

ỨNG DỤNG QSPR: SO SÁNH DỰ BÁO HẰNG SỐ BỀN CỦA PHỨC THIOSEMICARBAZONE VỚI ION KIM LOẠI SỬ DỤNG MÔ HÌNH HỒI QUY ĐA BIẾN, BÌNH PHƯƠNG TỐI THIỂU RIÊNG PHẦN VÀ HỒI QUY THÀNH PHẦN CHÍNH CÙNG VỚI CÁC THAM SỐ MÔ TẢ PHÂN TỬ

Nguyễn Minh Quang^{1,3}, Trần Xuân Mậu¹, Phạm Văn Tất^{*}

¹Khoa Hóa học, Trường Đại học khoa học, Đại học Huế

²Khoa Khoa học và Kỹ thuật, Trường Đại học Hoa Sen

³Khoa Kỹ thuật Hóa học, Trường Đại học Công nghiệp Tp. Hồ Chí Minh

*Email: vantat@gmail.com

Ngày nhận bài: 28/8/2018; ngày hoàn thành phản biện: 26/9/2018; ngày duyệt đăng: 10/12/2018

TÓM TẮT

Trong nghiên cứu này, các mô hình liên quan định lượng giữa cấu trúc và tính chất (QSPRs) của các phức chất thiosemicarbazone và ion kim loại được xây dựng dựa trên phương pháp hồi quy đa biến, bình phương tối thiểu riêng phần và hồi quy thành phần chính. Chất lượng các mô hình được đánh giá dựa vào các giá trị hệ số tương quan, sai số chuẩn trung bình và chuẩn Fisher. Kết quả nhận được mô hình QSPR_{MLR} với các giá trị $R^2_{train} = 0,908$; $R^2_{CV} = 0,850$; $Q^2_{test} = 0,8542$; $MSE = 0,852$; mô hình QSPR_{PLS} với $R^2_{train} = 0,908$; $R^2_{CV} = 0,888$; $Q^2_{test} = 0,8972$; $MSE = 0,661$; mô hình QSPR_{PCR} với $R^2_{train} = 0,914$; $R^2_{CV} = 0,948$; $Q^2_{test} = 0,8842$; $MSE = 0,827$. Các mô hình QSPR_{MLR}, QSPR_{PLS} và QSPR_{PCR} có khả năng dự đoán phù hợp với thực nghiệm.

Từ khóa: QSPR_{MLR}, QSPR_{PLS}, QSPR_{PCR}, hằng số bền, thiosemicarbazone.

1. MỞ ĐẦU

Dẫn xuất thiosemicarbazone và phức của nó với các ion kim loại được ứng dụng rộng rãi trong nhiều lĩnh vực. Chúng có nhiều hoạt tính sinh học [1] như kháng khuẩn, kháng nấm, chống ung thư... nên được nghiên cứu nhiều trong lĩnh vực dược học [2]. Do khả năng tạo phức mạnh với nhiều ion kim loại nên chúng có thể được sử dụng trong phân tích trắc quang [3]. Phức của thiosemicarbazone cũng được nghiên cứu ứng dụng trong kỹ thuật xúc tác [4].

Đối với phức chất, hằng số bền là một thông số quan trọng. Từ hằng số bền có thể tính nồng độ cân bằng của các thành phần trong dung dịch. Nó cũng có thể dự đoán sự thay đổi của cấu trúc điện tử phức tạp trong dung dịch từ nồng độ ban đầu

của ion trung tâm và phối tử. Trong những năm gần đây, hằng số bền của phức được nghiên cứu nhiều trong phân tích trắc quang [5].

Trong những năm qua, sự phát triển mạnh mẽ của khoa học máy tính đã cho ra đời các công cụ tính toán lượng tử như Hyperchem, Mopac, Gaussian [6] cùng với các phần mềm mô phỏng trong hóa học được sử dụng rộng rãi trong nghiên cứu lý thuyết như QSARIS, Spartan, MOE, Materials Studio, Dragon [6] nhằm xây dựng các mô hình dự đoán tính chất của các hợp chất hóa học dựa trên mối quan hệ định lượng cấu trúc – tính chất (QSPR) [6,7]. Trong các công trình này, các tác giả đã phát triển các mô hình QSPR trên các đối tượng khác nhau với các tính chất khác nhau như nhiệt độ sôi [6], độ hòa tan [6], tính kỵ nước (logP) [6], hệ số phân tán nước-hợp chất hữu cơ [6], tính axit của các hợp chất chứa nhóm xeton [8], chỉ số thời gian lưu của pha đảo trong phân tích sắc ký lỏng của các hợp chất hydrocacbon thơm đa vòng [9]. Mặc dù, việc nghiên cứu phát triển mô hình QSPR trên các hợp chất khác nhau với các tính chất khác nhau nhưng điểm chung của các nghiên cứu này là sử dụng các phương pháp hồi quy đa biến và mạng thần kinh nhân tạo để phát triển mô hình [7]. Tuy nhiên, trong các công trình đã được công bố chúng tôi nhận thấy rằng chưa có một công bố nào phát triển mô hình QSPR trên đối tượng phức chất giữa các dẫn xuất thiosemicarbazone với các ion kim loại với đại lượng đặc trưng là hằng số bền của phức.

Trong nghiên cứu này, chúng tôi định hướng thiết kế các hợp chất thiosemicarbazone làm ligand sử dụng trong kỹ thuật phân tích các ion kim loại trong các mẫu thực phẩm và môi trường. Các kỹ thuật đa biến được sử dụng để xây dựng mô hình biểu diễn mối quan hệ định lượng giữa cấu trúc và hằng số bền của các phức thiosemicarbazone với các ion kim loại. Kết quả từ các mô hình QSPR cũng được so sánh với thực nghiệm.

