Tạp chí Khoa học Đại học Quốc gia Hà Nội:
Khoa học Tự nhiên và Công nghệ
Website: https://js.vnu.edu.vn/NST

Xây dựng mô hình QSAR mô tả hoạt tính Estrogen của Bisphenol A và các dẫn xuất sử dụng lý thuyết hóa lượng tử và phép hồi quy đa biến tuyến tính

Vũ Văn Đạt[1,*], Lâm Ngọc Thiềm[1], Lê Kim Long[2],
Đoàn Văn Phúc[3], Nguyễn Hoàng Trang[2], Nguyễn Văn Tráng[4]

[1]*Khoa Hóa học, Trường Đại học Khoa học Tự nhiên, ĐHQGHN, 19 Lê Thánh Tông, Hà Nội, Việt Nam*
[2]*Trường Đại học Giáo dục, ĐHQGHN, 144 Xuân Thủy, Cầu Giấy, Hà Nội, Việt Nam*
[3]*Viện Hóa học Vật liệu, Viện Khoa học và Công nghệ Quân sự, 17 Hoàng Sâm, Hà Nội, Việt Nam*
[4]*Viện Kỹ thuật Nhiệt đới, Viện Hàn lâm KH&CN Việt Nam,18 Hoàng Quốc Việt, Hà Nội, Việt Nam*

Nhận ngày 20 tháng 8 năm 2018
Chỉnh sửa ngày 29 tháng 8 năm 2018; Chấp nhận đăng ngày 30 tháng 8 năm 2018

Tóm tắt: Bài báo trình bày kết quả nghiên cứu và xây dựng mô hình QSAR mô tả hoạt tính estrogen đối với Bisphenol A và các dẫn xuất. Các thông số lượng tử (cấu trúc, năng lượng) được tính toán theo phương pháp phiếm hàm mật độ (DFT), phiếm hàm tương quan trao đổi meta GGA và bộ hàm cơ sở 6-31+G*. Các tham số hóa lượng tử thu được – đóng vai trò là các biến độc lập, kết hợp với các giá trị sinh hóa thực nghiệm được chọn lọc – đóng vai trò là các biến phụ thuộc, tạo ra bộ dữ liệu số để xây dựng mô hình trên cơ sở sử dụng phương pháp hồi quy đa biến tuyến tính. Kết quả cho thấy, đã xây dựng được mô hình QSAR với 10 biến độc lập, mô tả/dự đoán tốt hoạt tính estrogen của hệ chất nghiên cứu. Mô hình dự đoán QSAR thu được có R^2> 0,9, $Q^2_{LOO} = 0,512$ và $R^2_{predicted} = 0,5438$. Các chỉ số thống kê cho thấy, mô hình xây dựng bằng phương pháp hồi quy đa biến tuyến tính sử dụng các tham số hóa lượng tử trong nghiên cứu này có thể ứng dụng như một mô hình dự đoán hoạt tính estrogen cho các dẫn xuất và các chất đồng dạng của BPA với độ tin cậy trung bình.

Keywords: Bisphenol A, Estrogen, DFT, QSAR, hồi quy đa biến tuyến tính.

1. Đặt vấn đề

Bisphenol A (BPA) là một trong những hóa chất phổ biến nhất trên thế giới, được phát hiện lần đầu tiên vào những năm 1890 và được sử dụng thương mại từ năm 1950. Tuy nhiên, đến đầu những năm 1990, các nghiên cứu chỉ ra rằng BPA tồn dư trong các sản phẩm nhựa polycarbonate và nhựa epoxy là nguyên nhân gây ra những rối loạn nội tiết trong cơ thể, góp phần gia tăng rủi ro đối với các bệnh liên quan đến tim mạch, béo phì [1] tiểu đường [2, 3], ảnh

*Tác giả liên hệ. ĐT.: 84-934277732.
Email: vvdat@most.gov.vn
https://doi.org/10.25073/2588-1140/vnunst.4778

hưởng khả năng phát triển trí não của trẻ em [4]; ảnh hưởng hoạt động tuyến tiền liệt, gây ra ung thư vú, u nang buồng trứng …[5].

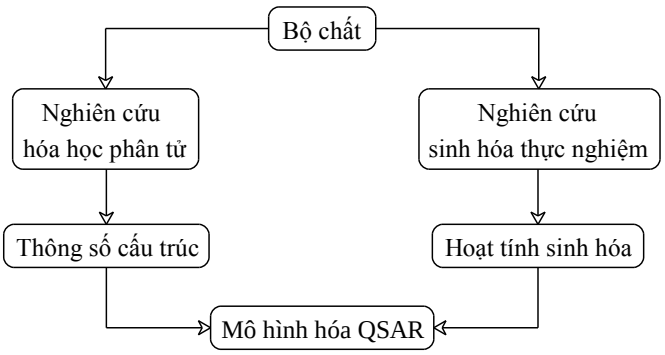
Hiện nay, rất nhiều hãng đã chuyển sang sản xuất các sản phẩm không chứa BPA, thay thế BPA bằng BPS (bisphenol – S), BPF (bisphenol – F) hoặc các dẫn suất khác của BPA. Tuy nhiên, những nghiên cứu gần đây cho thấy kể cả một số lượng nhỏ BPS và BPF cũng có thể ảnh hưởng đến chức năng của các tế bào giống như BPA, mặc dù liều lượng tiếp xúc an toàn của chúng không giống nhau [6].

