Modélisation de molécules organiques hétérocycliques biologiquement actives par des méthodes QSAR/QSPR. Recherche de nouveaux médicaments

Samir Chtita

جامعة مولاي إسماعيل
UNIVERSITÉ MOULAY ISMAÏL

كلية العلوم
FACULTÉ DES SCIENCES

THÈSE

Présentée par :

Samir CHTITA

En vue de l'obtention de Doctorat en **Chimie**

Spécialité : **Chimie Physique et Modélisation**

Modélisation de molécules organiques hétérocycliques biologiquement actives par des méthodes QSAR/QSPR. Recherche de nouveaux médicaments

Soutenue le **08 Juillet 2017** devant le jury composé de :

Mr. EL IDRISSI M.	PES, Faculté des Sciences, Meknès	Président
Mr. KHALIL F.	PES, Faculté des Sciences et Techniques, Fès	Rapporteur
Mr. GUENOUNE F.	PES, Faculté des Sciences, Meknès	Rapporteur
Mr. BELYAS M.	PES, Faculté des Sciences et Techniques, Er-rachidia	Rapporteur
Mr. EL ASRI M.	PES, Faculté des Sciences et Techniques, Fès	Examinateur
Mr. BOUACHRINE M.	PES, Ecole Supérieure de Technologie, Meknès	Examinateur
Mr. LAKHLIFI T.	PES, Faculté des Sciences, Meknès	Directeur de thèse

« Aucun de nous, en agissant seul, ne peut atteindre le succès »

***Nelson Mandela**, 10 mai 1994*

Production scientifique à l'issu de ce travail

Publications 2013

1. A. Zahlou, ***S. Chtita***, M. Ghamali, L. Bejjit, T. Lakhlifi and M. Bouachrine, Electronic and photovoltaic properties of new materials based on Imidazo[1.2-a] pyrazine - Computational investigations, ***Functional Materials, 2013,*** *20(4):504-509.*
2. ***S. Chtita***, M. Ghamali, M. Larif, A. Adad, R. Hmammouchi, M. Bouachrine and T. Lakhlifi, Studies of two different cancer cell lines activities (MDAMB-231 and SK-N-SH) of imidazo[1,2-a] pyrazine derivatives by combining DFT and QSAR results, ***International Journal of Innovative Research in Science, Engineering and Technology, 2013,*** *2(11):6586-6601.*
3. ***S. Chtita***, M. Ghamali, M. Larif, A. Adad, R. Hmammouchi, M. Bouachrine and Tahar Lakhlifi, Prediction of biological activity of imidazo[1,2-a] pyrazine derivatives by combining DFT and QSAR results, ***International Journal of Innovative Research in Science, Engineering and Technology, 2013,*** *2(12):7951-7962.*
4. M. Larif, A. Zahlou, ***S. Chtita***, L. Bejjit, M. Bouachrine and Tahar Lakhlifi, New innovation in renewable energy provided by the organic solar cells based on 3-aryl-4-hydroxyquinolin-2-(1H)-one: Correlation-Structure/electronic properties, ***International Journal of Innovative Research in Science, Engineering and Technology, 2013,*** *2(12):8061-8071.*
5. M. Larif, ***S. Chtita***, A. Adad, R. Hmammouchi, M. Bouachrine and T. Lakhlifi, Predicting biological activity of anticancer molecules 3-aryl 4-hydroxyquinoline-2-(1H)-one by DFT-QSAR models, ***International Journal of Advanced Research in Computer Science and Software Engineering, 2013,*** *3(12):32-42.*

Publications 2014

6. T. Abram, ***S. Chtita***, L. Bejjit, M. Bouachrine and T. Lakhlifi, Electronic and photovoltaic properties of new materials based on 6-monosubstituted and 3,6-disubstituted acridines and their application to design novel materials for organic solar cells***, Journal of Computational Methods in Molecular Design, 2014,*** *4 (3):19-27.*
7. H. Sadki, ***S. Chtita***, M. N. Bennani, T. Lakhlifi and M. Bouachrine, New Materials Based on Acridine: Correlation Structure – Properties and Optoelectronic Applications, ***Journal of Chemistry and Materials Research, 2014,*** *1(4):112-122.*
8. A. Ousaa, B. Elidrissi, M. Ghamali, ***S. Chtita***, M. Bouachrine and T. Lakhlifi, Acute toxicity of halogenated phenols: Combining DFT and QSAR studies, ***Journal of Computational Methods in Molecular Design, 2014,*** *4(3):10-18.*
9. B. Elidrissi, A. Ousaa, M. Ghamali, ***S. Chtita***, M. A. Ajana, M. Bouachrine and T. Lakhlifi, Toxicity in-vivo of nitro-aromatic compounds: DFT and QSAR results, ***Journal of Computational Methods in Molecular Design, 2014,*** *4(3):28-37.*
10. B. Elidrissi, A. Ousaa, M. Ghamali, ***S. Chtita***, M. A. Ajana, M. Bouachrine and T. Lakhlifi, Combining DFT and QSAR result for predicting the biological activity of 1-(2-ethoxyethyl)-1H-pyrazolo[4,3-d] pyrimidines as phosphodiesterase V inhibitors, ***Journal of Computational Methods in Molecular Design, 2014,*** *4(4):140-149.*

11. M. Ghamali, ***S. Chtita***, A. Adad, R. Hmammouchi, M. Bouachrine and T. Lakhlifi, Biological activity of molecules based on benzylpiperidine inhibitors of human acetylcholinesterase (HuAChE), Predicting by Combining DFT and QSAR calculations, ***International Journal of Advanced Research in Computer Science and Software Engineering, 2014,*** *4(1):536-546.*
12. M. Ghamali, ***S. Chtita***, A. Adad, R. Hmamouchi, M. Bouachrine and T. Lakhlifi, Combining DFT and QSAR result for predicting the toxicity of a series of substituted phenols, ***Journal of Computational Methods in Molecular Design, 2014,*** *4(4):46-53.*
13. M. Larif, ***S. Chtita***, A. Adad, R. Hmammouchi, M. Bouachrine and T. Lakhlifi, Anticancer Activity of Novel Molecules Based on Imidazo [4, 5-B] Pyridine: 3D-QSAR Study, ***International Journal of Advanced Research in Computer Science and Software Engineering, 2014,*** *4(9):34-43.*

Publications 2015

14. ***S. Chtita***, T. Abram, L. Bejjit, M. Bouachrine and T. Lakhlifi, Optoelectronic properties and molecular design of new materials based on 2-thienyl-4-furyl-6-arylpyridine, ***Journal of Chemical and Pharmaceutical Research, 2015,*** *7(1):556-567.*
15. ***S. Chtita***, M. Larif, M. Ghamali, M. Bouachrine and T. Lakhlifi, Quantitative structure–activity relationship studies of dibenzo[a,d]cycloalkenimine derivatives for non-competitive antagonists of N-methyl-D-aspartate based on density functional theory with electronic and topological descriptors, ***Journal of Taibah University for Science, 2015,*** *9(2):143-154.*
16. ***S. Chtita***, M. Larif, M. Ghamali, M. Bouachrine and Tahar Lakhlifi, QSAR Studies of toxicity towards monocytes with (1,3-benzothiazol-2-yl) amino-9-(10H)-acridinone derivatives using electronic descriptors, ***Orbital: The Electronic Journal of Chemistry, 2015,*** *7(2):176-184.*
17. M. Larif, ***S. Chtita***, A. Adad, R. Hmammouchi, M. Bouachrine and T. Lakhlifi, Predicting biological activity of chalcone (1,3-diphenyl-2-propen-1-one) derivatives cytotoxicity against HT-29 human colon adenocarcinoma cell lines by DFT-QSAR models, ***Journal of Computational Methods in Molecular Design, 2015,*** *4(4):121-130.*
18. B. Elidrissi, A. Ousaa, ***S. Chtita***, M. A. Ajanaa, M. Bouachrine and T. Lakhlifi, The biological activity of pyrazine carboxamides derivatives as an herbicidal agent: combining DFT and QSAR studies, ***Journal of Computational Methods in Molecular Design, 2015,*** *5(2):83-91.*
19. B. Elidrissi, A. Ousaa, M. Ghamali, ***S. Chtita***, M. A. Ajana, M. Bouachrine and T. Lakhlifi, The acute toxicity of nitrobenzenes to Tetrahymena pyriformis: Combining DFT and QSAR studies, ***Moroccan Journal of Chemistry, 2015,*** *3(4):848-860.*
20. A. Ousaa, B. Elidrissi, M. Ghamali, ***S. Chtita***, M. Bouachrine and T. Lakhlifi, Acute toxicity of phenol derivatives: Combining DFT and QSAR studies, ***Journal of Computational Methods in Molecular Design, 2015,*** *5(3):16-24.*
21. M. Ghamali, ***S. Chtita***, A. Adad, R. Hmamouchi, M. Bouachrine and T. Lakhlifi, Combining DFT and QSAR results for predicting the cytotoxicity of a series of orthoalkyl substituted 4-X-phenols, ***Journal of Materials and Environmental Science, 2015,*** *6(1):280-288.*

Publications 2016

22. R. Hmamouchi, M. Larif, ***S. Chtita***, A. Adad, M. Bouachrine and T. Lakhlifi, Predictive modeling of the activity of coumarin derivatives DL_{50} using neural statistical approaches: Electronic descriptors, ***Journal of Taibah University for Science, 2016,*** *10(4):451-461.*
23. A. Aouidate, A. Ghaleb, M. Ghamali, ***S. Chtita***, M. Choukrad, A. Sbai, M. Bouachrine and T. Lakhlifi, Combining DFT and QSAR studies for predicting psychotomimetic activity of substituted phenethylamines using statistical methods, ***Journal of Taibah University for Science, 2016,*** *10(6):787-796.*
24. A. Mouadili, ***S Chtita***, A. Elouafi, M. Bouachrine, A. Zarrouk and R. Touzani, Biomimetic catecholase activities by prepared in-situ complexes: development of a quantitative structure–properties relationship (QSPR), ***Journal of Materials and Environmental Science, 2016,*** *7(1):210-221.*
25. ***S. Chtita***, R. Hmamouchi, M. Larif, M. Ghamali, M. Bouachrine and Tahar Lakhlifi, QSPR studies of 9-aniliioacridine derivatives for their DNA drug binding properties based on density functional theory using statistical methods: Model, validation and influencing factors***, Journal of Taibah University for Science, 2016***, *10(6):451-461.*
26. ***S. Chtita***, M. Bouachrine and T. Lakhlifi, Basic approaches and applications of QSAR/QSPR methods, ***Revue Interdisciplinaire, 2016***, *1(1).*
27. M. Ghamali, S. Chtita, M. Bouachrine and T. Lakhlifi, Méthodologie générale d'une étude RQSA/RQSP, ***Revue Interdisciplinaire, 2016,*** *1(1).*
28. M. Ghamali, ***S. Chtita***, A. Adad, R. Hmamouchi, M. Bouachrine and T. Lakhlifi, Combining DFT and QSAR computation for predicting the soil sorption coefficients of substituted phenols and anilines, ***Journal of Materials and Environmental Science, 2016,*** *7(8):3027-3034.*
29. M. Ghamali, ***S. Chtita***, R. Hmamouchi, A. Adad, M. Bouachrine and T. Lakhlifi, The inhibitory activity of aldose reductase of flavonoids compounds. Combining DFT and QSAR calculations, ***Journal of Taibah University for Science, 2016,*** *10(4):534-542.*
30. ***S. Chtita***, M. Ghamali, R. Hmamouchi, B. Elidrissi, M. Bourass, M. Bouachrine and T. Lakhlifi, Research of Antileishmanial activities against Promastigotes and Amastigotes form of parasites drug research using Quantitative Structure Activity Relationship (QSAR) Analysis of Acridines derivatives, ***Advances in Physical Chemistry, 2016***, *1-16.*
    http://dx.doi.org/10.1155/2016/5137289
31. R. Hmamouchi, M. Larif, ***S. Chtita***, M. Bouachrine and T. Lakhlifi, Density Functional Theory Based Quantitative Structure-Activity Relationship Study of Cycloguanil Derivatives Acting as Plasmodium falciparum, ***Moroccan Journal of Chemistry, 2016***, *4(4):1061-1075.*

Publications 2017

32. M. Ghamali, ***S. Chtita***, A. Ousaa, B. Elidrissi, M. Bouachrine and T. Lakhlifi, QSAR analysis of the toxicity of phenols and thiophenols using MLR and ANN, ***Journal of Taibah University for Science, 2017,*** *11(1):1-10.*
33. ***S. Chtita***, M. Ghamali, M. Larif, R. Hmamouchi, M. Bouachrine and T. Lakhlifi, Quantitative structure–activity relationship studies of anticancer activity for Isatin (1H-

indole-2,3-dione) derivatives based on density functional theory with electronic and topological descriptors, ***International Journal of Quantitative Structure-Property Relationships (IJQSPR), 2017,*** *2(2):90-115.*

34. A. Ousaa, B. Elidrissi, M. Ghamali, ***S. Chtita***, A. Aouidate, M. Bouachrine and T. Lakhlifi, QSTR analysis and combining DFT of the toxicity of heterogeneous phenols, ***Journal of Materials and Environmental Science***, ***2017***, *8 (2):476-484.*
35. A.I. Taourati, M. Ghamali, ***S. Chtita***, H. Zaki, M. Benlyass, F. Guenoun, T. Lakhlifi and M. Bouachrine, QSAR Studies of the Inhibitory Activity of a Series of Substituted Indole and Derivatives Againt Isoprenylcysteine Carboxyl Methyltransferase (Icmt), ***International Journal of Pharmaceutical Science Invention***, ***2017***, *6(1):6-13.*
36. M. Ghamali, ***S. Chtita,*** A. Aouidate, A. Ghaleb, M. Bouachrine and T. Lakhlifi, Combining DFT and QSAR computation to predict the interaction of flavonoids with the GABA (A) receptor using electronic and topological descriptors, ***Journal of Taibah University for Science, 2017***, *11(3):422-433.*
37. A. Aouidate, A. Ghaleb, M. Ghamali, ***S. Chtita***, M. Choukrad, A. Sbai, M. Bouachrine and T. Lakhlifi, Combined 3D-QSAR and Molecular Docking Study on 7,8 dialkyl-1,3-diaminopyrrolo-[3,2-f] Quinazoline Series Compounds to understand the binding mechanism of DHFR Inhibitor, ***Journal of Molecular Structure, 2017***, *11(39):319-327.*
38. A. Aouidate, A. Ghaleb, M. Ghamali, ***S. Chtita***, M. Choukrad, A. Sbai, M. Bouachrine and T. Lakhlifi, QSAR studies on PIM1 and PIM2 inhibitors using statistical methods: A rustic strategy to screen for 5-(1H-indol-5-yl)-1,3,4-thiadiazol analogues and predict their PIM inhibitory activity, ***Chemistry Central Journal***, **2017**, *11(41):1-10*.
39. B. Elidrissi, A. Ousaa, M. Ghamali, ***S. Chtita***, M.A. Ajana, M. Bouachrine and T. Lakhlifi, Study of 5-hydroxy-6-oxo-1,6-dihydropyrimidine-4-carboxamide derivatives as HIV-1 integrase inhibitors using QSAR and DFT calculations, ***International Journal of Quantitative Structure-Property Relationship, (IJQSPR), 2017, article accepté***.
40. R. Hmamouchi, M. Larif, ***S. Chtita***, M. Bouachrine and T. Lakhlifi, Application of artificial neural networks on a series of anilids molecules based on the density functional theory (DFT), ***Moroccan Journal of Chemistry, 2017***, ***article accepté***.
41. A. Belhassan, ***S. Chtita***, T. Lakhlifi and Mohammed Bouachrine, QSPR study of the retention/release property of odorant molecules in water using statistical methods, ***Orbital: The Electronic Journal of Chemistry, 2017, article accepté***.
42. A. Belhassan, ***S. Chtita***, T. Lakhlifi and M. Bouachrine, QSPR study of the retention/release property of odorant molecules in pectin gels, ***Journal of Taibah University for Science, 2017, Journal of Taibah University for Science, article accepté***.
https://doi.org/10.1016/j.jtusci.2017.05.004
43. B. Elidrissi, A. Ousaa, M. Ghamali, ***S. Chtita***, M.A. Ajana, M. Bouachrine and T. Lakhlifi, QSPR and DFT studies on the Melting point of carbocyclic nitroaromatic compounds, ***article soumis***.
44. A. Belhassan, ***S. Chtita***, T. Lakhlifi and Mohammed Bouachrine, QSPR study of the retention/release property of odorant molecules in dairy gels using statistical methods, ***article soumis***.
45. A. Aouidate, A. Ghaleb, M. Ghamali, ***S. Chtita***, A. Ousaa, M. Choukrad, A. Sbai, M. Bouachrine and T. Lakhlifi, Combined 3D-QSAR and Molecular Docking Study on

derivatives of 7-Azaindole as inhibitors of Trk A: a strategic design in novel anticancer agents, ***article soumis***.

46. A. Belhassan, ***S. Chtita***, T. Lakhlifi and Mohammed Bouachrine, QSRR study of linear retention indices for volatile compounds using statistical methods, ***article soumis***.

Livre

S. Chtita, M. Bouachrine and T. Lakhlifi, Book: Modeling of Organic Heterocyclic Molecules using QSAR/QSPR Analysis, ***NOOR Publishing, Editor: Aelgargouh, 1st edition, ISBN: 978-3-330-84483-4, 2017***.

Communications orales

1. ***S. Chtita***, "Étude de la relation quantitative structure-activité (QSAR) des dérivés de l'imidazo [1,2-a] pyrazine sur les lignées cellulaires tumorales humaines à l'aide des méthodes statistiques et les réseaux de neurones", ***2ème édition des journées « Jeunes Chercheurs » JJC 02,*** *Faculté des sciences Kenitra, 20-21 décembre* ***2013.***
2. ***S. Chtita***, "Étude de dérivés de MK801 pour les antagonistes non compétitifs du récepteur NMDA, en combinant les résultats DFT et QSAR", ***3ème édition doctorales 2014,*** *Faculté des sciences Rabat, 6-8 Février* ***2014.***
3. M. Ghamali, ***S. Chtita***, A. Adad, R. Hmamouchi, M. Bouachrine et T. Lakhlifi, "Biological activity of molecules based on benzylpiperidine inhibitors of human acetylcholinesterase : Predicting by Combining DFT and QSAR calculations", ***1ère journée sur l'impact de la pollution des eaux, de l'air et du sol sur la population de la région du Gharb-Chrarda-Beni-Hssen (JIPG-I)***, *Chambre de Commerce, d'Industrie et de Services de Kénitra, 15 Mars* ***2014***.
4. ***S. Chtita***, "Étude de dérivés de MK801 pour les antagonistes non compétitifs du récepteur NMDA, en combinant les résultats DFT et QSAR", ***La 1ère rencontre internationale de chimie moléculaire, chimiométrie et applications (RICMCA-2014),*** *Faculté des sciences et techniques Béni-Mellal, 29-30 Mai* ***2014***.
5. ***S. Chtita***, "Antileishmanial activities of several substituted acridines : QSAR Studies using electronic descriptors", ***3ème édition des journées « Jeunes Chercheurs » JJC 03***, *Faculté des sciences Rabat, 20-22 Novembre* ***2014***.
6. ***S. Chtita***, "Activité anti-tumorale pour des dérivés d'acridines : Etudes QSAR avec des descripteurs électroniques", ***1er congrès international : substances naturelles & modélisation,*** *Faculté poly-disciplinaire Taza, 15-16 décembre* ***2014***.
7. ***S. Chtita***, M. Ghamali, M. Larif, R. Hmamouchi, M. Bouachrine et T. Lakhlifi, "QSAR studies of anticancer activity for Isatin (1H-indole-2,3-dione) based on density functional theory using Statistical Methods", ***2ème édition des journées doctorales JDOC'15,*** *Faculté des sciences et techniques Béni-Mellal, 26-28 mars* ***2015***.
8. ***S. Chtita***, "QSAR studies of Antileishmanial activities of acridines derivatives", ***4ème édition des journées « Jeunes Chercheurs » JJC 04,*** *Faculté des sciences El-Jadida, 19-20 Novembre* ***2015***.

9. ***S. Chtita***, M. Bouachrine et T. Lakhlifi, "Les approches de base et les applications des méthodes QSAR/QSPR", ***2^{éme} édition des journées doctoriales,*** *Faculté des sciences et techniques Er-Rachidia, 14-16 Avril* ***2016****.*
10. ***S. Chtita***, M. Ghamali, M. Larif, M. Bouachrine et T. Lakhlifi, "Etudes QSAR de l'activité Antileishmanienne des dérivés de l'acridine", ***2^{éme} édition des journées doctoriales,*** *Faculté des sciences et techniques Er-Rachidia, 14-16 Avril* ***2016****.*
11. B. Elidrissi, A. Oussa, M. Ghamali, ***S. Chtita***, A. Ajana, M. Bouachrine and T. Lakhlifi, "The biological activity of Pyrazinecarboxamides derivatives as an herbicidal agent : Combinig DFT and QSAR studies", ***1ère édition des journées doctorales****, Faculté des Sciences de Meknès, 2-3 Juin* ***2016****.*

Communications affichées

1. ***S. Chtita***, "Étude de la relation quantitative structure-activité (QSAR) des dérivés de l'imidazo[1,2-a] pyrazine sur les lignées cellulaires tumorales humaines à l'aide des méthodes statistiques et les réseaux de neurons", ***3ème édition doctorales 2014,*** *Faculté des sciences Rabat, 6-8 Février* ***2014.***
2. ***S. Chtita***, "QSAR Studies of MK801 derivatives for noncompetitive antagonists of NMDA using electronic and topological descriptors: Model, validation and influencing factors", ***3ème édition doctorales 2014,*** *Faculté des sciences Rabat, 6-8 Février* ***2014.***
3. ***S. Chtita***, "QSAR studies of imidazo [1,2-a] pyrazine derivatives on human tumor cell-lines using Statistical Methods and Neural Network", ***La 1ère rencontre internationale de chimie moléculaire, chimiométrie et applications (RICMCA-2014),*** *Faculté des sciences et techniques Béni-Mellal, 29-30 Mai* ***2014****.*
4. ***S. Chtita***, "Studies of two different cancer cell-lines activities (MDAMB-231 and SK-N-SH) of imidazo[1,2-a] pyrazine derivatives by combining DFT and QSAR results", ***La 4ème école de chimie quantique "Capzeo2014",*** *Faculté des sciences Rabat, 9-12 Juin* ***2014***
5. ***S. Chtita***, "DFT-based QSAR Studies of MK801 derivatives for noncompetitive antagonists of NMDA using electronic and topological descriptors : Model, validation and influencing factors", ***La 4ème école de chimie quantique "Capzeo2014",*** *Faculté des sciences Rabat, 9-12 Juin* ***2014****.*
6. M. Ghamali, ***S. Chtita***, A. Adad, R. Hmamouchi, M. Bouachrine et T. Lakhlifi, "Activité biologique des dérivés de la benzylpipéridine inhibteurs de l'acétycholnestérase humaine. Prédiction en utilisant les méthodes DFT et RQSA", ***1ère édition des journées doctorales sciences et techniques de Beni Mellal,*** *Faculté des sciences et techniques Béni-Mellal, 10-11 Juin* ***2014****.*
7. ***S. Chtita***, "3D-QSAR study of (1,3-benzothiazol-2-yl) amino-9-(10H)-acridinone derivatives as antiproliferative towards human monocytes agents", ***3ème édition des journées « Jeunes Chercheurs » JJC 03,*** *Faculté des sciences Rabat, 20-22 Novembre* ***2014****.*
8. ***S. Chtita***, "3D-QSAR study of (1,3-benzothiazol-2-yl) amino-9-(10H)-acridinone derivatives as antiproliferative towards human monocytes agents", ***1er congrès international : substances naturelles & modélisation****, Faculté Poly-disciplinaire Taza, 15-16 Décembre* ***2014****.*

9. ***S. Chtita***, M. Ghamali, M. Larif, R. Hmamouchi, M. Bouachrine et T. Lakhlifi, "QSPR studies of 9-aniliioacridine derivatives for their drug binding to DNA propriety based on density functional theory using Statistical Methods : Model, validation and influencing factors", ***2ème édition des journées doctorales JDOC'15,*** *Faculté des sciences et techniques Béni-Mellal, 26-28 Mars* ***2015.***
10. A. Ousaa, B. Elidrissi, M. Ghamali, ***S. Chtita***, M. Bouachrine et T. Lakhlifi, "Activité biologique des phénols dérivatives : Prédiction en utilisant les méthodes DFT et RQSA", ***2ème édition des journées doctorales JDOC'15***, *Faculté des sciences et techniques Béni-Mellal, 26-28 Mars* ***2015.***
11. M. Ghamali, ***S. Chtita***, A. Adad, R. Hmamouchi, M. Bouachrine et T. Lakhlifi, "Combining DFT and QSAR results for predicting the cytotoxicity of a series of orthoalkyl substituted 4-X-phenols", ***1ère édition des journées doctorales sciences et techniques***, *Faculté des sciences et techniques Er-Rachidia,* 28-29 Mai ***2015***.
12. ***S. Chtita***, "QSAR studies of anticancer activity for Isatin (1H-indole-2,3-dione) based on density functional theory (DFT) using Statistical Methods", ***4ème édition des journées « Jeunes Chercheurs » JJC 04,*** *Faculté des sciences El-Jadida, 19-20 Novembre* ***2015***.
13. ***S. Chtita***, M. Ghamali, M. Larif, R. Hmamouchi, M. Bouachrine et T. Lakhlifi, "QSPR studies of 9-aniliioacridine derivatives for their drug binding to DNA propriety based on density functional theory using Statistical Methods: Model, validation and influencing factors"**.** ***2éme édition des journées doctoriales,*** *Faculté des sciences et techniques Er-Rachidia, 14-16 Avril* ***2016****.*
14. ***S. Chtita***, M. Bouachrine et T. Lakhlifi, "QSAR studies of anticancer activity for Isatin (1H-indole-2, 3-dione) derivatives based on density functional theory with electronic and topological descriptors", ***2éme édition des journées doctoriales,*** *Faculté des sciences et techniques Er-Rachidia, 14-16 Avril* ***2016****.*
15. A. Ousaa, B. Elidrissi, M. Ghamali, ***S. Chtita***, M. Bouachrine et T. Lakhlifi, "QSAR analysis of the toxicity of heterogeneous phenol derivatives : Combining DFT and QSAR studies", ***1ère édition des journées doctorales***, *Faculté des Sciences de Meknès, 2-3 Juin* ***2016****.*
16. M. Ghamali, ***S. Chtita***, A. Adad, R. Hmamouchi, M. Bouachrine et T. Lakhlifi, "Combining DFT and QSAR computation for predicting the soil sorption coefficients of substituted phenols and anilines", ***1ère édition des journées doctorales***, *Faculté des Sciences de Meknès, 2-3 Juin* ***2016****.*

Résumé

L'expérience est un moyen direct pour obtenir des données de l'activité/propriété des composés organiques. Une telle expérience peut être déficiente en termes d'exigence de grands organismes expérimentaux, coûte beaucoup d'argent et prenant beaucoup de temps, en plus de la différence entre les valeurs mesurées par différents chercheurs selon les conditions expérimentales. Par conséquent, il serait impossible que l'expérience fournisse les valeurs des activités de tous les composés organiques. Il est donc crucial d'utiliser des méthodes théoriques pour compenser les inconvénients de l'expérience et pour prédire les données (activités ou propriétés) exactes des composés.

Le développement significatif de l'informatique ainsi que des études théoriques de la chimie quantique permettent aux chercheurs d'obtenir des paramètres physicochimiques et quantiques plus précis des composés en un temps plus court.

Les paramètres structuraux ainsi que l'introduction de la relation quantitative structure-activité (ou propriété) RQSA/RQSP peuvent augmenter l'interprétation et prédire l'activité/propriété pour de nouveaux composés organiques. Ceci est utilisé pour examiner la relation entre des descripteurs moléculaires d'un ensemble de composés et leurs activités biologiques ou propriété physicochimique.

Dans ce travail, l'auteur se concentrera sur l'étude de « l'activité antileishmanienne », « la propriété d'association des médicaments avec l'ADN », « l'activité anticancéreuse », et « l'activité antagoniste vis-à-vis du récepteur NMDA » pour certains dérivés organiques hétérocycliques, tels que l'acridine, l'isatine et la dizocilpine (MK801).

Les calculs de la chimique quantique en utilisant la théorie de la fonctionnelle de la densité DFT avec la fonctionnelle hybride de Becke à trois paramètres en combinaison avec l'énergie d'échange de Becke B3 et de l'énergie de corrélation de Lee-Young-Parr LYP ont été effectués pour calculer les paramètres électroniques et quantiques des composés étudiés.

Divers descripteurs moléculaires sont calculés avec les logiciels Gaussian, ChemSketch, Marvin Sketch et ChemOffice. L'ensemble de données sont soumis à des études statistiques : l'analyse en composantes principales ACP, la régression linéaire multiple RLM, la régression non linéaire multiple RNLM, la régression par les moindres carrés partiels PLS et les réseaux de neurones artificiels RNA. Les modèles obtenus, linéaires et non linéaires, ont été validés selon les cinq principes établis par *l'organisation de coopération et de développement économiques* (OCDE). Le domaine d'applicabilité des modèles est étudié à l'aide du diagramme de William pour détecter les composés extrêmes.

Pour appliquer les modèles développés avec succès en vue de prédire des acticités/propriétés de nouveaux composés, des validations rigoureuses ont été utilisées. Les effets des différents descripteurs sur les activités/propriétés sont décrits et utilisés pour examiner et proposer de nouveaux composés avec des valeurs d'effet (activité ou propriété) plus importantes.

Abstract

The experiment is a straight way to obtain the activity/propriety data of organic compounds. Such experiment may be deficient in terms of requiring various sample organs, costing much money, taking much time as well different measured values used by different researchers. Consequently, it would be almost impossible for these experiments to provide the activity values of all organic compounds. Hence, it is crucial to use theoretical methods to make up for the disadvantages of the experiment and to predict the exact data of compounds.

The significant development of computer science as well as the theoretical quantum of chemical studies enables researchers to get more precise physicochemical and quantum parameters of compounds in a shorter time.

The structural parameters along with the introduction of the quantitative structure activity/propriety relationship QSAR/QSPR methods can increase the interpretability and predict the activity/propriety of new organic compounds. This is used to examine the relationship between molecular descriptors of a set of compounds and their biological activity or physicochemical propriety.

Therefore, the author will elaborate, in this dissertation, on the antileishmanial activity, DNA drug binding proprieties, anticancer activity and antagonistic activity against the NMDA receptor of some organic heterocyclic, such as Acridine, Isatin, and Dizocilpine (MK801) derivatives.

Quantum chemical calculation use density functional theory, DFT, with Becke's three-parameters hybrid function, B3, and Lee-Young-Parr, LYP, exchange correlation functional methods. These methods are performed on the studied compounds and are used to calculate the electronic and quantum chemical parameters.

A variety of molecular descriptors are computed with Gaussian, ACD/ChemSketch, Marvin Sketch, and ChemOffice programs. The datasets are subject to multivariate statistical analyses, i.e. principal components analysis PCA, multiple linear regression MLR, multiple nonlinear regression MNLR, partial least squares PLS, and artificial neural network ANN. Both obtained linear and nonlinear models are proposed and validated according to the principles that are set up by the *Organization for Economic Co-operation and Development* (OECD). The applicability domain of models is investigated using William's plot to detect outlier and outside compounds.

To successfully apply the developed models in order to predict new compounds, rigorous validations have been used in this direction. The effects of different descriptors in the activities/proprieties are described and used to examine and form new compounds with larger effect values (activity or propriety).

Table de matières

Introduction générale

L'identification et la découverte, aussi tôt et de manière aussi fiable que possible, de nouvelles molécules susceptibles de devenir des médicaments, représente un objectif principal dans le domaine de la recherche pharmaceutique, elle constitue un enjeu majeur pour les années à venir. Le coût, le temps nécessaire et même la disponibilité des laboratoires équipés pour la réalisation des synthèses et des tests rendent le processus particulièrement difficile.

En moyenne, pour 10 000 molécules synthétisées et testées une molécule qui arrive sur le marché en tant que médicament innovant. De plus, le développement d'un médicament demande généralement entre 10 et 15 ans de recherche. Il s'agit en effet de trouver une molécule qui doit à la fois présenter des propriétés thérapeutiques particulières, et posséder le minimum d'effets secondaires indésirables. Le prix de revient d'un médicament est essentiellement dû à ces synthèses longues, coûteuses et finalement inutiles. Par conséquent, l'industrie pharmaceutique s'oriente vers de nouvelles méthodes de recherche, qui consistent à prédire les propriétés et les activités des molécules avant même que celles-ci ne soient synthétisées.

Dans les dernières années, le recours à des technologies permettant de synthétiser un très grand nombre de molécules simultanément et de tester leurs actions sur des cibles thérapeutiques a donné des résultats très attirants. C'est l'objectif principal des études des relations quantitatives structure-activité RQSA, et des relations quantitatives structure-propriété RQSP. Ces études se basent essentiellement sur la recherche de similitudes entre molécules dans de grandes bases de données de molécules existantes dont les activités ou les propriétés sont connues. La découverte d'une telle relation permet de prédire les activités et les propriétés des nouveaux composés, et, par conséquent, de guider les synthèses de nouvelles molécules, sans avoir à les réaliser. Les relations entre les structures des molécules et leurs activités ou propriétés sont généralement établies à l'aide de méthodes de modélisation moléculaire et des méthodes statistiques. Les techniques usuelles reposent sur la caractérisation des molécules par un ensemble de descripteurs, nombres réels mesurés ou calculés à partir des structures moléculaires. Il est alors possible d'établir une relation entre ces descripteurs et la grandeur modélisée.

Le développement continu de la chimie hétérocyclique désormais incontournable en synthèse organique, l'hétérocycle constitue le squelette de base dans une grande variété de composés d'intérêt chimique, biologique, pharmacologique et industriel. Les hétérocycles contenant un atome d'azote, tels que: les quinoléines, les dizocilpines, les pyrimidines, les acridines, les phénothiazines, les indoles..., se trouvent dans de nombreux produits naturels et

sont parmi les éléments les plus présents dans le domaine de la chimie médicinale. La recherche de nouveaux hétérocycles azotés à potentiel biologique est cruciale pour le développement de nouveaux composés répondent à une demande toujours croissante de molécules originales.

Le manuscrit de cette thèse est articulé sur cinq chapitres :

- Le premier chapitre est consacré à une étude bibliographique sur les approches de base, les méthodologies, le développement, les techniques de validation et les applications des méthodes RQSA/RQSP. Une description des différents outils nécessaires à la mise en œuvre de ces méthodes sera ainsi détaillée (les descripteurs, les méthodes statistiques, les principes OCDE de validité des modèles...).

Dans les autres chapitres, nous présenterons et nous discuterons les résultats obtenus pour les études RQSA/RQSP que nous avons effectué au cours de ces quatre années et demie. La stratégie générale adoptée pour élaborer les modèles RQSA/RQSP dans ces études est celle qui utilise un grand nombre de descripteurs en se basant sur la nature des molécules de nos bases de données et les différents mécanismes possibles pour expliquer l'activité ou la propriété étudiée et en respectant tous les critères concernant un modèle RQSA/RQSP fiable, robuste et prédictif. La plupart des modèles proposés ont une bonne interprétation statistique.

- Dans le deuxième chapitre nous effectuerons une étude RQSA de deux types d'activité Antileishmanienne pour une série de 60 dérivés d'acridine à l'aide de divers types des descripteurs moléculaires.
- Dans le troisième chapitre nous effectuerons une étude RQSP de la constante d'association avec l'ADN pour une série de 31 dérivés d'acridine à l'aide de 12 descripteurs moléculaires.
- Dans le quatrième chapitre nous effectuerons une étude RQSA de l'activité anticancéreuse pour une série de 40 dérivés de l'Isatine à l'aide de 14 descripteurs moléculaires.
- Dans le cinquième chapitre nous effectuerons une étude RQSA de l'activité antagoniste vis-à-vis du récepteur NMDA pour une série de 48 dérivés de MK801 à l'aide de 16 descripteurs moléculaires.

Enfin, nous terminerons par une conclusion générale et les perspectives envisagées pour ce travail, et à la fin de ce manuscrit, une annexe est consacrée à la description des méthodes de la chimie quantique utilisées pour l'optimisation des structures moléculaires.

Chapitre 1: Approches de base, développement, validation et application des méthodes RQSA/RQSP

Le développement de nouvelles techniques de modélisation a permis la mise en place de nombreuses méthodes RQSP (en anglais QSPR : *Quantitative Structure Property Relationships*) et RQSA (en anglais QSAR : *Quantitative Structure-Activity Relationships*) ; elles reposent pour la plupart sur « la recherche d'une relation entre un ensemble de nombres réels, appelés descripteurs moléculaires, et la propriété ou l'activité que l'on souhaite prédire ». Ces méthodes permettent de justifier les données expérimentales disponibles et de prédire les propriétés/activités pour des nouveaux composés ou des composés pour lesquels les données expérimentales ne sont pas disponibles.

Dans ce chapitre, une étude bibliographique sur les différentes méthodologies RQSA/RQSP a été présentée, les différentes étapes de développement, de validation et d'application de ces méthodes sont aussi mises en œuvre.

1. Relations quantitatives structures activités/propriétés RQSA/RQSP

1.1. Historique

Il y a plus d'un siècle et demi, en 1863, *Cros* **[1]** a observé que le point d'ébullition et le point de fusion des alcanes augmente avec le nombre d'atomes de carbone et la masse moléculaire. Il a observé également une diminution de la solubilité dans l'eau des alcools avec l'augmentation du nombre d'atomes de carbone et la masse moléculaire, cela est considéré depuis comme la première formulation générale en RQSP.

Cinq ans après, en 1868, *Crum-Brown* et *Fraser* **[2]** postulèrent que « l'activité biologique d'une molécule est une fonction de sa constitution chimique ».

Quelques décennies plus tard, en 1893, *Richet* **[3]** a montré que la cytotoxicité de certains composés organiques était inversement proportionnelle à leur solubilité dans l'eau.

A la fin du 19ème siècle, *Meyer* en 1899 et *Overton* en 1901 **[4-6]**, ont indépendamment observé « une relation linéaire entre l'activité des narcotiques et leur coefficient de partage huile-eau ».

Six ans après, en 1907, *Fühner* et *Neubauer* **[7]** ont montré pour une série de narcotiques homologues, que l'activité augmentait en fonction de la progression géométrique de la série de composés, ceci montrant l'importance de la contribution d'additivité de groupements fonctionnels pour l'activité biologique.

En 1962, *Hansen* **[8]** a montré l'existence d'une corrélation entre la toxicité des acides benzoïques substitués et les constantes électroniques « σ » des substituants.

L'année 1964 est considérée comme le début des méthodes RQSA modernes. *Hansch* et *Fujita* ont établi les premières corrélations entre les propriétés physico-chimiques (log P, pKa,

paramètres stériques et électroniques) et l'activité biologique (activité enzymatique, pharmacologique), Ces méthodes seront appelées par la suite l'analyse de *Hansch* et l'analyse de *Free Wilson* **[9-10]**). Sept ans plus tard, *Hansch* et *Lien* ont réalisé une étude RQSA sur différentes familles d'antifongiques : benzoquinones, sels d'alkylpyridinium, imidazoles et phénols. Ils ont observé que quels que soient la famille et le champignon utilisé, l'activité antifongique dépend du coefficient de partage Eau-Octanol, expérimental ou calculé **[11]**.

Ces études ont été extrapolées aux techniques séparatives en corrélant les propriétés physico-chimiques des analytes avec les temps de rétention obtenus expérimentalement : c'est l'étude quantitative des relations structure temps de rétention noté RQSR **[12].**

Maintenant, des méthodes 3D comme l'étude CoMFA (*Comparative Molecular Field Analysis*) et CoMSIA (*Comparative Molecular Similarity Indices Analysis*) **[13-14]** permettent de traiter les relations structure-activité en trois dimensions, 3D-RQSA/RQSP.

1.2. Définition

Les méthodes RQSA/RQSP sont basées sur l'hypothèse que l'activité ou la propriété d'un composé chimique est liée à sa structure, plus précisément cette approche affirme que l'activité (ou la propriété) et la structure d'un composé chimique sont liées d'un certain algorithme mathématique, cela est basé sur le postulat de base « les composés chimiques similaires ont des activités similaires ». De plus, lorsque les paramètres moléculaires sont exprimés par des chiffres, on peut proposer une relation mathématique, ou relation quantitative structure activité/propriété, entre les deux.

Par définition, Une RQSA/RQSP est un modèle mathématique qui associe un ou plusieurs paramètres quantitatifs dérivés de la structure chimique, à une mesure quantitative d'une propriété ou d'une activité.

1.3. Principe

Le principe d'une étude RQSA/RQSP (Figure 1), consiste à trouver une relation mathématique reliant de manière quantitative une activité biologique, ou une propriété, mesurée pour une série de composés similaires dans les mêmes conditions expérimentales, avec des descripteurs moléculaires à l'aide des méthodes statistiques. L'objectif de ces études est d'analyser les données structurales afin de détecter les facteurs déterminants pour l'activité ou la propriété étudiée. Pour ce faire, différents types de méthodes statistiques peuvent être employées (voir plus loin : les méthodes statistiques).

L'expression mathématique obtenue peut alors être utilisée comme moyen prédictif de l'activité/propriété étudiée pour de nouvelles molécules ou des molécules pour lesquelles les données expérimentales ne sont pas disponibles.

Ceci peut être traduit par l'équation suivante :

$$\text{Activité/Propriété} = f(\text{descripteurs moléculaires})$$

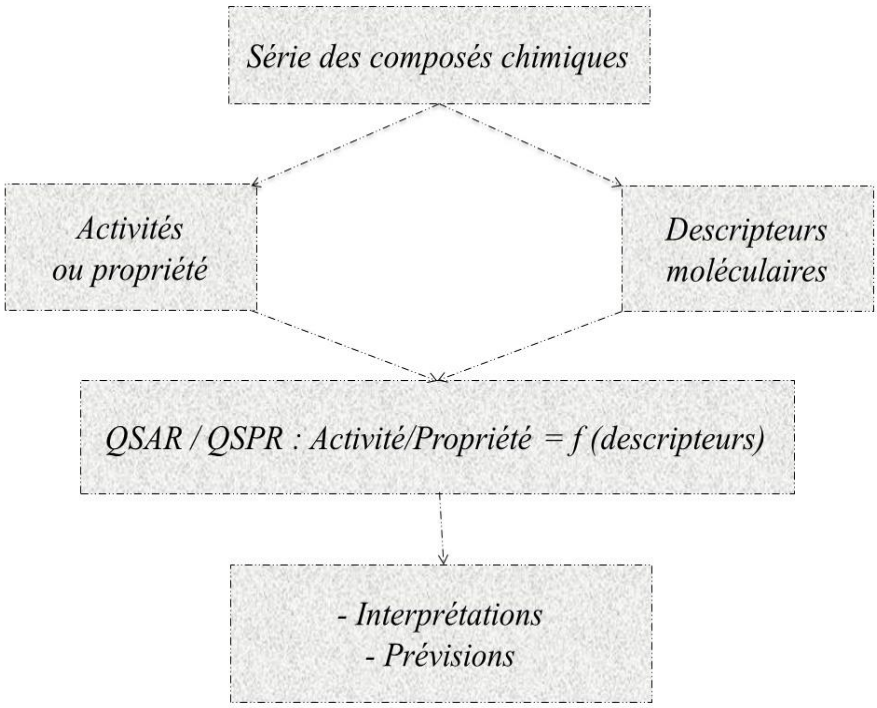


Figure 1 : Modèle de l'étude de relation structure activité/propriété

1.4. Stratégie globale

Le développement d'un modèle débute par la recherche du maximum possible des données expérimentales fiables. Ensuite, le développement d'une série de descripteurs qui caractérisent les structures moléculaires des composés de la base de données en vue de les relier à l'activité/propriété expérimentale étudiée. Une fois développé, le modèle doit être validé en termes de corrélation (sur le jeu de données d'entraînement). L'influence des composés du jeu d'entraînement sur le modèle (robustesse du modèle) est estimée par des méthodes de validation interne. Pour estimer le pouvoir prédictif du modèle, il est nécessaire de disposer de données expérimentales supplémentaires (jeu de données de validation externe) afin de déterminer la capacité du modèle à prédire ces valeurs. Enfin, pour tout modèle, il est important de savoir pour quel type de molécules il est utilisable ou non, c'est-à-dire connaître son domaine d'applicabilité.

Un modèle RQSA/RQSP relie, d'une manière qualitative ou quantitative, la structure des molécules à une activité ou propriété donnée. La stratégie de développement de tels modèles, en respectant les cinq règles mises en place par l'OCDE (*Organisation de Coopération et de Développement Economique*) pour la validation des modèles RQSA/RQSP (voir plus loin : les principes OECD de validité des modèles RQSA/RQSP), suit les étapes suivantes :

- Constituer la base de données structure – activité (ou propriété) à partir de mesures quantitatives, fiables et normalisées de l'activité (ou propriété) cible, pour chaque composé, et sélectionner des descripteurs moléculaires en relation avec l'activité (ou la propriété) cible afin de traduire de manière numérique la structure des molécules ;
- Diviser ce jeu de données en un jeu d'apprentissage et un jeu de test ;
- Construire des modèles à partir de jeu d'apprentissage à l'aide des méthodes statistiques ;
- Caractériser ces modèles par leurs indices statistiques et par une validation interne ;
- Valider les modèles avec le jeu de test et calculer leur indice de corrélation externe ;
- Répéter l'opération de division pour obtenir d'autres jeux d'apprentissage et de test, et répéter les mêmes étapes (facultative) ;
- Définir le domaine d'applicabilité des modèles proposés afin d'éviter des extrapolations hasardeuses ;
- Explorer et exploiter les modèles validés pour comprendre les mécanismes possibles et faire des prévisions d'activité/propriété pour de nouvelles molécules, si cela est possible.

2. Base de données

2.1. Source de données

Le choix de la base de données expérimentale initiale est une étape critique pour le développement des modèles RQSA/RQSP. Généralement, les composés testés ont deux origines possibles (dans la plupart des cas sont issues de la littérature), soit des produits de synthèse ou bien des produits d'extraction à partir de plantes. Quelle que soit son origine, il arrive qu'un échantillon ne soit pas pur mais corresponde à un mélange racémique. Le résultat du test d'un tel échantillon pose problème : il est impossible de savoir quelle est la contribution de chaque énantiomère dans l'activité observée. Les structures dont la propriété étudiée est mesurée sur un mélange racémique ne peuvent pas être utilisées dans les études RQSA/RQSP **[15]**.

Pour être de qualité, une base de données doit être composée de données expérimentales fiables, puisque les barres d'erreurs sur celles-ci se propageront dans le modèle final. Il est donc important de choisir des données présentant de faibles incertitudes afin de limiter les barres d'erreur expérimentales.

D'autre part, l'homogénéité des données est fondamentale. Si l'on veut comparer l'activité/propriété d'une série de molécules, il faut s'assurer, si cela est possible, qu'elle est le résultat de leur interaction avec une seule et même cible et plus précisément avec le même site actif, et l'activité doit être mesurée par un seul et même test, avec des conditions expérimentales identiques pour chaque molécule.

En fin, la diversité des structures est un facteur important dans la qualité des modèles construits, elle définit l'espace chimique que l'analyse va couvrir.

2.2. Homogénéité de la distribution des valeurs

L'homogénéité de la distribution des valeurs mesurées doit être contrôlée. En effet, la plupart des méthodes statistiques reposent sur l'hypothèse que la distribution des valeurs observées suit une loi normale. Il est donc nécessaire de contrôler la normalité de cette distribution. Il existe pour cela des tests statistiques de normalité mais la simple représentation des données sur un histogramme de distribution permet d'évaluer cette caractéristique.

Dans le cas défavorable, des transformations mathématiques (test de *Box-Cox* **[16]**, transformation de *Logit de x* **[17]**, ...) permettent, parfois, de retrouver une distribution normale sans que l'information contenue dans le jeu de données ne soit modifiée.

2.3. Les activités/propriétés ciblées dans ce travail

Les activités biologiques ou les propriétés physicochimiques des molécules peuvent être exprimées de manière quantitative par des chiffres (valeurs numériques). Ces valeurs sont habituellement exprimés sur une échelle logarithmique ou logarithmes inverses dans laquelle les grands nombres sont comprimés, rapprochés de 1 et facilement représentés, en revanche les nombres inférieurs à 1 sont dilatés et très vite renvoyés vers l'infini négatif.

Dans ce travail, nous avons présenté quatre études parmi d'autres que nous avons effectué au long du parcours de cette thèse :

- L'activité antileishmanienne.
- La constante d'association avec l'ADN.
- L'activité anticancéreuse.
- L'activité antagoniste vis-à-vis du récepteur NMDA.

3. Principes OECD de validité des modèles RQSA/RQSP

Afin d'accompagner le développement des méthodes alternatives, des règles ont été mises en place récemment par l'OCDE (*Organisation de Coopération et de Développement Economique*) pour la validation des modèles RQSA/RQSP **[18, 19]**. L'évaluation de chacun des cinq principes est une condition importante afin de proposer des modèles applicables dans le plan expérimental, ce qui était le but de cette thèse.

D'après l'OCDE, la validation des modèles repose sur cinq grands principes **[20]** :

- **Un effet défini** : la base de données (l'activité/propriété ciblée) doit être fiable et définie avec un protocole expérimental identifié.

- **Un algorithme non ambigu** : l'algorithme sur lequel repose le modèle doit garantir la transparence et la reproductibilité du calcul. Les prévisions issues d'un modèle utilisant un algorithme qui ne permet pas de vérifier son fonctionnement et dont les prévisions ne peuvent pas être reproduites peuvent difficilement être acceptées. Il convient notamment d'être prudent lorsque des méthodes non transparentes et difficilement reproductibles ont été utilisées pour élaborer le modèle RQSA/RQSP.

- **Un domaine d'applicabilité défini** : le domaine d'applicabilité et les limitations du modèle doivent être décrits pour permettre l'évaluation de l'espace chimique dans lequel on peut faire les prévisions avec confiance. Les méthodes les plus utilisées pour décrire le domaine d'applicabilité consistent à prendre en compte l'intervalle des descripteurs individuels et la présence de fragments structuraux dans l'ensemble de formation (training set). Les prévisions issues d'un modèle ne contenant aucune information sur le domaine d'applicabilité ne peuvent être acceptées.

- **Des mesures appropriées du degré d'ajustement, de la robustesse et de la prévisibilité** : ce principe traduit la nécessité d'une validation statistique du modèle. Des statistiques relatives à la validation interne (degré d'ajustement et robustesse) et la validation externe (prévisibilité) doivent être disponibles. Par exemple, les statistiques portant sur le modèle de régression peuvent être consignées en utilisant le coefficient de corrélation, le coefficient de corrélation de la validation croisée, l'erreur quadratique moyen du modèle... ; la validation externe doit avoir été effectuée dans le cadre d'une prévision des composés issus d'un ensemble externe (test set). Les statistiques relatives à la validation externe permettent d'estimer l'incertitude associée aux prévisions.

- **Une interprétation mécanistique, lorsque cela est possible** : une justification du lien de causalité entre les descripteurs moléculaires utilisés dans le modèle et l'effet prévu renforce la fiabilité des prévisions.

4. Descripteurs moléculaires

4.1. Introduction

Un descripteur moléculaire est un paramètre (une valeur numérique) propre à une structure chimique donnée. Ces valeurs peuvent être obtenues expérimentalement ou calculées à partir de la structure de la molécule. Les descripteurs calculés, permettent d'effectuer des prédictions sans avoir à synthétiser les molécules, ce qui est l'un des objectifs de la modélisation moléculaire.

Les descripteurs moléculaires jouent un rôle fondamental dans les études de la relation quantitative structure activité/propriété. Ils sont utilisés en tant que variables indépendantes pour prédire une variable dépendante (activité ou propriété).

L'utilisation des descripteurs moléculaires dans le développement de modèles RQSA/RQSP n'est pas une tâche aisée. Tout d'abord, un très grand nombre de descripteurs moléculaires, de différentes complexités et de conceptions diverses a été introduit au cours des dernières années. Ensuite, pendant ce temps, aucune règle stricte n'a été établie pour la sélection de descripteurs adaptés parmi le grand nombre de descripteurs disponibles. Ce choix a souvent été basé sur l'intuition chimique des chercheurs, ou en se pliant à la tradition. **[21]**

4.2. Types de descripteurs

L'importance du nombre des descripteurs (plus de 6000 descripteurs répertoriés **[22]**) pouvant décrire une molécule rend toute classification ou présentation de ces descripteurs non exhaustive.

Dans ce qui suit, nous allons présenter que les descripteurs moléculaires les plus utilisés et ceux qui ont été utilisés dans l'ensemble de nos travaux, en commençant par les descripteurs les plus simples, qui nécessitent peu de connaissances sur la structure moléculaire, Nous verrons ensuite comment les progrès de la modélisation moléculaire ont permis d'accéder à la structure 3D de la molécule, et de calculer des descripteurs à partir de cette structure.

Historiquement, deux grands schémas pour la classification des descripteurs moléculaires ont été établis : l'un en fonction de leur origine (constitutionnel, topologique, géométrique, quantique, thermodynamique...), et un autre sur leur dimensionnalité (1D, 2D, 3D ou 4D) **[23]**.

4.2.1. Les descripteurs 1-D

Ces descripteurs sont calculés à partir de la formule brute de la molécule à l'aide de la composition moléculaire, c'est-à-dire les atomes qui la constituent, et ils représentent des propriétés générales telles que : les pourcentages massiques des atomes, la masse molaire, le poids moléculaire...

Dans nos travaux nous avons utilisé :

- Le poids moléculaire, noté MW (appelé aussi le poids de formule), mesuré en daltons (Da). C'est la somme des poids atomiques des différents atomes constituant la molécule. Il est utilisé dans l'étude de transport dont la diffusion et le mode de fonctionnement. Les composés avec des poids plus élevés sont moins susceptibles d'être absorbés et donc ne peuvent pas atteindre le site d'action. Ainsi, essayer de garder des poids moléculaires aussi bas que possible devrait être l'objectif pour établir un médicament **[24]**. Pour les médicaments délivrés par voie orale le poids moléculaire doit être inférieur ou égal à 500 daltons (optimum autour de 300 daltons) **[25]**.

- Le pourcentage massique, défini par la formule suivante :

$$\%\ \text{massique} = \frac{\text{la masse de l'élément dans une mole du composé}}{\text{la masse d'une mole du composé}} * 100$$

Les descripteurs 1D sont facile à calculer, leurs valeurs sont précises, essentielles et interviennent régulièrement dans les modèles RQSA/RQSP, mais ils ne permettent pas de distinguer les isomères de constitution et ne permettent pas d'élaborer des modèles plus complexes, c'est-à-dire, si on développe des modèles avec ce type de descripteurs seulement, on aura des problèmes au niveau de l'interprétation des mécanismes d'interaction mis en jeu pour l'activité ou la propriété étudiée **[26]**. Or, pour la grande majorité des propriétés, la position d'un substituant modifie la valeur de celle-ci, les descripteurs 1D sont, dans de tels cas, défaillants. Il faut alors recourir à d'autres classes de descripteurs.

4.2.2. Les descripteurs 2-D

Les descripteurs 2D sont obtenus à partir de la structure plane de la molécule. Dans cette catégorie on trouve principalement les descripteurs topologiques.

Les descripteurs topologiques (ou indices topologiques) décrivent les connectivités atomiques dans la molécule. Ce sont des descripteurs plus "sophistiqués" qui n'ont pas forcément un sens chimique évident mais ils contiennent en leur sein des informations sur la taille globale du système, sa forme globale et ses ramifications **[21]**. Le principe est de trouver une valeur différente pour chaque squelette moléculaire.

Ces descripteurs sont faciles à calculer, leurs valeurs sont généralement précises, Ils interviennent souvent dans les modèles. Ils sont issus de la théorie des graphes développée par *Euler* en 1736 **[27]** ; cette théorie est appliquée à la table de connectivité, qui est une représentation compacte de la connectivité interatomique au sein de la molécule.

Un graphe est un ensemble de point, certains reliés par des lignes ; il permet de représenter la topologie de la molécule sans se soucier de la géométrie spatiale exacte de cette dernière **[28]**.

Cette théorie est inventée à partir de quelques lois simples : deux points adjacents sont reliés par des lignes et deux lignes avec un point commun sont adjacentes. L'ordre ou le degré d'un point est le nombre de lignes reliées à ce point. Un pas est un enchaînement de points et de lignes adjacentes débutant par un point et se terminant par un point. Un chemin est un pas dans lequel aucun point n'est utilisé plus d'une fois. D'un point de vue moléculaire, un point représente un atome et une ligne une liaison covalente. Les chemins sont caractéristiques de l'architecture de l'ensemble des atomes constituant la molécule. Les atomes d'hydrogène sont exclus du graphe pour simplifier les calculs.

Une molécule de l'acridine, par exemple, peut être représentée soit à partir de sa formule développée soit par un graphe lié (Figure 2).

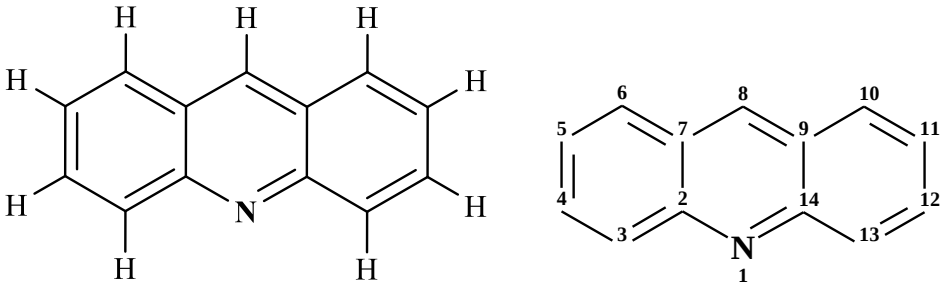


Figure 2 : Formule développée et graphe lié de l'acridine.

A partir de ce graphe lié, une matrice de connectivité notée C et une matrice de distance D peuvent être établies. Les éléments C_{ij} de la matrice C prennent la valeur de 1 si les points i et j sont adjacents et la valeur de 0 si les points i et j ne sont pas adjacents, et les éléments D_{ij} de la matrice D sont égaux à la longueur du plus petit chemin joignant i et j. Les matrices de connectivité et de distance l'acridine sont représentées sur la figure ci-dessous (Figure 3).

$C =$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	1	0	0	0	0	0	0	0	0	0	0	0	1
2	1	0	1	0	0	0	1	0	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	0	0	0	0	0	0
4	0	0	1	0	1	0	0	0	0	0	0	0	0	0
5	0	0	0	1	0	1	0	0	0	0	0	0	0	0
6	0	0	0	0	1	0	1	0	0	0	0	0	0	0
7	0	1	0	0	0	1	0	1	0	0	0	0	0	0
8	0	0	0	0	0	0	1	0	1	0	0	0	0	0
9	0	0	0	0	0	0	0	1	0	1	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0	1	0	0	0
11	0	0	0	0	0	0	0	0	0	1	0	1	0	0
12	0	0	0	0	0	0	0	0	0	0	1	0	1	0
13	0	0	0	0	0	0	0	0	0	0	0	1	0	1
14	1	0	0	0	0	0	0	0	1	0	0	0	1	0

$D =$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	1	2	3	4	3	2	3	2	3	4	3	2	1
2	1	0	1	2	3	2	1	2	3	4	5	4	3	2
3	2	1	0	1	2	3	2	3	4	5	6	5	4	3
4	3	2	1	0	1	2	3	4	5	6	7	6	5	4
5	4	3	2	1	0	1	2	3	4	5	6	7	6	5
6	3	2	3	2	1	0	1	2	3	4	5	6	5	4
7	2	1	2	3	2	1	0	1	2	3	4	5	4	3
8	3	2	3	4	3	2	1	0	1	2	3	4	3	2
9	2	3	4	5	4	3	2	1	0	1	2	3	2	1
10	3	4	5	6	5	4	3	2	1	0	1	2	3	2
11	4	5	6	7	6	5	4	3	2	1	0	1	2	3
12	3	4	5	6	7	6	5	4	3	2	1	0	1	2
13	2	3	4	5	6	5	4	3	2	3	2	1	0	1
14	1	2	3	4	5	4	3	2	1	2	3	2	1	0

Figure 3 : Matrices de connectivité C et de distance D de l'acridine.

Ces matrices permettent de calculer de nombreux indices. Parmi ceux, que nous avons utilisé dans nos travaux, on trouve :

- L'indice de Balaban [29], noté (J), est l'un des indices topologiques les plus importants. Il est défini par la formule suivante :

$$J = \frac{q}{\mu + 1} \sum_{i,j}^{N,N} (\delta_i \delta_j)^{-\frac{1}{2}}$$

Avec : i et j sont les atomes voisins (autres que l'hydrogène), N est le nombre d'atomes de la molécule, δ_i et δ_j sont les connectivités des atomes i et j, q est le nombre des liaisons, et μ est le nombre des cycles.

La détermination de l'indice de Balaban suppose, dans une première étape, de compter tous les atomes dont la connectivité est égale à 1 ; la valeur obtenue est élevée au carré. Dans

une deuxième étape, ces atomes sont éliminés du graphe et on répète ce procédé jusqu'à la disparition de tous les atomes. L'indice de Balaban est obtenu en additionnant les valeurs issues de chacune des étapes. Cet indice décrit très bien le degré de ramification des molécules non cycliques.

- L'indice de Wiener [30], noté (W), permet de caractériser le volume moléculaire et la ramification d'une molécule, l'indice de Wiener est égal à la demi-somme de toutes les distances topologiques entre les différentes paires d'atomes de la molécule. Il est défini par la formule suivante :

$$W = \frac{1}{2}\sum_{i,j}^{N,N} d_{i,j}$$

Avec : N est le nombre des atomes (autres que l'hydrogène) de la molécule ; $d_{i,j}$ est le plus petit nombre de liaisons séparant les deux atomes i et j.

- L'indice topologique moléculaire [39], noté (MTI), en plus des matrices de connectivité (C) et de distance (D), le MTI utilise une troisième matrice de valence (V) appelé aussi le vecteur de degrés des sommets. Ce vecteur se calcule de la façon suivante : le résultat de la somme des matrices (C) et (D) est lui-même multiplié par la matrice (V). La somme des éléments résultants de ce produit donne l'indice.

L'indice MTI est défini par la formule suivante :

$$MTI = \sum_{i}^{N} E_i \text{ avec } E_i = V_i(C_i + D_i) \text{ ou encore } E = V(C + D)$$

La matrice de valence (V) peut être remplacée par : V = (1, 1, 1, ..., 1) (C) [31].

Avec : C est la matrice de connectivité ; D est la matrice de distance D ; V est la matrice de valence.

- La somme des degrés de valence [32], notée (SVD), est la somme de tous les degrés de valence de la molécule représentée par un graphe, le degré d'un point correspondant au nombre de lignes se terminant par ce point. Ce paramètre dépend donc principalement de la ramification de la molécule.

- Le coefficient de forme [33], noté (I), caractérise la forme globale de la molécule. Il est calculé à partir du carré de la matrice des distances D^2.

- La superficie de la surface polaire [34], notée (PSA), en ($Å^2$), est un paramètre très utile pour la prédiction des propriétés du transport des médicaments. Elle est définie comme la somme des surfaces des atomes polaires (habituellement, l'oxygène, l'azote, le soufre, le chlore et l'hydrogène ci-jointes) dans une molécule.

- La connectivité totale de valence, notée (TVCon), définit la connectivité de valence des hétéroatomes constituant la molécule.

- La connectivité totale, notée (TCon), définit la connectivité des hétéroatomes constituant la molécule.

Les descripteurs 2-D sont employés pour obtenir des modèles RQSA/RQSP plus simples, mais leur défaillance, comme pour les descripteurs 1-D, qu'il ne permet pas la bonne interprétation des mécanismes d'interaction mis en jeu pour l'activité/propriété étudiée.

4.2.3. Les descripteurs 3-D

Ce type de descripteurs nécessitent une conformation 3D de la molécule ; Ils sont évalués à partir des positions relatives de leurs atomes dans l'espace et décrivent des caractéristiques plus complexes ; leurs calculs nécessitent donc de connaître, le plus souvent par « modélisation moléculaire empirique » ou « ab-initio », la géométrie 3D de la molécule. La plupart de ces descripteurs s'avèrent relativement coûteux en temps de calcul, mais apportent davantage d'informations et sont nécessaires à la modélisation de propriétés ou d'activités qui dépendent de la structure 3D. On distingue plusieurs familles de descripteurs 3D :

4.2.3.1. Les descripteurs géométriques

Ces descripteurs peuvent être obtenus expérimentalement ou par modélisation moléculaire, empirique ou ab-initio. Ils sont basés sur l'arrangement spatial des atomes constituant la molécule et sont définis par les coordonnées des noyaux atomiques et la grosseur de la molécule représentée. Ces descripteurs incluent des informations sur la surface moléculaire obtenue par les aires de Van Der Waals et leur superposition **[35]**. Les volumes moléculaires peuvent être obtenus par les volumes de Van Der Waals **[36]**. Parmi ceux qui sont les plus importants, que nous avons utilisé dans nos travaux, on trouve :

- Le volume moléculaire, noté MV, en cm^3, est défini par la formule suivante :

$$MV = \frac{MW}{d}$$

Avec : MW est le poids moléculaire et d la densité.

- Le nombre de liaisons rotatives *:* La liaison rotative est définie comme une liaison d'un composé non cyclique, associée à un atome non lourd (qui n'est pas l'hydrogène). Les liaisons CN (amide) ne sont pas considérées en raison de leur barrière d'énergie de rotation élevée. Le nombre de liaisons rotatives, noté N_{ROT}, est utilisé pour identifier la flexibilité de la molécule, il a été montré pour être un descripteur de très bonne biodisponibilité orale de médicaments, et pour qu'une structure chimique puisse présenter de bons effets inhibiteurs et

être similaire aux médicaments, selon la règle de Lipinski, il faut que le nombre de liaisons rotatives soit inférieur ou égal à 5 **[25]**.