2. PHƯƠNG PHÁP TÍNH TOÁN

2.1. Phản ứng tạo phức

Phản ứng tạo phức giữa ion kim loại (M) với thiosemicarbazone (L) xảy ra như sau [10]



Hằng số bền của phức $M_p L_q$ là hằng số cân bằng (β_{pq}) của phản ứng được xác định theo công thức (2)

$$\beta_{pq} = \frac{[M_p L_q]}{[M]^p [L]^q} \quad (2)$$

Trong trường hợp phản ứng tạo phức chỉ xảy ra một nấc tức là khi $p = 1$ và $q = 1$. Cấu trúc của phức trong nghiên cứu này được mô tả như Hình 1.

2.2. Hồi quy tuyến tính bội

Mục đích của hồi quy tuyến tính bội (Multivariate Linear Regression – MLR) là xây dựng mô hình liên quan giữa hai hay nhiều biến độc lập và một biến phụ thuộc bằng cách xây dựng một phương trình tuyến tính cho các số liệu quan sát [11]. Mỗi giá trị biến độc lập x được liên hệ với một giá trị biến phụ thuộc Y . Mô hình hồi quy bội MLR được biểu diễn ở công thức (3) [11-13]

$$y = b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_mx_m + \varepsilon \quad (3)$$

trong đó m là số biến độc lập; $b_1, b_2, \dots, b_m$ các hệ số hồi quy và y biến phụ thuộc; ε là sai số. Các hệ số hồi quy đặc trưng cho sự đóng góp độc lập của mỗi tham số mô tả phân tử. Mô hình MLR được xác định bằng phương trình ma trận [11-13]

$$y = Xb + e \quad (4)$$

Khi X là dãy đầy đủ thì lời giải cực tiểu hóa là [11-13]

$$\hat{b} = (X^T X)^{-1} X^T y \quad (5)$$

trong đó $\hat{b}$ là giá trị ước lượng cho hệ số hồi quy. Mô hình MLR được xây dựng từ một tập luyện, tập đánh giá và dự đoán ngoại.

2.3. Bình phương tối thiểu riêng phần

Bình phương tối thiểu riêng phần (Partial Least Square – PLS) được sử dụng để xây dựng các mô hình hồi quy dựa vào sự phân tích biến ẩn liên quan đến hai khối ma trận X và Y , bao gồm các biến độc lập x và biến phụ thuộc y . Các ma trận được phân chia thành tổng của các biến ẩn f , như [14-17]

$$X = TP^T + E = \sum t_f p_f' + E \quad (6)$$

$$Y = UQ^T + F = \sum u_f q_f' + F \quad (7)$$

trong đó T và U là các ma trận đặc trưng cho X và Y tương ứng; P và Q là ma trận hệ số tương ứng; E và F là các ma trận sai số; hai ma trận X và Y tương quan theo T và U đặc trưng đối với mỗi biến ẩn [14-17]

$$u_f = b_f t_f \quad (8)$$

trong đó b_f là hệ số hồi quy cho biến ẩn f . Ma trận Y có thể được tính từ u_f , hằng số bên của các phức chất mới có thể được tính toán từ các T mới thế vào phương trình (9) dẫn đến phương trình (10) [14-17]

$$Y = TBQ^T + F \quad (9)$$

$$Y_{\text{moi}} = UBQT \quad (10)$$

Trong phần tính toán này, cần tìm số biến ẩn tốt nhất mà nó được thực hiện chuẩn hóa bằng kỹ thuật đánh giá chéo dựa vào sai số dự đoán cực tiểu. Mô hình PLS được thảo luận ở nhiều công trình.

2.4. Hồi quy thành phần chính

Từ một tập dữ liệu $\{X, y\}$, trong đó X là một ma trận với n quan sát và p biến số; y là vector biến phụ thuộc tương ứng. Các số liệu được tập trung và không được xử lý trước, hồi quy tuyến tính bộ chuẩn MLR được dựa vào ma trận [18-22]

$$Y = Xb + \varepsilon \quad (11)$$

trong đó b là các hệ số và ε là vector sai số. Đặc trưng chính của hồi quy thành phần chính (Principal Component Regression – PCR) là giá trị đáp ứng y không tương quan trực tiếp với X nhưng với thành phần chính của nó. Các thành phần chính nhận được bằng cách phân tách X bằng phương pháp phân tích thành phần chính (Principal Component Analysis - PCA) [18-22]

$$X = \sum t_i p_i = TP' \quad (12)$$

trong đó $T = [t_1, t_2, \dots, t_q]$, $P = [p_1, p_2, \dots, p_q]$, t_i là thành phần chính thứ i ; p_i là vector riêng thứ i của $X'X$; q là hàng của X và $T = XP$.

Trong hồi quy thành phần chính, nếu k thành phần đầu ($k < q$) được sử dụng cho hồi quy, thì phương trình hồi quy được viết như sau [18-22]

$$Y = T_1 P_1' b + e = T_1 \alpha_1 + e \quad (13)$$

trong đó $T_1 = [t_1, t_2, \dots, t_k]$, $P_1 = [p_1, p_2, \dots, p_k]$, $\alpha_1 = P_1' b$ và e là vector sai số.

Thực hiện bình phương tối thiểu phương trình (13), nhận được phương trình [18-22]

$$\hat{\alpha}_1 = (T_1' T_1)^{-1} T_1' y \quad (14)$$

Đánh giá phương trình hồi quy thành phần chính PCR của b ($\hat{b}_p$), loại bỏ $(q-k)$ thành phần sau đó, có thể nhận được phương trình (15) [18-22]

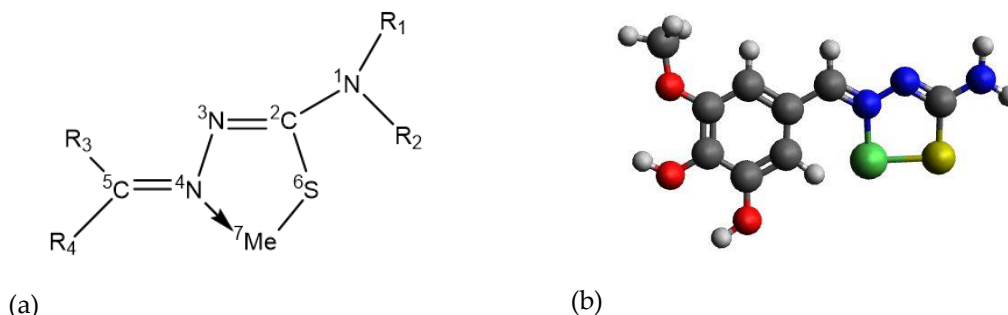
$$\hat{b}_p = P_1 (T_1' T_1)^{-1} T_1' y \quad (15)$$

