دانشکده شیمی

پایان نامه کارشناسی ارشد شیمی تجزیه

مطالعه ارتباط کمی ساختار-فعالیت مشتقات پیریدین و پیریمیدین به عنوان بازدارنده‌ی HIV

نگارنده:

زینب مظفری

استاد راهنما:

دکتر منصور عرب چم جنگلی

استاد مشاور:

دکتر محمد آرشی

بهمن ۱۳۹۵

شماره ۴۱۱۴۰
تاریخ ۹۵/۱۲/۱۶
ویرایش:

باسمه تعالی

مدیریت تحصیلات تکمیلی

پیوست شماره ۲

دانشکده : شیمی

گروه : شیمی تجزیه

پایان نامه کارشناسی ارشد آقای/ خانم **زینب مظفری** به شماره دانشجویی: **۹۴۱۶۷۱۴**

تحت عنوان: **مطالعه ارتباط کمی ساختار-فعالیت مشتقات پیریدین و پیریمیدین به عنوان بازدارنده‌ی HIV**

در تاریخ **۱۳۹۵/۱۱/۱۶** توسط کمیته تخصصی زیر جهت اخذ مدرک کارشناسی ارشد مورد ارزیابی و با درجه ..عالی.. مورد پذیرش قرار گرفت.

اساتید راهنما	امضاء	اساتید مشاور	امضاء
دکتر منصور عرب چم جنگلی		دکتر محمد آرشی	

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء
دکتر ناصر گودرزی		دکتر اسماعیل سلیمانی	
دکتر قدمعلی باقریان			
-			
-			

اگر شایسته تقدیم باشد

تقدیم به:

«زیباترین نقش های عالم هستی»

پدر بزرگوارم؛

مادر عزیز و دلسوزم؛

فروزان و ابراهیم عزیزم؛

و

به تمام آزادمردانی که نیک می اندیشند و عقل و منطق را پیشه خود نموده و جز رضای الهی و پیشرفت و سعادت

جامعه، هدفی ندارند.

سپاسگزاری

لطف بی‌دریغ خالقم عالم گیر است و من شکرگزار مدام عنایتش. کنون نیز که در این عرصه از وادی پویای دانستن و تامل قدم پیش نهادم، حیّ همراهم را می‌ستایم که از روزن بخشش و امید بر من باوری از جنس "شدن" و "جان گرفتن" را پروراند و همسو با اهداف مطلوبم شعله‌ی تلاشم را روز به روز گرم‌تر بر دامنه‌ی پایندگی افروخت. مسیری که آغاز و استمرارش را مدیون حامیان پرمهر خویش، آینه‌ی دل و جان من و نزدیکترینانم حتی از پشت مسافت‌ها هستم و ماحصل تعالیمم را نثار صبوری و چشمان همیشه مشتاق و کلام همواره مشوق پدر و مادر عزیزم می‌نمایم و بی‌شک گام نهادن و طی این طریق در سایه‌ی رهنمون‌های شایسته‌ی استاد راهنمای گرانقدرم جناب آقای دکتر منصور عرب چم جنگلی و توجه مجدانه‌ی ایشان در پیشبرد هرچه بهتر و ارزنده‌تر آن به ثمر رسید؛ و نیز حسن دقت و زحمات خالصانه‌ی استاد مشاور ارجمندم جناب آقای دکتر محمد آرشی که در رسیدن به نتیجه‌ی والامرا یاری رساندند؛ همچنین اساتید محترم جناب آقایان دکتر قدمعلی باقریان دهقی و دکتر ناصر گودرزی که زحمت داوری پایان‌نامه‌ی حاضر را داشته و از محضر ایشان در کلاس‌های خویش بهره‌ی وافر برده‌ام. از راحله دوستی عزیزم نیز تشکر می‌کنم که در طی این طریق مراهمراهی نمود و پشتکار و تلاش بی‌وقفه را به من آموخت، تشکر می‌کنم تا باشد یادم که دوستی دارم به اندازه یک کوه. از همگی آن عزیزان و فرهیختگان نهایت سپاس و قدردانی را دارم و آرزومندم که در پرتوی نگاه لایزال ایزدی، تلالوء حضور و اندیشه‌ی آن بزرگواران فروزان و برقرار باشد.

تعهد نامه

نویسنده پایان نامه **مطالعه ارتباط کمی ساختار-فعالیت مشتقات پیریدین و پیریمیدین به عنوان بازدارنده‌ی HIV** تحت راهنمایی **دکتر منصور عرب چم جنگلی و** مشاوره **دکتر محمد آرشی** متعهد می شوم :

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ **امضای دانشجو**

مالکیت نتایج و حق نشر

- **کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود .**
- **استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.**

۱ چکیده

برای ساخت مدل، دو نوع توصیف‌کننده تحت عنوان توصیف‌کننده‌های داکینگ مولکولی و توصیف‌کننده‌های ساختاری محاسباتی استفاده شدند. توصیف‌کننده‌های جدید داکینگ مولکولی با توجه به برهم کنش بین ترکیبات (لیگاندها) و پروتئین (گیرنده) از داکینگ استخراج شدند. توصیف‌کننده‌های ساختاری با استفاده از ساختار ترکیبات محاسبه شدند. پس از تولید توصیف‌کننده‌ها، مهم‌ترین توصیف‌کننده‌ها با به کارگیری روش‌های انتخاب متغیر رگرسیون گام به گام و روش منقبض کننده حداقل قدر مطلق و عملگر انتخاب کننده (لاسو) انتخاب شدند. توصیف‌کننده‌های انتخاب شده به‌عنوان ورودی برای ساخت مدل‌های QSAR با استفاده از مدل‌های شبکه عصبی مصنوعی به‌عنوان مدل غیر خطی و جنگل‌های تصادفی به‌عنوان مدل خطی مورد استفاده قرار گرفتند. برای مدل شبکه عصبی، سری داده‌ها به سری آموزش، سری ارزیابی و سری آزمون شامل ۵۱، ۱۱ و ۱۱ ترکیب، تقسیم شدند. برای مدل جنگل‌های تصادفی، سری داده‌ها به دو بخش سری آموزش و سری آزمون شامل ۶۲ و ۱۱ ترکیب تقسیم بندی شدند. برای رسیدن به مدل‌های مناسب همه پارامترهای مدل‌های شبکه عصبی و جنگل‌های تصادفی بهینه شدند. ارزیابی مدل‌های شبکه عصبی مصنوعی و جنگل‌های تصادفی با پیش‌بینی فعالیت ترکیبات سری آزمون مورد مطالعه قرار گرفت. نتایج به دست آمده برای ارزیابی مدل‌ها به ترتیب ضریب همبستگی ۰/۹۲۷۴ و ۰/۸۹۶۸ را برای پیش‌بینی فعالیت ترکیبات سری آزمون مدل‌های شبکه عصبی مصنوعی و جنگل‌های تصادفی نشان داد. هم‌چنین میانگین توان‌های دوم خطا برای پیش‌بینی سری آزمون مدل‌های شبکه عصبی مصنوعی و جنگل‌های تصادفی به ترتیب ۰/۰۵۹۷ و ۰/۰۸۴۹ به دست آمد.

