

سورة



پایان نامه دوره کارشناسی ارشد در رشته شیمی گرایش تجزیه

عنوان:

مدل سازی QSPR پتانسیل اکسایش- کاهش  
آنتی اکسیدان های فنولی

استاد راهنما:

دکتر محمد حسین فاطمی

استاد مشاور:

دکتر محمدرضا حاج محمدی

اساتید داور:

دکتر محمد جواد چایچی      دکتر سید ناصر عزیزی

دانشجو:

زهره قره چاهی

شهریور ماه ۱۳۹۰

## پاسکزاری

ستایش و سپاس بیکران بایسته آن ایزد دانایی است که چراغ دانش را در اندیشه انسان فروزان می‌دارد، تا در پرتو آن خود را از بردگی جمل برهنده و به آزادی و بهروزی دست یابد.

از پدر و مادر عزیزم

که گوهر وجودشان، نسیم کلامشان، باران محبتشان و دستان همیشه یاریگرشان همواره راهنمای راه زندگی ام بوده است، پاسکزارم و بردستان پر مهرشان بوسه می‌زنم.

از برادر مهربانم، احسان، که در دوران تحصیل همواره راهنما و پشتیبانم بوده، بی‌نیات پاسکزارم و بهترین‌ها را برایش آرزو مندم. در اینجا وظیفه‌ناگر دی خود می‌دانم تا از زحمات استاد بزرگوارم جناب آقای دکتر محمد حسین فاطمی، که مراب‌ه عنوان عضوی از گروه خود پذیرفتند و در تمام مراحل انجام پایان‌نامه صبورانه راهنما و پشتیبان من بودند، نیات قدردانی را داشته باشم. همچنین صمیمانه‌ترین تشکرات خود را به استاد مشاور محترم جناب آقای دکتر محمد رضا حاج محمدی تقدیم می‌دارم. از اساتید داور جناب آقای دکتر محمد جواد چایچی و جناب آقای دکتر سید ناصر عزیز می‌گردد که زحمت مطالعه پروژه بنده را قبل نمودند، بسیار متشکرم.

در پایان از دوست بسیار مهربانم، مانیه‌الستی که آشنایی با وی دنیای جدیدی را پیش روی چشمانم گشود و در این دوره از تحصیل همواره مشوق و حامی من بوده، بی‌نیات پاسکزارم.

تقدیم به

عزیزترین موبیت های زندگیم

پدر و مادر

و دردانه های همیشه قلمم

احسان، ناامید

ایمان و نادیا

## چکیده

روابط کمی ساختارفعالیت یا ویژگی (QSAR/QSPR) یکی از فنون نویدبخش در زمینه روش‌های مجازی به منظور پیش‌بینی ویژگی‌های شیمیایی است. این روش‌ها، با استفاده از توصیف‌کننده‌هایی که از ساختار مولکولی منتج می‌شوند، به جستجوی الگویی در داده‌ها می‌پردازند تا فعالیت یا ویژگی مواد شیمیایی جدیدی را که ویژگی‌های مولکولی مشابهی دارند، پیش‌بینی کنند. در بخش اول این پروژه، از روش QSPR، جهت پیش‌بینی پتانسیل اکسایش-کاهش ۴۲ آنتی‌اکسیدان فنولی استفاده شد. به منظور انتخاب مهم‌ترین توصیف‌کننده‌ها از روش رگرسیون خطی چندگانه مرحله‌ای استفاده گردید. سپس، روش رگرسیون خطی چندگانه (MLR) و شبکه عصبی پرسپترون چند لایه (MLP NN) جهت ساخت مدل‌های QSPR خطی و غیرخطی به کار گرفته شد. مقایسه نتایج آماری این دو مدل نشان داد که مدل MLP NN در پیش‌بینی پتانسیل اکسایش-کاهش آنتی‌اکسیدان‌های فنولی از اعتبار بیشتری برخوردار است. بعلاوه، بررسی توصیف‌کننده‌های موجود در مدل‌های QSPR نشان داد که خصوصیات الکترونی مشتقات فنولی، نقش مهمی در ویژگی‌های آنتی‌اکسیدانی آن‌ها دارد.

