
بسم الله

دانشکده شیمی

پایان‌نامه دوره کارشناسی‌ارشد در رشته شیمی گرایش تجزیه

عنوان:

مدل‌سازی QSPR پتانسیل اکسایش-کاهش آنتی‌اکسیدان‌های فنولی

استاد راهنما:

دکتر محمد حسین فاطمی

استاد مشاور:

دکتر محمدرضا حاج‌محمدی

اساتید داور:

دکتر محمدجواد چایچی دکتر سید ناصر عزیزی

دانشجو:

زهره قره‌چاهی

شهریور ماه ۱۳۹۰

سپاسگزاری

ستایش و سپاس بیکران بایسته آن ایزد دانایی است که چراغ دانش را در اندیشه انسان فروزان می دارد، تا در پرتو آن خود را از بردگی جهل برهاند و به آزادگی و بهروزی دست یابد.

از پدر و مادر عزیزم

که گوهر وجودشان، نسیم کلامشان، باران محبت‌شان و دستان همیشه یاریگرشان هموار کننده راه زندگی ام بوده است، سپاسگزارم و بر دستان پر مهرشان بوسه می زنم.

از برادر مهربانم، احسان، که در دوران تحصیل همواره راهنما و پشتیبانم بوده، بی نهایت سپاسگزارم و بهترین ها را برایش آرزومندم.

در اینجا وظیفه شاگردی خود می دانم تا از زحمات استاد بزرگوارم جناب آقای دکتر محمد حسین فاطمی، که مرا به عنوان عضوی از گروه خود پذیرفتند و در تمام مراحل انجام پایان نامه صبورانه راهنما و پشتیبان من بودند، نهایت قدردانی را داشته باشم. همچنین صمیمانه ترین تشکرات خود را به استاد مشاور محترم جناب آقای دکتر محمدرضا حاج محمدی تقدیم می دارم. از اساتید داور جناب آقای دکتر محمدجواد چاپچی و جناب آقای دکتر سید ناصر عزیزی که زحمت مطالعه پروژه بنده را تقبل نمودند، بسیار متشکرم.

در پایان از دوست بسیار مهربانم، هانیه الستی که آشنایی با وی دنیای جدیدی را پیش روی چشمانم گشود و در این دوره از تحصیل همواره مشوق و حامی من بوده، بی نهایت سپاسگزارم.

تقدیم به

عزیزترین موهبت های زندگیم

پدر و مادر

و دردانه های همیشه قلبم

احسان، ناهید

ایمان و نادیا

چکیده

روابط کمی ساختارـفعالیت 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 اکسایش‌ـکاهش 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 اکسایش‌ـکاهش 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 Q^2_{LOO} و R^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 الکترونگاتیویته از عوامل اصلی در تعیین نیمه‌عمر پالایشی ترکیبات بای‌فنیل‌های چندکلره می‌باشد.

واژه‌های کلیدی: پتانسیل اکسایش‌ـکاهش، آنتی‌اکسیدان‌های فنولی، نیمه‌عمر پالایشی، بای‌فنیل‌های چندکلره، رابطه کمی ساختار ویژگی، شبکه عصبی مصنوعی