- La surface de Van Der Waals, noté SVDW, est décrite comme résultant de l'ensemble des surfaces atomiques définies par le rayon de Van Der Waals de chaque atome composant la molécule (Figure 4). Plus cette surface est grande et plus importantes sont les possibilités d'interactions.

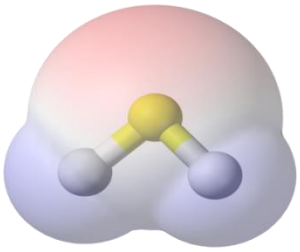


Figure 4 : surface (enveloppe) de Van Der Waals

- Le volume de Van Der Waals, noté VVDW, est le volume occupé par l'enveloppe de Van Der Waals, ces valeurs numériques dépendent de la méthode de calcul et des rayons de Van Der Waals (RVDW) atomique. Ces derniers déterminent la position la plus favorable d'un atome par rapport à un autre, la distance adéquate où les potentiels répulsifs et attractifs des atomes s'équilibrent. Ils sont particulièrement utilisés pour modéliser comment les molécules organiques "s'approchent" les unes des autres.

4.2.3.2. Les descripteurs physico-chimiques

Les descripteurs physicochimiques, (ou indices physicochimiques) certains d'entre eux reflètent la composition moléculaire du composé (le nombre et le type d'atomes et de liaisons présents dans la molécule, le nombre de cycle, les propriétés donneur/accepteur de liaison H, cation, anion, etc.…) **[37]**. D'autres représentent le caractère hydrophile ou lipophile de la molécule généralement évalué à partir du coefficient de partage Octanol/eau représenté par le log P **[38]**. Parmi ceux que nous avons utilisé dans nos travaux, on trouve :

- Le coefficient de partage Octanol/Eau : Le transport, le passage à travers les membranes et l'activité pharmacologique d'une molécule peuvent être conditionnés par son partage entre une phase lipidique et une phase aqueuse, c'est-à-dire son caractère hydrophile. Celui-ci peut être quantifié par le coefficient de partage Octanol-Eau, noté (log P), qui mesure la solubilité différentielle d'un soluté dans ces deux solvants non miscibles **[39]**. C'est une mesure importante pour l'identification de la similarité médicamenteuse, selon la règle de Lipinski, les médicaments délivrés par voie orale doivent avoir des valeurs de log P supérieures ou égales à -2 et inférieures ou égales à 5) **[25]**. Il est défini par la formule suivante :

$$\log P = \log \frac{[Octanol]}{[H_2O]}$$

[Octanol] et [H_2O] sont les concentrations du soluté dans l'Octanol et l'eau.

Les composés qui ont les valeurs de log P > 0 sont dites lipophiles, et les composés qui ont les valeurs de log P <0 sont dites hydrophiles. Si le Log P est positif et très élevé, cela exprime le fait que la molécule est plus soluble dans l'Octanol que dans l'eau, ce qui reflète son caractère lipophile, et inversement, si le Log P est négatif cela signifie que la molécule est hydrophile. Un Log P nul signifie que la molécule est aussi soluble dans un solvant que dans l'autre.

- La réfractivité moléculaire, notée (MR), en m^3/mol, est le volume de la substance absorbée par mole de cette substance. Elle est définie par Lorentz-Lorenz **[40]** par la formule suivante :

$$MR = \frac{n^2 - 1}{n^2 + 2}\frac{MW}{d} = \frac{n^2 - 1}{n^2 + 2} MV$$

Où : MW est le poids moléculaire ; d est la densité ; n est l'indice de réfraction ; MV est le volume molaire.

La réfractivité moléculaire est également proportionnelle à la polarisabilité αe, par la relation suivante **[41]** : MR =4/3π NA αe

Où : NA est le nombre d'Avogadro qui est, le nombre de molécules dans une mole de substance, NA = 6.022 10^{23}.

- L'indice de réfraction, noté n, est défini par la formule de Lorentz suivante **[40]** :

$$n = \sqrt{\frac{2\,MR + MW}{MV - MR}}$$

- La polarisabilité, notée (αe), en (m^3), est l'aptitude à la déformation du nuage électronique de la molécule sous l'influence d'un champ électrique uniforme. C'est l'un des paramètres qui traduisent les propriétés moléculaires liées à l'hydrophobie et par conséquent aux activités biologiques **[42-44]**. Elle est calculée à partir de la réfractivité molaire ou du volume molaire comme suit :

$$\alpha e = 0.3964308 \times MR = 0.3964308 \times \frac{n^2 - 1}{n^2 + 2} MV$$

- La densité, notée (d), en (kg/m^3), est liée à la masse et la taille de la molécule. C'est le rapport du poids moléculaire MW au volume moléculaire MV :

$$d = \frac{MW}{MV}$$

L'augmentation de la pression augmente la densité, alors que l'augmentation de la température diminue généralement la densité, mais il y a des exceptions (par exemple, l'eau).

- Le parachor, noté (Pc), en (m^3), est un paramètre moléculaire qui définit la nature stérique de la molécule. Il est calculé à partir de la densité d, le poids moléculaire MW et la surface de tension γ par l'équation suivante **[45]** :

$$Pc = \left(\frac{MW}{d}\right)\gamma^{\frac{1}{4}}$$

- La surface de tension, notée γ, en dyne/cm, est calculée à partir du parachor Pc et du poids moléculaire MW selon la formule suivante **[45]** :

$$\gamma = (\frac{Pc}{MW})^4$$

- Le nombre de donneurs de liaisons hydrogène, noté (NHD), calcule le nombre de donneurs de liaison hydrogène dans la molécule. Il s'agit du nombre d'atomes possédant une case quantique vide et contenant un hydrogène acide, c'est-à-dire un atome d'hydrogène lié à un hétéroatome (comme dans les amines, alcools, thiols).

- Le nombre d'accepteurs de liaisons hydrogène, noté (NHA), calcule le nombre d'accepteurs de liaison hydrogène dans la molécule. Il s'agit du nombre d'atomes possédant des doublets non liants (azote, oxygène ou fluor) et capable de se lier par liaisons hydrogène à d'autres molécules.

Selon la « règle des cinq de Lipinski » **[25]**, lors de l'identification de la similarité médicamenteuse, les médicaments délivrés par voie orale doivent avoir un nombre d'accepteurs de liaisons hydrogène (NHA) inférieur ou égal à 10 (optimum d'environ 5) et un nombre de donneurs de liaisons hydrogène (NHD) inférieur ou égal à 5 (optimum de 2).

4.2.3.3. Les descripteurs quantiques/électroniques

Ces descripteurs caractérisent la distribution de charge des molécules (polarité des molécules) mais aussi les paramètres de la chimie quantique qui, pour être calculés de manière fiable, nécessitent des calculs plus sophistiqués.

Les approches de la chimie quantique nous donnent accès à des informations supplémentaires telles que des données structurales, énergétiques, électroniques et spectroscopiques des systèmes étudiés **[46]**.

Les structures étudiées dans ce travail ont été optimisées en utilisant la base 6-31G(d) de la fonctionnelle B3LYP qui est une sorte de la méthode de théorie de la fonctionnelle de la densité DFT, détaillée dans l'annexe. Le calcul des descripteurs commence par le dessin des molécules dans le logiciel GaussView 3.09 **[47]** puis l'ouverture de ces structures dans le programme Gaussian 03 et ensuite l'exécution de l'optimisation (les calculs). A la fin de ces calculs, des propriétés électroniques seront obtenues.

Parmi ces propriétés, que nous avons utilisé dans nos travaux, on trouve :

- L'énergie totale *:* Pour une molécule isolée à l'état fondamental, l'énergie totale calculée, notée Et, mesurée en eV, peut être utilisée comme descripteur moléculaire quantique. Cette énergie approximative a été calculée pour une conformation optimisée de la géométrie la plus stable dont la structure d'énergie est minimale. Les expressions de l'énergie totale de l'état fondamental d'un système sont décrites en détails dans l'annexe.

- Le moment dipolaire, noté μ, mesuré en debye (D), mesure la polarité nette moléculaire, et décrit la séparation de charge dans une molécule où la densité d'électrons est partagée inégalement entre les atomes. L'existence d'un moment dipolaire dans une molécule a son origine dans la différence d'électronégativité entre les atomes. La densité électronique est plus élevée au voisinage de l'atome le plus électronégatif. Ceci entraîne une dissymétrie dans la répartition des électrons de liaison. Ainsi, plus le moment dipolaire d'une molécule est élevé, plus la dissymétrie dans la molécule est importante.

- Les énergies des orbitales frontières, jouent un rôle majeur dans de nombreuses réactions chimique et dans les mécanismes réactionnels. Les énergies de ces orbitales sont des paramètres très populaires dans la chimie quantique et dans les études RQSA/RQSP **[48, 49]** :

L'énergie HOMO, notée E_{HOMO}, mesurée en eV, est le niveau d'énergie le plus élevé dans la molécule qui contient des électrons, il est directement lié au potentiel d'ionisation. Lorsqu'une molécule agit comme une base de Lewis (un doublet d'électrons donneur) dans la formation d'une liaison, les électrons sont alimentés à partir de cette orbite. Il mesure la nucléophilie d'une molécule et caractérise la susceptibilité de la molécule à l'attaque par des électrophiles **[50]**.

L'énergie LUMO, notée E_{LUMO}, mesurée en eV, est le niveau d'énergie le plus bas dans la molécule qui ne contient pas d'électrons, il est directement lié à l'affinité d'électron. Lorsqu'une molécule agit comme un acide de Lewis (un doublet d'électrons accepteur) dans la formation de liaisons, des doublets d'électrons entrants sont reçus dans cette orbite. Il mesure l'électrophilicité d'une molécule et caractérise la susceptibilité de la molécule à l'attaque par les nucléophiles **[50]**.

- Le Gap énergétique, ou l'écart HOMO-LUMO, noté E_{gap}, mesuré en eV, traduit l'énergie comprise entre l'orbitale moléculaire la plus haute occupée et la plus basse vacante, C'est un indice de stabilité important. Cette différence d'énergie sert de mesure de l'excitabilité d'une molécule. Ainsi, plus l'intervalle d'énergie sera faible et plus la molécule pourra interagir avec l'environnement.

Un grand écart HOMO-LUMO implique une grande stabilité pour la molécule dans le sens de sa faible réactivité dans les réactions chimiques, et de même, un faible écart implique une grande réactivité de la molécule **[51, 52]**. L'écart HOMO-LUMO a également été utilisé comme une approximation de la plus faible énergie d'excitation de la molécule **[53]**.

- L'électronégativité, notée χ, mesurée en eV, est l'opposé du potentiel chimique qui mesure la tendance du nuage électronique à s'échapper de la molécule, c'est un paramètre global du système moléculaire égal à la pente de l'énergie en fonction du nombre d'électrons N à potentiel externe v(r) constant telle que définie par Parr et Mulliken **[54, 55]** :

$$\chi = -\mu = -\left(\frac{\partial E}{\partial N}\right)_{v(r)} = -\frac{(E_{LUMO} + E_{HOMO})}{2}$$

Avec :$E = E[N, v(r)]$

- La dureté et la mollesse, notée η, et son inverse la mollesse, notée S, peuvent être obtenues à partie de la première dérivée du potentiel chimique **[56, 57]** :

$$\eta = \left(\frac{\partial \mu}{\partial N}\right)_{v(r)} = \left(\frac{\partial^2 E}{\partial N^2}\right)_{v(r)} = \frac{1}{S} = \frac{E_{LUMO} - E_{HOMO}}{2}$$

$$\mu = \left(\frac{\partial E}{\partial N}\right)_{v(r)} = -\frac{PI + AE}{2} = -\chi$$

Où : PI est le potentiel d'ionisation et AE est l''affinité électronique

La définition qualitative de la dureté est étroitement liée à la polarisabilité, car une diminution de l'écart d'énergie conduit généralement à faciliter la polarisation de la molécule. Ce descripteur conduit à une distinction entre les vitesses de réaction à différents sites dans la molécule **[52, 58]**.

- L'indice d'électrophilicité, notée ω, utilisée pour caractériser la capacité d'une molécule à engendrer un transfert d'électron, elle est calculée selon la formule suivante **[59]** :

$$\omega = \frac{\chi^2}{2\eta}$$

- La charge négative totale, notée TNC, est la somme de toutes les charges négatives des atomes dans une molécule, elle est calculée selon la formule suivante :

$$TNC = \sum_i q_i^-$$

Où : q_i^- sont les charges négatives atomiques nettes.

- La longueur d'onde du maximum d'absorption et l'énergie d'activation : Les transitions électroniques correspondent au passage des électrons des orbitales moléculaires liantes et non

liantes remplies, vers des orbitales moléculaires anti-liantes non remplies. L'absorption provient de ce passage (transition) entre les deux orbitales moléculaires. La longueur d'onde d'absorption et l'énergie d'activation, notée E_a, (de transition) dépendent de la nature des orbitales mises en jeu.

- La force d'oscillation, notée $f_{s.o.}$, est la probabilité pour que la transition électronique soit permise.

4.2.3.4. Les descripteurs thermodynamiques

Ce sont des descripteurs peu utilisés dans les études RQSA/RQSP. Ils peuvent être exprimés par la fonction de partition Q de la molécule utilisée en thermodynamique statistique ainsi que de ses dérivées **[60–62]**. Cette fonction décrit la façon avec laquelle l'énergie d'un système de molécules est répartie parmi les individus moléculaires. Sa valeur dépend du poids moléculaire, de la température, du volume moléculaire, des distances inter nucléaires, des mouvements moléculaires et des forces intermoléculaires. La fonction de partition est le point le plus commode entre les propriétés microscopiques des molécules indépendantes (niveaux d'énergie, moments d'inertie) avec les propriétés macroscopiques (point de fusion, point d'ébullition, entropie). L'expression de cette fonction s'écrite :

$$Q = Q_{éle} * Q_{trans} * Q_{rot} * Q_{vibr}$$

Avec : $Q_{éle} = \sum_i g_i \exp(-\varepsilon_i/kT)$ Fonction de partition électronique ; $Q_{vibr} = \prod_i (1 - \exp(-h\nu_i/kT))^{-1}$ Fonction de partition de vibration ; $Q_{rot} = (8\pi^2(8\pi^3 ABC)^{\frac{1}{2}}(kT)^{\frac{3}{2}})/\sigma h^3$ Fonction de partition de rotation ; $Q_{trans} = ((2\pi mkT)^{3/2}V)/h^3$ Fonction de partition de translation.

T est la température en °K, k est la constante de Boltzmann k=1.38 10^{-23} J/K, gi représente la dégénérescence du niveau d'énergieε_i, h est la constante de Planck h = 6.62 10^{-34} J.s, ν_i les fréquences de vibration de la molécule, σ est le degré de symétrie, A, B et C sont les trois moments d'inertie par rapport aux axes x, y et z, m la masse de la particule, V le volume de la molécule.

Les descripteurs thermodynamiques que nous avons utilisé dans nos travaux sont :

- Le point d'ébullition, en °K, est la température à laquelle les phases liquide et gazeuse d'une substance pure sont en équilibre à une pression donnée, c'est la température à laquelle la substance change d'état du liquide au gaz à une pression donnée. Le point d'ébullition normal est le point d'ébullition à la pression atmosphérique normale (1.013 10^{05} kPa).

En termes d'interactions intermoléculaires, le point d'ébullition représente la température à laquelle les molécules possèdent l'énergie thermique suffisante pour surmonter les attractions

intermoléculaires liant les molécules dans le liquide (par exemple des liaisons hydrogène, les attractions dipôle-dipôle...).

Le point d'ébullition d'un composé pur augmente avec la taille, la ramification de la molécule, et avec la présence des liaisons hydrogènes et des interactions dipôle-dipôle.

- La constante de Henry, notée K_H, est issue de la loi de Henry qui établit une relation entre la pression partielle pi d'un corps pur gazeux et sa concentration c dans un solvant K_H=Pi/c. La constante de Henry traduit la volatilité de la molécule. La constante de Henry dépend du soluté, du solvant, et de la température. Une molécule est considérée comme volatile si sa constante est supérieure à 1.10^{-5} $Pa.m^3.mol^{-1}$.

- La température critique, notée Tc, est la température au-dessus de laquelle les phases liquide et gazeuse d'une substance n'existent pas, autrement dit, la température au-dessus de laquelle un gaz ne peut être liquéfié par une augmentation de la pression. Lorsqu'on rapproche de la température critique, les propriétés des phases gazeuse et liquide, deviennent les mêmes et se transforment en une seule phase fluide **[63]**.

- La pression critique, notée Pc, est la pression de vapeur à la température critique et au volume critique, c'est la pression minimale pour liquéfier un gaz à la température critique **[64]**.

- La chaleur de formation (appelé aussi l'enthalpie), notée H, mesurée en KJ/mol, est l'énergie résultant de la formation d'une mole d'une substance à partir de ses éléments constitutifs à l'état standard (T à 298.15 °K et P à 1 atm).

- L'énergie libre de Gibbs, mesurée en KJ/mole est définie par la formule suivante :

$$G(p, T) = U + p\,V - TS \text{ ou encore : } G(p, T) = H - TS$$

Avec : U l'énergie interne (en J) ; p est la pression (en Pa) ; V est le volume (en m^3) ; T est la température (en °K) ; S est l'entropie (en J/°K) ; H est l'enthalpie (en J).

- Le point de fusion (Melting Point) **[65]**, est la température à laquelle la substance passe de l'état solide à liquide à la pression atmosphérique normale. La taille de la molécule et de sa symétrie augmente habituellement le point de fusion ; Cependant, contrairement au point d'ébullition, le point de fusion est relativement insensible à la pression. Les points de fusion sont souvent utilisés pour caractériser la pureté des composés organiques. Le point de fusion d'une substance pure est toujours supérieur au point de fusion de cette substance quand une petite quantité d'impureté est présente. Il est utilisé pour prédire la solubilité des composés. La formule pour calculer le point de fusion est :

$$T = \frac{\Delta H}{\Delta S}$$

Où : T est la température au point de fusion en °K ; ΔS est la variation d'entropie en J/°K ; ΔH est la variation d'enthalpie en J.

Liste des descripteurs utilisés dans nos travaux

Descripteurs		Type
Le pourcentage massique de l'azote	$N_{\%}$	**Constitutionnel**
Le pourcentage massique de l'hydrogène	$H_{\%}$	
Le pourcentage massique de l'oxygène	$O_{\%}$	
Le pourcentage massique du carbone	$C_{\%}$	
Le poids moléculaire	MW	
L'indice de Balaban	J	**Topologique**
L'indice de Wiener	WI	
L'indice topologique moléculaire	MTI	
La somme des degrés de valence	SVD	
Le coefficient de forme	I	
La superficie de la surface polaire	PSA	
La connectivité totale de valence	TVCon	
La connectivité totale	TCon	
Le coefficient de partage Octanol/Eau	Log P	**Physico-chimique**
La réfractivité molaire	MR	
La densité	d	
Le parachor	Pc	
La polarisabilité	αe	
Le nombre de donneurs de liaisons H	NHA	
Le nombre d'accepteurs de liaisons H	NHD	
La surface de tension	γ	
L'indice de réfraction	n	
Le volume moléculaire	MV	**Géométrique**
Le nombre de liaisons rotatives	N_{ROT}	
Le rayon de Van Der Waals	R_{DWV}	
Le volume Van Der Waals	V_{DWV}	
La surface de Van Der Waals	S_{DWV}	
L'énergie totale	Et	**Quantique**
L'énergie HOMO	E_{HOMO}	
L'énergie LUMO	E_{LUMO}	
Le Gap énergétique	E_{Gap}	
Le moment dipolaire	μ	
La dureté	η	
La mollesse	S	
L'électronégativité	χ	
L'indice d'électrophilicité	ω	
La charge négative totale	TNC	
La longueur d'onde du maximum d'absorption	λmax	
L'énergie d'activation	Ea	
La force d'oscillation	f s.o.	
L'énergie libre de Gibbs	G	**Thermodynamique**
La chaleur de formation	H°	
Le point d'ébullition	T	
La température critique	CT	
Le point de fusion	TB	
La pression critique	CP	
La constante de Henry	K_H	

4.2.4. Les descripteurs 4-D

Ils correspondent à la mesure des propriétés 3D (potentiel électrostatique, d'hydrophobicité, de liaison hydrogène...) d'une molécule en tout point de l'espace. Ils permettent d'avoir l'information sur la structure de la cible (protéine). On pourra ainsi distinguer les descripteurs 4D qui nécessitent un alignement de la molécule guidé par l'étude des complexes ligand-cible (ou, au moins, par des contraintes visant d'optimiser le recouvrement spatial des champs électriques et stériques des ligands, faute d'information sur le vrai mode de fixation dans la cible) avant d'être calculés. Ces descripteurs sont obtenus par le calcul des champs d'interactions moléculaires (CoMFA, CoMSIA) entre une molécule et une sonde représentée par une autre molécule (eau, amide, ...). **[13, 26, 66, 67]**

4.3. Logiciels de calcul des descripteurs moléculaires

Plusieurs logiciels sont disponibles pour faire les calculs des descripteurs, parmi ceux, que nous avons utilisé dans nos travaux, on trouve : Gaussian **[68]**, ChemOffice **[69]**, ChemSketch **[70]**, Marvin Sketch **[71]**. Mais il existe plusieurs autres logiciels comme : QSARIS **[72]**, Cerius2 **[73]**, Vol Surf **[74]**, Dragon **[75]**, ...

4.4. Sélection et réduction du nombre de descripteurs

Un grand nombre de descripteurs différents sont collectés pour la modélisation d'une grandeur donnée (activité ou propriété), car les facteurs déterminants du processus étudié ne sont *a priori* pas connus. Cependant, les descripteurs envisagés n'ont pas tous une influence significative sur la grandeur modélisée, et les variables ne sont pas toujours indépendantes. De plus, le nombre de descripteurs, c'est-à-dire la dimension de la base de données d'entée, détermine la dimension du vecteur des paramètres à ajuster. Si cette dimension est trop importante par rapport au nombre des observations (molécules) de la base d'apprentissage, le modèle risque d'être sur-ajusté à ces exemples, incapable de prédire la grandeur modélisée sur de nouvelles molécules et peut contenir des informations redondantes. Les descripteurs moléculaires employés doivent être porteurs de sens et interprétables d'un point de vue chimique. Et par conséquent, lorsque les descripteurs sélectionnés sont pertinents, ils offrent des idées sur les mécanismes, et les modèles RQSA/RQSP seront simples, transparents et compréhensibles **[76]**.

Il ne s'agit d'utiliser alors que le minimum de descripteurs pour expliquer la propriété ciblée, mais il ne faut pas avoir de perdre d'information. Comme Einstein a dit : "Tout devrait être fait aussi simple que possible, mais pas plus simple.". Dans une modélisation

RQSA/RQSP, ce principe d'Einstein signifie que le modèle doit avoir le moins de paramètres possibles tout en traduisant au mieux l'information contenue dans la propriété **[77]**.

Avant d'entamer le développement effectif des équations de régression RQSA/RQSP, il est hautement recommandé d'examiner la qualité statistique des données de départ, à la fois les données à corréler (variable dépendante) et les descripteurs utilisés dans la corrélation (variables indépendantes).

On distingue habituellement dans un tel prétraitement des données les analyses univariées des analyses bivariées [**78-83]**.

Dans l'analyse univariée, il est recommandé de vérifier la conformité des données à la distribution normale. Une précaution particulière doit être prise lors de la procédure de régression subséquente si les valeurs de la propriété étudiée, ou d'un descripteur, ne suivent pas la loi de Laplace-Gauss **[84]**.

Pour un ensemble de descripteurs différents, il est nécessaire d'effectuer une analyse des données bivariée, c'est-à-dire de calculer le coefficient de corrélation entre chacune des paires de l'ensemble des descripteurs. Si ce coefficient est statistiquement significatif (R>0,95), ces deux descripteurs sont considérés comme fortement corrélés et ne peuvent être utilisés simultanément lors de l'analyse RQSA/RQSP **[85]** et en pratique, ils seront alors enlevés dans le procédé de sélection. Ce type d'analyse est appelé l'analyse objective qui permet de réduire le nombre de descripteurs sans faire participer la variable dépendante (l'activité ou la propriété).

Les méthodes statistiques, qu'on discutera dans la partie suivante de ce chapitre, sont aussi utilisées pour l'élimination des descripteurs qui n'interviennent pas dans les modèles proposés, c'est-à-dire qui n'influencent pas l'activité ou la propriété étudiée dans ce qu'on appelle l'analyse subjective (spécialement l'analyse en composantes principales et la régression linéaire descendante).

Finalement, pour que les relations RQSA/RQSP ne soient pas statistiquement non significatives ou en cas d'erreur ponctuelles, il faut que le rapport composés/descripteurs doive être supérieur à 5 **[86, 87]**.

5. Méthodes statistiques

5.1. Définition

Par définition, la statistique est « la science dont l'objet est de recueillir, de traiter et d'analyser des données issues de l'observation de phénomènes dans lesquels le hasard intervient (phénomène aléatoire) ». Par conséquent, l'objectif principal de la statistique est de

maîtriser au mieux l'incertitude pour extraire des informations utiles des données, par l'intermédiaire de l'analyse des variations dans les observations. En outre, l'analyse des données est utilisée pour décrire, comprendre et gérer les phénomènes étudiés, faire des prévisions et prendre des décisions.

5.2. Domaines d'application

Les données étudiées peuvent être de toute nature, ce qui rend la statistique utile dans la plupart des disciplines : économie, sociologie, psychologie, agronomie, biologie, médecine, chimie, physique, géologie, sciences de l'ingénieur, sciences de l'information et de la communication, etc...

5.3. Méthodes statistiques

Faire de la statistique suppose que l'on étudie un ensemble d'objets équivalents sur lesquels on observe des caractéristiques appelées « variables ». Dans notre cas, les objets (ou individus) sont les molécules et les variables sont les descripteurs moléculaires précédemment décrits dans ce chapitre.

Après le recueil des descripteurs, la démarche statistique consiste à traiter et interpréter les informations recueillies sur ces molécules. Cette démarche comporte deux grandes classes : la statistique descriptive et la statistique décisionnelle ou prédictive.

5.3.1. La statistique descriptive

La statistique descriptive (appelée aussi l'analyse des données), a pour but d'extraire le maximum de l'information contenue dans les données d'une façon efficace, simple et compréhensible. Elle permet de résumer les caractéristiques essentielles du phénomène étudié et de suggérer des hypothèses pour des études plus sophistiquées. Elle utilise pour cela des représentations de données sous forme de graphiques, de tableaux et d'indicateurs statistiques. Elle est utilisée aussi pour diviser et classer les données dans des classes homogènes.

Dans l'ensemble de nos travaux, nous avons principalement utilisé l'analyse en composantes principales (ACP) comme technique pour l'analyse des données, et la méthode du partitionnement en k-moyennes (ou k-means en anglais) et la classification ascendante hiérarchique (CAH) pour la classification des données.

a. L'analyse en composantes principales

L'analyse en composantes principales (ou ACP) **[88-89]**, est une méthode très efficace d'analyse de données quantitatives utilisée pour réduire la dimension de l'espace de représentation des données. Les variables initiales sont remplacées par de nouvelles variables,

appelées composantes principales, deux à deux non corrélées, et telles que les projections des données sur ces composantes soient de variance maximale.

Considérons un ensemble de M observations, représentées chacune par N données. Ces observations forment un nuage de M points dans $\mathbb{R}^N$. Le principe de l'ACP est d'obtenir une représentation approchée des variables dans un sous-espace de dimension K plus faible, par projection sur des axes bien choisis ; ces axes principaux sont ceux qui maximisent l'inertie du nuage projeté. La maximisation de l'inertie permet de préserver au mieux la répartition des points. Par conséquent, les N composantes principales peuvent être représentées dans l'espace sous-tendu par ces axes, par une projection orthogonale des N vecteurs d'observations sur les K axes principaux.

Ces composantes peuvent être classées par ordre d'importance. Puisqu'elles sont des combinaisons linéaires des variables initiales, l'interprétation du rôle de chacune de ces composantes reste possible. Il suffit en effet de déterminer quels descripteurs d'origine leur sont le plus fortement corrélés. Les variables obtenues peuvent ensuite être utilisées en tant que nouvelles variables du modèle.

L'analyse en composantes principales est généralement utilisée pour visualiser et analyser rapidement les corrélations entre les variables et pour visualiser et analyser les observations initialement décrites par les variables sur un graphique à deux ou trois dimensions, construit de manière à ce que la dispersion entre les données soit aussi bien préservée que possible...

Les limites de l'ACP viennent du fait que c'est une méthode de projection, et que la perte d'information induite par la projection peut entraîner des interprétations erronées.

b. Classification des données

Les méthodes de classification, aussi appelées de partition des données, sont appliquées à l'analyse de bases de données et à la classification des composés. Ces méthodes permettent de grouper des objets (observations ou individus) dans des classes (clusters) de manière à ce que les objets appartenant au même cluster soient plus similaires entre eux qu'aux objets appartenant aux autres clusters et partagent des caractéristiques communes, qui correspondent le plus souvent à des critères de proximité (similarité informatique). En général, ces méthodes sont divisées en techniques non-hiérarchique et hiérarchique, les dernières étant elles-mêmes subdivisées en descendante ou ascendante. Les classifications non-hiérarchiques organisent les composés en un nombre initial défini de clusters disjoints, qui sont la plupart du temps effectués par des analyses du plus proche voisin dans l'espace des descripteurs **[90]**.

Dans l'ensemble de nos travaux, l'analyse de bases de données de composés chimiques s'est effectuée soit à l'aide de la méthode de k-means **[91]** qui représente l'approche non-hiérarchique la plus courante, soit par la méthode ascendante hiérarchique CAH **[92]**.

Dans la méthode de classification k-means, deux molécules sont inclues dans le même cluster si elles partagent un nombre spécifique minimum prédéfini de plus proches voisins **[90]**. Même si cette méthode a été très usuelle, l'approche CAH a démontré des résultats plus constants et des clusters plus homogènes **[93]**.

- La classification ascendante hiérarchique

La classification ascendante hiérarchique (CAH) est une méthode de classification qui consiste à regrouper une collection d'objets (individus) en groupes (sous-ensembles), de telle sorte que les objets au sein de chaque groupe sont liés les uns aux autres que les objets dans les différents groupes **[94]**. Le principe de cette méthode est simple, on commence par le calcul de la similarité entre les N objets, puis on regroupe les deux objets dont le regroupement minimise un critère d'agrégation donné, créant ainsi une classe comprenant ces deux objets. On calcule ensuite la similarité entre cette classe et les N-2 autres objets en utilisant le critère d'agrégation. Puis on regroupe les deux objets ou classes d'objets dont le regroupement minimise le critère d'agrégation. On continue ainsi jusqu'à ce que tous les objets soient regroupés.

Ces regroupements successifs produisent un arbre de classification, appelé dendrogramme, dont la racine correspond à la classe regroupant l'ensemble des individus. Ce dendrogramme représente une hiérarchie de partitions. On peut alors choisir une partition en tronquant l'arbre à un niveau donné, le niveau dépendant soit des contraintes de l'utilisateur (l'utilisateur sait combien de classes il veut obtenir), soit de critères plus objectifs.

- Le partitionnement k-moyennes

Le partitionnement k-moyennes (k-means) est une méthode non hiérarchique de classification qui peut être utilisée lorsque le nombre de groupes présents dans les objets ou les cas est connu. En général, la méthode k-means produit exactement k différents clusters.

Dans l'ensemble de nos travaux, ces deux méthodes de classification sont utilisées pour la division/le regroupement des données en « ensemble d'apprentissage » et « ensemble de test », le premier ensemble est utilisé pour la formation des modèles obtenus par les méthodes statistiques et le deuxième pour la validation externe. De telle sorte que dans chaque cluster/groupe des molécules obtenues, on choisit au hasard un composé pour l'ensemble de

test (test set) et les autres composés du même cluster pour l'ensemble de formation (training set) **[95]**. (Voir 1.3.2)

5.3.2. La statistique décisionnelle ou prédictive

Contrairement à la statistique descriptive, dans ce type des statistiques les probabilités jouent un rôle fondamental. Cette statistique a pour but de prendre des décisions et de faire des prévisions au vu des observations. En général, il faut pour cela proposer des modèles probabilistes du phénomène aléatoire étudié et savoir gérer les risques d'erreurs. Dans notre cas, il faut rechercher une relation approximative entre une activité ou propriété et plusieurs variables quantitatives (descripteurs moléculaires), la forme de cette relation peut être linéaire ou non linéaire.

Dans l'ensemble de nos travaux, nous avons utilisé la régression linéaire multiple MLR, la régression non linéaire multiple MNLR, la régression des moindres carrés partiels PLS et les réseaux de neurones artificiels ANN pour la construction des modèles RQSA/RQSP.

5.3.2.1. La régression linéaire multiple

La régression linéaire multiple MLR est l'une des méthodes de modélisation les plus populaires grâce à sa simplicité d'utilisation et facilité d'interprétation. L'avantage important de la régression linéaire multiple est qu'elle est très transparente, puisque l'algorithme est disponible, et que les prédictions peuvent être réalisées facilement. Dans la plupart de nos travaux, cette méthode a été utilisée aussi pour la sélection des descripteurs moléculaires utilisés dans les autres méthodes statistiques **[96]**.

La méthode MLR se base sur l'hypothèse que la propriété y dépend linéairement des différentes variables (les descripteurs) $x_1, x_2, \ldots x_i$, selon la relation :

$$y = a_0 + \sum_{i=1}^{n} a_i x_i$$

Avec : y est la variable dépendante (à expliquer ou à prédire) ; x_i sont les variables indépendantes (explicatives) ; n est le nombre de variables explicatives ; a_0 est la constante de l'équation du modèle ; a_i sont les coefficients de descripteurs dans l'équation du modèle ;

La taille de ces coefficients indique le degré d'influence des descripteurs moléculaires correspondants sur l'activité/propriété cible. Un coefficient positif indique que le descripteur moléculaire correspondant contribue positivement à la cible, tandis qu'un coefficient négatif suggère la contribution négative.

On distingue divers types de MLR, les plus utilisés sont :

– La MLR progressive ascendante, qui consiste à incorporer les variables au modèle une à une, en sélectionnant, à chaque étape, la variable dont la corrélation partielle avec la grandeur modélisée est la plus élevée. À l'inverse, lors de MLR progressive descendante, on débute la modélisation avec l'ensemble des descripteurs, en les éliminant un par un jusqu'à obtenir le meilleur jeu de composantes, c'est-à-dire l'obtention d'un modèle valide (voir la partie validation) ayant la bonne corrélation.

– La MLR pas à pas (Stepwise), est une combinaison des deux méthodes évoquées précédemment. Les variables sont incorporées une à une dans le modèle, par sélection progressive. Cependant, à chaque étape, on vérifie que les corrélations partielles des variables précédemment introduites sont encore significatives.

5.3.2.2. La régression non linéaire multiple

La régression non linéaire multiple MNLR est une méthode non linéaire (exponentielle, logarithmique, polynomiale, …) qui permet de déterminer le modèle mathématique qui permet d'expliquer non-linéairement au mieux la variabilité d'une propriété ou d'une activité y en fonction des descripteurs moléculaires. Dans l'ensemble de nos travaux nous avons utilisé le modèle polynomial en nous basant sur les descripteurs proposés par le modèle linéaire qui seront élevés à la puissance 2 selon l'équation suivante :

$$y = a_0 + \sum_{i=1}^{n} a_i x_i + b_i x_i^2$$

Avec : y est la variable dépendante (à expliquer ou à prédire) ; x_i sont les variables indépendantes (explicatives) ; i est le nombre de variables explicatives ; a_0 est la constante de l'équation du modèle ; a_i et b_i sont les coefficients de descripteurs dans l'équation du modèle ;

5.3.2.3. La régression des moindres carrés partiels PLS

La régression des moindres carrés partiels PLS, est une généralisation de la régression linéaire multiple, elle peut être utilisée lorsque le nombre de descripteurs est élevé et que ceux-ci sont fortement corrélées **[97, 98]**. Cette méthode utilise à la fois des principes de l'ACP et de la régression multilinéaire. Elle permet de trouver par une transformation linéaire, les axes qui représentent au mieux les données dans l'espace. En d'autres termes, cette méthode va permettre de trouver les axes qui expliquent au mieux la dispersion du nuage de points. Si les données sont représentées en fonction de n descripteurs, la PLS va donc permettre de trouver au maximum n axes classés en fonction de la variance qu'ils représentent. Cette méthode consiste à remplacer une matrice des données prédictives X

comprenant n lignes et m colonnes, par une nouvelle matrice, dérivée de X, qu'on désigne par T, comprenant le même nombre de lignes (molécules) que X, mais un nombre de colonnes k très inférieur à m. On impose, de plus, que les colonnes de la matrice T soient des combinaisons linéaires des variables d'origine. Sous forme matricielle, la relation s'écrit :

$$T = XW$$

Avec : $W(m * k)$ est la matrice des coefficients définissant les combinaisons linéaires ; T est la nouvelle matrice dont les colonnes forment des « variables artificielles », obtenues par combinaison linéaire des variables d'origine ;

Après cette transformation, la régression linéaire multiple est appliquée sur le tableau T à la place de X.

5.3.2.4. Les réseaux de neurones artificiels ANN

- Les neurones biologiques :

Le cerveau humain est constitué d'un très grand nombre de cellules nerveuses appelées neurones, environs 100 milliards, avec 1000 à 10000 synapses (connexions) par neurone **[99]**. Le neurone biologique (Figure 5) est une cellule nerveuse spécialisée dans le traitement de l'information (signaux électriques). Il est constitué de trois composantes principales :

Les dendrites : fines prolongations du corps cellulaire entourant celui-ci en une sorte de filet qui capte les oscillations et les informations issues d'autres cellules nerveuses et les transmettent au corps cellulaire.

Le corps cellulaire : qui a pour fonction de recevoir les excitations, les intégrer et les transmettre ou non. Il contient également le noyau qui assure la vie du neurone.

L'axone : Les axones conduisent les signaux électriques de la sortie d'un neurone vers l'entrée à un autre neurone. Le point de contact entre l'axone d'un neurone et la dendrite d'un autre neurone s'appelle la synapse **[100,101]**.

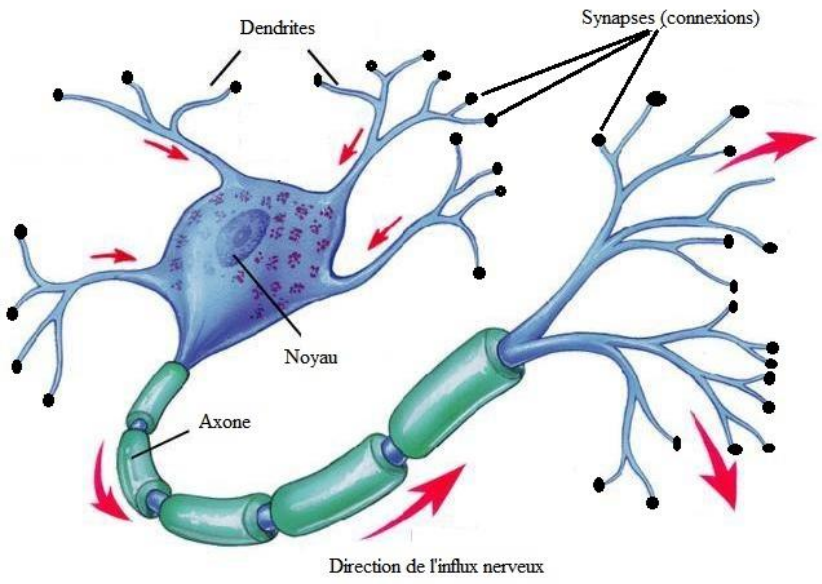


Figure 5 : Le neurone biologique

Au niveau du neurone se produit une intégration (sommation) des signaux reçus et si cette somme dépasse un certain seuil le neurone émet à son tour un signal électrique vers d'autres neurones. Ce signal peut renforcer ou diminuer l'activité des neurones qui le reçoivent selon que les synapses soient excitatrices ou inhibitrices **[102]**.

- Les réseaux de neurones artificiels ANNs

Historique :

Les réseaux de neurones étaient à l'origine une tentative de modélisation mathématique simplifiée des systèmes nerveux biologiques, initiée dès 1943 avec Mc-Culloch et Pitts qui inventent le premier neurone formel **[103, 104]**. Ce n'est qu'en 1958 qu'apparaît le premier réseau de neurones artificiels grâce aux travaux de Rosenblatt **[105]** qui a développé le modèle du Perceptron. Ce dernier est constitué d'une couche de neurones d'entrée appelée couche de perception (sert à recueillir les entrées) et d'une couche de neurones de sortie appelée couche de décision. Ce réseau est le premier système artificiel présentant la capacité d'apprendre par l'expérience. En 1960, Widrow et Hoff **[106]** ont proposé un modèle inspiré du perceptron, le modèle de l'Adaline (Adaptive Linear Element). Ce dernier sera, par la suite, le modèle de base des réseaux de neurones multicouches.

Néanmoins, en 1969, Minsky et Papert **[107]** démontrent dans leur livre « Perceptrons » les limites des réseaux de neurones à une seule couche, en particulier, l'impossibilité de traiter les problèmes non linéaires par ce modèle. Il faut attendre 1982 et les travaux de Hopfield **[108]** pour susciter à nouveaux l'intérêt des scientifiques en proposant les neurones associatifs. Dans le même temps, Werbos **[109]** conçoit l'algorithme de rétro-propagation de

l'erreur offrant un mécanisme d'apprentissage pour les réseaux multicouches de type Perceptron et qui permet d'entraîner les neurones des couches cachées. Cependant, cet algorithme ne deviendra connu qu'après 1986 grâce à Rumelhart **[110]**. Ce type de réseau est capable de résoudre des problèmes non linéaires. Toutefois, en 1984 c'est la découverte des cartes de Kohonen **[111]** avec un algorithme non supervisé basé sur l'auto-organisation et suivi une année plus tard par la machine de Boltzman.

Enfin en 1989 Moody et Darken **[112]** ont proposé le réseau à Fonctions de Base Radiales (RFR), connu sous l'appellation anglophone Radial Basis Function network (RBF).

Principe :

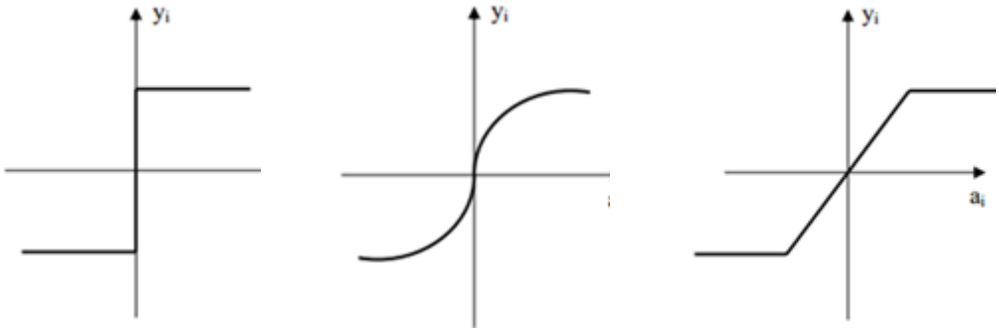
L'approche par les ANNs est analogue aux systèmes de neurones biologiques qui permettent de traiter et de transmettre des informations en faisant circuler des signaux électriques dans un réseau constitué d'axones. Chaque neurone artificiel est un processeur élémentaire. Il est donc avant tout un opérateur mathématique avec des « entrées » (variables de la fonction mathématique) et des « sorties » (valeurs de la fonction). L'intérêt des neurones réside dans les propriétés qui résultent de leur association en réseaux, c'est-à-dire de la composition des fonctions réalisées par chacun des neurones. Il reçoit un nombre variable d'entrées en provenance de neurones en amont ou des capteurs composant la machine dont il fait partie. A chacune de ses entrées est associé un poids (*wi*) représentatif de la force de la connexion. Chaque processeur élémentaire est doté d'une sortie unique, qui se ramifie ensuite pour alimenter un nombre variable de neurones en aval. Le neurone renvoie un signal de sortie si la somme pondérée des entrées dépasse un certain seuil.

Un réseau de neurones est constitué de multiples couches : une couche d'entrée représentée par les descripteurs moléculaires, une ou plusieurs couches cachées et une couche de sortie représentée par les propriétés à modéliser. Les neurones d'une couche sont interconnectés avec les neurones d'une couche voisine.

Chaque neurone de la couche cachée réalise des opérations de sommations pondérées, à l'issu desquelles le neurone peut être activé ou non. Chaque neurone de la couche d'entrée est relié par des synapses à chacun des neurones de la couche cachée, et au niveau de ces synapses virtuelles, se trouvent des poids (*wi*) permettant de moduler l'importance relative de chacun des descripteurs. La couche de sortie compte autant de neurones que de propriétés modélisées. Dans notre cas une seule propriété/activité a été modélisée. Pendant la phase d'apprentissage du modèle par un réseau de neurones, les molécules sont présentées une par une aux neurones de la couche d'entrée. Les poids (*wi*) associées aux neurones d'entrée sont

ajustés itérativement, afin de minimiser l'erreur entre la propriété calculée et la propriété expérimentale.

La sortie d'un neurone, donc, dépend de l'entrée du neurone et de sa fonction de transfert. Il existe essentiellement trois types de fonction de transfert qui sont les fonctions à seuil, les fonctions sigmoïdes et les fonctions linéaires (Figure 6). La fonction sigmoïde est la plus utilisée car elle représente un bon compromis entre les fonctions seuils et linéaires.



La fonction à seuil La fonction sigmoïde La fonction linéaire

Figure 6 : Différents types de fonction de transfert pour le neurone artificiel.

Il existe deux types de réseaux de neurones : les réseaux non bouclés et les réseaux bouclés. Nous ne parlerons que des premiers. Les réseaux de neurone non bouclés réalisent une (ou plusieurs) fonction algébrique de ses entrées, par composition des fonctions réalisées par chacun de ses neurones. Il s'agit donc d'un ensemble de neurones connectés entre eux, l'information circulant des entrées vers les sorties sans retour en arrière possible. On parle souvent de perceptron multicouche à cause de la présence de neurones cachés (Figure 7).

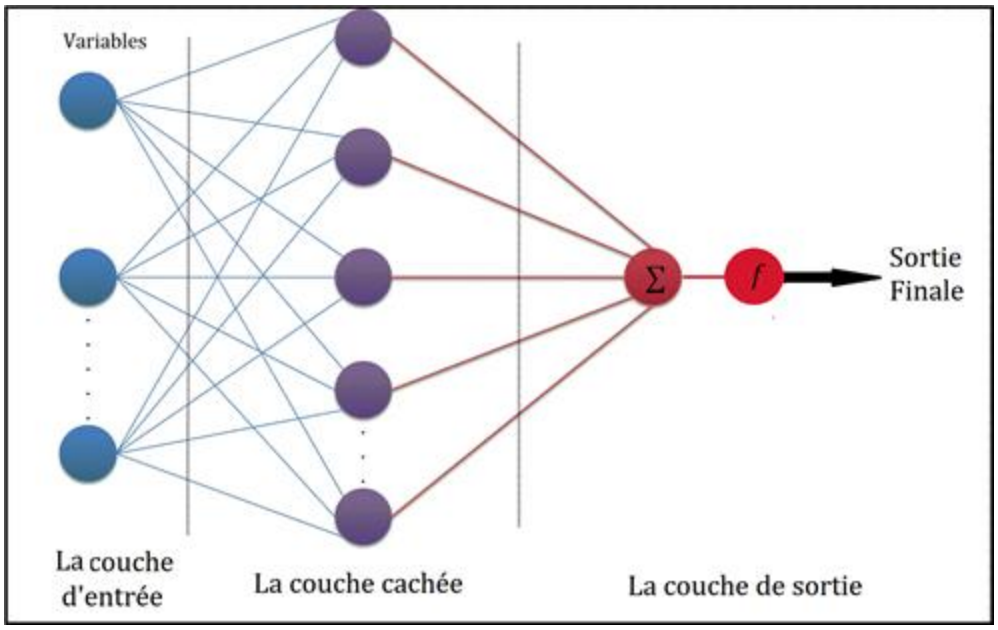


Figure 7 : Topologie d'un réseau de neurones à n entrées et une seule sortie.

- Apprentissage des réseaux de neurones artificiels ANNs

Dans le domaine des réseaux de neurones, l'apprentissage est une phase très importante qui désigne la procédure ou la façon qui consiste à déterminer l'architecture et les paramètres du réseau. En effet, une des propriétés fondamentales d'un réseau neuronal est sa capacité à s'adapter et améliorer sa performance en ajustant les connexions des neurones face à une source d'informations par la procédure d'apprentissage **[113]**.

L'apprentissage des réseaux de neurones artificiels se fait grâce à des algorithmes d'apprentissage. Dans la majorité des algorithmes actuels, l'apprentissage consiste à modifier les poids de connexions pour que la réponse du réseau s'accorde aux exemples de l'expérience **[102, 113]**.

Après une initialisation aléatoire des poids, des exemples expérimentaux sous formes de couples de vecteurs d'entrées et de sorties sont présentés au réseau. Les poids sont modifiés graduellement à l'aide des algorithmes d'apprentissage en vue de minimiser l'écart entre les sorties calculées (estimées) par le réseau et les sorties expérimentales (observations).

Mise en œuvre :

La base des données est divisée en deux parties :

- L'ensemble d'apprentissage : sur lequel se fait l'optimisation des poids.
- L'ensemble de test : sur lequel on teste la capacité de généralisation du réseau de façon à ce que les poids retenus soient ceux pour lesquels l'erreur obtenue sur cette base est faible.

En effet, si les poids sont ajustés sur toutes les données de l'ensemble d'apprentissage (70% de la base de données globale), on risque d'avoir le « sur-apprentissage » ou l'apprentissage par cœur, dans ce cas le réseau apprend très bien les données présentées dans la phase d'apprentissage sans pour autant être capable de généraliser le modèle à des données nouvelles.

Pour éviter le « sur-apprentissage » on introduit un nouvel ensemble de données appelé l'ensemble de validation (15% de la base de données globale). Comme pour l'ensemble de test (15% de la base de données globale), les éléments de cet ensemble ne participent pas à l'apprentissage. De plus, cet ensemble doit bien sûr avoir les mêmes contraintes que l'ensemble de test quant à sa représentativité et à sa taille.

L'ensemble de validation est utilisé de la façon suivante : dès que l'on s'aperçoit que l'erreur sur l'ensemble de validation stagne ou augmente, alors on arrête la procédure d'apprentissage **[113]**.

Dans la pratique, dans un premier temps, il faut calculer les poids du réseau c'est-à-dire estimer les paramètres essentiels. Pour cela, il faut construire un réseau reliant directement les neurones représentant les descripteurs moléculaires choisis avec les neurones de sortie. Chaque descripteur est alors affecté d'un poids en fonction de l'importance de chacun d'entre eux dans la propriété/l'activité étudiée.

Dans un second temps, il faut choisir l'architecture du réseau d'apprentissage c'est-à-dire choisir les entrées externes, le nombre de neurones cachés et l'agencement des neurones entre eux. Le nombre d'unités cachées joue un rôle important dans la qualité du réseau. Si le nombre est trop petit, le réseau possède trop peu de paramètres et ne peut interpréter les dépendances servant à modéliser et prévoir. Si le nombre de neurones cachés est trop grand, le réseau risque de s'ajuster au bruit présent dans les données de l'ensemble d'apprentissage.

Certains auteurs **[114, 115]** ont proposé un paramètre ρ, conduisant à déterminer le nombre de neurones cachés, qui joue un rôle majeur dans la détermination de la meilleure architecture du réseau (le meilleur nombre des couches cachées). Il est défini comme suit :

$$\rho = \frac{\text{Nombre de données dans le jeu d'entraînement}}{\text{Somme de nombre de connexions}}$$

Par conséquent, afin d'éviter le sur-ajustement ou le sous-ajustement, il est recommandé que la valeur de ρ soit comprise entre $1 < \rho < 3$ [**116**].

Enfin, il faut estimer la qualité du réseau obtenu en lui présentant des données ne faisant pas partie de l'apprentissage. Il s'agit d'une validation croisée (voir la partie : techniques de validation).

- Les réseaux de neurones artificiels dans MATLAB

Pour les réseaux de neurones artificiels (RNA), nous avons utilisé, dans l'ensemble de nos travaux, le logiciel MATLAB version 7.12 **[117]**.

Matlab est un logiciel de calcul matriciel à syntaxe simple. Avec ses fonctions spécialisées, Matlab peut être aussi considéré comme un langage de programmation adapté pour les problèmes scientifiques. Il est organisé en boîte à outils (*Neural Network Toolbox* **[118]**) spécialisés. Cette boîte offre de nombreuses architectures et fonctions d'apprentissage qui permettent de modéliser en toute simplicité des systèmes complexes non linéaires à l'aide de systèmes artificiels. Les applications de cette boite permettent de concevoir, d'effectuer l'apprentissage, de visualiser et simuler le réseau de manière interactive pour ensuite générer le code MATLAB équivalent et ainsi automatiser le processus.

6. Techniques de validation

Afin d'évaluer l'importance des modèles RQSA/RQSP et, par conséquent, ces capacités de prédiction des activités/propriétés d'autres (nouveaux) composés, la validation des modèles RQSA/RQSP reste une étape très sensible dans les études statistiques. Un modèle étant le résultat d'une analyse statistique, son interprétation et son application doivent se faire dans le cadre très précis du domaine couvert par l'analyse **[119]**. Toute application hors de ce cadre exige beaucoup de précautions et elle est d'autant plus hasardeuse qu'on s'éloigne du cadre. Pour éviter les erreurs, tant au moment de la validation qu'au moment de l'exploitation, les limites du modèle doivent être clairement établies : le non-robuste du modèle doit être vérifié, les pouvoir de prévision interne et externe doivent être déterminés, et l'espace chimique de l'application du modèle doit être limité.

6.1. Coefficients et tests statistiques standards

Afin de déterminer la qualité d'un modèle, différents paramètres statistiques sont employés, tels que les erreurs quadratiques moyennes (*Mean Square Errors*), les coefficients de corrélation qui sont régulièrement utilisés dans les études RQSA/RQSP, sont décrits en détail dans cette partie.

6.1.1. Coefficient de corrélation r (et coefficient de détermination r^2)

C'est l'indicateur statistique le plus répandu est le coefficient de corrélation qui évalue la part de la variance de l'activité / la propriété cible expliquée par le modèle.

$$r = \sqrt{1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}}$$

Avec : r est le coefficient de corrélation ; y_i et $\hat{y}_i$ sont, respectivement, les valeurs observées et calculées de la variable dépendante ; $\bar{y}$ est la valeur moyenne des valeurs observées.
Ces coefficients ne sont pas affectés par l'unité de mesure choisie et traduisent une bonne corrélation entre l'activité cible et l'activité initiale si r^2 est proche de 1 (cas idéal).

Le jugement sur la valeur de r ou r^2 est très subjectif. Bien que ce coefficient soit très facile à comprendre, il faut se garder d'y attacher trop d'importance car il est loin de fournir un critère suffisant pour juger la qualité d'une régression. Il n'est pas recommandé d'utiliser r^2 pour comparer des modèles avec un nombre différent de descripteurs, le coefficient r^2 nous dira toujours de choisir le modèle avec le plus grand nombre de descripteurs car son r^2 sera plus important (on projette sur un espace plus grand), même si les variables sont sans effets sur la réponse (l'activité ou la propriété étudiée).

La valeur de r^2 dépend de la taille de l'échantillon et le nombre de variables prédictives dans l'équation. Il garde la même valeur ou augmente lors d'une nouvelle variable de prédiction est ajoutée à l'équation de régression, même si la variable ajoutée ne contribue pas à la réduction de la variance inexpliquée. Par conséquent, un autre paramètre statistique peut être utilisé, appelé r^2 ajusté (r^2adj). Bien entendu, un autre indicateur est l'erreur quadratique moyenne (*MSE*, pour *Mean Square Error*), à laquelle est parfois préférée la déviation standard *s*.

6.1.2. Le coefficient de détermination ajusté r^2adj

Ce coefficient est utilisé en régression multiple par ce qu'il tient compte du degré de liberté :

$$r^2\text{adj} = \sqrt{\frac{r^2(\text{n}-1)-\text{p}}{\text{n}-\text{p}-1}}$$

Avec : n est le nombre des observations (les molécules) ; p est le nombre de variables indépendantes (les descripteurs) ; r^2 est le coefficient de détermination du modèle.

6.1.3. L'erreur quadratique moyenne « MSE » et l'erreur type résiduel « s »

$$MSE = \frac{\sum|(\hat{y}_\text{i} - \text{y}_\text{i})^2|}{\text{n}}$$

Ou encore, l'erreur type résiduel « *s* » :

$$s = \sqrt{\frac{\sum(\hat{y}_\text{i} - \text{y}_\text{i})^2}{\text{n}-\text{p}-1}}$$

Avec : y_i et $\hat{\text{y}}_\text{i}$ sont, respectivement, les valeurs observées et calculées de la variable dépendante ; n est le nombre des observations ; p est le nombre de variables indépendantes.

Ces paramètres mesurent la variation de l'activité cible non expliquée par le modèle RQSA/RQSP. En particulier, plus la déviation standard est petite et plus la corrélation est meilleure. Sa valeur est toujours fonction de l'unité de mesure de l'activité cible et tient également compte des erreurs expérimentales ce qui explique qu'une valeur trop petite n'ait aucune signification.

6.1.4. Le facteur d'inflation de la variance VIF

C'est un paramètre qui permet de détecter la colinéarité entre les descripteurs utilisés dans un modèle statistique, Il est défini par :

$$\text{VIF} = \frac{1}{1-r_\text{i}^2}$$

Avec : r_i^2 est le coefficient de détermination de la régression de la variable x_i sur les autres variables. Plus x_i est linéairement proche des autres variables, plus r_i^2 est proche de 1 et le VIF est grand. L'avantage du VIF par rapport à la matrice de corrélation est qu'il prend en compte des corrélations multiples.

6.1.5. Le test de Fisher F

L'indice de Fisher *F-test* est employé afin de mesurer le niveau de signifiance statistique du modèle à « x% » (le niveau usuel est 95%), c'est-à-dire la qualité du choix du jeu de paramètres. La conclusion obtenue ne doit pas nous faire penser que la corrélation a « x % » de chances d'être vraie mais seulement que la corrélation est vérifiée pour « x% » des composés pris pour référence et qu'une abstraction est faite pour les autres.

Hypothèses :

H_0: les variances des échantillons sont homogènes

H_1: les variances des échantillons ne sont pas homogènes

La valeur à calculer est :

On calcule le F (observé) à partir de la formule :

$$F(observé) = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \hat{y}_i)^2} \frac{n - p - 1}{p}$$

Avec : F est l'indice de Fisher ; y_i et $\hat{y}_i$ sont, respectivement, les valeurs observées et calculées de la variable dépendante ; $\bar{y}$ est la valeur moyenne des valeurs prédites ; n est le nombre des observations (les molécules) ; p est le nombre de variables indépendantes (les descripteurs).

Après le calcul de F (observé) on le compare avec le F théorique obtenu à partir des tables statistiques usuelles (la table de Fisher).

Si F observé est plus grand que le F théorique : refus de l'hypothèse nulle H_0 et cela signifie que les variances des échantillons sont trop différentes pour être considérées comme homogènes.

Si F observé est plus petit que le F théorique : acceptation de l'hypothèse nulle H_1 et cela signifie que les deux variances ont des valeurs suffisamment proches pour qu'on accepte l'idée qu'elles soient homogènes.

6.1.6. Le test de Student

L'indice de Student (le *t-test* de Student) est employé afin d'évaluer la pertinence des descripteurs dans un modèle. Il s'agit de tester l'hypothèse considérant le descripteur comme

non significatif. Pour une régression multilinéaire, cela revient à supposer le coefficient a_i qui lui est associé comme nul.

$$|t_i| = \left|\frac{a_i}{s(a_i)}\right| > t_{1-\frac{\alpha}{2}}^{n-p-2}$$

Avec : t_i est le *t-test* pour le descripteur « i » ; a_i est le coefficient associé au descripteur « i » dans le modèle ; $s(a_i)$ est l'erreur type associé au descripteur « i » ; α est l'intervalle de confiance ; n est le nombre des observations (les molécules) ; p est le nombre de variables indépendantes (les descripteurs).

Cette hypothèse est rejetée (avec un intervalle de confiance α) si le ratio t_i entre a_i et son erreur type s (a_i) atteint la valeur du fractile d'ordre (1- α/2) de la loi de Student à (n-p-2) degrés de liberté.

L'indice de Student (le t-test de Student) est employé aussi pur évaluer la significativité du modèle complet. Le test s'écrit : H_0 : r = 0 et H_1 : r≠0

Si le coefficient de corrélation est différent de zéro on rejette l'hypothèse H_0 (l'hypothèse nulle) et on accepte H_1 donc le modèle est significatif.

Sous H_0, la loi de Student à (n-p-1) degré de liberté t_{calc}s'écrit :

$$t_{calc} = \left[\frac{r}{\sqrt{\frac{1-r^2}{(n-p-1)}}}\right]$$

On rejette H_0 d'après (l'hypothèse nulle) lorsque :

$$t_{calc} > t_{\left(1-\frac{\alpha}{2}\right),(n-p-1)}$$

$t_{(1-\frac{\alpha}{2}),(n-p-1)}$: La valeur de la loi de Student, à (n-p-1) degré de liberté, à une Probabilité $(1-\frac{\alpha}{2})$ **[120-126]**.

6.2. Pouvoir de prévision interne

Afin de déterminer la stabilité prédictive d'un modèle et de tester l'influence de chaque échantillon (composé) sur le modèle final, des procédures de validation croisée (en anglais : cross-validation) sont souvent utilisées **[119]**.

Généralement, Ces techniques de validation permettent l'évaluation de la robustesse du modèle, autrement dit la stabilité des paramètres du modèle RQSA/RQSP vis-à-vis des molécules du jeu d'entraînement. Cela dit, qu'elles ne permettent en aucun cas de démontrer le pouvoir prédictif des modèles **[119, 116]**.

Le principe de ces méthodes consiste à extraire un certain nombre de molécules du jeu d'apprentissage et à construire un nouveau modèle avec les molécules restantes à l'aide des descripteurs choisis (seules les constantes de la régression changent). Ce nouveau modèle est alors utilisé pour la phase de prédiction sur les molécules retirées. Ce processus est ensuite répété pour retirer et prédire les valeurs de toutes les molécules du jeu d'entraînement (Figure 8). Le coefficient de corrélation q^2 (ou r^2_{cv}) entre les activités ainsi calculées et les activités observées exprime le pouvoir de prévision interne du modèle, plus la valeur du coefficient se rapproche de 1 plus le pouvoir de prévision sera meilleur. Pour que le modèle soit acceptable le pouvoir de prévision interne doit être supérieur à 0,5 **[127]**.

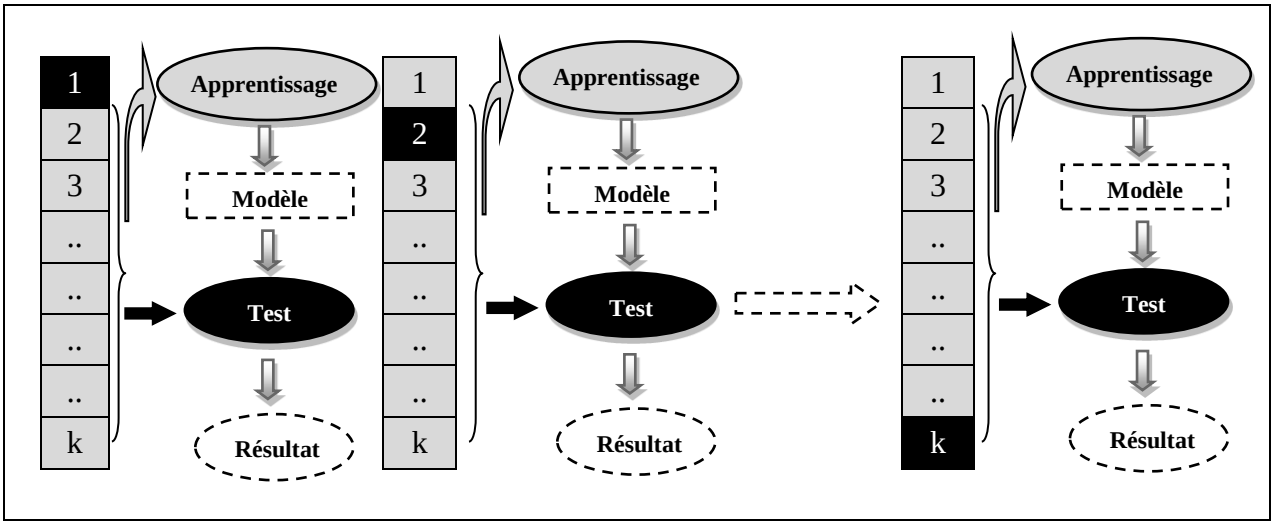


Figure 8 : Procédure de la validation croisée

6.2.1. Validation croisée "leave-many-out"

La procédure « k-fold cross-validation » correspond à un découpage du jeu d'apprentissage en k parties. On sélectionne un des k échantillons comme ensemble de validation et les (k-1) autres échantillons constitueront l'ensemble d'apprentissage. On construit un modèle RQSA/RQSP à partir de l'ensemble d'apprentissage, et on prédit les activités/propriétés de l'ensemble de validation. L'opération est répétée k fois pour qu'en fin de compte chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation. Le coefficient de corrélation q^2 (ou r^2_{cv}) entre les activités ainsi prédites et les activités observées est calculée pour estimer le pouvoir de prévision interne du modèle.

6.2.2. Validation croisée "leave one out cross-validation"

Cette méthode est un cas particulier de la validation croisée « k-paquets » où k=n. c'est-à-dire que l'on apprend sur (n-1) observations pour construire le modèle RQSA/RQSP puis on le valide sur la $n^{ième}$ observation et l'on répète cette opération n fois pour qu'en fin de compte chaque observation ait été utilisée exactement une fois comme ensemble de validation.

