Universidad Nacional Autónoma de México

Programa de Maestría y Doctorado en Ciencias Químicas

Estudio QSAR basado en la densidad electrónica de compuestos anticancerígenos

Tesis para optar por el grado de :

Maestro en Ciencias

Presenta:

Q. F. B. Rodrigo Galindo Murillo

Tutor: Dr. Fernando Cortés Guzmán 2008

Agradecimientos

Agradecemos el soporte de los proyectos PAPIIT: IN216205, IN207107 y IN500107 y el proyecto CONACyT 52619. Agradecemos especialmente a la Dirección General de Servicios de Cómputo Académico (DGSCA), UNAM, por el tiempo brindado en la super computadora KanBalam.

Esta tesis fue realizada en el Laboratorio 203 y 210, Edificio F del Departamento de Química Orgánica de la Facultad de Química de la UNAM.

Parte del trabajo de esta tesis se presentó como poster en la Gordon Research Conference 2007, Electron Distribution and Chemical Bonding, realizada en Boston, MA.

Dedicatoria

Sandrita, Sebastián, Papá, Mamá, Quique y Elena, Ana y Alex, Ale y Ale, todos mis sobrinos.

Dr. Fernando Cortés, Dr. Jesús Hernández, Dr. Rafael Moreno, Dra. Isabel Gracia, Dra. Lena Ruiz, Marco Revilla, Maestra María Isabel Vazquez, Rosy Viñas, Mary Salazar.

Benito Mastropiero, Arsenio Díaz, Raymundo Jauregui, Xavier Moreno, Francisco Bustamante, David Villalobos, Sonia Santana, Jorge Flores.

Luis Miguel de los Ríos, Enrique Ramos, Maurizio Tazzer, Federico Miranda, Erick Sánchez, Manuel Reynaga, Juan Carlos Gómez, Jorge Vieyra, Gilberto Cardoso.

Jose Tenopala, promesa cumplida.

A Dios.

Índice general

Introducción

La búsqueda de nuevos compuestos o fármacos capaces de generar una respuesta terapéutica ha encontrado muchos métodos que varían desde estudios *in vivo*, mediciones de diferentes propiedades físicas y químicas, ensayos *in vitro* y cálculos teóricos. La generación de nuevas estructuras que produzcan la mayor actividad terapéutica posible con mayor especificidad es fundamental.

Actualmente, el poder de cómputo ha alcanzado un nivel donde es fácil medir y simular propiedades químicas con gran precisión. Es por esto que una opción muy viable para la búsqueda de nuevos fármacos es por medio de cálculos teóricos.

Para agilizar la búsqueda de relaciones estructura-actividad en el diseño de fármacos, se utilizan descriptores electrónicos sencillos como tipos de fuerzas intermoleculares (enlace covalente, iónico, interacciones de van der Waals, etc.), hidrofobicidad, interacciones electrostáticas, etc. Estos descriptores, junto con las propiedades de actividad biológica de las moléculas, son procesados con métodos estadísticos especiales para establecer las características más importantes de la molécula, que generarán actividad biológica.

Los descriptores usados en relaciones cuantitativas estructura-actividad tienen que codificar una variedad de propiedades físicas y químicas para poder construir modelos confiables. Este trabajo utiliza la densidad electrónica para obtener descriptores electrónicos que ayuden a relacionar la estructura con la actividad biológica. El uso de la densidad electrónica y las propiedades que se puedan extraer de ella nos permiten comparar una gran cantidad de compuestos con posible actividad biológica, siempre y cuando consten de una estructura en común. El análisis se realizó utilizando la teoría de átomos en moléculas sobre 3 familias de compuestos: 39 ácidos benzoicos, 21

Casiopeínas® y 121 fenilbencimidazoles.

Los ácidos benzoicos sustituidos son un conjunto ampliamente estudiado. Utilizamos su índice de acidez, pKa como descriptor de la actividad. Los fenilbencimidazoles y las Casiopeínas®, complejos de cobre patentados por la UNAM, ambos son antineoplásicos.

En los tres casos se emplea la medición de la cantidad mínima de substancia para causar un efecto biológico. Al extraer los descriptores electrónicos de ρ a los ácidos benzoicos y realizar el análisis estadístico de mínimos cuadrados parciales encontramos la región de la molécula que explica la actividad con altos índices de confiabilidad y predicción. Esto nos ayudó a comprobar que, efectivamente, la densidad electrónica nos da información para explicar el comportamiento del sistema. Después se aplicó la misma metodología para los fenilbencimidazoles y las Casiopeínas®. Un análisis de componentes principales nos indica las relaciones entre las moléculas y después se efectuó el análisis de mínimos cuadrados parciales sobre cada una de las familias identificadas por medio del análisis de componentes principales. Comprobamos que la región activa de los ácidos benzoicos es en el carbono del ácido. Para los fenilbencimidazoles se detectó que la región activa de la molécula pertenece a la región de los nitrógenos en el fenilbencimidazol, y para las Casiopeínas®, la actividad se reparte entre los lugares directamente coordinados con el metal con el acetilacetonato y la fenantrolina.

Capítulo 1

Antecedentes

1.1. Relación estructura-actividad

En la industria farmacéutica, uno de los objetivos más importantes de la investigación es identificar estructuras químicas, que tengan el potencial de convertirse en medicamentos para el uso público. Típicamente, esta investigación trata de encontrar relaciones entre propiedades químicas y actividades terapéuticas. Una manera de investigar estas relaciones es con el uso de modelos matemáticos semi-empíricos en donde una función biológica de una serie de compuestos es expresada como una función de sus propiedades fisicoquímicas. Este tipo de expresiones es conocido como relacion cuantitativa entre la estructura y la actividad. (Quantitative structure-activity relationship, QSAR).

Un modelo QSAR es capaz de predecir la actividad o respuesta de compuestos aún no probados. Además es capaz de revelar qué propiedad química regula cierto tipo de actividad biológica, y cómo se tiene que modificar esa propiedad para incrementar la actividad.

La metodología involucrada en los estudios QSAR fue introducida por Hansch a principios de la década de los 60's [1, 2]. Esta estrategia está fundamentada en que la diferencia estructural de los compuestos es responsable de la diferencia de la actividad biológica (o la propiedad molecular) y los parámetros estructurales se obtienen a través de un análisis de regresión múltiple.

El desarrollo de un modelo QSAR se puede formalizar de la siguiente manera: Se monitorean las actividades biológicas para cierto tipo de compuestos, estas actividades constituyen la matriz de datos $\mathbf{Y} = (NxM)$, donde N es el número de compuestos y M son las variables de respuesta. También, para el mismo conjunto de compuestos, se obtienen descriptores que reflejen las propiedades estructurales y químicas. Este nuevo conjunto de datos forma la matriz $\mathbf{X} = (NxK)$, donde N es el mismo número de compuestos y K es el número de descriptores. Los datos biológicos ($\mathbf{Y}$) son modelados con los datos estructurales ($\mathbf{X}$) en términos de la ecuación (1.1).

$$Y = F(X, \beta) + E \tag{1.1}$$

donde $F(\mathbf{X}, \beta)$ representa la parte sistemática de los datos, la correlación entre los descriptores químicos y las respuestas biológicas, y E los residuales, como errores de medición o imperfecciones en el modelo. En la ecuación (1.1), β corresponde a los coeficientes de regresión, que indican hasta qué grado un descriptor influye en el modelado de cierta respuesta. De manera similar, podemos expresar los modelos QSAR como $P_i = k(D_1, D_2, \cdots, D_n)$, donde P_i son las actividades biológicas (o alguna otra propiedad de interés de las moléculas). $D_1, D_2, \cdots, D_n$ son propiedades estructurales calculadas (o experimentales) de los compuestos. En la expresión, k, representa alguna transformación matemática establecida empíricamente que habría que aplicar a los descriptores para calcular el valor de la propiedad P_i de todas las moléculas. La relación entre los descriptores D y la propiedad P puede ser lineal, donde la propiedad puede calcularse directamente a partir del valor de los descriptores, o no lineal, donde el valor de los descriptores se utilizan para caracterizar la similitud química entre las moléculas, lo cual se usa para predecir la propiedad P de todos los compuestos [3].

Desde los primeros trabajos de Hansch, se han desarrollado diferentes estrategias para hacer estudios QSAR [4, 5]. Los diferentes métodos que se han desarrollado pueden analizarse desde dos puntos de vista: 1) los tipos de parámetros estructurales que se utilizan para caracterizar a las moléculas, desde simples fórmulas químicas a estructuras tridimensionales y, 2) los procedimientos matemáticos que se emplean para obtener las relaciones cuantitativas entre los descriptores moleculares y la actividad biológica.

De acuerdo al origen de los descriptores moleculares usados en los cálculos, los métodos QSAR pueden dividirse en tres grupos. Un grupo usa un

número relativamente pequeño de propiedades fisicoquímicas y parámetros que describen, por ejemplo, efectos hidrofóbicos, estéricos y electrostáticos. Estos métodos son conocidos en la literatura con el nombre de análisis de Hansch. Métodos más recientes se basan en características cuantitativas de gráficas moleculares. Debido a que las gráficas moleculares o fórmulas estructurales son bidimensionales, a estos métodos se les refiere como estudios QSAR en dos dimensiones o QSAR-2D [6].

Un tercer grupo de método se basa en los descriptores obtenidos de la contrucción tridimensional de las estructuras de las moléculas. A estos estudios se les conoce como QSAR en tres diensiones o QSAR-3D [6, 7]. Estos métodos se han hecho cada vez más populares con el desarrollo rápido y preciso de métodos computacionales para generar conformaciones tridimensionales y alineaciones de las estructuras químicas. Probablemente el ejemplo más conocido de QSAR-3D es CoMFA (Comparative Molecular Field Analysis) [8] y el CoMSIA (Comparative Molecular Similarity Indices Analysis) [9], explicados brevemente en las siguientes secciones.

1.1.1. CoMFA

La metodología utilizada en el análisis CoMFA es una técnica QSAR-3D que permite diseñar y predecir la actividad de moléculas. La base de datos con las propiedades de las moléculas en estudio, el data-set, se alinea en el espacio 3D siguiendo diferentes metodologías. Las técnicas de superimposición incluyen la maximización estérica de las moléculas [10], aquellas basadas en datos cristalográficos [11], basados en farmacóforos [12, 13], algoritmos de alineamiento estérico y electrostático [14], automatización de ajuste de campos [15] y aquellos utilizando programas de mapeo de farmacóforos [16]. Después de que se selecciona la técnica de superimposición, se calculan las cargas de cada una de las moléculas con un nivel de teoría aceptable. Los campos estéricos y electrostáticos se calculan por medio de interacciones con un átomo de prueba (probe atom) en una serie de puntos alrededor de la base de datos en el espacio tridimensional. Después se correlacionan las energías del campo con las propiedades de interés por medio de mínimos cuadrados parciales. El análisis estadístico arroja las relaciones de la actividad o lo que se haya seleccionado con las cargas detectadas por las interacciones del átomo de prueba.

1.1.2. CoMSIA

De la misma manera que para CoMFA, el método CoMSIA utiliza un átomo de prueba, pero en esta ocasión se detecta la similitud entre los efectos estéricos, electrostáticos, hidrofóbicos, lugares electroatrayentes y electrodonadores. CoMSIA detecta índices de similaridad para una molécula j con i átomos en una referencia q de acuerdo a la expresión:

$$A^q_{F,k}(j) = -\sum \omega_{probe,k}\omega_{i,k}e^{-\alpha r^2_{iq}}$$

donde q es el punto de referencia para la molécula j; ω_{ik} es el valor de la propiedad fisicoquímica k del átomo i; $\omega_{probe,k}$ son las características del átomo de prueba; α es un factor de atenuación; r_{iq} es la distancia entre el átomo de prueba en el punto de referencia q y el átomo i de la molécula en estudio. CoMSIA utiliza el mismo procedimiento estadístico de mínimos cuadrados parciales para evaluar los resultados.

Los estudios QSAR pueden también ser clasificados por el tipo de métodos de correlación usados en el desarrollo de los modelos. Estos pueden ser métodos lineales, como es el caso de la regresión lineal o regresión lineal múltiple, y no lineales [17].

Los diferentes modelos QSAR tienen sus propias ventajas y desventajas. Por ejemplo, los métodos QSAR-3D tienen la ventaja de mostrar sus resultados en forma gráfica que pueden interpretarse fácilmente en términos de las interacciones estéricas y electrostáticas importantes para que el ligando se una al receptor. Sin embargo, estos métodos requieren de la alineación de las estructuras en tres dimensiones, lo cual es subjetivo, consume tiempo y dificulta el análisis de bases de datos grandes. Por otra parte, los métodos QSAR-2D son mucho más rápidos y más fácilmente adaptables a la automatización. Esto se debe a que no requieren búsqueda conformacional y alineación de las estructuras. De esta forma, los métodos QSAR-2D son mejores para analizar a una gran cantidad de compuestos y hacer búsquedas computacionales en bases de datos. Sin embargo, la interpretación de los modelos en términos químicos comunes es, con frecuencia, difícil o prácticamanente imposible [6].

Después de desarrollar un modelo que correlacione la actividad biológica con los descriptores, la siguiente etapa de un estudio QSAR es la validación de los modelos. Esta etapa es necesaria debido a la posibilidad de que las correlaciones encontradas sean producto de la casualidad [18]. La validación consiste en evaluar la capacidad que tiene el modelo QSAR para predecir con exactitud la propiedad de interés (por ejemplo, la actividad biológica) de los compuestos que no fueron utilizados en el desarrollo del modelo. La predicción es una de las características más importantes de los modelos QSAR [19]. Al grupo de compuestos con los que se desarrolla el modelo se denomina *grupo de entrenamiento*, mientras que, al grupo con que se valida, se le llama *grupo de prueba*. Se ha sugerido que el *grupo de prueba* debe contener al menos cinco compuestos [20]. En el caso de que, al momento de desarrollar los modelos QSAR, no se cuente con un grupo de prueba, se puede dividir el grupo de datos original en los *grupos de entrenamiento* y *de prueba*. Una alternativa a esto es la llamada validación cruzada [5]. En esta estrategia se desarrollan varios modelos QSAR en donde se han eliminado uno o varios compuestos del grupo de datos que después son predichos por los correspondientes modelos. La validación cruzada más común es la llamada dejar-uno-afuera en donde cada compuesto se excluye una vez. Si se tienen n

compuestos, se derivan entonces n modelos y las n predicciones se comparan con los valores experimentales. Como resultado se obtiene un coeficiente de correlación cruzado q^2 [5]. La capacidad de predicción de un método QSAR puede considerarse como el criterio más importante para evaluar la eficiencia de este método.[21].

Para evaluar qué tan robusto es un modelo QSAR, se puede usar la estrategia de asignar valores aleatorios a la actividad biológica y repetir las predicciones. Si el modelo es robusto, las predicciones con las actividades reales deben ser estadísticamente superiores que cuando se derivaron los modelos con valores aleatorios [21].

Las diferentes metodologías QSAR han tenido aplicaciones exitosas al diseño de fármacos. Numerosos ejemplos de estas aplicaciones pueden encontrarse en la literatura [5, 21] Hoy en día, una de las aplicaciones más importante que tienen los estudios QSAR es la predicción de una gran cantidad de compuestos en forma rápida. Esto se debe al creciente desarrollo de la química combinatoria y a las pruebas biológicas de alto rendimiento. De esta manera se puede predecir en forma virtual la actividad de compuestos que podrán ser sintetizados posteriormente, de manera que los estudios QSAR sirven como guía para la creación de bases de datos. Todo esto hace que en la era moderna de la química farmacéutica, los estudios QSAR constituyan una de las herramientas más importantes del diseño de fármacos asistido por computadora [21].

1.2. Métodos para el análisis de la función de onda

Durante mucho tiempo, los conceptos químicos han surgido de la observación, de la percepción de la experiencia humana y ha existido una brecha entre los conceptos químicos y las teorías de la física. Hoffmann en uno de sus libros menciona que a pesar de que los campos de la química y la física están tan cercanos (la división tiene fines más educativos que reales), *"hay conceptos químicos que no son reducibles a la física, o que si se reducen pierden mucho del interés que tienen. Yo pediría al lector químico que considerara ideas como la de aromaticidad, acidez y basicidad, el concepto de grupo funcional, o de efecto del sustituyente. Los límites de esos conceptos tienden a desvanecerse cuando uno trata de definirlos sin ambigüedad, pero son de una utilidad fantástica para nuestra ciencia"* [22].

El problema planteado por Hoffmann radica en la aproximación teórica utilizada en la definición de los conceptos. Hasta hace algún tiempo, los análisis en química cuántica se hacían sólo a partir de energías, geometrías, características de los orbitales moleculares y frecuencias. Pero se estaba lejos de acercarse a los conceptos útiles en la química experimental. Surge entonces la pregunta: ¿existe una unión entre los conceptos químicos y la mecánica cuántica? Encontrar esta unión es uno de los retos actuales de la química.

Existen varios métodos para obtener una solución aproximada a la ecuación de Schrödinger con el fin de predecir y entender las propiedades de algunos sistemas de interés. En la mayoría de los casos, sin embargo, los químicos están interesados en algo más que la energía, la geometría y otras propiedades que pueden ser obtenidas directamente de dichos cálculos.

Se conoce que la función de onda Ψ contiene toda la información que se puede conocer de un sistema. Desafortunadamente, Ψ no puede ser analizada directamente ya que al menos se presentan dos problemas si uno quiere interpretar y entender Ψ para un sistema de partículas arbitrario. El primer problema es que Ψ es compleja y que no reside en el espacio real. Este problema se resuelve introduciendo el producto de Ψ con su complejo conjugado $\Psi\Psi^*$ o $|\Psi|^2$ la probabilidad de carga electrónica. El segundo problema es que Ψ se encuentra en un espacio multidimensional, donde cada electrón está descrito por cuatro coordenadas. Este problema se puede resolver integrando Ψ sobre todo el espacio, excepto el conjunto de tres coordenadas espaciales que

describen a un electrón y sumando sobre todas las coordenadas de espín. Esto da como resultado una función de probabilidad de densidad, que si se multiplica por el número de electrones, se obtiene la densidad electrónica $\rho(r)$.

Existen varios métodos para extraer información de ρ, entre los cuales se encuentra la Teoría de Átomos en Moléculas (AEM). A continuación se presenta una breve introducción de como la teoría de AEM traduce la información cuántica en información química.

1.3. Teoría de Átomos en moléculas

La teoría de Átomos en moléulas (AEM) fue desarrollada por el profesor Richard Bader en la Universidad de McMaster, Canadá, desde la década de los 70. Ha sido utilizada ampliamente para encontrar la explicación de muchos problemas químicos. Esta teoría es una extensión de la química cuántica para un átomo en una molécula, donde la densidad electrónica ρ es el medio para predecir las propiedades del átomo. Define a un átomo dentro de una molécula en el espacio real, lo que permite extraer información química a la densidad electrónica. En la figura 1.1 se muestra la densidad electrónica del etano, la densidad de la molécula en 3D y sobre un plano.

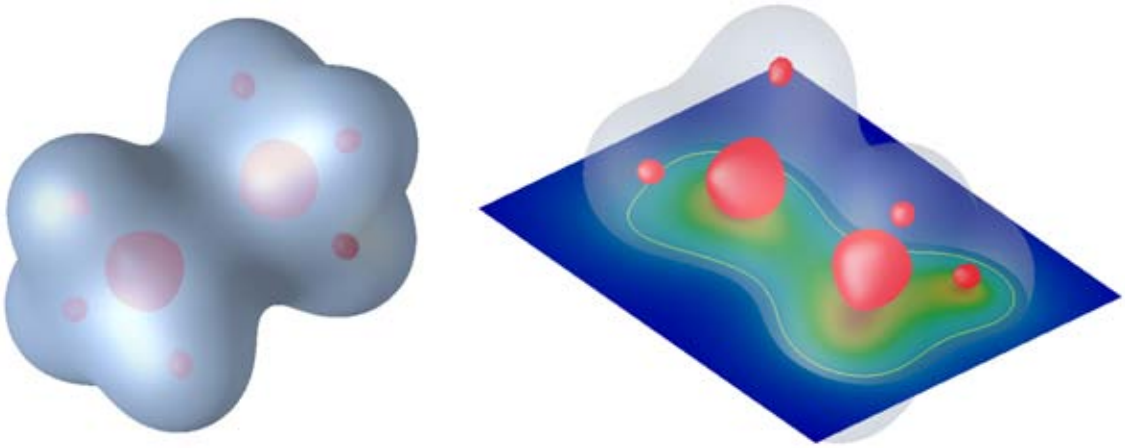

Figura 1.1: Densidad electrónica del etano (izquierda) y en un plano (derecha).

Para extraer informacion a la densidad electrónica se necesita estudiar el gradiente de la densidad ($\nabla\rho$) y no sólo la densidad directamente (ρ). El campo vectorial del gradiente es en esencia una infinita colección de gradientes. Un ejemplo del campo de gradiente de la densidad electrónica, ρ, en el plano molecular del metanal se muestra en la Figura 1.3

$\nabla\rho$ como todo campo vectorial tiene tres características:

- $\nabla\rho$ apunta en la dirección en la cual crece ρ.
- $\nabla\rho$ es perpendicular en cualquier lugar a una isosuperficie de ρ.

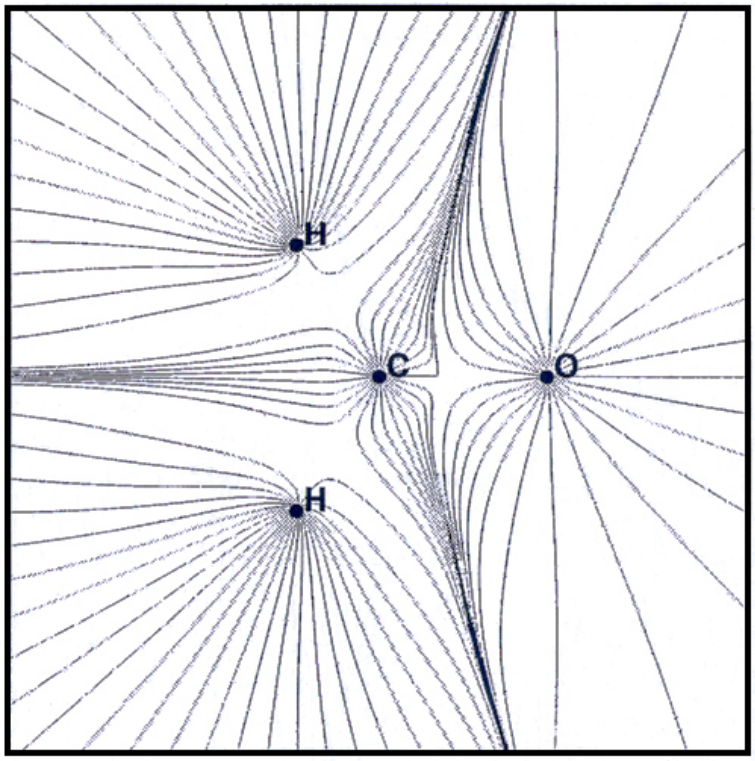

Figura 1.2: Un mapa del gradiente de la densidad electrónica para el plano que contiene los núcleos del metanal. Cada línea representa el camino del gradiente de $\nabla\rho$. Se muestran sólo los gradientes que terminan en cada núcleo.

- Al $\nabla\rho$ se le asocian líneas de flujo que tienen su origen y fin en puntos determinados del $\nabla\rho$.

Las líneas de flujo del $\nabla\rho$ no se cruzan pero se pueden encontrar en puntos donde $\nabla\rho = 0$. Esos puntos se conocen como puntos críticos (PC). Como la densidad electrónica es mayor en los núcleos que en sus alrededores las líneas de $\nabla\rho$ se originan en el infinito para terminar en el núcleo. Se puede decir que las líneas del $\nabla\rho$ son atraídas al núcleo, y es la razón por lo que se les denominan atractores nucleares y dominan una porción del espacio. Estas regiones dividen a las moléculas en segmentos que pueden identificarse con el concepto químico de átomo. La región dominada por el núcleo se conoce como contenedor atómico. Lo anterior se observa en la Figura 1.3. Con lo anterior se puede definir al átomo como: "La unión de un atractor (núcleo) y

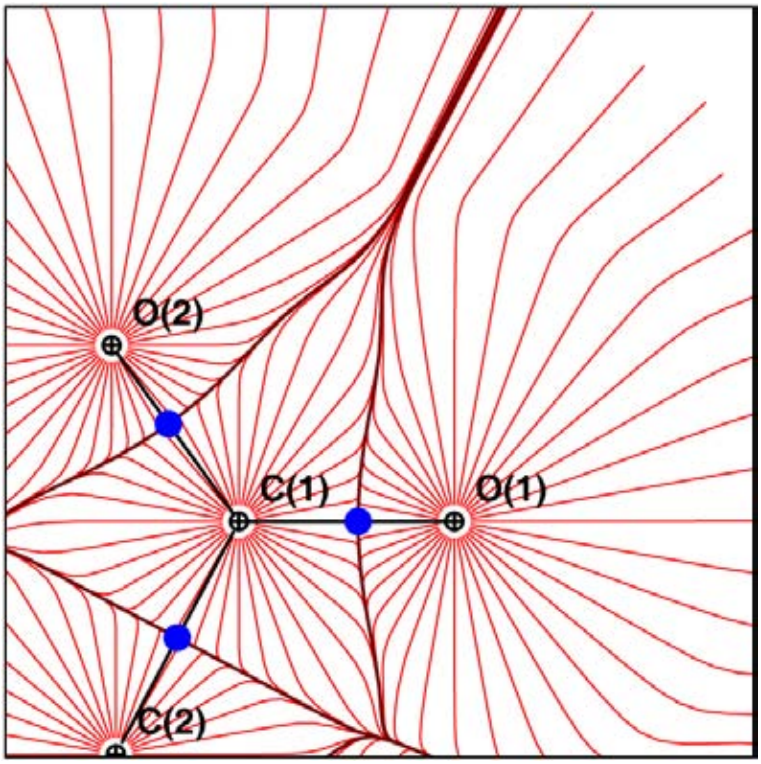

Figura 1.3: Las líneas rojas representan el gradiente de la densidad electrónica. Cuando dos líneas de gradiente se cruzan, forman el punto crítico de enlace entre dos átomos (en azul).

su contenedor atómico asociado" [23]. Cada átomo refleja las características del ambiente químico particular.

Además existe un conjunto de líneas de $\nabla\rho$ que inician en el infinito y terminan en un PC entre dos átomos, este conjunto de líneas constituyen una superficie interatómica (SIA), también llamada superficie de flujo cero. Esta superficie se distingue de cualquier otra en que no hay líneas del $\nabla\rho$ que crucen las SIA, como lo muestra la ecuación (1.2).

$$\nabla\rho(r) \cdot n(r) = 0 \; \forall \; r \; \epsilon \; S(r) \tag{1.2}$$

Además hay líneas del $\nabla\rho$ que se originan en este punto crítico y terminan en los núcleos. A este punto se conoce como punto crítico de enlace (PCE).

Este PCE es un punto de silla en la densidad electrónica, ya que es un máximo en la dirección de la SIA y un mínimo en la dirección de los núcleos. Las líneas que se originan en el PCE y terminan en los núcleos se conocen como lineas de interacción atómica (LIA). Las LIA se encuentran en cada par de núcleos que comparten una SIA (Figura 1.5). De esta manera, el enlace se puede definir de la siguiente manera: "Dos átomos están enlazados cuando comparten una superficie interatómica (SIA), y existe entre ellos un punto crítico de enlace (PCE) y una línea de interacción atómica (LIA) que los une"[23].

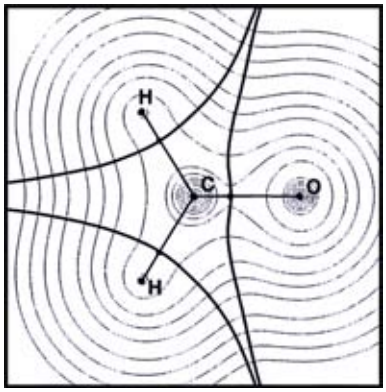

Figura 1.4: Mapa de contorno de la densidad electrónica del metanal. Se muestra también los puntos críticos de enlace entre los átomos, la superficie interatómica y las líneas de interacción atómica.

La LIA entre los átomos reproduce la conectividad encontrada experimentalmente. Cada característica topológica de la $\rho(r)$, ya sea un máximo, un mínimo o un punto de silla, está asociado con un punto crítico (PC) donde $\nabla\rho = 0$.

Uno puede diferenciar los máximos, mínimos y los puntos de silla considerando las segundas derivadas por medio del Hessiano de $\rho(r)$ que cuando se evalúa en un punto crítico localizado en r se escribe como:

$$A(r) = \left(\begin{array}{ccc} \frac{\partial^2\rho}{\partial x^2} & \frac{\partial^2\rho}{\partial x\partial y} & \frac{\partial^2\rho}{\partial x\partial z} \\ \frac{\partial^2\rho}{\partial y\partial x} & \frac{\partial^2\rho}{\partial y^2} & \frac{\partial^2\rho}{\partial y\partial z} \\ \frac{\partial^2\rho}{\partial z\partial x} & \frac{\partial^2\rho}{\partial z\partial y} & \frac{\partial^2\rho}{\partial z^2} \end{array} \right)_{r=r_c} \tag{1.3}$$

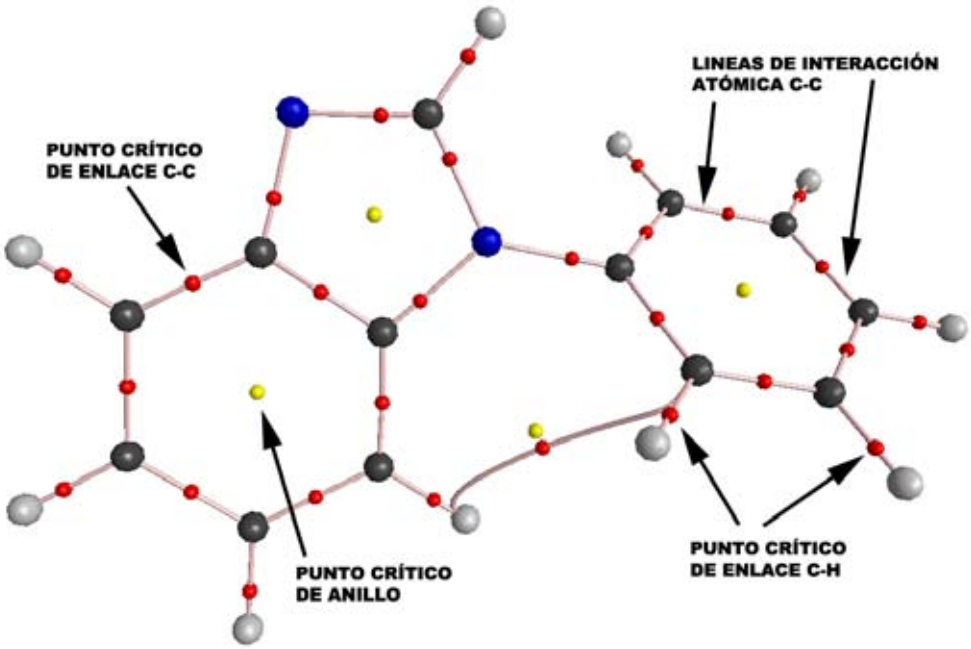

Figura 1.5: Gráfica molecular de un fenilbencimidazol mostrando las líneas de interacción atómica (líneas) y diferentes puntos críticos: nucleares (C = negros, N = azules, H = grises), de enlace (puntos rojos) y de anillo (puntos amarillos).