در بخش دوم این پروژه، مدل‌های QSPR بر پایه روش‌های MLR، MLP NN و رگرسیون بردار پشتیبان (SVR)، جهت پیش‌بینی نیمه‌عمر پالایشی ۶۲ مولکول بای‌فنیل چندکلره در ماهی قزل‌آلای رنگین‌کمان ساخته شد. در این بخش از پروژه، از الگوریتم ژنتیک به عنوان روش انتخاب متغیر استفاده شد. مقادیر عددی بزرگ  $R^2$  و  $Q_{LOO}^2$  و مقدار کم RMSE، برتری مدل MLP NN و همچنین وابستگی غیرخطی ویژگی‌های ساختاری مولکولی به نیمه‌عمر پالایشی بای‌فنیل‌های چندکلره را تأیید می‌کند. تجزیه و تحلیل توصیف‌کننده‌های موجود در مدل‌ها حاکی از این است که ویژگی‌های ساختاری ۲ بعدی مولکول، تراکم و الکترون‌نگاتیویته از عوامل اصلی در تعیین نیمه‌عمر پالایشی ترکیبات بای‌فنیل‌های چندکلره می‌باشد.

**واژه‌های کلیدی:** پتانسیل اکسایش-کاهش، آنتی‌اکسیدان‌های فنولی، نیمه‌عمر پالایشی، بای‌فنیل‌های چندکلره،

رابطه کمی ساختار ویژگی، شبکه عصبی مصنوعی

## فهرست مطالب

صفحه	عنوان
۱	فصل اول: مقدمه.....
۳	۱-۱- روش های QSAR/QSPR.....
۴	۲-۱- اهداف روش های QSAR/QSPR.....
۵	۳-۱- روش های مدل سازی.....
۷	فصل دوم: تئوری.....
۹	۱-۲- جمع آوری و انتخاب سری داده ها.....
۱۰	۲-۲- رسم و بهینه سازی ساختارهای مولکولی.....
۱۱	۳-۲- محاسبه توصیف کننده های مولکولی.....
۱۱	۱-۳-۲- توصیف کننده های تجربی.....
۱۲	۲-۳-۲- توصیف کننده های تئوری.....
۱۲	۱-۲-۳-۲- توصیف کننده های ساختاری.....
۱۳	۲-۲-۳-۲- توصیف کننده های توپولوژیکی.....
۱۳	۱-۲-۲-۳-۲- توصیف کننده های جزء.....
۱۳	۲-۲-۲-۳-۲- شاخص های توپولوژی.....
۱۴	۳-۲-۲-۳-۲- توصیف کننده های زیرساختاری.....
۱۴	۴-۲-۲-۳-۲- توصیف کننده های محیطی.....
۱۵	۳-۲-۳-۲- توصیف کننده های هندسی.....
۱۶	۴-۲-۳-۲- توصیف کننده های الکترونی.....
۱۷	۴-۲- تجزیه و تحلیل آماری توصیف کننده ها و انتخاب مؤثرترین آن ها.....
۱۸	۱-۴-۲- رگرسیون خطی چندگانه مرحله ای.....
۱۹	۲-۴-۲- الگوریتم ژنتیک.....
۲۱	۱-۲-۴-۲- مفاهیم و اصول الگوریتم ژنتیک.....
۲۶	۵-۲- ایجاد مدل های آماری.....
۲۷	۱-۵-۲- رگرسیون خطی چندگانه.....
۲۸	۲-۵-۲- شبکه های عصبی مصنوعی.....
۲۸	۱-۲-۵-۲- ساختار سلول عصبی انسان.....
۳۰	۲-۲-۵-۲- عنصر پردازش در شبکه های عصبی مصنوعی.....
۳۳	۳-۲-۵-۲- ساختار شبکه های عصبی.....
۳۶	۴-۲-۵-۲- انواع شبکه های عصبی.....