فهرست مطالب

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فهرست جدول‌ها

عنوان **صفحه**

لیست علایم و اختصارات

QSAR	رابطه کمی ساختار-فعالیت (Quantitative Structure-Activity Relationship)
QSPR	رابطه کمی ساختار-ویژگی (Quantitative Structure-Propety 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	توصیف‌کننده‌های تحلیل مقدار ویژه (Eigen Value Analysis descriptors)
WHIM	توصیف‌کننده‌های مولکولی کلی نامتغیر وزن‌دار شده (Weighted Holistic Invariant Molecular descriptors)
3D-MoRSE	توصیف‌کننده‌های نمایش ساختار سه بعدی مولکول بر پایه پراش الکترونی (3D-Molecular Representation of Structure based on Electron diffraction descriptors)
GETAWAY	توصیف‌کننده‌های ترکیب هندسه، توپولوژی و وزن‌های اتمی (GEometry, Topology and Atom-Weights Assembly descriptors)
HOMO	بالاترین اوربیتال مولکولی اشغال‌شده (Highest Occupied Molecular Orbital)
nOH	تعداد گروه‌های هیدروکسیل (number of hydroxyl group)
LUMO	پایین‌ترین اوربیتال مولکولی اشغال‌نشده (Lowest Unoccupied Molecular Orbital)
KKT	کاروش-کان-تاکر (Karush-Kuhn-Tucker)
RSS	مجموع مربعات باقی‌مانده‌ها (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 آروماتیسیته (Harmonic Oscillator Model of Aromaticity)
IP	پتانسیل یونیزاسیون (Ionization Potential)
PCBs	بای‌فنیل‌های چندکلره (PolyChlorinated Biphenyls)
BEHm3	بالاترین مقدار ویژه n.3 از ماتریس بردن/ وزن‌دار 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/ وزن‌دار شده با الکترونگاتیویته اتمی ساندرسون (Geary autocorrelation lag-8/weighted by atomic Sanderson electronegativities)

فصل اول

مقدمه

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

"شیمی‌سنجی 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 شیمی‌سنجی بهبود فرآیندهای اندازه‌گیری و استخراج اطلاعات شیمیایی مفید از داده‌های اندازه‌گیری شده فیزیکی و شیمیایی می‌باشد. اگر چه شاخه شیمی‌سنجی 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،

---

[۱] Chemometrics

[۲] Wold

[۳] Kowalski

زیست‌شناسی (ژنومیکس[۱]، پروتئومیکس[۲] و متابولومیکس[۳]) و سم‌شناسی جهت تفسیر داده‌ها به علم شیمی‌سنجی نیازمندند. روش‌های شناخت الگوها[۴]، پردازش علائم[۵]، دسته‌بندی[۶]، بهینه‌سازی[۷]، رابطه کمی ساختار-فعالیت[۸] (QSAR) و رابطه کمی ساختار-ویژگی[۹] (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 اندازه‌گیری‌شده آن‌ها برقرار می‌گردد [۶]. به منظور ایجاد این رابطه کمی، ویژگی‌های فیزیکوشیمیایی (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[۱۰] نامیده شده و به عنوان متغیرهای مستقل در مدل QSAR/QSPR مورد استفاده قرار می‌گیرند. بدین ترتیب، با بهره‌گیری از داده‌های تجربی گروهی از

---

[۱] Genomics

[۲] Proteomics

[۳] Metabolomics

[۴] Pattern recognition

[۵] Signal processing

[۶] Classification

[۷] Optimization

[۸] Quantitative Structure-Activity Relationship

[۹] Quantitative Structure-Property Relationship

[۱۰] 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[۱] 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[۲] 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[۳] (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]:

۱ـ پیش‌بینی فعالیت‌های زیستی و ویژگی‌های فیزیکوشیمیایی مولکول‌ها

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

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

---

[۱] Cros

[۲] Hansch

[۳] Multiple Linear Regression

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

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

۳ـ دستیابی به اهداف شیمی سبز به منظور افزایش کارآیی 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[۱] را در روش‌های غیرخطی آسان می‌کند [۱۰]. ماشین فراگیرنده، شاخه‌ای از هوش مصنوعی[۲] 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[۳] (ANN) و ماشین بردار پشتیبان[۴] (SVM) اشاره کرد.

[۱] Machine learning

[۲] Artificial intelligence

[۳] Artificial Neural Network

[۴] Support Vector Machine

در بخش اول این پروژه، کارآیی دو روش مدل‌سازی MLR و شبکه عصبی پرسپترون چند لایه[1] (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[2] (SVR) در مدل‌سازی و پیش‌بینی نیمه‌عمر گروه وسیعی از آلاینده‌های بای‌فنیل‌های چندکلره مورد ارزیابی قرار گرفته است.

[1] Multi-Layer Perceptron Neural Network

[2] 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 قابل‌قبول 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[۱] و سری آزمون[۲] 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

[۱] Training set

[۲] Test set