La matriz Hessiana puede diagonalizarse porque es real y simétrica. La diagonalización de $A(r_c)$ es equivalente a la rotación del sistema de coordenadas $r(x, y, z) \rightarrow r(x', y', z')$ substituyendo los ejes nuevos x', y', z' con los ejes de la curvatura del punto crítico. La rotación del sistema de coordenadas se lleva a cabo por medio de una transformación unitaria $r' = rU$, donde U es una matriz unitaria construida con valores propios de la expresión $Au_i = \lambda_i U_i (i = 1, 2, 3)$. Esto transforma a la matriz Hessiana a su forma diagonalizada expresada como:

$$A(r) = \begin{pmatrix} \frac{\partial^2\rho}{\partial x^2} & 0 & 0 \\ 0 & \frac{\partial^2\rho}{\partial y^2} & 0 \\ 0 & 0 & \frac{\partial^2\rho}{\partial z^2} \end{pmatrix}_{r=r_c} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \tag{1.4}$$

en donde λ_1, λ_2 y λ_3 son las curvaturas de la densidad con respecto a los tres ejes principales x', y', z'.

Una propiedad importante del Hessiano es que su traza es independiente de las rotaciones del sistema de coordenadas. El trazo del Hessiano de la densidad es conocido como el Laplaciano de la densidad $[\nabla^2\rho(r)]$ y cuando $x = x', y = y'$ y $z = z'$, se expresa con:

$$\nabla^2\rho(r) = \nabla \cdot \nabla\rho(r) = \underbrace{\frac{\partial^2\rho(r)}{\partial x^2}}_{\lambda_1} + \underbrace{\frac{\partial^2\rho(r)}{\partial y^2}}_{\lambda_2} + \underbrace{\frac{\partial^2\rho(r)}{\partial z^2}}_{\lambda_3} \tag{1.5}$$

La colección de las nueve segundas derivadas de $\rho(r)$, el Hessiano de la ρ, produce un conjunto de valores y vectores propios asociados.

Los primeros corresponden a las tres principales curvaturas de la $\rho(r)$ en el punto crítico, y los segundos a los ejes asociados. La curvatura, la segunda derivada de $\rho(r)$, es negativa en un máximo y positivo en un mínimo.

Los puntos críticos pueden clasificarse por medio de dos valores: w y σ. w es el número de curvaturas diferentes de cero de ρ en un cierto punto crítico. Un punto crítico que es $w < 3$ es matemáticamente inestable, va a desaparecer o bifurcarse bajo pequeñas perturbaciones de la densidad causadas por movimientos del nucleo. La presencia de un punto crítico con $w < 3$ no se encuentra generalmente en equilibrio, lo que se tiene en la mayoría de los casos es $w = 3$. El valor de σ corresponde a la suma algebraica de los signos de las curvaturas, por ejemplo, cada una de las tres curvaturas contribuye con ± 1 dependiendo si es una curvatura negativa o positiva. Resumiendo:

- w número de curvaturas diferentes de cero.
- σ la suma algebraica de los signos de las curvaturas.

La clasificación de un punto crítico está dada por los valores (w, σ). De esta manera existen cuatro tipos de puntos críticos estables con valores propios diferentes a cero en la densidad electrónica:

- 3,-3 Todas las curvaturas de $\rho(r)$ en el punto crítico son negativas y por tanto $\rho(r)$ es un máximo local en r_c. Este punto se asocia a un atractor, a un núcleo.

- 3,-1 Dos curvaturas son negativas y $\rho(r)$ es un máximo en r_c en el plano definido por sus dos ejes asociados. La tercera curvatura es positiva y por tanto $\rho(r)$ es un mínimo en el eje perpendicular a ese plano. Este punto se asocia a un punto crítico de enlace.

- 3,+1 Dos curvaturas son positivas y $\rho(r)$ es un mínimo en r_c en el plano definido por sus ejes. La curvatura restante es negativa y por lo cual $\rho(r)$ es un máximo en r_c a lo largo del eje perpendicular a este plano. Este punto se asocia a un punto crítico de anillo.

- 3,+3 Todas las curvaturas son positivas y $\rho(r)$ es un mínimo local en r_c. Este punto se asocia a un punto crítico de caja.

Cuando se analiza la topología de la densidad de una molécula es necesario encontrar todos los puntos críticos de los diferentes tipos que tiene que satisfacer la relación de Poincaré-Hopf, $n - b + r - c = 1$, donde n es el número de núcleos, b es el número de enlaces, r es el número de anillos y c es el número de cajas en la molécula.

1.3.1. El Laplaciano de la Densidad Electrónica en el punto crítico de enlace ($\nabla^2\rho_b$)

El Laplaciano en el punto crítico de enlace es la suma de las tres curvaturas de la densidad electrónica en el punto crítico (ecuación (1.5)), siendo negativos los dos valores perpendiculares a la línea de interacción atómica, λ_1 y λ_2 (por convención, $|\lambda_1| > |\lambda_2|$) y el tercer valor, λ_3, que se encuentra en la línea de interacción atómica, positivo. Las curvaturas negativas miden hasta dónde la densidad electrónica está concentrada en la línea de interacción atómica y la curvatura positiva mide hasta qué punto se ha agotado la densidad en la región de la superficie interatómica y se ha concentrado en los puntos de silla atómicos.

En los enlaces covalentes, las curvaturas negativas son dominantes y $\nabla^2\rho_b < 0$, por ejemplo, $\nabla^2\rho_b = -1.1$ au para un típico enlace C-H. En contraste, para un enlace iónico o interacciones tipo van der Walls, la interacción se caracteriza por el agotamiento de la densidad electrónica en la región de contacto de los dos átomos y $\nabla^2\rho_b > 0$. Un puente de hidrógeno N-(H$\cdots$O)=C se caracteriza por $\nabla^2\rho_b = +0.03$ au. En un enlace polar (C-X,

donde X=O, N, F) hay una considerable acumulación de densidad electrónica entre los núcleos, como en todas las interacciones compartidas, pero el valor del Laplaciano en este tipo de enlace puede ser de cualquier signo.

1.3.2. La elepticidad de enlace (ε)

La elepticidad mide hasta dónde preferentemente se acumula la densidad electrónica en el plano de la interacción atómica. Se define como:

$$\varepsilon = \frac{\lambda_1}{\lambda_2} - 1 \text{ (donde } |\lambda_1| \geq |\lambda_2|) \tag{1.6}$$

Si $\lambda_1 = \lambda_2$, entonces $\varepsilon = 0$, y el enlace es cilíndrico (simétrico), por ejemplo, enlaces sencillo tipo C-C en el etano o el triple enlace en el acetileno. Podemos decir entonces que ε es una medida de qué tanto caracter π tiene el enlace hasta convertirse en un doble enlace donde la elepticidad alcanza el máximo. Cuando el enlace cambia de doble a triple enlace, la tendencia se invierte y la elepticidad disminuye. La elepticidad de un enlace aromático es 0.23 en el benceno y el de un doble enlace es 0.45 en el etileno.

1.3.3. Las densidades energéticas en el punto crítico de enlace

Las densidades de energía (potencial, cinética y total) se usan para resumir la mecánica de la interacción del enlace. La densidad de energía potencial, $\mathscr{V}(\mathbf{r})$ (teorema Virial), es el promedio efectivo del campo potencial que "siente" un electrón en un punto r en un sistema multipartícula. La energía potencial evaluada en cualquier punto en el espacio siempre es negativa y su integral sobre todo el espacio nos da la energía potencial total de la molécula. El teorema Virial relaciona el campo Virial (virial field), la densidad de la energía cinética y el Laplaciano, que descrito para un estado estacionario, se escribe como:

$$\left(\frac{\hbar^2}{4m}\right)\nabla^2\rho(\mathbf{r}) = 2G(\mathbf{r}) + \mathscr{V}(\mathbf{r}) \tag{1.7}$$

donde

$$G(\mathbf{r}) = \frac{\hbar^2}{2m} N \int d\tau' \nabla\Psi^* \cdot \nabla\Psi \tag{1.8}$$

y donde $G(\mathbf{r})$ es la densidad de la energía cinética y Ψ es una función antisimétrica de muchos electrones.

Para comparar las densidades de energía cinética y potencial de igual manera, Cremer y Kraka [24] proponen evaluar la densidad de energía electrónica $[H(\mathbf{r}) = G(\mathbf{r}) + \mathscr{V}(\mathbf{r})]$ en el punto crítico de enlace como:

$$H_b = G_b + \mathscr{V}_b \tag{1.9}$$

La densidad de energía total nos da entonces la energía electrónica total cuando se integra en todo el espacio. H_b es negativa para interacciones que comparten muchos electrones, su magnitud refleja entonces lo "covalente" de la interacción.

1.3.4. El uso de las propiedades de los puntos críticos de enlace QTAIM

QTAIM (*Quantum theory of atoms in molecules*) provee a los químicos de herramientas para interpretar, entender y predecir los resultados de la química experimental. Varias de las propiedades descritas por QTAIM se han correlacionado con propiedades moleculares experimentales. Por ejemplo, la densidad electrónica en los puntos críticos de enlace, ρ_b, se relaciona fuertemente con las energías de enlace, así que se puede usar como una medida del orden del enlace (Bond Order= $exp[A(\rho_b - B)]$) [25]. La energía potencial de la densidad en el punto crítico de enlace se ha mostrado que esta altamente correlacionada con las energías de los enlaces H-H [26]; se ha encontrado relación entre las interacciones $\pi - \pi$ en dímeros de benceno y en empaquetamientos de bases nitrogenadas en ADN con puntos críticos de enlace y puntos críticos de caja entre monómeros π-apilados [27, 28, 29].

El uso de las propiedades extraídas de los puntos críticos de enlace en el campo del diseño de fármacos ha sido desarrollada principalmente por Paul Popelier y su grupo de trabajo. Estos autores han propuesto la construcción de un espacio vectorial formado principalmente por las propiedades evaluadas en los puntos críticos de enlace, por ejemplo, un punto en este espacio es

especificado por las propiedades del enlace [30, 31, 32, 33]. Este espacio ha servido como base para comparar moléculas con esqueletos similares, mientras más pequeña es la distancia entre dos moléculas en este espacio, mayor similitud presentan. La similitud molecular se puede cuantificar de esta manera y tiene enormes ventajas sobre otras medidas de similitud, por ejemplo, el índice de Carbó [34]:

1. Es mucho más rápido pues no involucra una integración espacial (la densidad de cada molécula se obtiene de las posiciones de los puntos críticos de enlace);

2. No es dominado por núcleos, más bien se enfoca en las regiones de los enlaces químicos de la molécua; y

3. No tiene problemas de alineación entre una u otra molécula para poderlas comparar.

El uso de la metodología QTAIM permite hacer correlaciones lineales con varias características obtenidas de la densidad electrónica y valores experimentales de grupos de moléculas. Gross y sus colaboradores [35, 36] han utilizado QTAIM para estudiar la relación del pKa en fenoles y anilinas *meta* y *para* substituidas. Chaudry y Popelier [33] realizaron estimaciones de los valores de pKa utilizando diferentes moléculas, incluyendo anilinas *meta* y *para* substituidas. Estas estimaciones incluyen el cálculo de la relación entre los valores del pKa de anilinas en solución y diferentes propiedades de la densidad electrónica de las anilinas en estado gaseoso. Grana [37] complementó el estudio utilizando anilinas *orto*, *meta* y *para* substituidas, valores de pKa con diferentes propiedades extraídas directamente de la densidad electrónica. QTAIM produce muy buenas correlaciones cuando se compara con otros métodos electrónicos para relacionar propiedades experimentales con teóricas [36]. Este método ha sido probado satisfactoriamente para predecir propiedades de varias series de moléculas [30, 31, 32, 33].

1.4. Análisis de Datos Multivariable

1.4.1. ¿Porqué Análisis de Datos Multivariable? (MVDA)

Para entender el mundo que nos rodea necesitamos medir muchas cosas, muchas variables, muchas propiedades de los sistemas y procesos que estudiamos. Las colecciones de datos en ciencia, tecnología y casi en cualquier aplicación, son datos multivariables, con muchas variables medidas en cientos de muestras o en diferentes intervalos de tiempo.

Se puede extraer una gran cantidad de información de datos que provienen de muchas variables siempre y cuando las observaciones y las variables se seleccionen cuidadosamente, es por esto que es necesaria una correcta caracterización como primer paso al análisis de los datos. Esto aplica tanto para la investigación científica como para el desarrollo tecnológico.

Para analizar un conjunto de datos con un método multivariable no sólo basta observar los datos. Estos datos se tienen que expresar de una manera que facilite encontrar interrelaciones y tendencias para así generar conclusiones de nuestros experimentos.

1.4.2. Métodos de Proyección (Análisis de componentes principales, PCA y Mínimos cuadrados parciales, PLS)

El análisis multivariable se puede explicar de una manera muy sencilla utilizando los métodos conocidos como "de proyección". Este acercamiento representa cada caso como un "enjambre" de puntos en un espacio dimensional K (siendo K = número de variables), y proyecta el punto hacia un plano. Las coordenadas de los puntos en este plano proveen una representación de las observaciones, y la dirección de los vectores con respecto al plano proveen la representación de las variables.

Las proyecciones se pueden adaptar a diferentes objetivos de análisis de datos, por ejemplo: (i) resumen y visualización de un conjunto de datos, (ii) clasificación y discriminación en los datos, y (iii) buscar relaciones cuantitativas entre las variables. Esto aplica a cualquier conjunto de datos sin importar

el número de variables que contengan, o si tienen muchas o pocas observaciones o si hay datos incompletos. Las proyecciones pueden manejar matrices de datos con más variables que observaciones, incluso los datos pueden incluir mucho ruido y tener una correlación lineal.

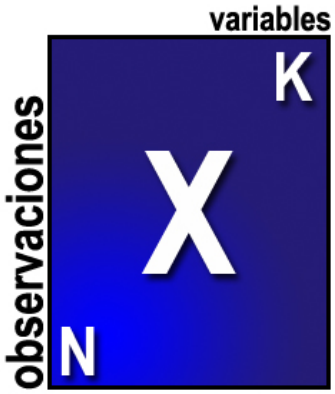

Figura 1.6: Un conjunto de datos con N observaciones y K variables. Las observaciones son las moléculas a analizar en el análisis PCA. Las variables son los descriptores electrónicos obtenidos de la densidad electrónica ρ para capturar las propiedades de las observaciones.

Los métodos de proyección tienen la capacidad de identificar datos que están fuera del rango aceptado, manejar relaciones no-lineales, y adaptar tendencias a partir de pequeñas modificaciones.

1.4.3. Aplicaciones del Análisis Multivariable

La flexibilidad que tienen los métodos de proyección para análisis de datos con muchas variables los han hecho muy útiles para el análisis y el modelado de conjuntos de datos desordenados y complicados; estos métodos se utilizan en una gran gama de aplicaciones.

Áreas de aplicación para el MVDA

- Monitoreo de procesos
- Tecnología analítica de procesos (PAT)

- Control de calidad
- Minería de datos e integración
- Relación de composición-propiedades
- Relación de estructura-actividad
- Calibración multivariable
- Caracterización multivariable
- Análisis de imágenes

Sectores de la industria donde se utiliza el MVDA

- Química básica
- Petroquímica
- Polímeros
- Plásticos
- Fibras
- Resinas, pinturas y aditivos
- Automóviles
- Medicamentos
- Biotecnología
- Papel
- Comida y alimentos
- Metales y materiales
- Semiconductores
- Telecomunicaciones

- Análisis de datos de mercado

MVDA proporciona una excelente herramienta, flexible y versátil para el análisis de datos. Hay 3 tipos de problemas a los cuales aplicar el análisis multivariable (Figura 1.7): (i) Resumen general del conjunto de datos, (ii) clasificación y/o discriminación entre los grupos de las observaciones, y (iii) creación de un modelo de regresión entre los bloques de datos (**X** y **Y**). Estos 3 pasos reflejan los pasos para aplicar el análisis multivariable.

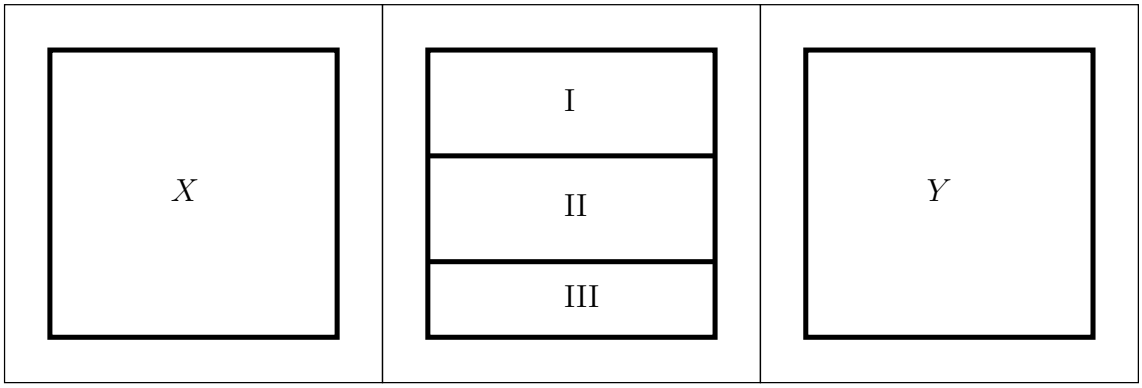

Figura 1.7: *MVDA* provee de herramientas para 3 tipos de problemas del análisis de datos, el resumen de los datos (izquierda), la clasificación y la discriminación (centro), y la elaboración de un modelo entre dos bloques de datos (derecha). En el resumen de los datos se utiliza el análisis de componentes principales, PCA y no se realiza ninguna distinción entre los grupos de las observaciones y las variables. PCA también se usa para separar si existen clases dentro de las observaciones. Para unir los bloques de las variables, usualmente nombrados **X** y **Y**, se utiliza el análisis de mínimos cuadrados parciales, PLS. PLS hace posible conectar muchas variables **X** y muchas variables **Y** al mismo tiempo, resultando en un perfil de respuesta de las observaciones que permite predecir los datos.

Resumen de los datos Al inicio de un proyecto, cuando no se sabe mucho del problema, uno requiere un resumen simple de la información que se está analizando. Este resumen se puede obtener con la ayuda del análisis de componentes principales, PCA [38, 39]. PCA produce un resumen mostrando como las observaciones se relacionan y si hay relaciones o grupos de observaciones en los datos. De especial interés en el procesamiento de los datos es la habilidad del análisis de componentes principales, PCA, de descubrir tendencias suaves o cambios bruscos entre las variables. Adicionalmente, PCA

muestra una relación inicial entre las variables: qué variables contribuyen de manera similar al modelo y qué variables proveen de información única a las observaciones. PCA describe la correlación entre la estructura en $\mathbf{X}$.

Clasificación y/o discriminación Es frecuente que el análisis inicial con PCA revele grupos dentro de las observaciones. Dos o tres de estos grupos de observaciones es común. La presencia de grupos indica la necesidad de un análisis PCA adicional de cada uno de los grupos para ajustar finamente el análisis y entender las características de cada grupo.

Creación de un modelo El último paso del análisis de datos es la creación de un modelo de regresión entre dos bloques de datos, usualmente designados X y Y, con el objetivo de predecir Y a partir de X. Esto se cumple con el método llamado Mínimos Cuadrados Parciales (PLS) [40, 39, 41, 42]. Se puede considerar al análisis PLS como una extensión del análisis PCA. Llamaremos entonces a las variables X *factores* o *predictores* y a las variables Y *respuestas*. En un modelo que sigue un proceso industrial, los factores son señales que se obtienen a intervalos de tiempo específico, para monitorear el estatus del proceso. Las respuestas pueden ser cantidad o calidad del producto. Las respuestas también pueden ser laboriosas, que requieran mucho tiempo o difíciles de obtener a comparación con los factores.

El objetivo de crear un modelo de datos complejos utilizando PLS es tener predicciones certeras, rápidas y cuantitativas de respuestas complejas (por ejemplo, calidad de un producto, sabor de un vino, impurezas en una muestra, etc) basadas en una colección de datos X. Como consecuencia, esta fase también puede llamarse la *cuantificación y la predicción*. Con datos apropiados y un modelo PLS funcional es posible descubrir cómo los factores influyen en las respuestas, cómo las respuestas correlacionan unas con otras y cómo se tienen que ajustar los factores para obtener una respuesta en específico.

1.4.4. PCA

Explicaremos ahora cómo funciona el análisis de componentes principales (PCA) utilizando una descripción geométrica y después un acercamiento algebraico formal.

Espacio multidimensional *K*

Supongamos una matriz $\mathbf{X}$ con N observaciones y K variables. Construimos para esta matriz, un espacio con tantas dimensiones como variables (Figura 1.8). Cada variable se representa con un eje coordenado. En la figura se muestra un conjunto de datos con 3 variables. El largo de los ejes se ajusta de acuerdo a la varianza de los datos.

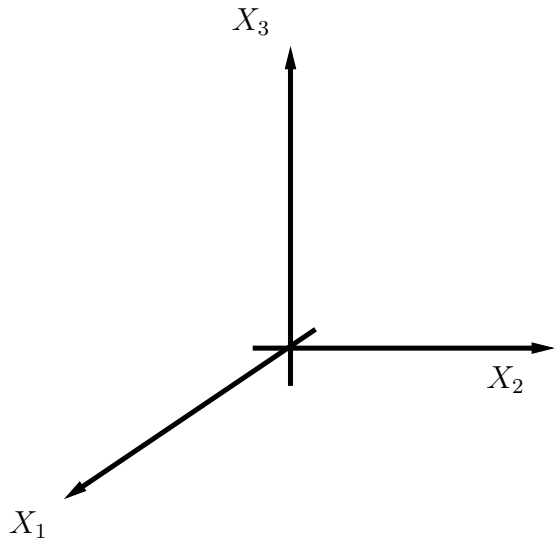

Figura 1.8: Un espacio K-dimensional. Para hacer el diagrama más sencillo, sólo se muestran 3 ejes que representan 3 variables. El largo de cada uno de los ejes representa la escala de la variación.

Graficando las observaciones en el espacio multidimensional *K*

En el siguiente paso, cada observación (cada renglón en nuestro conjunto de datos) de la matriz $\mathbf{X}$ se grafica en el espacio multidimensional K. Los

renglones de la tabla de los datos forman un "enjambre" de puntos en el espacio (Figura 1.9).

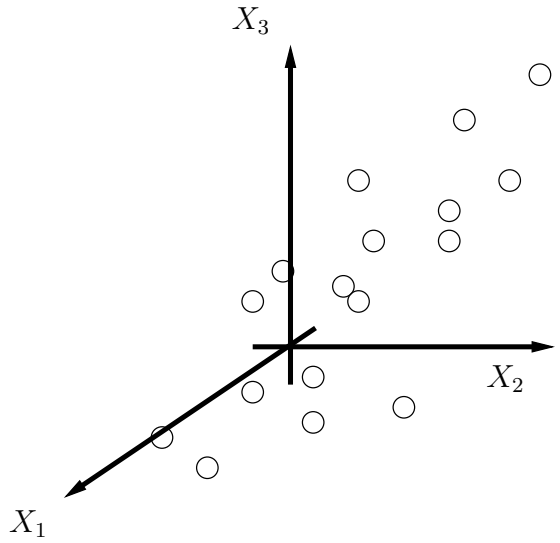

Figura 1.9: Las observaciones (renglones) en la matriz de datos **X** se presentan como un "enjambre" de puntos en nuestro espacio multidimensional K.

Calibración del espacio multidimensional K

Es necesario en este punto realizar un ajuste a los datos para asegurar que los rangos de los valores sean homogéneos. Tomando de partida el promedio de todos los valores de la matriz de datos **X**, se mueve el espacio multidimensional K para que los ejes coordenados coincidan en el centro. Esto se hace restando los promedios de las variables de los datos. Este vector de promedios corresponde a un punto en el espacio K (Figura 1.10).

Restando ahora los promedios de los datos originales resulta en un reposicionamiento de los ejes coordenados del espacio multidimensional K. Ahora el punto de origen de los ejes coordenados coincide con el promedio del "enjambre" de datos (Figura 1.11).

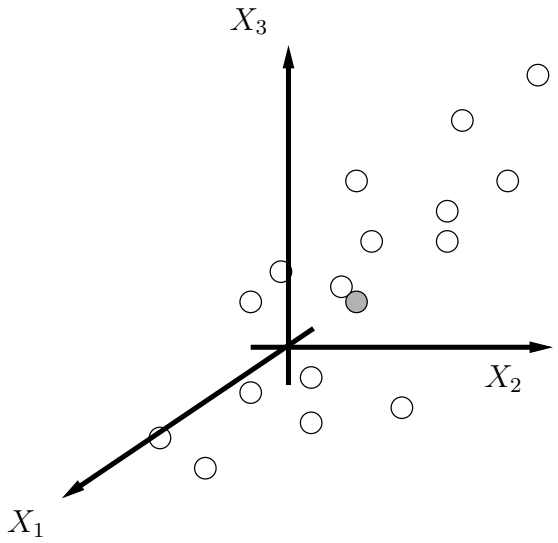

Figura 1.10: En la calibración del espacio multidimensional K, primero se extraen los promedios de las variables. Esto resulta en un punto que representa el promedio de todos los valores en el conjunto de datos (punto gris) y está situado en el centro del "enjambre" de datos.

El primer componente principal

Una vez ajustado el centro de acuerdo a la media de los datos, se calcula el primer componente principal (PC1). Este componente es la línea en el espacio multidimensional K que mejor se ajusta a los datos. La línea cruza los promedios del "enjambre" de datos (Figura 1.12). Utilizando esta línea, se puede generar una coordenada de cada observación hacia la línea del componente principal. Esta nueva coordenada se conoce como *score*.

Extendiendo el modelo

Generalmente, un componente principal no es suficiente para modelar variaciones sistemáticas dentro del conjunto de datos. Es necesario calcular el segundo componente principal, PC2. Este segundo componente principal también es representado por una línea dentro del espacio multidimensional K, y que es ortogonal al primer componente principal (Figura 1.13). Esta línea cruza también los promedios de las observaciones y ayuda a mejorar la

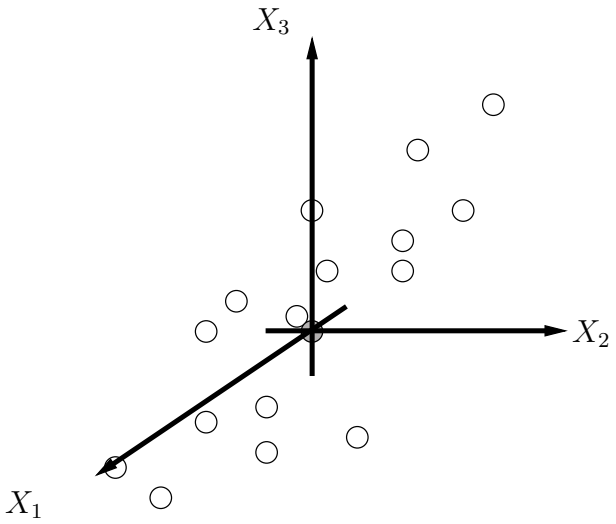

Figura 1.11: La calibración del espacio multidimensional K resulta en un reposicionamiento del origen de los ejes coordenados de manera que el promedio de los datos coincida en el centro de estos ejes (punto gris).

aproximación de los datos de la matriz $\mathbf{X}$ lo más posible.

Dos componentes principales definen un plano dentro del espacio multidimensional K (Figura 1.14). Proyectando todas las observaciones hacia este nuevo sub-espacio, es posible visualizar la estructura del conjunto de datos que se está analizando. Los valores de las coordenadas de las observaciones hacia este plano son los *scores*.

Muchas veces es necesario más de dos componentes principales para generar un modelo que explique correctamente la variación dentro del conjunto de datos. Estos componentes principales se crean de igual manera, generando planos en el espacio multidimensional K, ortogonales con respecto al componente principal anterior. Para cada una de las observaciones se genera una sub-coordenada con cada uno de los componentes principales. Por ejemplo, para un conjunto de datos donde se generaron 4 componentes principales, existirá para cada observación, los valores t_1, t_2, t_3 y t_4, donde cada uno de estos valores son las sub-coordenadas para cada componente principal (PC1, PC2, PC3 y PC4). Estos valores explican diferentes porcentajes de variación del modelo y ayudan a descubrir las observaciones o valores dentro de la matriz de datos $\mathbf{X}$ que influyen positiva o negativamente en la variación.

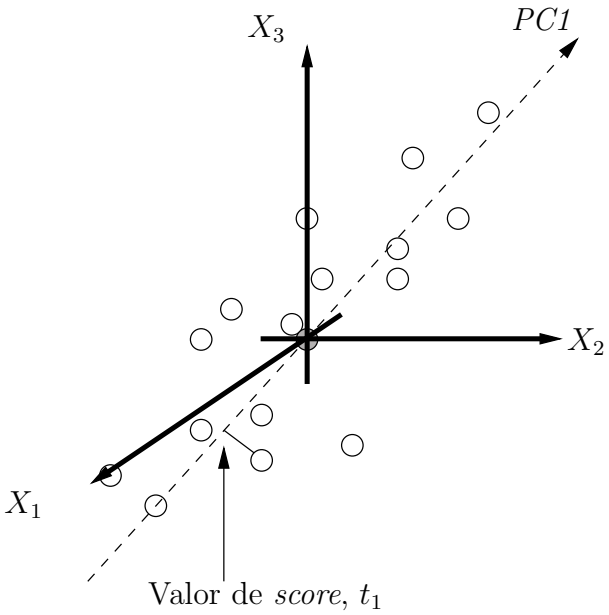

Figura 1.12: El primer componente principal, PC1, está representado por una línea que mejor se ajusta en el "enjambre" de datos. Cada observación puede ser proyectada hacia esta línea para obtener una nueva coordenada. Esta nueva coordenada se conoce como *score*.

¿Cuántos componentes principales se necesitan?

Una pregunta importante que se debe realizar es *¿cuántos componentes se necesitan incluir en el modelo?* La pregunta esta ligada a la diferencia entre el grado de ajuste y la habilidad predictiva. El ajuste nos dice qué tan bien podemos reproducir matemáticamente los datos del conjunto. Una medida cuantitativa del ajuste esta dada por el parámetro R^2X. Más importante que el ajuste del modelo, es el poder predictivo del mismo. Esto puede ser estimado con qué tanto podemos predecir los datos en $\mathbf{X}$, ya sea internamente con los datos existentes o con datos externos como si fuera una validación de las observaciones. El poder predictivo del modelo esta resumida con el poder de predicción Q^2X. En este caso, usamos validación cruzada para estimar la habilidad de predicción del modelo.

La validación cruzada es una herramienta para comprobar la eficiencia de modelos de análisis creados ya sea con PCA o PLS. La idea es reservar

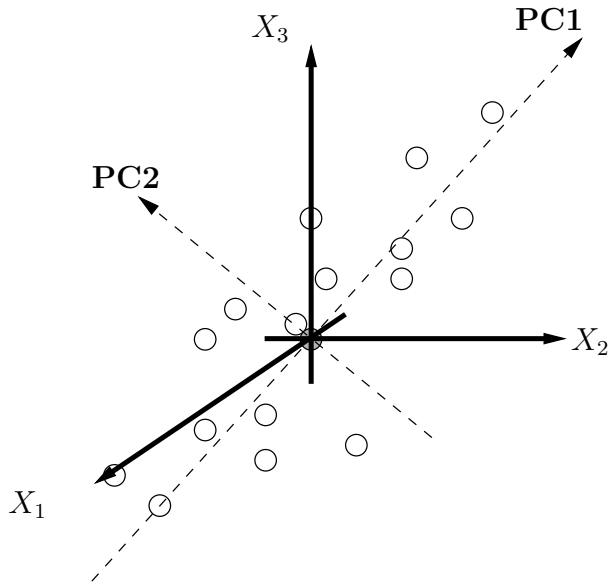

Figura 1.13: El segundo componente principal, PC2, se genera de tal manera que refleje la segunda fuente de variación principal dentro de la matriz de datos **X**. Esta línea es ortogonal al primer componente principal y también cruza el conjunto de datos por el promedio.

una porción de los datos fuera del desarrollo del modelo, generar el modelo con los datos no reservados y luego comprobar el modelo con los datos que se quedaron fuera. Finalmente se comparan los datos predichos con los datos originales.