Nhiệm vụ cấp thiết đặt ra trước mắt cho các nhà khoa học hiện nay là xác định liều lượng tối thiểu gây ra các tác dụng sinh học tiêu cực của BPA và nhóm dẫn xuất đối với con người cũng như môi trường sinh thái, từ đó, có hướng khai thác và sử dụng hiệu quả các chế phẩm từ BPA mà vẫn bảo đảm an toàn cho người sử dụng. Theo đó, nghiên cứu QSAR (quantitative structure-activity relationship - nghiên cứu mối quan hệ định lượng giữa cấu trúc phân tử và hoạt chất sinh học) được ứng dụng rộng rãi như một giải pháp tối ưu để kiểm tra, đánh giá và sàng lọc về khả năng đáp ứng sinh học của bộ chất cần khảo sát cũng như để thiết lập, kiểm nghiệm và cho đề xuất về các hợp chất mới có khả năng đáp ứng các hoạt tính sinh học nhất định. Các mô hình QSAR được xây dựng trên quy trình nghiên cứu hóa học phân tử và phân tích sinh hóa thực nghiệm để tham số hóa các đặc điểm cấu trúc và hoạt tính sinh học của một bộ chất, kết hợp với quy trình xử lý số liệu bằng các phương pháp thống kê để thiết lập mối quan hệ định lượng cấu trúc - hoạt tính dưới dạng các mô hình toán học. Mỗi mô hình QSAR được thực hiện trên một bộ chất, có phạm vi ứng dụng nhất định, trong đó, có thể dự đoán sinh học đối với nhiều hợp chất khác mà không cần thực hiện các phép kiểm tra sinh học phức tạp bằng thực nghiệm, cho phép tiết kiệm thời gian, chi phí và nhân lực. Hiện nay, nghiên cứu QSAR được phát triển mạnh mẽ trên toàn cầu, tạo ra một bộ cơ sở dữ liệu khổng lồ về các hoạt tính thực nghiệm của rất nhiều hợp chất cũng như xây dựng thành công rất nhiều chương trình cho phép khảo sát và tính toán các tham số cấu trúc đặc trưng cho phân tử.

Mục đích của nghiên cứu này là khai thác các dữ liệu sẵn có, kết hợp với việc chọn lọc các công cụ tính toán và xử lý số liệu hiệu quả, để mô hình hóa mối quan hệ giữa hoạt tính sinh học với các tham số cấu trúc hóa lượng tử đối với Bisphenol A và nhóm dẫn xuất bằng phương pháp hồi quy đa biến tuyến tính, đồng thời đánh giá chất lượng dự đoán của mô hình.

2. Đối tượng và phương pháp nghiên cứu

Sơ đồ tổng quát của nghiên cứu QSAR được trình bày ở Hình 1.

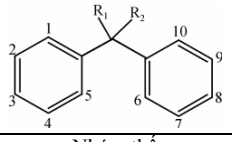


Hình 1. Sơ đồ tổng quát trong nghiên cứu QSAR.

2.1. Đối tượng nghiên cứu

Bộ dữ liệu được sử dụng trong nghiên cứu này gồm 23 hợp chất được tổng hợp và nghiên cứu hoạt tính sinh học bởi nhóm nghiên cứu của trường Đại học Minnesota và trường Đại học New Orleans, Hoa Kỳ [7]. Cấu tạo của các phân tử trong bộ dữ liệu được trình bày ở Bảng 1.

Bảng 1. Cấu tạo phân tử của bộ chất dữ liệu [7]

STT	Hợp chất											
		Nhóm thế										
		1	2	3	4	5	6	7	8	9	R_1	R_2
1	DMB Bis A	H	H	OH	H	H	H	H	OH	H	CH_3	$CH_2CH(CH_3)_2$
2	HPTE	H	H	OH	H	H	H	H	OH	H	H	CCl_3
3	MM4	H	H	OH	H	H	H	H	OH	H	C_2H_5	C_2H_5
4	DM DMB Bis A	H	CH_3	OH	H	H	H	CH_3	OH	H	CH_3	$CH_2CH(CH_3)_2$
5	HF Bis A	H	H	OH	H	H	H	H	OH	H	CF_3	CF_3
6	Bis B	H	H	OH	H	H	H	H	OH	H	CH_3	C_2H_5
7	DM Bis A	H	CH_3	OH	H	H	H	H	OH	CH_3	CH_3	CH_3
8	P Bis A	H	H	OH	H	H	H	H	OH	H	CH_3	C_6H_5
9	MM2	H	H	OH	H	H	H	H	OH	H	H	C_2H_5
10	Bis A	H	H	OH	H	H	H	H	OH	H	CH_3	CH_3
11	PCP	H	H	H	H	H	H	H	OH	H	CH_3	CH_3
12	TM Bis A	H	CH_3	OH	CH_3	H	H	CH_3	OH	CH_3	CH_3	CH_3
13	MH MM1	H	H	OH	H	H	H	H	H	H	CH_3	H
14	o,p'-Bis A	H	H	H	H	OH	H	H	OH	H	CH_3	CH_3
15	MH Bis F	H	H	H	H	H	H	H	OH	H	H	H
16	MM1	H	H	OH	H	H	H	H	OH	H	H	CH_3
17	Bis F	H	H	OH	H	H	H	H	OH	H	H	H
18	DM HPTE	H	CH_3	OH	H	H	H	CH_3	OH	H	H	CCl_3
19	1844-00-44	H	H	OH	H	H	H	H	OH	H	H	$CH(CH_3)_2$
20	Mono Mxy Bis A	H	H	OH	H	H	H	H	OCH_3	H	CH_3	CH_3
21	TC Bis A	H	Cl	OH	Cl	H	H	Cl	OH	Cl	CH_3	CH_3
22	TB Bis A	H	Br	OH	Br	H	H	Br	OH	Br	CH_3	CH_3
23	Mxy Bis A	H	H	OCH_3	H	H	H	H	OCH_3	H	CH3	CH3