۳۶	..... آموزش شبکه‌های عصبی .....
۳۹	..... ماشین بردار پشتیبان .....
۴۰	..... اصول رگرسیون بردار پشتیبان .....
۴۴	..... انواع توابع کرنل .....
۴۵	..... آموزش رگرسیون بردار پشتیبان .....
۴۷	..... تحلیل و ارزیابی آماری مدل‌ها و انتخاب بهترین مدل .....
۵۱	..... نرم‌افزارهای مورد استفاده .....
۵۱	..... HyperChem نرم‌افزاری .....
۵۲	..... DRAGON نرم‌افزاری .....
۵۲	..... MOPAC نرم‌افزاری .....
۵۳	..... CODESSA نرم‌افزاری .....
۵۳	..... SPSS نرم‌افزاری .....
۵۴	..... STATISTICA نرم‌افزاری .....
۵۴	..... MATLAB نرم‌افزاری .....

## فصل سوم: مدل‌سازی QSPR پتانسیل اکسایش-کاهش آنتی‌اکسیدان‌های فنولی..... ۵۵

۵۹	..... ۱-۱-۳ روش کار .....
۵۹	..... ۱-۱-۳ سری داده‌ها .....
۶۱	..... ۲-۱-۳ محاسبه و پیش‌پردازش توصیف‌کننده‌ها .....
۶۲	..... ۳-۱-۳ انتخاب توصیف‌کننده‌ها و مدل‌سازی خطی .....
۶۶	..... ۴-۱-۳ مدل‌سازی غیرخطی با شبکه‌ی عصبی پرسپترون چند لایه .....
۶۷	..... ۲-۳ بحث و نتیجه‌گیری .....
۶۷	..... ۱-۲-۳ بررسی نتایج .....
۷۲	..... ۲-۲-۳ بررسی و تفسیر توصیف‌کننده‌ها .....
۷۳	..... ۳-۳ نتیجه‌گیری کلی .....

## فصل چهارم: مدل‌سازی و پیش‌بینی نیمه‌عمر آلاینده‌های بای‌فنیل‌های چند کلره با بهره‌گیری از رویکرد QSPR..... ۷۴

۷۸	..... ۱-۴ روش کار .....
۷۸	..... ۱-۱-۴ سری داده‌ها .....
۸۰	..... ۲-۱-۴ محاسبه و پیش‌پردازش توصیف‌کننده‌ها .....
۸۱	..... ۳-۱-۴ انتخاب توصیف‌کننده‌ها و مدل‌سازی خطی .....
۸۳	..... ۴-۱-۴ مدل‌سازی غیرخطی .....
۸۳	..... ۱-۴-۱-۴ مدل‌سازی غیرخطی با شبکه‌ی عصبی پرسپترون چند لایه .....

۸۵	..... ۲-۴-۱-۴ مدل‌سازی غیرخطی با رگرسیون بردار پشتیبان
۸۶	..... ۲-۴-۲-۴ بحث و نتیجه‌گیری
۸۶	..... ۲-۴-۱-۲ بررسی نتایج
۸۹	..... ۲-۴-۲-۲ ارزیابی داخلی مدل‌ها
۹۰	..... ۲-۴-۳-۲ بررسی و تفسیر توصیف‌کننده‌ها
۹۲	..... ۲-۴-۳-۴ نتیجه‌گیری کلی
۹۳	..... پیشنهادات برای پژوهش‌های آینده