Các giá trị đáp ứng dự đoán dựa vào X bằng $\hat{b}_p$ được cho bằng ma trận (16) [18-22]

$$\hat{y} = X \hat{b}_p \quad (16)$$

2.5. Dữ liệu

Cấu trúc các phức chất và các giá trị hằng số bền $\log\beta_{11}$ thực nghiệm thu nhận từ các công trình đã được công bố như Hình 1 [1] và Bảng 1.



Hình 1. Cấu trúc của phức giữa thiosemicarbazone và ion kim loại;

a) Cấu trúc tổng quát; b) Phức Ni^{2+} and 3,4-dihydroxy-5-methoxy benzaldehyde thiosemicarbazone [23]

Từ các phức chất thực nghiệm, tiến hành xây dựng lại cấu trúc bằng BIOVA Draw 2017 R2 [24]. Sau đó, các cấu trúc được tối ưu hóa và tính toán tham số lượng tử theo phương pháp bán thực nghiệm SCF PM7 của MoPac2016 [25, 26]. Tham số 2D, 3D được tính bằng QSARIS [27, 28]. Các dữ liệu sau khi tính toán được sử dụng để xây dựng mô hình đa biến QSPR cấu trúc – hằng số bền ($\log\beta_{11}$) bằng MLR, PCR, PLS và phân tích thống kê khác bằng XLSTAT2016 [29], Regression [11] và MS-EXCEL [12, 13, 27]. Các mô hình đa biến được biểu diễn bằng ba mô hình tương ứng QSPR_{MLR} , QSPR_{PLS} và QSPR_{PCR} .

Bảng 1. Giá trị thực nghiệm $\log\beta_{11}$ và cấu trúc các phức chất nghiên cứu

STT	Thiosemicarbazone				Ion kim loại	$\log\beta_{11}$	Tài liệu tham khảo
	R ₁	R ₂	R ₃	R ₄			
1	H	H	H	$-\text{C}_6\text{H}_2(\text{OH})_2\text{OCH}_3$	Co(II)	6,382	[30]
2	H	H	H	$-\text{C}_6\text{H}_3\text{OHCH}_3$	Cu(II)	13,330	[31]
3	H	H	H	$-\text{C}_6\text{H}_3\text{OHCH}_3$	Ni(II)	12,620	[31]
4	H	H	H	$-\text{C}_6\text{H}_3\text{OHCH}_3$	Co(II)	11,097	[31]
5	H	H	H	$-\text{C}_6\text{H}_3\text{OHCH}_3$	Mn(II)	10,550	[31]
6	H	H	H	$-\text{C}_6\text{H}_3\text{OHCH}_3$	Pb(II)	6,830	[32]
7	H	H	H	$-\text{C}_6\text{H}_3\text{OHCH}_3$	Cd(II)	7,070	[32]
8	H	H	H	$-\text{C}_6\text{H}_3\text{OHCH}_3$	Zn(II)	7,420	[32]
9	H	H	H	$-\text{C}_6\text{H}_3\text{OHCH}_3$	Fe(II)	7,990	[32]
10	H	CH ₃	CH ₃	$-\text{C}_5\text{H}_4\text{N}$	Cu(II)	6,114	[33]
11	H	H	CH ₃	$=\text{N}-\text{NH}-\text{C}_6\text{H}_5$	Cu(II)	11,700	[34,35]
12	H	H	H	$-\text{C}_6\text{H}_3\text{OHCH}_3$	Cr(VI)	4,842	[36]
13	H	H	H	$-\text{C}_6\text{H}_4\text{N}(\text{CH}_3)_2$	Ag(I)	17,200	[37]
14	H	H	H	$-\text{C}_6\text{H}_4\text{N}(\text{CH}_3)_2$	Cu(II)	15,300	[38]
15	H	H	CH ₃	$-\text{C}_6\text{H}_4\text{OH}$	Mn(II)	4,510	[39]
16	H	H	CH ₃	$-\text{C}_6\text{H}_4\text{OH}$	Ni(II)	5,310	[39]

17	H	H	CH ₃	-C ₆ H ₄ OH	Cu(II)	5,910	[39]
18	H	H	-	-C ₉ H ₈ NO	Cu(II)	8,714	[40]

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Chọn tập luyện và tập kiểm tra

Tập dữ liệu luyện thể hiện một vai trò quan trọng trong việc phát triển các tính chất mô hình như độ đúng và khả năng phù hợp của mô hình trong thực tế dự đoán. Việc chọn dữ liệu luyện, dữ liệu đánh giá và kiểm tra ngoại một cách ngẫu nhiên là một trong trong các giai đoạn quan trọng nhất. Sau khi tiến hành sàng lọc phân loại các dữ liệu và loại bỏ dữ liệu bất thường, kết quả nhận được là một bộ dữ liệu gồm 105 biến với 62 quan sát thực nghiệm để chuẩn bị cho quá trình xây dựng mô hình.

3.2. Xây dựng các mô hình QSPR

Quá trình xây dựng và đánh giá mô các mô hình QSPR từ tập dữ liệu mẫu gồm 62 quan sát được chia ngẫu nhiên thành nhóm luyện 80 %, nhóm đánh giá 20 % và nhóm đánh giá ngoại gồm 10 hợp chất không thuộc nhóm 62 quan sát. Các mô hình QSPR được xây dựng từ nhóm luyện để dự đoán giá trị hằng số bền của các phức chất trong nhóm đánh giá và nhóm kiểm tra. Chất lượng các mô hình QSPR_{MLR}, QSPR_{PCR} và QSPR_{PLS} thể hiện ở hệ số tương quan R^2_{train} , R^2_{cv} và Q^2_{test} cũng như các giá trị thống kê MSE và F_{stat} . Các biến số độc lập X được chọn đưa vào mô hình dựa vào hệ số tương quan nội và sự thay đổi của các giá trị thống kê MSE, R^2_{train} , R^2_{cv} , F_{stat} khi sử dụng kỹ thuật loại dần hoặc nhập dần biến số vào mô hình. Kết quả được dẫn ra ở Bảng 2 và Bảng 3.

