenida a partir de un cálculo de estructura electrónica, se les conoce como descriptores cuánticos.

En particular, en este estudio se hace una distinción entre los descriptores cuánticos que se obtienen directamente de un cálculo de estructura electrónica, y los que se obtienen a partir de un análisis posterior de la densidad electrónica, empleando la teoría de átomos en moléculas. A pesar de que estrictamente la densidad electrónica queda definida a partir de la función de onda (ecuación 3.3) y, por lo tanto, ambos descriptores cabrían en la misma categoría, se decidió hacer la distinción, ya que el análisis tipo átomos en moléculas nos brinda información más selectiva sobre ciertas propiedades asociadas a la densidad electrónica en diferentes regiones de la molécula (describe propiedades locales de la densidad electrónica), lo cual a su vez es importante para la descripción del efecto del disolvente sobre estas propiedades, ya que éste modifica de una manera local a la densidad electrónica.

$$\rho\left(\vec{r}_1,\vec{r}_2,\ldots\vec{r}_N\right) = N\int\cdots\int d\vec{r}_1\, d\vec{r}_1\, \ldots\, d\vec{r}_{N-1}\,\psi\left(\vec{r}_1,\vec{r}_2,\ldots\vec{r}_N\right)\psi^*\left(\vec{r}_1,\vec{r}_2,\ldots\vec{r}_N\right) \tag{3.3}$$

Algo importante a considerar es que todos los descriptores cuánticos se calcularon a partir de las geometrías del estado basal (S0), es decir, estado inicial en el proceso de absorción y el estado final en el proceso de emisión. Además, ninguno de los cálculos de estructura electrónica realizados implementa algún modelo de solvatación implícito o explicito, ya que el efecto del disolvente se incluye en el modelo de solvatación propuesto en la sección 1.1.

Descriptores Cuánticos: Función de onda

En un afán por conocer los descriptores mínimos necesarios para describir el comportamiento de cada propiedad, se decidió subclasificar este tipo de descriptores según su naturaleza física. Un primer grupo son los descriptores que están asociados a las energías orbitales, las cuales llamaremos energías orbitales. Ejemplo de estos son: la energía del último orbital molecular ocupado "highest occupied molecular orbital" (HOMO), la energía del primer orbital molecular desocupado "lowest unoccupied molecular orbital" (LUMO) y la diferencia de energía entre ambos. Una segunda categoría son los descriptores provenientes de la distribución de los electrones alrededor de la molécula, los cuales llamaremos descriptores elec-

trónicos. Ejemplo de estos son las cargas de los átomos dadas por el análisis poblacional de Mulliken y momento dipolar total. Y una última categoría de descriptores asociados a un análisis vibracional (descriptores vibracionales), como por ejemplo la energía de punto cero.

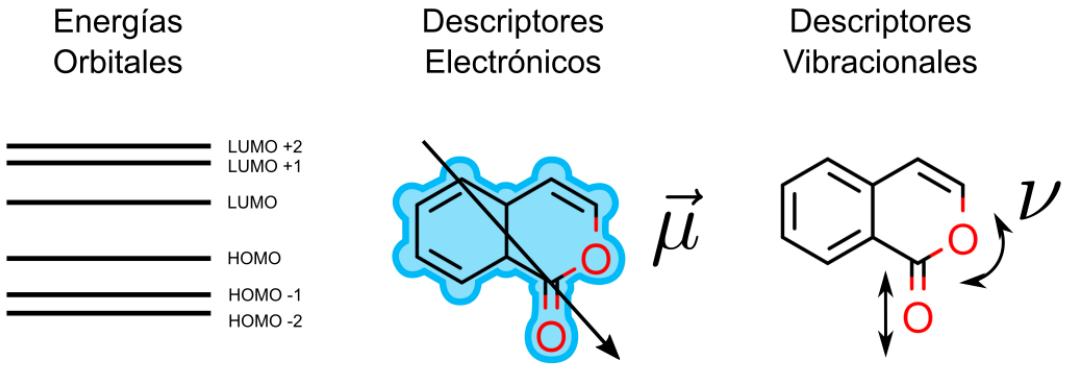


Figura 3.1. Figura que muestra las tres subcategorías diferentes para los descriptores cuánticos: energías orbitales, descriptores electrónicos y descriptores vibracionales.

Los descriptores cuánticos para los sistemas 1 y 2 se calcularon empleando la misma metodología, únicamente cambiando el nivel de teoría del cálculo, de estructura electrónica y de optimización. Para el sistema 1 se empleó el programa AMPAC [57] bajo una aproximación semiempírica AM1 debido a que este método ha mostrado ser eficiente y preciso para predecir las geometrías del estado basal, además de haber mostrado ser muy eficiente para la predicción de momentos dipolares [58]. Con este método se reduce el costo computacional (comparado con DFT), debido a que para este sistema se requiere optimizar la geometría del estado basal para 500 cromóforos, de los cuales una gran parte de ellas son cianinas (30 %) (moléculas computacionalmente costosas de optimizar debido a su gran cantidad de grados de libertad).

Para el sistema 2 se utilizó el programa GAUSSIAN 09 con un nivel de teoría DFT/B3LYP/6-31G(d,p). Para este sistema se optó por usar un nivel de teoría computacionalmente más costoso, ya que éste consta de 30 moléculas, todas cumarinas, las cuales son moléculas generalmente pequeñas y con menos grados de libertad. En ambos casos se realizó una optimización y un análisis del Hessiano para asegurarse de tener la estructura del mínimo de energía y poder obtener información de los modos vibracionales de la molécula.

La selección del funcional (B3LYP) y la base (6-31G(d,p)) se realizó con el objetivo de obtener una metodología eficiente para el cálculo de las geometrías de optimización para moléculas orgánicas [59] y [60]. No importando mucho la precisión del método para la predicción de energías de transición, ya que el algoritmo de aprendizaje automatizado tiene como objetivo compensar los errores que pudiera haber cometido el cálculo de estructura electrónica.

El procedimento que se siguió para el cálculo de este tipo de descriptores fue el siguiente:

- Se partió de la estructura 3D obtenida previamente y se realizaron los correspondientes cálculos de estructura electrónica junto con la optimización, empleando los niveles de teoría ya mencionados.
- Los archivos de salida de dichos cálculos se importaron al programa CODESSA, el cual lee estos archivos y crea una base de datos con los descriptores calculados, exportando la base de datos en un formato (.csv).
- Se importó la base de datos obtenida por CODESSA al programa en Python para el posterior entrenamiento del modelo.

Descriptores cuánticos: Átomos en moléculas

A los descriptores que hacen uso de un análisis de átomos en moléculas para describir propiedades atómicas los llamaremos descriptores QTAIM. Debido a la ventaja proporcionada por el análisis de átomos en moléculas de particionar la densidad electrónica en regiones atómicas, se propuso construir dos tipos de descriptores. Un tipo de descriptores que usan propiedades atómicas, promediando sobre cada tipo de átomo (QTAIM por átomos). Un segundo grupo de descriptores emplean propiedades por grupos (QTAIM por grupos) en donde se suman las propiedades atómicas correspondientes a los átomos pertenecientes a cada grupo (figura (3.2)). De este modo se seleccionan los grupos de átomos que tengan una relevancia química para la molécula. Ejemplo de estos puede ser el grupo carbonilo presente en el núcleo de una cumarina.

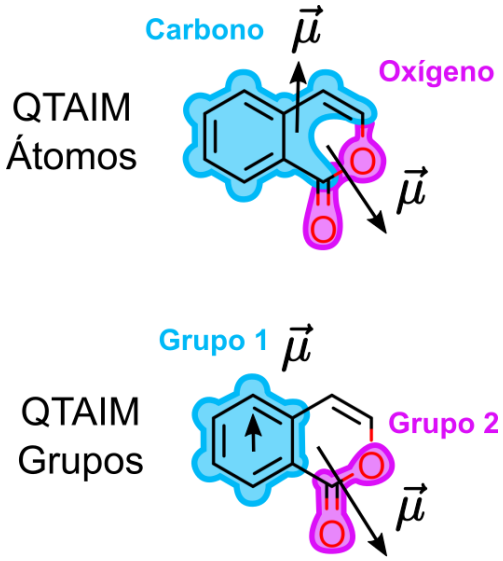


Figura 3.2. Figura en donde se muestran, de una manera esquemática, los dos diferentes tipos de descriptores que hacen uso de un análisis de átomos en moléculas, en donde los descriptores QTAIM por átomos promedian sobre cada tipo de átomo, obteniendo por ejemplo el momento dipolar de los átomos de oxígeno o el momento dipolar de los átomos de carbono. Por otro lado, tenemos a los descriptores QTAIM por grupos, los cuales permiten el cálculo de propiedades promedio de cada grupo. Ejemplo de esto son, el momento dipolar para el grupo carbonilo o un momento dipolar para el anillo heterocíclico.

Específicamente, los momentos dipolares por grupos se calcularon empleando la siguiente ecuación.

$$\mu_\alpha(Grupo) = \sum_A q(A)\, r_\alpha(A) + \mu_\alpha(A) \tag{3.4}$$

donde A son los átomos pertenecientes a cada grupo, $q(A)$ la carga del átomo A, $r_\alpha(A)$ las diferentes componentes de la posición (α = x, y, z) para el átomo A y $\mu_\alpha(A)$ los momentos dipolares intrínsecos del átomo A, estos están asociados únicamente a la distribución de la densidad electrónica en las regiones correspondientes a dicho átomo (A). Los momentos dipolares calculados se realizaron tomando en cuenta el eje como el centro de masa del "núcleo" de la cumarina. Evitando de esta forma algún problema, ya que para todas las cumarinas esa coordenada es constante.

Para implementar este último tipo de descriptores (QTAIM por grupos) es necesario hacer algunas consideraciones. La primera es que sólo se pueden estudiar cromóforos del mismo tipo, ya que se requiere que para cada molécula se conserven los mismos grupos representativos. Por ejemplo, en el caso de las cumarinas, estos pueden ser el grupo carbonilo, el anillo homoaromático o el anillo heteroaromático (Figura 3.3). Por otro lado, si se tratase de una cianina se tendrían distintos grupos característicos. Ejemplo de estos pueden ser la cadena polimetínica o los dos centros de nitrógeno. Para este estudio, los grupos que se emplearon para describir al núcleo de cada cumarina se muestran en la figura 3.3.

Cabe mencionar que esta es la primera ocasión en que se consideran este tipo de descriptores para estudiar propiedades espectroscópicas. Sin embargo, éstas solo se calcularon para el Sistema 2, ya que un análisis de átomos en moléculas como éste no se puede realizar partiendo de un cálculo semiempírico, como fue el caso para el Sistema 1.

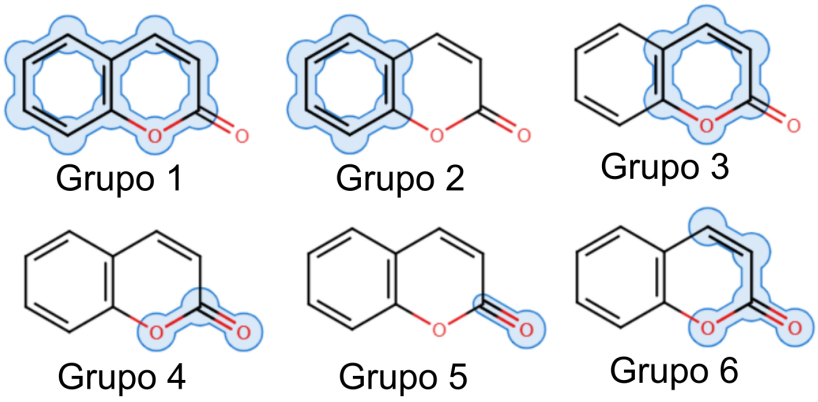


Figura 3.3. Figura en donde se muestran los grupos de la cumarina que se emplearon para el cálculo de los descriptores QTAIM por grupos. Los grupos 1,2 y 3 hacen referencia a los anillos aromáticos, mientras que los grupos 4, 5 y 6 al grupo carbonilo y a sus respectivos átomos adyacentes.

Para el cálculo de los descriptores QTAIM por átomos, se partió de los archivos de salida de GAUSSIAN (formato .wfn). Estos archivos contienen a la función de onda (optimizada) resultante del cálculo de estructura electrónica. Posteriormente, se realizó una rutina para el análisis de la densidad electrónica empleando el programa AIMA11 [61]. Como resultado de este análisis se obtuvieron los archivos (formato .sum) con la información de las diferentes propiedades calculadas para cada región atómica (átomo). Una vez que ya se tuvieron todos los archivos (.sum) para todas las moléculas, estos se exportaron al programa CODESSA, el cual analizó estos archivos y exportó todas las propiedades categorizadas por tipos de átomos. Esta rutina además calculó la suma, el promedio, el valor máximo y el valor mínimo de cada una de estas propiedades. Finalmente, se exportó una tabla (formato .csv) con toda esta información de cada descriptor para cada molécula. Esto incluyendo a los cromóforos y a los solventes.

Para el cálculo de los descriptores QTAIM por grupos se parte de la salida del programa AIMA11 [61] (archivos con formato .sum). Estos archivos se leen empleando un "script" en Python (desarrollado para este trabajo), el cual realiza las correspondientes sumas de cada propiedad para cada grupo. Además, es necesario proporcionarle al "script" la información de los átomos que pertenecen a cada grupo en cada una de las moléculas.

Para saber qué átomos pertenecen a cada grupo es necesario abrir el archivo de la estructura de la molécula empleando algún visualizador e identificar el índice de los átomos pertenecientes a cada grupo. Por último, y como es de esperarse, el "script" exporta una tabla (formato .csv) con la información de cada descriptor, etiquetándolos con respecto a cada grupo.

3.2.3. Descriptores empíricos del disolvente

Por último, se añadieron descriptores empíricos para las moléculas del disolvente. Estos se obtuvieron a partir de tablas de propiedades fisicoquímicas [62], en donde se incluyeron la constante dialéctica ε y el índice de refracción n. Adicionalmente, con estos valores se calcularon descriptores basados en las teorías de solvatocromismo, como lo son la de Lippert-Mataga, y Bilot-Kawski [63] las cuales son funciones del índice de refracción y la constante dieléctrica del disolvente. También se implementaron parámetros empíricos que incluyen efectos de dispersión y electrostáticos, así como parámetros que incluyen los efectos de la interacción soluto-disolvente de Lewis y por puentes de hidrógeno. En el Apéndice B se encuentra la lista completa de los descriptores empíricos implementados.

3.3. Aprendizaje automatizado

Antes de empezar con el preprocesador se requiere hacer unos pasos previos a éste, los cuales constan de preparar la entrada (vector de características) junto con su correspondiente salida. En otras palabras, requiere de juntar en una sola matriz todos los descriptores calculados (figura 3.4).

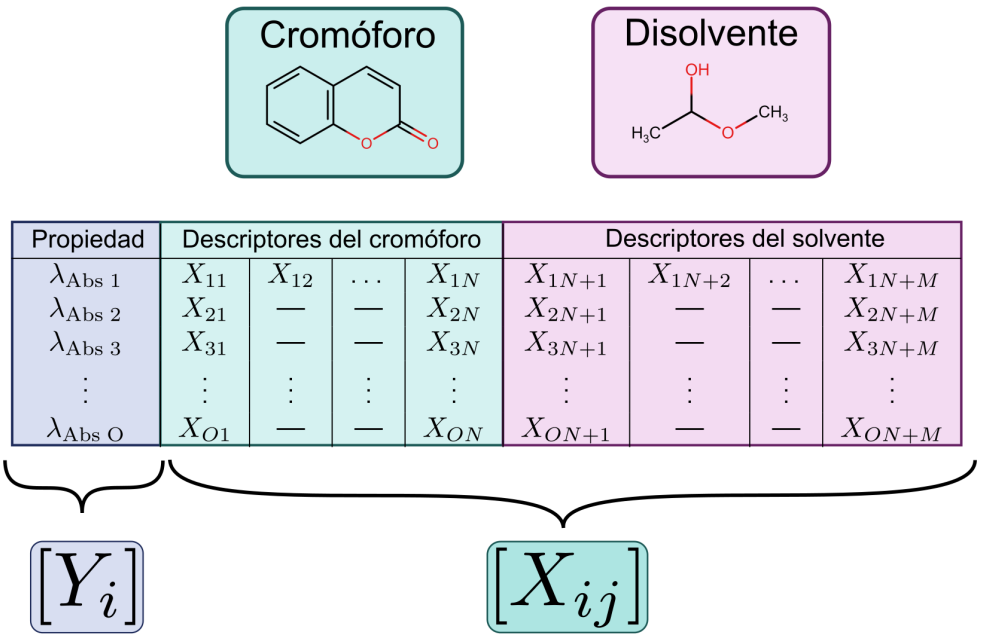


Figura 3.4. Esquema en donde se muestra el conjunto de datos después de añadir los descriptores, en donde la columna azul corresponde a la propiedad a estudiar (salida) y las columnas verdes y rosas a los descriptores empleados (entrada) tanto del cromóforo (verde) como del disolvente (rosa)

Para el preprocesado se realizó un escalado estándar siguiendo la metodología mostrada en la sección 1.3.2. Posteriormente, se imputaron los datos faltantes empleando el método de "vecinos cercanos" implementado por la librería "scikit-learn" [64]. Este método le asigna un valor al dato faltante promediando los valores de los k vecinos más cercanos al dato faltante, los vecinos se encuentran a partir de la distancia de los datos en el plano que se define a partir de la salida $[Y_i]$ y cada una de las diferentes características $[X_j]$.

Además, durante el preprocesado también se realizó una reducción de dimensión empleando el algoritmo de "SelectKBest" proporcionado por la librería "scikit-learn", en donde para cada prueba se fijó el número de descriptores resultantes después de la reducción de dimensión. Este método, en particular, funciona realizando una prueba estadística "f – regression" la cual le asocia un coeficiente de correlación a cada característica $[X_j]$, organizándolos de los que presentan una mayor correlación, a los que tienen una menor correlación para finalizar seleccionando únicamente a las "k" características que obtuvieron un valor de correlación mayor.

Para obtener el valor del número de descriptores mínimos necesarios para tener una buena predicción, se validaron los modelos empleando diferentes valores de "k" en la reducción de dimensión. Posteriormente, se graficaron los valores de las métricas (r^2 y MAE) obtenidas en el proceso de validación. En estas gráficas se buscó una región de saturación, en la que por más que se añadieron más descriptores no se observaron cambios significativos en los valores de las métricas (Figura 3.5-a y -b). Manualmente, se seleccionó el valor de la saturación (número de descriptores óptimo). Una vez seleccionado el valor de la reducción de dimensión óptimo "k", se visualizaron las gráficas de valor predicho contra el valor experimental para asegurarse de tener una distribución correcta de los datos (Figura 3.5-c y -d).