## فهرست شکل‌ها

عنوان	صفحه
شکل ۱-۲- نمایی از نقاط بهینه محلی و کلی .....	۲۱
شکل ۲-۲- طرحواره‌ای از ترکیب تک‌نقطه در الگوریتم ژنتیک .....	۲۴
شکل ۳-۲- طرحواره‌ای از عملگر جهش در الگوریتم ژنتیک .....	۲۵
شکل ۴-۲- نمودار گردش الگوریتم ژنتیک .....	۲۶
شکل ۵-۲- اجزای تشکیل‌دهنده یک نرون زیستی .....	۲۹
شکل ۶-۲- شبیه‌سازی نرون مصنوعی از روی نرون زیستی .....	۳۰
شکل ۷-۲- طرحواره‌ای از عنصر پردازش در شبکه‌های عصبی .....	۳۱
شکل ۸-۲- تابع فعال‌سازی سیگموئیدی (S-شکل) .....	۳۳
شکل ۹-۲- طرحواره‌ای از ساختار شبکه عصبی سه لایه .....	۳۴
شکل ۱۰-۲- ساختار شبکه‌های (الف) پیشخور (ب) برگشتی .....	۳۵
شکل ۱۱-۲- طرحواره‌ای از ساختار آموزش با ناظر .....	۳۸
شکل ۱۲-۲- تابع اتلاف و موقعیت بردارهای پشتیبان بر روی آن .....	۴۲
شکل ۱۳-۲- نگاشت داده‌ها به فضایی با ابعاد بالاتر ( $\emptyset$ ) توسط تابع کرنل .....	۴۴
شکل ۱۴-۲- تأثیر افزایش $E$ بر روی تعداد بردارهای پشتیبان .....	۴۶
شکل ۱-۳- طرحواره‌ای از شبکه عصبی پرسپترون سه لایه بهینه شده .....	۶۷
شکل ۲-۳- نمودار میله‌ای پارامترهای آماری مدل MLP NN و MLR .....	۶۸
شکل ۳-۳- نمودار مقادیر پیش‌بینی شده پارامتر $E_7$ با استفاده از مدل MLP NN بر حسب مقادیر تجربی .....	۶۹
شکل ۴-۳- نمودار باقیمانده مقادیر پارامتر $E_7$ پیش‌بینی شده با استفاده از مدل MLP NN بر حسب مقادیر تجربی .....	۷۰
شکل ۵-۳- نتایج تحلیل حساسیت .....	۷۱
شکل ۱-۴- طرحواره‌ای از شبکه عصبی پرسپترون سه لایه بهینه شده .....	۸۴
شکل ۲-۴- نمودار میله‌ای پارامترهای آماری مدل MLP NN و MLR .....	۸۷
شکل ۳-۴- نمودار مقادیر پیش‌بینی شده لگاریتم نیمه‌عمر با استفاده از مدل MLP NN بر حسب مقادیر تجربی .....	۸۸
شکل ۴-۴- نمودار مقادیر باقیمانده لگاریتم نیمه‌عمر با استفاده از مدل MLP NN بر حسب مقادیر تجربی .....	۸۸
شکل ۵-۴- نتایج تحلیل حساسیت .....	۸۹

## فهرست جدول‌ها

عنوان	صفحه
جدول ۱-۳- ترکیبات سری داده و مقادیر تجربی پتانسیل اکسایش-کاهش آن‌ها بر حسب ولت ( $E_7$ )	۶۰
جدول ۲-۳- مشخصات توصیف کننده‌های انتخاب شده جهت مدل سازی $E_7$	۶۳
جدول ۳-۳- ماتریس ضرایب همبستگی بین توصیف کننده‌های انتخاب شده	۶۳
جدول ۴-۳- مقادیر عددی توصیف کننده‌ها جهت مدل سازی پارامتر $E_7$	۶۴
جدول ۵-۳- مقادیر تجربی و پیش‌بینی شده $E_7$ با استفاده از دو مدل MLR و MLP NN به همراه مقادیر باقیمانده آن‌ها	۶۵
جدول ۶-۳- پارامترهای آماری حاصل از مدل‌های MLR و MLP NN	۶۸
جدول ۷-۳- نتایج آزمون درهم آمیختگی -Y	۷۰
جدول ۱-۴- مقادیر تجربی و پیش‌بینی شده نیمه عمر پالایشی ترکیبات بای فنیل‌های چند کلره	۷۹
جدول ۲-۴- نام و علامت اختصاری توصیف کننده‌های انتخاب شده	۸۲
جدول ۳-۴- ماتریس ضرایب همبستگی بین توصیف کننده‌های انتخاب شده	۸۲
جدول ۴-۴- پارامترهای آماری حاصل از مدل‌های MLR، MLP NN و SVR	۸۶