Bảng 2. Các mô hình QSPR_{MLR} số biến k từ 4 đến 11, và các giá trị thống kê tương ứng

Số biến k	Biến số trong mô hình QSPR _{MLR}	MSE	R^2_{train}	R^2_{adj}	R^2_{cv}	F_{stat}
4	$x_1/x_2/x_3/x_4$	1,610	0,639	0,613	0,557	25,1928
5	$x_1/x_2/x_3/x_4/x_5$	1,549	0,672	0,642	0,552	22,8981
6	$x_1/x_2/x_3/x_4/x_5/x_6$	1,341	0,758	0,732	0,636	28,7568
7	$x_1/x_2/x_3/x_4/x_5/x_6/x_7$	1,196	0,811	0,787	0,696	33,1253
8	$x_1/x_2/x_3/x_4/x_5/x_6/x_7/x_8$	1,147	0,829	0,804	0,715	32,2235
9	$x_1/x_2/x_3/x_4/x_5/x_6/x_7/x_8/x_9$	0,852	0,908	0,892	0,850	56,8949
10	$x_1/x_2/x_3/x_4/x_5/x_6/x_7/x_8/x_9/x_{10}$	0,836	0,913	0,896	0,855	53,4399
11	$x_1/x_2/x_3/x_4/x_5/x_6/x_7/x_8/x_9/x_{10}/x_{11}$	0,839	0,914	0,895	0,853	48,2420
Ký hiệu các biến số						
Năng lượng tổng	x_1	$\log P$	x_5	Năng lượng elctron	x_9	
Knotp	x_2	Diện tích Me ⁷	x_6	Diện tích Cosmo	x_{10}	
Thể tích Cosmo	x_3	ΔH_f	x_7	Thế ion hóa	x_{11}	
pH	x_4	LUMO	x_8			

Các biến số chọn lựa đưa vào các mô hình QSPR_{MLR} (Bảng 2) đã cho thấy các giá trị R^2_{train} , Q^2_{test} và F_{stat} thay đổi và tăng theo số biến k. Khi các giá k tăng từ 9 đến 11, thì

các giá trị thống kê tương ứng tăng thêm không đáng kể và còn có xu hướng giảm xuống như giá trị F_{stat} . Như vậy, chọn số biến $k = 9$ là phù hợp cho xu hướng thay đổi tối ưu này. Các biến số từ x_1 đến x_9 được kiểm tra mối tương quan nội giữa hai hay nhiều biến dựa vào ma trận hệ số tương quan Pearson, xác định mối tương quan có ý nghĩa đối với $\log_{11}$. Ma trận tương quan được đưa ra ở Bảng 3.

Từ Bảng 3, hệ số tương quan của 9 biến số độc lập và 1 biến phụ thuộc $\log_{11}$ cho thấy các biến được chọn vào mô hình $QSPR_{MLR}$ với $k = 9$ là phù hợp và chấp nhận về mặt thống kê tương quan và kiểm định student (t -test) đặc trưng cho các biến số.

Từ Bảng 2, mô hình $QSPR_{MLR}$ với số biến $k = 9$ được xây dựng lại và nhận được các hệ số tương ứng cùng các giá trị thống kê mô hình $QSPR_{MLR}$ như sau:

$$\log\beta_{11} = 8,402 + 0,0195x_1 + 13,690x_2 - 0,066x_3 + 0,885x_4 + 3,871x_5 - 3,184x_6 - 0,050x_7 + 2,961x_8 - 0,005x_9 \quad (17)$$

$$\text{với } n = 62; R^2_{train} = 0,908; R^2_{CV} = 0,850; MSE = 0,852$$

Như vậy, tập dữ liệu luyện dùng để xây dựng mô hình $QSPR_{MLR}$ đạt yêu cầu về thống kê, có khả năng dự đoán tốt. Khả năng dự đoán của mô hình $QSPR_{MLR}$ phù hợp tốt đối với nhóm các phức chất. Các tham số lựa chọn trong mô hình không có sự tương quan nào giữa các biến đã chọn. Dữ liệu xây dựng mô hình này sẽ được sử dụng để xây dựng các mô hình $QSPR_{PCR}$ và $QSPR_{PLS}$.

Bảng 3. Ma trận tương quan Pearson của các biến số trong mô hình $QSPR_{MLR}$ với $k = 9$

Biến số	$\log_{11}$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
$\log_{11}$	1	0,237	-0,222	-0,638	0,423	-0,246	-0,289	-0,258	0,982	1
x_1	0,237	1	0,305	-0,421	-0,226	0,078	0,286	0,236	0,283	0,237
x_2	-0,222	0,305	1	0,294	-0,681	-0,339	0,328	0,083	-0,245	-0,222
x_3	-0,638	-0,421	0,294	1	-0,526	-0,163	0,496	0,189	-0,678	-0,638
x_4	0,423	-0,226	-0,681	-0,526	1	0,517	-0,565	-0,128	0,473	0,423
x_5	-0,246	0,078	-0,339	-0,163	0,517	1	-0,307	0,232	-0,135	-0,246
x_6	-0,289	0,286	0,328	0,496	-0,565	-0,307	1	0,257	-0,332	-0,289
x_7	-0,258	0,236	0,083	0,189	-0,128	0,232	0,257	1	-0,130	-0,258
x_8	0,982	0,283	-0,245	-0,678	0,473	-0,135	-0,332	-0,130	1	0,982
x_9	1	0,237	-0,222	-0,638	0,423	-0,246	-0,289	-0,258	0,982	1