Los parámetros R^2X y Q^2X muestran comportamientos diferentes del modelo en la medida que van aumentando su valor. El ajuste del modelo, R^2X, varía entre 0 y 1, donde 1 significa un modelo perfectamente ajustado, y el 0 significa ningún tipo de ajuste. R^2 es inflacionario y se va acercando a la unidad al mismo tiempo que la complejidad del modelo (número de parámetros, número de componentes...) aumenta. Por otro lado, el índice de predicción, Q^2X, es menos inflacionario y no se acercará a 1 mientras más complejo sea el modelo.

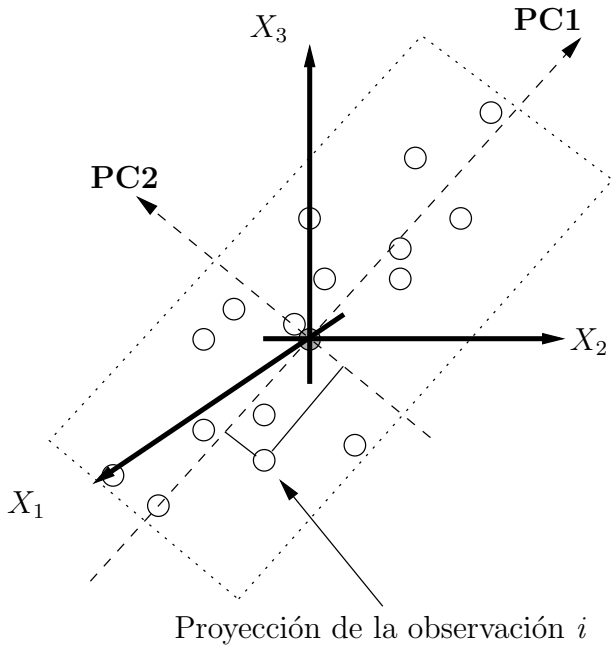

Figura 1.14: Dos componentes principales forman un plano. Este plano es una ventana a un nuevo espacio dimensional dentro del espacio multidimensional K. Cada observación se puede proyectar hacia este plan resultando en una nueva coordenada conocida como *valor de score*.

1.4.5. PLS

En su forma más simple, el análisis de cuadrados mínimos parciales (*projections to latent structures by means of partial least squars*), PLS, es un método para relacionar dos matrices de datos, $\mathbf{X}$ y $\mathbf{Y}$, por medio de un modelo multivariable lineal [40, 43, 41] (Figura 1.15). Permite de una manera eficaz analizar datos que tienen mucho ruido, que son colineales y hasta conjuntos de datos incompletos en ambas matrices $\mathbf{X}$ y $\mathbf{Y}$. Utilizando parámetros que estén relacionados con las observaciones (moléculas, muestras, compuestos, etc), la precisión de un modelo PLS aumenta mientras más número de variables $\mathbf{X}$ se tengan.

Se puede considerar la técnica PLS como una regresión particular para modelar relaciones entre $\mathbf{X}$ y $\mathbf{Y}$. Historicamente, PLS tiene 3 grandes aplicaciones: (i) relaciones estructura-actividad cuantitativa (QSAR), (ii) calibra-

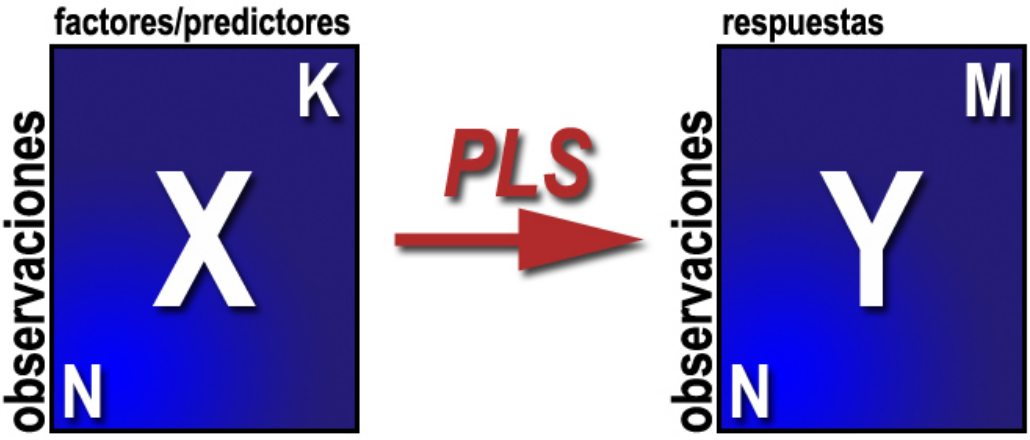

Figura 1.15: Un conjunto de datos con N observaciones y K variables forman la matriz $\mathbf{X}$ y el mismo número de observaicones N con M variables de respuesta forman la matriz $\mathbf{Y}$. Las observaciones son las moléculas a analizar en el análisis PCA. Las variables son los descriptores electrónicos obtenidos de la densidad electrónica ρ para capturar las propiedades de las observaciones y se utiliza el análisis PLS para relacionar $\mathbf{X}$ con $\mathbf{Y}$. PLS permite predecir $\mathbf{X}$ a partir de $\mathbf{Y}$

ción multivariable, y (iii) monitoreo de procesos y optimización. En QSAR, PLS se usa para modelar la relación entre, por un lado, variables que caracterizan la variación estructural de un conjunto de N compuestos, y por el otro lado, respuestas biológicas de las mismas N substancias. Lo primero forma la matriz $\mathbf{X}$ de variables de predicción y lo segundo forma la matriz $\mathbf{Y}$ con las variables de respuesta. Estas matrices tienen dimensiones (NxK y NxM) respectivamente.

Espacio multidimensional K con una respuesta M=1

Continuando con la explicación geométrica de las técnicas de análisis multivariable, retomemos el espacio multidimensional K utilizado para explicar el análisis PCA (Figura 1.8). La geometría de la técnica PLS ha sido explorada a profunidad por Phatak y DeJong [44]. Consideremos una aplicación con N observaciones, K=3 variables $\mathbf{X}$, y M=1 variable $\mathbf{Y}$. Ambos espacios pueden ser representados por la figura 1.16.

Asi como en PCA, cada observación puede ser representada gráficamente.

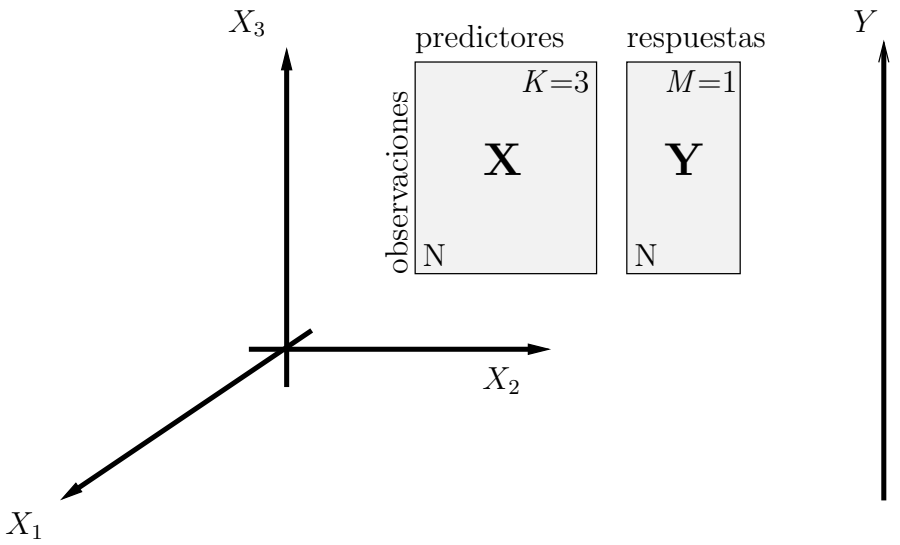

Figura 1.16: Un conjunto de datos con K=3 variables **X** y M=1 variable **Y**. La longitud de cada eje coordenado ha sido ajustada de acuerdo a la media de los datos.

Sin embargo, la gran diferencia en PLS es que cada renglón de datos en la tabla corresponde a dos puntos en lugar de uno. Uno en el espacio **X** y otro en el espacio **Y**. Consecuentemente, mientras más observaciones se tengan en el conjunto de datos, más puntos aparecen en estos espacios. La figura 1.17 muestra cómo se podría ver el espacio multidimensional cuando K=3 y M=1. La tarea del análisis de datos es describir la relación entre las posiciones de las observaciones en el espacio predictor (**X**) y sus posiciones en el espacio de las variables de respuesta (**Y**).

El primer componente PLS (M=1)

Después de ajustar los centros de los ejes coordenados, se calcula el primer componente PLS. Este componente es una línea en el espacio **X** que cruza todas las observaciones de la mejor manera posible, y que tiene una buena correlación con el vector en el espacio **Y**. La coordenada de alguna observación se obtiene proyectando la muestra en esta línea. Esta coordenada se conoce como *score,* t_{il}, de la observación i. Los *scores* de todas las observaciones forman el primer vector de *scores* en **X**, t_1 (Figura 1.18).

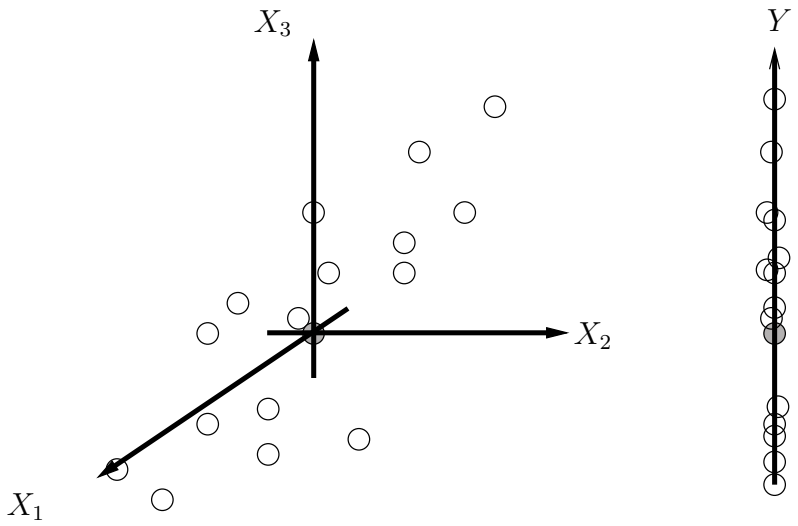

Figura 1.17: En un problema de regresión PLS, las observaciones se distribuyen en dos regiones del espacio multidimensional. Uno con los predictores (el espacio **X**) y otro con las variables de respuesta (el espacio **Y**). Los datos se han ajustado a la media de todo el conjunto de datos y los sistemas de coordenadas de ambos espacios se han ajustado a este valor (punto obscuro)

El vector de *score* t_1 se puede considerar como una nueva variable, una variable latente. Este valor refleja la información en la variable **X** original que es de importancia para modelar y predecir la variable de respuesta **Y**. Este valor, t_1, se puede utilizar también para sacar un estimado de **Y** después del primer componente PLS. Las diferencias entre las respuestas medidas y estimadas se llaman *residuales*. Los residuales del vector y representan la variación que se quedó sin explicar por el primer componente PLS. Un buen modelo tiene residuales pequeños. Los puntos alrededor de la diagonal en la figura 1.18 (parte derecha) es una manera gráfica de evaluar el modelo. Cuando todos los puntos están situados sobre la diagonal, tenemos un modelo ideal (aunque no muy realista) con cero residuales.

Extendiendo el modelo

Usualmente, no es suficiente con un solo componente PLS para adecuar correctamente un modelo a la variación de los datos en **Y**. El poder de la

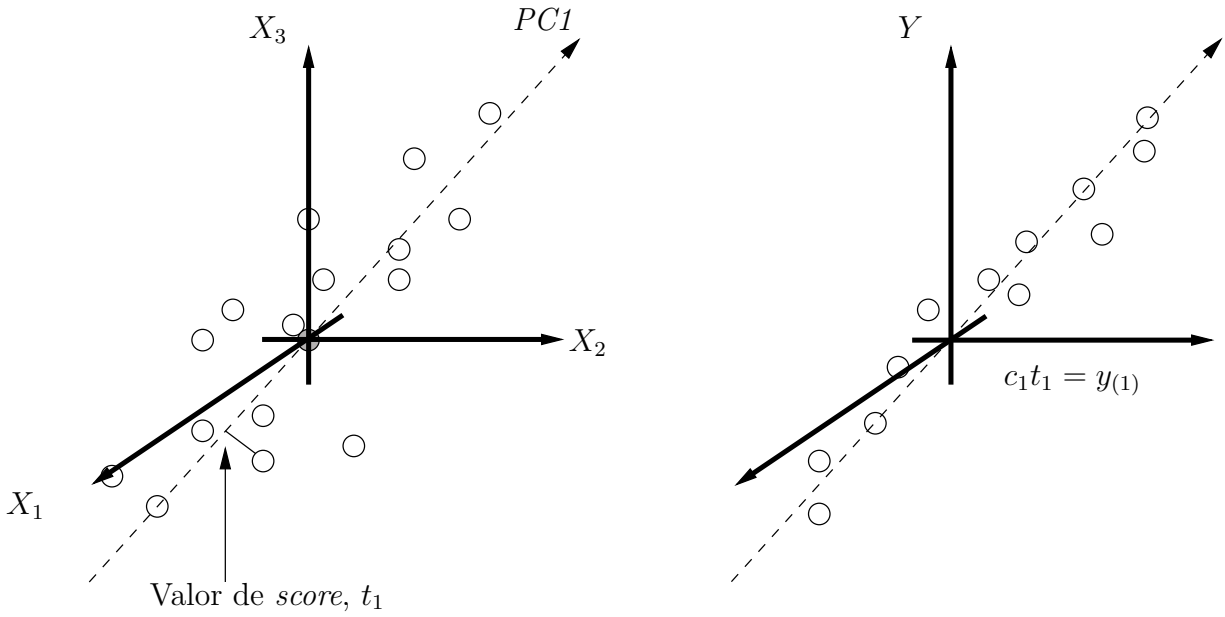

Figura 1.18: Con una sola variable en **Y**, el espacio se reduce a un solo vector de una dimensión. El primer componente en el modelo PLS entonces se orienta de manera que describa lo mejor posible los puntos en el espacio **X** mientras que de una buena correlación con el vector de **Y**. Las proyecciones de las observaciones hacia la línea del espacio en **X** da los valores *score* de cada observación.

herramienta PLS se incrementa expandiéndolo con un segundo componente. El segundo componente también es una línea en el espacio **X**, que cruza por el origen y es ortogonal al primer componente (figura 1.19). Este componente encuentra la dirección en el espacio **X** y mejora la descripción de los datos lo mejor posible, mientras que provee de una buena correlación con los residuales en **Y** que quedaron del componente anterior.

Como se puede ver en la parte izquierda de la figura 1.19, el segundo conjunto de valores *score* de las observaciones sale a partir de las coordenadas de la segunda proyección en el espacio **X**. Este segundo vector *score* se designa t_1. En la parte derecha de la figura 1.19 podemos ver como multiplicando el valor de t_1 con el valor del vector de la línea que cruza los puntos en el espacio **Y** obtenemos los residuales de **Y**. La interpretación de la parte derecha de la figura 1.19 es similar a la figura 1.18. Con esto podemos decir que mientras más juntos estén los puntos a la línea punteada, mejor será la correlación

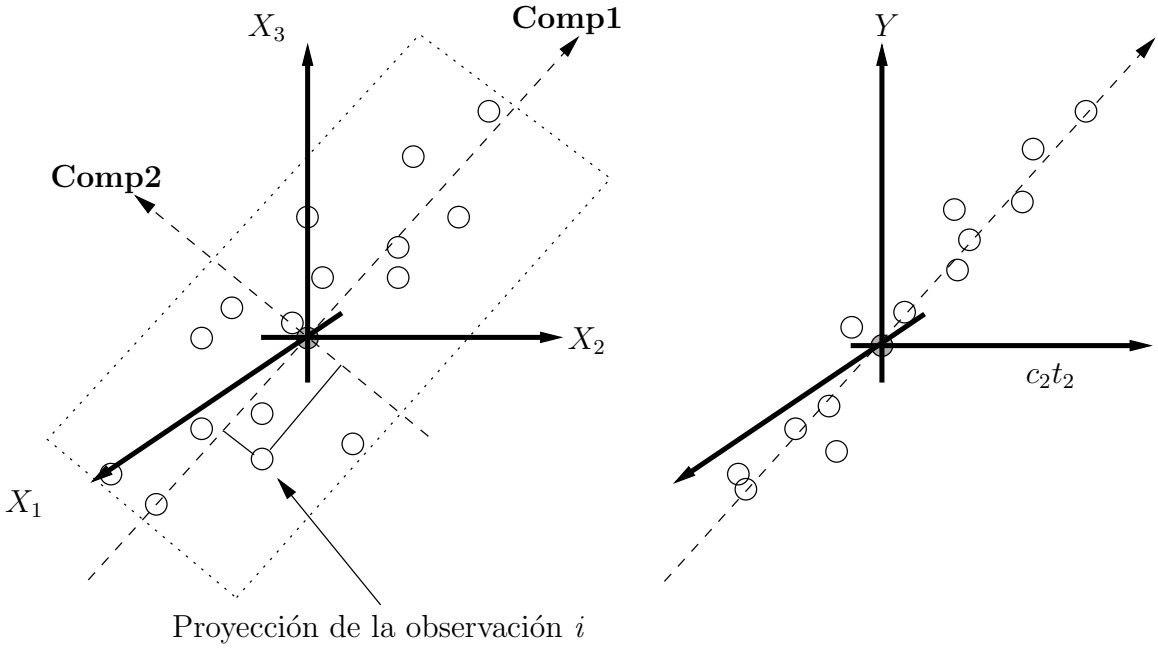

Figura 1.19: La segunda proyección en el espacio **X** es ortogonal a la primera. Si se proyectan las observaciones en este plano, se tiene el vector *score* t_1. Los dos componentes forman un plano en el espacio **X**.

entre **X** y **Y** en la segunda dimensión del modelo PLS.

Interpretando un modelo PLS

PLS consta de muchos diagnósticos que ayudan a la interpretación de los datos que se están analizando. Entre las herramientas que más se utilizan son los *pesos*, los *coeficientes*, los valores *VIP* (Variable Influence on Projection) así como el valor q^2.

Pesos Los pesos en el análisis PLS se definen como w^*c y dan información de cómo las variables **X** se combinan para formar los índices de *scores*. Estos índices son la base de la relación cuantitativa entre **X** y **Y**

Diagrama de influencia El diagrama de influencia o *loadings* muestra la relación entre variables. Indican qué tanto influyen las variables **X** so-

bre **Y**. Se podría decir que el diagrama de influencias y los pesos son complementarias y se pueden super-poner o encimar, lo que significa que un patrón identificado en la gráfica de los pesos se puede interpretar también observando en la misma dirección en el diagrama de influencias.

Coeficientes de regresión Los coeficientes son de especial interés pues simplifican la interpretación del modelo cuando hay varios componentes (4-5). La ventaja es que se obtiene un solo vector con la información del modelo por cada respuesta, a diferencia de muchos vectores utilizando los pesos. La desventaja de los coeficientes es que la información que corresponde a la correlación de la estructura de las respuestas se pierde. Esa información está guardada por los valores de los pesos.

VIP's Interpretar un modelo PLS con muchas componentes y una multitud de respuestas puede ser un trabajo complicado. Un parámetro que hace un resumen de la importancia de las variables **X**, para ambos modelos **X** y **Y**, se llama VIP (Variable Influence on Projection). Este índice fue desarrollado por Wold en 1993 [45]. Los valores VIP son la suma de los cuadrados de los pesos PLS, w^*, tomando en cuenta la cantidad de variación explicada en cada una de las dimensiones de **Y**. Para cada modelo y problema sólo existirá un solo vector VIP, explicando todos los componentes y todas las variables **Y**. Los valores VIP mayores a uno se consideran que tienen mucha influencia en la creación del modelo.

q^2 Este valor nos dice la fracción de la variación total de las variables **Y** que se pueden predecir por un componente después de aplicar validación cruzada al modelo.

Capítulo 2

Planteamiento del problema y objetivos

2.1. Planteamiento del problema

El estudio que constituye el proyecto de investigación de la presente tesis, buscó responder las siguientes preguntas:

1. ¿Es posible encontrar el sitio farmacofórico a partir de las propiedades de la densidad electrónica de una molécula?

2. ¿Es posible cuantificar la influencia de las propiedades de la densidad electrónica en la actividad biológica de una molécula?

3. ¿Es posible predecir la actividad biológica de una molécula a partir de las propiedades de su densidad electrónica?

Para poder responder estas preguntas se realizó el estudio en 3 sistemas (ácido benzoico, Casiopeínas® y fenilbencimidazoles) con los siguientes objetivos.

2.2. Objetivos

1. Buscar farmacóforos en las familias de compuestos estudiados, por medio del análisis topológico de la densidad electrónica.

2. Obtener descriptores electrónicos específicos de cada molécula, a partir de la densidad electrónica para establecer una relación de estos descriptores con la actividad biológica de cada compuesto.

3. Identificar los sitios activos o los posibles lugares que influyan directamente en la respuesta biológica.

En este estudio se utilizan 3 familias de compuestos:

1. *Ácidos benzoicos.*

   Los ácidos benzoicos permitirán comparar el análisis topológico de la densidad electrónica con propiedades electrónicas ya conocidas. Se utilizarán los valores de la constante de acidez, pKa.

2. *Casiopeínas®*

   Se buscará la relación estructura-actividad en la familia de complejos de cobre conocida como Casiopeínas®. Se utilizarán 21 compuestos multisubstituidos y su correspondiente actividad biológica. Se analizará, posteriormente, esta actividad biológica con sus descriptores electrónicos extraídos de la densidad electrónica de cada uno de los compuestos para proponer una relación entre la estructura química y su respuesta.

3. *Fenilbencimidazoles*

   En este grupo de antineoplásicos se mostrará la relación de los descriptores electrónicos con la actividad biológica reportada. Se utilizará una familia de 123 compuestos multisubstituidos y se analizarán estadísticamente las relaciones encontradas entre sus respuestas biológicas y los valores de los descriptores electrónicos extraídos directamente de la densidad electrónica de cada compuesto.

Capítulo 3

Metodología

3.1. Análisis Topológico QTMS

Con la disponibilidad actual del poder de cómputo y nuevos algoritmos de cálculo, el campo de la relación estructura-actividad cuantitativa (QSAR) se ha visto enormemente beneficiada con cálculos químico-cuánticos, tanto *ab-initio* como semi-empíricos. Este trabajo se basa en el principio de que la densidad electrónicia ρ, contiene:

- Toda la información que podemos extraer de una molécula.
- Es una propiedad que existe en el espacio 3D real.
- Puede ser determinada experimentalmente por medio de difracción de rayos-X [31] o calculada utilizando métodos *ab-initio*.

La metodología utilizada para el estudio de la relación estructura-actividad utilizando la densidad electrónica se llama QTMS (Quantum Topological Molecular Similarity), y se basa principalmente en los trabajos de Popelier y O'Brien [31, 30].

El método es "cuántico" pues extrae los datos de los esquemas computacionales que explícitamente incorporan la naturaleza cuántica de las moléculas. Es "topológico" pues utiliza la teoría de átomos en moléculas creada por Bader [25] para extraer toda la información química contenida en un sistema molecular. QTMS utiliza las propiedades de los puntos críticos de enlace

(BCP) para representar de una manera compacta una molécula. La parte topológica del análisis permite obtener representaciones moleculares discretas basadas en propiedades de enlace después de particionar ρ vía su gradiente. QTMS permite comparar moléculas y evaluar su similitud.

QTMS involucra los siguientes pasos:

1. Optimización.
2. Análisis de la función de onda.
3. Análisis estadístico.

En la Figura 3.1 se representan visualmente los pasos a seguir para la aplicación del análisis QTMS.

3.1.1. Optimización

El primer paso para realizar el análisis QSAR es la optimización de la geometría de cada una de las moléculas que se ocupan en el proceso. El método para la optimización de todas las moléculas estudiadas fue B3LYP con la base 6-311++G(2d,2p), utilizando el programa GAUSSIAN 03 [46].

3.1.2. Análisis de la función de onda

Una vez que se tienen todas las funcionas de onda, se procesan con la aplicación ext94b, que se incluye en el paquete AIMPAC [47]. Esta aplicación localiza los puntos críticos de enlace dentro de la molécula. En esta etapa, se extraen todos los parámetros que van a formar la huella digital cuántica de cada molécula y que van a intervenir en el estudio QSAR.

3.1.3. Análisis estadístico

El siguiente paso es realizar el análisis estadístico de los parámetros electrónicos obtenidos, y la propiedad que se va a utilizar como variable de respuesta. Todo el análisis estadístico se realizó utilizando el paquete **SIMCA-P** (umetrics.com). El análisis estadístico que permite correlacionar

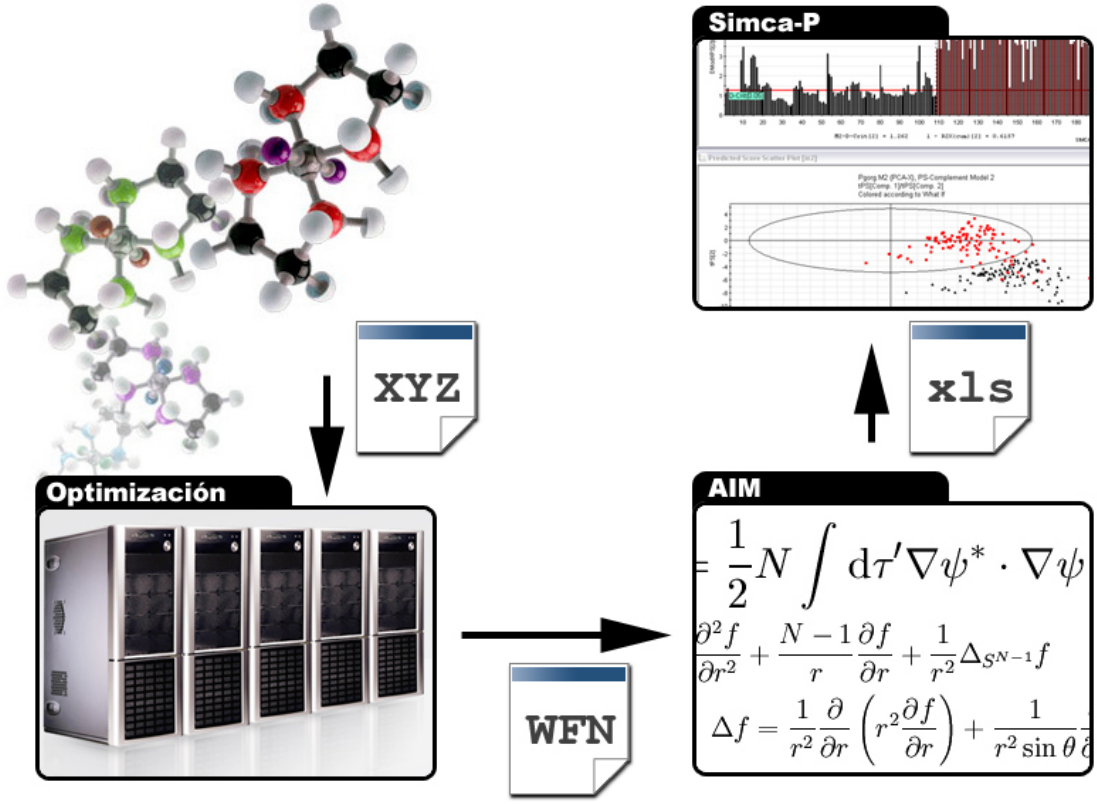

Figura 3.1: Diagrama de flujo explicando el método QTMS. La molécula a analizar se construye generando un archivo con coordenadas XYZ (GaussView), se optimiza la geometría y se extrae la función de onda (GAUSSIAN), se procesa utilizando las aplicaciones específicas de AIM (AIMPAC, AIM2000) y el resultado se analiza utilizando el paquete de estadística (SIMCA-P).

observaciones, factores de respuesta y descriptores son los mínimos cuadrados parciales (Partial Least Squares, PLS), así como el análisis de componentes principales (PCA).

Al inicio del análisis se sabe muy poco o nada de si una familia de moléculas se relacionarán entre sí, es por esto que se requiere un resumen inicial de la información del conjunto de datos. Este resumen lo proporciona el análisis de componentes principales, PCA, y muestra cómo las observaciones están relacionadas y si hay alguna observación o grupos de observaciones en los datos que difieran mucho. PCA también descubre tendencias suaves o cam-

bios bruscos en la información analizada. Adicionalmente, PCA nos ayuda a entender la relación entre variables: qué variables contribuyen con información similar al modelo, y nos ofrece información única de esas variables. PCA describe la correlación dentro de la estructura de los datos ($\mathbf{X}$).

Es común que un análisis inicial con PCA muestre agrupaciones dentro de las observaciones. Dos o tres grupos en un conjunto de datos es usual. Esto puede indicar la necesidad de un segundo análisis PCA para cada grupo (o clase) para afinar el análisis y entender los detalles de cada grupo.

El análisis de componentes principales permite ver la relación entre un grupo de datos para descubrir relaciones o familias dentro del grupo. El análisis de mínimos cuadrados relaciona dos o más grupos de datos y extrae la relación de la variación entre los elementos.

El paso final para el análisis de datos es un modelado de regresión entre dos bloques de datos, usualmente llamados X y Y, con el objetivo de predecir Y a partir de X. Este tipo de modelo se denomina PLS y se puede considerar como una extensión del análisis PCA. Las variables X son los factores o predictores, y las variables Y son las respuestas. Un modelo PLS debe de ser rápido, preciso y tener el poder de predecir respuestas complejas basado en la colección de datos X. Además, es posible con este análisis encontrar cómo los factores influyen en las respuestas, cómo las respuestas se correlacionan unas con otras y cómo los factores se ajustan para obtener un modelo que pueda predecir bien las respuestas.

Capítulo 4

Ácidos Benzoicos

4.1. Introducción

Los ácidos benzoicos y sus derivados son moléculas de gran importancia química y biológica. La familia de ácidos benzoicos multi-substituidos es compacta y confiable para obtener información relevante a partir de su función de onda extraída con métodos *ab-initio*. Paul Popelier y su grupo de investigación utilizaron un grupo de 23 ácidos benzóicos para elaborar la metodología de *Quantum Topological Molecular Similarity* relacionando las constantes de Hammet con las propiedades electrónicas extraídas de la densidad electrónica [31, 30, 32, 33]. Utilizando ácidos benzoicos *meta-* y *para*-substituidos, demuestran la eficiencia de la metodología QTMS comparando datos experimentales de la σ de Hammett con datos extraídos de la densidad electrónica [31]. Estudios similares de la predicción de valores de pKa para los ácidos benzoicos han sido desarrollados por el grupo de investigación de Polanski [48] utilizando 41 moléculas *orto- meta-* y *para*-substituidas. Ellos obtienen los valores experimentales de los pKa's [49] y realizaron análisis CoMFA y CoMSA para los mismos compuestos, obteniéndose una q^2 de 0.75 y 0.90, respectivamente.

