# UNIVERSITAT

# ROVIRA I VIRGILI

*Escola Tècnica superior d'Enginyeria Química*

# MODELOS QSPR/QSAR/QSTR BASADOS EN SISTEMAS NEURONALES COGNITIVOS

**Memoria presentada por:**

Gabriela Espinosa Porragas

**Para optar por el grado de Doctor en Enginyeria Química**

**Tarragona, Junio 2002**

# Índice

**ANEXOS**

I. Espinosa G. Yaffe D. Cohen Y. Arenas A. Giralt F. Neural Network Based Quantitative Structural Property Relations (QSPRs) for Predicting Boiling Points of Aliphatic Hydrocarbons. *J. Chem. Inf. Comput. Sci*. **2000**, 40, 859.

II. Espinosa G. Yaffe D. Arenas A. Cohen Y. Giralt F. A Fuzzy ARTMAP based Quantitative Structure-Property Relationships (QSPRs) for Predicting Physical Properties of Organic Compounds, *Ind. Eng. Chem. Res*. **2001**, 40(12), 2757

III. Espinosa G. Arenas A. Giralt F. Prediction of Boiling Points of Organic Compounds from Molecular Descriptors by using Back-propagation Neural Networks, in *Fundamentals of molecular similarity*, Ed. R. Carbo-Dorca, Kluwer: Academic, **2001**, 1

IV. Yaffe D., Cohen Y., Espinosa G., Arenas A., Giralt F., A fuzzy ARTMAP based quantitative structure-property relationships (QSPRs) for predicting Aqueous Solubility of Organic Compounds, *J. Chem. Inf. Compt. Sci*., **2001**, 41, 5, 1150

V. Espinosa G. Arenas A.. Giralt F. A. An Integrated SOM-fuzzy ARTMAP Neural System for the Evaluation of Toxicity, *J. Chem. Inf. Compt. Sci*., **2002**, 42, 2, 343

VI. Yaffe D., Cohen Y., Espinosa G., Arenas A., Giralt F., A fuzzy ARTMAP based quantitative structure-property relationships (QSPRs) for Octanol/Water Partition Coefficients of Organic Compounds, *J. Chem. Inf. Compt. Sc*i., **2002**, 42, 2, 162-183

VII. Espinosa G. Arenas A. Giralt F. Amat L. Girones X. Carbó-Dorca R. Fuzzy ARTMAP-Based QSAR for $TD_{50}$ of Aromatic Compounds. Ed. R. Carbó-Dorca, Kluwer: Academic, *In press*

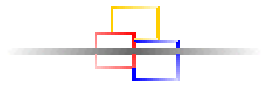
VIII. Giralt F., Espinosa G., Arenas A., Ferre-Gine J., Amat L., Gironés X., Carbó-Dorca R., and Cohen Y. Prediction of Activity Coefficients of Diverse Organic Compounds in Water at Infinite Dilution with an Integrated Som-fuzzy ARTMAP Neural System. AICHE Journal, *Submitted*

IX. Yaffe D., Cohen Y., Espinosa G., Arenas A., Giralt F. A fuzzy ARTMAP based Quantitative Structure-Property Relationships (QSPRs) for the Henry's Law Constant of Organic Compounds. J.Chem. Inf. Compt. Sci. *In press*.

***A mi familia ....***

*Tu casa puede sustituir al mundo; el mundo jamás sustituirá tu casa*
*(Proverbio alemán).*

*Somos lo que hacemos cada día.*
*De modo que la excelencia no es un acto, sino un hábito.*
***Aristóteles***



# *AGRADECIMIENTOS*

**Agradezco a todas aquellas personas que han hecho posible la realización de este trabajo.**

Mi agradecimiento al Dr. Francesc Giralt por aceptar la dirección de esta tesis. Por las horas de discusión, su paciencia y el ánimo que impulsó el desarrollo de cada etapa.

De la misma forma le doy las Gracias al Dr. Alex Arenas por co-dirigir este trabajo siendo además de un puente entre la Ingeniería Química y la Informática un soporte fundamental en la realización de esta tesis.

Quisiera agradecer también a todos los miembros del grupo de Fenómenos de Transporte del Departamento de Ingeniería Química de la URV por el soporte y sus oportunos comentarios.

Mi sincero agradecimiento a los miembros del tribunal por haber aceptado formar parte del mismo.

Mi agradecimiento al Prof. Ramón Carbó y a su grupo de investigación del Instituto de Química Computacional por su ayuda y cálida acogida.

Agradezco la colaboración del Prof. Yoram Cohen por su colaboración en el desarrollo de este trabajo.

Agradezco la colaboración de los integrantes del RIETSE, quienes a lo largo de los años dieron soporte *urgente* a mi medio de trabajo, el ordenador.

Me gustaría recordar a mis compañeros del programa de doctorado de Ingeniería Química por los momentos compartidos.

Al club de la picudo que ha estado presente durante la escritura de la tesis y en las sobremesas nos ha dado de que hablar y nos ha hecho sonreír despertando la imaginación del qué menos.

A la sección femenina de Tarragona que la han hecho un sitio inolvidable: a Vessy, Cata, Montse, Mónica, Susana, Isabela, Magda, Silvia, Eleana.

No podía olvidar a aquellos que no forman parte de la sección anterior y que me han hecho sobre todo reir, , Ulian, Jorge, Orlando, Albert, Mohammad, Josep María, Roger, Frank, Bonet, Carlitos, Thanos, Juan, Ivan, Javier y de todos aquellos que por mala memoria he dejado.

Evidentemente, no puedo olvidar a Alvaro por estar ahí y por ser como es.

Por último quiero agradecer a la Universidad Rovira i Virgili por el soporte económico proporcionado.

***A mis padres y hermano por su apoyo incondicional en los buenos y no tan buenos momentos***

# *SUMMARY*

One of the most attractive applications of computer-aided techniques in molecular modeling stands on the possibility of assessing certain molecular properties before the molecule is synthesized. The field of Quantitative Structure Activity/Property Relationships (QSAR/QSPR) has demonstrated that the biological activity and the physical properties of a set of compounds can be mathematically related to some " simple" molecular structure parameters.

Artificial neural network (ANN) approaches provide an alternative to established predictive algorithms for analyzing massive chemical databases, potentially overcoming obstacles arising from variable selection, multicollinearity, specification of important parameters, and sensitivy to erroneous values. In most instances, ANN's have proven to be better than MLR, PCA or PLS because of their ability to handle non-linear associations.

In the last years there has been a growing interest in the application of neural networks to the development of QSAR/QSPR. The mayor advantage of ANN lies in the fact QSAR/QSPR can be developed without having to a priori specify an analytical form for the correlation model. The NN approach is especially suited for mapping complex non-linear relationships that exists between model output (physicochemical or biological properties) and input model (molecular descriptors). The NN approach could also be used to classify chemicals according to their chemical descriptors and used this information to select the most suitable Índices capable of characterize the set of molecules. Existing neural networks based QSAR/QSPR for estimating properties of chemicals have relied primarily on backpropagation architecture. Backpropagation are an error based learning system in which adaptive weights are dynamically revised so as to minimize estimation errors of target values. However, since chemical compounds can be classified into various structural categories, it is also feasible to use cognitive classifiers such as fuzzy ARTMAP cognitive system, for unsupervised learning of categories, which represent structure and properties simultaneously. This class of neural networks uses a match-based learning, in that it actively searches for recognition categories or hypotheses whose prototype provides an acceptable match to input data.

The current study have been proposed a new QSAR/QSPR fuzzy ARTMAP neural network based models for predicting diverse physical properties such as phase transition temperatures (boiling and melting points) and critical properties (temperature and pressure) and the biological activities (toxicity indicators) of diverse set of compounds. In addition, traditional pre-screening methods to determine de minimum set of inputs parameters have been compared with novel methodology based in self organized maps algorithms.

The most suitable set of molecular descriptor was obtained by choosing a representative from each cluster, in particular the index that presented the highest correlation with the target variable, and additional Índices afterwards in order of decreasing correlation. The selection process ended when a dissimilarity measure between the maps for the different sets of descriptors reached a minimum valued, indicating that the inclusion of more descriptors did not add supplementary information. The optimal subset of descriptors was finally used as input to a fuzzy ARTMAP architecture modified to effect predictive capabilities.

The proposed QSPR/QSAR model predicted physicochemical or biological activities significantly better than backpropagation neural networks or traditional approaches such as group contribution methods when they applied.

**Key words:** QSAR, QSPR, molecular modeling, physicochemical properties, toxicity, carcinogenicity, computational chemistry, molecular descriptors, graph theory, neural networks, self-organizing maps, fuzzy ART, fuzzy ARTMAP, backpropagation, group contribution methods

# *RESUMEN*

Un área sumamente interesante dentro del modelado molecular es el diseño de nuevos compuestos. Los métodos QSPR/QSAR han demostrado que las relaciones entre la estructura molecular y las propiedades físico químicas o la actividad biológica de los compuestos se pueden cuantificar matemáticamente a partir de parámetros estructurales simples.

Las redes neuronales (ANN) constituyen una alternativa para el desarrollo de algoritmos predictivos aplicados en diversos campos. En el análisis masivo de bases de datos, para subsanar los obstáculos derivados de la selección o la multicolinealidad de variables, así como la sensibilidad de los modelos a la presencia de ruido en los datos de entrada al sistema neuronal. En la mayoría de los casos, las redes neuronales han dado mejores resultados que los métodos de regresión multilineal (MLR), el análisis de componentes principales (PCA), o los métodos de mínimos cuadrados parciales (PLS) debido a la no linealidad inherente en los modelos de redes.

En los últimos años el interés por los modelos QSPR/QSAR basados en redes neuronales se ha incrementado. La principal ventaja de los modelos de redes recae en el hecho que un modelo QSAR/QSPR puede desarrollarse sin especificar a priori la forma analítica del modelo. Las redes neuronales son especialmente útiles para establecer las complejas relaciones existentes entre la salida del modelo (propiedades físico químicas o biológicas) y la entrada del modelo (descriptores moleculares). Además, permiten clasificar los compuestos de acuerdo a sus descriptores moleculares y usar esta información para seleccionar el conjunto de índices capaz de caracterizar mejor al conjunto de moléculas. Los modelos QSPR basados en redes utilizan principalmente algoritmos del tipo backpropagation. Backpropagation emplea un sistema de aprendizaje por minimización del error. Sin embargo, ya que los compuestos químicos pueden clasificarse en grupos de acuerdo a su similitud molecular, es factible usar un clasificador cognitivo como fuzzy ARTMAP para crear una representación simultánea de la estructura y de la propiedad objetivo. Este tipo de sistema cognitivo usa un aprendizaje competitivo, en el cual hay una activa búsqueda de la categoría o la hipótesis cuyos prototipos provean una mejor representación de los datos de entrada (estructura química).

En el presente trabajo se propone y se estudia una metodología que integra dos sistemas cognitivos SOM y fuzzy ARTMAP para obtener modelos QSAR/QSPR. Los modelos estiman diferentes propiedades como las temperaturas de transición de fase (temperatura de ebullición, temperatura de fusión) y propiedades críticas (temperatura y presión), así como la actividad biológica de compuestos orgánicos diversos (indicadores de toxicidad). Dentro de este contexto, se comparan la selección de variables realizados

por métodos tradicionales (PCA, o métodos combinatorios) con la realizada usando mapas auto-organizados (SOM).

El conjunto de descriptores moleculares más factible se obtiene escogiendo un representante de cada categoría de índices, en particular aquel índice con la correlación más alta con respecto a la propiedad objetivo. El proceso continúa añadiendo índices en orden decreciente de correlación. Este proceso concluye cuando una medida de disimilitud entre mapas para los diferentes conjuntos de descriptores alcanza un valor mínimo, lo cual indica que el añadir descriptores adicionales no provee información complementaria a la clasificación de los compuestos estudiados. El conjunto de descriptores seleccionados se usa como vector de entrada a la red fuzzy ARTMAP modificada para poder predecir.

Los modelos propuestos QSPR/QSAR para predecir propiedades tanto físico químicas como biológicas predicen mejor que los modelos obtenidos con métodos basados en backpropagation o en métodos de contribución de grupos (en los casos en los que se apliquen dichos métodos).

**Palabras clave:** QSAR, QSPR, modelado molecular propiedades físico químicas, toxicidad, carcinogenicidad, química computacional, descriptores moleculares, teoría de grafos, redes neuronales, mapas auto-organizados, fuzzy ART, fuzzy ARTMAP, backpropagation, métodos de contribución de grupos

# Lista de Tablas

# Lista de Figuras

| | | |
|---|---|---|
| | (A) 80 compuestos; (B) con 80 compuestos y la información adicional de los 49 vectores prototipos. Los descriptores seleccionados son kappa, DI002, HLB, DI008, Sol, $\chi(1)$, HBD, $\Delta$Hf, TX35, DI009, TX33, TX15, TX37, RE, SA, $\rho$, TX39, CX45, TE, DX1007. *el conjunto de compuestos propuesto por Jalali et al., (Jalali, M. et al., 2000) es usado para entrenar y probar el modelo | |
| **Figura 4.37.** | Comparación de los valores experimentales de la actividad inhibitoria del HIV y los valores estimados con fuzzy ARTMAP basados en 20 descriptores moleculares seleccionados con SOM y entrenados con (A) 80 compuestos; (B) con 80 compuestos y la información adicional de los 49 vectores prototipos; (C) MLR, Jalali et al.; (D) Backpropagation 6-6-1, Jalali, M. et al., 2000).*el conjunto de compuestos propuesto por Jalali et al., (Jalali, M. et al., 2000) es usado para entrenar y probar el modelo | 123 |

# Lista de Símbolos

| | |
|---|---|
| $\alpha_o$ | Velocidad de aprendizaje inicial (ecuación 3.9) |
| $\alpha_T$ | Velocidad de aprendizaje final (ecuación 3.9) |
| $\alpha(t)$ | Velocidad de aprendizaje (ecuación 3.7) |
| $\chi(4)$ | Conectividad de valencia de orden 4 |
| $\chi(3)$ | Conectividad de valencia de orden 3 |
| $\chi(2)$ | Conectividad de valencia de orden 2 |
| $\chi(1)$ | Conectividad de valencia de orden 1 |
| $\chi(0)$ | Conectividad de valencia de orden 0 |
| $\boldsymbol{\delta}$ | Delta molecular (números de vecinos en el esqueleto molecular) (ecuación 2.2) |
| $\delta^v$ | Delta molecular de valencia (ecuación 2.3) |
| $\gamma^\infty$ | Coeficiente de actividad a dilución infinita |
| $\eta(t)$ | Función que define la región influencia que el vector de entrada tiene sobre SOM (ecuación 3.7) |
| $\boldsymbol{\eta}$ | Velocidad de aprendizaje (ecuación 3.2) |
| $\kappa$ | Kappa |
| $\mu$ | Momento dipolar |
| $\rho a$ | Parámetro de vigilancia |
| $\rho$ | Densidad |
| $\sigma(t)$ | Radio que define la vecindad de t (ecuación 3.89 |
| $\sigma$ or s | Desviación estándar |
| $\boldsymbol{\sigma}$ | Cantidad de electrones en un orbital sigma (ecuación 2.2) |
| $\boldsymbol{\sigma}$ | Constante de los substituyentes del anillo aromático (ecuación 1.2) |
| $\Delta\varepsilon$ | Barrera energética |
| $\Delta S_m$ | Entropía de transición |
| $\Delta S$ | Delta de entropía (ecuación 1.3) |
| $\Delta H_m$ | Entalpía de transición |
| $\Delta Hf$ | Calor de formación |
| $\Delta H$ | Delta de entalpía (ecuación 1.3) |
| $\boldsymbol{\Pi}$ | Operador hermítico no diferencial |
| $\Delta G^\circ$ | Energía libre de Gibbs (ecuación 1.3) |
| $\Psi(r)$ | Funciones de densidad electrónica |
| A(G) | Matriz de adyacencia molecular |
| AM1 | Austin Model |
| ANN | Redes neuronales artificiales |
| AP | Polarizabilidad promedio |

| | |
|---|---|
| ART | Adaptive Resonance Theory |
| artA | Modulo artA en fuzzy ARTMAP |
| artB | Modulo artB en fuzzy ARTMAP |
| ASA | Atomic Shell Approximation |
| Bal | Índice de Balaban |
| BMU | Best Matching Unit o vector ganador |
| C3Sh | Índice 3D de Schultz |
| CHo1 | Contribución de orden 1 al índice de de Hosoya |
| CHo2 | Contribución de orden 2 |
| CHo3 | Contribución de orden 3 |
| CHo4 | Contribución de orden 4 |
| CHo5 | Contribución de orden 5 |
| CHo6 | Contribución de orden 6 |
| CHo7 | Contribución de orden 7 |
| CHo8 | Contribución de orden 8 |
| CHo9 | Contribución de orden 9 |
| Chos | Índice de Hosoya |
| CNDO | Complete Neglect of Differential Overlap |
| Cou | Índice de autosimilitud cuántica de Coulomb |
| CPMET | Correlated Pair Many Electron Theory |
| Cran | Índice de Randic |
| CSch | Índice de Schultz |
| CX10 | Un índice Chi de: 1 P orden 0 |
| CX11 | Un índice Chi de: 1 P orden 1 |
| CX12 | Un índice Chi de: 1 P orden 2 |
| CX13 | Un índice Chi de: 1 P orden 3 |
| CX14 | Un índice Chi de: 1 P orden 4 |
| CX15 | Un índice Chi de: 1 P orden 5 |
| CX16 | Un índice Chi de: 1 P orden 6 |
| CX17 | Un índice Chi de: 1 P orden 7 |
| CX18 | Un índice Chi de: 1 P orden 8 |
| CX19 | Un índice Chi de: 1 P orden 9 |
| CX23 | Un índice Chi de: 2 C orden 3 |
| CX25 | Un índice Chi de: 2 C orden 5 |
| CX27 | Un índice Chi de: 2 C orden 7 |
| CX29 | Un índice Chi de: 2 C orden 9 |
| CX33 | Un índice Chi de: 3 CH orden 3 |
| CX34 | Un índice Chi de: 3 CH orden 4 |
| CX35 | Un índice Chi de: 3 CH orden 5 |
| CX36 | Un índice Chi de: 3 CH orden 6 |
| CX37 | Un índice Chi de: 3 CH orden 7 |
| CX38 | Un índice Chi de: 3 CH orden 8 |
| CX39 | Un índice Chi de: 3 CH orden 9 |

| | |
|---|---|
| CX44 | Un índice Chi de: 4 PC orden 4 |
| CX45 | Un índice Chi de: 4 PC orden 5 |
| CX46 | Un índice Chi de: 4 PC orden 6 |
| CX47 | Un índice Chi de: 4 PC orden 7 |
| CX48 | Un índice Chi de: 4 PC orden 8 |
| CX49 | Un índice Chi de: 4 PC orden 9 |
| D(G) | Matriz de distancias molecular |
| DI0001 | Índices pseudo-electrotopológicos del átomo 1 |
| DI0002 | Índices pseudo-electrotopológicos del átomo 2 |
| DI0003 | Índices pseudo-electrotopológicos del átomo 3 |
| DI0004 | Índices pseudo-electrotopológicos del átomo 4 |
| DI0005 | Índices pseudo-electrotopológicos del átomo 5 |
| DI0006 | Índices pseudo-electrotopológicos del átomo 6 |
| DI0007 | Índices pseudo-electrotopológicos del átomo 7 |
| DI0008 | Índices pseudo-electrotopológicos del átomo 8 |
| DI0009 | Índices pseudo-electrotopológicos del átomo 9 |
| DI0010 | Índices pseudo-electrotopológicos del átomo 10 |
| DI0011 | Índices pseudo-electrotopológicos del átomo 11 |
| DI0012 | Índices pseudo-electrotopológicos del átomo 12 |
| DI0013 | Índices pseudo-electrotopológicos del átomo 13 |
| DI0014 | Índices pseudo-electrotopológicos del átomo 14 |
| DI0015 | Índices pseudo-electrotopológicos del átomo 15 |
| DI0016 | Índices pseudo-electrotopológicos del átomo 16 |
| DI0017 | Índices pseudo-electrotopológicos del átomo 17 |
| DI0018 | Índices pseudo-electrotopológicos del átomo 18 |
| DI0019 | Índices pseudo electrotopológicos del átomo 19 |
| DI0020 | Índices pseudo-electrotopológicos del átomo 20 |
| $D_{ij}$ | Distancia entre los vértices i, j de en G (ecuación 2.1) |
| DIPPR | Design Institute for Chemical Properties Data |
| E | Caras de una gráfica molecular o tipo de enlace |
| E | Error |
| EE | Energía de intercambio |
| EER | Repulsión electrón-electrón |
| EHT | Extended Hückel Theory |
| ENA | Atracción electrón-núcleo |
| exp | Función exponencial |
| f | Función primitiva evaluada en el cuerpo de la neurona (función de activación) |
| $F^{ab}$ | Módulo Inter-ART, map field |
| Fig. | Figura |
| G | Entidad molecular (gráfica) |
| $g_h$ | El número de pares no ordenados de vértices cuya distancia es h (ecuación 2.1) |

| | |
|---|---|
| GLC | Cromatografía de gas líquido |
| H | Número de átomos hidrógeno |
| HBA | Enlace hidrógeno aceptor |
| HBD | Enlace hidrógeno donador |
| HIV-1 | Virus de inmuno deficiencia adquirida |
| HLB | Balance hidrofóbico-lipofíbico |
| HOMO | Highest Occupied Molecular Orbital |
| i(x) | El nodo que mejor represente al patrón x |
| $I^D$ | Índice de la potencia inhibitoria de un compuesto |
| INDO | Intermediate Neglect of Differential Overlap |
| K | Contante de equilibrio (ecuación 1.3) |
| k | Constante de Boltzmann |
| K | Grado Kelvin |
| Kin | Índice de autosimilitud cuántica cinético |
| $LC_{50}$ | Índice de concentración letal media |
| $LD_{50}$ | Índice de dosis letal media |
| LSER | Linear Solvation Energy Relationship |
| LUMO | Unoccupied Molecular Orbital |
| LVQ | Learning Vector Quantization |
| MC SCF | Configurational Method Self Consistent Field |
| Md | Profundidad molecular |
| MEE | Métodos de Estructura Electrónica |
| mg/Kg | Miligramos por Kilogramos |
| Ml | Longitud molecular |
| MLR | Análisis de regresión multilineal |
| MM | Mecánica Molecular |
| MO | Orbitales Moleculares |
| MP2 | Moller-Plesset Theory |
| MPa | Mega Pascales |
| MQS | Medidas de similitud cuántica |
| MQSM | Matrices de similitud cuántica |
| MR | Parámetro de los substituyentes |
| Mv | Volumen molecular |
| Mw | Peso molecular |
| Mwd | Anchura molecular |
| ${}^m\chi^t$ | Índice de conectividad molecular (de orden m y tipo de fragmento caracterizado (cadenas, clusters) |
| N | Número de entradas a una red neuronal |
| N | Suma de números atómicos |
| N x N | Dimensión de la matriz (número de vértices en una matriz simétrica) |
| NFL | Número de niveles ocupados |
| Nj(t) | Función de vecindad (ecuación 3.6) |

| | |
|---|---|
| NNR | Repulsión núcleo-núcleo |
| $O_{pi}$ | Salida actual de la neurona i correspondiente al patrón p |
| Ove | Índice de autosimilitud cuántica de Overlap |
| Pc | Presión crítica |
| PCA | Análisis de componentes principales |
| PHS | Porcentaje de superficie hidrofílica |
| PI | Potencial de Ionización |
| PLS | Mínimos cuadrados parciales |
| PM3 | Parametric Model |
| PR | Parachor |
| $Q_k$ | Grupos de superficie |
| QSAR | Quantitative Structure Activity Relationships |
| QSPR | Quantitative Structure Property Relationships |
| QSTR | Quantitative Structure Toxicity Relationships |
| R | Constante de los gases |
| r | Posición de las neuronas sobre el mapa |
| $r^2$ | Coeficiente de correlación |
| Ran | Índice de Randic |
| RE | Energía de resonancia |
| $R_k$ | Grupos de volumen |
| RT | Reverse Transcriptasa |
| S | Correlación (ecuación 3.3) |
| S3Sh | Índice 3D de Schultz |
| SA | Superficie molecular |
| SAR | Relaciones estructura actividad |
| SCF | Self Consistent Field |
| SHo1 | Contribución de orden 1 al índice de de Hosoya |
| SHo2 | Contribución de orden 2 |
| SHo3 | Contribución de orden 3 |
| SHo4 | Contribución de orden 4 |
| SHo5 | Contribución de orden 5 |
| SHo6 | Contribución de orden 6 |
| SHo7 | Contribución de orden 7 |
| SHo8 | Contribución de orden 8 |
| SHo9 | Contribución de orden 9 |
| SHos | Índice de Hosoya |
| Sol | Parámetro de solubilidad |
| SOM | Mapas auto-organizados |
| SPR | Relaciones estructura propiedad |
| SRan | Índice de Randic |
| SSch | Índice de Schultz |
| SX10 | Un índice Chi de: 1 P  orden 0 |
| SX11 | Un índice Chi de: 1 P  orden 1 |

| | |
|---|---|
| SX12 | Un índice Chi de: 1 P orden 2 |
| SX13 | Un índice Chi de: 1 P orden 3 |
| SX14 | Un índice Chi de: 1 P orden 4 |
| SX15 | Un índice Chi de: 1 P orden 5 |
| SX16 | Un índice Chi de: 1 P orden 6 |
| SX17 | Un índice Chi de: 1 P orden 7 |
| SX18 | Un índice Chi de: 1 P orden 8 |
| SX19 | Un índice Chi de: 1 P orden 9 |
| SX23 | Un índice Chi de: 2 C orden 3 |
| SX25 | Un índice Chi de: 2 C orden 5 |
| SX27 | Un índice Chi de: 2 C orden 7 |
| SX29 | Un índice Chi de: 2 C orden 9 |
| SX33 | Un índice Chi de: 3 CH orden 3 |
| SX34 | Un índice Chi de: 3 CH orden 4 |
| SX35 | Un índice Chi de: 3 CH orden 5 |
| SX36 | Un índice Chi de: 3 CH orden 6 |
| SX37 | Un índice Chi de: 3 CH orden 7 |
| SX38 | Un índice Chi de: 3 CH orden 8 |
| SX39 | Un índice Chi de: 3 CH orden 9 |
| SX44 | Un índice Chi de: 4 PC orden 4 |
| SX45 | Un índice Chi de: 4 PC orden 5 |
| SX46 | Un índice Chi de: 4 PC orden 6 |
| SX47 | Un índice Chi de: 4 PC orden 7 |
| SX48 | Un índice Chi de: 4 PC orden 8 |
| SX49 | Un índice Chi de: 4 PC orden 9 |
| T | Número de etapas o epochs de entrenamiento (ecuación 3.9) |
| T | Temperatura |
| T3Ba | Índice 3D de Balaban |
| T3DHar | Número de 3D Harary |
| T3Sh | Índice 3D de Schultz |
| Tb | Temperatura de ebulllición |
| TBal | índice de Balaban |
| Tc | Temperatura crítica |
| $TD_s$ | Índices de dosis tóxicas (s representa el umbral de toxicidad (10, 50, 90)) |
| te | conjunto de prueba |
| TE | Energía total |
| TE | Energía total |
| THar | Número de Harary |
| THo1 | Contribución de orden 1 a el índice de Hosoya |
| THo2 | Contribución de orden 2 |
| THo3 | Contribución de orden 3 |
| THo4 | Contribución de orden 4 |

| | |
|---|---|
| THo5 | Contribución de orden 5 |
| THo6 | Contribución de orden 6 |
| THo7 | Contribución de orden 7 |
| THo8 | Contribución de orden 8 |
| THo9 | Contribución de orden 9 |
| THos | Índice de Hosoya |
| Tm | Temperatura de fusión |
| $t_{pi}$ | Valor real de salida de la neurona i correspondiente al patrón p |
| TPM | Topology Preserving Map |
| tr | conjunto de entrenamiento |
| TRan | Índice de Randic |
| TSch | Índice de Schultz |
| TW3D | Índice 3D de Wiener |
| Twiener | índice de de Wiener |
| TWph | Número de caminos topológicos de Wiener |
| TX10 | Un índice Chi de: 1 P orden 0 |
| TX12 | Un índice Chi de: 1 P orden 2 |
| TX13 | Un índice Chi de: 1 P orden 3 |
| TX14 | Un índice Chi de: 1 P orden 4 |
| TX15 | Un índice Chi de: 1 P orden 5 |
| TX16 | Un índice Chi de: 1 P orden 6 |
| TX17 | Un índice Chi de: 1 P orden 7 |
| TX18 | Un índice Chi de: 1 P orden 8 |
| TX19 | Un índice Chi de: 1 P orden 9 |
| TX23 | Un índice Chi de: 2 C orden 3 |
| TX25 | Un índice Chi de: 2 C orden 5 |
| TX27 | Un índice Chi de: 2 C orden 7 |
| TX29 | Un índice Chi de: 2 C orden 9 |
| TX33 | Un índice Chi de: 3 CH orden 3 |
| TX34 | Un índice Chi de: 3 CH orden 4 |
| TX35 | Un índice Chi de: 3 CH orden 5 |
| TX36 | Un índice Chi de: 3 CH orden 6 |
| TX37 | Un índice Chi de: 3 CH orden 7 |
| TX38 | Un índice Chi de: 3 CH orden 8 |
| TX39 | Un índice Chi de: 3 CH orden 9 |
| TX44 | Un índice Chi de: 4 PC orden 4 |
| TX45 | Un índice Chi de: 4 PC orden 5 |
| TX46 | Un índice Chi de: 4 PC orden 6 |
| TX47 | Un índice Chi de: 4 PC orden 7 |
| TX48 | Un índice Chi de: 4 PC orden 8 |
| TX49 | Un índice Chi de: 4 PC orden 9 |
| V | Vértice de una gráfica molecular |
| $V_j$ | Unidades intermedias en la red (unidades |

| | |
|---|---|
| | ocultas) |
| W | Índice de Wiener (ecuación 2.1) |
| $w_i$ | Sinapsis o pesos de entre la neurona i a la neurona de la capa siguiente |
| $x_i$ | Valor real |
| $Z^v$ | Número de electrones de valencia |
| $^o$ C | Grado centígrado |
| $^{3D}W$ | Índice de Wiener 3D |
| 3D | Tridimensional |
| 2D | Bidimensional |
| \|\| \|\| | Norma euclídea del argumento |

# I. INTRODUCCIÓN

## 1.1. Motivación

En el diseño de procesos tanto a escala industrial como de laboratorio se requieren datos termodinámicos y de las propiedades físicas de los compuestos químicos que intervienen en los mismos. El acceso a parámetros básicos de temperaturas de transición de fase o de las propiedades críticas de los compuestos, resulta crucial en procesos petroquímicos, de destilación y de tratamiento de desechos, entre otros. Las bases de datos existentes (DIPPR, Belstein, Detherm) constituyen la fuente de información primaria de dichos datos. Sin embargo, en los casos relacionados con la actividad tóxica o inhibitoria de diversos fármacos o con la predicción del impacto medioambiental, la información existente es escasa o en muchos casos inexistentes. Las razones pueden ser diversas, partiendo de las dificultades implícitas en el método experimental utilizado para obtener los parámetros cuantitativos, o la inviabilidad del método, por que requiera una considerable inversión de tiempo y/o dinero. Es por ello, que la búsqueda de métodos alternativos para estimar propiedades físicas, químicas y biológicas es un campo ampliamente explorado. Uno de los métodos utilizados recientemente con este fin, son los métodos basados en la obtención de las relaciones estructura-propiedad, QSPR, estructura-actividad, QSAR, o estructura-toxicidad, QSTR (Balaban A. et al., 1984; Hall L. et al., 1984; Egolf L. et al., 1993; Katritzky R. et al., a) 1996; b) 1998); Basak S. et al., 1997; Bünz P. et al., 1998; Basak S. et al., 1999; Espinosa G. et al., a) 2000; b) 2001.

## 1.2. Hipótesis

Si partimos de la hipótesis que las propiedades macroscópicas de una sustancia están directamente relacionadas con su estructura molecular, la cual determina tanto la magnitud como el tipo de fuerza intermolecular predominante, cualquier tipo de modelo para predecir las propiedades de una sustancia requiere la caracterización del comportamiento molecular.

La similitud entre las estructuras de diversos compuestos y sus propiedades sugiere que una propiedad macroscópica puede ser calculada a partir de la contribución de los grupos que forman a la molécula (Lee M. et. al, 1993). Las relaciones estructura-propiedad y estructura-actividad (QSPR, QSAR respectivamente) constituyen una alternativa viable para calcular las propiedades físicas, químicas y biológicas de los compuestos. Es claro, que por muchas razones, incluyendo las limitaciones físicas de la tecnología en los ordenadores y en ausencia de una base teórica única, llevar a cabo cálculos cuantitativos de las propiedades de los compuestos a

partir de primeros principios no será posible en un futuro próximo. Como consecuencia, el estudio de métodos alternativos es un campo ampliamente explorado y en el se han desarrollado diversos modelos termodinámicos y correlaciones que pueden ser usados conociendo el rango de aplicabilidad de los mismos (muchos de ellos se encuentran incorporados a diversos simuladores comerciales).

## 1.3. Antecedentes

La similitud entre las estructuras de diversos compuestos y sus propiedades sugiere que una propiedad macroscópica puede ser calculada a partir de la contribución de los grupos que forman parte de la estructura molécular (Fredenslund A. et al., 1975; Kier L. et al., 1976; Joback K. et al.,1987; Lee M. et al., 1993). Cada vez que un científico comienza a cuantificar las propiedades físicas o biológicas del medio se buscan patrones o relaciones entre las medidas realizadas Sin embargo, no fue hasta 1933 cuando Louis Hammet observa que la velocidad de la aminólisis en ésteres aromáticos es directamente proporcional a la constante de ionización de los ácidos carboxílicos. Las velocidades de reacción y las constantes de equilibrio están relacionadas con los cambios de energía libre (lo que se conoce como relaciones lineales de energía libre, LFER). La ecuación derivada por Hammet es para una reacción,

$$\log k = rho \log K + c; \log(k_x / k_H) = rho \log(K_x / K_H) \qquad (1.1)$$

donde la constante de los substituyentes del anillo aromático denominada sigma, ($\sigma$) es igual a

$$\sigma = \log(K_x / K_H) \qquad (1.2)$$

derivada a partir del estado de referencia de la reacción (ionización del ácido en agua), puede ser usada para extrapolar el efecto del substituyente en otros sistemas. Lo que en realidad hizo, fue proveer un trabajo pionero que mostraba la utilidad de procedimientos paramétricos en la descripción de propiedades empíricas (constantes de equilibrio, velocidades de reacción) en términos de un parámetro que describe la estructura molecular (sigma, también la pendiente, rho). La relación propuesta por Hammet realmente proveía una base termodinámica para todo el trabajo posterior de QSAR,

$$\Delta G^o = -RT \ln K = \Delta H^o - T\Delta S^o \qquad (1.3)$$

al asumir que la constante de equilibrio es una función de la estructura de la molécula, la cual afecta a $\Delta G^o$. Los descriptores desarrollados por Hammet

y posteriormente por Taft, asumen las relaciones anteriores, y de hecho las verifican.

Sorprendentemente aunque se realizaron diversos estudios relacionados con el potencial de una droga o su toxicidad, pocos intentos hubo por conectar la actividad biológica con las propiedades físicas de los compuestos. No fue hasta 1950 cuando Hansch desarrollo un parámetro hidrofóbico y empleó métodos de regresión para correlacionar la actividad biológica con las propiedades moleculares. Desde entonces se han usado diversos métodos estadísticos y se han promovido otras técnicas de reconocimiento de patrones como el análisis de categorías o grupos, el análisis factorial y el análisis de componentes principales en la búsqueda de relaciones explícitas entre la actividad y las propiedades moleculares. La ecuación de Hasch se define, (Hansch et al., 1964)

$$\log(1/C) = a(\log P)^2 + b(\log P) + c\sigma + dE_s + e + \ldots \qquad (1.4)$$

donde a-e son constantes, determinadas para una reacción particular mediante un análisis de regresión. Log P, $\pi$, $\sigma$, F, R, $E_s$, etc, son variables independientes cuyos valores se obtienen directamente del experimento.

Las técnicas de QSAR/QSPR asumen que existe una correlación implícita entre las propiedades y la estructura molecular y tratan de establecer relaciones matemáticas simples para describir y luego extrapolar una o varias de esas propiedades a un conjunto de compuestos que usualmente pertenecen a una misma familia (Randic M., 1984). Dichas relaciones pueden determinarse a través de métodos matemáticos, ya sean, regresiones multilineales o métodos no lineales. Los métodos QSAR con métodos de regresión múltiple fueron aplicados exitosamente en diversos problemas de diseño de nuevos fármacos (Kier L. et al., 1976; Wagener M. et al., 1995; Anzali S. et al., 1998; Warne M., et al., 1999).

El campo de los métodos QSAR/QSPR's se encuentra bien establecido. Se ha demostrado que la actividad biológica de un conjunto de compuestos que actúan vía el mismo mecanismo de acción puede ser modelado matemáticamente a partir de parámetros estructurales simples. De ahí que en la última década, los modelos QSAR/QSPR hayan encontrado un auge como una alternativa viable para estimar un sin fin de propiedades químicas y de actividades biológicas. Los métodos QSAR/QSPR tienen algunas ventajas sobre métodos tradicionales como puede ser su rapidez y simplicidad. Sin embargo, la metodología clásica de este tipo de métodos tiene limitaciones, como la imposibilidad de distinguir estéreo isómeros, o la imposibilidad de encontrar modelos lineales cuando la base estructural tiene una variación considerable (conjuntos heterogéneos).

Esencialmente, los modelos QSAR/QSPR requieren de: i) la definición de descriptores moleculares capaces de caracterizar satisfactoriamente diferentes conjuntos de compuestos químicos y de ii) modelos matemáticos capaces de generar un modelo predictivo, por ejemplo, el análisis de regresión multilineal (MLR), mínimos cuadrados parciales (PLS) o las redes neuronales (ANN). Tradicionalmente, los modelos QSPR/QSAR seleccionan *a priori* la forma del modelo (lineal, polinomial o log-lineal). De esta manera cuantifican la correlación entre los índices moleculares seleccionados y la propiedad que se desea modelar, seguido de un análisis de regresión en el que se determinan los parámetros del modelo (Kier L. et al., 1985; Balaban A. et al., 1994; Pogliani P., 1995; Ivanciuc O. et al., 1996; Basak S. et al., 1996; Katritzky A. et al.,1996; Ivanciuc O., 1998). En términos generales podemos distinguir dos tipos de métodos, i) los bidimensionales (2D), en los cuales características topológicas son usadas como descriptores moleculares y ii) los tridimensionales (3D) en los cuales se calculan propiedades moleculares en un conjunto de puntos alrededor de las moléculas estudiadas, ej. Índices de similitud cuántica, (Carbó-Dorca R. et al., 1980; Carbó-Dorca R. et al., 1988; Amat L. et al., a) 1997; b)1999).

Las moléculas pueden representarse como una colección de entidades individuales ensambladas, en la cual, cada una de las partes que la constituyen, por ejemplo, átomos, están conectados para formar a la especie química. La forma más básica de ensamble de una entidad puede representarse a través de una relación binaria, definida por una gráfica G=(V,E), donde V representa el conjunto de vértices o átomos de la estructura molecular y E simboliza a las caras, o el tipo de enlace (usualmente covalente) que integran a la entidad. Realmente ha sido durante la segunda mitad del siglo XX, que los matemáticos y químicos teóricos han propuesto diversos métodos para caracterizar la topología de las gráficas moleculares (Wiener H. 1947; Randic M. 1974; Kier L. and Hall L. 1986; 1990; Basak S.et al., 1999). La caracterización matemática de las gráficas moleculares se logra a través de una gráfica invariante (Trinajsti• N. 1992). Una gráfica invariante es una propiedad teórica de una gráfica o un parámetro el cual es siempre idéntico, o tiene el mismo valor para gráficas isomórficas. Una gráfica invariante puede ser un polinomio, una secuencia de números o un simple índice numérico, este último se conoce como índice topológico.

Los índices de conectividad molecular son un ejemplo de índice topológico, son índices simples y uno de los más utilizados en diversos modelos QSAR/QSPR desde hace algunas décadas (Randic M., a) 1975; b) 1984; Randic M., 1984; Kier L. et al., 1985; Pogliani P., 1995; Balaban A. et al., 1994). Por ejemplo, el índice de conectividad molecular de primer orden fue utilizado en 1982 para correlacionar la solubilidad de hidrocarburos en agua (Medir M. et al., 1982). Este tipo de índices describe las similitudes y diferencias entre las moléculas, toman en cuenta

características como el tamaño, la ramificación, la instauración, la presencia de heteroátomos y la ciclicidad

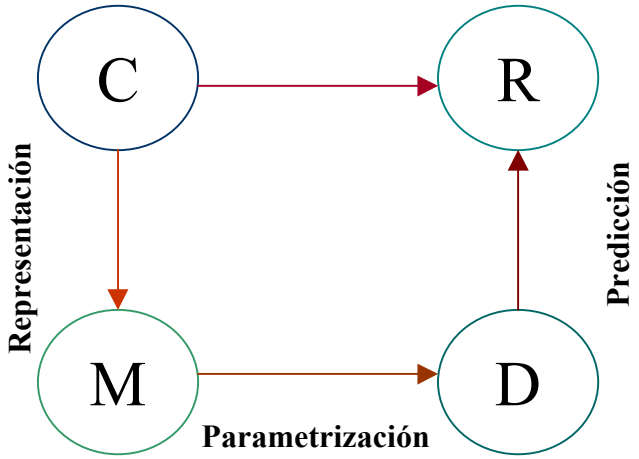



**Fig.1.1.** Representación del proceso para estimar teóricamente propiedades a partir de relaciones estructura propiedad (SPR). La C representa el conjunto de compuestos, la R el conjunto de números reales, la M el objeto y el conjunto de descriptores moleculares queda representado en la letra D (Basak S. et al., 1996)

El elemento clave de tales estudios es encontrar explícitamente la relación estructura-propiedad. Esto involucra la representación de la estructura de los compuestos, usando diversos descriptores (o índices) moleculares y el método utilizado para la para la construcción del modelo. (Jurs P. et al.,1997). Estos métodos no son algo nuevo, han sido ampliamente utilizados por las principales compañías farmacéuticas para esclarecer el efecto de la estructura sobre diversas propiedades biológicas. Sin embargo, la aplicación de las relaciones estructura-propiedad para predecir propiedades físicas y químicas ha sido poco explorada (Katrizky A. et. Al., 1997).

Uno de los principales problemas de las aproximaciones QSPR/QSAR lo constituye la selección de los descriptores moleculares necesarios para obtener un buen modelo. Diversas técnicas estadísticas de preselección se utilizan para seleccionar las variables de entrada de los modelos QSAR/QSPR (Stanton D. et al., a) 1991; b) 1992; Katritzky A. et al., 2000; Gute, B. et al., 1997).

Se propone una nueva metodología capaz de seleccionar de un conjunto de descriptores moleculares aquellos que mejor caractericen la relación estructura-propiedad o estructura-actividad de un conjunto diverso de compuestos químicos y cuantificar esa relación en un modelo extrapolable a compuestos ajenos a los usados para crear el modelo. Dicha metodología integra dos sistemas cognitivos: mapas autoorganizados (SOM) para la selección de los índices moleculares y un clasificador cognitivo, fuzzy ARTMAP (Carpenter G. et al., 1988; Carpenter G. et al., 1991; Carpenter G. et al., 1992) para la construcción explícita del modelo. Dicho sistema se ha aplicado a diversas propiedades físico químicas como las temperaturas de transición de fases o propiedades críticas de compuestos diversos, (Espinosa et al., a) 2000; b) 2001). Los modelos fuzzy ARTMAP-QSAR/QSPR han

demostrado ser robustos y tener una capacidad de extrapolación mayor comparada con modelos similares obtenidos a partir redes backpropagation o algunos modelos generados por métodos de contribución de grupos.

## 1.4. Objetivos y estructura de la tesis

El objetivo general de la presente tesis es el de predecir propiedades físicas, químicas y biológicas de compuestos orgánicos a partir de las relaciones estructura propiedad (QSPR) o relaciones estructura actividad (QSAR). Este tipo de técnica nos lleva a subdividir los objetivos en dos grandes bloques:

i) *Selección de variables.* Las variables seleccionadas son claves para obtener un buen modelo predictivo. Con lo cual, nos encontramos ante una pregunta fundamental, ¿Cuáles son las variables adecuadas para modelar un problema?, ¿Cómo podemos determinar estás variables?. La gran cantidad de descriptores existentes en la literatura hace imposible calcularlos todos y evaluar su desempeño en cada caso, lo cual lleva a utilizar métodos de preselección o realizar una reducción dimensional de las variables (descriptores moleculares) de entrada. Como cualquier modelo, el primer paso será utilizar índices simples o aquellos usados en trabajos previamente publicados (ej. Índices de conectividad molecular). Posteriormente, cuando la relación estructura propiedad no pueda ser evidenciada por este tipo de índices, se tomarán en cuenta aquellos que clarifiquen aspectos más específicos como la distribución electrónica, o la polarizabilidad de la molécula. Una vez determinadas las variables de entrada (diferentes tipos de índices) encontramos el segundo objetivo:

ii) *La modelización*, es decir, seleccionar el método matemático capaz de capturar la relación estructura propiedad o estructura actividad (métodos de regresión multilineal, o redes neuronales). Se estudian diferentes métodos y se realiza una comparación cuantitativa de la capacidad predictiva de los modelos QSPR-QSAR con redes neuronales (backpropagation, fuzzy ARTMAP) con respecto a los métodos tradicionales (contribución de grupos o algún método QSAR regresión multilineal).

En los siguientes capítulos se presentan las ideas básicas de las técnicas aplicadas en el presente trabajo, así como algunos ejemplos prácticos de los modelos obtenidos. En el capítulo 2 se describen brevemente los tipos de descriptores moleculares usados a lo largo de la memoria, tanto topológicos como cuánticos. Además de las diferentes propiedades planteadas como objetivo de estudio (fisicoquímicas y biológicas).

En el capítulo 3 se describen diferentes algoritmos de redes neuronales, sus ventajas e inconvenientes frente a los métodos más

tradicionales como, el análisis de regresión multilineal, MLR. Así mismo, se realizará una comparación entre la capacidad de diferentes paradigmas de redes neuronales dentro del campo de las relaciones estructura propiedad o actividad.

En el capítulo 4, se lleva a cabo la discusión de los resultados expuestos. Dicha discusión evolucionará de la misma manera que lo fue haciendo nuestra metodología a lo largo de los 4 años de tesis. Bajo la concepción de la técnica de QSPR usando redes neuronales, se presentan los primeros ejemplos con propiedades físicas, punto de ebullición y temperatura y presión críticas comparando los modelos obtenidos con backpropagation y fuzzy ARTMAP con los métodos tradicionales de contribución de grupos. Una vez determinado el mejor sistema cognitivo, capaz de extraer la naturaleza no lineal de las relaciones estructura propiedad, se refina la metodología integrando un sistema que nos ayuda en la selección de variables. Haciendo uso de la metodología completa se expondrán algunas aplicaciones, como la estimación del punto de fusión, coeficientes de actividad a dilución infinita y diferentes índices de toxicidad. Finalmente, un poco fuera del marco de propiedades fisicoquímicas o toxicidades se expondrá un caso particular, el modelado de la capacidad inhibitoria de un conjunto de la familia de los HEPT frente al virus del SIDA (HIV-1). Este último ejemplo, nos sirve como punto de partida para plantear una línea de trabajo muy prometedora dentro del ámbito de la química médica, y sin lugar a duda con una gran repercusión dentro del desarrollo de nuevos fármacos.

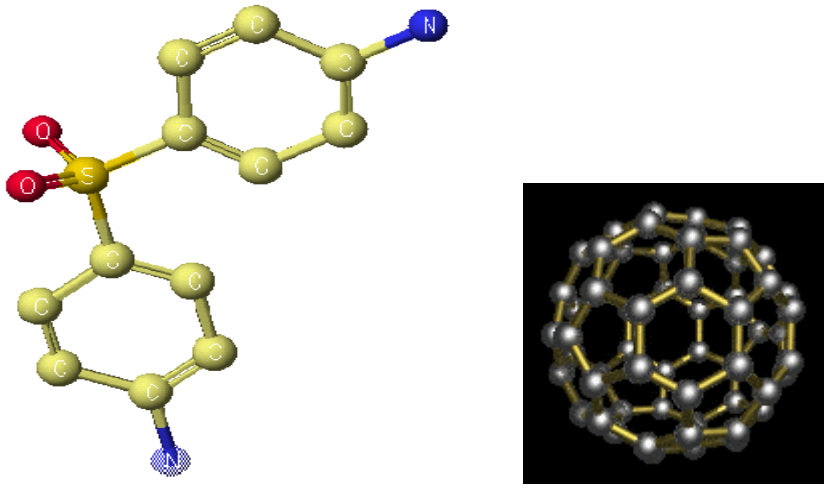
En el capítulo 5 se exponen las conclusiones derivadas de la presente investigación y junto con el capítulo 6 en el cual se dan las recomendaciones para trabajos futuros, engloban la síntesis del presente trabajo. La información complementaria a esta memoria se encontrará en las referencias citadas en el índice de la misma.

# 2. PROPIEDADES Y MÉTODOS

## 2.1. Métodos

### 2.1.1 Índices Topológicos y Cuánticos

En los últimos años se han propuesto un sin número de descriptores moleculares, una de las clasificaciones más aceptadas hace referencia a tres tipos de índices: (i) *índices topológicos*. Los cuales dan idea de la adyacencia y distancia de los átomos dentro de la estructura molecular, por ejemplo, el índice de Wiener (Wiener H., et al.,1947), los índices de conectividad molecular (Randic M., a) 1975; b)1984; Kier L. et al., 1976; Hall L. et al., 1991), los índices de conectividad de enlace definidos por Basak y Margunson (Basak S. et al.,1987), entre otros; (ii*) índices geométricos o índices de forma* los cuales guardan información acerca de las características espaciales de los átomos en la molécula, por ejemplo, el volumen de Van der Waals, el índice 3-D de Wiener; (iii) *índices cuánticos* expresan en principio, todas las propiedades tanto electrónicas como geométricas de las moléculas y sus interacciones, por ejemplo, polarizabilidad, los índices de similitud cuántica (Carbó-Dorca R. et al., a) 1988; b) 1998).



a) b)

**Fig. 2.1.** Ejemplos de gráficas moleculares, a) Dapsone. b) $C_{60}$

A continuación daremos una breve reseña de los índices más representativos en la mayoría de los modelos. Un resumen de los índices y algunas de sus referencias se encuentran en la Tabla 2.1.

La mayoría de los índices topológicos pueden obtenerse a partir de matriz de adyacencia A(G) y la matriz de distancias D(G) de una gráfica molecular, Fig. 2.1. Estas matrices se obtienen etiquetando los esqueletos moleculares sin considerar a los átomos de hidrógeno. Es decir, para una

gráfica G con un conjunto de vértices $\{\nu_1, \nu_2, ....,\nu_n\}$, A(G) se define como la matriz de n x n ($a_{ij}$), donde $a_{ij}$ tomará únicamente dos valores:

$a_{ij}$ = 1, si los vértices $\nu_i$ y $\nu_j$ son adyacentes en G

$a_{ij}$ = 0, de otra manera

La matriz de distancia, D(G) de una gráfica molecular (*G)* con *n* vértices es una matriz simétrica ($d_{ij}$) n x n, donde $d_{ij}$ es igual a la distancia entre los vértices $\nu_i$, y $\nu_j$ en G. Cada elemento de la diagonal $d_{ij}$ de D(G) es igual a cero.

El Prof. Harry Wiener fue el primero en crear un índice estructural (índice topológico) para estimar propiedades de moléculas a partir de su estructura (Wiener H., 1947). Este índice se conoce popularmente como índice de Wiener (W). Se calcula a partir de la matriz de distancias D(G) de una gráfica G (sin tomar en cuenta los enlaces hidrógeno) como la suma de los elementos de la matriz triangular superior.

$$W = \frac{1}{2}\sum_{ij} d_{ij} = \sum_{h} h\ g_h \qquad (2.1)$$

donde $g_h$ es el número de pares no ordenados de vértices cuya distancia es h.

Los índices de conectividad molecular, $\chi$, son los descriptores moleculares más populares en estudios estructura-propiedad (QSPR) y/o estructura-actividad (QSAR). Cuantifican la estructura molecular al incorporar fragmentos de sus sub-estructuras en índices numéricos. Dentro de cada índice se cuantifican características como tamaño, grado de insaturación, presencia de heteroátomos y la ciclicidad de la molécula El cálculo comienza con la reducción de la molécula a una estructura carente de hidrógenos. A cada átomo se le asignan dos descriptores atómicos basados en la cantidad de electrones sigma o de electrones de valencia. El primer descriptor, $\delta$, se define mediante la siguiente ecuación

$$\delta = \sigma - h \qquad (2.2)$$

donde, $\sigma$ es la cantidad de electrones en un orbital sigma y h es el número de átomos de hidrógeno. El valor de $\delta$ para un átomo es igual al número de átomos vecinos en el esqueleto molecular. Estos valores son la base para el cálculo de los índices de conectividad molecular. El valor correspondiente al número de electrones de valencia, $\delta^v$ se define como:

$$\delta^v = Z^v - h \qquad (2.3)$$

donde $Z^v$ representa el número de electones de valencia.

Los índices de conectividad molecular se representan por ${}^{m}_{t}\chi$. Donde m representa el orden del índice y t el tipo de fragmento caracterizado (cadenas, cluster). Las sub-estructuras pertenecientes al esqueleto molecular están definidas por la descomposición de la molécula en diferentes fragmentos, Fig. 2.2, como átomos (orden cero, m=0), enlaces simples (primer orden, m=1), fragmentos con dos enlaces contiguos (orden dos, m=2), fragmentos con tres enlaces contiguos (orden tres, m=3, t=P) y otros fragmentos incluyendo grupos de predefinidos de átomos, y sus cadenas. Para cada orden y tipo de fragmento es posible calcular un índice de conectividad molecular.

$$ {}^{m}C_i = \prod_{k=1}^{m+1} (\delta)^{-0.5} \qquad (2.4) $$

$$ {}^{m}_{t}\chi = \sum_{i=1}^{Ns} {}^{m}C_i \quad (2.5) $$

Los índices conectividad molecular de valencia se calculan de la misma forma, únicamente se sustituyen las $\delta$'s de la ecuación 2.2 por las correspondientes $\delta^{v}$ definida mediante la ecuación 2.3. El índice de conectividad de primer orden ${}^{1}\chi^{v}$ tiene relación con el grado de ramificación del compuesto y el tamaño de la molécula expresada en términos del número de átomos diferentes al hidrógeno. El índice de segundo orden representa la disección del esqueleto molecular en fragmentos con dos enlaces contiguos.

La diversidad de gráficas moleculares es ilimitada, lo que justifica el gran número de descriptores moleculares. Por lo tanto, el arte en el modelado de estructuras moleculares está en el diseño, la construcción o detección de los descriptores moleculares que nos proporcionen una fuerte correlación con la propiedad seleccionada. Aunque no existe una relación explícita causa-efecto, es posible identificar los componentes dominantes de las estructuras y de esta manera generalizar los modelos a moléculas con estructuras similares. Una de las tendencias ha sido la búsqueda de índices altamente discriminatorios. Esta clase de estudios se ha enfocado en el buscar racionalizar el papel que juegan descriptores en la caracterización de una propiedad específica. A pesar de que tales trabajos son legítimos, están limitados, no podemos negar la importancia de buscar el mejor descriptor para correlacionar diversas propiedades moleculares, tal descriptor puede considerarse dominante, pero tomando en cuenta el alto grado de correlación entre muchos de los descriptores es frecuente encontrar varios descriptores dominantes de calidad comparable.