Sử dụng ma trận dữ liệu với số biến độc lập $k = 9$ và biến phụ thuộc $\log_{11}$, thực hiện xây dựng mô hình $QSPR_{PCR}$. Kết quả phân tích thành phần chính PCA cho thấy 9 thành phần chính có ý nghĩa thống kê. Mô hình $QSPR_{PCR}$ được biểu diễn như sau:

$$\log\beta_{11} = 6,209 + 0,0214x_1 + 13,513x_2 - 0,065x_3 + 0,786x_4 + 3,867x_5 - 3,100x_6 - 0,052x_7 + 3,307x_8 - 0,006x_9 \quad (18)$$

$$\text{với } n = 62; R^2_{train} = 0,914; R^2_{CV} = 0,948; MSE = 0,827$$

Tương tự, từ kết quả xây dựng mô hình QSPR_{MLR}, tiến hành xây dựng mô hình QSPR_{PLS} dựa trên ma trận dữ liệu với 9 biến độc lập. Chất lượng mô hình QSPR_{PLS} được đánh giá dựa vào các chỉ số thống kê với các giá trị thống kê tích lũy $Q^2_{cum} = 0,147$; $R^2_{Ycum} = 0,858$ và $R^2_{Xcum} = 0,916$. Ngoài ra, đại lượng mức độ quan trọng của các biến số X (Variable Importance for the Projection – VIP) ảnh hưởng đến $\log\beta_{11}$ được sử dụng để lựa chọn biến trong mô hình QSPR_{PLS}. Theo đề xuất của Word [17] và Ericksson [20], các biến số trong mô hình được lựa chọn phải có giá trị VIP lớn hơn 0,8. Trên cơ sở đó, mô hình QSPR_{PLS} có dạng như sau:

$$\log\beta_{11} = 6,102 + 0,023x_1 + 13,467x_2 - 0,062x_3 + 0,802x_4 + 3,884x_5 - 2,984x_6 - 0,049x_7 + 3,266x_8 - 0,006x_9 \quad (19)$$

$$\text{với } n = 62; R^2_{train} = 0,908; R^2_{CV} = 0,888; MSE = 0,661$$

Trong các mô hình QSPR, giá trị R^2_{train} là hệ số tương quan bội được nhân với 100 cho phương sai giải thích hằng số bền $\log\beta_{11}$. Khả năng dự đoán của các mô hình QSPR được đánh giá bằng R^2_{CV} và Q^2_{test} . Giá trị thống kê F_{stat} phản ánh tỷ lệ phương sai giải thích bởi mô hình và phương sai từ sai số hồi quy. Giá trị F_{stat} cao cho thấy mô hình có ý nghĩa về mặt thống kê. Giá trị MSE thấp cũng cho thấy mô hình có ý nghĩa về mặt thống kê. Khả năng dự báo của mô hình được thể hiện qua giá trị Q^2_{test} đánh giá ngoại đối với nhóm hợp chất không thuộc nhóm luyện ban đầu.

3.3. Đánh giá khả năng dự đoán $\log\beta_{11}$

Khả năng dự đoán của các mô hình QSPR đều được đánh giá cẩn thận bằng kỹ thuật đánh giá chéo và đánh giá ngoại; đánh giá khả năng dự đoán của các mô hình QSPR ở trên đối với 10 hợp chất chọn ngẫu nhiên từ các kết quả nghiên cứu thực nghiệm được dẫn ra ở Bảng 4.

Bảng 4. Hằng số bền $\log\beta_{11}$ của 10 phức chất chọn đánh giá dự đoán ngoại các mô hình QSPR

Ligand				Ion	$\log\beta_{11,exp}$	QSPR _{MLR}		QSPR _{PLS}		QSPR _{PCR}	
R ₁	R ₂	R ₃	R ₄			$\log\beta_{11,cal}$	ARE, %	$\log\beta_{11,cal}$	ARE, %	$\log\beta_{11,cal}$	ARE, %
H	H	CH ₃	– C ₇ H ₇ N ₂	Ni(II)	10,940 [35]	12,402	13,366	12,264	12,101	11,894	8,724
H	H	CH ₃	– C ₇ H ₇ N ₂	Mn(II)	9,720 [35]	11,584	19,172	11,780	21,195	11,236	15,594
H	H	CH ₃	– C ₇ H ₇ N ₂	Ni(II)	10,790 [35]	12,402	14,942	12,264	13,659	11,894	10,236
H	H	CH ₃	– C ₇ H ₇ N ₂	Co(II)	9,900 [35]	13,374	35,087	13,823	39,628	13,322	34,561
H	H	CH ₃	– C ₇ H ₇ N ₂	Mn(II)	9,600 [35]	11,584	20,662	11,780	22,710	11,236	17,039
H	H	–	– C ₉ H ₈ NO	Zn(II)	8,160 [40]	8,409	3,055	8,644	5,929	8,317	1,921

H	H	-	$\text{C}_9\text{H}_8\text{NO}$	Cd(II)	6,611 [40]	6,048	8,520	6,010	9,085	5,595	15,370
H	H	-	$\text{C}_9\text{H}_8\text{NO}$	Mn(II)	6,230 [40]	6,450	3,527	6,622	6,287	6,339	1,746
H	H	CH ₃	-C ₅ H ₄ N	Cu(II)	5,491 [41]	7,754	41,216	6,621	20,574	6,688	21,806
H	H	CH ₃	-C ₅ H ₄ N	Cu(II)	5,924 [41]	7,324	23,627	5,979	0,924	6,136	3,581
MARE, %:						18,317		15,209		13,058	

Con đường tốt nhất để đánh giá chất lượng mô hình là thực hiện đánh giá nội. Giá trị thống kê đặc trưng cho đánh giá nội là [11-22]

$$R_{cv}^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (20)$$

trong đó $\hat{y}_i$, y_i và $\bar{y}$ là các giá trị hằng số bền log₁₁ dự đoán, thực nghiệm và trung bình.