En este capítulo se utilizan los valores de pKa como descriptores electrónicos y, dado que se conoce ampliamente el sitio activo que explica el comportamiento de la molécula, es un excelente caso para comprobar el método QTMS, descrito en el capítulo 3. Se utilizan datos obtenidos de la referencia [49] que incluyen 38 ácidos benzoicos *orto-*, *meta-* y *para*-substituidos. Se

utilizará la información generada por estos compuestos para comprobar que la metodología QTMS utilizada en el presente trabajo es válida. Los resultados se comprobarán con los obtenidos por Popelier [31]. Realizando esto se puede continuar con el análisis QSAR con los anticancerígenos propuestos.

4.2. Resultados y discusión

El uso de estas moléculas es muy útil para la comprobación del método QTMS debido a su sencilla estructura. Como las moléculas son principalmente planas, no se presentaron problemas de conformación. A cada uno de los 38 ácidos benzoicos se les optimizó la energía, de acuerdo a lo descrito en la sección 3, se obtuvo una función de onda y posteriormente se analizó cada enlace para obtener 5 propiedades de cada uno: ε, ρ, ρ^2, G y V. Esta matriz de datos se procesó después con el software SIMCA-P [50]. En el Cuadro 4.1 se muestran los substituyentes, los pKa's experimentales y los predichos por el análisis PLS.

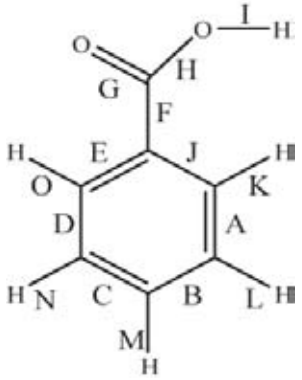

Figura 4.1: Diagrama indicando el nombre de los enlaces para los ácidos benzoicos.

Los enlaces de cada uno de los ácidos benzoicos se nombraron de acuerdo a la Figura 4.1.

A cada enlace le corresponden 5 propiedades electrónicas extraídas directamente de la densidad electrónica (ε, ρ, $\nabla^2\rho$, G, V) y por cada compuesto, le corresponde un valor de pKa. Este último valor se usó como variable de respuesta, y las 5 propiedades de cada enlace, como variables controladas. Con estos datos se realizó un análisis con mínimos cuadrados parciales (PLS) de los 15 enlaces por 5 propiedades = 75 variables a considerar. De este análisis se obtuvo un modelo y unos valores acumulativos de r^2 y q^2 de 0.91 y 0.84, respectivamente, utilizando los 5 descriptores electrónicos. El análisis de los índices VIP (Variable Importance in the Projection) de los componentes principales mayores a uno, se muestran en la Figura 4.2

En la Figura 4.2 se muestran aquellas variables que más explican **X** (los

Cuadro 4.1: Substituyentes para los ácidos benzoicos y su correspondiente valor de pKa

No.	Substituyente	$pK_{(exp)}$	$pK_{(pred)}$
1	3-NH_2	4.78	4.20
2	2-NH_2	4.95	4.65
3	4-NH_2	4.85	4.96
4	2-$C(CH_3)_3$	3.54	2.98
5	3-Br	3.90	3.84
6	2-Br	2.85	2.84
7	4-Br	3.97	3.96
8	4-$C(CH_3)_3$	4.40	4.54
9	3-Cl	3.82	3.81
10	2-Cl	2.94	2.88
11	4-Cl	3.99	3.96
12	3-CN	3.60	3.78
13	4-CN	3.55	3.91
14	2-OH, 3-OH	2.94	3.21
15	2-OH, 4-OH	3.29	3.63
16	2-OH, 5-OH	2.97	3.11
17	2-OH, 6-OH	1.30	1.59
18	3-OH, 4-OH	4.48	4.22
19	3-OH, 5-OH	4.04	3.73
20	2-C_2H_5	3.79	3.76
21	4-C_2H_5	4.35	4.44
22	3-F	3.87	3.73
23	2-F	3.27	2.96
24	4-F	4.14	3.83
25	2-OH	3.00	3.38
26	4-OH	4.58	4.45
27	3-OCH_3	4.09	4.08
28	2-OCH_3	4.09	3.45
29	4-OCH_3	4.47	4.63
30	3-CH_3	4.27	4.26
31	2-CH_3	3.92	3.82
32	4-CH_3	4.36	4.44
33	3-NO_2	3.45	3.59
34	2-NO_2	2.17	2.90
35	4-NO_2	3.44	3.44
36	2-OH, 4-OH, 6-OH	1.68	1.81
37	2-NO_2, 4-NO_2, 6-NO_2	0.65	0.27
38	H	4.18	4.16

Los valores de $pK_{(exp)}$ vienen de: Physical and Chemical Data Compendium; Poradnik fizykochemiczny, WNT: Warsaw, 1974; pp 347-351. Los valores calculados de $pK_{(pred)}$ se obtienen del análisis PLS utilizando (N)=38, (K)=76 (X=75, Y=1).

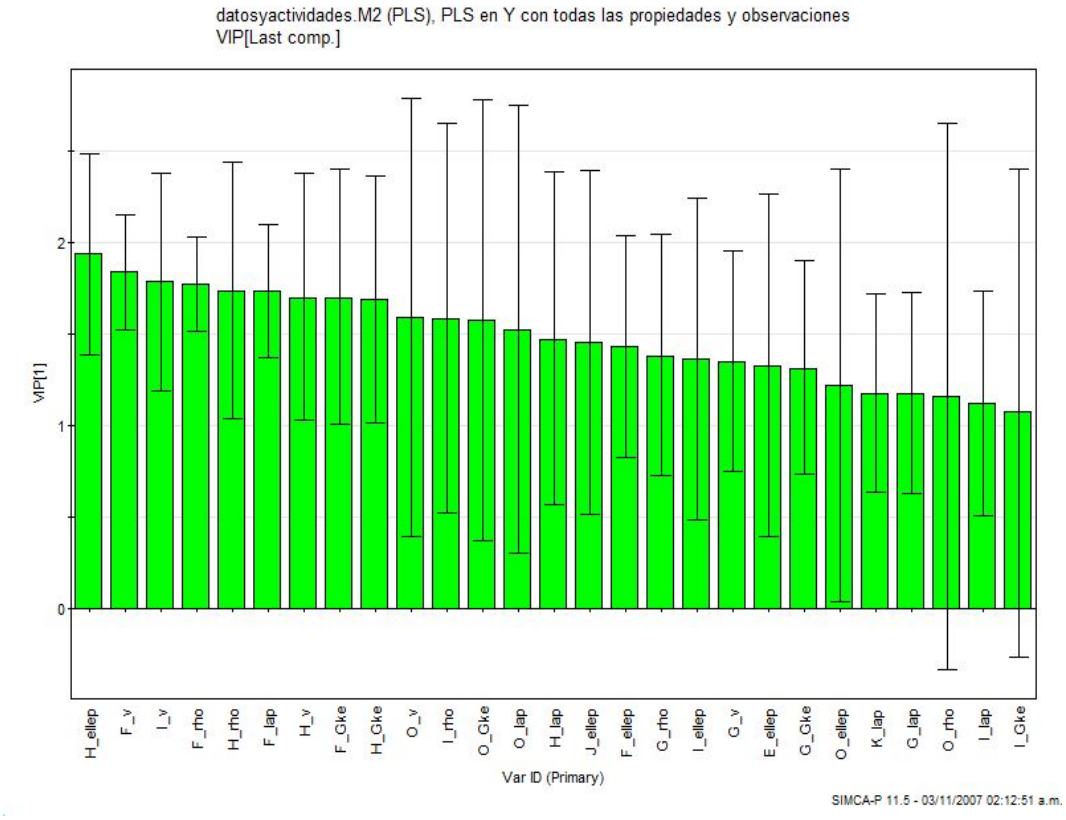

Figura 4.2: Gráfica de los valores VIP para los ácidos benzóicos. Solo se muestran aquellos valores mayores de uno.

descriptores electrónicos) y se correlacionan con $\mathbf{Y}$ (las variables de respuesta, en este caso, el pKa). Las variables que más influyen en el modelo son los del enlace H, F e I. Si nos referimos a la Figura 4.1, podemos ver que son las variables de ρ que forman el grupo carboxílico. El siguiente enlace que aparece es el F, que forma la unión C-C del grupo carboxílico con el anillo aromático. Estos cuatro enlaces (H, F, I y G) se pueden ligar directamente al tipo de reactividad química esperada en estos compuestos. Los cambios en las propiedades de ρ en esta región, reflejan las variaciones en la actividad, en este caso, representada con los diferentes valores de pKa.

Podemos notar también en la Figura 4.2, que mientras baja el nivel de VIP, van apareciendo los BCP de los enlaces C-H y más adelante, los enlaces

C-C del anillo aromático.

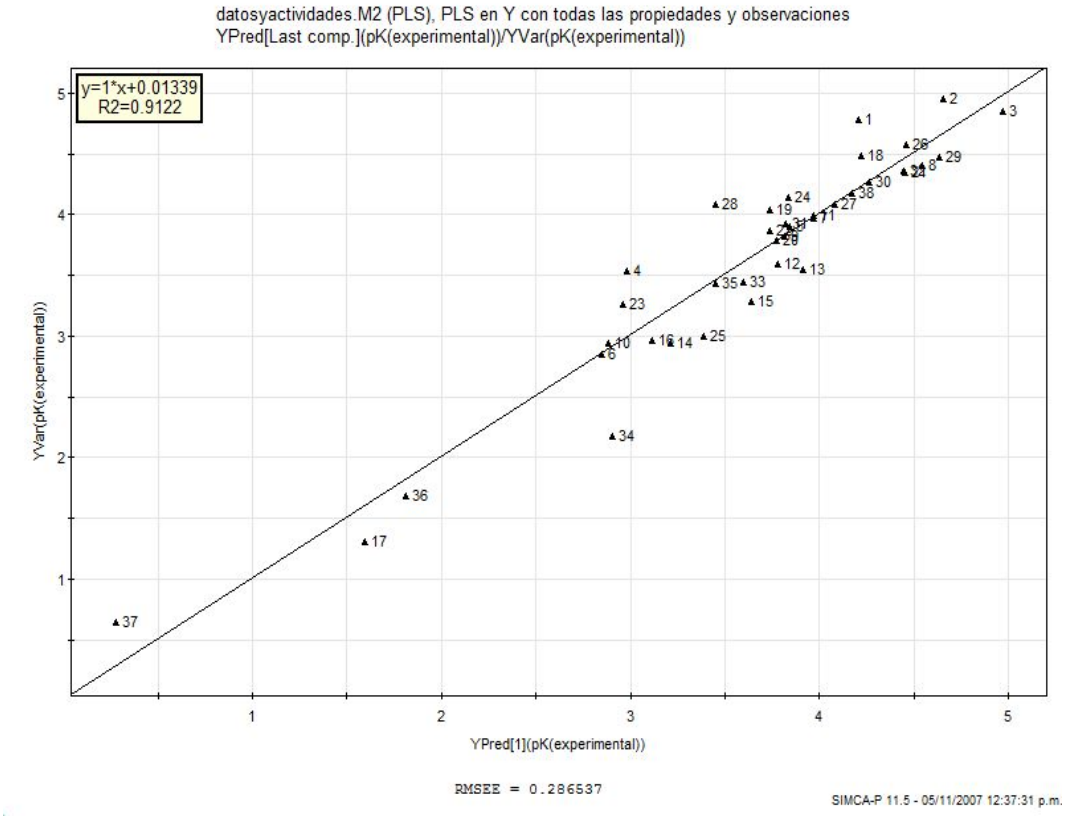

Figura 4.3: Gráfica de los valores VIP para los ácidos benzoicos. Sólo se muestran aquellos valores mayores de uno.

La Figura 4.3 muestra los valores observados de la respuesta contra los valores predichos. Podemos ver que existe una alta correlación, concluyendo que el modelo efectivamente predice los valores de manera confiable. Se enlistan cada uno de los valores de pKa que el modelo predice en el cuadro 4.1. Pocos valores están alejados de la línea central de regresión como lo indica el valor de RMSEE = 0.2865 (Root Mean Square Error). Ambos diagramas (4.2 y 4.3) nos ayudan a comprobar que la metodología QTMS efectivamente tiene el poder de crear un modelo que prediga una actividad y que explique cuales son los factores observables que se relacionan con las variables de respuesta.

Propiedad	R^2X	R^2Y	Q^2(cum)
ε	0.561	0.908	0.817
ρ	0.573	0.930	0.808
$\nabla^2\rho$	0.475	0.858	0.745
G	0.445	0.930	0.848
V	0.246	0.899	0.845

Cuadro 4.2: Valores de R^2X, R^2Y y Q^2(cum) analizando las propiedades de la densidad electrónica por separado para los ácidos benzoicos.

Se realizaron análisis independientes para cada descriptor. En el Cuadro 4.2 se observa la influencia de cada uno de los 5 descriptores electrónicos cuando se utilizan para realizar el análisis estadístico por separado. Sólo se usa uno de los 5 descriptores como matriz **X** usando el mismo listado de variables de respuesta (pKa) como matriz **Y**. Aquellos descriptores que más posibilidad tienen de describir las variables de respuesta **Y** a partir de **X** son ρ, ε y G respectivamente, lo cual concuerda con la Figura 4.2 de los índices VIP. Los descriptores que más influyen en la predicción de la actividad son G y ε respectivamente, que son las variables con mayor valor de VIP (Figura 4.2). El índice VIP de mayor influencia es el descriptor electrónico ε del enlace H. El enlace H es el que se encuentra entre el carbono y el oxígeno del grupo carbonilo (figura 4.1). La propiedad electrónica ε tiene la capacidad de decir cuantitativamente la deformación de la densidad electrónica en un plano dentro de un enlace (consultar sección 1.3.2).

Se puede analizar el comportamiento de ε en los ácidos benzoicos y su relación con los substituyentes utilizando la molécula 3 (substituyente *para*-NH_2) y la molécula 35 (substituyente *para*-NO_2). En este caso, un electrodonador y un electroatractor. Comparando los valores de ε (ver cuadro 4.3) para estas dos moléculas, y para el ácido benzoico sin substituyentes, se observa que al sacar densidad electrónica del anillo por medio de un electroatractor, el enlace toma más caracter π, se deforma en mayor cantidad que cuando ingresa densidad electrónica proporcionada por un electrodonador.

Para visualizar directamente la deformación de la densidad electrónica de las moléculas 3, 35 y 38, se graficó la densidad electrónica en el plano perpendicular al enlace entre los átomos de carbono y oxígeno del enlace H

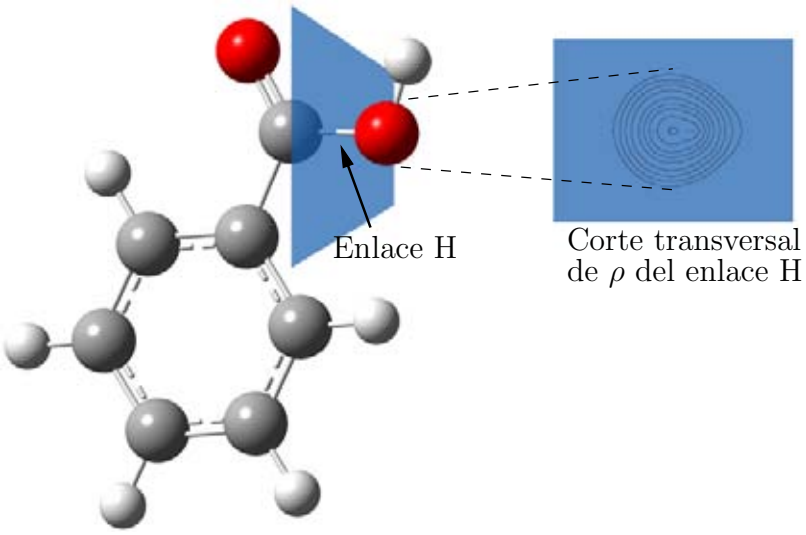

Figura 4.4: Para las moléculas 3, 35 y 38, se calculó los valores de ε entre el átomo de carbono y el oxígeno del enlace H. Este valor se graficó utilizando como centro entre los dos átomos, las coordenadas del punto crítico de enlace.

de las moléculas mencionadas. Se utilizarán como centro entre ambos átomos las coordenadas del punto crítico de enlace (Figura 4.4). Una vez teniendo las gráficas de los cortes transversales de las 3 moléculas, se sobreponen para observar mejor la diferencia (Figura 4.6). El contorno de menor tamaño es el del ácido benzoico sin substituyente en el anillo. A partir de lo anterior se puede encontrar una relación empírica entre ε y los valores de pKa. A mayor ε, mayor pKa. A mayor deformación del enlace C-O, la acidez disminuye. La relación directa de esto se puede observar en la Figura 4.5 en donde los valores de las elepticidades del enlace H se comparan con los valores de pKa de cada compuesto. En esta Figura se utilizan los compuestos *orto-*, *meta-* y *para-* tanto para los electrodonadores como los electroactractores, confirmando el aumento de pKa al ingresar densidad electrónica en este enlace teniendo muy poca influencia la posición del substituyente dentro del anillo.

En la Figura 4.7 podemos ver las influencias sobre el pKa de cada uno de los descriptores electrónicos por enlace. Esta gráfica como se vio en el capítulo 1 nos indica la influencia de las variables **X** sobre la respuesta **Y**. Las propiedades que influyen positivamente al aumento del pKa son el laplaciano y las energías G y V del enlace O. Las propiedades que negativamente influyen

	ε-enlace H	pKa
orto-NH_2	0.0217228	4.95
meta-NH_2	0.0242562	4.78
para-NH_2	0.0250017	4.85
orto-NO_2	0.0414555	2.17
meta-NO_2	0.0329026	3.45
para-NO_2	0.0319568	3.44
H	0.0259094	4.18

Cuadro 4.3: Valores de ε del enlace H para compuestos electronodadores, compuestos electroatrayentes y el ácido benzoico sin substituyente. Se observa un aumento del valor de ε con la presencia de la densidad electrónica ingresada por el electrodonador, deformando el enlace y haciéndolo más π.

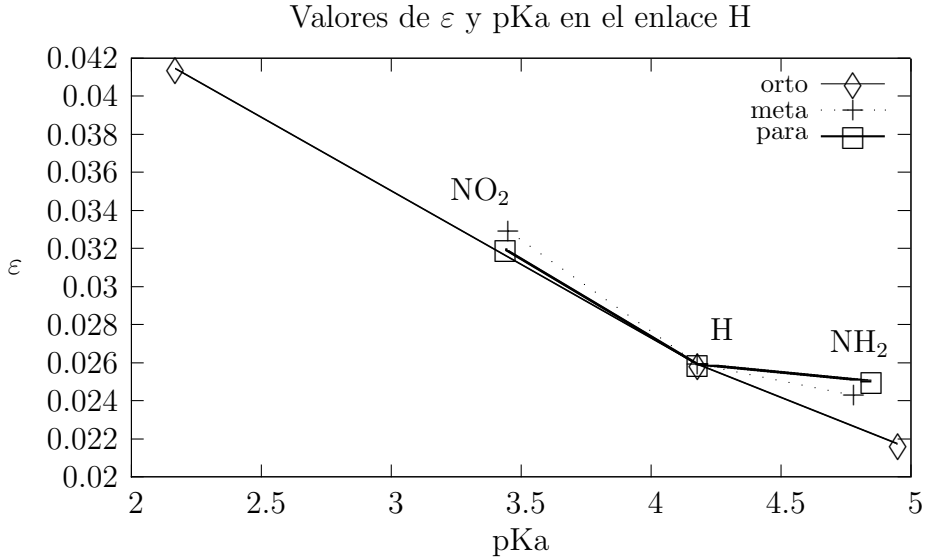

Figura 4.5: Comparación de los compuestos *orto*-, *meta*-, y *para*- electrodonadores NH_2 y electroatrayentes NO_2. Se incluye el valor de pKa del ácido benzoico sin substituente como referencia.

Figura 4.6: Se presentan las gráficas de los valores de ε para los compuestos 3 (4-NH_2), 35 (4-NO_2) y 38 (H). En la esquina inferior derecha están los 3 contornos superpuestos donde se observa la deformación del enlace.

a la respuesta son F_lap, F_v, los 5 descriptores del enlace G y H. La gráfica está formada por los pesos de cada una de las variables, y al ser un modelo PLS con un solo componente, sólo se compara con el valor de r^2x.

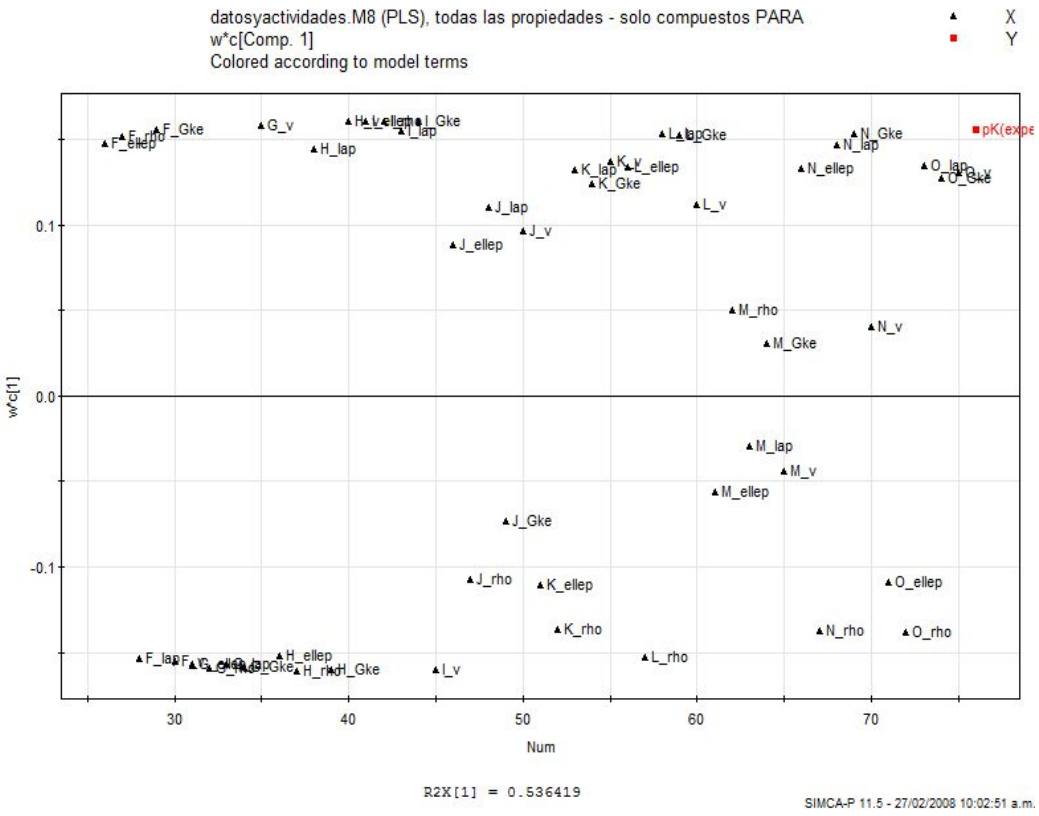

Figura 4.7: Gráfica de los valores VIP para los ácidos benzoicos. Sólo se muestran aquellos valores mayores de uno.

4.3. Conclusiones

Tomando como referencia la Figura 4.2 se observa que los componentes principales más importantes están en los enlaces H, F e I. Si nos referimos al esqueleto molecular de la Figura 4.1, estas variables corresponen al área de ρ que incluye al grupo carboxílico. Tanto el OH como el enlace C-C unido al

anillo aromático influyen en el valor de pKa. Este enlace afecta directamente el carbono del carbonilo. El área de la molécula que contiene los componentes principales de mayor influencia de acuerdo a los VIP's, constituye el sitio activo. Esta información se ajusta a lo que se conoce de la reactividad de los ácidos benzoicos obtenida con métodos tradicionales. La Figura 4.3 nos ayuda a comprobar el poder de predicción del modelo. Comprobamos entonces que los descriptores electrónicos utilizados en el análisis QTMS funcionan tanto para elucidar el sitio reactivo de la molécula como para crear un análisis QSAR a partir de la densidad electrónica ρ. El análisis por separado de cada descriptor electrónico (cuadro 4.2), nos dice que se puede obtener un buen poder de predicción utilizando sólo aquellos enlaces que influyan directamente con la variación de las variables de respuesta.

Capítulo 5

Casiopeínas®

5.1. Introducción

El tratamiento del cáncer ha tenido muchos avances con técnicas de radioterapia, así como el diseño de nuevos medicamentos. Sin embargo, los fármacos más eficientes también tienen efectos secundarios igual de potentes que la acción terapéutica que provocan. Las antraciclinas como la adriamicina (doxorubicina) muestran una acción antineoplásica muy elevada, así como una acción cardiotóxica. El cisplatino es otro ejemplo de fármacos antineoplásicos muy eficientes para el tratamiento de cánceres sólidos, así como el cáncer testicular. Desafortunadamente, los estudios clínicos del cisplatino están limitados por sus efectos a los riñones y por la resistencia en varias líneas celulares tumorales. Nuevos medicamentos basados en complejos metálicos se han desarrollado, buscando incrementar las propiedades terapéuticas comprobadas con el cisplatino [51, 52].

La familia de las Casiopeínas® son unos complejos de cobre con fórmula general [Cu(NN)(ON)]NO y [Cu(NN)(OO)]NO).

Han demostrado la inhibición de crecimiento celular *in vitro* e *in vivo* en líneas celulares humanas [53]. Algunos compuestos de esta familia han mostrado hasta un 50 % de inhibición al crecimiento celular en células epiteliales del cervix humano (HeLa), en células del colon (CaLo) así como en células de leucemia murina (L1210) con dosis de 10 a 100 veces más bajas que utilizando cisplatino. Se ha encontrado la inducción de apoptosis en células de

leucemia murina (L1210) y en carcinoma ovárico (CH1). Similar a la adriamicina, las Casiopeínas® II-gly (CasIIgly) y III-i-a (CasIII-ia) mostraron un potente efecto de inhibición en las funciones mitocondriales en experimentos con mitocondrias aisladas y con células completas del hepatoma AS-30D [52, 53].

En el sistema SMART de ensayos con *Drosophila*, Casiopeínas® I y II generaron mutaciones puntuales. Se encontró una correlación entre la respuesta y las dosis de los fármacos [53].

5.2. Resultados

En esta sección se realizará el análisis de la relación estructura-actividad de los compuestos de cobre conocidos como Casiopeínas® siguiendo la metodología QTMS descrita ya en el capítulo 3. Se contó con 21 compuestos que han comprobado su actividad antineoplásica [51, 52, 53]. Los compuestos empleados son los derivados de Casiopeínas® II-gly [(4,7-dimetil-1,10-fenantrolina) (glicinato) cobre (II) nitrato] y III-i-a [(4,4-dimetil- 2,2-bipiridina) (acetylacetonato) cobre (II) nitrato]. La Figura 5.1 muestra las estructuras de los tres tipos de moléculas que se utilizaron para este análisis. Se contaron con compuestos cuyos ligantes eran: 2 bipiridinas con acetilacetonato, 10 fenantrolinas con acetilacetonato y 9 fenantrolinas con glicina. En el cuadro 5.1 se enlistan los diferentes compuestos.

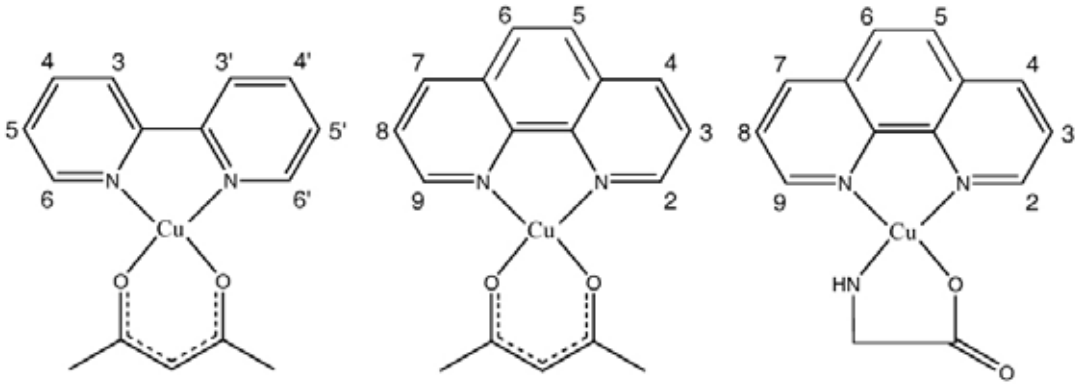

Figura 5.1: Grupo de bipiridina (izquierda) y fenantrolinas.

Bipiridinas			
No.	X	Ligante secundario	IC50 HeLa[a]
1	H	Acac	42.0
2	4,4'-diMe	Acac	18.2
Fenantrolinas			
3	H	Acac	10.7
4	4-Me	Acac	4.1
5	5-Me	Acac	6.2
6	4,7-diMe	Acac	3.7
7	5,6-diMe	Acac	3.4
8	3,4,7,8-tetraMe	Acac	1.9
9	5-fenil	Acac	2.8
10	4,7-difenil	Acac	4.2
11	5-Cl	Acac	5.9
12	5-NO_2	Acac	21.3
13	H	Gli	13.9
14	4-Me	Gli	8.7
15	5-Me	Gli	6.3
16	4,7-diMe	Gli	5.5
17	5,6-diMe	Gli	3.9
18	3,4,7,8-tetraMe	Gli	2.3
19	4,7-difenil	Gli	5.7
20	5-Cl	Gli	20.3
21	5-NO_2	Gli	20.8

[a]IC50(μM): concentración mínima del compuesto capaz de inhibir el crecimiento del 50 % de las células cancerígenas de la línea HeLa

Cuadro 5.1: Substituyentes para los 21 complejos de cobre y valores de actividad biológica en dos líneas de cáncer cérvico-uterino.

Los enlaces de cada uno de los compuestos se nombraron de acuerdo a la Figura 5.2. La geometría de cada molécula se optimizó de acuerdo al procedimiento descrito en el capítulo 3 y se obtuvo la función de onda correspondiente. Para cada enlace se obtuvieron 5 propiedades electrónicas extraídas directamente de la densidad electrónica (ϵ, ρ, $\nabla^2\rho$, G, V). El conjunto de datos empleado fue de 21 observaciones (N=21), 19 enlaces por 5 descriptores electrónicos = 95 variables (**X**=95) y la actividad biológica (**Y**=1).

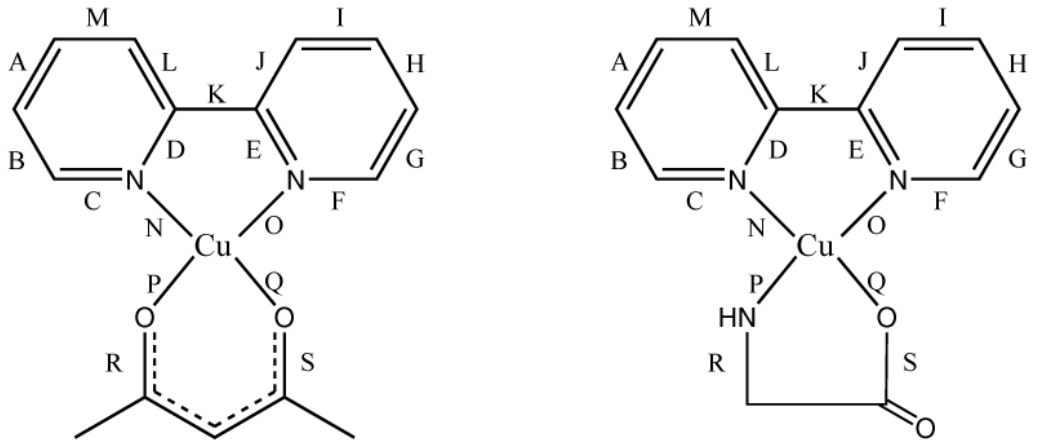

Figura 5.2: Diagrama mostrando los nombres de los enlaces .