6.2.3. Validation par le test d'hasardisation (Y-randomisation test)

Le pouvoir de prévision interne par la procédure de la validation croisée a tendance à être surestimé. Une valeur élevée du coefficient de corrélation q^2 (ou r^2_{cv}) peut résulter d'une corrélation due au hasard **[128]** ou résulter de la redondance des structures lorsque les différences entre les composés du jeu d'apprentissage sont minimes (l'autocorrélation).

Pour contrôler la robustesse d'un modèle on utilise souvent le test d'hasardisation [**129, 130]**. Dans cette approche de validation, les valeurs de la variable cible sont redistribuées aléatoirement sur l'ensemble du jeu d'apprentissage et un nouveau modèle est dérivé. L'opération est répétée plusieurs fois et si la moyenne des coefficients de corrélations résultants reste élevée, on peut conclure qu'aucun modèle acceptable ne peut être obtenu par cette méthode statistique sur ce jeu de données.

6.3. Pouvoir de prévision externe

Un modèle avec des valeurs élevées des indices internes q^2 (ou r^2_{cv}) n'est pas encore dit valide, par conséquent, la validation interne est nécessaire mais insuffisante.

La puissance prédictive réelle d'un modèle RQSA/RQSP est de tester leur capacité à prédire parfaitement l'activité/propriété des composés à partir d'un ensemble de test externe (composés non utilisés pour le développement du modèle). Le but d'un bon modèle RQSA/RQSP est non seulement de prédire l'activité des composés d'ensemble d'apprentissage, mais aussi de prévoir les activités des molécules de test **[131]**. Le modèle RQSA/RQSP est bâti sur l'ensemble d'apprentissage et validé sur l'ensemble de test. La capacité prédictive du modèle est basée sur le coefficient de corrélation r^2_{test} entre les activités observées et les activités prédites pour l'ensemble de test, la valeur plus élevée de r^2_{test} (> 0,5) indique la bonne productivité du modèle.

6.4. Domaine d'applicabilité

Un modèle RQSA/RQSP ne peut pas être considéré comme un modèle universel, parce qu'il est développé sur un nombre limité de composés qui ne couvrent pas tout l'espace chimique. Par conséquent, l'activité/propriété prédite d'un composé, chimiquement dissimilaire au jeu d'apprentissage, ne pourra pas être considérée fiable **[132, 133]**.

Un modèle idéal est celui qui est capable de prédire l'acticité ou la propriété de n'importe quelle molécule imaginable. Cependant cela est souvent loin d'être possible. La taille limitée du jeu d'entrainement rend l'espace chimique des modèles construits limité. Et par conséquent, lorsqu'une molécule se situe en dehors de cet espace chimique, la prédiction ne sera plus fiable **[20]**.

Pour éviter cette extrapolation hasardeuse et prévenir ce type des problèmes, un domaine d'applicabilité (DA), qui permet de définir la zone dans laquelle un composé pourra être prédit avec confiance, doit être déterminé. Le DA correspond donc à la région de l'espace chimique incluant les composés du jeu d'apprentissage et les composés similaires, proches dans ce même espace **[134]**. Cette stratégie permet d'éliminer du jeu de test les molécules se situant en dehors de l'espace chimique du jeu d'entrainement. Cette partie de l'analyse est d'ailleurs explicitement demandée dans les démarches de validation mises en place au niveau de l'OCDE **[20, 135]**.

Il existe plusieurs méthodes pour la détermination du domaine d'applicabilité d'un modèle RQSA/RQSP, parmi ces méthodes on trouve la méthode de leviers (en anglais : *leverage*) qui est la plus utilisée, elle est basée sur la variation des résiduels standardisés de la variable dépendante avec la distance entre les valeurs des descripteurs et leurs moyennes appelée leviers h_i. Si un composé a un résiduel et un levier qui dépasse le seuil h*=3p/n (où p est le nombre de descripteurs plus 1 et n le nombre d'observations), ce composé est considéré en dehors du domaine d'applicabilité du modèle élaboré.

Le domaine d'applicabilité sera discuté à l'aide du diagramme de Williams qui représente les résidus de prédiction standardisés en fonction des valeurs des leviers h_i **[119, 136]**.

Pour chaque composé *i* dans l'espace original des variables indépendantes (xi), la valeur de h_i est calculée par la relation suivante **[137]** : $h_i = x_i^T (X^T X)^{-1} x_i \ (i = 1,2, \dots, n)$

Avec : xi est le vecteur ligne des descripteurs du composé i ; X (n*k-1) est la matrice du modèle déduit des valeurs des descripteurs de l'ensemble de d'entrainement ; L'indice T désigne la matrice transposée. La valeur critique du levier (h*) est fixée à : $h^* = 3(k+1)/n$ **[134]**.

Avec n est le nombre de composés utilisés de test ; k est le nombre des descripteurs du modèle.

Si hi<h*, la probabilité d'accord entre les valeurs mesurée et prédite du composé « i » est aussi élevée que celle des composés de la base de données. Les composés avec hi>h* renforcent le modèle quand ils appartiennent à l'ensemble d'entrainement, mais auront, sinon, des valeurs prédites douteuses sans pour autant être forcément aberrantes, les résidus pouvant être bas **[138]**.

7. Logiciels utilisés dans nos études RQSA/RQSP

Il existe plusieurs logiciels libres ou commerciaux disponibles dans les études RQSA/RQSP. Ceux-ci comprennent des logiciels spécialisés pour dessiner les structures chimiques, générant des structures 3D, le calcul des descripteurs moléculaires et le développement de modèles RQSA/RQSP. Les logiciels utilisés dans nos travaux sont :

Le dessin des molécules a été fait par ChemOffice, ChemSketch et Marvin Sketch **[69-71]** ;

Les structures 3D des molécules ont été générées par GaussView et ChemOffice **[46, 69]** ;

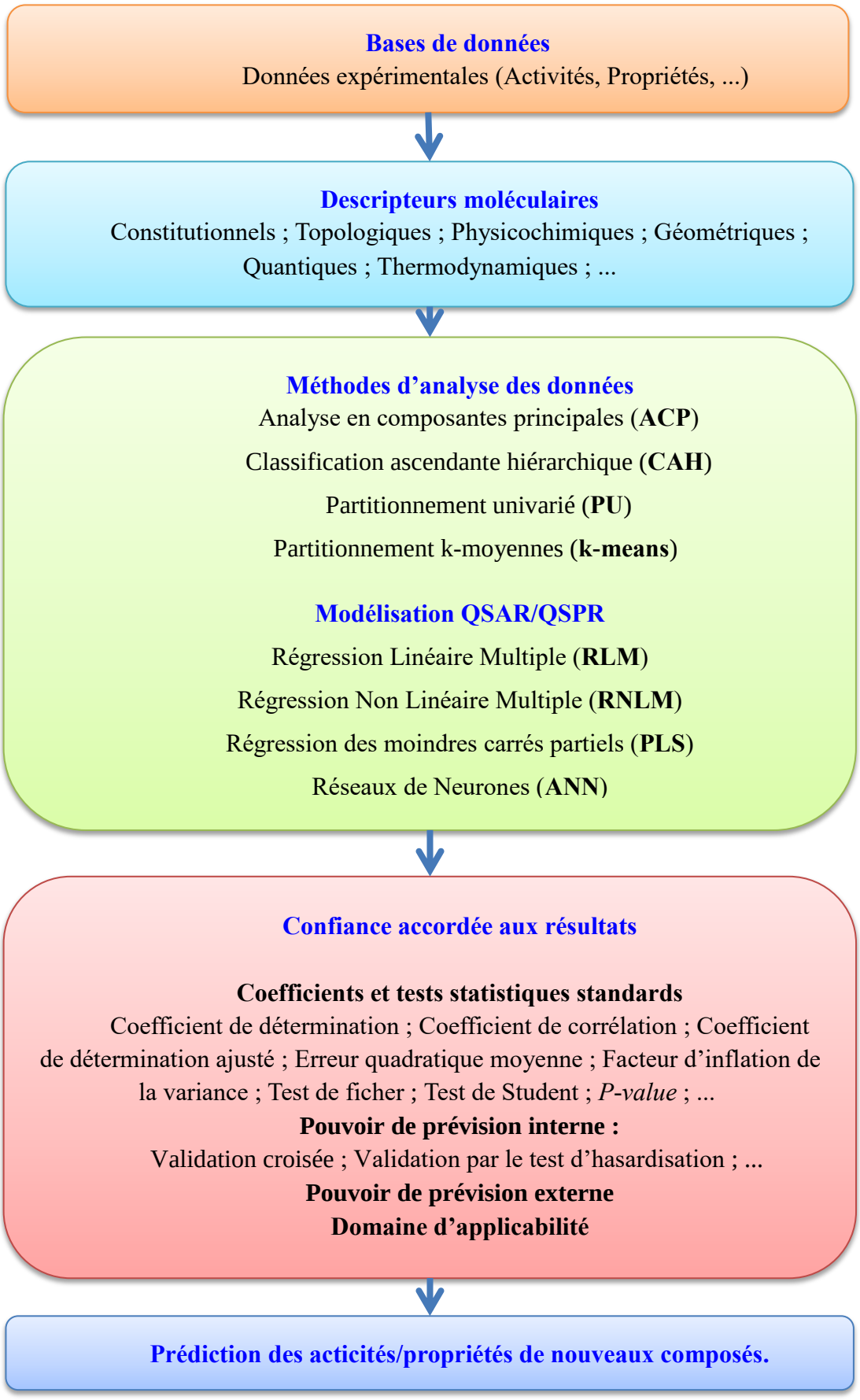
Les descripteurs ont été calculés par Gaussian 03, ChemSketch et Marvin Sketch **[68-71]** ;

L'analyse descriptive et la validation des modèles ont été faites par XLSTAT **[139]** ;

Les modèles RQSA/RQSP ont été générés en utilisant les logiciels suivants :

- Pour la régression linéaire multiple, la régression non linéaire multiple et la régression des moindres carrés partiels, nous avons utilisé XLSTAT.
- Pour les réseaux de neurones artificiels et les domaines d'applicabilité, nous avons utilisé MATLAB **[117]**.

8. Schéma de la méthodologie utilisée dans nos travaux



Références

[1] A.F.A Cros, "Action de l'alcool amylique sur l'organisme", *Thèse de Doctorat, Faculté de Médecine, Université Strasbourg, Strasbourg*, **1863**.

[2] A.C. Crum-Brown and T.R. Fraser, "On the Connection Between Chemical Constitution and Physiological Action, Part I: On the Physiological Action of the Salts of the Ammonium Bases, Derived from Strychnia, Brucia, Thebia, Codeia, Morphia, Nicotia", *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 25, **1868**, 151–203; "On the Connection between Chemical Constitution and Physiological Action. Part II: On the Physiological Action of the Ammonium Bases derived from Atropia and Conia", *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 25, **1869**, 693–739.

[3] M.C. Richet, "Noté sur le rapport entre la toxicité et les propriétés physiques des corps", *Comptes rendus des séances de la Société de biologie et de ses filiales, Paris,* 45, **1893**, 775–6

[4] H. Meyer, "Zur Theorie der Alkoholnarkose. Erste Mittheilung. Welche Eigenschaft der Anästhetica bedingt ihre narkotische Wirkung", *Archiv für experimentelle Pathologie und Pharmakologie*, 42, **1899**, 109–118.

[5] E. Overton, "Studien über die Narkose zugleich ein Beitrag zur allgemeinen Pharmakologie", *Ed. G. Fischer, Jena,* **1901**.

[6] a- R.L. Lipnick, "Charles Ernest Overton: narcosis studies and a contribution to general pharmacology", *Trends in Pharmacological Sciences,* 7, **1986**, 161–164.
b- R.L. Lipnick, "Hans Horst Meyer and the lipoid theory of narcosis", *Trends in Pharmacological Sciences,* 10(7), **1989**, 265–269.

[7] H. Fühner and E. Neubauer, "ămolyse durch Substanzen homologen Reihen", *Archiv für experimentelle Pathologie und Pharmakologie*, 56, **1907**, 333–345.

[8] O.R. Hansen, "Hammett Series with Biological Activity", *Acta Chemica Scandinavica,* 16, **1962**, 1593–1600.

[9] C. Hansch and T. Fujita, "p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure", *Journal of the American Chemical Society*, 86(8), **1964**, 1616–1626.

[10] S.M. Free and J.W. Wilson, "A Mathematical Contribution to Structure-Activity Studies", *Journal of Medicinal Chemistry*, 7(4), **1964**, 395–399.

[11] C. Hansch and E.J. Lien, "Structure-activity relationships in antifungial agents. A survey", *Journal of Medicinal Chemistry*, 14(8), **1971**, 653–670.

[12] S.Y. Tham and S. Agatonovic-Kustrin, "Application of the artificial neural network in quantitative structure-gradient elution retention relationship of phenylthiocarbamyl amino acids derivatives", *Journal of Pharmaceutical and Biomedical Analysis*, 28(3), **2002**, 581-590.

[13] R.D. Cramer, D.E. Patterson, and J.D. Bunce, "Comparative molecular field analysis (CoMFA): Effect of shape on binding of steroids to carrier proteins", *Journal of the American Chemical Society*, 110(18), **1988**, 5959-5967.

[14] G. Klebe, U. Abraham, and T. Mietzner, "Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity", *Journal of Medicinal Chemistry*, 37(24), **1994**, 4130-4146.

[15] A. Fortuné, "Techniques de Modélisation Moléculaire appliquées à l'étude et à l'optimisation de Molécules Immunogènes et de Modulateurs de la Chimiorésistance", *Thèse de Doctorat, Université Joseph Fourier – Grenoble I, France*, **2006**.

[16] G.E.P. Box and D.R Cox., "An analysis of distributions", *Journal of the royal statistical society, Series B,* 26(2), **1964**, 211-243.

[17] P. Armitage, G. Berry, "Statistical Methods in Medical Research", *3rd ed., Blackwell*, **1994**.

[18] Organisation de Coopération et de Développement Economiques (OCDE), 20 octobre **2009**.

[19] "Final Report of the OECD Workshop on Harmonization of Validation and Acceptance Criteria for Alternative Toxicological Test Methods", *Organisation de Coopération et de Développement Economique, Paris,* **2009**.

[20] "Principles for the Validation, for Regulatory Purposes, of Structure-Activity Relationship Models", *Organisation de Coopération et de Développement Economique, Paris,* 2009.
[21] M. Karelson, "Molecular descriptors in QSAR/QSPR", *Wiley, New York,* **2000**.
[22] R. Todeschini, V. Consonni, and R. Mannhold "Molecular Descriptors for Chemoinformatics", *Drug Discovery & Development*, 41(2), **2009**.
[23] A.Z. Dudek, T. Arodz, and J. Gàlvez, "Computational methods in developing quantitative structure-activity relationships (QSARs): A review", *Combinatorial Chemistry & High Throughput Screening*, 9, **2006**, 213–228.
[24] S. Rekkab, "Drug Design et synthèse de nouveaux calix[8]arènes sulfoniques flexibles à activités anticorrosive et anticoagulante", *Thèse de Doctorat, Univ. Constantine, Alger,* **2014**.
[25] C.A. Lipinski, F. Lombardo, B.W. Dominy, and P.J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings", *Advanced Drug Delivery Reviews*, 6(1–3), **1997**, 3–25.
[26] A. Goulon-Sigwalt-Abram, "Une nouvelle méthode d'apprentissage de données structurées : applications à l'aide à la découverte de médicaments", *Thèse de Doctorat, Université Pierre et Marie Curie (Paris 6), France*, **2008**.
[27] L. Euler, "the solution of a problem relating to the geometry of position", *Commentarii academiae scientiarum Petropolitanae*, 8, **1741**, 128–140.
[28] H.P. Schultz, "Topoligical organic chemistry: Graph theory and topological indices of alkanes", *Journal of Chemical Information and Modeling*, 29, **1989**, 227–228.
[29] A.T. Balaban, "Highly discriminating distance-based topological index", *Chemical Physics Letters*, 89, **1982**, 399–404
[30] H. Wiener, "Structural determination of paraffin boiling points", *Journal of Chemical Information and Modeling*, 69, **1947**, 17–20.
[31] I. Gutman, "Selected properties of the Schultz molecular topological index", *Journal of Chemical Information and Modeling*, 34(5), **1994**, 1087–1089.
[32] D. Jaiswal, C. Karthikeyan, and P. Trivedi, "Rationalization of physicochemical properties of alkanoic acid derivaties towards histone deacetylase inhibition", *Internet Electronic Journal of Molecular Design*, 5, **2006**, 13–26.
[33] M. Randic, "Molecular shape profiles", *Journal of Chemical Information and Modeling*, 35, **1995**, 373–382.
[34] R.A. Saunders, J.A. Platts, "Scaled polar suface area descriptors: development and application to three sets of partition coefficients", *New Journal of Chemistry*, 28, **2004**, 166–172.
[35] P. Labute, "A widely applicable sef of descriptors", *Journal of Molecular Graphics and Modelling*, 18, **2000**, 464–477.
[36] J. Higo and N. Go, "Algorithme for rapid calculation of excluded volume of large molecules", *Journal of Computational Chemistry*, 10, **1989**, 376–379.
[37] R. Bosque, J. Sales, E. Bosch, M. Rosès, M.C. Garcia-Alvarez-Coque, and J.R. Torres-Lapasio, "A QSPR study of the p-solute polarity parameter to estimate retention in HPLC", *Journal of Chemical Information and Modeling*, 43, **2003**, 1240–1247.
[38] V.N. Viswanadhan, A.K. Ghose, G.R. Revankar and R.K. Robins, "Atomic physicochemical parameters for 3D structure directed quantitative structure-activity relationships: Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occuring nucleoside antibiotics", *Journal of Chemical Information and Modeling*, 29, **1989**, 163–172.
[39] P.J. Taylor, "Hydrophobic Properties of Drugs, In Quantitative Drug Design", *Pergamon Press, Oxford (UK),* 4, **1990**, 241–294.
[40] H.A. Lorentz, "Über die Beziehungzwischen der Fortpflanzungsgeschwindigkeit des Lichtes der K€orperdichte", *Wiedemann's Annalen der Physik*, 9, **1880**, 641–665.

[41] C. Hansch, B.R. Telzer, and L.T. Zhang, "Comparative QSAR in toxicology: examples from teratology and cancer chemotherapy of aniline mustards", *Critical Reviews in Toxicology*, 25, **1995**, 67–89.
[42] A. Cammarata, "An Apparent Correlation between the in Vitro Activity of Chloramphenicol Analogs and Electronic Polarizability", *Journal of Medicinal Chemistry*, 10, **1967**, 525–527.
[43] A. Leo, C. Hansch, and C. Church, "Comparison of parameters currently used in the study of structure-activity relationships", *Journal of Medicinal Chemistry*, 12(5), **1969**, 766–771.
[44] C. Hansch and E. Coats, "α-Chymotrypsin: A Case Study of Substituent Constants and Regression Analysis in Enzymic Structure—Activity Relationships", *Journal of Pharmaceutical Sciences*, 59(6), **1970**, 731–743.
[45] S. Sudgen, "The variation of surface tension with temperature and some related functions", *Journal of the Chemical Society*, 125, **1924**, 32–41.
[46] F. Neese, "A critical evaluation of DFT, including time-dependent DFT, applied to bioinorganic chemistry", *Journal of Biological Inorganic Chemistry*, 11(6), **2006**, 702–711.
[47] I.I.R. Denning, T. Keith, J. Millam, K. Eppinnett, W.L. Hovell, R. Gilliland, GaussView Version 3.09, *Semichem Shawnee Mission, KS, USA*, **2003**.
[48] K. Fukui, "Theory of Orientation and Stereoselection", *Reactivity and Structure Concepts in Organic Chemistry*, 2, **1975**, 34–39.
[49] R. Franke, "Theoretical Drug Design Methods", *Elsevier Amsterdam*, **1984**, 115–123.
[50] P.W. Atkins and J. de Paula, "Atkins' Physical Chemistry", *7th ed., Oxford University Press, Oxford*, **2002**.
[51] D.F.V. Lewis, C. Ioannides, and D.V. Parke, "Interaction of a series of nitriles with the alcohol-inducible isoform of P450: Computer analysis of structure—activity relationships", *Xenobiotica*, 24(5), **1994**, 401–408.
[52] Z. Zhou and R.G. Parr, "Activation hardness: new index for describing the orientation of electrophilic aromatic substitution", *Journal of the American Chemical Society*, 112(15), **1990**, 5720–5724.
[53] O. Kikuchi, "Systematic QSAR Procedures with Quantum Chemical Descriptors", *Molecular Informatics*, 6(4), **1987**, 179–184.
[54] R.G. Parr, R.A. Donnelly, M. Levy, and W.E. Palke, "Electronegativity: The density functional viewpoint", *The Journal of Chemical Physics*, 68(8), **1978**, 3801–3807.
[55] R.S. Mulliken, "A new electroaffinity scale; Together with data on valence states and on valence ionization potentials and electron affinities", *The Journal of Chemical Physics*, 2, **1934**, 782–793.
[56] R.G. Parr and R.G. Pearson, "Absolute hardness: companion parameter to absolute electronegativity", *Journal of the American Chemical Society*, 105(26), **1983**, 7512–7516.
[57] W. Yang and R.G. Parr, "Hardness, softness, and the fukui function in the electronic theory of metals and catalysis", *Proceedings of the National Academy of Sciences of the United States of America*, 82(20), **1985**, 6723–6726.
[58] R.G. Pearson, "Absolute electronegativity and hardness: applications to organic chemistry", *The Journal of Organic Chemistry*, 54(6), **1989**, 1423–1430.
[59] R.G. Parr, L.V. Szentpaly, and S. Liu, "Electrophilicity Index", *Journal of the American Chemical Society*, 121(9), **1999**, 1922–1924.
[60] D.A. McQuarrie, "Statistical Thermodynamics", *harper row publishers, New* York, **1973**.
[61] A.I. Akhiezer and S.V., "Peltminskii, Methods of Statistical Physics", *Pergamon Press, Oxford*, **1981**.
[62] P.W. Atkins, "Physical Chemistry", *2nd edition, W.H. Freeman and Company, San Francisco*, **1982**.

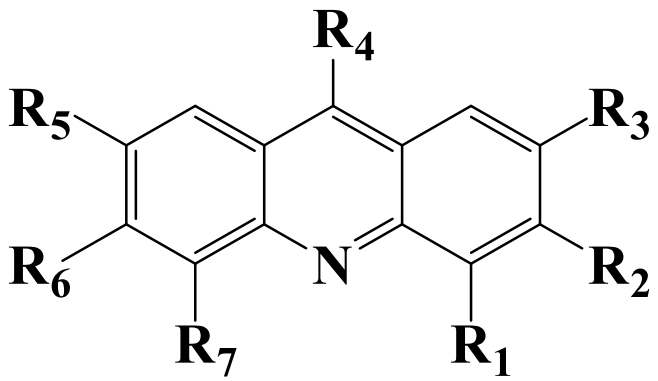
[63] A.R Katritzky, V.S. Lobanov, and M. Karelson, "Normal boiling points for organic compounds: correlation and prediction by a quantitative structure–property relationship", *Journal of Chemical Information and Modeling*, 38, **1998**, 28–41.
[64] W.A. Wakeham, G.S. Cholakov, and R.P. Stateva, "Liquid density and critical properties of hydrocarbons estimated from molecular structure", *Journal of Chemical & Engineering Data*, 47, **2002**, 559–570.
[65] A.R. Katritzky, U. Maran, M. Karelson, and V.S. Lobanov, "Prediction of melting points for the substituted benzenes: a QSPR approach", *Journal of Chemical Information and Modeling*, 37, **1997**, 913–919.
[66] C. Navajas, A. Poso, K. Tuppurainen, and J. Gynther, "Comparative Molecular Field Analysis (CoMFA) of MX Compounds using different Semi-Empirical Methods: LUMO Field and its Correlation with Mutagenic Activity", *Molecular Informatics*, 15(3), **1996**, 189–193.
[67] F. Bonachera, "Les triplets pharmacophoriques flous : développement et applications", *Thèse de Doctorat, Université Lille 1 Sciences et Téchnologies, France*, **2011**.
[68] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, T. Vreven, K. N. Jr., Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez and J.A. Pople, "Gaussian 03, Revision B.04", *Gaussian, Inc. Pittsburgh PA*, **2003**.
[69] ChemOffice, *PerkinElmer Informatics*, **2010**.
[70] ACDLABS 10, *Advanced Chemistry Development Inc., Toronto, ON, Canada*, **2015**.
[71] Marvin Sketch 5.11.4, *Chem Axon*, **2012**.
[72] QSARIS. www.scivision.com/qsaris.html.
[73] Cerius2. www.accelrys.com/products/cerius2.
[74] Vol Surf. www.moldiscovery.com/softvolsurf.php.
[75] DRAGON. http://www.talete.mi.it/products/dragondescription.htm.
[76] S. Mannan, "Lee's Loss Prevention in Process Industries: Hazard Identification, Assessment and Control", *Elsevier Butterworth-Heinemann, Burlington*, **2005**.
[77] M.J. Crawley, "Statistics: an introduction using R", *Wiley, Chichester, UK*, **2005**.
[78] A.R. Katritzky, V.S. Lobanov, and M. Karelson, "CODESSA Reference Manual", *University of Florida, Gainesville*, **1994**.
[79] V.Y. Nalimov, "the Application of Mathematical Statistics to Chemical Analysis", *Addison-Wesley, Reading, MA*, **1962**.
[80] R. Calcutt and R. Body, "Statistics for Analytical Chemists", *Champman & Hall, New York*, **1983**.
[81] J.C. Miller and J.N. Miller, "Statistics for Analytical Chemistry", *Ellis Horwood, New York*, **1988**.
[82] P.C. Meier and R.E. Zund, "Statistical Methods in Analytical Chemistry", *Wiley, New York*, **1993**.
[83] P. Dagnélie, « Statistique théorique et appliquée. Tomes 1 et 2", *De Boeck & Larcier*, **1998**.
[84] S. Stigler, "Statistics on the Table: The History of Statistical Concepts and Methods", *Harvard University Press*, **1999**.

[85] N. Trinajstiæ, S. Nikoliæ, S.C. Basak, and I. Lukovits, "Distances indices and their hypercounterparts: Intercorrelation and use in the structure-property modeling", *SAR and QSAR in Environmental Research,* 12, **2001**, 31–54.
[86] P.P. Roy, S. Paul, I. Mitra, and K. Roy, "Two novel parameters for validation of predictive QSAR models", *Molecules,* 14, **2009**, 1660–701.
[87] J.G. Topliss and R.P. Edwards, "Chance factors in studies of quantitative structure-activity relationships", *Journal of Medicinal Chemistry,* 22(10), **1979**, 1238–1244.
[88] I.T. Jolliffe, "Principal Component Analysis", *New-York, NY: Springer, 2ème édition,* **2002**.
[89] A. Morineau and T. Aluja-Banet, "Analyse en composantes principales Centre international de statistique et d'informatique appliquée", *Saint-Mandé CISIA-CERESTA,* 1998.
[90] F.L. Stahura and J. Bajorath, "New methodologies for ligand-based virtual screening", *Curr Pharm Des,* 11(9), **2005**, 1189–1202.
[91] R.A. Jarvis and E. A. Patrick, "Clustering using a similarity measure based on shared near neighbors", *IEEE Transactions on Computers,* C-22(11), **1973**, 1025–1034.
[92] J. H. Ward, "Hierarchical grouping to optimize an objective function", *Journal of the American Statistical Association,* 58(301), **1963**, 236–244.
[93] D.T. Stanton, T.W. Morris, S. Roychoudhury, and C.N. Parker, "Application of nearest-neighbor and cluster analyses in pharmaceutical lead discovery", *Journal of Chemical Information and Modeling,* 39(1), **1999**, 21–7.
[94] J. Han, M. Kamber and J. Pei, "Data Mining: Concepts and Techniques, Chapter 8: Classification: Basic Concepts", *3rd edition, Morgan Kaufmann Publishers,* **2011**.
[95] F.R. Burden, M.G. Ford, D.C. Whitley and D.A. Winkler, "Use of automatic relevance determination in QSAR studies using Bayesian neural networks", *Journal of Chemical Information and Modeling,* 40(6), **2000**, 1423–1430.
[96] K. Roy, S. Kar and R. Narayan Das, "Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment, Chapter 6 - Selected Statistical Methods in QSAR", *Academic Press, Boston,* **2015**, 191–229.
[97] T. Puzyn, J. Leszczynski and M.T. Cronin, "Recent Advances in QSAR Studies: Methods and Applications: Part I Theory of QSAR", *Challenges and Advances in Computational Chemistry and Physics*, 8, **2010**.
[98] R.D. Tobias, "An Introduction to Partial Least Squares Regression", *Statistical Analysis System Institute Inc., Cary, USA,* **2002**.
[99] J. Rolland and P. Blouch, "Les bouées météorologiques : L'exemple de Météo-France", *La Météorologie,* 39, **2002**, 83–88.
[100] J.M. Torres-Moreno, ''Apprentissage et généralisation par des réseaux de neurones : étude de nouveaux algorithmes constructifs'', *Thèse de Doctorat, Institut Nationale Polytechnique de Grenoble,* **1992**.
[101] N. Fadlallah, "Contribution à l'optimisation de la synthèse du lobe de rayonnement pour une antenne intelligente. Application à la conception de réseaux à déphasage", *Thèse de Doctorat, Université de Limoges, Facultés des Sciences et Techniques,* **2005**.
[102] S. Chabaa, "Identification des Systèmes non Linéaires en Utilisant les Techniques d'Intelligences Artificielles et les Bases de Fonctions de Laguerre pour la Modélisation des Données du Trafic dans les Réseaux Internet", *Thèse de Doctorat, Université Cadi Ayyad, Faculté des Sciences Semlalia - Marrakech,* **2011**.
[103] W.S. Mc Culloch and W. Pitts, "A logical calculus of ideas immanent in nervous activity", *The Bulletin of Mathematical Biophysics*, 5, **1943**, 115–133.
[104] G. Dreyfus, J-M. Martinez, M. Samuelides, M.B. Gordon, F. Badran, S. Thiria, and L. Hérault, "Réseaux de neurones, méthodologie et applications", *2nd ed., Paris : Eyrolles,* **2004**.
[105] F. Rosenblatt, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain", *Psychological Review*, 65, **1958**, 386–408.

[106] B. Widrow and M.E. Hoff, "Adaptive Switching Circuits", *IRE WESCON Convention Record*, 4, **1960,** 96–104.
[107] M. Minsky and S. Papert, "Perceptrons: An Introduction to Computational Geometry, 2nd edition with corrections, first edition 1969", *The MIT Press; Cambridge MA*, **1972**.
[108] J.J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities, Biophysics", *Proceedings of the National Academy of Sciences of the United States of America*, 79(8), **1982**, 2554–2558.
[109] P. Werbos, "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences", *PhD Thesis, Harvard University, Cambridge*, **1974**.
[110] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning representations by back-propagating errors", *Nature*, 323, **1986**, 533–536
[111] T. Kohonen, "Self-Organization and Associative Memory", *Springer, Berlin*, **1984**.
[112] J. Moody and C. Darken, "Fast learning in networks of locally-tuned processing units", *Neural Computation*, 1, **1989**, 281–294.
[113] M.Y. Ammar, "Mise en œuvre de réseaux de neurones pour la modélisation de cinétiques réactionnelles en vue de la transposition Batch/Continu", *Thèse de Doctorat, Institut Nationale Polytechnique de Toulouse*, **2007**.
[114] S.S. So and W.G. Richards, "Application of neural networks: quantitative structure-activity relationships of the derivatives of 2,4diamino (substituted-benzyl) pyrimidines as DHFR inhibitors", *Journal of Medicinal Chemistry*, 35(17), **1992**, 3201–3207.
[115] T.A. Andrea and H. Kalayeh, "Applications of neural networks in quantitative structure-activity relationships of dihydrofolatereductase inhibitors", *Journal of Medicinal Chemistry*, 34(9), **1991**, 2824–2836.
[116] A. Golbraikh and A. Tropsha, "Beware of q^2!", *Journal of Molecular Graphics and Modelling*, 20(4), **2002**, 269–276.
[117] MATLAB 7.9.0 (R2009b) and Statistics Toolbox Release, "The Math Works", *Inc., Natick, Massachusetts, United States*, **2011**.
[118] H. Demuth, M. Beale, and M. Hagan, "Neural Network Toolbox™ 6 *User's Guide*", Available at: https://filer.case.edu/pjt9/b378s10/nnet.pdf
[119] A. Tropsha, P. Gramatica, and V.K. Gombar, "the importance of Being Earnest: Validation is the Absolute Essential for Successful Application and interpretation of QSPR Models", *QSAR and Combinatorial Sciences*, 22(1), **2003**, 69–77.
[120] J. Jacques, "Modélisation Statistique".
Available at: http://eric.univ-lyon2.fr/~jjacques/Download/Cours/Mod-Cours.pdf
[121] P. Besse, "Pratique de la modélisation statistique", *Publications du laboratoire de statistique et Probabilités*, **2003**.
[122] P. Besse, "Apprentissage Statistique Data mining", *Publications du laboratoire de statistique et Probabilités*, **2009**.
[123] G.J. Mclachlan, "Discriminant analysis and Statistical Pattern Recognition" *Wiley, New-York*, **1992**.
[124] J.P. Nakache and J. Confais, "Statistique explicative appliquée", Edition Technip, **2003**.
[125] G. Saporta, "Probabilité, Analyse de données et statistique", Edition Technip, **2006**.
[126] D. Laffly, "Régression multiple : principes et exemples d'application", *Université de Pau et des Pays de l'Adour*, **2006**.
[127] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation. Encyclopedia of Database Systems", *Editors: M. Tamer Ã-zsu and Ling Liu, Springer*, **2009**.
[128] M. Clark and R.D. Cramer, "The probability of chance correlation using partial least squares (PLS)", *Molecular Informatics*,12(2), **1993**, 137–45.
[129] S. Wold, and L. Eriksson, "Statistical validation of QSAR results. Validation tools", *Methods and Principles in Medicinal Chemistry*, 2, **1995**, 309–18.

[130] H. van der Voet, “Comparing the predictive accuracy of models using a simple randomization test”, *Chemometrics and Intelligent Laboratory Systems,* 25(2), **1994**, 313–23.
[131] S. Ekins, G. Bravi, S. Binkley, J.S. Gillespie, B.J. Ring, J.H. Wikel, and S.A Wrighton, “Three- and four-dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2C9 inhibitors”, *Drug Metabolism & Disposition,* 28(8), **2000**, 994–1002.
[132] I.V. Tetko, I. Sushko, A.K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, and A. Varnek, “Critical assessment of QSAR models of environmental toxicity against Tetrahymena-pyriformis: focusing on applicability domain and overfitting by variable selection”, *Journal of Chemical Information and Modeling,* 48, **2008**, 1733–1746.
[133] J. Jaworska, N.N. Jeliazkova, T. Aldenberg, “QSAR applicabilty domain estimation by projection of the training set descriptor space: a review”, *Alternatives to Laboratory Animals,* 33, **2005**, 445–459.
[134] T.I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. Van De Sandt, W. Tong, G. Veith, and C. Yang, “Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships”, *Alternatives to Laboratory Animals*, 33(2), **2005**, 155–173
[135] J. Tunkel, K. Mayo, C. Austin, A. Hickerson, and P. Howard, “Practical considerations on the use of predictive models for regulatory purposes”, *Environmental Science & Technology*, 39(7), **2005,** 2188–2199.
[136] L. Eriksson, J. Jaworska, A. Worth, M. Cronin, R.M. Mc Dowell, and P. Gramatica, “Methods for reliability, uncertainty assessment, and applicability evaluations of regression based and classification QSARs”, *Environmental Health Perspectives*, 111(10), **2003**, 1361–1375.
[137] P. Gramatica, “Principles of QSAR models validation: internal & external”, *QSAR and Combinatorial Sciences*, 26(5), **2007**, 694–701.
[138] J.C. Dearden, “the History & Development of Quantitative Structure-Activity Relationships (QSARs)”, *International Journal of Quantitative Structure-Property Relationships*, 1(1) **2016**.
[139] XLSTAT **2009** Add-in software, *XLSTAT Company*. www.xlstat.com.

Chapitre 2 : Etude de la RQSA de l'activité Antileishmanienne pour des dérivés de l'acridine



Hindawi Publishing Corporation
Advances in Physical Chemistry
Volume 2016, Article ID 5137289, 16 pages
http://dx.doi.org/10.1155/2016/5137289

Research Article

Investigation of Antileishmanial Activities of Acridines Derivatives against Promastigotes and Amastigotes Form of Parasites Using QSAR Analysis

Samir Chtita [1], Mounir Ghamali [1], Rachid Hmamouchi [1], Bouhya Elidrissi [1], Mohamed Bourass [2], Majdouline Larif [3], Mohammed Bouachrine [4] and Tahar Lakhlifi [1]

[1] *MCNSL, Faculty of Science, Moulay Ismail University, Meknes, Morocco*
[2] *Faculty of Sciences, Sidi Mohamed Ben Abdellah University, Fez, Morocco*
[3] *Separation Process Laboratory, Faculty of Science, Ibn Tofail University, Kenitra, Morocco*
[4] *High School of Technology, Moulay Ismail University, Meknes, Morocco*

Correspondence should be addressed to Samir Chtita; samirchtita@gmail.com

Received 13 August 2016; Revised 22 September 2016; Accepted 27 September 2016

Academic Editor: Michael D. Sevilla

Abstract

In a search of newer and potent antileishmanial (against promastigotes and amastigotes form of parasites) drug, a series of 60 variously substituted acridines derivatives were subjected to a quantitative structure activity relationship (QSAR) analysis for studying, interpreting, and predicting activities and designing new compounds by using multiple linear regression and artificial neural network (ANN) methods. The used descriptors were computed with Gaussian 03, ACD/ChemSketch, Marvin Sketch, and ChemOffice programs. The QSAR models developed were validated according to the principles set up by the Organization for Economic Co-operation and Development (OECD). The principal component analysis (PCA) has been used to select descriptors that show a high correlation with activities. The univariate partitioning (UP) method was used to divide the dataset into training and test sets. The multiple linear regression (MLR) method showed a correlation coefficient of 0.850 and 0.814 for antileishmanial activities against promastigotes and amastigotes forms of parasites, respectively. Internal and external validations were used to determine the statistical quality of QSAR of the two MLR models. The artificial neural network (ANN) method, considering the relevant descriptors obtained from the MLR, showed a correlation coefficient of 0.933 and 0.918 with 7-3-1 and 6-3-1 ANN models architecture for antileishmanial activities against promastigotes and amastigotes forms of parasites, respectively. The applicability domain of MLR models was investigated using simple and leverage approaches to detect outliers and outsides compounds. The effects of different descriptors in the activities were described and used to study and design new compounds with higher activities compared to the existing ones.

1. Introduction

For hundreds of years, Leishmaniasis, a disease caused by a number of species of protozoan parasites belonging to the genus *Leishmania*, is recognized as an important public health problem throughout the world [1–3]. Currently, the leishmaniases are considered to be endemic in 88 countries and, according to World Health Organization (WHO) [4], twelve million people are infected, with about two to three million new cases each year, and 350 million people are under risk of infection; it is a major public health problem particularly in Latin America, Africa, and Asia [5–9]. To date, no vaccine against any clinical form of Leishmaniasis has been commercialized and treatment relies only on chemotherapy, which has been based on the use of pentavalent antimonial drugs. Other medications, such as pentamidine and amphotericin B, have been used as alternative drugs for resistant parasites.

With the emergence of some resistant strains, the toxicity of current drugs, severe side effects, and high cost and/or restricted therapeutic spectrum, a need for development of new and safer drugs is warranted **[2, 3, 10]**. A great number of natural and synthetic compounds have been tested in the past years in antileishmanial assays. Their structures are diverse and often contain nitrogen heterocycles such as quinolines, pyrimidines, acridines, phenothiazines, and indoles **[11–13]**.

Many experiments have been performed with the compounds bearing the heterocyclic ring structures to explore their effectiveness against *Leishmania*. These studies suggested their similar pharmacophoric feature of the heterocyclic scaffold as a potential target for drug discovery of antileishmanial drugs **[14]**.

Leishmania parasites exist in two forms, one is promastigotes and the other is amastigotes. The promastigotes are flagellated and found in sand fly, while the amastigotes are ovoid and nonflagellated form of Leishmania **[15]**. Antileishmanial activity is performed against promastigotes and then amastigotes form of parasites. Heterocyclic system may also be formed by fusion with other rings, either carbocyclic or heterocyclic.

Since their discovery in the 1880s, acridines families have demonstrated a broad spectrum of pharmacological properties **[16]**. The first employed as antibacterial agents during the beginning of the twentieth century **[17]**. They have been rapidly revealing interesting antiproliferative activities against both protozoa and tumor cells **[18, 19]**. Consequently, they have been extensively used in antiparasitic chemotherapy and a wide range of new acridines derivatives have been synthesized and successfully assessed for their antileishmanial activities **[20, 21]**.

In order to open a new way in antileishmanial drug research, a series of sixty acridines derivatives were synthesized **[22–24]** and studied for their antileishmanial (against promastigotes and amastigotes form of parasites) activities. The aim of this study was to develop a QSAR model able to correlate the structural features of the acridines derivatives with their biological activities.

In general, the QSAR methods are based on the assumption that the activity of a certain chemical compound related to its structure through a certain mathematical algorithm. This relationship can be used in the prediction, interpretation, and assessment of new compounds with desired activities reducing and rationalizing time, efforts, and cost of synthesis and new product development.

The basic assumption to drive a QSAR model is presented due to a mathematical function of the chemical properties which is related to the effect (activity). Therefore, the effect is like the function "f" of the chemical properties "x": $y = (x)$. To find this algorithm, we use a number of chemical compounds with known values of the studied effect (y). For each chemical compound, we calculate a series of parameters (called chemical descriptors). Then, we find an algorithm that provides a quite accurate value, similar to the real experimental value. The final step is to check if the obtained algorithm is able to predict the activity values for other chemicals not used to build up the model (external validation).

Indeed, it is very important to generate a model which worked not only for the chemical substances used within the training set but also for other similar chemicals. Consequently, the challenge is to define the correct statistical properties of the model.

2. Material and Methods

The current QSAR study investigates prediction and interpretation of the studied compounds and was also used for designing new proposed compounds by using linear and nonlinear methods. It consists of four stages: selection of dataset and generation of molecular descriptors, descriptive analysis, statistical analysis (prediction and evaluation), and suggestion of novel compounds.

A flow chart for the development of the QSAR model along with the various validation methods used in this work is demonstrated in figure 1.

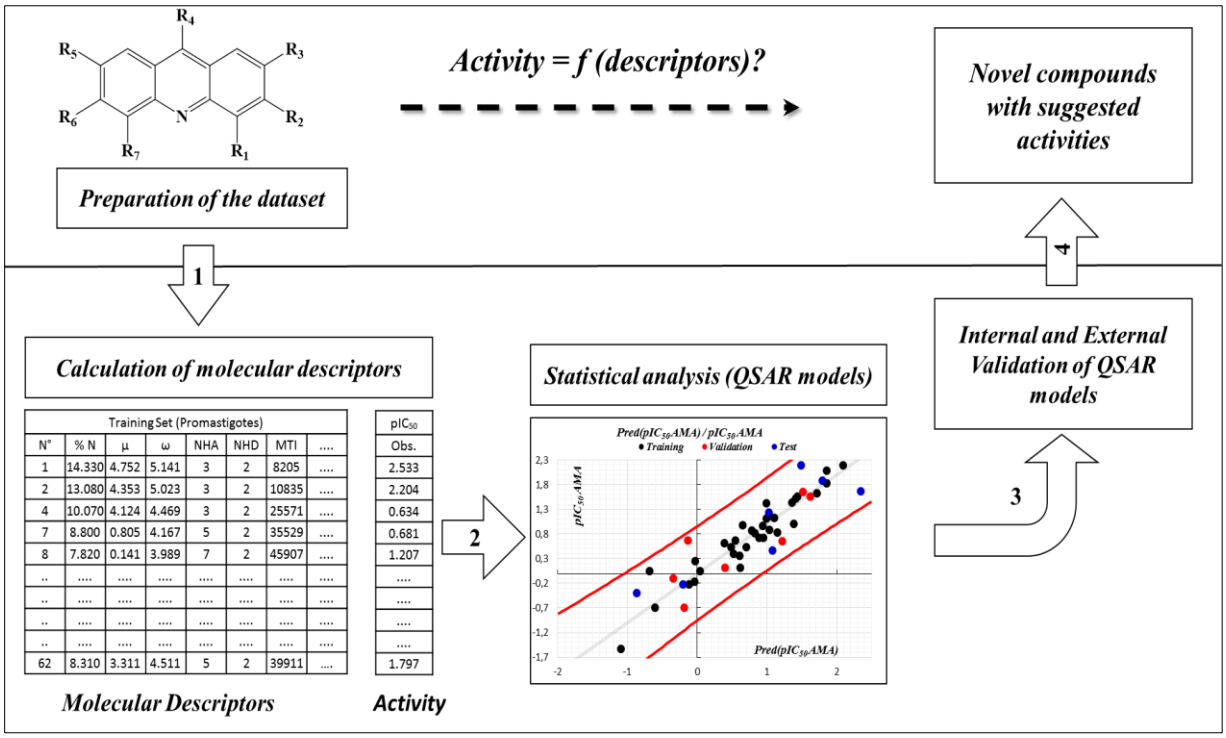


Figure 1: Flowchart of the methodology used in this work

2.1. Selection of dataset and generation of molecular descriptors

2.1.1. Selection of dataset

In this stage, the dataset of the antileishmanial activities (against promastigotes and amastigotes forms of parasites) of various acridine derivatives (4-monosubstituted acridines, 3,6-disubstituted acridines, 4,5-disubstituted acridines, and 7-monosubstituted 9-chloro and 9-amino-2-methoxy acridines) were collected from previous works **[22–24]**. The molecular structures of the studied molecules with their antileishmanial activities are presented in table 1. All experimental IC_{50} antileishmanial activity values (μM) were converted to the negative logarithm of IC_{50} ($pIC_{50} = -\log_{10}(IC_{50})$).

2.1.2. Molecular descriptors generation

A wide variety of molecular descriptors was calculated using Gaussian 03, ACD/ChemSketch, Marvin Sketch, and ChemOffice programs **[25–28]** to predict the correlation between these descriptors for the studied molecules with their antileishmanial activities and to develop linear (multiple linear regression (MLR)) and nonlinear (artificial neural network (ANN)) models. Tables 3 and 4 show the selected descriptors (using the PCA method; see more in descriptive analysis results) to be used in this study.

2.2. Descriptive analysis

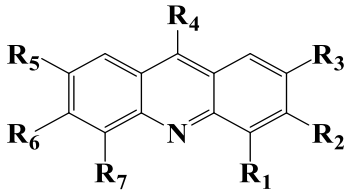
In this stage, the principal component analysis (PCA) was used to determine the nonlinearity among variables (descriptors) and to select descriptors that correlate with the activity. After that, the univariate partitioning (UP) method was used to form dissimilar clusters of compounds, to which the query compounds would be compared for determining the degree of similarity and dividing the dataset into training and test sets.

2.3. Statistical analysis (Models development and evaluation)

In this stage, linear and nonlinear QSAR models were developed and evaluated to predict the activities of the test compounds. The study we conducted consists of the multiple linear regression (MLR) available in the *XLSTAT* software and the artificial neural network (ANN) available in the *Matlab* software.

In order to propose models and to evaluate quantitatively the physicochemical effects of the substituents on the activities of molecules, we submitted the data matrix constituted obviously from the used variables (descriptors) corresponding to the dataset molecules to a descendant MLR analysis and to an ANN. We use the coefficients r, r^2, r^2_{adj}, MSE, and P_{value} to select the best regression performance **[29]**, where r is the correlation coefficient; r^2 is the coefficient of determination; r^2_{adj} is the coefficient adjusted for degrees of freedom.

Table 1: Chemical structure and antileishmanial activities of studied compounds



N°	R_1	R_2	R_3	R_4	R_5	R_6	R_7	pIC_{50} [b]	
								PRO[c]	AMA[d]
1	H	$NHCOCH_3$	H	H	H	$NHCOCH_3$	H	-2.533	-0.653
2	H	$NHCOC_2H_5$	H	H	H	$NHCOC_2H_5$	H	-2.204	-0.886
3	H	$NHCOC_3H_7$	H	H	H	$NHCOC_3H_7$	H	-1.669	-1.107
4	H	NHCOph	H	H	H	NHCOPh	H	-0.634	-0.041
5	H	NHCO-p-PhCl	H	H	H	NHCO-p-PhCl	H	---	0.699
6	H	NHCO-p-PhF	H	H	H	NHCO-p-PhF	H	-0.230	1.523
7	H	NHCO-p-PhOMe	H	H	H	NHCO-p-PhOMe	H	-0.681	-0.041
8	H	$NHCO\text{-}m,p\text{-}Ph(OMe)_2$	H	H	H	$NHCO\text{-}m,p\text{-}Ph(OMe)_2$	H	-1.207	0.097
9	H	NHCOMe	H	H	H	NHCOph	H	-2.270	Tox [a]
10	H	NHCOMe	H	H	H	NHCO-p-PhCl	H	-1.061	-0.462
11	H	NHCOMe	H	H	H	NHCO-p-PhF	H	-2.123	0.174
12	H	NHCOMe	H	H	H	NHCO-p-PhOMe	H	-1.939	Tox [a]
13	H	NHCOMe	H	H	H	$NHCO\text{-}m,p\text{-}Ph(OMe)_2$	H	---	-0.114
14	H	H	CH_3	NH_2	OMe	H	H	0,301	0,398
15	H	H	CH_2OH	NH_2	OMe	H	H	-0,732	-0,613
16	H	H	CH_2Br	NH_2	OMe	H	H	-0,556	-0,663
17	H	H	$(CH_2)_2OCOOMe$	NH_2	OMe	H	H	-0,380	-0,114
18	H	H	$(CH_2)_2OCO(CH_2)_2CH_3$	NH_2	OMe	H	H	-0,491	-0,398
19	H	H	$(CH_2)_2OCOCH_2CH(CH_3)_2$	NH_2	OMe	H	H	-0,342	Tox [a]
20	H	H	$(CH_2)_2OCOPh$	NH_2	OMe	H	H	0,398	0,699
21	H	H	$(CH_2)_2OCOPhF$	NH_2	OMe	H	H	-0,114	0,222
23	H	H	$(CH_2)_2OCOPhOMe$	NH_2	OMe	H	H	-0,279	Tox [a]
24	H	H	CH_3	Cl	OMe	H	H	-0,041	-0,362
25	H	H	CH_2OH	Cl	OMe	H	H	-2,262	-0,959
26	H	H	CH_2Br	Cl	OMe	H	H	-1,703	-1,170
27	H	H	$(CH_2)_2OCOOMe$	Cl	OMe	H	H	-2,178	-1,877
28	H	H	$(CH_2)_2OCO(CH_2)_2CH_3$	Cl	OMe	H	H	-1,707	-1,628

N°	R_1	R_2	R_3	R_4	R_5	R_6	R_7	pIC_{50} [b]	
								PRO[c]	AMA[d]
29	H	H	$(CH_2)_2OCOCH_2CH(CH_3)_2$	Cl	OMe	H	H	-1,446	-1,561
30	H	H	$(CH_2)_2OCOPh$	Cl	OMe	H	H	-2,135	Tox [a]
31	H	H	$(CH_2)_2OCOPhF$	Cl	OMe	H	H	-2,295	-1,645
32	H	H	$(CH_2)_2OCOPhCl$	Cl	OMe	H	H	-2,194	-2,191
33	H	H	$(CH_2)_2OCOPhOMe$	Cl	OMe	H	H	-2,098	-2,088
34	CH_2NH_2	H	H	H	H	H	H	-0,230	-0,531
35	CH_2OH	H	H	H	H	H	H	Tox [a]	-1,515
36	$CH_2NHCO(CH_2)_3Cl$	H	H	H	H	H	H	-1,655	-1,427
37	$CH_2OCO(CH_2)_3Cl$	H	H	H	H	H	H	-2,200	-2,185
38	$CH_2NHCOCH=CH_2$	H	H	H	H	H	H	-2,245	-0,833
39	$CH_2OCOCH=CH_2$	H	H	H	H	H	H	-1,464	---
40	$CH_2NHCOPh$	H	H	H	H	H	H	-1,654	-1,121
41	CH_2OCOPh	H	H	H	H	H	H	-1,885	Tox [a]
42	CH_2NHCO-p-PhF	H	H	H	H	H	H	-1,815	---
43	CH_2OCO-p-PhF	H	H	H	H	H	H	-1,741	Tox [a]
44	CH_2NHCO-p-PhCl	H	H	H	H	H	H	Tox [a]	-1,004
45	CH_2OCO-p-PhCl	H	H	H	H	H	H	-1,790	-1,433
46	CH_2NHCO-p-PhOMe	H	H	H	H	H	H	-1,513	-0,973
47	CH_2OCO-p-PhOMe	H	H	H	H	H	H	-1,471	-0,869
48	CH_2NHCO-p-PhNMe$_2$	H	H	H	H	H	H	-1,539	-0,672
49	CH_2OCO-p-PhNMe$_2$	H	H	H	H	H	H	-1,819	Tox [a]
50	CH_2NH_2	H	H	H	H	H	CH_2NH_2	-0,820	-0,531
51	CH_2OH	H	H	H	H	H	CH_2OH	Tox [a]	0,222
52	$CH_2NHCO(CH_2)_3Cl$	H	H	H	H	H	$CH_2NHCO(CH_2)_3Cl$	-0,663	-0,813
54	$CH_2NHCOCH=CH_2$	H	H	H	H	H	$CH_2NHCOCHCH_2$	-1,061	Tox [a]
56	$CH_2NHCOPh$	H	H	H	H	H	$CH_2NHCOPh$	-0,556	-1,236
57	CH_2OCOPh	H	H	H	H	H	CH_2OCOPh	Tox [a]	-0,716
58	CH_2NHCO-p-PhF	H	H	H	H	H	CH_2NHCO-p-PhF	-0,756	-0,255
60	CH_2NHCO-p-PhCl	H	H	H	H	H	CH_2NHCO-p-PhCl	---	-1,562
62	CH_2NHCO-p-PhOMe	H	H	H	H	H	CH_2NHCO-p-PhOMe	-1,797	Tox [a]
63	CH_2OCO-p-PhOMe	H	H	H	H	H	CH_2OCO-p-PhOMe	Tox [a]	-1,825
64	CH_2NHCO-p-PhNMe$_2$	H	H	H	H	H	CH_2NHCO-p-PhNMe$_2$	-0,940	-0,724
65	CH_2OCO-p-PhNMe$_2$	H	H	H	H	H	CH_2OCO-p-PhNMe$_2$	Tox [a]	-1,667

[a] Toxic: toxicity observed on human macrophages at concentrations that did not display antileishmanial activity; [b] $pIC_{50} = -\log(IC50)$; [c] pIC_{50} AMA: antileishmanial activity against Promastigote parasite form; [d] pIC_{50} PRO: antileishmanial activity against Amastigote parasite form; -p-: para; -m-: meta

Mean squared error (*MSE*) measures the average of the squares of the errors or deviations of the predictions from the true values; P_{value} is the probability (*P*) of Fisher statistics (*F*), which gives an indication of the probability that a QSAR is a chance correlation.

In order to assess the significance of the models and accurate prediction ability for new compounds:

(i) We use an internal validation procedure (leave-one-out cross-validation), whereby one compound is removed and the rebuilt model with the remaining molecules is used to predict the response of the eliminated compound. This one is then returned and a second is removed, and the cycle is repeated, and so on, until all compounds have been removed one by one, and an overall correlation coefficient r_{cv} is computed;

(ii) After the model is built, an external prediction is necessary. In this one, the obtained model was used to predict the activities of a test set comprising compounds that are similar to those used in the training set. This is usually performed by splitting a dataset into a training set and a test set, typically in a 1/5 ratio. Further, before performing the external validation of a model, it is very important to check for the presence of systematic error that violates the basic assumptions of the least squares regression model. If high systematic error (bias) is present in the model, then such model should be discarded and performing any external validation test is of no use on such biased model. *Xternal Validation Plus* is a tool that checks the presence of systematic errors in the model and further computes all the required external validation parameters, while judging the performance of actual prediction quality of a QSAR model based on recently proposed MAE-based criteria **[30]**;

(iii) A model is valid only within its training domain and new molecules must be considered as belonging to the domain before the model is applied (OECD Principle 3 **[31]**). Without applicability domain (AD), each model can predict the activity of any compound, even with a completely different structure from those included in the study. Therefore, the AD is a tool to find out compounds that are outside the applicability domain of the built QSAR model and it detects outliers present in the training set compounds. There are several methods for defining the applicability domain (AD) of QSAR models **[32]**, but the most common one is determining the leverage values h_i ($h_i = x_i^T (X^T X)^{-1} x_i$ ($i = 1,2,\dots,n$)) for each compound, where x_i is the descriptor row-vector of query compound, X is $n * (k-1)$ matrix of k model descriptor values for n training set compounds, and the superscript T refers to the transpose of matrix/vector **[32, 33]**. In this study, we use Williams plot; in this plot, the applicability domain is established inside a squared area within standard deviation $\pm x$ (in this study $x = 3$;

"three-sigma rule" [34]) and a leverage threshold h^* ($h^* = 3 * (k + 1)/n$) [35], where n is the number of training set compounds; k is the number of model descriptors.

The leverage (h) greater than the warning leverage h^*, suggested that the compound was very influential on the model [36]. The results of the leverage approach were compared with that of the simple approach introduced by Roy et al. [37].

2.3.1. Multiple Linear Regression (MLR)

The descendent multiple linear regression (MLR) analysis, based on the elimination of non-significant descriptors (one by one) until a valid model (including the critical probability: $P_{value} < 0.05$ for all descriptors and the model complete), was employed to find a linear model of the activity of interest, which takes the following form:

$$Y = a_0 + \sum_{i=1}^{n} a_i x_i \quad (1)$$

Where: Y is the studied activity (the dependent variable); a_0 is the intercept of the equation; x_i are the molecular descriptors; a_i are the coefficients of those descriptors.

This method is one of the most popular methods of QSAR/QSPR thanks to its simplicity in operation, reproducibility, and ability to allow easy interpretation of the features used. The important advantage of the linear regression analysis is that it is highly transparent; therefore, the algorithm is available and predictions can be made easily [38]. It has served also to select the descriptors used as the input parameters in the artificial neural network (ANN).

2.3.2. Artificial Neural Networks (ANN)

The artificial neural networks (ANN) are used in order to increase the probability of characterizing the compounds and to generate a predictive QSAR model between the set of molecular descriptors obtained from the MLR models and the observed activities values. The ANN model is done on the *Matlab* software. It consists of three layers of neurons, called input layer, hidden layer, and output layer (Figure 2). The input layer formed by a number of neurons equal to the number of descriptors obtained in the multiple linear regression models and the output layer represents the calculated activity values. For determination of the number of hidden neurons in the hidden layer, where all calculations of parameter optimization of neural networks are made, a parameter ρ has been proposed. The parameter ρ plays a major role in determining the best artificial neural network architecture [39, 40]; ρ is defined as follows:

$$\rho = \frac{\text{Number of data points in the training set}}{\text{Sum of the number of connections in the ANN}} \quad (2)$$

In order to avoid over-fitting or under-fitting, it is recommended that the value of ρ should be between 1.00 and 3.00; if ρ < 1, the network simply memorizes the data, whereas if ρ > 3, the network is not able to generalize **[41]**.

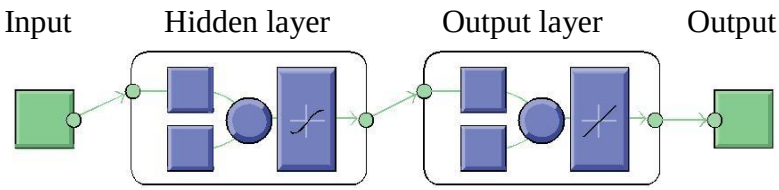


Figure 2: Architecture used in our study of the artificial neural network.

2.4. Software packages used in this QSAR development study

There are various, free and commercial, software available for QSAR development. These include specialized software for drawing chemical structures, generating 3D structures, calculating chemical descriptors, and developing QSAR models. The software packages used in our works are represented in table 2.

Table 2: Software packages used in this work

Drawing chemical structures	Marvin Sketch, ACD/ChemSketch, and ChemBioDraw
Generating 3D structures	Gauss View 3.0 and ChemBio3D
Calculating chemical descriptors	Gaussian 03, Marvin Sketch 6.2, ChemSketch, and ChemBio3D
Developing QSAR models	XLSTAT 2009 and Matlab 7.9.0 (version 2011)

3. Results and Discussions

3.1. Principal Component Analysis (PCA)

PCA is a useful statistical technique for summarizing all the information encoded in the structures of the compounds. It is very helpful for understanding the distribution of the compounds. This is an essentially descriptive statistical method, which aims to extract the maximum of information contained in the dataset compounds **[42, 43]**. In this work, the PCA is used to overview the examined compounds for similarities and dissimilarities and to select descriptors that show a high correlation with the response activity; this one gives extra weight because it will be more effective at prediction **[44].**

Tables 3 and 4 present the descriptors with a correlation coefficient with the activity higher, in absolute value, than 0.1. The absence of any serious colinearity between the descriptors present in the model was confirmed by the correlation matrix.

Table 3: The absolute values of correlation coefficients with descriptors and activities

Promastigotes			
Descriptors	*r*	Descriptors	*r*
Henry's law constant (K_H)	**0.130**	Dipole moment (μ)	**0.213**
Ideal gas thermal capacity (IGTC)	**0.108**	Electronegativity (χ)	**0.599**
Number of H-Bond acceptors NHA	**0.299**	Electrophilicity index (ω)	**0.630**
Number of H-Bond donors (NHD)	**0.250**	Molecular topological index (MTI)	**0.116**
Balaban index (*J*)	**0.133**	Polar surface area (PSA)	**0.282**
Shape coefficient (*I*)	**0.135**	Total connectivity (TC)	**0.269**
Total valence connectivity (TVC)	**0.292**	Wiener index (*W*)	**0.124**
Percent ratios of Nitrogen N%	**0.238**	E_{HOMO}	**0.554**
Percent ratios of Hydrogen H%	**0.260**	E_{LUMO}	**0.632**
Gibbs free energy (G)	**0.119**	E_{Gap}	**0.227**
Amastigotes			
Descriptors	*r*	Descriptors	*r*
Total energy, *E*	**0.160**	Dipole moment (μ)	**0.224**
Electrophilicity index (ω)	**0.335**	E_{LUMO}	**0.338**
Electronegativity (χ)	**0.316**	E_{HOMO}	**0.286**
Polar surface area (PSA)	**0.391**	Percent ratios of Nitrogen N%	**0.436**
Percent ratios of Oxygen O%	**0.105**	Percent ratios of Carbon C%	**0.155**
Index of refraction (*n*)	**0.548**	Surface tension (γ)	**0.553**
Density (d)	**0.242**	Boiling point (TB)	**0.254**
Critical pressure (CP)	**0.176**	Critical temperature (CT)	**0.276**
Gibbs free energy (G)	**0.195**	Heat of formation (*H*°)	**0.152**
Henry's law constant (KH)	**0.387**	Partition coefficient Log *P*	**0.248**
Number of H-Bond acceptors (NHA)	**0.198**	Melting point (*T*)	**0.314**
Number of H-Bond donors (NHD)	**0.587**	Partition coefficient (PC)	**0.163**
Number of rotatable bonds (NRB)	**0.139**	Shape coefficient *(I)*	**0.176**
Sum of valence degrees (SVD)	**0.146**	Winner index *(W)*	**0.103**

Table 4: Descriptors selected by the principal component analysis (PCA) and software packages used in the calculation of descriptors

Software	Descriptors
Gaussian 03	Highest occupied molecular orbital energy E_{HOMO}(eV) ; lowest unoccupied molecular orbital energy E_{LUMO}(eV) ; hardness η(eV)=((E_{LUMO}-E_{HOMO}))/2; electronegativity χ(eV)=-(E_{LUMO}+E_{HOMO})/2; electrophilicity index ω(eV) =χ^2/2η; total energy E(eV); dipole moment μ(Debye); energy gap between E_{HOMO} and E_{LUMO}values E_{Gap}(eV)
ChemOffice	Heat of formation $H°$ (kJ mol^{-1}); Gibbs free energy G (kJ mol^{-1}); ideal gas thermal capacity (IGTC) (J mol^{-1} K^{-1}); melting point (T) (Kelvin); critical temperature (CT) (Kelvin); boiling point (TB) (Kelvin); critical pressure (CP) (Bar); Henry's law constant (KH); total valence connectivity (TVC); partition coefficient (PC); molecular topological index (MTI); number of rotatable bonds (NRB); shape coefficient (I); sum of valence degrees (SVD); total connectivity (TC);
ChemSketch	Percent ratios of nitrogen, hydrogen, oxygen and carbon atoms (N%; H%; O% and C%); surface tension γ (dyne/cm); index of refraction (n); density (d)
Marvin Sketch	Log P; Winner index (W); number of H-Bond acceptors (NHA); number of H-Bond donors (NHD); Balaban index (J); polar surface area (PSA) ($A°$)^2

3.2. Univariate Partitioning (UP)

The aim of the UP was the recognition of groups of objects based on their similarity; this method is based on the criterion of partitioning proposed by Fisher **[45]**. In this study, the division of the dataset into training and test sets has been performed. In this one, from each obtained cluster, one compound for the training set was selected randomly to be used as test set compound. The partitioning results are given in table 5.