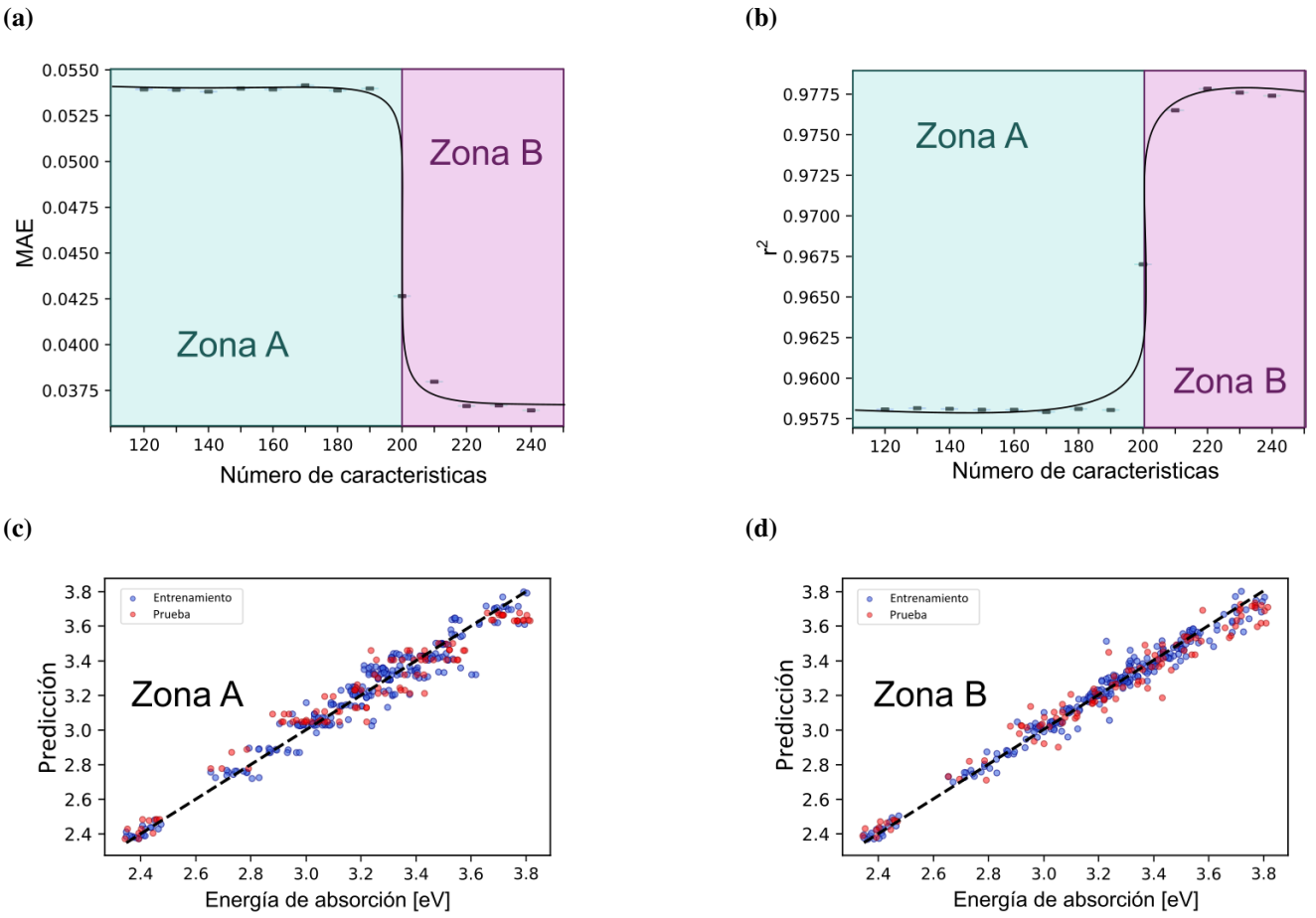


Figura 3.5. Figura en donde se muestra el proceso para la selección del valor óptimo para la reducción de dimensión (número de dimensiones apropiado). (a) y (b) muestran el barrido de las respectivas métricas de evaluación con respecto al número de reducción de dimensión (el barrido se realizó variando el número "k" de descriptores a acotar empleando el algoritmo "SelectKBest"), en donde para ambas métricas (r^2 y MAE) se observa una región de saturación y un punto de inflexión. Lo anterior permite identificar dos regiones, la Zona A (Verde) antes del punto de inflexión y la Zona B (Rosa) después del punto de inflexión. (c) y (d) son las gráficas de valor predicho contra experimental empleando la reducción de dimensión correspondiente a la Zona A y Zona B. Se observa que en la zona A antes del punto de inflexión, la distribución no es homogénea, mientras que para la Zona B se observa una distribución homogénea, por lo que el número mínimo de descriptores para el entrenamiento se encuentra justamente después del punto de inflexión.

Por último, se realizó el entrenamiento, la validación y la prueba del modelo. Para ello se emplean las funciones implementadas por la librería “sk-learn”, las cuales parten de las entradas $[X_{i\,j}]$ (reprocesadas) y salidas $[Y_i]$ obtenidas de los pasos previos. El proceso de entrenamiento, validación y prueba se realiza empleando la metodología explicada en la sección 1.3.2, en donde la validación cruzada se realizó con 5 iteraciones para el Sistema 1 y 10 iteraciones para el Sistema 2.

Resultados y discusiones

Debido a que no se encontraron diferencias significativas (en el comportamiento y la predicción) de los modelos empleados: "GradienteBoostingRegressor" y "RandomForestRegressor" (referirse a la sección 1.3.3 para ver en qué consiste cada uno de estos modelos), se presentarán únicamente los resultados obtenidos para el modelo de "Gradiente Boosting Regressor", mientras que los resultados obtenidos para el modelo de "Random Forest Regressor" se muestran en el Anexo A.

4.1. Resultados para el Sistema 1

Como se mencionó en secciones previas, al no tener una teoría formal que respalde la metodología empleada, es necesario estar seguros de que se tenga el conjunto de datos correcto. Esto incluye estar seguros de que se está haciendo una selección adecuada de cromóforos, así como si se están implementando los descriptores correctos. Con este fin, se realizaron una serie de pruebas en donde se evaluó el desempeño de diferentes subconjuntos de datos. Las dos siguientes subsecciones presentan los resultados obtenidos para una selección de distintos tipos de cromóforos y una selección de distintos tipos de descriptores (Sistema 1).

4.1.1. Comparación entre diferentes tipos de cromóforos

El primer conjunto de pruebas se realizó para evaluar qué tan general es el modelo. Con ello, nos referimos a que si el modelo es capaz de desempeñarse de una manera correcta para todos los tipos de cromóforos. Con este fin, se entrenaron modelos empleando únicamente ciertos tipos de cromóforos (cianinas y bodipys) debido a que estos se encuentran en mayor proporción en la base de datos. El primer subconjunto constó de 156 bodipys diferentes (453 parejas cromóforo -disolvente) y el segundo de 113 cianinas diferentes (507 parejas cromóforo -disolvente). Los resultados de estas pruebas se muestran en la tabla 4.1.

Para la mayoría de las propiedades estudiadas, se observó que los modelos correlacionaron mejor (obtuvieron un valor de r^2 mayor) empleando todos los tipos de cromóforos (cianinas, bodipys, cumarina, rodaminas, etc.), mientras que en las pruebas individuales de cada tipo de cromóforos (cianinas y bodipys) se obtuvieron errores (MAE) menores (figura 4.1). La diferencia obtenida al emplear los diferentes subconjuntos de cromóforos fue particularmente mayor para la predicción de tiempo de vida de fluorescencia, obteniendo un error absoluto medio de 0.382 ns para la prueba que incluye únicamente cianinas, el cual es considerablemente menor que el que se obtuvo para el conjunto con todos los tipos de cromóforos (0.969 ns). Sin embargo, al no obtener un valor razonable de coeficiente de correlación ($r^2 = 0.562$), para las pruebas con puras cianinas, es posible que esa disminución en el error obtenido sea asociado a otros factores. Por ejemplo, el hecho de que las cianinas presentan relativamente tiempos de vida más cortos que los otros cromóforos.

El hecho de que la correlación aumente al implementar todos los tipos de cromóforos (aumentar la variabilidad estructural) sugiere que la variabilidad estructural es un factor clave para la correcta descripción del sistema. Sin embargo, mientras más general sea el modelo (incluya más variedad de cromóforos) el error también será más grande. Lo anterior se debe a que al tener mayor variabilidad estructural también se tiene una mayor variabilidad en las propiedades espectroscópicas. Un ejemplo claro de esto es el hecho de que las cianinas tienen longitudes de onda de absorción mayores que las cumarinas o los bodipys, debido a la alta conjugación que presentan en su cadena polimetínica [65].

Tabla 4.1. Tabla en donde se muestra una comparación del desempeño de los modelos para los diferentes subconjuntos de cromóforos (Sistema 1). Para el entrenamiento se empleó el modelo de "GradientBoostingRegressor" junto con todos los tipos de descriptores calculados (estructurales, cuánticos y empíricos). Para la evaluación se reportan dos métricas de evaluación, el coeficiente de correlación r^2 y el error absoluto medio MAE (descritas más a detalle en la sección 1.3.2), tanto para el conjunto de validación como para el de prueba. Además, se calculó el error relativo a la media (% Error). Para el entrenamiento de los modelos se emplearon todos los tipos de descriptores implementados para este sistema, tanto para el cromóforo (estructurales y cuánticos) como para el disolvente (estructurales, cuánticos y empíricos). En la tercera columna se muestran los resultados de la reducción de dimensión, donde el primer valor corresponde al número de descriptores antes de la reducción de dimensión y el segundo valor corresponde al número de descriptores después de la reducción de dimensión.

Propiedad	Tipo de Cromóforos	Reducción de dimensión		r^2 (Prueba)	r^2 (Validación)	MAE (Prueba)	MAE (Validación)	% Error (Prueba)
E_{abs}	Todos	1820	50	0.925	0.955 ± 0.045	0.124 eV	0.079 ± 0.025 eV	5.02 %
E_{abs}	Bodipys	1292	250	0.891	0.909 ± 0.199	0.073 eV	0.041 ± 0.028 eV	2.96 %
E_{abs}	Cianinas	1678	200	0.945	0.909 ± 0.113	0.064 eV	0.048 ± 0.044 eV	2.59 %
E_{em}	Todos	1820	100	0.870	0.894 ± 0.062	0.104 eV	0.089 ± 0.025 eV	4.68 %
E_{em}	Bodipys	1292	50	0.832	0.920 ± 0.129	0.060 eV	0.039 ± 0.024 eV	2.70 %
E_{em}	Cianinas	1678	150	0.835	0.747 ± 0.330	0.067 eV	0.058 ± 0.028 eV	3.02 %
ΔStokes	Todos	1820	200	0.873	0.873 ± 0.132	0.058 eV	0.052 ± 0.017 eV	21.48 %
ΔStokes	Bodipys	1292	150	0.732	0.779 ± 0.213	0.033 eV	0.024 ± 0.013 eV	12.22 %
ΔStokes	Cianinas	1678	400	0.764	0.489 ± 1.029	0.038 eV	0.030 ± 0.020 eV	14.07 %
τ_{flu}	Todos	1820	300	0.651	0.814 ± 0.113	0.969 ns	0.610 ± 0.205 ns	142.5 %
τ_{flu}	Bodipys	1292	350	0.545	0.663 ± 0.223	0.941 ns	0.729 ± 0.287 ns	138.4 %
τ_{flu}	Cianinas	1678	400	0.562	0.849 ± 0.197	0.382 ns	0.207 ± 0.087 ns	56.18 %

(a) **(b)**

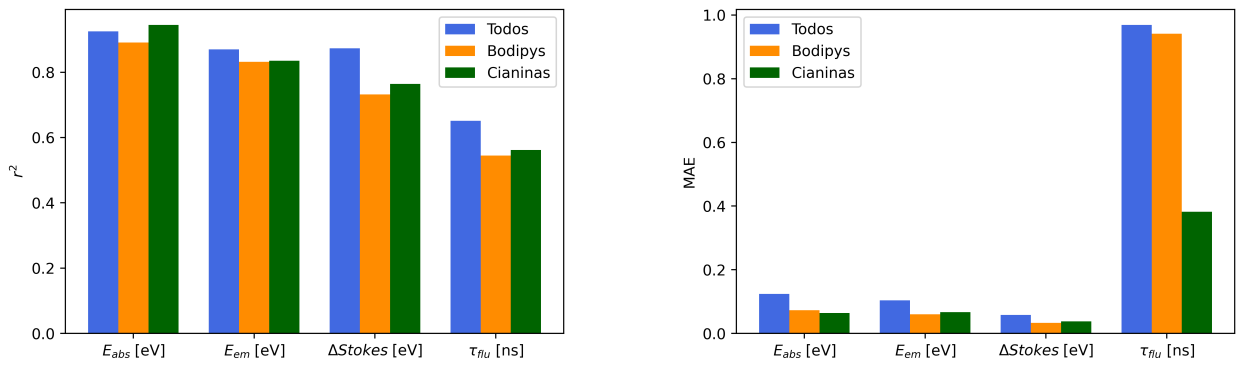


Figura 4.1. Gráficas en donde se muestran los valores obtenidos del coeficiente de correlación r^2 (a) y del error cuadrático medio MAE (b), para las diferentes pruebas realizadas con diferentes subconjuntos de cromóforos (Sistema 1). Las barras de color azul son para todos los tipos de cromóforos (empleando toda la base de datos), las barras amarillas para el subconjunto de bodipys y las verdes para el subconjunto de cianinas.

4.1.2. Comparación entre diferentes tipos de descriptores

Con el objetivo de conocer qué descriptores son los mejores para la predicción de las propiedades estudiadas para el Sistema 1, se evaluó el desempeño de diferentes combinaciones de descriptores. Para ello, se realizó una primera prueba empleando únicamente descriptores estructurales; una segunda, usando únicamente descriptores cuánticos y, por último, una tercera, la cual emplea los dos tipos de descriptores simultáneamente. Los descriptores empíricos del disolvente fueron empleados para los tres casos.

En las pruebas individuales se observó un mejor desempeño de los descriptores cuánticos, sin embargo, esta diferencia es muy pequeña y queda contenida en la incertidumbre obtenida por la validación cruzada. Con esto se puede concluir que el pequeño aumento en la predicción no compensa el costo computacional extra requerido para el cálculo de los descriptores cuánticos. Por otro lado, un resultado interesante es el hecho de que, para todas las propiedades, los modelos se desempeñaron mejor combinando ambos tipos de descriptores. Esto sugiere que el modelo es capaz de sacarle provecho a la información que proporcionan cada uno de los diferentes tipos de descriptores. Por ejemplo, los modelos que predicen la energía de emisión empleando únicamente descriptores estructurales y empleando únicamente descriptores cuánticos, obtuvieron un valor del coeficiente de correlación de 0.846 y 0.824 respectivamente, mientras que el modelo que implementa ambos tipos de descriptores obtuvo un valor de coeficiente de correlación de 0.870 (mayor que el que se obtuvo para los casos individuales).

Un análisis más a fondo de qué descriptores son los más importantes para la predicción de las correspondientes propiedades estudiadas, se presenta en la sección 4.1.4 en donde se muestran las gráficas de la importancia relativa de los diferentes descriptores.

Tabla 4.2. Tabla en donde se muestra una comparación del desempeño de los modelos para los diferentes subconjuntos de descriptores (Sistema 1). Para el entrenamiento se empleó el modelo de "GradientBoostingRegressor" en donde se reportan dos métricas de evaluación, coeficiente de correlación r^2 y el error absoluto medio MAE (descritas más a detalle en la sección 1.3.2), tanto para el conjunto de validación como para el de prueba. Además, se calculó el error relativo a la media (% Error). Para el entrenamiento de los modelos se emplearon descriptores estructurales y empíricos para el disolvente, así como los descriptores del cromóforo que se especifican en las columnas dos y tres. En la cuarta columna se muestran los resultados de la reducción de dimensión, en donde el primer valor corresponde al número de descriptores antes de la reducción y el segundo valor corresponde al número de descriptores después de la reducción de dimensión.

Propiedad	Descriptores estructurales	Descriptores cuánticos	Reducción de dimensión		r^2 (Prueba)	r^2 (Validación)	MAE (Prueba)	MAE (Validación)	% Error (Prueba)
E_{abs}	✓	✓	1820	50	0.925	0.955 ± 0.045	0.124 eV	0.079 ± 0.025 eV	5.02 %
E_{abs}	✓		292	150	0.914	0.943 ± 0.072	0.124 eV	0.086 ± 0.035 eV	5.02 %
E_{abs}		✓	1678	50	0.925	0.955 ± 0.044	0.109 eV	0.083 ± 0.017 eV	4.41 %
E_{em}	✓	✓	1820	100	0.870	0.894 ± 0.062	0.104 eV	0.089 ± 0.025 eV	4.68 %
E_{em}	✓		292	150	0.846	0.926 ± 0.037	0.113 eV	0.073 ± 0.020 eV	5.09 %
E_{em}		✓	1678	100	0.824	0.897 ± 0.046	0.124 eV	0.087 ± 0.010 eV	5.59 %
ΔStokes	✓	✓	1820	200	0.873	0.873 ± 0.132	0.058 eV	0.052 ± 0.017 eV	21.48 %
ΔStokes	✓		292	250	0.786	0.869 ± 0.106	0.093 eV	0.050 ± 0.015 eV	34.44 %
ΔStokes		✓	1678	300	0.818	0.903 ± 0.074	0.069 eV	0.047 ± 0.009 eV	25.56 %
τ_{flu}	✓	✓	1820	300	0.651	0.814 ± 0.113	0.969 ns	0.610 ± 0.205 ns	142.5 %
τ_{flu}	✓		292	150	0.655	0.780 ± 0.129	0.998 ns	0.703 ± 0.153 ns	146.8 %
τ_{flu}		✓	1678	250	0.737	0.795 ± 0.147	0.767 ns	0.650 ± 0.144 ns	112.8 %

(a) **(b)**

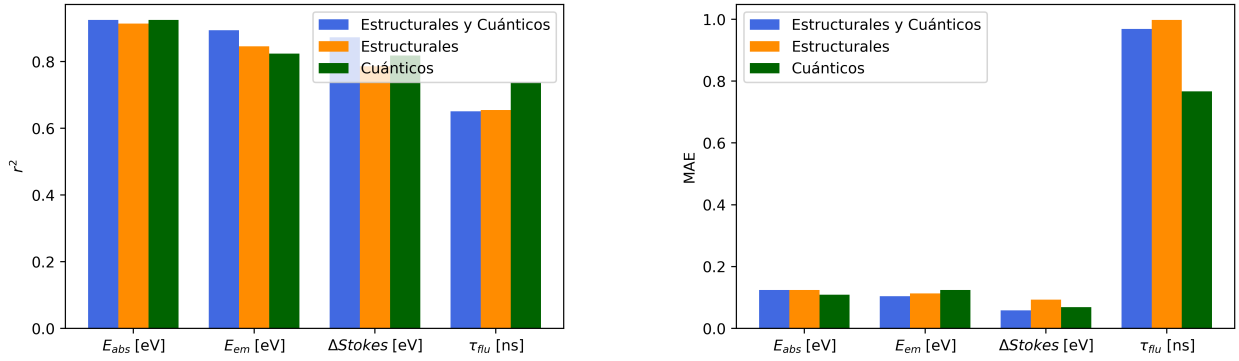


Figura 4.2. Gráfica en donde se muestran los valores obtenidos del coeficiente de correlación r^2 (a) y el error cuadrático medio MAE (b) para las diferentes pruebas realizadas con diferentes subconjuntos de descriptores. Las barras de color azul son para todos los descriptores (estructurales y cuánticos), las de color naranja únicamente para descriptores estructurales y las verdes únicamente para descriptores cuánticos.

4.1.3. Resultados para cada una de las propiedades espectroscópicas

Como se mencionó en la sección 1.3.2, para evaluar si el modelo se desempeñó correctamente, es necesario verificar que las gráficas de valor predicho contra valor experimental estén correctas. Para ello, los datos (puntos en la gráfica) deben de correlacionar y mostrar poca dispersión alrededor la recta ideal, la cual es una línea recta con pendiente igual a uno y ordenada al origen de cero (línea negra punteada).

En esta sección se muestran las gráficas de distribución (figura 4.3) para las diferentes propiedades estudiadas. En donde se puede observar que las distribuciones para las energías de absorción y emisión es relativamente simétrica, comparándolas con las distribuciones para el tiempo de vida de fluorescencia y el desplazamiento de Stokes, las cuales son distribuciones completamente asimétricas. Esto a su vez puede explicar el porqué para estas propiedades no se encontró un modelo capas de predecirlas (esto se explica más a detalle en esta misma sección), ya que una distribución homogénea y simétrica es deseable para un mejor desempeñó del los modelos de aprendizaje automatizado. Además, se presentan las gráficas de valor predicho contra valor experimental, las cuales se obtuvieron empleando todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente, a modo de presentar los mejores modelos obtenidos.

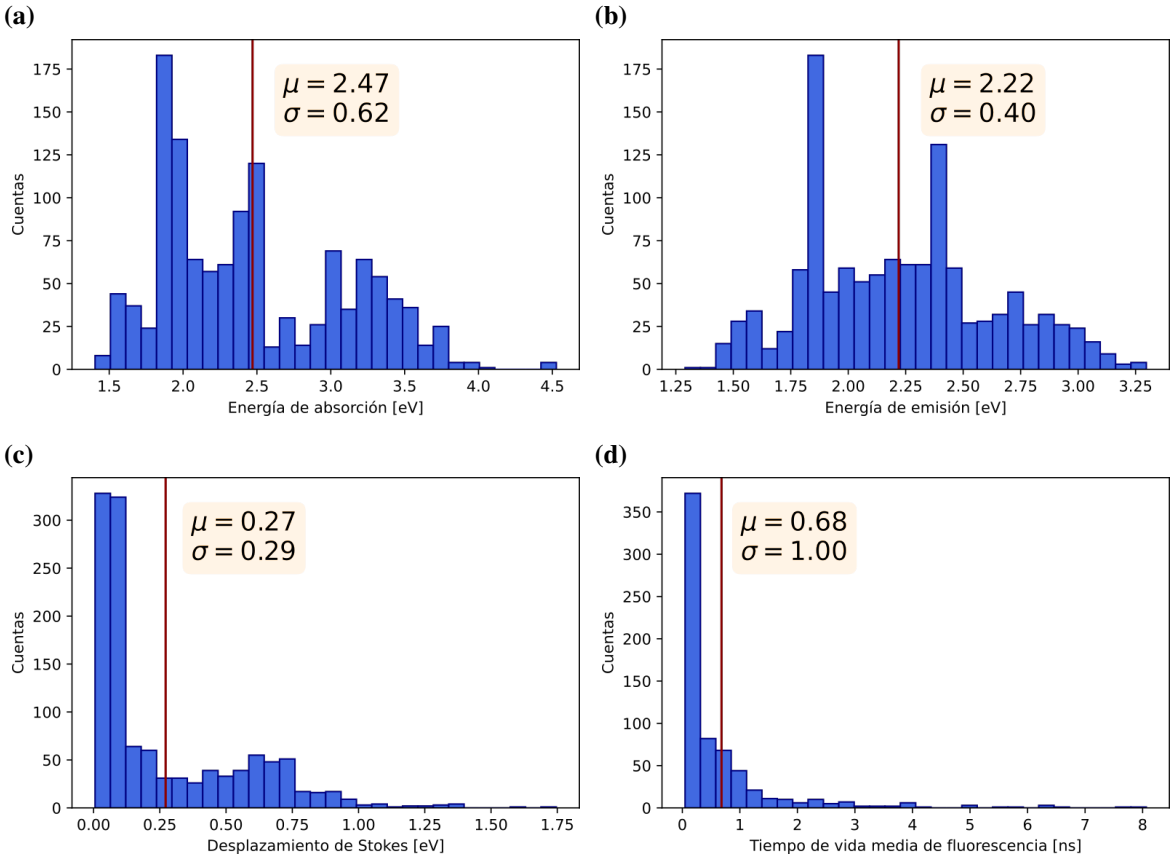
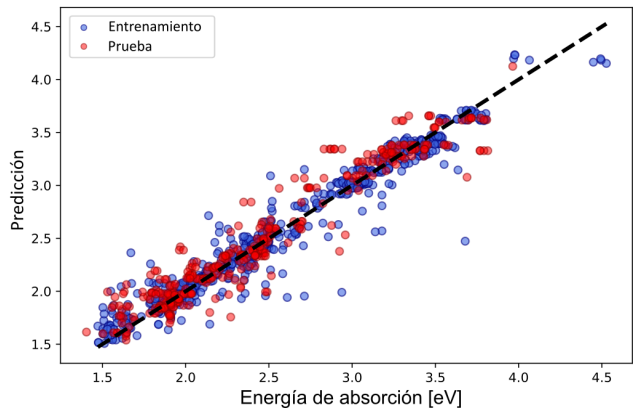


Figura 4.3. Figura en donde se muestra la distribución de datos para las diferentes propiedades estudiadas. Energía de absorción (a), energía de emisión (b), el desplazamiento de Stokes (c) y el tiempo de vida de fluorescencia (d). Además, se muestra el valor medio de la distribución μ (línea roja) junto con la desviación σ.