Utilizando inicialmente un análisis de componentes en el conjunto de datos completos (95 variables en **X** más la actividad biológica), se encontró un modelo de 6 componentes con $R^2X = 0.976$ y $Q^2X = 0.906$. Los últimos dos componentes describen hasta el 97 % de la variación. La Figura 5.3 muestra 3 grupos de compuestos. Los compuestos 1 y 2 forman el primer grupo y se ve que se salen del intervalo límite de aceptación para el análisis. Los compuestos 3 al 12 forman el segundo grupo del 13 al 21 el tercero. Esto va de acuerdo con las tres posibles estructuras químicas de los compuestos. El primer grupo (compuestos 1 y 2) están formados de un anillo bipirimídico con acetilacetonato. Los compuestos 3 al 21 tienen un anillo de fenantrolina, pero varían en el segundo ligante de acuerdo con el cuadro 5.1. La separación tan marcada de los 3 tipos de compuestos nos sugiere continuar el análisis de manera separada, aplicando un segundo análisis PCA por grupo y posteriormente, el análisis PLS relacionando ahora las variables de respuesta.

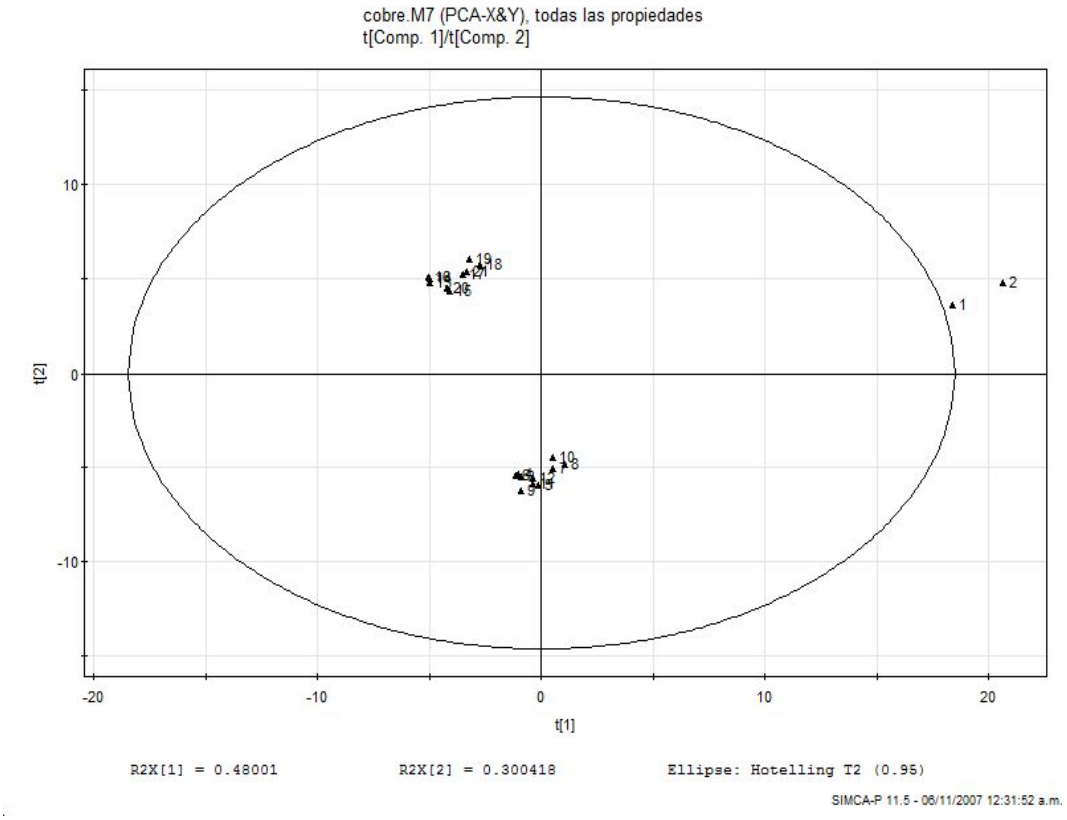

Figura 5.3: Gráfica de *scores* (t1/t2) mostrando tres grupos de compuestos que influyen en la respuesta.

En el segundo análisis de componentes principales, se tomaron ahora por separado los compuestos 3-12 y 13-21. Los compuestos 1 y 2 se descartaron por estar fuera del intervalo de confianza. En la Figura 5.4 se muestra la gráfica *score* para los compuestos 3 al 12 y en la Figura 5.5 vemos la gráfica *score* para los compuestos 13 al 21. En ambas imágenes vemos que ya no existen grupos definidos ni hay valores fuera de rango y todos los compuestos influyen por igual en la respuesta. El grupo II (3-12) mostró un modelo con 3 componentes principales, con una $R^2X = 0.846$ y $Q^2X = 0.507$ (Figura 5.6), para el grupo III (13-21) también se obtuvo un modelo de 3 componentes principales con $R^2X = 0.841$ y $Q^2X = 0.459$ (Figura 5.7) lo que nos sugiere que el grupo II de los complejos de cobre tiene posibilidades ligeramente

mayores de crear un modelo confiable, aunque el grupo III es suficientemente predictivo como para tomarse en cuenta según se explicó en el capítulo 1.

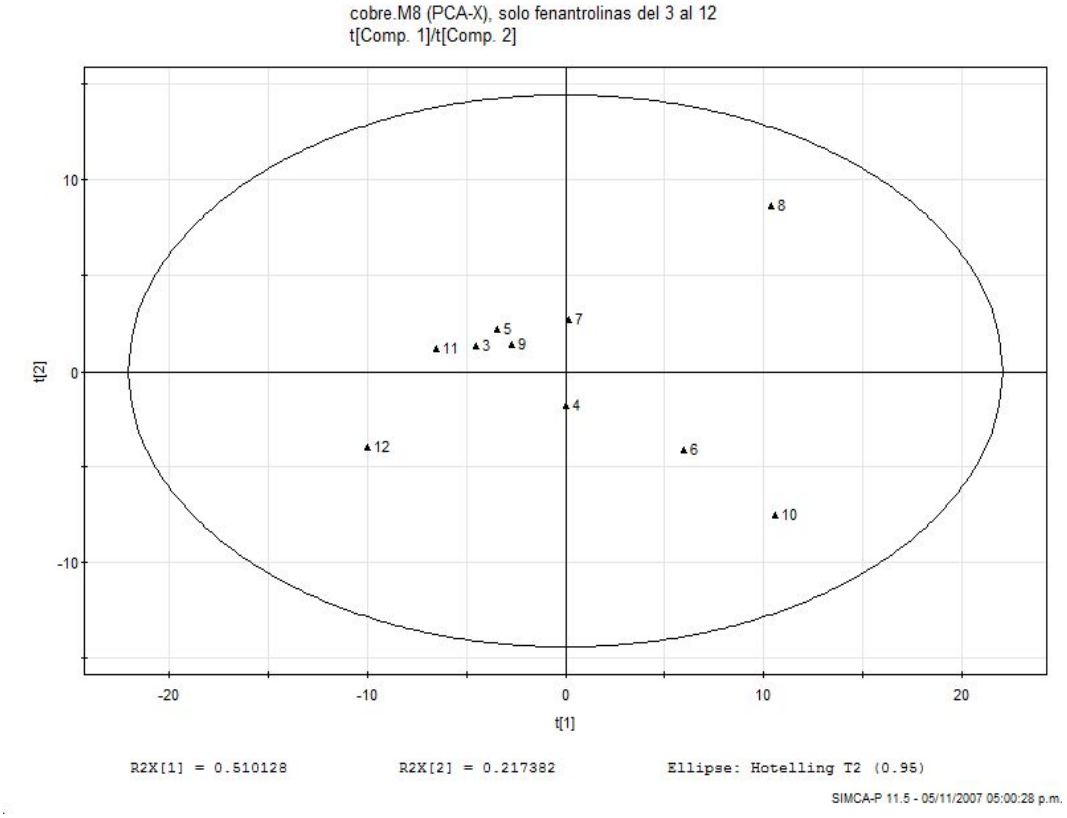

Figura 5.4: Gráfica de *scores* (t1/t2) para los compuestos 3 al 12.

Al aplicar el análisis por mínimos cuadrados parciales al grupo II, se obtuvo un modelo de 3 componentes capaz de predecir el 91 % de la varianza en la actividad biológica. El modelo usó 82 % ($R^2X = 0.828$) de la variación en $\mathbf{X}$ para describir el 91 % ($R^2Y = 0.912$) y predecir 65 % ($Q^2Y = 0.65$) de la variación de $\mathbf{Y}$. Existe muy buena relación entre los 3 componentes lo que revela que los descriptores químicos proveen información relevante para modelar la respuesta biológica.

El análisis de mínimos cuadrados parciales del grupo III obtuvo un modelo con 2 componentes capaz de predecir el 86 % de la varianza en la actividad

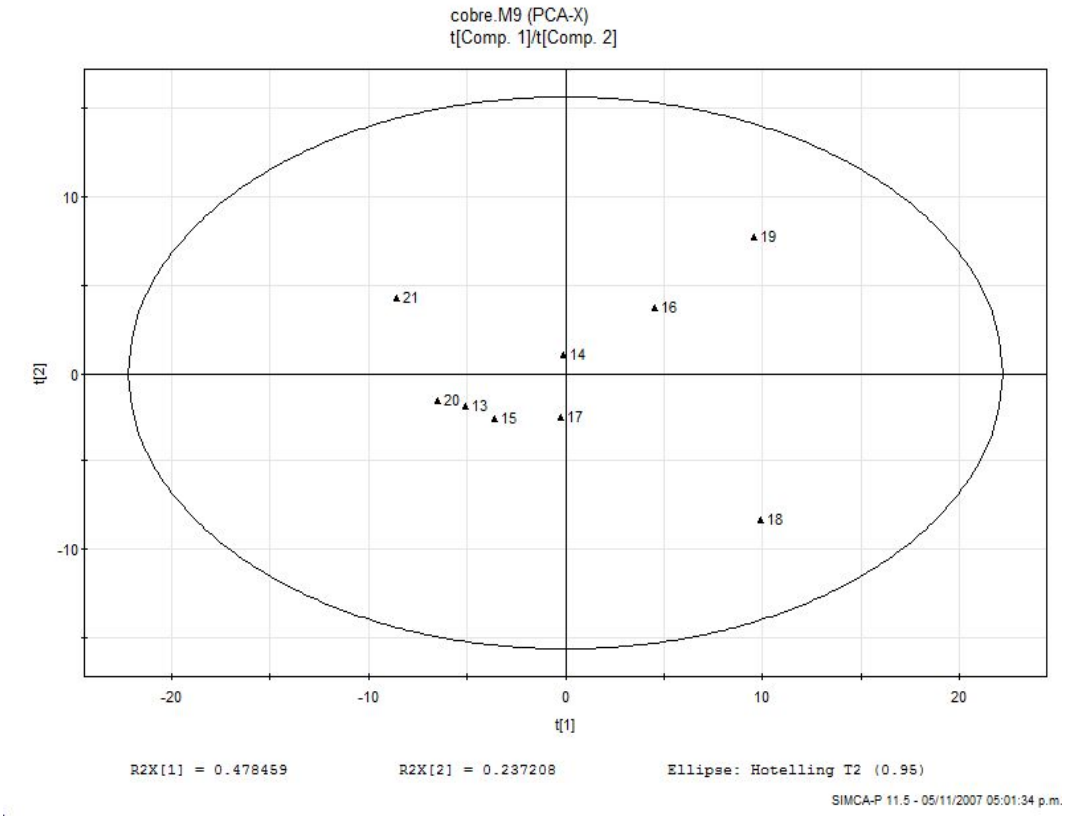

Figura 5.5: Gráfica de *scores* (t1/t2) para los compuestos 13 al 21.

biológica. El modelo usó 64 % ($R^2X = 0.643$) de la variación en $\mathbf{X}$ para describir el 86 % ($R^2Y = 0.863$) y predecir 66 % ($Q^2Y = 0.663$) de la variación de $\mathbf{Y}$ (referirse al capítulo 1).

La Figura 5.8 muestra los valores VIP para el grupo II de las Casiopeínas®. Sólo se muestran aquellos valores mayores de uno, pues son los que más influyen en la proyección de las variables. Podemos ver que los VIP's de mayor influencia involucran a los enlaces E, S, F y K. La Figura 5.9 nos muestra los VIP's para el grupo III de Casiopeínas®. Una vez más, sólo se muestran los valores significativos y encontramos que los índices con mayor influencia son los relacionados con los enlaces S, R y P.

Se realizó la comparación de los datos de actividad biológica observados y

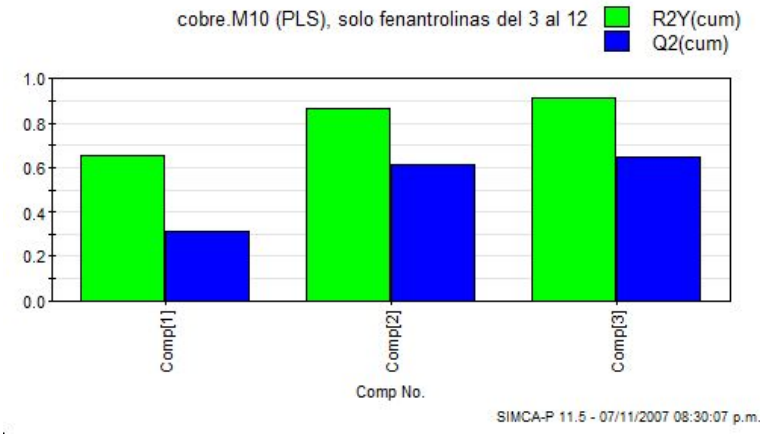

Figura 5.6: Resumen de valores de R^2Y y Q^2Y para los compuestos 3 al 12 (Grupo II).

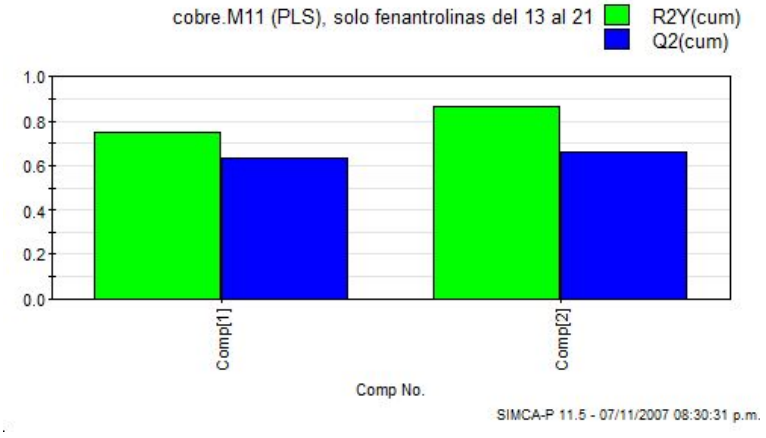

Figura 5.7: Resumen de valores de R^2Y y Q^2Y para los compuestos 13 al 21 (Grupo III).

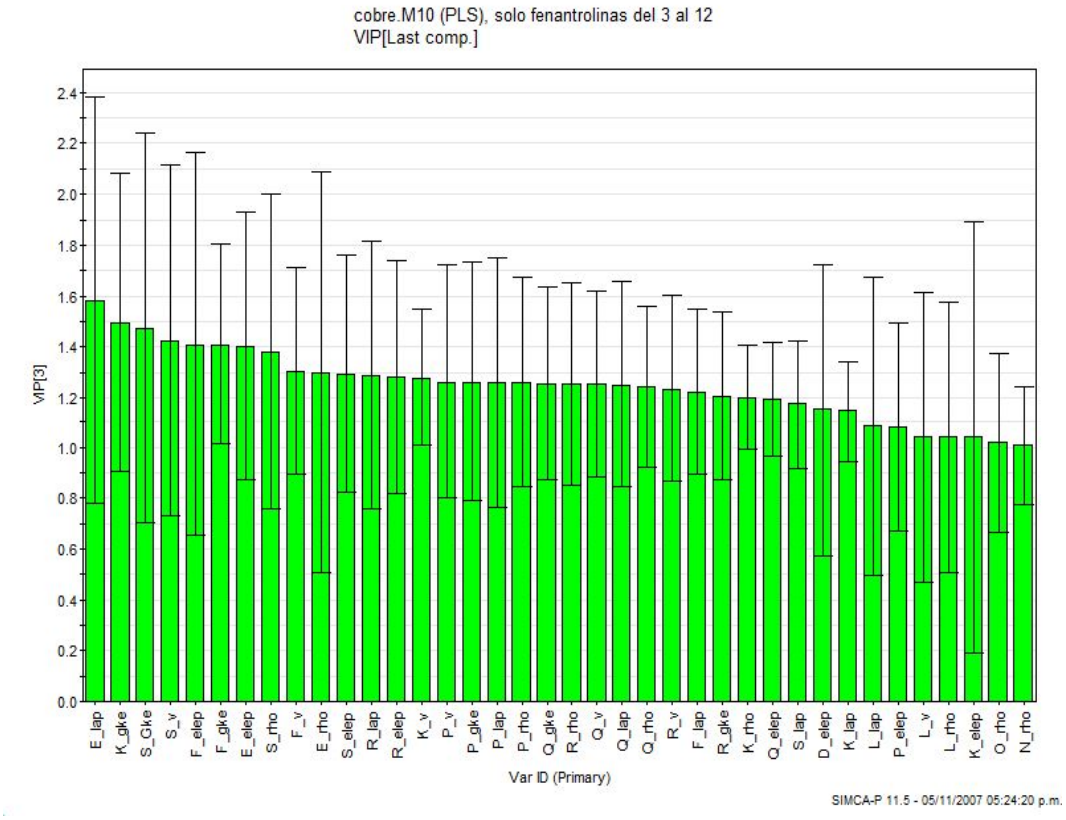

Figura 5.8: Gráfica de variables importantes en la proyección (VIP) para la regresión PLS de los compuestos 3 al 12 (Grupo II).

los datos de actividad biológica calculados con el modelo. Las Figuras 5.10 y 5.11 nos muestran los valores observados de actividad biológica y los valores predichos por el modelo

$$\mathbf{Y} = \mathbf{Y}_{avg} + \mathbf{X}B + F$$

donde B son los coeficientes que se muestran en las Figuras 5.12 y 5.13.

Se realizó una regresión lineal para ver el grado de correlación entre los datos donde tenemos valores para el grupo II de $r^2 = 0.912$ y para el grupo III de $r^2 = 0.8634$.

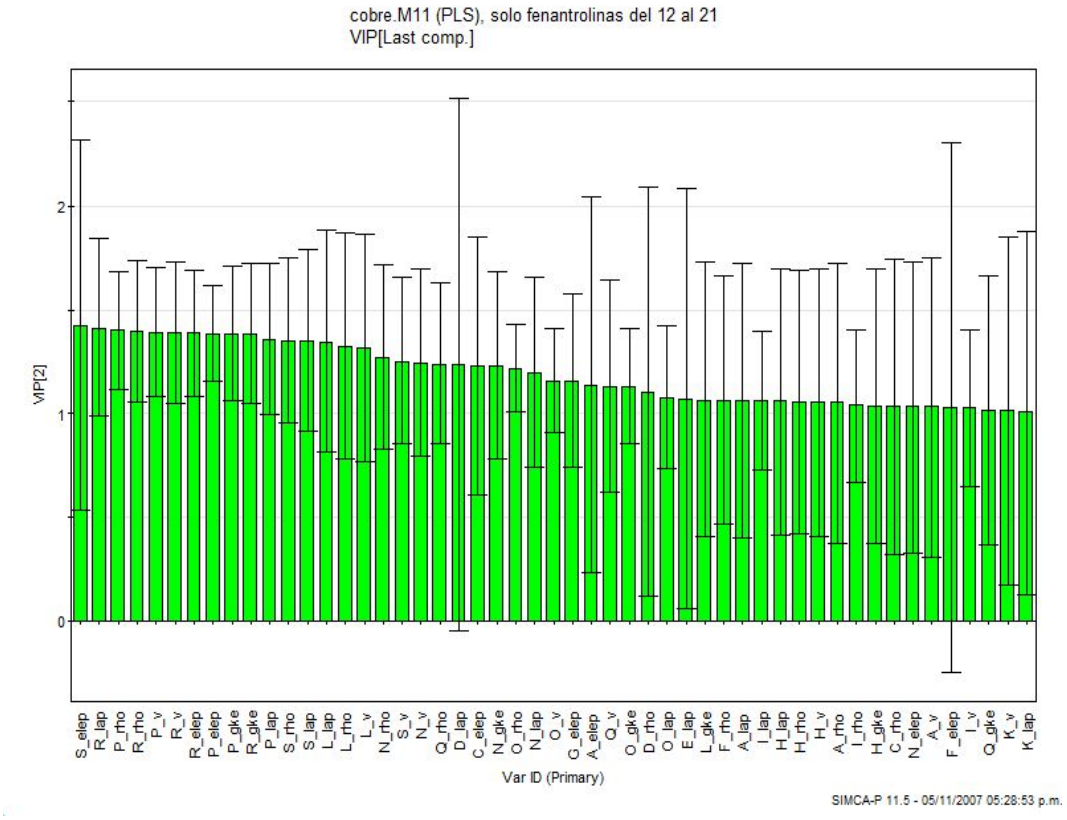

Figura 5.9: Gráfica de variables importantes en la proyección (VIP) para la regresión PLS de los compuestos 13 al 21 (Grupo III).

Para analizar cuál de los descriptores de la densidad electrónica se relacionan más con la respuesta biológica se analizó por separado cada una de las propiedades que se obtienen de ρ. Con esto, se realizó un análisis PLS para el grupo II y el grupo III de los complejos de cobre. Los resultados de este análisis se muestran en la Tabla 5.2. Vemos cómo para el grupo II las propiedades que más se ajustan a los datos son las energías (G y V) mientras que para el grupo III es la densidad electrónica ρ. Las fracciones del modelo R^2Y que más se ajustan con respecto a la variable de respuesta $\mathbf{Y}$ son también las energías cinética y potencial para el grupo II y también la densidad electrónica para el grupo III. Esto es de esperarse pues el modelo se ajusta de acuerdo a como se relacionan las variables y las observaciones. El poder

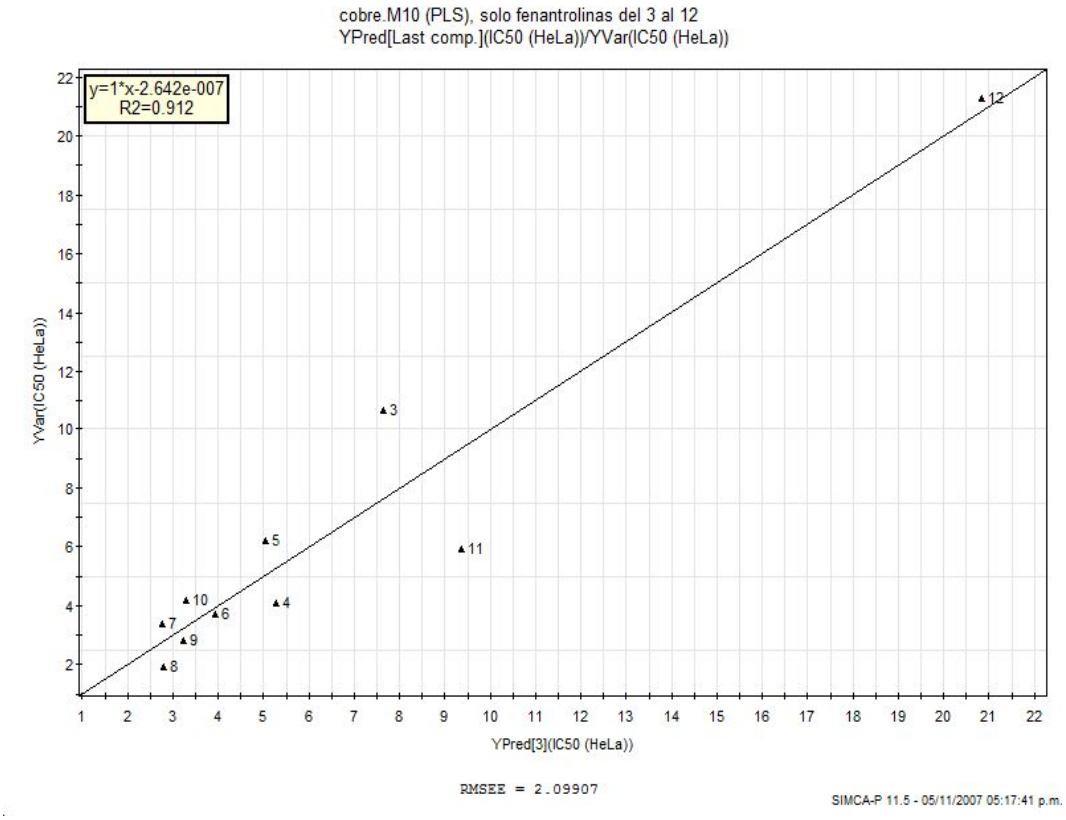

Figura 5.10: Gráfica de datos de actividad biológica observados contra datos de actividad biológica predichos por el modelo, para los compuestos 3 al 12 (grupo II).

de predicción para el grupo II está dado también por la energía cinética G y después por el laplaciano $\nabla^2\rho$. En el caso del grupo III, la construcción del modelo que mejor predice está dado por la densidad electrónica ρ. Ambos casos son congruentes con los valores altos de R^2X y R^2Y.

Las energías extraídas de la densidad electrónica para las Casiopeínas® del grupo II son las que mejor pueden describir la actividad, en especial, G. Si comparamos los enlaces que más influyen en la proyección del análisis PLS realizado incluyendo todas las variables $\mathbf{X}$ (ver Figura 5.8) con la Figura donde vemos las variables importantes en la proyección utilizando unicamente el descriptor G, Figura 5.14, encontramos prácticamente los mismos enlaces en donde está distribuida la actividad. Los enlaces presentes en ambos análisis

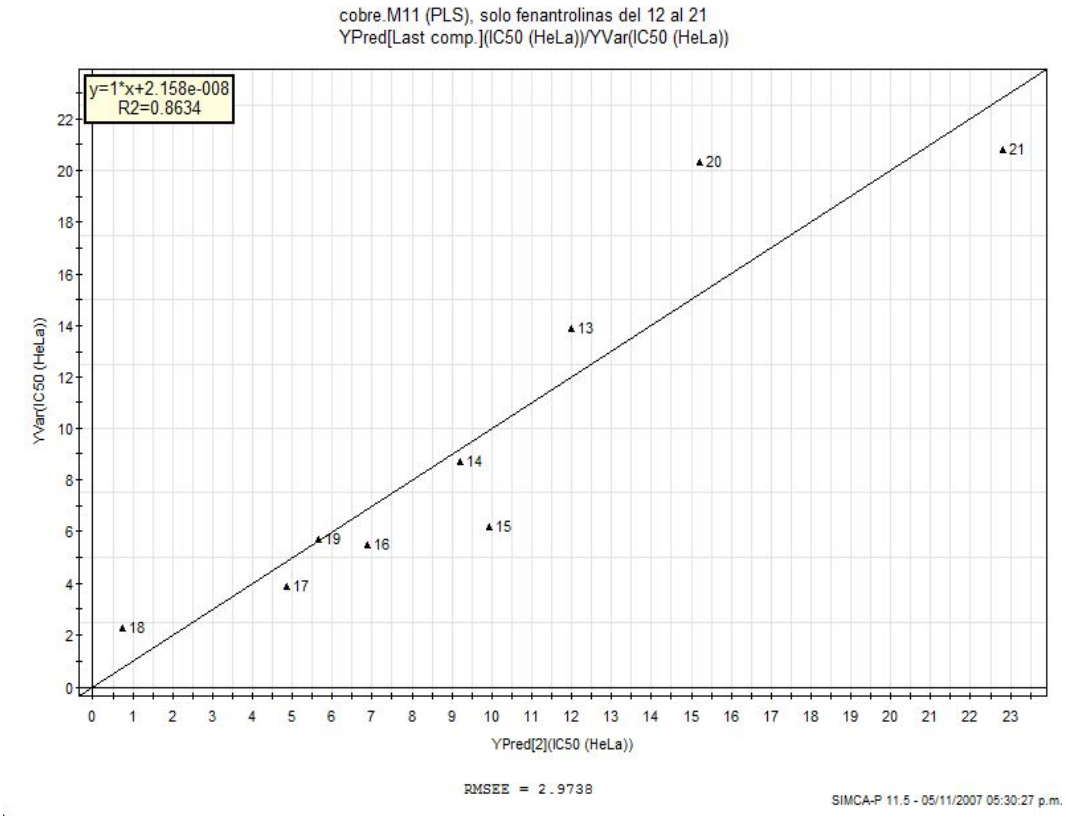

Figura 5.11: Gráfica de datos de actividad biológica observados contra datos de actividad biológica predichos por el modelo, para los compuestos 13 al 21 (Grupo III).

son los S, K y F. Los enlaces que también influyen en ambos casos son aquellos que se encuentran directamente coordinados con el metal.

En el caso de las Casiopeínas® del grupo III, los enlaces que más influyen en la proyección, considerando únicamente la densidad electrónica ρ, estan dados en la Figura 5.15. De acuerdo al análisis PLS inicial donde se consideraron todos los descriptores electrónicos, Figura 5.9, podemos ver que los enlaces son también similares en ambos casos, S, R, P, L, N y Q.

Los enlaces involucrados en los VIP's con mayor peso constituyen la parte de la estructura que más influye en la actividad biológica; por tanto, el índice VIP es una herramienta para encontrar el farmacóforo de cada familia.

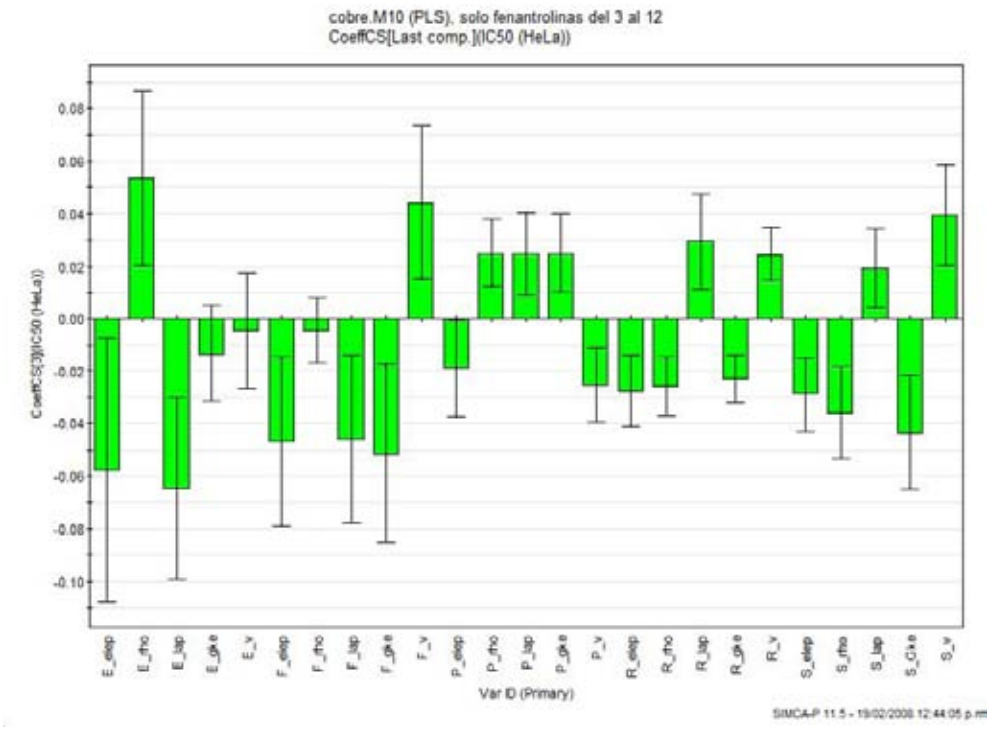

Figura 5.12: Coeficientes de regresión para los descriptores electrónicos con mayor influencia en la predicción. grupo II.