Table 5: The partitioning results of compounds into training and test sets

Classes	The test set compounds	
	Promastigotes	Amastigotes
	3, 6, 10, 14, 17, 21, 37, 38, 42, 64	**5, 6, 26, 27, 29, 36, 37, 44, 46, 50**
1	1, 8, **37**, 40, 56	1, 10, 15, 16, 34, 48, **50**, 57, 64
2	2, **14**, 29, 34, 48	2, 25, 38, **44**, 47, 51, 52
3	**3**, 18, 19	3, **26**, 40, 56
4	4, 24, 25, 27, 30, **42**, 58	4, 7, 8, 11, 13, 17, 21, **46**
5	**6**, 26	**5**, 14, 20
6	7, 9, 12, **21**, 33, 39, 41, 47	**6**
7	**10**, 11, 16, 31, 32, 43, 45	18, **29**, 58
8	15, **17**, 23, 28, 36, 49, 52, 54	**27**, 63
9	20, **38**, 46, 62	24, 28, 31, 35, **36**, 45, 60, 65
10	50, **64**	32, 33, **37**

3.3. Multiple Linear Regression (MLR)

The QSAR analysis was performed using the values of the chemical descriptors selected by the PCA method and, on the other hand, the experimental values of the antileishmanial activities for 60 of the acridines derivatives (effect). **Table S1** in Supplementary Material shows the value of each molecular descriptor that is configured in established MLR models and the QSAR model built is represented by the following equations and the values of the statistical parameters.

- MLR model for pIC_{50} promastigotes:

$$pIC_{50}\ PRO = 3.030 - 8.214\ 10^{-02}\ N\% + 0.239\ \mu - 1.264\ \omega - 0.233\ NHA + 0.732\ NHD + 3.311\ 10^{-05}\ MTI + 3.184\ 10^{04}\ TVC \quad (3)$$

Statistical parameters: r^2=0.723; r=0.850; r^2_{adj}=0.664; *MSE*=0.189; P_{value} <10^{-04}; F=12.293

- MLR model for pIC_{50} amastigotes:

$$pIC_{50}\ AMA = -7.314 + 1.191\ E_{HOMO} + 2.543\ 10^{-02}\ CT - 0.350\ K_H - 1.632\ 10^{-02}\ T + 0.652\ NHA + 1.262\ NHD \quad (4)$$

Statistical parameters: r^2=0.663; r=0.814; r^2_{adj}=0.598; *MSE*=0.208; P_{value} <10^{-04}; F=10.195

(i) For the two models, P_{value} is lower than 0.0001; it means that we would be taking a lower than 0.01% risk in assuming that the null hypothesis (no effect of the explanatory variables) is wrong and that the regressions equations have statistical significance. Therefore, we can conclude with confidence that the selected variables do bring a significant amount of information.

(ii) Higher value of r^2 and r^2_{adj} and lower mean squared error (MSE) indicate that the two proposed models are predictive and reliable.

(iii) Our obtained models were validated internally by the leave-one-out cross-validation technique; the cross-validation coefficient r^2_{cv} for the two models was determined based on the predictive ability of the model. The value of r^2_{cv} is higher than 0.5 (r^2_{cv}=0.536 for the MLR pIC_{50} PRO model and r^2_{cv}=0.525 for the MLR pIC_{50} AMA model); indicating good predictability of the models.

(iv) True predictive power of these models is to test their ability to predict perfectly pIC_{50} of compounds from an external test set (compounds that were not used for the developed model); pIC_{50} of the remaining set of 10 compounds are deduced from the quantitative models proposed with the compounds used in training set by MLR. These models will be able to predict the activities of test set molecules in agreement with the experimentally determined value. The observed and calculated pIC_{50} values are given in table 6. The predictive capacity of the models was judged, the higher value of r^2_{test} (r^2_{test}=0.660 for the MLR pIC_{50} (PRO) model and

r^2_{test}=0.718 for the MLR pIC_{50} (AMA) model) indicates the improved predictability of the model.

(v) *Xternal Validation Plus* indicates the absence of systematic errors in the model and a moderate performance of prediction quality of a QSAR model based on proposed MAE-based criteria (**Table S2** in Supplementary Material).

(vi) The values of calculated activities from (3) and (4) are given in **table S1** and the correlations of calculated and observed activities values are illustrated in figure 3.

In the first model (pIC_{50} PRO), the descriptors influencing negatively the activities are the percent ratio of nitrogen (N%), the number of H-Bond acceptors (NHA), and the electrophilicity index (ω), and the parameters influencing positively the activities are the dipole moment (μ), the number of H-Bond donors (NHD), the molecular topological index (MTI), and the total valence connectivity (TVC).

(i) The number of H-Bond acceptors (NHA) has a negative sign in the model, which suggests that increased activity can be achieved by decreasing the number of heteroatoms (nitrogen or oxygen atoms), mostly the nitrogen ones for decreasing the ratio of nitrogen (N%) having also a negative sign in the model.

(ii) The electrophilicity index ω has a negative sign in the model, which suggests that increased activity can be achieved by decreasing the electrophilicity of the acridine derivatives (a high value of electrophilicity describes a good electrophile, while a small value of electrophilicity describes a good nucleophile).

(iii) The dipole moment μ has a positive sign in the model, which suggests that increased activity can be achieved by increasing the polarity of the acridine derivatives.

(iv) The number of H-Bond donors (NHD) has a positive sign in the model, which suggests that increased activity can be achieved by increasing of the heteroatom with one or more hydrogen atoms.

(v) The molecular topological index (MTI) coefficient (topological parameter) has a positive sign in the model, which suggests that increased activity can be achieved by increasing the flexibility of the substituent side chain.

(vi) The total valence connectivity (TVC) has a positive sign in the model, which suggests that increased activity can be achieved by increasing in branching of the acridine derivatives, because the overall connectivity increases with both molecule size and complexity, as expressed in branching and cyclicity of molecular skeleton **[46]**.

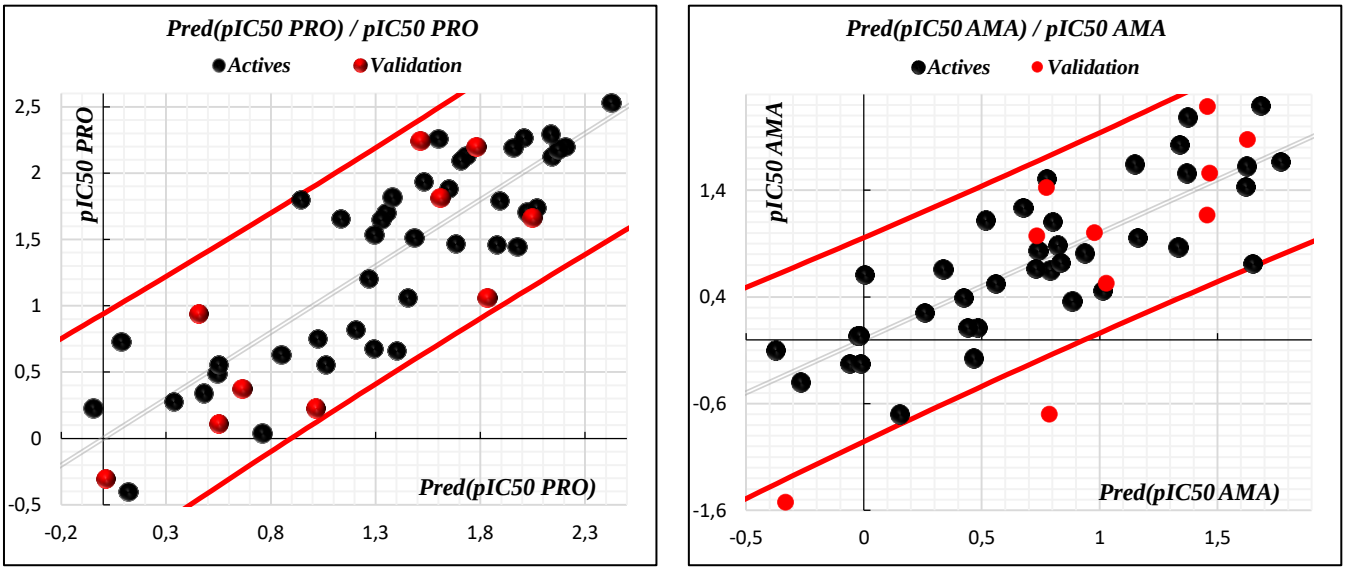


Figure 3: Correlations of observed and predicted activities (training set in black and test set in red) values calculated using MLR models.

Table 6: The values of chemical descriptors and the observed and predicted activities using the MLR models for the test set.

Test set (promastigotes)								pIC_{50} PRO	
N°	N%	μ	ω	NHA	NHD	MTI	TVC	Obs.	MLR
3	12.030	4.259	5.024	3	2	14075	$3.891\ 10^{-07}$	−1.669	−2.047
6	9.270	5.640	4.654	5	2	28763	$1.716\ 10^{-10}$	−0.230	−1.014
10	10.780	5.730	5.267	3	2	16459	$4.902\ 10^{-08}$	−1.061	−1.832
14	11.760	2.698	3.062	3	1	4387	$3.050\ 10^{-05}$	0.301	−0.012
17	8.580	4.969	3.431	4	1	10329	$5.188\ 10^{-07}$	−0.380	−0.664
21	7.180	5.214	3.498	5	1	18600	$1.334\ 10^{-08}$	−0.114	−0.551
37	4.460	6.939	4.779	2	0	8245	$4.076\ 10^{-06}$	−2.200	−1.781
38	10.680	5.122	4.374	2	1	6388	$5.083\ 10^{-06}$	−2.245	−1.512
42	8.480	5.260	4.457	3	1	12005	$1.307\ 10^{-07}$	−1.815	−1.608
64	13.170	5.297	4.388	5	2	47653	$1.201\ 10^{-10}$	−0.940	−0.455
Test set (amastigotes)								pIC_{50} AMA	
N°	E_{HOMO}	K_H	T	NHA	NHD	CT		Obs.	MLR
5	−5.850	1064.61	1300.866	3	2	19.110		0.699	−0.787
6	−5.785	1005.95	1247.456	5	2	18.715		1.523	0.332
26	−5.665	583.700	926.651	2	0	7.815		−1.170	−1.456
27	−5.603	622.270	930.050	3	0	8.827		−1.877	−1.628
29	−5.640	618.850	938.380	3	0	9.007		−1.561	−1.467
36	−5.966	676.650	1037.977	2	1	12.201		−1.427	−0.774
37	−6.025	552.770	930.967	2	0	8.348		−2.185	−1.457
44	−5.919	749.400	1083.064	2	1	12.828		−1.004	−0.978
46	−5.790	752.980	1078.419	3	1	13.925		−0.973	−0.733
50	−5.461	625.750	913.563	3	2	13.449		−0.531	−1.028

In the second model (pIC_{50} AMA), the descriptors influencing negatively the activity are Henry's law constant (K_H) and the melting point (T), and the parameters influencing positively the activities are E_{HOMO} energy, the critical temperature (CT), the number of H-Bond acceptors (NHA), and the number of H-Bond donors (NHD).

(i) Henry's law constant (K_H) has a negative sign in the model, which suggests that increased activity can be achieved by decreasing solubility; consequently, by decreasing the polarity, $K_H = c/p$ (c: solubility; p: pressure).

(ii) The melting point (T) has a negative sign in the model, which suggests that increased activity can be achieved by decreasing the polarity and the branching of molecule (increasing branching makes the molecule less compact. As the surface area of the molecule decreases, it will become more compact and thus easier to pack).

(iii) The highest occupied molecular orbital energy E_{HOMO} (negative values) has a positive sign in the model, which suggests that the higher of E_{HOMO}, the strong donating electron ability, is showing the fact that the nucleophilic reaction occurs more easily and the activity of the compound is higher.

(iv) The critical temperature (CT) has a positive sign in the model, which suggests that increased activity can be achieved by increasing the critical temperature.

(v) The number of H-Bond acceptors (NHA) has a positive sign in the model, which suggests that increased activity can be achieved by increasing the number of heteroatoms (nitrogen or oxygen atoms).

(vi) The number of H-Bond donors (NHD) has a positive sign in the model, which suggests that increased activity can be achieved by increasing of the heteroatoms with one or more hydrogen atoms.

Comparing the importance of each descriptor on pIC_{50} of acridines, we must know the standardized coefficient and the t-test values of them in the MLR equations. The bigger the absolute value of the t-test value is, the greater the influence of the descriptor is.

In (3), the t-test values are 1.773, −4.035, 7.408, 1.720, −3.361, −2.429, and −4.570 for N%, μ, ω, NHA, NHD, MTI, and TVC, respectively. Moreover, in (4), the t-test values are 3.196, 3.556, −3.764, −3.802, −3.838, and −6.012 for T, K_H, CT, E_{HOMO}, NHA, and NHD, respectively.

This meant that the t-test values of ω, μ, NHD, and TVC are both larger than those of the other descriptors, which indicates that, in this model, the influence of these descriptors on activity is stronger than that of the others. It shows also the importance of electrophilicity index in the prediction of pIC_{50} PRO. Moreover, the t-test value of NHD is larger than that of the other descriptors, which indicates that the influence of this descriptor, in this model, on activity is stronger than that of the others.

In the conclusion, these results illustrate that, to increase the antileishmanial activity against promastigotes parasites, we will decrease the electrophilicity and increase the branching, the polarity, and the number of hydrogen atoms attached in the heteroatom of the acridine derivatives. Moreover, to increase the antileishmanial activity against amastigotes parasites, we will decrease the solubility, the polarity, and the branching and increase the electrophilicity and the number of heteroatoms mostly attached in the hydrogen atoms and the critical temperature of the acridine derivatives.

3.4. Applicability Domain (AD)

The applicability domain (AD) of these models was evaluated by leverage analysis expressed as Williams plot (Figures 4 and 5), in which the standardized residuals and the leverage threshold values (h^* = 0.585 and 0.553 for pIC_{50} (PRO) and pIC_{50} (AMA), resp.) were plotted. Any new value of predicted pIC_{50} data must be considered reliable only for those compounds that fall within this AD on which the model was constructed.

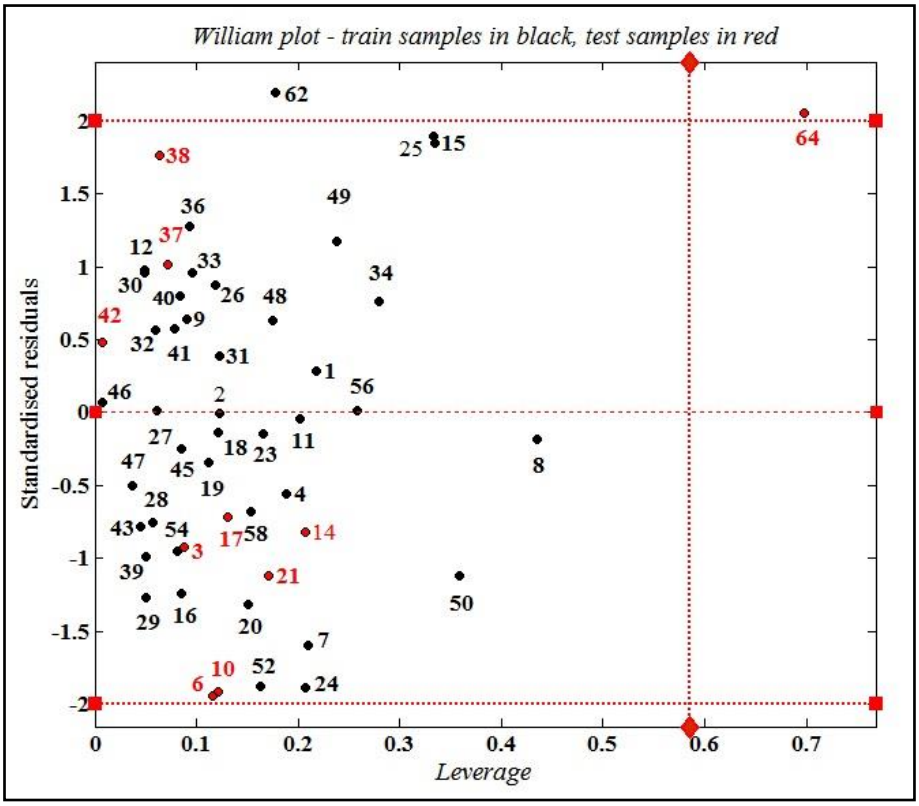


Figure 4: Williams plot of standardized residual versus leverage for pIC_{50} (PRO) MLR model (with $h*$ = 0.585 and residual limits= ±3).

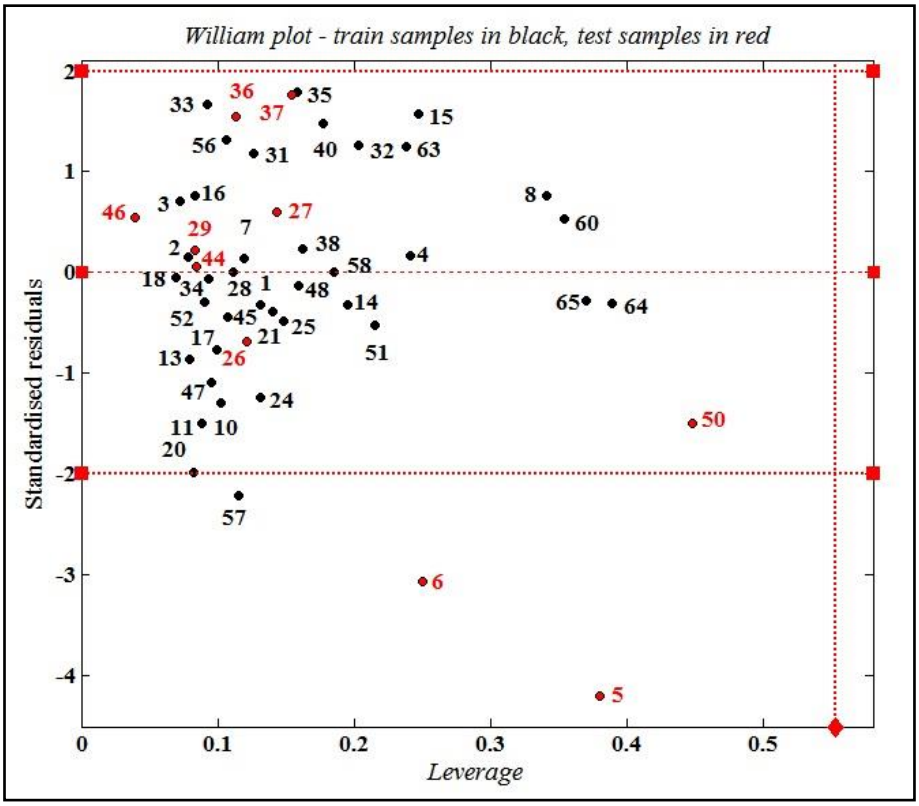


Figure 5: Williams's plot of standardized residual versus leverage for pIC_{50} (AMA) MLR model (with $h* = 0.553$ and residual limits = ±3).

From these figures, it is obvious that there is no response outlier in training set and no response outside in test set. Only one chemical is identified as outside for MLR model for pIC_{50} PRO and two chemicals for MLR model for pIC_{50} AMA; these outsides compounds are given as follows.

For pIC_{50} PRO: Compound number 64 in test set has higher leverage which is greater than h^* value of 0.585 and all the compounds have a standard deviation into $\pm x$ interval (x =3).

For pIC_{50} AMA: Compounds numbers 5 and 6 in test set have a standard deviation value greater than the $\pm x$ interval (x = 3).

These results are confirmed using the simple approach (**Tables S3 and S4**), the chemical number 8 is identified as outlier and the chemical number 64 is identified as outside for MLR model for pIC_{50} PRO; and any chemical is identified as outside or outlier for MLR model for pIC_{50} AMA.

These erroneous predictions could probably be attributed to wrong experimental data or to the structure of these outsides; Cl, F, or NMe_2 substitutes the tree compounds at the para position of the phenyl ring (Figure 6); maybe the selected descriptors do not pay much attention to these special substructures. The predictions of tree compounds are extrapolations of the model, but fortunately they are all "good leverage" chemicals.

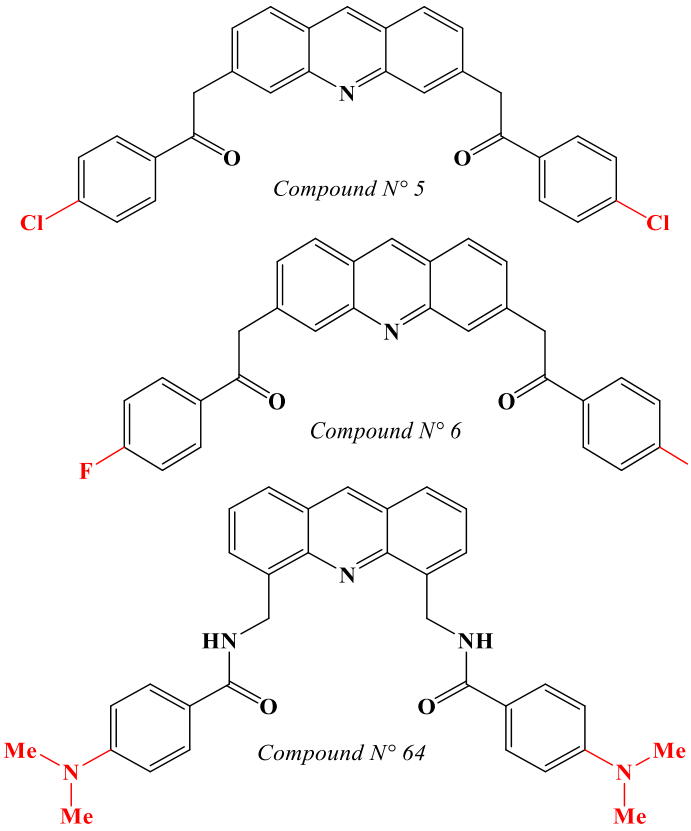


Figure 6: Chemical structures of the outsides compounds.

3.5. Artificial Neural Network (ANN)

In this study, we used a feed forward network with two layers, with a sigmoid transfer function in the hidden layer and a linear transfer function in the output layer (Figure 2). The architecture of the artificial neural network used in this work was 7-3-1 and ρ = 1.82 for pIC_{50} (PRO) and 6-3-1 and ρ = 2 for pIC_{50} (AMA).

The output layer represents the calculated activity values (predicted pIC_{50}). The calculation result and the performances of established models are recorded in the output layer.

To justify the predictive quality of models, total data are distributed randomly into three groups. The first group (70% of the total data) used to drive the system. The second group (15% of the total data) will be used to validate the network; and the remaining 15% that did not participate in the learning models will be used as an independent test of network generalization. The distribution of the total data into training, validation, and test sets is shown in table 7.

The correlation between the experimental and calculated values using the artificial neural network models is highly significant, as illustrated in figure 7 and as indicated by better r and r^2 and the small MSE values for all three phases: training, validation, and test

(Table 7). The predicted activities calculated with the artificial neural network and the observed values are given in **table S5**.

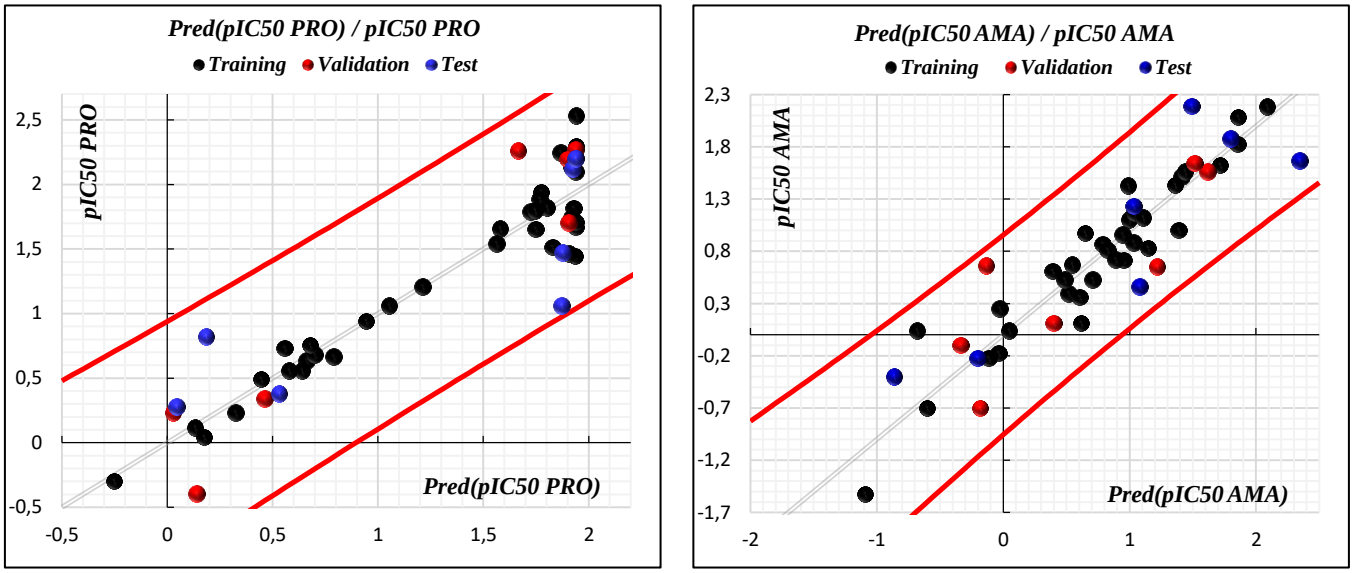


Figure 7: Correlations between observed and predicted activities values calculated using ANN models (training set in black, validation set in red, and test set in blue).

Table 7: The coefficient of determination, the coefficient of correlation, and the MSE obtained by the model established by the ANN for the three phases: Training, validation, and test.

pIC_{50} PRO		
Model	**n= 51; MSE = 0.079; r^2 = 0.870; r = 0.933**	
Training set	[1; 3; 4; 5; 6; 7; 11; 12; 13; 14; 16; 19; 21; 25; 26; 27; 28; 30; 32; 34; 35; 36; 37; 38; 39; 40; 41; 43; 44; 46; 47; 48; 49; 50; 51]	MSE = 0.044; r^2 = 0.914; r = 0.956
Validation set	[8; 17; 18; 22; 23; 24; 29; 31]	MSE = 0.125; r^2 = 0.922; r = 0.960
Test set	[2; 9; 10; 15; 20; 33; 42; 45]	MSE = 0.186; r^2 = 0.725; r = 0.851
pIC_{50} AMA		
Model	**n= 48; MSE = 0.104; r^2 = 0.918; r = 0.843**	
Training set	[2; 3; 4; 5; 6; 7; 10; 11; 13; 16; 18; 19; 20; 21; 23; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 41; 43; 44; 45; 46; 47]	MSE = 0.060; r^2 = 0.897; r = 0.947
Validation set	[1; 8; 14; 15; 17; 24; 25]	MSE = 0.198; r^2 = 0.710; r = 0.843
Test set	[9; 12; 22; 26; 40; 42; 48]	MSE = 0.228; r^2 = 0.791; r = 0.890

The results obtained by MLR and ANN are very sufficient to conclude the performance of the models. A comparison of the quality of the statistical terms of these models shows that the ANN has substantially better predictive capability.

ANN was able to establish a more satisfactory relationship between the molecular descriptors and the activity of the studied compounds compared to MLR; but the most negative side of this method is the fact that it is poorly transparent, whereas transparency of MLR approach gives the most interpretable results and gives a good explanation of activities with descriptors. Consequently, we can design new compounds with improved values of activity compared to the studied compounds using the MLR models. Taking into account the above results, we added suitable substitutions and then we moved to calculate their activities using the proposed model's equations ((3) and (4)). Therefore, the suggested models will reduce the time and cost of synthesis as well as the determination of the antileishmanial activities against promastigotes and amastigotes forms of parasites for the acridine derivatives.

3.6. Proposed novel compounds with higher Antileishmanial activities values

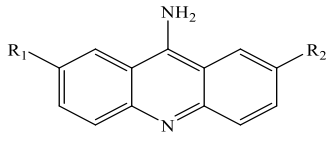
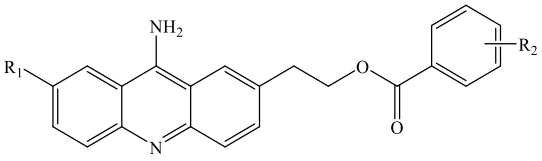
According to the above discussions, the MLR models could be applied to other acridines derivatives according to table 1 and could add further knowledge in the improvement of new way in antileishmanial drug research. If we develop a new compound with better values than the existing ones, it may give rise to the development of more active compounds than those currently in use.

In this way, we carried out structural modification starting from compounds having the highest pIC_{50} values as template (number 14 and number 20). The structures of the designed compounds and their parameter values calculated by the same methods as well as pIC_{50} values theoretically predicted by the MLR models are listed in table 8.

From the predicted activities, it has been observed that the designed compounds have higher pIC_{50} values than the existing compounds in the case of the 60 studied compounds (Table 1).

The leverage values (h) for the new designed compounds are 0.268, 0.602, 0.775, 0.527, 0.386, and 0.369 for $X1$, $X2$… $X6$, respectively, and 0.649, 0.797, 0.201, 0.251, 0.235, 0.166, 0.176, 0.339, 0.169, and 0.166 for $Y1$, $Y2$… $Y10$, respectively (Table 8). Only compounds $X2$, $X3$, $Y1$, and $Y2$ are defined as outliers and consequently they are not being considered, because they have higher leverage which is greater than h^* (0.585 for pIC_{50} (PRO) and 0.553 for pIC_{50} (AMA)); we suggest all other compounds as candidates that will be synthetized and evaluated as antileishmanial drugs.

Table 8: Values of descriptors, antileishmanial activity pIC_{50} for the new designed compounds

Designed Compounds	N°	Descriptors values	pIC_{50} (PRO)	N°	Descriptors values	pIC_{50} (AMA)
	X_1	R1= OH; R2=CH3 $N_{\%}$ = 12.490; μ = 2.232; ω = 3.099 NHA = 3; NHD = 2; MTI = 3651; TVC = 3.34 10^{-05}	**0.571**	Y_1	R_1=OCH_3; R_2=CF_3 E_{HOMO} = -5.235; CT = 885.365; KH = 9.302; T = 581.430 ; NHA = 6 ; NHD = 1	**1.395**
				Y_2	R_1=OH; R_2=CF_3 E_{HOMO} = -5.312; CT = 903.801; KH = 12.056; T = 647.130 ; NHA = 6 ; NHD = 2	**0.998**
				Y_3	R_1=OCH_3 ; R_2=F E_{HOMO} = -5.034; CT = 893.897; KH = 10.174; T = 566.560 ; NHA = 4 ; NHD = 1	**0.485**
				Y_4	R_1=R_2=OC_2H_5 E_{HOMO} = -4.696; CT = 927.241; KH = 11.222; T = 622.010; NHA = 4 ; NHD = 1	**0.464**
	X_2	R1= OH; R_2=OH $N_{\%}$ = 12.380; μ = 1.080; ω = 3.131 NHA = 4; NHD = 3; MTI = 3497; TVC = 1.49 10^{-05}	**0.170**	Y_5	R_1=R_2=OCH_3 E_{HOMO} = -4.731; CT = 916.967; KH = 11.469; T = 599.470 ; NHA = 4 ; NHD = 1	**0.442**
				Y_6	R_1=OCH_3; R_2=$OC(CH_3)_3$ E_{HOMO} = -4,872; CT = 928,656; KH = 11,099; T = 635,700 ; NHA = 4 ; NHD = 1	**0.109**
				Y_7	R_1=OCH_3 ; R_2=CH_2NH_2 E_{HOMO} = -4.801 ; CT = 934.561; KH = 14.185 ; T = 660.500 ; NHA = 4 ; NHD = 2	**0.122**
	X_3	R1=OH; R_2=meta-NH_2 $N_{\%}$ = 11.250; μ = 5.944; ω = 3.392 NHA = 5; NHD = 3; MTI = 17128; TVC = 2.23 10^{-08}	**0.838**	Y_8	R_1= OCH_3; R_2 = para, meta -F: E_{HOMO} = -5.081 ; CT = 1006.097; KH = 13.267 ; T = 762.370 ; NHA = 6 ; NHD = 1	**0.308**
	X_4	R1= OCH_3; R2=para-NH_2 $N_{\%}$ = 10.850; μ = 7.114; ω = 3.295 NHA = 5; NHD = 2; MTI = 19138; TVC = 2.04 10^{-08}	**0.607**	Y_9	R_1= OCH_3; R_2 = para NH_2: E_{HOMO} = -4.966; CT = 1066.417; KH = 16.853; T = 831.93 ; NHA = 5 ; NHD = 2	**0.199**
	X_5	R1= OCH_3; R2=meta-NH_2 $N_{\%}$ = 10.850; μ = 6.193; ω = 3.353 NHA=5; NHD = 2; MTI = 18992; TVC = 2.04 10^{-08}	**0.309**			
	X_6	R1= OH; R2=H $N_{\%}$ = 7.820; μ = 5.631; ω = 3.483 NHA = 4; NHD = 2; MTI = 15585; TVC = 4.47 10^{-08}	**0.381**	Y_{10}	R_1= OCH_3; R_2= meta NH_2: E_{HOMO} = -4.993 ; CT = 1066.417 ; KH = 16.853 ; T = 831.930 ; NHA = 5 ; NHD = 2	**0.166**

4. Conclusion

Following the five principles of the Organization for Economic Co-operation and Development (OECD) for the validation of QSAR models, two different modeling methods, multiple linear regression (MLR) and artificial neural network (ANN), were used in the construction of QSAR models for the antileishmanial activities against two forms of parasites (promastigotes and amastigotes) of acridines derivatives. The accuracy and predictability of the proposed models were proven by the comparison of key statistical terms of models. The good results obtained with the internal and external validations show that the proposed models in this paper are able to predict activities with a great performance and that the selected descriptors are pertinent. The applicability domain (AD) of the MLR model was defined. The resulting models have shown that we have established a relationship between some descriptors and the activities in satisfactory manners. The ANN results have substantially better predictive capability than the MLR, but the latter gives the most important interpretable results.

The obtained results show that, to increase antileishmanial activity against promastigotes parasites, we will increase electrophilicity and decrease branching, polarity, and number of hydrogen atoms attached in the heteroatom of the acridine derivatives. Moreover, to increase antileishmanial activity against amastigotes parasites, we will increase solubility, polarity, and branching and decrease electrophilicity and number of heteroatoms mostly attached in the hydrogen atoms and critical temperature of acridine derivatives. Comparing t-test and standardized coefficient values of descriptors indicates that the influences of the electrophilicity index (ω) on pIC_{50} (PRO) and of the number of H-Bond donors (NHD) on pIC_{50} (AMA) are stronger than those of the others.

The most important finding from this research is that we have designed and proposed new compounds with higher values of activities compared to existing ones by adding suitable substituents and calculating their activity using the regression equations. Consequently, the proposed models will reduce the time and cost of synthesis as well as the determination of the antileishmanial activities against promastigotes and amastigotes forms of parasites of acridine derivatives.

References

[1] M.R. Moein, R.S. Pawar, S.I. Khan, B.L. Tekwani, and I.A. Khan, "Antileishmanial, antiplasmodial & cytotoxic activities of 12, 16-dideoxyaegyptinone B from Zhumeria majdae Rech. f. & Wendelbo", *Phytotherapy Research*, 22, **2008**, 283–285.

[2] L.G. Rocha, J.R. Almeida, R.O. Macedo, and J.M. Barbosa-Filho, "A review of natural products with antileishmanial activity", *Phytomedicine*, 12(6-7), **2005**, 514–535.

[3] O. Kayser, A.F. Kiderlen, and S.L. Croft, "Natural products as potential antiparasitic drugs", *Studies in Natural Products Chemistry*, 26, **2002**, 779–848.

[4] WHO—World Health Organization, **2011**. www.who.int/leishmaniasis/disease epidemiology/en/index.html.

[5] S.M.B. Jeronimo, P. Duggal, R.F.S. Braz, C. Cheng, G.R.G. Monteiro, E.T. Nascimento, D.R.A. Martins, T.M. Karplus, M.F.F.M. Ximenes, C.C.G. Oliveira, V.G. Pinheiro, W. Pereira, J.M. Peralta, J.M.A. Sousa, I.M. Medeiros, R.D. Pearson, T.L. Burns, E.W. Pugh, and M.E. Wilson, "An emerging peri-urban pattern of infection with Leishmania chagasi, the protozoan causing visceral leishmaniasis in northeast Brazil", *Scandinavian Journal of Infectious Diseases*, 36(6-7), **2004**, 443–449.

[6] M.V.L. Marlet, D.K. Sang, K. Ritmeijer, R.O. Muga, J. Onsongo, and R.N. Davidson, "Emergence or re-emergence of visceral leishmaniasis in areas of Somalia, northeastern Kenya, and south-eastern Ethiopia in 2000-2001", *Transactions of the Royal Society of Tropical Medicine & Hygiene*, 97(5), **2003**, 515–518.

[7] J. Querido, "Emergency initiative to reduce leishmaniasis in Afghanistan", *The Lancet Infectious Diseases*, 4(10), **2004**, 599.

[8] B.L. Herwaldt, "Leishmaniasis", *The Lancet*, 354(9185), **1999**, 1191–1199.

[9] Control of the leishmaniases, *Geneva, Switzerland*, **2010**. www.apps.who.int/iris/bitstream/10665/44412/1/WHOTRS 949eng.pdf

[10] P. Desjeux, "The increase in risk factors for Leishmania-sis worldwide", *Transactions of the Royal Society of Tropical Medicine & Hygiene*, 95(3), **2001**, 239–243.

[11] G. Chakrabarti, A. Basu, P.P. Manna, S.B. Mahato, N.B. Mandal, and S. Bandyopadhyay, "Indolylquinoline derivatives are cytotoxic to Leishmania donovani promastigotes and amastigotes in vitro and are effective in treating murine visceral leishmaniasis", *Journal of Antimicrobial Chemotherapy*, 43(3), **1999**, 359–366.

[12] S.A. Gamage, D.P. Figgitt, S.J. Wojcik, R.K. Ralph, A. Ransijn, J. Mauel, V. Yardley, D. Snowdon, S.L. Croft, and W.A. Denny, "Structure-activity relationships for the antileishmanial and antitrypanosomal activities of 1'-substituted 9-anilinoacridines", *Journal of Medicinal Chemistry*, 40(16), **1997**, 2634–2642.

[13] M.O.F. Khan, S.E. Austin, C. Chan, H. Yin, D. Marks, S.N. Vaghjiani, H. Kendrick, V. Yardley, S.L. Croft, and K.T. Douglas, "Use of an additional hydrophobic binding site, the Z site, in the rational drug design of a new class of stronger trypanothione reductase inhibitor, quaternary alkylammonium phenothiazines", *Journal of Medicinal Chemistry*, 43(16), **2000**, 3148–3156.

[14] D. Pathak, M. Yadav, N. Siddiqui, and S. Kushawah, "Antileishmanial agents: an updated review", *Der Pharma Chemica*, 3(1), **2011**, 39–249.

[15] L. Gupta, A. Talwar, Nishi, S. Palne, S. Gupta, and P. M. S. Chauhan, "Synthesis of marine alkaloid: 8,9-Dihydrocoscin-amide B and its analogues as novel class of antileishmanial agents", *Bioorganic and Medicinal Chemistry Letter*s, 17(14), **2007**, 4075–4079.

[16] W.A. Denny, "Acridine derivatives as chemotherapeutic agents", *Current Medicinal Chemistry*, 9(18), **2002**, 1655–1665.

[17] M. Wainwright, "Acridine—a neglected antibacterial chromophore", *Journal of Antimicrobial Chemotherapy*, 47(1), **2001**, 1–13.

[18] I. Antonini, "DNA-binding antitumor agents: From pyrimido[5,6,1-de] acridines to other intriguing classes of acridine derivatives", *Current Medicinal Chemistry*, 9(18), **2002**, 1701–1716.

[19] M. Demeunynck, F. Charmantray, and A. Martelli, “Interest of acridine derivatives in the anticancer chemotherapy”, *Current Pharmaceutical Design*, 7(17), **2001**, 1703–1724.
[20] S.A. Gamage, D.P. Figgitt, S.J. Wojcik, R.K. Ralph, A. Ransijn, J. Mauel, V. Yardley, D. Snowdon, S.L. Croftand, and W.A. Denny, “Structure-activity relationships for the antileishmanial and antitrypanosomal activities of 1 -substituted 9-anilinoacridines”, *Journal of Medicinal Chemistry*, 40(16), **1997**, 2634–2642.
[21] C.M. Mesa-Valle, J. Castilla-Calvente, M. Sanchez-Moreno, V. Moraleda-Lindez, J. Barbe, and A. Osuna, “Activity and mode of action of acridine compounds against Leishmania donovani”, *Antimicrobial Agents and Chemotherapy*, 40(3), **1996**, 684– 690.
[22] C. Di Giorgio, K. Shimi, G. Boyer, F. Delmas, and J.P. Galy, “Synthesis and antileishmanial activity of 6-mono-substituted and 3,6-di-substituted acridines obtained by acylation of proflavine”, *European Journal of Medicinal Chemistry*, 42(10), **2007**, 1277–1284.
[23] C. Di Giorgio, F. Delmas, N. Filloux, M. Robin, L. Seferian, N. Azas, M. Gasquet, M. Costa, P. Timon-David, and J.P. Galy, “In vitro activities of 7-substituted 9-chloro and 9-amino-2-methoxyacridines and their bis- and tetra-acridine complexes against Leishmania infantum”, *Antimicrobial Agents and Chemotherapy*, 47(1), **2003**, 174–180.
[24] C. Di Giorgio, D.M. Michel, C. Julien, D. Florence, N. Anna, J. Séverine, D. Gérard, T.D. Pierre, G. Jean-Pierre, “Synthesis and antileishmanial activities of 4,5-di-substituted acridines as compared to their 4-mono-substituted homologues”, *Bioorganic & Medicinal Chemistry*, 13(19), **2005**, 5560–5568.
[25] Adamo and Baron, **2000**, Parac & Grimme, Gaussian 03, **2003**.
[26] ACDLABS 10, *Advanced Chemistry Development, Inc., Toronto, ON, Canada*, **2015**, http://www.acdlabs.com.
[27] MarvinSketch 5.11.4, *Chem Axon*, **2012**, http://www.che-maxon.com.
[28] ChemBioOffice, *PerkinElmer Informatics*, 2010, http://www.cambridgesoft.com.
[29] J.C. Dearden, M.T.D. Cronin, and K.L.E. Kaiser, “How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR)”, *SAR and QSAR in Environmental Research*, 20(3-4), **2009**, 241–266.
[30] K. Roy, R.N. Das, P. Ambure, and R.B. Aher, “Be aware of error measures. Further studies on validation of predictive QSAR models”, *Chemometrics and Intelligent Laboratory Systems*, 152, **2016**, 18–33.
[31] OECD, “Guidance Document on the Validation of QSAR Models”, *Organization for Economic Co-Operation & Development, Paris, France*, **2007**.
[32] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, and P. Gramatica, “Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs”, *Environmental Health Perspectives*, 111(10), **2003**, 1361–1375.
[33] P. Gramatica, “Principles of QSAR models validation: internal and external”, *QSAR and Combinatorial Science*, 26(5), **2007**, 694–701.
[34] G.E. Batista and D.F. Silva, “How k-nearest neighbor parameters affect its performance”, in *Proceedings of the Argentine Symposium on Artificial Intelligence, Instituto de Ciencias Matematicase de Computa cao, Sao Carlos, Brazil*, **2009**, 1–12.
[35] T.I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. van de Sandt, Weida Tong, G. Veith, and C. Yang, “Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships”, *Alternatives to Laboratory Animals*, 33(2), **2005**, 1–19.
[36] J.C. Dearden, “The history and development of quantitative structure-activity relationships (QSARs)”, *International Journal of Quantitative Structure-Property Relationships*, 1(1), **2016**, 1–44.

[37] K. Roy, S. Kar, and P. Ambure, "On a simple approach for deter-mining applicability domain of QSAR models", *Chemometrics and Intelligent Laboratory Systems*, 145, **2015**, 22–29.
[38] K. Roy, S. Kar, and R.D. Narayan, "Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences & Risk Assessment, in Chapter 6-Selected Statistical Methods in QSAR", *Academic Press, Boston, Mass, USA*, **2015**, 191–229.
[39] S.S. So and W.G. Richards, "Application of neural networks: quantitative structure-activity relationships of the derivatives of 2,4-diamino-5-(substituted-benzyl) pyrimidines as DHFR inhibitors", *Journal of Medicinal Chemistry*, 35(17), **1992**, 3201–3207.
[40] T.A. Andrea and H. Kalayeh, "Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors", *Journal of Medicinal Chemistry*, 34(9), **1991**, 2824–2836.
[41] A. Golbraikh and A. Tropsha, "Beware of q^2!", *Journal of Molecular Graphics and Modeling*, 20(4), **2002**, 269–276.
[42] S. Chtita, R. Hmamouchi, M. Larif, M. Ghamali, M. Bouachrine, and T. Lakhlifi, "QSPR studies of 9-aniliioacridine derivatives for their DNA drug binding properties based on density functional theory using statistical methods: model, validation and influencing factors", *Journal of Taibah University for Science*, 10(6), **2016**, 868–876.
[43] S. Chtita, M. Larif, M. Ghamali, M. Bouachrine, and T. Lakhlifi, "Quantitative structure–activity relationship studies of dibenzo[a,d]cycloalkenimine derivatives for non-competitive antagonists of N-methyl-d-aspartate based on density functional theory with electronic and topological descriptors", *Journal of Taibah University for Science*, 9(2), **2015**, 143–154.
[44] R.G. Brereton, "Chemometrics: Data Analysis for the Laboratory and Chemical Plant", *John Wiley & Sons, Chichester, UK*, **2003**.
[45] W.D. Fisher, "On grouping for maximum homogeneity", Journal of the American Statistical Association, 53(284), **1958**, 789–798.
[46] A.N. Choudhary, A. Kumar, and V. Juyal, "Quantitative structure activity relationship (QSAR) analysis of substituted 4-oxothiazolidines and 5-arylidines as lipoxygenase inhibitors", *Mini-Reviews in Medicinal Chemistry*, 10(8), **2010**, 705–714.

Supplementary Materials

Table S1: the values of chemical descriptors, the observed and predicted activities using the MLR models, and the cross-validation results.

Training Set (**Promastigotes**)								pIC_{50} PRO			Training Set (**Amastigotes**)							pIC_{50} AMA		
N°	$N_{\%}$	μ	ω	NHA	NHD	MTI	TVC	Obs.	MLR	CVLOO	N°	E_{HOMO}	CT	KH	T	NHA	NHD	Obs.	MLR	CVLOO
1	14.330	4.752	5.141	3	2	8205	$1.556\ 10^{-06}$	-2.533	-2.423	-2.386	1	-6.057	1126.804	17.443	814.190	3	2	-0.653	-0.789	-0.815
2	13.080	4.353	5.023	3	2	10835	$7.782\ 10^{-07}$	-2.204	-2.204	-2.204	2	-6.013	1134.579	17.196	836.730	3	2	-0.886	-0.820	-0.813
4	10.070	4.124	4.469	3	2	25571	$1.601\ 10^{-09}$	-0.634	-0.849	-0.909	3	-6.008	1146.159	16.950	859.270	3	2	-1.107	-0.802	-0.769
7	8.800	0.805	4.167	5	2	35529	$2.002\ 10^{-10}$	-0.681	-1.291	-1.514	4	-5.687	1267.085	18.849	979.730	3	2	-0.041	0.025	0.049
8	7.820	0.141	3.989	7	2	45907	$2.502\ 10^{-11}$	-1.207	-1.266	-1.317	7	-5.551	1302.123	21.305	1071.770	5	2	-0.041	0.020	0.030
9	11.820	4.629	5.100	3	2	15328	$4.992\ 10^{-08}$	-2.270	-2.007	-1.971	8	-5.469	1353.818	23.761	1163.810	7	2	0.097	0.374	0.536
11	11.250	5.405	5.234	4	2	16459	$1.634\ 10^{-08}$	-2.123	-2.140	-2.145	10	-6.077	1210.767	18.276	939.400	3	2	-0.462	-1.013	-1.095
12	10.900	2.935	4.367	4	2	18767	$1.765\ 10^{-08}$	-1.939	-1.529	-1.489	11	-6.041	1183.448	18.079	910.070	4	2	0.174	-0.465	-0.548
15	11.020	3.080	3.126	4	2	4981	$9.645\ 10^{-06}$	-0.732	-0.086	0.276	13	-5.632	1223.280	20.602	989.000	5	2	-0.114	-0.484	-0.528
16	8.830	4.319	3.675	3	1	4877	$1.567\ 10^{-06}$	-0.556	-1.063	-1.142	14	-4.879	911.054	10.198	577.240	3	1	0.398	0.269	0.232
18	8.280	5.542	3.519	4	1	12303	$6.354\ 10^{-07}$	-0.491	-0.545	-0.555	15	-4.913	926.848	14.635	638.060	4	2	-0.613	-0.002	0.228
19	7.950	5.464	3.515	4	1	13957	$5.188\ 10^{-07}$	-0.342	-0.480	-0.503	16	-5.119	949.837	11.136	637.040	3	1	-0.663	-0.335	-0.295
20	7.520	6.037	3.443	4	1	17337	$4.076\ 10^{-08}$	0.398	-0.121	-0.240	17	-5.037	954.980	12.149	675.610	4	1	-0.114	-0.439	-0.486
23	6.960	4.940	3.358	5	1	21159	$1.441\ 10^{-08}$	-0.279	-0.336	-0.351	18	-5.075	961.797	12.452	675.920	4	1	-0.398	-0.422	-0.424
24	5.430	1.289	4.141	2	0	4259	$5.990\ 10^{-05}$	-0.041	-0.759	-1.159	20	-5.038	1022.386	13.401	736.150	4	1	0.699	-0.152	-0.256
25	5.120	1.260	4.198	3	1	4841	$1.894\ 10^{-05}$	-2.262	-1.599	-1.218	21	-5.064	1013.830	13.334	749.260	5	1	0.222	0.061	0.028
26	4.160	3.893	4.755	2	0	4737	$4.236\ 10^{-05}$	-1.703	-1.352	-1.227	24	-5.450	887.835	6.876	523.900	2	0	-0.362	-0.883	-0.980
27	4.050	5.216	4.575	3	0	10103	$1.019\ 10^{-06}$	-2.178	-2.171	-2.170	25	-5.461	902.268	11.313	584.720	3	1	-0.959	-1.160	-1.203
28	3.910	6.025	4.677	3	0	12055	$1.248\ 10^{-06}$	-1.707	-2.022	-2.051	28	-5.642	936.409	9.130	622.580	3	0	-1.628	-1.624	-1.623
29	3.770	5.943	4.670	3	0	13687	$1.019\ 10^{-06}$	-1.446	-1.975	-2.017	31	-5.630	988.148	10.013	695.920	4	0	-1.645	-1.148	-1.058
30	3.570	6.072	4.580	3	0	17023	$8.006\ 10^{-08}$	-2.135	-1.734	-1.702	32	-5.646	1015.008	10.210	725.250	3	0	-2.191	-1.683	-1.532
31	3.420	5.503	4.649	4	0	18260	$2.620\ 10^{-08}$	-2.295	-2.137	-2.110	33	-5.568	1015.236	11.308	728.830	4	0	-2.088	-1.375	-1.279
32	3.290	5.447	4.689	3	0	18260	$7.861\ 10^{-08}$	-2.194	-1.957	-1.933	34	-5.554	888.744	9.505	518.700	2	1	-0.531	-0.557	-0.560
33	3.320	6.105	4.494	4	0	20791	$2.830\ 10^{-08}$	-2.098	-1.706	-1.650	35	-5.600	874.078	9.955	496.260	2	1	-1.515	-0.776	-0.609
34	13.450	1.247	3.791	2	1	3263	$7.044\ 10^{-05}$	-0.230	0.048	0.367	38	-5.854	1007.518	12.272	633.700	2	1	-0.833	-0.739	-0.717
36	8.960	7.120	4.612	2	1	8334	$4.992\ 10^{-06}$	-1.655	-1.133	-1.059	40	-5.845	1068.691	12.697	706.960	2	1	-1.121	-0.517	-0.363
39	5.320	5.908	4.556	2	0	6308	$4.151\ 10^{-06}$	-1.464	-1.878	-1.917	45	-5.946	976.240	8.975	625.520	2	0	-1.433	-1.619	-1.647
40	8.970	5.208	4.359	2	1	11068	$3.994\ 10^{-07}$	-1.654	-1.326	-1.275	47	-5.813	973.029	10.073	629.100	3	0	-0.869	-1.333	-1.397
41	4.470	5.786	4.431	2	0	10972	$3.261\ 10^{-07}$	-1.885	-1.648	-1.613	48	-5.185	1085.273	15.464	774.490	3	1	-0.672	-0.728	-0.741
43	4.230	5.341	4.530	3	0	11903	$1.067\ 10^{-07}$	-1.741	-2.068	-2.097	51	-5.549	900.947	11.349	580.870	3	2	0.222	0.013	-0.053
45	4.030	5.299	4.586	2	0	11903	$3.202\ 10^{-07}$	-1.790	-1.893	-1.910	52	-6.229	1230.146	18.841	941.650	3	2	-0.813	-0.936	-0.952
46	8.180	4.332	4.253	3	1	13906	$1.412\ 10^{-07}$	-1.513	-1.484	-1.483	56	-5.990	1281.779	19.834	1002.270	3	2	-1.236	-0.675	-0.589
47	4.080	5.446	4.305	3	0	13797	$1.153\ 10^{-07}$	-1.471	-1.683	-1.699	57	-6.036	1079.798	12.128	754.510	3	0	-0.716	-1.649	-1.803
48	11.820	5.905	4.214	3	1	15827	$1.547\ 10^{-07}$	-1.539	-1.294	-1.231	58	-6.081	1266.256	19.700	1028.490	5	2	-0.255	-0.255	-0.255
49	7.860	7.360	4.231	3	0	15711	$1.263\ 10^{-07}$	-1.819	-1.379	-1.222	60	-6.129	1319.107	20.094	1087.150	3	2	-1.562	-1.368	-1.249
50	17.710	0.470	3.636	3	2	4395	$2.490\ 10^{-05}$	-0.820	-1.205	-1.486	63	-5.918	1128.097	14.584	846.550	5	0	-1.825	-1.338	-1.163
52	9.410	6.623	5.258	3	2	19223	$1.251\ 10^{-07}$	-0.663	-1.401	-1.605	64	-5.211	1350.653	25.367	1137.330	5	2	-0.724	-0.832	-0.909
54	12.170	5.656	4.775	3	2	13049	$1.297\ 10^{-07}$	-1.061	-1.452	-1.498	65	-5.354	1154.774	17.662	889.570	5	0	-1.667	-1.767	-1.833
56	9.430	5.991	4.723	3	2	29193	$8.007\ 10^{-10}$	-0.556	-0.551	-0.550										
58	8.730	6.320	4.927	5	2	32645	$8.578\ 10^{-11}$	-0.756	-1.024	-1.132										
62	8.310	3.311	4.511	5	2	39911	$1.001\ 10^{-10}$	-1.797	-0.942	-0.720										

Table S2: Output file summarize the information including presence/absence of systematic error, and all the external validation parameters that are required to judge the performance of prediction quality of the QSAR model

User Input File Info.	File Name	Sample_TestSet.xlsx
Model biasness test	Systematic Error Result	Absent
	nPE / nNE	1,1000
	nNE / nPE	0,9091
	|MPE / MNE|	1,0131
	|MNE / MPE|	0,9870
	AAE - |AE|	0,5295
	R^2 (Residuals; serial correlation)	0,0226
	R^2 (Residuals and Yobs values)	0,1492
Classical Metrics (for 100% data)	R^2Test (100% data)	0,6919
	R0^2Test (100% data)	0,6797
	R0^'2Test (100% data)	0,6418
	Q2F1 (100% data)	0,6814
	Q2F2 (100% data)	0,6796
	Scaled Avg.Rm^2(100% data)	0,5788
	Scaled DeltaRm^2(100% data)	0,0734
	CCC (100% data)	0,8299
Classical Metric (after removing 5% data with high residuals)	R^2Test (95% data)	0,7562
	R0^2Test (95% data)	0,7459
	R0^'2Test (95% data)	0,6382
	Q2F1 (95% data)	0,7466
	Q2F2 (95% data)	0,7458
	ScaledAvgRm2 (95% data)	0,6608
	ScaledDeltaRm2 (95% data)	0,0364
	CCC (95% data)	0,8687
Error-based metrics (for 100% data)	RMSEP (100% data)	0,7127
	SD (100% data)	0,4432
	SE (100% data)	0,0432
	MAE (100% data)	0,5598
Error-based metric (after removing 5% data with high residuals)	RMSEP (95% data)	0,6127
	SD (95% data)	0,3636
	SE(95% data)	0,0365
	MAE (95% data)	0,4945
	MAE+3*SD (95% data)	1,5853
Number of test set compounds, Range and Mean (train and test)	N Comp Test	105
	Train range	7,8195
	Train YMean	6,7604
	Test range	6,8474
	Test YMean	6,6664
Distribution of observed response values of Test set around Test mean (in %)	%Y (+/-0.5) Test Mean	26,6667
	%Y (+/-1.0) Test Mean	52,3810
	%Y (+/-1.5) Test Mean	82,8571
	%Y (+/-2.0) Test Mean	92,3810
Distribution of observed response values of Test set around Train mean (in %)	%Y (+/-0.5) Test Mean	25,7143
	%Y (+/-1.0) Train Mean	52,3810
	%Y (+/-1.5) Train Mean	84,7619
	%Y (+/-2.0) Train Mean	93,3333
Distribution of prediction errors (in %)	% NComp > (0.1*TR)	28,5714
	% NComp > (0.15*TR)	10,4762
	% NComp > (0.2*TR)	3,8095
	% NComp > (0.25*TR)	0,0000
Threshold values utilized to judge the model predictions	(0.1* Training Set Range)	0,7820
	(0.15* Training Set Range)	1,1729
	(0.2* Training Set Range)	1,5639
	(0.25* Training Set Range)	1,9549
RESULT (MAE-based criteria applied on 95% data)	Prediction Quality	MODERATE

Table S3: Outliers and outsides with simple approach for pIC_{50} (PRO)

N°	N%	μ	ω	NHA	NHD	MTI	TVC	pIC_{50} PRO	Outlier Info.
1	14.330	4.752	5.141	3	2	8205	$1.556\ 10^{-06}$	-2,423	-
2	13.080	4.353	5.023	3	2	10835	$7.782\ 10^{-07}$	-2,204	-
4	10.070	4.124	4.469	3	2	25571	$1.601\ 10^{-09}$	-0,849	-
7	8.800	0.805	4.167	5	2	35529	$2.002\ 10^{-10}$	-1,291	-
8	7.820	0.141	3.989	7	2	45907	$2.502\ 10^{-11}$	-1,266	Outlier
9	11.820	4.629	5.100	3	2	15328	$4.992\ 10^{-08}$	-2,007	-
11	11.250	5.405	5.234	4	2	16459	$1.634\ 10^{-08}$	-2,14	-
12	10.900	2.935	4.367	4	2	18767	$1.765\ 10^{-08}$	-1,529	-
15	11.020	3.080	3.126	4	2	4981	$9.645\ 10^{-06}$	-0,086	-
16	8.830	4.319	3.675	3	1	4877	$1.567\ 10^{-06}$	-1,063	-
18	8.280	5.542	3.519	4	1	12303	$6.354\ 10^{-07}$	-0,545	-
19	7.950	5.464	3.515	4	1	13957	$5.188\ 10^{-07}$	-0,48	-
20	7.520	6.037	3.443	4	1	17337	$4.076\ 10^{-08}$	-0,121	-
23	6.960	4.940	3.358	5	1	21159	$1.441\ 10^{-08}$	-0,336	-
24	5.430	1.289	4.141	2	0	4259	$5.990\ 10^{-05}$	-0,759	-
25	5.120	1.260	4.198	3	1	4841	$1.894\ 10^{-05}$	-1,599	-
26	4.160	3.893	4.755	2	0	4737	$4.236\ 10^{-05}$	-1,352	-
27	4.050	5.216	4.575	3	0	10103	$1.019\ 10^{-06}$	-2,171	-
28	3.910	6.025	4.677	3	0	12055	$1.248\ 10^{-06}$	-2,022	-
29	3.770	5.943	4.670	3	0	13687	$1.019\ 10^{-06}$	-1,975	-
30	3.570	6.072	4.580	3	0	17023	$8.006\ 10^{-08}$	-1,734	-
31	3.420	5.503	4.649	4	0	18260	$2.620\ 10^{-08}$	-2,137	-
32	3.290	5.447	4.689	3	0	18260	$7.861\ 10^{-08}$	-1,957	-
33	3.320	6.105	4.494	4	0	20791	$2.830\ 10^{-08}$	-1,706	-
34	13.450	1.247	3.791	2	1	3263	$7.044\ 10^{-05}$	0,048	-
36	8.960	7.120	4.612	2	1	8334	$4.992\ 10^{-06}$	-1,133	-
39	5.320	5.908	4.556	2	0	6308	$4.151\ 10^{-06}$	-1,878	-
40	8.970	5.208	4.359	2	1	11068	$3.994\ 10^{-07}$	-1,326	-
41	4.470	5.786	4.431	2	0	10972	$3.261\ 10^{-07}$	-1,648	-
43	4.230	5.341	4.530	3	0	11903	$1.067\ 10^{-07}$	-2,068	-
45	4.030	5.299	4.586	2	0	11903	$3.202\ 10^{-07}$	-1,893	-
46	8.180	4.332	4.253	3	1	13906	$1.412\ 10^{-07}$	-1,484	-
47	4.080	5.446	4.305	3	0	13797	$1.153\ 10^{-07}$	-1,683	-
48	11.820	5.905	4.214	3	1	15827	$1.547\ 10^{-07}$	-1,294	-
49	7.860	7.360	4.231	3	0	15711	$1.263\ 10^{-07}$	-1,379	-
50	17.710	0.470	3.636	3	2	4395	$2.490\ 10^{-05}$	-1,205	-
52	9.410	6.623	5.258	3	2	19223	$1.251\ 10^{-07}$	-1,401	-
54	12.170	5.656	4.775	3	2	13049	$1.297\ 10^{-07}$	-1,452	-
56	9.430	5.991	4.723	3	2	29193	$8.007\ 10^{-10}$	-0,551	-
58	8.730	6.320	4.927	5	2	32645	$8.578\ 10^{-11}$	-1,024	-
62	8.310	3.311	4.511	5	2	39911	$1.001\ 10^{-10}$	-0,942	-

N°	N%	μ	ω	NHA	NHD	MTI	TVC	pIC_{50} PRO	Outside Info.
3	12.030	4.259	5.024	3	2	14075	$3.891\ 10^{-07}$	-2,047	-
6	9.270	5.640	4.654	5	2	28763	$1.716\ 10^{-10}$	-1,014	-
10	10.780	5.730	5.267	3	2	16459	$4.902\ 10^{-08}$	-1,832	-
14	11.760	2.698	3.062	3	1	4387	$3.050\ 10^{-05}$	-0,012	-
17	8.580	4.969	3.431	4	1	10329	$5.188\ 10^{-07}$	-0,664	-
21	7.180	5.214	3.498	5	1	18600	$1.334\ 10^{-08}$	-0,551	-
37	4.460	6.939	4.779	2	0	8245	$4.076\ 10^{-06}$	-1,781	-
38	10.680	5.122	4.374	2	1	6388	$5.083\ 10^{-06}$	-1,512	-
42	8.480	5.260	4.457	3	1	12005	$1.307\ 10^{-07}$	-1,608	-
64	13.170	5.297	4.388	5	2	47653	$1.201\ 10^{-10}$	-0,455	Outside AD

Table S4: Outliers and outsides with simple approach for pIC_{50} (AMA)

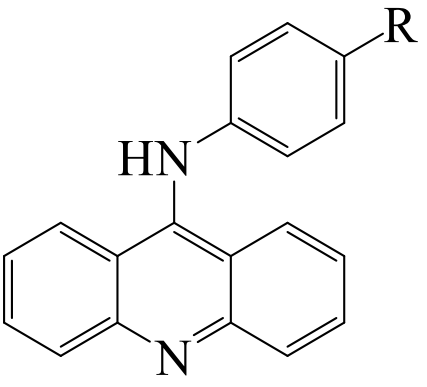
N°	E_{HOMO}	CT	KH	T	NHA	NHD	pIC_{50} AMA	Outlier Info.
1	-6.057	1126.804	17.443	814.190	3	2	-0.789	-
2	-6.013	1134.579	17.196	836.730	3	2	-0.820	-
3	-6.008	1146.159	16.950	859.270	3	2	-0.802	-
4	-5.687	1267.085	18.849	979.730	3	2	0.025	-
7	-5.551	1302.123	21.305	1071.770	5	2	0.020	-
8	-5.469	1353.818	23.761	1163.810	7	2	0.374	-
10	-6.077	1210.767	18.276	939.400	3	2	-1.013	-
11	-6.041	1183.448	18.079	910.070	4	2	-0.465	-
13	-5.632	1223.280	20.602	989.000	5	2	-0.484	-
14	-4.879	911.054	10.198	577.240	3	1	0.269	-
15	-4.913	926.848	14.635	638.060	4	2	-0.002	-
16	-5.119	949.837	11.136	637.040	3	1	-0.335	-
17	-5.037	954.980	12.149	675.610	4	1	-0.439	-
18	-5.075	961.797	12.452	675.920	4	1	-0.422	-
20	-5.038	1022.386	13.401	736.150	4	1	-0.152	-
21	-5.064	1013.830	13.334	749.260	5	1	0.061	-
24	-5.450	887.835	6.876	523.900	2	0	-0.883	-
25	-5.461	902.268	11.313	584.720	3	1	-1.160	-
28	-5.642	936.409	9.130	622.580	3	0	-1.624	-
31	-5.630	988.148	10.013	695.920	4	0	-1.148	-
32	-5.646	1015.008	10.210	725.250	3	0	-1.683	-
33	-5.568	1015.236	11.308	728.830	4	0	-1.375	-
34	-5.554	888.744	9.505	518.700	2	1	-0.557	-
35	-5.600	874.078	9.955	496.260	2	1	-0.776	-
38	-5.854	1007.518	12.272	633.700	2	1	-0.739	-
40	-5.845	1068.691	12.697	706.960	2	1	-0.517	-
45	-5.946	976.240	8.975	625.520	2	0	-1.619	-
47	-5.813	973.029	10.073	629.100	3	0	-1.333	-
48	-5.185	1085.273	15.464	774.490	3	1	-0.728	-
51	-5.549	900.947	11.349	580.870	3	2	0.013	-
52	-6.229	1230.146	18.841	941.650	3	2	-0.936	-
56	-5.990	1281.779	19.834	1002.270	3	2	-0.675	-
57	-6.036	1079.798	12.128	754.510	3	0	-1.649	-
58	-6.081	1266.256	19.700	1028.490	5	2	-0.255	-
60	-6.129	1319.107	20.094	1087.150	3	2	-1.368	-
63	-5.918	1128.097	14.584	846.550	5	0	-1.338	-
64	-5.211	1350.653	25.367	1137.330	5	2	-0.832	-
65	-5.354	1154.774	17.662	889.570	5	0	-1.767	-

N°	E_{HOMO}	KH	T	NHA	NHD	CT	pIC_{50} AMA	Outside Info.
5	-5.850	1064.61	1300.866	3	2	19.110	-0.787	-
6	-5.785	1005.95	1247.456	5	2	18.715	0.332	-
26	-5.665	583.700	926.651	2	0	7.815	-1.456	-
27	-5.603	622.270	930.050	3	0	8.827	-1.628	-
29	-5.640	618.850	938.380	3	0	9.007	-1.467	-
36	-5.966	676.650	1037.977	2	1	12.201	-0.774	-
37	-6.025	552.770	930.967	2	0	8.348	-1.457	-
44	-5.919	749.400	1083.064	2	1	12.828	-0.978	-
46	-5.790	752.980	1078.419	3	1	13.925	-0.733	-
50	-5.461	625.750	913.563	3	2	13.449	-1.028	-

Table S5: the values of the observed, the predicted activities and the errors using the ANN models results for the studied compounds (training set validation and test set).