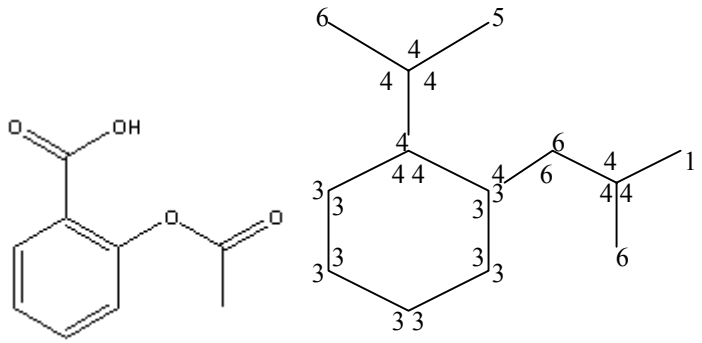



**Fig. 2.2.** Esqueleto molecular de la aspirina y contribución de cada sub-fragmento molecular

En el la primera parte del presente trabajo, se utilizan los índices de conectividad molecular y de conectividad de valencia de manera sistemática para caracterizar las diversas propiedades estructurales y electrónicas de cada molécula y correlacionarlas con las diversas propiedades fisicoquímicas de diversos compuestos orgánicos. Dichos índices se aplican como un conjunto ordenado, es decir, se utilizan los mismos descriptores en el mismo orden en cada prueba realizada. Lo cual restringe la manera de usar los índices de conectividad (Randic M., 1984). Lo cual nos permitirá evaluar la capacidad y la aplicabilidad de los índices de conectividad como conjunto base en estudios estructura-propiedad.

Las propiedades fisocoquímicas de una molécula y aún su comportamiento en reacciones o interacciones intermoleculares, son dependientes del contenido energético y de las carácterísticas tridimensionales de su estructura molecular, forma y tamaño. El rápido avance en la tecnología de los ordenadores y el desarrollo de algoritmos eficientes, ha facilitado la aplicación de nuevos métodos denominados semi-empíricos en el área de química cuántica. Estos métodos intentan proporcionar información molecular con un coste de cálculo reducido. Los métodos semi-empíricos han dado lugar a una fuente atractiva de nuevos descriptores moleculares, los cuales pueden en principio, expresar todas las propiedades geométricas y electrónicas de las moléculas y sus interacciones. Es por ello que en trabajos recientes se emplean este tipo de descriptores, ya sea, solos o conjuntamente con los descriptores convencionales.

La químicа computaciónal está compuesta de dos grandes áreas basadas en principios físicos distintos. Por una parte, la *mecánica molecular* (MM) que se basa en la mecánica clásica, considerando los átomos como particulas puntuales (dotadas de masa y carga) interaccionando unos con otros mediante enlaces que se asimilan mayormente a resortes. Por otro parte, los *métodos de estructura electrónica* (MEE) basados en la aplicación de la mecánica cuántica a los sistemas atómicos y moleculares. Tanto la MM como los MEE permiten calcular un conjunto básico de propiedades, por ejemplo la energía de un determinado arreglo de núcleos (conformación de la molécula), la

geometría óptima de un sistema molecular (es decir, el arreglo geométrico de los núcleos que brinde la energía más baja) y otras propiedades moleculares como las frecuencias vibracionales y el momento dipolar.

Los distintos métodos dentro de la MM comparten el hecho de tomar sus fundamentos de las leyes de la física clásica. Se diferencia en: (i) el tipo de ecuaciones que definen como varía la energía potencial de una molécula con la geometría de sus átomos; (ii) el tipo de átomos, es decir el tipo de características que definen a un átomo de número atómico dado en el entorno molecular en el que se encuentre (con lo cual, un oxígeno carboxílico no es equivalente necesariamente a un hidroxílico); y/o (iii) el conjunto de parámetros que ajustan las ecuaciones aplicadas a los átomos definidos por su tipo, a los valores experimentales que se emplean para la parametrización. Estos tres componentes definen lo que se conoce como *campo de fuerza* y definen completamente un método de MM.

Los MEE, emplean la mecánica cuántica para estudiar el comportamiento de núcleos y electrones ( y no sólo núcleos como en MM) que se consideran como partículas puntuales con carga y masa fijas e invariables, interaccionando según la ley de Coulomb. Los MEE se dividen en tres grupos principales: ab initio, funcionales de la densidad y semi-empíricos en función que no usen más que las constantes atómicas fundamentales o empleen datos experimentales para parametrizar parte de los cálculos. Todos los MEE se basan en el planteo y solución aproximada de la ecuación de Schrödinger.

El modelo ab initio del Halmiltoniano provee una representación completa de todas las interacciones no relativistas entre el núcleo y los electrones en una molécula. Sin embargo, la solución a la ecuación de Schrödinger es necesariamente una aproximación cuyo tiempo de cálculo es proporcional al número de electrones de la molécula. Muchos de los cálculos ab initio se basan en la aproximación de los orbitales moleculares (Hartree-Fock). En general este método provee buenos resultados para la energía electrónica total de la molécula. Otros métodos ab initio son ampliamente utilizados y comercializados lo que permite el cálculo de correlaciones electrónicas de las moléculas de una forma fácil. Entre ellos se encuentran, *configurational self consistent field* (MC SCF), *correlated pair many electron theory* (CPMET) y *perturbation theory* (Moller-Plesset theory (MP2, MP3, MP4)), de los cuales el MP2 es el método que se ha utilizado para el cálculo de la minimización de la energía de las moléculas. Hay que tener en cuenta que la mayoría de estos métodos no son prácticos y fallan en la convergencia en el caso de moléculas mayores de diez átomos.

Los métodos ab initio Hartree-Fock y post-Hartree-Fock tienen el enorme inconveniente que son muy costosos desde el punto de vista computacional y, consecuentemente aplicables sólo a moléculas de tamaño

reducido. El factor limitante en este tipo de métodos es el tiempo necesario para calcular el gran número de integrales bielectrónicas sobre las funciones bases y/o para transformarlas en integrales sobre orbitales moleculares. Los métodos de la MM por el contrario, son aplicables a grandes moléculas, pero no tienen en cuenta en froma explícita la estructura electrónica. Es por ello necesario encontrar una metodología intermedia que permita tratar la estructura electrónica de las moléculas de forma aproximada, sin los costosos requisitos de los métodos ab initio.

Como métodos alternativos, se han desarrollado métodos semi-empíricos, estos métodos están basados en la teoría de orbitales moleculares (SCF MO) y usan simplificaciones que reducen dramáticamente el tiempo de cálculo al eliminar gran parte o una parte de las integrales mencionadas anteriormente. Algunos parámetros se obtienen a partir de los datos experimentales de los átomos y/o de los sistemas moleculares prototipos. Entre los métodos más difundidos se encuentran, *extended Hückel theory* (EHT*), complete neglect of differential overlap* (CNDO), *intermediate neglect of differential overlap* (INDO), *Austin model 1* (AM1), y *Parametric model 3* (PM3).

MNDO, AM1 y PM3 parten de la correcta inclusión de la superposición de un centro. MNDO y AM1 presentan la ventaja de tiempos cortos de cálculo y que se encuentran parametrizados para un elevado número de átomos. En contraste con MNDO, AM1 provee una mejor descripción para sistemas con aniones o puentes de hidrógeno. MNDO tiende a sobrestimar la repulsión electrónica. Las cargas, dipolos y longitud de los enlaces obtenidas por AM1 son más realistas que aquellas obtenidas con métodos ab initio de baja calidad.

Es necesario hacer notar que, los resultados obtenidos por diferentes métodos semi-empíricos no son comparables de manera general, aunque dan tendencias similares. Es decir, la carga electrónica neta calculada por AM1, MNDO, o INDO puede ser muy diferente en valor absoluto, pero consistentes en sus tendencias, (Karelson M. et al., 1996).

A manera de ejemplo, el momento dipolar es una medición de la distribución y la fuerza de las cargas parciales en una molécula. Algunas moléculas tienen polos aparentes positivos o negativos. Algunas, presentan una distribución de carga parcial hacia uno de los lados de las mismas. Estas tendrán un momento dipolar mayor que aquellas con una distribución de carga media centralizada. Una medida de similitud cuántica molecular (MQS) definida por Carbó-Dorca y colaboradores (Carbó-Dorca R. et al., a) 1988; b) 1998), es útil para establecer medidas cuantitativas de semejanza entre las estructuras moleculares por medio de la projección de sus funciones de densidad. Las matrices de similitud cuántica molecular (MQSM) están basadas en postulados mecánico-cuánticos y usan a la función de densidad de primer

orden como descriptor molecular, el cual provee una representación 3D coherente. Los detalles teóricos y su implementación pueden encontrarse en diversas referencias (Amat L. et al.; 1998; Amat L. et al., 1997; Carbó-Dorca R. et al., 1998).

Una medida de MQS se define como una integral de volumen producto de dos funciones densidad multiplicadas por un operador hermitico no diferencial.

$$MQS_{ij} = \int \psi_i(\mathbf{r}) \Pi \psi_j^*(\mathbf{r}) d\mathbf{r} \quad (2.6)$$

En la ecuación (2.6) $MQS_{ij}$ representa la similitud cuántica molecular entre las moléculas i y j, $\psi(r)$ son las funciones densidad y $\Pi$, el operador hermítico no diferencial. De acuerdo con la forma de este operador se pueden definir diferentes medidas de similitud cuántica. El cálculo de las matrices de similitud cúantica se realizó siguiendo los tres primeros puntos del protocolo MQSM-QSAR y adaptando los dos últimos a la metología propuesta en la presente memoria. El procedimiento implica los siguientes pasos: i) se calculan las geometrías moleculares usando el método semi-empírico (PM3), ii) se calculan las densidades moleculares. Para evitar tiempos de cálculo excesivos se ha utilizado la aproximación ASA (atomic shell approxiamation) (Amat L. et al., 1997; Amat L. et al., 1998; Carbó-Dorca R. et al., 1998).; iii) se selecciona un operador hermítico de Overlap y Coulomb; y iv) se calculan las matrices de similitud cuántica, MQSM. Todos estos cálculos han sido implementados en el laboratorio de química computacional de la Universidad de Girona (Amat L. et al., 1997; Amat L., et al., 1998; Carbó-Dorca R. et al., 1998)

Los cálculos de las integrales dependen de la orientación relativa de las moléculas por lo tanto, las estructuras deben alinearze de tal forma que se maximice la medida. Así al analizar un conjunto de estructuras moleculares, todos los pares de MQS se recogen en una matrix, que se denominará matriz de similitud cuántica, MQSM, los elementos de su diagonal, represental los valores de auto-similitud. Estos valores son usados como descriptores en estudios de QSAR (Amat L. et al., 1997; Amat L., et al., 1998; Carbó-Dorca R. et al., 1998).

Los descriptores moleculares para cada compuestos se obtienen de la estructura molecular. Dichas estructuras son dibujadas con ayuda de un software comercial, Molecular Modeling Pro 3.01 (ChemSW Software Inc.) y convertidas a estructuras 3-D cuando los cálculos así lo requieren usando el Software CAChe (Oxford Molecular Ltd.) y Spartan Pro (Wavefunction, Inc.).

## 2.2. Propiedades

### 2.2.1 Propiedades fisicoquímicas

El presente trabajo se establecen métodos predictivos para las siguientes propiedades físicas: la temperatura de ebullición, el punto de fusión, la temperatura crítica, la presión crítica, y el coeficiente de actividad a dilución infinita. A continuación daremos una breve descripción de cada una de ellas.

#### 2.2.1.1 Punto de ebullición

El punto de ebullición es una propiedad sumamente útil en la identificación de sustancias desconocidas, y a su vez se emplea para estimar otras propiedades físicas (Reid R., 1987). Existen diferentes métodos para predecir el punto de ebullición entre los más populares se encuentran los métodos de contribución de grupos (Joback K. et al., 1987; Lymman W., 1990). En los cuales, se asume que las fuerzas de cohesión en el líquido son predominantemente de corto alcance y provienen de la división de una molécula en grupos estructuralmente predefinidos, cada uno de los cuales añade un incremento constante al valor de la propiedad para el compuesto estudiado. Aunque se encuentran limitados al tipo de compuestos para los cuales se ha pre-establecido el aporte de los grupos. Los métodos de relación estructura-propiedad son una alternativa a dichas limitaciones, (Katritzky A. et al., 1997; Jurs P. et al., 1997). Uno de los trabajos pioneros en este campo, realizado por Wiener (Wiener H., 1947), predice el punto de ebullición de un conjunto de parafinas. Otros índices topológicos, como los índices de conectividad molecular (Hall L. et al., 1995, Katrizky A. et al., 1997) y los índices de Randic (Randic M., 1993) se aplicaron con éxito en la correlación de puntos de ebullición de familias de compuestos orgánicos. Una revisión más detallada de los trabajos relativos al punto de ebullición la encontraremos en los apéndices, 1 y 2.

#### 2.2.1.2 Punto de fusión

La fusión es un proceso de transformación de un cristal o un sólido amorfo en un líquido. A la temperatura de fusión, la energía libre de la transición de fase es igual a cero, y de ahí la ecuación termodinámica para el punto de fusión,

$$T_m = \frac{\Delta H_m}{\Delta S_m} \qquad (2.7)$$

Siendo, $\Delta H_m$ la entalpía de transición y $\Delta S_m$ la entropía de fusión. El punto de fusión tiene numerosas aplicaciones en bioquímica y en ciencias

medioambientales debido a su relación con la solubilidad. Los primeros en realizar estudios de estructura propiedad para estimar la solubilidad de un compuesto a partir de su punto de fusión, coeficiente de partición y la entropía de fusión fueron Yalkowsky y Valvini (Yalkowsky S. et al., 1980). Sin embargo, no se ha desarrollado aún un método general basado en la estructura química de los compuestos para predecir la temperatura de fusión. Algunos de los trabajos previos en QSPR para determinar el punto de fusión están confinados a conjuntos reducidos de hidrocarbonos y a familias de aromáticos dando lugar a modelos relativamente satisfactorios. El modelo Needham et al. reportado en la literatura para alcanos normales y ramificados usando índices topológicos, presenta una desviación estándar, s=23.8K (Needham D. et al., 1994).

Las correlaciones de puntos de fusión de compuestos aromáticos, se han desarrollado con un éxito moderado. Por ejemplo, para el conjunto de 443 benzenos mono y di sustituidos, combinando descriptores tanto cuánticos como tradicionales, Katritzky obtiene un coeficiente de correlación, r=0.83 (Katritzky A. et al., 1997).

#### 2.2.1.3 Propiedades críticas

La temperatura y presión críticas son propiedades significativas que revelan aspectos importantes de las relaciones intermoleculares. La temperatura crítica puede entenderse fácilmente en términos de las fuerzas intermoleculares. Suponiendo que la Fig. 2.3 representa el potencial energético entre una molécula y una única molécula vecina, en función de la distancia que las separa.

La energía térmica de cada molécula es del orden de kT, donde k es la constante de Boltzmann's y T es la temperatura absoluta. De ahí que una molécula pueda escapar de sus vecinas si, $kT>\Delta\varepsilon$ o $T>\Delta\varepsilon/k$, eq. 2.6. La sustancia es por lo tanto no condensable para temperaturas superiores a,

$$T_c = \frac{\Delta\varepsilon}{k} \qquad (2.8)$$

De acuerdo a lo establecido en el principio de estados correspondientes, todas las sustancias obedecen a la misma ecuación en términos de sus variables reducidas, la temperatura, la presión y el volumen críticos. Las variables reducidas caracterizan un estado específico de una sustancia. Por lo tanto, las propiedades críticas son los factores de escala que expresan las características individuales de cada sustancia y se determinan por las diferencias en la estructura molecular de las mismas.

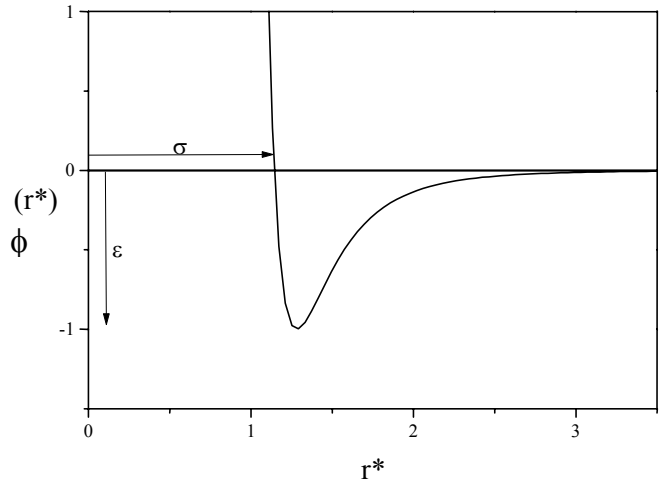



**Fig.2.3.** Energía potencial entre una molécula y su vecina como función de la distancia de separación

Los trabajos previos publicados en la literatura para propiedades críticas basados en la estructura molecular son escasos, algunos de ellos como el Langmuir en 1925, propone expresar las interacciones intermoleculares en el estado líquido usando la energía de superficie. La energía de superficie de un hidrocarburo puede expresarse en función del área de su superficie molecular total debido a que la energía de superficie por unidad de superficie puede considerarse uniforme. Algunos otros trabajos han intentado modelizar moléculas polares introduciendo para ello descriptores que tienen en cuenta la distribución de carga no uniforme creada por presencia de heteroátomos, (Turner B. et al., 1998), otros correlacionan la temperatura crítica de 165 compuestos orgánicos, combinando 19 índices electrotopológicos con los que caracterizan cada tipo de átomo (Hall L. et al., 1996). Katritzy (Katritzky A. et al.,1998) correlaciona la temperatura crítica de 76 hidrocarburos usando una combinación de índices cuánticos y topológicos. Las correlaciones obtenidas son aceptables, pero tienen un rango de aplicabilidad reducido a familias homogéneas de compuestos orgánicos.

#### 2.2.1.4 Coeficiente de actividad a dilución infinita, ln $\gamma^{\infty}$

El coeficiente de actividad a dilución infinita, $\gamma^{\infty}$, está directamente relacionado con el comportamiento termodinámico de soluciones acuosas, la predicción de azeótropos, el cálculo de solubilidades, las constantes de Henry y los coeficientes de partición.

El agua es una sustancia única tanto en su habilidad para disolver electrolitos y compuestos polares como en su aversión hacia los compuestos no polares. Las mezclas acuosas se encuentran en diversas operaciones industriales, como el procesamiento del petróleo, la gasificación del

carbón, la separación y el tratamiento de desechos. El éxito limitado de modelos simples para caracterizar compuestos orgánicos polares se explica a partir de lo que se denomina asociación hidrofóbica y da lugar a una autoorganización de microestructuras como las micelas.

Se han desarrollado diversos métodos para medir los coeficientes de actividad a dilución infinita, $\gamma\infty$ los más utilizados son: *la cromatografía gas-líquido* (GLC), *non-steady-state gas-liquid chromatography*, *differential ebulliometry*, *static methods and the dilutor method*. Los métodos cromatográficos permiten la determinación de los coeficientes de actividad a dilución infinita, $\gamma^{\infty}$ de solutos volátiles en alta (clásica GLC) y baja ebullición (non-steady state GLC) de solventes. Para solutos que son apreciablemente inmiscibles en el solvente, como lo es el caso de muchos solutos orgánicos en agua, la adsorción en la interface vapor-líquido es el proceso de retención dominante lo que da lugar a una sobrecarga de la columna. El método *dilutor* permite determinar $\gamma^{\infty}$ en mezclas de solventes.

El comportamiento no ideal de las soluciones acuosas hace difícil extender los métodos existentes como UNIFAC, o ASOG para estimar los $\gamma^{\infty}$ de sistemas acuosos diluidos donde las variaciones de dichos coeficientes, $(\gamma^{\infty})$son grandes. La viabilidad de los métodos de contribución de grupos depende de la existencia de los grupos de volumen (Rk), de los grupos de superficie (Qk), y tanto de la existencia como de la precisión de los parámetros de interacción. *Linear solvation energy relationship* (LSER) asume que las interacciones soluto-solvente son debido a de las interacciones específicas del enlace hidrógeno. Este tipo de técnicas falla al correlacionar $\gamma^{\infty}$ para soluciones diluidas de alcanos donde el tamaño soluto-solvente varía y la formación de cavidades es la contribución primaria. Simulaciones de energía libre basada en el método de perturbaciones es una técnica cuyo costo y tiempo de cálculo es elevado. En los últimos años, las relaciones estructura propiedad (QSPR) se han presentado como una alternativa asequible. Su ámbito de aplicación, en la mayoría de los casos, es más amplio que el de una ecuación de estado tradicional o un método de contribución de grupos, y cuyo tiempo de cálculo es muy inferior al de técnicas más sofisticadas como la simulación molecular.

Entre los trabajos previos en los que se correlacionan las características estructurales y el $\gamma^{\infty}$ se encuentran los de Mackay y Shiu (Mackay D. et al., 1977) que relacionan el $\gamma^{\infty}$ de hidrocarburos polinucleares aromáticos en agua con el número de átomos de carbono, Yalkowsky y Valvani (Yalkowsky S. et al., 1979) relacionaron la solubilidad con el área molecular superficial. Medir y Giralt (Medir M. et al., 1982) correlacionaron el $\gamma^{\infty}$ con parámetros moleculares, Mitchell y Jurs (Mitchell

B. et al., 1998) estiman el $\gamma^{\infty}$ a partir de doce descriptores moleculares: tres descriptores topológicos, cuatro descriptores CPSA, dos descriptores para el enlace hidrógeno, el calor de formación y dos descriptores TLSER para describir la basicidad del enlace hidrógeno.

### 2.2.1.5 Índices de toxicidad y propiedades biológicas

Por otro lado, la metología se aplica a diferentes tipos de índices de toxicidad con el fin de extrapolar los alcances de la misma a estudios estructura-actividad (QSAR) entre ellos tenemos: el índice de concentración letal, $LC_{50}$, el índice correspondiente a la dosis letal $LD_{50}$ y el índice de la dosis letal $TD_{50}$. A continuación daremos una breve descripción de cada una de ellos. Finalmente, se presenta una pequeña introducción a la actividad anti-HIV-1 de los derivados HEPT.

#### 2.2.1.5.1 Toxicidad

La *toxicología* estudia los daños causados al organismo por la exposición a los agentes tóxicos que se encuentran en el ambiente. Su objetivo principal, es evaluar los impactos que producen en la salud pública la exposición de la población a los tóxicos ambientales presentes en un sitio contaminado. Es conveniente recalcar que se estudian los efectos sobre los humanos, aunque pudieran existir, en el sitio de estudio, otros blancos de los tóxicos tales como microorganismos, plantas, animales, etc.

Los *tóxicos* son los xenobióticos que producen efectos adversos en los organismos vivos. Un *xenobiótico* es cualquier substancia que no ha sido producida por la biota, tales como los productos industriales, drogas terapéuticas, aditivos de alimentos, compuestos inorgánicos, etc. La *biota* son todos los seres vivos; sean plantas o animales superiores o microorganismos.

Por otro lado cabe definir el término *exposición*, como el contacto de una población o individuo con un agente químico o físico. La magnitud de la exposición se determina midiendo o estimando la cantidad (concentración) del agente en la superficie expuesta durante un período de tiempo específico. Esta cantidad cuando se expresa por unidad de masa corporal del individuo expuesto se le denomina *dosis.*

Paracelsus estudió el envenenamiento de plantas o animales causado por sustancias químicas específicas. De la misma forma citó la relación entre la dosis recibida y la respuesta del cuerpo. Sus estudios enfatizan como pequeñas dosis pueden ser inofensivas o quizá benéficas mientras que dosis mayores pueden resultar tóxicas. Esto es lo que se conoce como

relación *dosis-respuesta*. La cual constituye uno de los conceptos básicos en toxicología. La relación *dosis-respuesta* correlaciona la exposición y el espectro de efectos inducidos. Generalmente, a mayor dosis la respuesta es más severa. Las relaciones *dosis-respuesta* están basadas principalmente en observaciones experimentales sobre animales, estableciendo i) la causalidad de que un compuesto químico induzca el efecto observado y ii) la dosis mínima donde tiene lugar el efecto inducido, además iii) determina la velocidad a la cual el daño se magnifica (efecto umbral).

La curva de dosis-respuesta presenta una forma sigmoidea. El primer punto de la curva correspondería a un umbral (punto thershold), es decir, el punto donde comienza a excederse la habilidad del cuerpo para detoxificar y/o reparar el daño tóxico causado por un xenobiótico. Las dosis tóxicas (TDs) indican la dosis que causa un efecto tóxico adverso, entre ellas encontramos a $TD_0$, $TD_{10}$, $TD_{50}$, $TD_{90}$, el subíndice nos indica el porcentaje de la población para la cual ha resultado tóxica una sustancia.

De la misma forma de las curvas *dosis-respuesta* se obtienen dosis estimadas de los efectos tóxicos. Un estimador común es la dosis letal para el 50% de la población, $LD_{50}$. Es la dosis a la cual el 50% de los individuos mueren o se espera que lo hagan. Cuando hablamos de toxicidad por inhalación, se usa la concentración del aire para calcular los niveles de exposición. El índice se denomina, $LC_{50}$, y representa la concentración letal para el 50% de la población. Basak y colaboradores (Basak S. et al., 1998) presentaron un estudio comparativo usando métodos estadísticos y redes neuronales para predecir diferentes modos tóxicos de acción de 283 compuestos diversos. La mayoría de estas moléculas habían sido previamente identificadas como narcóticos. Índices topológicos y pares de átomos son los descriptores usados para desarrollar los modelos. El rango de certeza en la clasificación fue de 65 a 95% para el conjunto global de compuestos. Gute y Basak (Gute B., y Basak S., 1997) desarrollaron un modelo para predecir la toxicidad ($LC_{50}$) de un conjunto de 69 derivados del benceno a partir de parámetros que ellos denominan topo estructurales, topo químicos, geométricos y cuánticos, el mismo conjunto había sido estudiado por Hall et al. (Hall L. et al., 1984) quienes construyeron un modelo basados en la contribución de los diferentes substituyentes que forman los compuestos. Siendo este conjunto uno de los seleccionados para aplicar la metodología fuzzy ARTMAP QSAR los resultados serán discutidos con mayor detalle en la sección de resultados. Randic y Basak (Randic M. y Basak S. 2001) publicaron un modelo para 21 alquil éteres utilizando el número de átomos hidrógeno (N) como descriptor molecular obteniendo un coeficiente de correlación de 0.9751, en comparación con 0.9548 que citan usando el índice de conectividad de primer orden en lugar de N. Basak et al., (Basak S. et al., 1999) estiman la potencia inhibitoria complementaria de 105 benzamidinas ($I^D$) a partir del índice de Wiener 3D ($^{3D}W$) obteniendo un coeficiente de regresión de 0.944, y una desviación estándar de 0.0196.

### 2.2.1.5.2 Actividad Anti-HIV-1

El diseño de nuevos fármacos es uno de los desafíos del siglo XXI. Dentro de este contexto, el estudio de diferentes inhibidores de las proteínas que son una pieza clave en el desarrollo del virus de inmuno deficiencia adquirida (HIV) y de los organismos responsables de las infecciones relacionadas con el mismo se presenta como un reto multi-disciplinario apasionante.

La *proteasa* es una de las enzimas del HIV, es necesaria para propagar la infección. Su trabajo comienza cerca del final del proceso de replicación, después de que el virus del HIV ha entrado en el núcleo de la célula y ha formado largas cadenas de proteínas y enzimas que darán lugar a la replicación mismo, pero antes de que ellas puedan comenzar a trabajar correctamente. Es en este punto donde las largas cadenas tienen que ser divididas en pequeñas piezas. La proteasa del HIV es como una tijera ya que corta las largas cadenas en pequeñas piezas ( Wei X. et al., 1995; Korber B., et al., 1996), Fig. 2.4.

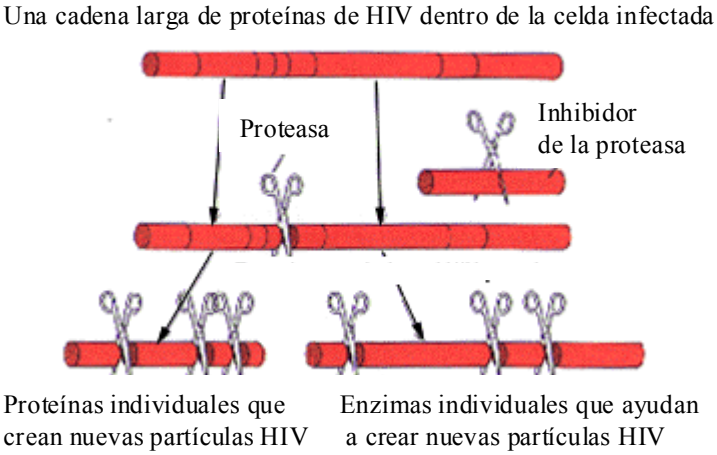



**Fig.2.4.** HIV puede crear nuevas copias de si mismo dentro de la célula infectada. Uno de los pasos es cortar cadenas largas de proteínas y enzimas en trozos de cadenas cortas. Las "tijeras" utilizadas es una enzima llamada proteasa. Referencia: International Association of Physicians in AIDS Care, Web:http://www.iapac.org

Los principales inhibidores de la proteasa difieren con respecto a otro anti-fármacos en su fuerza. A este tipo de inhibidores se les denomina inhibidores de la *reverse-transcriptasa* debido a que inhiben o perturban el trabajo de la enzima del HIV que lleva el mismo nombre, y se denomina comúnmente como RT. La RT es la enzima que el HIV usa para cambiar su mensaje genético químico a una forma que pude ser fácilmente insertada dentro del núcleo de la célula infectada (Bukrinsky M. et al., 1993).

Uno de los objetivos favoritos del HIV es la célula blanca de la sangre llamada célula auxiliadora T o célula CD4. Estas células de la

sangre son importantes porque ellas dicen a otras células combatientes de la infección cuando comenzar a trabajar. El HIV destruye las células CD4, y cuando el número de CD4 cae de cierto nivel, la infección por HIV continua y el sistema inmune del cuerpo se debilita. Como resultado, ciertos organismos como hongos, otros virus y parásitos pueden causar infecciones serias en personas con HIV. Cuando estas infecciones ocurren, o cuando el número de CD4 cae de cierto límite, se dice que una persona infectada de HIV tiene SIDA (Ho D. et al., 1995).

Otra clase de inhibidores HIV-RT son los HEPT (Kireev D. et al., 1997; Luco J. et al., 1997) que han demostrado ser uno de los inhibidores más potentes y selectivos del HIV-1. El diseño de nuevos fármacos derivados del HEPT requiere un conocimiento más detallado del mecanismo de inhibición de la RT para esta clase de compuestos. Los métodos QSAR son frecuentemente usados para predecir la actividad y definir los modelos fármaco cinéticos, de igual forma que pueden usarse para comprender y estudiar el comportamiento de un inhibidor en un sistema biológico. Particularmente, un modelo QSAR puede proporcionar información sobre los tipos de interacciones intermoleculares e intramoleculares de las moléculas activas expuestas durante su incorporación en la enzima.

Una radiografía reciente de la estructura cristalina de la enzima y del complejo inhibidor de la misma ha sido publicada recientemente por Kelly y colaboradores (Kelly T. et al., 1995), con lo cual se han derivado QSAR convencionales y modelos 3D-QSAR para el mismo conjunto de compuestos. La conclusión global que puede derivarse de estos estudios experimentales muestra que el carácter hidrofóbico y las características estéricas de los substituyentes juegan un papel predominante en la actividad inhibitoria del HIV-1.

Hasta el momento hemos encontrado en la literatura dos trabajos previos en los que utilizan las redes neuronales en el diseño de inhibidores HIV-1-RT. El primero sobre un conjunto de 44 moléculas derivadas del AZT (Tetko I. et al., 1994) y el segundo en el que utilizado técnicas convencionales de redes neuronales junto con técnicas de regresión multilineal modelan a los inhibidores del HIV-1 de 107 derivados del HEPT (Jalali-Heravi M. et al., 2000).

**Tabla 2.1.** Símbolos y definiciones de los parámetros topológicos y cuánticos utilizados

| *Símbolo* | *Definición* | *Descripción* | *Referencias* |
|---|---|---|---|
| | | **Índices 2-D** | |
| ${}^{m}_{t}\chi$ | Índices de conectividad molecular, orden m=0-4 | Cuantifican fragmentos moleculares, a partir de los electrones en los orbitales σ | Randic M. 1984<br>Kier B. et al., 1986<br>Hall L. et al.,. 1991 |
| ${}^{m\ v}_{t}\chi$ | Índices de conectividad molecular de valencia, orden m=0-4 | Cuantifican fragmentos moleculares, tomando en cuenta los electrones de valencia | Kier B. et al., 1986<br>Hall L. et al., 1991 |

| | | | |
|---|---|---|---|
| **$^{n}\kappa$** | El índice kappa o índice de forma | Cuantifica los atributos de la forma molecular en 3 valores, los cuales se derivan de los fragmentos con uno, dos y tres enlaces | Kier L. 1989<br>Hall L. et al., 1991 |
| **N** | Suma de números atómicos | La cantidad de protones existentes en el núcleo del átomo | Mackay D., et al., 1993<br>Lyman W., 1990 |
| **$M_w$** | Peso molecular | Es la masa de 1 mol de moléculas del compuesto. Es la suma de las masas atómicas de los átomos componentes. | Mackay D., et al., 1993<br>Lyman W., 1990 |
| **SA** | Superficie molecular | Area de las esferas de radio Van der Waals menos su superposición | ChemSM Inc., 1988<br>Stanton D. et .al., 1990 |
| **HD** | Dispersion de Hansen | Se define a a partir de una suma de fragmentos. Los valores fueron determinados por el trabajo de Hansen | ChemSM Inc., 1988<br>Hasch C. et al., 1964 |
| **HP** | Polaridad de Hansen | Se define a a partir de una suma de fragmentos. Los valores fueron determinados por el trabajo de Hansen | ChemSM Inc., 1988<br>Hasch C. et al., 1964 |
| **HH** | Hidrógeno de Hansen | Se define a a partir de una suma de fragmentos. Los valores fueron determinados por el trabajo de Hansen | ChemSM Inc., 1988<br>Hasch C. et al., 1964 |
| **Randic** | Índice de Randic | Cuantifican la conectividad molecular | Randic M., 1975<br>Randic M., 1984 |
| **Balam** | Índice de Balaban | Cuantifican la conectividad molecular | Balaban A., 1982<br>Balaban A., 1988 |
| **$\mu$** | Momento dipolar | El momento dipolar puede considerarse como el primer término del desarrollo del campo eléctrico generado por la molécula, siendo el término que le sigue el momento cuadrupolar (cuya consideración es importante para aquéllas moléculas que siendo suficientemente simétricas tienen momento dipolar nulo).i). Calculado a partir de las cargas parciales de los átomos | ChemSM Inc., 1988<br>Lyman W., 1990<br>Dewar, M., et al., 1985 |
| | **Índices 3-D obtenidos a partir del método semi-empírico PM3** | | |
| **$\mu$** | Momento dipolar | ii) * Calculado usando un método semi-empírico, PM3 | Stewart J., 1989<br>McWeeny R. 1999 |
| **AP** | Average molecular polarizability | La propiedad más significativa de la polarizabilidad molecular es su relación con el seno molecular o volumen molar. | 53<br>Stewart J., 1989<br>McWeeny R. 1999 |
| **NFL** | Número de niveles llenos | número de niveles en los orbitales moleculares doblemente ocupados ocupados | Stewart J., 1989<br>McWeeny R. 1999 |
| **TE** | Energía total | La energía total, es la suma de los términos que involucran uno dos centros, respectivamente en de la teoría de orbitales moleculares. Los términos de energía de un centro incluyen la EER y ENA. Los términos de dos centros incluyen a la RE, EE, EER, ENA, NNR | Stewart J., 1989<br>McWeeny R. 1999 |
| **EER** | Energía de repulsión eléctrónica | | Stewart J., 1989<br>McWeeny R. 1999 |
| **ENA** | Energía de atracción electrón-núcleo | | Stewart J., 1989<br>McWeeny R. 1999 |
| **RE** | Energía de resonancia | Corresponde a la diferencia entre los electrones pi completamente deslocalizados localizados en el doble enlace | Stewart J., 1989<br>McWeeny R. 1999 |

| | | | |
|---|---|---|---|
| **EE** | Energía de intercamio | La energía de atracción es calculada en términos del núcleo y la superposición de la carga en el enlace | Stewart J., 1989<br>McWeeny R. 1999 |
| **HOMO** | Homo | Orbitales moleculares mayormente ocupados | Stewart J., 1989<br>McWeeny R. 1999 |
| **LUMO** | Lumo | Orbitales moleculares menormente ocupados | Stewart J., 1989<br>McWeeny R. 1999 |
| **$W^{3D}$** | Índice de Wiener 3D | Toma en cuenta las ramificaciones en la estructura molecular | Wiener, H., 1947<br>Bondanov B., et al., 1989 |
| **Ove** | Índice de similitud cuántica de overlap | Cuantifica la simulutud entre las estructuras moleculares por medio de la proyección de sus funciones de densidad usando un operador hermítico, Overlap | Carbó-Dorca R., et al., 1998<br>Amat L. et al., 1997 |
| **Cou** | Índice de similitud cuántica de Coulomb | Cuantifica la simulutud entre las estructuras moleculares por medio de la proyección de sus funciones de densidad usando un operador hermítico, Coulomb | Carbó-Dorca R., et al., 1998<br>Amat L. et al., 1997 |
| **Kin** | Índice de similud cuántica cinético | Cuantifica la simulutud entre las estructuras moleculares por medio de la proyección de sus funciones de densidad usando un operador hermítico, Cinético | Carbó-Dorca R., et al., 1998<br>Amat L. et al., 1997 |

# 3. REDES NEURONALES

## 3.1. Algoritmos, prestaciones y preprocesado de datos

La búsqueda cuantitativa de la relación entre la estructura de un compuesto y una propiedad específica (SPR/SAR) es en esencia un problema de regresión o un proceso de reconocimiento de patrones. Históricamente, la regresión multilineal (MLR), el análisis de componentes principales (PCA) y de mínimos cuadrados parciales (PLS) han sido los métodos más empleados en el desarrollo de los modelos QSAR/QSPR. Sin embargo, uno de los inconvenientes lo encontramos al tener que asumir *a priori* la propia forma del modelo. El hecho de fijar una relación lineal entre los descriptores moleculares que caracterizan a la estructura molecular y la propiedad que se está modelando, no implica que esta se cumpla o que el modelo obtenido sea el óptimo.

Las relaciones estructura propiedad QSAR/QSPR pueden modelarse con otras técnicas de reconocimiento de patrones como las *redes neuronales*, (ANN). En muchos casos, ANN han demostrado dar mejores resultados que los métodos tradicionales de MLR, PLS debido a i) su habilidad de tratar con las no-linealidades inherentes en las relaciones estudiadas; ii) a ser capaces de producir resultados razonables aún cuando se carezca de alguno de los datos de entrada o dichos datos presente ruido; iii) a su capacidad de generalizar los resultados a nuevos datos no presentados durante la fase de entrenamiento de la red.

Existen numerosas formas de definir a una red neuronal, desde las definiciones cortas y genéricas hasta las que intentan explicarlas más detalladamente. Una de las más populares es la propuesta por Kohonen, (Kohonen,T., 1988): *Redes neurales artificiales son redes interconectadas masivamente en paralelo compuestas de elementos simples (usualmente adaptativos) y con una organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico*.

En las redes neurales biológicas, las células neuronales (neuronas) representan a los elementos de procesamiento. Las interconexiones se realizan por medio de las ramas de salida (axones) que producen un número de variable de conexiones (sinapsis) con otras neuronas, Fig. 3.1.

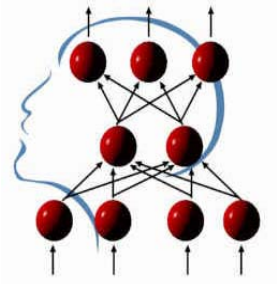


**Fig. 3.1.** Arquitectura de una red neuronal. Cuatro unidades en la capa de entrada, dos neuronas en una única capa intermedia y tres neuronas a la salida

La neurona artificial pretende mimetizar las características más importantes de las neuronas biológicas. En la Fig. 3.2 se muestra la estructura de una neurona artificial con *n* entradas. Cada canal de entrada *i*, pude transmitir un valor real $\mathbf{x}_i$. La función primitiva *f* evaluada en el cuerpo de la neurona se selecciona de forma arbitraria. Usualmente los canales de entrada tienen un peso asociado a ellos, lo cual significa que la información de los vectores de entrada $\mathbf{x}_i$, se multiplica por el correspondiente peso $\mathbf{w}_i$. La información transmitida se integra en la neurona para evaluar a la función primitiva.

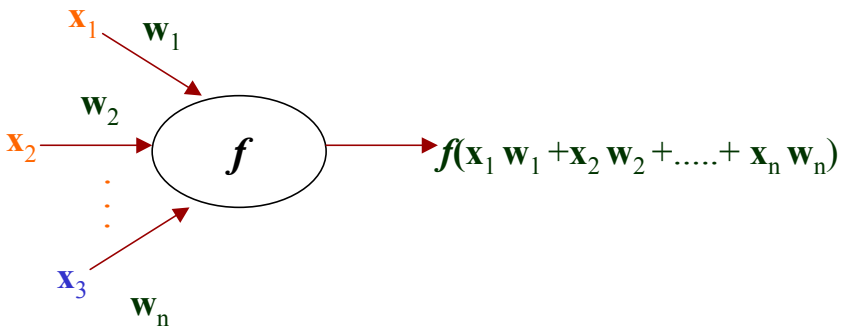



**Fig. 3.2.** Una neurona artificial

Si se concibe cada nodo en una red neuronal como una función primitiva capaz de transformar sus entradas en una salida definida, entonces se puede decir que las redes neuronales no son otra cosa que una red de funciones primitivas. Los diferentes modelos de redes difieren la estructura de los nodos, la topología de la red y el algoritmo de aprendizaje empleado para encontrar los pesos de la misma.

Las redes neuronales presentan un gran número de características semejantes a las del cerebro. Por ejemplo, son capaces de aprender basándose en la experiencia, de generalizar de casos anteriores a nuevos casos, de abstraer características esenciales a partir de entradas que presentan información irrelevante, etc. Esto hace que ofrezcan numerosas ventajas y que este tipo de tecnología se esté aplicando a múltiples áreas. Estas ventajas incluyen:

*Aprendizaje adaptativo*. Capacidad de aprender a realizar tareas basada en un entrenamiento o en una experiencia inicial.

*Autoorganización*. Una red neuronal puede crear su propia organización o representación de la información que recibe mediante una etapa de aprendizaje.

*Tolerancia a fallos*. La destrucción parcial de una red conduce a una degradación de su estructura; sin embargo, algunas capacidades de la red se pueden retener, incluso sufriendo un gran daño.

*Paralelismo*. Los computadores neuronales pueden ser procesados en paralelo.

El diseño de una red neuronal consiste básicamente en: (i) determinar el arreglo de las neuronas (o unidades) en cada capa, (ii) decidir el tipo de conexiones (pesos) entre las neuronas de las diferentes capas, como entre la neuronas de la misma capa, (iii) decidir la forma como la neurona recibe las entradas y produce la salida, (iv) determinar la fuerza de las conexiones con la red al permitir que la red aprenda los valores apropiados de los pesos durante el entrenamiento del sistema.

Los sistemas de redes neurales se han aplicado con éxito en diversas áreas como:

- *predicción*, a partir de determinadas entradas se genera una salida.
- *clasificación*, usa los valores de entrada para determinar la categoría a la cual pertenece.
- *asociación de datos*, semejante a la clasificación pero también reconoce datos que contienen errores.
- *conceptualización de datos*, analiza los datos de entrada para agruparlos de acuerdo a las relaciones que pueda inferir de los mismos.
- *filtro de datos*, genera señales de entrada sin ruido.

### 3.1.1 Topología de las redes

Biológicamente, las redes neuronales están construidas de forma 3D por componente microscópicos. Estas neuronas parecen ser capaces de interconectarse sin ninguna restricción. Esto no es posible en las redes neurales artificiales las cuales son un simple conglomerado de neuronas artificiales primitivas (ver descripción previa, Fig. 3.2). Las neuronas se distribuyen creando capas, las cuales se conectan unas con otras. Como es está conexión varía de un tipo de red a otra. La capa de entrada consiste en una serie de neuronas que reciben la información de entrada del exterior. La capa de salida consiste en una serie de neuronas que comunican la respuesta del sistema al usuario o ambiente exterior. Usualmente hay un número de capas intermedias (ocultas) entre estas dos capas. Los parámetros a determinar referentes a la topología de una red neuronal son, por tanto:

el número de capas, el número de neuronas por capa, el grado de conectividad y el tipo de conexión entre las neuronas. Para ello, se puede recurrir al método de prueba y error, teniendo en cuenta que si se incrementan demasiado el número de neuronas ocultas es posible que ocurra un "*over fitting*" en el sistema, que hará que la red tenga problemas durante la generalización ( o fase de prueba) de la red. Es decir el conjunto de entrenamiento será memorizando dando lugar a una red inservible para predecir nuevos casos.

Si clasificamos a las redes desde un punto de vista topológico, podemos hablar de redes monocapa y las redes multicapas.

*Redes monocapas.* Una red formada por sucesivas capas tendrá como forma más simple una capa de nodos de entrada y otra de nodos de salida, Fig. 3.2

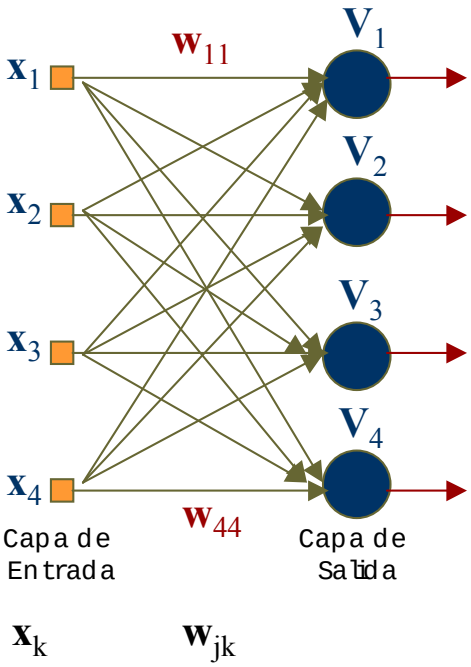



**Fig. 3.3.** Redes monocapa

donde $\mathbf{x}_i$ representa el vector de entrada del patrón *k*, $V_j$ a las unidades intermedias, $\mathbf{w}_{jk}$ las conexiones o pesos que relacionan a la neurona i de la capa de entrada con la neurona *j* de la capa intermedia.

*Redes multicapas.* Una red multicapa consiste en una extensión de una red monocapa, con tantas capas ocultas como se desee. Existen dos variantes, totalmente conectadas, y las parcialmente conectadas. Este subgrupo se caracteriza porque no todas sus unidades de la capa oculta están conectadas a todas las unidades de entrada. Donde $\mathbf{x}_i$ representa el vector de entrada del patrón *k*, $V_j$ a las unidades intermedias, $\mathbf{w}_{jk}$ las conexiones o pesos que relacionan a la neurona i de la capa de entrada con la neurona *j* de la capa

intermedia y $w_{ij}$ de la capa intermedia a las unidades de salida. El índice *i* siempre se referirá a la unidad de salida, *j* al de las unidades o neuronas ocultas, y *k* al nodo o neurona de entrada.

Normalmente, todas las neuronas de una capa reciben señales de entrada de otra capa anterior, más cercana a la entrada de la red. A este tipo de conexiones se les denomina conexiones hacia delante o *feedforward*. Las redes más conocidas son: *perceptron, backpropagation, cascade correlation*, siendo especialmente útiles en problemas de clasificación y reconocimiento de patrones. Por otro lado, cabe la posibilidad de conectar las salidas de las neuronas de capas posteriores a las entradas de las capas anteriores, a estas conexiones se les denomina conexiones hacia atrás o *feedback.*

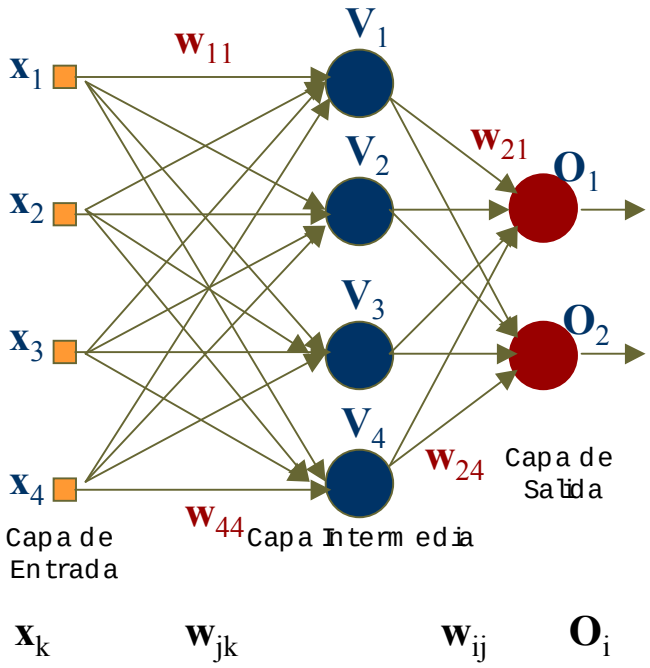



**Fig. 3.4.** Redes multicapas

Además, existen redes que combinan ambos tipos de conexiones, es decir, son redes con conexiones hacia delante y hacia atrás (*feedforward/feedback*). En general, suelen ser redes bicapas, existiendo por lo tanto, dos conjuntos de pesos correspondientes cada uno de ellos a un tipo de conexión. Este tipo de estructura es particularmente útil para asociar la información contenida en el patrón de entrada con la información del patrón de salida, proceso conocido como *heteroasociación*. Algunas redes de este tipo tienen un funcionamiento basado en lo que se conoce como *resonancia*, de tal forma que la información tanto de la primera como de la segunda capa interactúan entre sí hasta alcanzar un estado estable. Uno de los modelos más conocidos ART (*Adaptive Resonance Theory*) se presenta en el siguiente sub-apartado.

### 3.1.2 Mecanismo de aprendizaje

Un algoritmo de aprendizaje es un método adaptativo por el cual una red adapta los parámetros de acuerdo con la experiencia previa (ejemplos de casos presentados previamente) hasta que se alcanza una solución, si esta existe.

Los algoritmos de aprendizaje pueden ser divididos en, Fig. 3.5: i) supervisados y ii) no supervisado. En el aprendizaje supervisado, la desviación de la salida producida por la red con respecto a la observación real se usa para corregir o modificar los pesos de la red de acuerdo con la regla definida por el algoritmo de aprendizaje. El aprendizaje no supervisado, se usa cuando para cada vector de entrada dado, no se conoce el valor de salida.

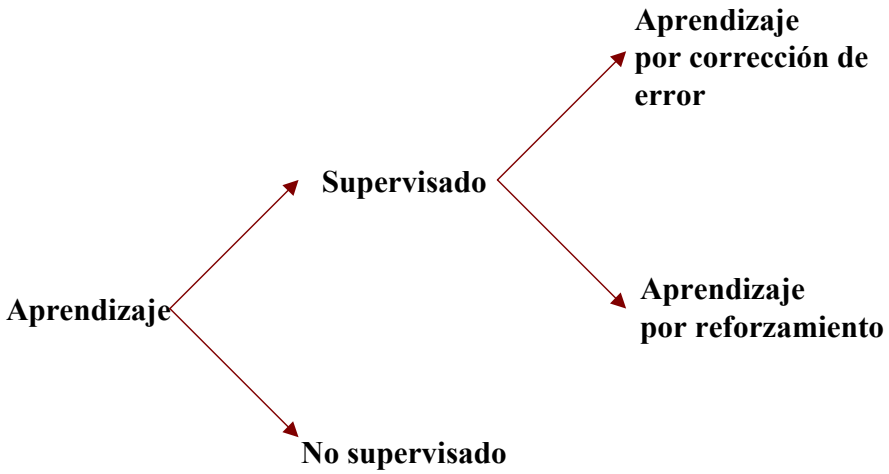



**Fig. 3.5.** Clases de algoritmos de aprendizaje

La diferencia fundamental entre ambos tipos estriba en la existencia o no de un agente externo al que se denomina *supervisor* que controle el proceso de aprendizaje de la red.

Hay cuatro reglas básicas de aprendizaje:

i) *Regla de Hebb*. La primera regla y la más conocida fue introducida por Donald Hebb. Su descripción aparece en su libro *The organization of Behaviour* in publicado en 1949. Se basa en el siguiente postulado: Si una neurona recibe una entrada de otra neurona, y si ambas están suficientemente activas (matemáticamente tienen el mismo signo), los pesos entre las neuronas deben reforzarse.

(ii) *Regla Delta*. La regla delta es una variación de la regla de *Hebb*, y es una de las más utilizadas. Esta regla se basa en la idea de modificar continuamente las conexiones de la entrada a la red (pesos) para reducir la diferencia (la delta) entre la salida deseada (variable objetivo) y la salida actual de la neurona. Esta regla cambia los pesos de forma que se

minimice el error cuadrático medio de la red. Consiste en ajustar los pesos de la red en función de la diferencia entre los valores deseados y los obtenidos a la salida de la red. El error es propagado hacia las capas previas, una a una hasta alcanzar la primera capa. Esta regla también se conoce como regla de aprendizaje de *Windrow-Hoff*.

(iii) *Aprendizaje competitivo*. Presenta un mecanismo que permite a las neuronas competir por el derecho a responder a las entradas, de forma que sólo una neurona, o una por grupo activo, podrá hacerlo. La neurona que gana la competición se le denomina neurona vencedora (*winner take all-unit*), quedando el resto anuladas forzándolas a sus valores de respuesta mínimos. La competición entre las neuronas se realiza en todas las capas de la red, existiendo en estas neuronas conexiones recurrentes de autoexitación y conexiones de inhibición (signo negativo) por parte de las neuronas vecinas, Fig. 3.6

(iv) *Aprendizaje de Boltzmann*. Debe su nombre a Boltzmann. En una máquina de *Boltzmann*, las neuronas constituyen una estructura recurrente y operan de forma binaria esto es +1/-1 son estados conocidos como *on/off*. Además, se caracteriza por una función energía *E* que está definida por los estados ocupados por cada neurona en la máquina. La máquina funciona de forma que se elige una neurona aleatoriamente y se le cambia de estado a una temperatura *T* con una probabilidad determinada. Si esta regla se aplica sucesivamente la máquina alcanzará *un equilibrio térmico*.

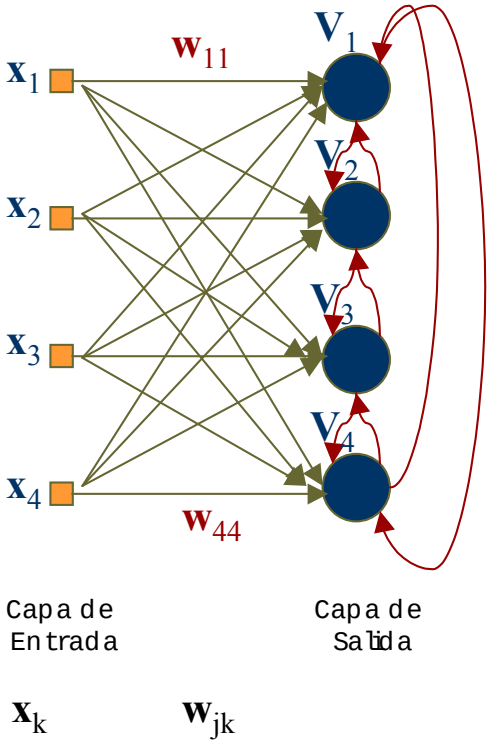



**Fig. 3.6.** Redes con aprendizaje competitivo

En la siguiente sección presentaremos una breve descripción de los algoritmos utilizados para estimar los modelos QSAR/QSPR. Comenzando con el *multilayer perceptron* de estructura predefinida y entrenado mediante el algoritmo de *backpropagation*; los resultados serán comparados con una red de tipo *cascade correlation*, capaz de crear su propia estructura a medida que se entrena la red y utilizando *backpropagation* como algoritmo de aprendizaje. Dos algoritmos de tipo competitivo, mapas auto-organizados de Kohonen (SOM) y fuzzy ARTMAP son utilizados para la selección de índices y para la construcción de los modelos SPR/SAR, respectivamente.

## 3.2. Redes de perceptrones con aprendizaje supervisado

### 3.2.1 Backpropagation

Una red neuronal de tipo *backpropagation* es de hecho una red *multilayer perceptron*. Esta clase de redes se caracteriza por una estructura multicapa, cada capa recibiendo su entrada de la capa precedente y enviando su salida a la capa siguiente.