La gráfica 4.4-(a) muestra los resultados obtenidos para predecir la energía de absorción, en donde se puede ver que los datos predichos correlacionan bien, ajustándose a la recta ideal, además de no presentar errores sistemáticos de algún tipo. Como puede observarse, a pesar de que el error absoluto medio es relativamente pequeño, hay algunos datos que se salen completamente de la tendencia. Considerando una energía de absorción media de 2.5 eV (figura 4.3-(a)), se obtiene un error relativo a la media de 4.8 %. Como ya se menciona en la sección 1.3.2, este error relativo se calcula a partir del valor medio de la distribución μ (figura 4.3) y el error absoluto medio con el fin de obtener una medida de error porcentual. Por otro lado, la gráfica 4.4-(b) muestra los resultados obtenidos para la predicción de la energía de emisión. En ésta se observa que los datos predichos correlacionan y ajustan adecuadamente a la recta ideal. Nuevamente, para este caso se observaron algunos puntos que se salen completamente de la tendencia. Además de que en general estos resultados presentan una mayor dispersión (alrededor de la línea punteada) con respecto al modelo que predice la energía de absorción. Considerando una energía de emisión media de 2.2 eV (figura 4.3-(b)), se obtuvo un error relativo a la media de 4.8 %.

(a)

	r^2	Error absoluto medio	Error cuadrático medio
Validación	0.954 +/- 0.045	0.079 +/- 0.026	0.13 +/- 0.055
Prueba	0.92	0.121	0.173



(b)

	r^2	Error absoluto medio	Error cuadrático medio
Validación	0.921 +/- 0.052	0.075 +/- 0.019	0.113 +/- 0.039
Prueba	0.859	0.105	0.152

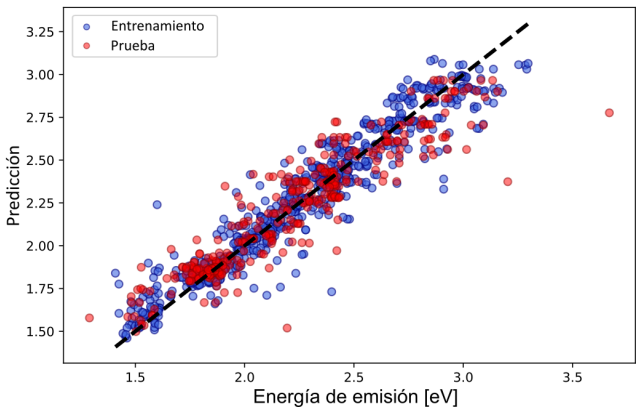


Figura 4.4. Gráficas de valor predicho contra el valor experimental para la energía de absorción (a) y emisión (b) (Sistema 1). Para ello se empleó el modelo de "GradientBoostingRegressor" junto con todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente. Previo a entrenar el modelo, se realizó una reducción de dimensión, teniendo un total de 50 (a) y 500 (b) descriptores después de la reducción de dimensión. En la parte superior de la figura se incluye una tabla con los resultados obtenidos para las tres métricas: coeficiente de correlación (r^2), error absoluto medio y error cuadrático medio, tanto para la validación como para la prueba. En la gráfica se muestra el valor predicho contra el valor experimental para el conjunto de datos de entrenamiento (círculos azules) y para el conjunto de datos de prueba (círculos rojos).

Esto quiere decir, que empleando esta metodología es posible predecir los valores de energías de las transiciones electrónicas S0-S1 en cromóforos, con una precisión del orden de 0.1 eV. Este resultado es considerablemente mejor que el que se obtiene con las metodologías alternativas como lo es TD-DFT, las cuales presentan error del orden de 0.2 eV (en estudios de más de 500 cromóforos) [35] después de haber realizado un ajuste lineal. Además, esta metodología es capaz de predecir no sólo la energía de la transición S0-S1 (energía de absorción), sino también la energía de la transición S1-S0 (energía de

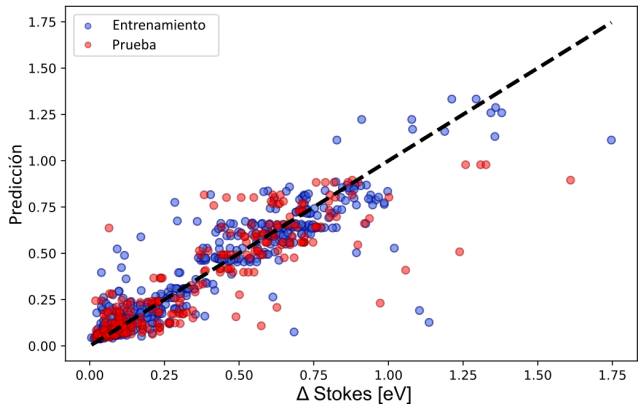
emisión) con la misma precisión (0.1 eV). Este resultado es particularmente importante debido a que la energía de esta transición es difícil de calcular empleando metodologías como TD-DFT, ya que se requiere de hacer una optimización de la geometría del estado excitado S1, para incluir el efecto de la relajación vibracional del estado excitado S1. La grafica 4.5-(a) muestra los resultados obtenidos para predecir el desplazamiento de Stokes, si bien se observa que existe una correlación con respecto a la recta ideal, este modelo no se desempeñó de manera la adecuada, debido a que hay varias regiones en donde obtuvo el mismo valor en el eje de las ordenadas para diferentes valores en el eje de las abscisas (se observa una gráfica escalonada). Además de esto, también se observan una gran cantidad de valores completamente fuera de la tendencia. Si se considera un desplazamiento de Stokes medio de 0.27 eV (figura 4.3-(c)), se obtiene un error relativo a la media de 27.0 %.

Un factor muy importante a considerar, y una posible causa de porque este modelo no es capaz de predecir el desplazamiento de Stokes, es el hecho de que la distribución de datos (figura 4.3-(c)) no es homogénea. Específicamente, la distribución tiene en un intervalo de 0 a 0.25 eV el 60 % de los datos, mientras que el resto de datos se encuentran distribuidos de 0.25 hasta 1.5 eV. La distribución es de esa forma debido a que los diferentes tipos de cromóforos presentan diferentes mecanismos de estabilización de los estados excitados, por lo tanto, presentan corrimientos de Stokes muy diferentes. Por ejemplo, las cumarinas casi no presentan cambios geométricos en su "núcleo" cuando estas pasan del estado base al estado excitado, cosa que no sucede con otros tipos de cromóforos. Por ejemplo, algunos cromóforos pueden presentar fenómenos de transferencia de protones e incluso foto-isomerización, lo cual apunta al valor de crear bases de datos con el mismo tipo de cromóforos. Como se verá más adelante, el Sistema 2 no presenta este problema, ya que en éste se conserva el mismo tipo de cromóforo (cumarinas)

Por último, la gráfica 4.5-(b) muestra los resultados obtenidos para predecir el tiempo de vida de fluorescencia τ_{flu}. Nuevamente, este modelo presenta correlación con la recta ideal (línea negra punteada). Sin embargo, éste también presenta una gran cantidad de valores completamente fuera de la tendencia. Considerando un tiempo de vida de fluorescencia medio de 0.68 ns (figura 4.3-(d)), se obtuvo un error relativo a la media de 118.6 %. Como se puede ver de este último resultado, no se logró encontrar un modelo capaz de predecir el valor del tiempo de vida de fluorescencia con precisión. Sin embargo, y debido al hecho de que no existe alguna teoría formal (de primeros principios) que lo pueda calcular de manera precisa, el simple hecho de que al menos se observe alguna correlación es un resultado prometedor, ya que esto indica que se va por el camino correcto para la predicción de esta propiedad. Es posible que otro tipo de métodos más sofisticados (redes neuronales) o modelos que emplean otro tipo de descriptores logren predecir esta propiedad con una precisión más apropiada.

(a)

	r^2	Error absoluto medio	Error cuadrático medio
Validación	0.892 +/- 0.093	0.048 +/- 0.01	0.092 +/- 0.041
Prueba	0.807	0.073	0.129



(b)

	r^2	Error absoluto medio	Error cuadrático medio
Validación	0.789 +/- 0.133	0.648 +/- 0.116	1.004 +/- 0.223
Prueba	0.719	0.807	1.166

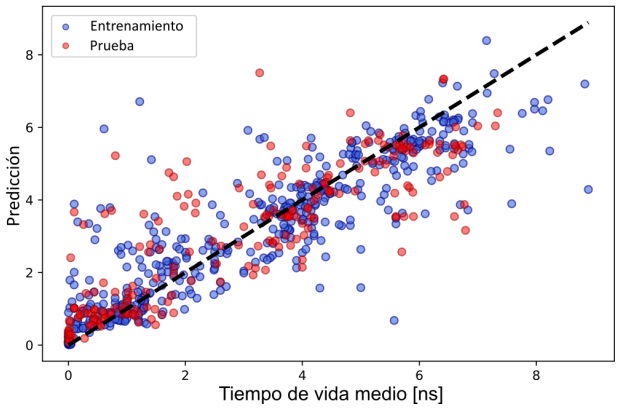


Figura 4.5. Gráficas de valor predicho contra el valor experimental para desplazamiento de Stokes (a) y del tiempo de vida de fluorescencia (b) (Sistema 1). Para ello se empleó el modelo de "GradientBoostingRegressor" junto con todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente. Previo a entrenar el modelo, se realizó una reducción de dimensión, teniendo un total de 500 (a) y 300 (b) descriptores después de la reducción de dimensión. En la parte superior de la figura se incluye una tabla con los resultados obtenidos para las tres métricas: coeficiente de correlación (r^2), error absoluto medio y error cuadrático medio, tanto para la validación como para la prueba. En la gráfica se muestra el valor predicho contra el valor experimental para el conjunto de datos de entrenamiento (círculos azules) y para el conjunto de datos de prueba (círculos rojos).

4.1.4. Importancia de las características

La importancia relativa de las características (descriptores) para predecir la energía de absorción, de emisión y el tiempo de vida de fluorescencia se muestran en las figuras 4.6, y 4.7. En todas estas gráficas se observa un mayor predominio de los descriptores estructurales sobre los descriptores cuánticos. Si bien este hecho sugiere que los descriptores cuánticos no ofrecen mayores ventajas sobre los descriptores estructurales, resultados posteriores del Sistema 2 mostraron (ver más adelante) que el problema está en el empleo de cálculos semiempíricos y no en los descriptores cuánticos por sí mismos.

Las energías orbitales toman una clara y esperada relevancia principalmente para predecir la energía de absorción y de emisión. Siendo la diferencia entre la energía del orbital HOMO y LUMO junto con las energías de estos orbitales las más importantes. Además, se observó que los descriptores del disolvente (incluyendo los empíricos) no figuran en los descriptores más importantes, reafirmando el hecho de que el efecto del disolvente sobre la energía de absorción es considerablemente menor que el dado por la estructura del cromóforo. No se realizó un análisis más detallado de los descriptores estructurales debido a que estos carecen de sentido físico directo, debido a que este tipo de descriptores se construyeron pensado en usarse más con un objetivo predictivo que con un objetivo descriptivo.

Si bien no se obtuvo un modelo capaz de predecir el tiempo de vida de fluorescencia, sí se obtuvo un modelo que correlacionó con esta propiedad, en donde los descriptores más importantes fueron predominantemente estructurales. Además, se observó que las energías orbitales no figuran para la predicción del tiempo de vida de fluorescencia, siendo esto de esperarse, ya que este no depende directamente de la energía de los estados electrónicos, si no más bien este se define a través de una competencia cinética entres diferentes estados electrónicos y vibracionales de la molécula (como ya se mencionó en la sección 1.1). Además, el hecho de que los descriptores del disolvente no figuren en la predicción del tiempo de vida es muy antiintuitivo, ya que claramente el disolvente puede desempeñar un papel importante para la descripción de la dinámica de los estados excitados. Esto sugiere la posibilidad del que Sistema 1 no incluya de una manera acertada el efecto del disolvente sobre las propiedades espectroscópicas. Debido a que no se obtuvieron resultados prometedores en la predicción del desplazamiento de Stokes para el Sistema 1, se decidió dejar el análisis de la importancia de las características para el Sistema 2, ya que éste presenta resultados significativamente más confiables.

Una de las razones de emplear los modelos basados en árboles de decisión es que estos nos proporcionan una forma de analizar qué características son las más importantes. Esto con la idea de relacionar estas características (descriptores) con propiedades físicas del cromóforo y de este modo tener herramientas de diseño para nuevos cromóforos. Es por esta razón que se decidió hacer uso de los descriptores cuánticos, ya que estos se relacionan más directamente con alguna propiedad física de la molécula (por ejemplo el momento dipolar), hecho que no sucede con los descriptores estructurales, debido a que estos carecen de sentido físico directo. Por ejemplo, los índices de conectividad, Kappa, o MoRSE (ver Apéndice B).

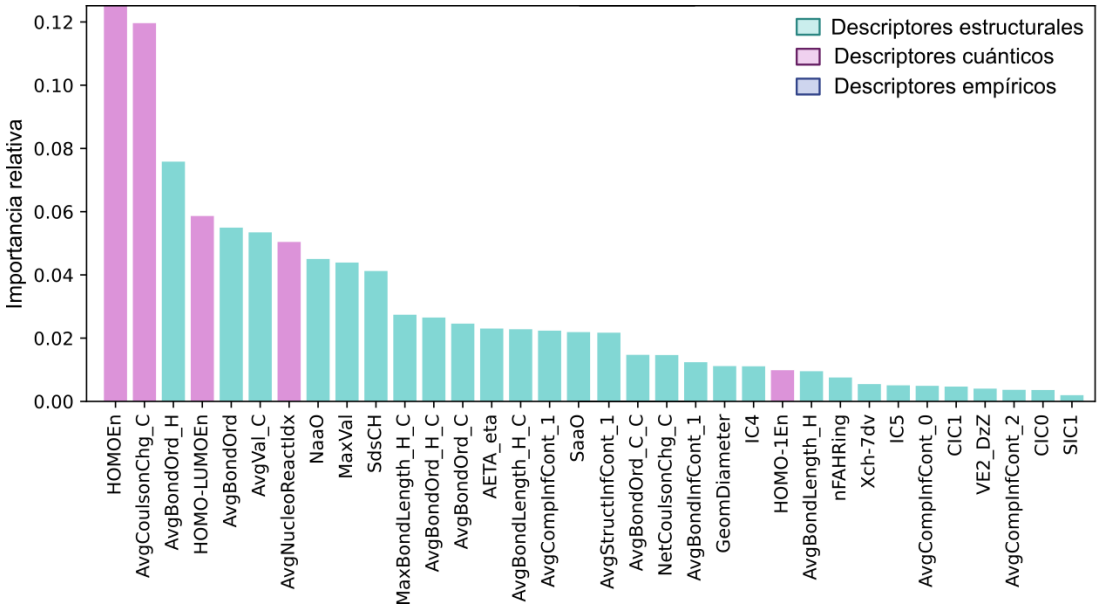
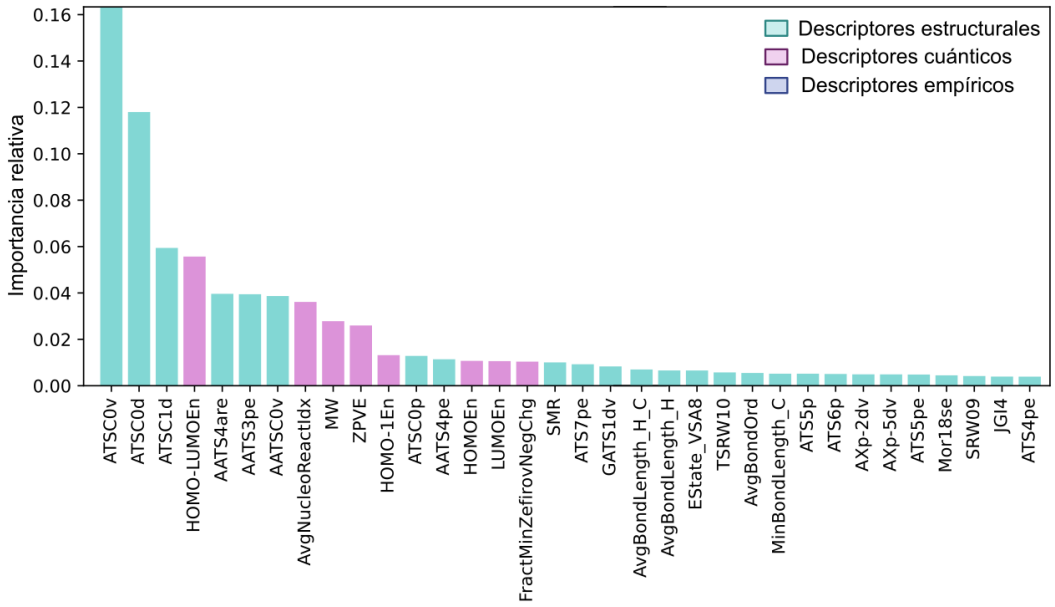


Figura 4.6. Figura en donde se muestra la importancia relativa de los 34 descriptores más importantes para predecir la energía de absorción. Para el entrenamiento se empleó el modelo de "GradientBoostingRegressor" con una reducción de dimensión a 50 descriptores. En donde las barras de color verde corresponden a descriptores estructurales, las barras color rosa a descriptores cuánticos y las barras color azul a descriptores empíricos. Además, un asterisco (*) sobre las barras indica que éstas corresponden a un descriptor del disolvente.

(a)



(b)

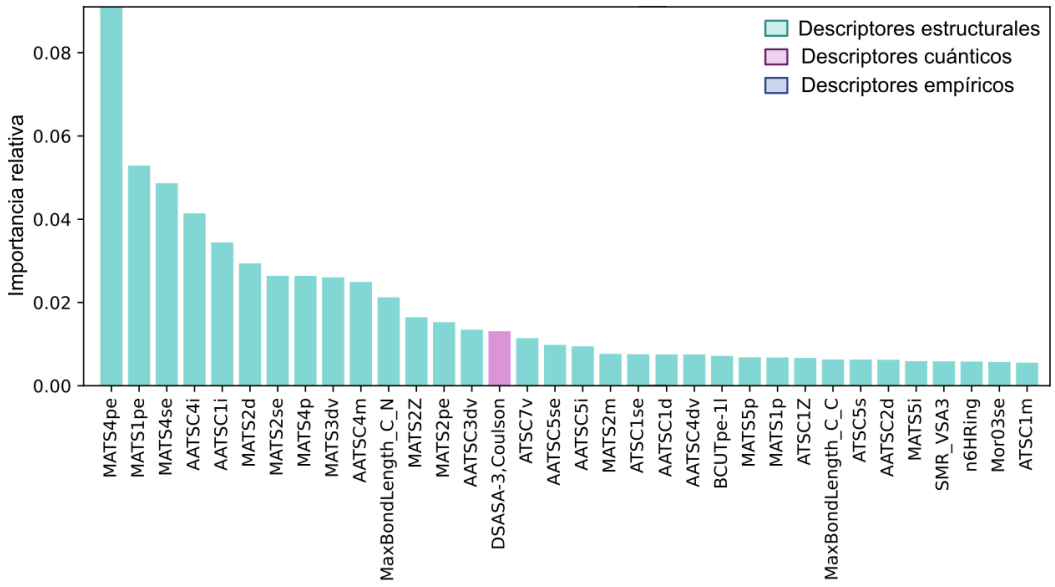


Figura 4.7. Figura en donde se muestra la importancia relativa de los 34 descriptores más importantes para predecir la energía de emisión (a) y el tiempo de vida de fluorescencia (b). Para el entrenamiento se empleó el modelo de "GradientBoostingRegressor" con una reducción de dimensión a 500 descriptores para (a) y 300 descriptores para (b). En donde las barras de color verde corresponden a descriptores estructurales, las barras color rosa a descriptores cuánticos y las barras color azul a descriptores empíricos. Además, un asterisco (*) sobre las barras indica que éstas corresponden a un descriptor del disolvente.