Training Set (PRO)				Training Set (AMA)				Validation and test set			
N°	Obs.	Pred.	Error	N°	Obs.	Pred.	Error	N°	Obs.	Pred.	Error
1	-2.533	-1.939	0.594	2	-0.886	-1.034	0.148	**Validation Set (PRO)**			
3	-1.669	-1.937	-0.268	3	-1.107	-0.994	-0.113	9	-2.270	-1.937	0.333
4	-0.634	-0.661	-0.027	4	-0.041	-0.044	0.003	19	-0.342	-0.461	-0.119
6	-0.230	-0.324	-0.094	5	0.699	0.603	0.096	20	0.398	-0.139	-0.537
7	-0.681	-0.699	-0.018	6	1.523	1.096	0.427	25	-2.262	-1.664	0.598
8	-1.207	-1.212	-0.005	7	-0.041	0.682	-0.723	26	-1.703	-1.904	-0.201
12	-1.939	-1.774	0.165	11	0.174	0.035	0.139	27	-2.178	-1.928	0.250
14	0.301	0.254	-0.047	13	-0.114	-0.614	0.500	32	-2.194	-1.899	0.295
15	-0.732	-0.555	0.177	15	-0.613	-0.392	-0.221	34	-0.230	-0.028	0.202
16	-0.556	-0.640	-0.084	18	-0.398	-0.518	0.120	**Test set (PRO)**			
18	-0.491	-0.446	0.045	21	0.222	0.117	0.105	2	-2.204	-1.940	0.264
21	-0.114	-0.132	-0.018	24	-0.362	-0.605	0.243	10	-1.061	-1.873	-0.812
24	-0.041	-0.174	-0.133	25	-0.959	-0.941	-0.018	11	-2.123	-1.922	0.201
28	-1.707	-1.938	-0.231	26	-1.170	-1.034	-0.136	17	-0.380	-0.532	-0.152
29	-1.446	-1.934	-0.488	28	-1.628	-1.715	0.087	23	-0.279	-0.043	0.236
30	-2.135	-1.920	0.215	33	-2.088	-1.855	-0.233	37	-2.200	-1.937	0.263
31	-2.295	-1.939	0.356	34	-0.531	-0.706	0.175	47	-1.471	-1.876	-0.405
33	-2.098	-1.937	0.161	35	-1.515	-1.409	-0.106	50	-0.820	-0.183	0.637
36	-1.655	-1.580	0.075	36	-1.427	-0.987	-0.440	**Validation Set (AMA)**			
38	-2.245	-1.866	0.379	37	-2.185	-2.087	-0.098	1	-0.653	-1.218	0.565
39	-1.464	-1.906	-0.442	38	-0.833	-1.147	0.314	8	0.097	0.341	-0.244
40	-1.654	-1.748	-0.094	40	-1.121	-1.105	-0.016	16	-0.663	0.138	-0.801
41	-1.885	-1.765	0.120	44	-1.004	-1.385	0.381	17	-0.114	-0.399	0.285
42	-1.815	-1.927	-0.112	45	-1.433	-1.360	-0.073	20	0.699	0.186	0.513
43	-1.741	-1.919	-0.178	46	-0.973	-0.649	-0.324	29	-1.561	-1.618	0.057
45	-1.790	-1.724	0.066	47	-0.869	-0.785	-0.084	31	-1.645	-1.514	-0.131
46	-1.513	-1.830	-0.317	48	-0.672	-0.544	-0.128	**Test set (AMA)**			
48	-1.539	-1.562	-0.023	50	-0.531	-0.486	-0.045	10	-0.462	-1.079	0.617
49	-1.819	-1.802	0.017	52	-0.813	-0.822	0.009	14	0.398	0.865	-0.467
52	-0.663	-0.790	-0.127	57	-0.716	-0.952	0.236	27	-1.877	-1.799	-0.078
54	-1.061	-1.052	0.009	58	-0.255	0.029	-0.284	32	-2.191	-1.489	-0.702
56	-0.556	-0.577	-0.021	60	-1.562	-1.439	-0.123	51	0.222	0.206	0.016
58	-0.756	-0.678	0.078	63	-1.825	-1.853	0.028	56	-1.236	-1.030	-0.206
62	-1.797	-1.745	0.052	64	-0.724	-0.886	0.162	65	-1.667	-2.341	0.674
64	-0.940	-0.943	-0.003	---	---	---	---	---	---	---	---

Chapitre 3 : Etude de la RQSP de la constante d'association avec l'ADN pour des dérivés de l'acridine



Available online at www.sciencedirect.com

Science Direct

Journal of Taibah University for Science 10 (2016) 868–876

Journal of Taibah University for Science

www.elsevier.com/locate/jtusci

QSPR studies of 9-aniliioacridine derivatives for their DNA drug binding properties based on density functional theory using statistical methods: Model, validation and influencing factors

Samir Chtita [a,*], Rachid Hmamouchi [a], Majdouline Larif [b], Mounir Ghamali [a], Mohammed Bouachrine [c], Tahar Lakhlifi [a,*]

[a] *Molecular Chemistry and Natural Substances Laboratory, Faculty of Science, University Moulay Ismail, Meknes, Morocco*

[b] *Separation Process Laboratory, Faculty of Science, University Ibn Tofail, Kenitra, Morocco*

[c] *High School of Technology, University Moulay Ismail, Meknes, Morocco*

Available online 3 June 2015

Abstract

As a continuation of our research on the development and optimization of the biological activities/proprieties of acridine derivatives, a series of 31 molecules based on 9-aniliioacridines (25 training set and 6 test set) were subjected to 3D quantitative structure propriety relationship QSPR analyses for their drug-DNA binding proprieties using multiple linear regression (MLR) and multiple non-linear regression (MNLR). Quantum chemical calculations using density functional theory (B3LYP/6-31G (d) DFT) methods was performed on the studied compounds and used to calculate the electronic and quantum chemical parameters.

The models were used to predict the association constant of the DNA drug binding of the test set compounds, and the agreement between the experimental and predicted values was verified. The descriptors determined by QSPR studies were used for the study and design of new compounds. The applicability domain of proposed models was investigated using William's plot to detect outliers and outsides compounds. The statistical results indicate that the predicted values were in good agreement with the experimental results (r = 0.935 and r = 0.936 for MLR and MNLR, respectively). To validate the predictive power of the resulting models, the external validation multiple correlation coefficients were 0.932 and 0.939 for the MLR and the MNLR, respectively. These results show that both models possess a favorable estimation stability and good prediction power.

Keywords: QSPR; DFT; MLR; MNLR; Acridine; Anti-tumor

1. Introduction

The ease of synthesis, attractive coloration and crystallinity of acridine derivatives has long attracted the attention of medicinal chemists. The acridine family includes a wide range of planar tri-cyclic aromatic molecules with various biological properties and consists of a nitrogen atom (N-atom) in its hetero-cyclic nucleus.

* Corresponding author. Tel.: +212 660005554. *E-mail address:* samirchtita@gmail.com (S. Chtita).

Peer review under responsibility of Taibah University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jtusci.2015.04.007

The natural and synthetic compounds of the acridine family are well known therapeutic agents due to their wide range of pharmacological and biological activities, including antileishmanial **[1, 2]**, antimicrobial **[1]**, antioxidant **[3]**, antimalarials **[4]**, anti-inflammatory **[5]**, analgesic **[6]**, antiparasitic **[7]**, anti-tumoral **[8]**, antibacterial or anticancer chemotherapy **[9–12]** activities, among others.

The diversity of the biological and pharmacological activities has given acridines a respectable reputation in chemotherapy in the 20th century **[13]**.

One of the important phenomena of DNA is their ability to reversibly bind planar molecules that can be inserted between the base pairs of the double helix **[14]**. Acridines are fused linear tri-cyclic aromatic molecules with planar geometry that bind tightly, but reversibly, to DNA by intercalating between adjacent base pairs **[15, 16]**. The driving force for this binding comes primarily from stacking interactions between the acridine nucleus and the DNA bases and is sufficiently large to physically unwind the DNA double helix to accommodate the inserted ligand. The majority of the biological effects of acridine derivatives are considered to result from this mode of non-covalent interactions with DNA **[17]**.

Generally, hetero-aromatic molecules bind to DNA by intercalating (i.e., a non-covalent interaction in which the drug is held rigidly and perpendicular to the helix axis) and stacking between the base pairs of the double helix. The principal driving forces for the intercalation are stacking and charge-transfer interactions, but hydrogen bonding and electrostatic forces also play a role in stabilization **[18]**. 9-Anilinoacridines, as intercalators of double-stranded (duplex) DNA, have been explored extensively as antitumor agents. In particular, m-AMSA and CI-921 are, in fact, used clinically for the treatment of leukaemia **[19–22]**.

To discover new active antitumor compounds, we examined the DNA-ethidium fluorescence quenching effect of these compounds and found a substance that exhibits a stronger fluorescence quenching effect than m-ASMA. Moreover, it has been established that the fluorescence quenching demonstrates a very good correlation with antitumor activity **[14, 23–24]**.

In an investigation of the structure propriety relationships in the AMSA tumor inhibitory analogues, the DNA binding properties of a series of 9-aniliioacridines was determined by drug competition with the fluorochrome ethidium for available sites. The decrease in fluorescence of a DNA-ethidium complex by the addition of a drug is due to both the drug displacement of the bound ethidium and the quenching of the fluorescence of the bound ethidium by the bound drug. The measurement of both factors allows the drug-DNA association constants (K) to be determined **[25, 26]**.

The experiment is a direct method of obtaining the activity/propriety data of organic compounds. However, this approach suffers from many deficiencies, including the requirement of myriads of trial organisms, high cost, long period of time, significant variations in the measured values between laboratories, and so on. Consequently, it would be impossible to determine the drug-DNA association constants of all of the organic compounds by experimentation. As new compounds are emerging, other difficulties will follow. Therefore, it is necessary to use theoretical research to compensate for the disadvantages of experimentation and to predict the data for compounds quickly and precisely. In this research paper, we focus on the drug-DNA binding constants of some acridine derivatives.

With the rapid development of computer science and theoretical quantum chemical studies, the quantum chemical parameters of compounds can be obtained quickly and precisely by computation. These structural parameters, along with the introduction of quantitative structure activity/propriety relationship (QAPR/QSPR) models, can increase the interpretability and predictability of the activities/proprieties of new organic compounds.

In this work, we attempt to establish a quantitative structure propriety relationship for the association constant (K) of drug binding to DNA by studying a series of 31 substituted 9-aniliioacridine derivatives. We accordingly propose quantitative models and try to interpret the propriety of the compounds relying on multivariate statistical analyses. Thus, we can predict the association constant (K) for drug binding to DNA.

2. Material and methods

2.1. Experimental dataset

To determine a quantitative structure function relationship, we studied a series of 31 selected 9-aniliioacridine derivatives that were synthesized and that had their antitumor activity evaluated by Bruce et al. **[25]**. Twenty-five molecules were selected to propose the quantitative model (training set) as well as 6 compounds that were not used in the training set were selected randomly served to test the performance of the proposed model (test set). In reality, Bruce et al. proposed 65 compounds; the remaining compounds had structures that differed from the structures required for this study.

Figure 1 shows the chemical structures of the studied compounds, and the experimental association constants for the drug binding to DNA (the concentration of drug needed to displace 50% of the ethidium) of the studied compounds were taken from Ref. **[25]** (Table 1).

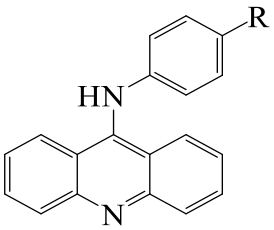


Figure 1: Chemical structure of the studied compounds

Table 1: Observed Log *K* for the 9-aniliioacridine derivatives

Compound	R	Log *K*	Compound	R	Log *K*
1	NO_2	5,48	17	CH_3	5,99
2	SO_2CH_3	5,86	18	$NHCOC_6H_5$	6,20
3	CN	5,70	19	$NHCONHCH_3$	6,24
4	SO_2NHCH_3	5,82	20	$NHCONHC_6H_5$	6,37
5	SO_2NH_2	5,96	21 [(a)]	OCH_3	6,12
6 [(a)]	$COCH_3$	5,87	22	OH	6,28
7	$COOCH_3$	5,88	23	$NH(CH_2)_5CH_3$	6,43
8	$CONH_2$	5,83	24	$NH(CH_2)_3CH_3$	6,45
9	F	5,90	25	$NH(CH_2)_2CH_3$	6,35
10	Cl	5,99	26 [(a)]	$NHCH_2CH_3$	6,44
11 [(a)]	Br	6,02	27	NH_2	6,31
12	$NHSO_2CH_3$	6,15	28	$N(CH_3)_2$	6,51
13	$NHSO_2C_6H_5$	6,20	29	$NHCH_3$	6,55
14	H	5,86	30	$NCH_3SO_2CH_3$	5,89
15	$NHCOCH_3$	6,30	31 [(a)]	$NHSO_2C_6H_4$-p-NH_2	6,30
16 [(a)]	$NHCOOCH_3$	6,36	*K: Association constant for drug binding to DNA*		

(a) Tested compounds (test set).

2.2. Computational methods

An attempt has been made to correlate the propriety of these compounds with various physicochemical parameters. DFT (density functional theory) and TD-DFT methods were used in this study. 3D structures of the molecules were generated using the Gauss View 3.0, and then, all of the calculations were performed using the Gaussian 03 W program series. Geometry optimization of the 31 compounds was carried out by a B3LYP function employing a 6–31G (d) basis set **[27, 28]**. The geometry of all of the species under investigation was determined by optimizing all of the geometrical variables without any symmetry constraints **[29]**.

2.3. Calculation of molecular descriptors

From the results of the DFT calculations, the quantum chemistry descriptors were obtained for the model building as follows: total energy, E_T (eV); highest occupied molecular orbital energy, E_{HOMO} (eV); lowest unoccupied molecular orbital energy, E_{LUMO} (eV); difference in absolute value Gap (eV); dipole moment, μ (Debye); absolute hardness, η (eV);

absolute electronegativity, χ (eV); electrophilicity index, ω (eV); and sum of the negative charges on the molecule (TNC) were deduced from the stable structure of the neutral form.

η, χ and ω were determined from [29]:

$$\eta = \frac{E_{LUMO} - E_{HOMO}}{2}, \chi = \frac{E_{LUMO} + E_{HOMO}}{2} \; and \; \omega = \frac{\chi^2}{2\eta}$$

The transition energies were calculated in the ground-state with excited-state geometries using TD-DFT calculations on the fully optimized geometries. The results obtained gave us the absorption maximum, λ_{max} (nm), their corresponding activation energy, Ea (eV), and the factor oscillation strengths, S.O.

2.4. Statistical analysis

To explain the structure-activity relationship, these 12 descriptors were calculated for the 31 molecules using the Gaussian 03W and Gauss View software. The study that we conducted consists of multiple linear regression (MLR) and nonlinear regression (MNLR), which are available in the XLSTAT software [30].

The multiple linear regression statistical techniques used to study the relationship between one dependent variable and several independent variables. It is a mathematical technique that minimizes the differences between actual and predicted values. It has also served to select descriptors that are used as input parameters in multiple nonlinear regression (MNLR). The MLR and MNLR techniques were used to predict the association constant for drug binding to DNA values, Log *K*. The equations were justified by the correlation coefficient (*r*), the Mean Squared Error (*MSE*), the Fishers F-statistic (F), and the significance level (*p-value*) [31].

The final stage of this QSPR analysis consists of applicability domain, a model is valid only within its training domain and new molecules must be considered as belonging to the domain before the model is applied (OECD Principle 3 [32]). Without applicability domain (AD), each model can predict the propriety of any compound, even with a completely different structure from those included in the study. Therefore, the AD is a tool to find out compounds that are outside the applicability domain of the built QSAR model and it detects outliers present in the training set compounds. There are several methods for defining the applicability domain (AD) of QSAR models, but the most common one is determining the leverage values h_i ($h_i = x_i^T (X^T X)^{-1} x_i$ ($i = 1,2, \ldots, n$)) for each compound [33].

Where xi is the descriptor row-vector of query compound, X is the n*k-1 matrix of k model descriptor values for n training set compounds, and the superscript “T” refers to the transpose of matrix/vector.

In this study, we use the Williams plot; in this plot, the applicability domain is established inside a squared area within standard deviation ±x (in this study x = 2.5 (“three sigma rule”) and a leverage threshold h* (h*= 2.5*(k+1)/n) **[34, 35]**.

“n” is the number of training set compounds and “k” is the number of model descriptors.

The leverage (h) greater than the warning leverage (h*) suggested that the compound was very influential on the model **[36]**.

3. Results and discussions

3.1. Dataset for analysis

The QSPR analysis was performed using the experimental association constant for drug binding to DNA values of the 31 selected molecules as reported by C. Bruce et al. **[25]**, the values of the 12 chemical descriptors are shown in table 2.

Table 2: Values of the parameters obtained by DFT/TD-DFT calculation for the training set

N°	E_T	μ	E_{HOMO}	E_{LUMO}	Gap	η	χ	ω	λ_{max}	S.O.	Ea	TNC
1	-26428.147	4.957	-5.552	-2.493	3.059	1,530	-4,023	5.289	423.01	0.257	2.931	-4.264
2	-38906.463	5.218	-5.744	-2.314	3.430	1.715	-4.029	4.733	410.00	0.185	3.024	-7.178
3	-25420.014	5.174	-5.740	-2.343	3.397	1.698	-4.041	4.808	415.77	0.191	2.982	-4.442
4	-40412.372	4.552	-5.696	-2.176	3.420	1.710	-3.986	4.645	412.38	0.185	3.007	-7.643
5	-39342.774	3.910	-5.681	-2.266	3.416	1.708	-3.973	4.622	412.51	0.187	3.006	-7.346
7	-29110.685	3.367	-5.524	-2.157	3.367	1.683	-3.841	4.381	420.84	0.211	2.946	-5.624
8	-27500.440	4.854	-5.496	-2.134	3.363	1.681	-3.815	4.328	420.74	0.201	2.947	-5.625
9	-25610.109	2.013	-5.345	-2.011	3.333	1.667	-3.678	4.058	422.63	0.167	2.934	-4.499
10	-35415.609	2.172	-5.449	-2.105	3.344	1.672	-3.777	4.265	423.13	0.179	2.930	-4.127
12	-40412.465	5.953	-5.416	-2.104	3.312	1.656	-3.760	4.268	429.74	0.193	2.885	-7.770
13	-45629.550	6.683	-5.392	-2.088	3.304	1.652	-3.740	4.234	431.85	0.210	2.871	-8.251
14	-22910.013	2.378	-5.298	-1.957	3.340	1.670	-3.627	3.939	422.14	0.168	2.937	-4.316
15	-28569.981	3.238	-5.386	-2.108	3.278	1.639	-3.747	4.284	435.87	0.192	2.845	-6.095
17	-23979.846	2.724	-5.218	-1.913	3.304	1.652	-3.565	3.847	428.57	0.181	2.893	-4.780
18	-33787.105	2.924	-5.368	-2.095	3.273	1.637	-3.731	4.254	438.24	0.205	2.829	-6.547
19	-30076.151	2.353	-5.247	-2.014	3.233	1.616	-3.631	4.078	442.79	0.199	2.800	-6.577
20	-35293.400	2.335	-5.263	-2.016	3.248	1.624	-3.639	4.079	440.90	0.198	2.812	-7.251
22	-24956.583	3.981	-5.110	-1.852	3.258	1.629	-3.481	3.718	434.13	0.178	2.856	-4.846
23	-30834.393	5.174	-4.785	-1.705	3.080	1.540	-3.245	3.419	467.03	0.186	2.655	-7.512
24	-28694.939	5.124	-4.789	-1.707	3.082	1.541	-3.248	3.424	466.68	0.184	2.657	-6.615
25	-27625.214	5.098	-4.794	-1.710	3.084	1.542	-3.252	3.429	466.32	0.181	2.659	-6.167
27	-24416.119	4.655	-4.900	-1.751	3.149	1.575	-3.326	3.512	453.49	0.176	2.734	-4.964
28	-25485.689	4.925	-4.813	-1.722	3.091	1.545	-3.267	3.454	464.81	0.174	2.667	-5.244
29	-26555.202	4.973	-4.743	-1.709	3.034	1.517	-3.226	3.430	474.87	0.181	2.611	-5.522
30	-41482.000	2.886	-5.439	-2.131	3.308	1.654	-3.785	4.330	431.82	0.188	2.871	-8.063

3.2. Multiple Linear Regressions (MLR)

To propose a mathematical model and to quantitatively evaluate the substituent’s physicochemical effects on log *K* for the entire set consisting of 31 molecules, we submitted the data matrix that was composed of the 12 variables that corresponded to the 25 molecules (training set) to a descendent multiple regression analysis.

The decreasing study of MLR based on the elimination of descriptors aberrant until a valid model (including the critical probability: *p-value*< 0.05 for all descriptors and the model complete). This method used the coefficients *r*, r^2, MSE and *p-value* to select the best regression performance, where *r* is the correlation coefficient; r^2 is the coefficient of determination; MSE is the mean squared error; *p-value* is the significance level and *F* is the Fisher F-statistic.

Treatment with multiple linear regressions is more accurate because it allows for the structural descriptors for each drug-DNA propriety of the 25 molecules to be connected to quantitatively evaluate the effect of the substituent. The selected descriptors are: the lowest unoccupied molecular orbital energy, E_{LUMO}; the electrophilicity index, ω; and the activation energy, Ea.

The QSPR model built using multiple linear regression (MLR) method is represented by the following equation (Equation 1):

$$\text{Log } K = 11.484 - 7.898\ E_{LUMO} - 3.495\ \omega - 2.461\ \text{Ea}$$

$$N = 25;\ r = 0.935;\ r^2 = 0.873;\ F = 48.256;\ \text{MSE} = 0.011;\ p\text{-value} < 0.0001$$

A higher correlation coefficient, *r*, and lower mean squared error, MSE, indicate that the model is more reliable. The Fisher F-test is also used. Given that the *p-value* is much smaller than 0.05, we are taking less than a 0.01% risk in assuming that the null hypothesis is wrong. Therefore, we can conclude, with confidence, that the model brings a significant amount of information.

The elaborated QSPR model reveals that the association constant for the drug-DNA could be explained by a number of electronic factors (E_{LUMO}, ω and Ea). The negative correlation of these factors with the association constant for the drug-DNA constants in equation 1 shows that an increase in the values of these factors implies a decrease in the value of Log *K*, i.e., the variation in Log *K* with the descriptor values, which are illustrated in figure 2, show that the lowest unoccupied molecular orbital, E_{LUMO}, varies in the same way as Log *K*, so the activation energy and the electrophilicity index vary inversely.

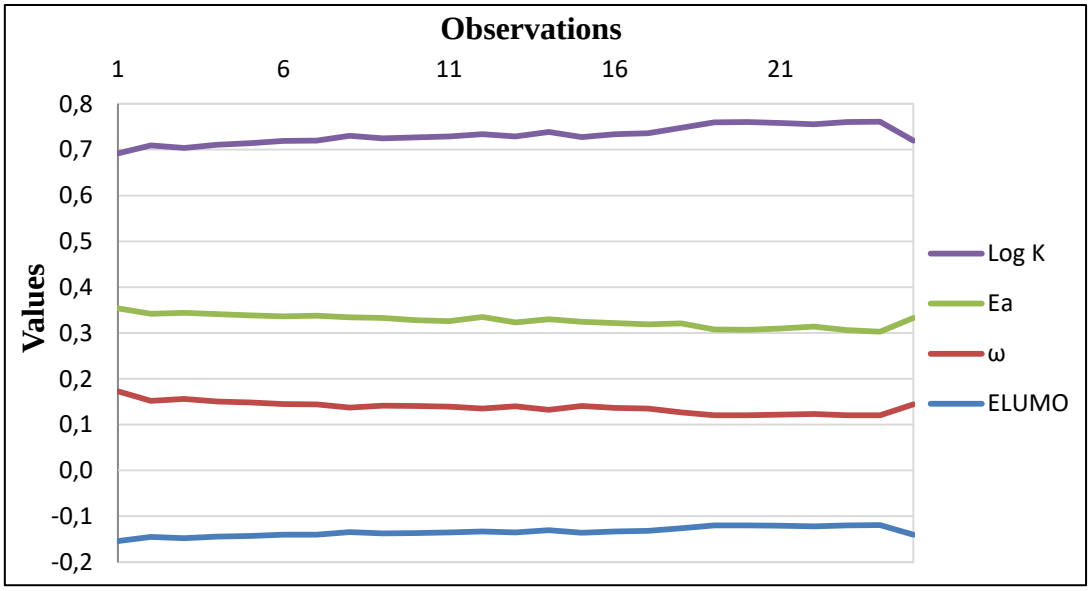


Figure 2: Variation of the Log *K* with the selected descriptors by the RLM

The predicted Log *K* values calculated from Eq. (1), using the optimal MLR model, are given in table 3 in comparison to the observed values. The correlation between the predicted and observed Log *K* (training set and test set) is illustrated in figure 3. The descriptors proposed in equation 1 by MLR were therefore used as the input parameters in the multiple non-linear regression (MNLR).

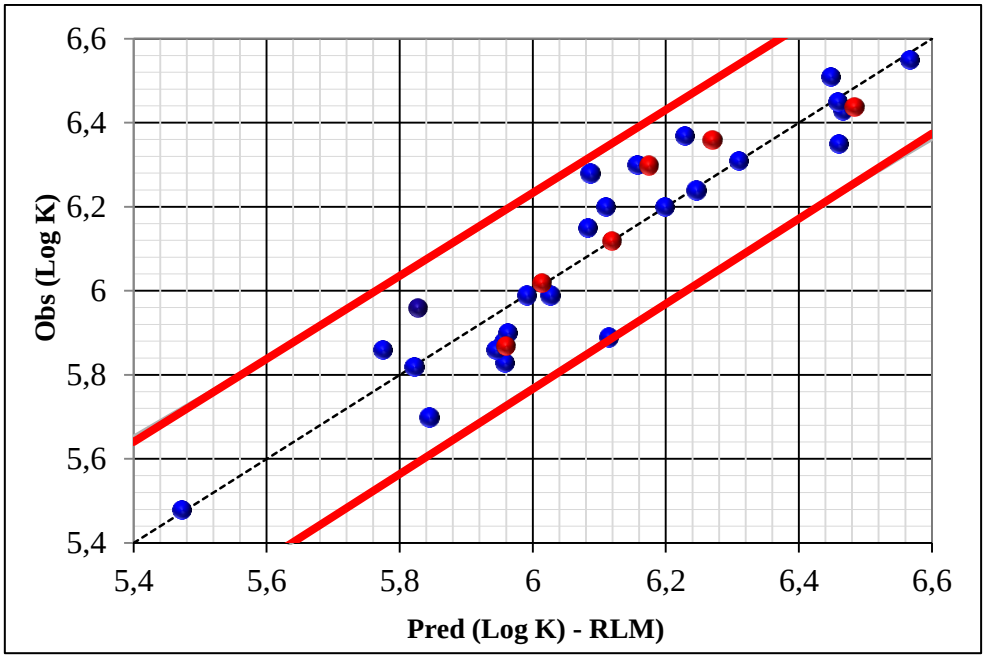


Figure 3: Correlations of observed and predicted Log *K* with MLR (training set is in blue and test set is in red)

3.3. Multiples Nonlinear regression (MNLR)

We also used the non-linear regression model to improve the structure-propriety relationship to quantitatively evaluate the effect of the substituent. We applied the descriptors

proposed by the MLR corresponding to the 25 molecules (training set) to the data matrix. The coefficients, r and r^2, were used to select the best regression performance. We used a pre-programmed function of XLSTAT following:

$$Y = a + (b\,X_1 + c\,X_2 + d\,X_3 + e\,X_4\ldots) + (f\,X_1^2 + g\,X_2^2 + h\,X_3^2 + i\,X_4^2\ldots)$$

Where a, b, c, d... represent the parameters and X_1, X_2, X_3, X_4... represent the variables.

The resulting equation (Equation 2):

$$\text{Log } K = 5.381 + 22.320\,E_{LUMO} + 10.627\,\omega + 3.151\,Ea + 5.445\,E_{LUMO}^2 - 1.272\,\omega^2 - 0.926\,Ea^2$$

$$N = 25;\ r = 0.936;\ r^2 = 0.875;\ \text{MSE} = 0.013$$

The predicted Log K values calculated from equation 2 are given in table 3 in comparison to the observed values. The correlation between the predicted and observed Log K values is shown in figure 4.

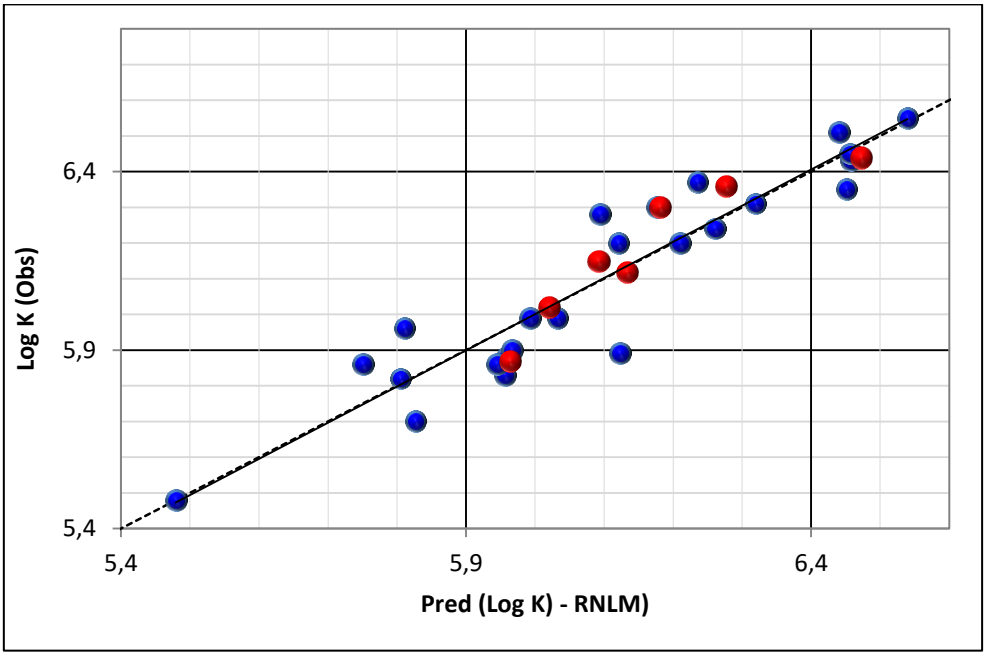


Figure 4: Correlations of observed and predicted Log K with MNLR (training set in bleu and test set in red).

The true predictive power of a QSPR model is to test their ability to accurately predict the Log K of compounds from an external test set (compounds that were not used for the model development). The Log K values for the remaining set of 6 compounds were deduced from the quantitative model proposed using the 25 molecules (training set) by MLR and MNLR. Their structures are given in table 4; the observed and calculated log K values are given in table 5.

Table 3: The observed, the predicted Log *K*, and residue according to RLM and RNLM

Compound	Obs	RLM		RNLM	
	Log K_{obs}	Log K_{RLM}	Residue	Log K_{RNLM}	Residue
1	5.480	5.472	0.008	5.479	0.001
2	5.860	5.774	0.086	5.751	0.109
3	5.700	5.844	-0.144	5.827	-0.127
4	5.820	5.821	-0.001	5.805	0.015
5	5.960	5.827	0.133	5.811	0.149
7	5.880	5.956	-0.076	5.959	-0.079
8	5.830	5.957	-0.127	5.957	-0.127
9	5.900	5.962	-0.062	5.967	-0.067
10	5.990	5.990	0.000	5.993	-0.003
12	6.150	6.083	0.067	6.092	0.058
13	6.200	6.110	0.090	6.121	0.079
14	5.860	5.944	-0.084	5.945	-0.085
15	6.300	6.157	0.143	6.176	0.124
17	5.990	6.026	-0.036	6.032	-0.042
18	6.200	6.198	0.002	6.210	-0.010
19	6.240	6.245	-0.005	6.260	-0.020
20	6.370	6.228	0.142	6.235	0.135
22	6.280	6.086	0.194	6.094	0.186
23	6.430	6.465	-0.035	6.457	-0.027
24	6.450	6.458	-0.008	6.456	-0.006
25	6.350	6.460	-0.110	6.451	-0.101
27	6.310	6.309	0.001	6.319	-0.009
28	6.510	6.447	0.063	6.440	0.070
29	6.550	6.566	-0.016	6.539	0.011
30	5.890	6.114	-0.224	6.123	-0.233

Table 4: The values of the parameters obtained by DFT calculation for the test set compounds

N°	E_{LUMO}	ω	Ea
6	-2.195	4.469	2.943
11	-2.109	4.275	2.918
16	-1.962	3.964	2.787
21	-1.831	3.676	2.836
26	-1.700	3.408	2.646
31	-2.013	4.070	2.838

Table 5: The observed, the predicted Log *K*, and residue according to MLR and MNLR for the 6 tested compounds (test set).

Compound	Obs.	RLM		RNLM	
	Log K_{obs}	Log K_{RLM}	Residue	Log K_{RNLM}	Residue
6	5.870	5.958	0.088	5.964	0.094
11	6.020	6.013	-0.007	6.020	0.000
16	6.360	6.269	-0.091	6.276	-0.084
21	6.120	6.118	-0.002	6.133	0.013
26	6.440	6.483	0.043	6.473	0.033
31	6.300	6.174	-0.126	6.180	-0.120

A comparison of the log (*K*-test) to the log (*K*-obs) values shows that the model made good predictions for the 6 compounds:

For the MLR: N= 6; r_{test}= 0.932; r^2_{test} = 0.869

For the MNLR: N= 6; r_{test}= 0.939; r^2_{test} = 0.881

From the results obtained by MLR and MNLR, we can conclude that the model performs well, as further supported by the results obtained from testing the 6 test compounds. Even if this good predictive power is the result of chance, we can claim that this is a positive result. Accordingly, this model could be applied to all 9-aniliioacridine derivatives in table 1 and add further knowledge to improve the search of antitumor drugs.

3.4. Applicability Domain (AD)

The applicability domain (AD) of these models was evaluated by leverage analysis expressed as Williams plot (Figure 5), in which the standardized residuals and the leverage threshold values (h*= 0.400) were plotted. Any new value of predicted Log *K* data must be considered reliable only for those compounds that fall within this AD on which the model was constructed.

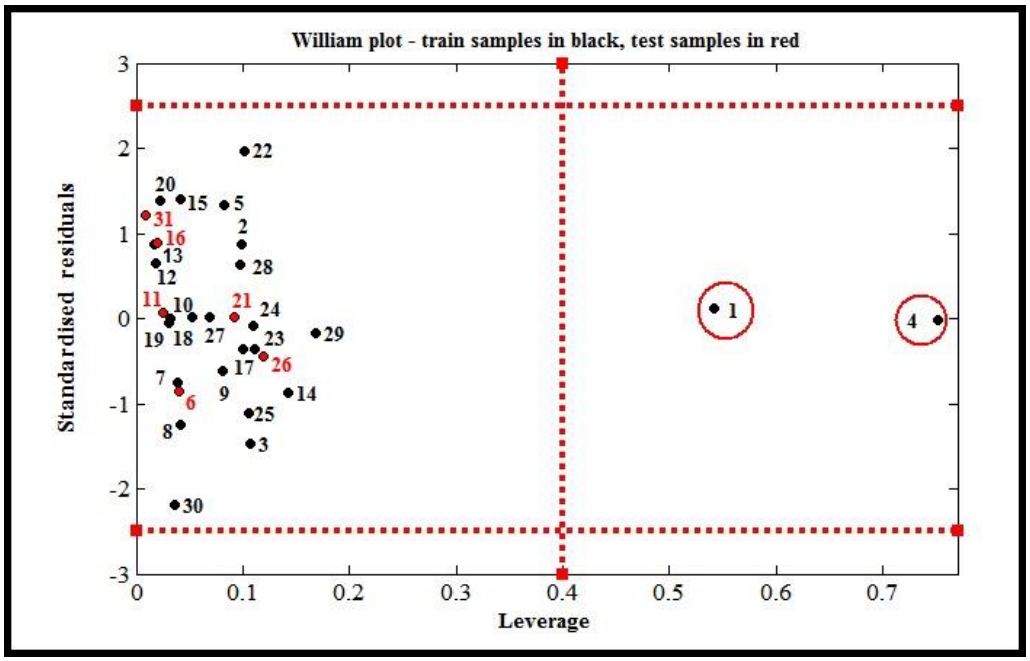


Figure 5: Williams plot of standardized residual versus leverage for the MLR model.

From the figure 5, it is obvious that there are two responses outliers both in the training set and no response outside in test set; and no compound have a standard deviation in the out of the ±x interval (x=2.5). Only two chemicals are identified as an outlier for MLR model; these outliers are the compounds N° 1 and N° 4 in training set has a higher leverage which is greater than h* value of 0.400. These erroneous predictions could probably be attributed to wrong experimental data or to the structural of theses outliers.

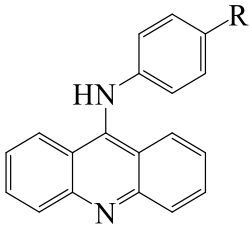
A comparison of the quality of the MLR and MNLR models shows that the 2 approaches have a better predictive capability as they give better results. MLR and MNLR were able to establish a satisfactory relationship between the molecular descriptors and the drug-DNA propriety of the studied compounds.

QSPR correlates propriety data with the physicochemical and/or structural properties of a group of compounds. It has been frequently used to predict the proprieties of new compounds and to design compounds with desired properties.

The developed equations can be used for the design of new 9-aniliioacridine derivatives with improved association constants for DNA drug binding properties (log *K*). For example, Eq. (1) (for RLM) and Eq. (2) (for RNLM) indicated the negative correlation of valance first order E_{LUMO}, ω and Ea. If we develop a new compound with higher values than the existing compounds, it may give rise to the development of more active compounds than those currently in use. In this way, we have designed new compounds (Table 6) by adding suitable substituents and calculated their propriety using Equations (1) and (2).

3.5. Proposed novel compounds with higher DNA-drug binding propriety values

Table 6: The proposed novel compounds



N°	*R*	E_{LUMO}	*ω*	*Ea*	*Log K - RLM*	*Log K - RNLM*	*hi*
X_1	$N(C_2H_2C_2H_4)_2$	-2.075	4.365	3.707	3,487	3,617	2.386
X_2	CN	-2.343	4.808	2.982	5.843	5.826	0.147
X_3	CF_3	-2.204	4.482	2.997	5.848	5.841	0.063
X_4	CCl_3	-2.248	4.585	2.979	5.878	5.874	0.088
X_5	CHO	-2.294	4.696	2.953	5.917	5.914	0.128
X_6	CBr_3	-2.231	4.550	2.949	5.944	5.945	0.091
X_7	CH_2F	-2.012	4.059	2.929	5.979	5.979	0.053
X_8	CMe_3	-1.912	3.845	2.890	6.034	6.038	0.078
X_9	CMe_2Ph	-1.905	3.830	2.883	6.047	6.053	0.078
X_{10}	OPh	-1.987	4.012	2.853	6.136	6.141	0.047
X_{11}	Ph	-1.998	4.038	2.815	6.220	6.226	0.049
X_{12}	$N(PhCl)_2$ortho	-1.935	3.920	2.651	6.543	6.536	0.055
X_{13}	$N(PhCl)_2$meta	-2.125	4.378	2.582	6.611	6.645	0.206
X_{14}	$N(PhCl)_2$para	-2.115	4.368	2.532	6.687	6.721	0.230
X_{15}	$N(o\text{-}C_6H_4CH_3)_2$	-1,880	3,824	2,470	6,883	6,835	0.083
X_{16}	NPh_2	-1.963	4.032	2.415	6.947	6.926	0.164
X_{17}	$N(m\text{-}C_6H_4CH_3)_2$	-1,941	3,987	2,389	6,997	6,965	0.159
X_{18}	$N(p\text{-}C_6H_4CH_3)_2$	-1.913	3.929	2.356	7.061	7,010	0.153

The values of the parameters obtained by DFT calculations for the proposed compounds with an association constant for DNA-drug binding properties (log *K*) based on the information derived from Equations (1) and (2), and the leverage values for the new designed compounds (calculated using ($h_i = x_i^T (X^T X)^{-1} x_i$)) are given in table 6.

From the table 6, it has been observed that the designed compounds (X_{13}, X_{14}, X_{15}, X_{16}, X_{17} and X_{18}) have higher Log *K* values than the existing compounds in the case of the 31 studied compounds (Table 1). These compounds are having a higher leverage which is greater than h*=0.400.

We suggested all these six compounds as candidates that will be synthesized and evaluated as compounds with higher association constant for drug-DNA propriety values.

4. Conclusion

Multiple linear and nonlinear regressions were used to construct a quantitative structure-propriety relation model of 9-aniliioacridine derivatives for their DNA drug binding proprieties. The two regression methods were compared and had a substantially better predictive capability with a greater power. The results show that the models proposed in this paper can predict the association constant for drug-DNA values accurately and that the selected electronic parameters (lowest unoccupied molecular orbital energy, E_{LUMO}, electrophilicity index, ω, and activation energy, Ea), which are sufficiently rich in electronic information to encode structural features, could be used with other descriptors in the development of predictive QSPR models. The accuracy and predictability of the proposed models were illustrated by comparing the key statistical terms *r* or r^2 for the two models (Table 3), and the predictive powers of the equations were validated by an external test set (Table 5). The applicability domain of the MLR models was investigated using William's plot to detect the subspace of chemical structures that can be predicted reliably by models.

We conclude that the most important finding from this research is that we have been able to design and propose new compounds with higher or lower values than existing compounds (Table 6) by adding suitable substituents by calculating their propriety using the regression equations. Consequently, the proposed models will reduce the time and cost of synthesis as well as the determination of the DNA drug binding capacity of 9-aniliioacridine derivatives.

References

[1] S.A. Gamage, D.P. Figgitt, S.J. Wojcik, R.K. Ralph, A. Ransijn, J. Mauel, V. Yardley, D. Snowdon, S.L. Croftand, and W.A. Denny, "Structure-activity relationships for the antileishmanial and antitrypanosomal activities of 1 -substituted 9-anilinoacridines, *Journal of Medicinal Chemistry*, 40(16), **1997**, 2634-2642.

[2] V. Nadaraj, S.T. Selvi, and S. Mohan, "Microwave-induced synthesis and anti-microbial activities of 7,10,11,12-tetrahydrobenzo[c]acridin-8(9H)-one derivatives", *European Journal of Medicinal Chemistry*, 44(3), **2009**, 976–980.

[3] B.F. Dickens, W.B. Weglicki, P.A. Boehme, and T. I. Mak, "Antioxidant and Lysosomotropic Properties of Acridine-propranolol: Protection against Oxidative Endothelial Cell Injury", *Journal of Molecular and Cellular Cardiology*, 34(2), **2002**, 129–137.

[4] S.A. Gamage, N. Tepsiri, P. Wilairat, S.J. Wojcik, D.P. Figgitt, R.K. Ralph, and W.A. Denny, "Synthesis and in vitro Evaluation of 9-Anilino-3,6-diaminoacridines Active Against a Multidrug-Resistant Strain of the Malaria Parasite Plasmodium falciparum", *Journal of Medicinal Chemistry*, 37(10), **1994**, 1486–1494.

[5] Y.L. Chen, C.M. Lu, I.L. Chen, L.T. Tsao, and J.P. Wang, "Synthesis and Antiinflammatory Evaluation of 9-Anilinoacridine and 9-Phenoxyacridine Derivatives", *Journal of Medicinal Chemistry*, 45(21), 2002, 4689–4694.

[6] S. M. Sondhi, M. Johar, N. Singhal, R. Shukla, R. Raghubir, and S.G. Dastidar, "Synthesis of sulpha drug acridine derivatives and their evaluation for anti-inflammatory, analgesic and anticancer activity", *Indian Journal of Chemistry - Section B Organic and Medicinal Chemistry*, 41(12), **2002**, 2659–2666.

[7] I. Antonin, "DNA-binding Antitumor Agents: from Pyrimido[5,6,1-de]acridines to Other Intriguing Classes of Acridine Derivatives", *Current Medicinal Chemistry*, 9(18), **2002**, 1701–1716.

[8] M. Demeunynck, A. Charmantray, and A. Martelli, "Interest of acridine derivatives in the anticancer chemotherapy", *Current Pharmaceutical Design*, 7(17), **2001**, 1703–1724.

[9] K.A. Werbovetz, P.G. Spoors, R.D. Pearson, and T.L. MacDonald, "Cleavable complex formation in Leishmania chagasi treated with anilinoacridines", *Molecular and Biochemical Parasitology*, 65(1), **1994**, 1–10.

[10] C.S. Rouvier, J.M. Barret, C.M. Farrell, D. Sharples, B.T. Hill, and J. Barbe, "Synthesis of 9-acridinyl sulfur derivatives: sulfides, sulfoxides and sulfones. Comparison of their activity on tumour cells", *European Journal of Medicinal Chemistry*, 39(12), **2004**, 1029–1038.

[11] K. Rastogi, J. Y. Chang, W. Y. Pan, C.H. Chen, T.C. Chou, L.T. Chen, and T.L. Su, "Antitumor AHMA Linked to DNA Minor Groove Binding Agents: Synthesis and Biological Evaluation", *Journal of Medicinal Chemistry*, 45(20), **2002**, 4485–4493.

[12] K.M. Chen, Y.W. Sun, Y.W. Tang, Z.Y. Sun, and C.H. Kwon, "Synthesis and Antitumor Activity of Sulfur-Containing 9-Anilinoacridines", *Molecular Pharmaceutics*, 2(2), **2005**, 118–128.

[13] A. Albert, "The Acridines", *2nd ed. Edward Arnold: London*, **1966**. "Selective Toxicity", *7th ed. Chapman & Hall: London*, **1985**.

[14] M. Kimura and I. Okabayashi, "Formation and molecular structure of the novel acridine substituted uracil derivatives", *Journal of Heterocyclic Chemistry*, 23(3), **1986**, 965–967

[15] M. Demeunynck, F. Charmantray, and A.Martelli, "Interest of acridine derivatives in the anticancer chemotherapy", *Current Pharmaceutical Design*, 7, **2001**, 1703–1724.

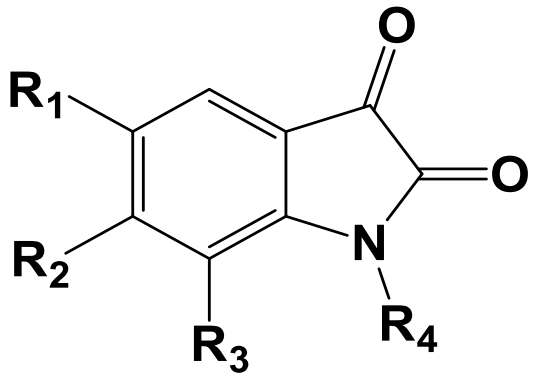
[16] M.K. Goftar, N.A. Rayeni, and N. Mohamadi, "Spectroscopic studies on the interaction between acridines permine conjugate with DNA", *International Journal of Biosciences*, 5(4), **2014**, 27–33.

[17] R.F. Lynnette and A.D. William, "The genetic toxicology of acridines", *Mutatton Research*, 258, **1991**, 123–160

[18] R.B. Silverman and M.W. Holladay, "DNA-Interactive Agents", *Organic Chemistry of Drug Design and Drug Action*, **2014**, 275–331.

[19] K. Drlica and R.J. Franco, "Inhibitors of DNA topoisomerases", *Biochemistry*, 27(7), **1988**, 2253–2259
[20] M.J. Waring, "DNA Modification and Cancer", *Annual Review of Biochemistry*, 50, **1981**, 159–192
[21] E.M. Nelson, K.M. Tewey, and L.F. Liu, "Mechanism of antitumor drug action: poisoning of mammalian DNA topoisomerase II on DNA by 4'-(9-acridinylamino)-methanesulfon-m-anisidide", *Proceedings of the National Academy of Sciences of the United States of America*, 81(5), **1984**, 1361–1365
[22] M.J. Waring, "DNA-binding characteristics of acridinylmethanesulfonanilide drugs: comparison with antitumor properties", *European Journal of Cancer*, 12, **1976**, 995–1001.
[23] M. Kimura, A. Kato, and I. Okabayashi, "Acridine derivatives. IV. Synthesis, molecular structure, and antitumor activity of the novel 9-anilino-2, 3-methylenedioxyacridines", *Journal of Heterocyclic Chemistry*, 29 (1), **1992**, 73–80.
[24] M. Kimura, "Quenching of ethidium-DNA fluorescence by novel acridines with antitumor activities", *Yakugaku Zasshi*, 112 (12), **1992**, 914–918
[25] C.B. Bruce, A.D. William, J.A. Graham, and F.C. Bruce, "Potential Antitumor Agents: Quantitative Relationships between DNA Binding and Molecular Structure for 9-Anilinoacridines Substituted in the Anilino Ring", *Journal of Medicinal Chemistry*, 24, **1981**, 170–177.
[26] B.C. Baguley, W.A. Denny, G.J. Atwell, and B.F. Cain, "Potential antitumor agents, Quantitative relationships between DNA binding and molecular structure for 9-anflinoacridines substituted in the anilino ring", *Journal of Medicinal Chemistry*, 34, **1981**, 107–177.
[27] C. Adamo and V. Barone, "A TDDFT study of the electronic spectrum of s-tetrazine in the gas-phase and in aqueous solution", *Chemical Physics Letters*, 330, **2000**, 152–160.
[28] L. Becker, K. Hinrichs, and U. Finke, "A New Algorithm for Computing Joins with Grid Files, In Proc. of the 9th International Conference on Data Engineering", *Vienna, Austria.*, **1993**, 190-197.
[29] S. Chtita, M. Ghamali, M. Larif, A. Adad, R. Hmammouchi, M. Bouachrine, and T. Lakhlifi, "Prediction of biological activity of imidazo[1,2-a] pyrazine derivatives by combining DFT and QSAR results", *International Journal of Innovative Research in Science, Engineering and Technology*, 2 (12), **2013**, 7951–7962.
[30] XLSTAT **2009** Add-in software, *XLSTAT Company*. www.xlstat.com
[31] S. Chtita, M. Larif, M. Ghamali, M. Bouachrine, and T. Lakhlifi, "Quantitative structure–activity relationship studies of dibenzo[a,d]cycloalkenimine derivatives for non-competitive antagonists of N-methyl-D-aspartate based on density functional theory with electronic and topological descriptors", *Journal of Taibah University for Science*, 9, **2015**, 143–154.
[32] OECD, "Guidance Document on the Validation of QSAR Models", *Organization for Economic Co-Operation & Development, Paris, France*, **2007**.
[33] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, and P. Gramatica, "Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs", *Environmental Health Perspectives*, 111(10), **2003**, 1361–1375.
[34] G.E. Batista, D.F. Silva, "How k-nearest neighbor parameters affect its performance, in Proceedings of the Argentine Symposium on Artificial Intelligence", *Instituto de Ciencias Matematicase de Computa cao, Sao Carlos, Brazil*, **2009**, 1–12.
[35] T.I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. van de Sandt, Weida Tong, G. Veith, and C. Yang, "Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships", *Alternatives to Laboratory Animals*, 33(2), **2005**, 1–19.
[36] J.C. Dearden, "The history and development of quantitative structure-activity relationships (QSARs)", *International Journal of Quantitative Structure-Property Relationships*, 1(1), **2016**, 1–44.

Chapitre 4 : Etude de la RQSA de l'activité anticancéreuse pour des dérivés de l'Isatine



International Journal of Quantitative Structure-Property Relationships
Volume 2 • Issue 2 • January-June 2017

Quantitative structure–activity relationship studies of anticancer activity for Isatin (1H-indole-2,3-dione) derivatives based on density functional theory

Samir Chtita [(a) (*)], Mounir Ghamali [(a)], Majdouline Larif [(b)], Rachid Hmamouchi [(a)], Mohammed Bouachrine [(c)] and Tahar Lakhlifi [(a)]

[(a)] MCNS Laboratory, Faculty of Science, University Moulay Ismail, Meknes, Morocco
[(b)] Separation Process Laboratory, Faculty of Science, University Ibn Tofail, Kenitra, Morocco
[(c)] High school of technology, University Moulay Ismail, Meknes, Morocco
[(*)] Corresponding Author: E-mail: samirchtita@gmail.com; Tel. +212 660.005.554

Abstract:

To establish a QSAR of anticancer activity for Isatin derivatives, a series of Isatin derivatives were analyzed by principal component analysis, multiple linear regression, partial least squares and multiple nonlinear regression analysis. We proposed linear and nonlinear models and we interpreted the activity of the compounds by multivariate statistical analysis. The proposed models were used to predict the activity of test set compounds, and an agreement between experimental and predicted values was verified. The applicability domain of MLR models was investigated using William's plot to detect outliers and outsides compounds. For the successful application of the developed models to predict new compounds, rigorous validation tests have been used in this direction. Additionally, the r_m^2 metrics have been used to ensure the close agreement of predicted response data with observed ones. The developed models have been used for designing some new Isatin derivatives with high predicted values of anticancer effect.

Keywords: Quantitative structure–activity relation, Density functional theory, Anticancer, Isatin, Cross Validation.

1. Introduction

Isatin (1H-indole-2,3-dione) is an indole derivative and an important class of heterocyclic compounds found in many plants, such as Isatistinctoria, Calanthe discolor and Couroupitaguianensis. It is also found in humans as a metabolic derivative of adrenaline **[1-2]**. Isatin was first obtained as a product from chromic acid oxidation of indigo dye by Erdmann and Laurent in 1841 **[3-4]**, and their synthetic derivatives are important substrates used for the synthesis of a variety of heterocyclic compounds, and used as raw materials for drug synthesis. Isatin and its derivatives are well known therapeutic agents due to their wide range of pharmacological and biological activities including anticancer **[5-6]**, anticonvulsant **[7]**, antiviral **[8-9]**, antibacterial and antifungal **[10]**, anti-HIV and anti-inflammatory activities **[11-13]**. Isatin is used as a starting point in the synthesis of oligomeric or polymeric structures which are used in the field of solar energy **[14]**, organic memory devices creation **[15]** and organic field effect transistors **[16-17]**.

Many methodologies have been adopted to synthesize Isatin derivatives and to explore their possible role in the treatment of various diseases. Among these protocols, the method developed by Sand-Meyer is the oldest and the most frequently used for the synthesis of Isatin **[18]**. This method involves the reaction of aniline with chloral hydrate and hydroxylamine hydrochloride in aqueous sodium sulfate to form an isonitrosoacetanilide, which after isolation, when treated with concentrated sulfuric acid produces Isatin.

Looking at the literature, it can be seen that in particular, halogenated isatin derivatives have been reported to exhibit anticancer activity. 5-Bromo-3-o-nitrophenyl isatinhydrazone and 5-bromo-(2-oxo-3-indolinyl) thiazolidine-2,4-diones substituted by various Mannich bases were found to exhibit anticancer activity against Walker carcinoma-256 and P388 lymphocytic leukemia in mice, respectively **[19-20]**. 6-Bromo-2- methylthio-3H-indol-3-one, tyrindoleninone, a brominated precursor to Tyrian purple, isolated from the egg masses of the Australian mollusk Dicathaisorbita has been reported as a cytotoxic marine compound **[21]**. 6-bromoisatin, a major decomposition product formed through the oxidation of tyriverdin (precursor of Tyrian purple), has been shown to have a weaker anti-cancer activity against a human lymphoma cell line in comparison with 6-Bromo-2- methylthio-3H-indol-3-one **[21-22]**. 5,7-Dibromoisatin, a significantly more potent as a cytotoxin than Isatin against U937 (human monocyte-like histiocytic lymphoma) cells, its N-benzyl derivatives with more cytotoxicity toward these lymphoma cells and activity against a range of human cancer cell lines including a metastatic breast adenocarcinoma cell line (MDA-MB-231) **[23]**, and cytotoxic N-alkylhaloisatins are some examples of reported anticancer halogenated Isatins in recent researches **[23-24]**. In the recently approved drugs by FDA, a 5-fluoro-3-substituted-2-oxoindole, SU11248 (Sutent) is provided for the treatment of gastrointestinal stromal tumors and advanced renal-cell carcinoma **[25-26]**.

The experiment is a direct way to obtain the activity data of organic compounds, which has many deficiencies, such as the requirement of large trial organisms, high expense, long time duration, difference in measured value between different researchers. Consequently, it would be impossible to search the activity values of all organic compounds by experiments. As new compounds are springing up, other difficulties will also arise. Therefore, it is necessary to use the theoretical methods to make up the disadvantages of the experiment and to predict the data of compounds exactly.

With the rapid development of computer science and theoretical quantum chemical studies, one can speedily and precisely obtain the quantum chemical parameters of compounds by computation. These structural parameters along with the introduction of the

quantitative structure-activity relationship (QSAR) models can increase the interpretability and predict the activity of new organic compounds. In this paper, the authors will focus on the anticancer activity of some Isatin derivatives.

In order to open a new way in anticancer drug research, a series of 40 Isatin derivatives were studied for their anticancer activity against U937 cells **[23-24, 27]**. The aim of this study is to develop QSAR models able to correlate the structural features of the derivatives of Isatin with their anticancer activity. A variety of molecular descriptors were calculated to develop models with the studied activities of a set of molecules. The principal component analysis (PCA), the descendant and the stepwise multiple linear regression (MLR), the partial least squares (PLS) and the multiple nonlinear regression (MNLR) were used as statistical techniques. The authors accordingly propose quantitative models, and they try to interpret the activity of the compounds relying on the multivariate statistical analyses. The prediction models were subjected to rigorous independent testing via internal and external tests.

2. Materiel and Methods

2.1. Experimental dataset

The reliability of the QSAR analysis depends on the available data set, the method of analysis and the validation tests. In the present analysis, a series of 40 selected Isatin derivatives that have been synthesized and evaluated for their anticancer activity against U937 cells, as demonstrated by *K.L. Vine* et al. **[23-24, 27]** were considered to carry out the QSAR analysis (32 molecules were selected to propose the quantitative model (training set), and remaining 8 compounds were selected to test the performance of the proposed model (test set)). The division of the dataset into training and test sets has been performed using the k-means clustering technique **[28]**. In this method, from each obtained cluster at least one compound for the training set was randomly selected. Table 1 shows the chemical structures of studied compounds, the experimental anticancer activity values of the studied compounds have been collected from the references **[23-24, 27]**. All experimental IC_{50} activity values (µM) were converted to the negative logarithm of IC_{50}, ($pIC_{50} = -\log_{10} (IC_{50})$). The biological data used in this study were the anticancer activity against U937 (human monocyte-like histiocytic lymphoma) cells.

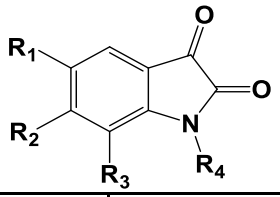
2.2. Computational methods

An attempt has been made to correlate the studied activity of these compounds with various physicochemical parameters. Density functional theory (DFT) was used in this study, the 3D structures of the molecules were generated using the Gauss View 3.0 software **[29]**, and then, all calculations were performed using Gaussian 03W software **[30]**. Geometry

optimization of the 40 compounds was carried out using the Becke's three-parameter hybrid method and the Lee-Yang-Parr B3LYP functional [31] employing 6–31G (d) basis set [32]. The geometry of all species under investigation was determined by optimizing all geometrical variables without any symmetry constraints [33].

The use of DFT to calculate different descriptors is justified for the reason that some comparative QSAR studies have shown that the descriptors calculated using the density functional theory (DFT) method can improve the accuracy of the results and lead to more reliable QSARs [34-37].

Table 1: Chemical structures of Isatin derivatives used in this study and their experimental activity for anticancer activity against U937 cells



N°	R_1	R_2	R_3	R4	pIC_{50}	N°	R_1	R_2	R_3	R_4	pIC_{50}
A1	Br	H	Br	$H_2CCH=CH_2$	5.18	**A21**	F	H	H	H	4.01
A2	Br	H	Br	$H_2CCH_2OCH_3$	5.46	**A22**	NO_2	H	H	H	3.88
A3	Br	H	Br	$H_2CCH_2CH(CH_3)_2$	5.62	**A23**	OCH_3	H	H	H	3.38
A4	Br	H	Br	$H_2CC_6H_5$	5.94	**A24**	Br	H	Br	H	4.98
A5	Br	H	Br	$H_2CC_6H_4CH_3$ [b]	6.31	**A25**	Br	Br	H	H	4.94
A6	Br	H	Br	$H_2CC_6H_4OCH_3$ [b]	5.74	**A26**	Br	H	NO_2	H	3.59
A7	Br	H	Br	$H_2CC_6H_4OCH_3$ [c]	5.75	**A27**	Br	Br	Br	H	5.17
A8	Br	H	Br	$H_2CC_6H_4NO_2$ [b]	6.05	**A28**	H	H	H	CH_3	3.62
A9	Br	H	Br	$H_2CC_6H_4NO_2$ [d]	5.64	**A29**	Br	H	Br	$H_2CCH_2C_6H_5$	6.11
A10	Br	H	Br	$H_2CC_6H_4Cl$ [b]	6.01	**A30**	Br	H	Br	$H_2CCH_2C_6H_4Br$ [c]	6.11
A11	Br	H	Br	$H_2CC_6H_4Br$ [b]	6.20	**A31**	Br	H	Br	$H_2CCH_2C_6H_4Br$ [b]	6.06
A12	Br	H	Br	$H_2CC_6H_4CF_3$ [b]	6.10	**A32**	Br	H	Br	$H_2CCH_2C_6H_4OCH_3$ [c]	5.97
A13	H	Br	H	$H_2CC_6H_4CF_3$ [b]	5.28	**A33**	Br	H	Br	$H_2CCH_2C_6H_4OCH_3$ [b]	5.63
A14	Br	H	Br	$H_2CC_6H_4COOCH_3$ [b]	5.92	**A34**	Br	H	Br	$CH_2C_{10}H_7$ [e]	6.72
A15	Br	H	Br	$H_2CC_6H_4C(CH_3)_3$ [b]	5.95	**A35**	Br	H	Br	$CH_2C_{10}H_7$ [f]	6.13
A16	Br	H	Br	$H_2CC_6H_4C_6H_5$ [b]	6.12	**A36**	Br	H	Br	$CH_2COC_6H_5$	5.00
A17	H	H	H	H	3.25	**A37**	Br	H	Br	$CH_2COC_6H_4Br$ [c]	5.20
A18	Br	H	H	H	4.19	**A38**	Br	H	Br	$CH_2COC_6H_4Br$ [b]	5.04
A19	H	Br	H	H	4.13	**A39**	Br	H	Br	$CH_2COC_6H_4OCH_3$ [c]	5.33
A20	H	H	Br	H	4.08	**A40**	Br	H	Br	$CH_2COC_6H_4OCH_3$ [b]	5.27

[a] pIC_{50}= -log (IC_{50}).
[b] Substitutions at Para position. [c] Substitutions at Meta position. [d] Substitutions at Ortho position.
[e] 1-naphthylmethyl. [f] 2-naphthylmethyl.

2.2.1. Calculation of molecular descriptors

From the results of the DFT calculations, the quantum chemistry descriptors were obtained for the building model as follows: the total energy E_T (eV), the highest occupied molecular orbital energy E_{HOMO}(eV), the lowest unoccupied molecular orbital energy E_{LUMO} (eV) and their difference in absolute value Gap (eV), the dipole moment μ (Debye), the absolute hardness η (eV), the absolute electronegativity χ (eV) and the electrophilicity index

ω (eV) were deduced from the stable structure of the neutral form [38]. η, χ and ω were determined from:

$$\eta = \frac{E_{LUMO} - E_{HOMO}}{2}$$

$$\chi = -\frac{(E_{LUMO} + E_{HOMO})}{2}$$

$$\omega = \frac{\chi^2}{2\eta}$$

Six other chemical descriptors of the studied molecules were calculated using Marvin Sketch Software [39]: Octanol/Water partition coefficient, Log P, Balaban Index, J, Wiener Index, WI, polarizability, α (Å^3), Van Der Waals volume, VDWV (Å^3), and molar refractivity, MR (Å^3/mole).