*Backpropagation* es una generalización de la regla de aprendizaje delta (corrección de error). La técnica original genera una señal de error a partir de la diferencia entre la salida actual de la red y la salida real. Los pesos (fuerzas sinápticas) se actualizan de forma proporcional al producto de la señal de error y la señal de entrada a la red, con lo cual se disminuye el error en dirección del gradiente (la dirección de cambio más rápido). En una red multicapa con unidades ocultas, el problema es mucho más complejo. La señal de error puede calcularse de la misma forma que en la regla delta original, pero para las capas ocultas el error no puede determinarse directamente. Es así como se define la regla delta generalizada, que da las pautas para ajustar los pesos basándose en la corrección del error en la capa de salida. Las unidades internas (ocultas) deben decir que tan grande es el error, y que tan fuertemente conectadas las unidades ocultas están conectadas a la unidad de salida en base al error (si una unidad oculta no contribuye al error, no será necesario modificar su peso).

De forma simplificada, este algoritmo es utilizado para el entrenamiento de una red multicapa. La red se inicializa con una serie de pesos aleatorios y tras hacerlos pasar a través de la estructura de la red se compara la salida con el patrón que deseamos que aprenda. La diferencia entre entradas-salidas, estimación de error o valores delta se utilizan para derivar los errores de los pesos de las sucesivas capas de la red haciendo pasar estos valores retrógradamente a la dirección de activación de la red, Fig. 3.2

Existe una función matemática para realizar esta modificación de los pesos. El objeto de la ecuación es buscar un mínimo del espacio de la función -n- dimensional de los pesos. En un perceptrón la técnica del

gradiente descendente da lugar a una solución única. En las redes multicapa nos encontramos con el problema de los mínimos locales.

A diferencia de la regla delta en el caso del perceptrón la técnica de *backpropagation* o generalización de la regla delta, requiere el uso de neuronas cuya función de activación sea continua, y por lo tanto, diferenciable. Generalmente, la función utilizada será de tipo sigmoidal. A continuación se resumen los pasos del algoritmo de entrenamiento, los detalles y su demostración puede encontrarse en la referencia original (Rumelhart D., 1986):

1. Se inicializan los pesos de la red con valores aleatorios
2. Se presenta un patrón entrada y se especifica la salida deseada
3. Se calcula la salida actual de la red
4. Se calculan los términos de error para todas las neuronas, mediante la ecuación 3.1

$$E = \frac{1}{2}\sum_{p}\sum_{i}\left(o_{pi} - t_{pi}\right)^2 \quad (3.1)$$

5. Se actualizan los pesos, usando la ecuación 3.2. Los pesos son modificados proporcional al producto del error por la señal de entrada, de esta manera el error decrece en la dirección del gradiente descendente.

$$\Delta w_{ijk} = -\eta \frac{\partial E}{\partial w_{ijk}} = \sum_{p} \frac{\partial E}{\partial o_{pi}} \frac{\partial o_{pi}}{\partial w_{ijk}} \quad (3.2)$$

El tamaño del paso, controla el cambio en las conexiones, ($w_{ij}$ representan la unión entre la unidad *i* y la unidad *j*) está determinado por el parámetro $\eta$ denominado *velocidad de aprendizaje*.

6. El proceso se repite hasta que el término del error alcanza un mínimo con respecto a cada uno de los patrones aprendidos.

El algoritmo de *backpropagation* es muy útil, pero desafortunadamente, no está exento de problemas. Uno de ellos es el relacionado con el mínimo local. Una de las razones por las que en la mayoría de los casos no se puede asegurar la convergencia a la solución óptima es la inicialización aleatoria de los pesos (siendo una condición común en muchos de los algoritmos de redes neuronales). Sin embargo, aún cuando se alcanzara un mínimo global, la red puede no ser útil. Un ejemplo de Boers & Kuiper's dará una idea más clara de lo que pasa (Boers & Kuiper, 1992). Suponemos que queremos entrenar una red para reconocer dígitos escritos a mano. Obviamente, no es posible enseñar a la red todos los posibles patrones, por lo cual se escoge un subconjunto, diremos 100 patrones de '0' y '1'

dibujados por 10 personas distintas, y se entrena la red con ellos. Este conjunto de puede ser aprendido perfectamente por la red, pero esto no dice nada de cómo la red responderá a nuevos patrones. Esta propiedad de ser capaz de responder correctamente a patrones de entrada nuevos se denomina *generalización* y es una característica muy útil en el comportamiento humano. *Backpropagation* generalmente generaliza bien, sin embargo algunas veces falla. Una razón es lo que se conoce como sobreentrenamiento de la red, lo cual ocurre cuando el conjunto de entrenamiento no es suficientemente amplio para cubrir el dominio y la red se entrena por periodos largos. Esto es más frecuente que pase en redes grandes, debido a que en ellas es posible memorizar fácilmente cada uno de los pares de entrada/salida. Una posible solución es usar redes pequeñas (pocas unidades ocultas) de forma que se fuerza a la red a comprimir la información del espacio de entrada, es decir a detectar características de los patrones de entrada mas que reservar una unidad oculta para cada una de ellas. Para hablar del siguiente inconveniente es necesario distinguir entre representación *local y distribuida*. Si al presentar un patrón de entrada activa una única unidad oculta (la activación de esta unidad es significativamente más grande que la activación del resto de las unidades) se dice que el patrón está representado localmente por esa unidad. De otra forma, si al presentar el patrón activa más de una unidad, la activación de todas las unidades da lugar a una representación distribuida del patrón. *Backpropagation* está diseñado para ser un detector de características, definidas por la activación del patrón sobre el conjunto de unidades ocultas y no por una unidad en sí misma, esto puede resultar una desventaja en aplicaciones como el reconocimiento de formas en donde cada patrón es constituido por la unión de características invariantes y en donde es necesario que el sistema neuronal sea capaz de reconocer cada estructura de forma tanto local como distribuida.

El comportamiento de una red depende tanto de su topología (estructura y organización) y de la regla de aprendizaje utilizada (modo en como se actualizan sus pesos). Las redes en las que la estructura no es una restricción pueden sufrir serios problemas. Tales redes asumen una interconectividad total entre todos sus nodos (Hopfield J. 1982) o una estructura multicapa (Rumelhart D. et al., 1986) en la cual cada nodo en una capa está conectado con todos los nodos en las capas posteriores y precedentes a ella. La ventaja de tener una arquitectura inicial con pocas restricciones es lo que se conoce como plasticidad de una red. La plasticidad de una red permite el aprendizaje de cada relación entrada/salida, si se proporcionan suficientes unidades ocultas. Sin embargo, hay una barrera entre la plasticidad de un sistema y su estabilidad. Demasiada plasticidad puede derivar en lo que se conoce como interferencia catastrófica: es decir si una red es entrenada con un conjunto de relaciones entrada/salida y posteriormente se le presenta un nuevo conjunto de relaciones entrada/salida, la matriz de pesos quizá

cambie sustancialmente, lo que se conoce como un lavado de memoria. Hay dos factores que pueden causar interferencias: (i) la superposición y la separación entre conjuntos de patrones presentados de forma secuencial, por ejemplo, si el segundo conjunto de patrones forma un conjunto interpolado del primero, aparece una interferencia si la red es entrenada con el primer conjunto primero y posteriormente con el segundo; (ii) el algoritmo de aprendizaje, *backpropagation* es susceptible de interferencias.

De ahí que se abra una nueva cuestión. ¿Cómo puede un sistema de aprendizaje ser diseñado para permanecer plástico, o adaptativo, en respuesta a eventos significativos y permanecer estable en respuesta a eventos irrelevantes. La plasticidad es necesaria par la incorporación de nuevas representaciones en la red. La estabilidad significa guardar intactas las representaciones antiguas. Una solución a este dilema será encontrar un mecanismo el cual sea capaz de distinguir entre las representaciones viejas y nuevas y pueda usar esta infamación para controlar el proceso de aprendizaje.

### 3.2.2 Cascade correlation

Las unidades ocultas en una red permite reducir el error o la función de costo mejorando la actuación del sistema neuronal. Sin embargo, no es posible saber *a priori* cuantas unidades ocultas son necesarias en cada capa intermedia de la red. Al considerar muy pocas unidades se obtiene una mala generalización, y si por el contrario se consideran muchas se obtiene un sistema que requerirá mayor tiempo de entrenamiento, a este tipo de problema se le llama *efecto caja negra*. Para resolverlo, existen algoritmos que modifican su topología durante el entrenamiento. Algunos de estos algoritmos son constructivos, es decir, comienzan con una red mínima y tanto las unidades ocultas como los pesos se van añadiendo durante el proceso de aprendizaje, un ejemplo bastante utilizado de este tipo de redes es el *cascade correlation*.

Las redes tipo *backpropagation* pueden tener algunos inconvenientes: i) un aprendizaje lento. Al tener que ajustar todos los pesos en las capas ocultas en cada etapa o *epoch* para generar una salida mejor. Si la capa oculta contiene muchas unidades, el sistema se vuelven realmente lento; ii) el objetivo móvil. Si suponemos que tenemos un determinado número de unidades ocultas para resolver dos tareas. Debido a que cada unidad no se comunica una con la otra, ellas tienden a resolver la misma tarea. Si una tarea genera una señal de error mucho mayor que la otra, todas las unidades tenderán a tratar de fijar este problema primero. Una vez que este problema esta resuelto, todas las unidades se concentran en el siguiente problema. Como ninguna neurona oculta esta trabajando sobre el primer problema, este vuelve a aparecer.

La arquitectura de *cascade correlation* fue propuesta por Fahlman y Lebier (Fahlman, S. et al., 1990). Es un algoritmo autoconstructivo cuyo objetivo es generar una red neuronal en donde cada unidad aprende una tarea específica lo más rápido posible, evitando algunos de los problemas inherentes en el algoritmo estándar de *backpropagation*. La redes *cascade correlation* crecen como su nombre lo indica en cascada. Comienzan sin unidades intermedias y se van agregando automáticamente una por una durante el entrenamiento hasta que el error total es suficientemente pequeño. Cada unidad nueva recibe la activación de todas las unidades previamente instaladas. Su propia activación, sin embargo, no pasa a las unidades instaladas previamente, de ahí que resulte una estructura en cascada. Cada unidad en este tipo de redes puede considerarse un nivel intermedio o una capa oculta y es un detector especializado de características en los patrones de entrada. Debido a que las unidades una vez instaladas no pueden modificarse, el entrenamiento sucesivo es acumulativo. Este enfoque elimina la necesidad de elegir el tamaño de la red neuronal de antemano y provee una topología multicapa adecuada para cada problema, Fig.3.7.

La Fig. 3.7 muestra una red entrenada y estructurada usando *cascade correlation*. Se inicia con una unidad de salida y las correspondientes unidades de entrada. El algoritmo comienza sin unidades ocultas y añade una a una. La unidad de salida es entrenada minimizando el error cuadrático. El entrenamiento se detiene cuando el error alcanza un umbral. Si el error cuadrático medio es mayor que un cierto límite deseado, se debe añadir una nueva unidad oculta y re-entrenar la red. El error medio de la red se representa por • y el error por patrón por $E_i$, donde i=1,2,...,p.

Cada unidad escondida se entrena de forma aislada del resto de la red, para maximizar el valor absoluto de la correlación entre $V_i - \bar{V}$ y $E_i - \bar{E}$, donde $V_i$ hace referencia a la unidad de salida para el patrón i y $\bar{V}$ a la salida promedio, respectivamente. La cantidad a maximizarse es:

$$S = \left| \sum_{i=1}^{p} \left(V_i - \hat{V}\right)\left(E_i - \hat{E}\right) \right| \qquad (3.3)$$

El objetivo de este paso es utilizar la unida oculta para determinar el error residual de la red. Usualmente se emplea el algoritmo de *backpropagation* en este paso. Una vez que la unidad oculta ha sido entrenada, es decir cuando la correlación no pude mejorarse, la unidad se añade a la red, como se muestra en la Fig. 3.7b. Los pesos de la unidad oculta se fija (congelan). La unidad de salida recibe la información de las entradas y de la unidad oculta. Todos los pesos de la unidad de salida se re-entrenan hasta que los niveles de error caen en el rango de aceptabilidad. Se prueba si es necesario añadir una nueva unidad. Cualquier

nueva unidad oculta recibirá la señal tanto de las entradas a la red como de cualquier unidad previamente añadida a la misma. El algoritmo continúa añadiendo unidades ocultas hasta que se logra minimizar el error.

La ventaja del algoritmo es la reducción considerable del tiempo de entrenamiento de la red, ya que en cada iteración se entrena una única capa de pesos.

**Fig 3.7 (a)**

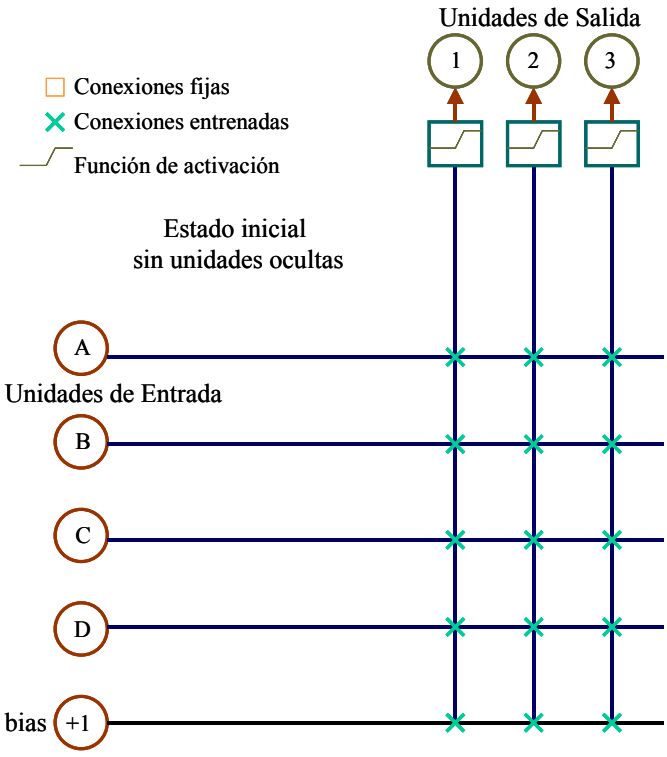
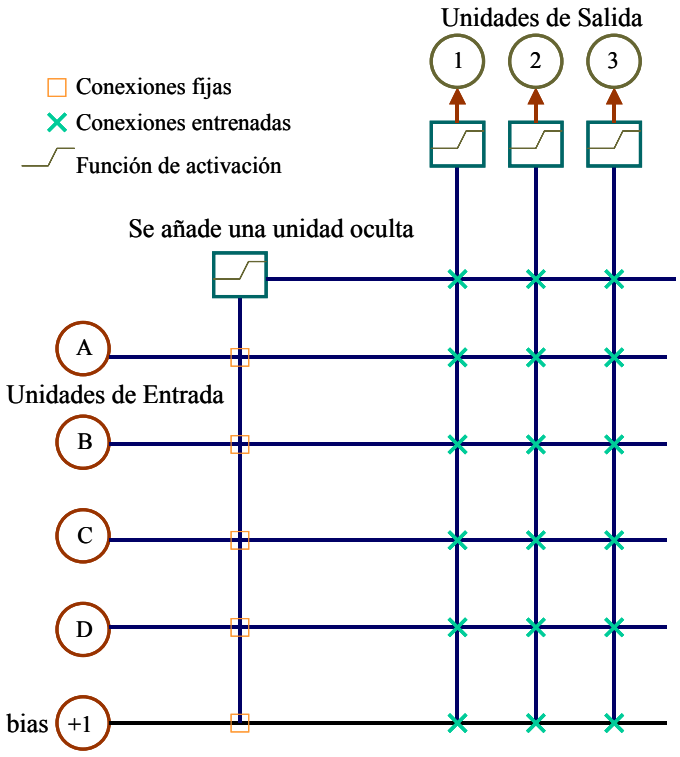



**Fig 3.7 (b)**



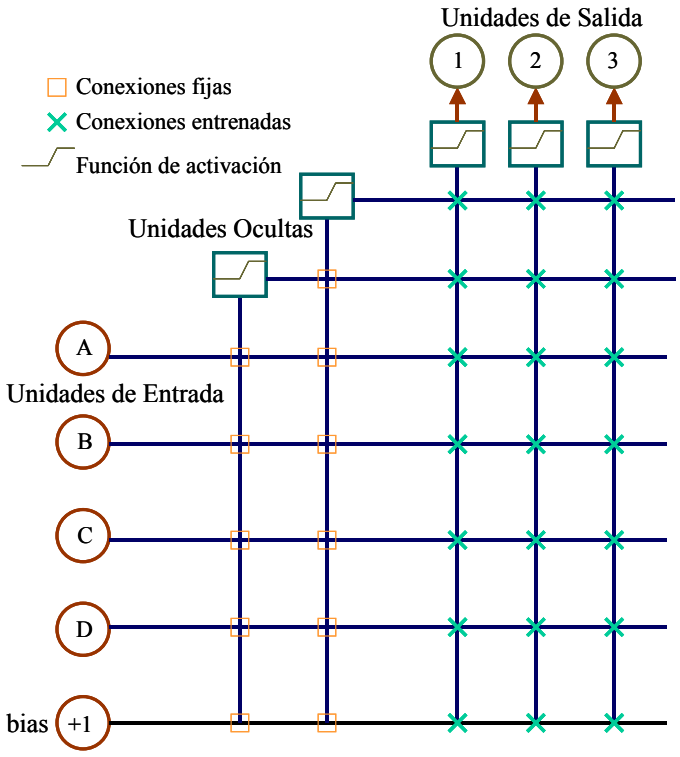




**Fig. 3.7 (c)**

**Fig. 3.7.** Arquitectura de *cascade correlation,* (a) el estado inicial, (b) después de añadir la primera unidad oculta, (c) estado final con dos unidades ocultas. En las líneas verticales se suman la activación de las entradas. Los cuadrados representan a las conexiones fijas, las X a las conexiones que se re-entrenan de forma repetida

## 3.3. Sistemas neuronales de mapas auto-organizados (SOM)

En el cerebro existen neuronas que se organizan en muchas zonas, de manera que la información captada del entorno, a través de los órganos sensoriales se representa internamente en forma de capas prácticamente bidimensionales. Por ejemplo, en el sistema visual se han detectado mapas del espacio visual en la zona del *cortex*. Este sugiere, que el cerebro podría poseer la capacidad inherente de formar mapas topológicos de la información que recibe del exterior.

Kohonen en 1982, presentó un modelo de red neuronal con capacidad de formar mapas o auto organizar la información de manera similar a como ocurre en el cerebro. Su objetivo, era demostrar que un estímulo externo (información de entrada) por sí solo, suponiendo una estructura propia y una descripción funcional del comportamiento de la red, era suficiente para forzar la formación de mapas. Este modelo tiene dos variantes denominadas LVQ (*Learning Vector Quantization*) y TPM (*Topology-Preserving Map*) o SOM (*Self-Organizing Map*). Ambos tienen como objetivo crear mapas topológicos para establecer las características comunes entre los vectores de entrada a la red, aunque difieren en las dimensiones de éstos, de una sola dimensión en el caso de LVQ, y bidimensional, en el caso del SOM.

SOM, establece una correspondencia entre los datos de entrada y un espacio bidimensional de salida, creando mapas topológicos, de tal forma que ante datos de entrada con características comunes se activan neuronas situadas en la vecindad del mapa. Su funcionamiento es relativamente simple. Cuando se presenta el conjunto de variables de entrada a la red, cada una de las M neuronas de la capa de salida la recibe a través de las conexiones *feedforward*. Al igual, estas neuronas reciben también las correspondientes entradas debidas a las conexiones laterales con el resto de las neuronas de salida y cuya influencia dependerá de la distancia a la que se encuentre. Es evidente que se trata de una red de tipo competitivo, ya que al presentar una entrada la red evoluciona hasta una situación estable en la que se activa una neurona de salida, la vencedora.

Lo que hace la red de Kohonen en definitiva, es realizar una tarea de clasificación, ya que la neurona de salida que se activa ante una entrada representa la clase a la que pertenece la información de entrada. Además, como ante otra entrada parecida se activa la misma neurona de salida, u otra cercana a la anterior, debido a la semejanza entre las clases, se garantiza que las neuronas topológicamente próximas sean sensibles a entradas físicamente similares. Es por ello, que este tipo de red es especialmente útil para establecer las relaciones existentes entre conjunto de datos, Fig. 3.8.

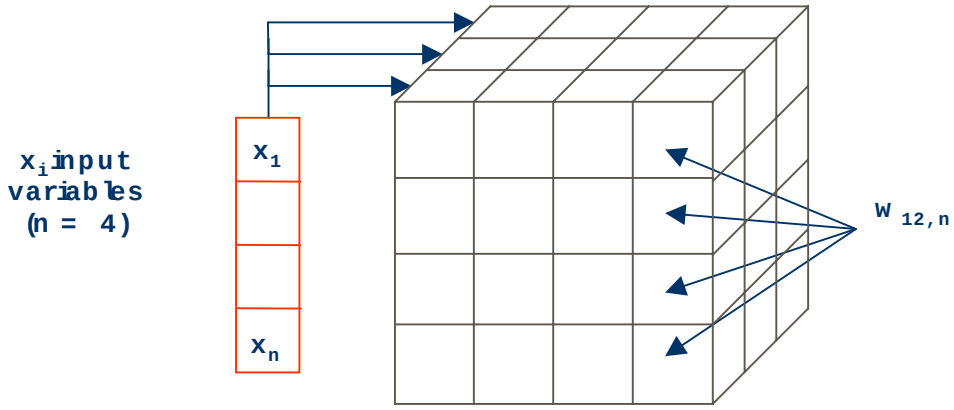



**Fig. 3.8.** Arquitectura de *SOM*

El entrenamiento de SOM es un proceso iterativo, un resumen del mismo se presenta a continuación. El vector **x** se refiere al vector de entrada, ecuación 3.4.

$$\vec{\mathrm{x}} = \left[\mathrm{x}_1, x_2, \ldots, x_p\right]^T \qquad (3.4)$$

El vector de pesos de la neurona j queda representado por:

$$\vec{\mathrm{w}}_j = \left[\mathrm{w}_{\mathrm{j1}}, w_{j2}, \ldots, w_{jp}\right]^T \quad j = 1,2,\ldots N \qquad (3.5)$$

Para encontrar el vector de pesos $\mathbf{w}_j$ que mejor represente a un patrón **x**, simplemente se compara el producto interno $\mathbf{w}_j^T\mathbf{x}$ para j=1,2,...,N y se selecciona aquel de valor mayor. Esto es equivalente a seleccionar la distancia Euclídea mínima entre los dos vectores. Si utilizamos i(**x**) para identificar al nodo que mejor represente al patrón **x**, i(**x**) se determina mediante,

$$\mathrm{i}(\vec{\mathrm{x}}) = \arg\min_{\mathrm{j}} \left\| \vec{x} - \vec{w}_j \right\| \quad j = 1,2,\ldots, N \qquad (3.6)$$

||·|| refiere a la norma euclídea del argumento. El nodo que satisface la ecuación 3.6 se denomina el nodo ganador del vector **x** o best matching unit (BMU). Mediante la ecuación 3.6 el espacio contínuo de las entradas a la red se proyecta sobre un conjunto discreto de nodos. Dependiento de la aplicación de interés la respuesta a la red, puede ser el índice del nodo ganador, es decir su posición en el mapa o el vector peso de la sinápsis que más cerca esté del patrón de entrada en el sentido euclídeo (vector prototipo).

Para llevar a cabo la auto-organiación en necesario que el vector de pesos del nodo j $\mathbf{w}_j$ cambie en relación al vector de entrada **x**.
4. Se actualizan los pesos del nodo *j** y de sus vecinos, sujetos a la condición de vecindad $Nj_*(t)$,

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)\left(x_i(t) - w_{ij}(t)\right) \qquad (3.7)$$

para $j \in Nj_*(t)$ e $1<i<N$. Donde $\eta(t)$ es una función monótonicamente decreciente, define la región de influencia que el vector de entrada tiene sobre SOM. La función $\eta(t)$ está definida por la función de vecindad y la velocidad de aprendizaje $\alpha(t)$ de acuerdo con la ecuación 3.8.

$$\eta(t) = \eta_o\left(\left\|r_c - r\right\|, t\right)\alpha(t) \qquad (3.8)$$

donde r es la posición de las unidades o neuronas sobre el mapa mapa.

La función de vecindad más simple es la función *burbuja (bubble),* la cual es constante sobre todo el radio de vecindad de la neurona ganadora y cero en cualquier otro sitio del mapa. La función de vecindad gausiana definida por la ecuación 3.9,

$$\eta_o = \exp\left(\frac{-\left\|r_c - r\right\|^2}{2\sigma^2(t)}\right) \qquad (3.9)$$

Donde $\sigma(t)$ es el radio que define la vecindad en t, el cual se adapta después de cada *epoch*. El tipo de función de vecindad y el número de neuronas usadas determina la sensitividad y la granularidad del mapa, respectivamente.

La velocidad de aprendizaje $\alpha(t)$ en la ecuación 3.8 es una función decreciente en t en el rango [0,1]. Una función tipo serie de potencias es usada comúnmente, ecuación 3.10.

$$\alpha(t) = \alpha_o\left(\frac{\alpha_T}{\alpha_o}\right)^{t/T} \qquad (3.10)$$

donde $\alpha_o$ y $\alpha_T$ son las velocidades de aprendizaje inicial y final, respectivamente y T es el número de etapas o *epoch* seleccionadas para el entrenamiento. El entrenamiento se lleva a cabo regularmente en dos etapas. La primera etapa de entrenamiento es *burdo*, donde se utilizan valores altos

de $\alpha$, la cual se va reduciendo monótonicamente mediante la ecuación 3.10. Este primer entrenamiento se refina manteniendo un radio constante y disminuyendo el valor de $\alpha$. El entrenamiento de un mapa es de naturaleza estocástica, esto significa que la precisión del mapa depende del número de iteraciones utilizadas para entrenar la red. Sin embargo, los factores limitantes son en sí la velocidad de aprendizaje y la función de vecindad. La velocidad de aprendizaje debe variar con el tiempo. En particular durante las primeras iteraciones se mantiene en un valor cercano a la unidad, que va decreciendo gradualmente hasta alcanzar un valor cercano a cero. La forma en como varía la velocidad de aprendizaje no es crítica, puede ser de forma lineal, exponencial, o inversamente proporcional a t. Es durante la fase inicial conocida como fase de ordenamiento que se obtiene un orden topológico o una distribución inicial, que se irá refinando durante la fase más larga llamada fase de convergencia. Se debe tener especial cuidado en la selección de la función de vecindad, lo más simple es definir una sección cuadrada que incluya a los vecinos que rodean a la neurona ganadora (BMU). Por ejemplo, con un radio de uno, se actualizan los pesos de la BMU y de sus 8 vecinas. Sin embargo, es más adecuado usar funciones gaussianas, o una gaussiana truncada, en este caso el radio de actualización varía durante el entrenamiento. Es en definitiva la función de vecindad lo que hace que SOM no tenga unidades sub-utilizadas aún cuando alguno de sus nodos no se convierta en una unidad ganadora.

## 3.4. Sistemas neuronales cognitivos derivados de ART

Una de las características de la memoria humana consiste en su habilidad para aprender nuevos conceptos sin necesidad por ello, de olvidar otros aprendidos en el pasado. Sería deseable que esta misma capacidad se pudiera conseguir en las redes neuronales. Sin embargo, muchas de estas redes tienden a olvidar informaciones pasadas al tratar de enseñarles otras nuevas.

Cuando se desarrolla una red para realizar una tarea de clasificación de patrones, se suelen reunir un conjunto de ejemplares que serán utilizados durante la fase de aprendizaje o entrenamiento de la red. Durante esta etapa, la información es registrada en el sistema mediante el ajuste de los valores de los pesos de las conexiones entre las neuronas. Una vez concluido el aprendizaje, la red está lista para funcionar y no se permite ningún cambio adicional en los pesos.

Este procedimiento es factible si el problema que se pretende resolver por la red está bien limitado y puede definirse el conjunto de entrada adecuado que permita entrenar a la red para resolver el problema.

Sin embargo, en muchas situaciones reales los problemas a resolver no tienen unos límites claros.

Como ejemplo, imaginemos que se pretende entrenar una red (por ejemplo, *backpropagation*) para reconocer las siluetas de cierto tipo de objeto. Se podrían reunir las imágenes correspondientes y utilizarlas para entrenar a la red. Después, la red entraría en funcionamiento y no se permitiría ningún tipo de modificación en los pesos. Si en un futuro se diseña otro objeto del mismo tipo y se desea que la red reconozca su silueta, se deberá re-entrenar dicha red utilizando la nueva silueta, pero también todas las que se usaron el aprendizaje anterior. Si se entrenase la red sólo con la nueva silueta, podría ocurrir que la red aprendiera esta información, pero olvidará las aprendidas previamente.

Stephan Grossberg denominó a este fenómeno el dilema de estabilidad plasticidad del aprendizaje (Grossber S. et al.,1980). Este dilema plantea las siguientes interrogantes: ¿Cómo puede una red aprender nuevos patrones? (*plasticidad* del aprendizaje). ¿Cómo puede una red retener los patrones previamente aprendidos? (*estabilidad* del aprendizaje). En respuesta a estos dilemas, Grossberg y Carpenter (Carpenter G. et al., 1988) desarrollaron la denominada *teoría de resonancia adaptativa* (*Adaptative Resonance Theory*: ART). Esta teoría se aplica a sistemas competitivos (redes con aprendizaje competitivo) en los cuales cuando se presenta cierta información de entrada sólo una de las neuronas de salida de la red se activa alcanzando su valor de respuesta máximo después de competir con las otras. Esta neurona recibe el nombre de vencedora (*winner-take-all unit*)

La teoría de resonancia adaptativa se basa en la idea de hacer *resonar* la información de entrada con los representantes o prototipos de las categorías que reconoce la red. Si entra en resonancia con alguno, es suficientemente similar, la red considera que pertenece a dicha categoría y únicamente realiza una pequeña adaptación del prototipo almacenado representante de la categoría para que incorpore alguna de las características del dato presentado. Cuando no resuena con ninguno, no se parece a ninguno de los existentes, de los recordados por la red hasta ese momento, la red se encarga de crear una nueva categoría con el dato de entrada como prototipo de la misma.

Lo que se pretende es categorizar los datos que se introducen en la red. La información similar es clasificada formando parte de la misma categoría, y, por tanto, deben activar la misma neurona de salida, la neurona vencedora. Las clases o categorías deben ser creadas por la propia red, puesto que se trata de un aprendizaje no supervisado, a través de las correlaciones entre los datos de entrada.

Para solucionar el dilema de estabilidad plasticidad, el modelo *ART* propone añadir a las redes un mecanismo de realimentación entre las

neuronas competitivas de la capa de salida de la red y de la capa de entrada. Este mecanismo facilita el aprendizaje de nueva información sin destruir la ya almacenada.

La arquitectura denominada *fuzzy ARTMAP* es una síntesis de la lógica difusa y la teoría de resonancia adaptativa (ART) al explorar una cercana similitud formal entre el procesamiento de subconjunto fuzzy, la elección de la categoría ART, la resonancia y el aprendizaje. Fuzzy ARTMAP también realiza una nueva regla de aprendizaje *minimax* que conjuntamente minimiza el error predictivo y maximiza la comprensión del código o generalización. Esto es logrado a través de un proceso de *match-tracking* que incrementa el parámetro de vigilancia de ART en una cantidad mínima necesaria para corregir el error predictivo. Como resultado el sistema aprende automáticamente un número mínimo de categorías, o unidades escondidas, de acuerdo al criterio de precisión permitido.

El aprendizaje es estable debido a que la adaptación de los pesos únicamente puede decrecer en el tiempo. Este decrecimiento de los pesos corresponde a un incremento en el tamaño de las " cajas" correspondientes a cada categoría. Valores pequeños en el parámetro de vigilancia, $\rho$, conllevan a categorías grandes.

El sistema fuzzy ARTMAP incorpora dos módulos ART, artA, y artB, que están unidos vía módulo inter-ART, $F^{ab}$ llamado *map field*. El *map-field* es usado como una asociación entre las módulos artA y artB, mientras que el parámetro de vigilancia de artA, $\rho_a$, se incrementa en respuesta a la mala predicción en artB. El *match-tracking* reorganiza la estructura de la categoría de manera que el error predictivo no vuelva a repetirse en subsecuentes presentaciones de la misma entrada. El *map-field* se activa si una de las categorías artA o artB está activa, Fig.3.9. El entrenamiento puede resumirse de la siguente forma:

1. Los vectores de entrada y salida son preprocesados en forma de (a, $a^c$) y (b,$b^c$) a través de una normalización [0-1] y la creación del código complemento de la variable
2. El vector de entrada se distribuye por $F_1$ hacia todos los nodos existentes en $F_2$, el cual realimenta los pesos de la categoría aprendida hacia $F_1$
3. La clase asignada como la mejor en $F_2$ es enviada hacia $F^{ab}$

3.1 Si la predicción se confirma, la categoría $F_2$ es modificada para que aprenda las características adicionales del nuevo patrón presentado dentro de la categoría ganadora existente.

3.2 Si la predicción de artA se desconfirma, la activación del *map-field* induce una nueva búsqueda, incrementando el parámetro $\rho_a$.

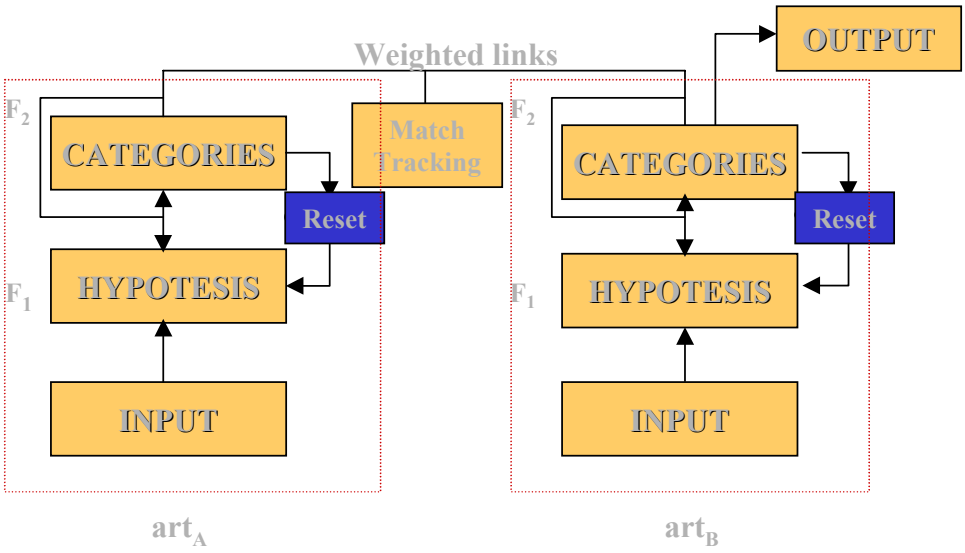



**Fig. 3.9.** Arquitectura de *fuzzy ARTMAP*

Esta nueva búsqueda de artA desencadena la selección de una nueva categoría, y así sucesivamente hasta encontrar aquella que prediga correctamente a artB, o si no existe, se asigna el patrón de entrada a un nodo de artA en $F_2$ que no esté comprometido.

Fuzzy ARTMAP fue diseñado para clasificar datos, con lo cual, no es capaz de genera un patrón de salida después de la fase de entrenamiento. El algoritmo original siempre estaba listo para aprender, así Giralt et al., introdujeron una modificación (Giralt, F. et al., 2001). El módulo artB se desconecta durante la fase predictiva, y las categorías de artA formadas durante la fase de entrenamiento se usan para obtener la respuesta de artB a través del *map-field*. De esta forma, los vectores de entrada se proyectan en la salida. Para cada vector de entrada presentado a la red durante la generalización, únicamente una categoría de artB se activa para proveer la salida deseada.

# 4. RESULTADOS Y DISCUSIÓN

## 4.1. Evaluación de modelos QSPR mediante backpropagation y fuzzy ARTMAP

La capacidad de generalización a nuevos ejemplos de los modelos fuzzy ARTMAP-QSPR se pone de manifiesto con su aplicación a diferentes casos prácticos. Comenzaremos estudiando el punto de ebullición de compuestos orgánicos. Esta propiedad ha sido utilizada en un sin número de referencias como punto de partida en diversas metodologías. Debido principalmente al carácter bidimensional de la propiedad, es decir, el punto de ebullición de cualquier sustancia tiene una relación con la estructura molecular que puede fácilmente codificarse en índices (topológicos, químicos, o topográficos) relacionados con la forma y la conectividad del compuesto. Dichos descriptores, caracterizan la similitud/disimilitud en términos de propiedades empíricas o semi-empíricas de las sustancias. La similitud intermolecular puede definirse en términos de características estructurales y de la correspondencia mutua entre dos especies químicas, como se ha demostrado en diversos trabajos citados en la literatura (Balaban A. et al., 1984; Bünz P. et al., 1998; Elgolf L. et al., 1993; Espinosa G. et al., 2001; Espinosa G. et al., 2002; Katritzky R. et a., 1996; Tetteh J. et al., 1999).

No debemos perder de vista que estamos tratando de solucionar una doble cuestión, i) los descriptores que mejor capten el carácter de cada propiedad (bidimensional o tridimensional), y ii) el modelo de red neuronal capaz de captar las no linealidades de la relación entre la estructura y las propiedades del compuesto. Partiendo de estas premisas, se comienza por un lado, con una propiedad simple y ampliamente estudiada como lo es el punto de ebullición, y por otro lado con uno de los algoritmos de redes neuronales más utilizados como lo es el backpropagation. Dentro de estos dos puntos se estudiará la calidad de los descriptores moleculares como variables de entrada al modelo, por ejemplo, los índices de conectividad molecular en la caracterización de conjuntos homogéneos y heterogéneos de compuestos orgánicos.

Los casos se presentaran a continuación cronológicamente, como se fueron desarrollando, sin embargo, alguno de ellos fue publicado posteriormente. Es importante conservar este orden ya que es a través de la experiencia que se fue estableciendo una metodología más detallada, cuya evolución se evidenciará en cada caso.

En el primer caso se usan dos conjuntos de descriptores moleculares, formados por cuatro índices de conectividad molecular y los

correspondientes índices de valencia para estimar el punto de ebullición de un conjunto de 1116 compuestos orgánicos diversos. En el segundo caso, se busca un modelo para un conjunto homogéneo de 327 hidrocarburos alifáticos, incrementando el número de descriptores moleculares y comparando la capacidad de generalización entre el backpropagation y el una red de fuzzy ARTMAP.

### 4.1.1 Punto de ebullición

#### 4.1.1.1 Backpropagation vs cascade correlation

Tres conjuntos de compuestos se utilizan para evaluar el punto de ebullición de compuestos orgánicos. Los datos experimentales fueron tomados de la base de datos DIPPR (DIPPR, 1997). Además, dos subconjuntos de compuestos uno de alcoholes y el otro de alcanos referenciados por Kier y Hall y Hall y Story (Hall L. et al., 1995; Hall L. et al., 1996), respectivamente, se usan para validar la calidad de los modelos obtenidos. El conjunto global es un conjunto muy heterogéneo formado por hidrocarburos tanto saturados como insaturados, compuestos aromáticos y halogenados, grupos ciano, amino, éster, éter, carbonilos, hidroxilos y carboxilos. Las estructuras y los índices de conectividad de orden uno a cuatro ($^{0\text{-}4}\chi$) y los correspondientes índices de valencia se obtienen con el software comercial Molecular Modeling Pro, ver 3.01.

El primer conjunto de compuestos está formado por 242 compuestos entre alcoholes con 1 a 10 átomos de carbono y alcanos con 1 a 5 átomos de carbono. El rango del punto de ebullición estudiado se encuentra entre 9.5 °C y 231 °C. 42 de los compuestos fueron seleccionados para evaluar la generalización del modelo. Los resultados obtenidos con una red de tipo backpropagation muestran que una expansión en la capa intermedia de la red da mejores resultados que una contracción en el número de neuronas de la capa de entrada a la red. Esto muestra que al aumentar la dimensionalidad es posible representar mejor las características del conjunto de entrenamiento, esta dimensionalidad adicional contribuye favorablemente a la generalización de la red. La mejor configuración de la red es 8-12-1 (cuatro índices de conectividad más sus correspondientes, doce neuronas en la capa intermedia y una sola capa intermedia. El error absoluto medio es de 4.3 °C (s= 3.3 °C) que corresponde a un 2.9%. Para el mismo conjunto de índices, un análisis de regresión multilínea, MLR da un error absoluto medio de 10.3 °C (s=5.5 °C) que equivale a un 7.7% de error relativo medio. Unicamente dos compuestos, el heptano y el nonanol, presentan un error residual mayor a 10 °C. Comparando estos resultados con el trabajo previamente publicado de Hall y Kier (Hall, L. et al., 1995) y a pesar de que una comparación formal no puede establecerse debido a que el tipo de descriptores utilizados no es el mismo, el error citado en dicho trabajo es de 5.9 °C, es decir un 4.1% de error relativo medio. La representación gráfica de los resultados se encuentra en la Fig. 4.1. Los datos

presentados en la Fig. 4.1 muestran un buen ajuste, el coeficiente de correlación es, $r^2$= 0.983.

El segundo conjunto le integran 220 compuestos orgánicos entre hidrocarburos saturados e insaturados, ésteres, éteres, grupos carbonilos, carboxilos e hidroxilos (entre 3 y 19 átomos de carbono). El rango del punto de ebullición estriba entre –47.64 °C y 334.85 °C. 30 de estos compuestos fueron reservados para la fase de prueba de la red. La mejor configuración obtenida para el backpropagation coincide con la anterior, 8-12-1. El error absoluto medio es de 21.80 °C (s=11.3 °C) equivalente a un 5.38% de error relativo, comparado con el trabajo publicado previamente por Hall y Story (Hall, L. et al., 1996) para una red backpropagation 19-5-1 con un error medio de 4.57 °C (1.12%), utilizando 19 índices electrotopológicos como entrada a la red. Esta diferencia tan notable es debida al tipo de índices utilizados para obtener el modelo, los cuales les permite una caracterización de todos los grupos funcionales existentes, en la Fig. 4.2, se representan los resultados obtenidos en contraste con los experimentales. El coeficiente de correlación en este caso $r^2$=0.8. Este coeficiente de correlación tan bajo nos indica que al incrementar la diversidad de los compuestos sin incrementar la información capaz de caracterizar los distintos grupos funcionales, decrece la capacidad de extrapolación de las redes neuronales basada sen unidades de procesamiento.

Partiendo de los resultados anteriores, podemos concluir que los índices de conectividad molecular, son una herramienta muy útil para caracterizar conjuntos homogéneos de compuestos químicos, pero no contienen la información suficiente para distinguir entre las diferentes clases de compuestos cuando hay más de un tipo de grupo funcional, o cuando están presentes isómeros estructurales dentro del conjunto estudiado. Además, la baja correlación obtenida en el segundo caso, plantea la necesidad de tener un conjunto de compuestos más completo, es decir que cada familia este perfectamente bien representada y se cuente con el mayor número de elementos dentro de la misma, con lo cual se evitaría que algún elemento del conjunto de prueba estuviera fuera del rango del conjunto con el que se formuló el modelo.

Con la intención de resolver las cuestiones planteadas, se utiliza un conjunto suficientemente amplio formado por 1116 compuestos diversos. El rango del punto de ebullición comprende desde los 111.7 K hasta los 715 K. Alrededor del 60% de los compuestos fueron usados para el entrenamiento de la red. La mejor configuración coincide una vez más, 8-12-1. El error absoluto medio es de 28 K, que refieren un 7% de error relativo medio, los resultados pueden visualizarse en la Fig. 4.3. Los compuestos con mayores residuales comúnmente llamados *outliers* corresponden a los polifluorados, aromáticos substituidos y compuestos con nitrógeno, N, como substituyente. La selección aleatoria del conjunto de entrenamiento y prueba, puede dar

lugar a errores importante a la hora de evaluar los resultados obtenidos. Es por ello que se utilizó una algoritmo publicado en la literatura (Tamburini F. et al., 1994) aplicado a problemas de clasificación de imágenes. El cual escoge el mínimo número de patrones que contengan la máxima cantidad de información acerca del conjunto. La mejor configuración en este caso es 6-12-1 con un error absoluto medio de 11.6 K (s=11. 3 K) que corresponde a un error relativo medio de 4.33%, Fig. 4.4. Finalmente, se utiliza una red del tipo cascade correlation, en la cual se probaron nodos con diferentes tipos de funciones de activación, sin embargo, en ningún caso se obtienen errores menores a los presentados a partir de backpropagation. El error absoluto medio es de 33 K comparado con el 11.6 K obtenido con backpropagation.

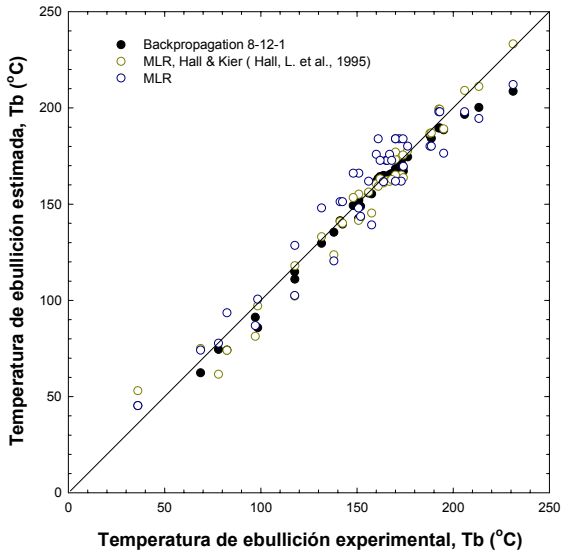



**Fig. 4.1.** Comparación de modelos para el punto de ebullición de alcoholes y alcanos usando a) una arquitectura backpropagation 8-12-1, b) MLR de Hall & Kier (Hall, L. et al., 1995) y c) MLR usando índices de conectividad

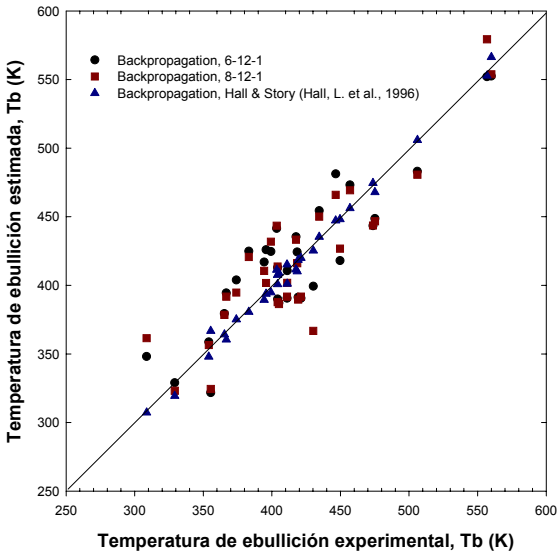



**Fig. 4.2.** Comparación de modelos para el punto de ebullición de compuestos orgánicos diversos con diversas arquitecturas backpropagation a) 6-12-1, b) 8-12-1 y c) 19-5-1 de Hall & Story (Hall, L. et al., 1996) basándose en índices electrotopológicos

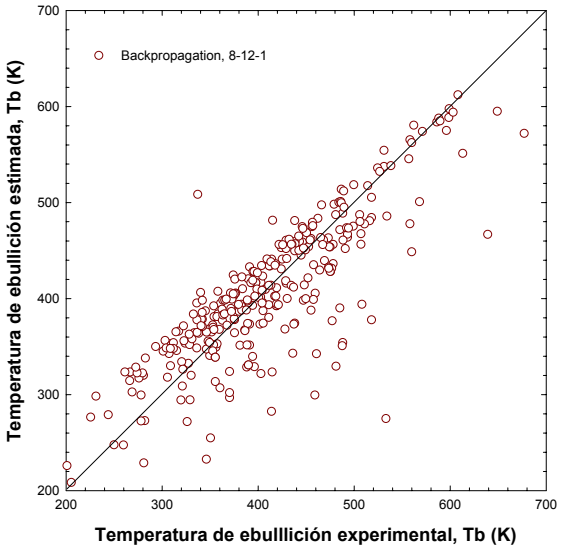



**Fig. 4.3.** QSPR para el punto de ebullición de compuestos orgánicos diversos con una arquitectura backpropagation 8-12-1

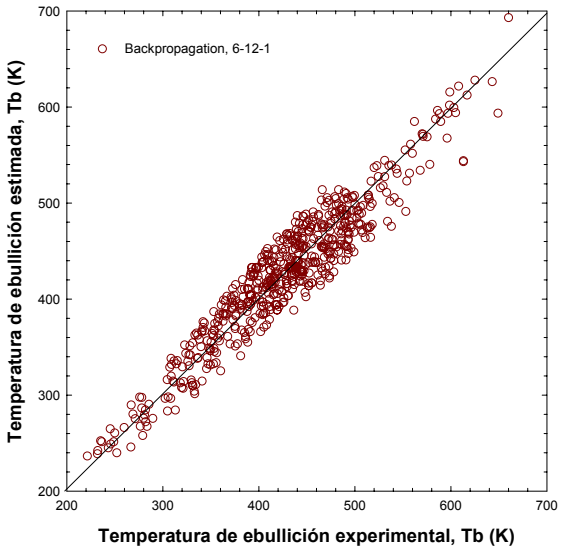



**Fig. 4.4.** QSPR para el punto de ebullición de compuestos orgánicos diversos con una arquitectura backpropagation 6-12-1

#### 4.1.1.2 Backpropagation vs fuzzy ARTMAP

En está sección el objetivo global es evaluar modelos QSPR para el punto de ebullición de hidrocarburos alifáticos a partir de dos arquitecturas neuronales: backpropagation y fuzzy ARTMAP modificado y un conjunto simple de descriptores moleculares. Los modelos de QSPRs se obtienen a partir de cuatro índices de conectividad molecular de valencia ($^{1}\chi^{v}$, $^{2}\chi^{v}$, $^{3}\chi^{v}$, $^{4}\chi^{v}$), el índice de forma kappa de segundo orden ($^{2}\kappa$), el momento dipolar ($\mu$) y el peso molecular ($M_{w}$). El momento dipolar ($\mu$), resulta ser de gran utilidad en la distinción de isómeros estructurales cis y trans. La metodología seguida en el presente caso se encuentra representada en la Fig. 4.5. La Tabla 1 (Tabla 1, apéndice 1; Espinosa et al., 2000) del mismo apéndice contiene los descriptores moleculares utilizados como parámetros de entrada, así como el punto de ebullición experimental de cada uno de los compuestos. Los modelos QSPR individuales para los 140 alcanos, los 144 alquenos y los 43 alquinos, así como un modelo global se calculan para el conjunto de hidrocarburos alifáticos lineales y ramificados, con compuestos que contienen entre uno y diez átomos de carbonos. En las Tablas 1-3 (Tabla 1-3, apéndice 1; Espinosa et al., 2000) se encontrarán los puntos de ebullición experimentales y los descriptores moleculares correspondientes a cada conjunto.

Los datos experimentales y los descriptores moleculares se dividieron en tres conjuntos: el de entrenamiento, el de prueba y el de validación para los modelos backpropagation. Siendo estos conjuntos distintos en cada modelo. En el modelo QSPR de los alcanos 92, 28 y 20 compuestos se usan

durante el entrenamiento, la validación y la generalización, respectivamente. En el caso de los alquenos, cada uno de los tres conjuntos de compuestos está formado por 97, 26 y 21, siguiendo el orden previamente establecido. Siendo 43 únicamente los alquinos de los que se disponía de datos experimentales, se les consideró un conjunto muy pequeño para crear un modelo independiente QSPR. Sin embargo, el modelo que integra los 327 compuestos, quedó dividido en 228 compuestos para la fase de entrenamiento, 67 usados durante la validación y 32 para evaluar la generalización del modelo. En el conjunto de entrenamiento 97 eran alcanos, 101 alquenos y 30 alquinos, en el de validación 15 eran alcanos, 13 alquenos y 4 alquinos, para quedar el conjunto de prueba constituido por 28 alcanos, 30 alquenos y 9 alquinos. Esta división se realiza de forma aleatoria después de normalizar los datos entre 0 y 1 usando el software NeuralSim.

El desempeño general de los modelos basados en las arquitecturas de backpropagation y fuzzy ARTMAP se resumen en la Tabla 4.1. Los modelos QSPRs para alcanos basados en fuzzy ARTMAP y backpropagation (7-4-1), respectivamente, se presentan en las Figs. 4.6 y 4.7, Tablas 5 y 6 (Tablas 5 y 6, apéndice 1; Espinosa et al., 2000). La arquitectura óptima para el backpropagation fue 7-4-1, es decir siete neuronas en la capa de entrada, cuatro en la capa intermedia y el punto de ebullición representado en la neurona de salida. Dicha arquitectura predice el punto de ebullición con una precisión del 99.6%, con un error máximo de 2.51%(10.5 K). El error promedio para todo el conjunto de datos es de 0.37% (1.54 K). Fuzzy ARTMAP predice el punto de fusión de los alcanos con una precisión un poco mayor 99.7%, con un error máximo del 0.88% (3.36 K). Cabe enfatizar que los errores experimentales típicos son de alrededor el 1%, por lo que los resultados obtenidos caen dentro de este rango. Una comparación de los resultados anteriores y los modelos QSPR con redes neuronales obtenidos por Ivanciuc (Ivanciuc O. 1998) y Gakh et al., (Gakh A. et al., 1994) se muestra en la Tabla 6 (Tabla 6,apéndice 1; Espinosa et al., 2000). Ivanciuc propuso modelos QSPR con redes neuronales basados en el modelo de red MolNet (Ivanciuc, O. 1997) y dos tipos de descriptores topológicos: *degree (DEG),* el cual se basa en la matriz de adyacencia entre los átomos de una estructura molecular y el recíproco de la suma de las distancias entre los átomos, RDS. Los errores de generalización en modelos del punto de ebullición con los índices DEG y RDS, para 25 alcanos usados por Ivanciuc son de 0.74% (3.0 K) y 0.42% (1.71 K), respectivamente. El modelo compuesto propuesto por Gakh et al., (Gakh A. et al., 1994) predice simultáneamente seis propiedades fisicoquímicas usando una arquitectura formada por siete entradas, ocho unidades ocultas y seis salidas. Para el mismo conjunto de 25 alcanos, el modelo de Gakh et al., (Gakh A. et al., 1994) tiene un error absoluto medio de 1.19%.

En este caso el modelo para alcanos generado con una backpropagation 7-4-1 y con fuzzy ARTMAP dan lugar a errores absolutos medios del orden de

0.40% (1.65 K) y 0.30% (1.3 K), respectivamente. Ambos resultados son comparables o mejoran los modelos con los cuales los hemos comparado. Al comparar las predicciones de los modelos de alcanos, QSPR con backpropagation y fuzzy ARTMAP con los generados con DEG-MolNet, RDS-MolNet y los modelos de Gakh et al., (Gakh A. et al., 1994) para los 125 compuestos usados en ambos estudios, los errores absolutos encontrados son de 0.38% (1.59 K), 0.31% (1.29 K), 0.65% (2.66 K), 0.43% (1.75 K) y 11.03% (2.63 K), respectivamente. Esta comparación sugiere que modelos QSPR para propiedades individuales, en este caso para el punto de ebullición de alcanos son claramente más precisos que los modelos QSPR generados para estimar simultáneamente varias propiedades.

Los resultados generados por los modelos QSPR con backpropagation y fuzzy ARTMAP se encuentran en la Tabla 7 (Tabla 7, apéndice 1; Espinosa et al., 2000). En las Figs. 4.8 y 4.9 se comparan los valores estimados para el conjunto de alquenos con sus valores experimentales para cada modelo de red. El modelo QSPR con una arquitectura backpropagation 7-10-1, es capaz de distinguir entre isómeros estructurales lo cual sugiere que el conjunto seleccionado es adecuado para estimar el punto de ebullición de hidrocarburos insaturados. El error absoluto medio global es de 1.25% (4.4 K, s= 1.13%), siendo el error máximo de 7.11% (19.57 K). El error medio para el conjunto de validación es de 1.83% (6.45 K, s=1.21%), con un error máximo de 4.39% (18.18 K). La predicción fue mejorada significativamente usando fuzzy ARTMAP para construir el modelo QSPR. El error absoluto medio para el conjunto de prueba, así como el global es de 0.19% (0.73 K) y 0.25% (0.95 K), respectivamente. Los resultados derivados del modelo QSPR con backpropagation se compararon con el estudio para 85 alquenos de Zhang et al., (Zhang R. et al., 1997) el cual obtiene un error absoluto medio del 2.3% (2 K) siendo el máximo de 10% (5.7 K). Los resultados anteriores reflejan con claridad que el presente modelo para alquenos usando una arquitectura backpropagation la cual genera un error global de 1.25% (4.42 K) y con la habilidad de distinguir cuantitativamente entre diasterómeros es aceptable dando errores similares al error experimental.