## لیست علایم و اختصارات

QSAR	رابطه کمی ساختار-فعالیت (Quantitative Structure-Activity Relationship)
QSPR	رابطه کمی ساختار-ویژگی (Quantitative Structure-Property Relationship)
MLR	رگرسیون خطی چندگانه (Multiple Linear Regression)
ANN	شبکه عصبی مصنوعی (Artificial Neural Network)
SVM	ماشین بردار پشتیبان (Support Vector Machine)
MLP NN	شبکه عصبی پرسپترون چند لایه (Multi-Layer Perceptron Neural Network)
SVR	رگرسیون بردار پشتیبان (Support Vector Regression)
EVA	توصیف کننده‌های تحلیل مقدار ویژه (Eigen Value Analysis descriptors) توصیف کننده‌های مولکولی کلی نامتغیر وزن دار شده
WHIM	(Weighted Holistic Invariant Molecular descriptors) توصیف کننده‌های نمایش ساختار سه بعدی مولکول بر پایه پراش الکترونی
3D-MoRSE	(3D-Molecular Representation of Structure based on Electron diffraction descriptors) توصیف کننده‌های ترکیب هندسه، توپولوژی و وزن‌های اتمی
GETAWAY	(GEometry, Topology and Atom-Weights Assembly descriptors)
HOMO	بالترین اوربیتال مولکولی اشغال شده (Highest Occupied Molecular Orbital)
nOH	تعداد گروه‌های هیدروکسیل (number of hydroxyl group)
LUMO	پایین ترین اوربیتال مولکولی اشغال نشده (Lowest Unoccupied Molecular Orbital)
KKT	کاروش-کان-تاکر (Karush-Kuhn-Tucker)
RSS	مجموع مربعات باقی مانده‌ها (Residual Sum of Squares)
TSS	مجموع مربعات کل (Total Sum of Squares)
RMSE	ریشه میانگین مربع خطا (Root Mean Square Error)
AAE	میانگین قدر مطلق خطا (Average Absolute Error)
HOMA	مدل نوسانگر هماهنگ آروماتیسیت (Harmonic Oscillator Model of Aromaticity)
IP	پتانسیل یونیزاسیون (Ionization Potential)
PCBs	بای فنیل‌های چند کلره (PolyChlorinated Biphenyls)
BEHm3	بالترین مقدار ویژه n.3 از ماتریس بردن/ وزن دار شده با جرم‌های اتمی (Highest eigenvalue n.3 of Burden matrix/weighted by atomic masses)
Lop	شاخص مرکزی جدا شده (Lopping centric index)
JGI1	میانگین شاخص بار توپولوژیکی از مرتبه ۱ (Mean topological charge index of order 1)
GATS8	خودهمبستگی گری با فاصله ۸/ وزن دار شده با الکترون‌گاتیویته اتمی ساندerson (Geary autocorrelation lag-8/weighted by atomic Sanderson electronegativities)

# فصل اول

## مقدمه

واژه کمومتریکس<sup>۱</sup> یا شیمی سنجی برای اولین بار در سال ۱۹۷۲ توسط شیمیدانی سوئدی به نام ولد<sup>۲</sup> پیشنهاد گردید و برای نخستین بار در سال ۱۹۷۵ در مجله "علوم کامپیوتر و اطلاعات شیمیایی" توسط کوالسکی<sup>۳</sup> معرفی شد [۱,۲]. کوالسکی تعریف جامعی از علم شیمی سنجی به صورت زیر ارائه داد [۳]:

"شیمی سنجی شاخه‌ای از علم شیمی است که از علوم ریاضی، آمار و منطق جهت طراحی و انتخاب روش‌های بهینه آزمایشگاهی، تحلیل داده‌های شیمیایی به منظور استخراج حداکثر اطلاعات شیمیایی و دستیابی به اطلاعات در مورد سیستم‌های شیمیایی استفاده می‌کند."