2.2.2. Statistical analysis

To explain the structure-activity relationship, these 14 descriptors calculated for the 40 molecules (Table 2) were subjected to a principal component analysis (PCA), to a stepwise and a descendant multiple linear regression (MLR), to a partial least squares (PLS) and to a nonlinear regression (MNLR) available in the XLSTAT software [40].

The principal component analysis is a statistical technique used for summarizing information encoded in the structures of compounds and for understanding the distribution. It is essentially a descriptive statistical method for presenting the maximum information contained in table 1 and table 2 in graphical form.

The multiple linear regression statistic techniques are used to study the relation between one dependent and several independent variables. It is a mathematic technique that minimizes differences between actual and predicted values. It has served also to select the descriptors used as the input parameters in the multiple nonlinear regression (MNLR).

MLR, PLS, and MNLR were generated to predict anticancer activity values (pIC_{50}). Equations were justified by the conventional approach: correlation coefficient (r), determination coefficient (r^2), adjusted (r_{adj}^2), Mean Squared Error (*MSE*), and significance level (p-value) [41]. These various statistical metrics has been defined in the following manners:

$$r^2 = 1 - \frac{\sum(Y_{obs} - Y_{cal})^2}{\sum(Y_{obs} - \bar{Y}_{cal})^2} \qquad (1)$$

$$r_{adj}^2 = \frac{(N-1)\times r^2 - p}{N-1-p} \qquad (2)$$

$$MSE = \frac{1}{n}\sum(Y_{obs} - Y_{calc})^2 \qquad (3)$$

The p-value, the significance level of the test, is used in the context of null hypothesis testing in order to quantify the idea of statistical significance of evidence, the threshold value for p-value is, traditionally, 5% or 1% **[42]**.

Here: $\boldsymbol{Y_{obs}}$ is the observed response value; $\boldsymbol{Y_{cal}}$ is the predicted response value; $\boldsymbol{\bar{Y}_{cal}}$ is average of the observed response values; p is the number of predictor variables used in the model development.

The next stage of this QSAR analysis consists of a statistical validation in order to assess the significance of the model and hence its ability to predict biological activities of other (novel) compounds.

In this paper, the models were validated internally by leave-one-out cross-validation followed by the calculation of the r_{cv}^2 cross-validated squared coefficient. This method involves removal of one of the compounds from the training set followed by the development of the QSAR model based on the reduced dataset. The model thus built with the remaining molecules is used to predict the response of the deleted compound. This cycle is repeated until all the molecules of the dataset have been deleted once. The cross-validated squared coefficient, r_{cv}^2 computed by the following equation:

$$r_{cv}^2 = 1 - \frac{\sum(Y_{obs}(\text{train}) - Y_{cal(train)})^2}{\sum(Y_{obs(train)} - \bar{Y}_{obs(train)})^2} \qquad (4)$$

With: $\mathbf{Y_{obs(train)}}$ is the observed response value; $\mathbf{Y_{obs(train)}}$ is the LOO predicted response value; $\mathbf{\bar{Y}_{train}}$ is the average of the observed response values;

A high value of r_{cv}^2(>0.5) indicates a better predictivity of the model.

The authors include the y-randomization approach (randomization of response, i.e., activity in our case) to ensure the robustness of a predictive model. Often, it is used along with the cross-validation. It consists of repeating the calculation procedure with randomized activities and subsequent probability assessment of the resultant statistics. The dependent variable vector is randomly shuffled and a new predictive model is developed using the original independent variable matrix. The new predictive models (after several repetitions) are expected to have low r^2 and r_{cv}^2 values. If the opposite happens, then an acceptable model cannot be obtained for the specific modeling method and data **[43]**.

Table 2: Values of parameters calculated for the 40 Isatin derivatives used in this study

N°	E_T	μ	E_{HOMO}	E_{LUMO}	Gap	η	χ	ω	Log P	J	WI	α	VDWV	MR
A1	-157055.270	4.103	-6.541	-3.087	3.454	1.727	4.814	6.710	3.084	1.921	393	25.718	200.329	67.999
A2	-159135.153	4.159	-6.563	-3.109	3.454	1.727	4.836	6.772	2.366	1.942	588	28.554	234.057	74.745
A3	-159228.457	4.172	-6.473	-3.046	3.426	1.713	4.760	6.612	3.964	1.986	573	25.558	242.006	77.259
A4	-161236.330	4.117	-6.519	-3.093	3.426	1.713	4.806	6.743	4.078	1.561	763	31.699	244.640	83.449
A5	-162306.179	4.250	-6.453	-3.066	3.387	1.693	4.760	6.689	4.591	1.467	892	33.434	261.433	88.490
A6	-164352.540	5.165	-6.060	-3.033	3.026	1.513	4.547	6.830	3.920	1.650	1042	34.262	270.714	89.912
A7	-164352.492	5.027	-6.110	-3.018	3.092	1.546	4.564	6.737	3.920	1.692	1014	34.262	270.831	89.912
A8	-166800.787	5.228	-6.822	-3.362	3.461	1.730	5.092	7.492	4.018	1.566	1194	33.805	267.090	90.774
A9	-166800.679	4.817	-6.712	-3.243	3.469	1.734	4.977	7.142	4.018	1.682	1110	33.805	266.935	90.774
A10	-173741.968	4.090	-6.621	-3.186	3.435	1.717	4.903	6.999	4.682	1.489	878	33.679	258.531	88.254
A11	-231196.120	4.048	-6.596	-3.183	3.413	1.706	4.889	7.005	4.846	1.489	878	37.775	262.830	91.072
A12	-170407.106	4.073	-6.698	-3.248	3.450	1.725	4.973	7.167	4.955	1.564	1292	33.216	276.511	89.423
A13	-100447.447	5.129	-6.678	-3.028	3.649	1.825	4.853	6.454	4.187	1.594	1164	30.092	258.159	81.800
A14	-167436.930	5.325	-6.616	-3.167	3.449	1.724	4.891	6.937	4.081	1.543	1313	36.237	289.876	95.474
A15	-165515.285	4.013	-6.443	-3.046	3.396	1.698	4.745	6.628	5.623	1.564	1292	38.863	312.994	102.115
A16	-167523.405	4.175	-6.187	-3.097	3.090	1.545	4.642	6.973	5.725	1.221	1660	42.364	315.160	108.585
A17	-13960.457	5.934	-6.550	-2.651	3.899	1.950	4.600	5.427	1.602	2.054	138	14.544	119.585	40.475
A18	-83920.143	5.203	-6.565	-2.923	3.642	1.821	4.744	6.180	2.370	1.894	178	17.508	137.827	48.098
A19	-83920.180	4.394	-6.775	-2.873	3.902	1.951	4.824	5.964	2.370	1.884	179	17.508	137.818	48.098
A20	-83920.202	5.034	-6.720	-2.891	3.830	1.915	4.806	6.030	2.370	1.943	174	17.513	137.813	48.098
A21	-16660.503	5.358	-6.539	-2.853	3.686	1.843	4.696	5.983	1.744	1.894	178	14.318	124.421	40.691
A22	-19524.873	5.483	-7.318	-3.308	4.011	2.005	5.313	7.038	1.541	2.010	284	16.563	141.599	47.800
A23	-17076.542	7.367	-5.972	-2.566	3.406	1.703	4.269	5.350	1.444	1.720	230	17.091	145.625	46.938
A24	-153879.861	4.099	-6.725	-3.131	3.595	1.797	4.928	6.756	3.139	1.818	218	20.648	156.051	55.721
A25	-153879.780	4.221	-6.717	-3.061	3.656	1.828	4.889	6.537	3.139	1.780	222	20.649	156.112	55.721
A26	-89484.550	2.769	-7.165	-3.520	3.645	1.823	5.343	7.830	2.310	2.022	326	19.650	159.862	55.423
A27	-223839.428	3.677	-6.803	-3.203	3.600	1.800	5.003	6.953	3.908	2.166	265	23.868	174.428	63.343
A28	-15030.218	5.897	-6.363	-2.593	3.770	1.885	4.478	5.318	0.816	1.971	172	16.377	137.147	43.591
A29	-162306.181	4.030	-6.471	-3.094	3.378	1.689	4.783	6.772	4.366	1.415	920	33.506	261.647	88.204
A30	-232265.945	2.279	-6.568	-3.164	3.404	1.702	4.866	6.956	5.135	1.619	1053	36.563	279.699	95.827
A31	-232265.935	3.573	-6.514	-3.180	3.334	1.667	4.847	7.047	5.135	1.599	1068	36.563	279.784	95.827
A32	-165422.349	4.829	-6.040	-3.048	2.992	1.496	4.544	6.900	4.209	1.537	1208	36.069	287.668	94.667
A33	-165422.344	3.594	-5.936	-3.057	2.879	1.440	4.497	7.024	4.209	1.503	1238	36.069	287.658	94.667
A34	-165416.912	4.375	-6.025	-3.082	2.942	1.471	4.553	7.047	5.067	1.390	1245	38.971	287.953	99.899
A35	-165417.022	4.033	-5.980	-3.081	2.899	1.449	4.531	7.081	5.067	1.327	1301	38.970	287.718	99.899
A36	-164320.049	5.890	-6.432	-2.988	3.444	1.722	-4.710	6.443	3.585	1.686	1014	33.669	263.714	88.738
A37	-234279.760	5.464	-6.528	-3.070	3.458	1.729	-4.799	6.659	4.354	1.615	1152	36.753	281.998	96.361
A38	-234279.784	4.667	-6.519	-3.063	3.455	1.728	-4.791	6.643	4.354	1.595	1168	36.753	282.028	96.361
A39	-167436.202	7.101	-6.385	-2.947	3.439	1.719	-4.666	6.332	3.428	1.538	1313	36.233	289.850	95.201
A40	-167436.290	6.916	-6.341	-2.916	3.425	1.712	-4.629	6.256	3.428	1.505	1345	36.233	289.821	95.201

An acceptable value of r_{cv}^2 does not inevitably indicates that the predicted activity data lie in close propinquity to the observed ones although there may exist a good overall correlation between the values. Thus, to obviate this problem and to better indicate the model predictability, the r_m^2 metrics (r_m^2, *Average* r_m^2 ($\overline{r_m^2}$) and *delta* r_m^2 (Δr_m^2)) introduced by *Roy* et al. **[44-45]** as shown in the following equations **[46]** were used:

$$Average\ r_m^2 = \overline{r_m^2} = \frac{(r_m^2 + r_m'^2)}{2} \quad (5)$$

$$Delta\ r_m^2 = \Delta r_m^2 = |r_m^2 - r_m'^2| \quad (6)$$

$$r_m^2 = r^2 \times \left(1 - \sqrt{(r^2 - r_0^2)}\right) \quad (7)$$

$$r_m'^2 = r^2 \times \left(1 - \sqrt{(r^2 - r_0'^2)}\right) \quad (8)$$

The parameter r^2 is the squared correlation coefficient between observed and (leave-one-out) predicted values of the compounds with intercept.

The parameter r_0^2 is the squared correlation coefficient between observed and (leave-one-out) predicted values of the compounds without intercept.

The parameter $r_0'^2$ (consequently $r_m'^2$) bears the same meaning but uses the reversed axes.

The parameter r_m^2 is a novel validation metric calculated according to Eq. (7) using observed and (leave-one-out) predicted values.

The parameter $r_m'^2$ is a novel validation metric calculated according to Eq. (8) using observed and (leave-one-out) predicted values.

The parameter $\overline{r_m^2}$ is the average value of r_m^2 and $r_m'^2$.

The parameter Δr_m^2 is the absolute difference between r_m^2 and $r_m'^2$.

The model with *average* r_m^2 values above the threshold of 0.5 and with a Δr_m^2 value less than 0.2 are considered to be predictive and reliable ones.

The final stage of this QSAR analysis consists of applicability domain, a model is valid only within its training domain, and applicability domain of new molecules must be checked before the model is applied (OECD Principle 3) **[46]**. Without applicability domain (AD), each model can unreliably predict the activity of any compound, probably with a completely different structure from those included in the study. Therefore, the AD is a tool to find out compounds that are outside the applicability domain of the built QSAR model and it detects outliers present in the training set compounds. There are several methods for defining the applicability domain (AD) of QSAR models **[47]**, but the most common one is determining the leverage values h_i ($h_i = x_i^T (X^T X)^{-1} x_i$ ($i = 1,2, \dots, n$)) for each compound **[48]**.

Where: xi is the descriptor row-vector of query compound; X is the n*k-1 matrix of k model descriptor values for n training set compounds, and the superscript "T" refers to the transpose of matrix/vector.

In this study, the authors use the Williams plot; in this plot, the applicability domain is established inside a squared area within standard deviation ±x (in this study x = 2.5 ("three sigma rule" **[49]**) and a leverage threshold h* (h*= 2.5*(k+1)/n) **[50]**.

Here: *n* is the number of training set compounds and *k* is the number of model descriptors.

The leverage (h) greater than the warning leverage (h*) suggested that the compound was very influential on the model **[51]**. A flowchart for the method of development of the QSAR models along with the various validation methods used in this work is demonstrated in figure 1.

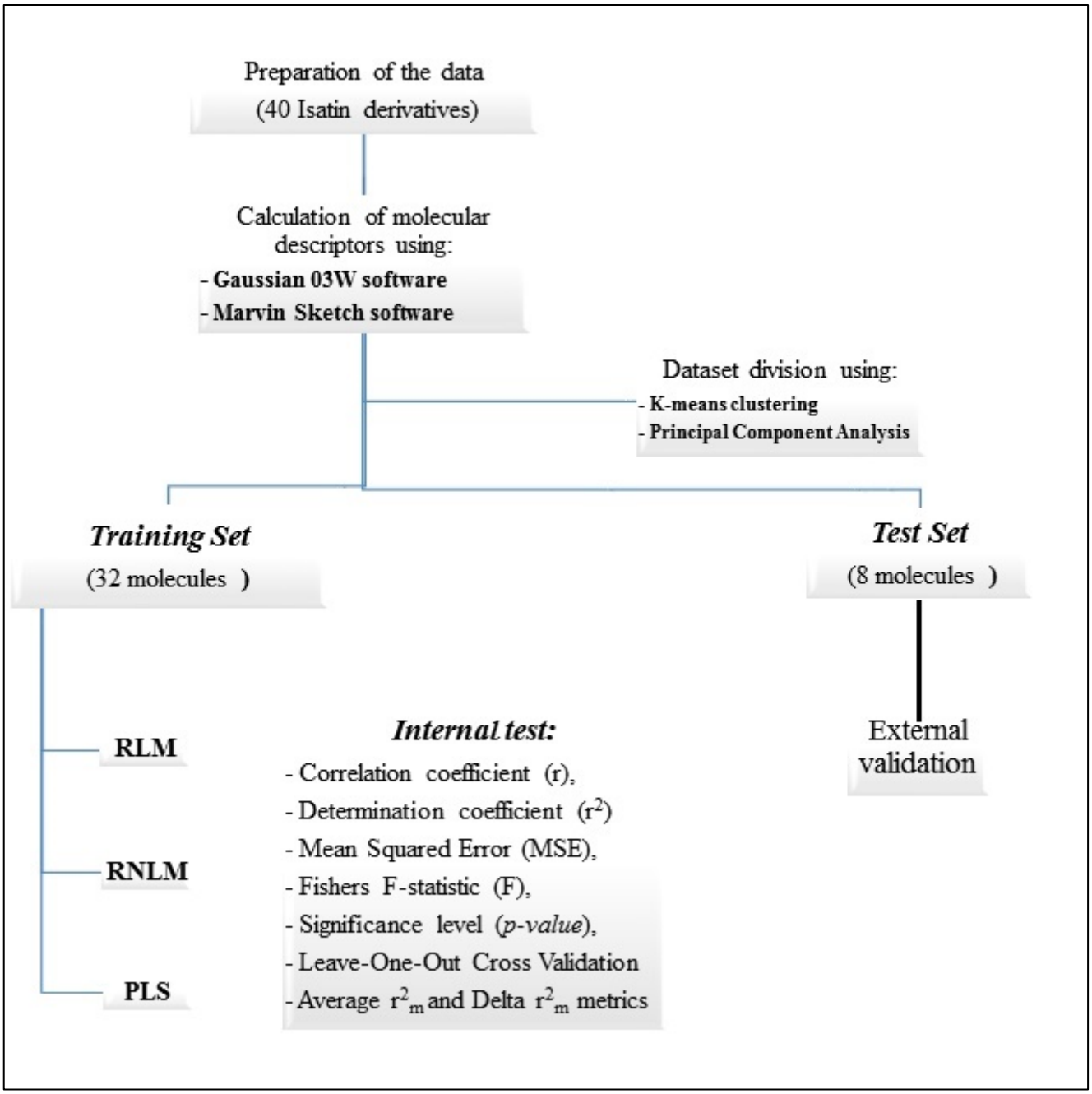


Figure 1: Flowchart of the methodology used in this work

3. Results

3.1. Dataset for analysis

The QSAR analysis was performed using the experimental anticancer activity values of the 40 selected molecules and the values of the 14 descriptors as shown in table 2.

The workflow involved a principal component analysis (PCA), which allowed us to eliminate descriptors that are highly correlated, followed by multiple linear regression based on the elimination of descriptors until a valid model was obtained (including the critical probability: p-value<0.05 for all descriptors and the model complete) and a stepwise multiple

linear regression based on the addition of the descriptors, saving those with largest contribution (*t* is less than the "Probability for entry", *t*: *student's t statistic*) and removing those with smallest contribution to the model. The selected descriptors by the MLR model were used as the input parameters in the multiple nonlinear regression methods (MNLR).

3.2. Principal Component Analysis (PCA)

All the 14 descriptors (variables) encoding the 40 molecules were submitted to a principal component analysis, and 12 components were obtained (Figure 2).

The first three axes, F1, F2, and F3, contributed 60.39%, 24.32% and 6.79%, respectively, to the total variance, and the total information was estimated to be 91.51%. The Pearson correlation coefficients are summarized in table 3; the matrix provides information on the negative and positive correlations between variables. Correlations among the 14 descriptors are shown in table 3 as a correlation matrix; in figure 3, these descriptors are represented in a correlation circle.

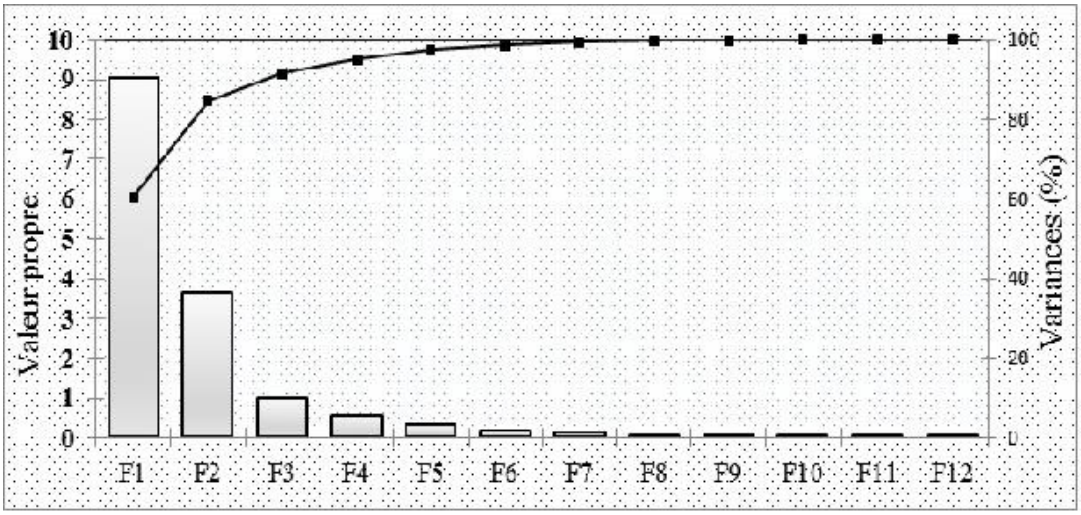


Figure 2: Principal components and their variances

Table 3: Correlation matrix (Pearson (n)) between different obtained descriptors

	WI	α	VDWV	MR	E_T	μ	E_{HOMO}	E_{LUMO}	Gap	η	χ	ω	Log P	BI
WI	1													
α	0.949	1												
VDWV	**0.965**	**0.983**	1											
MR	**0.957**	**0.995**	**0.994**	1										
E_T	-0.633	-0.798	-0.753	-0.794	1									
μ	-0.049	-0.207	-0.163	-0.207	0.448	1								
E_{HOMO}	0.426	0.439	0.437	0.420	-0.151	0.203	1							
E_{LUMO}	-0.324	-0.385	-0.367	-0.403	0.529	0.617	0.476	1						
Gap	-0.702	-0.758	-0.744	-0.751	0.535	0.198	-0.794	0.156	1					
η	-0.702	-0.758	-0.744	-0.751	0.535	0.198	-0.794	0.156	**1.000**	1				
χ	-0.162	-0.144	-0.151	-0.123	-0.124	-0.416	-0.923	-0.779	0.498	0.498	1			
ω	0.481	0.548	0.525	0.562	-0.609	-0.632	-0.220	-0.961	-0.417	-0.417	0.578	1		
Log P	0.810	0.891	0.864	0.894	-0.823	-0.478	0.289	-0.486	-0.660	-0.660	0.007	0.614	1	
BI	-0.849	-0.841	-0.810	-0.824	0.499	0.051	-0.524	0.163	0.701	0.701	0.302	-0.334	-0.759	1

In the correlation matrix, we have:

Gap and η are perfectly correlated ($r = 1$), the hardness (η) is removed.

E_{HOMO} and ω are perfectly correlated ($r \approx 1$), the electrophilicity index (ω) is removed.

(VDWV; α), (VDWV; MR), (VDWV; WI), (WI; MR), (MR; WI), (MR; α) are perfectly correlated ($r \approx 1$). Consequently, polarizability (α), Van Der Waals volume (VDWV) and molar refractivity (MR) are removed.

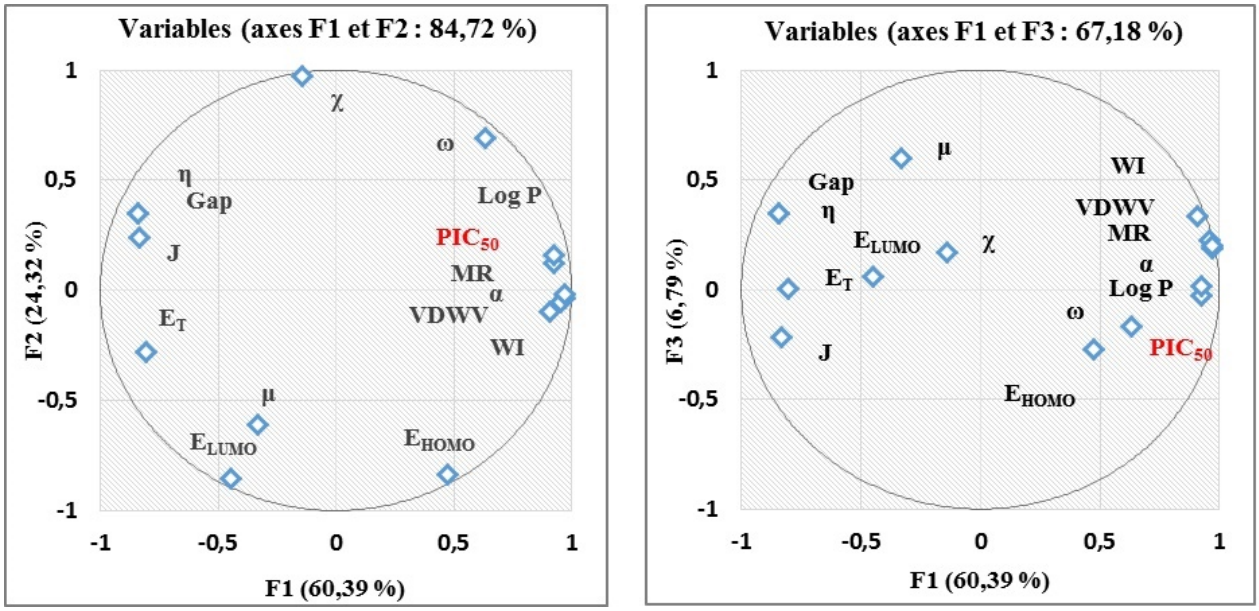


Figure 3: Representation of the descriptors in correlation circles.

The k-means method is used, in this study, to divide the observations into homogeneous clusters, based on their description by the used descriptors. For the division of dataset into subsets of a training set (80% of the compounds) and a test set (20% of the compounds), we used a combination of the k-means clustering results (Table 4). The test set consists of eight molecules (a molecule of each cluster is chosen: A1, A4, A8, A11, A13, A17, A18 and A27), the remaining molecules (32 molecules) are used for the training set.

Table 4: the k-means clustering results

1	A1, A2, A3, A24, A25
2	A4, A5, A6, A7, A15, A29, A32, A33, A34, A35, A36
3	A8, A9, A10, A12, A14, A16, A39, A40
4	A11, A30, A31, A37, A38
5	A13
6	A17, A21, A22, A23, A28
7	A18, A19, A20, A26
8	A27

The projections of the compounds (training and test sets) in the first three axes, F1, F2, and F3 are illustrated in the figure 4.

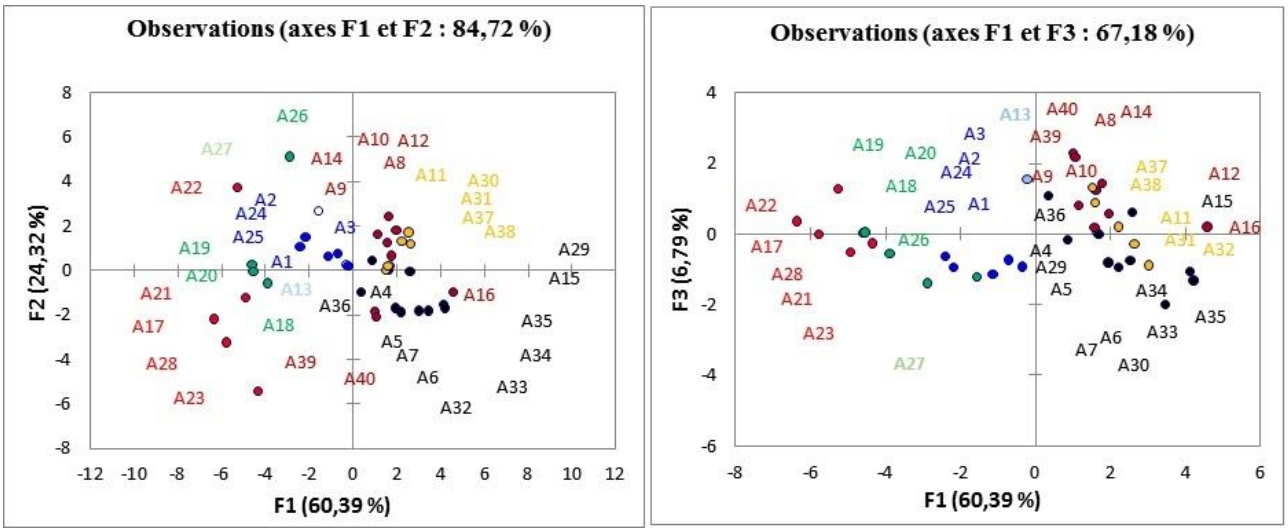


Figure 4: Projection of the compounds in the first three axes, F1, F2 and F3.

3.3. Multiple Linear Regression (MLR)

In order to propose a mathematical model and to evaluate quantitatively the substituent's physicochemical effects on the pIC_{50} of the set of these 40 molecules, the authors submitted the data matrix constituted from the 9 remainder variables corresponding to the 32 molecules (training set), to a stepwise and to a descendant multiple regression analysis.

The study of MLR (descendant) based on the elimination of descriptors until a valid model was obtained and the stepwise multiple linear regression procedures based on the forward selection and backward elimination methods were employed to determine the best regression models. These methods used the coefficients *r*, r^2, r_{adj}^2, *MSE*, *p-value*, *Average* r_m^2 ($\overline{r_m^2}$) and *Delta* r_m^2 (Δr_m^2) to select the best regression performance.

The QSAR models built using descendant and stepwise multiple linear regression methods are represented by the following equations:

Descendant MLR: $pIC_{50} = 4.692 + 0.744\ E_{HOMO} - 1.164\ E_{LUMO} + 0.503\ Log\ P$ (eq. 1)

Statistical parameters:

$r = 0.910$; $r^2 = 0.828$; $r_{adj}^2 = 0.810$; *MSE* = 0.155; *p-value* < 0.0001; $\overline{r_m^2} = 0.676$; $\Delta r_m^2 = 0.174$

Stepwise MLR: $pIC_{50} = 2.944 + 0.638\ Log\ P$ (eq.2)

Statistical parameters:

$r = 0.895$; $r^2 = 0.801$; $r_{adj}^2 = 0.794$; *MSE* = 0.168; *p-value* < 0.0001; $\overline{r_m^2} = 0.720$; $\Delta r_m^2 = 0.162$

The VIF was defined as $\frac{1}{1-r^2}$; where *r* was the correlation coefficient for one independent variable against all the other descriptors in the model. Variables with a VIF greater than 5 are unstable and should be eliminated, models with a VIF values between 1 and 4 means the models can be accepted.

Table 5: The variance inflation factors (VIF) of descriptors in QSAR model.

Statistic	E_{HOMO}	E_{LUMO}	*Log P*
Tolerance	0.367	0.344	0.419
VIF	2.725	2.906	2.384

As can be seen from table 5, the VIF values of all three descriptors are smaller than 5.0, indicating that there is no collinearity among the selected descriptors and the resulting model has good stability.

A high value of r^2, r^2_{adj} and lower mean squared error *MSE* indicate that the two proposed models are reliable. Given the fact that the probability corresponding to the *p-value* is much smaller than 0.0001; we would be taking a lower than 0.01 % risk in assuming that the null hypothesis is wrong. The value of $\overline{r_m^2}$ is above the threshold of 0.5 and Δr_m^2 is less than 0.2. Therefore, the authors conclude with confidence that the proposed models are predictive and reliable.

With the MLR models, the values of predicted pIC_{50} calculated from eq.1 and eq. 2 and the observed values are given in table 7. The correlations of predicted and observed pIC_{50} values are illustrated in figure 5.

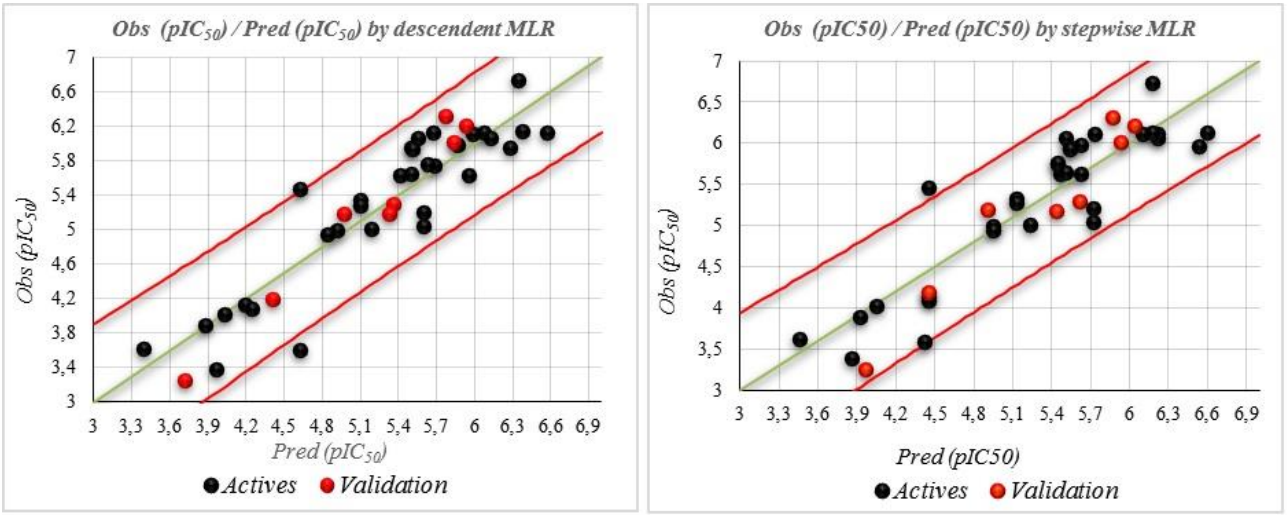


Figure 5: Correlations of observed and predicted pIC_{50} by descendant and stepwise MLR.

The authors use leave-one-out cross validation (LOOCV) as an internal test of the quality of the two MLR models. The values of predicted activities calculated using LOOCV and the observed values are given in table 6. The model's performance was good and was characterized by r^2_{cv}, *Average* $r^2_{m(loocv)}$ ($\overline{r^2_{m(loocv)}}$) and *Delta* $r^2_{m(loocv)}$ ($\Delta r^2_{m(loocv)}$):

For the Descendant MLR: $r^2_{cv} = 0.701$; $\overline{r^2_{m(loocv)}} = 0.513$; $\Delta r^2_{m(loocv)} = 0.182$

For the stepwise MLR: $r^2_{cv} = 0.772$; $\overline{r^2_{m(loocv)}} = 0.683$; $\Delta r^2_{m(loocv)} = 0.154$

The values of r_{cv}^2 and $\overline{r_{m(loocv)}^2}$ are above the threshold of 0.5 and $\Delta r_{m(loocv)}^2$ are less than 0.2 for the stepwise MLR model and just above the acceptance threshold values for the descendant MLR. These results obtained with the cross-validation show that the models proposed by the descendant and the stepwise MLR are able to predict activity with a great performance and that the selected descriptors are pertinent.

All calculations were repeated with randomized activities of the training set compounds as well to evaluate model robustness (y-randomization test). In the present case, 10 random trials were run for the MLR models. None of the random trials could match the original model (Table 6).

Table 6: Randomization test results for MLR models.

	For the Descendant RLM			*For the stepwise RLM*		
Model	r	r^2	r_{cv}^2	r	r^2	r_{cv}^2
Original	0.910	0.828	0.701	0.895	0.801	0.772
Random 1	0.354	0.126	-0.075	0.005	0.000	-0.147
Random 2	0.303	0.092	-0.134	0.246	0.060	-0.105
Random 3	0.156	0.024	-0.316	0.196	0.038	-0.098
Random 4	0.299	0.089	-0.122	0.022	0.000	-0.109
Random 5	0.235	0.055	-0.265	0.042	0.002	-0.124
Random 6	0.212	0.045	-0.231	0.407	0.165	0.073
Random 7	0.278	0.077	-0.167	0.014	0.000	-0.135
Random 8	0.492	0.242	0.056	0.069	0.005	-0.113
Random 9	0.302	0.091	-0.355	0.272	0.074	-0.046
Random 10	0.118	0.014	-0.209	0.134	0.018	-0.116
Average	0.275	0.086	-0.182	0.141	0.036	-0.092

3.4. Applicability Domain (AD)

The applicability domain (AD) of these models was evaluated by the leverage analysis expressed as Williams plot (Figure 6), in which the standardized residuals (*r*) and the leverage threshold values (h*=0.313 and 0.156 for descendant and stepwise MLR models respectively) were plotted. Any new value of predicted pIC_{50} data must be considered reliable only for those compounds that fall within this AD on which the model was constructed.

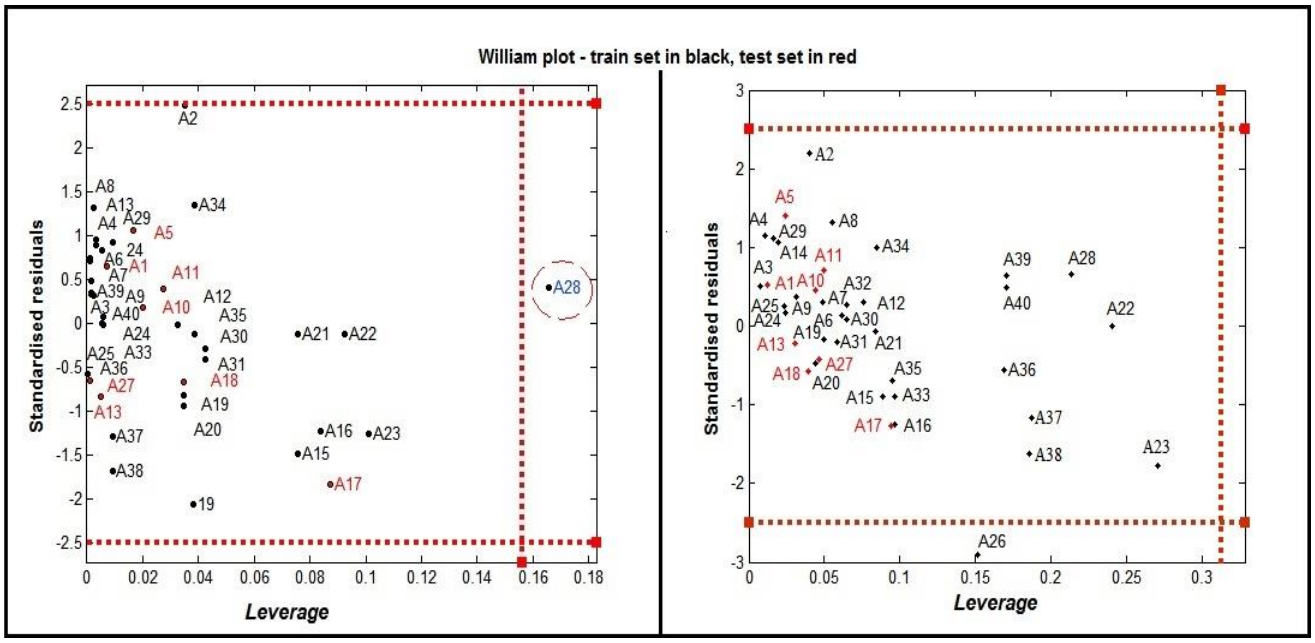


Figure 6: Williams plot of standardized residual versus leverage for and for stepwise MLR model (left figure: h*=0.156; residual limits = ±2.5) and descendant MLR model (right figure: h*=0.313; residual limits = ±2.5).

3.5. Partial Least Squares (PLS)

In PLS analysis, the descriptor data matrix is decomposed to orthogonal matrices with an inner relationship between the dependent and independent variables. Therefore, unlike MLR analysis, the multicollinearity problem in the descriptors is omitted by PLS analysis. As a minimal number of latent variables are used for modeling in PLS, this modeling method coincides with noisy data better than MLR **[52]**.

The QSAR model built using Partial Least Squares method is represented by the following equation:

$$pIC_{50} = 8.104 + 3.567\ 10^{-04}\ WI + 0.146\ Log\ P - 0.741\ J - 0.103\ \chi - 2.678\ 10^{-06}\ E_T - 6.641\ 10^{-02}\ \mu + 0.239\ E_{HOMO} - 0.429\ E_{LUMO} - 0.519\ Gap \quad \text{(eq. 3)}$$

Statistical parameters: $r = 0.903$; $r^2 = 0.815$; $MSE = 0.146$; $\overline{r_m^2} = 0.738$; $\Delta r_m^2 = 0.150$

The values of predicted pIC_{50} calculated from eq. 3, and the observed values are given in table 7. The correlations of predicted and observed pIC_{50} values are illustrated in figure 7. The values of predicted activities calculated using LOOCV and the observed values are given in table 7.

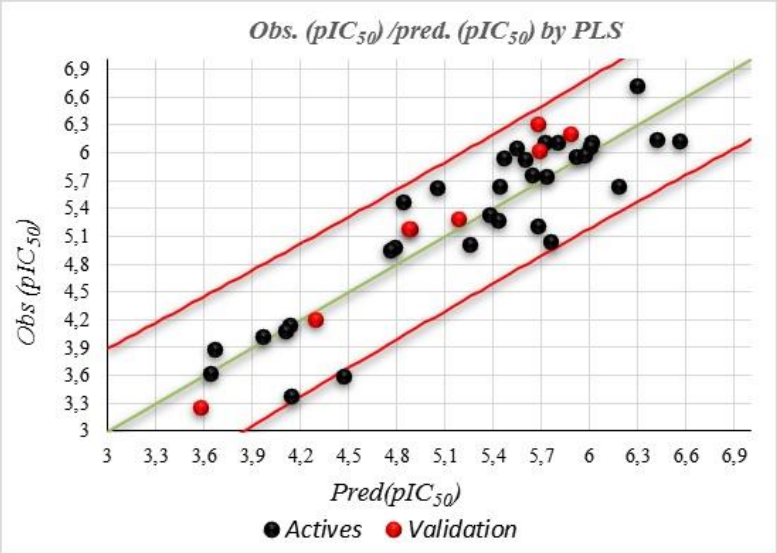


Figure 7: Correlations of observed and predicted pIC_{50} by PLS.

The authors use leave-one-out cross validation as an internal test of the quality of the PLS model. The model's performance was good and was characterized by r_{cv}^2, *Average* $r_{m(loocv)}^2$ ($\overline{r_{m(loocv)}^2}$) and *Delta* $r_{m(loocv)}^2$ ($\Delta r_{m(loocv)}^2$)):

$$r_{cv}^2 = 0.728;\ \overline{r_{m(loocv)}^2} = 0.623;\ \Delta r_{m(loocv)}^2 = 0.210$$

The values of r_{cv}^2 and $\overline{r_{m(loocv)}^2}$ are above the threshold of 0.5, but the value of $\Delta r_{m(loocv)}^2$ is higher than 0.2 for this model. These bad results obtained with the cross-validation show that the model proposed by the PLS is not able to predict anticancer activity satisfactorily.

3.6. Multiples Nonlinear Regression (MNLR)

The authors have used also the technique of nonlinear regression model to improve the structure-activity relationship to quantitatively evaluate the effect of the substituent and they have applied to the data matrix constituted obviously from the descriptors proposed by MLR corresponding to the 32 molecules (training set). The coefficients *r*, r^2, *MSE*, *Average* r_m^2 ($\overline{r_m^2}$) and *Delta* r_m^2 (Δr_m^2)) are used to select the best regression performance. The authors used a pre-programmed function of XLSTAT following:

$$Y = a + (b\ X_1 + c\ X_2 + d\ X_3 + e\ X_4 \ldots) + (f\ X_1^2 + g\ X_2^2 + h\ X_3^2 + i\ X_4^2 \ldots)$$

Where a, b, c, d... represent the parameters and X_1, X_2, X_3, X_4... represent the variables.

The proposed descriptors in eq. 1 and eq. 2 by MLR were, therefore, used as the input parameters in the MNLR method. The QSAR models built using multiple nonlinear regression methods are represented by the following equations:

With the descriptors proposed by the descendant MLR:

pIC_{50} = -92.182 – 23.708 E_{LUMO} – 18.954 E_{HOMO} – 0.172 Log P – 3.392 E_{LUMO}^2 – 1.571 E_{HOMO}^2 + 5.055 (Log P) ^{2} (eq. 4)

Statistical parameters: r = 0.945; r^2 = 0.892; *MSE* = 0.109 ; $\overline{r_m^2}$ = 0.845; Δr_m^2 = 0.093

With the descriptors proposed by the stepwise MLR:

pIC_{50} = 2.288 + 1.934 Log P – 6.749 10^{-02} (Log P) ^{2} (eq. 5)

Statistical parameters: r = 0.904; r^2 = 0.818; *MSE* = 0.158; $\overline{r_m^2}$ = 0.743; Δr_m^2 = 0.149

The values of predicted pIC_{50} calculated from equations 4 and 5 and the observed values are given in table 7. The correlations of predicted and observed pIC_{50} values are illustrated in figure 8.

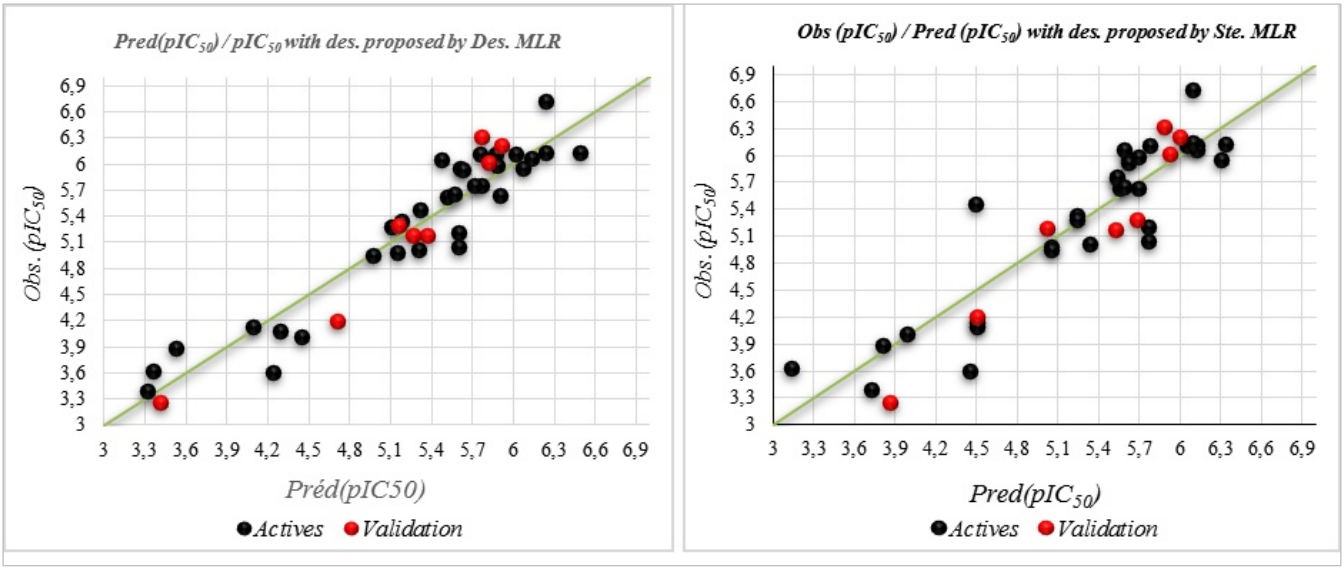


Figure 8: Correlations of observed and predicted pIC_{50} by MNLR (with descriptors proposed by descendant and stepwise MLR).

The authors use leave-one-out cross validation as an internal test of the quality of the RNLM models. The model's performance was good and was characterized by r_{cv}^2, *Average* $r_{m(loocv)}^2$ ($\overline{r_{m(loocv)}^2}$) and *Delta* $r_{m(loocv)}^2$:

For the model with descriptors proposed by the descendant MLR (Log P, E_{LUMO} and E_{HOMO}):

$$r_{cv}^2 = 0.677;\ \overline{r_{m(loocv)}^2} = 0.590;\ \Delta r_{m(loocv)}^2 = 0.037$$

For the model with descriptor proposed by the stepwise MLR (Log P):

$$r_{cv}^2 = 0.772;\ \overline{r_{m(loocv)}^2} = 0.686;\ \Delta r_{m(loocv)}^2 = 0.076$$

The values of r_{cv}^2 and $\overline{r_{m(loocv)}^2}$ are above the threshold of 0.5 and $\Delta r_{m(loocv)}^2$ are less than 0.2 for the two nonlinear models. These good results obtained with the cross-validation show that the models proposed by the two MNLR models are able to predict activity with a great

performance. The values of predicted activities calculated using LOOCV and the observed values are given in table 7.

Table 7: Observed and predicted anticancer activity according to RLM, PLS, and RNLM for the training set

	pIC_{50}										
		RLM-des		**RLM-step**		**PLS**		**RNLM-des**		**RNLM-step**	
N°	Obs	Pred.	CV	Pred.	CV	Pred.	CV	Pred.	CV	Pred.	CV
A2	5.460	4.625	4.383	4.455	4.383	4.843	4.802	5.322	5.192	4.498	4.425
A3	5.620	5.423	5.415	5.475	5.470	5.055	5.010	5.517	5.510	5.562	5.559
A4	5.940	5.500	5.483	5.548	5.534	5.467	5.449	5.612	5.592	5.625	5.609
A6	5.740	5.692	5.685	5.447	5.437	5.735	5.737	5.773	5.779	5.538	5.527
A7	5.750	5.637	5.625	5.447	5.437	5.649	5.643	5.717	5.713	5.538	5.526
A8	6.050	5.557	5.491	5.509	5.491	5.545	5.492	5.476	5.297	5.592	5.569
A9	5.640	5.501	5.492	5.509	5.505	5.440	5.427	5.576	5.569	5.592	5.590
A12	6.100	5.988	5.975	6.108	6.108	5.803	5.772	5.877	5.843	6.050	6.046
A14	5.920	5.515	5.496	5.550	5.536	5.602	5.583	5.632	5.612	5.627	5.612
A15	5.950	6.279	6.604	6.534	6.604	5.918	5.923	6.076	6.128	6.303	6.408
A16	6.120	6.579	6.644	6.599	6.661	6.565	6.635	6.492	6.645	6.336	6.418
A19	4.130	4.195	4.480	4.457	4.480	4.141	4.147	4.090	4.077	4.501	4.529
A20	4.080	4.256	4.484	4.457	4.484	4.113	4.118	4.289	4.330	4.501	4.533
A21	4.010	4.032	4.035	4.058	4.063	3.970	3.964	4.447	4.617	3.990	3.988
A22	3.880	3.880	3.880	3.928	3.935	3.668	3.757	3.528	1.872	3.813	3.801
A23	3.380	3.968	4.320	3.866	3.940	4.144	4.543	3.322	3.229	3.727	3.801
A24	4.980	4.918	4.915	4.948	4.947	4.794	4.781	5.147	5.164	5.056	5.061
A25	4.940	4.843	4.837	4.948	4.949	4.759	4.746	4.975	4.979	5.056	5.064
A26	3.590	4.627	5.379	4.419	4.481	4.469	4.893	4.244	5.318	4.454	4.521
A28	3.620	3.393	3.298	3.465	3.427	3.645	3.746	3.358	2.990	3.136	2.746
A29	6.110	5.681	5.661	5.732	5.715	5.722	5.704	5.758	5.738	5.776	5.760
A30	6.110	6.078	6.074	6.222	6.231	6.019	6.008	6.020	6.007	6.124	6.125
A31	6.060	6.136	6.143	6.222	6.235	6.010	6.006	6.141	6.153	6.124	6.131
A32	5.970	5.869	5.618	5.631	5.618	5.974	5.973	5.890	5.876	5.695	5.682
A33	5.630	5.957	6.039	5.631	5.631	6.185	6.242	5.903	5.990	5.695	5.698
A34	6.720	6.352	6.138	6.179	6.138	6.295	6.245	6.243	6.152	6.096	6.038
A35	6.130	6.384	6.426	6.179	6.183	6.420	6.449	6.237	6.262	6.096	6.093
A36	5.000	5.194	5.202	5.233	5.240	5.259	5.268	5.304	5.330	5.341	5.362
A37	5.200	5.605	5.746	5.724	5.746	5.685	5.718	5.604	5.636	5.770	5.797
A38	5.040	5.604	5.753	5.724	5.753	5.760	5.807	5.598	5.644	5.770	5.804
A39	5.330	5.102	5.092	5.133	5.126	5.381	5.358	5.185	5.171	5.244	5.238
A40	5.270	5.099	5.090	5.133	5.128	5.438	5.430	5.113	5.093	5.244	5.242

True predictive power of a QSAR model is to test their ability to predict accurately the pIC_{50} of compounds from an external test set (compounds which were not used for the developed model), the pIC_{50} of the remaining set of 8 compounds are deduced from the quantitative models proposed with the 32 molecules (training set) by MLR and MNLR. The observed and calculated pIC_{50} values are given in tables 8.

Table 8: Observed and predicted anticancer activity according to MLR (descendant and stepwise), PLS, and MNLR for the eight tested compounds (test set).

N°	Obs.	RLM		PLS	RNLM	
	$pIC_{50\ obs.}$	$pIC_{50\ des}$	$pIC_{50\ step}$	$pIC_{50\ PLS}$	$pIC_{50\ des}$	$pIC_{50\ step}$
A1	5.180	4.977	4.913	4.890	5.373	5.019
A5	6.310	5.776	5.875	5.681	5.767	5.886
A10	6.010	5.836	5.933	5.686	5.827	5.929
A11	6.200	5.934	6.038	5.884	5.917	6.002
A13	5.280	5.362	5.617	5.186	5.164	5.684
A17	3.250	3.716	3.967	3.581	3.412	3.867
A18	4.190	4.409	4.457	4.294	4.715	4.501
A27	5.170	5.332	5.439	4.882	5.269	5.531

The comparison of the values of pIC_{50}-test to pIC_{50}-obs shows that a good prediction has been obtained for the 8 compounds (r_{test} and r^2_{test} showed in table 9).

Table 9: Statistical coefficients for the MLR, PLS, and RNLM studies

	MLR descendant	**MLR** stepwise	**PLS**	**MNLR** descendant	**MNLR** stepwise
r	0.910	0.895	0.903	0.944	0.904
r^2	0.828	0.801	0.815	0.892	0.818
$\overline{r_m^2}$	0.676	0.720	0.738	0.845	0.743
Δr_m^2	0.174	0.162	0.150	0.093	0.149
r^2_{cv}	0.652	0.772	0.728	0.677	0.772
$r^2_{m(loocv)}$	0.513	0.683	0.623	0.590	0.686
$\Delta r^2_{m(loocv)}$	0.182	0.154	0.210	0.037	0.076
r_{test}	0.992	0.963	0.990	0.967	0.930
r^2_{test}	0.984	0.927	0.981	0.936	0.964

4. Discussions

In this work, the PCA technique is used to get an overview the examined compounds for similarities and dissimilarities, to eliminate independent descriptors that are highly correlated by examining multicollinearity between descriptors (consequently, polarizability (α), Van Der Waals volume (VDWV) and molar refractivity (MR) are removed) and to select descriptors that show a high correlation with the response activity; that give extra weight because they will be more effective for predictions. The presence of the multicollinearity between descriptors was confirmed from the correlation matrix and VIF values.

In the PLS model, the number of observations is larger than the number of descriptors by five times. The number of observations used in order to obtain eq. 3 is thirty-two, whereas the number of the descriptor is nine. Therefore, the model can have a chance correlation **[53]**. On the other hand, the value of $\Delta r^2_{m(loocv)}$ is higher than 0.2 for this model. These bad results

obtained with the cross-validation confirm that the model proposed by the PLS is not able to predict anticancer activity satisfactorily.

For the two MLR models, the p-value is lower than 0.0001, it means that we would be taking lower than 0.01% risk in assuming that the null hypothesis (no effect of the explanatory variables) is wrong and that the regressions equations have statistical significance. Therefore, the authors conclude with confidence that the selected variables do carry a significant amount of information.

The higher value of r^2 (r^2=0.828 for the descendant MLR model, r^2=0.801 for the stepwise MLR model, r^2=0.892 for MNLR (with descriptors proposed by descendant MLR) and r^2=0.818 for MNLR (with descriptors proposed by stepwise MLR)) and the higher value of r^2_{adj} and the lower mean squared error *MSE* indicate that the two proposed models are predictive and reliable.

The obtained models were validated internally by the leave one cross-validation technique, the cross-validation coefficient r^2_{cv} for models was determined based on the predictive ability of the model. The value of r^2cv is higher than 0.5 (r^2_{cv}=0.652 for the descendant MLR model, r^2_{cv}=0.772 for the stepwise MLR model, r^2_{cv}=0.677 for MNLR (with descriptors proposed by descendant MLR), and r^2_{cv}=0.772 for MNLR (with descriptors proposed by stepwise MLR)), indicates the better predictivity of models.

The true predictive power of these models can be tested from their ability to predict perfectly the pIC_{50} of compounds from an external test set (compounds that were not used for the developed model), the pIC_{50} of the remaining set of 8 compounds are deduced from the quantitative models proposed with the compounds used in training set by MLR. These models were able to predict the activities of test set molecules in agreement with the experimentally determined value. The observed and calculated pIC_{50} values are given in table 8. The predictive capacity of the models was judged: the higher values of r^2_{test} (r^2_{test}=0.984 for the descendant MLR model, r^2_{test}=0.927 for the stepwise MLR model, r^2_{test}=0.936 for MNLR (with descriptors proposed by descendant MLR), and r^2_{test}=0.964 for MNLR (with descriptors proposed by descendant MLR)) indicate the improved predictivity of the models.

In this work, MLR and MNLR are complementary methods; MNLR has added corrections of a second degree (polynomial form) for the linear model.

A comparison of the quality of MLR and MNLR models shows that the four approaches have the good predictive capability; which is sufficient to conclude the performance of these models and to establish a satisfactory relationship between selected descriptors and anticancer activity. Furthermore, the results obtained by MNLR are relatively better than those obtained

by MLR, but the latter approach is more transparent and gives the most interpretable results and a good explanation of the descriptors associated with activities.

Consequently, with MLR approaches, the authors can design new compounds with improved values of activity than the studied compounds. Taking into account the above results, the authors added suitable substitutions and then calculated the activities of the new compounds using the proposed models (eq. 1 and eq. 2).

In the first model (descendant MLR model), the descriptors influencing positively the activities are: the highest occupied molecular orbital energy E_{HOMO} and the Octanol/Water partition coefficient Log P, and the parameter influencing negatively the activities is the lowest unoccupied molecular orbital energy E_{LUMO}.

In the second model, stepwise MLR model, the descriptor influencing positively the activities is the Octanol/Water partition coefficient Log P. These results illustrate that to increase anticancer activities, we will decrease the lowest unoccupied molecular orbital energy E_{LUMO} (negative values), increase the highest occupied molecular orbital energy E_{HOMO} (negative value) and the Octanol/Water partition coefficient Log P.

HOMO and LUMO refer to highest occupied molecular orbital and lowest unoccupied molecular orbital. According to the frontier orbital theory **[54]**, the nucleophilic attack occurs by electron flow from the HOMO of the nucleophile into the LUMO of the electrophile. In stable molecules, occupied electrons always reside into orbitals with negative energies and unoccupied orbitals have positive energies. The energies of HOMO and LUMO are related to the reactivity of the molecule: molecules with electrons at accessible (near-zero) HOMO levels tend to be good nucleophiles because it does not cost much to donate these electrons toward making a new bond. Similarly, molecules with lower LUMO energies tend to be good electrophiles because it does not cost much to place an electron into such an orbital. The distributions of charge of HOMO and LUMO orbital are similar for all Isatin derivatives. The authors remark that the HOMOs of all compounds are almost delocalized over the whole molecule, especially for the donor units which are electrons-rich, while for the LUMO orbital there is a large contribution over groups which are electrons-deficient (Figure 9).

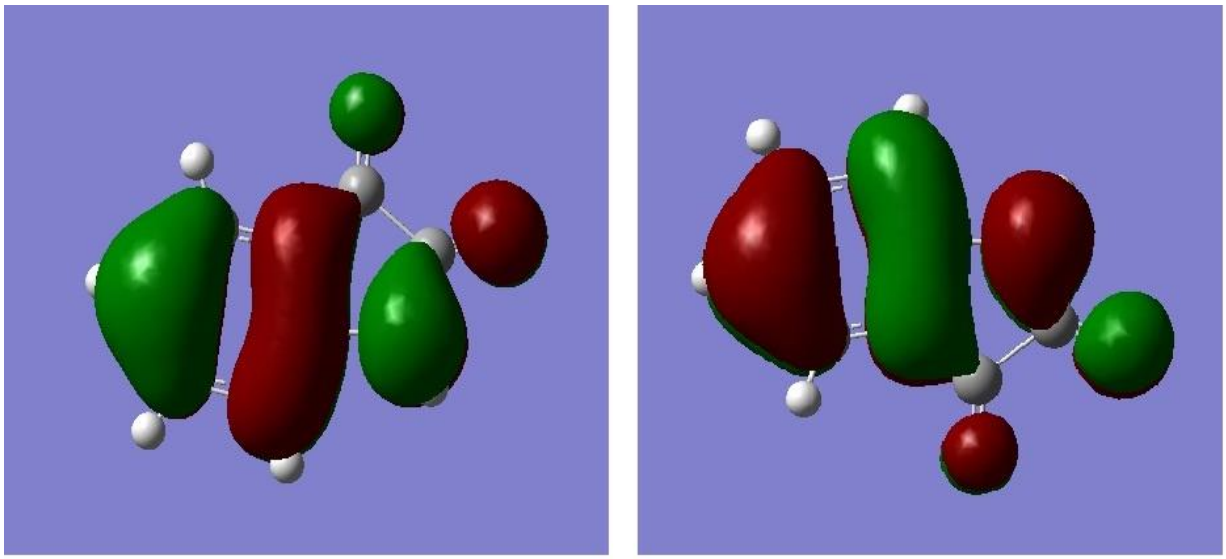


Figure 9: The contour plots of HOMO and LUMO orbitals of the Isatin.

According to the above discussions:

Highest occupied molecular orbital energy E_{HOMO} (negative values) has a positive sign in the model, which suggests that the substitution of the Isatin derivatives with a stronger donating electron ability group (such as phenyl) may lead to high activity values.

Lowest unoccupied molecular orbital energy E_{LUMO} (negative values) has a negative sign in the model, which suggests the substitution of the Isatin with a stronger accepting electron ability group (such as halogens) may lead to high activity values.

Octanol/Water partition coefficient Log P has a positive sign in the model, which suggests that a higher value of the Log P (stronger lipophilic / weaker hydrophilic ability) and a substitution of the Isatin with a nonpolar group may lead to high activity values.

The importance of each descriptor on the pIC_{50} values can be compared from the standardized coefficients and the t-test values in the model equation. The bigger the absolute value of the t-test value, the greater is the influence of the descriptor. In eq. 1, the t-test values are 2,094, -1,814 and 5,833 for E_{HOMO}, E_{LUMO} and Log P, respectively. This means that the t-test value of Log P is larger than that of other two descriptors, which indicate that in this model, the influence of Log P on activity is stronger than that of the other two.

Analysis MLR models shows that descendant MLR, with a strong influence of the Log P on the activity, gives slightly better results than the stepwise one, but MLR with just log P (stepwise model) seems to be pretty simple.

5. Detection of the outliers

From figure 6, it is obvious that there is one response outlier in the training set and no response outlier is present in the test set. Only one chemical is identified as an outlier for

stepwise MLR model; this outlier is the compound N° 28 in the training set and it has a leverage which is greater than h* value of 0.156 and only one compound has a residual out of the ±x times standard deviation interval (x=2.5). These erroneous predictions could probably be attributed to wrong experimental data or to the structural peculiarity of this outlier; a methyl substitution at the ring nitrogen ring; the selected descriptors might noy pay much attention to this special substructure. The prediction for this compound is an extrapolation of the model, but fortunately, it is a ‘‘good leverage’’ chemical.

6. Proposed novel compounds with higher anticancer activity values

QSAR correlates activity data with the physicochemical and/or structural properties of a group of compounds; it has been frequently used to predict activities/proprieties of new compounds and to design compounds with desired activities/properties. The developed equations can be used for the designing of new Isatin derivatives with improved anticancer activity (pIC_{50}). If we develop a new compound with higher values of activities than the existing compounds, it may give rise to the development of more active compounds than those currently in use.

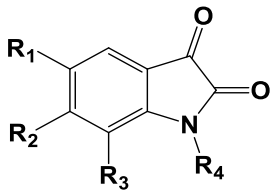
In this way, for the descendant MLR, eq. 2 indicates a positive correlation of Octanol/Water partition coefficient Log P with the activity. If we want to increase the value of the anticancer activity, it is necessary to increase the value of Log P, for which we must substitute the Isatin for lipophilic substituents. This study is in agreement with previous QSAR study **[27]** which revealed the importance of lipophilic factors as R_3 and R_5 substituents on the anticancer activity.

The present authors carried out structural modification starting from compounds having the highest pIC_{50} values as a template (such as N° 34 with pIC_{50} = 6.72). The structures of the designed compounds and their parameter values calculated by the same methods, as well as the pIC_{50} values theoretically predicted by the MLR models are listed in table 10. The authors have designed seven new compounds by adding suitable substituents and calculated their predicted anticancer activity using equations 1 and 2. From the predicted activity, it has been observed that the designed compounds have higher pIC_{50} values than all the 40 studied compounds (Table 1).

The leverage values for the new designed compounds (calculated using ($\mathbf{h_i} = \boldsymbol{x_i^T (X^T X)^{-1} x_i}$) **[48]**) are: 0.295, 0.202, 1.322, 0.278, 0.280, 0.488 and 0.284 for X1, X2… and X7, respectively, for the descendant model and are: 0.093, 0.083, 0.135, 0.083, 0.104, 0.097 and 0.105 for X1, X2… and X7, respectively, for the stepwise model. Compounds X3 and X6 are defined as outliers and consequently not been considered because they have a

higher leverage which is greater than h* (0.313 for descendant model and 0.156 for the stepwise model). The authors suggested other five compounds as candidates that can be synthesized and evaluated as anticancer drugs.

Table 10: Proposed compounds, value of calculated descriptors, and predicted anticancer activities using RLM models.



								pIC_{50} RLM	
	R_1	**R_2**	**R_3**	**R4**	**E_{HOMO}**	**E_{LUMO}**	**Log P**	**Desc.**	**Step.**
X1	Br	Ph	Br	1-naphthylmethyl	-5.970	-2.984	6.714	7.101	7.228
X2	Br	$(CH)_3CH_2$	Br	1-naphthylmethyl	-5.971	-3.025	6.331	6.955	6.983
X3	Ph	$(CH)_3CH_2$	Ph	1-naphthylmethyl	-5.787	-2.625	8.088	7.510	8.104
X4	CH_2Br	H	CH_2Br	1-naphthylmethyl	-6.095	-2.966	6.338	6.798	6.988
X5	Br	Br	Br	1-naphthylmethyl	-6.041	-3.156	7.098	7.442	7.473
X6	CH_2Br	CH_3	CH_2Br	1-naphthylmethyl	-6.068	-2.891	6.851	6.988	7.315
X7	CH_2Br	CH_2Br	CH_2Br	1-naphthylmethyl	-6.068	-3.156	7.110	7.428	7.480

7. Conclusion

Multiple linear regression, partial least squares and multiple nonlinear regression (based on the descendant or stepwise methods) were used to construct quantitative structure-activity relation models of Isatin derivatives for their anticancer activity. These regression methods were compared and it was found that MNLR models had substantially better predictive capability than the MLR models. The results show that the models proposed in this paper can predict anticancer activity accurately and that the selected parameters (highest occupied molecular orbital energy E_{HOMO}, lowest unoccupied molecular orbital energy E_{LUMO}, and the Octanol/Water partition coefficient Log P) are pertinent. The accuracy and predictability of the proposed models were illustrated by a comparison of the key statistical terms r^2 and r^2_{CVLOO}. The predictive power of the equations was validated by an internal test (Cross-validation) and an external test set. The applicability domain of the MLR models was investigated using William's plot to detect the subspace of chemical structures that can be predicted reliably by models.