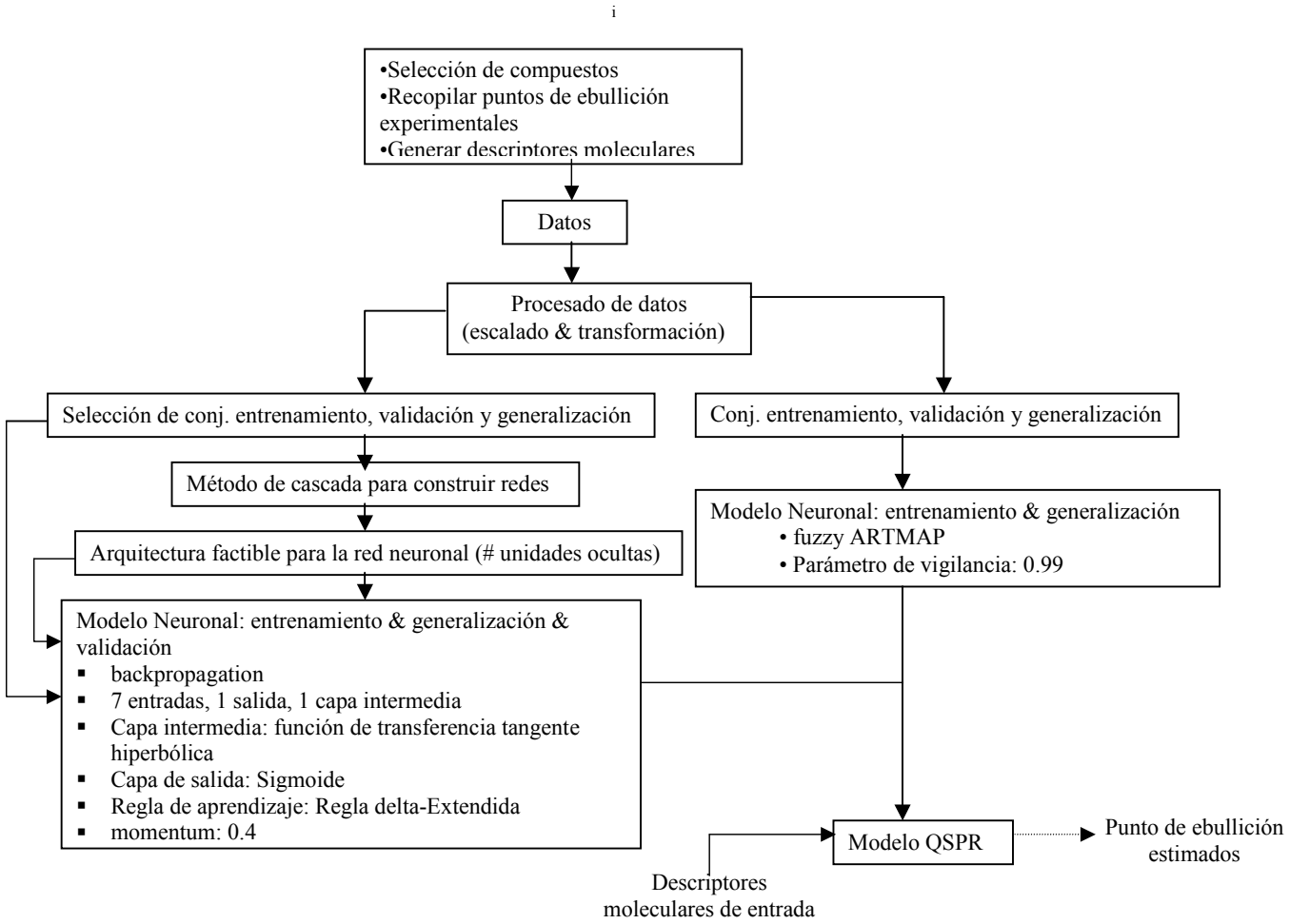
La relación cuantitativa QSPR para el conjunto general de hidrocarburos alifáticos se muestra en las Figs. 4.10, 4.11, 4.12 y en las Tablas 4.2-4.5 (apéndice 1, Espinosa et al., 2000) respectivamente, con un resumen de los errores en la Tabla 4.1. En el modelo global para los hidrocarburos alifáticos con una backpropagation de estructura 7-9-1 es: 2.86 K (0.75%), 6.16 K (1.77%) y 6.85 K (2.04%) para alcanos, alquenos y alquinos, respectivamente, así como un error global de 4.85 K (1.37%). El error máximo es de 9.75% (20.92 K). El error medio para cada uno de los conjuntos de entrenamiento, prueba y validación son 1.28% (4.51 K), 1.74% (6.1 K), 1.25% (4.68 K), respectivamente, siendo la desviación estándar en cada caso de 1.41% (4.2 K), 1.74% (5.33 K) y 1.0% (3.68 K), respectivamente. Cabe destacar que el error absoluto medio y el error

máximo asociado con los 30 alquinos del conjunto de entrenamiento es de 1.79% (6.08 K) y 9.74% (18.33 K), respectivamente, con una desviación estándar de 1.65% (3.69 K). A pesar de que el error para el conjunto de alquinos es relativamente alto, los alquinos constituyen únicamente el 13% del conjunto de entrenamiento. Sin embargo contribuyen en un 17.7% al error medio de entrenamiento. Los alcanos representan al 44.5% del conjunto de entrenamiento y contribuyen en un 57.9% a su error medio.

Una mejora sustancial en las predicciones se obtiene usando fuzzy ARTMAP, el error absoluto medio es de 1.35 K (0.35%), aunque cabe decir que la diferenciación entre isómeros es inconsistente.

Los resultados obtenidos en el este caso, nos llevan a las siguientes conclusiones, los modelos fuzzy ARTMAP-QSPR generados a partir de un conjunto simple de índices, relativamente fáciles de obtener, dan lugar a estimaciones precisas. Por otro lado, hay que mencionar que las relaciones QSPR usando backpropagation fueron capaces de discernir entre isómeros estructurales y en cambio aquellas basadas en fuzzy ARTMAP presentaron algunas inconsistencias al respecto. Sin embargo, el modelo con fuzzy ARTMAP mejora substancialmente respecto a los modelos obtenidos con otras técnicas. Así pues, en este punto se plantea el expandir este trabajo a diversas propiedades fisicoquímicas, no perdiendo de vista la necesidad de probar diferentes descriptores moleculares, en cuanto las interacciones presentes no pueda modelarse únicamente con la forma de la molécula como en el caso del punto de ebullición.

**Fig. 4.5.** Diagrama de flujo para los modelos QSPR basado en redes neuronales



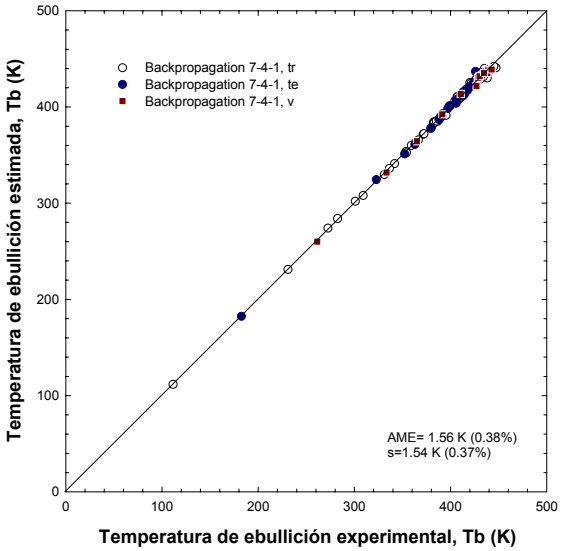



**Fig. 4.6.** QSPR para el punto de ebullición de alcanos usando una arquitectura backpropagation 7-4-1

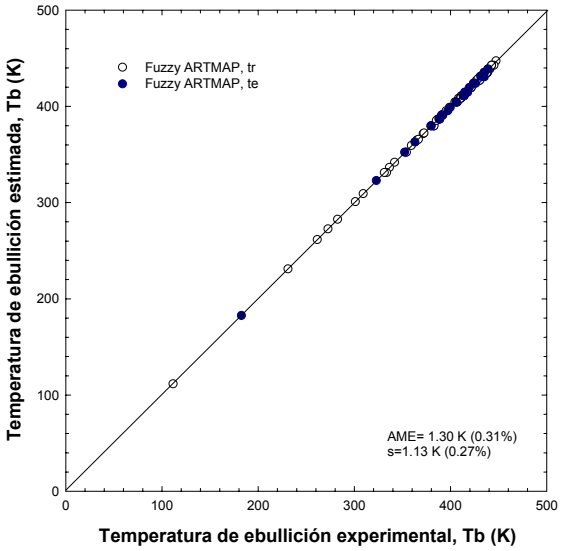



**Fig. 4.7.** QSPR para el punto de ebullición de alcanos usando una red neuronal fuzzy ARTMAP

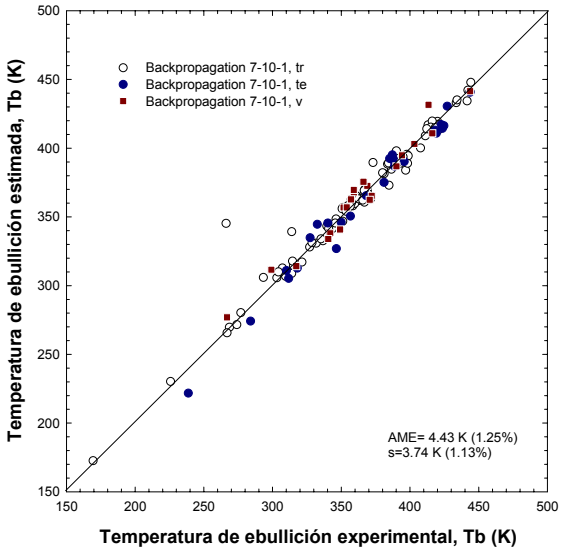



**Fig. 4.8 QSPR** para el punto de ebullición de alquenos usando una arquitectura 7-10-1 backpropagation

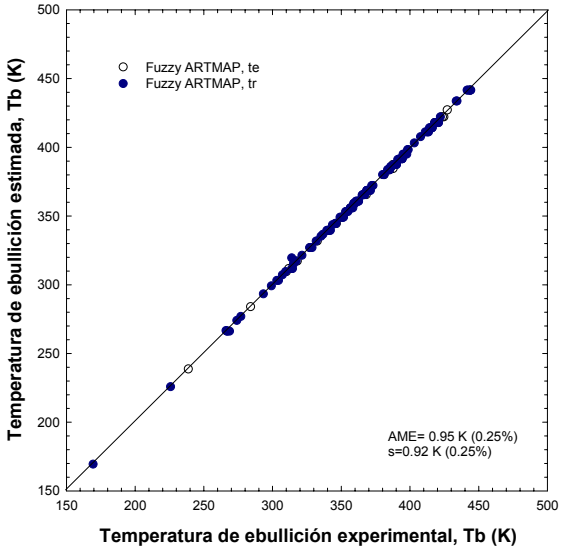



**Fig. 4.9.** QSPR para el punto de ebullición de alquenos usando una red neuronal fuzzy ARTMAP

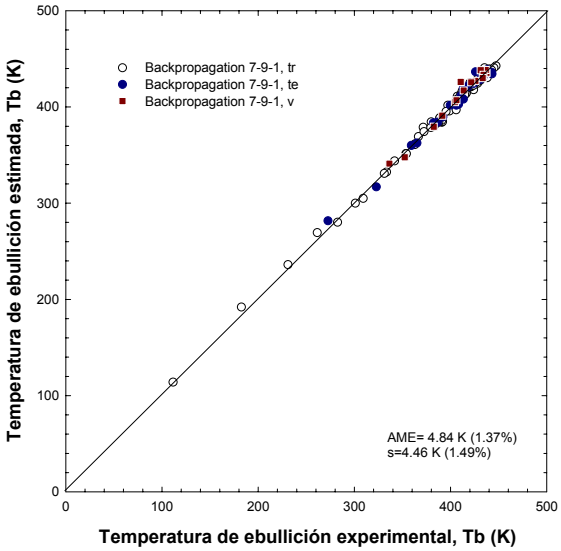



**Fig. 4.10** QSPR para el punto de ebullición de 327 hidrocarburos alifáticos usando una arquitectura 7-9-1 backpropagation

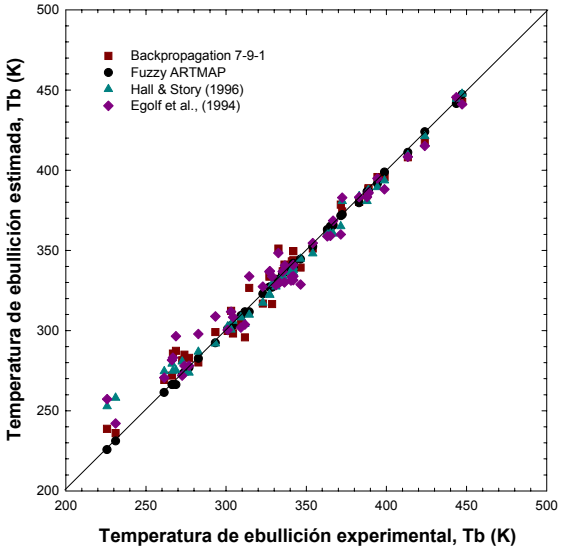



**Fig. 4.11** Comparación de diferentes modelos QSPR para el punto de ebullición un subconjunto de alcanos y alquenos

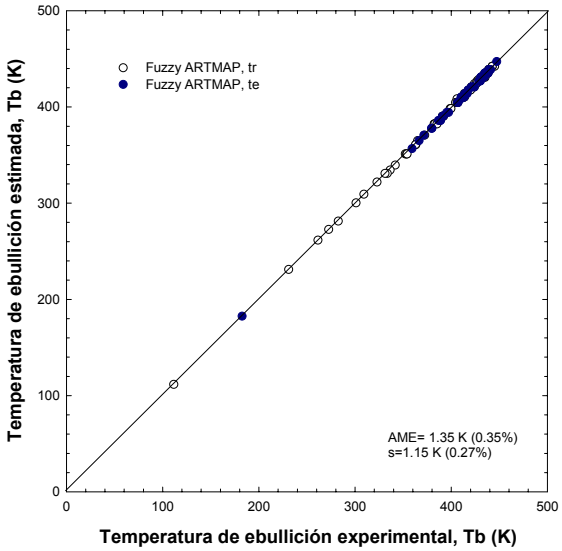



**Fig. 4.12.** QSPR para el punto de ebullición de 327 hidrocarburos alifáticos usando una red neuronal fuzzy ARTMAP

**Tabla 4.1.** Desempeño de los modelos QSPR/NN para predecir el punto de ebullición de hidrocarburos alifáticos (327 compuestos) usando backpropagation y fuzzy ARTMAP

| Conjunto | # de comp.. | Contribución al error absoluto total | | Desviación estándar | | Error absoluto medio | | Error máximo | |
|---|---|---|---|---|---|---|---|---|---|
| | | K | Porcentaje, % | K | Porcentaje, % | K | Porcentaje, % | K | Porcentaje, % |
| Modelo para 140 alcanos con Backpropagation 7-4-1 | | | | | | | | | |
| Global | 140 | | | **1.53** | **0.35** | **1.54** | **0.37** | **10.71** | **2.51** |
| Tr | 92 | | | 1.44 | 0.33 | 1.47 | 0.56 | 8.17 | 1.86 |
| Te | 28 | | | 1.96 | 0.46 | 1.65 | 0.40 | 10.71 | 2.51 |
| V | 20 | | | 1.24 | 0.28 | 1.73 | 0.42 | 5.01 | 1.17 |
| Modelo para 140 alcanos con Fuzzy ARTMAP | | | | | | | | | |
| Global | 140 | | | **1.13** | **0.27** | **1.30** | **0.31** | **3.36** | **0.88** |
| Tr | 112 | | | 1.10 | 0.27 | 1.39 | 0.31 | 3.36 | 0.88 |
| Te | 28 | | | 1.26 | 0.27 | 0.81 | 0.19 | 3.3 | 0.77 |
| Modelo para 144 alquenos con Backpropagation: 7-10-1 | | | | | | | | | |
| Global | 144 | | | **3.76** | **1.14** | **4.42** | **1.25** | **19.57** | **7.11** |
| Tr | 97 | | | 3.00 | 8.84 | 3.34 | 0.93 | 16.33 | 4.37 |
| Te | 26 | | | 4.17 | 1.50 | 6.79 | 1.96 | 19.57 | 7.11 |
| V | 21 | | | 4.18 | 1.21 | 6.45 | 1.83 | 18.18 | 4.39 |
| Modelo para 144 alquenos con Fuzzy ARTMAP | | | | | | | | | |
| Global | 144 | | | **0.92** | **0.25** | **0.95** | **0.25** | **2.73** | **0.91** |
| Tr | 116 | | | 0.91 | 0.25 | 0.99 | 0.26 | 2.73 | 0.91 |
| Te | 26 | | | 0.94 | 0.24 | 0.73 | 0.19 | 2.71 | 0.71 |
| Modelo para 327 hidrocarburos alifáticos con Backpropagation: 7-9-1 | | | | | | | | | |
| Global | 327 | total = 4.85 K | total = 1.37% | **4.45** | **1.46** | **4.85** | **1.37** | **20.92** | **9.75** |
| Alcanos | 140 | 1.22 | 0.32 | 2.46 | 0.76 | 2.86 | 0.75 | 15.5 | 5.18 |
| Alquenos | 144 | 2.72 | 0.78 | 5.06 | 1.59 | 6.16 | 1.77 | 20.92 | 7.05 |
| Alquinos | 43 | 0.89 | 0.27 | 4.76 | 1.89 | 6.85 | 2.04 | 18.99 | 9.74 |
| Tr | 228 | total = 4.51 K | total = 1.28% | 4.20 | 1.41 | 4.51 | 1.28 | 20.92 | 9.74 |
| Alcanos | 97 | 1.10 | 0.30 | 2.05 | 0.72 | 2.60 | 0.71 | 9.45 | 5.18 |
| Alquenos | 101 | 2.61 | 0.75 | 5.06 | 1.61 | 5.85 | 1.68 | 20.92 | 7.05 |
| Alquinos | 30 | 0.80 | 0.23 | 3.69 | 1.65 | 6.08 | 1.79 | 18.33 | 9.74 |
| Te | 67 | total = 6.09 K | total = 1.74 % | 5.33 | 1.74 | 6.09 | 1.74 | 19.99 | 7.44 |

| Conjunto | # de comp.. | Contribución al error absoluto total | | Desviación estándar | | Error absoluto medio | | Error máximo | |
|---|---|---|---|---|---|---|---|---|---|
| | | K | Porcentaje, % | K | Porcentaje, % | K | Porcentaje, % | K | Porcentaje, % |
| Alcanos | 28 | 1.43 | 0.36 | 2.76 | 0.77 | 3.41 | 0.87 | 10.35 | 3.25 |
| Alquenos | 30 | 3.45 | 0.99 | 5.40 | 1.67 | 7.72 | 2.21 | 19.99 | 6.95 |
| Alquinos | 9 | 1.21 | 0.38 | 7.14 | 2.60 | 9.00 | 2.88 | 18.89 | 7.44 |
| V | 32 | total = 4.68 K | total = 1.25% | 3.68 | 1.00 | 4.68 | 1.25 | 15.54 | 3.78 |
| Alcanos | 15 | 1.67 | 0.41 | 3.68 | 0.90 | 3.56 | 0.88 | 15.54 | 3.78 |
| Alquenos | 13 | 2.03 | 0.59 | 3.16 | 0.94 | 5.00 | 1.45 | 10.27 | 3.03 |
| Alquinos | 4 | 0.98 | 0.25 | 3.24 | 0.97 | 7.82 | 2.02 | 12.36 | 3.45 |
| Modelo para 327 hidrocarburos alifáticos Fuzzy ARTMAP | | | | | | | | | |
| All data | 327 | total =1.35 K | total = 0.35 % | **1.15** | **0.27** | **1.35** | **0.35** | **3.45** | **0.98** |
| alkanes | 140 | 0.54 | 0.15 | 1.06 | 0.29 | 1.25 | 0.34 | 3.42 | 0.98 |
| alkenes | 144 | 0.66 | 0.16 | 1.10 | 0.26 | 1.51 | 0.37 | 3.45 | 0.83 |
| alkynes | 43 | 0.16 | 0.04 | 1.16 | 0.31 | 1.20 | 0.32 | 3.41 | 0.94 |
| Training | 228 | total =1.44 K | total = 0.35 % | 1.09 | 0.27 | 1.44 | 0.35 | 3.45 | 0.92 |
| alkanes | 112 | 0.66 | 0.17 | 1.03 | 0.27 | 1.34 | 0.35 | 3.42 | 0.98 |
| alkenes | 114 | 0.78 | 0.19 | 1.11 | 0.26 | 1.56 | 0.37 | 3.45 | 0.83 |
| alkynes | 34 | .20 | 0.05 | 1.19 | 0.31 | 1.33 | 0.35 | 3.41 | 0.94 |
| Testing | 67 | total =1.15 K | total = 0.84 % | 1.08 | 0.31 | 1.15 | 0.84 | 3.42 | 0.98 |
| alkanes | 28 | 0.39 | 0.13 | 1.10 | 0.35 | 0.94 | 0.30 | 2.99 | 0.98 |
| alkenes | 30 | 0.56 | 0.16 | 1.02 | 0.29 | 1.24 | 0.36 | 2.98 | 0.83 |
| alkynes | 9 | 0.08 | 0.02 | 0.8 | 0.24 | 0.56 | 0.17 | 2.2 | 0.67 |

[a] tr = conjunto de entrenamiento, te = conjunto de generalización, v = conjunto de validación;[b] Reid et al. (1977), DIPPR (1996) and Properties of Organic Compounds, CRC Press, Inc. (1996)

Pero antes de modelar otras propiedades fisicoquímicas o biológicas, sería conveniente utilizar la herramienta neuronal que da los mejores resultados para modelar el punto de ebullición de un conjunto suficientemente heterogéneo de compuestos orgánicos, para posteriormente extrapolarlo a otras propiedades de interés como las propiedades críticas (temperatura y presión).

El punto de ebullición y en el caso en que estuvieran disponibles en la literatura la temperatura y la presión crítica de 1168 compuestos orgánicos diversos constituyen el conjunto de datos para los siguientes casos de QSPRs. El conjunto se divide en cuatro subconjuntos cada uno para una propiedad, temperatura de ebullición, $T_b$, temperatura crítica, $T_c$, presión crítica, $P_c$ y otro en el que se incluyen las tres propiedades simultáneamente. El rango para cada una de las propiedades experimentales oscila entre, 111.6 K a 771 K para el punto de ebullición, entre 190.5 K y 926 K para la temperatura crítica (530 compuestos), y entre 1.02 a 8.95 MPa (465 compuestos). 436 de los compuestos forman el conjunto común para las tres propiedades con el cual se desarrolla un modelo global y simultáneo para cada una de ellas.

Se utilizan dos conjuntos de índices, el primero formado por siete descriptores topológicos, y se denomina tset. El cual, se integra por cinco índices de conectividad de valencia ($^{0\text{-}4}\chi^{v}$), el índice de forma kappa ($^{2}\kappa$) y la suma de números atómicos (N). Además, se toma en cuenta el momento

dipolar ($\mu$), de forma que hubiera algún índice que cuantificara las interacciones espaciales, con lo cual, el segundo conjunto de descriptores queda constituido por ocho índices y se denomina wset.

Una forma de comparar la calidad de los modelos QSPR propuestos es, contrastando los resultados con los métodos convencionales más utilizados, como los métodos de contribución de grupos. El método de Meissner para el punto de ebullición y el método de Joback para la temperatura crítica, son usados con este propósito. Una breve descripción de estos métodos la encontraremos en la literatura (Lymman W. et al., 1990; Reid R. et al., 1987). Un resumen de los errores para cada propiedad usando tanto backpropagation como fuzzy ARTMAP se encuentra en las Tablas 4.2-4.5 así como los resultados obtenidos con los métodos de contribución de grupos para las temperaturas de ebullición y críticas.

La metodología empleada en este caso se representa en la Fig. 4.13. Una diferencia con respecto a los casos anteriores estriba en el método empleado para seleccionar el conjunto de entrenamiento. En este caso empleamos el algoritmo de fuzzy ART. Como entrada a la red se presentan tanto el conjunto de descriptores moleculares como la propiedad objetivo, obteniéndose un conjunto de clases de dichas variables. Un porcentaje de sus elementos se toma como elementos del conjunto de prueba, garantizando con ello, la homogeneidad en ambos conjuntos (tr y te).

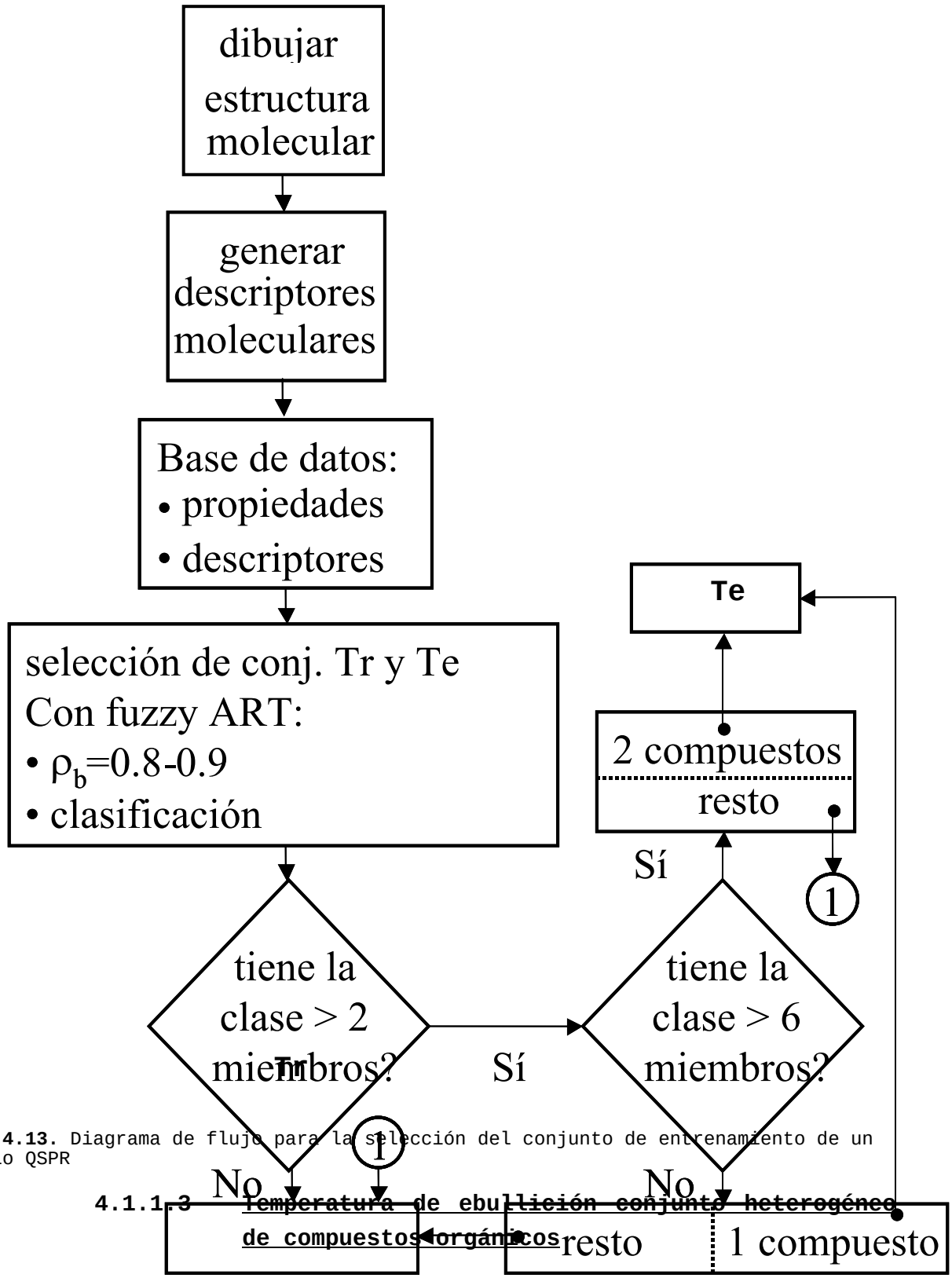



**Fig. 4.13.** Diagrama de flujo para la selección del conjunto de entrenamiento de un modelo QSPR

**4.1.1.3 Temperatura de ebullición conjunto heterogéneo de compuestos orgánicos**

En el entrenamiento de fuzzy ARTMAP se utilizan 1015 compuestos. Los modelos QSPR-fuzzy ARTMAP con ocho descriptores moleculares (wset) dieron

mejores resultados al compararlos con modelos que se obtienen usando backpropagation o métodos de contribución de grupos. Los errores absolutos para el conjunto global, el conjunto de entrenamiento y el de generalización son: 2.0 K (0.49%), 0.26 K (0.06%) y 13.5 K (3.3%), respectivamente, con una desviación estándar de 8.6 K (2.1%), 1.3 K (0.30%) y 19.9 K (4.8%). Estos errores se ven incrementados a 4.0 K (1.0%), 1.6 K (0.4%), y 20.5 K (5.0%), respectivamente, para el modelo basado en el conjunto de descriptores tset, las desviaciones estándar correspondientes a cada conjunto de errores son 14.2 K (3.5%), 7.2 K (1.8%) y 30.2 K (7.4%). El aporte de la información 3D dado por el momento dipolar se ve reflejado en los resultados anteriores. Por otro lado fuzzy ARTMAP fue capaz de establecer explícitamente la relación no-linealidad entre los descriptores moleculares y el punto de ebullición de este grupo de compuestos orgánicos. Las predicciones que se obtienen con la mejor arquitectura de backpropagation 8-12-1 y el wset (Tabla 4.2) dan lugar a un error medio absoluto de 27.7 K (7.0%) y una desviación estándar de 38.3 K (10.3%). Los errores del conjunto de entrenamiento y de generalización son 28.7 K (7.2%) y 20.8 K (5.4%), cuya desviación estándar es 39.6 K (10.8%) y 25.7 K (10.2%), Fig. 4.14a. La mejor arquitectura en el caso del tset es de 7-14-1, la cual genera unos errores equivalentes a 28 K (7.1%), 30 K (8.0%) y 23.7 K (5.2%) para el conjunto global, el de entrenamiento y el de prueba, respectivamente. La desviación estándar en cada caso fue de 23.7 K (5.2%), 29.3 K (8.1%), 27.2 K (5.8%). Cabe destacar que el conjunto de entrenamiento seleccionado por fuzzy ART con cada conjunto de índices cambia, lo cual puede en parte explicar la diferencia en los porcentajes en cada modelo QSPR.

En la Fig. 4.14b se comparan los resultados de método de Meissner y fuzzy ARTMAP, el cual genera un error medio de 28.7 K (6.95%). El gráfico 4.14 da una idea clara de la capacidad de un modelo QSPR basado en fuzzy ARTMAP ya que este modelo fue creado en un amplio rango de puntos de ebullición, y los resultados presentan sólo unos cuantos compuestos con errores mayores que el error medio. De ello puede extraerse la capacidad de fuzzy ARTMAP para dividir el espacio de entrada lo cual permite el desarrollo de modelos QSPR más generales y flexibles comparados con los modelos basados en sistemas de activación global como lo es un backpropagation. De hecho solo seis compuestos presentan una clasificación errónea al usar el wset y cuatro de ellos deben considerarse como *outliers* al presentar errores mayores al 15%. Los compuestos que fueron clasificados incorrectamente son: el ciclopentileno, piridina, metil-ciclohexano, 3-etil-pentano, 2-metil-3-etil-pentano y el 4-cloro-fenol. Las razones pueden entenderse mejor al explicarlo desde el punto de vista de la preclasificación de fuzzy ART y la realizada por fuzzy ARTMAP en su tarea de clasificación/predicción. El sistema fuzzy ART asigna al ciclopentileno a una clase formado por compuestos cíclicos con 5-6 átomos de carbono durante la etapa de preclasificación, la cual se lleva a cabo utilizando un

parámetro de vigilancia ρa=0.9. Este compuesto, el cual es el único hidrocarburo cíclico con doble enlace y cinco átomos de carbono de la base de datos, no fue seleccionado por fuzzy ART para entrenar la red, de acuerdo con el procedimiento descrito en la Fig. 4.13. Una red fuzzy ARTMAP es entrenada con un parámetro de vigilancia mayor ρa=0.997 lo cual permite captar las características de las relaciones estructura-propiedad siendo capaces de discernir entre una clase y otra de compuestos de forma detallada. El vector prototipo de las clases creadas por el módulo artA está formado por compuestos muy similares. Como resultado, ninguna de las clases creadas durante el entrenamiento, contiene información suficientemente semejante para caracterizar al ciclopentileno, el cual es llevado a una clase anómala durante la generalización. En el caso de la piridina y en algunos otros casos los compuestos que contienen nitrógeno y oxígeno presentaran clasificaciones erróneas, y esto se debe a que la con la información proporcionada a la red (descriptores) la red es incapaz de distinguir entre compuestos que contengan un u otro tipo de átomo. La justificación de ello la dio Randic (Randic M. 1993) al señalar que los índices de conectividad no son suficientes para captar las diferencias entre algunos heteroátomos como el nitrógeno y el oxígeno, y ello debido a su definición. Ambos átomos están en el mismo grupo de la tabla periódica, por lo cual su número de electrones de valencia es equivalente, si recordamos la definición de las deltas para construir los índices de conectividad molecular, vemos que compuestos con el mismo número de átomos de carbono, cuya diferencia sea la presencia de nitrógeno u oxígeno, serán prácticamente equivalentes. Argumentos similares, se pueden aplicar para el resto de los compuestos. El máximo error que se observa (24%) corresponde al 4-cloro-fenol, el cual es asignado a una clase de compuestos aromáticos, pero que no presentan heteroátomos. Cabe puntualizar que el número de compuestos mal clasificados y que pueden considerarse aumenta a 20 y 18 respectivamente al utilizar el tset para construir el modelo QSPR.

### 4.1.2 Temperatura crítica

Fuzzy ARTMAP estima la temperatura crítica a partir de wset con un error medio absoluto de 1.4 K (0.24%) y una desviación estándar de 5.6 K (0.96%). Si presentamos a la red únicamente información estructural (tset) el error se incrementa a 2.1 K (0.96%) con una desviación estándar de 8.9 K (1.5%), como se observa en la Tabla 4.3. Tendencias similares se observan, pero con errores mayores durante la generalización al utilizar el tset o wset. La predicción media decrece de 13.8 K (2.3%) a 8.5 K (1.5%) cuando el momento dipolar es considerado como parámetro de entrada a la red. Hay dos compuestos el 2-etil-1-hexanol y el dipropil-amina no son clasificados correctamente, el error relativo de ambos es menor al 15% con respecto al valor experimental, por lo que no podemos hablar de *outliers* en este modelo. El alcohol cae dentro de la clase de las anilinas y su predicción general el error mayor de los observados, 12.5%. De forma inversa la amina

es clasificada en una clase que contiene éteres. En contraste con estos resultados, el número de compuestos cuya clase puede considerarse no apta se incrementa a doce al no incluir el momento dipolar (tset). Las razones se repiten, compuestos con nitrógeno no pueden diferenciarse de aquellos que contienen oxígeno y viceversa. Y es nuevamente el 2-etil-1-hexanol el compuesto con el mayor error relativo durante la generalización. Una representación gráfica de los resultados se muestra en la Fig. 4.15.

El contraste de los resultados anteriores con los obtenidos con la mejor configuración de backpropagation, 8-12-1, Fig. 4.15a y el método de contribución de grupos de Joback se presenta en la Fig. 4.15b. El modelo QSPR-backpropagation estima la temperatura crítica de este conjunto con un error medio de 30.2 K (5.6%), cuya desviación estándar es de 29.2 K (6.0%). El error de entrenamiento y del conjunto de prueba es: 31.4 K (5.8%) y 21.6 K (3.9%), respectivamente, y su desviación estándar es 30.3 K (6.3%) y 18.9 K (3.9%). El error más grande, corresponde a la molécula más pequeña el metano, 52.2%. Este mismo comportamiento lo observaron Hall y Kier ( Hall, L. y Kier, L., 1995) al usar únicamente información 2D para modelar el punto de ebullición de alcoholes y alcanos. La notable diferencia del desempeño de fuzzy ARTMAP con respecto al método de Joback es evidente la Fig. 4.15b. Este último presenta un error medio de 19.5 K (3.4%) con una desviación estándar del 23.9 K (4.1%).

### 4.1.3 Presión crítica

En el modelo para la presión crítica fuzzy ARTMAP QSPR se utilizaron 409 compuestos para la etapa de entrenamiento de la red, el error global derivado para el conjunto total de 463 compuestos es de 0.02 MPa (0.52%) y 0.02 MPa (0.65%), si se utiliza wset o tset. La desviación estándar en cada caso es de 0.08 MPa (2.5%) y 0.09 MPa (2.9%), respectivamente, Tabla 4.4. Unicamente cuatro compuestos se clasificaron incorrectamente al usar wset como conjunto de entrada para el modelo QSPR. El máximo error relativo fue para el trans-4-metil-2-pentileno (34.9%). Este alqueno fue agrupado junto a compuestos aromáticos. Además se observa que el isobutileno presenta un error superior al 15%, sin embargo no es debido a una mala clasificación. El número de compuestos mal clasificados se incrementa a diez si no se toma en cuenta el momento dipolar en los parámetros de entrada (tset), y el máximo error relativo asociado al 3-metil-pentano por una mala clasificación dentro de una categoría de aromáticos es de 33.6%.

El desempeño del modelo QSPR para la presión crítica usando fuzzy ARTMAP se compara con la mejor arquitectura obtenida con backpropagation 8-10-1 en la Fig. 4.16. De este último modelo se obtiene un error medio para el conjunto global de 0.31 MPa (7.7%), con una desviación estándar de 0.39 MPa (7.4%), tal como aparece en la Tabla 4.4. Los errores correspondientes al conjunto de entrenamiento y de generalización son 0.33 MPa (8.0%) y 0.19

MPa (6.0%), respectivamente, con su correspondiente desviación estándar de 0.40 MPa (7.6%) y 0.21 MPa (5.4%). Es interesante recalcar que en este caso el error de generalización para el modelo fuzzy ARTMAP es menor que el error generado durante el entrenamiento con una arquitectura backpropagation 8-10-1.

### 4.1.4 Predicciones simultáneas de diversas propiedades

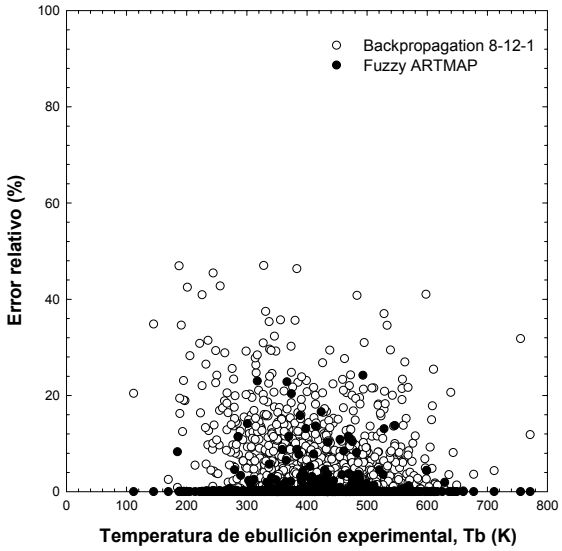
Finalmente, la capacidad de fuzzy ARTMAP para descubrir las relaciones estructura propiedad se probo en un sistema de predicción simultánea de las tres propiedades (temperatura de ebullición, temperatura y presión críticas). La intersección de los tres conjuntos de datos experimentales da lugar a un subconjunto formado por 436 elementos comunes. De los cuales 386 fueron seleccionados por fuzzy ART como conjunto de entrenamiento para el modelo múltiple y el resto se usa para medir la generalización del modelo QSPR. Los errores globales usando wset y tset se resumen en la Tabla 4.5, siendo 1.0 K (0.28%) y 1.6 K (0.44%) para la temperatura de ebullición, 1.3 K (0.24%) y 2.1 K (0.38%) para la temperatura crítica y 0.02 MPa (0.63%) y 0.02 MPa (0.63%) para la presión crítica. La desviación estándar correspondiente es 4.7 K (1.43%), 7.8 K (2.3%), 5.8 K (1.1%) y 9.03 K (1.8%) y 0.01 MPa (2.3%) y 0.1 MPa (2.7%), en cada caso. El momento dipolar como parámetro de entrada, vuelve a mejorar el desempeño de la red como en el caso de los modelos individuales para ambas temperaturas, pero no se observa un efecto significativo en el caso de la presión crítica. La arquitectura óptima de backpropagation es 8-15-3 da lugar a errores mayores en comparación con el modelo fuzzy ARTMAP como puede observarse en la Tabla 4.5.

En general, los errores estimados para el modelo de predicción múltiple son similares a los obtenidos para los modelos individuales. Y a pesar de que se impone una demanda mayor a fuzzy ARTMAP para reconocer una relación más compleja entre la estructura de los compuestos y las tres propiedades objetivo, esta se ve compensada por la información adicional proporcionada por la Interrelación entre esas propiedades dada por el modulo artB durante el proceso de *match tracking*. La información adicional no es un problema para un clasificador neuronal ya que es capaz de correlacionar múltiples entradas con múltiples salidas y al mismo tiempo descubrir las no linealidades implícitas entre la estructura molecular y las propiedades objetivo. Como consecuencia, aún el modelo múltiple es superior a los métodos tradicionales de contribución de grupos que predicen propiedades únicas con ecuaciones lineales.

Son cuatro los compuestos mal clasificados durante la generalización de la red fuzzy ARTMAP. El 1-5-hexadieno el cual se clasifica con el 2-5-dioxano. El isobutileno que se clasifica junto al tri-butil amina, y el 2-metil-butanol se clasifico erróneamente con el 2-metil-butano. Con lo cual

se general algunos elementos con errores superiores al 15%. Para el punto de ebullición el isobutileno presenta un error de 20.9%, sin embargo los errores para su temperatura y presión crítica caen dentro de la desviación estándar individual para cada propiedad. En el caso del -metil-2-butanol sus errores son mayores en los tres casos que la desviación estándar de cada modelo, pero inferiores al 15%. La presión crítica presenta un máximo error relativo de 32% correspondiente al ciclobutano, pero los errores en las otras dos propiedades son menores que el promedio. 9.4% fue el mayor error encontrado en el caso de la temperatura crítica para el dimetil oxalato. Este error y algunos otros en cualquiera de los tres casos tiene lugar para compuestos poco comunes en la base datos, es decir aquellos que no pueden crear clases con características suficientemente significativas para identificar a un colectivo. En el caso del modelo basado en tset 11 compuestos fueron clasificados de forma errónea y coincidentemente el máximo error se obtiene para el punto de ebullición del ciclobutano (35%).

Un par de aplicaciones adicionales en las que se aplica la metodología expuesta hasta el momento, puede encontrarse en la literatura (Yaffe D. et al., 2001; Yaffe D. et al., 2002). Los modelos QSPR usando tanto backpropagation como fuzzy ARTMAP y diferentes tipos de índices dan resultados satisfactorios tanto para la solubilidad de 515 compuestos orgánicos, como para el coeficiente de partición octanol/agua de 442 compuestos.



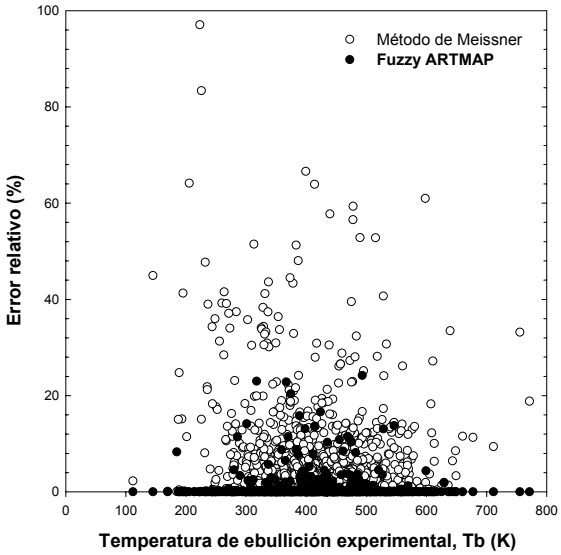
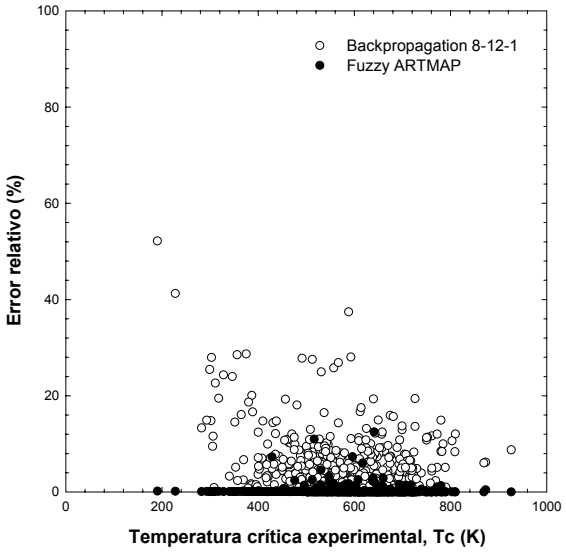



**Fig. 4.14.** Comparación de los errores relativos usando fuzzy ARTMAP para estimar el punto de ebullición con respecto a (a) la arquitectura 8-12-1 backpropagation y (b) método de contribución de grupos de Meissner's.



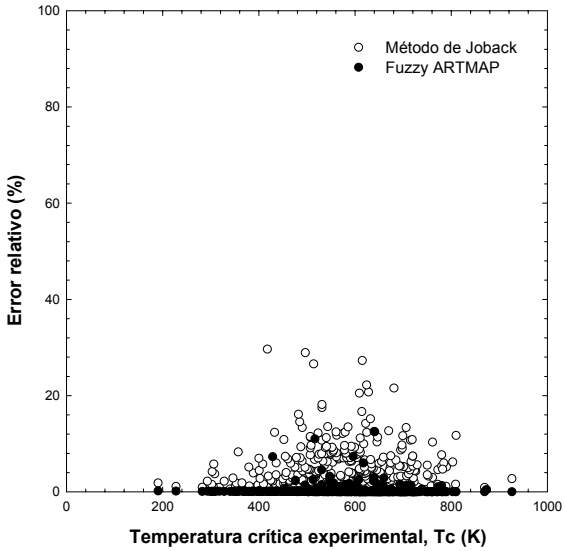



**Fig. 4.15.** Comparación de los errores relativos usando fuzzy ARTMAP para estimar la temperatura crítica con respecto a (a) la arquitectura 8-12-1 backpropagation y (b) método de contribución de grupos de Joback.

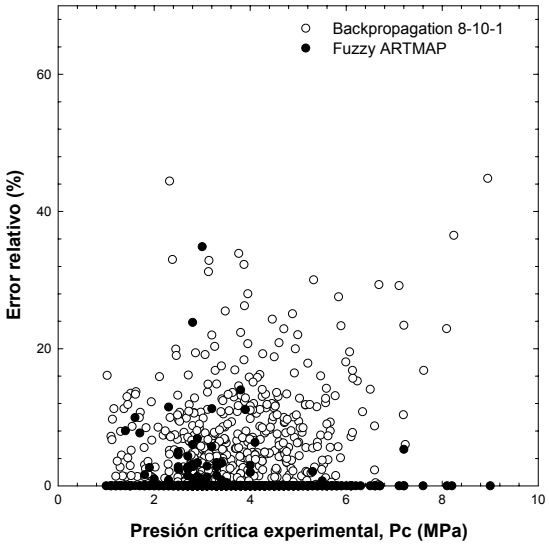



**Fig. 4.16.** Comparación de los errores relativos usando fuzzy ARTMAP para estimar la presión crítica con respecto a (a) la arquitectura 8-10-1 backpropagation.

**Tabla 4.2.** Desempeño de modelos QSPR para el punto de ebullición de un conjunto heterogéneo de compuestos orgánicos usando Fuzzy ARTMAP y Backpropagation como arquitecturas neuronales y Meissner como método de contribución de grupos

| Conjunto | # de patrones | Error medio | | Error medio | | Desviación estándar | |
|---|---|---|---|---|---|---|---|
| | | K | % | K | % | K | % |
| Modelo Fuzzy ARTMAP con N, μ, χ(0-4), $\kappa^2$ | | | | | | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Global | 1168 | 2.00 | 0.49 | | | 8.61 | 2.09 |
| Tr | 1015 | | | 0.26 | 0.06 | 1.33 | 0.30 |
| Te | 153 | | | 13.47 | 3.27 | 19.86 | 4.82 |
| Modelo Fuzzy ARTMAP con N, $\chi$(0-4), $\kappa^2$ | | | | | | | |
| Global | 1168 | 4.04 | 1.00 | | | 14.21 | 3.50 |
| Tr | 1015 | | | 1.61 | 0.40 | 7.23 | 1.79 |
| Te | 153 | | | 20.47 | 5.03 | 30.20 | 7.42 |
| Modelo Backpropagation (8-12-1) con N, $\mu$, $\chi$ (0-4), $\kappa^2$ | | | | | | | |
| Global | 1168 | 27.70 | 7.02 | | | 38.30 | 10.30 |
| Tr | 1015 | | | 28.70 | 7.20 | 39.60 | 10.80 |
| Te | 153 | | | 20.80 | 5.40 | 26.70 | 10.20 |
| Método de contribución de grupos de Meissner's | | | | | | | |
| Global | 1168 | 28.70 | 6.95 | | | 38.37 | 10.51 |

**Tabla 4.3.** Desempeño de modelos QSPR para la temperatura crítica de un conjunto heterogéneo de compuestos orgánicos usando Fuzzy ARTMAP y Backpropagation como arquitecturas neuronales y Joback como método de contribución de grupos

| Conjunto | # de | Error medio | | Error medio | | Desviación | |
|---|---|---|---|---|---|---|---|
| | | K | % | K | | | K |
| Modelo FuzzyARTMAP con N, $\mu$, $\chi$(0-4), $\kappa^2$ | | | | | | | |
| Global | 530 | 1.37 | 0.24 | | | 5.59 | 0.96 |
| Tr | 461 | | | 0.27 | 0.05 | 0.21 | 0.04 |
| Te | 69 | | | 8.51 | 1.45 | 13.29 | 2.30 |
| Modelo FuzzyARTMAP con N, $\chi$(0-4), $\kappa^2$ | | | | | | | |
| Global | 530 | 2.05 | 0.35 | | | 8.94 | 1.48 |
| Tr | 461 | | | 0.32 | 0.06 | 0.69 | 0.12 |
| Te | 69 | | | 13.84 | 2.30 | 21.62 | 3.58 |
| Modelo Backpropagation (8-12-1) con N, $\mu$, $\chi$(0-4), $\kappa^2$ | | | | | | | |
| Global | 530 | 30.20 | 5.60 | | | 29.20 | 6.00 |
| Tr | 463 | | | 31.40 | 5.80 | 30.30 | 6.30 |
| Te | 68 | | | 21.60 | 3.90 | 18.90 | 3.90 |
| Método de contribución de groupo de Joback's | | | | | | | |
| Global | 530 | 19.50 | 3.40 | | | 23.90 | 4.10 |

**Tabla 4.4.** Desempeño de modelos QSPR para la presión crítica de un conjunto heterogéneo de compuestos orgánicos usando Fuzzy ARTMAP y Backpropagation como arquitecturas neuronales QSPR

| Conjunto | # de | Error medio | | Error medio | | Desviación | |
|---|---|---|---|---|---|---|---|
| | | MPa | % | MPa | % | MPa | % |
| Modelo FuzzyARTMAP con N, μ, χ(0-4), $\kappa^2$ | | | | | | | |
| Global | 463 | 0.02 | 0.52 | | | 0.08 | 2.48 |
| Tr | 409 | | | 0.00 | 0.01 | 0.00 | 0.02 |
| Te | 54 | | | 0.13 | 4.31 | 0.19 | 6.02 |
| Modelo FuzzyARTMAP con N, χ(0-4), $\kappa^2$ | | | | | | | |
| Global | 463 | 0.02 | 0.65 | | | 0.09 | 2.91 |
| Tr | 409 | | | 0.00 | 0.08 | 0.04 | 0.77 |
| Te | 54 | | | 0.15 | 4.90 | 0.21 | 6.87 |
| Modelo Backpropagation model (8-10-1) con N, μ, χ(0-4), $\kappa^2$ | | | | | | | |
| Global | 463 | 0.31 | 7.71 | | | 0.39 | 7.38 |
| Tr | 409 | | | 0.33 | 7.96 | 0.40 | 7.58 |
| Te | 54 | | | 0.19 | 6.00 | 0.21 | 5.43 |

**Tabla 4.5.** Desempeño de modelos QSPR para la temperatura de ebullición, la temperatura crítica y la presión crítica de un conjunto heterogéneo de compuestos orgánicos usando fuzzy ARTMAP y Backpropagation como arquitecturas neuronales

| Conjunto | # de | Error medio | | Error medio | | Desviación estándar | |
|---|---|---|---|---|---|---|---|
| | | K / MPa | % | K / MPa | % | K / MPa | % |
| Modelo fuzzy ARTMAP para el punto de ebullición con N, μ, χ(0-4), $\kappa^2$ | | | | | | | |
| Global | 436 | 1.01 | 0.28 | | | 4.72 | 1.43 |
| Tr | 386 | | | 0.02 | 0.01 | 0.08 | 0.02 |
| Te | 50 | | | 8.31 | 2.31 | 11.32 | 3.57 |
| Modelo fuzzy ARTMAP para la temperatura crítica con N, μ, χ(0-4), $\kappa^2$ | | | | | | | |
| Global | 436 | 1.34 | 0.24 | | | 5.83 | 1.06 |
| Tr | 386 | | | 0.02 | 0.00 | 0.01 | 0.00 |
| Te | 50 | | | 11.10 | 2.01 | 13.40 | 2.43 |
| Modelo fuzzy ARTMAP para la presión crítica con N, μ, χ(0-4), $\kappa^2$ | | | | | | | |
| Global | 436 | 0.02 | 0.63 | | | 0.11 | 2.74 |
| Tr | 386 | | | 0.00 | 0.01 | 0.00 | 0.00 |
| Te | 50 | | | 0.18 | 5.23 | 0.27 | 6.29 |
| Modelo fuzzy ARTMAP para el punto de ebullición con N, χ(0-4), $\kappa^2$ | | | | | | | |
| Global | 436 | 1.59 | 0.44 | | | 7.81 | 2.33 |
| Tr | 376 | | | 0.21 | 0.06 | 1.30 | 2.01 |
| Te | 50 | | | 11.79 | 3.29 | 19.66 | 5.98 |
| Modelo fuzzy ARTMAP para la temperatura crítica con N, χ(0-4), $\kappa^2$ | | | | | | | |
| Global | 436 | 2.07 | 0.38 | | | 9.03 | 1.75 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Tr | 376 | | | 0.30 | 0.06 | 2.01 | 0.37 |
| Te | 50 | | | 15.09 | 2.81 | 21.66 | 4.28 |
| Modelo fuzzy ARTMAP para la presión crítica con con N, $\chi$(0-4), $\kappa^2$ | | | | | | | |
| Global | 436 | 0.02 | 0.63 | | | 0.09 | 2.30 |
| Tr | 376 | | | 0.00 | 0.12 | 0.04 | 0.88 |
| Te | 50 | | | 0.15 | 4.44 | 0.19 | 4.75 |
| Modelo backpropagation para el punto de ebullición (8-15-3) con N, $\mu$, $\chi$(0-4), $\kappa^2$ | | | | | | | |
| Global | 436 | 19.45 | 5.48 | | | 22.54 | 6.32 |
| Tr | 376 | | | 20.95 | 5.92 | 23.63 | 6.63 |
| Te | 50 | | | 10.03 | 2.72 | 9.51 | 2.58 |
| Modelo backpropagation para la temperatura crítica (8-15-3) con N, $\mu$, $\chi$(0-4), $\kappa^2$ | | | | | | | |
| Global | 436 | 27.65 | 5.23 | | | 30.46 | 5.99 |
| Tr | 376 | | | 29.84 | 5.64 | 31.92 | 6.28 |
| Te | 50 | | | 13.90 | 2.62 | 11.83 | 2.38 |

| Conjunto | # de | Error medio | | Error medio | | Desviación estándar | |
|---|---|---|---|---|---|---|---|
| | | K / MPa | % | K / MPa | % | K / MPa | % |
| Modelo backpropagation para la presión crítica (8-15-3) con N, $\mu$, $\chi$(0-4), $\kappa^2$ | | | | | | | |
| Global | 436 | 0.31 | 7.64 | | | 0.40 | 6.99 |
| Tr | 376 | | | 0.33 | 8.03 | 0.42 | 7.23 |
| Te | 50 | | | 0.18 | 5.39 | 0.18 | 4.68 |

## 4.2. Evaluación de QSPR, QSAR y QSTR mediante una metodología integrada SOM/Fuzzy ARTMAP

Hasta el momento es posible justificar que una parte del proceso está bien cimentado. Los resultados muestran como fuzzy ARTMAP es una herramienta eficaz con una buena capacidad de interpolación y extrapolación. Pero aún nos queda por resolver una cuestión, ¿Cuáles y cuántos índices se requieren para caracterizar una propiedad?. Al intentar obtener modelos para otro tipo de propiedades como el punto de fusión, los resultados no fueron nada satisfactorios. Las clases a las cuales iban a parar los compuestos durante la fase de generalización eran en su mayoría, inconsistentes. Esto podía deberse a la mala caracterización de la propiedad en sí. Es decir, la topología de la molécula carece la información intrínseca en ciertas propiedades. En estos casos es preciso recurrir a características más fundamentales, como la configuración electrónica, la polarizabilidad o algún tipo de índice que permita cuantificar interacciones que se dan no sólo a través del enlace entre los distintos átomos que conforman una estructura molecular, si no aquellas debidas a efectos estéricos e interacciones entre los diferentes grupos funcionales presentes en la

molécula. Este tipo de índices, ya sea geométricos o cuánticos, llamados comúnmente como índices 3-D sólo pueden calcularse a partir de métodos semi-empíricos o puramente cuánticos. Con lo cual al aumentar el número de índices se plantea un nuevo problema, encontrar un método para seleccionar dichos índices sin tener que hacer un proceso de prueba y error o combinatorio/discriminatorio de los mismos, por razones obvias de tiempo y coste computacional. Una vía factible para este nuevo problema se desarrolla en los siguientes casos.