Khả năng dự đoán của mô hình được đánh giá tương tự bằng hệ số tương quan Q_{test}^2 [11-22]

$$Q_{test}^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_{i,test} - y_{i,test})^2}{\sum_{i=1}^n (y_{i,test} - \bar{y})^2} \quad (21)$$

trong đó $\hat{y}_{i,test}$, $y_{i,test}$ và $\bar{y}$ là các giá trị hằng số bền log₁₁ dự đoán, thực nghiệm trong nhóm kiểm tra ngoại và trung bình của nhóm kiểm tra.

Kết quả dự đoán của các mô hình QSPR được đánh giá bằng giá trị tuyệt đối của các sai số tương đối ARE (absolute value of relative error), % tính theo công thức (22)

$$ARE, \% = \frac{|y_{i,exp} - y_{i,cal}|}{y_{i,exp}} 100 \quad (22)$$

Giá trị trung bình tuyệt đối của các sai số tương đối MARE (mean absolute values of relative error), % được sử dụng để đánh giá tổng quát sai số của các mô hình QSPR tính theo công thức (23)

$$MARE, \% = \frac{\sum_{i=1}^n ARE_i, \%}{n} \quad (23)$$

trong đó $n = 10$ là số hợp chất, log_{11,exp} là giá trị hằng số bền thực nghiệm, log_{11,cal} là giá trị hằng số bền dự đoán của các phức chất trong nhóm kiểm tra ngoại.

Như vậy, kết quả đánh giá giữa ba mô hình QSPR_{MLR}, QSPR_{PLS} và QSPR_{PCR} dựa vào các giá trị $MARE$ (%) (Bảng 4) cho thấy mô hình QSPR_{MLR} có khả năng dự đoán kém nhất, sau đó là mô hình QSPR_{PLS} và cuối cùng là mô hình QSPR_{PCR} tương ứng với các giá trị 18,317 %, 15,209 % và 13,058 %. Đồng thời, các giá trị Q^2_{test} của ba mô hình QSPR lần lượt là $Q^2_{test,MLR} = 0,8542$; $Q^2_{test,PLS} = 0,8972$ và $Q^2_{test,PCR} = 0,8842$ cho thấy kết quả dự đoán $\log\beta_{11}$ nhận được từ ba mô hình rất gần với thực nghiệm và các mô hình này có khả năng ứng dụng trong thực tiễn.

Sự khác biệt giữa các giá trị $\log\beta_{11}$ thực nghiệm và $\log\beta_{11}$ dự đoán từ ba mô hình trên được đánh giá bằng phương pháp ANOVA một yếu tố. Sự chênh lệch giữa các giá trị thực nghiệm và tính toán các hằng số bền $\log\beta_{11}$ ở cả ba mô hình là không đáng kể ($F = 0,0655 < F_{0,05} = 3,354$). Vì vậy, có thể khẳng định rằng khả năng dự báo của cả ba mô hình QSPR phù hợp với dữ liệu thực nghiệm.

4. KẾT LUẬN

Công trình này đã xây dựng thành công quan hệ định lượng cấu trúc – hằng số bền $\log\beta_{11}$ (QSPRs) sử dụng các phương pháp hồi quy tuyến tính bội (QSPR_{MLR}), bình phương tối thiểu riêng phần (QSPR_{PLS}) và hồi quy thành phần chính (QSPR_{PCR}). Bộ dữ liệu xây dựng các mô hình đã được tạo ra thành công từ các tính toán lượng tử bán thực nghiệm và cơ học phân tử kết hợp với các tham số thực nghiệm. Các mô hình đều được đánh giá nội và đánh giá ngoại thành công bằng các giá trị thống kê R^2_{CV} , Q^2_{test} , và $MARE$, % và phương pháp ANOVA. Các mô hình QSPR_{MLR}, QSPR_{PLS} và QSPR_{PCR} đều đáp ứng yêu cầu dự đoán thực tế. Mô hình QSPR_{PCR} cho kết quả dự đoán tốt nhất.

Kết quả nhận được từ công trình này cho phép dự đoán và định hướng thực nghiệm tổng hợp các dẫn xuất thiosemicarbazone có khả năng tạo phức tốt với các ion kim loại mở ra hướng nghiên cứu mới và có nhiều hứa hẹn trong lĩnh vực phân tích môi trường và kỹ thuật xúc tác.

TÀI LIỆU THAM KHẢO

- [1]. R. B. Singh, B. S. Garg, and R. P. Singh (1978). Analytical applications of thiosemicarbazones and semicarbazones: A review, *Talanta*, Vol. 25, (11–12), pp. 619–632.
- [2]. B. H. Patel, J. R. Shah, and R. P. Patel (1976). Stability constants of complexes of 2-hydroxy-5-methylacetophenone-thiosemicarbazone with Cu(II), Ni(II), Co(II), Zn(II) and Mn(II), *J. Ind. Chem. Soc.*, Vol. 53, pp. 9–10.
- [3]. E. B. Seenaa, R. Bessy, M. R. Prathapachandra Kurup, and I. E. Suresh (2006). A crystallographic study of 2-hydroxyacetophenone *N* (4) cyclohexyl thiosemicarbazone, *J. Chem. Crystallogr.*, Vol. 36, pp. 189.