4.2. Resultados para el Sistema 2

4.2.1. Comparación entre diferentes tipos de descriptores

El hecho de que el Sistema 2 sea menos general, es decir que sólo incluya un tipo de cromóforo y además que estos sean cumarinas (cromóforos relativamente pequeños), nos permitió realizar cálculos de estructura electrónica empleando DFT. Lo anterior además nos permitió calcular un nuevo tipo de descriptores, los descriptores QTAIM por tipo de átomos y QTAIM por grupos. Estos últimos prometen capturar mejor el efecto de la relajación vibracional al incluir de una forma más directa el efecto de los sustituyentes y el efecto de la interacción del cromóforo con el disolvente. Con lo anterior en mente y con el objetivo de contestar la pregunta sobre ¿Cuáles descriptores son los más importantes?, y ¿cuáles son indispensables para la predicción de las propiedades espectroscópicas?, se realizó una serie de pruebas en donde se entrenó a los modelos frente a diferentes conjuntos de estos descriptores, estas pruebas se muestran en las tablas 4.3, 4.4 y 4.5. A continuación, se mencionarán las filas a las que corresponden cada una de las pruebas, así como el objetivo de las mismas.

El primer par de pruebas (fila 1 y 2) tienen como objetivo ver si efectivamente los descriptores QTAIM por grupo son mejores que los descriptores QTAIM por tipo de átomos. Para ello se emplearon todos los descriptores previamente calculados (estructurales, orbitales moleculares, densidad electrónica, vibracionales y empíricos) y únicamente se empleó un diferente tipo de descriptores QTAIM. El segundo par de pruebas (fila 3 y 4) tienen como objetivo ver el poder predictivo de los descriptores estructurales, los cuales del estudio anterior mostraron ser muy eficientes al predecir muy bien con un costo computacional muy bajo. Para ello se utilizaron únicamente descriptores estructurales, con y sin descriptores empíricos.

El tercer conjunto de pruebas (fila 4 a 7), tiene como objetivo ver el desempeño de los descriptores cuánticos en diferentes combinaciones de estos. Finalmente, las últimas dos pruebas (filas 10 y 11) tienen como objetivo mostrar si los descriptores QTAIM por grupos son capaces de predecir las propiedades espectroscópicas sin necesidad de emplear algún otro tipo de descriptor, a excepción de las energías orbitales, las cuales son claramente importantes para la predicción de la energía de absorción y de emisión.

Energía de Absorción

Los resultados para la predicción de la energía de absorción se muestran en la tabla 4.3. En dicha tabla se observa que la mayor predictibilidad se obtuvo empleando todos los tipos de descriptores ($r^2 = 0.947$ y MAE $= 0.059$eV). Además, no se observó una diferencia significativa entre emplear los descriptores QTAIM por grupos o por tipos de átomos (fila 1 y 2). El segundo par de pruebas (fila 3 y 4) mostró nuevamente que, emplear los descriptores estructurales junto con los descriptores empíricos es más que suficiente para predecir la energía de absorción con el menor costo computacional posible ($r^2 = 0.932$ y MAE $= 0.074$eV).

El mejor resultado obtenido empleando únicamente descriptores cuánticos lo podemos observar en la fila 8 ($r^2 = 0.935$ y MAE $= 0.068$eV), en donde se emplearon los descriptores QTAIM por grupos junto con los otros descriptores cuánticos. Este resultado es ligeramente mayor que el obtenido empleando descriptores estructurales y empíricos. Sin embargo, el costo computacional para el cálculo de estos descriptores es considerablemente mayor para el pequeño aumento en la predicción. Por último (fila 10 y 11), se mostró que empleando únicamente las energías orbitales y los descriptores QTAIM por grupos es más que suficientes para tener una buena predictibilidad de la energía de absorción ($r^2 = 0.929$ y MAE $= 0.064$eV). Además, el hecho de que se haya logrado obtener una muy buena predicción ($r^2 = 0.920$ y MAE $= 0.071$) con únicamente 40 descriptores (fila 10) muestra que efectivamente este tipo de descriptores incluye de una manera más compacta los efectos de las sustituciones y el efecto del disolvente sobre la energía de absorción.

Tabla 4.3. Tabla en donde se muestra una comparación del desempeño del modelo "GradientBoostingRegressor" para predecir la energía de absorción (E_{abs}) empleando diferentes descriptores tanto para el cromóforo como para el disolvente. Para la evaluación se reportan dos métricas, coeficiente de correlación r^2 y el error absoluto medio MAE (descritas más a detalle en la sección 1.3.2) para el conjunto de prueba. Para el entrenamiento de los modelos se emplearon el mismo tipo de descriptores tanto para el cromóforo como para el disolvente, a excepción de los descriptores empíricos, los cuales sólo están definidos para el disolvente. En la novena columna se muestran los resultados de la reducción de dimensión, donde el primer valor corresponde al número de descriptores antes de la reducción y el segundo valor corresponde al número de descriptores después de la reducción de dimensión.

	Estructurales	Orbitales moleculares	Densidad electrónica	Vibracionales	QTAIM Átomos	QTAIM Grupos	Empíricos (Disolvente)	Reduccion de Dimensión		r^2 (Prueba)	MAE [eV] (Prueba)
1	✓	✓	✓	✓	✓		✓	622	190	0.947	0.059
2	✓	✓	✓	✓		✓	✓	771	250	0.959	0.061
3	✓							368	190	0.873	0.079
4	✓						✓	378	195	0.932	0.074
5		✓	✓	✓				158	100	0.902	0.071
6					✓			96	80	0.856	0.088
7		✓	✓	✓	✓			249	160	0.886	0.079
8		✓	✓	✓		✓		340	100	0.935	0.068
9						✓		167	40	0.897	0.094
10		✓				✓		172	40	0.920	0.071
11		✓				✓		172	100	0.929	0.064

Algo importante a considerar, es el hecho de que si bien los descriptores estructurales y cuánticos contienen la información necesaria para poder predecir las correspondientes propiedades (es por ello que en las pruebas individuales ambos correlacionan bien) sigue habiendo información extra que no está contenida en un tipo de descriptor, pero sí en el otro. Lo anterior explica el por qué al combinar ambos tipos de descriptores se tiene la mejor predictibilidad posible. En otras palabras, el modelo es capaz de sacarle provecho a cada uno de los tipos de descriptores, capturando la información que está contenida en un tipo de descriptores, pero no en el otro.

Energía de emisión

Los resultados para la predicción de la energía de absorción se muestran en la tabla 4.4. En dicha tabla se observa que la mayor predictibilidad se obtuvo empleando todos los tipos de descriptores ($r^2 = 0.922$ y MAE $= 0.064$eV). Para esta propiedad sí se observó una diferencia considerable al usar los descriptores QTAIM por grupos contra usar los descriptores QTAIM por tipos de átomos (fila 1 y 2). El segundo par de pruebas (fila 3 y 4) mostró nuevamente que, emplear los descriptores estructurales junto con los descriptores empíricos es más que suficiente para predecir la energía de absorción con el menor costo computacional posible ($r^2 = 0.919$ y MAE $= 0.065$eV).

El mejor resultado obtenido empleando únicamente descriptores cuánticos lo podemos observar en la fila 8 ($r^2 = 0.842$ y MAE $= 0.084$eV), en donde se emplearon los descriptores QTAIM por grupos junto con los otros descriptores cuánticos. Este resultado es ligeramente menor que el obtenido empleando descriptores estructurales y empíricos, mostrando una vez más que el cálculo de estos descriptores es innecesario, si el objetivo es únicamente predictivo. Por último, se mostró (fila 11) que empleando únicamente descriptores QTAIM por grupos es más que suficiente para tener una buena predictibilidad con la menor cantidad de descriptores posible, no necesitando el uso de las energías orbitales, como en el caso de predecir la energía de absorción ($r^2 = 0.880$ y MAE $= 0.079$). El hecho de que esta prueba se haya realizado empleando únicamente 35 descriptores muestra, nuevamente, que los descriptores QTAIM por grupos incluye de una manera más compacta los efectos de las sustituciones y el efecto del disolvente sobre la energía de emisión.

Tabla 4.4. Tabla en donde se muestra una comparación del desempeño del modelo "GradientBoostingRegressor" para predecir la energía de emisión (E_{em}) empleando diferentes descriptores tanto para el cromóforo como para el disolvente. Para la evaluación se reportan dos métricas, coeficiente de correlación r^2 y el error absoluto medio MAE (descritas más a detalle en la sección 1.3.2) para el conjunto de prueba. Para el entrenamiento de los modelos se emplearon el mismo tipo de descriptores tanto para el cromóforo como para el disolvente, a excepción de los descriptores empíricos, los cuales sólo están definidos para el disolvente. En la novena columna se muestran los resultados de la reducción de dimensión. Donde el primer valor corresponde al número de descriptores antes de la reducción y el segundo valor corresponde al número de descriptores después de la reducción de dimensión.

	Estructurales	Orbitales moleculares	Densidad electrónica	Vibracionales	QTAIM Átomos	QTAIM Grupos	Empíricos (Disolvente)	Reduccion de Dimensión		r^2 (Prueba)	MAE [eV] (Prueba)
1	✓	✓	✓	✓	✓		✓	622	120	0.885	0.077
2	✓	✓	✓	✓		✓	✓	771	200	0.922	0.064
3	✓							368	180	0.892	0.071
4	✓						✓	378	175	0.919	0.065
5		✓	✓	✓				158	80	0.823	0.089
6					✓			96	80	0.770	0.107
7		✓	✓	✓	✓			249	160	0.834	0.089
8		✓	✓	✓		✓		340	100	0.842	0.084
9						✓		167	100	0.840	0.094
10		✓				✓		172	100	0.840	0.094
11						✓		167	35	0.880	0.079

Además, el hecho de que los descriptores QTAIM por grupos presenten una mejoría considerable al predecir la energía de emisiones (comparado con predecir la energía de absorción) es consistente con el hecho de que la energía de emisión depende más los efectos de las sustituciones y de la interacción con el disolvente. Esto debido a que la energía de emisión es la energía del estado excitado después de una relajación vibracional, la cual es mediada por los modos vibracionales del cromóforo (modificados por los sustituyentes) y la interacción con el disolvente.

Desplazamiento de Stokes

Los resultados para la predicción del desplazamiento de Stokes se muestran en la tabla 4.5, en donde se observa que la mayor predictibilidad no se obtuvo empleando todos los tipos de descriptores, sino empleando los descriptores QTAIM por grupos (fila 9) ($r^2 = 0.949$ y MAE $= 0.045$eV). Este hecho es muy importante, ya que el desplazamiento de Stokes por definición es la diferencia de la energía de absorción menos la energía de emisión, teniendo como resultado la fracción de energía asociada a la dinámica del cromóforo en el estado excitado. Esta dinámica se rige principalmente por la relajación vibracional y la interacción con otros estados electrónicos. De modo que si se conserva el mismo tipo de cromóforo (como es el caso para el Sistema 2) el cambio en esta dinámica solamente puede estar asociada a dos factores: el efecto de los sustituyentes sobre el cromóforo y la interacción con su entorno de solvatación. Es aquí donde toma justificación el uso de estos nuevos descriptores, los cuales se construyeron especialmente para representar estas interacciones.

Tabla 4.5. Tabla en donde se muestra una comparación del desempeño del modelo "GradientBoostingRegressor" para predecir el desplazamiento de Stokes empleando diferentes descriptores tanto para el cromóforo como para el disolvente. Para la evaluación se reportan dos métricas, coeficiente de correlación r^2 y el error absoluto medio MAE (descritas más a detalle en la sección 1.3.2) para el conjunto de prueba. Para el entrenamiento de los modelos se emplearon el mismo tipo de descriptores tanto para el cromóforo como para el disolvente, a excepción de los descriptores empíricos, los cuales sólo están definidos para el disolvente. En la novena columna se muestran los resultados de la reducción de dimensión. En donde el primer valor corresponde al número de descriptores antes de la reducción y el segundo valor corresponde al número de descriptores después de la reducción de dimensión.

	Estructurales	Orbitales moleculares	Densidad electrónica	Vibracionales	QTAIM Átomos	QTAIM Grupos	Empíricos (Disolvente)	Reduccion de Dimensión		r^2 (Prueba)	MAE [eV] (Prueba)
1	✓	✓	✓	✓	✓		✓	622	160	0.906	0.048
2	✓	✓	✓	✓		✓	✓	771	250	0.943	0.042
3	✓							368	165	0.811	0.061
4	✓						✓	378	100	0.926	0.042
5		✓	✓	✓				158	55	0.798	0.069
6					✓			96	55	0.883	0.054
7		✓	✓	✓	✓			249	150	0.874	0.063
8		✓	✓	✓		✓		340	200	0.890	0.049
9						✓		167	120	0.949	0.045
10		✓				✓		172	100	0.850	0.062

Una vez más, se mostró que emplear los descriptores estructurales junto con los descriptores empíricos (fila 3 y 4) es más que suficiente para predecir el desplazamiento de Stokes con el menor costo computacional posible ($r^2 = 0.929$ y MAE $= 0.042$eV). Además, los descriptores cuánticos de función de onda (orbitales moleculares, densidad electrónica y vibracionales) mostraron no ser tan efectivos ($r^2 = 0.798$ y MAE $= 0.069$eV) para la predicción del desplazamiento de Stokes, esto si se comparan con los descriptores estructurales o los descriptores QTAIM por grupos, los cuales se desempeñaron mucho mejor.

Resumen tipo de descriptores

El resumen de los resultados obtenidos para las diferentes pruebas específicas que evalúan el desempeño de los modelos, empleando diferentes tipos de descriptores, se muestran en las figuras 4.8 y 4.9. Estas gráficas muestran tanto el valor obtenido para el coeficiente de correlación (r^2) y para el error absoluto medio (MAE), en donde se observa una pequeña mejoría al emplear únicamente descriptores cuánticos sobre emplear únicamente descriptores estructurales. Además, esta pequeña mejoría se ve más notoria para la predicción del desplazamiento de Stokes. Estos resultados sugieren que (como también se observó para Sistema 1) el emplear todos los tipos de descriptores (estructurales, cuánticos y empíricos) sigue siendo la mejor opción si se desea tener la mayor predictibilidad. Sin embargo, y como se mencionó previamente, si lo que se busca es eficiencia, es más recomendable emplear los descriptores estructurales.

Por último, las gráficas presentes en la figura 4.9 muestran que el emplear los descriptores QTAIM por grupos sí presenta una mejoría cuantitativa contra emplear los descriptores QTAIM por tipo de átomos. Siendo esta diferencia más notoria para el caso del desplazamiento de Stokes y la energía de emisión. Esto es consistente con lo ya mencionado anteriormente, en donde los descriptores QTAIM por grupos fueron construidos especialmente para capturar el efecto de la interacción del "núcleo" del cromóforo con el disolvente y con los grupos sustituyentes vecinos a éste, lo cual a su vez define los valores tanto de la energía de emisión como del desplazamiento de Stokes.

(a) **(b)**

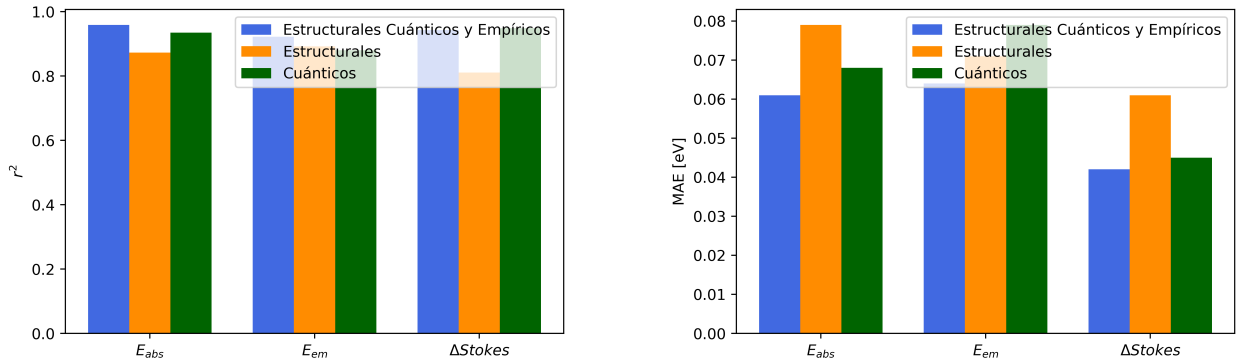


Figura 4.8. Gráficas en donde se muestran los valores obtenidos para el coeficiente de correlación r^2 (a) y para el error absoluto medio MAE (b). En ambas gráficas se muestran los resultados obtenidos para los diferentes subconjuntos de descriptores. Color azul para todos los descriptores (estructurales, cuánticos y empíricos), color naranja únicamente para descriptores estructurales y color verde únicamente para descriptores cuánticos.

(a) **(b)**

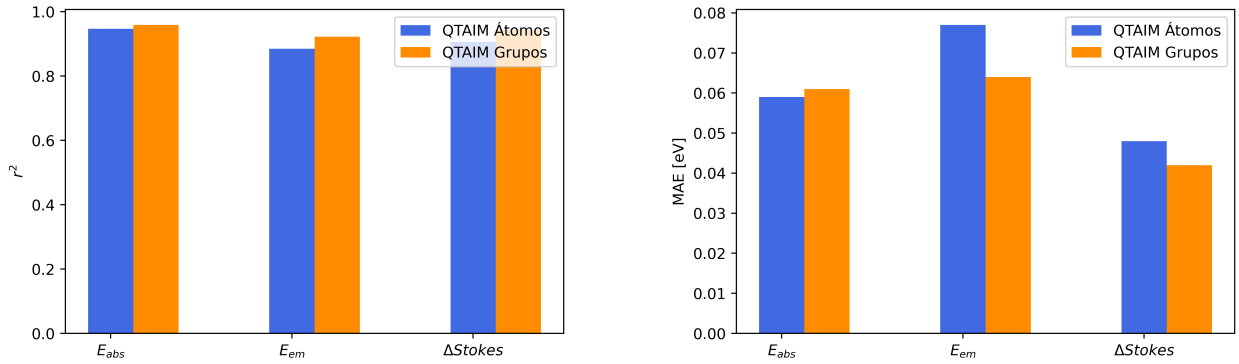


Figura 4.9. Gráficas en donde se muestran los valores obtenidos para el coeficiente de correlación r^2 (a) y para el error absoluto medio MAE (b). En ambas gráficas se muestran los resultados obtenidos para los diferentes subconjuntos de descriptores. Color azul para todos los descriptores (empleando descriptores QTAIM por tipo de átomos) y color amarillo para todos los tipos de descriptores (empleando descriptores QTAIM por grupos).

4.2.2. Resultados para cada una de las propiedades espectroscópicas

En esta sección se muestran las gráficas de distribución (figura 4.10) para las diferentes propiedades estudiadas. En donde se puede observar que las distribuciones para las tres propiedades (energía de absorción, energía de emisión y desplazamiento de Stokes) son relativamente simétricas, comparándolas con las distribuciones para el caso del Sistema 1. Este hecho es de esperarse, debido a que el Sistema 1 contempla una variedad de distintos tipos de cromóforos, lo cual ocasiona que se tenga un mayor intervalo en el valor de sus propiedades, mientras que el sistema 2 se concentra únicamente en un tipo de cromóforo reduciendo de este modo dicho intervalo. Además, esta sección también presenta las gráficas de valor predicho contra valor experimental, las cuales se obtuvieron empleando todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente.