اساساً هدف شیمی سنجی بهبود فرآیندهای اندازه‌گیری و استخراج اطلاعات شیمیایی مفید از داده‌های اندازه‌گیری شده فیزیکی و شیمیایی می‌باشد. اگر چه شاخه شیمی سنجی در علم شیمی تجزیه ایجاد شد اما امروزه قلمرو کاربرد آن بسیار گسترده شده است [۴]. بسیاری از شاخه‌های شیمی مانند شیمی آلی، شیمی معدنی، شیمی محیط‌زیست، شیمی غذایی، شیمی کشاورزی و حتی سایر علوم مانند طراحی دارو،

---

<sup>۱</sup> Chemometrics

<sup>۲</sup> Wold

<sup>۳</sup> Kowalski

زیست‌شناسی (ژنومیکس<sup>۱</sup>، پروتئومیکس<sup>۲</sup> و متابولومیکس<sup>۳</sup>) و سم‌شناسی جهت تفسیر داده‌ها به علم شیمی‌سنجی نیازمندند. روش‌های شناخت الگوها<sup>۴</sup>، پردازش علائم<sup>۵</sup>، دسته‌بندی<sup>۶</sup>، بهینه‌سازی<sup>۷</sup>، رابطه کمی ساختار-فعالیت<sup>۸</sup> (QSAR) و رابطه کمی ساختار-ویژگی<sup>۹</sup> (QSPR)، برخی از کاربردهای مهم شیمی‌سنجی در علوم مختلف می‌باشد [۴,۵].

## ۱-۱- روش‌های QSAR/QSPR

در روش‌های QSAR/QSPR به کمک یک مدل محاسباتی، رابطه‌ای ریاضی بین خصوصیات ساختاری ترکیبات و یک ویژگی اندازه‌گیری شده آن‌ها برقرار می‌گردد [۶]. به منظور ایجاد این رابطه کمی، ویژگی‌های فیزیکو-شیمیایی (از قبیل چربی‌دوستی، آبگریزی، حلالیت، قطبش‌پذیری) و ویژگی‌های فضایی، الکترونیکی و ساختاری ترکیبات با استفاده از الگوریتم‌های مختلفی به مقادیر عددی تبدیل می‌شود. این کمیت‌های عددی، توصیف‌کننده‌های مولکولی<sup>۱۰</sup> نامیده شده و به عنوان متغیرهای مستقل در مدل QSAR/QSPR مورد استفاده قرار می‌گیرند. بدین ترتیب، با بهره‌گیری از داده‌های تجربی گروهی از

---

<sup>۱</sup> Genomics

<sup>۲</sup> Proteomics

<sup>۳</sup> Metabolomics

<sup>۴</sup> Pattern recognition

<sup>۵</sup> Signal processing

<sup>۶</sup> Classification

<sup>۷</sup> Optimization

<sup>۸</sup> Quantitative Structure-Activity Relationship

<sup>۹</sup> Quantitative Structure-Property Relationship

<sup>۱۰</sup> Molecular descriptors

ترکیبات و توصیف‌کننده‌های مولکولی آنها، مدلی جهت پیش‌بینی رفتار مولکول‌های مشابه و جدید که خصوصیات ساختاری مشترکی دارند، ساخته می‌شود.