Hoạt tính sinh học được chọn lựa cho nghiên cứu này là *mức độ hoạt động estrogen* được đánh giá dưới dạng *biểu hiện sinh học của các gen tín hiệu* (reporter gene) được gắn vào trong tế bào. Theo đó, thông số hoạt tính được chọn là EC50 (half maximal effective concentration), được đo bằng đơn vị mol/l (M), là nồng độ gây 50% hiệu ứng sinh học tối đa, trong thời gian phơi nhiễm thực nghiệm là 72 giờ. Dữ liệu thực nghiệm về hoạt tính sinh học của các chất nghiên cứu được trình bày trong Bảng 2.

2.2. Tính toán tham số cấu trúc và chọn lọc biến

Các thông số hóa lượng tử đặc trưng cho cấu trúc phân tử được tính toán dựa trên lý thuyết phiếm hàm mật độ (Density Functional Theory- DFT) thực hiện trên phần mềm Gaussian 09 [8]. Các tính toán sử dụng phiếm hàm lai hóa meta GGA và bộ hàm cơ sở 6-31+G*. Phương pháp này đã được chứng minh là phù hợp với hệ nghiên cứu [9].

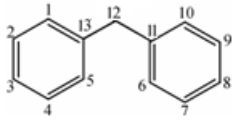
Không phải tất cả các tham số cấu trúc được tính toán đều có ý nghĩa thống kê đối với mô hình, do đó, phải tiến hành đánh giá chọn lọc các biến tiềm năng để xây dựng bộ dữ liệu cấu trúc. Việc chọn biến được thực hiện bằng cách khảo sát mức độ tương quan lẫn nhau giữa các tham số cấu trúc và mức độ tương quan của chúng so với các tham số hoạt tính, thông qua ma trận hệ số tương quan Pearson. Theo đó, các tham số cấu trúc không có sự tương quan (hoặc

kém tương quan) so với tham số hoạt tính sẽ bị loại bỏ; đối với các tham số cấu trúc không chỉ có mối tương quan với hoạt tính mà còn có sự tương quan lẫn nhau, chỉ giữ lại 1 tham số có sự tương quan lớn nhất so với hoạt tính. Danh sách các thông số cấu trúc được chọn lọc tính toán trình bày trong Bảng 3.

Bảng 2. Dữ liệu thực nghiệm về hoạt tính sinh học của bộ chất nghiên cứu [7]

STT	Hợp chất	lgEC_{50}(Gene induction)
1	DMB Bis A	- 2.03
2	HPTE	- 3.37
3	MM4	- 2.28
4	DM DMB Bis A	- 1.99
5	HF Bis A	- 2.79
6	Bis B	- 3.28
7	DM Bis A	- 3.31
8	P Bis A	- 4.05
9	MM2	- 3.57
10	Bis A	- 2.56
11	PCP	- 4.05
12	TM Bis A	- 3.80
13	MH MM1	- 4.05
14	o,p'-Bis A	- 3.96
15	MH Bis F	- 4.05
16	MM1	- 3.15
17	Bis F	- 3.28
18	DM HPTE	- 2.91
19	1844-00-44	- 3.38
20	Mono Mxy Bis A	- 4.04
21	TC Bis A	– 6.04
22	TB Bis A	– 6.04
23	Mxy Bis A	– 6.04

Bảng 3. Các thông số lượng tử lựa chọn để tính toán

STT	Ký hiệu	Ý nghĩa	Đơn vị
1	DDLK	Độ dài liên kết	Å
2	GLK	Góc liên kết	Độ
3	GND	Góc nhị diện	Độ
4	E_{HOMO}	Năng lượng của orbital phân tử bị chiếm cao nhất	eV
5	E_{LUMO}	Năng lượng của orbital phân tử chưa bị chiếm thấp nhất	eV
6	μ	Moment lưỡng cực phân tử	Debye
7	E_{sp}	Năng lượng phân tử	eV
8	ΔE	$\Delta E = E_{LUMO} - E_{HOMO}$ đặc trưng độ chênh lệch giữa hai mức năng lượng orbital phân tử HOMO và LUMO	eV
9	χ	$\chi = \frac{-(E_{LUMO} + E_{HOMO})}{2}$ độ âm điện của phân tử (trong khuôn khổ thuyết Koopman E_{LUMO} = -A, E_{HOMO}= -I, trong đó A: ái lực electron, I- năng lượng ion hóa)	eV
10	η	$\eta = \frac{E_{LUMO} - E_{HOMO}}{2}$ độ cứng của phân tử (trong khuôn khổ thuyết Koopman E_{LUMO} = -A, E_{HOMO}=-I, trong đó A: ái lực electron, I- năng lượng ion hóa)	eV
11	ω	$\omega = \mu^2 / 2\eta$	eV
12	C1, C2, … C13	Mật độ điện tích tại các nguyên tử cacbon trên khung phân tử như ký hiệu trên hình vẽ 	