کلمات کلیدی: QSAR، ضدHIV، داکینگ مولکولی، روش رگرسیون لاسو، شبکه عصبی مصنوعی، جنگل‌های تصادفی

نتایج حاصل از این پایان‌نامه در دو پوستر تحت عنوان:

1. "QSAR study of inhibitory activities (EC_{50}) of azabenzenes derivatives using Bayesian regularized artificial neural network (BR-ANN) and calculated descriptors"
2. "Application of random forest regression and artificial neural network in the modeling of the Anti-HIV activity of azabenzenes derivatives, as potent HIV-1 reverse transcriptase"

در بیست و سومین کنفرانس شیمی تجزیه ایران دانشگاه صنعتی شریف تهران در شهریور ماه ۱۳۹۵ پذیرفته و ارائه شد.

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

۱ فصل اول

مقدمه

مقدمه

یکی از مشکلاتی که جامعه بشری همیشه با آن روبه‌رو بوده، مقابله با انواع بیماری‌هایی است که سلامت انسان‌ها را به مخاطره انداخته و همواره یکی از مهم‌ترین دغدغه‌های محققان یافتن داروهای موثر، برای رفع و یا کاهش عوارض این بیماری‌ها بوده است. بروز انواع بیماری‌ها از قبیل ایدز، سرطان، آلزایمر و بیماری‌های مشترک انسان و دام و به علاوه مقاوم شدن ویروس‌ها در برابر آنتی بیوتیک‌ها، همه از جمله مواردی می‌باشند که ذهن دانشمندان را در جهت یافتن داروهای موثر و کارآمد برای مقابله با این بیماری‌ها معطوف به خود نموده‌اند. در گذشته، روند کشف و توسعه داروهای جدید، به‌روش آزمون و خطا صورت می‌گرفت که روشی وقت‌گیر و هزینه بر است. محدودیت دیگری که در این راه دانشمندان با آن مواجه می‌باشند، عدم اطلاع آن‌ها از فعالیت دارویی ترکیبات، قبل از انجام سنتز و بررسی تجربی آن‌ها بوده و به‌همین دلیل یکی از مهم‌ترین اهداف شیمی‌دان‌ها و محققان دارویی پیش‌بینی فعالیت ترکیبات، قبل از سنتز و یا انجام آزمایش بر روی آن‌ها می‌باشد. چرا که انجام بسیاری از آزمایشات مستلزم صرف زمان و هزینه‌های زیادی است. از این‌رو نیاز به استفاده از روش‌های تئوری و محاسباتی که بدون انجام آزمایش بتواند ویژگی و یا فعالیت ترکیبات جدید را پیش‌بینی کند، ضروری به نظر می‌رسد. کمومتریکس در مطالعات مربوط به یافتن ارتباط کمی بین ساختار و فعالیت، مدل‌سازی کیفی و دسته بندی سیستم‌های شیمیایی مورد استفاده قرار می‌گیرد و توانسته راه حلی برای رفع محدودیت‌ها باشد.

۱-۱ پروتئین

ا سیدهای آمینه، الفبای پروتئین‌ها را ت شکیل می‌دهند. به‌طور کلی پروتئین‌ها از تعدادی ا سید آمینه تشکیل شده‌اند که در یک ر شته پشت سر هم قرار گرفته‌اند. ترتیب و ماهیت ا سیدهای آمینه، ویژگی های هر پروتئین را تعیین می‌کند. تعداد کل اســـیدهای آمینه موجود در طبیعت ۲۰ عدد می‌باشـــد که هر کدام از ترکیب‌های متفاوتی به‌وجود آمده‌اند و با توجه به ترکیب و ترتیب اتم‌ها، هر اســید آمینه خواص خود را داراســت. این بیســت اســید آمینه می‌توانند با هر ترکیب و به هر تعداد در ساختار یک پروتئین وجود داشته باشند.

پروتئین‌ها در بخشی از درون سلول به نام ریبوزوم توسط RNA ساخته می‌شوند. رفتار سلولی و تمام فعالیت‌هایی که در ســـلول انجام می‌شـــود بر عهده پروتئین‌ها اســـت. همه پروتئین‌ها با هم بر هم‌کنش دارند و تقریبا می‌توان گفت که همه پروتئین‌ها اثر خود را با همکاری پروتئین‌های دیگر در ســلول اعمال می‌کنند و هیچ پروتئینی نیســت که در ســلول به تنهایی عمل کند. مطالعه پروتئین‌ها و عملکرد آن‌ها در واقع موجب فهم بهتر خود ســـلول‌ها و موجودات می‌شـــود. در ســـطح مولکولی، کلیه مکانیسم‌های زیستی سلول‌ها، توسط پروتئین‌ها انجام می‌شود. پروتئین‌ها در ارتباط با یک‌دیگر به‌طور دقیق و بسیار کنترل شده‌ای وظایف خود را انجام می‌دهند. اساس مولکولی اغلب بیماری‌ها، بروز نقص یا تداخل در کارکرد عادی پروتئین‌هایی اســت که از طریق آن مکانیســم‌های زیســتی ســلولی را انجام می‌دهند. اگر فعالیت پروتئین ها از حالت طبیعی خارج شـــود می‌توانند باعث بیماری های مختلف شوند[۱].

۱-۲ دارو

در بیو شیمی، گیرنده[1] به یک مولکول پروتئینی گفته می شود که بر سطح و یا در عرض غشای

[1] Receptor

پلاسمایی سلول یا درون سیتوپلاسم جاسازی شده‌است تا لیگاند به آن‌ها بچسبند. مولکولی که به گیرنده می‌چسبد لیگاند نام دارد و ممکن است یک پپتید، هورمون، یک دارو یا یک سم باشد و وقتی که اتصال انجام شود، گیرنده تغییر شکل داده و باعث شروع یک زنجیره پاسخ‌ها توسط سلول می‌شود. گیرنده‌ها پیام‌های خارج سلولی را دریافت کرده و اثر این پیام‌ها یا خود پیام را به درون سلول منتقل می‌کنند، ضمن اینکه به‌صورت اختصاصی عمل می‌کنند به این معنا که هر لیگاند، گیرنده مخصوص به خود را دارد[۲].

اتصال دارو به گیرنده از طریق پیوندهای شیمیایی صورت می‌گیرد که شامل پیوندهای کوالانسی، یونی، هیدروفوبی و هم‌چنین هیدروژنی می‌باشد. نیروهایی که دارو را به گیرنده اش جذب می‌کنند باید یه اندازی کافی قوی و بادوام باشند تا منجر به پاسخ فیزیولوژیک گردد.