En la Figura 5.8 observamos que los VIP's que más influyen para el grupo II son los enlaces E, K, S y F. Refiriéndonos al diagrama del esqueleto molecular (5.2) vemos que los enlaces K, E y F pertenecen al anillo de la fenantrolina, mientras que el S pertenece a uno de los oxígenos del acetilacetonato.

Para el grupo III, la Figura 5.9 de los VIP nos dan como componentes principales los enlaces S, R y P, que en este caso corresponden a la parte del segundo ligando (una glicina en el caso del grupo III) que se coordina directamente con el metal. El enlace P es la unión de la amina con el cobre, y los enlaces S y R están en el ligante directamente, uno es el enlace de cobre con oxígeno y el otro con el nitrógeno respectivamente (Figura 5.18).

Para el grupo II, el sitio reactivo está en ambos ligantes del cobre y para el grupo III, principalmente está dado en el ligante secundario (Figura 5.17).

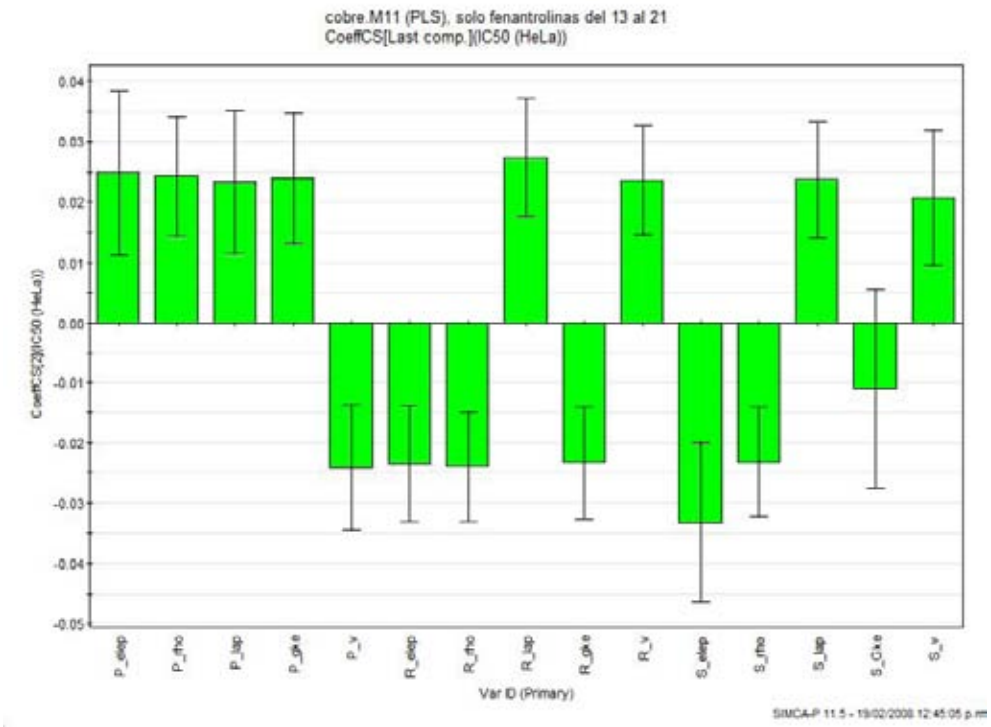

Figura 5.13: Coeficientes de regresión para los descriptores electrónicos con mayor influencia en la predicción. grupo III.

	grupo II			grupo III		
Propiedad	R^2X	R^2Y	Q^2(cum)	R^2X	R^2Y	Q^2(cum)
ϵ	0.680	0.842	0.428	0.485	0.705	0.568
ρ	0.661	0.821	0.552	0.920	0.951	0.762
$\nabla^2\rho$	0.694	0.842	0.588	0.670	0.866	0.660
G	0.871	0.926	0.698	0.464	0.704	0.586
V	0.871	0.883	0.511	0.470	0.737	0.642

Cuadro 5.2: Valores de R^2X, R^2Y y Q^2(cum) analizando las propiedades de la densidad electrónica por separado en los complejos de cobre.

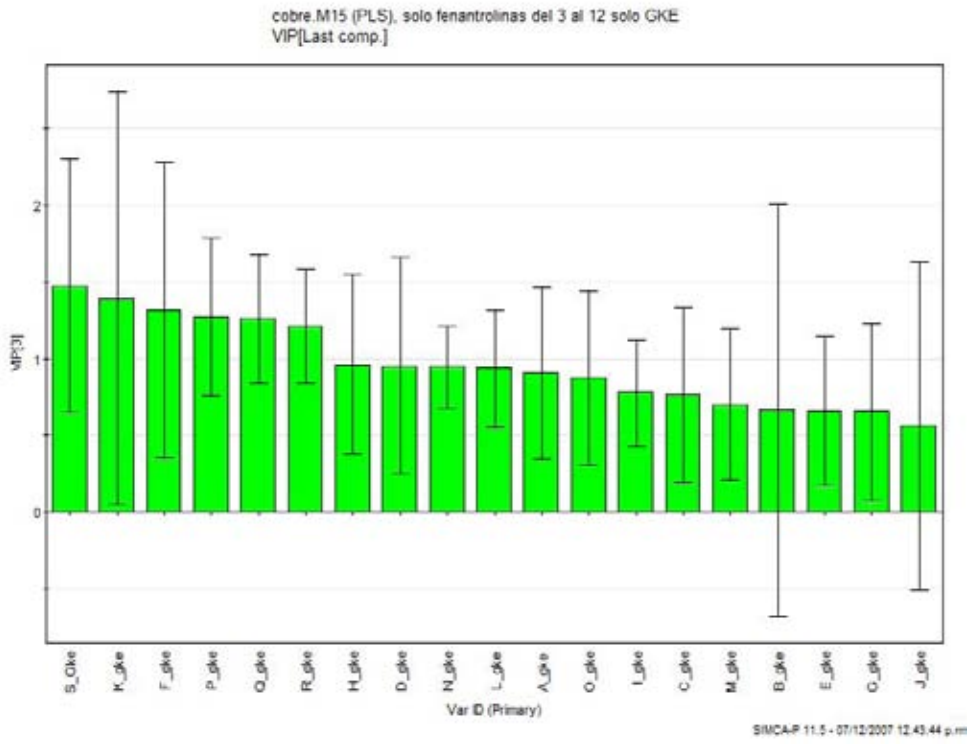

Figura 5.14: Gráfica de variables importantes en la proyección (VIP) para la regresión PLS de los compuestos 3 al 12 (grupo II) únicamente utilizando el descriptor electrónico *G*.

Los enlaces P, R y S son los que cuentan con el mayor índice VIP y menor error para el grupo III (Figura 5.16). Comparando estos enlaces con el diagrama de influencia (positiva o negativa) o *loading* (Figura 5.18) podemos ver que los enlaces que tienen influencia positiva en el valor de la actividad, están en el mismo cuadrante y son: P_elep, P_rho, P_lap, R_lap, R_v, S_lap y S_v. Los enlaces que influyen inversamente a la actividad son: P_v, R_elep, R_rho, R_gke y S_rho. S_gke y S_elep se descartan del análisis pues el margen de error es muy alto, como se puede ver en 5.16. Es importante comentar que los enlaces que influyen directamente en el valor de la IC_{50}, en realidad influyen disminuyendo la actividad, pues a mayor valor de IC_{50}, menor actividad biológica. La IC_{50} nos está indicando la concentración mínima del compuesto capaz de inhibir el crecimiento del 50 % de las células cancerígenas, por lo

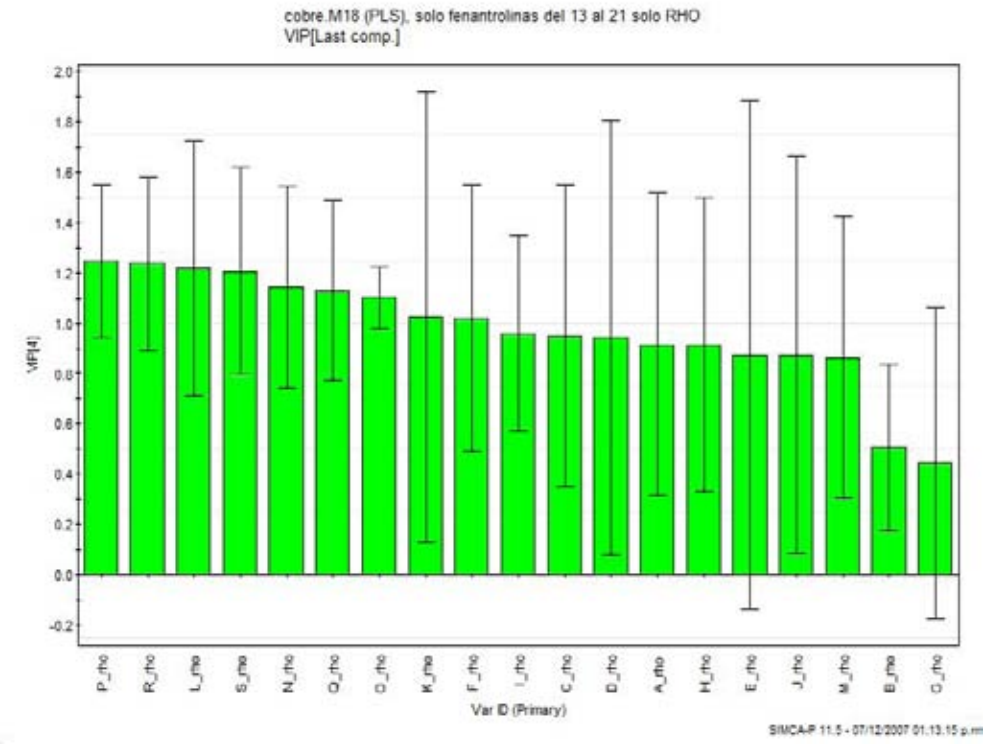

Figura 5.15: Gráfica de variables importantes en la proyección (VIP) para la regresión PLS de los compuestos 13 al 21 (grupo III) únicamente utilizando el descriptor electrónico ρ.

que el descriptor electrónico R_lap, aumenta este valor, aumentando el valor de la IC_{50}, disminuyendo la actividad biológica. Incrementando el valor para P_v, disminuirá el valor de la IC_{50}, incrementando la actividad biológica.

Al estudiar por separado cada propiedad en las Figuras 5.19 y 5.20 comprobamos como el aumento de ambas propiedades, ε y ρ, influyen directamente con el aumento del valor de la IC_{50}. En las Figuras 5.21 y 5.22 comprobamos la disminución en el valor de la IC_{50} causada por la influencia inversa de los descriptores electrónicos en el enlace R. Ambos comportamientos se observaron en el diagrama de influencias de la Figura 5.18, donde los puntos que más se acercan a la variable $\mathbf{Y}$ mayor influencia positiva tienen; y los puntos que más se alejan de la variable, mayor influencia negativa o inversa tienen.

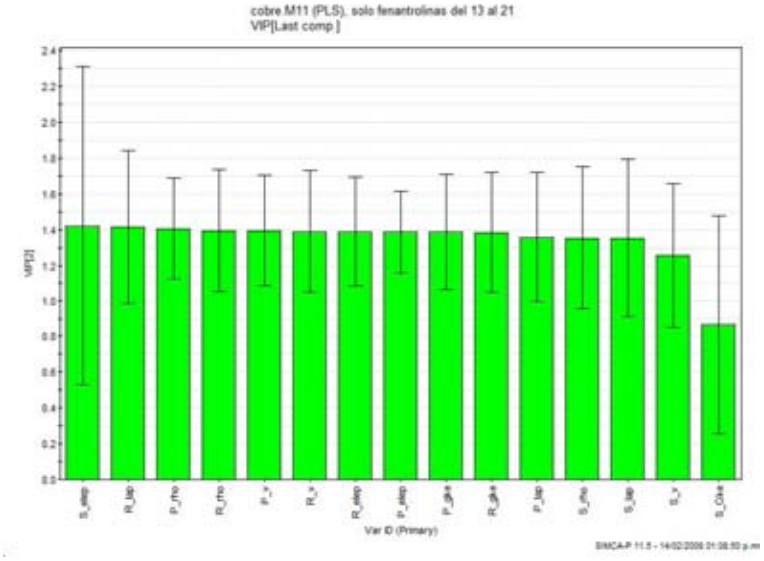

Figura 5.16: Valores VIP para los enlaces que más influyen en la respuesta.

Perfiles del enlace Con la ayuda de las funciones de onda obtenidas, se graficó el valor del menos laplaciano ($-\nabla^2\rho$) entre el átomo de cobre y el átomo de oxígeno que forma el enlace S del ligante para los grupos II y III. El valor del laplaciano determina en dónde está localmente concentrada la carga electrónica ($-\nabla^2\rho > 0$) o dónde está disminuida ($-\nabla^2\rho < 0$). El perfil mostrado en la Figura 5.23 muestra las zonas de acumulación o deficiencia en la región de enlace entre el cobre y el oxígeno (compuesto 3) o el nitrógeno (13). No existe diferencia entre los perfiles de cada uno de los compuestos del grupo II ni entre los compuestos del grupo III; sin embargo, entre grupos existe una diferencia que se puede observar en la Figura 5.23. En esa Figura tenemos el valor de la distancia entre el átomo de cobre y el heteroátomo (oxígeno o nitrógeno) en el eje X en unidades atómicas, y su valor de $-\nabla^2\rho$ en el eje Y. Los 3 máximos que se observan de 0 a 0.6 corresponden al *core* (núcleo y electrones de primera capa) del átomo de cobre y el máximo que se encuentra a partir de 3.5 corresponde al *core* del oxígeno.

Entre 1 y 3.7 se observa la concentración que se forma por los electrones de valencia de los átomos involucrados en el enlace. Se observa alrededor de 3 una concentración de densidad de 3.6. Para el caso del oxígeno y de 2.03 para

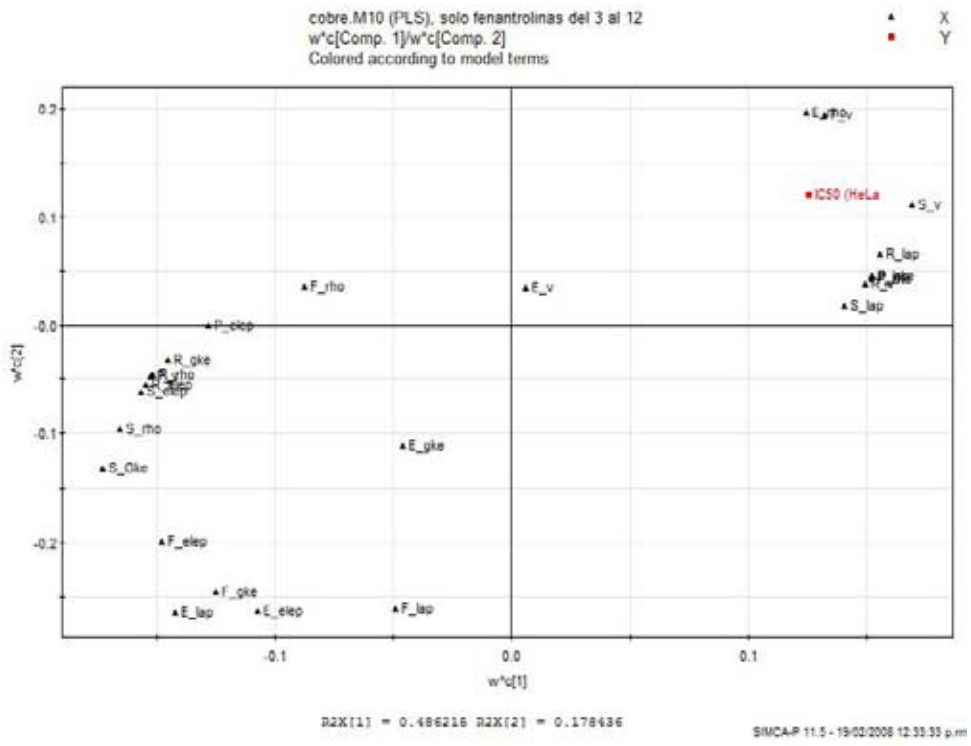

Figura 5.17: Diagrama de influencias (*loadings*) para los enlaces con mayor valor VIP y la actividad biológica del grupo II.

el caso de nitrógeno. Esta acumulación es la aportación del O o N al enlace de coordinación con el cobre. Esta diferencia de acumulación de densidad electrónica es la que determina la agrupación de las moléculas 3 a 21 en dos grupos.

Diferencias en el valor de la densidad electrónica por átomo Se analizó la influencia en la densidad electrónica contenida en cada átomo o grupo de los substituyentes en las moléculas de los grupos II y III. Utilizando el programa AIMALL97 [54] se calculó el valor de la población electrónica por cada átomo y se calculó la diferencia entre la molécula sin substituyentes (átomo-base) y el mismo átomo con substituyentes (átomo-substituido). El programa AIMALL97 nos proporciona el valor de q (donde $q = Z - N$ por lo que Δq es:

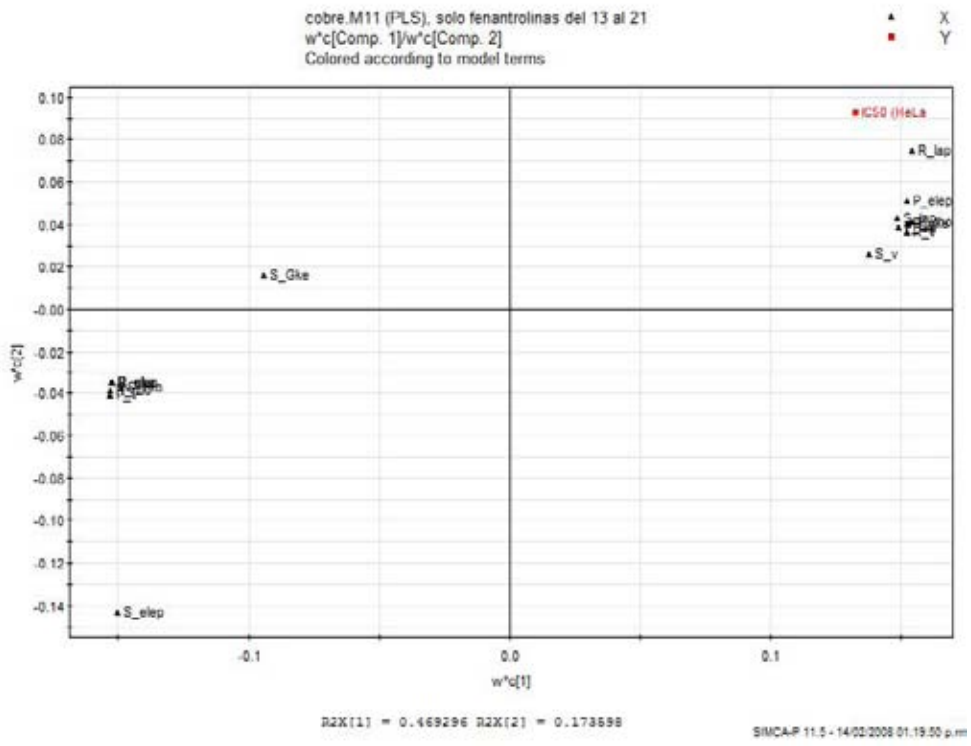

Figura 5.18: Diagrama de influencias (*loadings*) para los enlaces con mayor valor VIP y la actividad biológica del grupo III.

$$\Delta q = q_{\text{atomo-base}} - q_{\text{atomo-substituido}}$$
$$\Delta q = (Z_{\text{atomo-base}} - N_{\text{atomo-base}}) - (Z_{\text{atomo-substituido}} - N_{\text{atomo-substituido}})$$

siendo Z el número atómico del átomo y N el valor de la población electrónica. Cada molécula se dividió en 3 secciones para comparar las diferencias de densidad electrónica de los ligantes y el átomo de cobre por separado de acuerdo a la Figura 5.24. Los valores de Δq indican si aumenta o disminuye la densidad electrónica en el átomo o grupo analizado. Si el valor de $\Delta q > 0$, aumenta ρ en la sección analizada y si $\Delta q < 0$, ρ disminuye. Los resultados se muestran en el Cuadro 5.3. Se observa que el mayor cambio se presenta en la sección A. La sección C muestra un cambio menor mientras que el átomo de Cu permanece sin cambio.

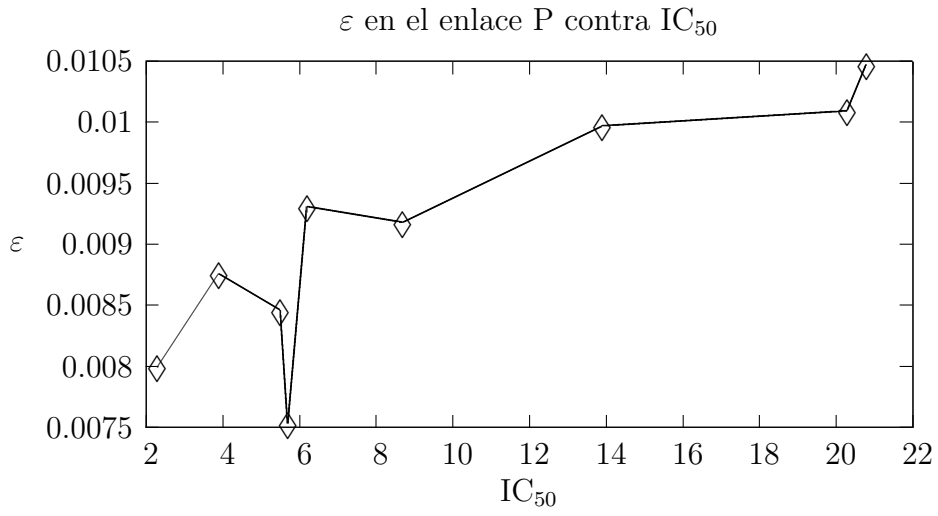

Figura 5.19: Aumento del descriptor electrónico ε al aumentar el valor de la IC_{50} en el enlace P del grupo III de los compuestos de cobre.

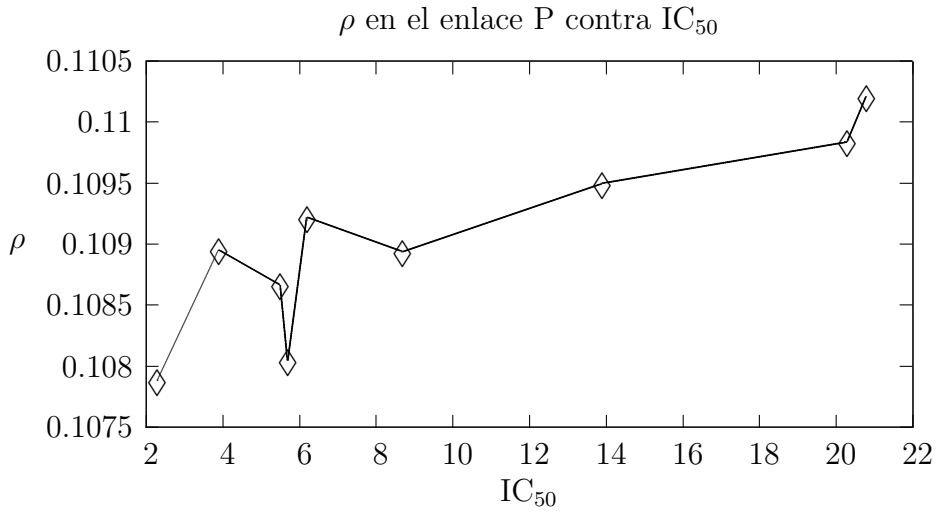

Figura 5.20: Aumento del descriptor electrónico ρ al aumentar el valor de la IC_{50} en el enlace P del grupo III de los compuestos de cobre.

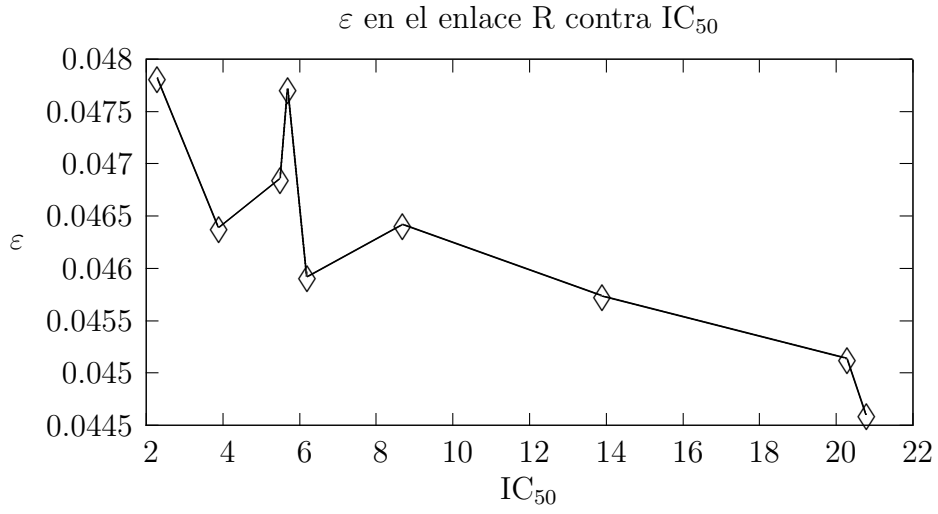

Figura 5.21: Disminución del descriptor electrónico ε al aumentar el valor de la IC_{50} en el enlace R del grupo III de los compuestos de cobre.

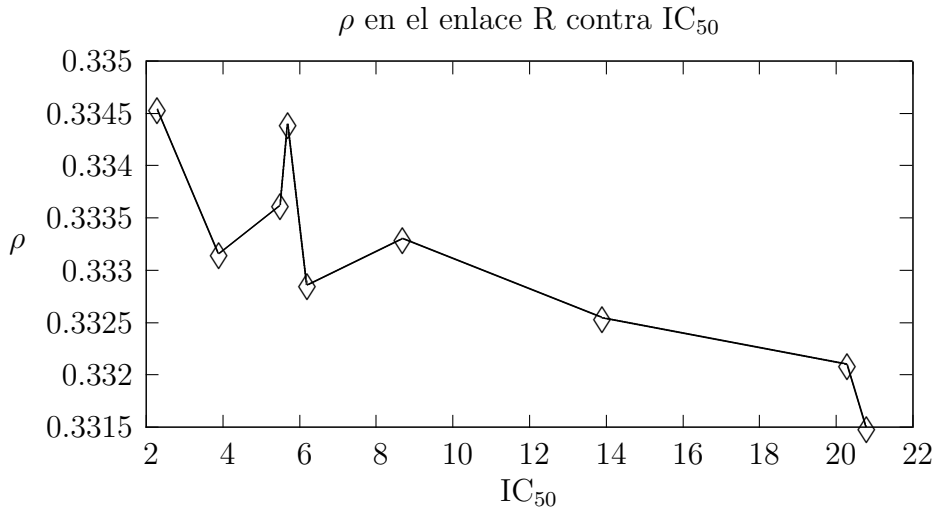

Figura 5.22: Disminución del descriptor electrónico ρ al aumentar el valor de la IC_{50} en el enlace R del grupo III de los compuestos de cobre.

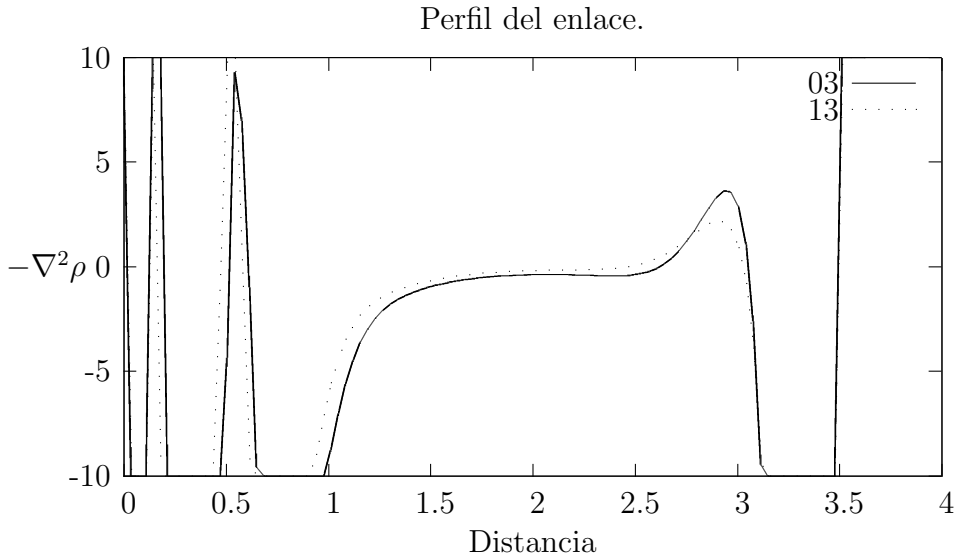

Figura 5.23: Valores de $-\nabla^2\rho$ entre el átomo de Cu y oxígeno para el compuesto 03 del grupo II y entre el átomo de Cu y nitrógeno para el compuesto 13 del grupo III. Estos dos compuestos no tienen substituyentes en su ligante.

	grupo II				grupo III		
IC_{50}	Seccion A	Seccion B	Seccion C	IC_{50}	Sección A	Sección B	Sección C
1.9	0.0323	0.0005	0.0179	2.3	0.1201	-0.0017	0.0205
2.8	0.1127	0.0016	0.0053	3.9	0.0581	-0.0005	0.0082
3.4	0.0655	0.0003	0.0069	5.5	0.0713	-0.0011	0.01272
3.7	0.0570	-0.0001	0.0112	5.7	0.4217	-0.0026	0.0208
4.1	0.1114	0.0003	0.0055	6.2	0.0318	-0.0006	0.0456
4.2	-0.1309	0.0007	0.0174	8.7	-0.0304	-0.0005	0.0064
5.9	0.0263	0.0004	-0.0037	20.3	-0.1256	-0.000004	-0.0038
6.2	0.0452	0.0004	0.0032	20.8	-0.2547	0.0021	-0.0086
21.3	-0.3182	0.0008	-0.0104				

Cuadro 5.3: Suma de todos los valores de Δq de cada átomo contra el valor de IC_{50} para los complejos de cobre, grupo II y III. Las secciones se refieren a las 3 partes de la estructura común de acuerdo a la Figura 5.24.

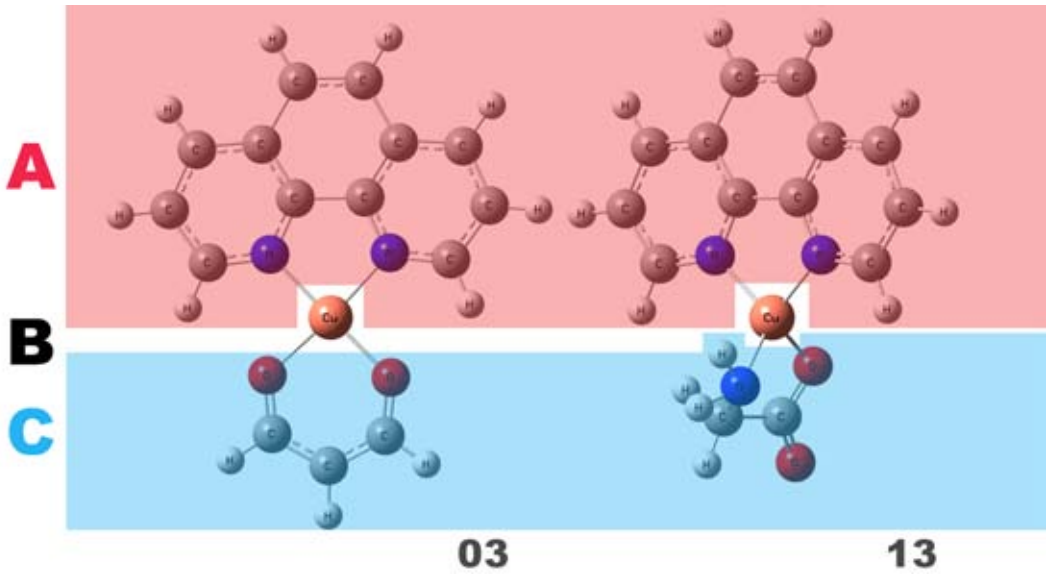

Figura 5.24: Los complejos de cobre se seccionaron en 3 partes para el análisis de la influencia de los ligantes en la densidad electrónica. A es el ligante superior, B es el átomo de cobre y C es el ligante inferior.