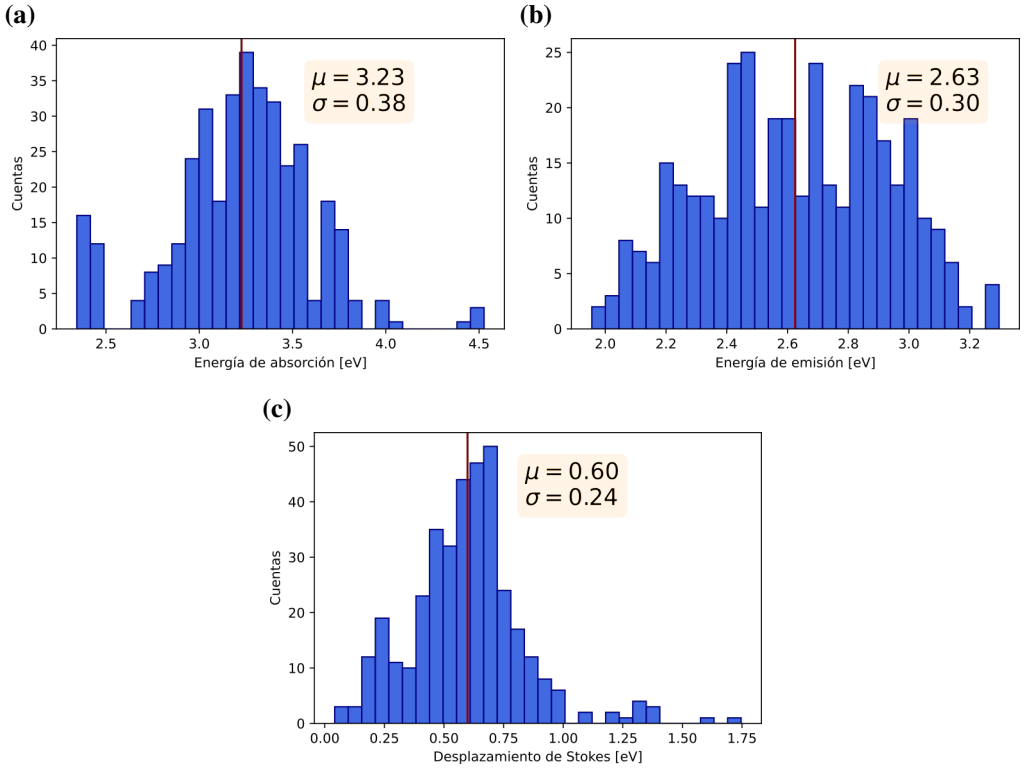


Figura 4.10. Figura en donde se muestra la distribución de datos para las diferentes propiedades estudiadas (Sistema 2). Energía de absorción (a), energía de emisión (b) y el desplazamiento de Stokes (c). Además, se muestra el valor medio de la distribución μ (línea roja) junto con la desviación σ.

La gráfica de valor predicho contra valor experimental para la predicción de la energía de absorción se muestra en la figura 4.11-(a), en donde se observa una muy buena correlación y una muy pequeña desviación de los datos alrededor de la recta ideal (línea negra punteada). Considerando un valor medio de energía de absorción de 3.32 eV (figura 4.10-(a)) se obtiene un error relativo medio de 1.8 %. Lo anterior corresponde a un error considerablemente menor que el obtenido para el Sistema 1.

Haciendo una comparación entre las gráficas del valor predicho contra valor experimental para el Sistema 1 (figura 4.4) y el Sistema 2 (figura 4.11), se puede ver que la cantidad de puntos fuera de la tendencia se redujo significativamente. Esto sugiere que los puntos fuera de la tendencia observados para el Sistema 1 eran ocasionados por una falta de variación en los medios de solvatación, en otras palabras, para evitar la dispersión observada en el sistema 1 se requiere incluir, para un mismo cromóforo, una mayor cantidad de mediciones espectroscópicas en diferentes disolventes.

Por otro lado, la figura 4.11-(b) muestra las gráficas de valor predicho contra valor experimental para la predicción de la energía de emisión, en donde se observa una correcta distribución de los datos alrededor de la recta ideal (línea negra punteada) junto con una dispersión de datos pequeña. Una vez más, considerando un valor medio de energía de emisión de 2.63 eV (figura 4.10-(b)) se obtuvo un error relativo medio de 2.4 %.

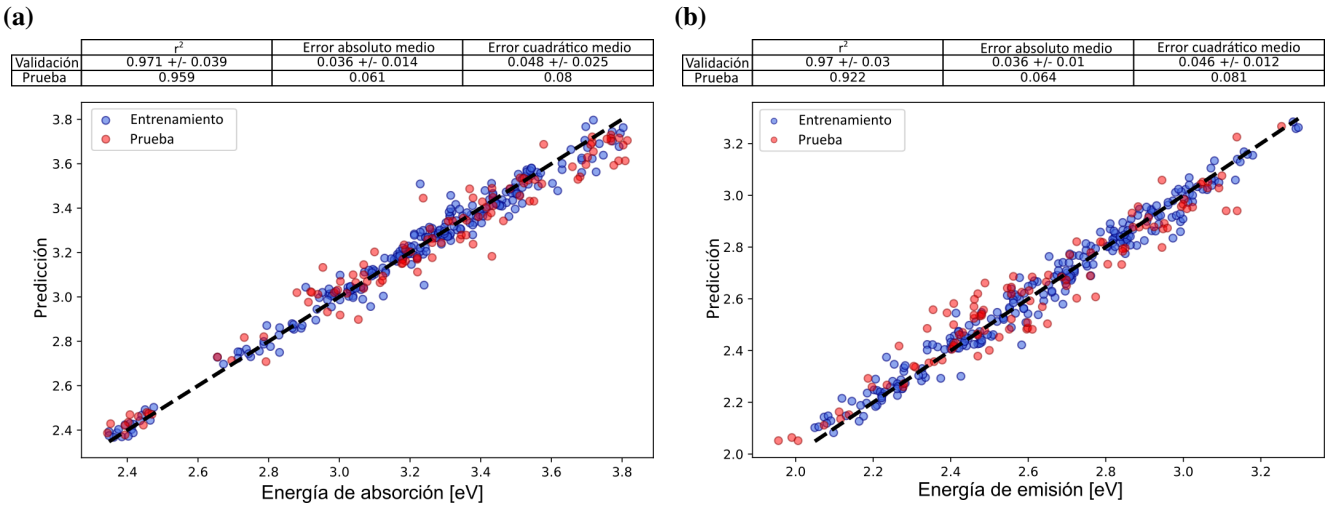


Figura 4.11. Gráficas de valor predicho contra el valor experimental para la energía de absorción (a) y emisión (b) (Sistema 1). Para ello se empleó el modelo de "GradientBoostingRegressor" junto con todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente. Previo a entrenar el modelo, se realizó una reducción de dimensión, teniendo un total de 250 (a) y 200 (b) descriptores después de la reducción de dimensión. En la parte superior de la figura se incluye una tabla con los resultados obtenidos para las tres métricas: coeficiente de correlación (r^2), error absoluto medio y error cuadrático medio, tanto para la validación como para la prueba. En la gráfica se muestra el valor predicho contra el valor experimental para el conjunto de datos de entrenamiento (círculos azules) y para el conjunto de datos de prueba (círculos rojos).

Los resultados anteriores muestran que es posible (empleando modelos de aprendizaje automatizado) predecir tanto la energía de absorción como la energía de emisión con una muy buena precisión ($\simeq$ 0.6 eV). Siendo, el resultado obtenido para la predicción de la energía de emisión particularmente interesante debido a que este no es sencillo de calcular empleando las metodologías convencionales (TD-DFT), esto debido a que para ello se requiere de contemplar el efecto de la relajación vibracional. Además, los resultados obtenidos para predecir la energía de absorción (en cumarinas) mostraron ser mejores que los obtenidos empleando las metodologías convencionales (TD-DFT) [34], debido a que estos no presentan los errores sistemáticos reportados empleando TD-DFT. Estos errores consisten en que los datos de las gráficas de valor predicho contra valor experimental reportados no se ajustan a la línea ideal

(línea recta con pendiente uno ordenada al origen de cero), sino más bien se reporta un ajuste a una línea que no sigue la tendencia adecuada.

Por último, la figura 4.12 muestra las gráficas de valor predicho contra valor experimental para la predicción del desplazamiento de Stokes, en donde a diferencia del Sistema 1, aquí sí se observa una correcta distribución de los datos alrededor de la recta ideal (línea negra punteada). Una vez más, considerando un valor medio de desplazamiento de Stokes de 0.6 eV (figura 4.10-(c)) se obtiene un error relativo medio de 7.0 %, el cual es bastante más pequeño que el obtenido para el caso del Sistema 1.

(a)

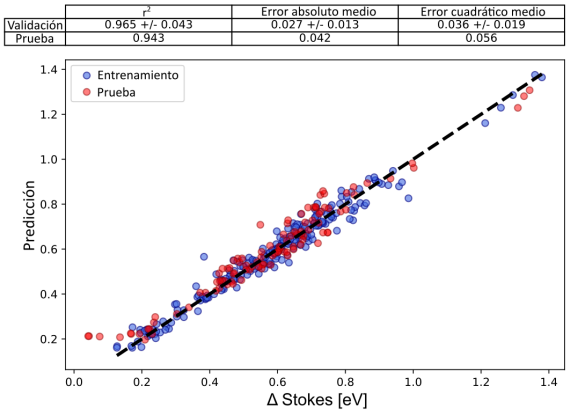


Figura 4.12. Gráficas de valor predicho contra el valor experimental para desplazamiento de Stokes. Para ello se empleó el modelo de "GradientBoostingRegressor" junto con todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente. Previo a entrenar el modelo, se realizó una reducción de dimensión, teniendo un total de 250 descriptores después de la reducción de dimensión. En la parte superior de la figura se incluye una tabla con los resultados obtenidos para las tres métricas: coeficiente de correlación (r^2), error absoluto medio y error cuadrático medio, tanto para la validación como para la prueba. En la gráfica se muestra el valor predicho contra el valor experimental para el conjunto de datos de entrenamiento (círculos azules) y para el conjunto de datos de prueba (círculos rojos).

Este último resultado, es particularmente importante debido a que, una vez más, para el cálculo de esta propiedad empleando TD-DFT se requiere de hacer el cálculo de la energía perdida a causa de la relajación vibracional de los estados excitados. Mientras que empleando esta metodología se puede calcular el desplazamiento de Stokes con una precisión del orden de 0.4 eV sin tener que hacer una optimización del estado excitado. Por otro lado, el hecho de que únicamente para el Sistema 2 se haya encontrado un modelo capaz de predecir el desplazamiento de Stokes sugiere que la variabilidad en los disolventes empleados, junto con no combinar cromóforos con diferentes mecanismos de desactivación, es crucial para la correcta descripción de esta propiedad. El hecho que empleando únicamente descriptores estructurales se haya logrado predecir esta propiedad, sugiere que el factor más importante fue, efectivamente, añadir más variedad de disolvente y no el haber implementado los descriptores QTAIM por Grupos.

A pesar de que este modelo ha mostrado ser competente con los modelos actuales para la predicción de las propiedades espectroscópicas, sigue habiendo un largo camino en desarrollo de mejores y más eficientes métodos que predigan estas propiedades (energía de absorción, energía de emisión y desplazamiento de stokes), ya que los resultados reportados en trabajos previos junto con los obtenidos por este estudio (($\simeq$ 0.6 eV)) están muy lejos de alcanzar la precisión experimental, la cual es del orden de 0.01 a 0.001 eV para un intervalo de 1.5 eV (270 nm) a 4.5 eV (830 nm). Este valor es obtenido a partir de la precisión proporcionada por los espectrómetros modernos, la cual es del orden de 2.0 nm a 0.2 nm, dependiendo del modelo del espectrómetro en cuestión.

4.2.3. Importancia de las características

La importancia relativa de las características (descriptores) para describir la energía de absorción, la energía de emisión y el desplazamiento de Stokes (para el Sistema 2) se muestran en las figuras 4.13 y 4.14. En estas gráficas se observa un mayor predominio de los descriptores cuánticos sobre los estructurales, resultado opuesto al obtenido para el Sistema 1. Este hecho sugiere que el implementar un cálculo de estructura electrónico más sofisticado (como lo es DFT) implica una mejora significativa en los descriptores cuánticos obtenidos, sobre emplear un cálculo semiempírico.

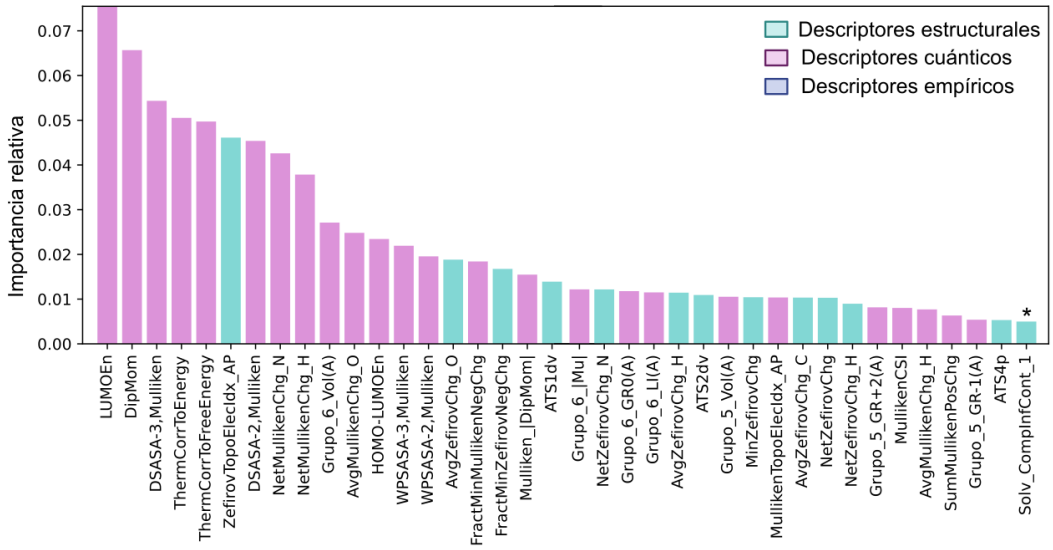


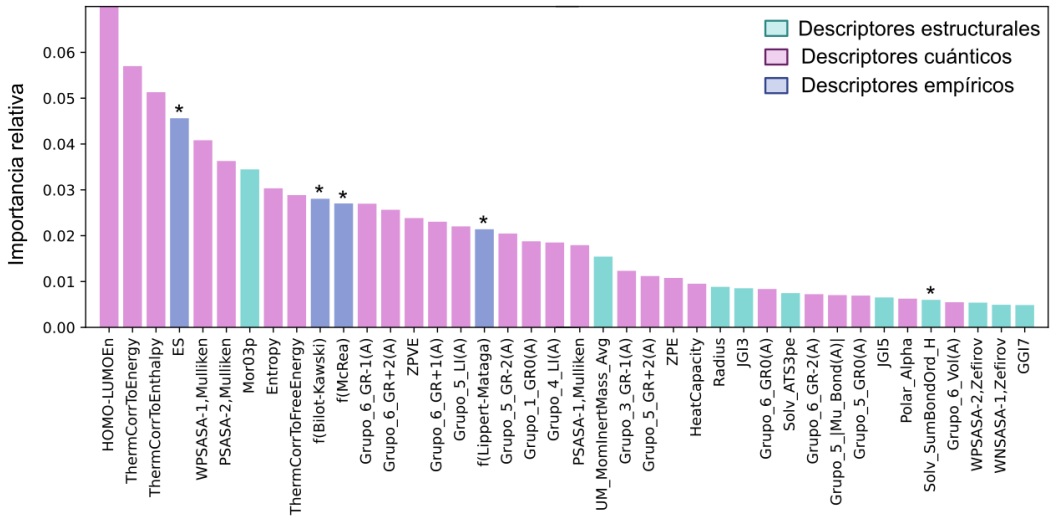
Figura 4.13. Figura en donde se muestra la importancia relativa de los 38 descriptores más importantes para predecir la energía de absorción. Para el entrenamiento se empleó el modelo de "GradientBoostingRegressor" con una reducción de dimensión a 350 descriptores. En donde las barras de color verde corresponden a descriptores estructurales, las barras rosas a descriptores cuánticos y las barras azules a descriptores empíricos, además un asterisco (*) indica que es un descriptor del disolvente.

Una vez más, las energías orbitales toman una clara y esperada relevancia, principalmente para predecir la energía de absorción y emisión. Otras propiedades, calculadas a partir del análisis poblacional de Mulliken, además de propiedades termodinámicas y vibracionales, presentan una relevancia significativa. Para el caso de la energía de emisión y el desplazamiento de Stokes se observó una mayor cantidad de descriptores vibracionales comparándolos con el caso de la energía de absorción.

Si bien, la mayoría de los 38 más importantes descriptores son descriptores del cromóforo (para todas las propiedades), se observó un incremento en la importancia de los descriptores del disolvente para predecir la energía de emisión y el desplazamiento de Stokes, esto comparándolo con el caso para la energía de absorción. Este hecho, junto con el incremento de la importancia en los descriptores vibracionales explica por qué ambas propiedades (energía de emisión y desplazamiento de Stokes) están ligadas a la dinámica de solvatación de los estados excitados, misma que se ve modificada por la interacción con el disolvente y la interacción con los modos normales de vibración del cromóforo.

Con respecto a los descriptores QTAIM por grupos, se observó que los grupos que figuraron más a la hora de predecir la energía de absorción y emisión son los descriptores de los grupos 5 y 6 (ver figura 3.3), los cuales contienen a los átomos de oxígeno del grupo carbonilo, con una pequeña participación de los grupos 3 y 4 (ver figura 3.3), los cuales también integran al grupo carbonilo. Por otro lado, para la predicción del desplazamiento de Stokes, los grupos más relevantes fueron el 1, 2 y 3 (ver figura 3.3), los cuales corresponden a los átomos de los anillos aromáticos. Este hecho indica que una modificación, la densidad electrónica del grupo carbonilo, modifica más fuertemente la energía de absorción y la energía de emisión. Mientras que una modificación sobre los anillos aromáticos modificará en mayor medida el desplazamiento de Stokes.

(a)



(b)

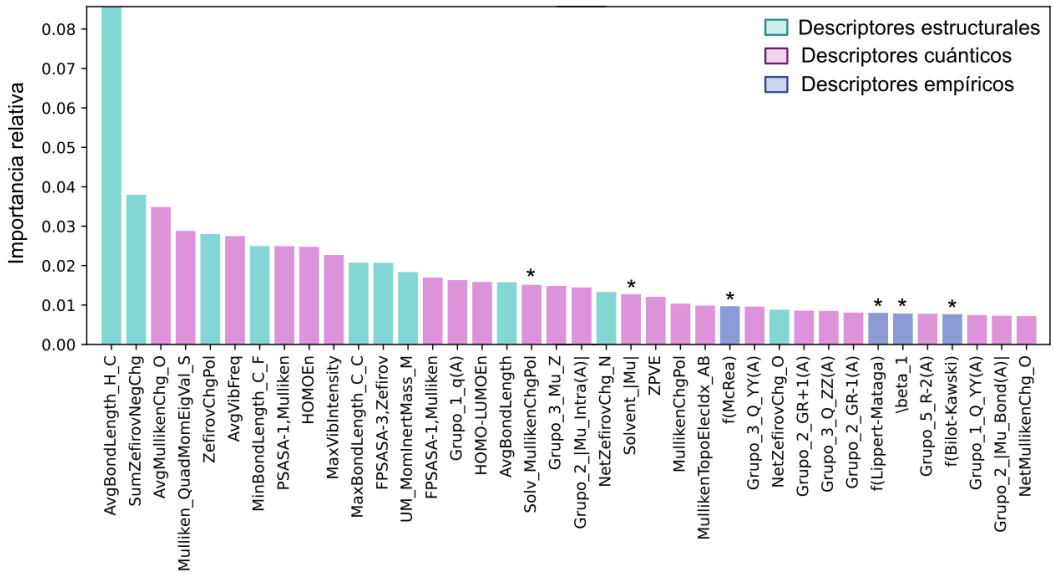


Figura 4.14. Figura en donde se muestra la importancia relativa de los 38 descriptores más importantes para predecir la energía de emisión (a) y el desplazamiento de Stokes (b). Para el entrenamiento se empleó el modelo de "GradientBoostingRegressor" con una reducción de dimensión a 250 descriptores para (a) y 250 descriptores para (b). En donde las barras de color verde corresponden a descriptores estructurales, las barras rosas a descriptores cuánticos y las barras azules a descriptores empíricos, además un asterisco (*) indica que es un descriptor del disolvente.

4.3. Mejores modelos para el Sistema 1 y para el Sistema 2

A modo de resumen, esta sección presenta los "mejores" modelos tanto para el Sistema 1 como para el Sistema 2. Para el Sistema 1 en todos los casos, el mejor modelo encontrado fue empleando todos los tipos de descriptores (estructurales, cuánticos y empíricos). Mientras que para el Sistema 2 el mejor modelo encontrado (para predecir la energía de absorción y emisión) fue empleando los descriptores estructurales, las energías orbitales, los descriptores electrónicos y vibracionales así como los descriptores QTAIM por grupos. Por último, y un caso excepcional, el mejor modelo encontrado para predecir el desplazamiento de stokes fue usando únicamente los descriptores QTAIM por grupos. Los resultados obtenidos para estos modelos se muestran en la figura 4.15.