- [4]. Ezhilarasi et al. (2012). Synthesis Characterization and Application of Salicylaldehyde Thiosemicarbazone and Its Metal Complexes, *Int. J. Res. Chem. Environ.*, Vol. 2, 4, pp. 130–148.
- [5]. G. Pelosi (2010). Thiosemicarbazone Metal Complexes: From Structure to Activity, *J. Open Crystallogr.*, Vol. 3, pp. 16–28.
- [6]. B. Chen, T. Zhang, T. Bond and Y. Gan (2015). Development of quantitative structure activity relationship (QSAR) model for disinfection byproduct (DBP) research: A review of methods and resources, *Journal of Hazardous Materials*, Vol 299, pp. 260–279.
- [7]. S. Yousefinejad and B. Hemmateenejad (2015). Chemometrics tools in QSAR/QSPR studies: A historical perspective, *Chemometrics and Intelligent Laboratory Systems*, Vol 149, pp. 177–204
- [8]. Y. Yuan, P. D. Mosier and Y. Zhang (2012). Quantitative structure-property relationship (QSPR) model for predicting acidities of ketones, *Journal of Biophysical Chemistry*, Vol.3, 1, pp. 49–57.
- [9]. F. A. Ribeiro, M. M. Ferreira (2003). QSPR models of boiling point, octanol–water partition coefficient and retention time index of polycyclic aromatic hydrocarbons, *Journal of Molecular Structure (Theochem)*, Vol. 663, 109–126.
- [10]. D. Harvey (2000). “Equilibrium Chemistry”, *Modern analytical Chemistry, the second ed.*, Mc.Graw-Hill, pp. 144.
- [11]. D. D. Steppan, J. Werner, and P. R. Yeater (1998). *Essential Regression and Experimental Design for Chemists and Engineers*. <http://www.jowerner.homepage.tonline.de/index.html>.
- [12]. E. J. Billo (2007). *Excel For Scientists And Engineers: Numerical Methods*, John Wiley and Sons, Inc., Hoboken, New Jersey, USA.
- [13]. E. J. Billo (1997). *Excel for chemists*, Wiley-VCH, Weinheim.
- [14]. M. Tenenhaus (1998). *La Régression PLS, Théorie et Pratique. Technip*, Paris.
- [15]. M. Tenenhaus, J. Pagès, L. Ambroisine, and C. Guinot (2005). PLS methodology for studying relationships between hedonic judgements and product characteristics, *Food Quality and Preference*, Vol. 16, 4, pp. 315–325.
- [16]. S. Wold, H. Martens, and H. Wold (1983). *The Multivariate Calibration Problem in Chemistry solved by the PLS Method*, Springer Verlag, Heidelberg, pp. 286 – 293.
- [17]. S. Wold (1995). “PLS for multivariate linear modelling”, *QSAR: Chemometric Methods in Molecular Design*, Vol 2, Wiley-VCH Weinheim, Germany, pp. 195–218.
- [18]. T. Amemiya (1980). Selection of regressors, *Int. Eco. Rev.*, Vol. 21, pp. 331–354.
- [19]. A. P. Dempster (1969). *Elements of Continuous Multivariate Analysis*, Addison-Wesley, Reading, MA.
- [20]. L. Eriksson, E. Johansson, N. Kettaneh-Wold, and S. Kettaneh-Wold (2001). Multi- and Megavariate Data Analysis: Principles and Applications, *Journal of Chemometrics*, Vol. 16, 5, 261–262.
- [21]. S. Kotz, and N. L. Johnson (1992). “Breakthroughs in Statistics”, Vol. 1. *Foundations and Basic Theory*, New York: Springer, pp. 610–624.

- [22]. G. Schwarz (1978). Estimating the dimension of a model, *Annals of Statistics*, Vol. 6, pp. 461–464.
- [23]. M. Hymavathi, C. Viswanatha, and N. Devanna (2014). A Study on Synthesis of Novel Chromogenic Organic Reagent 3,4-dihydroxy-5-methoxy benzaldehyde thiosemicarbazone and Spectrophotometric Determination of Nickel (II) in Presences of Triton X-100, *Res. J. Pharma., Bio. and Chem. Sci.*, Vol. 5, 5, 625–630.
- [24]. BIOVA Draw 2017 R2, version: 17.2.NET (2016). Dassault Systèmes, France.
- [25]. James. J. P. Stewart (2016). MOPAC2016, version: 17.240W, Stewart Computational Chemistry, USA.
- [26]. James J. P. Stewart (2013). Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters, *J. Mol. Model.*, Vol. 19, 1–32.
- [27]. Pham Van Tat (2009). *Development of QSAR and QSPR*, Publisher of Natural sciences and Technique, Ha Noi.
- [28]. QSARIS 1.1. (2001). Statistical Solutions Ltd., USA.
- [29]. XLSTAT2016, version 2016.02.28451 (2016). Addinsoft, USA.
- [30]. M. Hymavathi, N. Devanna, and C. Viswanatha (2014). A study on synthesis of novel chromogenic organic reagent 3,4-dihydroxy-5-methoxy benzaldehyde thiosemicarbazone and spectrophotometric determination of Cobalt (II) in presences of Triton X-100, *J. Chem. Pharm. Res.*, Vol. 6, 7, pp. 2787–2791.
- [31]. D. K. Singh, P. K. Jha, R. K Jha, P. M. Mishra, A. K. Jha, S. K. Jha, and R. P. Bharti (2009). Equilibrium Studies of Transition Metal Complexes with Tridentate Ligands Containing N, O, S as Donor Atoms, *Asian Journal of Chemistry*, Vol. 21, 7, pp. 5055–5060.
- [32]. B. S. Garg, and V. K. Jain (1989). Determination of thermodynamic parameters and stability constants of complexes of biologically active o-vanillinthiosemicarbazone with bivalent metal ions, *Thermochimica Acta*, Vol. 146, pp. 375–379.
- [33]. K. V. Reddy, D. N. Reddy, S. V. Babu and K. H. Reddy (2011). Spectrophotometric determination of copper (II) in Biological samples by using 2-acetylpyridine 4-methyl-3-thiosemicarbazone (APMT), *Der Pharmacia Sinica*, Vol. 2, 4, 176–183.
- [34]. M. Aljahdali, and A. A. EL-Sherif (2013). Synthesis, characterization, molecular modeling and biological activity of mixed ligand complexes of Cu(II), Ni(II) and Co(II) based on 1,10-phenanthroline and novel thiosemicarbazone, *Inorganica Chimica Acta.*, Vol. 407, pp. 58–68.
- [35]. A. T. A. El-Karim, and A. Ahmed, El-Sherif (2016). Potentiometric, equilibrium studies and thermodynamics of novel thiosemicarbazones and their bivalent transition metal(II) complexes, *J. Mol Liq.*, Vol. 219, 914–922.
- [36]. I. Sreevania, P. Raveendra Reddy, and V. Krishna Reddy (2013). A Rapid and Simple Spectrophotometric Determination of Traces of Chromium (VI) in Waste Water Samples and in Soil samples by using 2-Hydroxy, 3-Methoxy Benzaldehyde Thiosemicarbazone (HMBATSC), *J. Applied Physics.*, Vol. 3, 1, 40–45.
- [37]. M. A. Jiménez, M. D. Luque De Castro, and M. Valcárcel (1980). Potentiometric Study of Silver(I)-Thiosemicarbazones, *J. Microchemical.*, Vol. 25, pp. 301–308.