5.3. Discusión

Este estudio pretendió responder a las siguientes preguntas para relacionar la estructura con la actividad de las Casiopeínas®:

¿Por qué se forman grupos en el conjunto de Casiopeínas® analizadas?

Los complejos de cobre utilizados están formados por dos ligantes unidos al átomo de cobre por medio de enlaces de coordinación. El ligante superior o ligante A es de dos tipos, un anillo de bipiridina y un anillo de fenantrolina. El ligante inferior o ligante C consta de una molécula de acetilacetonato o el aminoacidato glicina. El analisis estadístico PCA de todos los componentes sugiere que el grupo I tiene valores que están fuera de intervalo para una correcta formación de modelos PLS. Los grupos II y III estan dentro del intervalo, y con valores que permiten agruparlos de acuerdo a sus descriptores electrónicos. El origen de la diferencia entre el grupo II y III, se debe a la distinta aportación de densidad electrónica que hace el nitrógeno de la glicina o el oxígeno del

acetilacetonato, como se observa en los perfiles de $-\nabla^2\rho$ de la densidad electrónica del enlace R.

¿Cuál es el sitio farmacofórico? El sitio estructural de los complejos de cobre que interviene más en la actividad biológica está formado por los enlaces del metal y la parte C. Esto se cumple para el grupo II (enlaces E, F, K, P, R y S) y el grupo III (enlaces P, R y S). En el grupo II, el sitio con mayor actividad está distribuido entre los enlaces de ambos ligantes hacia el cobre. Para el grupo III, los enlaces con más influencia en la actividad son los que directamente unen la parte C (glicinato) con el metal.

¿Qué propiedades electrónicas influyen en la IC_{50}? El analisis PLS para el grupo II generó un modelo de 3 componentes capaz de describir el 91 % y predecir el 91 % de la variación. En el grupo III, el modelo PLS de 2 componentes describe un 86 % y predice un 66 %. De estos casos, la propiedad electrónica que más influye en la generación del modelo es G para el grupo II, y ρ para el grupo III.

¿Cómo influyen esas propiedades electrónicas? El análisis estadístico PLS mide la influencia de cada descriptor electrónico en la respuesta biológica. En el grupo II las propiedades electrónicas del enlace E disminuyen el valor del IC_{50} con excepción de E_rho. Todos los descriptores del enlace F también disminuyen el valor, con excepción de F_v. Los descriptores de P aumentan el IC_{50} con excepción de P_elep. Los descriptores de los enlaces R y S tienen un comportamiento variado. Para el grupo III los descriptores del enlace P aumentan el IC_{50} con excepción de P_v. Los descriptores del enlace R disminuyen el IC_{50} con excepción de R_lap y finalmente el enlace S tienen descriptores con tendencia variable. Si se quisiera generar una estructura con mayor o menor IC_{50} se tendrían que introducir substituyentes a la molécula que modificaran los descriptores de los enlaces que forman el farmacóforo en la dirección adecuada.

¿Se puede predecir la IC_{50}? Utilizando los valores de los coeficientes generados por los modelos PLS de los grupos II y III, se puede construir una ecuación que nos permita predecir la actividad biológica para un nuevo compuesto. En las Figuras 5.12 y 5.13 podemos ver los valores en los

que influyen para el modelo, de cada uno de los descriptores electrónicos por enlace. Estos valores permiten re-expresar las variables **Y** como modelos de regresión de las variables **X**:

$$\mathbf{Y} = \mathbf{X}B$$

X se refiere a la matriz **X**, incluyendo todos sus términos, y **Y** es la matriz con las variables de respuesta. Los valores de los coeficientes de regresión B dependen de los valores de **X** y **Y** [55]. La correlación entre los valores calculados y predichos de IC_{50} para el grupo II es de 0.91 y para el grupo III es de 0.86.

5.4. Conclusiones

El refinamiento del conjunto de datos inicial nos mostró que se cuentan con 3 familias de moléculas. Los compuestos 1 y 2 se descartaron por no caer dentro del intervalo de confianza establecido por la aplicación estadística. Se trabajó con los compuestos 3 al 12 (Grupo II) y 13 al 21 (Grupo III). Al elaborar los análisis PLS, ambos grupos mostraron buenas correlaciones para R^2X y R^2Y. Los modelos predictivos generados de estos análisis mostraron un 65 % y un 66 %, respectivamente, para el grupo II y el grupo III. Se realizó el análisis para investigar cuál es la propiedad electrónica que más influye en la actividad biológica encontrando que para el grupo II son las energías G y V y para el grupo III la densidad electrónica ρ. Este análisis más refinado generó modelos capaces de predecir el 69 % y 76 % de la actividad, respectivamente. Para el grupo II de compuestos se comprobó que la parte de la molécula que más influye en la respuesta está dada directamente con los átomos coordinados el metal (enlaces S, R, Q y P), así como parte de la estructura de la fenantrolina. En el caso del grupo III, se puede decir que la densidad electrónica alrededor del metal y el glicinato (enlaces P, R y S) es la que más influye en la actividad biológica.

Para el grupo II, contribuyen positivamente a la variable biológica **Y** las variables S_v, S_lap, R_lap, R_v, F_v, E_rho y las propiedades E_lap, E_elep, F_G, F_lap, F_elep contribuyen negativamente. Para el grupo III, las propiedades P_elep, P_rho, P_lap, P_G, R_lap, R_v, S_lap y S_v contribuyen positivamente a la variable biológica, y las propiedades S_elep, P_v, R_elep, R_rho,

R_G y S_rho contribuyen negativamente. Se podría proponer una nueva Casiopeína® con una mayor actividad que las ahora estudiadas si se agregan grupos a la estructura que aumentaran dichas propiedades de la densidad electrónica en los enlaces E, P, R y S del grupo II y P, R y S del grupo III.

Capítulo 6

Fenilbencimidazoles

6.1. Introducción

La necesidad para el desarrollo de nuevos agentes antineoplásicos, más efectivos, más selectivos y menos tóxicos cada vez es mayor. La inhibición de la angiogénesis continúa siendo uno de los principales temas de investigación para la producción de medicamentos. En el estudio de la biología de la angiogénesis tumoral, se han identificado moléculas capaces de inhibir el desarrollo del tumor. De especial interés se encuentra el factor de crecimiento derivado de plaquetas (platelet-derived growth factor, PDGF) como uno de los principales reguladores del crecimiento celular [56, 57, 58]. Se ha descubierto que este factor, así como su receptor (platelet-derived growth factor receptor, PDGFR), intervienen en la inhibición del crecimiento de células endoteliales, el principal componente de los vasos sanguíneos. La posibilidad de estos factores de promover proliferación celular, migración y de inducir cambios enzimáticos los hace de gran importancia para tratar condiciones patofisiológicas en donde los factores de crecimiento están ausentes o inhibidos. La unión del PDGF con su receptor transmembranal, PDGFR, resulta en la fosforilación de tirosina de sustratos activos que intervienen en diferentes caminos metabólicos, incluyendo la 3-fosfatidilinositolcinasa [59]. Se han reportado varios grupos de moléculas capaces de inhibir el PDGFR.

Una de estas familias corresponde a las 3-arilquinolinas que muestran valores de IC_{50} 80nM, bloqueando la unión de ATP de la autofosforilación del PDGFR en células de músculo liso [60, 61]. Algunas 3-arilquinolinas muestran

un IC_{50} de 300nM para la inhibición de la autofosforilación del PDGFR en células 3T3 [62]. Otros han reportado que inhiben el funcionamiento del PDGFR y se están haciendo pruebas clínicas para el tratamiento del glioma [63].

También se han reportado como inhibidores de la autofosforilación del PDGFR a concentraciones de nanomoles las fenilaminopirimidinas [64, 65]. Se busca utilizar estos compuestos como agentes anticancerígenos.

Los 1-fenilbencimidazoles se han reportado recientemente como inhibidores selectivos del PDGFR, con una clara relación entre las propiedades moleculares y su actividad inhibitoria [56, 66]. Existe poca literatura con respecto al desarrollo de modelos QSAR para los fenilbencimidazoles y los modelos reportados no son del todo satisfactorios [67, 68, 69, 70]. Estudios con bencimidazoles utilizando CoMFA se han realizado generando modelos de predicción con q^2 muy altos. [71].

Shen y otros [72] propusieron varios modelos QSAR para determinar la actividad inhibitoria de los 1-fenilbencimidazoles sobre los inhibidores de la PDGFR utilizando como descriptores electrónicos las cargas de densidad calculados con AM1 para ciertos átomos dentro de las moléculas. Consideraron compuestos con diferentes substituyentes en el anillo del bencimidazol (Figura 6.1). Las ecuaciones de correlación que obtuvieron tienen una r^2 de 0.66 a 0.73 para un conjunto de datos de 75 compuestos. Zhong y otros [73] obtuvieron modelos QSAR incluyendo los compuestos utilizados en la referencia [72] y utilizaron los índices de conectividad como principales descriptores para su modelo de relación estructura-actividad. Obtuvieron $r^2 = 0.78$ para un conjunto de datos de 55 derivados del 1-fenilbencimidazoles. Este trabajo utiliza estos conjuntos de datos para realizar el estudio QSAR.

6.2. Resultados y discusión

Esta sección investigará el sitio activo o parte reactiva de una familia de 123 fenilbencimidazoles empleando la metodología descrita en el capítulo 3. Los datos experimentales son las concentraciones mínimas para inhibir la fosforilación de copolímeros de tirosina o glutamato por proteínas del PDFGR (IC_{50}) [67]. El juego de datos consiste de 123 fenilbencimidazoles y sus valores de IC_{50} con diferentes substituyentes de acuerdo al Cuadro A.1. La estructura común de todos los compuestos se muestra en la Figura 6.1.

No.	R	X	Log(I/IC_{50}[1])
1	H	H	5.030
2	4-OMe	H	4.301
3	4-OH	H	4.854
4	5-Me	H	5.357
5	5-OMe	H	6.367
6	5-OH	H	6.357
7	5-Cl	H	5.398
8	5-COOH	H	5.030
9	5-COOMe	H	6.081
10	5-$CONH_2$	H	4.796
11	5-NO_2	H	4.796
12	5-COMe	H	6.066
13	5-CHO	H	6.367
14	5-OC_2H_7	H	6.602
15	5-OC_2H_5	H	6.620
16	5-$OCH(Me)_2$	H	5.509
17	5-OC_4H_9	H	5.886
18	5-$OCH_2CH{=}CH_2$	H	6.215
19	5-$O(CH_2)_4OH$	H	6.347
20	5-OCH_2(oxyranyl)	H	6.497
21	5-$OCH_2CH(OH)CH_2OH$	H	6.509
22	5-$O(CH_2)_2OH$	H	6.187
23	5-$O(CH_2)_2N(Me)_2$	H	5.824
24	5-$O(CH_2)_3N(Me)_2$	H	6.824

[1]IC50(μM): concentración mínima del compuesto capaz de lnhibir la autofosforilación del PDGFR en células 3T3

25	5-O$(CH_2)_4$N$(Me)_2$	H	6.796
26	5-O$(CH_2)_2$Nmorph	H	6.137
27	5-O$(CH_2)_3$Nmorph	H	6.770
28	5-O$(CH_2)_4$Nmorph	H	6.569
29	5-SH	H	5.482
30	5-SMe	H	6.131
31	5-OCSN$(Me)_2$	H	5.337
32	6-Me	H	4.398
33	6-OMe	H	5.194
34	6-OH	H	5.678
35	6-Cl	H	5.268
36	6-COOH	H	4.301
37	6-COOMe	H	4.886
38	6-$CONH_2$	H	4.602
39	6-NO_2	H	4.301
40	6-NH_2	H	4.638
41	7-OMe	H	4.432
42	4,5DiOH	H	4.602
43	4-OH,5-OMe	H	5.149
44	4-CH_2CH(Me)O-5	H	4.538
45	5,6-DiOH	H	4.638
46	5,6-DiMe	H	5.921
47	5,6-OCH_2O	H	5.658
48	5-OMe,6-Me	H	6.000
49	5-OH,6-Me	H	5.602
50	5-OMe,6-COOH	H	4.678
51	5-OH,6-COOH	H	5.367
52	5-OMe,6-COOMe	H	6.061
53	5-OMe,6-CH_2OH	H	6.432
54	5-OMe,6-CHO	H	6.000
55	5-NH_2	H	5.569
56	5-Aza	H	5.000
57	7-Aza	H	4.553
58	H	3'-Me	4.553
59	H	3'-OMe	4.602
60	H	3'-OH	5.420
61	H	3'-Cl	4.328
62	H	3'-NO_2	4.796

63	H	3'-NH_2	5.444
64	H	3'-COMe	4.721
65	H	3'-CHO	5.168
66	H	4'-OMe	4.886
67	H	4'-OH	5.745
68	H	4'-Cl	4.300
69	H	4'-COOMe	5.143
70	H	4'-$CONH_2$	4.638
71	H	4'-NO_2	4.523
72	H	4'-NH_2	5.252
73	H	4'-COMe	4.620
74	H	4'-CHO	4.886
75	H	4'-CN	4.796
76	H	4'-Aza	4.921
77	5-OMe	2'-Tienil	5.602
78	5-OMe	3'-Tienil	6.155
79	5-OMe	4'-NH_2	6.553
80	4-COOH	H	<4.3
81	4-COOMe	H	<4.3
82	4-$CONH_2$	H	<4.3
83	4-NO_2	H	<4.3
84	4-NH_2	H	<4.3
85	7-Me	H	<4.3
86	7-OH	H	<4.3
87	7-Cl	H	<4.3
88	7-COOH	H	<4.3
89	7-COOMe	H	<4.3
90	7-$CONH_2$	H	<4.3
91	7-NO_2	H	<4.3
92	7-NH_2	H	<4.3
93	4-OMe,5-OH	H	<4.3
94	4,5-DiOMe	H	<4.3
95	4-Br,5-Oh	H	<4.3
96	4-Br,5-$OCH_2CH{=}CH_2$	H	<4.3
97	4-$CH_2CH{=}CH_2$,5-OH	H	<4.3
98	5-$S(CH_2)_3$Nmorph	H	<4.3
99	4-Me	H	<4.3
100	4-Cl	H	<4.3

101	2-Me	H	<4.3
102	2-OH	H	<4.3
103	2-NH_2	H	<4.3
104	H	2'-Me	<4.3
105	H	2'-OMe	<4.3
106	H	2'-OH	<4.3
107	H	2'-Cl	<4.3
108	H	2'-COOH	<4.3
109	H	2'-COOEt	<4.3
110	H	2'-$CONH_2$	<4.3
111	H	2'-NO_2	<4.3
112	H	2'-NH_2	<4.3
113	H	2'-COMe	<4.3
114	H	2'-CHO	<4.3
115	H	2'-CN	<4.3
116	H	3'-COOH	<4.3
117	H	3'-COOEt	<4.3
118	H	3'-$CONH_2$	<4.3
119	H	3'-CN	<4.3
120	H	4'-Me	<4.3
121	H	4'-COOH	<4.3
122	H	2'-Aza	<4.3
123	H	3'-Aza	<4.3

Cuadro 6.1: Substituyentes para los fenilbencimidazoles y los valores de las actividades biológicas.

Los enlaces de cada uno de los fenilbencimidazoles se nombraron de acuerdo a la Figura 6.1.

A cada enlace le corresponde 5 propiedades electrónicas extraídas directamente de la densidad electrónica (ϵ, ρ, $\nabla^2\rho$, G, V) y, por cada compuesto, le corresponde un valor de IC_{50}. Este último valor se usó como variable de respuesta, y las 5 propiedades de cada enlace, como variables controladas.

El primer paso fue efectuar un análisis de componentes principales sobre todas las observaciones del juego de datos para encontrar si existen familias o compuestos con tendencias comunes y compuestos que se tuvieran que

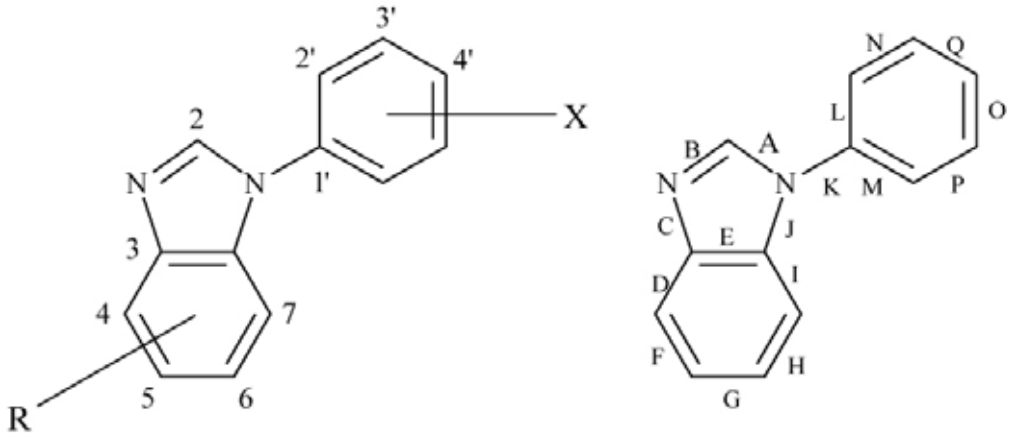

Figura 6.1: Estructura común para los fenilbencimidazoles y diagrama donde se muestran las etiquetas de cada enlace involucrado en el análisis.

desechar por estar fuera de los límites esperados. El análisis nos da un modelo con cuatro componentes que explica el 46 % de la variación. Podemos observar en la Figura 6.2 cómo la mayor parte de los compuestos son homogéneos y no detectamos grupos o separaciones notables. Encontramos también 5 compuestos que se salen por completo de los límites del análisis por lo que se desecharon. Esto puede ser causado por errores en el proceso de cálculo o a la hora de extraer las propiedades de la densidad electrónica.

Se realizó un segundo análisis de componentes principales, excluyendo ahora las observaciones 57, 76, 98, 122 y 123, y se obtiene un modelo con 22 componentes principales. Aunque se explica con este modelo el 97 % de la variación, analizando las gráficas de los *scores* del primer componente con el segundo, todavía encontramos observaciones que están lejos de influir en la variación en **X**. Sabemos que el primer componente explica la mayor variación del espacio **X** (las observaciones). Comparando el primer componente con los 21 restantes vemos que son los mismos compuestos en todos los componentes que se tienen que eliminar. Descartando las observaciones alejadas, se refinó el modelo hasta obtener una gráfica de componentes donde todas las observaciones cayeran dentro del límite. Al final utilizaremos 80 observaciones para obtener un modelo con 7 componentes explicando el 61 % de la variación. Se puede ver cómo las observaciones que se utilizarán para continuar el análisis todas influyen prácticamente en un mismo grado para

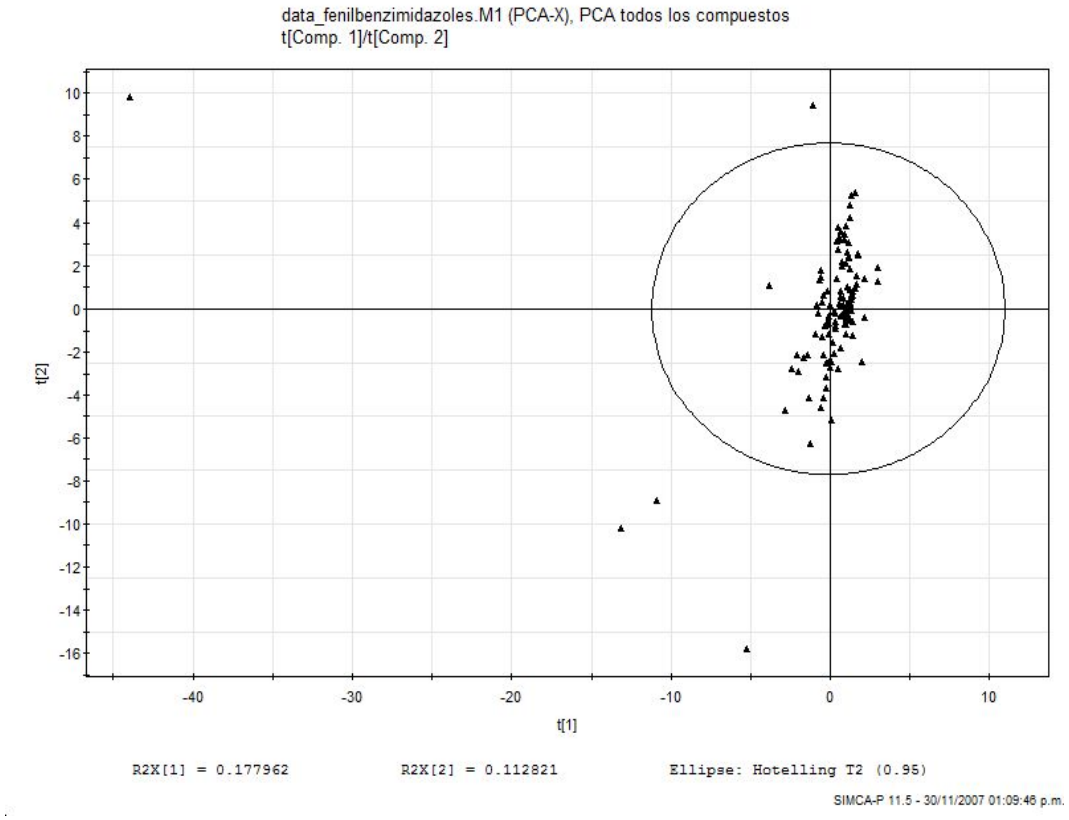

Figura 6.2: Análisis inicial de los 123 fenilbencimidazoles donde se muestra la homogeneidad de las observaciones, así como 5 compuestos que son descartados por no entrar en el rango aceptado.

explicar la variación de **X** y ninguna se aleja de los límites (Figura 6.3). En el apéndice A se encuentra el listado de los compuestos que entraron dentro de los límites para la generación del modelo estadístico.

Se realizó el análisis incorporando ahora las variables de respuesta. Para esta familia, se utilizará la concentración mínima inhibitoria (IC_{50}) que tienen los fenilbencimidazoles de bloquear el receptor del factor de crecimiento derivado de las plaquetas (PDGFR) como ya se vió en la introducción. El análisis de mínimos cuadrados parciales se realizó utilizando las 80 observaciones que ya comprobamos en el análisis de componentes principales como el más adecuado. Inicialmente obtenemos un modelo con 2 componen-

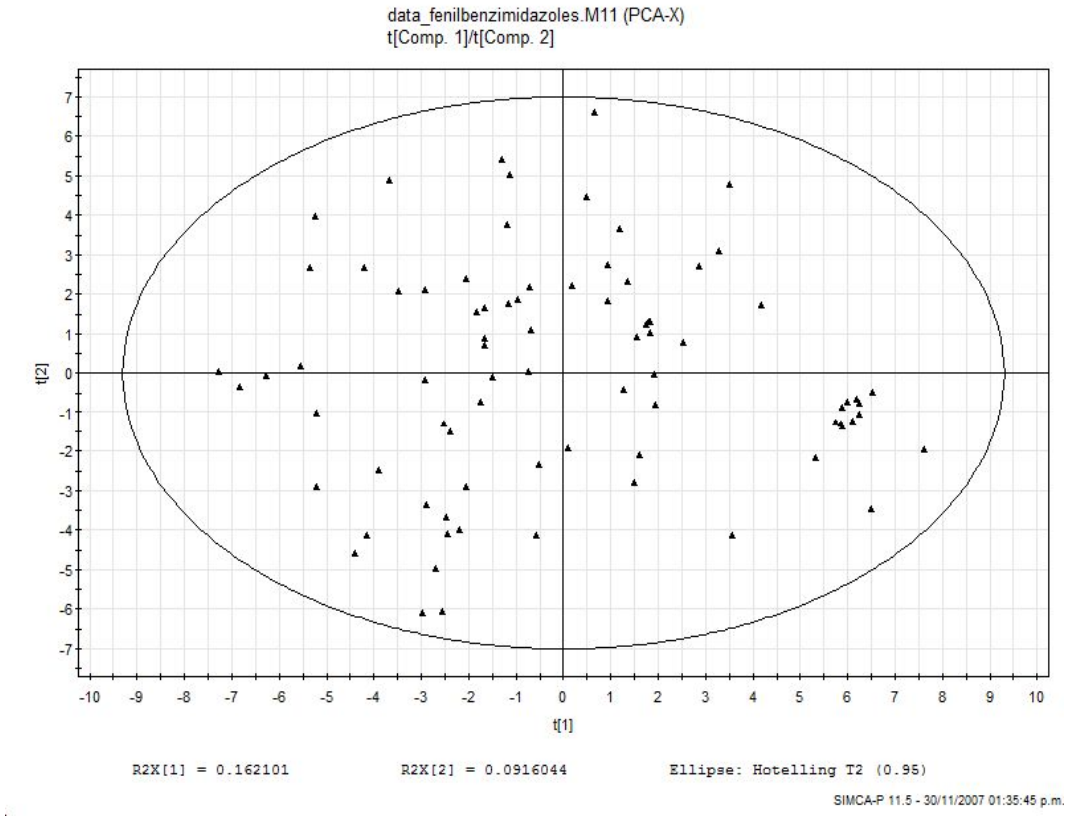

Figura 6.3: Figura del análisis de los 2 primeros componentes principales utilizando sólo las observaciones que son homogéneas y que no se alejan de los límites establecidos. Se utilizarán 80 observaciones que forman un modelo de 7 componentes explicando el 61 % de la variación.

tes explicando el 21 % de la variación y prediciendo en un 51 %. Sin embargo, analizando la gráfica donde se compara el primer componente de **X** con el primer componente de Y, vemos datos que tienen exactamente el mismo valor en la variable de respuesta. Esto era de esperarse pues nuestro listado de actividades presenta el mismo valor de la actividad biológica del compuesto 80 al 123 (Figura 6.4).

Eliminando los compuestos cuyo valor de actividad biológica no es funcional, realizamos un análisis de mínimos cuadrados parciales. El modelo se generó con 72 observaciones utilizando todos los descriptores electrónicos y

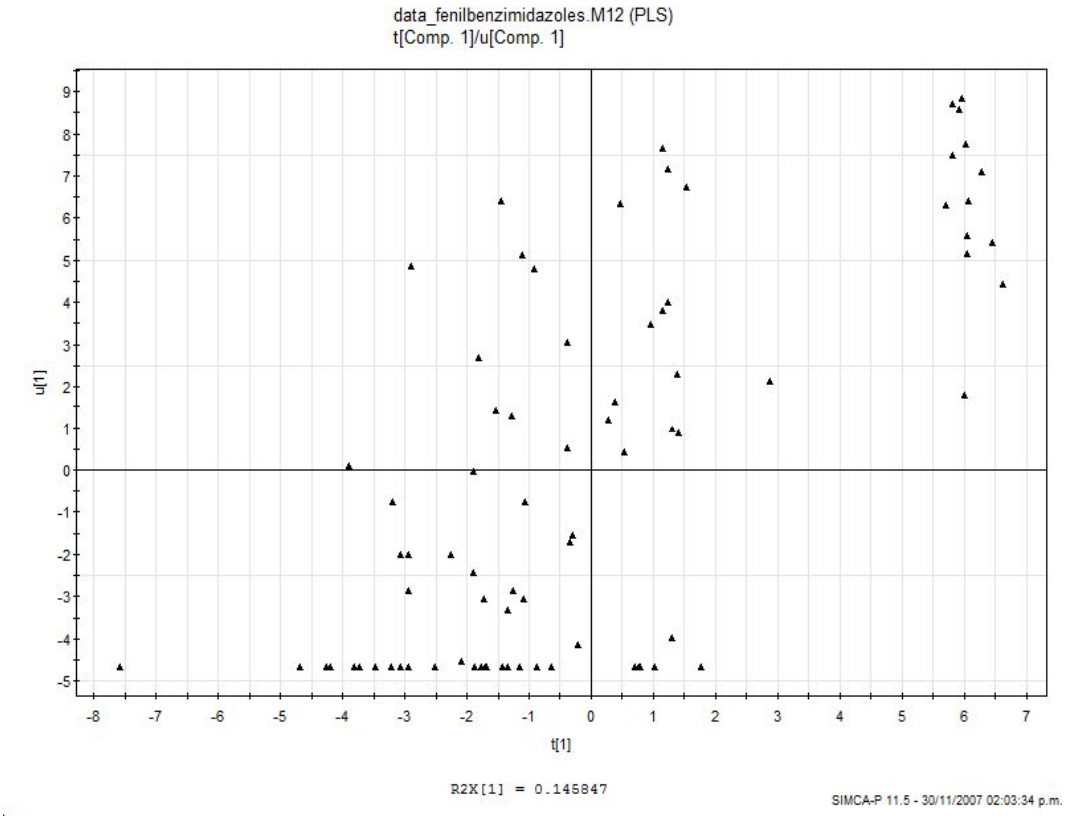

Figura 6.4: Primer análisis PLS donde encontramos las observaciones que tienen el mismo valor de variable de respuesta (actividad biológica) y se ven agrupados en el eje Y.

la variable de actividad biológica. Se obtiene un modelo con un solo componente principal. El modelo explica el 45 % de la variación y predice el 30 % ($R^2Y = 0.455$ y $Q^2 = 0.306$). Para conocer cuáles son los enlaces o parte de la molécula que describe la actividad biológica, se emplea la gráfica de VIP (Variable Importance on Projection, Figura 6.5). Se puede observar que el enlace B (referirse a la Figura 6.1) tiene sus 5 descriptores electrónicos con un índice de importancia mayor a uno. Los valores de B_Rho y B_V son los más altos, seguidos por D_Lap y B_Gke. El valor de B_Lap y B_ellep también están presentes en el diagrama VIP. Los enlaces A y E aparecen con 4 descriptores electrónicos (A_Gke, A_V, A_Rho, A_Lap, E_Rho, E_Lap, E_V y E_Gke) seguido por el enlace H con 3 descriptores (H_Gke, H_V y H_Rho).

La siguiente influencia está dada por el enlace D (D_Lap y D_V), el enlace G (G_ellep y G_Lap) y el enlace K (K_Gke y K_V). Otros descriptores presentes son F_ellep, C_Gke J_ellep, J_Rho e I_Lap.