مولکول دارو پس از مصرف و عبور از غشاها وقتی در نزدیک سطح گیرنده قرار گرفت جاذبه الکترواستاتیک اولین نیرویی است که داروها را به طرف سطح گیرنده با بار مخالف می‌کشد. این نوع پیوند از فا صله ی دورتر می‌تواند ایجاد جاذبه کند که پیوند هیدروژنی از این فا صله قادر به ایجاد جاذبه نمی‌باشند. عموما پیوند یونی (الکترواستاتیک) باید قبل از آنکه گیرنده فعال شود از طریق ایجاد پیوند هیدروژنی یا وان در والس یا هر دو تقویت شود در غیر این صورت به آسانی و سریع توسط انرژی گرمایی شکسته می‌شود. حتی اگر پیوندهای ثانویه متعددی ایجاد شود کمپلکس دارو-گیرنده باز هم تفکیک می‌شود مگر اینکه پیوند کوالانسی ایجاد شده باشد.

در بحث تعامل دارو با گیرنده، داروی آگونسیت[1] پس از اتصال به گیرنده با ایجاد تغییراتی در سطح گیرنده آن‌را فعال می‌کند و سبب ایجاد پاسخ می‌شود، در بعضی گیرنده‌ها مولکول اثر کننده درون خود گیرنده واقع شده و با اتصال دارو مستقیما فعال می‌شود. مانند باز شدن یک کانال یونی و در برخی دیگر مولکول اثر کننده یک مولکول واسطه مجزا می‌باشد که گیرنده پس از

[1] Agonist

اتصال دارو با آن ارتباط برقرار می‌کند. داروهای آنتاگونیست[1] وقتی روی گیرنده می‌نشینند باعث بلوک یا مهار آن می‌شوند و در واقع از اتصال مولکول های دیگر به گیرنده جلوگیری می‌کند[۲].

۱-۳ مکانیسم اتصال دارو-گیرنده

نقاط هدف دارو، اغلب پروتئین‌هایی هم‌چون آنزیم و یا گیرنده‌های سطح سلول می‌باشند. برای انتخاب داروی مناسب دانستن مکانیسم مهار کنندگی آن الزامی است. مکانیسم مهارکنندگی شامل موارد زیر می‌باشد:

۱. مهار کنندگی رقابتی
۲. مهار کنندگی غیر رقابتی

اغلب داروها از راه مکانیسم رقابتی عمل می‌کنند. به این معنا که به طور برگشت پذیری به جایگاه فعال مولکول هدف متصل می شوند. هنگامی که دارو در جایگاه فعال است، سوبسترای اصلی نمی‌تواند متصل شود و این مساله کارایی پروتئین را کاهش می‌دهد، بدون این‌که آن‌را کاملا از بدن حذف نماید. موادی وجود دارند که ساختار شیمیایی آن‌ها عیناً مثل سوبسترای طبیعی یک آنزیم است. این امر ممکن است آن‌ها را قادر سازد تا با جایگاه فعال موجود روی آنزیم پیوند برقرار کنند و به همین دلیل از طریق کاهش تعداد جایگاه‌های فعال در دسترس سوبسترای اصلی، در کار آن‌ها اختلال ایجاد کند. این مواد بر سر اشغال جایگاه فعال با سوبسترا رقابت می‌کنند؛ بنابراین به عنوان مهارکننده‌های رقابتی شناخته می‌شوند. واکنش آنزیمی با مهار کنندگی رقابتی از طریق ایجاد کمپلکس را می‌توان به صورت زیر نمایش داد:

[1] Antagonist

$$E + S \underset{}{\overset{K_s}{\rightleftharpoons}} ES \xrightarrow{K_p} E + P$$

$$+ I \qquad \qquad K_i = \frac{[E][I]}{[EI]}$$

$$\rightleftharpoons K_i$$

$$EI$$

E، S، I، EI، ES و P به ترتیب مربوط به غلظت‌های آنزیم، سوبسترا، بازدارنده، کمپلکس سوبسترا-آنزیم، کمپلکس سوبسترا-بازدارنده و محصول می‌باشد. K_s و K_p به ترتیب مربوط به ثابت‌های تفکیک آنزیم-سوبسترا و آنزیم محصول است و K_i ثابت بازداری نام دارد. مهارکننده‌های غیر رقابتی به جایگاه‌هایی به جز جایگاه‌های آنزیم متصل می‌شوند که در این صورت جایگاه فعال آنزیم برای سوبسترا آزاد می‌باشد، اما این مهارکننده‌ها شکل پروتئین را تغییر می‌دهند و بنابراین فعالیت کاتالیتیک جایگاه فعال را کاهش می‌دهند[۳].

$$E + S \underset{}{\overset{K_s}{\rightleftharpoons}} ES \xrightarrow{K_p} E + P \qquad K_s = \frac{[E][S]}{[ES]} = \frac{[EI][S]}{[ESI]}$$

$$+ I \qquad + I$$

$$\rightleftharpoons K_i \qquad \rightleftharpoons K'_i \qquad K_i = \frac{[E][I]}{[EI]} = \frac{[ES][I]}{[ESI]}$$

$$EI + S \underset{}{\overset{K's}{\rightleftharpoons}} ESI$$

غلظت مهارکنندگی در برهم‌کنش لیگاند- پروتئین حائز اهمیت بوده و آگاهی از آن در بررسی فعالیت آنزیم ضروری است. IC_{50} یا غلظت مهارکنندگی نصف ماکسیمم غلظتی از دارو است، که ۵۰٪ اثر بازدارندگی دا شته با شد. به‌عبارت دیگر به غلظتی از دارو که از فعالیت بیولوژیکی و یا بیو شیمیایی آنزیم تا حدود ۵۰٪جلوگیری کند، IC_{50} گفته می‌شود. مقدار IC_{50} با استفاده از رابطه(۱-۱) به ثابت بازداری K_i تبدیل می‌شود.

$$K_i = \frac{IC_{50}}{1+\frac{s}{K_m}}$$

رابطه(۱-۱)

S غلظت بازدارندگی (دارو) بوده و K_m نیز ثابت متنا سب با هر آنزیم می‌با شد و در غلظت بازدارندگی (دارو) در نصف ماکسیمم فعالیت آنزیم می‌باشد[۴].