اولین مشاهدات در مورد روابط QSAR/QSPR در رشته سم‌شناسی توسط کراس<sup>۱</sup> صورت گرفت [۷]. وی در سال ۱۸۶۳ در پایان‌نامه دوره دکتری خود، گزارشی در خصوص وجود رابطه بین خصوصیات زیستی و مولکولی ارائه داد. کراس مشاهده کرد که سمیت الکل‌های آلیفاتیک نوع اول با کاهش حلالیت آنها در آب افزایش می‌یابد. به دنبال آن محققین دیگری نیز به وجود روابط خطی بین چربی‌دوستی و ویژگی‌های زیستی مثل سمیت پی بردند [۷]. در این میان، هانش<sup>۲</sup> را می‌توان پیشگام و پایه‌گذار روش QSAR مدرن و به تبع آن QSPR دانست [۸]. وی در اوایل دهه ۱۹۶۰ استفاده از روش رگرسیون خطی چندگانه<sup>۳</sup> (MLR) را به منظور پیش‌بینی فعالیت زیستی ترکیباتی که تا آن زمان سنتز نشده بودند ارائه کرد.

## ۱-۲- اهداف روش‌های QSAR/QSPR

همان‌گونه که گفته شد روش‌های QSAR/QSPR کاربرد گسترده‌ای در بسیاری از زمینه‌های علمی

دارند. در این روش‌ها به طور عمده دو هدف اصلی دنبال می‌شود [۶]:

۱- پیش‌بینی فعالیت‌های زیستی و ویژگی‌های فیزیکوشیمیایی مولکول‌ها

۲- درک و توجیه چگونگی عملکرد و مکانیسم مولکول‌ها در فرآیندهای مختلف

در کنار این اهداف، دلایل دیگری سبب توسعه این روش‌ها شده است [۶]:

---

<sup>۱</sup> Cros

<sup>۲</sup> Hansch

<sup>۳</sup> Multiple Linear Regression

۱- صرفه‌جویی در وقت و هزینه به منظور تولید و توسعه محصولات (به طور مثال فراورده‌های دارویی و حشره‌کش‌ها)

۲- کاهش نیاز به آزمایش‌های هزینه‌بر و طولانی مدت بر روی حیوانات و در نتیجه کاهش استفاده از حیوانات

۳- دستیابی به اهداف شیمی سبز به منظور افزایش کارایی و یا حذف مواد زائد

### ۱-۳- روش‌های مدل‌سازی

به منظور ساخت مدل‌های QSAR/QSPR از روش‌های آماری و ریاضی مختلفی استفاده می‌شود. روش MLR، یکی از انواع روش‌های متداول آماری تحلیل چند متغیره در مطالعات QSAR/QSPR است [۹]. در این روش تنها روابط خطی پدیده‌های مورد مطالعه بررسی و مدل‌سازی می‌شود. این درحالی است که بسیاری از ویژگی‌های مولکول‌ها از روابط غیرخطی پیروی می‌کنند. پیشرفت‌های اخیر در زمینه علم آمار منجر به توسعه الگوریتم‌های بسیار پیچیده‌ای شده است که استفاده از ابزارهای ماشین فراگیرنده<sup>۱</sup> را در روش‌های غیرخطی آسان می‌کند [۱۰]. ماشین فراگیرنده، شاخه‌ای از هوش مصنوعی<sup>۲</sup> است که به طراحی و کاربرد الگوریتم‌هایی که کامپیوتر را قادر به یادگیری از تجربه می‌سازد، می‌پردازد. از ابزارهای ماشین فراگیرنده می‌توان به شبکه عصبی مصنوعی<sup>۳</sup> (ANN) و ماشین بردار پشتیبان<sup>۴</sup> (SVM) اشاره کرد.

---

<sup>۱</sup> Machine learning

<sup>۲</sup> Artificial intelligence

<sup>۳</sup> Artificial Neural Network

<sup>۴</sup> Support Vector Machine



در بخش اول این پروژه، کارآیی دو روش مدل‌سازی MLR و شبکه عصبی پرسپترون چند لایه<sup>۱</sup> (MLP NN) جهت پیش‌بینی پتانسیل اکسایش-کاهش برخی آنتی‌اکسیدان‌های فنولی با یکدیگر مقایسه شده است. در بخش دوم نیز، عملکرد روش‌های MLR، MLP NN و رگرسیون بردار پشتیبان<sup>۲</sup> (SVR) در مدل‌سازی و پیش‌بینی نیمه‌عمر گروه وسیعی از آلاینده‌های بای‌فینیل‌های چندکلره مورد ارزیابی قرار گرفته است.