### 4.2.1 Toxicidad

Todos los días el cuerpo está rodeado de substancias químicas potencialmente peligrosas, por ello el estudio de la respuesta tóxica en función con lo que sucede en un organismo es un área ampliamente explorada dentro del campo de la toxicología ambiental. La relación entre la cantidad de un tóxico y la magnitud del efecto es lo que se conoce como relación dosis-respuesta y es uno de los conceptos centrales de la toxicología. A continuación, se da una breve descripción de algunos conceptos introductorios, antes de formular los siguientes casos, con lo cual se busca reafirmar la importancia del tema y sirve como punto de partida en la discusión de resultados.

La toxicidad de una substancia se puede incrementar o disminuir por la exposición simultánea o consecutiva de otra substancia. Los efectos combinados pueden ser aditivos, sinérgicos, potenciantes o antagónicos. Los efectos de dos tóxicos administrados de manera simultánea pueden producir una simple respuesta aditiva, la cual es la suma de las respuestas individuales, por ejemplo, dos insecticidas organofosforados producen una inhibición aditiva de la colinesterasa. La respuesta puede ser sinérgica, cuando es mayor que la esperada por la adición de las respuestas individuales, por ejemplo, el tetracloruro de carbono y el etanol son hepatóxicos, es decir, producen una lesión hepática mucho mayor cuando son administrados juntos que la suma de las respuestas que cada uno produce cuando se administra individualmente. Una respuesta se potencia cuando una substancia que no es tóxica en un determinado órgano blanco, pero que cuando se agrega a otra hace que ésta se vuelva mucho más tóxica, por ejemplo el isopropanol no es tóxico para el hígado pero cuando se administra junto con el tetracloruro de carbono, incrementa la actividad hepatóxica de este último compuesto. Una respuesta es antagónica cuando dos substancias administradas simultáneamente se interfieren mutuamente en sus acciones o una se interfiere con la acción de la otra. Las respuestas antagónicas son la base de muchos antídotos. En el antagonismo funcional cada substancia produce efectos contrarios sobre la misma función fisiológica contrabalanceándose mutuamente, por ejemplo, la administración de norepinefrina (vasodepresor) para bajar la alta presión producida por las intoxicaciones severas con barbitúricos. El antagonismo químico, que también se le llama inactivación, es simplemente una reacción química entre

los dos compuestos que da lugar a un producto menos tóxico, por ejemplo, la quelación del dimercaptol con iones metálicos. El antagonismo disposicional es la alteración de ya sea la absorción, distribución, metabolismo o excreción de un compuesto para disminuir su concentración o duración en el sitio blanco.

La relevancia de este tipo de estudios que en muchas ocasiones se denominan QSTR (relaciones estructura toxicidad) es evidente, cada uno de ellos podría convertirse en un interesante caso de estudio. Sin embargo, se han escogido los que se refieren a la predicción de diferentes índices de toxicidad (diferentes índices hacen referencia a diferentes caminos de acción o diferentes organismos sobre los cuales se realiza el estudio). Además, en este subapartado se propone una metodología más general con el fin de resolver las dos partes de un modelo QSAR/QSTR. El procedimiento integra dos sistemas cognitivos. El primero, un mapa auto-organizado, SOM cuyo objetivo será resolver la parte de reconocimiento de patrones, es decir cuales y cuantas son las variables necesarias para caracterizar al conjunto estudiado, en especial a la propiedad que estamos modelando. Fuzzy ARTMAP será segundo sistema neuronal, el cual como se ve en los ejemplos anteriores, tiene una buena capacidad de interpolación y de generalización. El procedimiento general se resume en el diagrama de flujo de la Fig. 4.13. Se han utilizado dos medidas de toxicidad distintas, $LC_{50}$ y $LD_{50}$ la concentración y la dosis letal en el que el 50% de la población muere. La geometría de cada compuesto se optimizó utilizando un software comercial MOPAC 6.0 y el método PM3 para calcular los diferentes parámetros cuánticos.

Una vez generados descriptores tanto topológicos como geométricos y cuánticos, se usan mapas auto-organizados (SOM) con el fin de analizar la similitud entre los diferentes compuestos. Y de esta forma, seleccionar el conjunto de índices más relevantes para dicha propiedad. Es así, como se define un mapa global (el número de nodos en la red) y con todos los descriptores disponibles, además de la propiedad objetivo, SOM se encarga de clasificar dichas variables preservando la topología y generando así una distribución de los compuestos en familias. Los resultados pueden analizarse usando los planos de cada componente (C-planes), en los cuales se pone de manifiesto la contribución de la variable en la el peso de cada nodo, con lo cual al aplicar una método de proyección lineal, es decir al calcular la correlación lineal entre los planos, es posible reorganizarlos e identificar que variables se comportan de forma semejante. Además la congruencia de los grupos de C-planes obtenidos es consistente con las varianzas entre pares de variables obtenidas al realizar una análisis de Pearson. El conjunto óptimo de descriptores se obtiene escogiendo de cada clase el índice con la correlación más alta con respecto a la variable objetivo, con la única restricción que su correlación sea mayor que la correlación media del conjunto con la variable objetivo. Una vez que la

información más relevante y no interrelacionada ha sido considerada como parte del conjunto base, se sigue añadiendo uno a uno los descriptores en orden decreciente de correlación de forma independiente a la categoría de la cual formen parte. Se genera un mapa por cada *nuevo* conjunto de índices. La disimilitud entre mapas se obtiene a partir de la relación  Este proceso termina cuando la disimilitud entre pares de mapas se incrementa, lo cual nos indica que el añadir un nuevo índice no contribuye a mejorar la distribución espacial de los compuestos. Siendo este conjunto de descriptores tomado como parámetro de entrada para el modelo QSTR/QSAR.

La viabilidad de la presente metodología se constata a partir de dos casos de estudio. Uno es el índice de toxicidad $LC_{50}$ para un conjunto homogéneo de 69 derivados del benzeno estudiado sobre una variedad de pez y el segundo conjunto está integrado por la toxicidad en ratas, vía ingestión oral de 155 compuestos orgánicos diversos. Los valores experimentales de toxicidad y sus correspondientes descriptores asociados a cada estructura en ambos conjuntos son asignados al conjunto de entrenamiento o de generalización a partir de la clasificación derivada de fuzzy ART. Seleccionar los elementos de acuerdo a su clasificación y de forma aleatoria da lugar a una distribución homogénea de los datos en ambos conjuntos y permite evaluar la capacidad de generalización de la red a elementos desconocidos pero que estén dentro del rango de los vistos previamente durante la fase de entrenamiento de la misma. Los histogramas de las Figs.4.17-4.18 muestran la distribución de los valores de toxicidad para ambos conjuntos (para los 69 datos de toxicidad –log ($LC_{50}$) y para los 155 correspondientes a log($LD_{50}$)), respectivamente. De dichas figuras podemos concluir que existe una distribución homogénea dentro de los mínimos y máximos de cada conjunto, a pesar observarse algunos valores de toxicidad carentes de elementos representativos, esto reafirma la necesidad de hacer una buena selección del conjunto de entrenamiento. En las Tablas 1 y 2 (Tabla 1 y 2 apéndice 3; Espinosa G. et al., 2002) muestran los compuestos seleccionados para el entrenamiento de la red (tr), 85% aproximadamente, 59 para el $LC_{50}$ y 135 para el $LD_{50}$, considerando el 15% restante como conjunto de generalización (te). Finalmente, partiendo de los conjuntos seleccionados para entrenamiento en cada caso, se construyen los dos modelos QSAR basados en el sistema neuronal fuzzy ARTMAP.

El siguiente paso es llevar a cabo la selección de índices mediante el uso de SOM, los 69 compuestos orgánicos se agruparon de acuerdo con la Fig. 4.19. En la figura se muestra la distribución de las seis familias (diferentes tipos de grupos funcionales  en el anillo aromático) proyectados sobre el plano de la variable objetivo $LC_{50}$. Cada categoría se identifica con una letra: (A) grupos halógenos, (B) hidroxilos, (C) nitro, (D) halógenos + hidroxilos, (E) alquilos, (F) combinaciones de los anteriores. Los compuestos se distribuyen de acuerdo a la familia a la cual pertenecen y a la similitud molecular identificado por el SOM y los índices

de los cuales se dispone para caracterizar cada estructura química, es decir, familias similares se localizan en posiciones similares o cercanas, por ejemplo, la familia A, la cual se forma de derivados halogenados, algunas veces forma interfaces con la familia D, la cual se integra de derivados cuyos substituyentes combinan halógenos más grupos hidroxilos.

En la Fig. 4.20. se muestran los diferentes planos (C-planes) de las variables agrupados en distintos grupos, de acuerdo al comportamiento o el aporte de la variable a la distribución de los compuestos en el mapa. Como información adicional en la Tabla 3 (Tabla 3 apéndice 3, Espinosa G. et al., 2002) se incluye la matriz de covarianzas para el conjunto de descriptores moleculares y la variable objetivo. De ambos (tabla y figura) se encontraron las siguientes concordancias:

(i) Los índices de conectividad de orden uno, dos, tres y cuatro, junto con la suma de números atómicos, forman el primer grupo de índices. Ellos muestran en la Tabla 3 (Tabla 3, apéndice 3, Espinosa G. et al., 2002) una covarianza elevada entre ellos.
(ii) Ambos tipos de índices de polarizabilidad, uno calculado por la suma de grupos propuestos por Hansen y el otro determinado a través de un método semi-empírico (PM3), se muestran altamente correlacionados con el número de niveles ocupados, el índice de repulsión núcleo-núcleo, y el índice de forma kappa. Formando el segundo grupo de índices.
(iii) Los índices cuánticos, que caracterizan la energía de resonancia, la atracción electrón-núcleo, y la energía de intercambio, forman la tercera categoría de índices.
(iv) El último grupo de índices queda integrado por el índice de conectividad de orden cero, el momento dipolar y el índice de hidrógeno de Hansen.

En la Tabla 4. (Tabla 4, apéndice 3: Espinosa G. et al., 2002) se muestran las medidas de disimilitud entre trece conjuntos de índices formados de acuerdo a la metodología descrita anteriormente. El primer conjunto de índices está formado por un representante de cada grupo y los otros doce, mediante la adición uno a uno de los índices restantes en orden decreciente de su correlación con la variable objetivo. Los representativos de cada grupo son: los índices de conectividad de orden cero y uno, la polarizabilidad media, y el índice de atracción electrón-núcleo. La disimilaridad promedio entre mapas alcanza un mínimo en 0.1386, al añadir al conjunto base, los siguientes seis índices en orden decreciente de correlación con la variable objetivo $LC_{50}$. Siendo estos índices, el índice de conectividad de orden tres, la suma de números atómicos, el índice de conectividad de orden dos, el número de niveles ocupados, la energía de repulsión núcleo-núcleo y el índice de conectividad de orden cuatro. En este caso específico, si introducimos los índices por orden de correlación,

solo encontramos un índice que cambia, el índice de orden cero es substituido por el índice de forma kappa. Este cambio que parece insignificante afecta a la clasificación de algunos compuestos como se discutirá a continuación.

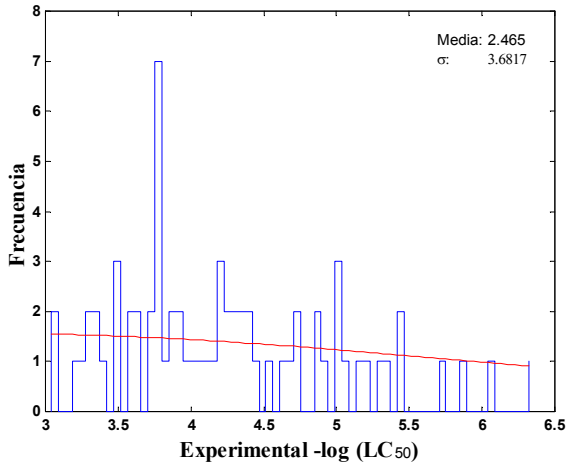



**Fig. 4.17** Histograma de los valores experimentales del índice de toxicidad -log($LC_{50}$) de un conjunto de derivados del benceno

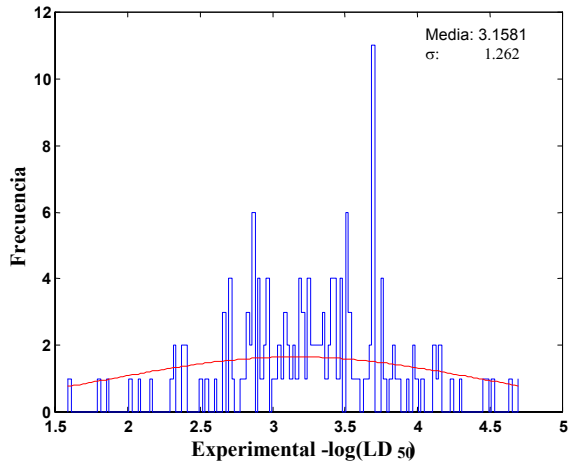



**Fig. 4.18.** Histograma de los valores experimentales del índice de toxicidad log($LD_{50}$) para un conjunto de compuestos orgánicos diversos

**Fig. 4.19**. Distribución de las seis familias del grupo de derivados del benceno generada con SOM: (A) halógenos, (B) hidroxilos, (C) nitro, (D) halógenos e hidroxilos, (E) alquilos, y (F) substituyentes adicionales Los colores indican las distancias relativas entre los elementos de cada categoría

El modelo fuzzy ARTMAP-QSAR para el $LC_{50}$ predice la toxicidad de los 69 compuestos derivados del benzeno sin producir clasificaciones erróneas, con un error absoluto medio y su correspondiente desviación estándar de 0.02 log (0.46%), 0.06 log (1.35%) y un error de generalización de 0.15 log (3.50%) y 0.11 log (2.04%) de desviación estándar, Fig. 4.21 y 4.22a-c. En la Fig. 4.21 se representan también los resultados usando para fuzzy ARTMAP los 10 descriptores seleccionados por orden decreciente de correlación, como se mencionó anteriormente, este hecho modifica el conjunto más favorable de índice en un sólo descriptor, es decir se substituye el índice de conectividad de orden cero por el índice kappa, este hecho modifica la clasificación del 3-meil-2-4-dinitroanilina y el 2-6-dimetil fenol durante la generalización del modelo, hecho que se identifica en la Fig. 4.21 con el número 1 y 2 respectivamente. Aunque de forma global, la repercusión de este hecho es mínima con respecto al error global del modelo QSAR-fuzzy ARTMAP. Las consecuencias son importantes desde el punto de vista de la generalización individual. En contraste, las categorías obtenidas con el conjunto más apropiado de índices no muestran ninguna clasificación errónea. El compuesto que presenta el error relativo más elevado 8.1% (el p-cresol) es clasificado junto con su homólogo (m-cresol).

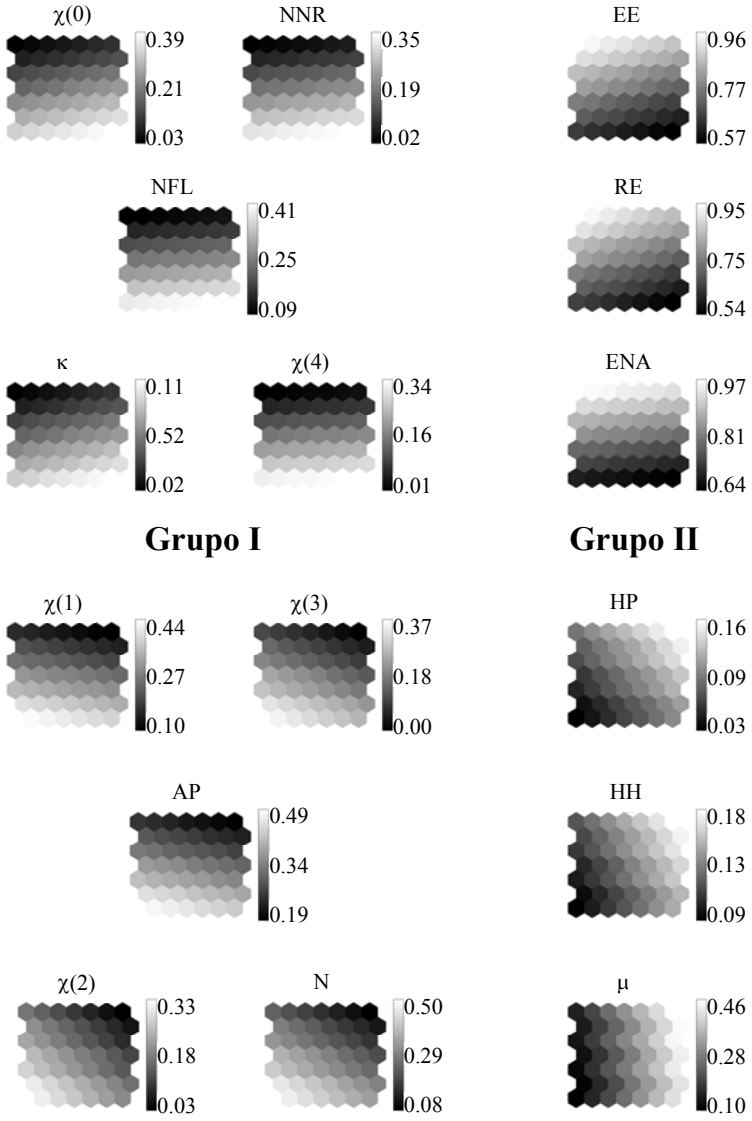



**Fig. 4.20**. Agrupación de los planos de descriptor proyectado sobre el mapa de $LD_{50}$. Los tonos de grises representan distancias relativas

El desempeño del modelo fuzzy ARTMAP-QSAR es significativamente superior que los modelos de regresión multilineal (MLR) publicados previamente como se presenta en la Fig. 4.21. Los errores absolutos y sus correspondientes desviaciones estándar para los modelos del $LC_{50}$ de Hall y Kier (Hall L. et al., 1984) , y Gute y Basak (Gute B. et al., 1997) son similares, alrededor de 0.22 (5.2%) y 0.2 (4.3%) log, comparados con 0.02 (0.05%) y 0.02 (0.05%) y 0.06 (1.4%) log, para el modelo actual. Una comparación directa de los mayores errores relativos muestra la superioridad de fuzzy ARTMAP con respecto a los modelos tradicionales. Mientras el error relativo mayor de fuzzy ARTMAP es 7.2% para el 1,2,4,5-tetraclorobenceno, el mayor error para el modelo propuesto por Gute y Basak (Gute B., et al., 1997) es de 21.1% para el 5-metil-2-4-nitroanilina.

Es de gran utilidad analizar el desempeño del modelo actual y de los modelos previos con respecto a los diferentes grupos funcionales presentes,

los errores por familias de compuestos se presentan en la Tabla 4.6 y en la Fig. 4.22a-c. La influencia de la posición de los substituyentes sobre la predicción de la toxicidad, tal como la propuso Hall y Kier (Hall L. et al., 1984) se presenta en la Tabla 4.6 para el grupo de clorobencenos, los derivados halogenados con substituyentes hidroxilos, alquilos y los fenoles con algún grupo adicional. Para la familia de clorobencenos, el error absoluto medio para el modelo de Gute y Basak (Gute B., et al., 1997) y el de Hall y Kier (Hall L. et al., 1984) son 0.11 (2.2%) y 0.04 (1.0%) log, respectivamente, comparado con el error del modelo actual que es de 0.09 log (1.6%). Las desviaciones estándar correspondientes son 0.11 (2.3%), 0.08 (1.7%) y 0.17 (3.2%) log. El buen desempeño de los modelos multilineales (MLR) para la familia de clorobencenos contrasta con los errores relativamente mayores obtenidos con fuzzy ARTMAP. El cual, sin embargo, tiene una buena capacidad de generalización global, sin especializarse en un conjunto determinado de compuestos, esto se pone de manifiesto tomando en cuanta que de los 59 compuestos utilizados para el entrenamiento de la red neuronal, siete eran los que pertenecían a la familia de los clorobencenos. De hecho fuzzy ARTMAP claramente supera los modelos de regresión multilineal para el resto de las familias de compuestos, tal y como se presenta en la Tabla 4.6.

En la Fig. 4.22a-c se comparan gráficamente los valores estimados vs los experimentales de acuerdo al grupo funcional adherido al anillo bencénico, (a) con halógenos, (b) hidroxilos, (c) nitro. La primera familia es una extensión de la de los clorobencenos (Tabla 4.6). El error absoluto medio para la primera Fig. 4.22a es de 0.15 (3.5%), 0.14 (3.0%) y 0.07 (1.4%) log, para Gute y Basak (Gute B., et al., 1997), Hall y Kier (Hall L. et al., 1984) y el modelo fuzzy ARTMAP, respectivamente. Los resultados para la familia de bencenos con grupos hidroxilos de la Fig. 4.22b muestra una mejora importante del modelo fuzzy ARTMAP con un error de 0.09 (2.4%) log comparado con los modelos de Gute y Basak (Gute B., et al., 1997) y el de Hall y Kier (Hall L. et al., 1984) , 0.15 (4.8%) y 0.19 (2.4%), respectivamente. Finalmente en la Fig. 4.22c muestra el desempeño del modelo con respecto a los compuestos que presenta grupos nitro adheridos al anillo, los resultados con fuzzy ARTMAP mejoran a los de modelos anteriores en un factor de diez.

La enorme diferencia entre los modelos QSAR previos con MLR y el modelo actual con fuzzy ARTMAP no puede atribuirse a diferencias en calidad y cantidad de información molecular (descriptores) en ningún caso, pero si a la naturaleza no lineal y al desempeño de un clasificador cognitivo como fuzzy ARTMAP da a los modelos QSAR. Gute y Basak (Gute B., et al., 1997) construyen el MLR para el $LC_{50}$ con siete parámetros, en cambio Hall y Kier (Hall L. et al., 1984) usan el análisis MLR para obtener el coeficiente de la ecuación de Wilson para cada grupo funcional.

Con el objetivo de discernir el efecto de los descriptores moleculares de la capacidad propia de fuzzy ARTMAP, fue desarrollado un modelo QSAR-fuzzy ARTMAP para los índices propuestos por Gute y Basak (Gute B., et al., 1997). El error absoluto para el conjunto global de 69 compuestos presenta un error absoluto medio de 0.02 (0.58%) log que es comparable con 0.02 (0.46%) para el modelo propuesto previamente. Sin embargo, la generalización con para el modelo QSAR-fuzzy ARTMAP con los 10 índices propuestos es mejor para los 10 compuestos que forman el conjunto de prueba de este modelo. El error absoluto medio y la desviación estándar para el modelo QSAR-fuzzy ARTMAP con los índices de Gute y Basak (Gute B., et al., 1997) son 0.17 (3.84%) y 0.13 (2.67%), comparados con 0.14 (3.18%) y 0.11 (2.04%) log, para el modelo actual. Estas diferencias son debidas a cambios en la clasificación del 1-3-diclorobenceno, 3-nitrotolueno, y el p-creso. Ejemplificando, el modelo con 10 índices clasifica correctamente al p-cresol con su homólogo o-cresol y predice la toxicidad del primero con un error del 5.31%, mientras que el modelo usando los índice de Gute y Basak (Gute B., et al., 1997) clasifica este compuesto erróneamente en la familia de 3-clorotolueno y consecuentemente el error de predicción se incremente al 7.26%, como se muestra en la Tabla 4.6.

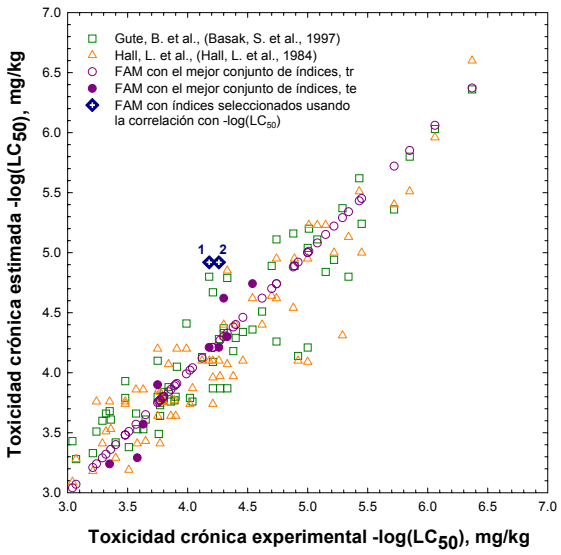



**Fig. 4.21.** Comparación de los valores estimados y predichos para el índice de toxicidad $-\log(LC_{50})$ de un conjunto de compuestos derivados del benceno

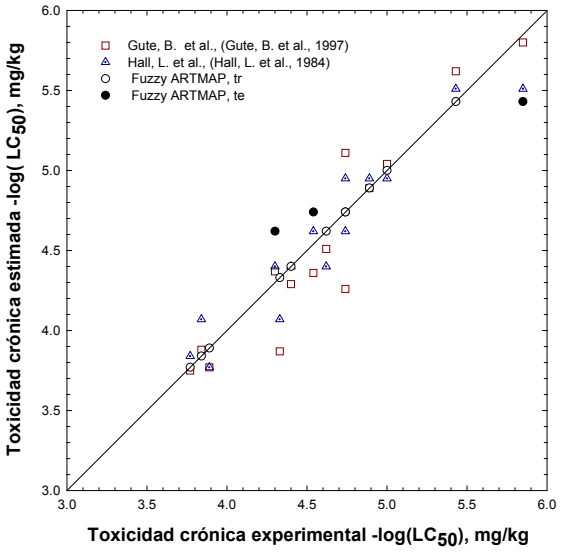



**Fig. 4.22a**

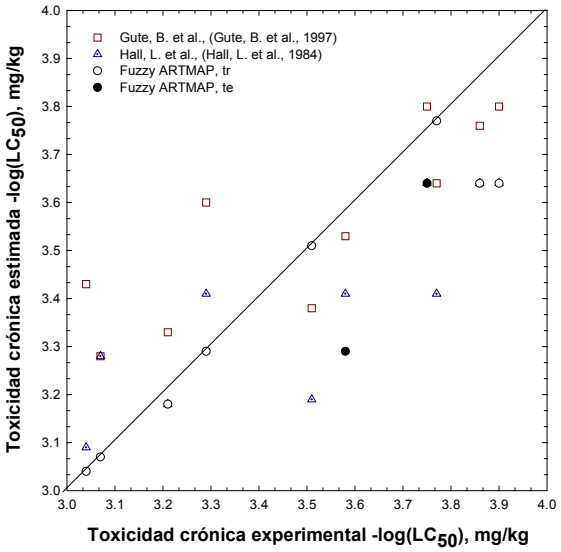



**Fig. 4.22b**

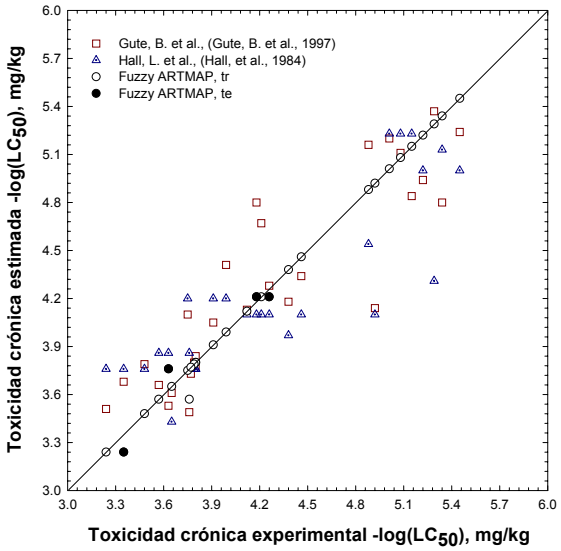



**Fig. 4.22c**

**Fig. 4.22.** Comparación de la influencia que ejerce la posición de los diferentes grupos funcionales en los valores del índice de toxicidad -log($LC_{50}$) experimentales y estimados para tres familias de compuestos derivados del benceno con grupos substituyentes (a) halógenos, (b) hidroxilos y (c) nitro

**Tabla 4.6.** Influencia de la posición de diferentes grupos funcionales sobre la generalización en el modelo de $LC_{50}$

Influencia sobre la toxicidad para los clorobencenos

| Nombre | fórmula | Exp. -log ($LC_{50}$) | Gute, B. et al. (1997) | Hall, L. et al. (1984) | -log ($LC_{50}$) χ(0-4), N, AP, NFL, ENA, NN |
|---|---|---|---|---|---|
| chlorobenzene | C 6 H 5 Cl 1 | 3.77 | 3.75 | 3.84 | 3.77 |
| 1-2-dichlorobenzene | C 6 H 4 Cl 2 | 4.40 | 4.29 | 4.40 | 4.40 |
| 1-3-dichlorobenzene | C 6 H 4 Cl 2 | 4.30 | 4.37 | 4.40 | 4.62 |
| 1-4-dichlorobenzene | C 6 H 4 Cl 2 | 4.62 | 4.51 | 4.40 | 4.62 |
| 1-2-3-trichlorobenzene | C 6 H 3 Cl 3 | 4.89 | 4.89 | 4.89 | 4.89 |
| 1-2-4-trichlorobenzene | C 6 H 3 Cl 3 | 5.00 | 5.04 | 5.00 | 5.00 |
| 1-3-5-trichlorobenzene | C 6 H 3 Cl 3 | 4.74 | 5.11 | 4.74 | 4.74 |
| 1-2-3-4-tetrachlorobenzene | C 6 H 2 Cl 4 | 5.43 | 5.62 | 5.43 | 5.43 |
| 1-2-4-5-tetrachlorobenzene | C 6 H 2 Cl 4 | 5.85 | 5.80 | 5.85 | 5.43 |
| Influencia sobre la toxicidad para los compuestos que contienen hidroxilos y halógenos | | | | | |
| Nombre | fórmula | Exp. -log ($LC_{50}$) | Gute, B. et al. (1997) | Hall, L. et al. (1984) | -log ($LC_{50}$) χ(0-4), N, AP, NFL, ENA, NN |
| 2-3-4-5-6-pentachlorophenol | C 6 H 1 O 1 Cl 5 | 6.06 | 6.03 | 5.96 | 6.06 |
| 2-3-4-5-tetrachlorophenol | C 6 H 2 O 1 Cl 4 | 5.72 | 5.36 | 5.40 | 5.72 |
| 2-4-6-tribromophenol | C 6 H 3 O 1 Br 3 | 4.70 | 4.89 | 4.64 | 4.70 |
| 2-4-6-trichlorophenol | C 6 H 3 O 1 Cl 3 | 4.33 | 4.79 | 4.85 | 4.30 |

| | | | | | |
|---|---|---|---|---|---|
| 2-4-dichlorophenol | C 6 H 4 O 1 Cl 2 | 4.30 | 4.33 | 4.30 | 4.30 |
| Influencia sobre la toxicidad de los derivados del benceno | | | | | |
| Nombre | fórmula | Exp. - log($LC_{50}$) | Gute, B. et al. (1997) | Hall, L. et al. (1984) | -log ($LC_{50}$) $\chi$(0-4), N, AP, NFL, ENA, NN |
| toluene | C 7 H 8 | 3.32 | 3.66 | 3.51 | 3.32 |
| 1-2-dimethylbenzene | C 8 H 10 | 3.48 | 3.93 | 3.74 | 3.48 |
| 1-4-dimethylbenzene | C 8 H 10 | 4.21 | 3.87 | 3.74 | 4.21 |
| 1-2-4-trimethyl-benzene | C 9 H 12 | 4.21 | 4.09 | 3.96 | 4.21 |
| benzene | C 6 H 6 | 3.40 | 3.42 | 3.29 | 3.40 |
| Influencia sobre la toxicidad de los derivados del benceno | | | | | |
| Nombre | fórmula | Exp. - log($LC_{50}$) | Gute, B. et al. (1997) | Hall, L. et al. (1984) | -log ($LC_{50}$) $\chi$(0-4), N, AP, NFL,ENA, NN |
| 4-chloro-3-methyl-phenol | C 7 H 7 O 1 Cl 1 | 4.27 | 3.87 | 3.97 | 4.27 |
| 2-4-dinitrophenol | C 6 H 4 N 2 O 5 | 4.04 | 3.76 | 3.87 | 4.04 |
| 2-methyl-4-6-dinitrophenol | C 7 H 6 N 2 O 5 | 5.00 | 4.21 | 4.09 | 5.00 |
| 4-nitrophenol | C 6 H 5 N 1 O 3 | 3.36 | 3.61 | 3.53 | 3.36 |
| 2-chlophenol | C 6 H 5 O 1 Cl 1 | 4.02 | 3.79 | 3.74 | 4.02 |

En el caso del modelo obtenido para el $LD_{50}$, los grupos de índices obtenidos mediante SOM a partir de la correlación lineal de los planos C para cada una de los descriptores moleculares, se presenta en la Fig. 4.23. Los grupos formados por SOM son consistentes con el comportamiento de las variables con respecto a su covarianza con la variable objetivo ($LD_{50}$), la matriz de covarianza se presenta el la Tabla 7 (Tabla 7, apéndice 3; Espinosa G. et al., 2002). El análisis de este grupo de compuestos muestra que:

(i) El primer grupo de índices se integra por el índice de conectividad de orden cero y cuatro, los cuales se correlacionan con el índice de repulsión núcleo-núcleo, el índice kappa, y el de número de niveles ocupados.

(ii) Los índices cuánticos de energía de intercambio, el índice de atracción electrón-núcleo y el índice de resonancia se correlacionan entre ellos y forman el segundo grupo de índices.

(iii) En el tercer grupo se agrupan los índices de conectividad de orden uno, dos y tres, junto con la polarizabilidad media y la suma de números atómicos.

(iv) Finalmente, el cuarto grupo integra a los índices propuestos por Hansen, de polaridad y él referente al hidrógeno, junto con el momento dipolar.

Es significativo establecer una comparación entre los grupos formados por SOM para el conjunto homogéneo como el del $LC_{50}$ (Fig. 4.20)y uno heterogéneo como el del $LD_{50}$ (Fig. 4.23). Para el grupo de compuestos derivados del benceno toda la información contenida en los índices de conectividad y en la suma de números atómicos (la mayoría agrupados en el grupo I de la figura 4.20) resultó muy relevantes en términos tanto de sus

covarianzas con el índice $LC_{50}$ en la información aportada al modelo QSAR (forman 6 de los 6 índices seleccionados como base para el modelo QSAR). Al trabajar con un conjunto más heterogéneo reduce el impacto de los índices bidimensionales en la caracterización y se hace más evidente la necesidad de información que caracterice las interacciones espaciales o la presencia de heteroátomos mediante índices de carácter cuántico.

Una vez más, la comparación entre disimilitud de pares de mapas formados por diferentes subconjuntos de índices a partir del conjunto de descriptores moleculares base permite determinar el conjunto más factible de descriptores capaces de caracterizar la relación entre las estructuras estudiadas, Tabla 8 (Tabla 8, apéndice 3, Espinosa G et al., 2002). El mejor conjunto de índices se compone por diez descriptores diversos, los índices de conectividad molecular de orden cero, uno y cuatro, el índice de forma kappa, el índice en energía de intercambio, la polarizabilidad de Hansen, el índice de repulsión núcleo-núcleo, el de energía de resonancia, el de número de niveles ocupados y el de atracción electrón-núcleo. Los cambios en el conjunto de índices seleccionados con respecto a los índices usados para el $LC_{50}$ son la sustitución de la conectividad de orden dos y tres, la suma de números atómicos y la polarizabilidad media por la energía de intercambio, la polarizabilidad de Hansen, la energía de resonancia, y el índice de forma kappa.

La predicción global del QSAR-fuzzy ARTMAP tiene un error medio de 0.02 log (0.53%), cuya desviación estándar es de 0.07 (1.84%) log. Los errores y sus desviaciones estándar relativos al conjunto de entrenamiento y generalización son 0.0 (0.06%) y 0.07 (1.84%) log y 0.13 (3.68%) y 0.15 (3.92%), respectivamente. El mayor error es del 12.5% corresponde al n-propil acetato el cual no es debido a un error en la clasificación ya que es identificado como parte de la familia del isopropil acetato. Los resultados se presentan gráficamente en la Fig. 4.24.

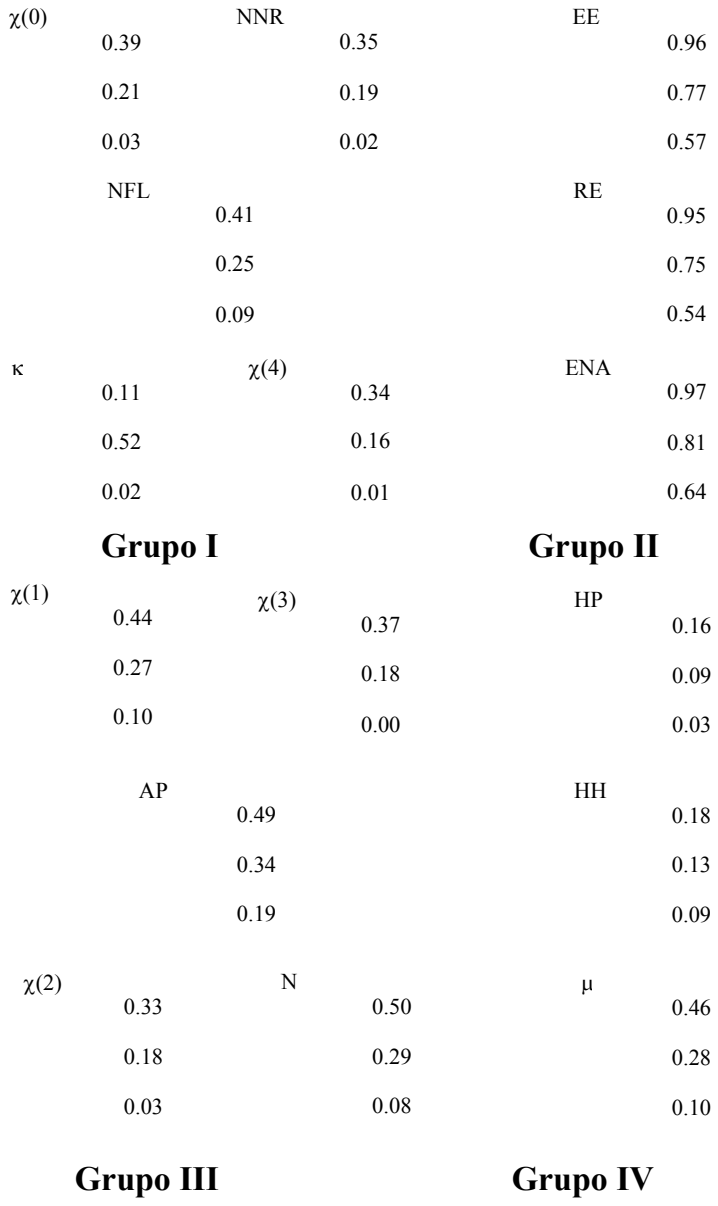



**Fig. 4.23.** Agrupación de los planos de descriptor proyectado sobre el mapa de $LD_{50}$. Los tonos de grises representan distancias relativas

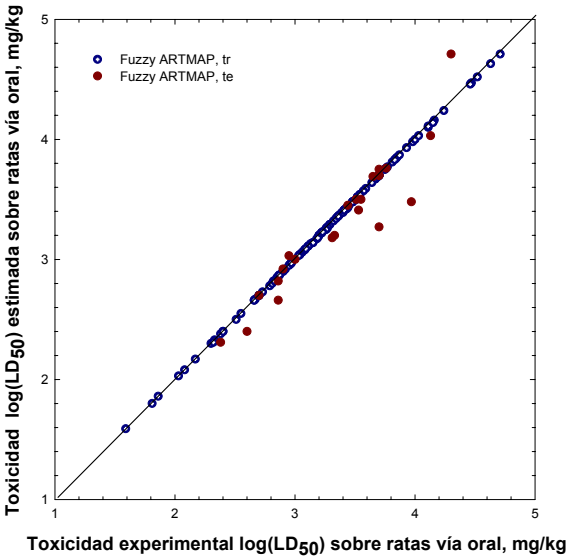



**Fig. 4.24.** Modelo QSAR fuzzy ARTMAP para estimar el índice de toxicidad administrado por vía oral sobre ratas log ($LD_{50}$) de un conjunto de compuestos orgánicos diversos

### 4.2.2 Carcinogénesis

El cáncer representa una de las principales causas de muerte en nuestro siglo. Es por ello necesario identificar los factores que participan en su origen y desarrollo posterior. En este sentido, se ha determinado que los *factores endógenos* -de la propia persona- son responsables de un 20% de los cánceres, mientras que en el 80% restante podrían *intervenir factores exógenos o medioambientales*, tales como: i) el tabaco (responsable del 20-30% de los tumores en el hombre y del 5-10% de los tumores en la mujer); ii) factores ambientales, laborales, virus, etc. (explican de un 20-30% de los cánceres).

La carcinogénesis o aparición del cáncer es el resultado de dos procesos sucesivos: el aumento descontrolado de un grupo de células que da lugar a un tumor o neoplasia, y la posterior, diseminación masiva desde su sitio natural en el organismo colonizando y proliferando en otros tejidos u órganos (metástasis).

Algunas substancias químicas peligrosas están relacionadas con el desarrollo del cáncer. Un individuo puede quedar expuesto a cancerígenos ambientales en su sitio de trabajo, debido a sus hábitos personales y por los alimentos que ingiere. Algunos riesgos son ubicuos (incremento en el desarrollo de cáncer de piel por exposición a la luz solar) y otros específicos de un estilo de vida.

Tras cincuenta años de estudios epidemiológicos, existe una larga serie de carcinógenos conocidos o sospechados, así como otros cuyo papel es mucho más discutido, como la radiación no ionizante o ciertos factores dietéticos. Ante la falta de evidencias concretas con éstos últimos el esfuerzo se está dirigiendo hacia la prevención de toda exposición intencionada e irresponsable de carcinógenos conocidos. Es por ello, que el siguiente caso de estudio se enfoca al estudio de la capacidad carcinogénica de 104 compuestos aromáticos nitrogenados, previamente estudiados por Gini et al., (Gini G. et al., 1999). Los valores de actividad carcinogénica, medidos a través del índice $TD_{50}$, representan a la dosis en mg/kg que al ser administrado sobre han causado tumores en la mitad de la población sobre la que se realizaron los estudios.

La elección de este conjunto de datos de hecho nos permite comparar el desempeño de los modelos fuzzy ARTMAP con los obtenidos por Gini et al. (Gini G. et al., 1999) con una arquitectura backpropagation. Además, la cualidad de los índices elegidos por SOM puede contrastarse con los 13 índices propuestos por estos autores después de una barrido cuidadosos del conjunto original formado por 34 índices entre electrostáticos, topológicos, químico cuánticos y fisicoquímicos, tal como se describe en la referencia original (Gini G. et al., 1999).

SOM se utiliza en la selección de los descriptores más relevantes para el modelo fuzzy ARTMAP QSAR. El conjunto de descriptores en este caso incorpora tanto información topológica como cuántica. Los mapas topológicos para cada descriptor molecular junto con la variable correspondiente a la actividad carcinogénica de cada compuesto permiten clasificar los compuestos en categorías, Fig. 4.25. Seleccionando el elemento de cada categoría con la correlación más alta con respecto a la actividad y posteriormente en orden decreciente de correlación. Nuevamente, el proceso concluye cuando la medida de disimilitud entre pares de mapas alcanza un mínimo, indicando que incluir un nuevo índice no incorporará información nueva a la organización de los compuestos. Este conjunto de índices será la base para el modelo final fuzzy ARTMAP-QSAR.

La lista completa de los compuestos junto con el valor experimental del índice de toxicidad, $TD_{50}$, se muestra en la Tabla 4.7. En esta tabla también se encuentran los valores estimados por Gini et al., (Gini G. et al., 1999). Se ha utilizado el método semi-empírico PM3 de MOPAC 6.0 para optimizar cada una de las estructuras y obtener un conjunto de índices cuánticos adicionales. Los índices utilizados son HOMO, LUMO, Wiener (W), kappa ($^{2}\kappa$), índices de conectividad molecular ($^{0}\chi^{v}$, $^{1}\chi^{v}$, $^{2}\chi^{v}$, $^{3}\chi^{v}$, $^{4}\chi^{v}$), la suma de número atómicos (N), la polarizabilidad promedio (Pol), el momento dipolar ($\mu$), y la energía total (TE). Los índices de Randic (Ran) y Balaban (Bal) tomados de la referencia de Gini et al., (Gini G. et al., 1999). Como se ha mencionado a lo largo del documento, hay un sin número de índices,

sin embargo hacer un estudio exhaustivo de los mismos no es el objetivo de este trabajo. En lugar de ello, intentaremos incrementar nuestro conjunto, con un par de índices cuyo fundamento teórico posea la mayor cantidad de información. Los índices de similitud cuántica nos proporcionan dicha información, al establecer las relaciones cuantitativas entre la similitud de las estructuras moleculares por medio de la proyección de sus funciones de densidad. Los operadores más utilizados son el operador Overlap (Ove), el operador de Coulomb (Cou) y Cinético (Kin). Al analizar un conjunto de moléculas se calculan todos los pares que son incorporados a la matriz correspondiente (MQS), los elementos de la diagonal de dicha matriz constituyen las auto similitudes, que serán usadas como descriptores moleculares. Aunque el número de descriptores contenidos en cada matriz podría parecer excesivo, no se utilizan todos ellos para describir una propiedad, en general, es posible combinar los elementos de la diagonal, con aquellos elementos de cada matriz que resulten más significativos.

Se realizan diferentes experimentos de forma que es posible cuantificar el aporte debido al sistema neuronal, a algún tipo de índice en particular, como los índices de similitud molecular, o a la combinación de los mismos. De la misma forma, se compara con los resultados de Gini et al., (Gini et al., 1999). Los experimentos pueden resumirse de la siguiente forma:

(i) Se utilizan los tres operadores cuánticos, Overlap, Coulomb, y cinético para entrenar una red fuzzy ARTMAP.
(ii) Un modelo fuzzy ARTMAP QSAR se construye usando un conjunto de índices seleccionados en orden decreciente de correlación, sin tomar en cuenta las categorías formadas por SOM.
(iii) Un modelo fuzzy ARTMAP QSAR se estima usando la metodología propuesta SOM-fuzzy ARTMAP.
(iv) Un modelo QSAR usando fuzzy ARTMAP se obtiene utilizando los índices propuestos por Gini et al., (Gini G. et al., 1999).
(v) Finalmente, se construye un backpropagation a partir de los mejores índices seleccionados con la metodología SOM-fuzzy ARTMAP.

Un resumen de cada experimento lo encontraremos en la Tabla 4.8 Algunos puntos importantes a recalcar de cada caso son los siguientes, el experimento uno pone de manifiesto que con tres índices de similitud cuántica y fuzzy ARTMAP se mejora en un factor de dos con respecto a las predicciones de Gini et al., (Gini G. et al., 1999) con 13 índices. Con lo cual se observa el aporte significativo que las tres auto similitudes proveen a fuzzy ARTMAP para cuantificar la relación estructura-actividad.

El experimento dos ignora la metodología propuesta y selecciona por prueba y error los índices para los cuales los resultados son los más satisfactorios, en ellos se toman en cuenta algunos elementos de las matrices de similitud cuántica y se van añadiendo índices tanto topológicos

como cuánticos de forma que se logre ir reduciendo el error absoluto medio entre los valores experimentales y los estimados. Los errores citados en la Tabla 4.8 dan una idea del aporte de información que le atorga al conjunto original de auto similitudes cuánticas, el incluir los elementos cruzados de Overlap, la polarizabilidad, la suma de números atómicos y el momento dipolar con lo cual la precisión del modelo se incrementa en un factor de dos con respecto al primer experimento planteado en esta sección y por un factor de tres con respecto a los resultados de Gini et al., (Gini G. et al., 1999).

Al aplicar la metodología completa SOM-fuzzy ARTMAP en el experimento tres, se obtiene una mejora significativa, los errores decrecen en un factor de diez con respecto al modelo referenciado. La Fig. 4.26, nos da una idea de la comparación propiamente dicha, la diferencia en los resultados con respecto al modelo publicado por Gini et al., (Gini G. et al., 1999) es de un factor de 10. El conjunto de índices que da los mejores resultados se forma por las tres medidas de auto similitud (Ove, Cou, Kin), y el elemento 101 (Ove101 [Phenacetin]), de la matriz de Overlap, además de la polarizabilidad (Pol), la energía total (TE), el índice 3D de Wiener (W), los índices de conectividad de orden cero a dos ($^{0\text{-}2}\chi$), el índice kappa ($^{2}\kappa$), el de Balaban (Bal) y el de Randic (Ran). Es interesante destacar, que el error más significativo en la predicción del $TD_{50}$ corresponde al 2-4-dinitrotolueno, para el cual su valor de toxicidad estimado es de $TD_{50}$=0.288 en lugar de cero. Siendo precisamente la clase de los compuestos totalmente anti-carcinogénicos (carcinogenicidad cero) en la cual el modelo de Gini et al., (Gini G. et al., 1999) obtienen los errores más significativos, incluyendo valores negativos den el índice de toxicidad. Esto sugiere que se necesitan mejores índices para un modelo fuzzy ARTMAP. Cabe mencionar, que la combinación de índices propuesta no es un conjunto universal, dependerá siempre del conjunto de índices disponibles, es decir, que los descriptores seleccionados, son aquellos que dentro de los índices disponibles aportan la mayor información para el conjunto estudiado. Esto se plasma en el experimento 4, en el cual con un conjunto e índices diferente, (el utilizado por Gini G et al., 1999 ) da resultados similares a los obtenidos usando la metodología SOM fuzzy ARTMAP. En este último experimento no se produce ninguna clasificación errónea de los compuestos con carcinogenicidad cercana a cero, aunque los valores estimados tienen una dispersión mayor a la observada en el experimento 3, Fig. 4.26. La similitud entre los resultados de los experimentos 3 y 4 ponen de manifiesto que la selección de índices propuesta por Gini et al., (Gini G. et al., 1999) usando PCA da lugar a un excelente grupo de índices. Es por ello que resulta de interés estudiar la intersección entre ambos conjuntos de variables, y estudiar su efecto en un nuevo modelo QSAR fuzzy ARTMAP.

Finalmente, el experimento 5, es un modo de comparación entre backpropagation y fuzzy ARTMAP y la calidad de ambos conjuntos de índices

independientemente del modelo de red neuronal elegido. La arquitectura de backpropagation es 14-7-1, con la cual se obtienen resultados similares a los sugeridos por Gini et al., con una arquitectura 13-6-1. Sin embargo, debe mencionarse que en este caso los errores citados por el modelo de Gini corresponden a los valores publicados en el trabajo original, el cual excluye 8 compuestos a los que denomina *outlieres*. Para una descripción más detallada de cada uno de los casos, se hace referencia ala referencia original, Fig. 4.26 (Espinosa G., et al., 2002).

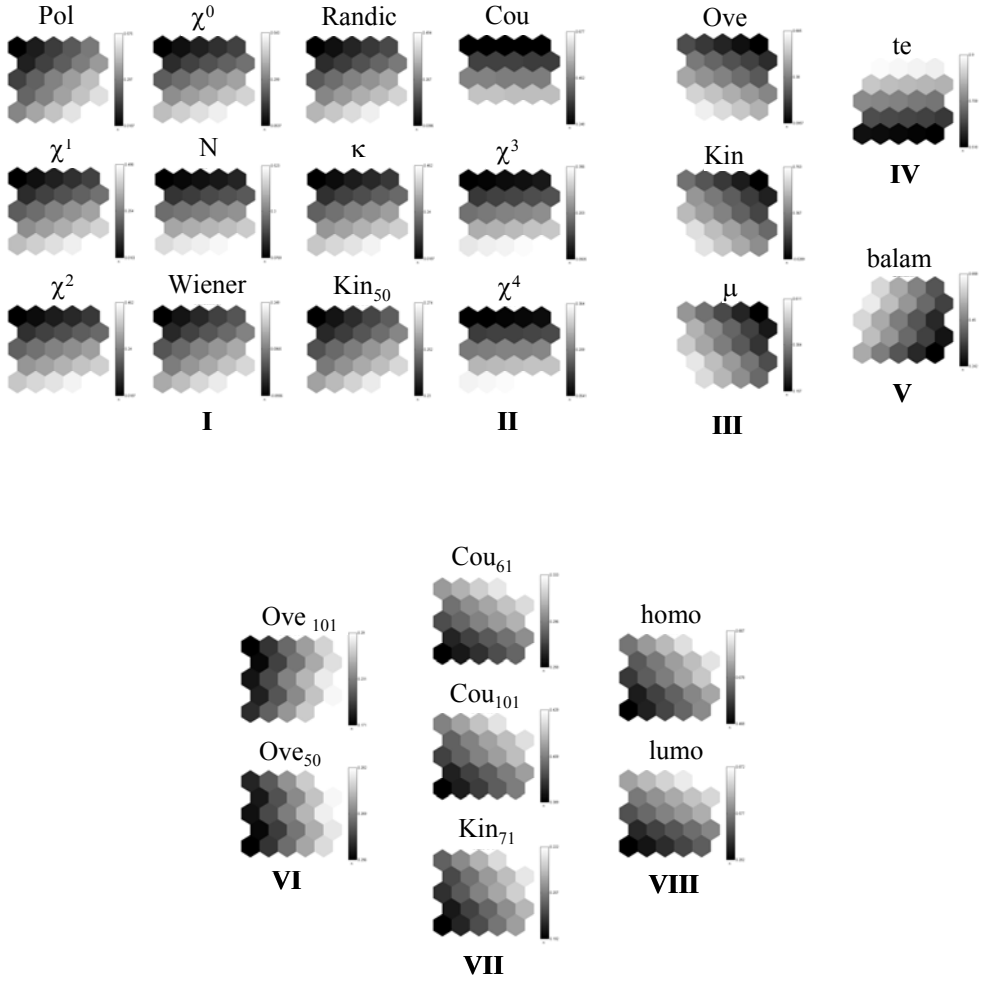



**Fig. 4.25.** Agrupación de los planos para cada descriptor proyectado sobre el mapa de $TD_{50}$. para un grupo de compuestos aromáticos nitrogenados.