2.3. Mô hình hóa QSAR

Trong bài báo này, chúng tôi nghiên cứu và xây dựng 2 mô hình, gồm: *Mô hình đánh giá* và *Mô hình dự đoán* bằng phương pháp hồi quy đa biến tuyến tính, tiến hành trên phần mềm STATGRAPHICS Centurion 15 [10]. Chất

lượng và khả năng dự đoán của các mô hình được đánh giá bằng các phép tham chiếu trong và ngoài (internal và external validation) thông qua các chỉ số thống kê.

Mô hình đánh giá được xây dựng trên toàn bộ tập dữ liệu và sử dụng phép tham chiếu chéo -LOO (Leave-one-out cross-validation) để tính toán các chỉ số dự đoán nội mô hình. Ở mỗi thủ tục của phép tham chiếu chéo LOO, một hợp chất sẽ bị loại bỏ khỏi tập dữ liệu, sau đó tiến hành xây dựng mô hình QSAR mới trên các hợp chất còn lại và sử dụng mô hình này để dự đoán hoạt tính của hợp chất bị loại bỏ. Thủ tục này được lặp lại cho đến khi tất cả các hợp chất của bộ dữ liệu đều 1 lần bị loại bỏ và được dự đoán hoạt tính. Sau khi kết thúc vòng lặp tham chiếu chéo, kết quả dự đoán của tất cả các hợp chất được tổng hợp để tính toán các chỉ số dự đoán nội mô hình [11, 12].

Tuy nhiên, theo Tropsha và các chuyên gia trong lĩnh vực nghiên cứu QSAR [13], các chỉ số dự đoán nội mô hình không đảm bảo khả năng dự đoán đối với các hợp chất mới nằm ngoài mô hình, một mô hình QSAR chất lượng và có khả năng dự đoán ổn định thực sự phải được xây dựng kèm theo các phép tham chiếu trong lẫn ngoài.

Mô hình dự đoán với các phép tham chiếu trong và ngoài được xây dựng trên cơ sở phân chia bộ dữ liệu thành bộ luyện (training set) và bộ kiểm (test set). Theo đó, mô hình dự đoán sẽ được xây dựng trên các hợp chất của bộ luyện. Phép tham chiếu chéo LOO thực hiện trên bộ luyện được sử dụng làm phép tham chiếu trong để cho các đánh giá về khả năng dự đoán nội mô hình. Phép tham chiếu ngoài của mô hình được thực hiện bằng cách sử dụng mô hình để dự đoán hoạt tính của các hợp chất trong bộ kiểm, sau đó, các giá trị dự đoán sẽ được tổng hợp để tính toán các chỉ số dự đoán ngoại mô hình.

Các chỉ số thống kê quan trọng của mô hình QSAR:

Hệ số xác định (R^2): bình phương hệ số tương quan giữa các giá trị hoạt tính sinh hóa tính theo mô hình hồi qui và các giá trị thực nghiệm.

$$R^2 = 1 - \frac{\sum (y_{obs} - y_{calc})^2}{\sum (y_{obs} - \bar{y}_{training})^2}, \qquad (2.1)$$

Nếu giá trị R^2 càng gần 1, mô hình mô tả càng tốt các số liệu thực nghiệm, các giá trị hoạt tính y_{calc} rất gần các giá trị hoạt tính thực nghiệm y_{obs}.

Hệ số hiệu chỉnh (R_a^2): được sử dụng để phản ánh sát hơn mức độ phù hợp của mô hình hồi quy đa biến tuyến tính (MLR). Giá trị này thường được dùng để so sánh mô hình hồi quy với số biến độc lập khác nhau.

$$R_a^2 = R^2 - \frac{k(1-R^2)}{N-k-1}, \qquad (2.2)$$

N – số hợp chất dùng xây dựng mô hình; k – số biến độc lập sử dụng trong mô hình;

Phương sai (MSE): sử dụng để phản ánh mức độ phân tán của một tập dữ liệu.

$$MSE = \frac{\sum (y_{obs} - y_{calc})^2}{n}, \qquad (2.3)$$

Tính tổng quát của mô hình (Q^2): được sử dụng để kiểm tra khả năng dự đoán nội mô hình bằng phép tham chiếu trong LOO (Leave-one-out cross-validation) thực hiện trên bộ khảo luyện (training set).

$$Q^2_{LOO} = 1 - \frac{\sum (y_{obs(\text{training})} - y_{pred(\text{training})})^2}{\sum (y_{obs(\text{training})} - \bar{y}_{\text{training}})^2}, \qquad (2.4)$$

3. Kết quả và thảo luận

Tính toán 50 tham số đặc trưng của 23 phân tử được tiến hành. Các tham số được đánh giá thông qua hệ số tương quan Pearson để chọn lọc ra các biến có ý nghĩa dự đoán đối với mô hình; kết quả đánh giá đã chọn ra 18 tham số có thể đóng vai trò là các biến độc lập để xây dựng mô hình, dữ liệu về các tham số này được cho ở Bảng 4.