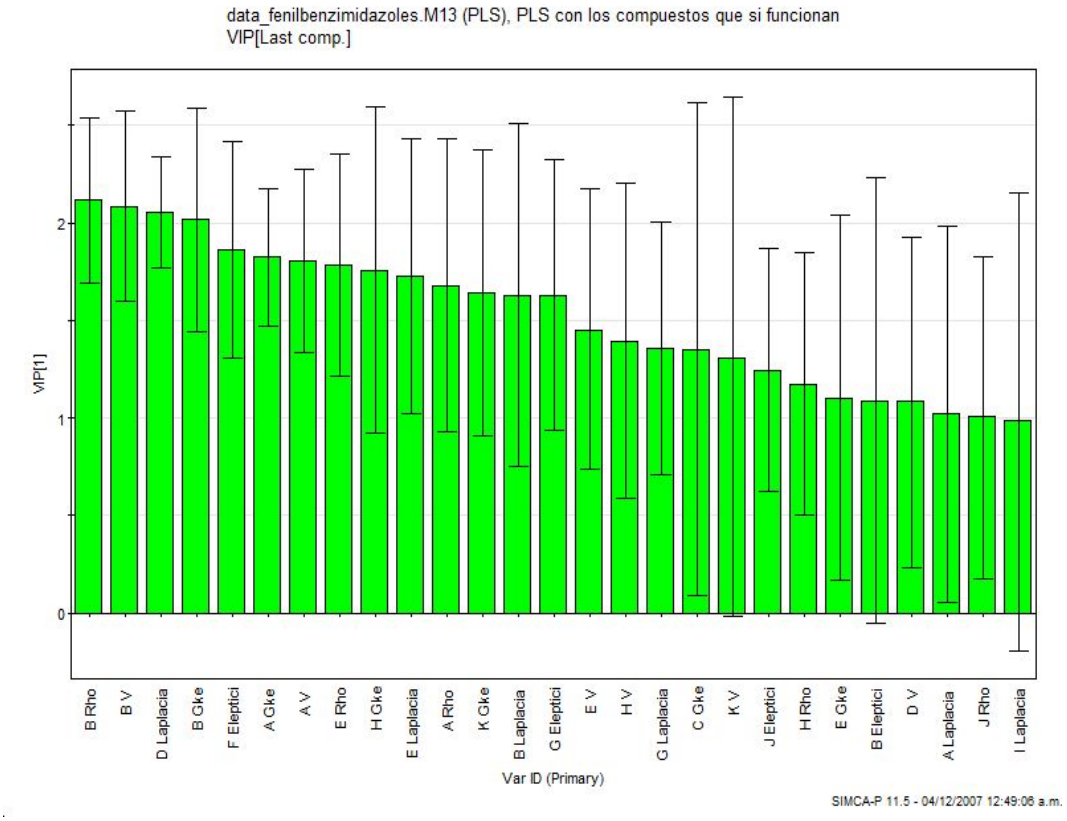

Figura 6.5: Gráfica de variables importantes en la proyección (VIP) para la regresión PLS (Sólo se muestras aquellos índices mayores a 1).

En la Figura 6.6 se tienen los valores de actividad contra los valores de actividad calculados a partir del modelo generado en el PLS. Se obtiene una regresión del $R^2 = 0.455$. Se puede ver en esta Figura como existen todavía muchas observaciones que no producen una buena respuesta e influyen sobre la baja correlación entre los datos.

Para encontrar mejores correlaciones y conocer la importancia de cada descriptor, se estudió la influencia de cada uno de manera independiente.

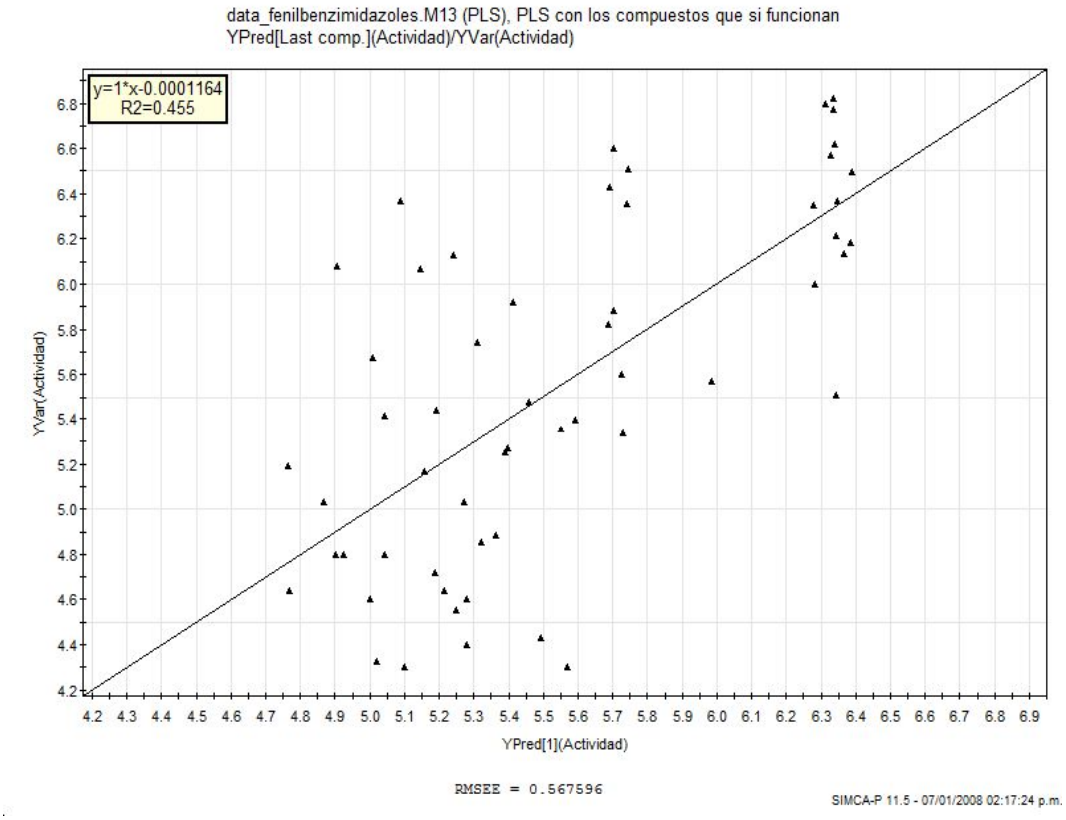

Figura 6.6: Gráfica de valores de actividades biológicas contra los valores de actividad biológica calculada a partir del modelo generado por el análisis PLS.

En el cuadro 6.2 se presenta el análisis PLS por cada uno de los 5 descriptores electrónicos utilizados como variables de respuesta. Se dan los valores de R^2X, R^2Y y Q^2(cum). La propiedad que más describe la variación con respecto a $\mathbf{X}$ sería la elepticidad ϵ con $R^2X = 0.261$ seguido por la densidad electrónica ρ con $R^2X = 0.234$. Con respecto a $\mathbf{Y}$, las energías son los descriptores que presentan el valor más alto, primero G con $R^2Y = 0.511$ y V con $R^2Y = 0.422$. La propiedad que mejor describe la actividad una vez más son las energías con 0.405 de Q^2 para G y 0.283 para V. Si nos referimos a la Figura de los índices que principalmente influyen en el modelo (VIP, Figura 6.5), podemos ver cómo las energías también están presentes en los enlaces que más influyen. Están presentes ambas energías de los enlaces B, A, E, H,

Propiedad	R^2X	R^2Y	Q^2(cum)
ϵ	0.261	0.322	0.160
ρ	0.234	0.371	0.249
$\nabla^2\rho$	0.197	0.399	0.247
G	0.210	0.511	0.405
V	0.222	0.422	0.283

Cuadro 6.2: Valores de R^2X, R^2Y y Q^2(cum) analizando las propiedades de la densidad electrónica por separado en 56 fenilbencimidazoles.

K, así como las energías de los enlaces D y C. El modelo que mejor describe la actividad es generado también por ambas energías, principalmente por G. La densidad electrónica ρ también tiene una contribución importante en la descripción de la variación. La gráfica de los índices VIP muestra como principal variable influyente la densidad electrónica del enlace B y después de las energías de enlace, es el descriptor que más contribuye para la creación del modelo PLS.

6.3. Conclusiones

El grupo farmacóforo de la familia de anticancerígenos inhibidores de la PDGFR se encuentra principalmente en los enlaces B, D, F y A del bencimidazol. Los descriptores electrónicos que más influyen en el modelo de predicción son B_rho, B_v, D_lap, B_Gke, F_elep, A_Gke, A_v, E_rho, H_Gke, E_lap, A_rho, K_Gke y B_lap. El enlace que mas contribuye al modelo es el B, dentro del anillo del bencimidazol, seguido por el enlace A. La actividad biológica de estos compuestos está relacionada con la densidad electrónica que contribuyen los substituyentes y se localizan en los enlaces A, B, D, E y F. La energía cinética (G) y pontencial (v) son los descriptores que contribuyen más a la formación del modelo PLS y los descriptores que mejor describen la actividad biológica. La baja correlacion y predicción del modelo posiblemente se deba a que los descriptores utilizados en este estudio son de caracter electrónico y es posible que otro tipo de efecto esté involucrado en el origen de la actividad biológica como puede ser el efecto estérico.

Capítulo 7

Conclusiones

El presente trabajo demuestra la eficiencia del metodo QTMS y el uso de los descriptores electrónicos basados en la topología de la densidad electrónica para elaborar estudios QSAR. Se usó eficientemente el poder predictivo de cálculos *ab-initio* combinados con metodología QSAR usando descriptores cuánticos y un análisis estadístico riguroso. Se llegó a identificar la influencia de los sustituyentes y el sitio activo relacionado con la acidez del ácido benzoico, así como la parte biológicamente activa en los fenilbencimidazoles y las Casiopeínas®. Los análisis PLS mostraron la estructura electrónica que describe la actividad encontrada experimentalmente.

Observamos que para cada grupo de moléculas no basta con un solo descriptor electrónico, sino es necesario utilizar los 5 descriptores que se manejaron en este trabajo para generar un modelo confiable. Además se estudió la influencia en el modelo de cada descriptor y se analizó cada propiedad por separado. En el caso del ácido benzoico se realizó un análisis de la deformación de la densidad electrónica en los enlaces ya identificados previamente como activos. Se encontró una fuerte relación entre la cantidad de densidad electrónica y el valor de pKa de la molécula. Para las Casiopeínas®, se realizó el análisis QTMS resultando en un modelo muy aceptable tanto para describir el comportamiento de la actividad, como para predecirlo. La actividad biológica se encontró relacionada a los enlaces coordinados con el metal en ambos grupos.

En las 3 familias se aplicó satisfactoriamente la metodología QTMS y se generaron modelos de predicción que satisfacen los alcances de un análisis

de la relación estructura-actividad, ya que se puede saber qué descriptores tienen mayor influencia y en qué dirección se da ésta. En el caso de los fenilbencimidazoles, a pesar de contar con un conjunto de datos muy grande (N = 121), pocos datos fueron suficientemente confiables para elaborar el modelo. Sin embargo, con el conjunto resultante (N = 56) se elaboró satisfactoriamente un modelo con un nivel de predicción aceptable. Utilizando QTMS se identificó el sitio activo en la región del bencimidazol que concuerda con trabajos previos [72, 73].

Apéndice A

Listado de los fenilbencimidazoles que se utilizaron para la generación del modelo PLS.

No.	R	X	Log(I/IC_{50}
1	H	H	5.030
2	4-OMe	H	4.301
3	4-OH	H	4.854
4	5-Me	H	5.357
5	5-OMe	H	6.367
6	5-OH	H	6.357
7	5-Cl	H	5.398
8	5-COOH	H	5.030
9	5-COOMe	H	6.081
10	5-$CONH_2$	H	4.796
11	5-NO_2	H	4.796
12	5-COMe	H	6.066
13	5-CHO	H	6.367
14	5-OC_2H_7	H	6.602
15	5-OC_2H_5	H	6.620
16	5-$OCH(Me)_2$	H	5.509
17	5-OC_4H_9	H	5.886
18	5-$OCH_2CH{=}CH_2$	H	6.215
19	5-$O(CH_2)_4OH$	H	6.347
20	5-OCH_2(oxiranil)	H	6.497
21	5-$OCH_2CH(OH)CH_2OH$	H	6.509
22	5-$O(CH_2)_2OH$	H	6.187

23	5-O$(CH_2)_2$N$(Me)_2$	H	5.824
24	5-O$(CH_2)_3$N$(Me)_2$	H	6.824
25	5-O$(CH_2)_4$N$(Me)_2$	H	6.796
26	5-O$(CH_2)_2$Nmorf	H	6.137
27	5-O$(CH_2)_3$Nmorf	H	6.770
28	5-O$(CH_2)_4$Nmorf	H	6.569
29	5-SH	H	5.482
30	5-SMe	H	6.131
31	5-OCSN$(Me)_2$	H	5.337
32	6-Me	H	4.398
33	6-OMe	H	5.194
34	6-OH	H	5.678
35	6-Cl	H	5.268
36	6-COOH	H	4.301
37	6-COOMe	H	4.886
38	6-$CONH_2$	H	4.602
39	6-NO_2	H	4.301
40	6-NH_2	H	4.638
41	7-OMe	H	4.432
42	4,5DiOH	H	4.602
43	4-OH,5-OMe	H	5.149
44	4-CH_2CH(Me)O-5	H	4.538
45	5,6-DiOH	H	4.638
46	5,6-DiMe	H	5.921
47	5,6-OCH_2O	H	5.658
48	5-OMe,6-Me	H	6.000
49	5-OH,6-Me	H	5.602
50	5-OMe,6-COOH	H	4.678
51	5-OH,6-COOH	H	5.367
52	5-OMe,6-COOMe	H	6.061
53	5-OMe,6-CH_2OH	H	6.432
54	5-OMe,6-CHO	H	6.000
55	5-NH_2	H	5.569
56	5-Aza	H	5.000
57	7-Aza	H	4.553
58	H	3'-Me	4.553
59	H	3'-OMe	4.602
60	H	3'-OH	5.420

61	H	3'-Cl	4.328
62	H	3'-NO_2	4.796
63	H	3'-NH_2	5.444
64	H	3'-COMe	4.721
65	H	3'-CHO	5.168
66	H	4'-OMe	4.886
67	H	4'-OH	5.745
68	H	4'-Cl	4.300
69	H	4'-COOMe	5.143
70	H	4'-$CONH_2$	4.638
71	H	4'-NO_2	4.523
72	H	4'-NH_2	5.252
73	H	4'-COMe	4.620
74	H	4'-CHO	4.886
75	H	4'-CN	4.796
76	H	4'-Aza	4.921
77	5-OMe	2'-Tienil	5.602
78	5-OMe	3'-Tienil	6.155
79	5-OMe	4'-NH_2	6.553
80	4-COOH	H	<4.3
81	4-COOMe	H	<4.3
82	4-$CONH_2$	H	<4.3
83	4-NO_2	H	<4.3
84	4-NH_2	H	<4.3
85	7-Me	H	<4.3
86	7-OH	H	<4.3
87	7-Cl	H	<4.3
88	7-COOH	H	<4.3
89	7-COOMe	H	<4.3
90	7-$CONH_2$	H	<4.3
91	7-NO_2	H	<4.3
92	7-NH_2	H	<4.3
93	4-OMe,5-OH	H	<4.3
94	4,5-DiOMe	H	<4.3
95	4-Br,5-Oh	H	<4.3
96	4-Br,5-$OCH_2CH{=}CH_2$	H	<4.3
97	4-$CH_2CH{=}CH_2$,5-OH	H	<4.3
98	5-$S(CH_2)_3$Nmorf	H	<4.3

99	4-Me	H	<4.3
100	4-Cl	H	<4.3
101	2-Me	H	<4.3
102	2-OH	H	<4.3
103	2-NH_2	H	<4.3
104	H	2'-Me	<4.3
105	H	2'-OMe	<4.3
106	H	2'-OH	<4.3
107	H	2'-Cl	<4.3
108	H	2'-COOH	<4.3
109	H	2'-COOEt	<4.3
110	H	2'-$CONH_2$	<4.3
111	H	2'-NO_2	<4.3
112	H	2'-NH_2	<4.3
113	H	2'-COMe	<4.3
114	H	2'-CHO	<4.3
115	H	2'-CN	<4.3
116	H	3'-COOH	<4.3
117	H	3'-COOEt	<4.3
118	H	3'-$CONH_2$	<4.3
119	H	3'-CN	<4.3
120	H	4'-Me	<4.3
121	H	4'-COOH	<4.3
122	H	2'-Aza	<4.3
123	H	3'-Aza	<4.3

Cuadro A.1: Substituyentes para los fenilbencimidazoles y los valores de las actividades biológicas.

Bibliografía

[1] C. Hansch, R. M. Muir, T. Fujita, P. P. Maloney, E. Geiger, and M. Streich. The correlation of biological activity of plant growth regulators and chloromycetin derivatives with hammet constants and partition coefficients. *Journal of the American Chemical Society*, 85:2817–2824, 1963.

[2] C. Hansch and T. Fujita. ρ-σ-π analysis, method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, 85:1616–1626, 1964.

[3] A. Tropsha and W. Zheng. *Computational Biochemistry and Biophysics*, chapter Computer aided drug design, pages 351–369. Marcel Dekker, New York, 2001.

[4] A. K. Debnath. *Combinatorial Library Design and Evaluation for Drug Discovery: Principles, Methods, Software Tools and Applications.*, chapter Quantitative structure-activity relationship (QSAR): a versatile tool in drug design., pages 73–129. Marcel Dekker, New York, 2001.

[5] H. Kubini. *Encyclopedia of Computational Chemistry*, chapter Quantitative structure-activity relationships in drug design., pages 2309–2320. John Wiley and Sons, West Sussex, UK, 1998.

[6] A. Tropsha. *Burger's Medicinal Chemistry and Drug Discovery*, volume 1, chapter Recent trands in quantitative structure-activity relationships., pages 49–76. John Wiley and Sons, Hoboken, New Jersey, 2003.

[7] G. Greco, E.Ñovellino, and I. C. Martin. *Reviews in Computational Chemistry*, volume 11, chapter Approaches to three-dimensional quanti-

tative structure-activity relationships, pages 183–240. VCH Publishers, 1997.

[8] Cramer III, R. D. Patterson, and J. D. Bunce. Comparative molecular field analysis (comfa). effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society*, 110:5959, 1988.

[9] G. Klebe, U. Abraham, and T. Mietzner. Molecular similarity indeces in a comparative analysis (comsia) of drug molecules to correlate and predict their biological activity. *Journal of Medical Chemistry*, 37:4130–4146, 1994.

[10] J. A. Calder. Comfa validation of the superposition of six classes of compunds which block GABA receptors non-competitively. *Journal of Computer Aided Molecular Design*, 7, 1993.

[11] S. A. DePriest. 3d-QSAR of angiotensin-converting enzyme and thermolysin inhibitors: a comparison of comfa models based on deduced and experimentally determined active site geometries. *Journal of the American Chemical Society*, 115:5372–84, 1993.

[12] G. Greco. Comparative molecular field analysis on a set of muscarinic agonists. *Quantitative Structure Activity Relationships.*, 10:289–299, 1991.

[13] K. Prendergast. Derivation of a 3d pharmacophore model for the angiotensin-II site on receptor. *Journal of Computer Aided Molecular Design*, 8:491–512, 1994.

[14] J. P. Horwitz. Comparative molecular field analysis of in vitro growth inhibition of L1210 and HCT-8 cells by some pyrazoloacridines. *Journal of Medical Chemistry*, 36:3511–3516, 1993.

[15] G. Klebe and U. Abraham. On the prediction of binding properties of drug molecules by comparative molecular field analysis. *Journal of Medical Chemistry*, 36:70–80, 1993.

[16] A. M. Myers. Conformational analysis, pharmacophore identification, and comparative molecular field analysis of ligands for the neuromodulatory σ3receptor. *Journal of Medical Chemistry*, 37:4109–4117, 1994.

[17] W. Zheng and A. Tropsha. Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle. *Journal of Chemical Information and Computational Science*, 40:185–194, 2000.

[18] J. G. Topliss and R. J. Costello. Chance correlations in structure-activity studies using multiple regression analysis. *Journal of Medicinal Chemistry*, 15:1066–1068, 1972.

[19] A. Golbraikh and A. Tropsha. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Journal of Computer-Aided Molecular Design*, 16:357–369, 2002.

[20] A. Golbraikh and A. Tropsha. Beware of q^2!. *Journal of Molecular Graphics and Modelling*, 20:269–276, 2002.

[21] A. Tropsha, P. Gramatica, and V. K. Gombar. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR and Combinatorial Science*, 22:69–77, 2003.

[22] R. Hoffmann. *Lo mismo y no lo mismo.* Fondo de cultura económica, 1990.

[23] Richard F. W. Bader. A bond path: A universal indicator of bonded interactions. *Journal of Physical Chemistry A*, 102:7314–7323, 1998.

[24] D. Cremer and E. Kraka. Chemical bonds without bonding electron density - does the difference electron-density analysis suffice for a description of the chemical bond? *Angewandte Chemie International Edition*, 23:627–628, 1984.

[25] Richard F. W. Bader. *Atoms in Molecules: A Quantum Theory.* Oxford University Press, USA, 1994.

[26] E. Espinosa, E. Molins, and C. Lecomte. Hydrogen bond strengths revealed by topological analyses of experimentally observed electron densities. *Journal of Physical Chemistry A*, 285:170–173, 1998.

[27] O. A. Zhikol, O. Shishkin, K. A. Lyssenko, and J. Leszczynski. Electron density distribution in stacked benzene dimers: A new approach towards the estimation of stacking interaction energies. *Journal of Chemical Physics*, 122:144104–144104–8, 2005.

[28] M. P. Waller, A. Robertazzi, J. A. Platts, D. E. Hibbs, and P. A. Williams. Hybrid density functional theory for π-stacking interactions: Application to benzenes, pyridines, and DNA bases. *Journal of Computational Chemistry*, 27:491–504, 2006.

[29] C. F. Matta, N. Castillo, and R. J. Boyd. Extended weak bonding interactions in DNA: π-stacking (base-base), base-backbone and backbone-backbone interactions. *Journal of Physical Chemistry B*, 110:563–578, 2006.

[30] S. E. O'Brien and P. L. A. Popelier. Quantum molecular similarity. part 2: The relation between properties in BCP space and bond length. *Canadian Journal of Chemistry*, 77:28–36, 1999.

[31] P. L. A. Popelier. Quantum molecular similarity. 1. BCP space. *Journal of Physical Chemistry A*, 103:2883–2890, 1999.

[32] S. E. O'Brien and P. L. A. Popelier. Quantum molecular similarity. 3. QTMS descriptors. *Journal of Chemical Information and Computer Sciences*, 41:764–775, 2001.

[33] U. A. Chaudry and P. L. A. Popelier. Estimation of pKa using quantum topological molecular similarity descriptors: Application to carboxylic acids, anilines and phenols. *Journal of Chemical Information and Computer Sciences*, 69:233–241, 2004.

[34] R. Carbó, L. Leyda, and M. Arnau. How similar is a molecule to another? an electron density measure of similarity between two molecular structures. *International Journal of Quantum Chemistry*, 17:1185–1189, 1980.

[35] Kevin C. Gross and Paul G. Seybold. Substituent effects on the physical properties and pKa of phenol. *International Journal of Quantum Chemistry.*, 85:569–579, 2001.

[36] Kevin C. Gross, Paul G. Seybold, and Christopher M. Comparison of different atomic charge schemes for predicting pKa variations in substituted anilines and phenols. *International Journal of Quantum Chemistry.*, 90:445–458, 2002.

[37] Ana M. Graña, José M. Hermida-Ramón, and Ricardo A. Mosquera. QTAIM interpretation of the basicity of substituted anilines. *Chemical Physics Letters.*, 412:106–109, 2005.

[38] J. E. Jackson. *A User's Guide to Principal components.* John Wiley, 1991.

[39] S. Wold, W. J. Dunn, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, E. Lindberg, and M. Sjöström. *Chemometrics: Mathematics and Statistics in Chemistry*, chapter Multivariate Data Analysis in Chemistry. Reidel Publishing Company, Dordrecht, Holland, 1984.

[40] A. Höskuldsson. *Prediction Methods in Science and Technology.* Thor Publishing, Copenhagen, Denmark, 1996.

[41] S. Wold and M. Josefson. *Encyclopedia of Analytical Chemistry*, chapter Multivariate Calibration of Analytical Data, pages 9710–9736. John Wiley and Sons, Ltd, 2000.

[42] S. Wold, J. Trygg, A. Brglund, and H. Antti. PLS-regression: A basic tool of chemometrics. *Journal of Chemometrics*, 58:109–130, 2001.

[43] S. Wold, M. Sjöström, and L. Eriksson. *Encyclopedia of Computational Chemistry*, chapter PLS in Chemistry, pages 2006–2020. John Wiley and Sons, Chichester, 1999.

[44] A. Phatak and S. DeJong. The geometry of PLS. *Journal of Chemometrics*, 11:311–338, 1997.

[45] S. Wold, E. Johansson, and M. Cocchi. *3D-QSAR in Drug Design, Theory, Methods and Applicaitons*, chapter PLS, pages 523–550. ESCOM Science, Ledien, 1993.

[46] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K.Ñ. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci,

M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H.Ñakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T.Ñakajima, Y. Honda, O. Kitao, H.Ñakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A.Ñanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004.

[47] http://www.chemistry.mcmaster.ca/aimpac/.

[48] J. Polanski, R. Gieleciak, and A. Bak. The comparative molecular surface analysis (comsa) - a nongrid 3d QSAR method by a coupled neural network and PLS system: Predicting pKa values of benzoic and alkanoic acids. *Journal of Chemical Information and Computer Science*, 42:184–191, 2002.

[49] Poradnik fizykochemiczny. *Physical and Chemical Data Compendium.* WNT: Warsaw, 1974.

[50] http://www.umetrics.com.

[51] A. De Vizcaya-Ruiz, A. Rivero-Müller, L. Ruiz-Ramírez, G.E.Ñ. Kass, L. R. Kelland, R. M. Orr, and M. Dobrota. Induction of apoptosis by a novel copper-based anticancer compound, casiopeina II, in L1210 murine leukaemia and ch1 human ovarian carcinoma cells. *Toxicology in Vitro*, 14:1–5, 2000.

[52] L. Hernandez-Esquivel, A. Martín-Hernandez, N. Pavon, K. Carvajal, and R. Moreno-Sanchez. Cardiotoxicity of copper-based antineoplastic drugs casiopeinas related to inhibition of energy metabolism. *Toxicology and Applied Pharmacology*, 212:79–88, 2006.

[53] A. De Vizcaya-Ruiz, A. Rivero-Müller, L. Ruiz-Ramirez, J. A. Howarth, and M. Dobrota. Hematotoxicity response in rats by the novel copper-anticancer agent: casiopeina II. *Toxicology*, 194:103–113, 2003.

[54] http://aim.tkgristmill.com/.

[55] L Eriksson, E Johansson, N Kettaneh-Wold, J Trygg, C Wikström, and S Wold. *Multi- and Megavariate Data Analysis, Part II, Advanced Applications and Method Extensions*. Umetrics Academy, 2006.

[56] Brian D. Palmer, Jeff B. Smaill, Maruta Boyd, Diane H. Boschelli, Annette M. Doherty, and James M. Hamby. Structure-activity relationships for 1-phenylbenzimidazoles as selective ATP site inhibitors of the platelet-derived growth factor receptor. *Journal of Medical Chemistry*, 41:5457–5465, 1998.

[57] L. Claesson-Welsh. PGDF receptors: structure and mechanism of action. *Cytokines*, 5:31–43, 1993.

[58] W. Meyer-Ingold and W. Eichner. Platelet-derived growth factor. *Cell Biology International*, 19:389–398, 1995.

[59] A. Iwama, M. Sawamura, Y.Ñara, and Y. Yamori. Phosphatidyl-inositol 3-kinase (PI3K) appears to have a crucial role in cellular proliferation induced by platelet-derived growth factor (PDGF). *Clinical and Experimental Pharmacology and Physiology.*, 22:S318–S320, 1995.

[60] M. Maguire, K. R. Sheets, K. McVety, A. P. Spada, and A. A. Zilberstein. A new series of PDGF receptor tyrosine kinase inhibitors: 3-substituted quinoline derivatives. *Journal of Medical Chemistry.*, 37:2129–2137, 1994.

[61] R. E. Dolle, J. A. Dunn, M. Bobko, B. Singh, J. E. Kuster, E. Baizman, A. L. Harris, D. G. Sawutz, D. Miller, S. Wang, C. R. Faltynek, W. Xie, J. Sarup, C. E. Bode, and E. D. Paganiand P. J. Silver. 5,7-dimethoxy-3-(4-pyridinyl)quinoline is a potent and selective inhibitor of human vascular β-type platelet-derived growth factor receptor tyrosine kinase. *Journal of Medical Chemistry.*, 37:2627–2629, 1994.

[62] M. Kovalenko, A. Gazit, A. Bohmer, C. Rorsman, L. Ronnstrand, C. Heldin, J. Waltenberger, F. Bohmer, and A. Levitzki. Selective platelet-derived growth factor receptor kinase blockers reverse sistransformation. *Cancer Research*, 54:6106–6114, 1994.

[63] M. Malkin, W. P. Mason, F. S. Liebermann, and A. L. Hannah. Phase I study of SU101, a novel signal transduction inhibitor in recurrent malignant gliomas. *Proceedings of the American Society for Clinical Oncology.*, 16:385a, 1997.

[64] E. Buchdunger, J. Zimmermann, H. Mett, M. Muller, U. Regenass, and L. B. Lydon. Selective inhibition of the platelet-derived growth factor signal transduction pathway by a protein-tyrosine kinase inhibitor of the 2- phenylaminopyrimidine class. *Proceedings of the National Academy of Sciences of the United States of America*, 92:2558–2562, 1995.

[65] J. Zimmermann, E. Buchdunger, H. Mett, T. Meyer, N. B. Lydon, and P. Traxler. Phenylamino-pyrimidine (PAP) . derivatives: a new class of potent and highly selective PDGF-receptor autophosphorylation inhibitors. *Bioorganic and Medicinal Chemistry Letters.*, 11:1221, 1996.

[66] Brian D. Palmer, A. J. Kraker, B. G. Hartl, A. D. Panopoulos, R. L. Panek, B. L. Batley, G. H. Lu, S. Trump-Kallmeyer, H. D. H. Showalter, and W. A. J. Denny. *Journal of Chemical Medicine*, 42:1373, 1999.

[67] Chongli Zhong, Jingtao He, Chunyu Xue, and Yajun Li. QSAR study on inhibitory activities of 1-phenylbenzimidazoles against the platelet-derived growth factor receptor. *Bioinorganic and Medicinal Chemistry*, 12:4009–4015, 2004.

[68] P. Ducrot, M. Legraverend, and D. Grierson. 3d-QSAR comfa on cyclin-dependent kinase inhibitors. *Journal of Medical Chemistry*, 43:4098–4108, 2000.

[69] L. L. Zhu, T. J. Hou, L. R. Chen, and X. J. Xu. 3d-QSAR analyses of novel tyrosine kinase inhibitors based on pharmacophore alignment. *Journal of Chemical Information and Computer Science*, 41:1032–1040, 2001.

[70] A. Kurup, R. Garg, and C. Hansch. Comparative QSAR study of tyrosine kinase inhibitors. *Chemical Reviews*, 101:2573–2600, 2001.

[71] Fabian López-Vallejo, Josó Luis Medina-Franco, Alicia Hernández-Campos, Sergio Rodríguez-Morales, Lilian Yepez, Roberto Cedillo, and Rafael Castillo. Molecular modeling of some 1H-benzimidazole derivatives with biological activity against entamoeba histolytica. a comparative molecular field analysis study. *Bioinorganic and Medicinal Chemistry*, 15:111–1126, 2007.

[72] Q. Shen, Q. Lu, J. H. Jiang, G. L. Shen, and R. Q. Yu. Quantitative structure.activity relationships (QSAR): studies of inhibitors of tyrosine kinase. *European Journal of Pharmaceutical Sciences.*, 20:63–71, 2003.

[73] C. Zhong, J. He, C. Xue, and Y. A. Li. A QSAR study on inhibitory activities of 1-phenylbenzimidazoles against the platelet-derived growth factor receptor. *Bioorganic Medical Chemistry*, 12:4009–4015, 2004.