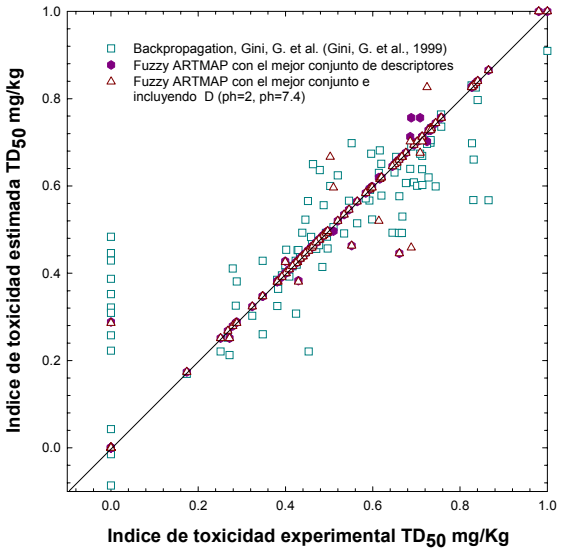



**Fig. 4.26.** Modelo QSAR fuzzy ARTMAP para estimar el índice de $TD_{50}$ de un conjunto de compuestos orgánicos diversos

**Tabla 4.7**. Lista de compuestos considerados en el presente caso de estudio, con los valores experimentales correspondientes al índice de toxicidad $TD_{50}$ (mg/kg) y el cuadrado de los errores absolutos medios

| Nombre | id | Exp. $TD_{50}$ | FA $TD_{50}$, Best set | AMSE | BP $TD_{50}$ Gini, G. et al. 1999 | AMSE |
|---|---|---|---|---|---|---|
| 5-nitro-o-anisidine | tr | 0.4276 | 0.4276 | 0.0000 | 0.4530 | 0.0006 |
| 2-amino-4-nitrophenol | tr | 0.4384 | 0.4384 | 0.0000 | 0.4929 | 0.0030 |
| 4-nitro-o-phenylenediamine | te | 0.0000 | 0.0000 | 0.0000 | 0.3094 | 0.0957 |
| o-aminoazotoluene | tr | 0.5936 | 0.5936 | 0.0000 | 0.5913 | 0.0000 |
| 4-chloro-o-phenylenediamine | tr | 0.4233 | 0.4234 | 0.0000 | 0.4286 | 0.0000 |
| 3-chloro-p-toluidine | tr | 0.3807 | 0.3807 | 0.0000 | 0.3849 | 0.0000 |
| Proflavine.HCl hemihydrate | tr | 0.6535 | 0.6535 | 0.0000 | 0.6667 | 0.0002 |
| 4-aminodiphenyl | tr | 0.8312 | 0.8306 | 0.0000 | 0.6604 | 0.0292 |
| 3,3'-dimethoxybenzidine-4,4'-diisocyanate | tr | 0.2791 | 0.2791 | 0.0000 | 0.4109 | 0.0174 |
| 2-naphthylamine | tr | 0.6557 | 0.6557 | 0.0000 | 0.6456 | 0.0001 |
| 2-sec-butyl-4,6-dinitrophenol | tr | 0.8360 | 0.8360 | 0.0000 | 0.8256 | 0.0001 |
| Cinnamyl anthranilate | tr | 0.4017 | 0.4017 | 0.0000 | 0.4539 | 0.0027 |
| 1-(1-naphthyl)-2-thiourea | tr | 0.8274 | 0.8266 | 0.0000 | 0.6979 | 0.0168 |
| n-nitrosodiphenylamine | tr | 0.4952 | 0.4952 | 0.0000 | 0.4837 | 0.0001 |
| Pentachloronitrobenzene | tr | 0.6161 | 0.6161 | 0.0000 | 0.6816 | 0.0043 |
| 1-amino-2-methylanthraquinone | te | 0.5516 | 0.4630 | 0.0078 | 0.6981 | 0.0215 |
| Dapsone | tr | 0.0000 | 0.0000 | 0.0000 | 0.4293 | 0.1843 |
| 1-[(5-nitrofurfurylidene)amino]hydantoin | tr | 0.4588 | 0.4588 | 0.0000 | 0.4831 | 0.0006 |
| 4,4'-methylene-bis(2-chloroaniline).2HCl | te | 0.6141 | 0.6194 | 0.0000 | 0.6300 | 0.0003 |
| m-toluidine.HCl | tr | 0.3831 | 0.3831 | 0.0000 | 0.3642 | 0.0004 |
| 2,4-diaminotoluene.2HCl | tr | 0.5643 | 0.5643 | 0.0000 | 0.5146 | 0.0025 |
| o-toluidine.HCl | te | 0.4296 | 0.3831 | 0.0022 | 0.4531 | 0.0006 |
| 2,4,6-trimethylaniline.HCl | tr | 0.6498 | 0.6498 | 0.0000 | 0.6310 | 0.0004 |
| Phenacetin | tr | 0.2859 | 0.2859 | 0.0000 | 0.3255 | 0.0016 |
| p-phenylenediamine.2HCl | tr | 0.3813 | 0.3807 | 0.0000 | 0.3249 | 0.0032 |
| 4-nitroanthranilic acid | tr | 0.2882 | 0.2882 | 0.0000 | 0.3812 | 0.0086 |
| 2-chloro-p-phenylenediamine sulfate | tr | 0.4001 | 0.4001 | 0.0000 | 0.4022 | 0.0000 |
| o-phenylenediamine.2HCl | tr | 0.4333 | 0.4333 | 0.0000 | 0.4379 | 0.0000 |
| 2,6-dichloro-p-phenylenediamine | tr | 0.4405 | 0.4405 | 0.0000 | 0.4430 | 0.0000 |
| 5-nitroacenaphthene | tr | 0.6194 | 0.6194 | 0.0000 | 0.6508 | 0.0010 |
| 5-nitro-2-furaldehyde semicarbazone | tr | 0.6600 | 0.6600 | 0.0000 | 0.4923 | 0.0281 |
| Methotrexate | tr | 0.6450 | 0.6450 | 0.0000 | 0.4927 | 0.0232 |
| Hydrochlorothiazide | tr | 0.4514 | 0.4514 | 0.0000 | 0.5654 | 0.0130 |
| Pyrimethamine | tr | 0.5199 | 0.5199 | 0.0000 | 0.6243 | 0.0109 |
| Furosemide | tr | 0.4876 | 0.4876 | 0.0000 | 0.5560 | 0.0047 |
| m-phenylenediamine.2HCl | tr | 0.4844 | 0.4844 | 0.0000 | 0.4144 | 0.0049 |
| 2,4-dimethoxyaniline.HCl | tr | 0.4257 | 0.4257 | 0.0000 | 0.4197 | 0.0000 |
| 2-acetylaminofluorene | tr | 0.7563 | 0.7563 | 0.0000 | 0.7638 | 0.0001 |
| Benzidine.2HCl | te | 0.7086 | 0.7563 | 0.0023 | 0.6738 | 0.0012 |
| 2-nitro-p-phenylenediamine | tr | 0.4532 | 0.4532 | 0.0000 | 0.2208 | 0.0540 |
| 2,5-xylidine.HCl | tr | 0.4458 | 0.4458 | 0.0000 | 0.5227 | 0.0059 |
| D & c red no. 9 | tr | 0.4336 | 0.4333 | 0.0000 | 0.4384 | 0.0000 |
| 4-chloro-m-phenylenediamine | tr | 0.4088 | 0.4088 | 0.0000 | 0.3924 | 0.0003 |
| 2,4-dinitrophenol | te | 0.0000 | 0.2882 | 0.0831 | -0.0145 | 0.0002 |
| fd & c red no. 4 | tr | 0.2512 | 0.2512 | 0.0000 | 0.2209 | 0.0009 |
| (n-6)-methyladenine | tr | 0.0000 | 0.0000 | 0.0000 | 0.4462 | 0.1991 |
| Metronidazole | tr | 0.4927 | 0.4927 | 0.0000 | 0.4924 | 0.0000 |
| Dacarbazine | tr | 0.8653 | 0.8653 | 0.0000 | 0.5674 | 0.0887 |
| 2,2,2-trifluoro-n-[4-(5-nitro-2-furyl)-2-thiazolyl]acetamide | tr | 0.7321 | 0.7321 | 0.0000 | 0.6992 | 0.0011 |
| 2,4-diaminoanisole sulfate | tr | 0.4965 | 0.4965 | 0.0000 | 0.4567 | 0.0016 |
| 2-amino-4-(5-nitro-2-furyl)thiazole | te | 0.7243 | 0.7018 | 0.0005 | 0.6966 | 0.0008 |
| 2-aminodiphenylene oxide | tr | 0.7344 | 0.7344 | 0.0000 | 0.7324 | 0.0000 |

| Nombre | id | Exp. $TD_{50}$ | FA $TD_{50}$, Best set | AMSE | BP $TD_{50}$ Gini, G. et al. 1999 | AMSE |
|---|---|---|---|---|---|---|
| AF-2 | tr | 0.5922 | 0.5922 | 0.0000 | 0.5664 | 0.0007 |
| Formic acid 2-[4-(5-nitro-2-furyl)-2-thiazolyl]hydrazide | tr | 0.7277 | 0.7277 | 0.0000 | 0.6196 | 0.0117 |
| 3-(3,4-dichlorophenyl)-1,1-dimethylurea | tr | 0.4788 | 0.4788 | 0.0000 | 0.6364 | 0.0248 |
| 4'-fluoro-4-aminodiphenyl | tr | 0.8306 | 0.8306 | 0.0000 | 0.5675 | 0.0692 |
| 4-chloro-o-toluidine.HCl | tr | 0.6942 | 0.6942 | 0.0000 | 0.6084 | 0.0074 |
| Mexacarbate | tr | 0.8264 | 0.8266 | 0.0000 | 0.8305 | 0.0000 |
| Chlorambucil | tr | 1.0000 | 1.0000 | 0.0000 | 0.9094 | 0.0082 |
| c.i. disperse yellow 3 | tr | 0.4769 | 0.4769 | 0.0000 | 0.4644 | 0.0002 |
| fd & c yellow no. 6 | te | 0.2717 | 0.2512 | 0.0004 | 0.2126 | 0.0035 |
| 2-hydrazino-4-(p-aminophenyl)thiazole | tr | 0.7018 | 0.7018 | 0.0000 | 0.6003 | 0.0103 |
| 2-hydrazino-4-(p-nitrophenyl)thiazole | te | 0.7134 | 0.7129 | 0.0000 | 0.6021 | 0.0124 |
| 2-hydrazino-4-(5-nitro-2-furyl)thiazole | te | 0.6857 | 0.7129 | 0.0007 | 0.6391 | 0.0022 |
| n-[5-(5-nitro-2-furyl)-1,3,4-thiadiazol-2-yl]acetamide | tr | 0.7440 | 0.7440 | 0.0000 | 0.5990 | 0.0210 |
| n-[4-(5-nitro-2-furyl)-2-thiazolyl]formamide | te | 0.7325 | 0.7277 | 0.0000 | 0.7051 | 0.0008 |
| 1,5-naphthalenediamine | tr | 0.5838 | 0.5838 | 0.0000 | 0.5713 | 0.0002 |
| n-phenyl-p-phenylenediamine.HCl | tr | 0.0000 | 0.0000 | 0.0000 | 0.4836 | 0.2339 |
| 2-biphenylamine.HCl | tr | 0.4241 | 0.4234 | 0.0000 | 0.3075 | 0.0136 |
| Fluometuron | tr | 0.5344 | 0.5344 | 0.0000 | 0.4913 | 0.0019 |
| 2,4-xylidine.HCl | te | 0.6608 | 0.4458 | 0.0462 | 0.5765 | 0.0071 |
| 2-amino-4-(p-nitrophenyl)thiazole | tr | 0.7133 | 0.7129 | 0.0000 | 0.6690 | 0.0020 |
| p-anisidine.HCl | tr | 0.0000 | 0.0000 | 0.0000 | 0.3522 | 0.1240 |
| Atrazine | te | 0.6881 | 0.7563 | 0.0047 | 0.6902 | 0.0000 |
| Nitrofen | tr | 0.6198 | 0.6194 | 0.0000 | 0.5780 | 0.0017 |
| 3-nitro-p-acetophenetide | te | 0.3995 | 0.4276 | 0.0008 | 0.4186 | 0.0004 |
| 2,2',5,5'-tetrachlorobenzidine | tr | 0.5963 | 0.5963 | 0.0000 | 0.6738 | 0.0060 |
| p-nitrosodiphenylamine | te | 0.5024 | 0.4952 | 0.0001 | 0.6000 | 0.0095 |
| Melphalan | te | 0.9803 | 1.0000 | 0.0004 | 1.0032 | 0.0005 |
| n-(1-naphthyl)ethylenediamine.2HCl | te | 0.0000 | 0.0000 | 0.0000 | 0.2226 | 0.0496 |
| Aniline.HCl | tr | 0.2679 | 0.2679 | 0.0000 | 0.2523 | 0.0002 |
| Nithiazide | tr | 0.4609 | 0.4609 | 0.0000 | 0.4735 | 0.0002 |
| 2,4,5-trimethylaniline | tr | 0.7129 | 0.7129 | 0.0000 | 0.6384 | 0.0056 |
| 4,4'-methylenedianiline.2HCl | tr | 0.6760 | 0.6760 | 0.0000 | 0.6068 | 0.0048 |
| o-anisidine.HCl | tr | 0.4162 | 0.4162 | 0.0000 | 0.4160 | 0.0000 |

| Nombre | id | Exp. $TD_{50}$ | FA $TD_{50}$, Best set | AMSE | BP $TD_{50}$ Gini, G. et al. 1999 | AMSE |
|---|---|---|---|---|---|---|
| Chloramben | tr | 0.3473 | 0.3473 | 0.0000 | 0.2602 | 0.0077 |
| 2-methyl-1-nitroanthraquinone | tr | 0.8404 | 0.8404 | 0.0000 | 0.7969 | 0.0019 |
| 2-amino-5-nitrophenol | tr | 0.3238 | 0.3238 | 0.0000 | 0.3026 | 0.0004 |
| 2-amino-5-nitrothiazole | tr | 0.0000 | 0.0000 | 0.0000 | -0.0862 | 0.0074 |
| 2,4-dinitrotoluene | te | 0.0000 | 0.2882 | 0.0831 | 0.3873 | 0.1500 |
| p-cresidine | tr | 0.5986 | 0.5986 | 0.0000 | 0.5234 | 0.0057 |
| 4-amino-2-nitrophenol | tr | 0.0000 | 0.0000 | 0.0000 | 0.2578 | 0.0665 |
| Anthranilic acid | tr | 0.1737 | 0.1737 | 0.0000 | 0.1693 | 0.0000 |
| 2-aminoanthraquinone | tr | 0.4630 | 0.4630 | 0.0000 | 0.6501 | 0.0350 |
| Melamine | tr | 0.3475 | 0.3473 | 0.0000 | 0.4286 | 0.0066 |
| p-chloroaniline | tr | 0.3917 | 0.3917 | 0.0000 | 0.3951 | 0.0000 |
| Acetaminophen | tr | 0.0000 | 0.0000 | 0.0000 | 0.3215 | 0.1034 |
| Azobenzene | tr | 0.7571 | 0.7563 | 0.0000 | 0.7360 | 0.0004 |
| m-cresidine | te | 0.5100 | 0.4965 | 0.0002 | 0.5057 | 0.0000 |
| 4,4'-oxydianiline | tr | 0.6680 | 0.6680 | 0.0000 | 0.5299 | 0.0191 |
| p-isopropoxydiphenylamine | tr | 0.4703 | 0.4703 | 0.0000 | 0.4558 | 0.0002 |
| 4,4'-methylenebis(n,n-dimethyl)benzenamine | tr | 0.5456 | 0.5456 | 0.0000 | 0.5662 | 0.0004 |
| Phenylhydrazine | tr | 0.0000 | 0.0000 | 0.0000 | 0.0428 | 0.0018 |
| (n-6)-(methylnitroso)adenine | tr | 0.6665 | 0.6665 | 0.0000 | 0.4923 | 0.0303 |

**Tabla 4.8**. QSAR usando NN para estimar la carcinogenicidad de compuestos aromáticos nitrogenados. Los modelos hacen referencia a los 5 casos de estudio presentados en el texto.

| *Datos* | *No. de muestras* | *Error promedio total* | *El cuadrado de los errores promedio* |
|---|---|---|---|
| *Experimento 1:Fuzzy ARTMAP usando [Ove, Cou, Kin]* | | | |
| **Todos los datos** | 104 | 0.0406 | 0.0186 |
| **Tr** | 85 | 0.0023 | 0.0000 |
| ***Te*** | 19 | 0.2116 | 0.1017 |
| *Experimento 2: Fuzzy ARTMAP usando [Ove, Cou, Kin, N, Pol, $^{1}\chi$, Ove50, Ove101]* | | | |
| **Todos los datos** | 104 | 0.0240 | 0.0071 |
| **Tr** | 85 | 0.0004 | 0.0000 |
| ***Te*** | 19 | 0.1326 | 0.0387 |
| *Experimento 3: Fuzzy ARTMAP con el mejor conjunto de descriptores [Ove, Cou, Kin, N, Pol, $^{0-2}\chi$, Ove101, Bal, W, Ran, $\kappa$, TE]* | | | |
| **Todos los datos** | 104 | 0.0088 | 0.0015 |
| **Tr** | 85 | 0.0001 | 0.000 |
| ***Te*** | 19 | 0.0476 | 0.0079 |
| *Experimento 4:Fuzzy ARTMAP con los índices de Gini et al. [1999] [MW, HUMO, LUMO, Bal, $\mu$, Pol, $^{3}\chi$, flex, log d (pH=2, pH=10), te, 3r eje de inercia, volumen elipsoidal]* | | | |
| **Todos los datos** | 104 | 0.0088 | 0.0015 |
| **Tr** | 85 | 0.0001 | 0.000 |
| ***Te*** | 19 | 0.0476 | 0.0079 |
| *Experimento 5: Back-propagation 14-7-1con el mejor conjunto de índices [Ove, Cou, Kin, N, Pol, $^{0-2}\chi$,Ove101, Bal, W, Ran, $\kappa$, TE]* | | | |
| **Todos los datos** | 104 | 0.1026 | 0.0181 |
| **Tr** | 85 | 0.0927 | 0.0141 |
| ***Te*** | 19 | 0.0749 | 0.0199 |
| *Gini et al. [20] Back-propagation 13-6-1 con sus índices [MW, Humo, Lumo, Bal, $\mu$, Pol, $^{3}\chi$, flex, log d (pH=2, pH=10), te, 3r eje de inercia, volumen elipsoidal]* | | | |
| **Todos los datos** | 104 | 0.0856 | 0.0183 |

### 4.2.3 Coeficientes de actividad a dilución infinita, ln $\gamma^{\infty}$

El comportamiento no ideal de las soluciones acuosas hace difícil extender métodos tradicionales de contribución de grupos como: UNIFAC o ASOG para estimar los coeficientes de actividad a dilución infinita, ln $\gamma^{\infty}$, de sistemas acuosos diluidos donde las variaciones de $\gamma^{\infty}$ son grandes. Los métodos QSPR se presentan como una alternativa asequible, cuyo ámbito de aplicación en la mayoría de los casos es más amplio que el de una ecuación de estado o un método de contribución de grupos y cuyo tiempo de cálculo es muy inferior al de técnicas más sofisticadas como la simulación molecular.

Los objetivos del presente caso se enumeran a continuación. En primer lugar, aplicar mapas auto-organizados, SOM, para (i) constituir un conjunto de índices con información relevante, topológica como cuántica (se tomarán en cuenta mediciones de similitud cuántica (Carbó-Dorca R. y Besalú E., (a) 1988, (b) 1988b) (MQSM)) que representen cuantitativamente el carácter tanto bidimensional como tridimensional de las diversas estructuras moleculares, por ejemplo, sus aspectos conformaciones, estero químicos o electrónicos. (ii) Determinar los prototipos de las clases de los 325 compuestos químicos obtenida al preservar la topología los vectores de entrada a la red mediante un mapa de Kohonnen. Además, se pretende identificar las características de estas clases con respecto a la información del conjunto características topo-cuánticas utilizadas. Y (iii) Seleccionar el conjunto más factible (menor y más relevante) de descriptores para establecer el modelo QSPR correspondiente a ln $\gamma^{\infty}$. En segundo lugar, se busca identificar por medio del sistema cognitivo fuzzy ART el conjunto más representativo de los 325 compuestos necesario para entrenar o modelar la relación estructura propiedad correspondiente. Se estudiará el efecto de incorporar la información obtenida de los prototipos (SOM) para favorecer la predicción (extrapolación a nuevas clases) de cualquier nueva información presentada al modelo durante la fase de evaluación del modelo QSPR fuzzy ARTMAP. Finalmente, el objetivo más general es construir el modelo QSPR para ln $\gamma^{\infty}$ de compuestos orgánicos en agua usando la arquitectura neuronal fuzzy ARTMAP (Carpenter G. et al., 1992) como en los casos presentados anteriormente. El integrar SOM con fuzzy ARTMAP no se busca tan sólo mejorar la capacidad de extrapolación del modelo QSPR sino tener una herramienta que permita explorar la contribución relativa de cada descriptor al modelo y la relación entre ellos con respecto tanto a la clasificación de los compuestos como a la estimación propiamente dicha de la propiedad en estudio.

Los datos experimentales correspondientes a los compuestos utilizados fueron previamente compilados y correlacionados por Sherman et al., (Sherman S et al., 1996) y posteriormente por (Mitchell B. et al., 1998).

El conjunto está formado por compuestos diversos desde hidrocarburos alifáticos, hasta alcoholes, éteres, ésteres, aldehidos, cetonas, ácidos, hidrocarburos halogenados, aminas, amidas, nitrilos, y compuestos conteniendo azufre. Un subconjunto de 280 compuestos se selecciona para la fase de entrenamiento (tr) de la red y los 45 restantes se reservan para probar la generalización del modelo (te) fuzzy ARTMAP QSPR. Ambos conjuntos se seleccionaron mediante el sistema neuronal fuzzy ART (Carpenter G. et al., 1991) siguiendo el procedimiento propuesto por Espinosa et al., (Espinosa G. et al., 2000). El conjunto de entrenamiento se complementa en una segunda fase con los 100 prototipos obtenidos de SOM (cabe recordar que para este caso se ha definido una malla de 10x10 nodos) con los índices que resultaron ser el mejor conjunto en relación con la clasificación de los mismos y a la propiedad objetivo.

SOM toma parte en diversos procesos y durante diferentes fases de la construcción del modelo. (i) En la selección de los elementos de las matrices cuánticas que mejor representen al resto de compuestos. (ii) En la identificación de las clases formadas durante el proceso de categorización de todos los compuestos en base a su estructura molecular (conjunto de descriptores) y la variable objetivo. Esto se logra a través de los vectores prototipos, los cuales emplearemos para definir pseudo compuestos que incluiremos en la fase de entrenamiento de la red. (iii) Para seleccionar mediante la medida de disimilitud entre mapas el conjunto más relevante de índices a partir de la información molecular disponible.

Como punto de partida se calcularon 23 descriptores moleculares, con información tanto topológica como cuántica, tomando en cuenta aquellos descriptores que habían resultado relevantes en estudios previos como los de Medir y Giralt (Medir M. y Giralt, F. 1982) y Yaffe et al., (Yaffe D. et al., 2001) para predecir ln $\gamma^{\infty}$ de hidrocarburos en agua y la solubilidad de diversos compuestos orgánicos, respectivamente. Los índices topológicos proveen información acerca de la adyacencia de los átomos en la estructura molecular. Entre los índices que con más frecuencia se usan en los estudios de QSPR están el índice de Wiener (Wiener H. 1947), los índices de conectividad de Randic (Randic M. 1984) y los de conectividad definidos por Basak et al., (Basak S. et al., 1997). En este contexto se consideran los índices de conectividad de valencia de orden cero a cuatro, el índice kappa de segundo orden, la suma de números atómicos, y los índices de definidos por Hansen de hidrógeno, de polaridad y de dispersión, (todos ellos calculados mediante un paquete comercial (Molecular Modeling Pro, V. 3.0).

Los descriptores cuánticos utilizados previamente por Yaffe et al., (Yaffe D. et al., 2001): polarizabilidad media, el momento dipolar, el número de niveles ocupados, la energía de atracción electrón núcleo, la energía de repulsión núcleo-núcleo, la energía de intercambio y la energía de resonancia, se incorporan al conjunto anterior. Este tipo de índices

cuánticos se han calcularon vía el método semi-empírico *Parametric Method 3* (PM3). La similitud tri-dimensional entre las estructuras moleculares puede extraerse a partir de las medidas de similitud cuántica propuestas por Carbó-Dorca et al., (Carbó-Dorca R. y Besalú E., 1988) determinadas via *Atomic Shell Aproximation* (ASA).

Con el objetivo de manejar toda la información contenida en cada una de las matrices (325x325) de similitud cuántica, tenemos que recurrir a métodos de reducción dimensional, o de selección de los elementos claves en cada una de ellas. La primera aproximación es el usar preferentemente los elementos de la diagonal de dichas matrices, a los cuales se denomina, medidas de auto-similitud. Con el objetivo de discriminar la influencia de los diferentes descriptores es posible proyectar los elementos de cada matriz de similitud cuántica (MQSM), en un SOM como se ha explicado anteriormente. Estos mapas retienen la topología (es decir, las relaciones entre las variables originales) de los elementos de la matriz, la clasificación resultante nos permite seleccionar los elementos más relevantes que representan a cada una de las clases formadas, mediante la definición de criterios heurísticos.

La selección de los elementos más relevantes de cada una de las dos matrices Overlap y Coulomb de 325x325 cada una, se hizo a partir del entrenamiento de dos SOMs definidos en una red cuadrada de 5x5 neuronas o unidades y entrenados usando los elementos de cada una de las matrices correspondientes. Después del entrenamiento, cada una de las clases queda definida o caracterizada por un vector prototipo (pesos correspondientes a dicha clases). Como resultado, además de los elementos de la diagonal de cada una de las matrices cuánticas, los siguientes elementos cruzados de la matriz fueron seleccionados: (i) aquellos que son centro de masas de cada nodo (prototipos comunes en ambos mapas), ya que esto significa que ellos representan a las moléculas del conjunto de $\ln \gamma^{\infty}$ que se comportan de forma similar delante de dos proyecciones distintas. (ii) Los prototipos de las clases con la máxima densidad de compuestos presentes ya que ellos representan al conglomerado con un número de características comunes. El resultado obtenido aplicando dichas restricciones identifica al 1-4-ciclohexadiene ($C_6H_8$) como el único prototipo común en ambos mapas. Las mediciones de similitud cuántica para cada operador, Overlap y Coulomb de todos los compuestos con el 1-4 ciclohexadiene se añadieron como un par de descriptores más $Ove_{C6H8}$ y $Cou_{C6H8}$ , respectivamente.

Dos prototipos adicionales se seleccionan siguiendo el criterio de las clases más pobladas: el N-metil-2-pirrolidona ($C_5H_9N_1O_1$) de la matriz de Overlap ($Ove_{C5H9N1O1}$) y el 1-cloropropano ($C_3H_7Cl_1$) de la matriz de Coulomb ($Cou_{C3H7Cl}$). Así pues, el conjunto de descriptores de tipo similitud cuántica que se incorporan al conjunto original es: Ove, Cou, $Ove_{C6H8}$ y $Cou_{C6H8}$, $Ove_{C5H9N1O1}$, $Cou_{C3H7Cl.}$ Cabe mencionar que el logaritmo natural de estos índices

MQS se emplea en todos los cálculos en lugar de sus valores originales, y que los valores que se citan en las tablas o en las figuras se expresa en dicho formato. De ahí el conjunto de índices con el cual contamos para construir un modelo QSPR para el ln $\gamma^{\infty}$ está formado por 23 descriptores: cinco índices de conectividad molecular de valencia (de orden cero a cuatro), el índice de forma kappa de segundo orden, la suma de números atómicos, los tres índices de Hansen, la polarizabilidad promedio, el momento dipolar, el número de niveles ocupados, la energía de atracción electrón-núcleo, la energía de repulsión núcleo-núcleo, la energía de intercambio, la energía de resonancia y las seis MQS definidas anteriormente.

Sin embargo, queda una cuestión adicional con respecto al tipo de modelo QSAR/QSPR desarrollado anteriormente, ¿Cuál sería el beneficio de usar únicamente información cuántica para construir un QSPR?. Sería fácil de argumentar en el caso de las MQSM, calculadas para cualquier conjunto de compuestos químicos usando las métricas definidas para los operadores cuánticos más relevantes, deberían contener en sí mismas toda la información estructural necesaria para establecer un modelo robusto QSPR. Para proveer una evidencia parcial pero tangible y cuantitativa, el mismo procedimiento descrito para seleccionar los índices más representativos del conjunto global de 23 descriptores moleculares, se aplicó al conjunto de MQS; formado por las auto similitudes y los elementos extraídos de cada una de las matrices de Overlap y Coulomb.

El papel principal que desempeña SOM es como en los casos anteriores el de ayudar a determinar cuales y cuantos son los índices requeridos para estimar en este caso ln $\gamma^{\infty}$. Se definió un mapa de 10x10, la configuración de los nodos fue hexagonal. Se obtuvieron 80 categorías diferentes y 20 nodos quedaron vacíos, como se ilustra en la Fig. 4.27. En está figura se identifican 13 familias de compuestos que ocupan los 80 nodos ocupados de acuerdo tanto a la etiqueta de su familia genérica como a su valor de similitud molecular. En la Fig. 4.27 se citan las características más importantes de las 80 categorías ocupadas. Dando una idea de cómo se agrupan los compuestos de acuerdo con la información molecular provista por los 23 descriptores podemos destacar los siguientes puntos:

(i) SOM distingue entre las 13 familias de compuestos orgánicos, con la presencia dominante de los nodos formados por las dos familias más pobladas: la D correspondiente a los hidrocarburos con substituyentes oxígeno, y el grupo E, correspondientes a los hidrocarburos alifáticos halogenados.

(ii) Aproximadamente el 70% de los 80 nodos ocupados contienen grupos homogéneos de compuesto químicos, el 30% restante presentan mezclas de familias como es el caso de categorías que albergan a hidrocarburos alifáticos y hidrocarburos cíclicos.

(iii) Se observan algunas diferencias importantes con respecto al rango de ln $\gamma^{\infty}$, a pesar de que los compuestos pertenecen a las mismas familias, sobre todo, dichas diferencias se acentúan con la presencia de heteroátomos principalmente halógenos dentro de las familias estudiadas.

Los 23 C-planes se muestran en la Fig. 4.28, ellos se agrupan en 6 familias de acuerdo con la contribución que los descriptores hacen a la organización del mapa. Las seis familias se identifican con las letras A-F. Las principales categorías creadas por SOM pueden resumirse de la siguiente forma:

(i) Los índices de conectividad molecular de orden uno, tres, junto con la polarizabilidad promedio forman la primera categoría.
(ii) El índice de conectividad de segundo orden y la suma de números atómicos se agrupan en la segunda categoría.
(iii) Los índices de conectividad de cero y cuarto orden, el índice de forma kappa, el número de niveles moleculares ocupados, y el índice de repulsión núcleo-núcleo integran el tercer grupo de índices.
(iv) El cuarto conjunto lo forman el momento dipolar los índices de Hansen correspondientes al enlace hidrógeno, a la polaridad y al índice de dispersión.
(v) La quinta categoría la conforman los índice cuánticos de atracción electrón-núcleo, junto con el de intercambio de energía y el de resonancia energética.
(vi) El último conjunto lo forman los índices de similitud cuántica, tanto Overlap, como Coulomb y los elementos cruzados seleccionados de cada una de sus matrices.

Los índices se van seleccionando de acuerdo al orden decreciente al valor absoluto de sus covarianzas y seleccionando cada vez un índice por familia, Fig. 4.28, por ejemplo, $^{1}\chi^{v}$, N, $^{4}\chi^{v}$, $\mu$, ENA y Cou . En este caso cada uno de estos elementos se seleccionan uno a uno al ser sus covarianzas mayor a la covarianza media del conjunto de índices. En la Tabla 4.9, se resumen el orden de presentación de los índices que conforman el proceso de selección. De esta tabla es posible observar que la adición sucesiva de índices en el proceso de selección produce un decremento en la función de disimilitud de 0.703 a 0.29 al incorporar los 6 índices antes mencionados. Este fuerte decremento pone de manifiesto la importancia de ir añadiendo índices de acuerdo a su impacto topológico en los estudios SPR. Al añadir los índices $^{3}\chi^{v}$, NFL, AP, $^{2}\chi^{v}$ y NNR seleccionados todos ellos de acuerdo al orden decreciente de sus covarianzas con ln $\gamma^{\infty}$, se obtiene un mínimo de disimilitud correspondiente a 0.190. Después de lo cual al añadir mayor información molecular el índice de disimilitud comienza a incrementarse,

con lo cual se concluye que dicha información no contribuye a mejorar la distribución dentro del mapa por lo cual la información que aportan es despreciable. El mismo proceso de selección fue aplicado a los seis descriptores MQS del grupo F en la Fig. 4.28, con el fin de evaluar la capacidad predictiva de estos descriptores. Como resultado de ello se obtienen dos categorías. Una formada, por los tres parámetros de Coulomb y la otra con la información de Overlap. El mejor grupo de índices (con disimilitud menor) fue Cou, $Cou_{C6H8}$, $Cou_{C3H7Cl}$, y Ove. Cabe decir que a pesar de la baja correlación del índice de Overlap, este es seleccionado para complementar la información electrostática proporcionada por Overlap, al no contar dentro del conjunto con información de forma adicional.

**Tabla 4.9**. Conjunto de descriptores, covarianzas y medidas de disimilitud acumuladas ordenadas de acuerdo al procedimiento de selección de variables usando SOM

| Descriptores | Abreviación | Grupo | Covarianza con variable objetivo | Disimilitud acumulada |
|---|---|---|---|---|
| Índice de conectividad de valencia de primer orden | $^{1}\chi^{v}$ | A | 0.632 | 0.703 |
| Suma de números atómicos | N | B | 0.583 | 0.553 |
| Índice de conectividad de valencia de cuarto orden | $^{4}\chi^{v}$ | C | 0.553 | 0.366 |
| Momento dipolar | $\mu$ | D | -0.535 | 0.276 |
| Atracción electrón-núcleo | ENA | E | -0.508 | 0.243 |
| Autosimilitud de Coulomb | Cou | F | 0.425 | 0.229 |
| Índice de conectividad de valencia de tercer orden | $^{3}\chi^{v}$ | A | 0.584 | 0.212 |
| Número de niveles ocupados | NFL | B | 0.543 | 0.204 |
| Polarizabilidad promedio | AP | A | 0.541 | 0.195 |
| Índice de conectividad de valencia de segundo orden | $^{2}\chi^{v}$ | B | 0.527 | 0.191 |
| Repulsión núcleo-núcleo | NNR | C | 0.496 | 0.190 |
| Polarizabilidad de Hansen | HP | D | -0.494 | 0.192 |
| Índice de enlace hidrógeno de | HH | D | -0.491 | 0.196 |
| Energía de intercambio | EE | E | -0.451 | 0.198 |
| Energía de resonancia | RE | E | -0.439 | 0.201 |
| Índice kappa de segundo orden | $^{2}\kappa$ | C | 0.437 | 0.211 |
| Índice de conectividad de valencia de orden cero | $^{0}\chi^{v}$ | C | 0.435 | 0.218 |
| MQS de Coulomb haciendo una referecia cruzada con el elemento | Cou222 | F | 0.419 | 0.234 |

| Descriptores | Abreviación | Grupo | Covarianza con variable objetivo | Disimilitud acumulada |
|---|---|---|---|---|
| 222 | | | | |
| MQS de Coulomb haciendo una referecia cruzada con el elemento 75 | Cou75 | F | 0.334 | 0.253 |
| MQS de Overlap | Ove | F | 0.170 | 0.257 |
| MQS de Overlap haciendo una referecia cruzada con el elemento 222 | Ove222 | F | 0.121 | 0.270 |
| MQS de Overlap haciendo una referecia cruzada con el elemento 154 | Ove154 | F | -0.037 | 0.377 |
| Dispersividad de | HD | D | -0.002 | 0.381 |

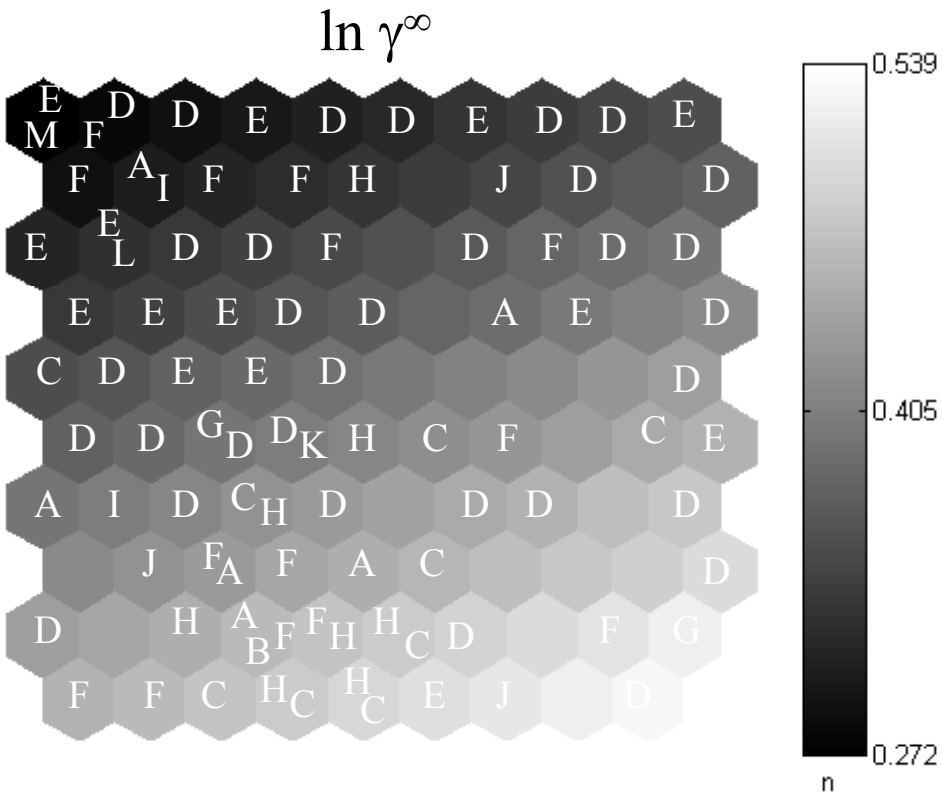



**Fig. 4.27.** Distribución de las trece familias de compuestos orgánicos generada usando SOM: (A) hidrocarburos monoaromáticos; (B) hidrocarburos poliaromáticos; (C) hidrocarburos alifáticos; (D hidrocarburos con oxígeno como substituyente; (E) hidrocarburos halogenados; (F) hidrocarburos con nitrógeno y oxígeno como substituyentes, (G) hidrocarburos con azufre y/o oxígeno como substituyentes, (H) hidrocarburos cíclicos, (I) hidrocarburos aromáticos con oxígeno como substituyentes, (J) hidrocarburos aromáticos halogenados, (K) hidrocarburos cíclicos con oxígeno como substituyentes, (L) hidrocarburos con halógenos y oxígeno como substituyentes, (M) hidrocarburos con grupos nitro y halógenos como substituyentes. Los tonos de grises indican las distancias relativas entre los elementos de cada categoría.

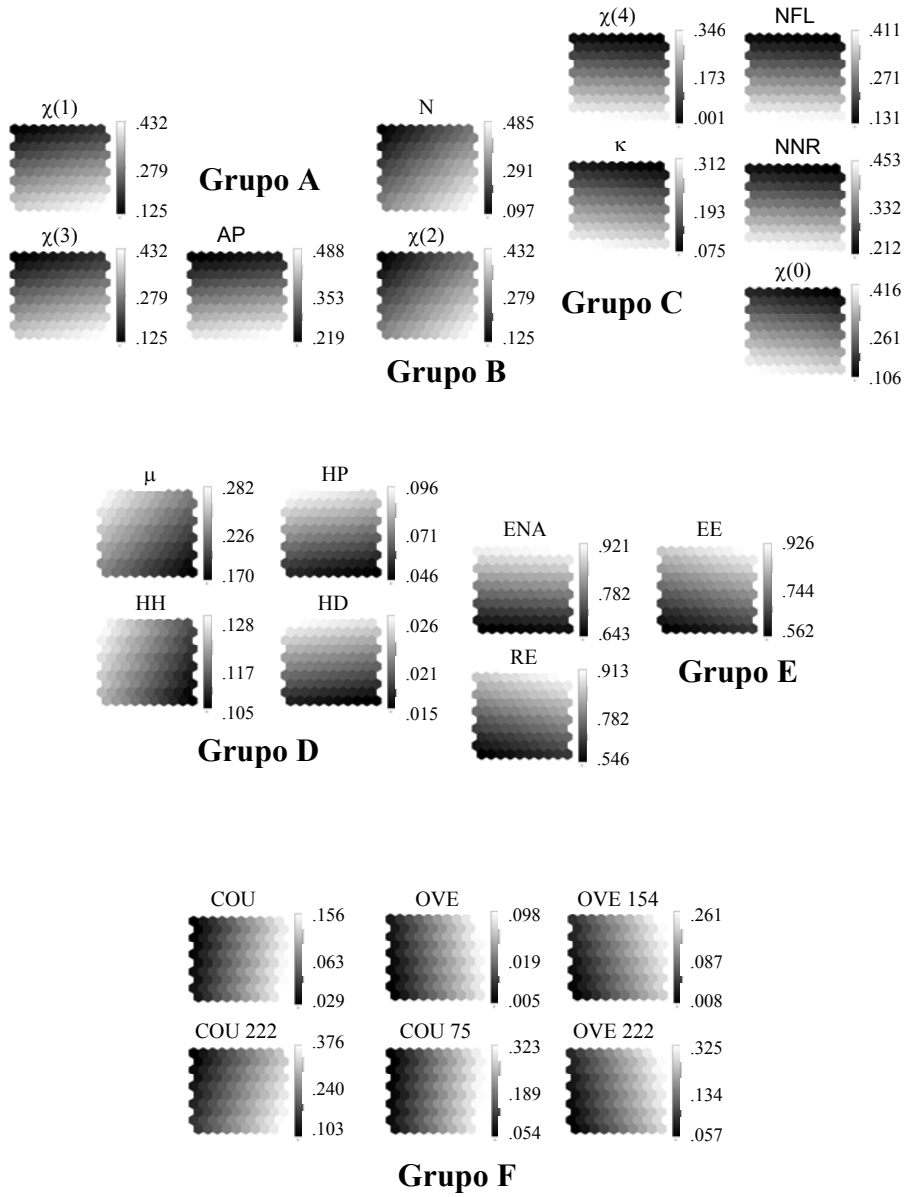



**Fig. 4.28** Agrupación de los planos para cada descriptor proyectado sobre el mapa de ln $\gamma^{\infty}$

Los modelos fuzzy ARTMAP-QSPR se calculan tanto para el conjunto de 11 descriptores (más representativo), $^{1}\chi^{v}$, N, $^{4}\chi^{v}$, μ, ENA, Cou, $^{3}\chi^{v}$, NFL, AP, $^{2}\chi^{v}$, NNR y usando únicamente los elementos más significativos de los MQSM (Cou, $Cou_{C6H8}$, $Cou_{C3H7Cl}$ y Ove). 280 compuestos fueron seleccionados con fuzzy ART para entrenar la red, los 45 restantes se usan para evaluar la generalización del modelo QSPR. En la Fig. 4.29, se representa la comparación entre los resultados obtenidos siguiendo la metodología SOM QSPR y los datos experimentales.

El desempeño del modelo QSPR-fuzzy ARTMAP con los 11 índices descritos anteriormente, Fig. 4.29a, presenta un error absoluto medio de 0.09 ln (1.21%) y una desviación estándar 0.29 ln. La generalización del modelo para los 45 elementos restantes es de 0.52 ln tanto para el error

como para la desviación estándar. Sin embargo, estos resultados mejoran al incluir en el subconjunto de 280 compuestos químicos, la información adicional de los 100 prototipos obtenidos del entrenamiento de SOM. En este caso el parámetro de vigilancia de ARTMAP se relajó de 0.999 a 0.995 para obtener una mayor generalización al beneficiarse de la información adicional provista al conjunto de entrenamiento. El error absoluto medio y la desviación estándar para el conjunto de prueba caen respectivamente de 0.40 a 0.48 ln en este caso como se muestra en la Fig. 4.29b. En la comparación entre los valores experimentales y estimados de las Figs. 4.29a-b se distinguen entre los valores del conjunto de entrenamiento y el de generalización. Ambos modelos tienen un desempeño excelente durante el entrenamiento. Al comparar las figuras 4.29 a y b se observa la mejora en la generalización del modelo QSPR, al tomar en cuenta los prototipos para en entrenamiento la clasificación de los 12 compuestos mejora, al asignar a 11 de ellos a clases formadas por prototipos de categorías ocupadas en SOM y 1 de ellos a una clase formada por una prototipo de una categoría de SOM que no tenía ningún elemento en ella.

Los valores de ln $\gamma^{\infty}$ estimados para los 280 compuestos durante el entrenamiento dan lugar a un error absoluto medio de 0.02 ln, como se muestra en la Fig. 4.29a. Este alto rendimiento en el modelo, en relación con los modelos MRL u otras técnicas estadísticas, no es sorprendente, ya que el parámetro de vigilancia controla la precisión con la cual cada compuesto es clasificado. En este caso un parámetro de vigilancia de 0.999 es útil para establecer un compromiso entre una buena precisión en el entrenamiento, y la posibilidad de generalizar durante la etapa de prueba de la red, Fig. 4.29a.

Para corroborar la hipótesis de que las medidas de similitud cuántica, MQS contienen la información suficiente para establecer modelos QSPR robustos, un modelo QSPR-fuzzy ARTMAP se entrena con 280 compuestos y se caracteriza únicamente por 4 MQSM (Cou, $Cou_{C6H8}$, $Cou_{C3H7Cl}$ y Ove). El error absoluto y la desviación estándar correspondiente para el conjunto de prueba de los 45 compuestos son de 0.92 y 1.09 ln, respectivamente, Fig. 4.30. A pesar de que estos valores doblan aquellos obtenidos con el modelo QSPR-fuzzy ARTMAP y los once índices, (Fig. 4.29a-b), la generalización sigue siendo buena tomando en cuenta que el número de variables es menos de la mitad. Además podemos corroborar que los índices de similitud cuántica constituyen una buena fuente de información en la caracterización de las estructuras moleculares. En este caso, al considerar los prototipos como parte del conjunto de entrenamiento el desempeño de la red no se vio beneficiado, debido a errores asociados con los grupos halógenos se propagan sobre el mapa.

Para ilustrar el potencial de los índices MQSM como una fuente fundamental de información molecular en la construcción de modelos QSPR,

aún en casos de modelos con compuestos heterogéneos, se genera un modelo adicional usando los 4 MQSM seleccionadas anteriormente y añadiendo información adicional de tamaño y volumen. Dicha información se obtiene de dos índices simples como la suma de números atómicos de los heteroátomos presentes en cada estructura molecular de tal forma que es fácil discernir la presencia o ausencia de un grupo funcional específico. El error absoluto medio y la desviación estándar para los 45 compuestos del conjunto de prueba son iguales a 0.57 ln. Estos valores son muy cercanos al 0.52 ln obtenido con los 11 índices usados en el primer modelo QSPR-fuzzy ARTMAP, Fig. 4.29a con lo cual corroboramos la versatilidad de este tipo de índices.

Parece interesante analizar en la generalización del modelo QSPR fuzzy ARTMAP a través de cada familia de compuestos que integra el conjunto global, como se muestra en la Tabla 4.10. La comparación indica que el modelo con medidas de MQS generaliza muy bien cuando el número de compuestos es elevado y el rango de los tamaños moleculares (número de átomos de carbono) es pequeño. Esto pone de manifiesto la necesidad de caracterizar la estructura molecular, no tan sólo las interacciones intermoleculares. Cabe señalar, que los errores y las desviaciones estándar correspondientes citadas en la Tabla 4.10 corresponden a medidas globales (entrenamiento+prueba), de ahí que las diferencias entre unos modelos y otros no sean tan evidentes como si comparamos la capacidad de generalización, como en las Figs. 4.29 y 4.30. Sin embargo, este formato facilitará la comparación con estudios previos y la discusión de los siguientes resultados.

Finalmente, se comparan los resultados obtenidos en el presente con aquellos publicados previamente el la literatura para el mismo conjunto de compuestos (Mitchell B. et al., 1998) y para familias específicas de compuestos (Medir M. et al., 1982). Los resultados muestran que los modelos generados con fuzzy ARTMAP mejoran sustancialmente los publicados previamente con modelos basados en el algoritmo de Broyden-Fletcher-GoldfarbShano, BFGS (Mitchell B. et al., 1998). La arquitectura y los resultados citados por Mitchell es 12-6-1, con un error medio de 0.27 ln para el conjunto de prueba y 0.32 ln para los elementos usados durante el entrenamiento. Estos autores utilizan doce índices como base del modelo QSPR, que incluyen tres índices topológicos, cuatro índices referidos a la superficie molecular con carga parcial, dos medias correspondientes al enlace hidrógeno, el calor de formación y dos relacionados con relaciones lineales teóricas de energías de solvatación. Cabe destacar, que en dicho trabajo han usado únicamente 25 compuestos como elementos de generalización. La comparación muestra claramente como los índices seleccionados caracterizan satisfactoriamente al conjunto de moléculas y son capaces de describir la interacción entre el soluto y el solvente que da lugar al comportamiento no lineal de las propiedades de equilibrio. Un

gráfico comparando el modelo SOM fuzzy ARTMAP con el presentado por Mitchell et al., (Mitchell B. et al., 1998) y con Medir et al., (Medir M. et al., 1982) se muestra en la Fig. 4.31. El dibujo, esquematiza una comparación por grupos funcionales, en la figura (a) se comparan los hidrocarburos mono aromáticos, desviaciones con respecto a los modelos previos se observan principalmente para valores grandes del coeficiente de actividad; mientras que en la (b) se esquematizan los resultados para los hidrocarburos aromáticos poli-nucleares, no muestran desviaciones considerables; la figura (c) representa a los hidrocarburos alifáticos, en este caso los modelos previamente estudiados muestran desviaciones en todo el rango de valores de ln $\gamma^{\infty}$, siendo el modelo de Mitchell et al., (Mitchell B. et al., 1998) en el que se observa mayor dispersión. En este caso, el modelo QSPR actual generaliza razonablemente bien, con excepción del caso del 1-octileno que puede ser considerado un *outlier*. El desempeño de los modelos QSPR-fuzzy ARTMAP es significativamente mejor que los modelos presentados por Mitchell y Jurs (Mitchell B. et al., 1998) y el modelo lineal presentado por Medir y Giralt (Medir M. et al., 1982) para cada una de las trece familias de compuestas, Tabla 4.10. Ambos modelos fuzzy ARTMAP uno entrenado con los 280 compuestos seleccionados de la base de datos original y el segundo con 100 elementos más tomados de los prototipos derivados de SOM, presentan un error medio y una desviación estándar del orden de siete veces menor que el modelo de Mitchell y Jurs (Mitchell B. et al., 1998). Esta mejora es aún más significativa al considerar que estos autores usan 300 compuestos durante el entrenamiento de la red, y presentan los errores de generalización para 21 compuestos, ya que 4 (de los 25 originales del conjunto de prueba) son excluidos por considerarse outliers -benzonitrilo, tri-iodometano, hexafluoroetano, y el ácido benzoico. Por otro lado, los modelos de Medir y Giralt estiman al coeficiente de actividad a dilución infinita con una precisión comparable al modelo de Mitchell et al., para la familia de hidrocarburos aromáticos, usando únicamente el índice de conectividad de primer orden y para la familia de alifáticos al incluir el momento dipolar al índice anterior en el modelo de correlación lineal

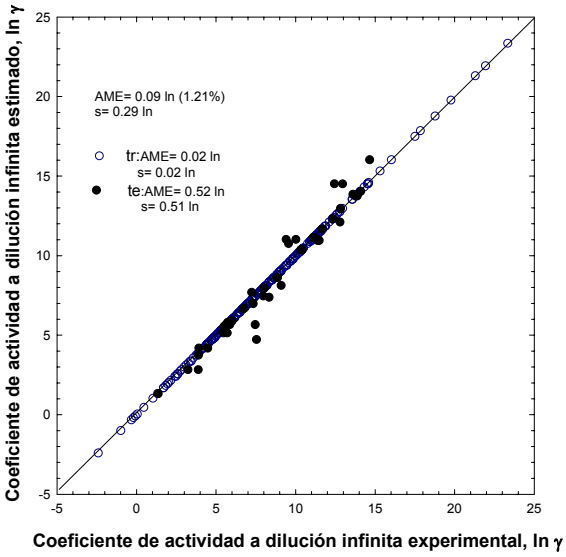



**Fig. 4.29a**

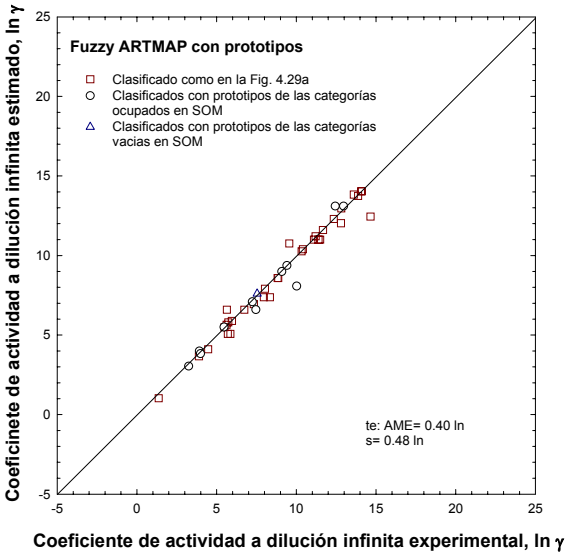



**Fig. 429b**

**Fig. 4.29** Comparación entre los valores experimentales de ln $\gamma^{\infty}$ y aquellos estimados con dos modelos fuzzy ARTMAP basados en los 11 descriptores seleccionados con SOM y entrenados con (a) 280 compuestos o (b) 280 compuestos y como información adicional 100 prototipos derivados de las categorías de SOM. Nota que (b) solo se presentan los resultados del conjunto de generalización

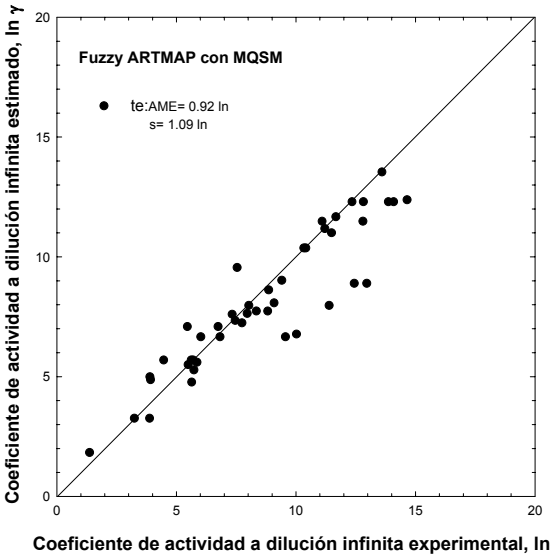



**Fig. 4.30.** Comparación entre los valores experimentales de ln $\gamma^{\infty}$ y los estimados con un modelo basado en fuzzy ARTMAP usando únicamente los cuatro índices de similitud cuántica más factibles para este y entrenado con 280 compuestos

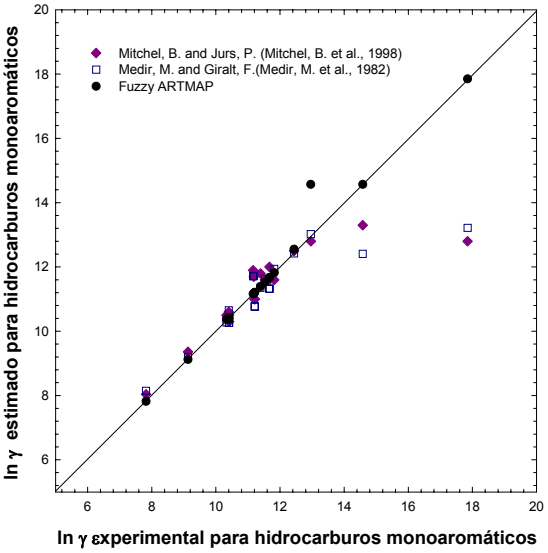



**Fig. 4.31a**

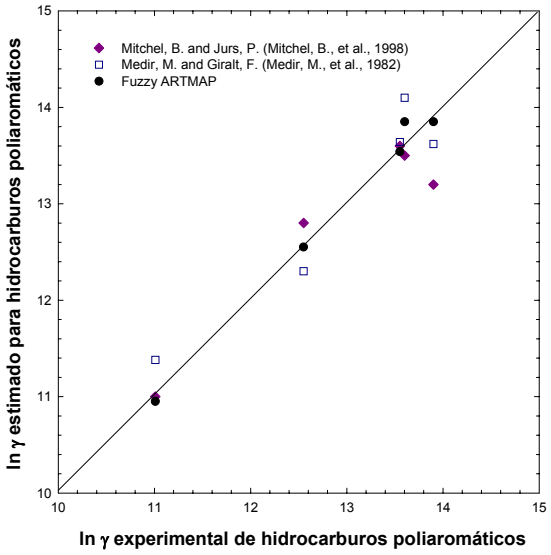



**Fig. 4.31b**

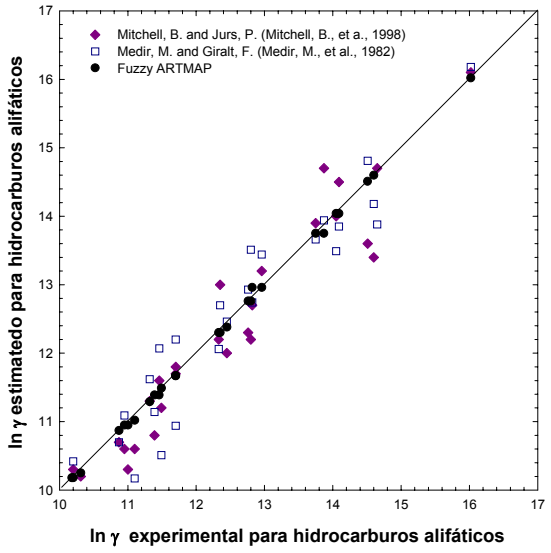



**Fig. 4.31c**

**Fig. 4.31** Comparación de los modelos QSPR para el ln $\gamma^{\infty}$ entre el modelo actual fuzzy ARTMAP (desarrollado con el conjunto de 11 índices más favorable y entrenado con 280 compuestos) y los modelos QSPR previos para diferentes familias (a) hidrocarburos monoaromáticos, (b) hidrocarburos poliaromáticos, e (c) hidrocarburos alifáticos.