۱-۴ بیماری و اهمیت پیش‌بینی داروهای ضد ایدز

۱-۴-۱ ایدز

۱ سندرم نقص ایمنی اکتسابی [1](AIDS) یک بیماری عفونی کشنده است که اولین بار در سال ۱۹۸۳ شناسایی شد. ویروس ایدز سلول‌های ایمنی موجود در خون (لنفوسیت‌ها) و سلول‌های ایمنی موجود در بافت‌ها مانند مغز استخوان، طحال، کبد و گره‌های لنفاوی را درگیر می‌سازد. این سلول‌ها در تولید پادتن برای مقابله با بیماری‌ها و سرطان‌ها نقش دارند. هنگامی که سیستم ایمنی آسیب ببیند نه تنها در برابر ویروس HIV (که در آغاز به آن صدمه زده) بلکه نسبت به سایر عفونت‌های فرصت طلب[2] که در حالت عادی بدن به راحتی با آن‌ها مقابله می‌کند، آسیب پذیر می‌شود. لذا با گذشت زمان افراد آلوده به HIV بیش‌تر و بیش‌تر بیمار می‌شوند و معمولاً سال‌ها پس از آلودگی با این ویروس، به یکی از بیماری‌های خاص (بیماری‌های متأثر از نقص دستگاه ایمنی) مبتلا شده و در این زمان گفته می‌شود که آن‌ها به ایدز مبتلا شده‌اند. بنابراین زمانی که فرد آلوده به ویروس HIVبرای اولین بار به یک بیماری جدی مبتلا شود و یا وقتی که تعداد سلول‌های ایمنی باقی‌مانده در بدن او از حد معینی کم‌تر شود (زیر ۲۰۰عدد در هر میکرولیتر)، مبتلا به بیماری ایدز در نظر گرفته می‌شود[۵].

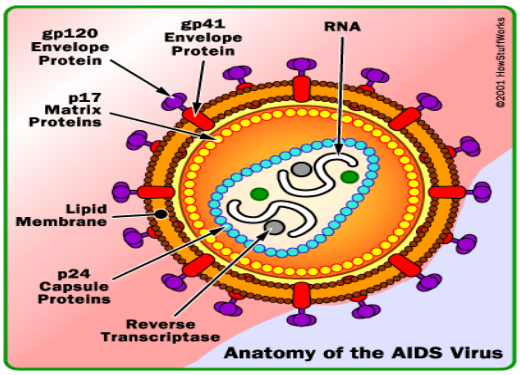
ویروس HIV از خانواده رترو- ویروس‌ها[3] می‌با شد. رتروویروس‌ها، گروهی از ویروس‌ها ه ستند که ژنوم آن‌ها از RNA تشکیل شده است. بنابراین برای تکثیر خود به آنزیمی به نام آنزیم نسخه بردار

[1] Acquired Immunodeficiency Syndrom
[2] Infection opportunistic
[3] Retrovirus

معکوس[1] وابســـته‌اند که ژنوم RNA آن‌ها را به DNA نســـخه برداری کند تا بعد بتواند آن‌را به کمک آنزیم اینتگراز[2] وارد ژنوم میزبان کند و به‌این ترتیب امکان تکثیر ویروس به‌وجود آید. کلمه رترو (معکوس) در نام این ویروس‌ها به‌همین خاطر اســت چرا که معمولا نســخه برداری از DNA به RNA انجام می‌شـــود اما در این دســـته از ویروس‌ها عکس این عمل صـــورت می‌گیرد. رتروویروس‌ها باعث کاهش توانایی سیستم ایمنی بدن میزبان می‌شوند. علایمی که در بیماری ایدز مشاهده می‌شود مربوط به بیماری‌هایی است که در اثر نقص دستگاه ایمنی بدن تولید می‌شوند[۵].

۱-۴-۲ ساختمان ویروس HIV

HIV یک ذره کاملاً کروی با قطر ۱۲۰-۱۰۰ نانومتر است که با غشای دو لایه لیپیدی پوشیده شده است. در سطح غشای لیپیدی دو نوع گلیکوپروتئین (GP41, GP120) وجود دارد که ویروس از طریق این گلیکوپروتئین‌ها قادر خواهد بود به گیرنده‌های CD_4^+ موجود در سطح لنفوسیت T کمک کننده[3] بچسبد. سطح داخلی غشای لیپیدی توسط ماتریکس احاطه شده و اطلاعات ژنتیکی مورد نیاز برای تکثیر ویروسی را در خود جای داده است. این اطلاعات شامل دو کپی یکسان از ژنوم RNA ویروسی و سه آنزیم اینتگراز، پروتئاز[4] و نسخه‌بردار معکوس می‌باشد[۲]. شکل(۱-۱) ساختمان ویروس HIV را نشان می‌دهد.



شکل(۱-۱) ساختمان ویروس HIV [۵]

[1] Reverse transcriptase
[2] Integrase
[3] T helper lymphocyte
[4] Protease

۱-۴-۳ چرخه تکثیر ویروس HIV

چرخه تولید و تکثیر ویروس شامل مراحل زیر است:

۱- چسبیدن ویروس به سلول‌های خونی و آلوده کردن سلول‌ها
۲- تولید DNA پرووِیروسی و اتصال به ژنوم میزبان
۳- رونویسی[1] ژن‌های ویروسی
۴- تولید قطعات ویروسی و رهاسازی ویروس

سلول‌هایی که مورد تهاجم این ویروس قرار می‌گیرند، نوعی از سلول‌های خونی انسان به نام لنفوسیت‌های نوع T کمک کننده می‌باشند. این سلول‌ها دارای گیرنده‌ای به نام CD_4^+ هستند که ویروس پس از اتصال به این گیرنده، توسط پروتئین GP41 (موجود در سطح ویروس HIV) به داخل غشای سلول نفوذ می‌کند و RNA ویروس و آنزیم‌های مختلف (از جمله نسخه‌بردار معکوس، اینتگراز و پروتئاز) به داخل سیتوپلاسم سلول تزریق می‌شود. مرحله بعد تولید DNA از روی RNA ویروس با کمک آنزیم نسخه بردار معکوس است. در صورت موفقیت این عمل، DNA ساخته شده که DNA پرووِیروسی نام دارد، با استفاده از آنزیم اینتگراز به DNA سلول میزبان ملحق می‌گردد. بعد از رونویسی در هسته سلول، RNA ناقل[2] به سیتوپلاسم سلول میزبان منتقل می‌شود و از اجزای سلولی میزبان برای ساختن پروتئین‌های ویروسی استفاده می‌کند. آنزیم پروتئاز فراهم شده به وسیله HIV وظیفه شکستن پلی پپتید ساخته شده به رشته‌های پروتئینی سازنده‌ی HIV را بر عهده دارد. به‌این ترتیب ویروس HIV می‌تواند یک ذره[3] کامل بسازد که به غشای میزبان مهاجرت کرده و غشای لیپیدی مورد نیازش را از غشای سلول میزبان می‌گیرد و آماده‌ی آلوده کردن سلول‌های دیگر است[۵].