Among the proposed models, MNLR based on descriptors proposed by the MLR (stepwise and descendant) have better predictive capability than the other models, but the stepwise MLR gives the more important interpretable and simple results. The bad results

obtained with the cross-validation show that the model proposed by the PLS is not able to predict anticancer activity satisfactorily.

The authors have designed and suggested, based on the QSAR studies, a few new compounds with higher theoretical anticancer activity than the studied compounds. Consequently, the proposed models can further be used in anticancer drug research for the Isatin derivatives.

References

[1] J.F.M. Da Silva, S. J. Garden, and A. C. Pinto, "The Chemistry of Isatins: A review from 1975 to 1999", *Journal of the Brazilian Chemical Society,* 12, **2001**, 273–324.

[2] R.P. Sonawane, R.R. Tripathi, "The chemistry and synthesis of 1H-indole-2,3-dione (Isatin) and its derivatives", International Letters of Chemistry, *Physics and Astronomy*, 12, **2013**, 30–36.

[3] O.L. Erdmann, "Untersuchungenüber den Indigo", *Journal für praktische Chemie*, 19, **1840**, 321–362.

[4] A. Laurent, "Recherches sur l'indigo", *Annales de chimie et de physique*, 3, **1840**, 393–434.

[5] D. Lee, S.A. Long, J.H. Murray, J.L. Adams, M.E. Nuttall, D.P. Nadeau, K. Kikly, J.D. Winkler, C.M. Sung, M.D. Ryan, M.A. Levy, P.M. Keller, and W.E. Jr DeWolf, "Potent and Selective Nonpeptide Inhibitors of Caspases 3 and 7", *Journal of Medicinal Chemistry*, 44, **2001**, 2015–2026.

[6] J.G. Chapman, W.P. Magee, H.A. Stukenbrok, G.E. Beckius, A.J. Milici, and W.R. Tracey, "A novel nonpeptidic caspase-3/7 inhibitor, (S)-(+)-5-[1-(2-methoxymethylpyrrolidinyl) sulfonyl] Isatin reduces myocardial ischemic injury", *European Journal of Pharmacology*, 456, **2002**, 59–68.

[7] M. Verma, S.N. Pandeya, and K.N. Singh, J.P. Stables, "Anticonvulsant activity of Schiff bases of isatin derivatives", *Acta Pharmaceutica*, 54, **2004**, 49–56

[8] D. Sriram, R.B. Tanushree, and P. Yogeeswari, "Design, synthesis and biological evaluation of novel non-nucleoside HIV-1 reverse transcriptase inhibitors with broad-spectrum chemotherapeutic properties", *Bioorganic & Medicinal Chemistry*, 12, **2004**, 5865–5873.

[9] C.P. Michael, V.P. Sunil, D.S. Koushik, A.K. Kathy, and R.K. Earl, "Combinatorial Optimization of Isatin-β-Thiosemicarbazones as Anti-Poxvirus Agents", *Journal of Medicinal Chemistry*, 48, **2005**, 3045–3050.

[10] Z.H. Chohan, H. Pervez, A. Rauf, K.M. Khan, and C.T., "Supuran Isatin-derived antibacterial and antifungal compounds and their transition metal complexes", *Journal of Enzyme Inhibition and Medicinal Chemistry*, 19, **2004**, 417–423.

[11] N. Lashgari and G.M. Ziarani, "Synthesis of heterocyclic compounds based on Isatin through 1,3-dipolar cycloaddition reactions", *Reviews and Accounts ARKIVOC*, 1, **2012**, 277–310.

[12] A. Mishra and P. Bauerle, "Small Molecule Organic Semiconductors on the Move: Promises for Future Solar Energy Technology", *Angewandte Chemie International Edition*, 51, **2012**, 2020–2067.

[13] B. Walker, C. Kim, and T.Q. Nguyen, "Small Molecule Solution-Processed Bulk Heterojunction Solar Cells", *Chemistry of Materials*, 23, **2011**, 470–482.

[14] G. Zhang, Y. Fu, Z. Xie, and Q. Zhang, "Synthesis and Photovoltaic Properties of New Low Bandgap Isoindigo-Based Conjugated Polymers", *Macromolecules*, 44, **2011**, 1414–1420.

[15] X. Xu, L. Li, B. Liu, and Y. Zou, "Organic semiconductor memory devices based on a low-band gap polyfluorene derivative with isoindigo as electron-trapping moieties", *Applied Physics Letters*, 98, **2011**, 1–3.

[16] T. Lei, Y. Cao, Y. Fan, C.J. Liu, S.C. Yuan, and J.J. Pei, "High-Performance Air-Stable Organic Field-Effect Transistors: Isoindigo-Based Conjugated Polymers", *Journal of the American Chemical Society*, 133, **2011**, 6099–6101.

[17] R.S. Ashraf, A.J. Kronemeijer, D.I. James, H. Sirringhaus, I. McCulloch, "A new thiophene substituted isoindigo based copolymer for high performance ambipolar transistors", *Chemical Communications*, 48, **2012**, 3939–3941.

[18] R. Stanley, "Synthesis of indole and oxindolederivaties incorporating pyrrolinino, pyrrolo or imidazolo moieties", *KarolinskaInstitutet, Stockholm, Sweden*, **2004**.

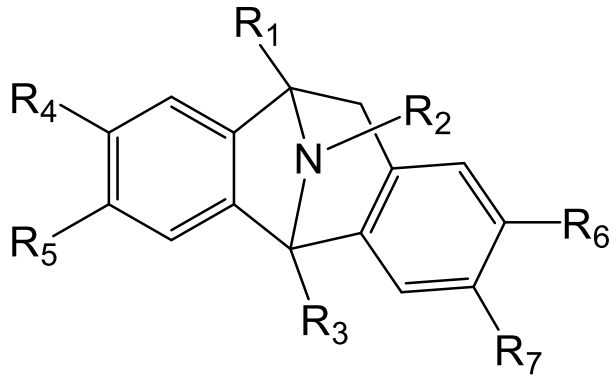
[19] N.H. Esheba and H.M. Salama, "5-(2-Oxo-3-indolinylidene) thiazolidine-2,4-dione-1,3-di-Mannich base derivatives: synthesis and evaluation for antileukemic activity", *Die Pharmazie*, 40, **1985**, 320–322.

[20] H. Pajouhesh, R. Parson, and F.D. Popp, "Potential anticonvulsants VI: condensation of isatin with cyclohexanone and other cyclic ketones", *Journal of Pharmaceutical Sciences*, 72, **1983**, 318–321.

[21] C.B. Westley, K.L. Vine, K. Benkendorff, L. Meijer, N. Guyard, and L.A. Skaltsounis, "Indirubin, the Red Shade of Indigo", *Life in Progress Editions, France*, **2006**, 31–44.
[22] K. Benkendorff, J. Bremner, and A. Davis, "Indole Derivatives from the Egg Masses of Muricid Molluscs", *Molecules*, 6(2), **2001**, 70–78.
[23] a) K.L. Vine, J.M. Locke, M. Ranson, K. Benkendorff, S.G. Pyne, and J.B. Bremner "Corrigendum to in vitro cytotoxicity evaluation of some substituted isatin derivatives", *Bioorganic and Medicinal Chemistry*, 15, **2007**, 931–938.
b) K.L. Vine, J.M. Locke, M. Ranson, S.G. Pyne, and J.B. Bremner, "An Investigation into the Cytotoxicity and Mode of Action of Some Novel N-Alkyl-Substituted Isatins", *Journal of Medicinal Chemistry*, 50, **2007**, 5109–5117.
[24] L. Matesic, J.M. Locke, J.B. Bremner, S.G. Pyne, D. Skropeta, M. Ranson, and K.L. Vine, "N-Phenethyl and N-naphthylmethyl Isatins and analogues as in vitro cytotoxic agents", *Bioorganic and Medicinal Chemistry*, 16, **2008**, 3118–3124.
[25] H. Prenen, J. Cools, N. Mentens, C. Folens, R. Sciot, P. Schoffski, A. Van Oosterom, P. Marynen, and M. Debiec-Rychter, "Efficacy of the kinase inhibitor SU11248 against gastrointestinal stromal tumor mutants refractory to imatinib mesylate", *Clinical Cancer Research*, 12(8), **2006**, 2622–2627.
[26] R.J. Motzer, M.D. Michaelson, B.G. Redman, G.R. Hudes, G. Wilding, R.A. Figlin, M.S. Ginsberg, S.T. Kim, C.M. Baum, S.E. DePrimo, J.Z. Li, C.L. Bello, C.P. Theuer, D.J. George, and B.I. Rini, "Activity of SU11248, a multitargeted inhibitor of vascular endothelial growth factor receptor and platelet-derived growth factor receptor, in patients with metastatic renal cell carcinoma", *Journal of Clinical Oncology*, 24(1), **2006**, 16–24.
[27] R. Sabet, M. Mohammad, A. Sadeghi, and A. Fassihi, "QSAR study of Isatin analogues as in vitro anti-cancer agents", *European Journal of Medicinal Chemistry*, 45, **2010**, 1113–1118.
[28] F.R. Burden, M.G. Ford, D.C. Whitley, and D.A. Winkler, "Use of automatic relevance determination in QSAR studies using Bayesian neural networks", *Journal of Chemical Information and Modeling*, 40(6), **2000**, 1423–1430.
[29] I.I.R. Denning, T. Keith, J. Millam, K. Eppinnett, W.L. Hovell, and R. Gilliland, GaussView Version 3.09, *Semichem Shawnee Mission, KS, USA*, **2003**.
[30] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, J.A. Montgomery, T. Vreven, K.N. Jr., Kudin, J.C. Burant, J.M. Millam, S.S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G.A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J.E. Knox, H.P. Hratchian, J.B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, P.Y. Ayala, K. Morokuma, G.A. Voth, P. Salvador, J.J. Dannenberg, V.G. Zakrzewski, S. Dapprich, A.D. Daniels, M.C. Strain, O. Farkas, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J.V. Ortiz, Q. Cui, A.G. Baboul, S. Clifford, J. Cioslowski, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, C. Gonzalez, and J.A. Pople, Gaussian 03, Revision C.02, *Gaussian, Inc., Wallingford CT*, **2004**.
[31] C. Lee, W. Yang, and R.G. Parr, "Development of the Colle-Salvetti correlation-energy for into functional of the electro density", *Physical Review B*, 37(2), **1988**, 785–789.
[32] P.C. Hariharan and J.A. Pople, "The influence of polarization functions on molecular orbital hydrogenation energies", *Theoretica Chimica Acta*, 28, **1973**, 213–222.
[33] S. Chtita, M. Larif, M. Ghamali, M. Bouachrine, and T. Lakhlifi, "DFT-based QSAR Studies of MK801 derivatives for noncompetitive antagonists of NMDA using electronic and topological descriptors", *Journal of Taibah University for Science*, 9(2), **2015**, 143–154.
[34] E. Eroglu and H. Türkmen, "A DFT-based quantum theoretic QSAR study of aromatic and heterocyclic sulfonamides as carbonic anhydrase inhibitors against isozyme CA-II", *Journal of Molecular Graphics and Modeling*, 26, **2007**, 701–708.

[35] C.G. Gu, X. Jiang, X.H. Ju, X.D. Gong, F. Wang, Y.R. Bian, C. Sun, "QSARs for congener-specific toxicity of poly-halogenated dibenzo-p-dioxins with DFT and WHIM theory", *Ecotoxicology and Environmental Safety*, 72, **2009**, 60–70.
[36] S. Arulmozhiraja, and M. Morita, "Structure–activity relationships for the toxicity of polychlorinated dibenzofurans: approach through density functional theory based descriptors", *Chemical Research in Toxicology*, 17, **2004**, 348–356.
[37] F.A. Pasha, H.K. Srivastava, and P.P. Singh, "Comparative QSAR study of phenol derivatives with the help of density functional theory", *Bioorganic and Medicinal Chemistry*, 13, **2005**, 6823–6829.
[38] U. Sakar, J. Padmanabhan, R. Parthasarathi, V. Subramanian, and P.K. Chattaraji, "Toxicity analysis of polychlorinated dibenzofurans through global and local electrophilicities", *Journal of Molecular Structure: THEOCHEM*, 758(2-3), **2006**, 119–125.
[39] Marvin 6.2.1 software, *Chem Axon*, **2014**. http://www.chemaxon.com
[40] XLSTAT software, *XLSTAT Company*, **2009**.
[41] S. Chtita, R. Hmamouchi, M. Larif, M. Ghamali, M. Bouachrine, and T. Lakhlifi, "QSPR studies of 9-aniliioacridine derivatives for their DNA drug binding properties based on density functional theory using statistical methods: Model, validation & influencing factors", *Journal of Taibah University for Science*, 10(6), **2016**, 868–876.
[42] B. Bhattacharya and D. Habtzghi, "Median of the pvalue under the alternative hypothesis", *American Statistical Association*, 56(3), **2002**, 202–206.
[43] S. Zhang, A. Golbraikh, S. Oloff, H. Kohn, and A. Tropsha, "A Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models", *Journal of Chemical Information and Modeling*, 46 (5), **2006**, 1984–1995.
[44] K. Roy, I. Mitra, S. Kar, P.K. Ojha, R.N. Das, and H. Kabir, "Comparative studies on some metrics for external validation of QSPR models", *Journal of Chemical Information and Modeling*, 52, 2012, 396–408.
[45] P.K. Ojha, I. Mitra, R. Das, and K. Roy, "Further exploring rm2 metrics for validation of QSPR models", *Chemometrics and Intelligent Laboratory Systems*, 107, **2011**, 194–205.
[46] OECD, "Guidance document on the validation of QSAR models Organization for Economic Co-operation & Development", *Paris*, **2007**.
[47] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, and P. Gramatica, "Methods for reliability & uncertainty assessment & for applicability evaluations of classification & regression-based QSARs", *Environmental Health Perspectives*, 111, **2003**, 1361–1375.
[48] P. Gramatica, "Principles of QSAR models validation: internal & external", *QSAR & Combinatorial Science*, 26, **2007**, 694–701.
[49] G.E. Batista and D.F. Silva, "How k-Nearest Neighbor Parameters Affect its Performance, Argentine Symposium on Artificial Intelligence", *Instituto de Ciencias Matemáticase de Computa cao*, Sao Carlos – SP – Brasil, **2009**, 1–12.
[50] T.I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. van de Sandt, W. Tong, G. Veith, and C. Yang, "Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships", *Alternatives to Laboratory Animals*, 33(2), **2005**, 155–173.
[51] J.C. Dearden, "The history and development of Quantitative Structure-Activity Relationships (QSARs)", *International Journal of Quantitative Structure-Property Relationships*, 1, **2016**, 1–44.
[52] J.H. Al-Fahemi, "The use of quantum-chemical descriptors for predicting the photo induced toxicity of PAHs", *Journal of Molecular Modeling*, 18, **2012**, 4121–4129.
[53] J.G. Topliss and R.P. Edwards, "Chance factors in studies of quantitative structure-activity relationships", *Journal of Medicinal Chemistry*, 22, **1979**, 1238–1244.
[54] K. Fukui, "Role of Frontier Orbitals in Chemical Reactions", *Science*, 218, **1982**, 747–754.

Chapitre 5 : Etude de la RQSA de l'activité antagoniste vis-à-vis du récepteur NMDA pour des dérivés de la Dizocilpine (MK801)



Available online at www.sciencedirect.com

Science Direct

Journal of Taibah University for Science 9 (2015) 143–154

Journal of Taibah University for Science

www.elsevier.com/locate/jtusci

Quantitative structure–activity relationship studies of dibenzo[*a*,*d*]cycloalkenimine derivatives for non-competitive antagonists of *N*-methyl-D-aspartate based on density functional theory with electronic and topological descriptors

Samir Chtita [a,*], Majdouline Larif [b], Mounir Ghamali [a], Mohammed Bouachrine [c], Tahar Lakhlifi [a]

[a] *Laboratoire de la Chimie Moléculaire et des Substances Naturelles, Faculté des Sciences, Université Moulay Ismail, Meknès, Maroc* [b] *Laboratoire des Procédés de Séparation, Faculté des Sciences, Université Ibn Tofail, Kénitra, Maroc*

[c] *Ecole Supérieure de Technologie Meknès, Université Moulay Ismail, Meknès, Maroc*

Available online 4 December 2014

Abstract

To establish a quantitative structure-activity relationship for non-competitive antagonists of N-methyl-D-aspartate receptor, 48 substituted dibenzo[a,d]cycloalkenimine derivatives were analyzed by principal components, a descendant multiple regression analyses, multiple non-linear regression and an artificial neural network. We propose non-linear and linear quantitative structure-activity models we interpret the activity of the compounds by the multivariate statistical analysis. Density functional theory with Becke's three-parameter hybrid function and Lee-Yang-Parr exchange correlation functional calculations were performed to define the structure, chemical reactivity and properties of the study compounds. The topological and the electronic descriptors were computed with ACD/ChemSketch and Gaussian 03W programs, respectively. The study shows that multiple regression and multiple non-linear regression analyses predict activity; however, predictions made with a 6-2-1 artificial neural network model were more accurate. This model gave statistically significant results and showed good stability to data variation in leave-one-out cross-validation. The applicability domain of models was investigated using William's plot to detect outliers and outsides compounds.

Keywords: *N*-Methyl-D-aspartate; Quantitative structure–activity relation; Density functional theory; Dibenzo[*a*,*d*]cycloalkenimines; Artificial neural network; Cross-validation

1. Introduction

Excitatory amino acids form the mainstay of synaptic transmission in the central nervous system. By the same token, dysfunctional toxic activity of excitatory amino acids can lead to or become instrumental in the progression of a number of neurological and neurodegenerative conditions, such as epilepsy, Huntington disease, Alzheimer disease and schizophrenia. Dementia due to Alzheimer disease is characterized by extracellular plaques containing amyloid (β peptide), which disrupt dendritic morphology and affect glutamate (α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid and N-Methyl-D-aspartate) receptor function to alter glutamatergic transmission. Huntington disease manifests as atrophy of the corpus striatum and cortex, with neurons containing the mutant Huntington protein which are perhaps more susceptible to excitotoxicity from corticostriatal inputs, as reflected by loss of the N-methyl-d-aspartate receptor and interactions with facilitatory group I metabotropic glutamate receptor.

* Corresponding author. Tel.: +212 660005554.
E-mail address: samirchtita@gmail.com (S. Chtita).

Peer review under responsibility of Taibah University.

http://dx.doi.org/10.1016/j.jtusci.2014.10.006
1658-3655 © 2015

In schizophrenia, abnormalities in brain (dendritic) development and synaptic plasticity may precipitate dysfunction of mesolimbic and mesocortical dopaminergic pathways. Here again, aberrations in glutamatergic transmission in the form of N-methyl-d-aspartate receptor hypofunction may be involved **[1]**.

Dizocilpine, a compound originally characterized as an anticonvulsant, is a potent *N*-Methyl-D-aspartate receptor antagonist **[2]**. It binds selectively and with high affinity to the receptors when they are open **[3]** and is therefore referred to as a use-dependent *N*-Methyl-D-aspartate receptor open channel blocker with a very slow off-rate. These properties can be exploited to 'pre-block' a population of receptors, such as synaptic ones, resulting in selective activation of a different population, such as extra-synaptic receptors. The usefulness of this approach depends on the stability of dizocilpine blockade after washout **[4]**.

Early electrophysiological and ligand binding studies revealed that blockade of *N*-Methyl-D-aspartate receptors by dizocilpine persisted long after the drug had been washed out **[5]**. This unusual property means that the blockade of receptors is highly stable and can be regarded as 'irreversible' over many experiments. It is therefore used to 'permanently' block a subpopulation of *N*-Methyl-D-aspartate receptors in order to study a separate population of non-blocked receptors. An example of such use is in the study of synaptic and extra-synaptic receptors **[6]**. Differential signaling by the neuroprotective phasic activation of synaptic *N*-Methyl-D-aspartate receptors as compared with the deleterious tonic activation of extra-synaptic receptors is a topic of much current interest **[7–19]**. Quantitative structure–activity relations are used to investigate the relations between molecular descriptors of the unique physicochemical properties of a set of compounds and their biological activity or chemical property **[20, 21]**. We attempted to establish a quantitative structure–activity relation for non-competitive antagonists of *N*-methyl-D-aspartate receptors by studying a series of 48 substituted dibenzo*[a,d]*cycloalkenimine derivatives.

2. Material and Methods

2.1. Experimental dataset

In order to determine quantitative structure–activity relations for non-competitive antagonists of *N*-methyl-D-aspartate receptors, we used 48 compounds that have been synthesized and evaluated for their ability to displace dibenzo*[a,d]*cycloalkenimines from their specific binding sites on rat cortical membranes and for their antagonistic activity against the *N*-methyl-D-aspartate receptor. As proposed by Thompson et al. **[22]**, 38 molecules were selected for the quantitative model (training set), and 10 were selected randomly to test the performance of the proposed model (test set). The molecular structures

and computational models for the 48 derivatives are shown in figure 1 and table 1 and described by their substituents as R_1, R_2, R_3, R_4, R_5, R_6 and R_7. Although Thompson et al. proposed 73 compounds; the structures of the remaining compounds are different from that required for this study.

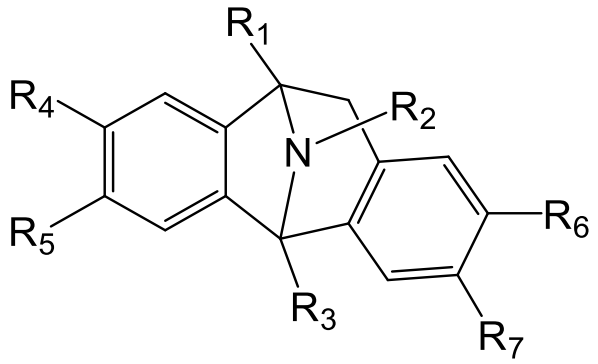


Figure1: Schematic diagram of MK801 skeleton

2.2. Computational methods

The activity of these compounds was correlated with various physicochemical parameters by density functional theory. The three-dimensional structures were generated with Gauss View 3.0, and all calculations were performed with Gaussian 03W programs. Geometrical optimization of 48 compounds was carried out by Lee–Yang–Parr exchange correlation functional with the 6-31G (d) basic set **[23–27]**. The geometry of the compounds was determined by optimizing all geometrical variables with no symmetry constraints **[28]**. ChemSketch **[29]** was used to calculate the other molecular descriptors.

The quantum chemistry descriptors were obtained for the model from the density functional theory calculations as follows: total energy (E), highest occupied molecular orbital (HOMO) energy (E_{HOMO}), lowest unoccupied molecular orbital (LUMO) energy (E_{LUMO}), the difference between LUMO and HOMO energy (Gap), the total dipole moment of the compound (μ, Debye), absolute hardness (η), absolute electronegativity (χ) and the reactivity index (ω) **[30]**. η, χ and ω were determined from:

$$\eta = \frac{E_{LUMO} - E_{HOMO}}{2}$$

$$\chi = \frac{E_{LUMO} + E_{HOMO}}{2}$$

$$\omega = \frac{\chi^2}{2\eta}$$

Table1: The structure and the observed activities for the 48 MK801 derivatives

N°	R1	R2	R3	R4	R5	R6	R7	Obs.
Test Set								
1	H	CH_3	CH_3	H	H	H	H	-0.215
2	H	H	CH_3	H	H	H	H	-1.252
3	CH_3	CH_3	CH_3	H	H	H	H	-0.149
4	H	OH	CH_3	H	H	H	H	1.279
5	H	H	CH_2CH_3	H	H	H	H	-1.347
6	H	CH_2CH_3	H	H	H	H	H	1.380
7	OH	OH	CH_2CO_2ET	H	H	H	H	3.653
8	OH	H	CH_2CO_2ET	H	H	H	H	0.556
9	H	H	CH_2CH_2OH	H	H	H	H	-0.585
10	H	H	CH_2CO_2ET	H	H	H	H	-0.260
Training Set								
11	H	H	CH_2OH	H	H	H	H	-0.456
12	H	H	CH_2F	H	H	H	H	-0.796
13	H	H	CH_2CH_2F	H	H	H	H	-0.759
14	H	H	$CH_2SC_6H_5$	H	H	H	H	1.863
15	H	H	$CH_2S(O)C_6H_5$	H	H	H	H	2.204
16	OH	H	CH_3	H	H	H	H	-1.114
17	F	H	CH3	H	H	H	H	-0.032
18	H	H	$CH=CH_2$	H	H	H	H	-1.060
19	OH	H	CH_2CO_2H	H	H	H	H	2.447
20	OH	H	CH_2CONH_2	H	H	H	H	0.806
21	Cl	H	CH_2CO_2H	H	H	H	H	3.653
22	Cl	H	CH_2CONH_2	H	H	H	H	1.954
23	H	H	$CH(CO)CO_2H$	H	H	H	H	3.000
24	Cl	H	CH_2CH_2Cl	H	H	H	H	1.724
25	H	H	CH_3	H	H	Cl	H	-1.076
26	H	H	CH_3	H	Cl	H	H	-1.959
27	H	H	CH_3	H	H	Br	H	-0.745
28	H	H	CH_3	H	H	OCH_3	H	-1.444
29	H	H	CH_3	H	H	OH	H	-1.638
30	H	H	CH_3	H	NH_2	H	H	-1.569
31	H	H	CH_3	H	Br	H	H	-1.097
32	H	H	CH_3	H	OCH_3	H	H	-1.337
33	H	H	CH_3	H	OH	H	H	-1.745
34	H	H	CH_3	H	CH_2OH	H	H	-0.863
35	H	H	CH_3	H	CH_3	H	H	-1.469
36	H	H	CH_3	H	$(CH_2)_3CH_3$	H	H	0.097
37	H	H	CH_3	H	C_6H_5	H	H	-1.495
38	H	H	CH_3	OCH_3	H	H	H	-0.215
39	H	H	CH_3	H	H	H	OCH_3	-1.481
40	H	H	CH_3	OH	H	H	H	-0.558
41	H	H	CH_3	H	H	H	OH	-1.310
42	H	H	CH_3	H	F	F	H	-1.509
43	H	CH_3	H	H	H	H	H	1.081
44	H	$(CH_2)_2OH$	H	H	H	H	H	1.854
45	H	H	$(CH_2)_2CH_3$	H	H	H	H	0.921
46	OH	H	$(CH_2)_2OH$	H	H	H	H	-1.328
47	H	H	$CH(OH)CO_2Et$	H	H	H	H	-0.408
48	Cl	H	$(CH_2)_2OH$	H	H	H	H	-0.553

The Advanced Chemistry Development ChemSketch program was used to calculate formula weight, molar volume (cm^3), molar refractivity (cm^3), parachor (cm^3), density (g/cm^3), refractive index (n), surface tension (γ(dyne/cm)) and polarizability (αe (cm^3)) **[31]**.

2.3. Statistical analysis

To explain the structure-activity relations, we used principal component analysis, multiple linear and non-linear regression in XLSTAT software **[32]**. The artificial neural network and leave-one-out cross-validation were performed with a program written in C language in MATLAB version 7.12.

Principal component analysis is a statistical technique used for summarizing information encoded in the structures of compounds and for understanding the distribution. It is essentially a descriptive statistical method for presenting the maximum information contained in tables 2 and 3 in graphical form.

Multiple linear regression is used to study the relation between one dependent and several independent variables. It minimizes differences between actual and predicted values and was used to select the descriptors to be used as inputs into multiple non-linear regression and the artificial neural network. Multiple linear and non-linear regression were used to predict effects on the activity of dibenzo*[a,d]*cycloalkenimines (K_i). Equations were justified by the correlation coefficient (*r*), the mean squared error, Fisher's F statistic and the significance level (F *value*) **[33]**.

Artificial neural networks are artificial systems that simulate the function of the human brain. A neural network has three components: the processing elements or nodes, the topology of the connections between the nodes, and the rule by which new information is encoded in the network. While there are a number of models, the most frequently used artificial neural network in quantitative structure–activity relation studies is a three-layered feed-forward network **[34]**. In this type of network, the neurons are arranged in layers, with an input layer, one hidden layer and an output layer. Each neuron in each layer is fully connected to the neurons of a succeeding layer, and there are no connections between neurons in the same layer. According to the supervised learning model, the networks are taught by giving them examples of input patterns and the corresponding target outputs. Through an iterative process, the connection weights are modified until the network gives the desired results for the training set of data. A back-propagation algorithm is used to minimize the error function. This algorithm has been described previously, with a simple example of application **[35]**. A detailed description of this algorithm is given elsewhere **[36]**. Cross-validation is used to explore the reliability of statistical models. In this technique, a number of modified datasets

are created by deleting one or a small group of molecules, known respectively as "leave-one-out" and "leave-some-out" [37–39]. For each dataset, an input–output model is prepared, and its accuracy in predicting the responses of the remaining data (those that were not used in the model) is evaluated. In this study, we used the leave-one-out procedure.

The final stage of this QSAR analysis consists of applicability domain, a model is valid only within its training domain and new molecules must be considered as belonging to the domain before the model is applied (OECD principle 3 [40]). Without applicability domain (AD), each model can predict the activity of any compound, even with a completely different structure from those included in the study. Therefore, the AD is a tool to find out compounds that are outside the applicability domain of the built QSAR model and it detects outliers present in the training set compounds. There are several methods for defining the applicability domain (AD) of QSAR models [41], but the most common one is determining the leverage values h_i ($h_i = x_i^T (X^T X)^{-1} x_i$ ($i = 1,2,\dots,n$)) for each compound [42].

Where: xi: the descriptor row-vector of query compound, X: the n*k-1 matrix of k model descriptor values for n training set compounds. The superscript "T" refers to the transpose of matrix/vector.

In this study, we use the Williams plot; in this plot, the applicability domain is established inside a squared area within standard deviation ±x (in this study x = 2.5 ("three sigma rule" [43]) and a leverage threshold h* (h*= 2.5*(k+1)/n) [44].

Where: *n* is the number of training set compounds and *k* is the number of model descriptors.

3. Results and Discussions

3.1. Dataset for analysis

The quantitative structure–activity relation analysis was performed using the Log K_i of the 48 selected molecules that have been synthesized and evaluated for their ability to displace dibenzo*[a,d]*cycloalkenimines (experimental values) as reported by Thompson et al. [22], The values of the 16 chemical descriptors as shown in table 2. The principle is to perform in the first time, a main component analysis (PCA), which allows us to eliminate descriptors that are highly correlated (dependent), then perform a decreasing study of multiple linear regression based on the elimination of descriptors aberrant until a valid model (including the critical probability: *p-value*<0.05 for all descriptors and the model complete).

Table 2: The values of the 16 chemical descriptors

	PM	MR	MV	Pc	n	γ	D	αe	Log -*E*	E_{HOMO}	E_{LUMO}	Gap	μ	χ	η	ω
Test Set																
1	235.324	73.95	208.3	535.7	1.628	43.7	1.129	29.31	4.288	-5.767	-0.157	5.610	0.805	2.962	2.805	1.564
2	221.297	69.02	193.3	510.8	1.632	48.7	1.144	27.36	4.263	-5.948	-0.184	5.764	1.079	3.066	2.882	1.631
3	249.350	78.56	223.1	574.1	1.621	43.8	1.117	31.14	4.311	-5.698	-0.133	5.564	0.850	2.915	2.782	1.527
4	237.296	70.83	185.3	512.7	1.689	58.5	1.280	28.08	4.309	-6.047	-0.190	5.857	0.578	3.119	2.928	1.661
5	235.324	73.64	210.9	555.6	1.615	48.1	1.115	29.19	4.288	-5.941	-0.195	5.745	1.055	3.068	2.873	1.638
6	235.324	73.96	211.2	537.4	1.617	41.9	1.114	29.32	4.288	-5.817	-0.176	5.641	0.795	2.997	2.820	1.592
7	325.358	87.98	231.1	673.8	1.686	72.1	1.407	34.88	4.473	-5.911	-0.226	5.684	1.043	3.068	2.842	1.656
8	309.359	86.17	239.1	669.7	1.640	61.4	1.293	34.16	4.442	-6.009	-0.153	5.856	1.987	3.081	2.928	1.621
9	251.323	75.18	208.4	572.3	1.640	56.8	1.205	29.80	4.331	-5.982	-0.251	5.731	2.315	3.116	2.866	1.695
10	293.360	84.68	247.4	659.9	1.600	50.5	1.185	33.57	4.408	-5.910	-0.130	5.780	2.357	3.020	2.890	1.578
Training Set																
11	237.296	70.54	191.9	532.5	1.656	59.2	1.236	27.96	4.309	-5.873	-0.137	5.736	1.140	3.005	2.868	1.574
12	239.287	69.17	199.8	525.3	1.608	47.7	1.197	27.42	4.323	-6.068	-0.307	5.761	1.379	3.188	2.880	1.764
13	253.314	73.81	216.3	216.3	1.597	46.5	1.170	29.26	4.344	-6.105	-0.369	5.736	2.442	3.237	2.868	1.827
14	329.458	102.25	259.4	723.9	1.717	60.5	1.260	40.53	4.550	-5.959	-0.582	5.377	1.954	3.271	2.689	1.989
15	345.457	103.26	276.7	758.4	1.668	56.3	1.248	40.93	4.574	-6.105	-0.698	5.407	1.549	3.402	2.704	2.140
16	237.296	70.51	185.0	520.6	1.687	62.5	1.282	27.95	4.309	-6.057	-0.231	5.826	0.526	3.144	2.913	1.697
17	239.287	69.67	191.1	505.1	1.649	48.7	1.250	27.62	4.323	-6.179	-0.370	5.810	2.012	3.274	2.905	1.845
18	233.308	75.10	194.2	526.2	1.700	53.9	1.201	29.77	4.287	-5.991	-0.234	5.756	1.113	3.112	2.878	1.683
19	281.306	76.70	197.3	587.2	1.705	78.4	1.425	30.40	4.407	-6.074	-0.257	5.816	1.150	3.165	2.908	1.723
20	280.321	78.70	203.4	599.5	1.700	75.3	1.377	31.20	4.397	-6.027	-0.241	5.786	2.947	3.134	2.893	1.698
21	299.752	80.67	208.1	597.1	1.702	67.7	1.440	31.98	4.556	-6.253	-0.500	5.753	2.145	3.376	2.877	1.982
22	298.767	82.75	214.3	609.9	1.699	65.5	1.390	32.80	4.549	-6.128	-0.341	5.787	3.421	3.234	2.894	1.808
23	281.306	76.65	201.0	574.8	1.687	66.7	1.398	30.38	4.407	-5.873	-0.524	5.349	2.147	3.199	2.674	1.913
24	304.214	83.85	225.1	613.9	1.667	55.2	1.350	33.24	4.647	-6.445	-0.727	5.718	3.110	3.586	2.859	2.249
25	255.742	73.92	205.3	546.6	1.639	50.2	1.245	29.30	4.489	-6.088	-0.427	5.661	3.107	3.258	2.831	1.874
26	255.742	73.92	205.3	546.6	1.639	50.2	1.245	29.30	4.489	-6.124	-0.488	5.636	1.569	3.306	2.818	1.939
27	300.193	76.71	209.5	561.3	1.653	51.4	1.432	30.41	4.946	-6.060	-0.437	5.623	3.026	3.249	2.812	1.877
28	251.323	75.70	217.3	567.4	1.613	46.4	1.156	30.01	4.331	-5.543	-0.118	5.425	1.235	2.830	2.712	1.477
29	237.296	70.91	191.7	525.8	1.661	56.5	1.237	28.11	4.309	-5.600	-0.146	5.455	1.498	2.873	2.727	1.513
30	236.312	73.82	204.1	546.0	1.642	51.1	1.150	29.26	4.297	-5.237	-0.082	5.154	2.809	2.659	2.577	1.372
31	300.193	76.71	209.5	561.3	1.653	51.4	1.432	30.41	4.946	-6.104	-0.488	5.616	1.479	3.296	2.808	1.934
32	251.323	76.27	215.5	46.4	1.625	46.4	1.160	30.23	4.331	-5.626	-0.168	5.459	2.299	2.897	2.729	1.538
33	237.296	71.73	197.9	533.3	1.644	52.6	1.190	28.43	4.309	-5.678	-0.221	5.458	2.176	2.949	2.729	1.594
34	251.323	75.47	207.1	564.3	1.649	55.0	1.213	29.92	4.331	-5.854	-0.145	5.709	1.931	2.999	2.854	1.576
35	235.324	73.85	209.6	548.4	1.622	46.8	1.122	29.27	4.288	-5.869	-0.158	5.711	1.341	3.014	2.856	1.590
36	277.403	87.84	259.1	666.9	1.593	43.8	1.070	34.82	4.354	-5.860	-0.149	5.711	1.395	3.004	2.855	1.581
37	297.393	93.62	258.6	684.1	1.643	48.9	1.149	37.11	4.391	-5.794	-0.684	5.109	0.978	3.239	2.555	2.053
38	251.323	76.27	215.5	562.6	1.625	46.4	1.160	30.23	4.331	-5.557	-0.121	5.436	2.180	2.839	2.718	1.483
39	251.323	75.70	217.3	567.4	1.613	46.4	1.156	30.01	4.331	-5.528	-0.165	5.363	0.961	2.847	2.681	1.511
40	237.296	71.73	197.9	533.3	1.644	52.6	1.190	28.43	4.309	-5.612	-0.164	5.448	1.980	2.888	2.724	1.531
41	237.296	70.91	191.7	525.8	1.661	56.5	1.237	28.11	4.309	-5.587	-0.203	5.383	1.154	2.895	2.692	1.557
42	257.278	69.01	201.7	525.0	1.599	45.8	1.275	27.36	4.375	-6.089	-0.501	5.588	2.061	3.295	2.794	1.943
43	221.297	69.34	193.6	497.4	1.635	43.5	1.142	27.49	4.263	-5.837	-0.185	5.652	0.839	3.011	2.826	1.604
44	251.323	75.49	208.7	554.4	1.643	49.8	1.204	29.93	4.331	-5.757	-0.162	5.595	1.968	2.960	2.798	1.565
45	249.350	78.28	227.4	595.3	1.604	46.9	1.096	31.03	4.311	-5.938	-0.196	5.742	1.079	3.067	2.871	1.638
46	267.322	76.67	200.1	582.1	1.691	71.5	1.335	30.39	4.371	-6.092	-0.295	5.797	1.587	3.194	2.899	1.759
47	309.359	86.12	242.9	658.3	1.627	53.9	1.273	34.14	4.442	-5.794	-0.340	5.454	3.325	3.067	2.727	1.725
48	285.768	80.55	211.1	592.0	1.688	61.7	1.350	31.93	4.531	-6.277	-0.559	5.718	3.765	3.418	2.859	2.043

Table 3: The correlation matrix (Pearson (n)) between different obtained descriptors

	PM	MR	MV	Pc	n	γ	D	αe	log-*E*	E_{HOMO}	E_{LUMO}	Gap	μ	χ	η	ω	log K_i
PM	1																
MR	0.862	1															
MV	0.720	**0.911**	1														
Pc	0.536	0.572	0.473	1													
n	0.368	0.243	-0.175	0.262	1												
γ	0.457	0.233	-0.107	0.297	0.834	1											
D	0.587	0.164	-0.136	0.219	0.722	0.782	1										
αe	0.862	**1**	**0.911**	0.572	0.243	0.233	0.164	1									
log-*E*	0.710	0.406	0.281	0.271	0.302	0.235	0.687	0.406	1								
E_{HOMO}	-0.398	-0.169	-0.026	-0.156	-0.351	-0.354	-0.551	-0.169	-0.484	1							
E_{LUMO}	-0.619	-0.486	-0.339	-0.263	-0.339	-0.199	-0.491	-0.486	-0.629	0.632	1						
Gap	-0.083	-0.249	-0.290	-0.053	0.124	0.260	0.235	-0.249	0.017	-0.673	0.148	1					
μ	0.395	0.186	0.130	-0.003	0.135	0.203	0.396	0.186	0.434	-0.232	-0.350	-0.037	1				
χ	0.544	0.337	0.177	0.223	0.382	0.317	0.580	0.337	0.603	**-0.930**	-0.872	0.355	0.312	1			
η	-0.082	-0.248	-0.289	-0.052	0.125	0.261	0.236	-0.248	0.019	-0.674	0.147	**1**	-0.036	0.356	1		
ω	0.610	0.446	0.289	0.261	0.367	0.255	0.539	0.446	0.635	-0.783	**-0.976**	0.068	0.339	**0.957**	0.069	1	
log K_i	0.520	0.437	0.245	0.345	0.473	0.489	0.455	0.437	0.219	-0.335	-0.243	0.196	0.044	0.327	0.196	0.291	1

3.2. Principal Component Analysis (PCA)

All 16 descriptors (variables) coding the 48 molecules were submitted to principal components analysis, and 17 components were obtained (Figure 2). The first three axes, F1, F2 and F3, contributed 41.5%, 21.6% and 11.1%, respectively, to the total variance, and the total information was estimated to be 74.2%. The Pearson correlation coefficients are summarized in table 3; the matrix provides information on the negative and positive correlations between variables. Correlations among the 16 descriptors are shown in table 3 as a correlation matrix; in figure 3, these descriptors are represented in a correlation circle.

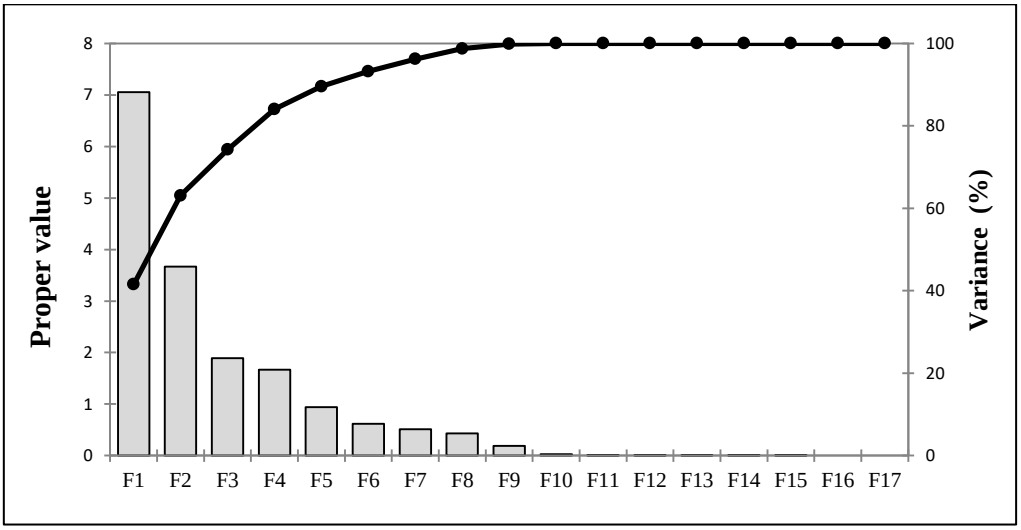


Figure 2: The principal components and their variances.

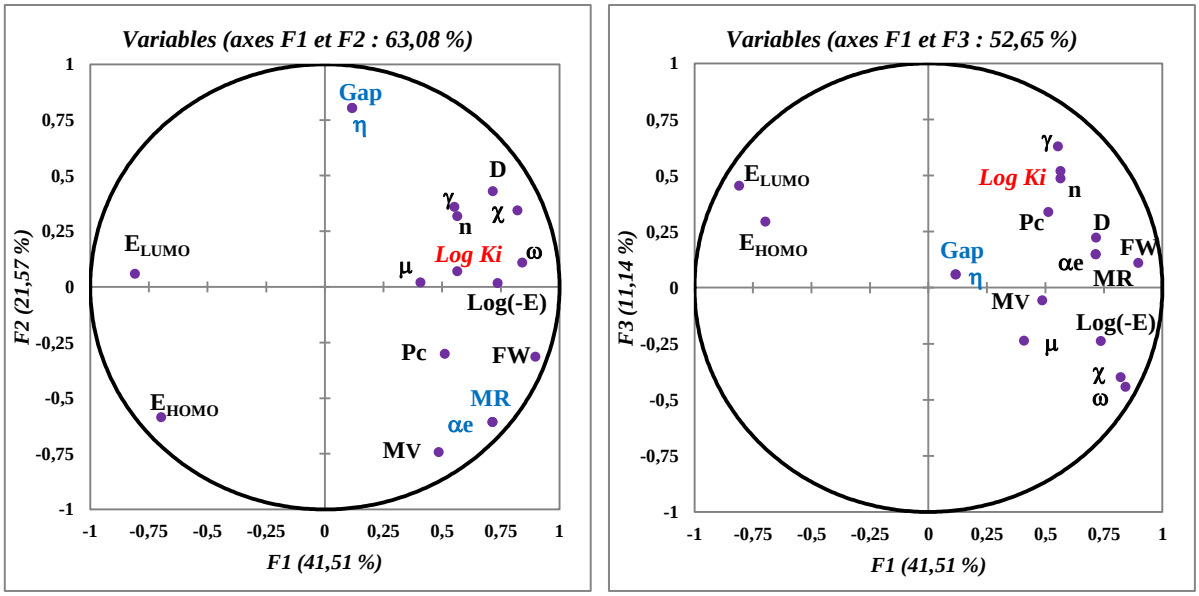


Figure 3: Correlation circles

As molar refractivity, polarizability and energy are perfectly correlated (r = 1); gap and absolute hardness are perfectly correlated (r = 1); both variables are redundant.

Molar refractivity, molar volume and polarizability are highly correlated (r (molar volume, molar refractivity) = 0.911, r (molar volume, polarizability) = 0.911.

E_{HOMO} and absolute electronegativity are strongly negatively correlated (r = −0.930), and E_{LUMO}, absolute electronegativity and reactivity index are strongly negatively correlated: r (E_{LUMO}, reactivity index) = −0.976; r (absolute electronegativity, reactivity index) = −0.957.

The variables polarizability, gap, molar volume and absolute electronegativity were therefore removed.

In the projection of the compounds in the plane of the three first axes, F1, F2 and F3 (Figure 4), they are distributed in three regions. Region 1 contains those with log (−E) values between 4.263 (E = −18323.14 eV) and 4.354 (E = −22594.36 eV), region 2 contains compounds with values between 4.371 (E = −23496.33 eV) and 4.489 (E = −30831.88 eV), and region 3 contains compounds with values between 4.531 (E = −33962.53 eV) and 4.946 (E = −88307.99 eV).

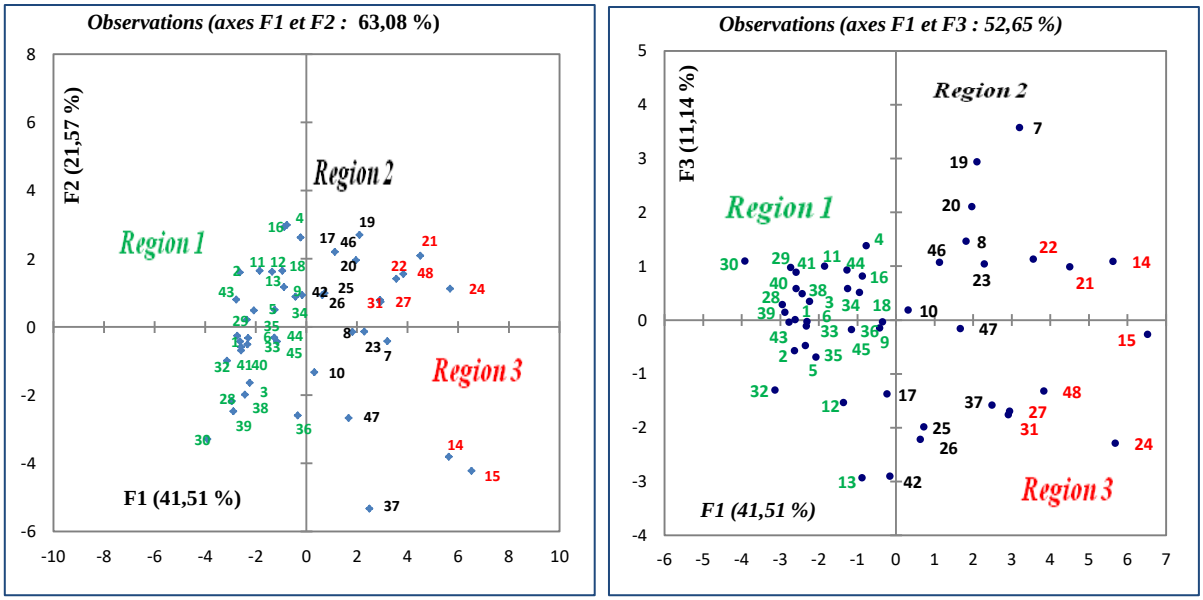


Figure 4: Cartesian diagram showing the separation between the tree regions and the dispersal of different molecules by groups

3.3. Multiple Linear Regressions (MLR)

In order to propose a mathematical model and to evaluate physicochemical effects on the activity of the entire set of 48 compounds quantitatively, we submitted the data matrix constituted from the 12 variables corresponding to the training set to descendent multiple regression analysis. The correlation coefficient, *r*, the coefficient of determination, r^2, mean squared error and F *values* were used to select the best regression performance. Multiple linear regression allowed connection of the structural descriptors for the activity of each of the 38 compounds in order to evaluate the effect of the substituent quantitatively. The descriptors selected were molar refractivity (MR), surface tension (γ), density, log ($-E$), E_{LUMO} and reactivity index (ω). The quantitative structure–activity relation model built with multiple linear regression is represented by:

$$\log K_i = -9.783 + 0.192\ MR - 0.134\ \gamma + 24.864\ d - 9.411 \log(-E) + 13.416\ E_{LUMO} + 9.651\ \omega$$

(Equation 1)

$$N = 38 \qquad r = 0.731 \qquad r^2 = 0.535 \qquad MSE = 1.276$$

A high correlation coefficient and a low mean squared error, indicate that the model is reliable. As the P_{value} is less than 0.05, we would be taking a less than 0.01% risk in assuming that the null hypothesis is wrong. Therefore, we can conclude that the model provides a significant amount of information. The quantitative structure–activity relation model showed that the *N*-Methyl-D-aspartate antagonist activity can be explained by a number of electronic and topological factors. The negative correlation between surface tension and total energy and the ability to displace the activity of dibenzo*[a,d]*cycloalkenimines results in a decrease in log

K_i, while the positive correlation between the topological descriptors density and molar refractivity and electronic descriptors E_{HOMO} and E_{LUMO} indicates the ability to displace dibenzo*[a,d]*cycloalkenimine activity, with an increase in log K_i. The predicted log K_i activities calculated from equation 1 in the optimal multiple linear regression model and the observed values are given in table 4. The correlations between the predicted and observed activities and the residue values are illustrated in figure 5. The descriptors proposed in Eq. (1) were therefore used as the input parameters for multiple nonlinear regression and the artificial neural network.

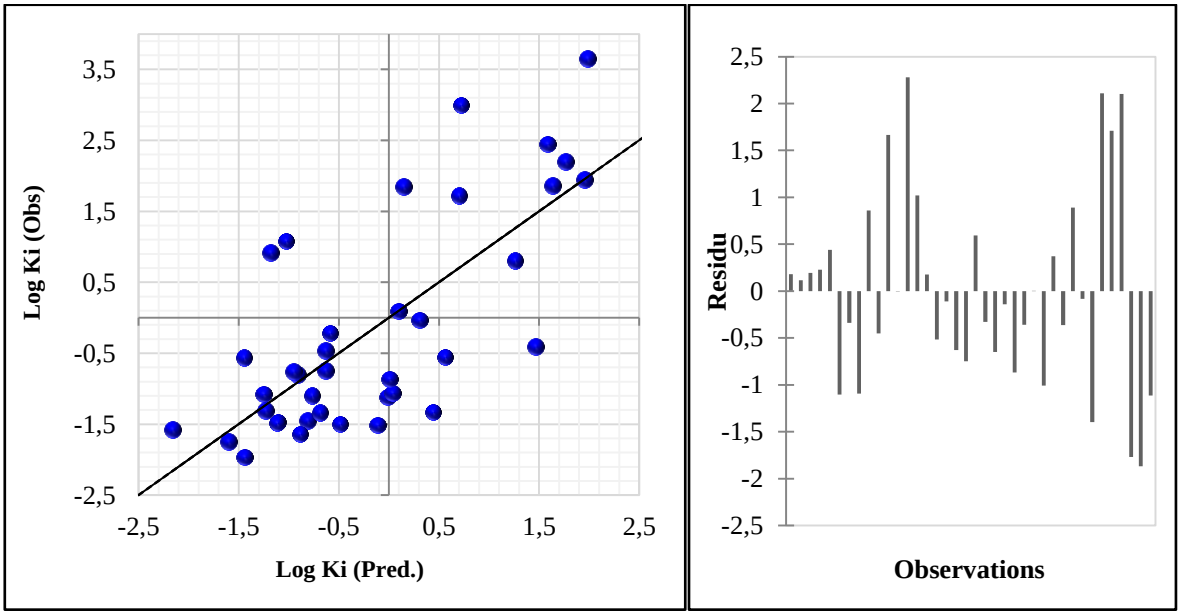


Figure 5: Correlations of observed and predicted activities and the residues values calculated using MLR.

3.4. Multiples Nonlinear Regression (MNLR)

We used also a non-linear regression model to improve the structure–activity relation and evaluate the effect of substituents quantitatively. We applied the descriptors proposed by multiple linear regression for the 38 molecules in the training set and used the coefficients *r*, r^2 and *MSE* to select the best regression performance. We used a pre-programmed function of XLSTAT as follows:

$$Y = a + (b\,X_1 + c\,X_2 + d\,X_3 + e\,X_4\ldots) + (f\,X_1^2 + g\,X_2^2 + h\,X_3^2 + i\,X_4^2\ldots)$$

Where *a, b, c, d*...: represent the parameters and X_1, X_2, X_3, X_4...: represent the variables.

The resulting equation was:

$$\text{Log } K_i = -97.482 - 0.893\ MR + 0.154\ \gamma - 220.100\ d + 117.475 \log(-E) + 18.508\ E_{LUMO} + 2.298\ \omega + 0.006\ (MR)^2 - 0.003\ (\gamma)^2 + 98.188\ d^2 - 14.042\ (\text{Log}(-E))^2 + 6.731(E_{LUMO})^2 + 1.963\ (\omega)^2 \quad \text{(Equation 2)}$$

$N = 38$ $r = 0.852$ $r^2 = 0.726$ $MSE = 0.933$

The predicted activities calculated from Eq. (2) and the observed values are given in table 4. The correlations of the predicted and observed activities and the residues values are illustrated in figure 6.

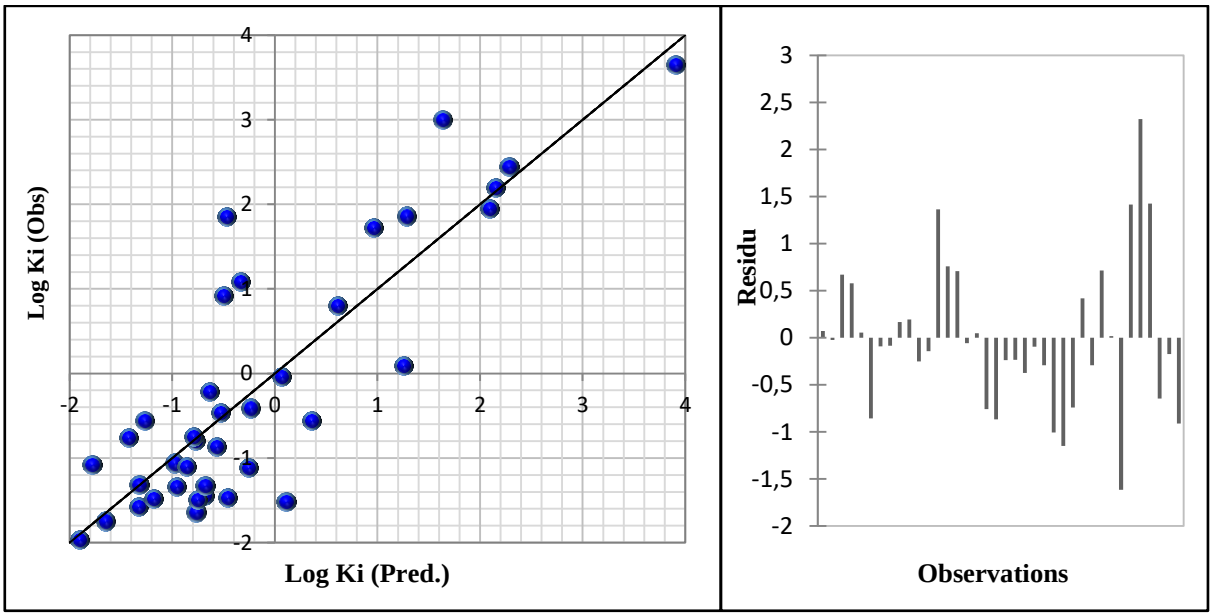


Figure 6: Correlations of observed and predicted activities and the residues values calculated using MNLR

The true predictive power of a quantitative structure–activity relation model is its ability to predict accurately the activities of compounds in an external test set (compounds not used in the model development). The activities of the remained 10 compounds are deduced from the training set by multiple linear and non-linear regression. Their structures and the observed and calculated log K_i values are given in tables 1 and 5. Comparison of the values of log (K_i-*test*) and log (K_i -*obs*.) shows that good predictions for the 10 compounds:

Multiple linear regression: $N = 10$, $r_{test} = 0.750$, $r^2_{test} = 0.563$

Multiple non-linear regression: $N = 10$, $r_{test} = 0.845$, $r^2_{test} = 0.715$

3.5. Artificial Neural Networks (ANN)

In order to increase the probability of characterizing the compounds well, artificial neural networks can be used to generate predictive models of quantitative structure–activity relations between a set of molecular descriptors obtained from multiple linear regression and the observed activities. The model was prepared with the properties of several of the compounds. A parameter, ρ, has been proposed for determination of the number of hidden neurons, which play a major role in determining the best artificial neural network architecture **[45, 46]**, defined as follows:

$$\rho = \frac{\text{Number of data points in the training set}}{\text{Sum of the number of connections in the NN}}$$

In order to avoid over-fitting or under-fitting, it is recommended that $1.8 < \rho < 2.3$ **[47]**. The output layer represents the calculated activity values log K_i. The architecture of the artificial neural network used in this work (6-2-1), with $\rho = 2.11$.

The correlation between the calculated and experimental artificial neural network and the residue values were highly significant, as indicated by the r and r^2 values (N = 38, r = 0.849, r^2 = 0.721). The predicted activities calculated with the artificial neural network and the observed values are given in table 4. The r value confirms that the results of the artificial neural network were the best for building quantitative structure–activity models.

'Leave-one-out' is an approach particularly well adapted for estimating the predictive ability of these models. The correlations between the observed activities, the cross-validation, calculated values and the residues are given in table 4. The good results obtained with cross-validation (r_{cv} = 0.836) and for the prediction of the activities of the 10-compound test set show that the model proposed in this paper can accurately predict activity and that the selected descriptors are pertinent. The results obtained by multiple linear and non-linear regression are sufficient to conclude that the model performs well. Although the good predictive ability might be due to chance, we consider it a positive result. This model could therefore be used for all the dibenzo*[a,d]*cycloalkenimine derivatives (Table 1) and could add to the search for non-competitive antagonists of *N*-methyl-D-aspartate receptors and their interaction with the receptor. Our tests with multiple linear and non-linear regression and artificial neural network models showed a substantially better predictive capability of the artificial neural network model. It showed a satisfactory relation between the molecular descriptors and the activity of the compounds and a good correlation with cross-validation (r_{cv}= 0.836).

Table 4: The observed, the predicted activities (Log K_i), and residue according to different methods for the 38 MK801 derivatives (training set).

N°	Obs.	MLR		MNLR		ANN		CV	
		Pred.	Resid.	Pred.	Resid.	Pred.	Resid.	Pred.	Resid.
11	-0.456	-0.635	0.179	-0.527	0.071	-0.818	-0.362	-0.161	0.295
12	-0.796	-0.911	0.115	-0.771	-0.025	-0.818	-0.022	-0.819	-0.023
13	-0.759	-0.954	0.195	-1.429	0.670	-0.818	-0.059	-0.820	-0.061
14	1.863	1.633	0.230	1.284	0.579	1.786	-0.077	-0.224	-2.087
15	2.204	1.764	0.440	2.148	0.056	2.348	0.144	1.531	-0.673
16	-1.114	-0.012	-1.102	-0.256	-0.858	-0.818	0.296	-0.807	0.307
17	-0.032	0.306	-0.338	0.060	-0.092	-0.818	-0.786	-0.172	-0.140
18	-1.060	0.034	-1.094	-0.976	-0.084	-0.818	0.242	-1.098	-0.038
19	2.447	1.587	0.860	2.282	0.165	2.825	0.378	3.327	0.880
20	0.806	1.259	-0.453	0.611	0.195	0.300	-0.506	1.038	0.232
21	3.653	1.987	1.666	3.905	-0.252	3.480	-0.173	2.998	-0.655
22	1.954	1.956	-0.002	2.096	-0.142	2.033	0.079	0.443	-1.511
23	3.000	0.720	2.280	1.636	1.364	2.874	-0.126	2.685	-0.315
24	1.724	0.702	1.022	0.964	0.760	1.394	-0.330	-0.220	-1.944
25	-1.076	-1.251	0.175	-1.782	0.706	-0.818	0.258	-0.809	0.267
26	-1.959	-1.443	-0.516	-1.900	-0.059	-0.818	1.141	-0.777	1.182
27	-0.745	-0.634	-0.111	-0.792	0.047	-0.793	-0.048	-0.796	-0.051
28	-1.444	-0.814	-0.630	-0.685	-0.759	-0.818	0.626	-0.795	0.649
29	-1.638	-0.888	-0.750	-0.770	-0.868	-0.818	0.820	-0.852	0.786
30	-1.569	-2.162	0.593	-1.331	-0.238	-0.818	0.751	-1.285	0.284
31	-1.097	-0.768	-0.329	-0.861	-0.236	-0.780	0.317	-0.760	0.337
32	-1.337	-0.688	-0.649	-0.961	-0.376	-0.818	0.519	-0.799	0.538
33	-1.745	-1.604	-0.141	-1.649	-0.096	-0.818	0.927	-0.784	0.961
34	-0.863	0.004	-0.867	-0.571	-0.292	-0.818	0.045	-0.817	0.046
35	-1.469	-1.109	-0.360	-0.462	-1.007	-0.818	0.651	-0.794	0.675
36	0.097	0.093	0.004	1.248	-1.151	-0.818	-0.915	-0.176	-0.273
37	-1.495	-0.487	-1.008	-0.752	-0.743	-0.817	0.678	-0.133	1.362
38	-0.215	-0.588	0.373	-0.635	0.420	-0.818	-0.603	-0.168	0.048
39	-1.481	-1.117	-0.364	-1.187	-0.294	-0.818	0.663	-0.794	0.687
40	-0.558	-1.447	0.889	-1.273	0.715	-0.818	-0.260	-0.836	-0.278
41	-1.310	-1.228	-0.082	-1.325	0.015	-0.818	0.492	-1.760	-0.450
42	-1.509	-0.113	-1.396	0.109	-1.618	-0.818	0.691	-0.792	0.717
43	1.081	-1.028	2.109	-0.333	1.414	-0.818	-1.899	0.990	-0.091
44	1.854	0.144	1.710	-0.471	2.325	-0.818	-2.672	-0.223	-2.077
45	0.921	-1.183	2.104	-0.504	1.425	-0.818	-1.739	-0.198	-1.119
46	-1.328	0.441	-1.769	-0.680	-0.648	-0.749	0.580	-0.697	0.631
47	-0.408	1.461	-1.869	-0.234	-0.174	-0.782	-0.374	-0.162	0.246
48	-0.553	0.560	-1.113	0.361	-0.914	0.100	0.653	-0.158	0.395

Table 5: The observed, the predicted activities (Log K_i), and residue according to MLR and MNLR for the 10 tested compounds (test set).

N°	Obs.	MLR		MNLR	
		Pred-test	Resid.	Pred-test	Resid.
1	-0.215	-0.740	-0.525	-0.400	-0.185
2	-1.252	-1.460	-0.208	-0.517	0.735
3	-0.149	-0.419	-0.270	-0.202	-0.053
4	1.279	0.737	-0.542	0.647	-0.632
5	-1.347	-1.530	-0.183	-0.766	0.581
6	1.380	-0.855	-2.235	-0.347	-1.727
7	3.653	3.291	-0.362	3.478	-0.175
8	0.556	2.472	1.916	1.220	0.664
9	-0.585	-0.765	-0.180	-1.611	-1.026
10	-0.260	1.170	1.430	0.242	0.502

3.6. Applicability Domain (AD)

The applicability domain (AD) of these models was evaluated by leverage analysis expressed as Williams plot (Figure7), in which the standardized residuals (r) and the leverage threshold value (h*= 0.553) were plotted. Any new value of predicted Log (K_i) data must be considered reliable only for those compounds that fall within this AD on which the model was constructed. From the figure 7, it is obvious that there is no response outlier both in the training set and no response outside in test set; and there is no compound have a standard deviation in the out of the ±x interval (x=2.5).