---

<sup>۱</sup> Multi-Layer Perceptron Neural Network

<sup>۲</sup> Support Vector Regression

# فصل دوم

## تئوری

امروزه از روش‌های پارامتری به طور گسترده‌ای جهت پیش‌بینی رفتار مولکول‌های جدید براساس رفتار مولکول‌های مشابه استفاده می‌شود. در این روش‌ها، بین یک سری توصیف‌کننده‌های مولکولی و فعالیت یا ویژگی مولکول‌های مورد نظر ارتباطی منطقی برقرار می‌شود. توصیف‌کننده‌های مولکولی، پارامترهایی هستند که جنبه‌های مختلف ساختاری مولکول را به صورت کمی نشان می‌دهند. پس از بیان خصوصیات ساختاری مولکول به صورت عدد، بین ساختار و فعالیت مولکول مورد بررسی رابطه‌ای ریاضی یا کمی برقرار می‌شود. به عبارت دیگر در این گونه مطالعات، توصیف‌کننده‌ها به عنوان متغیرهای مستقل و فعالیت یا ویژگی (پارامتر بیولوژیکی یا شیمیایی) مورد نظر به عنوان متغیر وابسته در نظر گرفته می‌شود. این رابطه می‌تواند برای پیش‌بینی پاسخ بیولوژیکی یا شیمیایی دیگر ساختارهای مشابه مورد استفاده قرار گیرد. در مرحله مدل‌سازی، مدلی از متغیر وابسته بر حسب متغیرهای مستقل ساخته شده و در مرحله پیش‌بینی مورد ارزیابی قرار می‌گیرد.

مراحل کلی مدل‌سازی به روش پارامتری به شرح زیر است:

- جمع‌آوری و انتخاب سری داده‌ها
- رسم و بهینه‌سازی ساختارهای مولکولی
- محاسبه توصیف‌کننده‌های مولکولی
- تجزیه و تحلیل آماری توصیف‌کننده‌ها و انتخاب مؤثرترین آن‌ها
- ایجاد مدل‌های آماری
- تحلیل و ارزیابی آماری مدل‌ها و انتخاب بهترین مدل

## ۲-۱- جمع‌آوری و انتخاب سری داده‌ها

اولین مرحله مدل‌سازی، جمع‌آوری و انتخاب یک سری مولکولی است که مقادیر تجربی ویژگی یا فعالیت مورد نظر آن‌ها صحیح و قابل اعتماد باشد. انتخاب سری داده‌ها از طریق جستجو در مجلات علمی و پایگاه داده‌ها صورت می‌گیرد. مدل‌سازی باید بر روی ترکیباتی که ساختار مشابهی دارند و فعالیت یا ویژگی آن‌ها در شرایط عملی یکسانی اندازه‌گیری شده است، انجام شود. این امر، لازمه دستیابی به مدلی قابل قبول است تا بتوان نتایج حاصل را به سایر ترکیبات مشابه تعمیم داد. به طور معمول، در مدل‌های خطی و برخی مدل‌های غیرخطی مانند رگرسیون بردار پشتیبان سری داده‌ها به دو گروه سری آموزش<sup>۱</sup> و سری آزمون<sup>۲</sup> تقسیم می‌شود. عملیات مدل‌سازی بر روی سری آموزش که در برگیرنده اکثر مولکول‌ها (۸۰٪) است، انجام می‌گردد. از این‌رو، سری آموزش باید نماینده مناسب و جامعی از خصوصیات ساختاری سری

---

<sup>۱</sup> Training set

<sup>۲</sup> Test set