[1] Transcription
[2] Messenger RNA

۱-۴-۴ اهداف استفاده از داروهای ضدHIV

۱. سرکوب ویروس و کاهش روند تکثیر آن‌ها
۲. تحکیم و بهبود عملکرد سیستم ایمنی که با شمارش سلول‌های TCD_4^+ ارزیابی می‌شود.
۳. بهبود کیفیت زندگی بیماران با کاهش عوارض بیماری
۴. بهبود امید به زندگی و افزایش طول عمر بیماران
۵. کاهش احتمال سرایت

۱-۴-۵ درمان ضد رتروویروسی HIV

درمان ضد رتروویرو سی نوع ا صلی درمان ایدز می‌با شد که البته شفابخش نی ست اما زندگی مبتلایان به این بیماری را طولانی‌تر می‌کند. درمان شامل داروهایی ا ست که باید ستی روزانه و تا آخر عمر مصرف شوند[۵].

۱-۴-۶ انواع داروهای ضد رتروویروسی

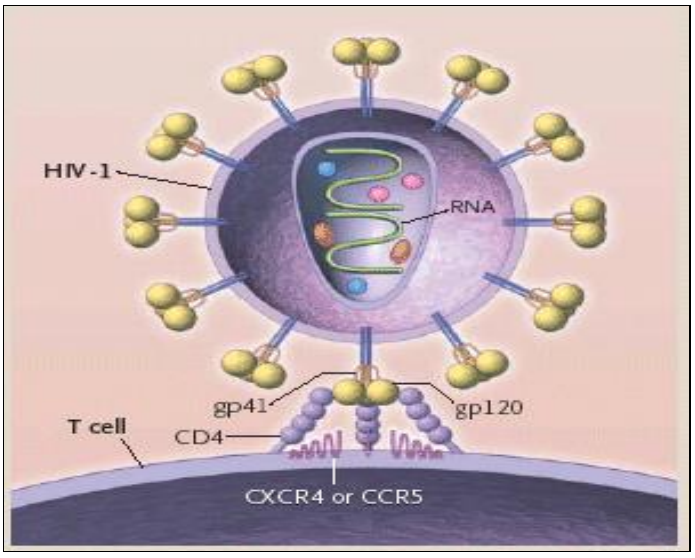
از زمان آغاز اپیدمی HIV مجموعه‌ای از داروها ساخته شده‌اند که از طریق مهار توانایی تکثیر این ویروس، عمر افراد مبتلا به این ویروس را به طور قابل ملاحظه‌ای افزایش می‌دهند. این داروها از سرعت نابودی CD_4^+ بیماران می‌کاهد و می‌توانند بروز بیماری ایدز را به تأخیر بیندازد، اما آن‌را معالجه نمی‌کند. در مجموع چهار دسته دارو وجود دارد که در مراحل مختلف آلودگی و رشد ویروس HIV عمل می‌کنند و با توجه به اینکه این ویروس از سه آنزیم نسخه‌بردار معکوس، پروتئاز و اینتگراز برای تکثیر خود و آلوده کردن سلول‌ها استفاده می‌کند، بنابراین داروهای سنتز شده بیش‌تر این نوع آنزیم‌ها را هدف قرار می‌دهند.

۱-۴-۶-۱ بازدارنده‌های ورودی[1]

در سطح بیرونی ویروس HIV گلیکوپروتئین‌های GP41, GP120 وجود دارد شکل(۱-۲) که به‌وسیله آن‌ها ویروس به گیرنده‌های CD_4^+ موجود در سطح لنفوسیت‌های T سلول میزبان متصل شده

[1] Fusion or Entry inhibitors

و وارد سلول می‌گردد. بازدارنده‌های ورودی به این پروتئین‌ها می‌چسبند و از پیوستن ویروس به سلول‌های CD_4^+ جلوگیری می‌کند. به‌این ترتیب با بلوکه کردن یکی از این پروتئین‌ها، تکثیر آهسته می‌شود.

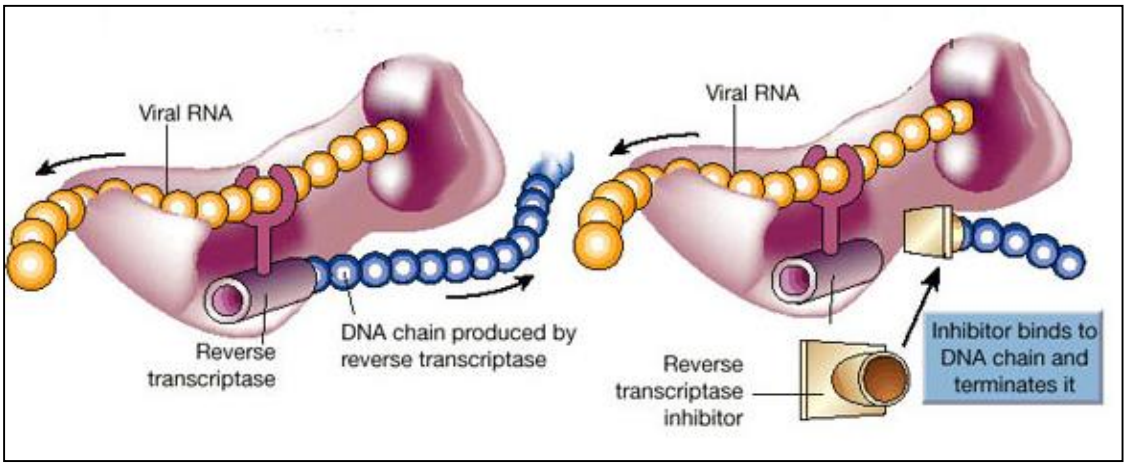


شکل(۱-۲) نحوه ورود ویروس HIV به دورن سلول میزبان [۵]

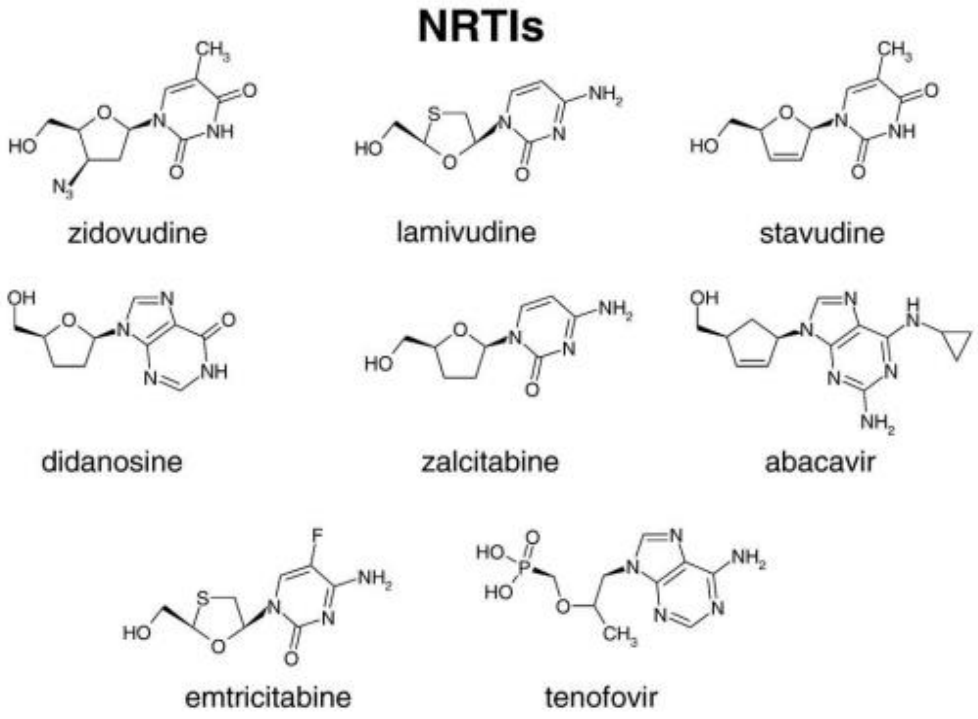
۱-۴-۶-۲ بازدارنده‌های نوکلئوزیدی آنزیم نسخه‌بردار معکوس[1]