Adicionalmente, para todos los caso (Sistema 1 y Sistema 2) los modelos más eficientes se encontraron empleando únicamente descriptores estructurales y empíricos, ya que el cálculo de estos dos tipos de descriptores conlleva un costo computacional muy bajo.

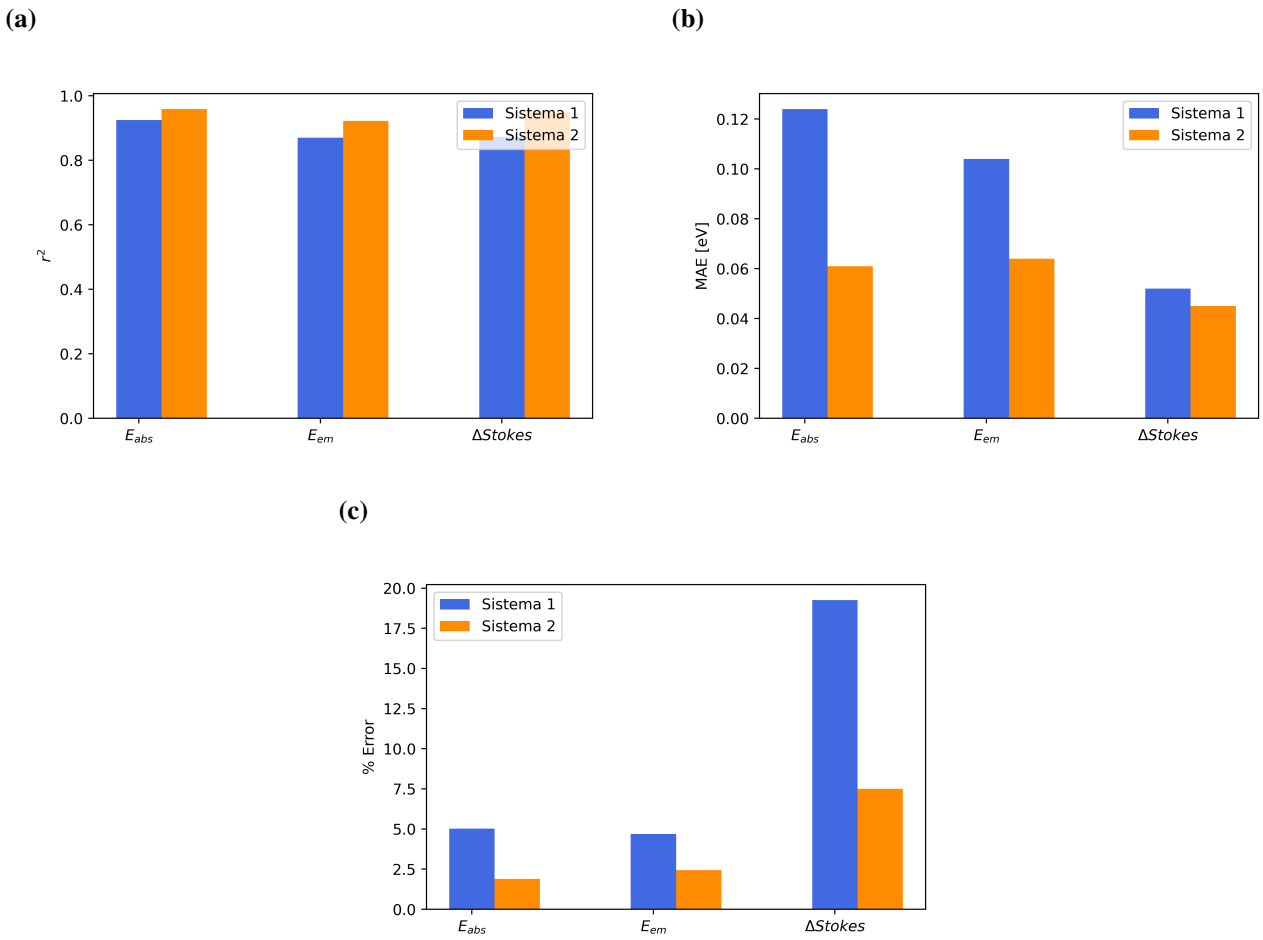


Figura 4.15. Figura en donde se muestran los resultados de los mejores modelos para todas las propiedades y para ambos sistemas. (a) presenta los valores de coeficiente de correlación r^2, (b) el error absoluto medio y (c) el error porcentual respecto a la media. Las barras de color azul corresponden a los resultados para el Sistema 1 y las de color amarillo para el Sistema 2.

Conclusiones

5.1. Conclusiones generales

- Los modelos de aprendizaje automatizado utilizados, junto con los respectivos descriptores empleados, mostraron ser capaces de predecir la mayoría de las propiedades espectroscópicas estudiadas, tales como: la energía de absorción, emisión y el desplazamiento de Stokes. Siendo en cierta medida menos adecuados para predecir otras, como fue el caso del tiempo de vida de fluorescencia.

- Se concluyó que el uso de descriptores estructurales, junto con descriptores empíricos, es lo más recomendable para un primer estudio, ya que con estos descriptores se pueden obtener resultados casi tan confiables como los resultados obtenidos empleando descriptores cuánticos, pero con un costo computacional mucho menor.

- Si bien, los descriptores cuánticos prometían proporcionar herramientas para el diseño de nuevos cromóforos, ya que estos se asocian directamente con propiedades físicas de la molécula, el hecho de que los modelos requieran de emplear más de 40 descriptores para la correcta descripción de los mismos, hace que estos no proporcionen las herramientas de diseño esperadas, limitando a los modelos a emplearse únicamente como una herramienta de predicción.

- Por último, el uso de modelos de aprendizaje automatizado basado en árboles de decisión nos proporciona una forma capturar ciertos comportamientos de las propiedades espectroscópicas, como lo es el hecho de que las energías orbitales son relevantes para la predicción de la energía de absorción y emisión, pero no para el desplazamiento de Stokes o el tiempo de vida de fluorescencia. Esto muestra el poder de las metodologías de aprendizaje automatizado, las cuales no solamente logran predecir, sino que también son capaces de reproducir cualitativamente algunos comportamientos de estas propiedades.

5.2. Conclusiones particulares

- Para el Sistema 1, subdividir al conjunto de cromóforos totales en conjuntos más específicos, no mostró una mejoría significativa para la predicción de las correspondientes propiedades espectroscópicas, sugiriendo que los modelos de aprendizaje automatizado empleados logran capturar los efectos ocasionados por la variabilidad estructural y los diferentes fenómenos que ocurren entre los estados electrónicos para los diferentes tipos de cromóforos, a excepción del caso de desplazamiento de Stokes en donde fue necesario realizar un estudio para un tipo de cromóforo en particular (cumarinas).

- Implementar una mayor variabilidad en los diferentes medios de solvatación, es decir, tener registro de las diferentes propiedades espectroscópicas en más medios de solvatación, sí mostró una mejora significativa, reduciendo en gran medida los puntos fuera de la tendencia que se observaban en el Sistema 1, los cuales ya no se observaron para el Sistema 2.

- El estudio sobre los diferentes tipos de descriptores cuánticos mostró, que efectivamente hay descriptores que correlacionan más con ciertas propiedades. Este es el caso de las energías orbitales, las cuales son cruciales para la predicción de la energía de absorción y poco relevantes para predecir el desplazamiento de Stokes.

- El uso de los descriptores QTAIM por grupos, los cuales implementan el efecto de los sustituyentes sobre el "núcleo" de algún tipo de cromóforo y la interacción con su entorno de solvatación, mostró tener una relevancia cuantitativa para la descripción de la dinámica de los estados excitados en cumarinas, al poder predecir con la mayor precisión los valores del desplazamiento de Stokes.

Bibliografía

[1] Eitan Lerner, Thorben Cordes, Antonino Ingargiola, Yazan Alhadid, Sang Yoon Chung, Xavier Michalet, and Shimon Weiss. Toward dynamic structural biology: Two decades of single-molecule förster resonance energy transfer. *Science*, 359(6373), 2018.

[2] Hoi Sung Chung and William A. Eaton. Protein folding transition path times from single molecule FRET. *Current Opinion in Structural Biology*, 48:30–39, 2018.

[3] Emmanuel Villatoro, Leonardo Muñoz-Rugeles, Jesús Durán-Hernández, Bernardo Salcido-Santacruz, Nuria Esturau-Escofet, Jose G. López-Cortés, M. Carmen Ortega-Alfaro, and Jorge Peón. Two-photon induced isomerization through a cyaninic molecular antenna in azo compounds. *Chemical Communications*, 57(25):3123–3126, 2021.

[4] Carlo Matera, Alexandre M.J. Gomila, Núria Camarero, Michela Libergoli, Concepció Soler, and Pau Gorostiza. Photoswitchable Antimetabolite for Targeted Photoactivated Chemotherapy. *Journal of the American Chemical Society*, 140(46):15764–15773, 2018.

[5] Willem A. Velema, Wiktor Szymanski, and Ben L. Feringa. Photopharmacology: Beyond proof of principle. *Journal of the American Chemical Society*, 136(6):2178–2191, 2014.

[6] Ruth Dorel and Ben L. Feringa. Photoswitchable catalysis based on the isomerisation of double bonds. *Chemical Communications*, 55(46):6477–6486, 2019.

[7] Dominik Wöll and Cristina Flors. Super-resolution fluorescence imaging for materials science. *Small Methods*, 1(10):1–12, 2017.

[8] Masafumi Minoshima and Kazuya Kikuchi. Photostable and photoswitching fluorescent dyes for super-resolution imaging. *Journal of Biological Inorganic Chemistry*, 22(5):639–652, 2017.

[9] Hafsah Klfout, Adam Stewart, Mahmoud Elkhalifa, and Hongshan He. BODIPYs for Dye-Sensitized Solar Cells. *ACS Applied Materials and Interfaces*, 9(46):39873–39889, 2017.

[10] Vigneshwari Subramanian, Ekaterina Ratkova, David Palmer, Ola Engkvist, Maxim Fedorov, and Antonio Llinas. Multisolvent Models for Solvation Free Energy Predictions Using 3D-RISM Hy-

dration Thermodynamic Descriptors. *Journal of Chemical Information and Modeling*, 60(6):2977–2988, 2020.

[11] Yipin Lu, Shankara Anand, William Shirley, Peter Gedeck, Brian P. Kelley, Suzanne Skolnik, Stephane Rodde, Mai Nguyen, Mika Lindvall, and Weiping Jia. Prediction of pKa Using Machine Learning Methods with Rooted Topological Torsion Fingerprints: Application to Aliphatic Amines. *Journal of Chemical Information and Modeling*, 59(11):4706–4719, 2019.

[12] Michael G. Taylor, Tzuhsiung Yang, Sean Lin, Aditya Nandy, Jon Paul Janet, Chenru Duan, and Heather J. Kulik. Seeing Is Believing: Experimental Spin States from Machine Learning Model Structure Predictions. *Journal of Physical Chemistry A*, 124(16):3286–3299, 2020.

[13] Fabienne Dioury, Arthur Duprat, Gérard Dreyfus, Clotilde Ferroud, and Janine Cossy. QSPR prediction of the stability constants of gadolinium(III) complexes for magnetic resonance imaging. *Journal of Chemical Information and Modeling*, 54(10):2718–2731, 2014.

[14] Brandon J. Jaquis, Ailin Li, Nolan D. Monnier, Robert G. Sisk, William E. Acree, and Andrew S.I.D. Lang. Using Machine Learning to Predict Enthalpy of Solvation. *Journal of Solution Chemistry*, pages 564–573, 2019.

[15] Mengshan Li, Suyun Lian, Fan Wang, Yanying Zhou, Bingsheng Chen, Lixin Guan, and Yan Wu. Prediction Model of Organic Molecular Absorption Energies based on Deep Learning trained by Chaos-enhanced Accelerated Evolutionary algorithm. *Scientific Reports*, 9(1):1–9, 2019.

[16] Andreas Schüller, Garrett Benjamin Goh, Hanjo Kim, Jun Seok Lee, and Young Tae Chang. Quantitative structure-fluorescence property relationship analysis of a large BODIPY library. *Molecular Informatics*, 29(10):717–729, 2010.

[17] Chia Hsiu Chen, Kenichi Tanaka, and Kimito Funatsu. Random Forest Approach to QSPR Study of Fluorescence Properties Combining Quantum Chemical Descriptors and Solvent Conditions. *Journal of Fluorescence*, 28(2):695–706, 2018.

[18] Cheng Wei Ju, Hanzhi Bai, Bo Li, and Rizhang Liu. Machine Learning Enables Highly Accurate Predictions of Photophysical Properties of Organic Fluorescent Materials: Emission Wavelengths and Quantum Yields. *Journal of Chemical Information and Modeling*, 61(3):1053–1065, 2021.

[19] Aditya R. Thawani, Ryan-Rhys Griffiths, Arian Jamasb, Anthony Bourached, Penelope Jones, William McCorkindale, Alexander A. Aldrick, and Alpha A. Lee. The Photoswitch Dataset: A Molecular Machine Learning Benchmark for the Advancement of Synthetic Chemistry. 2020.

[20] David I Ramírez-Palma, Cesar R García-Jacas, Pablo Carpio-Martínez, and Fernando Cortés-Guzmán. Predicting reactive sites with quantum chemical topology: carbonyl additions in multicomponent reactions. *Physical Chemistry Chemical Physics*, 22(17):9283–9289, 2020.

[21] David I. Ramírez-Palma, Cesar R. García-Jacas, Pablo Carpio-Martínez, and Fernando Cortés-Guzmán. Predicting reactive sites with quantum chemical topology: Carbonyl additions in multicomponent reactions. *Physical Chemistry Chemical Physics*, 22(17):9283–9289, 2020.

[22] Luis Gutiérrez-Arzaluz, Tomás Rocha-Rinza, and Fernando Cortés-Guzmán. Stilbene photoisomerization driving force as revealed by the topology of the electron density and QTAIM properties. *Computational and Theoretical Chemistry*, 1053:214–219, 2015.

[23] José Manuel Guevara-Vela, Miguel Gallegos, Mónica A. Valentín-Rodríguez, Aurora Costales, Tomás Rocha-Rinza, and Ángel Martín Pendás. On the relationship between hydrogen bond strength and the formation energy in resonance-assisted hydrogen bonds. *Molecules*, 26(14):1–11, 2021.

[24] José Manuel Guevara-Vela, Eduardo Romero-Montalvo, Víctor Arturo Mora Gómez, Rodrigo Chávez-Calvillo, Marco García-Revilla, Evelio Francisco, Ángel Martín Pendás, and Tomás Rocha-Rinza. Hydrogen bond cooperativity and anticooperativity within the water hexamer. *Physical Chemistry Chemical Physics*, 18(29):19557–19566, 2016.

[25] P. L.A. Popelier and P. J. Smith. QSAR models based on quantum topological molecular similarity. *European Journal of Medicinal Chemistry*, 41(7):862–873, 2006.

[26] P. J. Smith and P. L.A. Popelier. Quantum chemical topology (QCT) descriptors as substitutes for appropriate Hammett constants. *Organic and Biomolecular Chemistry*, 3(18):3399–3407, 2005.

[27] Mike Devereux, Paul L.A. Popelier, and Iain M. McLay. Quantum isostere database: A web-based tool using quantum chemical topology to predict bioisosteric replacements for drug design. *Journal of Chemical Information and Modeling*, 49(6):1497–1513, 2009.

[28] Bartosz Błasiak, Casey H. Londergan, Lauren J. Webb, and Minhaeng Cho. Vibrational Probes: From Small Molecule Solvatochromism Theory and Experiments to Applications in Complex Systems. *Accounts of Chemical Research*, 50(4):968–976, 2017.

[29] Christiane Albrecht. Joseph r. lakowicz: Principles of fluorescence spectroscopy, 2008.

[30] Junfeng Li, Zilvinas Rinkevicius, and Zexing Cao. A time-dependent density-functional theory and complete active space self-consistent field method study of vibronic absorption and emission spectra of coumarin. *Journal of Chemical Physics*, 141(1), 2014.

[31] Panwang Zhou. Why the lowest electronic excitations of rhodamines are overestimated by time-dependent density functional theory. *International Journal of Quantum Chemistry*, 118(23):1–11, 2018.

[32] Edward J. Beard, Ganesh Sivaraman, Álvaro Vázquez-Mayagoitia, Venkatram Vishwanath, and Jacqueline M. Cole. Comparative dataset of experimental and computational attributes of UV/vis absorption spectra. *Scientific Data*, 6(1):1–11, 2019.

[33] Mai Van Bay, Nguyen Khoa Hien, Phan Thi Diem Tran, Nguyen Tran Kim Tuyen, Doan Thi Yen Oanh, Pham Cam Nam, and Duong Tuan Quang. TD-DFT benchmark for UV-Vis spectra of coumarin derivatives. *Vietnam Journal of Chemistry*, 59(2):203–210, 2021.

[34] Amjad Ali, Muhammad Imran Rafiq, Zhuohan Zhang, Jinru Cao, Renyong Geng, Baojing Zhou, and Weihua Tang. TD-DFT benchmark for UV-visible spectra of fused-ring electron acceptors using global and range-separated hybrids. *Physical Chemistry Chemical Physics*, 22(15):7864–7874, 2020.

[35] Denis Jacquemin, Valérie Wathelet, Eric A Perpete, and Carlo Adamo. Extensive td-dft benchmark: singlet-excited states of organic molecules. *Journal of Chemical Theory and Computation*, 5(9):2420–2435, 2009.

[36] Amjad Ali, Muhammad Imran Rafiq, Zhuohan Zhang, Jinru Cao, Renyong Geng, Baojing Zhou, and Weihua Tang. Td-dft benchmark for uv-visible spectra of fused-ring electron acceptors using global and range-separated hybrids. *Physical Chemistry Chemical Physics*, 22(15):7864–7874, 2020.

[37] Mai Van Bay, Nguyen Khoa Hien, Phan Thi Diem Tran, Nguyen Tran Kim Tuyen, Doan Thi Yen Oanh, Pham Cam Nam, and Duong Tuan Quang. Td-dft benchmark for uv-vis spectra of coumarin derivatives. *Vietnam Journal of Chemistry*, 59(2):203–210, 2021.

[38] Edward Kim, Kevin Huang, Alex Tomala, Sara Matthews, Emma Strubell, Adam Saunders, Andrew McCallum, and Elsa Olivetti. Machine-learned and codified synthesis parameters of oxide materials. *Scientific data*, 4(1):1–9, 2017.

[39] Mohd Shahid Khan and Zahid H Khan. Ab initio and semiempirical study of structure and electronic spectra of hydroxy substituted naphthoquinones. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 61(4):777–790, 2005.

[40] P Lakshmi Praveen and Durga P Ojha. Optical absorption behavior and spectral shifts of fluorinated liquid crystals in ultraviolet region: A comparative study based on dft and semiempirical approaches. *Journal of Molecular Liquids*, 194:8–12, 2014.

[41] Mikhail Y Berezin and Samuel Achilefu. Fluorescence lifetime measurements and biological imaging. *Chemical reviews*, 110(5):2641–2684, 2010.

[42] Richard Bader. *Atoms in molecules : a quantum theory*. Clarendon Press, Oxford, 1990.

[43] Chia-Hsiu Chen, Kenichi Tanaka, and Kimito Funatsu. Random forest approach to qspr study of fluorescence properties combining quantum chemical descriptors and solvent conditions. *Journal of fluorescence*, 28(2):695–706, 2018.

[44] Andreas Schüller, Garrett Benjamin Goh, Hanjo Kim, Jun-Seok Lee, and Young-Tae Chang. Quantitative structure-fluorescence property relationship analysis of a large bodipy library. *Molecular Informatics*, 29(10):717–729, 2010.

[45] Cheng-Wei Ju, Hanzhi Bai, Bo Li, and Rizhang Liu. Machine learning enables highly accurate predictions of photophysical properties of organic fluorescent materials: Emission wavelengths and quantum yields. *Journal of Chemical Information and Modeling*, 61(3):1053–1065, 2021.

[46] Siddharth S. Matikonda, Joseph Ivanic, Miguel Gomez, Gabrielle Hammersley, and Martin J. Schnermann. Core remodeling leads to long wavelength fluoro-coumarins. *Chemical Science*, 11(28):7302–7307, 2020.

[47] K. Das, B. Jain, and H. S. Patel. Hydrogen bonding properties of coumarin 151, 500, and 35: The effect of substitution at the 7-amino position. *Journal of Physical Chemistry A*, 110(5):1698–1704, 2006.

[48] Xiaogang Liu, Jacqueline M. Cole, and Kian Sing Low. Solvent effects on the uv-vis absorption and emission of optoelectronic coumarins: A Comparison of three empirical solvatochromic models. *Journal of Physical Chemistry C*, 117(28):14731–14741, 2013.

[49] Sandip K. Lanke and Nagaiyan Sekar. Coumarin Push-Pull NLOphores with Red Emission: Solvatochromic and Theoretical Approach. *Journal of Fluorescence*, 26(3):949–962, 2016.