Bảng 4. Các thông số cấu trúc tính toán theo bộ hàm 6-31+G* được chọn để xây dựng mô hình

No.	Hợp chất	E_{HOMO} (eV)	ΔE (eV)	μ (Debye)	E_{sp} (Hartree)	ω	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
1	DMB Bis A	-5,922	-5,4223	1,6253	-849,632074	0,487157536	-0,222	-0,311	0,303	-0,277	-0,205	-0,209	-0,277	0,302	-0,31	-0,222	-0,048	-0,1	-0,049
2	HPTE	-6,234	-4,8139	1,8774	-2071,148444	0,732203073	-0,198	-0,311	0,315	-0,278	-0,194	-0,201	-0,275	0,311	-0,305	-0,213	-0,069	-0,304	-0,079
3	MM4	-5,986	-5,4357	2,2785	-810,318242	0,955091619	-0,218	-0,312	0,303	-0,281	-0,218	-0,218	-0,281	0,303	-0,312	-0,218	-0,052	-0,098	-0,052
4	DM DMB Bis A	-5,793	-5,4632	0,9222	-928,270067	0,15566126	-0,23	-0,304	0,308	-0,08	-0,206	-0,203	-0,08	0,308	-0,305	-0,23	-0,04	-0,09	-0,04
5	HF Bis A	-6,552	-5,5891	1,5767	-1327,16975	0,444788135	-0,197	-0,307	0,319	-0,277	-0,198	-0,198	-0,277	0,319	-0,307	-0,197	-0,095	-0,032	-0,095
6	Bis B	-5,966	-5,4090	2,0765	-771,007595	0,797168315	-0,221	-0,311	0,303	-0,279	-0,212	-0,214	-0,28	0,303	-0,312	-0,218	-0,055	-0,105	-0,052
7	DM Bis A	-5,776	-5,4044	1,9216	-810,333677	0,683251142	-0,215	-0,115	0,308	-0,272	-0,215	-0,215	-0,272	0,308	-0,115	-0,215	-0,045	-0,112	-0,045
8	P Bis A	-5,825	-5,2934	1,3032	-923,460706	0,320832033	-0,203	-0,307	0,349	-0,273	-0,203	-0,207	-0,275	0,349	-0,305	-0,198	-0,067	-0,077	-0,066
9	MM2	-5,953	-5,3396	2,0483	-731,698103	0,785737802	-0,219	-0,311	0,303	-0,278	-0,21	-0,215	-0,279	0,303	-0,311	-0,215	-0,063	-0,282	-0,061
10	Bis A	-5,933	-5,4131	1,7296	-731,697812	0,552645411	-0,222	-0,311	0,302	-0,277	-0,208	-0,208	-0,277	0,302	-0,311	-0,222	-0,052	-0,113	-0,052
11	PCP	-6,068	-5,5755	1,5457	-656,476699	0,428529031	-0,242	-0,235	-0,248	-0,236	-0,228	-0,208	-0,277	0,303	-0,311	-0,222	-0,052	-0,115	-0,021
12	TM Bis A	-5,669	-5,4354	1,3057	-888,971469	0,313657385	-0,223	-0,11	0,312	-0,073	-0,212	-0,212	-0,073	0,312	-0,11	-0,223	-0,037	-0,111	-0,037
13	MH MM1	-6,078	-5,5655	1,5352	-617,167107	0,423500989	-0,244	-0,235	-0,247	-0,235	-0,225	-0,21	-0,276	0,302	-0,311	-0,217	-0,057	-0,29	-0,028

14	o,p'-Bis A	-5,901	-5,4446	1,7663	-731,695221	0,573033095	-0,221	-0,264	-0,228	-0,307	0,318	-0,213	-0,279	0,299	-0,311	-0,212	-0,039	-0,111	-0,068
15	MH Bis F	-6,117	-5,5456	1,4452	-577,854902	0,376622292	-0,237	-0,235	-0,247	-0,235	-0,229	-0,208	-0,276	0,303	-0,311	-0,217	-0,069	-0,479	-0,039
16	MM1	-5,951	-5,4286	1,4398	-692,388119	0,381886079	-0,223	-0,311	0,303	-0,276	-0,205	-0,211	-0,277	0,302	-0,311	-0,217	-0,057	-0,287	-0,059
17	Bis F	-5,990	-5,4036	1,4703	-653,075856	0,400066241	-0,217	-0,312	0,302	-0,277	-0,209	-0,209	-0,277	0,302	-0,312	-0,217	-0,069	-0,476	-0,069
18	DM HPTE	-6,077	-4,7187	2,3026	-2149,78590	4,12357909	-0,192	-0,115	0,321	-0,273	-0,2	-0,197	-0,075	0,315	-0,299	-0,221	-0,062	-0,303	-0,073
19	1844-00-44	-5,928	-5,3184	0,6328	-771,014961	0,075292205	-0,218	-0,307	0,302	-0,276	-0,21	-0,212	-0,276	0,302	-0,309	-0,2313	-0,051	-0,269	-0,051
20	Mono Mxy Bis A	-5,675	-5,4329	0,9440	-771,019282	0,164011528	-0,21	-0,311	0,335	-0,257	-0,203	-0,199	-0,272	0,348	-0,307	-0,212	-0,066	-0,07	-0,063
21	TC Bis A	-6,558	-5,3271	3,5848	-2570,07511	8,412343891	-0,239	-0,103	0,283	-0,085	-0,224	-0,224	-0,085	0,283	-0,103	-0,239	-0,03	-0,1	-0,03
22	TB Bis A	-6,517	-5,2536	3,5184	-11016,2170	12,35630294	-0,239	-0,17	0,281	-0,155	-0,221	-0,221	-0,155	0,281	-0,17	-0,239	-0,029	-0,1	-0,029
23	Mxy Bis A	-5,799	-5,4014	0,9166	-810,308156	0,155544512	-0,223	-0,318	0,306	-0,267	-0,212	-0,212	-0,267	0,306	-0,318	-0,223	-0,051	-0,113	-0,051