این دسته از مهارکننده‌ها بازدارنده‌های رقابتی هستند. مهارکننده‌های نوکلئوزیدی به زنجیره نوکلئوزیدی DNA که توسط آنزیم نسخه‌بردار معکوس درحال تولید است، می‌چسبد و مانع از اتصال نوکلئوزیدهای بیشتر به DNA می شوند به این ترتیب یک DNA ناقص تولید می شود. شکل(۱-۳) عملکرد این مهار کننده‌ها و شکل(۱-۴) نمونه‌هایی از این دسته از بازدارنده‌ها را به خوبی نشان می‌دهد[۷،۶].

[1] Nucleoside Reverse Transcriptase Inhibitors (NRTIs)



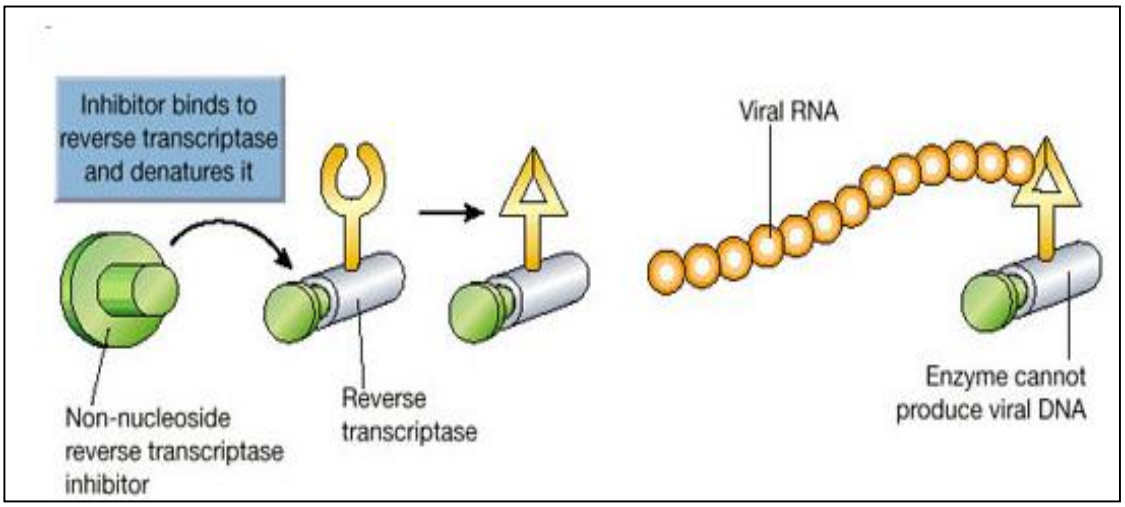
شکل(۱-۳) نحوه عملکرد بازدارنده‌های نوکلئوزیدی آنزیم نسخه بردار معکوس [۶]



شکل(۱-۴) نمونه‌هایی از مهار کننده‌های نوکلئوزیدی [۷]

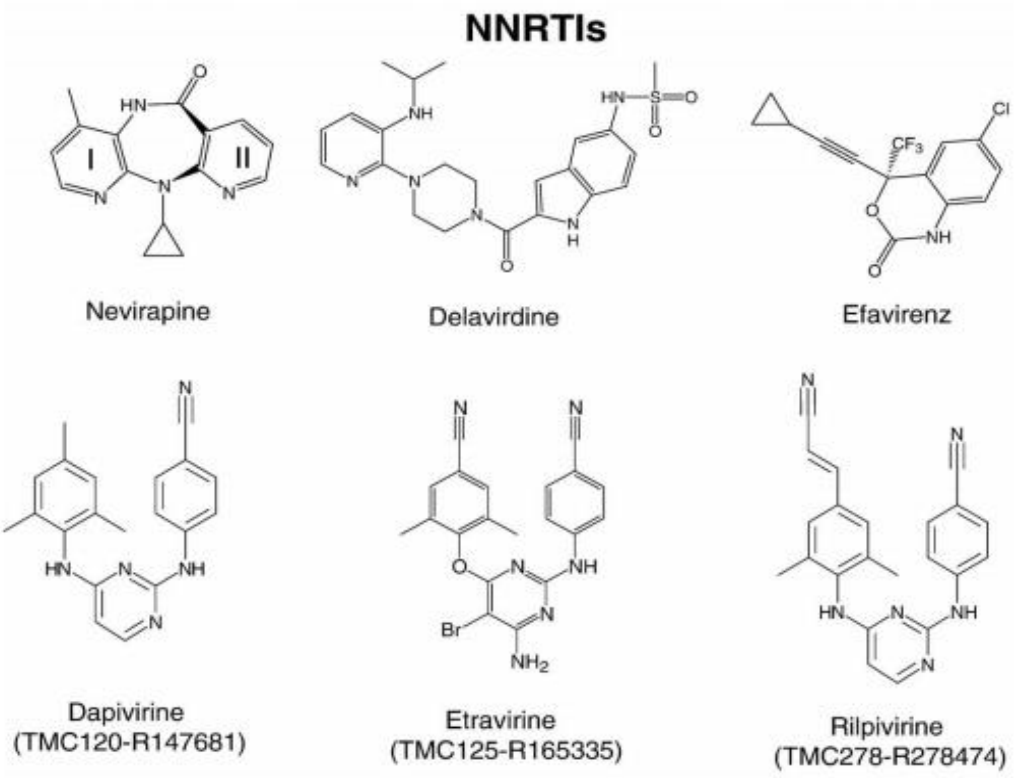
۱-۴-۶-۳ بازدارنده‌های غیرنوکلئوزیدی نسخه‌بردار معکوس[1]

این دسته از مهارکننده‌ها بازدارنده‌های غیر رقابتی هستند و به بخش هیدروفوبی جایگاه فعال متصل می‌شوند. نویراپین[2] نخستین گروه از این مهارکننده‌ها بود که در سال ۱۹۹۶ توسط وزارت بهداشت کانادا مورد تایید قرار گرفت. این نوع مهارکننده‌ها نیز فرایند رونویسی را مختل می‌کنند. با توجه به شکل(۱-۵) آن‌ها با چسباندن خود به آنزیم نسخه‌بردار معکوس و محدود کردن آنزیم، مانع از نسخه‌برداری می‌شوند. شکل(۱-۶) نمونه‌هایی از این دسته از بازدارنده‌ها را به خوبی نشان می‌دهد[۷،۶].