[50] Banibrata Maity, Aninda Chatterjee, and Debabrata Seth. Photophysics of a coumarin in different solvents: Use of different solvatochromic models. *Photochemistry and Photobiology*, 90(4):734–746, 2014.

[51] Bernardo Salcido-Santacruz. Spectroscopy data base of chromophores in different solvents. Available at https://github.com/BernardoSalcido/Chromophore-database, 2022.

[52] Bernardo Salcido-Santacruz. Machine Learning Molecule. Available at https://github.com/BernardoSalcido/ML_Molecule, 2022.

[53] V S.; Karelson M Katritzky A. R.; Lobanov. CODESSA, 1994.

[54] Hirotomo Moriwaki, Yu Shi Tian, Norihito Kawashita, and Tatsuya Takagi. Mordred: A molecular descriptor calculator. *Journal of Cheminformatics*, 10(1):1–14, 2018.

[55] G. Landrum. Open-source cheminformatics; http://www.rdkit.org, 2016.

[56] Sereina Riniker and Gregory A. Landrum. Better Informed Distance Geometry: Using What We Know to Improve Conformation Generation. *Journal of Chemical Information and Modeling*, 55(12):2562–2574, 2015.

[57] Steven M Bachrach. Ampac 6.6 with ampac 6.0 graphic interface semichem, inc., po box 1649, shawnee mission, ks 66222. http://www. semichem. com. list price: 11,000.educationalprice: 900, 2000.

[58] Ademola Soyemi and Tibor Szilvási. Benchmarking semiempirical qm methods for calculating the dipole moment of organic molecules. *The Journal of Physical Chemistry A*, 126(11):1905–1921, 2022.

[59] Lilin Lu, Hui Hu, Hua Hou, and Baoshan Wang. An improved B3LYP method in the calculation of organic thermochemistry and reactivity. *Computational and Theoretical Chemistry*, 1015:64–71, 2013.

[60] Julian Tirado-Rives and William L. Jorgensen. Performance of B3LYP density functional methods for a large set of organic molecules. *Journal of Chemical Theory and Computation*, 4(2):297–306, 2008.

[61] T A Keith. AIMAll (Version 19.10.12), 2019.

[62] Christian Laurence, Julien Legros, Agisilaos Chantzis, Aurélien Planchat, and Denis Jacquemin. A Database of dispersion-induction DI, electrostatic ES, and hydrogen bonding $\alpha 1$ and $\delta 1$ solvent parameters and some applications to the multiparameter correlation analysis of solvent effects. *Journal of Physical Chemistry B*, 119(7):3174–3184, 2015.

[63] Aneta Lewkowicz, Karolina Baranowska, Piotr Bojarski, and Marek Józefowicz. Solvent dependent spectroscopic properties of fingerprint reagent-1, 8-diazafluoren-9-one. *Journal of Molecular Liquids*, 285:754–765, 2019.

[64] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, September 1975.

[65] Hirofumi Mitekura, Tomoko No, Kazuyoshi Suzuki, Kyosuke Satake, and Masaru Kimura. Spectroscopic properties of meso-substituted cyanine dyes: evidences for intramolecular charge transfer from a julolidine moiety as a meso-substituent to the cyanine chromophore. *Dyes and pigments*, 54(2):113–120, 2002.

[66] FM Richards and T Richmond. and protein structure. *Molecular Interactions and Activity in Proteins*, (60):23, 1978.

Apéndice A

Sistema 1

Tabla 5.1. Tabla en donde se muestra una comparación del desempeño de los modelos para los diferentes subconjuntos de cromóforos (Sistema 1). Para el entrenamiento se empleó el modelo de "RandomForestRegressor" junto con todos los tipos de descriptores calculados (estructurales, cuánticos y empíricos). Para la evaluación se reportan dos métricas, coeficiente de correlación r^2 y el error absoluto medio MAE, tanto para el conjunto de validación como para el de prueba. Para el entrenamiento de los modelos se emplearon todos los tipos de descriptores implementados para este sistema, tanto para el cromóforo (estructurales y cuánticos) como para el disolvente (estructurales, cuánticos y empíricos). En la tercera columna se muestran los resultados de la reducción de dimensión, donde el primer valor corresponde al número de descriptores antes de la reducción de dimensión y el segundo valor corresponde al número de descriptores después de la reducción de dimensión.

Propiedad	Tipo de Cromóforos	Reduccion de dimensión		r^2 (Prueba)	r^2 (Validación)	MAE (Prueba)	MAE (Validación)
E_{abs}	Todos	1820	50	0.925	0.955 ± 0.045	0.077 eV	0.046 ± 0.026 eV
E_{abs}	Bodipys	292	250	0.875	0.897 ± 0.198	0.104 eV	0.089 ± 0.025 eV
E_{abs}	Cianinas	1678	200	0.921	0.910 ± 0.107	0.076 eV	0.050 ± 0.041 eV
E_{em}	Todos	1820	100	0.859	0.899 ± 0.059	0.105 eV	0.081 ± 0.023 eV
E_{em}	Bodipys	292	50	0.853	0.920 ± 0.135	0.054 eV	0.040 ± 0.027 eV
E_{em}	Cianinas	1678	150	0.848	0.768 ± 0.304	0.066 eV	0.058 ± 0.033 eV
ΔStokes	Todos	1820	200	0.868	0.881 ± 0.136	0.060 eV	0.049 ± 0.017 eV
ΔStokes	Bodipys	292	150	0.808	0.746 ± 0.231	0.028 eV	0.025 ± 0.014 eV
ΔStokes	Cianinas	1678	400	0.839	0.526 ± 1.139	0.034 eV	0.030 ± 0.022 eV
τ_{flu}	Todos	1820	300	0.726	0.801 ± 0.116	0.855 ns	0.616 ± 0.198 ns
τ_{flu}	Bodipys	292	350	0.628	0.625 ± 0.126	0.853 ns	0.779 ± 0.280 ns
τ_{flu}	Cianinas	1678	400	0.591	0.864 ± 0.159	0.349 ns	0.213 ± 0.100 ns

Tabla 5.2. Tabla donde se muestra una comparación del desempeño de los modelos para los diferentes subconjuntos de descriptores (Sistema 1). Para el entrenamiento se empleó el modelo de "RandomForestRegressor" en donde se reportan dos métricas, coeficiente de correlación r^2 y el error absoluto medio MAE, tanto para el conjunto de validación como para el de prueba. Para el entrenamiento de los modelos se emplearon descriptores estructurales y empíricos para el disolvente, así como los descriptores del cromóforo que se especifican en las columnas dos y tres. En la cuarta columna se muestran los resultados de la reducción de dimensión, en donde el primer valor corresponde al número de descriptores antes de la reducción y el segundo valor corresponde al número de descriptores después de la reducción de dimensión.

Propiedad	Descriptores estructurales	Descriptores cuánticos	Reducción de dimensión		r^2 (Prueba)	r^2 (Validación)	MAE (Prueba)	MAE (Validación)
E_{abs}	✓	✓	1820	50	0.926	0.957 ± 0.052	0.123 eV	0.067 ± 0.032 eV
E_{abs}	✓		292	150	0.918	0.946 ± 0.07	0.120 eV	0.069 ± 0.016 eV
E_{abs}		✓	1678	50	0.938	0.959 ± 0.034	0.092 eV	0.081 ± 0.023 eV
E_{em}	✓	✓	1820	100	0.859	0.899 ± 0.059	0.105 eV	0.062 ± 0.019 eV
E_{em}	✓		292	150	0.827	0.926 ± 0.051	0.111 eV	0.081 ± 0.012 eV
E_{em}		✓	1678	100	0.846	0.896 ± 0.048	0.114 eV	0.049 ± 0.017 eV
ΔStokes	✓	✓	1820	200	0.868	0.881 ± 0.136	0.060 eV	0.050 ± 0.015 eV
ΔStokes	✓		292	250	0.786	0.869 ± 0.106	0.093 eV	0.046 ± 0.011 eV
ΔStokes		✓	1678	300	0.799	0.903 ± 0.072	0.067 eV	0.046 ± 0.011 eV
τ_{flu}	✓	✓	1820	300	0.726	0.801 ± 0.116	0.855 ns	0.616 ± 0.198 ns
τ_{flu}	✓		292	150	0.663	0.783 ± 0.127	0.951 ns	0.686 ± 0.185 ns
τ_{flu}		✓	1678	250	0.745	0.787 ± 0.129	0.732 ns	0.642 ± 0.154 ns

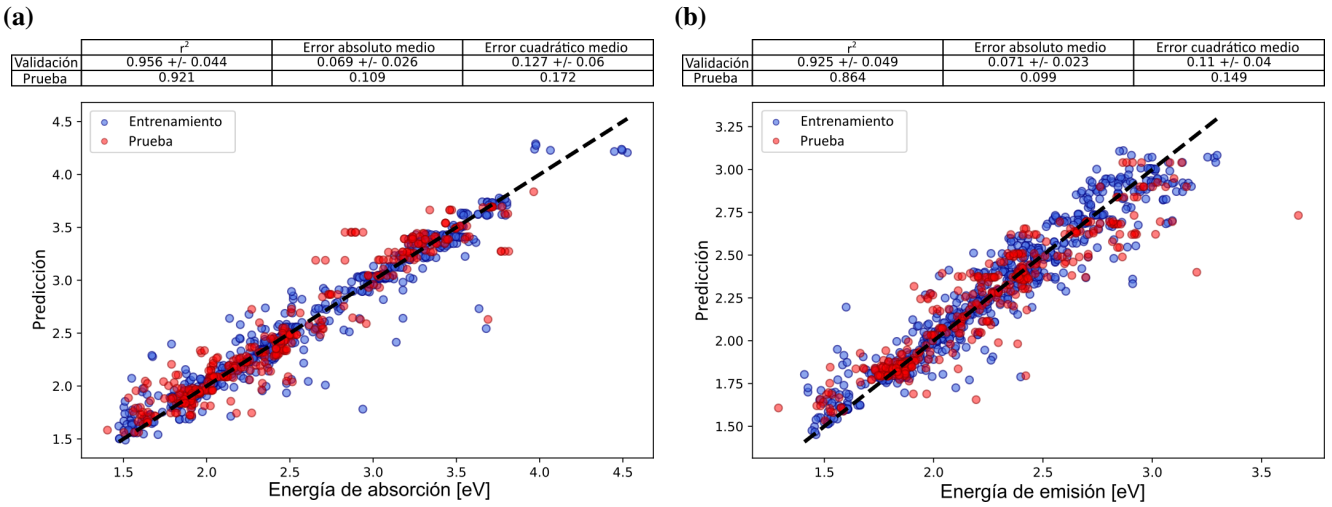
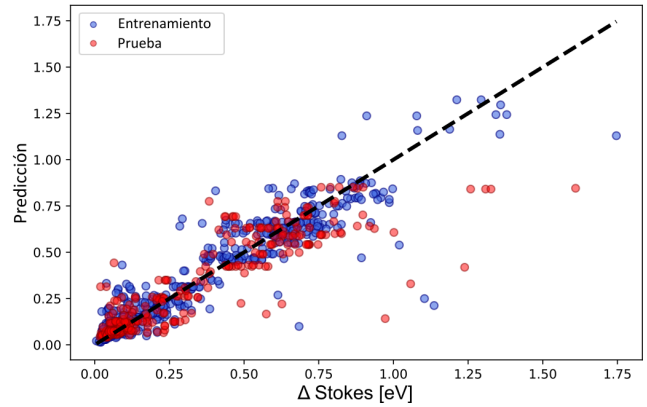


Figura 5.1. Gráficas de valor predicho contra el valor experimental para la energía de absorción (a) y emisión (b) (Sistema 1). Para ello se empleó el modelo de "RandomForestRegressor" junto con todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente. Previo a entrenar el modelo, se realizó una reducción de dimensión, teniendo un total de 50 (a) y 500 (b) descriptores después de la reducción de dimensión. En la parte superior de la figura se incluye una tabla con los resultados obtenidos para las tres métricas: coeficiente de correlación (r^2), error absoluto medio y error cuadrático medio, tanto para la validación como para la prueba. En la gráfica se muestra el valor predicho contra el valor experimental para el conjunto de datos de entrenamiento (círculos azules) y para el conjunto de datos de prueba (círculos rojos).

(a)

	r^2	Error absoluto medio	Error cuadrático medio
Validación	0.897 +/- 0.085	0.048 +/- 0.011	0.09 +/- 0.04
Prueba	0.793	0.072	0.133



(b)

	r^2	Error absoluto medio	Error cuadrático medio
Validación	0.788 +/- 0.125	0.642 +/- 0.158	1.008 +/- 0.24
Prueba	0.75	0.779	1.099

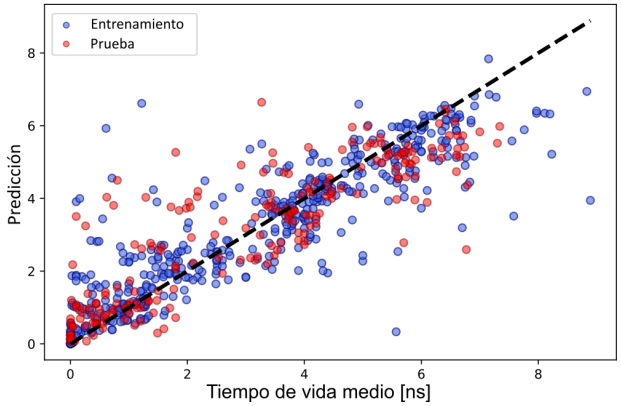


Figura 5.2. Gráficas de valor predicho contra el valor experimental para desplazamiento de Stokes (a) y del tiempo de vida de fluorescencia (b) (Sistema 1). Para ello se empleó el modelo de "RandomForestRegressor" junto con todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente. Previo a entrenar el modelo, se realizó una reducción de dimensión, teniendo un total de 500 (a) y 300 (b) descriptores después de la reducción de dimensión. En la parte superior de la figura se incluye una tabla con los resultados obtenidos para las tres métricas: coeficiente de correlación (r^2), error absoluto medio y error cuadrático medio, tanto para la validación como para la prueba. En la gráfica se muestra el valor predicho contra el valor experimental para el conjunto de datos de entrenamiento (círculos azules) y para el conjunto de datos de prueba (círculos rojos).

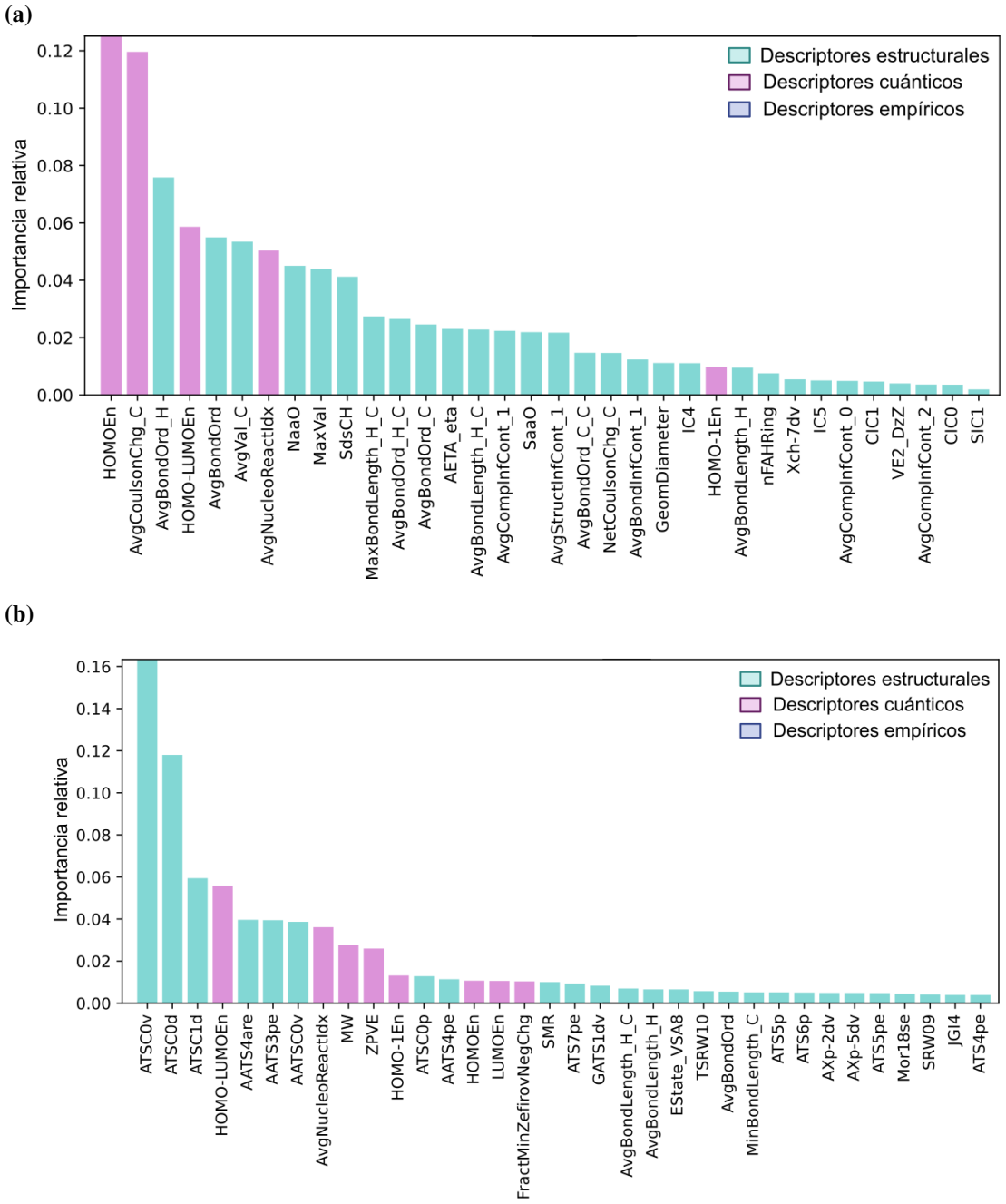


Figura 5.3. Figura en donde se muestra la importancia relativa de los 34 descriptores más importantes para predecir la energía de absorción (a) y emisión (b). Para el entrenamiento se empleó el modelo de "RandomForestRegressor" con una reducción de dimensión a 50 descriptores para (a) y 500 descriptores para (b). En donde las barras de color verde corresponden a descriptores estructurales, las barras color rosa a descriptores cuánticos y las barras color azul a descriptores empíricos. Además, un asterisco (*) sobre las barras indica que éstas corresponden a un descriptor del disolvente.

(a)

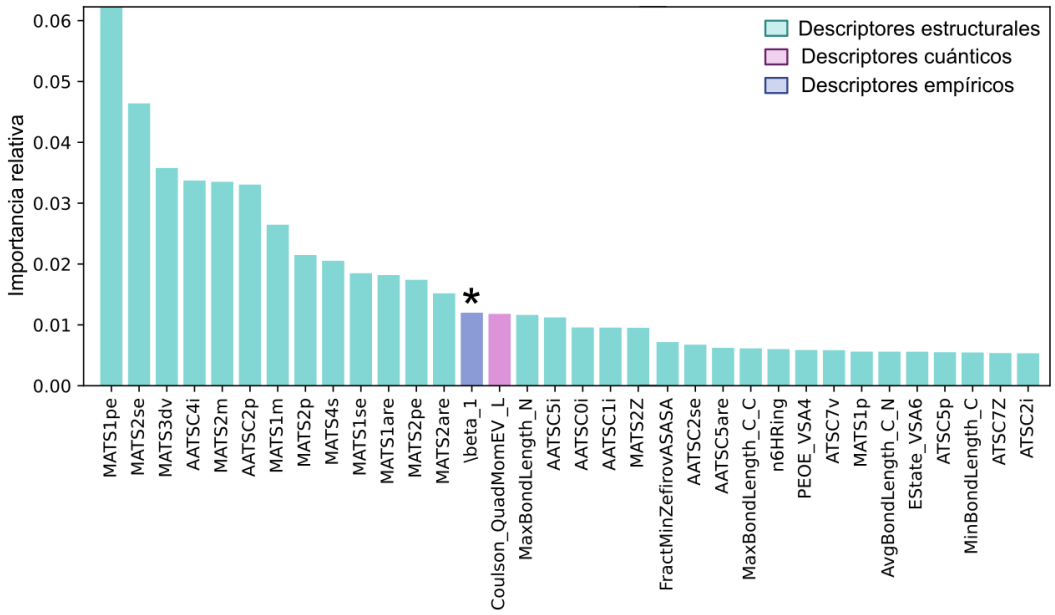


Figura 5.4. Figura en donde se muestra la importancia relativa de los 34 descriptores más importantes para predecir el tiempo de vida de fluorescencia. Para el entrenamiento se empleó el modelo de "RandomForestRegressor" con una reducción de dimensión a 300 descriptores. En donde las barras de color verde corresponden a descriptores estructurales, las barras color rosa a descriptores cuánticos y las barras color azul a descriptores empíricos. Además, un asterisco (*) sobre las barras indica que éstas corresponden a un descriptor del disolvente.