Từ 18 biến trên, thực hiện hồi quy chọn mô hình (regrassion model selection), kết quả thu được tổng cộng **89.692** mô hình chứa từ 3 đến 18 biến. Một vài mô hình có hệ số R_a^2 cao hơn được liệt kê ở Bảng 5, trong đó, mô hình 1 có hệ số R_a^2 cao nhất, chứa 11 biến độc lập, được chọn lựa để phát triển thành các mô hình nghiên cứu.

Bảng 5. Các thông số thống kê của một số mô hình có hệ số R_a^2 cao nhất

Mô hình	MSE	R^2	R_a^2	Variables
1	**0,24086**	**90,7284**	**81,4568**	**ACDEFJKNOQR**
2	0,330852	86,1065	74,5286	ABCDFLNOQR
3	0,335496	85,9115	74,171	ACDIJMNOQR
4	0,335571	85,9083	74,1652	ACDEFJLNPQ
5	0,34088	85,6854	73,7565	ABCDEFJLNQ
6	0,438909	80,0329	66,2095	ACDEFJLNQ
7	0,463702	78,905	64,3007	BDEGILOQR
8	0,470142	78,612	63,805	ACDFGIJNQ
9	0,48003	78,1622	63,0437	ACDIJMNQR
10	0,484112	77,9765	62,7294	BCDGILOQR

Trong đó, ký hiệu các biến như sau: A= C1; B= C10; C= C11; D= C12; E= C13; F= C2; G= C3; H= C4; I= C5; J= C6; K= C7; L= C8; M= C9; N=ΔE; O= E_{HOMO}; P= Esp; Q=μ; R=ω;

Tiến hành tính toán và đánh giá các chỉ số thống kê đối với bộ 11 biến độc lập trên, kết quả chỉ ra rằng, có 1 biến (biến K) có hệ số P-value kém và được loại bỏ khỏi mô hình. Như vậy, bộ thông số cấu trúc tối ưu được chọn để xây dựng các mô hình nghiên cứu bao gồm 10 thông số: C1, C11, C12, C13, C2, C6, ΔE, E_{HOMO}, μ, ω.

Mô hình đánh giá được xây dựng trên tất cả 23 hợp chất của bộ dữ liệu với 10 biến độc lập, kết hợp với phép tham chiếu trong LOO-CV, cho kết quả như sau:

LgEC50 = (31,6935 ± 8,77747) + (292,701 ± 51,3223) *C1 + (158,384 ± 27,5407) *C11 + (10,4344 ± 2,88221) *C12 + (131,757 ± 27,8454) *C13 + (15,1931 ± 2,88221) *C2 + (182,188 ± 37,9349) *C6 + (9,28815 ± 1,87743) *ΔE + (3,78452 ± 1,17537) *E_{HOMO} + (4,50892 ± 0,801006) *μ + (4,86291 ± 1,18376) *ω (3.1)

N = 23; k = 10; *R^2 = 0,8803; F-Ratio = 8,82; P-Value = 0,0004; P-value of DW statistic = 0,1519; Q^2_{LOO} = 0,5564;*

Giá trị *P-Value* của mô hình < 0,05 (=0,0004), cho thấy rằng, giữa biến phụ thuộc và các biến độc lập có mối liên hệ chặt chẽ với nhau với độ tin cậy thống kê lớn hơn 95%; chứng tỏ bộ biến độc lập được lựa chọn xây dựng mô hình là phù hợp. Giá trị *P-value of DW statistic* > 0,05, chứng tỏ rằng không tồn tại mối tương quan nào dựa vào thứ tự nhập biến trong tập số liệu. Hệ số xác định R^2 ~ 0.9 cho thấy mô hình có khả năng tái lập dữ liệu tốt. Hệ số khái quát hóa $Q^2_{LOO} = 0,5564$ cho thấy mô hình có khả năng dự đoán ở mức trung bình.

Hoạt tính của bộ dữ liệu ước tính theo mô hình (3.1) được cho trong Bảng 6 dưới dạng logEC50-estimated.