**Tabla 4.10** Influencia de la presencia de diferentes grupos funcionales en la generalización del modelo QSPR para $\ln\gamma^{\infty}$ en términos de los errores absolutos medios (desviaciones estándar) para las trece familias de compuestos orgánicos presentes

| ID | Familia | # comp.. | No. de átomos de C | SOM-Fuzzy ARTMAP mejor conjunto | SOM-Fuzzy ARTMAP con MQSM | Fuzzy ARTMAP Yaffe et al. (2001) | Mitchell & Jurs (1998) | Medir & Giralt (1984) |
|---|---|---|---|---|---|---|---|---|
| **A** | **Hidrocarburos monoaromáticos** | 20 | C6 – C10 | 0.22 (0.57) | 0.53 (1.34) | 0.31 (1.12) | 0.54 (1.13) | 0.56 (1.09) [(1)] |
| **B** | **hidrocarburos poliaromáticos** | 5 | C9 – C12 | 0.07 (0.10) | 0.16 (0.38) | 0.06 (0.13) | 0.22 (0.28) | 0.30 (0.15) [(2)] |
| **C** | **Hidrocarburos alifáticos** | 36 | C4 – C8 | 0.10 (0.25) | 0.32 (0.75) | 0.13 (0.42) | 0.45 (0.47) | 0.46 (0.32) [(3)] |
| **D** | **Hidrocarburos con oxígeno como substituyentes** | 126 | C1 – C18 | 0.06 (0.17) | 0.09 (0.31) | 0.11 (0.41) | 0.70 (0.78) | ------ |
| **E** | **Hidrocarburos halogenados** | 66 | C1 – C6 | 0.12 (0.34) | 0.12 (0.43) | 0.13 (0.41) | 0.82 (2.25) | ------ |
| **F** | **Hidrocarburos con nitrógeno y oxígeno como substituyentes** | 26 | C3 – C8 | 0.05 (0.12) | 0.06 (0.20) | 0.01 (0.03) | 1.12 (2.11) | ------ |
| **G** | **Hidrocarburos con azufre y/o oxígeno como substituyentes** | 7 | C1 – C4 | 0.09(0.14) | 0.26(0.61) | 0.04(0.11) | 0.18(0.12) | ------ |
| **H** | **Hidrocarburos cíclicos** | 15 | C5 – C9 | 0.07(0.17) | 0.10(0.34) | 0.06(0.20) | 0.51(0.52) | ------ |
| **I** | **Hidrocarburos aromáticos con oxígeno como substituyentes** | 7 | C7 – C8 | 0.04(0.03) | 0.03(0.03) | 0.02(0.04) | 0.25(0.23) | ------ |
| **J** | **Hidrocarburos aromáticos halogenados** | 9 | C6 – C7 | 0.01(0.02) | 0.06(0.12) | 0.00(0.01) | 0.27(0.24) | ------ |
| **K** | **Hidrocarburos cíclicos con oxígeno como substituyentes** | 5 | C5 – C8 | 0.02(0.03) | 0.00(0.00) | 0.00(0.00) | 0.36(0.25) | ------ |
| **L** | **Hidrocarburos con halógenos y oxígeno como substituyentes** | 2 | C2 – C4 | 1.41(1.99) | 1.00(1.42) | 0.04(0.05) | 0.15(0.11) | ------ |
| **M** | **Hidrocarburos con nitrógeno y grupos halógeno como substituyentes** | 1 | C1 | 0.01(0.00) | 0.01(0.00) | 0.01(0.00) | 0.08(0.00) | ------ |

(1) $\ln\gamma^{\infty} = 2.99+2.58\ {}^{1}\chi^{v}$
(2) $\ln\gamma^{\infty} = -0.054+3.1\ {}^{1}\chi^{v}$
(3) $\ln\gamma^{\infty} = 5.461+2.739\ {}^{1}\chi^{v}+5.884\ \mu$

### 4.2.4 Puntos de fusión

Al menos alguna de las temperaturas de transición de fase se conoce para la mayoría de los compuestos orgánicos. Sin embargo, a diferencia de lo que lo que sucede con el punto de ebullición, existen pocos trabajos que relacionen al punto de fusión directamente con la estructura molecular. En general, el trabajo previo encontrado para la predicción del punto de fusión es escaso. En 1879, comenzaron a extenderse las relaciones en las cuales el incremento del punto de fusión de moléculas se ligaba directamente con la simetría y que tan compacta era la estructura molecular, y en series homólogas este incremento venía ligado al aumento en el peso molecular. Si bien este hecho es cierto para alcanos lineales y ciclo-alcanos y puede justificarse basándose en el aumento de las interacciones atractivas VDW que aumenta con el aumento de la cadena, no ocurre lo mismo en ciclos, o compuestos orgánicos con presencia de substituyentes, en los cuales las interacciones estéricas juegan un papel importante en la determinación del estado de conformación más estable y por lo tanto en las interacciones que intervienen y determinan las propiedades físicas de los fluidos.

Recientemente Krzyzaniak y colaboradores propusieron un método simple para estimar tanto puntos de fusión como de ebullición de compuestos aromáticos (Krzyzaniak J. et al., 1995). En dicho trabajo se demostró que el punto de ebullición requiere solamente de propiedades aditivas constitutivas, mientras que el punto de fusión requiere tanto propiedades constitutivas cono no constitutivas como la simetría rotacional para caracterizar las relaciones estructuras propiedad. De forma semejante Katritzky et al., (Katritzky A. et al., 2000) al correlacionar el punto de fusión de 443 bencenos mono y di-substituidos encontró que los principales parámetros del modelo QSPR eran un descriptor para el enlace hidrógeno que denomina HDSA2 y descriptores cuánticos. En el presente caso de estudio se toman en cuenta 292 de los 325 compuestos originales usados para estimar el coeficiente de actividad a dilución infinita, $\ln \gamma^{\infty}$. Los datos experimentales fueron recopilados de la base de datos del DIPPP (DIPPR, 1997). Un subconjunto de 263 compuestos se selecciona mediante fuzzy ART para la fase de entrenamiento (tr) de fuzzy ARTMAP y los 29 restantes se usan durante la fase de generalización del modelo fuzzy ARTMAP QSPR (te). Los vectores prototipos de SOM de las clases que no fueron ocupadas por ningún elemento, con el conjunto más favorable de índices se utilizan para complementar la información proporcionada por el vector de entrada de los compuestos usados durante el entrenamiento de la red (destacando que en este caso se utiliza una malla 10x10 para construir cada SOM).

Inicialmente se usan 22 descriptores moleculares, entre los que se cuenta con información tanto topológica, como químico cuántico. Partiendo de la metodología propuesta los mapas topológicos para cada variable junto

con la correspondiente a la variable objetivo (punto de fusión) permiten obtener categorías de los compuestos estudiados, Fig. 4.32. Llevando a cabo una selección de un elemento de cada grupo de descriptores con la correlación más alta con respecto al punto de fusión, siempre que dicha correlación sea mayor que la correlación media del conjunto haciendo referencia a la variable objetivo. Y posteriormente, seleccionando uno a uno los índices en orden decreciente de correlación. Una vez más el proceso concluye cuando la medida de disimilitud entre pares de mapas alcanza un mínimo, con lo cuál se garantiza que la adición de más índices no aporta información adicional a la organización de los compuestos sobre el mapa auto-organizado. Este será el conjunto de descriptores moleculares base para el modelo final fuzzy ARTMAP QSPR.

El objetivo de SOM será como en los casos anteriores determinar cuales y cuantos son los descriptores necesarios para estimar el punto de fusión de compuestos orgánicos diversos. Se usa un mapa de 10x10 y tal como se observa en la Fig. 4.32, se obtienen 81 nodos en los cuales se agrupa al menos un compuesto y 19 nodos quedan distribuidos sobre el mapa sin poseer un elemento característico. Al ser un parte del conjunto de compuestos usados para estimar ln $\gamma^{\infty}$, se distinguen las 13 familias del caso anterior. Es posible enfatizar que la distribución de las familias no es tan homogénea como en el caso del coeficiente de actividad a dilución infinita, es decir, 17 de las 81 categorías tienen más de dos familias de compuestos formando el prototipo de la clase. Sin embargo, en la mayoría de los casos al menos uno de los integrantes de esa clase está rodeado de clases que poseen el mismo grupo.

Los diferentes planos (C-planes) de las variables agrupados en distintos grupos, de acuerdo al aporte de la variable a la categorización de los compuestos sobre el mapa. Algunos puntos que cabe destacar son:

(i) Los índices de atracción electrón-núcleo, de resonancia y de intercambio electrónico forman el primer conjunto de índices.
(ii) La auto similitud cuántica de Overlap, el índice de hidrógeno de Hansen y el momento dipolar forman el segundo grupo de índices.
(iii) En la tercera categoría de índices se agrupan los índices de similitud cuántica cruzados de Overlap y Coulomb ($Cou_{C6H8}$, $Ove_{C6H8}$, $Cou_{C3H5Cl}$, $Ove_{C6H9NO}$).
(iv) El cuarto conjunto lo forman el índice de autosemejanza de Coulomb, la suma de números atómicos, el índice de conectividad de valencia de orden dos y el índice de forma kappa.
(v) Finalmente, el último grupo lo forman los índices: el número de niveles ocupados, la polarizabilidad media, la superficie molecular y los índices de conectividad de valencia de orden cero, uno, tres y cuatro.

El primer conjunto de índices se integra por un representante de cada grupo, las auto similitudes de Overlap, Coulomb, el índice MQS $Cou_{C6H8}$, y el NFL. Los 17 restantes se forman mediante la adicción uno a uno de los descriptores en orden decreciente de correlación con la variable objetivo. La disimilitud promedio entre mapas alcanza un mínimo en 0.1222, al añadir al conjunto base, los siguientes índices, la suma de número atómicos (N), la energía de resonancia (RE), el índice MQS $Ove_{C6H8}$, el índice MQS $Ove_{C5H9NO}$, la polarizabilidad media (AP), la superficie molecular (SA) y el índice de conectividad de orden uno.

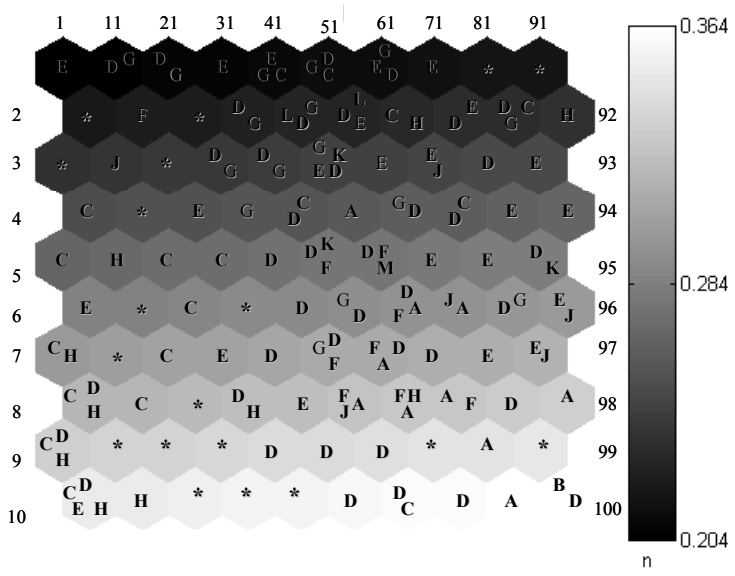



**Fig. 4.32.**Distribución de las trece familias de compuestos orgánicos generada usando SOM: (A) hidrocarburos monoaromaticos; (B) hidrocarburos poliaromaticos; (C) hidrocarburos alifáticos; (D hidrocarburos con oxígeno como substituyente; (E) hidrocarburos halogenados; (F) hdrocarburos con nitrogeno y oxígeno como substituyentes, (G) hidrocarburos con azufre y/o oxígeno como substituyentes, (H) hidrocarburos cíclicos, (I) hidrocarburos aromáticos con oxígeno como substituyentes, (J) hidrocarburos arómaticos halogenados, (K) hidrocarburos cíclicos con oxígeno como substituyentes, (L) hidrocarburos con halógenos y oxígeno como substituyentes, (M) hidrocarburos con grupos nitro y halógenos como substituyentes. Los tonos de grises índican las distancias relativas entre los elementos de cada categoría

La predicción global del modelo QSPR fuzzy ARTMAP presenta un error absoluto medio de 0.91 K (0.49%), con una desviación estándar de 3.49 K (1.96%). Los errores y sus desviaciones estándar relativos a los conjuntos de entrenamiento y generalización son 0.08 K (0.04%), 0.12 K (0.06%) y 8.43 K (4.55%), 8.15 K (4.40%), respectivamente. El máximo error relativo es de 18.97% y corresponde al hexano que se clasifica con un elemento de su misma familia estructural 2-3-dimetil-butano, sin embargo, como se comenta al inició del caso el en punto de fusión los efectos estéricos juegan un papel predominante sobre el comportamiento de la propiedad y este es altamente no lineal. Un análisis de la generalización del modelo se encuentra en la Tabla 4.11. Los resultados muestran que la clasificación es correcta y sólo

un elemento el isopropil propil éter se clasifica en un grupo incorrecto, junto a una amida. Los resultados tanto del conjunto de entrenamiento como de generalización se presentan en la Fig. 4.33. Estos resultados no mejoran al incluir en el subconjunto de 263 compuestos químicos usados durante el entrenamiento, la información adicional de los prototipos de las clases vacías generados por SOM. El error absoluto medio y la desviación estándar para el conjunto de prueba son de 8.89 K (4.94%) y 1.02 K (5.97%), respectivamente, Fig. 4.33. El error máximo es de 26.6% y sigue correspondiendo al hexano que en este caso se clasifica con un grupo de compuestos insaturados y/o cíclicos. Ningún compuesto de la fase de prueba se clasificó con un pseudo compuesto, esto lleva a pensar que las clases existentes tienen en este caso información suficientemente similar para generalizar a los compuestos propuestos para la generalización. Sin embargo, la poco diferencia entre uno y otro modelo, nos lleva a corroborar que es posible caracterizar estructuras inexistentes (hasta el momento) usando la información de los nodos vacíos de SOM.

Cabe hacer notar que 10 de los 14 índices seleccionados son de naturaleza cuántica, esto ratifica el trabajo de Simamora et al., (Simamora P. et al., 1993) en el cual propone que son necesarios tanto la simetría como la entalpía de fusión para determinar el punto de fusión de cualquier sustancia. La entalpía queda representada por los parámetros energéticos, y los parámetros de simetría en cambio por aquellos índices tanto cuánticos como topológicos referentes a la forma, MQS de Overlap, N, SA, y los índices de conectividad molecular.

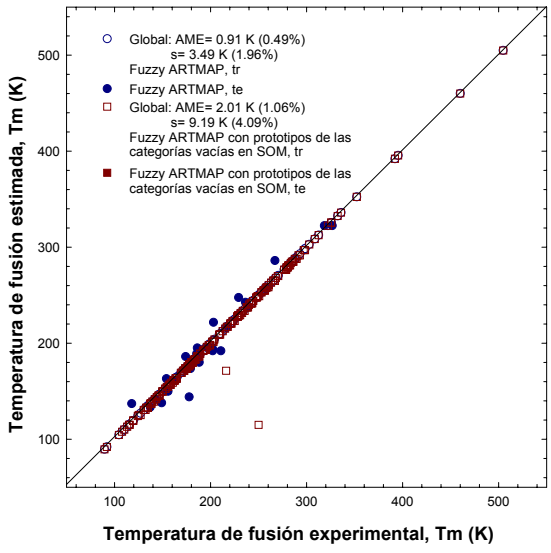



**Fig. 4.33.** Comparación de los valores experimentales de la temperatura de fusión $T_m$ y los estimados con fuzzy ARTMAP basados en 14 descriptores moleculares seleccionados con SOM y entrenados con (a) 262 compuestos (b) con 262 compuestos y la información adicional de los prototipos derivados de las clases vacías.

**Tabla 4.11**. Punto de fusión experimentales y estimados de 292 compuestos orgánicos diversos con un modelo fuzzy ARTMAP

| ID | Nombre | Fórmula | Exp. Punto de fusión K | Est. Punto de fusión K | Error absoluto K | Error relativo, % |
|---|---|---|---|---|---|---|
| **tr** | tetrachloromethane | C 1 Cl 4 | 250.00 | 250.01 | 0.01 | 0.01 |
| **tr** | trichlorofluoromethane | C 1 F 1 Cl 3 | 161.90 | 161.89 | 0.01 | 0.01 |
| **tr** | dichlorodifluoromethane | C 1 F 2 Cl 2 | 115.00 | 115.00 | 0.00 | 0.00 |
| **tr** | chlorotrifluoromethane | C 1 F 3 Cl 1 | 92.00 | 91.99 | 0.01 | 0.01 |
| **tr** | tetrafluoromethane | C 1 F 4 | 89.50 | 89.50 | 0.00 | 0.00 |
| **tr** | tribromomethane | C 1 H 1 Br 3 | 281.00 | 280.99 | 0.01 | 0.00 |
| **tr** | trichloromethane | C 1 H 1 Cl 3 | 209.40 | 208.98 | 0.42 | 0.20 |
| **tr** | dichlorofluoromethane | C 1 H 1 F 1 Cl 2 | 138.00 | 137.88 | 0.12 | 0.09 |
| **tr** | chlorodifluoromethane | C 1 H 1 F 2 Cl 1 | 115.60 | 115.58 | 0.02 | 0.02 |
| **tr** | triiodomethane | C 1 H 1 I 3 | 392.00 | 392.00 | 0.00 | 0.00 |
| **tr** | dibromomethane | C 1 H 2 Br 2 | 220.50 | 220.49 | 0.01 | 0.01 |
| **tr** | dichloromethane | C 1 H 2 Cl 2 | 177.90 | 177.79 | 0.11 | 0.06 |
| **tr** | difluoromethane | C 1 H 2 F 2 | 137.00 | 137.01 | 0.01 | 0.01 |
| **tr** | diidomethane | C 1 H 2 I 2 | 279.10 | 279.08 | 0.02 | 0.01 |
| **tr** | formaldehyde | C 1 H 2 O 1 | 181.00 | 180.99 | 0.01 | 0.01 |
| **tr** | formic acid | C 1 H 2 O 2 | 281.30 | 280.99 | 0.31 | 0.11 |
| **tr** | bromomethane | C 1 H 3 Br 1 | 179.30 | 179.00 | 0.30 | 0.17 |
| **tr** | chloromethane | C 1 H 3 Cl 1 | 175.30 | 175.30 | 0.00 | 0.00 |
| **tr** | fluoromethane | C 1 H 3 F 1 | 131.20 | 131.20 | 0.00 | 0.00 |
| **tr** | nitromethane | C 1 H 3 N 1 O 2 | 244.50 | 244.49 | 0.01 | 0.00 |
| **tr** | methanol | C 1 H 4 O 1 | 175.40 | 175.30 | 0.10 | 0.06 |
| **tr** | methyl mercaptane | C 1 H 4 S 1 | 150.00 | 149.88 | 0.12 | 0.08 |
| **tr** | nitrotrichloromethane | C 1 N 1 O 2 Cl 3 | 209.00 | 208.98 | 0.02 | 0.01 |
| **tr** | 1-2-4-5-tetramethylbenzene | C 10 H 14 | 352.30 | 352.30 | 0.00 | 0.00 |
| **tr** | butyl benzene | C 10 H 14 | 185.10 | 185.10 | 0.00 | 0.00 |
| **tr** | sec-butyl-benzene | C 10 H 14 | 198.00 | 198.02 | 0.02 | 0.01 |
| **tr** | tert-butyl-benzene | C 10 H 14 | 215.20 | 215.21 | 0.01 | 0.01 |
| **tr** | 1-decanol | C 10 H 22 O 1 | 279.90 | 279.58 | 0.32 | 0.11 |
| **tr** | 1-methyl naphthalene | C 11 H 10 | 242.60 | 242.41 | 0.19 | 0.08 |
| **tr** | 1-3-dimethyl naphthalene | C 12 H 12 | 267.00 | 267.00 | 0.00 | 0.00 |
| **tr** | 1-4-dimethyl naphthalene | C 12 H 12 | 280.60 | 280.58 | 0.02 | 0.01 |
| **tr** | 1-ethyl naphthalene | C 12 H 12 | 259.10 | 259.11 | 0.01 | 0.00 |
| **tr** | 1-dodecanol | C 12 H 26 O 1 | 297.00 | 296.98 | 0.02 | 0.01 |
| **tr** | 1-tetradecanol | C 14 H 30 O 1 | 312.50 | 312.52 | 0.02 | 0.01 |
| **tr** | 1-hexadecanol | C 16 H 34 O 1 | 322.30 | 322.32 | 0.02 | 0.01 |
| **tr** | 1-octadecanol | C 18 H 38 O 1 | 332.50 | 332.49 | 0.01 | 0.00 |
| **tr** | tetrachloroethene | C 2 Cl 4 | 250.70 | 250.72 | 0.02 | 0.01 |
| **tr** | hexachloroethane | C 2 Cl 6 | 460.00 | 459.99 | 0.01 | 0.00 |
| **tr** | 1-1-2-trichlorotrifluoroethane | C 2 F 3 Cl 3 | 238.00 | 238.01 | 0.01 | 0.00 |
| **tr** | tetrafluoroethene | C 2 F 4 | 130.50 | 130.49 | 0.01 | 0.01 |
| **tr** | 1-2-dichlorotetrafluoroethane | C 2 F 4 Cl 2 | 179.00 | 179.00 | 0.00 | 0.00 |
| **tr** | chloropentafluoroethane | C 2 F 5 Cl 1 | 173.60 | 173.60 | 0.00 | 0.00 |
| **tr** | hexafluoroethane | C 2 F 6 | 172.30 | 172.31 | 0.01 | 0.01 |
| **tr** | trichloroethene | C 2 H 1 Cl 3 | 188.30 | 188.01 | 0.29 | 0.15 |
| **tr** | pentachloroethane | C 2 H 1 Cl 5 | 244.00 | 243.99 | 0.01 | 0.00 |
| **tr** | 1-1-2-2-tetrabromoethane | C 2 H 2 Br 4 | 229.20 | 229.21 | 0.01 | 0.00 |

| ID | Nombre | Fórmula | Exp. Punto de fusión K | Est. Punto de fusión K | Error absoluto K | Error relativo, % |
|---|---|---|---|---|---|---|
| **tr** | cis-1-2-dichloroethene | C 2 H 2 Cl 2 | 193.00 | 192.99 | 0.01 | 0.00 |
| **tr** | trans-1-2-dichloroethene | C 2 H 2 Cl 2 | 223.20 | 223.10 | 0.10 | 0.04 |
| **tr** | 1-1-1-2-tetrachloroethane | C 2 H 2 Cl 4 | 202.80 | 202.79 | 0.01 | 0.00 |
| **tr** | 1-1-2-2-tetrachloroethane | C 2 H 2 Cl 4 | 229.20 | 229.21 | 0.01 | 0.00 |
| **tr** | chloroethene | C 2 H 3 Cl 1 | 119.30 | 119.32 | 0.02 | 0.02 |
| **tr** | 1-1-1-trichloroethane | C 2 H 3 Cl 3 | 242.60 | 242.41 | 0.19 | 0.08 |
| **tr** | acetonitrile | C 2 H 3 N 1 | 229.20 | 229.21 | 0.01 | 0.00 |
| **tr** | 2-2-2-trifluoroethanol | C 2 H 3 O 1 F 3 | 229.50 | 229.21 | 0.29 | 0.13 |
| **tr** | 1-2-dibromoethane | C 2 H 4 Br 2 | 282.90 | 282.91 | 0.01 | 0.00 |
| **tr** | 1-bromo-2-chloroethane | C 2 H 4 Cl 1 Br 1 | 256.30 | 256.28 | 0.02 | 0.01 |
| **tr** | 1-1-dichloroethane | C 2 H 4 Cl 2 | 176.10 | 176.09 | 0.01 | 0.01 |
| **tr** | 1-2-dichloroethane | C 2 H 4 Cl 2 | 237.50 | 237.51 | 0.01 | 0.01 |
| **tr** | acetaldehyde | C 2 H 4 O 1 | 150.00 | 149.88 | 0.12 | 0.08 |
| **tr** | ethylene oxide (oxirane) | C 2 H 4 O 1 | 161.30 | 161.10 | 0.20 | 0.13 |
| **tr** | acetic acid | C 2 H 4 O 2 | 289.60 | 289.59 | 0.01 | 0.00 |
| **tr** | methyl formate | C 2 H 4 O 2 | 174.00 | 173.60 | 0.40 | 0.23 |
| **tr** | bromoethane | C 2 H 5 Br 1 | 154.40 | 154.41 | 0.01 | 0.01 |
| **tr** | chloroethane | C 2 H 5 Cl 1 | 134.30 | 134.31 | 0.01 | 0.01 |
| **tr** | iodoethane | C 2 H 5 I 1 | 161.90 | 161.89 | 0.01 | 0.01 |
| **tr** | nitroethane | C 2 H 5 N 1 O 2 | 183.50 | 183.48 | 0.02 | 0.01 |
| **tr** | ethanol | C 2 H 6 O 1 | 158.90 | 158.90 | 0.00 | 0.00 |
| **tr** | dimethylsulfoxide | C 2 H 6 O 1 S 1 | 291.50 | 291.50 | 0.00 | 0.00 |
| **tr** | dimethyl sulfide | C 2 H 6 S 1 | 174.70 | 174.72 | 0.02 | 0.01 |
| **tr** | ethanethiol | C 2 H 6 S 1 | 125.20 | 125.22 | 0.02 | 0.01 |
| **tr** | 3-chloro-1-propene | C 3 H 5 Cl 1 | 138.50 | 138.51 | 0.01 | 0.00 |
| **tr** | 1-2-3-trichloropropane | C 3 H 5 Cl 3 | 258.30 | 257.99 | 0.31 | 0.12 |
| **tr** | propanenitrile | C 3 H 5 N 1 | 180.20 | 179.99 | 0.21 | 0.11 |
| **tr** | 1-2-dibromopropane | C 3 H 6 Br 2 | 217.80 | 217.50 | 0.30 | 0.14 |
| **tr** | 1-2-dichloropropane | C 3 H 6 Cl 2 | 172.60 | 172.31 | 0.29 | 0.17 |
| **tr** | acetone | C 3 H 6 O 1 | 178.20 | 177.79 | 0.41 | 0.23 |
| **tr** | propionaldehyde | C 3 H 6 O 1 | 193.00 | 192.99 | 0.01 | 0.00 |
| **tr** | methyl acetate | C 3 H 6 O 2 | 175.00 | 174.72 | 0.28 | 0.16 |
| **tr** | 1-bromopropane | C 3 H 7 Br 1 | 193.40 | 192.99 | 0.41 | 0.21 |
| **tr** | 2-bromopropane | C 3 H 7 Br 1 | 163.00 | 163.01 | 0.01 | 0.00 |
| **tr** | 1-chloropropane | C 3 H 7 Cl 1 | 150.20 | 149.88 | 0.32 | 0.21 |
| **tr** | 1-iodopropane | C 3 H 7 I 1 | 171.70 | 171.31 | 0.39 | 0.22 |
| **tr** | 2-iodopropane | C 3 H 7 I 1 | 183.00 | 182.98 | 0.02 | 0.01 |
| **tr** | 1-nitropropane | C 3 H 7 N 1 O 2 | 165.00 | 164.71 | 0.29 | 0.18 |
| **tr** | 2-nitropropane | C 3 H 7 N 1 O 2 | 181.70 | 181.49 | 0.21 | 0.12 |
| **tr** | N-N-dimethylformamide | C 3 H 7 N 1 O 2 | 212.60 | 212.59 | 0.01 | 0.00 |
| **tr** | 1-propanol | C 3 H 8 O 1 | 146.90 | 146.89 | 0.01 | 0.00 |
| **tr** | 2-propanol | C 3 H 8 O 1 | 183.50 | 183.48 | 0.02 | 0.01 |
| **tr** | octafluorocyclobutane | C 4 F 8 | 232.90 | 232.90 | 0.00 | 0.00 |
| **tr** | 2-methyl-propane | C 4 H 10 | 134.80 | 134.81 | 0.01 | 0.01 |
| **tr** | n-butane | C 4 H 10 | 134.80 | 134.81 | 0.01 | 0.01 |
| **tr** | 1-butanol | C 4 H 10 O 1 | 183.20 | 182.98 | 0.22 | 0.12 |
| **tr** | 2-butanol | C 4 H 10 O 1 | 158.30 | 158.32 | 0.02 | 0.01 |
| **tr** | 2-methyl-1-propanol | C 4 H 10 O 1 | 165.00 | 164.71 | 0.29 | 0.18 |
| **tr** | diethylether | C 4 H 10 O 1 | 156.70 | 156.70 | 0.00 | 0.00 |
| **tr** | tert-butanol | C 4 H 10 O 1 | 298.40 | 298.40 | 0.00 | 0.00 |
| **tr** | 1-butanethiol | C 4 H 10 S 1 | 157.30 | 157.32 | 0.02 | 0.01 |
| **tr** | diethyl sulfide | C 4 H 10 S 1 | 169.10 | 169.11 | 0.01 | 0.01 |

| ID | Nombre | Fórmula | Exp. Punto de fusión K | Est. Punto de fusión K | Error absoluto K | Error relativo, % |
|---|---|---|---|---|---|---|
| **tr** | diethyl amine | C 4 H 11 N 1 | 223.20 | 223.10 | 0.10 | 0.04 |
| **tr** | thiophene | C 4 H 4 S 1 | 233.60 | 233.61 | 0.01 | 0.00 |
| **tr** | butyronitrile | C 4 H 7 N 1 | 161.10 | 161.10 | 0.00 | 0.00 |
| **tr** | isobutyronitrile | C 4 H 7 N 1 | 201.50 | 201.51 | 0.01 | 0.00 |
| **tr** | 2-butanone | C 4 H 8 O 1 | 186.40 | 186.10 | 0.30 | 0.16 |
| **tr** | tetrahydrofuran | C 4 H 8 O 1 | 164.70 | 164.71 | 0.01 | 0.01 |
| **tr** | 1-4-dioxane | C 4 H 8 O 2 | 284.80 | 284.82 | 0.02 | 0.01 |
| **tr** | ethyl acetate | C 4 H 8 O 2 | 189.40 | 189.01 | 0.39 | 0.21 |
| **tr** | ethyl propenoate | C 4 H 8 O 2 | 196.50 | 196.11 | 0.39 | 0.20 |
| **tr** | propyl formate | C 4 H 8 O 2 | 180.10 | 179.99 | 0.11 | 0.06 |
| **tr** | 1-bromobutane | C 4 H 9 Br 1 | 160.60 | 160.60 | 0.00 | 0.00 |
| **tr** | 1-chlorobutane | C 4 H 9 Cl 1 | 149.90 | 149.88 | 0.02 | 0.01 |
| **tr** | 2-chlorobutane | C 4 H 9 Cl 1 | 141.80 | 141.79 | 0.01 | 0.01 |
| **tr** | N-n-dimethyl acetamide | C 4 H 9 N 1 O 1 | 253.00 | 253.00 | 0.00 | 0.00 |
| **tr** | 1-pentene | C 5 H 10 | 107.80 | 107.81 | 0.01 | 0.01 |
| **tr** | 2-methyl-2-butene | C 5 H 10 | 139.30 | 139.29 | 0.01 | 0.00 |
| **tr** | 2-pentene | C 5 H 10 | 132.80 | 132.82 | 0.02 | 0.01 |
| **tr** | 3-methyl-1-butene | C 5 H 10 | 104.50 | 104.49 | 0.01 | 0.01 |
| **tr** | cyclopentane | C 5 H 10 | 179.20 | 179.00 | 0.20 | 0.11 |
| **tr** | 2-pentanone | C 5 H 10 O 1 | 196.10 | 196.11 | 0.01 | 0.00 |
| **tr** | 3-methyl-2-butanone | C 5 H 10 O 1 | 181.00 | 180.99 | 0.01 | 0.01 |
| **tr** | 3-pentanone | C 5 H 10 O 1 | 234.00 | 233.61 | 0.39 | 0.17 |
| **tr** | pentanal | C 5 H 10 O 1 | 181.50 | 181.49 | 0.01 | 0.01 |
| **tr** | tetrahydropyran | C 5 H 10 O 1 | 228.00 | 228.00 | 0.00 | 0.00 |
| **tr** | ethyl propanoate | C 5 H 10 O 2 | 199.10 | 199.01 | 0.09 | 0.04 |
| **tr** | isopropyl acetate | C 5 H 10 O 2 | 199.60 | 199.60 | 0.00 | 0.00 |
| **tr** | methyl butyrate | C 5 H 10 O 2 | 187.20 | 187.22 | 0.02 | 0.01 |
| **tr** | n-propyl acetate | C 5 H 10 O 2 | 180.00 | 179.99 | 0.01 | 0.00 |
| **tr** | pentanoic acid | C 5 H 10 O 2 | 239.00 | 238.88 | 0.12 | 0.05 |
| **tr** | 1-bromopentane | C 5 H 11 Br 1 | 178.00 | 177.79 | 0.21 | 0.12 |
| **tr** | 1-chloropentane | C 5 H 11 Cl 1 | 174.00 | 173.60 | 0.40 | 0.23 |
| **tr** | 2-chloro-2-methyl-butane | C 5 H 11 Cl 1 | 199.00 | 199.01 | 0.01 | 0.01 |
| **tr** | 2-2-dimethyl-propane | C 5 H 12 | 256.40 | 256.28 | 0.12 | 0.05 |
| **tr** | 2-methylbutane | C 5 H 12 | 113.10 | 113.09 | 0.01 | 0.01 |
| **tr** | pentane | C 5 H 12 | 143.30 | 143.28 | 0.02 | 0.01 |
| **tr** | 1-pentanol | C 5 H 12 O 1 | 194.10 | 194.11 | 0.01 | 0.01 |
| **tr** | 2-2-dimethyl-1-propanol | C 5 H 12 O 1 | 325.50 | 325.51 | 0.01 | 0.00 |
| **tr** | 3-methyl-1-butanol | C 5 H 12 O 1 | 155.80 | 155.78 | 0.02 | 0.01 |
| **tr** | ethyl propyl ether | C 5 H 12 O 1 | 145.50 | 145.48 | 0.02 | 0.01 |
| **tr** | methyl butyl ether | C 5 H 12 O 1 | 228.40 | 228.00 | 0.40 | 0.17 |
| **tr** | terbutyl methyl ether | C 5 H 12 O 1 | 164.00 | 164.00 | 0.00 | 0.00 |
| **tr** | pyridine | C 5 H 5 N 1 | 231.40 | 231.41 | 0.01 | 0.00 |
| **tr** | 1-4-pentadiene | C 5 H 8 | 124.20 | 124.22 | 0.02 | 0.02 |
| **tr** | 1-pentyne | C 5 H 8 | 183.00 | 182.98 | 0.02 | 0.01 |
| **tr** | 2-methyl-1-3-butadiene | C 5 H 8 | 127.10 | 127.08 | 0.02 | 0.01 |
| **tr** | cyclopentene | C 5 H 8 | 137.90 | 137.88 | 0.02 | 0.01 |
| **tr** | cyclopentanone | C 5 H 8 O 1 | 221.70 | 221.61 | 0.09 | 0.04 |
| **tr** | pentanenitrile | C 5 H 9 N 1 | 176.80 | 176.80 | 0.00 | 0.00 |
| **tr** | N-methyl-2-pyrrolidone | C 5 H 9 N 1 O 1 | 249.00 | 249.02 | 0.02 | 0.01 |
| **tr** | hexachlorobenzene | C 6 Cl 6 | 504.80 | 504.80 | 0.00 | 0.00 |
| **tr** | 1-5-hexadiene | C 6 H 10 | 132.30 | 132.32 | 0.02 | 0.01 |
| **tr** | 1-hexyne | C 6 H 10 | 141.10 | 141.08 | 0.02 | 0.01 |
| **tr** | cyclohexene | C 6 H 10 | 169.50 | 169.11 | 0.39 | 0.23 |

| ID | Nombre | Fórmula | Exp. Punto de fusión K | Est. Punto de fusión K | Error absoluto K | Error relativo, % |
|---|---|---|---|---|---|---|
| **tr** | cyclohexanone | C 6 H 10 O 1 | 242.00 | 241.71 | 0.29 | 0.12 |
| **tr** | hexanenitrile | C 6 H 11 N 1 | 192.70 | 192.70 | 0.00 | 0.00 |
| **tr** | 1-hexene | C 6 H 12 | 133.30 | 133.31 | 0.01 | 0.01 |
| **tr** | 2-3-dimethyl-1-butene | C 6 H 12 | 115.70 | 115.58 | 0.12 | 0.10 |
| **tr** | 4-methyl-1-pentene | C 6 H 12 | 119.40 | 119.32 | 0.08 | 0.07 |
| **tr** | cyclohexane | C 6 H 12 | 279.60 | 279.58 | 0.02 | 0.01 |
| **tr** | methylcyclopentane | C 6 H 12 | 130.50 | 130.49 | 0.01 | 0.01 |
| **tr** | 1-hexanol | C 6 H 12 O 1 | 157.50 | 157.32 | 0.18 | 0.12 |
| **tr** | 3-3-dimethyl-2-butanone | C 6 H 12 O 1 | 220.50 | 220.49 | 0.01 | 0.01 |
| **tr** | 3-hexanone | C 6 H 12 O 1 | 217.50 | 217.50 | 0.00 | 0.00 |
| **tr** | 4-methyl-2-pentanone | C 6 H 12 O 1 | 189.00 | 189.01 | 0.01 | 0.00 |
| **tr** | cyclohexanol | C 6 H 12 O 1 | 298.40 | 298.40 | 0.00 | 0.00 |
| **tr** | hexanal | C 6 H 12 O 1 | 217.00 | 217.00 | 0.00 | 0.00 |
| **tr** | ethyl butyrate | C 6 H 12 O 2 | 175.00 | 174.72 | 0.28 | 0.16 |
| **tr** | hexanoic acid | C 6 H 12 O 2 | 270.00 | 269.99 | 0.01 | 0.00 |
| **tr** | isobutyl acetate | C 6 H 12 O 2 | 174.20 | 174.18 | 0.02 | 0.01 |
| **tr** | n-butyl-acetate | C 6 H 12 O 2 | 195.00 | 194.99 | 0.01 | 0.01 |
| **tr** | 2-2-dimethylbutane | C 6 H 14 | 174.00 | 173.60 | 0.40 | 0.23 |
| **tr** | 2-3-dimethyl-butane | C 6 H 14 | 144.20 | 143.99 | 0.21 | 0.15 |
| **tr** | 2-methyl-pentane | C 6 H 14 | 119.30 | 119.32 | 0.02 | 0.02 |
| **tr** | 3-methylpentane | C 6 H 14 | 110.10 | 110.10 | 0.00 | 0.00 |
| **tr** | 2-2-dimethyl-1-butanol | C 6 H 14 O 1 | 258.00 | 257.99 | 0.01 | 0.00 |
| **tr** | 2-methyl-2-pentanol | C 6 H 14 O 1 | 170.00 | 169.99 | 0.01 | 0.01 |
| **tr** | 3-3-dimethyl-2-butanol | C 6 H 14 O 1 | 278.60 | 278.50 | 0.10 | 0.03 |
| **tr** | 3-methyl-3-pentanol | C 6 H 14 O 1 | 249.40 | 249.02 | 0.38 | 0.15 |
| **tr** | 4-methyl-2-pentanol | C 6 H 14 O 1 | 183.00 | 182.98 | 0.02 | 0.01 |
| **tr** | di-isopropylether | C 6 H 14 O 1 | 186.20 | 186.10 | 0.10 | 0.05 |
| **tr** | di-propylether | C 6 H 14 O 1 | 146.90 | 146.89 | 0.01 | 0.00 |
| **tr** | butyl ethyl amine | C 6 H 15 N 1 | 195.00 | 194.99 | 0.01 | 0.01 |
| **tr** | dipropyl amine | C 6 H 15 N 1 | 210.00 | 210.02 | 0.02 | 0.01 |
| **tr** | tri-ethylamine | C 6 H 15 N 1 | 158.30 | 158.32 | 0.02 | 0.01 |
| **tr** | 1-2-dichlorobenzene | C 6 H 4 Cl 2 | 256.30 | 256.28 | 0.02 | 0.01 |
| **tr** | 1-3-dichlorobenzene | C 6 H 4 Cl 2 | 248.20 | 247.81 | 0.39 | 0.16 |
| **tr** | 1-4-dichlorobenzene | C 6 H 4 Cl 2 | 325.70 | 325.51 | 0.19 | 0.06 |
| **tr** | bromobenzene | C 6 H 5 Br 1 | 242.40 | 242.41 | 0.01 | 0.01 |
| **tr** | chlorobenzene | C 6 H 5 Cl 1 | 227.80 | 227.79 | 0.01 | 0.00 |
| **tr** | fluorobenzene | C 6 H 5 F 1 | 230.80 | 230.79 | 0.01 | 0.01 |
| **tr** | iodobenzene | C 6 H 5 I 1 | 241.70 | 241.71 | 0.01 | 0.00 |
| **tr** | nitrobenzene | C 6 H 5 N 1 O 2 | 278.70 | 278.50 | 0.20 | 0.07 |
| **tr** | benzene | C 6 H 6 | 278.50 | 278.50 | 0.00 | 0.00 |
| **tr** | 3-methyl pyridine | C 6 H 7 N 1 | 254.90 | 254.91 | 0.01 | 0.01 |
| **tr** | 4-methyl pyridine | C 6 H 7 N 1 | 276.66 | 276.68 | 0.02 | 0.01 |
| **tr** | aniline | C 6 H 7 N 1 | 267.00 | 267.00 | 0.00 | 0.00 |
| **tr** | 1,4-cyclohexadiene | C 6 H 8 | 223.80 | 223.81 | 0.01 | 0.00 |
| **tr** | 1-6-heptadiene | C 7 H 12 | 144.00 | 143.99 | 0.01 | 0.01 |
| **tr** | 1-heptyne | C 7 H 12 | 192.00 | 192.00 | 0.00 | 0.00 |
| **tr** | 1-methyl-cyclohexene | C 7 H 12 | 152.60 | 152.58 | 0.02 | 0.01 |
| **tr** | cycloheptene | C 7 H 12 | 217.00 | 217.00 | 0.00 | 0.00 |
| **tr** | 2-heptene | C 7 H 14 | 163.50 | 163.51 | 0.01 | 0.00 |
| **tr** | cycloheptane | C 7 H 14 | 265.00 | 265.01 | 0.01 | 0.00 |
| **tr** | methylcyclohexane | C 7 H 14 | 146.40 | 146.40 | 0.00 | 0.00 |
| **tr** | 2-4-dimethyl-3-pentanone | C 7 H 14 O 1 | 204.00 | 204.00 | 0.00 | 0.00 |
| **tr** | 2-heptanone | C 7 H 14 O 1 | 238.00 | 238.01 | 0.01 | 0.00 |

| ID | Nombre | Fórmula | Exp. Punto de fusión K | Est. Punto de fusión K | Error absoluto K | Error relativo, % |
|---|---|---|---|---|---|---|
| **tr** | 4-heptanone | C 7 H 14 O 1 | 240.00 | 240.00 | 0.00 | 0.00 |
| **tr** | 5-methyl-2-hexanone | C 7 H 14 O 1 | 199.00 | 199.01 | 0.01 | 0.01 |
| **tr** | heptanal | C 7 H 14 O 1 | 229.70 | 229.71 | 0.01 | 0.00 |
| **tr** | ethyl pentanoate | C 7 H 14 O 2 | 181.80 | 181.49 | 0.31 | 0.17 |
| **tr** | isopentyl acetate | C 7 H 14 O 2 | 194.50 | 194.11 | 0.39 | 0.20 |
| **tr** | 2-4-dimethyl-pentane | C 7 H 16 | 153.10 | 153.08 | 0.02 | 0.01 |
| **tr** | 2-methyl-hexane | C 7 H 16 | 154.80 | 154.41 | 0.39 | 0.25 |
| **tr** | 3-3-dimethylpentane | C 7 H 16 | 138.10 | 137.88 | 0.22 | 0.16 |
| **tr** | n-heptane | C 7 H 16 | 182.40 | 182.40 | 0.00 | 0.00 |
| **tr** | 1-heptanol | C 7 H 16 O 1 | 239.00 | 238.88 | 0.12 | 0.05 |
| **tr** | 2-2-dimethyl-3-pentanol | C 7 H 16 O 1 | 270.50 | 270.49 | 0.01 | 0.00 |
| **tr** | 2-3-dimethyl-3-pentanol | C 7 H 16 O 1 | 243.00 | 242.99 | 0.01 | 0.00 |
| **tr** | 2-4-dimethyl-2-pentanol | C 7 H 16 O 1 | 253.00 | 253.00 | 0.00 | 0.00 |
| **tr** | 2-4-dimethyl-3-pentanol | C 7 H 16 O 1 | 203.00 | 202.79 | 0.21 | 0.10 |
| **tr** | 3-ethyl-3-pentanol | C 7 H 16 O 1 | 260.50 | 260.31 | 0.19 | 0.07 |
| **tr** | benzonitrile | C 7 H 5 N 1 | 260.30 | 260.31 | 0.01 | 0.00 |
| **tr** | benzoic acid | C 7 H 6 O 2 | 395.40 | 395.41 | 0.01 | 0.00 |
| **tr** | benzylchloride | C 7 H 7 Cl 1 | 228.00 | 228.00 | 0.00 | 0.00 |
| **tr** | 2-nitrotoluene | C 7 H 7 N 1 O 2 | 263.00 | 263.01 | 0.01 | 0.00 |
| **tr** | 3-nitrotoluene | C 7 H 7 N 1 O 2 | 288.50 | 288.51 | 0.01 | 0.00 |
| **tr** | 2-methoxynitrobenzene (2-nitroanisole) | C 7 H 7 N 1 O 3 | 283.50 | 283.49 | 0.01 | 0.00 |
| **tr** | 1-6-heptadiyne | C 7 H 8 | 188.00 | 188.01 | 0.01 | 0.00 |
| **tr** | 1-3-5-cycloheptatriene | C 7 H 8 | 193.50 | 193.49 | 0.01 | 0.00 |
| **tr** | toluene | C 7 H 8 | 178.10 | 177.79 | 0.31 | 0.17 |
| **tr** | m-cresol (3-hydroxytoluene) | C 7 H 8 O 1 | 284.80 | 284.82 | 0.02 | 0.01 |
| **tr** | methoxybenzene(anisole) | C 7 H 8 O 1 | 235.50 | 235.52 | 0.02 | 0.01 |
| **tr** | o-cresol (2-hydroxytoluene) | C 7 H 8 O 1 | 302.80 | 302.80 | 0.00 | 0.00 |
| **tr** | p-cresol (4-hydroxytoluene) | C 7 H 8 O 1 | 308.50 | 308.49 | 0.01 | 0.00 |
| **tr** | 2-aminotoluene | C 7 H 9 N 1 | 256.70 | 256.70 | 0.00 | 0.00 |
| **tr** | ethyl benzene | C 8 H 10 | 178.10 | 177.79 | 0.31 | 0.17 |
| **tr** | m-xylene | C 8 H 10 | 225.20 | 225.22 | 0.02 | 0.01 |
| **tr** | o-xylene | C 8 H 10 | 247.80 | 247.81 | 0.01 | 0.00 |
| **tr** | p-xylene | C 8 H 10 | 286.20 | 286.02 | 0.18 | 0.06 |
| **tr** | ethoxybenzene | C 8 H 10 O 1 | 243.50 | 243.49 | 0.01 | 0.00 |
| **tr** | 4-ethenylcyclohexene | C 8 H 12 | 164.10 | 164.00 | 0.10 | 0.06 |
| **tr** | 1-octyne | C 8 H 14 | 193.70 | 193.49 | 0.21 | 0.11 |
| **tr** | 1-octene | C 8 H 16 | 171.30 | 171.31 | 0.01 | 0.01 |
| **tr** | cis-1-2-dimethylcyclohexane | C 8 H 16 | 223.10 | 223.10 | 0.00 | 0.00 |
| **tr** | cyclooctane | C 8 H 16 | 287.80 | 287.81 | 0.01 | 0.00 |
| **tr** | octanal | C 8 H 16 O 1 | 286.00 | 286.02 | 0.02 | 0.01 |
| **tr** | n-hexyl-acetate | C 8 H 16 O 2 | 192.10 | 192.00 | 0.10 | 0.05 |
| **tr** | octane | C 8 H 18 | 216.20 | 216.21 | 0.01 | 0.00 |
| **tr** | 1-octanol | C 8 H 18 O 1 | 257.50 | 257.49 | 0.01 | 0.00 |
| **tr** | 2-2-3-trimethyl-3-pentanol | C 8 H 18 O 1 | 267.00 | 267.00 | 0.00 | 0.00 |
| **tr** | di-butyl-ether | C 8 H 18 O 1 | 177.80 | 177.79 | 0.01 | 0.00 |
| **tr** | styrene | C 8 H 8 | 242.00 | 241.71 | 0.29 | 0.12 |
| **tr** | acetophenone | C 8 H 8 O 1 | 293.00 | 293.00 | 0.00 | 0.00 |
| **tr** | indan | C 9 H 10 | 221.60 | 221.61 | 0.01 | 0.00 |
| **tr** | m-methyl styrene | C 9 H 10 | 186.10 | 186.10 | 0.00 | 0.00 |
| **tr** | p-methyl styrene | C 9 H 10 | 238.90 | 238.88 | 0.02 | 0.01 |
| **tr** | 1-2-3-trimethyl-benzene | C 9 H 12 | 247.60 | 247.60 | 0.00 | 0.00 |
| **tr** | 1-3-5-trimethylbenzene | C 9 H 12 | 228.30 | 228.00 | 0.30 | 0.13 |

| ID | Nombre | Fórmula | Exp. Punto de fusión K | Est. Punto de fusión K | Error absoluto K | Error relativo, % |
|---|---|---|---|---|---|---|
| **tr** | o-ethyl-toluene | C 9 H 12 | 192.20 | 192.00 | 0.20 | 0.11 |
| **tr** | propyl benzene | C 9 H 12 | 173.50 | 173.52 | 0.02 | 0.01 |
| **tr** | 2-6-dimethyl-4-heptanone | C 9 H 18 O 1 | 231.50 | 231.41 | 0.09 | 0.04 |
| **tr** | 2-nonanone | C 9 H 18 O 1 | 265.50 | 265.50 | 0.00 | 0.00 |
| **tr** | nonanal | C 9 H 18 O 1 | 336.00 | 335.98 | 0.02 | 0.01 |
| **tr** | butyl pentanoate | C 9 H 18 O 2 | 180.20 | 179.99 | 0.21 | 0.11 |
| **tr** | 1-nonanol | C 9 H 20 O 1 | 268.00 | 268.00 | 0.00 | 0.00 |
| **te** | trifluoromethane | C 1 H 1 F 3 | 117.90 | 137.01 | 19.11 | 16.21 |
| **te** | 1-pentadecanol | C 15 H 32 O 1 | 318.50 | 322.32 | 3.82 | 1.20 |
| **te** | 1-heptadecanol | C 17 H 36 O 1 | 326.80 | 322.32 | 4.48 | 1.37 |
| **te** | 1-1-2-trichloroethane | C 2 H 3 Cl 3 | 236.40 | 242.41 | 6.01 | 2.54 |
| **te** | 3-bromo-1-propene | C 3 H 5 Br 1 | 154.00 | 163.01 | 9.01 | 5.85 |
| **te** | 1-3-dibromopropane | C 3 H 6 Br 2 | 238.80 | 237.51 | 1.29 | 0.54 |
| **te** | 1-3-dichloropropane | C 3 H 6 Cl 2 | 173.50 | 172.31 | 1.19 | 0.69 |
| **te** | 2-chloropropane | C 3 H 7 Cl 1 | 155.80 | 149.88 | 5.92 | 3.80 |
| **te** | butyraldehyde | C 4 H 8 O 1 | 174.00 | 186.10 | 12.10 | 6.95 |
| **te** | methyl propanoate | C 4 H 8 O 2 | 185.50 | 189.01 | 3.51 | 1.89 |
| **te** | 2-bromobutane | C 4 H 9 Br 1 | 161.10 | 160.60 | 0.50 | 0.31 |
| **te** | 1-chloro-2-methyl-propane | C 4 H 9 Cl 1 | 142.70 | 141.79 | 0.91 | 0.64 |
| **te** | methyl isobutyrate | C 5 H 10 O 2 | 188.30 | 179.99 | 8.31 | 4.41 |
| **te** | 2-methyl-1-butanol | C 5 H 12 O 1 | 203.00 | 221.61 | 18.61 | 9.17 |
| **te** | 2-pentanol | C 5 H 12 O 1 | 200.00 | 194.11 | 5.89 | 2.94 |
| **te** | 2-methyl-1-pentene | C 6 H 12 | 137.30 | 133.31 | 3.99 | 2.90 |
| **te** | 2-hexanone | C 6 H 12 O 1 | 217.50 | 217.50 | 0.00 | 0.00 |
| **te** | n-propyl propanoate | C 6 H 12 O 2 | 197.10 | 194.99 | 2.11 | 1.07 |
| **te** | 1-chlorohexane | C 6 H 13 Cl 1 | 179.00 | 173.60 | 5.40 | 3.02 |
| **te** | hexane | C 6 H 14 | 177.70 | 143.99 | 33.71 | 18.97 |
| **te** | 2-3-dimethyl-2-butanol | C 6 H 14 O 1 | 259.00 | 257.99 | 1.01 | 0.39 |
| **te** | isopropyl propyl ether | C 6 H 14 O 1 | 186.20 | 194.99 | 8.79 | 4.72 |
| **te** | methyl hexanoate | C 7 H 14 O 2 | 202.00 | 194.11 | 7.89 | 3.90 |
| **te** | n-pentyl-acetate | C 7 H 14 O 2 | 202.20 | 192.00 | 10.20 | 5.05 |
| **te** | 2-2-dimethyl-pentane | C 7 H 16 | 149.20 | 137.88 | 11.32 | 7.59 |
| **te** | 1-2-4-trimethyl-benzene | C 9 H 12 | 229.20 | 247.60 | 18.40 | 8.03 |
| **te** | isopropylbenzene | C 9 H 12 | 177.00 | 173.52 | 3.48 | 1.97 |
| **te** | p-ethyl-toluene | C 9 H 12 | 210.70 | 192.00 | 18.70 | 8.88 |
| **te** | 5-nonanone | C 9 H 18 O 1 | 267.10 | 286.02 | 18.92 | 7.08 |

### 4.2.5 Actividad de fármacos frente al virus del HIV-1

La proteasa es una enzima que el VIH necesita para completar su proceso de autorreplicación dando lugar a nuevos virus capaces de infectar a otras células. En el proceso de replicación el VIH produce largas cadenas de proteínas que necesitan fragmentarse en trozos pequeños dando lugar a las proteínas y enzimas que ayudan a construir nuevas copias del virus. La fragmentación de las cadenas más largas es producida por la proteasa, sus inhibidores impiden que la fragmentación tenga lugar con lo que las proteínas que se forman dan lugar a copias defectuosas del VIH que, si bien no puede destruir la célula que infectó, ya no puede infectar más células.