Sistema 2

Tabla 5.3. Tabla en donde se muestra una comparación del desempeño del modelo "RandomForestRegressor" para predecir la energía de absorción (E_{abs}) empleando diferentes descriptores tanto para el cromóforo como para el disolvente. Para la evaluación se reportan dos métricas, coeficiente de correlación r^2 y el error absoluto medio MAE para el conjunto de prueba. Para el entrenamiento de los modelos se emplearon el mismo tipo de descriptores tanto para el cromóforo como para el disolvente, a excepción de los descriptores empíricos, los cuales solo están definidos para el disolvente. En la novena columna se muestran los resultados de la reducción de dimensión, donde el primer valor corresponde al número de descriptores antes de la reducción y el segundo valor corresponde al número de descriptores después de la reducción de dimensión.

	Estructurales	Orbitales moleculares	Densidad electrónica	Vibracionales	QTAIM Átomos	QTAIM Grupos	Empíricos (disolvente)	Reduccion de Dimensión		r^2 (Prueba)	MAE [eV] (Prueba)
1	✓	✓	✓	✓	✓		✓	622	190	0.950	0.055
2	✓	✓	✓	✓		✓	✓	771	200	0.938	0.072
3	✓							368	190	0.0.854	0.081
4	✓						✓	378	195	0.930	0.065
5		✓	✓	✓				158	100	0.919	0.059
6					✓			96	80	0.847	0.086
7		✓	✓	✓	✓			249	160	0.894	0.073
8						✓		167	40	0.846	0.120
9		✓	✓	✓		✓		340	100	0.930	0.070
10		✓				✓		172	40	0.891	0.076
11		✓				✓		172	100	0.986	0.078

Tabla 5.4. Tabla en donde se muestra una comparación del desempeño del modelo "RandomForestRegressor" para predecir la energía de emisión (E_{em}) empleando diferentes descriptores tanto para el cromóforo como para el disolvente. Para la evaluación se reportan dos métricas, coeficiente de correlación r^2 y el error absoluto medio MAE para el conjunto de prueba. Para el entrenamiento de los modelos se emplearon el mismo tipo de descriptores tanto para el cromóforo como para el disolvente, a excepción de los descriptores empíricos, los cuales solo están definidos para el disolvente. En la novena columna se muestran los resultados de la reducción de dimensión. Donde el primer valor corresponde al número de descriptores antes de la reducción y el segundo valor corresponde al número de descriptores después de la reducción de dimensión.

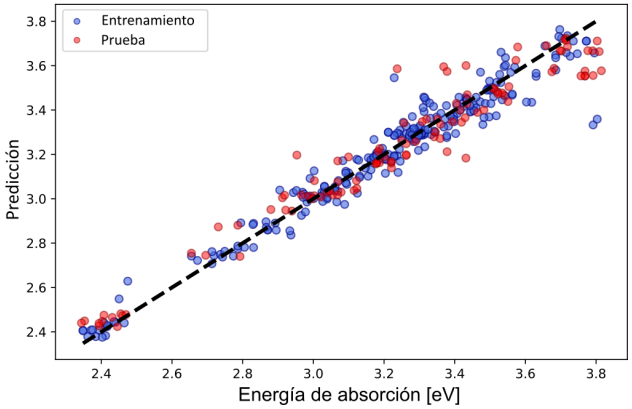
	Estructurales	Orbitales moleculares	Densidad electrónica	Vibracionales	QTAIM Átomos	QTAIM Grupos	Empíricos (disolvente)	Reduccion de Dimensión		r^2 (Prueba)	MAE [eV] (Prueba)
1	✓	✓	✓	✓	✓		✓	622	120	0.888	0.070
2	✓	✓	✓	✓		✓	✓	771	250	0.939	0.057
3	✓							368	180	0.862	0.082
4	✓						✓	378	175	0.902	0.070
5		✓	✓	✓				158	80	0.737	0.092
6					✓			96	80	0.776	0.107
7		✓	✓	✓	✓			249	160	0.755	0.110
8		✓	✓	✓		✓		340	100	0.851	0.082
9						✓		167	100	0.856	0.095
10		✓				✓		172	100	0.823	0.098
11						✓		167	35	0.856	0.094

Tabla 5.5. Tabla en donde se muestra una comparación del desempeño del modelo "RandomForestRegressor" para predecir el desplazamiento de Stokes empleando diferentes descriptores tanto para el cromóforo como para el disolvente. Para la evaluación se reportan dos métricas, coeficiente de correlación r^2 y el error absoluto medio MAE para el conjunto de prueba. Para el entrenamiento de los modelos se emplearon el mismo tipo de descriptores tanto para el cromóforo como para el disolvente, a excepción de los descriptores empíricos, los cuales sólo están definidos para el disolvente. En la novena columna se muestran los resultados de la reducción de dimensión. Donde el primer valor corresponde al número de descriptores antes de la reducción y el segundo valor corresponde al número de descriptores después de la reducción de dimensión.

	Estructurales	Orbitales moleculares	Densidad electrónica	Vibracionales	QTAIM Átomos	QTAIM Grupos	Empíricos (disolvente)	Reduccion de Dimensión		r^2 (Prueba)	MAE [eV] (Prueba)
1	✓	✓	✓	✓	✓		✓	622	160	0.911	0.046
2	✓	✓	✓	✓		✓	✓	771	250	0.923	0.050
3	✓							368	165	0.864	0.056
4	✓						✓	378	100	0.904	0.048
5		✓	✓	✓				158	55	0.912	0.047
6					✓			96	55	0.902	0.051
7		✓	✓	✓	✓			249	150	0.884	0.057
10		✓	✓	✓		✓		340	200	0.860	0.056
8						✓		167	120	0.928	0.050
9		✓				✓		172	100	0.860	0.058

(a)

	r^2	Error absoluto medio	Error cuadrático medio
Validación	0.941 +/- 0.101	0.047 +/- 0.021	0.069 +/- 0.05
Prueba	0.938	0.072	0.098



(b)

	r^2	Error absoluto medio	Error cuadrático medio
Validación	0.946 +/- 0.039	0.048 +/- 0.014	0.063 +/- 0.017
Prueba	0.939	0.057	0.072

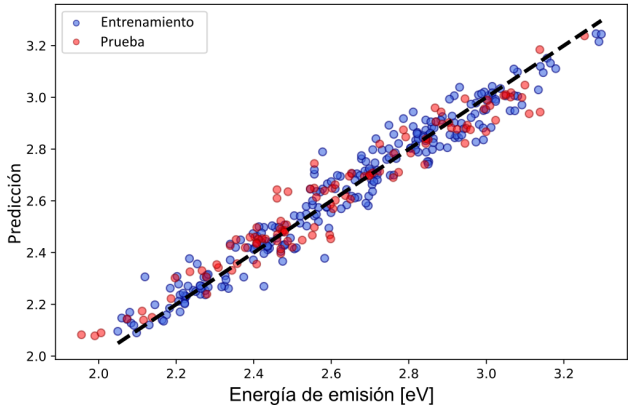


Figura 5.5. Gráficas de valor predicho contra el valor experimental para la energía de absorción (a) y emisión (b) (Sistema 1). Para ello se empleó el modelo de "RandomForestRegressor" junto con todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente. Previo a entrenar el modelo, se realizó una reducción de dimensión, teniendo un total de 250 (a) y 200 (b) descriptores después de la reducción de dimensión. En la parte superior de la figura se incluye una tabla con los resultados obtenidos para las tres métricas: coeficiente de correlación (r^2), error absoluto medio y error cuadrático medio, tanto para la validación como para la prueba. En la gráfica se muestra el valor predicho contra el valor experimental para el conjunto de datos de entrenamiento (círculos azules) y para el conjunto de datos de prueba (círculos rojos).

(a)

	r^2	Error absoluto medio	Error cuadrático medio
Validación	0.953 +/- 0.061	0.031 +/- 0.016	0.041 +/- 0.024
Prueba	0.923	0.05	0.065

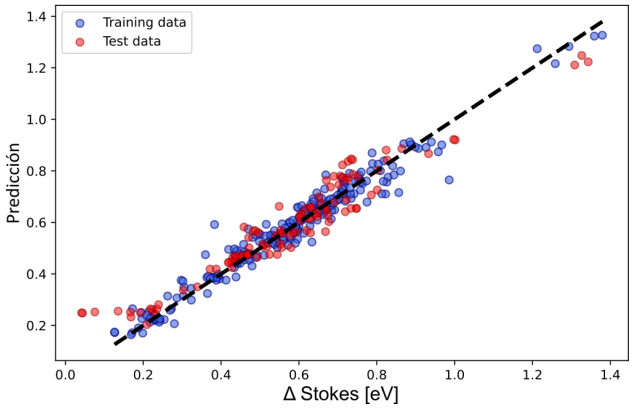
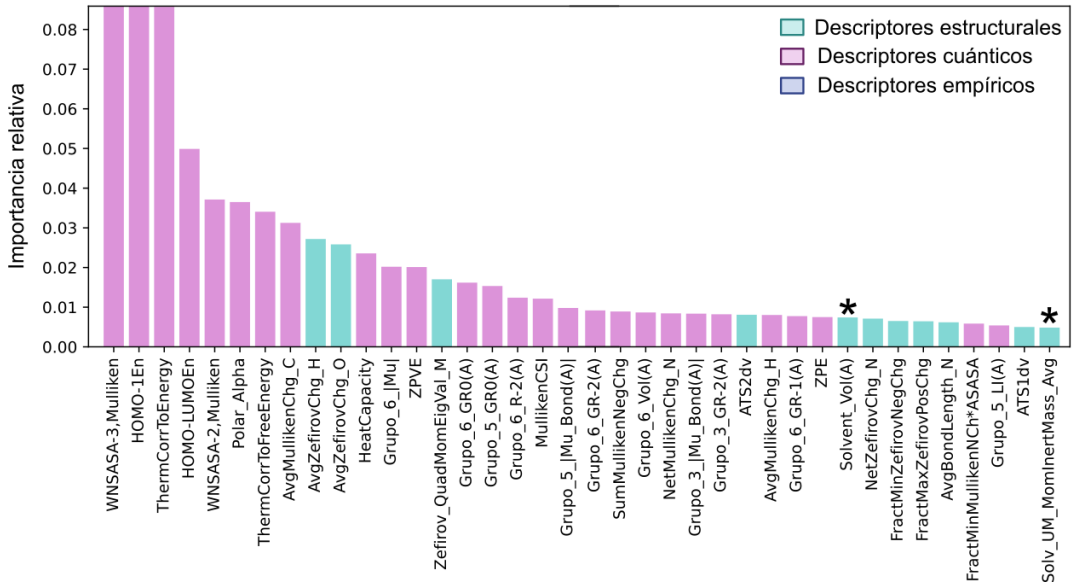


Figura 5.6. Gráficas de valor predicho contra el valor experimental para desplazamiento de Stokes. Para ello se empleó el modelo de "RandomForestRegressor" junto con todos los tipos de descriptores (estructurales, cuánticos y empíricos) tanto para el cromóforo como para el disolvente. Previo a entrenar el modelo, se realizó una reducción de dimensión, teniendo un total de 250 descriptores después de la reducción de dimensión. En la parte superior de la figura se incluye una tabla con los resultados obtenidos para las tres métricas: coeficiente de correlación (r^2), error absoluto medio y error cuadrático medio, tanto para la validación como para la prueba. En la gráfica se muestra el valor predicho contra el valor experimental para el conjunto de datos de entrenamiento (círculos azules) y para el conjunto de datos de prueba (círculos rojos).

(a)



(b)

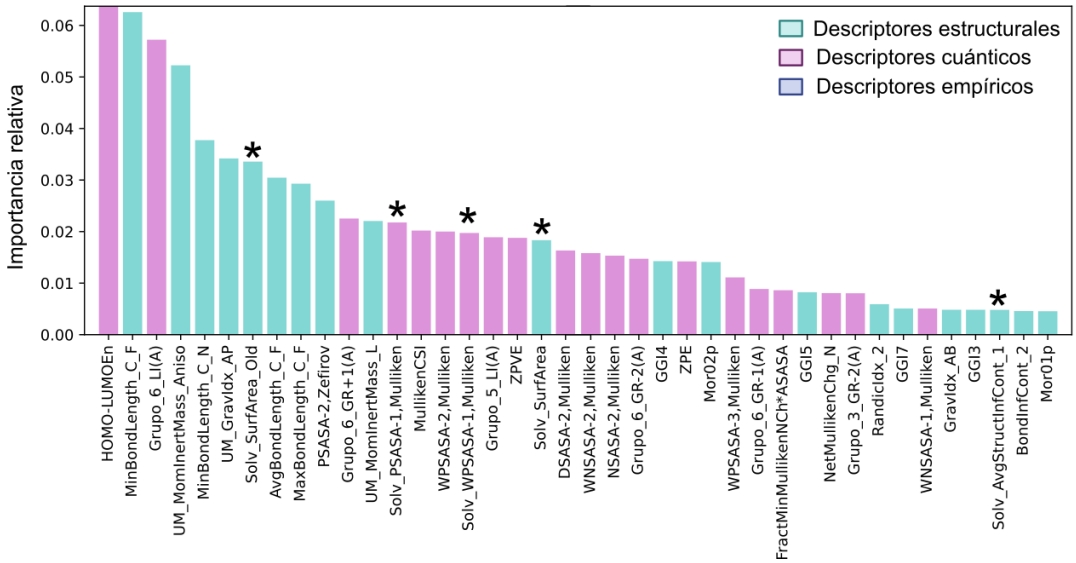


Figura 5.7. Figura en donde se muestra la importancia relativa de los 38 descriptores más importantes para predecir el desplazamiento de Stokes. Para el entrenamiento se empleó el modelo de "RandomForestRegressor" con una reducción de dimensión a 300 descriptores. En donde las barras de color verde corresponden a descriptores estructurales, las barras rosas a descriptores cuánticos y las barras azules a descriptores empíricos, además un asterisco (*) indica que es un descriptor del disolvente.

(a)

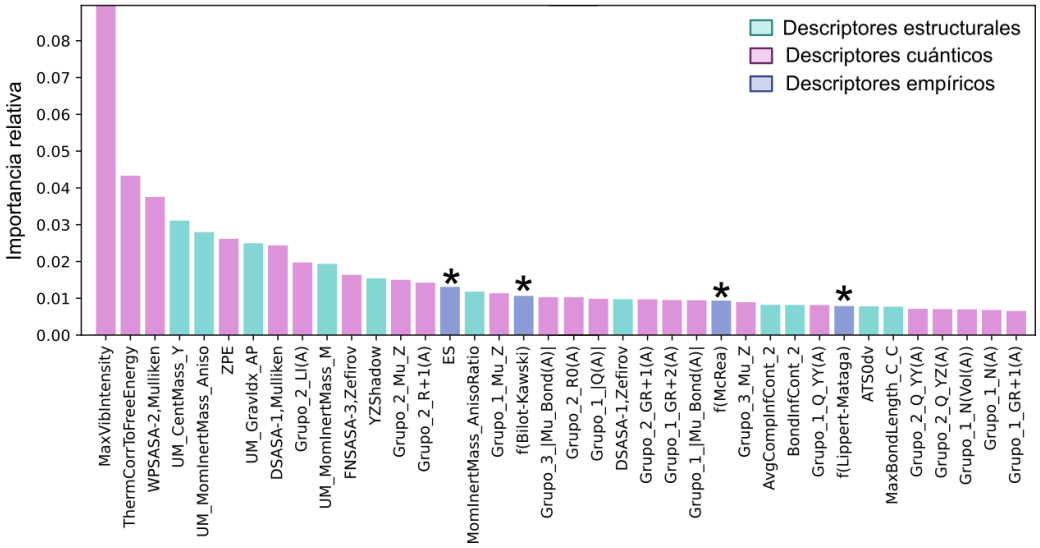


Figura 5.8. Figura en donde se muestra la importancia relativa de los 38 descriptores más importantes para predecir la energía de absorción (a) y emisión (b). Para el entrenamiento se empleó el modelo de "RandomForestRegressor" con una reducción de dimensión a 250 descriptores para (a) y 200 descriptores para (b). En donde las barras de color verde corresponden a descriptores estructurales, las barras rosas a descriptores cuánticos y las barras azules a descriptores empíricos, además un asterisco (*) indica que es un descriptor del disolvente.

Apéndice B

Descriptores estructurales

Los descriptores estructurales incluidos por el programa "CODESSA" se clasifican en los siguientes grupos:

- Cuenta de átomos (Número de átomos, número de átomos por elemento, etc.)
- Masa (Peso molecular, Índices gravimétricos, centro de masa, momentos de inercia, etc.)
- Cuenta de enlaces (Número de enlaces simples, dobles y triples, número de enlaces para cada elemento, etc.)
- Valencia (valencia mínima, máxima promedio y suma para todos los átomos y elemento por elemento, etc.)
- Topológicos
    - Índice Balaban J
    - Contenido informacional
    - Índice Kier, Hall, Randic y Wiener
- Superficie (Área y Volumen) Para ello se realizó la suma del área o volumen (van der Waals) correspondiente a cada átomo en la molécula y se le restó la suma se los traslapes entre átomos [66].
- Geométricos (Longitudes de enlace, indices gravimétricos, etc.)
- Físicos (Carga neta y multiplicidad)
- Nuclear (ESP nuclear para todos los átomos y por cada elemento)
- Electrostática (Cargas parciales empíricas de Zefirov's)

Los descriptores estructurales incluidos por el programa “MORDRED” se clasifican en los siguientes módulos:

- Índice ABC
- Acido Base
- Matriz Adyacente
- Cuenta de átomos
- Aromáticos
- Autocorrelación
- Índice Balaban J
- Matriz Barysz
- Cuenta de átomos
- Tipo de carbono
- Chi
- Constitucional
- Matriz Detour
- Matriz de distancias
- Índice de conectividad excéntrica
- Índices de “ExtendedTopochemicalAtom”
- Complejidad de fragmentos
- “Molecular Frameworks”
- Índices gravitacionales
- Enlaces de hidrógeno
- Índice Kappa
- Tipo de fragmento
- Distancias moleculares
- Molecularidad
- Momentos de inercia
- Cuenta de caminos
- Índices MoRSE
- Polarizabilidad
- Cuenta de anillos
- Enlaces rotables
- Entropia y grados de libertad
- Índices topológicos
- Carga topológica
- Peso
- Índice Wiener y Zagreb

Descriptores cuánticos (Función de onda)

La siguiente lista muestra las subcategorías de los descriptores cuánticos implementados.

- Electrostáticas
    - Cargas parciales atómicas de Mulliken
    - Orden de enlace de Mulliken (Orden de enlace mínimo, máximo, promedio y suma para todos

los átomos y elemento por elemento, etc.)

- Polarizabilidad e hiperpolarizabilidad
- Cargas parciales superficiales “Charged partial surface area” (CPSA)
- Orbitales moleculares (Energías HOMO-1, HOMO, LUMO+1, LUMO y HOMO - LUMO)
- Índices de reactividad Fukui (Nucleófilo, electrófilo y de un electrón)

- Vibracionales
  - Entalpia vibracional rotacional y traslacional
  - Entropia vibracional rotacional y traslacional
  - Energía de punto cero
  - Capacidad calorífica

Descriptores cuánticos (QTAIM)

Las propiedades calculadas para los descriptores QTAIM (para el átomo A en la molecula) tanto para el análisis por grupos como el tipo de átomos son los siguientes:

- Carga q(A)
- número de electrones localizados N() y deslocalizados LI(A)
- Área y Volumen
- Magnitud del momento dipolar total $|\mu|$, interno $|\mu\text{Intra}|$ y del enlace $|\mu\text{Bond}|$.
- Componentes del momento (X, y, Z) dipolar total $|\mu_i|$, interno $|\mu\text{Intra}_\text{i}|$ y del enlace $|\mu\text{Bond}_\text{i}|$
- Componentes del tensor cuadripolar $Q_{i,j}$
- Magnitud del tensor cuadripolar $||Q(\text{A})||$
- Momentos Radiales R+k(A)
- Distribución radial GR+k(A)

Donde $i, j = \{x, y, z\}$ y k = -2, -1, 0, 1, 2. Para el caso de los descriptores QTAIM por átomos, estas propiedades se calcularon para cada tipo de átomo haciendo el correspondiente suma, promedio máximo o mínimo. Mientras que para el caso de los descriptores QTAIM por grupos, se realizó la suma de estas propiedades para los átomos pertenecientes al grupo.

Descriptores empíricos

- Propiedades físicas
    - Constante dieléctrica (ε)
    - Índice de refracción (n)
- Parámetros empíricos
    - Dispersión he inducción (DI)
    - Interacciones electrostáticas (ES)
    - Interacción soluto disolvente de Lewis (α_1)
    - Interacción soluto disolvente por puente de hidrógeno (β_1)
- Teorías solvatocrómicas
    - $f(n,\varepsilon)$ Lippert-Mataga
    - $f(n,\varepsilon)$ Bilot-Kawski
    - $f(n,\varepsilon)$ McRea
    - $g(n,\varepsilon)$ Bliot-Kawski