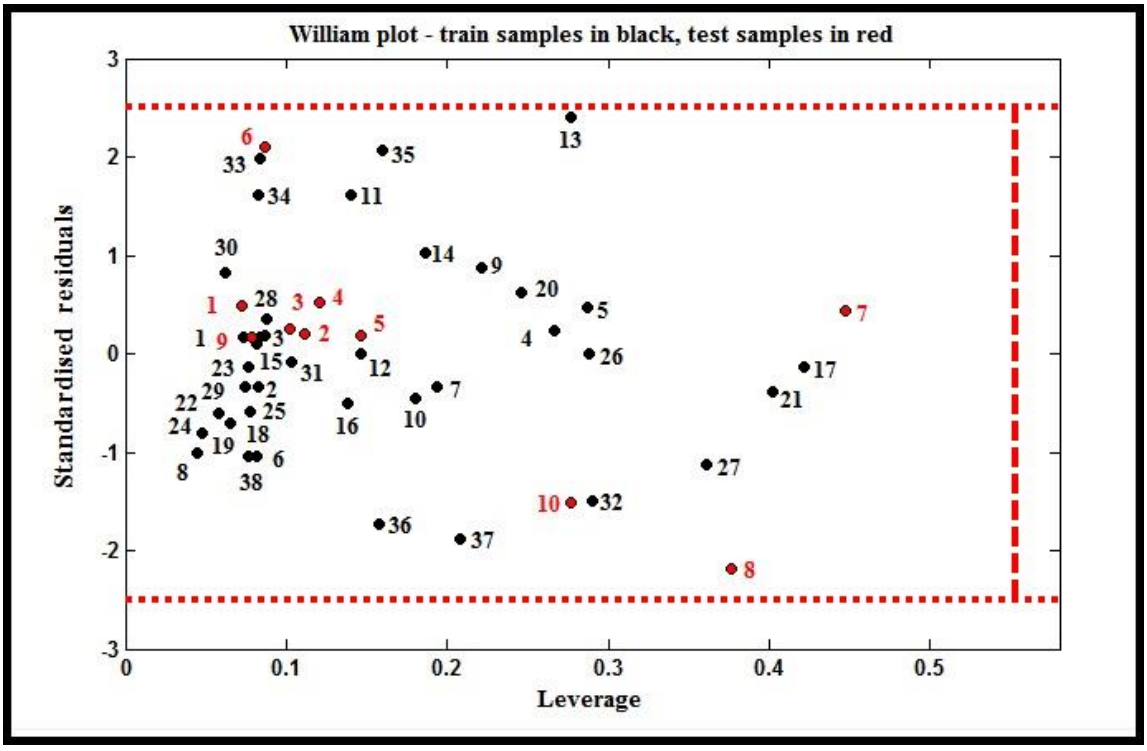


Figure 7: William's plot of standardized residual versus leverage (h*=0.553; x = ±2.5)

4. Conclusion

Multiple linear and non-linear regression and an artificial neural network were used to construct a quantitative structure–activity relation model for antagonists of *N*-Methyl-D-aspartate receptors and compared. The artificial neural network had substantially better predictive capability than the other two models, with greater predictive power. We established satisfactory relations between several descriptors and antagonist activity to the *N*-Methyl-D-aspartate receptor, with cross-validation. The results show that the model proposed in this paper can predict activity accurately and that the selected descriptors are pertinent. The accuracy and predictability of the proposed models were illustrated by comparison of the key statistical terms r or r^2 for the different models (Table 4). The applicability domain of the proposed models was investigated using William's plot to detect the subspace of chemical structures that can be predicted reliably by models. The proposed methods will reduce the time and cost of synthesis and determination of the activity of *N*-methyl-D-aspartate receptor antagonists based on dibenzo*[a,d]*cycloalkenimine derivatives. Furthermore, the descriptors are sufficiently rich in chemical, electronic and topological information to encode structural features that could be used with other descriptors in the development of predictive, quantitative structure–activity models.

References

[1] G. Flores, J.V. Negrete-Díaz, M. Carrión, Y. Andrade-Talavera, S.A. Bello, T.S. Sihra, A. Rodríguez-Moreno, and J.P.F. D'Mello, "Excitatory amino acids in neurological and neurodegenerative disorders – Chapter 25", J.P.F. D'Mello (Ed.), *Amino Acids in Human Nutrition and Health, Wallingford, UK*, **2012**, 427–453.

[2] E.H. Wong, J.A. Kemp, T. Priestley, A.R. Knight, G.N. Woodruff, L.L. Iversen, "The anticonvulsant MK-801 is a potent N-Methyl-D-aspartate antagonist", *Proceedings of the National Academy of Sciences, U.S.A.*, 83, **1986**, 7104–7108.

[3] J.E. Huettner and B.P. Bean, "Block of NMDA-activated current by the anticonvulsant MK-801: selective binding to open channels", *Proceedings of the National Academy of Sciences*, 85, **1988**, 1307–1311.

[4] S. McKay, C.P. Bengtson, H. Bading, D.J.A. Wyllie, and G.E. Hardingham, "Recovery of NMDA receptor currents from MK-801 blockade is accelerated by Mg^{2+} and memantine under conditions of agonist exposure", *Neuropharmacology*, 74, **2013**, 119–125.

[5] I.J. Reynolds, R.J. Miller, "Multiple sites for the regulation of the N-methyl-d-aspartate receptor", *Molecular Pharmacology*, 33, **1988**, 581–584.

[6] G.E. Hardingham, Y. Fukunaga, and H. Bading, "Extrasynaptic NMDARs oppose synaptic NMDARs by triggering CREB shut off and cell death pathways", *Nature Neuroscience*, 5, **2002**, 405–414.

[7] K. Bordji, J. Becerril-Ortega, O. Nicole, and A. Buisson, "Activation of extrasynaptic, but not synaptic, NMDA receptors modifies amyloid precursor protein expression pattern and increases amyloids production", *Journal of Neuroscience*, 30, **2010**, 15927–15942.

[8] O. Dick and H. Bading, "Synaptic activity and nuclear calcium signaling protect hippocampal neurons from death signal-associated nuclear translocation of FoxO3a induced by extrasynaptic NMDA receptors", *The Journal of Biological Chemistry*, 285, **2010**, 19354–19361.

[9] G.E. Hardingham and H. Bading, "Coupling of extrasynaptic NMDA receptors to a CREB shut-off pathway is developmentally regulated", *Biochimica et Biophysica Acta*, 1600, **2002**, 148–153.

[10] G.E. Hardingham and H. Bading, "Synaptic versus extrasynaptic NMDA receptor signaling, implications for neurodegenerative disorders", *Nature Reviews Neuroscience*, 11, **2010**, 682–696.

[11] A. Ivanov, C. Pellegrino, S. Rama, I. Dumalska, Y. Salyha, Y. Ben-Ari, and I. Medina, "Opposing role of synaptic and extrasynaptic NMDA receptors in regulation of the ERK activity in cultured rathippocampal neurons", *The Journal of Physiology*, 572, **2006**, 789–798.

[12] F. Leveille, F. El Gaamouch, E. Gouix, M. Lecocq, D. Lobner, O. Nicole, and A. Buisson, "Neuronal viability is controlled bya functional relation between synaptic and extrasynaptic NMDA receptors", *The FASEB Journal*, 22, **2008**, 4258–4271.

[13] A.J. Milnerwood, C.M. Gladding, M.A. Pouladi, A.M. Kaufman, R.M. Hines, J.D. Boyd, R.W.Y. Ko, O.C. Vasuta, R.K. Graham, M.R. Hayden, T.H. Murphy, and L.A. Raymond, "Early increase in extrasynaptic NMDA receptor signalling and expression contributes to phenotype onset in Huntington's disease in mice", *Neuron*, 65, **2010**, 178–190.

[14] S. Okamoto, M.A. Pouladi, M. Talantova, D. Yao1, P. Xia, D.E. Ehrnhoefer, R. Zaidi, A. Clemente, M. Kaul, R.K. Graham, D. Zhang, H.S.V. Chen, G. Tong, M.R. Hayden, and S.A. Lipton, "Balance between synaptic versus extrasynaptic NMDA receptor activity influences inclusions and neurotoxicity of mutant huntingtin", *Nature Medicine*, 15, **2009**, 1407–1413.

[15] S. Papadia, F.X. Soriano, F. Léveillé, M.A. Martel, K.A. Dakin, H.H. Hansen, A. Kaindl, M. Sifringer, J. Fowler, V. Stefovska, G. Mckenzie, M. Craigon, R. Corriveau, P. Ghazal, K. Horsburgh, B.A. Yankner, D.J.A. Wyllie, C. Ikonomidou, and G.E. Hardingham, "Synaptic NMDA receptor activity boosts intrinsic antioxidant defenses", *Nature Neuroscience*, 11, **2008**, 476–487.

[16] R.X. Soriano, P. Baxter, L.M. Murray, M.B. Sporn, T.H. Gillingwater, and G.E. Hardingham, "Transcriptional regulation of the AP-1 and Nrf2 target gene sulfiredoxin", *Molecules and Cells*, 27, **2009**, 279–282.
[17] W. Tu, X. Xu, L. Peng, X. Zhong, W. Zhang, M.M. Soundarapandian, C. Belal, M. Wang, N. Jia, W. Zhang, F. Lew, S.L.Chan, Y. Chen, and Y. Lu, "DAPK1 interaction with NMDA receptor NR2B subunits mediates brain damage in stroke", *Cell*, 140, **2010**, 222–234.
[18] A.S. Wahl, B. Buchthal, F. Rode, S.F. Bomholt, H.E. Freitag, G.E. Hardingham, L.C.B. Ronn, and H. Bading, "Hypoxic/ischemic conditions induce expression of the putative pro-death gene Clca1via activation of extrasynaptic N-methyl-d-aspartate receptors", *Journal of Neuroscience*, 158, **2009**, 344–352.
[19] S.J. Zhang, M.N. Steijaert, D.G. Lau, C. Schütz, D. Vivier, P. Descombes, and H. Bading, "Decoding NMDA receptor signaling: identification of genomic programs specifying neuronal survival and death", *Neuron*, 53, **2007**, 549–562.
[20] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, and V. Prachayasittikul, "A practical overview of quantitative structure–activity relationship", *EXCLI Journal*, 8, **2009**, 74–88.
[21] C. Nantasenamat, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, "Advances in computational methods to predict the biological activity of compounds", *Expert Opinion on Drug Discovery*, 5, **2010**, 633–654.
[22] W.J. Thompson, P.S. Anderson, S.F. Britcher, T.A. Lyle, J.E. Thies, C.A. Magill, S.L. Varga, J.E. Schwering, P.A. Lyle, M.E. Christy, B.E. Evans, M.D. Colton, M.K. Holloway, J.P. Springer, J.M. Hirshfield, R.G. Ball, J.S. Amato, R.D. Larsen, E.H.F. Wong, J.A. Kemp, M.D. Tricklebank, L. Singh, R. Oles, T. Priestly, G.R. Marshall, A.R. Knight, D.N. Middlemiss, G.N. Woodruff, and L.L. Iversen, "Synthesis and pharmacological evaluation of a series of dibenzo[a,d]cycloalkenimines as N-methyl-d-aspartate antagonists", *Journal of Medicinal Chemistry*, 33, **1990**, 789–808.
[23] C. Adamo, V. Barone, "A TDDFT study of the electronic spectrum of stetrazine in the gas-phase and in aqueous solution", *Chemical Physics Letters*, 330, **2000**, 152–160.
[24] M. Parac, S. Grimme, "Comparison of Multireference Møller−Plesset Theory and Time-Dependent Methods for the Calculation of Vertical Excitation Energies of Molecules", *The Journal of Physical Chemistry*, 106(29), **2003**, 6844–6850.
[25] Y. Yamaguchi, S. Yokoyama, S. Mashiko, "Strong coupling of the single excitations in the Q-like bands of phenylene-linked free-base and zinc bacteriochlorin dimers: A time-dependent density functional theory study", *The Journal of Chemical Physics*, 116(15), **2002**, 6541–6548.
[26] S. Chtita, M. Ghamali, M. Larif, A. Adad, R. Hmammouchi, M. Bouachrine, and T. Lakhlifi, "Prediction of biological activity of imidazo[1,2-a] pyrazine derivatives by combining DFT and QSAR results", *International Journal of Innovative Research in Science, Engineering and Technology*, 2 (12), **2013**, 7951–7962.
[27] L. Becker, K. Hinrichs, and U. Finke, "A New Algorithm for Computing Joins with Grid Files, In Proc. of the 9th International Conference on Data Engineering", *Vienna, Austria.*, **1993**, 190-197.
[28] ACDLABS 10, *Advanced Chemistry Development, Inc., Toronto, ON, Canada*, **2015**.
[29] S.J. Lee, J. Fink, A.B. Balantekin, M.R.A. Strayer, S. Umar, P.G. Reinhard, J.A. Maruhn, and W. Greiner, "Lee et al. reply", *Physical Review Letters*, 60(2), **1988**, 163.
[30] U. Sarkar, R. Parthasarathi, V. Subramanian, and P.K. Chattaraji, "Toxicity analysis of polychlorinated dibenzofurans through global", *Journal of Molecular Structure: THEOCHEM*, 758(2-3), **2006**, 119–125.
[31] ACD/ChemSketch Version 4.5, *User's Guide*.
[32] XLSTAT **2009** Add-in software, *XLSTAT Company*. www.xlstat.com

[33] M. Larif, A. Adad, R. Hmammouchi, A.I. Taghki, A. Soulaymani, A. Elmidaoui, M. Bouachrine, T. Lakhlifi, "Biological activities of triazine derivatives combining DFT and QSAR results", *Arabian Journal of Chemistry*, **2013**, in press.
[34] V.J. Zupan, J. Gasteiger, "Neural Networks for Chemists - An Introduction", *VCH Verlagsgesellschaft, Weinheim*, **1993**.
[35] D. Cherqaoui, D. Villemin, "Use of neural network to determine the boiling point of alkanes", *Journal of the Chemical Society*, 90, **1994**, 97–102.
[36] J.A. Freeman, D.M. Skapura, "Neural Networks Algorithms, Applications, and Programming Techniques", *Addition Wesley Publishing Company*, **1991**.
[37] B. Efron, "Estimating the error rates of a predictive rule: improvement on cross-validation", *Journal of the American Statistical Association*, 78, **1983**, 316–331.
[38] M.A. Efroymson, "Multiple regression analysis", A. Ralston, H.S. Wilf (Eds.), In *Mathematical Methods for Digital Computers*, Wiley, New York, **1960**.
[39] D.W. Osten, "Selection of optimal regression models via cross-validation", *Journal of Chemometrics*, 2, **1998**, 39–48.
[40] OECD, "Guidance Document on the Validation of QSAR Models", *Organization for Economic Co-Operation & Development, Paris, France*, **2007**.
[41] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, and P. Gramatica, "Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs", *Environmental Health Perspectives*, 111(10), **2003**, 1361–1375.
[42] P. Gramatica, "Principles of QSAR models validation: internal and external", *QSAR and Combinatorial Science*, 26(5), **2007**, 694–701.
[43] G.E. Batista and D.F. Silva, "How k-nearest neighbor parameters affect its performance", in *Proceedings of the Argentine Symposium on Artificial Intelligence, Instituto de Ciencias Matematicase de Computa cao, Sao Carlos, Brazil*, **2009**, 1–12.
[44] T.I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. van de Sandt, Weida Tong, G. Veith, and C. Yang, "Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships", **Alternatives to Laboratory Animals**, 33(2), **2005**, 1–19.
[45] S.S. So and W.G. Richards, "Application of neural networks: quantitative structure-activity relationships of the derivatives of 2,4-diamino-5-(substituted-benzyl) pyrimidines as DHFR inhibitors", *Journal of Medicinal Chemistry*, 35(17), **1992**, 3201–3207.
[46] T.A. Andrea and H. Kalayeh, "Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors", *Journal of Medicinal Chemistry*, 34(9), **1991**, 2824–2836.
[47] M. Elhallaoui, "Modélisatrice moléculaire et étude QSAR d'antagonistes non compétitifs du récepteur NMDA par les méthodes statistiques et le réseau de neurons", Doctoral thesis, Fez, *Morocco*, **2002**, 106.

Conclusion générale

L'objectif de cette thèse était de développer des modèles RQSA/RQSP fiables pour la prédiction de quelques propriétés et activités biologiques de certaines familles de molécules organiques hétérocycliques de structures variées appartenant aux trois familles chimiques : Acridine, Isatin et Dizocilpine (MK801). Un grand nombre de descripteurs moléculaires a été calculés (Descripteurs constitutionnels, électroniques, topologiques, géométriques, physico-chimiques, thermodynamiques, ...). Diverses méthodes statistiques ont été utilisées dans la construction de ces modèles (ACP, RLM, RNLM, PLS, ANN...). Les principales techniques de validation ont été utilisés (les tests statistiques standards, la validation interne, la validation externe, la Y-Randomisation, les domaines d'applicabilité...). Ces modèles ont été développés en accord avec les cinq principes de l'OCDE pour la validation des modèles RQSA/RQSP, (à savoir : une propriété ciblée définie avec un protocole expérimental identifié ; un algorithme sans équivoque ; un domaine d'applicabilité défini ; des mesures appropriées de la qualité d'ajustement, de robustesse et du pouvoir prédictif ; et si possible, une interprétation des mécanismes sous-jacents).

Dans ce cadre, nous avons présenté dans ce travail quatre applications, parmi d'autres, que nous avons accomplies au cours de ces quatre années :

Dans la première application nous avons établi des modèles reliant certains descripteurs moléculaires avec l'activité Antileishmanienne dans les deux stades de la Leishmaniose (promastigote et amastigote) pour des dérivés d'acridine. Dans cette étude deux méthodes de modélisation, la MLR et l'ANN, ont été utilisées, les résultats obtenus montrent que :

- L'ANN a une capacité prédictive nettement meilleure que la RLM, mais cette dernière donne des résultats facilement interprétables.
- Pour augmenter l'activité Antileishmanienne dans le stade promastigote, on augmente l'électrophilie et on diminue la ramification, la polarité et le nombre d'atomes d'hydrogène attachés à des hétéroatomes.
- Pour augmenter l'activité Antileishmanienne dans le stade amastigote, on augmente la solubilité, la polarité, et la ramification et on diminue l'électrophilie, la température critique et le nombre d'hétéroatomes attachés à des atomes d'hydrogène.

Le résultat le plus important de cette application est que nous avons proposé des modèles qui peuvent aider les chercheurs pour faire la synthèse de nouveaux dérivés d'acridine candidates d'avoir de bonnes activités Antileishmaniennes, et nous avons proposé de nouveaux composés avec des valeurs théoriques d'activités plus élevées par rapport aux composés existants.

Dans la deuxième application nous avons utilisé la RLM et la RNLM pour établir des modèles RQSP reliant des descripteurs électroniques et la constante d'association avec l'ADN pour des dérivés de la 9-aniliioacridine. Ces modèles peuvent prédire la constante d'association en calculant l'énergie E_{LUMO}, l'indice d'électrophilicité et l'énergie d'activation des dérivés de la 9-aniliioacridine. Les deux méthodes utilisées donnent de bonnes capacités prédictives. Le résultat le plus important est que nous avons pu concevoir et proposer de nouvelles molécules avec des valeurs théoriques plus élevées ou plus basses que celle des composés existants en se basant sur les résultats obtenus par la RLM.

Dans la troisième application nous avons utilisé la RLM, la PLS et la RNLM pour construire des modèles RQSA de dérivés d'Isatine pour leurs activités anticancéreuses contre les cellules U937. Les méthodes utilisées ont été comparées, et parmi eux, la MNLR basés sur des descripteurs proposés par le RLM (Stepwise et descendant) qui a présenté la meilleure capacité prédictive que les autres, même que la RLM Stepwise donne les résultats les plus simplement interprétables. Les mauvais résultats de la validation croisée pour la PLS montrent que le modèle proposé par cette méthode n'est pas capable de prédire l'activité de manière satisfaisante. Les résultats de cette application montrent que les modèles proposés peuvent prédire l'activité anticancéreuse avec une bonne précision et que les paramètres sélectionnés (E_{HOMO}, E_{LUMO} et Log P) sont pertinents. Nous avons conçu et suggéré, sur la base des résultats obtenus, quelques nouveaux composés ayant des activités anticancéreuses théoriques supérieures à celles des composés étudiés. Ces composés peuvent être synthétisés et testés dans le cadre des recherches des médicaments anticancéreux à base de l'Isatine.

Dans la quatrième application nous avons établi une RQSA de l'activité antagoniste vis-à-vis du récepteur N-méthyl-D-aspartate (NMDA) pour une série des dérivés de MK801. Dans cette application, la RLM, la RNLM et l'ANN ont été utilisés dans la construction des modèles. Nous avons établi des relations satisfaisantes entre plusieurs descripteurs et l'activité antagoniste du récepteur NMDA. L'ANN avait une capacité prédictive nettement meilleure que les deux autres modèles RQSA. Les résultats montrent que les modèles proposés dans cette application peuvent prédire l'activité avec une bonne précision et que les descripteurs sélectionnés sont pertinents.

D'après ces applications on peut conclure que :

L'emploi de modèles issus de l'approche combinée RQSA/RQSP-DFT utilisée dans ces travaux présente un fort potentiel. En effet, la prise en compte de divers descripteurs, dont les descripteurs issus de calculs de la chimique quantiques, en particulier ceux issus de la DFT conceptuelle, ont montré un grand intérêt dans l'ensemble des études que nous avons effectué.

Ces descripteurs contribuent, non seulement, à obtenir des modèles robustes et prédictifs mais ils rendent également ces derniers chimiquement plus interprétables.

La plupart des applications présentés dans cette thèse ont montré que la prédiction des activités/propriétés étudiées est faite principalement par une combinaison de ces descripteurs quantiques, à savoir, l'énergie E_{LUMO}, l'énergie E_{HOMO}, l'énergie totale E_T, le moment dipolaire μ, et l'indice d'électrophilie ω, ..., avec des descripteurs physico-chimiques, qui jouent un rôle très important pour l'identification de la similarité médicamenteuse plus précisément pour les médicaments délivrés par voie orale (selon les règles de Lipinski), à savoir l'indice d'hydrophobie log P (facteur de transport responsable du passage des molécules à travers les membranes), le nombre de sites accepteurs de liaison hydrogène NHA et le nombre de sites donneurs de liaison hydrogène NHD, (Les liaisons hydrogènes assurant l'affinité entre une drogue et son récepteur, L'établissement d'une liaison s'accompagne d'une diminution de l'énergie libre du système considéré ($\Delta G < 0$), et tout changement de cette énergie libre est lié à la constante d'équilibre de l'affinité drogue-récepteur par l'équation suivante: $\Delta G° = – RT \ln Keq$. Par exemple, à 37°C, une faible diminution de $\Delta G°$ de –2.7 kcal/mol change la valeur de la constante d'équilibre de 1 à 100 sachant que l'énergie libre d'une liaison hydrogène est entre $\Rightarrow \Delta G° = –3$ à –5 kcal/mol).

Les réseaux de neurones pour les différentes applications présentées dans cette thèse témoignent de l'existence d'une relation non-linéaire entre l'activité ou la propriété avec les structures moléculaires étudiées. Mais le problème majeur est qu'il n'existe pas une forme explicite expliquant et analysant la relation entre les entrées et les sorties pour les ANN. Cela cause des difficultés d'interprétation des résultats (modèles) obtenus par cette méthode.

La méthodologie basée sur la RLM, a été utilisée principalement dans la prédiction. Cette méthode permet d'extraire de manière efficace des modèles QSAR/QSPR transparents. Ces modèles sont à la fois fiables, explicatifs, prédictifs et interprétables en choisissant des descripteurs pertinents pour expliquer et interpréter l'activité/propriété des composés étudiés du point de vue statistique et chimique.

Les modèles proposés dans ces quatre applications donnent des orientations, dans le domaine de la recherche pharmaceutique, pour concevoir et synthétiser de nouvelles molécules susceptibles de devenir des médicaments.

Finalement, bien que les objectifs principaux de cette thèse aient été remplis, et afin de poursuivre notre chemin de recherche dans cette discipline nous prévoyons les perspectives, qui semblent divers, suivants :

- Nous avons l'intention de reprendre les mêmes bases de données et élaborer des modèles en utilisant d'autres méthodes telles que : les algorithmes génétiques (GA), les graphs machines, les séparateurs à vaste marge (SVM)).
- Nous avons déjà commencé des travaux avec l'utilisation d'autres méthodes de modélisation moléculaires, soit celles avec le développement de modèles QSAR-3D (les méthodes CoMFA et CoMSIA) avec l'utilisation du logiciel SYBYL ou aussi celles avec les études basées sur l'analyser l'affinité des molécules organiques pour une cible de nature protéique (Docking moléculaire).
- Au niveau expérimental, les perspectives pour le développement des modèles proposés seraient de poursuivre les efforts en vue de valider les modèles par la synthèse des nouvelles molécules proposées, ceci doit être fait en collaboration avec des autres laboratoires.

Annexe : La théorie de la fonctionnelle de la densité DFT

La théorie de la fonctionnelle de la densité (couramment abrégée par son acronyme anglais DFT) est une approche théorique basée sur des concepts de la chimie quantique qui connaît un essor spectaculaire depuis une vingtaine d'années, elle a été récompensée par un prix Nobel attribué à *Walter Kohn* en 1998. Elle a été l'objet de plus de 10 000 publications en 2012, dont une part importante concerne des applications en lien avec les expérimentateurs dans des domaines de recherche très variés.

Dans cette annexe, nous effectuons un rappel sur les notions théoriques dans le but de présenter la théorie de la fonctionnelle de la densité que nous avons utilisée, comme méthode de calculs, pour accéder à des propriétés physico-chimiques à l'échelle moléculaire (telles que : l'énergie totale, les énergies HOMO et LUMO, le moment dipolaire, la mollesse, la durée, l'électronégativité, l'indice d'électrophilicité...), dans la plupart de nos travaux présentés dans cette thèse.

1. Introduction

Traditionnellement, les molécules, en chimie, sont décrites comme un assemblage d'atomes joints par des liaisons. Cette description diffère de la conception de la mécanique quantique où ces mêmes molécules sont représentées comme un ensemble de noyaux et d'électrons soumis à des lois spécifiques qui prennent en compte les effets quantiques et non seulement les effets classiques. Même lorsqu'on considère les noyaux comme fixes, prévoir le comportement des électrons et toutes les propriétés qui en découlent constitue un problème compliqué à N corps, et par conséquent à N variables, qui passe en principe par la résolution de l'équation de Schrödinger **[1]** proposée par le physicien Erwin Schrödinger dans le cadre de la théorie quantique. Cependant, cette équation est généralement bien trop compliquée à résoudre exactement et le défi relevé par la chimie quantique depuis 1926, aidée par le développement des outils informatiques depuis 60 ans, est d'y trouver des solutions approchées.

La forme la plus courante de cette équation (indépendante du temps) s'écrit :

$$\hat{H}\Psi = E\Psi \qquad (1)$$

- L'écriture $\hat{H}\Psi$ doit être comprise comme l'application de l'opérateur $\hat{H}$ à la fonction Ψ, ce qui pourrait être noté plus explicitement par $\hat{H}(\Psi)$.
- Le terme $E\Psi$ désigne un simple produit entre le nombre réel E et la fonction Ψ.

Avec : E : L'énergie totale du système et $\hat{H}$ représente l'opérateur Hamiltonien du système ; ce dernier contient des termes relatifs à l'énergie cinétique des électrons et des noyaux

atomiques, ainsi que des termes décrivant l'interaction coulombienne électron-noyau, électron-électron, et noyau-noyau.

L'Hamiltonien $\hat{H}$ s'écrit sous la forme :

$$\hat{H} = \hat{T}_e + \hat{T}_N + \hat{V}_{NN} + \hat{V}_{ee} + \hat{V}_{eN} \qquad (2)$$

Avec :

$\hat{T}_e = -\sum_{k=1}^{n} \frac{\hbar^2}{2m_e} \nabla_k^2$ Opérateur énergie cinétique des n électrons.

$\hat{T}_N = -\sum_{A=1}^{N} \frac{h^2}{8\pi^2 M_A} \nabla_A^2$ Opérateur énergie cinétique des N noyaux.

$\hat{V}_{NN} = \sum_{A=1}^{N} \sum_{B>A}^{N} \frac{Z_A Z_B e^2}{4\pi\varepsilon_0 R_{AB}}$ Opérateur énergie de répulsion noyau- noyau.

$\hat{V}_{ee} = \sum_{k=1}^{n} \sum_{l>k}^{n} \frac{e^2}{4\pi\varepsilon_0 r_{kl}}$ Opérateur énergie de répulsion électron-électron.

$\hat{V}_{eN} = -\sum_{A=1}^{N} \sum_{k=1}^{n} \frac{Z_A e^2}{4\pi\varepsilon_0 r_{kA}}$ Opérateur énergie d'attraction électrons- noyaux.

Où : ħ représente la constante réduite de Planck ; h est la constante de Planck ; m_e est la masse de l'électron ; e est la charge de l'électron ; M_A est la masse du noyau A ; r_{KA} est la distance entre l'électron k et le noyau A ; R_{AB} est la distance entre les noyaux de l'atome A et de l'atome B dont les charges nucléaires sont respectivement Z_A et Z_B ; ∇_k^2 est le Laplacien du $K^{éme}$ électron défini de la manière suivante :

$$\nabla_k^2 = \frac{\partial^2}{\partial x_k^2} + \frac{\partial^2}{\partial y_k^2} + \frac{\partial^2}{\partial z_k^2} \qquad (3)$$

Pour simplifier l'écriture d'une telle équation, on adaptera le système d'unités atomiques de telle sorte que : ħ =1, me = 1 ; e = 1 et $4\pi\varepsilon_0 = 1$.

D'où l'expression de l'Hamiltonien devient :

$$\hat{H} = -\sum_{k=1}^{n} \frac{1}{2} \nabla_k^2 + -\sum_{A=1}^{N} \frac{1}{2M_A} \nabla_A^2 + \sum_{A=1}^{N} \sum_{B>A}^{N} \frac{Z_A Z_B}{R_{AB}} + \sum_{k=1}^{n} \sum_{l>k}^{n} \frac{1}{r_{kl}} + -\sum_{A=1}^{N} \sum_{k=1}^{n} \frac{Z_A}{r_{kA}} \quad (4)$$

L'équation de Schrödinger ainsi décrite correspond mathématiquement à une « équation aux valeurs propres » et conduit à une infinité de solutions appelées états quantiques. À chaque état quantique correspond une fonction d'onde Ψ et une énergie associée E. L'état le plus stable (de plus basse énergie) s'appelle l'état fondamental, et les autres états ayant des énergies plus grandes sont les états excités.

Cette équation décrit le mouvement des particules à travers des termes cinétiques et les interactions des noyaux et des électrons entre eux. Cette équation est en pratique très difficile à résoudre de façon exacte (sauf pour l'ion moléculaire H_2^+, ce qui est extrêmement limitant), d'autant plus lorsque le nombre d'électrons du système augmente, et le défi relevé par la chimie quantique a été de trouver des résolutions approchées de cette équation.

En raison de la grande différence entre la masse d'un électron (me) et la masse d'un noyau (M), ce dernier est considéré comme immobile, et par conséquent son énergie cinétique est alors nulle et l'énergie d'interaction noyau-noyau $\widehat{V}_{NN}$ est une constante. Ceci est connu sous le terme d'approximation de Born-Oppenheimer **[2]**.

Cependant l'équation de Schrödinger à n électrons et à N noyaux peut ainsi être séparée en deux parties, une partie nucléaire et une partie électronique. Puisque la fonction d'onde nucléaire dépend uniquement des coordonnées des noyaux, la fonction d'onde électronique sera alors calculée pour une position donnée des noyaux et dépendra de paramètres liés aux coordonnées nucléaires. La fonction d'onde du système, solution de l'équation de Schrödinger dans l'approximation de Born et Oppenheimer, peut donc s'écrire sous la forme d'un produit de deux fonctions :

$$\Psi(r, R) = \Psi_R(r).\Phi(R) \qquad (5)$$

Où $\Phi(R)$ représente la fonction d'onde nucléaire, $\Psi_R(r)$ est la fonction d'onde électronique correspondant à un jeu de positions R des noyaux figés, r et R étant respectivement les positions des électrons et des noyaux.

Malgré cela, l'équation de Schrödinger électronique reste insoluble analytiquement pour les systèmes à plus d'un seul électron. Deux grandes familles de méthodes de résolution approchées utilisant la puissance de calcul croissante des ordinateurs depuis les années 1950 ont été développées en chimie quantique :

- Les méthodes basées sur des approximations directes de la fonction d'onde **[3]**.
- Les méthodes contournant le calcul de la fonction d'onde à l'aide de fonctionnelles de la densité électronique **[4]**.

Ces méthodes sont aujourd'hui disponibles dans un grand nombre de logiciels de chimie quantique tels que le logiciel Gaussian que nous avons utilisé dans tous nos calculs.

2. Les méthodes basées sur des approximations directes de la fonction d'onde

Ce sont les plus anciennes, reposent sur le calcul de propriétés à partir de l'orbitale moléculaire poly-électronique ψ. Cette orbitale ne peut pas être construite de façon exacte, il est alors nécessaire d'utiliser plusieurs niveaux d'approximation pour y parvenir.

La première approximation appelée « approximation orbitale », introduite par Hartree en 1928 **[5]**, permet de calculer l'orbitale moléculaire poly-électronique comme étant le produit antisymétrique d'orbitales moléculaires mono-électroniques ϕ_i:

$$\Psi = \phi_1 * \phi_2 * \ldots * \phi_N \qquad (6)$$

Où ϕ_i est la fonction d'onde, appelées « orbitale », décrivant l'électron i.

En 1930, Fock démontre que la méthode de Hartree ne respecte pas le principe d'antisymétrie de la fonction d'onde **[6]**. En effet, d'après le principe d'exclusion de Pauli, deux électrons ne peuvent pas être simultanément dans le même état quantique.

La méthode de Hartree-Fock **[7]** permet une résolution approchée de l'équation de Schrödinger d'un système quantique à n électrons et N noyaux dans laquelle la fonction d'onde poly-électronique est écrite sous la forme d'un déterminant de Slater composé de spin orbitale mono-électronique qui respecte l'antisymétrie de la fonction d'onde :

$$\Psi(X_1, X_{2,\dots} X_{2n}) = \frac{1}{\sqrt{2n!}} \begin{bmatrix} \phi_1(X_1) & \phi_2(X_1) & \dots & \phi_{2n}(X_1) \\ \phi_1(X_2) & \phi_2(X_2) & \cdots & \phi_{2n}(X_2) \\ \vdots & \vdots & & \vdots \\ \phi_1(X_{2n}) & \phi_2(X_{2n}) & \cdots & \phi_{2n}(X_{2n}) \end{bmatrix} \qquad (7)$$

Les variables X_i représentent ici les coordonnées d'espace et de spin ; $\frac{1}{\sqrt{2n!}}$ est le facteur de normalisation ; n étant le nombre d'électrons.

- L'inversion de deux électrons correspond à la permutation de deux lignes (ou de deux colonnes), ce qui a pour effet de changer le signe du déterminant.
- Les spin-orbitales ϕ_i doivent être différentes les unes des autres, sinon le déterminant de Slater (7) s'annule.

Les orbitales sont déterminées en minimisant l'énergie associée à la fonction d'onde Ψ, ce qui conduit à une équation similaire à celle de Schrödinger mais pour un seul électron. Ainsi, l'orbitale i, notée ϕ_i, et son énergie εi associée sont déterminées par $\hat{F}\phi_i = \varepsilon_i\phi_i$; où $\hat{F}$ est l'opérateur de Hartree-Fock.

L'approximation Hartree-Fock revient ainsi à considérer que les électrons sont indépendants et à ne prendre en compte l'interaction électron-électron qu'à travers un champ moyen de répulsion coulombienne généré par les autres électrons. Le champ moyen dans l'opérateur $\hat{F}$ dépend de toutes les orbitales et doit être déterminé de façon itérative. On parle alors de champ auto-cohérent (en anglais SCF pour « self-consistent field »).

Les différentes orbitales moléculaires ϕ_i ne peuvent plus être construites strictement, c'est pour cela que l'on a recours à l'approximation de combinaison linéaire des orbitales atomiques (en anglais, Linear Combination of Atomic Orbital (LCAO)), proposée par Mulliken en 1941 **[8]**, qui permet de construire l'orbitale moléculaire ϕ_i à partir d'une combinaison linéaire d'orbitales atomiques mono-électroniques. A leur tour, les différentes orbitales atomiques mono-électroniques seront approximées afin de faciliter les calculs. Cette dernière approximation va permettre plus ou moins de bien décrire les orbitales atomiques mono-électroniques, et l'on parlera alors de bases de fonctions atomiques. L'utilisation d'une

base ayant un nombre fini de fonctions atomiques introduit une approximation supplémentaire, mais dont l'effet peut être systématiquement réduit en augmentant le nombre de fonctions afin de s'approcher de la limite d'une base complète (ce qui demande en principe une infinité de fonctions). Un grand nombre de bases atomiques de différentes « tailles » ont été développées et sont disponibles dans les logiciels.

Dans nos calculs nous avons utilisé la base 6-31 G* noté aussi 6-31 G (d) dont le 6 représente le nombre de fonctions atomiques (ici de type gaussiennes) utilisées pour représenter les orbitales de cœur, alors que le chiffre 31 représente les deux fonctions des orbitales atomiques de valences (de type gaussiennes) et la lettre d (ou aussi le symbole *) représente la fonction de polarisation utilisée pour avoir une flexibilité de calculs additionnelle, et qui signifie ainsi qu'un jeu de fonction d a été ajouté à tous les atomes (sauf H) dans la molécule.

3. Les méthodes contournant le calcul de la fonction d'onde à l'aide de la DFT

Historiquement, les premiers à avoir exprimé l'énergie en fonction de la densité furent Thomas (1927), Fermi (1927, 1928) et Dirac (1930) sur le modèle du gaz uniforme d'électrons qui englobe les noyaux de la molécule **[9-11]**. Cependant, la DFT a véritablement débuté avec les théorèmes fondamentaux de Hohenberg et Kohn.

En 1964 **[12]**, Hohenberg et Kohn ont repris la théorie de Thomas-Fermi et ont montré qu'il existe une fonctionnelle de l'énergie E[ρ(r)], associée à un principe variationnel, ce qui a permis de relancer les bases de la théorie de la fonctionnelle de la densité. Des applications pratiques ont ensuite été possibles grâce aux travaux de Kohn et Sham **[13]** qui ont proposé, en 1965, un ensemble d'équations mono-électroniques analogues aux équations de Hartree-Fock à partir desquelles il est en principe possible d'obtenir la densité électronique d'un système et donc son énergie totale.

Le but des méthodes DFT est de déterminer des fonctionnelles qui permettent de relier la densité électronique à l'énergie. Au lieu d'une description explicite de chaque électron à travers la fonction d'onde, la DFT choisit de se focaliser sur une grandeur plus simple, la densité électronique, que l'on peut directement représenter en trois dimensions et donc interpréter avec une vision classique.

4. Densité, fonctionnelle et théorie

4.1. Densité électronique

Par définition, la densité électronique ρ(r), est le nombre d'électronspar élément de volume autour de la position r **[14]**. C'est une observable (c'est-à-dire une grandeur mesurable, contrairement à la position de l'électron), dont l'intégrale sur tout l'espace est

égale au nombre d'électrons N. Ainsi, la valeur de ρ pour un volume de l'espace donné représente la probabilité de présence d'un électron dans ce volume. Elle contient en effet toute l'information nécessaire pour établir la structure atomique telle qu'on la dessine usuellement :

- La position des atomes : les électrons, chargés négativement, se concentrent autour des noyaux, chargés positivement, ce qui se traduit par un maximum local de la densité électronique au voisinage des noyaux.
- Le type d'atomes : la taille des maximas observés est directement reliée au type de noyau.

Il est donc possible de reconstruire complètement l'assemblage d'atomes observé uniquement à partir de la densité électronique.

La théorie de la fonctionnelle de la densité se propose d'exprimer les propriétés électroniques des systèmes à partir de leur densité par le biais d'une fonctionnelle.

4.2. Fonctionnelle

C'une fonction qui prend en argument une autre fonction F → F[f]. Prenons par exemple le nombre d'électrons N d'un système ; il dépend de la densité électronique ρ, qui est une fonction de la position. Il existe donc une fonctionnelle qui relie la densité de chaque système au nombre d'électrons correspondant. Dans ce cas, la fonctionnelle est très simple puisqu'il suffit juste d'intégrer la densité sur tout l'espace pour obtenir N.

En DFT, on cherche à exprimer une partie de l'énergie électronique comme une fonctionnelle de la densité. Cette approche est fondée sur une théorie exacte, bien qu'en pratique des approximations soient nécessaires.

4.3. Théorie

Dans les années 1960, Pierre Hohenberg et Walter Kohn ont démontré le théorème du même nom **[15]**. Ce théorème constitue vraiment le cœur de la théorie de la fonctionnelle de la densité. En effet, il prouve qu'il est possible d'obtenir, en principe, l'énergie de l'état fondamental (c'est-à-dire l'état le plus bas en énergie) d'un système d'électrons en ne connaissant que la densité électronique. On dispose donc de toute l'information nécessaire pour décrire le système et reconstruire l'équation de Schrödinger, puisque cette analyse permet d'obtenir d'une part les positions des noyaux et leur nature, et d'autre part le nombre d'électrons par intégration de la densité.

Le premier théorème de Hohenberg-Kohn montre que l'on peut exprimer l'énergie de l'état fondamental E_0, comme une fonctionnelle F, qui dépend uniquement de la densité électronique de cet étatρ_0, à laquelle on ajoute simplement un terme qui dépend du potentiel d'interaction entre électrons et noyaux v_{ne} :

$$E_0 = F[\rho_0] + \int v_{ne}(r)\,\rho_0(r)\,d^3r \qquad (8)$$

Si on remplace dans la partie droite de cette équation la densité ρ_0, inconnue *a priori*, par une densité ρ choisie arbitrairement, on obtient une énergie E plus élevée que E_0. Une procédure d'optimisation variationnel de l'énergie en fonction de la densité électronique permet donc d'obtenir $\rho_0$comme étant la densité minimisant l'énergie, ainsi que l'énergie de l'état fondamental correspondant E_0. Autrement dit, l'énergie calculée pour un système est minimale si et seulement si la densité électronique est celle de l'état fondamental. C'est le second théorème de Hohenberg-Kohn **[16]**.

Cependant ces deux théorèmes démontrent juste que cette approche est fondée, mais ne donnent pas de méthodologie pour résoudre les équations. Afin de résoudre ce problème, Walter Kohn et Lu Jeu Sham **[17]** ont proposé la décomposition des termes inconnus de la fonctionnelle $F[\rho]$ en deux parties. La première correspond au comportement des électrons dans un système sans interactions, notée $U[\rho]$. Il est possible d'obtenir une expression analytique exacte de ces termes. La seconde partie correspond donc aux phénomènes non pris en compte, c'est-à-dire les termes d'échange-corrélation, dont l'expression est inconnue. Les équations résultantes, dites de Kohn-Sham, sont intéressantes car elles permettent de réduire la partie inconnue à la connaissance de l'énergie d'échange et de corrélation **[18-20]** notée $E_{xc}[\rho]$. Comme pour Hartree-Fock, on est donc ramené à un problème mono-électronique.

4.3.1. La partie exacte

Lorsqu'on utilise la densité pour calculer l'énergie électronique, une des composantes de cette énergie est connue et s'exprime naturellement en fonction de ρ. Il s'agit de la composante classique des interactions entre électrons, c'est-à-dire le terme de répulsion électrostatique U. Par analogie avec la loi de Coulomb pour la répulsion de deux charges ponctuelles, le terme de répulsion U entre deux densités de charge s'écrit :

$$U[\rho] = \frac{1}{2}\iint \frac{\rho(r_1)\rho(r_2)}{r_{12}} d^3r_1 d^3r_2 \qquad (9)$$

Où r_1 et r_2 sont les positions *des deux électrons, et r_{12} leur distance respective.*

Cette énergie ainsi que toutes les grandeurs utilisées par la suite sont données en unités atomiques. Le choix de traiter de façon exacte cette partie de l'énergie électronique découle de la décision de décrire également de façon exacte l'interaction électrostatique attractive entre noyaux et électrons $\int v_{ne}(r)\rho(r)d^3r$.

En effet, ces termes électrostatiques sont de signes opposés et jouent un rôle fondamental dans la description du système. Une mauvaise description de l'interaction électrostatique pourrait briser l'équilibre recherché de ces deux contributions.

Le cas de l'interaction électrostatique étant réglé, il reste à trouver des approximations, dépendantes de ρ, pour décrire le mouvement des électrons et les effets quantiques.

Le mouvement des électrons, c'est-à-dire la partie cinétique de leur énergie, est particulièrement difficile à décrire à partir de leur densité. En effet, en mécanique quantique, les particules élémentaires sont classées en deux grandes familles selon leur spin, qui est une de leurs propriétés au même titre que leur charge ou leur masse :

Les bosons ont un spin nul ou entier, ont tendance à se trouver dans le même état quantique. Par conséquent, on peut aisément exprimer l'énergie cinétique pour les bosons en fonction de la densité grâce à l'occupation multiple du même état.

Les fermions (dont les électrons font partie) ont un spin demi-entier (c'est-à-dire 1/2, 3/2, 5/2...), doivent satisfaire le principe d'exclusion de Pauli qui les oblige à occuper des états quantiques différents ce qui rend les choses beaucoup plus complexes à cause des différents états à considérer.

À ce jour, il n'existe pas de méthode capable d'extraire avec précision, pour n'importe quel système, le mouvement des électrons à partir de la densité de façon générale. Cependant, des efforts continuent à être faits en ce sens.

La méthode de Kohn et Sham permet de contourner ce problème en introduisant un système auxiliaire dont l'énergie cinétique est calculable explicitement, tout en prenant en compte le principe de Pauli. Pour cela, on imagine un système modèle de N fermions qui n'interagissent pas entre eux, mais qui évoluent dans un potentiel effectif qui permet de garantir que le système a la même densité que le système réel de N électrons. Ces fermions possèdent le même spin que les électrons et respectent le principe d'exclusion de Pauli. Cependant, en éliminant l'interaction, chaque fermion peut être traité indépendamment, ce qui permet de se ramener à un problème à un corps. L'énergie cinétique de ce système fictif diffère de celle du système réel, mais a l'avantage de pouvoir s'exprimer directement à partir d'orbitales ϕi, dites de Kohn-Sham, que l'on peut directement relier à la densité.

L'énergie cinétique de ce système fictif, qui peut être calculée exactement, est alors :

$$T = \frac{1}{2}\sum_{i=1}^{N}|\nabla\varphi_i|^2 d^3r \qquad (10)$$

En revanche, les orbitales de Kohn-Sham n'ont pas de signification physique et seule la somme de leurs carrés est reliée en tout point à la densité électronique du système :

$$\rho(r) = \sum_i|\varphi_i(r)|^2 \qquad (11)$$

Le défaut majeur réside dans le fait qu'on ne connaît pas la fonctionnelle d'échange-corrélation. Il est donc nécessaire d'en faire une approximation. La variété de ces

fonctionnelles est telle qu'un traitement exhaustif dépasse le cadre de cette annexe. Aussi, nous nous limitons à la présentation des fonctionnelles les plus utilisées.

4.3.2. La partie approchée

Tout le raisonnement précédent, repose sur un formalisme exact, utilisé pour définir une énergie cinétique qui n'est pas celle du vrai système, mais qui permet d'introduire le principe de Pauli. Ceci permet d'alléger la difficulté de trouver des approximations pour la partie restante de l'énergie, privée de la partie cinétique et la partie électrostatique. Cette énergie est appelée énergie d'échange et de corrélation par analogie avec des définitions existantes (mais sans se recouvrir exactement avec elles) :

$$E_{xc}[\rho] = F[\rho] - U[\rho] - T[\rho] \qquad (12)$$

Elle doit donc décrire la partie quantique de l'interaction entre les électrons et la différence entre l'énergie cinétique du vrai système et celle du système fictif. L'expression de l'énergie dans l'approche de Kohn-Sham est par conséquent :

$$E = T[\rho] + U[\rho] + E_{xc}[\rho] + \int v_{ne}(r)\rho(r)d^3r \qquad (13)$$

À ce stade, il reste donc à définir une fonctionnelle permettant d'obtenir la quantité E_{xc} à partir de la fonction $\rho(r)$. C'est principalement au choix de cette fonctionnelle que se heurte l'utilisateur de la DFT, il faut en effet déterminer qu'elle est la plus pertinente à appliquer au système et à la grandeur à évaluer.

Approximation de la densité locale LDA

La fonctionnelle la plus simple à laquelle on puisse penser consiste à intégrer la densité sur tout l'espace. Cependant, cette intégrale donne uniquement le nombre d'électrons du système, ce qui est inutile dans le cas présent. Une approche légèrement plus complexe consiste à intégrer une fonction de la valeur de la densité en chaque point de l'espace $\int f(\rho(r))d^3r$. Une telle fonction f ne dépend plus de la totalité de la densité mais de sa valeur en un seul point de l'espace. Cette restriction sur le type de fonctionnelle constitue la classe d'approximations que l'on appelle locales ou LDA (Local Density Approximation).

Cette approche est fondée sur le modèle du gaz uniforme d'électron et constitue l'approche la plus simple pour exprimer l'énergie d'échange-corrélation. Celle-ci est décrite comme :

$$E_{xc}(\rho) = \int \rho(r)\varepsilon_{xc}(\rho)dr$$

Où $\varepsilon_{xc}(\rho)$ désigne l'énergie d'échange-corrélation pour une particule d'un gaz homogène d'électron qui peut être décomposée en une contribution d'échange $\varepsilon_x(\rho)$ et de corrélation $\varepsilon_c(\rho)$:

$$E_{xc}(\rho) = \varepsilon_x(\rho) + \varepsilon_c(\rho)$$

La contribution d'échange est déterminée analytiquement pour le gaz homogène :

$$\varepsilon_x[\rho] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{\frac{1}{3}}(r)^{\frac{1}{3}}$$

Enfin, Ceperley et Aller **[21]**, et plus récemment Ortiz et Ballone **[22]**, ont déterminé numériquement la contribution des corrélations par des simulations de type Monte-Carlo quantique. La recherche de fonctions analytiques se rapprochant le plus possible de ces résultats conduit à l'élaboration de diverses fonctionnelles au succès plus ou moins grand.

Une fois la fonction f définie, il suffit de connaître la valeur de la fonction $\rho(r)$ en chaque point r de l'espace pour calculer l'énergie. Cependant, on peut remarquer que le terme d'énergie électrostatique U ne peut pas s'écrire sous cette forme, ce qui est problématique.

Approximation des gradients généralisée GGA

L'approche LDA se fondait sur le modèle du gaz d'électrons et supposait donc une densité électronique uniforme. Cependant les systèmes atomiques ou moléculaires sont le plus souvent très différents d'un gaz d'électrons homogène et, de manière plus générale, on peut considérer que tous les systèmes réels sont inhomogènes c'est-à-dire que la densité électronique possède une variation spatiale. On peut généraliser l'approximation LDA en faisant dépendre la fonction f, non pas uniquement de la valeur de la densité $\rho(r)$ au point r, mais aussi au voisinage de r, cette variation de la densité exprime les énergies d'échanges et de corrélation en fonction de la densité mais également son gradient. En pratique, on utilise pour cela la dérivée de la densité par rapport à la position, au point r. On parle alors d'approximations semi-locales (parfois aussi appelées méthodes non locales), ou GGA (Generalized Gradient Approximation).

En général, l'énergie d'échange-corrélation est définie dans l'approximation GGA comme :

$$E_{xc}(\rho) = \int \rho(r)\varepsilon_{xc}(\rho, \nabla\rho)dr$$

L'introduction explicite de termes gradient de la densité $\nabla\rho$ améliore très sensiblement les performances de la méthode. À titre d'exemple, la GGA estime correctement les énergies de liaison dans les molécules alors que la LDA les surestime.

L'énergie d'échange associée à une fonctionnelle GGA s'exprime sous la forme :

$$E_x(\rho) = E_x^{LDA} - \int F(S)\rho^{4/3}(r)dr;$$

$$\text{Avec : } S(r) = \frac{|\nabla\rho(r)|}{\rho^{4/3}(r)}$$

S(r) est une quantité sans dimension appelée gradient de densité réduit. Deux classes de fonctions F (S) sont couramment utilisées pour l'échange ajustées sur des résultats expérimentaux.

Fonctions ajustées sur l'énergie d'échange des gaz rares

Une des plus connues est la fonctionnelle, proposée par Becke **[23]** en 1988, qui représente la correction de gradient à apporter à l'énergie d'échange LDA :

$$F^{B88} = \frac{\beta S^2}{1 + 6\beta S \sinh^{-1} S} \qquad \text{avec} \qquad \beta = 0{,}0042$$

Développement en fonctions rationnelles des puissances de S(r)

En 1986, Perdew a proposé le développement suivant à la fonctionnelle précédente **[24]** :

$$F^{B86} = [1 + 1.296(S/P)^2 + 14(S/P)^4 + 0.2(S/P)^6]^{1/15} \qquad \text{avec} \qquad p = (24\pi^2)^{\frac{1}{3}}$$

La fonctionnelle GGA combinant les deux corrections ci-dessus, celles de Becke pour l'échange et de Perdew pour la corrélation, constitue la fonctionnelle BP86. Il en existe d'autres, dont celle de Perdew-Wang (PW91).

Expression de l'énergie de corrélation proposée par Lee, Yang et Parr

Ces trois auteurs **[25]** adoptent une forme de $\varepsilon_c$basée sur la connaissance de l'énergie de corrélation de l'atome d'hélium. La fonctionnelle LYP est dérivée de la fonction d'onde corrélée réelle de ce système à deux électrons et non pas du modèle du gaz uniforme d'électrons. Elle contient également des termes de gradient de la densité électronique. Son expression comporte quatre paramètres « ajustés » pour reproduire l'énergie de corrélation de l'atome d'hélium.

La combinaison de l'énergie d'échange de Becke et de l'énergie de corrélation de LYP conduit à la fonctionnelle BLYP.

Fonctionnelles hybrides

Dans ces fonctionnelles, on introduit un certain pourcentage de l'échange « exact » calculable en théorie de HF. Becke **[26]** a proposé l'expression suivante de l'énergie d'échange et de corrélation, qualifiée d'hybride car elle prend en compte l'énergie d'échange exacte HF ainsi que l'énergie d'échange et de corrélation DFT :

$$E_x^{hybride} = C^{HF} E_x^{HF} + C^{DFT} E_{xc}^{DFT}$$

Les paramètres C^{HF} et C^{DFT} étant des constantes à déterminer.

Ainsi, par exemple, une fonctionnelle de type Becke à trois paramètres peut s'écrire :

$$E_{xc}^{B3LYP} = E_x^{LDA} + C_0(E_x^{HF} + E_x^{LDA}) + C_x\Delta E_x^{888} + E_c^{VWN3} + C_c(E_c^{LYP} - E_c^{VWN3})$$

Becke a déterminé les valeurs des trois paramètres de façon à reproduire au mieux les énergies de liaison d'une série de molécules de référence et propose les valeurs suivantes : $C_0 = 0{,}20$, $C_x = 0{,}72$ et $C_c = 0{,}81$.

Ces fonctionnelles sont notées B3XXX (Becke3LYP ou B3LYP, B3P86, B3PW91). La fonctionnelle B3LYP **[27]** est actuellement l'une des plus utilisées, et c'est la fonctionnelle que nous avons utilisée dans tous nos travaux de cette thèse.

5. Domaines d'application

L'approche de Kohn-Sham est générale appliquée à l'étude des propriétés électroniques de la matière dans différents états d'agrégation (de la molécule en phase gazeuse au cristal, en passant par les macromolécules biologiques, les polymères et les molécules en solution).

Elle est utilisée pour calculer l'énergie et la densité électronique de l'état fondamental, ainsi que des propriétés dépendant directement de celle-ci telles que le moment dipolaire. Elle est utilisée aussi pour déterminer les énergies des orbitales HOMO et LUMO.

Un certain nombre de propriétés supplémentaires sont accessibles en faisant varier le potentiel d'interaction noyau-électron v_{ne}. En particulier, des propriétés telles que la géométrie d'équilibre et les spectres vibrationnels peuvent être déterminées à partir des variations de l'énergie de l'état fondamental avec le déplacement des noyaux. De la même façon, on peut introduire un champ électrique externe et obtenir des polarisabilités et des hyper-polarisabilités qui sont importantes en optique linéaire et non linéaire. En ajoutant un peu de complexité, on peut généraliser la théorie pour permettre également de calculer les propriétés dans des champs magnétiques. Jusqu'à présent, toutes ces propriétés concernaient l'état fondamental du système ; cependant, il est possible d'accéder aux états électroniques excités en utilisant une extension de la théorie appelée DFT dépendante du temps.

La DFT permet aujourd'hui de clarifier un grand nombre de phénomènes chimiques dans des domaines très variés, allant de la production d'énergie (photovoltaïque, piles à combustible, nucléaire) à la géochimie (minéraux, noyau de la Terre), en passant par la biologie (enzymes, protéines, ADN). Enfin, mentionnons que le prix Nobel de chimie 2013, *Ariel Warshel*, a largement utilisé la DFT pour ses travaux théoriques dans le domaine de la modélisation de la catalyse enzymatique.

L'obtention des propriétés présentées jusqu'ici est fondée théoriquement. En pratique, on arrive également à utiliser avec un certain succès la méthode Kohn-Sham pour obtenir des propriétés supplémentaires sans avoir une véritable justification théorique, comme par exemple pour caractériser la diffusion inélastique du rayonnement d'un photon par les

électrons. En fonction des propriétés calculées et des systèmes considérés, il peut être intéressant de choisir une fonctionnelle plutôt qu'une autre, surtout dans le cas de fonctionnelles qui ont été paramétrées pour traiter un problème bien spécifique. Dans l'ensemble, les résultats obtenus sont satisfaisants et permettent une analyse qualitative, voire quantitative des phénomènes. Cependant, il existe un certain nombre de cas pathologiques où ces approximations sont incapables de décrire les phénomènes observés.

6. La précision des approximations DFT

De façon générale, les méthodes de chimie quantique ne sont pas capables d'estimer les erreurs introduites par les différentes approximations utilisées, et donc de donner des barres d'erreur sur les résultats obtenus. De par la procédure de minimisation utilisée dans la méthode de Kohn-Sham, on sait que l'énergie obtenue est nécessairement supérieure à la vraie énergie du système. On a donc une borne supérieure pour l'énergie. Cependant, on ne possède pas de borne inférieure.

Afin de juger la qualité des approximations proposées, il faut donc déterminer des données de référence auxquelles on pourra comparer les résultats obtenus par les méthodes approchées. Deux types de référence peuvent être utilisés :

- Des données théoriques calculées sur des petits systèmes par d'autres méthodes de chimie quantique considérées comme très précises.
- Des données expérimentales, bien que les mesures puissent être affectées par des perturbations extérieures qui ne sont pas prises en compte dans les calculs et risquent de fausser l'analyse.

7. Le futur de la DFT

Grâce au développement de fonctionnelles précises, la DFT connaît un succès considérable depuis une vingtaine d'années. La combinaison de sa facilité d'utilisation (elle ne nécessite que la géométrie de la molécule et le choix de la fonctionnelle), de la rapidité des calculs et de la fiabilité des résultats en a fait la méthode la plus utilisée pour étudier la structure électronique. Le succès de ces calculs attire à présent l'attention d'expérimentateurs, qui interagissent de plus en plus avec les théoriciens : une nette augmentation des travaux collaboratifs a été observée dans la littérature ces dernières années. Il faut cependant être vigilant pour que cet engouement pour la DFT n'occulte pas les petits « manques » théoriques des approximations et le caractère semi-empirique de certains modèles.

L'heure du bilan est arrivée et il convient de se poser quelques questions sur l'avenir de la DFT. L'explosion de la puissance de calcul va-t-elle la faire tomber dans l'oubli dans une dizaine d'années au profit de méthodes plus coûteuses mais plus précises ? Il y a plusieurs

raisons pour que cela ne soit pas le cas. Tout d'abord, à puissance de calcul égale, la DFT sera toujours capable de traiter des systèmes plus gros se rapprochant des systèmes réels étudiés par les expérimentateurs, ce qui est un atout non négligeable. De plus, un autre argument en faveur de la DFT est sa simplicité qui permet d'imaginer plus aisément de nouvelles variantes qui puissent gagner, soit en simplicité, soit en complexité. Finalement, la qualité exceptionnelle des résultats obtenus avec les approximations rudimentaires faites en DFT sous-tend l'existence d'une vision plus simple de ce problème compliqué à N particules et ce point seul mérite que l'on continue l'étude de la DFT.

Il existe actuellement un très grand nombre de logiciels de chimie quantique et de modélisation moléculaire largement utilisés aussi bien par le secteur de la recherche que de l'industrie. Plusieurs grandes entreprises dans le domaine de l'industrie pharmaceutique ou de la pétrochimie, par exemple, en font grand usage, et comportent souvent en leur sein une division de modélisation. De plus, indépendamment des revues spécialisées dans ce domaine, la plupart des articles publiés dans les grands journaux de chimie font de plus en plus appel à des résultats de la chimie théorique. C'est dire que la chimie quantique trouve toute sa place dans le développement scientifique et technologique actuel en tant que puissant moyen d'investigation ou appui de l'expérience. Le développement de la technologie des ordinateurs ne pourra qu'accentuer cette tendance.

References

[1] E. Schrödinger, "Quantisierungals Eigenwertproblem", *Annalen der Physik*, 79, **1926**, 361–376.

[2] M. Born, R. Oppenheimer, "Zur Quantentheorie der Molekeln", *Annalen der Physik*, 389, **1927**, 457–484.

[3] J.A. Pople, "Nobel Lecture: Quantum chemical models", *Review of Modern Physics*, 71, **1999**, 1267–1274.

[4] W. Kohn, "Nobel Lecture: Electronic structure of matter – Wave functions and density functionals", *Review of Modern Physics*, 71, **1999**, 1253–1266.

[5] D.R. Hartree, "The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods", *Mathematical Proceedings of the Cambridge Philosophical Society*, 24, **1928**, 89–110.

[6] V. Fock, "Näherungs methode zur Lösung des quanten mechanischen Mehrkörper problems", *Zeitschrift für Physik*, 61, **1930**, 126–148.

[7] J.L. Rivail, "Eléments de chimie quantique à l'usage des chimistes", 2^{éme} édition, CNRS Editions, **1999**.

[8] R.S. Mulliken, "The Assignment of Quantum Numbers for Electrons in Molecules. II. Correlation of Molecular and Atomic Electron States", *Physical Review*, 32, **1928**, 761–772.

[9] L.H. Thomas, "The calculation of atomic fields", *Proceedings of the Cambridge Philosophical Society*, 23, **1927**, 542–548.

[10] a- E. Fermi, "A Statistical Method for the Determination of Some Atomic Properties", *Rendiconti Lincei*, 6, **1927**, 602–607.
b- E. Fermi, "A Statistical Method for the Determination of Some Properties of the Atom and Its Application to the Theory of the Periodic System of the Elements", *Zeitschrift für Physik*, 48, **1928**, 73–75.
c- E. Fermi, "On the Statistical Deduction of Some Atomic Properties. Application to the Theory of the Periodic System of the Elements" *Rendiconti Lincei*, 7(6), **1928**, 342–346.

[11] a- P.A.M. Dirac, "The Quantum Theory of the Electron", *Proceedings of the Royal Society of London A*, 117, **1928**, 610–624.
b- P.A.M. Dirac, "The Quantum Theory of the Electron. Part II", *Proceedings of the Royal Society of London A*, 118, **1928**, 351–361.

[12] P. Hohenberg and W. Kohn, "Inhomogeneous Electron Gas", *Physical Review*, 136, **1964**, 864–871.

[13] W. Kohn and L. Sham, "Self-Consistent Equations Including Exchange and Correlation Effects", *Physical Review*, 140, **1965**, 1133–1138.

[14] R.G. Parr, R.A. Donnelly, M. Levy, and W.E. Palke, "Electronegativity: the density functional viewpoint", *The Journal of Chemical Physics*, 68, **1978**, 3801–3807.

[15] P. Hohenberg and W. Kohn, "Inhomogeneous Electron Gas", *Physical Review*, 136(3), **1964**, 864–871.

[16] R.G. Parr and W. Yang, "Density Functional Theory of Atoms and Molecules", *Oxford University Press New-York*, **1989**.

[17] W. Kohn and L.J. Sham, "Self-consistent equations including exchange and correlation effects", *Physical Review*, 140(4), **1965**, 1133–1138

[18] J. Harris and R.O. Jones, "The surface energy of a bounded electron gas", *Journal of Physics F: Metal Physics*, 4(8), **1974**, 1170–1186.

[19] R.A. Harris, "Induction and dispersion forces in the electron gas theory of interacting closed shell systems", *The Journal of Chemical Physics*, 81(5), **1984**, 2403–2405.

[20] O. Gunnarsson and B.I. Lundqvist, "Exchange and correlation in atoms, molecules, and solids by the spin-density-functional formalism", *Physical Review B*, 13(10), **1976**, 4274–4298.

[21] D.M. Ceperley and B.J. Alder, “Ground state of the electron gas by a stochastic method”, *Physical Review Letters*, 45(7), **1980**, 566–569.
[22] G. Ortiz and P. Ballone, “Correlation energy, structure factor, radial distribution function, and momentum distribution of the spin-polarized uniform electron gas”, *Physical Review B*, 50 **1994**, 1391–1405.
[23] A.D. Becke, “Density functional exchange energy approximation with correct asymptotic behavior”, *Physical Review*, 38, **1988**, 3098–3100.
[24] J.P. Perdew, “Density functional approximation for the correlation energy of the inhomogeneous electron gas”, *Physical Review*, 33, **1986**, 8822–8824.
[25] C. Lee, W. Yang, and R.G. Parr, “Development of the colle-salvetti correlation-energy formula into a functional of the electron density”, *Physical Review*, 37, **1988**, 785–789.
[26] A.D. Becke, “Density-functional thermochemistry.3: The role of exact exchange”, *The Journal of Physical Chemistry*, 98, **1993**, 5648–5652.
[27] P.J. Stephens, J.F. Devlin, C.F. Chabalowski, and M.J. Frisch, “ab initio calculations of vibrational absorption and circular dichroïsm spectra using SCF, MP2 and density functional theory force fields”, *The Journal of Physical Chemistry*, 98, **1994**, 11623–11627.