Por sí solos los inhibidores de la proteasa no eliminan completamente el VIH del organismo. Sin embargo, con este tipo de medicamentos se ha observado que pueden reducir la cantidad del VIH hasta en un 99% aunque algunos sigan latentes dentro de las células infectadas.

Al producir nuevos virus defectuosos se lograría que al menos la infección por el VIH no se propagase dentro del organismo con la misma rapidez que lo hace en la actualidad y teóricamente se podría llegar a una cronificación de la infección del VIH ya que al haber menos virus, menos células CD4 se infectarían y la persona infectada podría combatir mejor las infecciones y vivir durante más tiempo.

La importancia de modelar la capacidad inhibitoria de fármacos existentes contra el VIH y extrapolar los resultados proponiendo nuevas drogas potencialmente efectivas es un campo muy prometedor. Por ello ha sido seleccionado como el último caso a discutir en la presente disertación.

Los datos de la actividad inhibitoria se tomaron del trabajo de Jalali-Heravi et al, (Jalali-Heravi M. et al., 2000). El conjunto de datos consiste en 107 inhibidores de la transcriptasa inversa (RT) del virus HIV-1, derivados del 1-[2-hydroxyethoxy)-methyl]-6-(phenylthio)thymine) (HEPT). El valor de 1/C es el que se usa como variable dependiente, C representa la concentración molar del fármaco requerida para alcanzar el 50% de protección de las células CD4 contra el efecto citopático del HIV-1.

Es necesario enfatizar que estamos ante estructuras complejas, y aunque es estudio y la comprensión del ciclo de vida del virus HIV por parte de diversas disciplinas sugiere que la inhibición del la proteasa y de la transcriptasa inversa constituye vías favorables para contrarrestar la infección, la creación de nuevos inhibidores requiere ir más allá y estudiar a detalle las estructuras moleculares y sus interacciones. Para dar una idea de la complejidad de las estructuras a las que se pretende inhibir la proteasa del HIV está formada por 99 residuos de amino ácidos (cada unidad de amino ácidos en una proteína se denomina residuo) todos ellos unidos por enlaces péptidos (la unión entre dos amino ácidos) entre el ácido carboxílico de un residuo y el grupo amino del segundo residuo. En una proteína o péptido, el orden o secuencia de los amino ácidos se denomina *estructura primaria*. Las estructuras secundarias, terciarias y cuaternarias dan idea acerca de la distribución 3-D de las proteínas. Hay muchos grupos funcionales diferentes presentes en una proteína: grupos aromáticos, alquilo, alcoholes, aminas y tioles. Sin embargo, para entender a las proteínas se requiere caracterizar a un nuevo fragmento, el grupo carbonilo, sobre todo enfatizando su forma en amidas y ácidos carboxílicos. ¿Por qué?, Simplemente porque al condensar un ácido carboxílico con una amina se genera una amida. Es el enlace de la amida el cual mantiene los

amino ácidos juntos para formar la estructura de la proteína. Por lo cual, para comprender las estructuras, y las propiedades físicas y químicas de las proteínas como la proteasa del HIV, es necesario caracterizar a los ácidos carboxílicos y a las amidas.

La metodología general se resume en el diagrama de flujo de la Fig. 4.13. La geometría de cada compuesto se optimizó con MOPAC 6.0 y el método PM3 para el cálculo de los diferentes parámetros cuánticos. Una vez generados los descriptores tanto topológicos, geométricos y químico-cuánticos que en este caso se incrementan al aumentar la complejidad del problema hasta 173 descriptores, se usan SOM con el fin de analizar la similitud entre los diferentes compuestos y seleccionar los índices más relevantes para caracterizar a la actividad inhibitoria. Una lista de los 173 descriptores y su nomenclatura se encuentran dentro de la lista de símbolos de la presente memoria entre los que se encuentran el índice de Wiener, índice de Randic, el índice de Schultz, el número de Harary, el índice de Balaban, el índice de Hosoya y la contribución de orden 1,2,3 a este mismo índice, los índices $\chi$, además de la mezcla con los índices topológicos clásicos y las versiones topológico-quánticas de los mismos, se generan la combinación de índices de la tabla, por otra banda se obtienen los descriptores usados en los casos previos como las similitudes cuánticas, de Overlap y Coulomb, los índices de conectividad de valencia, el índice de forma kappa, el calor de formación y los índices de resonancia, intercambio energético, y atracción-repulsión entre núcleo y electrones, entre otros . Los mapas topológicos para cada descriptor molecular junto con la variable objetivo nos permiten categorizar a los compuestos. La representación gráfica de los planos correspondientes a cada variable (C-planes), no se presentará en está ocasión debido al elevado número de variables.

Se realizan diversos experimentos de forma que es posible seguir la metodología propuesta hasta el momento y en el segundo caso comparar con los resultados publicados en la literatura. Los experimentos pueden resumirse de la siguiente forma:

(i) Un modelo fuzzy ARTMAP QSAR se estima usando la metodología propuesta SOM fuzzy ARTMAP (seleccionando el 15% de compuestos de forma aleatoria mediante el sistema neuronal fuzzy ART).
(ii) Un modelo fuzzy ARTMAP-QSAR se construye usando el conjunto de índices seleccionados mediante SOM pero usando el conjunto de entrenamiento propuesto por Jalali-Heravi, M. et al., (Jalali-Heravi M. et al., 2000)
(iii) Un modelo fuzzy ARTMAP QSAR se construye de la misma forma que en el caso (i) añadiendo los prototipos de las clases vacías al conjunto de entrenamiento.

(iv) Un modelo fuzzy ARTMAP QSAR se construye de la misma forma que en el caso (ii) añadiendo los prototipos de las clases vacías al conjunto de entrenamiento.

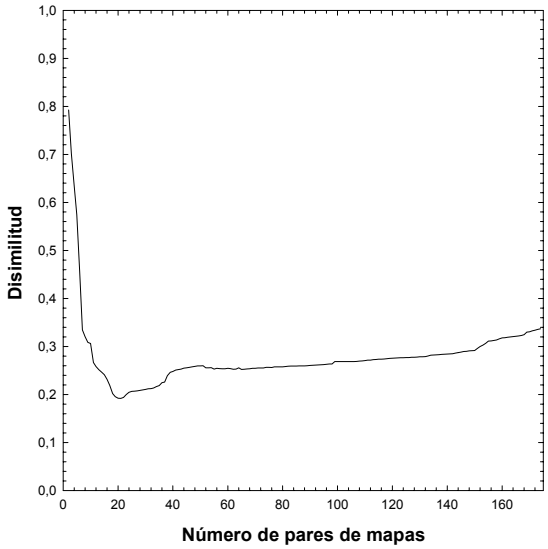



**Fig. 4.34.** Medida de disimilitud entre pares de SOM

Un resumen de los errores obtenidos en cada experimento se presenta en la Tabla 4.12. Hay en esencia dos diferencias entre cada experimento. El número de compuestos usados para el entrenamiento, y el usar o no los pseudo compuestos obtenidos de los vectores prototipos de SOM como parte del conjunto de entrenamiento. Un factor adicional además del puramente comparativo con respecto al trabajo de Jalali-Heravi et al., es el hecho de que en este trabajo se justifica al conjunto de prueba como aquellos derivados para los cuales se cita una actividad i*mprecisa*. Sin embargo, a lo largo de la memoria se ha puesto de manifiesto que la selección de un conjunto homogéneo de entrenamiento es la base de una buena generalización. Con lo cual además, será posible contrastar la selección de sistema autónomo como fuzzy ART para clasificar y proponer de forma aleatoria los compuestos de entrenamiento.

De SOM (definiendo un mapa de 10x10) se obtienen 19 grupos de variables. Los descriptores de cada grupo son en su mayoría del mismo tipo, es decir, los índices topológicos clásicos se distribuyen en dos grupos, mientras tanto los topológico-cuánticos o los electrotopológicos ocupan categorías distintas. El índice con la covarianza más elevada es el kappa (grupo A). En este caso, no se utiliza un conjunto básico de variables, es decir, el mapa inicial se construye partiendo de la variable kappa y el índice de actividad inhibitoria del HIV, y una a una se añaden las variables siguiendo criterio propuesto por Espinosa et al., (Espinosa G. et al., 2002). En este caso cada la mayoría de los grupos de variables, tenía

algún índice con la correlación más alta a la correlación media. De esta forma se van añadiendo una variable de cada grupo en el siguiente orden: el índice kappa (grupo A), DI0002 (grupo B), HLB (grupo C), DI0008 (grupo D), Sol (grupo E), $^{v}\chi^{1}$ (grupo F), HBD (grupo G), $\Delta$Hf (grupo H), DI0009 (grupo J), TX33 (grupo K), TX15 (grupo L), TX37 (grupo M), RE (grupo N) , $\rho$ (grupo P), TX39 (grupo Q), CX45 (grupo R), TE (grupo S) ,TX35 (grupo I), SA (grupo O) con lo se tenía un elemento representativo de cada grupo de índices, el proceso continúa añadiendo índices en orden decreciente de correlación con respecto a la variable objetivo (la actividad inhibitoria de los derivados del HEPT con respecto al virus del HIV).

La Fig. 4.34 presenta la disimilitud entre pares de mapas, el mínimo se observa al añadir la variable 20 (DI0009) al conjunto de variables presentadas anteriormente (una por grupo). En este caso la disimilitud mínima tiene un valor de 0.1874. Cabe destacar, que orden en el que se añaden los índices nos da una idea de la contribución de cada índice en el proceso de descripción de la propiedad, en este caso los factores de forma, son primordiales, sin dejar de ser por ello importantes aquellos que describen las interacciones espaciales y nos dan idea del tipo de heteroátomos presentes en la molécula.

En cada experimento se usan este conjunto de descriptores seleccionados con SOM, lo que cambia en esta ocasión son los compuestos usados para entrenar a la red y si se usa o no la información proporcionada por los vectores prototipos de SOM. El experimento uno pone de manifiesto el hecho de que la homogeneidad en ambos conjuntos (entrenamiento y generalización) es un factor determinante en el desempeño de un modelo QSAR. A pesar de que el número de compuestos utilizados para medir la generalización no es el mismo (aproximadamente el 15% para el caso uno), el error absoluto medio es la mitad en el caso uno que usando los mismos 20 índices y generalizando con el conjunto propuesto por Jalali-Heravi et al., (Jalali-Heravi M. et al., 2000), Tabla 4.12. Once de los 16 compuestos seleccionados por fuzzy ART no corresponden a los compuestos propuestos por Jalali-Heravi et al., (Jalali-Heravi et al., 2000). Los valores de la actividad inhibitoria del HIV estimados con 91 compuestos como conjunto de entrenamiento dan lugar a un error absoluto medio de 0.07 (1/C) (1.22%) y una desviación estándar del 0.21 (3.52%), la contribución al error absoluto medio del conjunto de entrenamiento y del de generalización es respectivamente 0.00 (1/C) (0.02%) y 0.47 (8.06%) en cada caso. Con una desviación estándar para cada conjunto de, 0.00 (0.05%) y 0.34 (5.39%), respectivamente. En comparación, usando el mismo conjunto de índices, pero el conjunto de entrenamiento propuesto por Jalali-Heravi, et al., (Jalali-Heravi M. et al., 2000) se obtiene un error absoluto medio de 0.16 (1/C) (3.92%), y una desviación estándar de 0.35 (8.70%), el error de generalización en este caso aumenta de un 8.0% para el caso anterior a un 15.5% aproximadamente para los 27 compuestos usados como conjunto de

prueba, Fig. 4.35 (A) y Fig. 4.36 (A). Realizando una rápida inspección a los grupos funcionales R1, R2 y R3 del conjunto de entrenamiento y comparándolos con los del conjunto de prueba, se identifican alguna ausencia de grupos característicos del conjunto de generalización en el conjunto de entrenamiento (no se incluyen compuestos con triples enlaces, ni R2= CONHPh, entre otros), lo cual puede en primera instancia disuadir la hipótesis de una mala medición o un grado de incertidumbre mayor en los valores experimentales del log (1/C) de estos compuestos, según lo propuesto por Jalali-Heravi et al., (Jalali-Heravi et al., 2000) se encuentras subestimados.

Con el objetivo de incrementar la generalización de los modelos QSAR-fuzzy ARTMAP, se entrenan dos redes fuzzy ARMAP correspondientes a los casos anteriores, tomando como información adicional al conjunto de entrenamiento los 49 prototipos derivados de SOM, como se explicó en el caso de $\ln \gamma^{\infty}$ . En este caso el parámetro de vigilancia se relaja de 0.999 a 0.997 para obtener un espectro de generalización más amplio, sacrificando el detalle de los vectores prototipos creados por artA y artB durante la fase de entrenamiento. El tercer experimento corresponde a las condiciones del experimento uno más 49 pseudo-compuestos creados con SOM a partir de los 20 índices seleccionados para entrenar el sistema fuzzy ARTMAP. El último experimento se realiza con las condiciones del experimento dos, más los 49 compuestos adicionales. Nuevamente una comparación formal entre ambos experimentos no es posible debido a la diferencia en el número de compuestos de entrenamiento. Sin embargo, es posible cuantificar el beneficio de incluir los vectores prototipos comparando los experimentos 1 y 3 y el 2 y 4. En el primer caso, el error absoluto medio se reduce de 0.07 log (1/C) (1.22%) a un 0.04 log (1/C) (0.61%). 3 compuestos del conjunto de prueba (el 73: $C_{21}H_{22}N_2O_2S_2$; el 75: $C_{16}H_{20}N_2O_3S_1$; y 107: $C_{12}H_{12}N_2O_2S_1$) se agruparon en clases pertenecientes a pseudo-compuestos, lo cual justifica la reducción de un 8.06% en el error de generalización para el experimento 1 con respecto al 5.72% que presenta el conjunto del experimento 3, en ambos casos se mantiene la desviación estándar del conjunto. Esta comparación se encuentra representada en la Fig. 4.35.

Una comparación similar se lleva a cabo entre el experimento 2 y 4, Fig. 4.36, Tabla 4.12. El error medio se reduce de un 0.16 log (1/C) (3.93%) a una 0.07 (1.55%) y el de generalización de un 0.63 log (1/C) (15.46%) a un 0.38 (8.79%), observándose una reducción considerable en la desviación estándar del conjunto de prueba de 11.31% (experimento 2) a 4.76% (experimento 4). Cuatro de compuestos fueron clasificados dentro de clases de pseudo-compuestos (82: $C_{14}H_{15}N_2O_4S_1Cl_1$; 84: $C_{14}H_{15}N_2O_4S_1F_1$; 85: $C_{14}H_{15}N_2O_4S_1Cl_1$; 102: $C_{13}H_{13}N_2O_4F_1S_1$). El hecho de que los compuestos químicos generados por SOM den lugar a categorías similares a compuestos totalmente ajenos al conjunto de entrenamiento aumenta la capacidad de generalización de la red. Cabe recordar que dentro de estos vectores prototipos se

encuentran algunos pertenecientes a categorías vacías, es que el vector prototipo de dichas categorías no es una combinación de los compuestos que pertenecen a esa clase, si no una extrapolación de las clases que rodean a dicho nodo.

Se lleva a cabo una comparación entre el modelo actual QSAR fuzzy ARTMAP con y sin vectores prototipos con los modelos publicados por Jalali-Heravi et al., (Jalali-Heravi M. et al., 2000). Jalali-Heravi y colaboradores presentan dos modelos uno de regresión multilineal con seis índices: el recíproco de la sombra del área sobre el plano YZ (1/S), que es considerado un índice geométrico, la relación de las cargas parciales sobre los átomos más positivos y más negativos (POS/NEG) y el calor de formación ($\Delta$Hf), que se consideran descriptores electrónicos, el cuadrado del número de átomos de carbono SP3 del substituyente R2 $(NCSP3\text{-}R2)^2$, el número de grupos hidroxilos sobre el substituyente R3 (NOH-R3) , el cubo de la suma de las posiciones de R1 sobre la constante del anillo aromático $(NS\text{-}R1)^3$ que se consideran índices topológicos. Los resultados se comparan con los modelos anteriores en la Fig. 4.37. El error medio global es de 0.64 log (1/C) (15.59% y una desviación estándar de 0.64 (15.80%). La no-linealidad de la relación entre la actividad inhibitoria y la estructura molecular se pone de manifiesto en el mismo trabajo de Jalali-Heravi, quien obtiene una reducción del error del 50% al utilizar un sistema neuronal para modelar la relación estructura actividad. El error absoluto medio para una backpropagation 6-6-1 es de 0.37 log (1/C) (7.75%) con una desviación estándar de 0.40 (10.19%). El error de generalización para el modelo MLR es de 0.61 log (1/C) (12.35%) que es equivalente al 0.63 log (15.46%) (1/C) del modelo fuzzy ARTMAP sin vectores prototipos y con el mismo conjunto de entrenamiento. Sin embargo la desviación estándar del modelo MLR es mayor que el modelo fuzzy ARTMAP, Tabla 4.12. El error medio del conjunto de prueba del modelo obtenido con una red backpropagation de architecutura 6-6-1 es de 0.45 log (1/C) (9.33%) con una desviación estándar de 0.49 log (1/C) (12.98%), que es mayor al mejor modelo fuzzy ARTMAP (experimento 4).

El desempeño de los modelos fuzzy ARTMAP-QSAR para estimar la actividad inhibitoria del virus HIV de compuestos derivados del HEPT es satisfactorio y demuestra una vez más la habilidad de un clasificador neuronal para predecir la actividad de compuestos químicos de forma bastante precisa. Paralelamente, SOM resultó ser una herramienta útil para discernir dentro de un conjunto predefinido de variables aquellas que resulten más relevantes para correlacionar una propiedad determinada.

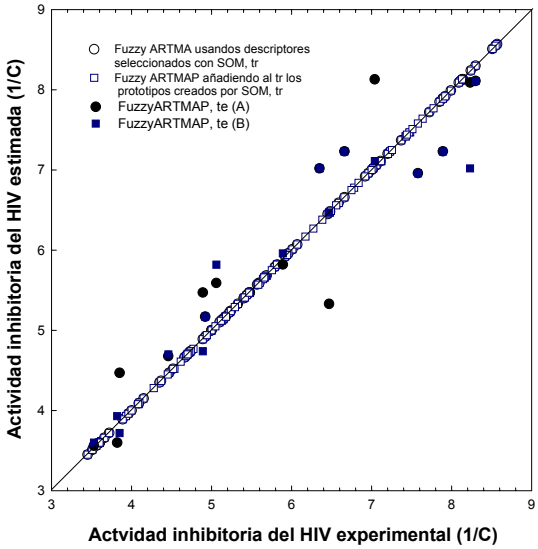



**Fig. 4.35.** Comparación de los valores experimentales de la actividad inhibitoria del HIV y los valores estimados con fuzzy ARTMAP basados en 20 descriptores moleculares seleccionados con SOM y entrenados con (A) 91 compuestos; (B) con 91 compuestos y la información adicional de los 49 vectores prototipos. Los descriptores seleccionados son kappa, DI002, HLB, DI008, Sol, $\chi(1)$, HBD, $\Delta$Hf, TX35, DI009, TX33, TX15, TX37, RE, SA, $\rho$, TX39, CX45, TE, DX1007 ***del conjunto original de compuestos, el 15% es seleccionado por fuzzy ART como conjunto de prueba**

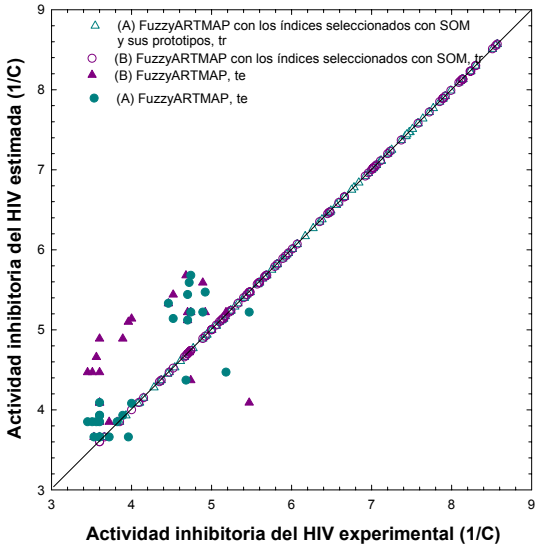



**Fig. 4.36.** Comparación de los valores experimentales de la actividad inhibitoria del HIV y los valores estimados con fuzzy ARTMAP basados en 20 descriptores moleculares seleccionados con SOM y entrenados con (A) 80 compuestos; (B) con 80 compuestos y la información adicional de los 49 vectores prototipos. Los descriptores seleccionados son kappa, DI002, HLB, DI008, Sol, $\chi(1)$, HBD, $\Delta$Hf, TX35, DI009, TX33, TX15, TX37, RE, SA, $\rho$, TX39, CX45, TE, DX1007. ***el conjunto de compuestos propuesto por Jalali et al., (Jalali-Heravi M. et al., 2000) es usado para entrenar y probar el modelo**

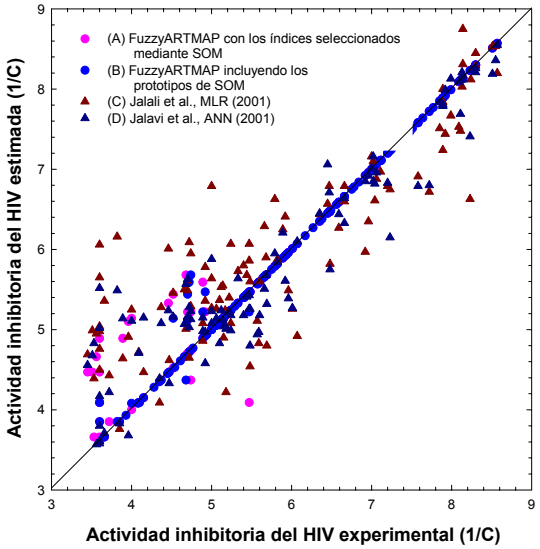



**Fig. 4.37.** Comparación de los valores experimentales de la actividad inhibitoria del HIV y los valores estimados con fuzzy ARTMAP basados en 20 descriptores moleculares seleccionados con SOM y entrenados con (A) 80 compuestos; (B) con 80 compuestos y la información adicional de los 49 vectores prototipos; (C) MLR, Jalali et al.; (D) Backpropagation 6-6-1, Jalali, M. et al., 2000).***el conjunto de compuestos propuesto por Jalali et al., (Jalali, M. et al., 2000) es usado para entrenar y probar el modelo***

**Tabla 4.12**. QSAR usando NN para estimar la actividad inhibitoria del HIV de derivados del HEPT. Los modelos hacen referencia a los 3 casos de estudio presentados en el texto.

| *Datos* | *No. de muestras* | *Error promedio total (%)* | *Desviación estándar (%)* |
|---|---|---|---|
| *Experimento 1:Fuzzy ARTMAP usando 20 índices y selección de índices con fuzzy ART* | | | |
| Todos los datos | 107 | 0.07 (1.22%) | 0.21 (3.52%) |
| Tr | 91 | 0.00 (0.02%) | 0.00 (0.05%) |
| *Te* | 16 | 0.47(8.06%) | 0.34 (5.39%) |
| *Experimento 2: Fuzzy ARTMAP usando 20 índices y el conjunto de entrenamiento propuesto por Jalali-Heravi, M (2000)* | | | |
| Todos los datos | 107 | 0.16 (3.92%) | 0.35 (8.70%) |
| Tr | 80 | 0.00 (0.02%) | 0.00 (0.05%) |
| *Te* | 27 | 0.63 (15.46%) | 0.45 (11.31%) |
| *Experimento 3: Fuzzy ARTMAP usando 20 índices + 49 vectores prototipos derivados de SOM y selección de índices con fuzzy ART* | | | |

| | | | |
|---|---|---|---|
| Todos los datos | 107 | 0.04 (0.61%) | 0.15 (2.27%) |
| Tr | 91 | 0.00 (0.03%) | 0.00 (0.06%) |
| *Te* | 16 | 0.36 (5.72%) | 0.34 (4.76%) |
| *Experimento 4 Fuzzy ARTMAP usando 20 índices + 49 vectores prototipos derivados de SOM y el conjunto de entrenamiento propuesto por Jalali-Heravi, M (2000)* | | | |
| Todos los datos | 107 | 0.07 (1.55%) | 0.18 (4.07%) |
| Tr | 80 | 0.00 (0.03%) | 0.00 (0.06%) |
| *Te* | 27 | 0.38 (8.79%) | 0.27 (5.71%) |
| *Jalali-Heravi, M. et al., (2000), MLR* | | | |
| Todos los datos | 107 | 0.64 (13.41%) | 0.64 (15.80%) |
| Tr | 80 | 0.64 (15.59%) | 0.63 (15.78%) |
| *Te* | 27 | 0.61 (12.35%) | 0.68 (16.36%) |
| *Jalali-Heravi, M. et al., (2000), ANN* | | | |
| Todos los datos | 104 | 0.37 (7.75%) | 0.40 (10.19%) |
| Tr | 80 | 0.36 (7.47%) | 0.38 (9.68%) |
| *Te* | 27 | 0.45 (9.33%) | 0.49 (12.98%) |

# *CONCLUSIONES*

En vista de los resultados presentados en la memoria, es posible afirmar que los modelos QSAR/QSPR/QSTR obtenidos integrando dos sistemas cognitivos, mapas auto-organizados (SOM) y fuzzy ARTMAP, son una alternativa viable para el cálculo de propiedades físico químicas y biológicas de los compuestos químicos.

A continuación se procede a enumerar las conclusiones particulares:

- Pudimos constatar que usar un algoritmo de estructura predeterminada como el *backpropagation* en contraste con un algoritmo auto-constructivo como *cascade correlation* no representa una limitación ya que la capacidad de generalización de backpropagation fue superior en los casos estudiados.

- Debemos tomar en cuenta que *backpropagation* no es el algoritmo ideal en muchos sentidos. Su popularidad se cimienta en la simpleza del mismo, a pesar de que en ocasiones es sumamente lento. Hay una restricción importante que no podemos pasar por alto y es su incapacidad de lidiar con el hecho de que patrones similares representen a propiedades muy distintas. La degeneración de los índices de conectividad molecular(índices similares para compuestos distintos), es común cuando se incrementa la diversidad de los compuestos usados. Este tipo de relaciones, desestabiliza el sistema y le impiden generalizar de una forma adecuada.

- Esta limitación puede resolverse usando un algoritmo no supervisado, cuyo objetivo será encontrar patrones o categorías a partir de las redundancias presentes en los datos de entrada. En este sentido, es la propia interrelación de los descriptores moleculares lo que proporcionará el conocimiento necesario a la red. Como pudimos constatar *fuzzy ARTMAP* es una opción viable para lograr este objetivo.

- La identificación de los descriptores moleculares más relevantes para construir los modelos QSAR/QSPR/QSTR pueden determinarse de forma sistemática con ayuda de mapas auto-organizados (SOM), estableciendo la influencia de los parámetros de entrada en el agrupamiento de los compuestos basándose en las similitudes encontradas.

- La selección del conjunto más factible de índices necesario para discernir entre clases de compuestos, por ejemplo los tóxicos de los no tóxicos se lleva a cabo al incorporar el índice más representativo de cada clase (el de correlación más alta con la variable objetivo).

- Los índices obtenidos se usan en para construir los modelos QSPR/QSAR o QSTR a usando fuzzy ARTMAP. Los resultados son equiparables en muchos casos al error experimental. Lo cual garantiza la bondad de la herramienta y los resultados proporcionados por la misma.

# *RECOMENDACIONES*

En base al trabajo realizado, es evidente que hay al menos un par de líneas de investigación prometedoras.

- Diseño de nuevos compuestos y optimización de los existentes
  Es deseable usar las relaciones estructura actividad no sólo para describir los posibles mecanismos de interacción, si no diseñar nuevas estructuras con propiedades específicas. Esto tiene una aplicación inmediata en la industria farmacéutica y agroquímica, donde se promueven fármacos cada vez más potentes y selectivos.

  La diversidad molecular requiere por supuesto caracterizar las estructuras en términos de sus propiedades físico químicas o cualquier otra propiedad deseable. Un área de interés se define como el calculo y la selección de descriptores que incluyan la funcionalidad química, el reconocimiento de receptores, la forma además de las características topológicas, geométricas y cuánticas tomadas en cuenta hasta el momento.

- Selectividad o bio-disponibilidad

  La selectividad en un fármaco, el control de la bio-disponibilidad, es crucial en el desarrollo de nuevos compuestos. La bio-disponibilidad de una substancia viene dada por diversos procesos como la absorción, distribución, el metabolismo y la excreción que se experimenta debido a un posible fármaco.

  Las propiedades de partición y la hidrofobicidad son las responsables de la mayor parte de las característicos de los procesos de distribución pasiva. De ahí que muchos de los trabajos en el modelado de fármacos se enfoquen en relacionar la estructura molecular con estos dos parámetros. Sin embargo son diversos factores los que intervienen en el transporte a través de membranas como la hidrofobicidad, el enlace hidrógeno, la ionización.

  De ahí que sería de gran utilidad encontrar y aplicar que variables diferencian el carácter de las membranas enfocado a aumentar su bio-disponibilidad y por ende su selectividad.

# 7. REFERENCIAS


1. Amat L. Robert D. Besalú E. and Carbó-Dorca R. Molecular Quantum Similarity Measures Tuned 3D QSAR: An Antitumoral Family Validation Study. *J. Chem. Inf. Compt. Sci*. **1998**, 38, 624.

2. Amat L. Carbó-Dorca R. and Ponec R. Simple Linear QSAR Models Based on Quantum Similarity Measures. *J. Med. Chem*. **1999**, 42, 5169.

3. Amat L. and Carbó-Dorca R. Quantum Similarity Measures under Atomic Shell Approximation: First Order Density Fitting using Elementary Jacobi Rotations*. J. of Compt. Chem*. **1997**, 18, 2023.

4. Anzali S. Gasteiger J. Holzgrabe U. Polanski J. Sadowski J. Teckentrup A. Wagener M. The Use of Self-Organizing Neural Networks in Drug Design. In *3D QSAR in Drug Design*, 2. H. Kubinyi G. Folkers Y. Martin Ed., Kluwer/ESCOM, Dordrecht, NL. **1998**, 273.

5. Balaban A. Basak C. Colburn T. Grunwald G. Correlation between Structure and Normal Boiling Points of haloalkanes C1-C4 using Neural Networks*. J. Chem. Inf. Comput. Sci*. **1984**, 34, 1118.

6. Bartfai B. Hierarchical Clustering with ART Neural Networks. *In proceedings of IEEE International Conference of neural networks*. **1994**, 2, 940.

7. Bartfai B. On the Match Tracking Anomaly of ARTMAP Neural Network. *Neural Networks*. **1996**, 9, 295.

8. Bartfai B. An ART-based Modular Architecture for Learning Hierarchical Clustering. *Neurocomputing*. **1996**, 13, 31.

9. Bartfai B. An Adaptive Resonance Theory-based Neural Network capable of Learning via Representational Redescription. *In proceedings of IEEE International Joint Conference of neural networks*. **1998**, 1137.

10. Basak S. Gute B. Grunwald G. Use of Topostructural, Topochemical and Geometrical parameters in the Prediction of Vapor Pressure: a Hierarchical QSAR Approach. *J. Chem. Inf. Compt. Sci*., **1997**, 37, 651.

11. Basak S. Magnunson V Niemi G. Regal R. Veith G.Topological Índices: their Nature, Mutual Relatedness, and Applications. *Mathematical Modeling,* **1997**, 8, 300.

12. Basak S. Gute B. and Ghatak S. Prediction of Complement-Inhibitory Activity of Benzamides Using Topological and Geometrical Parameters. *J. Chem. Inf. Compt. Sci*., **1999**, 39, 255.

13. Basak S. Gute B. Lu•i• B. Nikoli• S. and Trinajsti• N. A comparative QSAR Study of Benzamidines Complement-Inhibitory Activity and Benzene Derivatives Acute Toxicity. *SAR and QSAR in Environmental Research*. **1997**, 7, 117.

14. Basak S. Gute B. Grunwald G. Assessment of the Mutagenicity of Aromatic Amines from Theoretical Structural Parameters: A Hierarchical Approach. *SAR and QSAR in Enviromental Research*. **1999**, 10, 117.

15. Bünz P. Braun B. Janowsky R. Application of Quantitative Structure-Performance Relationship and Neural Network Models for the Prediction of Physical Properties from Molecular Structure. *Ind. Eng. Chem. Res*. **1998**, 37, 3043.

16. Bodor N. Gabanyi Z. and Wong C. A New Method for the Estimation of Partition Coefficient. *J. Am. Chem. Soc*. **1989**, 111, 3783.

17. Bukrinsky M. Haggerty S. Dempsey M. Sharova N. Adzhubei A. Spitz L. Lewis P. Goldfarb D. Emerman M. Stevenson M. A Nuclear Localization Signal within HIV Matrix Protein that Governs Infection on Non-Dividing Cells, *Nature*, **1993**, 365, 14, 666.

18. Carbó R. Arnau J. Leyda L. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int J. Quantum Chem*. **1980**, 17, 1185.

19. Carbó-Dorca R. and Besalú E. A general survey of molecular quantum similarity. *J. of Mol. Struct*., **1988**, 451, 11.

20. Carbó-Dorca R. Besalú E. A General Survey of Molecular Quantum Similarity. *Theo Chem*. **1998**, 45, 11.

21. Carpenter A. Grossberg S. The ART of Adaptive Pattern Recognition by a Self-organizing Neural Network. *Computer*. **1988**,77.

22. Carpenter G. Grossberg S. Marcuzon N. Reynolds J. Rosen D. Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps. *IEEE Trans. on Neural Networks*. **1992**, 3, 698.

23. Carpenter G. Grossberg S. Marcuzon N. Rosen D. Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System. *Neural Networks*. **1991**, 4, 759.

*24.* Carpenter G. Grossberg S. A Self-Organizing Neural Network for Supervised Learning, Recognition, and Prediction. *IEEE Communications Magazine*. **1992**, 38.

25. Carpenter A. Grossberg S. A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine. *Computer Vision, Graphics, and Image Processing*. **1987**, 37, 54.

26. Carpenter G. Grossberg S. *Pattern Recognition by Self-Organizing Neural Networks*; MIT Press: Cambridge (MA), **1991**.

27. Carpenter G. Grossberg S. *Neural Network for Vision and Image Processing*; MIT Press: Cambridge (MA), **1992**.

28. Cramer R. Patterson D. and Bunce J. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carried proteins. *J. Am. Chem. Soc*. **1988**, 110, 18, 5959.

29. Daubert T. Danner R. *Data Compilation Tables of Properties of Pure Compounds*. Design Institute for Physical Property Data, AICHE: New York,**1984**.

30. Design Institute for Physical Property Data (DIPPR) – Physical and thermodynamic property database from PRO/II Version. 4.02; SIMCI Simulation Sciences Inc. **1997**.

31. Egolf L. Jurs P. Prediction of Boiling Points of Organic Heterocyclic Compounds Using Regression and Neural Network Techniques Structure. *J. Chem. Inf. Comput. Sci*. **1993**, 33, 616.

32. Egolf L. Jurs P. Prediction of Boiling Points and Critical Temperatures of Industrially Important Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci*. **1996**, 34, 947.

33. Espinosa G. Yaffe D. Cohen Y. Arenas A. Giralt F. Neural Network Based Quantitative Structural Property Relations (QSPRs) for Predicting Boiling Points of Aliphatic Hydrocarbons. *J. Chem. Inf. Comput. Sci*. **2000**, 40, 859

34. Espinosa G. Yaffe D. Arenas A. Cohen Y. Giralt F. A Fuzzy ARTMAP based Quantitative Structure-Property Relationships (QSPRs) for Predicting Physical Properties of Organic Compounds, *Ind. Eng. Chem. Res.* **2001**, 40(12), 2757

35. Espinosa G. Arenas A. Giralt F. Prediction of Boiling Points of Organic Compounds from Molecular Descriptors by using Back-propagation Neural Networks, in Fundamentals of molecular similarity, Ed. R. Carbo-Dorca, Kluwer: Academic, **2001**, 1.

36. Espinosa G. Arenas A. Giralt F. An Integrated SOM-fuzzy ARTMAP Neural System for the Evaluation of Toxicity. *J. Chem. Inf. Compt. Sci., J. Chem. Inf. Compt. Sci.,* **2002***,.42, 2, 343*

37. Espinosa G. Arenas A. Giralt F. Amat L. Girones X. Carbó-Dorca R. Fuzzy ARTMAP-Based QSAR for $TD_{50}$ of Aromatic Compounds. *Submited*

38. Espinosa G. Arenas A. Giralt F. Amat L. Girones X. Carbó-Dorca R. QSAR for TD50 of Aromatic Compounds by using an Integrated SOM-Fuzzy ARTMAP based neural system with topological and Quantum Molecular Similarity Descriptors. *Submited.*

39. Estrada E. Ramirez A. Edge Adjacency Relationships and Molecular Topographic Descriptors. Definitions and QSAR Applications*. J. Chem. Inf. Comput. Sci.* **1996**, 36, 837.

40. Erwing E. Obermayer K. Schulten K. Self-organizing maps: Stationary states, metastability and convergence rate. *Biological Cybernetics*. **1992**, 67 (1), 35.

41. Fahlman, S. and Lebiere, C. The Cascade Correlation Learning Architecture. In Advances in Neural Information Processing Systems II (ed. S.S. Touretzky, (Denver 1989 ). San Mateo: Morgan Kaufmann, **1990**, 524.

42. Ferré-Gine J. Rallo R. Arenas A. Giralt F. Extraction of Structures Embedded in the Velocity Field of a Turbulent Wake, in Solving Engineering Problems with Neural Networks. *Proceedings of the International Conference on Enginnering Applications of Neural Networks* (EANN'969; Bulsari A. B. Kallio S. Tsaptsinos D. Turku, Eds., **1996**, 1, 17.

43. Fredenslund A. Jone R. Prausnitz J. Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures. *AICHE J.* **1975**, 21, 6, 1086.

44. Gasteiger J. Li X. Rudolph C. Sadowski J. and Zupan J. Representation of Molecular Electrostatic Potentials by Topological Feature Maps. *J. Am. Chem. Soc*. **1994**, 116, 4608.

45. Gasteiger J. Li X. and Uschold A. The Beauty of Molecular Surfaces as Revealed by Self-Organizing Neural Networks. *J. Mol. Graphics*, **1994**, 12, 90.

46. Gakh A. Gakh E. Sumpter B. Noid D. Neural Network-Graph Theory Approach to the Prediction of the Physical Properties of Organic Compounds. *J. Chem. Inf. Comput. Sci*. **1994**, 34, 832.

47. Gini G. Lorenzini M. Benfenati E. Grasso P. and Bruschi M. Predictive Carcinogenicity: A Model for Aromatic Compounds, with Nitrogen-Containing Substituents, based on Molecular Descriptors using Artificial Neural Network. *J. Chem. Inf. Compt. Sci.* **1999**, 39, 1076.

48. Giralt F. Arenas A. Ferre-Giné J. Rallo R. The Simulation and Interpretation of Turbulence with a Cognitive Neural System. *Physics of Fluids*. **2000**, 12, 1826.

49. Grossberg S. How does a brain build a cognitive code. *Physchological Review*. **1987**, 151.

50. Gute B. Basak S. Predicting Acute Toxicity (LC50) of Benzene Derivatives using Theoretical Molecular Descriptors: A Hierarchical QSAR Approach. *SAR and QSAR in Environmental Research*. **1997**, 7, 117.

51. Hall L. Story C. Boiling point and Critical Temperature of a Heterogeneous Data Set: QSAR Atom Type Electrotopological State Índices Using Artificial Neural Networks. *J. Chem.Inf. Compt. Sci*. **1996**, 36, 1004.

52. Hall L. Kier L. Electrotopological State Índices for Atom Types: A Novel Combination of Electronic, Topological and Valence State Information. *J. Chem. Inf. Comput*. Sci. **1995**, 35, 1039.

53. Hall L. Kier L. Phipps G. Structure-Activity Relationship Studies on the Toxicities of Benzene Derivatives I an Additive Model. *Environ. Toxicol. Chem*. **1984**, 3, 355.

54. Hansen C. The three dimensional solubility parameter- key to paint component affinities. II. Dyes, emulsifiers, mutual solubility and compatibility, and pigments. *J. Paint Technol*. **1979**, 39, 509.

55. Hansch C. Fujita T. A method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc*. **1964**, 86, 1616.

56. Hansen C. The three dimensional solubility parameter- key to paint component affinities. II. Dyes, emulsifiers, mutual solubility and compatibility, and pigments. *J. Paint Technol*. **1979**, 39, 509.

57. Hertz, J. Krogh, A. Palmer, R., Introduction to the Theory of Neural Computation. Santa Fe Institute Studies in the Sciences of Complexity. Ed. Addison-Wesley, 1991.

58. Ho D. Neumann A. Perelson A. Chen W. Leonard J. Markowitz M. Rapid Turnover of Plasma Viriosn and CD4 lymphocytes in HIV-1 Infection. *Nature*, **1995**, 373, 12.

59. Ivanciuc O. Taraviras S. Cabrol-Bass D. Quasi-orthogonal Basis Sets of Molecular Graph Descriptors as a Chemical Diversity Measure. *J. Chem. Inf. Comput. Sci*. **2000**, 40, 126.

60. Jalali-Heravi M. Parastar F. Use of Artificial Neural Networks in QSAR Study of Anti-HIV Activity for a Large Group of HEPT Derivatives. *J. Chem. Inf. Comput. Sci.* **2000**, 40, 147.

61. Joback K. Reid R. Estimation of Pure Component Properties from Group Contributions. *Chem. Eng. Common*. **1987**, 233.

62. Jurs P. Prediction of Chemical Properties of Organic Compounds from Molecular Structure. *214th ACS National Meeting*. **1997**.

63. Kaski S. and Kohonen T. Winner-take-all Networks for Physiological Models of Competitive Learning*. Neural Networks*. **1994**, 7, 973

64. Kaski S. and Lagus K. Comparing Self-organizing maps. In *Proc. of ICANN'96*. **1996**, 809

65. Katritzky A. Maran U. Lobanov V. Karelson M. Structurally Diverse Quantitative Structure-Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci*. **2000**, 40, 1.

66. Katritzky R. Mu L. Lobanov V. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem*. **1996**, 100, 10400.

67. Katrizky A. Mu L. Relationships of Critical Temperatures to Calculated Molecular Properties. *J. Chem. Inf. Compt. Sci*. **1998**, 38, 293.

68. Katritzky A. Karelson M. and Lobanov V. QSPR as Means of Predicting and Understanding Chemical and Physical Properties in Terms of Structure. *Pure & Appl. Chem*. **1997**, 69, 2, 245.

69. Katritzky A. Maran U. Karelson M. and Lobanov V, Prediction of Melting points for the Substituted Benzenes: A QSPR Approach. *J. Chem. Inf. Comput. Sci*., **1997,** 37, 913.

70. Katritzky A. Mu L. A QSPR study of the solubility of gases and vapors in water. *J. Chem. Comput. Sci*. **1996**, 36, 1162.

71. Karelson M. Lobanov S. and Katritzky A. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev*. **1996**, 96, 1027.

72. Kaski, S and Kohonen, T. Winner-take-all networks for physiological models of competitive learning*. Neural Networks*. **1994**, 7, 973.

73.Kelly T. Proudfoot J. McNeil D. Patel U. David E. Hargrave K. Grob P. Cardozo M. Agarwal A. Adams J. Novel Non-Nucleoside Inhibitors of Human Immunodeficiency Virus Type-1 Reverse Transcriptase. 5-4 substituted and 2-4 disubstituted Analogs of Nevirapine. *J. Med. Chem*. **1995**, 38, 4839.

74.Kier L. Hall L. Molecular Connectivity in Chemistry and Drug Research, Academic Press: New York, **1976**.

75.Kier L. A Shape Index from Molecular Graphs. *Quant. Struct-Act. Relat*. **1985**, 4, 109.

76.Kier L. Hall L. An Electrotopological State Index for Atoms in Molecules. *Pharm. Res*. **1990**, 7, 801.

77.Kireev D. Chrétien J. Grierson D. and Monneret C. 3D QSAR Study of a Series of HEPT Analogues: The Influence of Conformational Mobility on HIV-1 Reverse Transcriptase Inhibition. *J. Med. Chem*. **1997**, 40, 4257.

78.Kohonen T. Self-Organizing Formation of Topologically Correct Feature Maps. *Biological Cybernetics*. **1982**, 43, 59

79.Kohonen T. Physiological Interpretation of the Self-organizing map algorithm. *Neural Networks*. **1993**, 6, 895

80.Kohonen T. The Self-Organizing Map. *Proc*. *IEEE*. **1990**, 78, 1464

81.Korber B. Moore J. Souza P. Brander C. Walker B. Koup R. Haynes B. Myers G. HIV molecular inmuno database 1996 II: Parts I and III, **1996**. Publisher: Los Alamos National Laboratory Theoretical and Biophysics, website:hiv-web.lanl.gov/immuno.

82.Krzyzaniak J. Myrdal P. Simamora P. Yalkowsky S. Boiling Point and Melting Point Prediction for Aliphatic, Non-Hydrogen-Bonding Compounds. *Ind. Eng. Chem. Res*. **1995**, 34, 2530.

83.Lazaridis T. and Paulaitis M. Activity Coefficients in Dilute Aqueous Solutions from Free Energy Simulations. *AICHE J*. **1993**, 39, 6, 1051.

84.Lee M. Chen J. Fluid Property Predictions with the Aid of Neural Networks. *Ind. Eng. Chem. Res*. **1993**, 32, 995.

85.Lymman W. Reehl W. Rosenblatt D. *Handbook of Chemical Property Estimation Methods*, 3rd edition, American Chemical Society: Washington DC, **1990**.

86.Luco J. Ferretti F. QSAR based on Multiple Linear Regression and PLS Methods for the Anti-HIV Activity of a Large Group of HEPT Derivatives. *J. Chem. Inf. Comput. Sci*. **1997**, 37,392.

87. Mackay D. Shiu W. Y. and Ma K. Illustrated Handbook of Physical Chemical Properties and Environmental Fate for Organic Chemicals. 1, ed. Lewis Publishers, **1992**.

88. Mackay D.and Shiu Y. Aqueous Solubility of Polinuclear Aromatic Hydrocarbons. *J. Chem. Eng. Data*. **1977**, 22,399.

89. McWeeny, R. Methods of Molecular Quantum Mechanics. Academic Press, New York, **1989**.

90. Mandloi M. Sikarwar A. Sapre N. Karmarkar S. and Khadikar P. A Comparative QSAR Study Using Wiener, Szeged, and Molecular Connectivity Índices. *J. Chem. Inf. Comput. Sci*., **2000,** 40, 57.

91. Medir M. Giralt F. Correlation of Activity Coefficients of Hydrocarbons in Water at Infinity Dilution with Molecular Parameters. *AIChE Journal*. **1982**, 28, 341.

92. Miller E. Li L. and Desimone R. A neural mechanism for working and recognition memory in inferior temporal cortex*. Science*. **1991**, 254,1377.

93. Molecular Modeling Pro. Ver. 3.14. ChemSM Inc. **1998**.

94. Mitchell B. and Jurs P. Prediction of Infinite Dilution Activity Coefficients of Organic Compounds in Aqueous Solution from Molecular Structure*. J Chem. Inf. Comput. Sci*. **1998**, 38,200.

95. Needham D. Wei I. Seybold P. Molecular Modeling of the Physical Properties of the Alkanes. *J. Am. Chem. Soc*. **1988**, 110, 4186.

*96.* Pogliani L. Molecular Modeling by Linear Combinations of Connectivity Indexes. *J. Phys. Chem*. **1995**, 99, 925.

*97.* Pogliani L. Modeling with Special Descriptors Derived from a Medium-sized Set of Connectivity Índices. *J. Phys. Chem*. **1996**, 100, 18065.

98. *POC (Properties of Organic Compounds)* - Personal Edition, Version 5.1; CRC Press, Inc.: Boca Raton, 1996.

99. Randic M. On Characterization of Molecular Branching. *J. Am. Chem. Soc*. **1975**, 97, 6609.

100. Randic M. Trinajstic, N. Comparative Structure-Property Studies: the Connectivity Basis. *Journal of Molecular Structure*. **1993**, 284, 209.

101. Randic M. On Molecualr Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, 24, 164.

102. Randic M. Dobrowolski J. Optimal Molecular Connectivity Descriptors for Nitrogen-Containing Molecules. *International Journal of Quantum Chemistry*. **1998**, 70, 1209.

103. Randic, M. Basak, S. On Use of the Variable Connectivity Index $^{1}\chi^{f}$ in QSAR: Toxicity of Aliphatic Ethers. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 614.

104. Raymond J. Rogers T. Molecular Structure Disassembly Program (MOSDAP): A Chemical Information to Automate Structure-Based Physical Property Estimation. *J. Chem. Inf. Comput. Sci*. **1999**, 39, 463.

105. Reid R. Prausnitz J. Sherwood T. *The Properties of Gases and Liquids*; 3rd edition; McGraw-Hill: New York, 1987.

106. Rumelhart D. Hinton G. Williams, R. Learning Representations by Back-propagating Errors. *Nature*. **1986**, 323, 533.

107. Sabljic A. Horvatic D. Graph III: A computer Program from Calculation Molecular Connectivity Índices on Microcomputers. *J. Chem. Inf. Comput. Sci*. **1993**, 33, 837.

108. Sherman S. Trampe B. Bush D. Shiller M. Eckert C. Dallas A. Li J. and Carr P. Compilation and Correlation of Limiting Activity Coefficients of Nonelectrolytes in Water. *Ind. Eng. Chem. Res*. **1996**, 35,1044.

109. Simamora P. Miller A. Yallkowsky S. Melting Point and Normal Boiling Point Correlations: Applications to Rigid Aromatic Compounds. *J. Chem. Inf. Comput. Sci*. **1993**, 33, 437.

110. Stanton D. Jurs P. Development and Use of Charged Partial Surface Area Structural Descriptors for Quantitative Structural-Property Relationship Studies. *Anal. Chem*. **1990**, 62, 2323.

111. Stanton D. Egolf L. Jurs P. Hicks M. Computer-Assisted Prediction of Normal Boiling Points of Furans, Tetrahydrofurans and Thiophenes. *J. Chem. Inf. Comput. Sci*. **1991**, 31, 301.

112. Stanton D. Egolf L. Jurs P. Hicks M. Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles. *J. Chem. Inf. Comput. Sci*. **1992**, 32, 306.

113. Stewart J. MOPAC 6.0. Quantum Chemistry Program Exchange No. 455, Bloomington, IN, **1989**.

114. Tamayo P. Slonim D. Mesirov J. Zhu Q. Kitareewan S. Dmitrovshy E. Lander E. and Golub T. Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation*. Proc. Natl. Acad. Sci. USA*., **1996**, 96, 2907.

115. Tetko I. Tanchuk V. Chentsova N. Antonenko S. Poda G. Kikhar V. Luik A. HIV-1 Reverse Transcriptase Inhibitor Design using Artificial Neural Networks. *J. Med. Chem*. **1994**, 37, 2520.

116. Tetteh J. Suzuki T. Metcalfe E. Howells S. Quantitative Structure-Property Relationships for the Estimation of Boiling Point and Flash Point Using a Radial Basis Function Neural Network. *J. Chem. Inf. Compt. Sci*. **1999**, 39, 491.

117. Tochigi K. Tiegs D. Gmehling J. Kojima K. Determination of New ASOG Parameters. *J. Chem. Eng. Jpn*. **1990**, 23, 453.

118. Trinajsti•, N. Chemical Graph Theory, 2nd ed.; CRC Press: Boca Raton, Fl, **1992**.

119. Turner B. Costello C. Jurs P. Prediction of Critical Temperatures and Pressures of Industrially Important Organic Compounds from Molecular Structure. *J. Chem. Inf. Compt. Sci*. **1998**, 38, 639.

120. Vesanto, J. Data Mining Techniques Based on the Self-Organizing Map. Master thesis, Helsinki University of Technology. **1997**.

121. Vesanto J. SOM-Based Data Visualization Methods. *Intelligent Data Analysis*. **1999**, 6, 111.

122. Wagener M. Sadowski J. and Gasteiger J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc*. **1995**, 117, 7769.

123. Warne M. Boyd E. Meharg A. Osborn D. Killham K. Lindon J. and Nicholson J. Quantitative Structure-Toxicity Relationships for Chlorophenols to bioluminescent Lux-Marked Bacteria Using Atom-Based Semi-Empirical Molecular-Orbital Descriptors . SAR and *QSAR in Environmental Research*. **1999**, 10, 473.

124. Wei X. Ghosh S. Taylor M. Johnson V. Emini E. Deutsch P. Lifson J. Bonhoeffer S. Nowak M. Hahn B. Saag M. Shaw G. Virial Dynamics in Human Immunodeficiency Virus Type-1 Infection. *Nature*. **1995**, 373, 12, 117.

125. Wiener H. Prediction of Isomeric Difference in Paraffin Properties. *J. Am. Chem. Soc*. **1947**, 69, 17.

126. Wikel H. The Use of Neural Networks for Variable Selection in QSAR. *Bioorg. Med. Chem. Lett*. **1993**, 3, 645.

127. Yaffe D. Cohen Y. Espinosa G. Arenas A. Giralt F. A fuzzy ARTMAP based quantitative structure-property relationships (QSPRs) for

predicting Aqueous Solubility of Organic Compounds, *J. Chem. Inf. Compt. Sci*. **2001***, 41, 5, 1150*

128. Yaffe D. Cohen Y. Espinosa G. Arenas A. and Giralt F. A fuzzy ARTMAP based Quantitative Structure-Property Relationships (QSPRs) for Octanol/Water Partition Coefficients of Organic Compounds. *J. Chem. Inf. Compt. Sci., **2002**, 42, 2, 162*

129. Yalwosky S. and Valvani S. Solubilities and Partitioning 2. Relationship between Aqueous solubilities, Partition Coefficients, and Molecular Surface Areas of Rigid Aromatic Hydrocarbons. *J. Chem. Eng. Data*. **1979**, 24, 2,127.

130. Zhen W. and Tropsha A. Novel Variable Selection Quantitative Structure-Property Relationship Approach Based on the k-Nearest-Neighbor Principle. *J. Chem. Inf. Comput. Sci*. **2000**, 40, 185

i