Mô hình dự đoán được xây dựng trên bộ luyện (training set) gồm 18 hợp chất với 10 biến độc lập, kết hợp với phép tham chiếu trong LOO-CV thực hiện trên bộ luyện và phép tham chiếu ngoài thực hiện trên bộ kiểm (test set) gồm 5 hợp chất (số thứ tự 21, 13, 9, 17, 3 trong Bảng 1), cho kết quả như sau:

LgEC50 = (42,026 ± 6,00388) + (384,159 ± 42,158) *C1 + (211,482 ± 21,0181) *C11 – (14,3442 ± 1,75148) *C12 + (163,684 ± 20,8313) *C13 – (19,1046 ± 2,34106) *C2 + (238,662 ± 26,1866) *C6 – (12,6523 ± 1,52011) *ΔE – (4,79409 ± 0,8085) *E_{HOMO} + (4,68799 ± 0,563598) *μ - (4,85759 ± 0,8259) *ω (3.2)

N = 18; k = 10; R^2 = 0,9629; F-Ratio = 18,16; P-Value = 0,0004; P-value of DW statistic = 0,0732, Q^2_{LOO} = 0,512; $R^2_{predicted}$ = 0,5438

Giá trị R^2 = 0,9629 chứng tỏ rằng, mô hình dự đoán có khả năng tái lập dữ liệu tốt. Giá trị Q^2_{LOO} = 0,512; $R^2_{predicted}$ = 0,5438 cho thấy, độ ổn định của mô hình cũng như khả năng dự đoán ngoại mô hình ở mức trung bình.

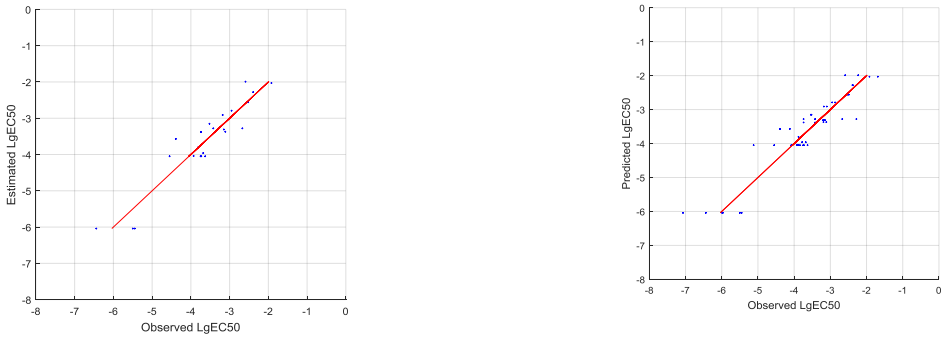
Hoạt tính của bộ dữ liệu dự đoán theo mô hình (3.1) được cho trong Bảng 6 dưới dạng logEC50-predicted. Mức độ tương quan giữa giá trị dự đoán (predicted), giá trị ước tính

(estimated) và giá trị thực nghiệm (observed) được biểu diễn trên các đồ thị ở Hình 2.

Ngoài ra, các hệ số ở cả 2 mô hình hồi quy cho thấy rằng, các thông số C1, C6, C11 và C13 có tác động mạnh nhất đến hoạt tính estrogen của nhóm chất nghiên cứu.

Bảng 6. Giá trị hoạt tính dự đoán

STT	Hợp chất	LgEC_{50} (Observed)	LgEC_{50} (Estimated)	LgEC_{50} (Predicted)
1	DMB Bis A	-2,03	-1,92	-1,69
2	HPTE	-3,37	-3,11	-3,19
3	MM4	-2,28	-2,39	-2,38
4	DM DMB Bis A	-1,99	-2,59	-2,23
5	HF Bis A	-2,79	-2,95	-2,87
6	Bis B	-3,28	-3,42	-3,74
7	DM Bis A	-3,31	-3,15	-3,20
8	P Bis A	-4,05	-3,75	-3,91
9	MM2	-3,57	-4,39	-4,12
10	Bis A	-2,56	-2,52	-2,49
11	PCP	-4,05	-3,63	-4,08
12	TM Bis A	-3,8	-3,81	-3,88
13	MH MM1	-4,05	-4,55	-5,12
14	o,p'-Bis A	-3,96	-3,68	-3,78
15	MH Bis F	-4,05	-3,73	-3,84
16	MM1	-3,15	-3,52	-3,54
17	Bis F	-3,28	-2,67	-2,28
18	DM HPTE	-2,91	-3,18	-3,10
19	1844-00-44	-3,38	-3,74	-3,42
20	Mono Mxy Bis A	-4,04	-3,93	-3,89
21	TC Bis A	-6,04	-6,44	-7,07
22	TB Bis A	-6,04	-5,45	-5,97
23	Mxy Bis A	-6,04	-5,50	-5,99



Hình 2. Đồ thị tương quan giữa các giá trị tính toán, dự đoán và thực nghiệm

4. Kết luận

Việc nghiên cứu QSAR đối với BPA và các dẫn xuất trong bài báo này được thực hiện thông qua các tính toán hóa lượng tử (sử dụng phương pháp phiếm hàm mật độ - DFT), phiếm hàm tương quan trao đổi meta GGA và bộ hàm cơ sở 6-31+G* kết hợp với phương pháp xử lý số liệu kinh điển hồi quy đa biến tuyến tính. Kết quả khảo sát thu được mô hình dự đoán QSAR với 10 biến (C1, C11, C12, C13, C2, C6, ΔE, E_{HOMO}, μ, ω) có khả năng tái lập tốt với hệ số xác định R^2= 96,29%; $Q^2_{LOO} = 0,512$ và khả năng dự đoán bên ngoài mô hình ở mức trung bình với hệ số xác định $R^2_{predicted} = 0,5438$. Do đó, có thể áp dụng được mô hình này trong thực tế để dự đoán hoạt tính estrogen của nhóm dẫn xuất và các chất đồng dạng của BPA chưa được nghiên cứu thực nghiệm.