شکل(۱-۵) نحوه عملکرد بازدارنده‌های غیر نوکلئوزیدی آنزیم نسخه بردار معکوس [۶]

[1] Non Nucleoside Reverse Transcriptase Inhibitors (NNRTIs)
[2] Nevirapin



شکل(۱-۶) نمونه هایی از مهار کننده‌های غیر نوکلئوزیدی [۷]

۱-۴-۶-۴ بازدارنده‌های نوع پروتئاز[1]

این داروها اولین بار در سال ۱۹۹۵ تایید شدند. پروتئاز یک آنزیم گوارشی است که پروتئین را تجزیه می‌کند و تقریباً در هر سلول زنده‌ای وجود دارد. آنزیم پروتئاز در HIV به زنجیره طویل آنزیم‌ها و پروتئین‌ها، در سلول حمله کرده و آن‌ها را به قطعات کوچک‌تری تقسیم می‌کند تا از این طریق پروتئین‌های لازم برای تولید یک ویروس بالغ تولید شود. ولی مهارکننده پروتئاز به مکان‌هایی از آنزیم که بریدن[2] زنجیره پلی‌پپتیدی از آن مکان رخ می‌دهد، می‌چسبد و مانع از تولید پروتئین‌های لازم برای تولید ویروس جدید می‌شود. به‌عبارتی، مهارکننده پروتئاز قادر است فرایند تبدیل ویروس نابالغ غیرعفونی را به ویروس بالغ عفونی، آهسته کند [۶].

[1] Protease inhibitors(PI)
[2] Cut

۱-۵ مشتقات آزابنزن‌ها[1] (آزین‌ها)

آزابنزن‌ها، مولکول‌هایی مشـــتق شـــده از بنزن هســـتند که گروه‌های CH با اتم‌های نیتروژن جایگزین می‌شوند. نیتروژن‌ها می‌توانند به صورت پیوسته (۱،۲-۱،۲،۳...) و گسسته (۱،۳-۱،۴،...) در حلقه توزیع شـــوند. پیریدین ترکیبی پرکاربرد با یک نیتروژن، ۱،۲-دی آزابنزن (پیریدازین)، ۱،۳-دی آزابنزن (پیریمیـدین)، ۱،۴-دی آزابنزن (پیرازین) بـا دو نیتروژن، بـه‌عنوان ترکیبـات پـایـدار و آروماتیک‌های شـــناخته شـــده هســـتند. ۱،۳،۵- تری آزین‌ها نیز با ســـه اتم نیتروژن در ســـال ۱۹۵۴ ساختارشان تصحیح و مورد استفاده قرار گرفت. سیستم‌هایی با چهار تا شش نیتروژن موجود در حلقه نیز مورد توجه قرار گرفته و نوع تترا آزابنزن کاملا شناخته شده ا ست اما پنتا آزابنزن و هگزا آزابنزن آن‌چنان شناخته شده نیستند[۸].

همان‌طور که گفته شـــد مهارکننده‌های غیر نوکلئوزیدی با هدف قرار دادن آنزیم نســـخه‌بردار معکوس، در درمان عفونت HIV مؤثر هستند. مشتقات آزین‌ها(پیریدین‌ها،پیریدازین‌ها، پیریمیدین‌ها، پیرازین‌ها و تری آزین‌ها و ...) به‌عنوان دســـته جدیدی از مهارکننده‌های غیرنوکلئوزیدی شـــناخته شده‌اند.

۱-۶ مروری بر کارهای انجام شده و ضرورت تحقیق

همانطور که گفته شـد بازدارنده‌های بیماری ایدز به ســه دســته کلی بازدارنده‌های نوکلئوزیدی نســـخه بردار معکوس و بازدارنده‌های غیر نوکلئوزیدی نســـخه بردار معکوس و بازدارنده‌های پروتئاز تقسیم بندی می‌شوند. مشتقات مورد استفاده در این تحقیق در دسته بازدارنده‌های غیر نوکلئوزیدی نســـخه بردار معکوس قرار دارند. از این رو در این بخش مقالات مربوط به بررســـی ســـاختار- فعالیت مربوط به بازدارنده‌های غیر نوکلئوزیدی نســـخه بردار معکوس در طی ســـال‌های ۲۰۱۷-۲۰۱۰ مورد

[1] Azabenzenes (azines)

بررسی قرار گرفته‌اند.

محققان مختلف در طول این سال‌ها توانستند با استفاده از مطالعات QSAR مدل‌های خطی و غیر خطی متفاوتی را برای پیش‌بینی داروهای غیر نوکلئوزیدی نسخه بردار معکوس ارائه دهند به طوری‌که از نظر آماری نیز مدل‌های قابل قبولی به شمار می‌آیند[۹-۳۱].

با مروری بر مقالات سال‌های ۲۰۱۷-۲۰۱۰ مشاهده شده است که برخی از محققان نیز به بررسی ارتباط ساختار-فعالیت داروهای غیر نوکلئوزیدی نسخه بردار معکوس و مدل سازی QSAR با استفاده از شبکه عصبی مصنوعی پرداخته‌اند که در این پروژه نیز یکی از روش‌های مدل سازی مورد استفاده برای بررسی ساختار فعالیت داروهای غیر نوکلئوزیدی نسخه بردار معکوس بوده است.

چرکائویی و همکارانش[1] در سال ۲۰۱۰ با استفاده از روش غیر خطی شبکه عصبی و خطی ماشین بردار پشتیبان مدل QSAR را برای فعالیت ضد HIV مشتقات بنزودیازپینون به عنوان مهارکننده‌های غیر نوکلئوزیدی نسخه بردار معکوس ارائه دادند. ضریب همبستگی برای مدل خطی برابر با ۰/۹۶ و برای مدل غیر خطی ۰/۹۰ به‌دست آمد.[۳۲]

ساریپینار و همکارانش[2] در سال ۲۰۱۲ با استفاده از مدل غیر خطی شبکه عصبی فعالیت ضد HIV مشتقات تیمین را به عنوان مهارکننده‌های غیر نوکلئوزیدی نسخه بردار معکوس ارائه دادند. ضریب همبستگی داده‌های سری آزمون برای مدل خطی برابر با ۰/۹۲ به‌دست آمد.[۳۳]

توروپوا و همکارانش[3] در سال ۲۰۱۴ با استفاده از مدل غیر خطی شبکه عصبی فعالیت ضد HIV مشتقات تیمین را به عنوان مهارکننده‌های غیر نوکلئوزیدی نسخه بردار معکوس ارائه دادند. ضریب همبستگی داده‌های سری آزمون برای مدل خطی برابر با ۰/۸۸ به‌دست آمد.[۳۴]

[1] Cherqaoui and et al.
[2] Saripinar and et al.
[3] Toropova and et al.