- [38]. T. Atalay, and E. Ozkan (1994). Thermodynamic studies of some complexes of 4'-morpholino-acetophenone thiosemicarbazone, *Thermochimica Acta.*, Vol. 237, pp. 369–374.
- [39]. B. S. Garg, S. Ghosh, V. K. Jain, and P. K. Singh (1990). Evaluation of thermodynamic parameters of bivalent metal complexes of 2-hydroxyacetophenone thiosemicarbazone (2-HATS), *Thermochimica Acta.*, Vol. 157, pp. 365–368.
- [40]. K. Sarkar, and B. S. Garg (1987). Determination of thermodynamic parameters and stability constants of the complexes of p-MITSC with transition metal ions, *Thermochimica Acta.*, Vol. 113, 7–14.
- [41]. D. Admasu, D. N. Reddy, and K. N. Mekonnen (2016). Spectrophotometric determination of Cu(II) in soil and vegetable samples collected from Abraha Atsbeha, Tigray, Ethiopia using heterocyclic thiosemicarbazone, *SpringerPlus*, Vol. 5, 1169.

**APPLICATION OF QSPR: COMPARISON OF PREDICTION OF STABILITY
CONSTANTS OF THIOSEMICARBAZONE COMPLEXES WITH METAL IONS
USING MULTIVARIATE LINEAR REGRESSION, PARTIAL LEAST SQUARE,
AND PRINCIPAL COMPONENT REGRESSION MODELS WITH MOLECULAR
DESCRIPTIVE PARAMETERS**

Nguyen Minh Quang^{1,3}, Tran Xuan Mau¹, Pham Van Tat^{2*}

¹ Faculty of Chemistry, University of Sciences, Hue University

² Faculty of Science and Technology, Hoa Sen University

³ Faculty of Chemical Engineering, Industry University of Ho Chi Minh City

* Email: vantat@gmail.com

ABSTRACT

In this study, the quantitative structure property relationships (QSPRs) of thiosemicarbazone complexes with metal ions were constructed based on multivariate linear regression, partial least square and principal component regression models. The quality of the models was evaluated based on coefficients of determination, mean standard error, and Fisher test. QSPR_{MLR} model had $R^2_{train} = 0,908$; $R^2_{CV} = 0,850$; $Q^2_{test} = 0,8542$; $MSE = 0,852$; QSPR_{PLS} model had $R^2_{train} = 0,908$; $R^2_{CV} = 0,888$; $Q^2_{test} = 0,8972$; $MSE = 0,661$; QSPR_{PCR} model had $R^2_{train} = 0,914$; $R^2_{CV} = 0,948$; $Q^2_{test} = 0,8842$; $MSE = 0,827$. These models could give a good prediction that agreed with the experiments.

Keywords: QSPR_{MLR}, QSPR_{PLS}, QSPR_{PCR}, stability constant, thiosemicarbazone.



Nguyễn Minh Quang sinh ngày 11/11/1977 tại Quảng Ngãi. Năm 2001, Ông tốt nghiệp ngành Công nghệ hóa học tại Trường Đại học Bách khoa Thành phố Hồ Chí Minh; tốt nghiệp thạc sỹ ngành Công nghệ hóa học năm 2007 tại Trường Đại học Bách khoa Thành phố Hồ Chí Minh. Năm 2016, ông là nghiên cứu sinh chuyên ngành Hóa lý và Hóa lý thuyết tại Trường Đại học Khoa học – Đại học Huế. Hiện nay, ông tham gia giảng dạy và nghiên cứu tại Trường Đại học Công nghiệp Thành phố Hồ Chí Minh.

Lĩnh vực nghiên cứu: Hóa học tính toán lượng tử, mô hình hóa QSAR, QSPR và tổng hợp vật liệu đã được công bố trên các Tạp chí trong nước và Quốc tế.



Trần Xuân Mậu sinh ngày 6/5/1958 tại Hà Tĩnh. Ông tốt nghiệp kỹ sư ngành Công nghệ hóa học năm 1982 tại Trường Đại học Kỹ thuật Slovakia tại Bratislava (Slovakia) và bảo vệ luận án phó tiến sĩ cùng ngành tại cùng trường vào năm 1986.

Lĩnh vực nghiên cứu: Hóa lý thuyết và hóa lý, Hóa học vật liệu.



Phạm Văn Tất sinh ngày 30/11/1966 tại Nam Định. Năm 1989 ông tốt nghiệp ngành Hóa học tại trường đại học Tổng hợp Hà Nội; tốt nghiệp Thạc sỹ ngành Hóa học Phân tích năm 2001 tại trường đại học Đà Lạt. Năm 2006 ông bảo vệ thành công luận án Tiến sĩ tại Viện Hóa lý và Hóa lý thuyết, trường đại học Cologne, Cộng hòa Liên bang Đức. Năm 2010 ông được nhà nước công nhận chức danh phó Giáo sư chuyên ngành Hóa lý và Hóa lý thuyết. Hiện nay là trưởng Bộ môn ngành Công nghệ Kỹ thuật Môi trường, trường đại học Hoa Sen.

Lĩnh vực nghiên cứu: Hóa học tính toán lượng tử và Mô phỏng Monte Carlo các trạng thái cân bằng, Hóa học Phân tích, Hóa học Môi trường và Quản lý và Đánh giá chất lượng Môi trường bằng GIS, Viễn thám và Hệ thống Trí tuệ Nhân tạo.