Hướng nghiên cứu tiếp theo của công trình này là hoàn chỉnh phương pháp luận nghiên cứu QSAR qua việc áp dụng các phương pháp xử lý số liệu hiện đại (như phương pháp mạng nơron nhân tạo) nhằm xây dựng các mô hình QSAR có khả dự đoán ngoại mô hình cao hơn.

Tài liệu tham khảo

[1] Rezg R, El-Fazaa S, Gharbi N, Mornagui B (March 2014). "Bisphenol A and human chronic diseases: Current evidences, possible mechanisms, and future perspectives".Environment International 2014, 64, 83–90.

[2] Melzer D, Rice NE, Lewis C, Henley WE, Galloway TS (2010). Zhang, Baohong, ed."Association of Urinary Bisphenol a Concentration with Heart Disease: Evidence from NHANES 2003/06". PLoS ONE 5 (1).

[3] Manikkam, M.; Tracey, R.; Guerrero-Bosagna, C.; Skinner, M. K. (January 24, 2013). "Plastics derived endocrine disruptors (BPA, DEHP and DBP) induce epigenetic transgenerational inheritance of obesity, reproductive disease and sperm epimutations".PLoS ONE 8 (1). 1–16.

[4] D.R. Doerge, N.C. Twaddle, M. Vanlandingham, R.P. Brown, J.W. Fisher, Toxicol. Appl. Pharmacol. 2011, 255, 261.

[5] Ho SM, Tang WY, Belmonte de Frausto J, Prins GS (2006). "Developmental exposure to estradiol and bisphenol A increases susceptibility to prostate carcinogenesis and epigenetically regulates phosphodiesterase type 4 variant 4". Cancer Res. 66 (11): 5624–32.

[6] Johanna R. Rochester and Ashley L. Bolden (2015 Jul) "Bisphenol S and F: A Systematic Review and Comparison of the Hormonal Activity of Bisphenol A Substitutes". Environ Health Perspect123(7):643-50.

[7] Kelly, P. C., William, A. T., Thomas, E. W., QSAR models of thein vitro estrogen activity of bisphenol A analogs, QSAR Comb.Sci., 2003, 22: 78—88.

[8] Frisch, M. J. T., G. W. et al , Gaussian 09, Revision D.01. Gaussian, Inc., Wallingford CT, 2009.

[9] Zhao, Y.; Truhlar, D., The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. Theor Chem Account 2008, 120 (1-3), 215-241.

[10] Statpoint, Inc. The User's Guide to STATGRAPHICS® Centurion XV, 2005.

[11] Roy K. On some aspects of validation of predictive QSAR models. Expert Opin Drug Discov 2007

[12] Wold S. Cross-validation estimation of the number of components in factor and principal components moldels. Technometrics, Vol. 20, No. 4, Part 1 (Nov., 1978), pp. 397-405.

[13] A. Golbraikh, A. Tropsha, Beware of q^2. J. Mol. Graphics Model. 20 (2002) 269–276.

Building a Quantitative Structure – Activity Relationship Model for Studying the Estrogenic Activities of Bisphenol A and Its Analogs

Vu Van Dat[1], Lam Ngoc Thiem[1], Le Kim Long[2],
Doan Văn Phuc[3], Nguyen Hoang Trang[2], Nguyen Van Trang[4]

[1]*Faculty of Chemitry, VNU University of Science, 19 Le Thanh Tong, Hanoi, Vietnam*
[2]*VNU University of Education, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam*
[3]*Institute of Chemistry and Materials, 17 Hoang Sam, Hanoi, Vietnam*
[4]*Institute for Tropical Technology, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Hanoi, Vietnam*

Abstract: This article presents the results of the QSAR study of bisphenol A and its analogs. A Molecular-chemical analysis of these substances was performed based on the Density Functional Theory (DFT) within the meta-GGA functional, using the basis sets 6-31+G*. The structure parameters of all the substances as well as their quantum parameters (orbital energies, dipole moment, electron density on atoms) were calculated. The obtained quantum parameters and known observable activities were used as input data for constructing a QSAR model, using the classical data processing method in statistical mathematics - the multivariable linear regression. The constructed model QSAR has $R^2 > 0.9$; $Q^2_{LOO} = 0.512$ and $R^2_{predicted} = 0.5438$. The statistical parameters show that the model, constructed by the method of multiple linear regression using the parameters of quantum chemistry, can be used as a predictive model of the activity of estrogens for unexplored derivatives and BPA analogs with moderate reliability.

Keywords: Bisphenol A, Estrogen, Density Functional Theory, M06 hybridmeta - GGA functional, Quantitative structure – activity relationship, multiple linear regression