نیظامی و همکارانش[1] در سال ۲۰۱۵ با استفاده از مدل غیر خطی شبکه عصبی فعالیت ضد HIV ۲۸۹ ترکیب از مشتقات پیریمیدین را به عنوان مهارکننده‌های غیر نوکلئوزیدی نسخه بردار معکوس ارائه دادند. ضریب همبستگی داده‌های سری آزمون برای مدل خطی برابر با ۰/۸۷ به‌دست آمد.[۳۵]

عرب چم جنگلی[2] و همکارانش در سال ۲۰۱۵ با استفاده از روش غیر خطی و خطی مدل QSAR را برای فعالیت ضد HIV مشتقات تیازول به عنوان مهارکننده‌های غیر نوکلئوزیدی نسخه بردار معکوس ارائه دادند. ضریب همبستگی برای مدل خطی چندگانه[3] برابر با ۰/۷۷ و برای مدل غیر خطی شبکه عصبی مصنوعی ۰/۹۱ به‌دست آمد.[۳۶]

با توجه به کارهای انجام شده برای مدل‌سازی با ترکیبات دارویی ضد ایدز به عنوان مهار کننده‌های غیر نوکلئوزیدی نسخه بردار معکوس اهمیت بر هم کنش لیگاند-پروتئین در نظر گرفته نشده است. در نظر گرفتن بر هم کنش بین لیگاند-پروتئین و محاسبه انرژی اتصال کمپلکس لیگاند-پروتئین ممکن است اطلاعات مفیدی را در اختیار طراح دارو قرار دهد از این رو استفاده از این روش کمک شگرفی را در امر طراحی دارو خواهد داشت.

در این پایان نامه، ترکیبات جدیدی به عنوان مهار کننده‌های غیر نوکلئوزیدی نسخه بردار معکوس، برای مدل‌سازی مورد استفاده قرار گرفته است و برای اولین بار بر هم کنش مشتقات آزابنزن (آزین‌ها) و پروتئین با استفاده از داکینگ مولکولی ارزیابی می‌شود و توصیف کننده‌های جدید پیشنهاد داده می‌شود و با توصیف‌کننده‌های حاصل از ساختار مشتقات مورد مطالعه ترکیب می‌شود. از روش انتخاب متغیر رگرسیون گام به گام و روش نوین منقبض کننده حداقل قدر مطلق و عملگر انتخاب کننده[4] (لاسو) برای انتخاب بهترین متغیرها از مجموع متغیرهای حاصل از داکینگ مولکولی و

[1] Nizami and et al.
[2] Arab chamjangali and et al.
[3] Multiple linear regression
[4] Least Absolute Shrinkage and Selection Operator

توصیف‌کننده‌های ساختاری محاسباتی استفاده می‌شود و مدل‌سازی به وسیله جنگل‌های تصادفی[1] و شبکه عصبی مصنوعی[2] با استفاده ازمتغیرهای حاصل از برهم کنش لیگاند و پروتئین و هم‌چنین متغیرهای ساختاری محاسباتی مشترک بین مشتقات که با روش انتخاب متغیر جدید انتخاب می‌شوند انجام می‌شود و مدل‌های QSAR با توصیف‌کننده‌های جدید ایجاد خواهد شد.

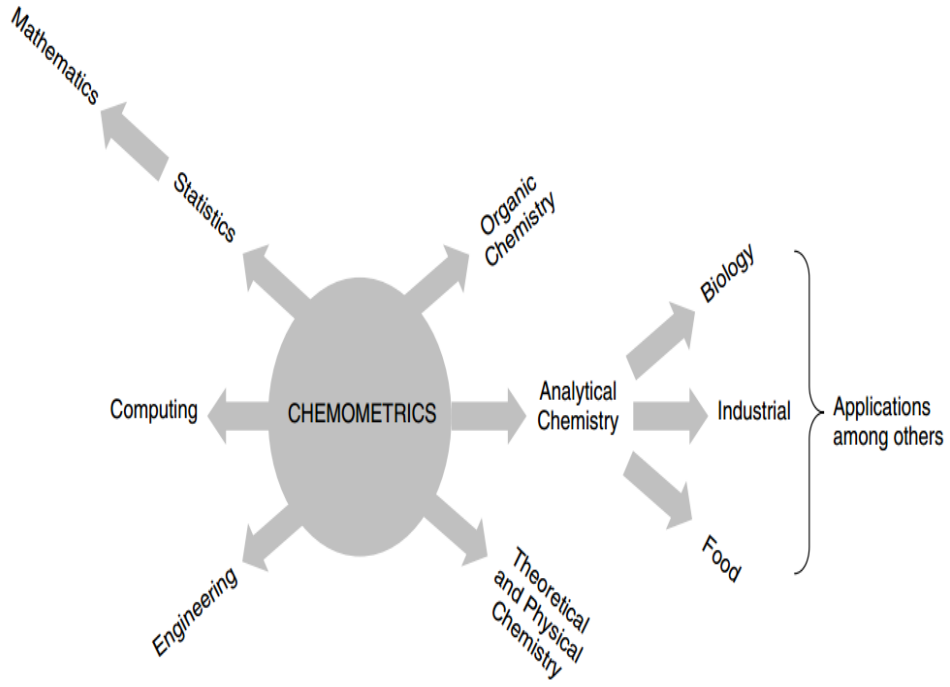
[1] Rndom Forest
[2] Artificial Neural Networks

۲ فصل دوم

کمومتریکس

۲-۱ مقدمه

هنر اســـتخراج اطلاعات مربوط به شـــیمی از داده‌های تولید شـــده از آزمایشـــات شـــیمیایی کمومتریکس نام دارد. کمومتریکس یکی از شـاخه‌های با اهمیت علمی اسـت که کاربردهای زیادی در بسیاری از شاخه‌های علوم مانند آمار چند متغیره، مدل سازی ریا ضی، علوم کامپیوتر و شیمی تجزیه دارا ست. این مو ضوع باعث شده ا ست تا در بـ سیاری از منابع علمی از کمومتریکس به عنوان فرزند ریا ضیات، کامپیوتر و شیمی نام برده شود[۳۸،۳۷]. بنابراین گ سترش سریع فناوری به خ صوص در قســـمت‌های مربوط به کاربرد کامپیوتر در شـــیمی تجزیه موجب پیشـــرفت‌های ســـریعی در زمینه کمومتریکس در چند سال اخیر شده ا ست. شکل(۲-۱) ارتباط کمومتریکس با سایر علوم را نـ شان می‌دهد[۳۹].



شکل(۲-۱) ارتباط کمومتریکس و علوم [۳۷]

طراحی دارو از جمله کاربرهای کمومتریکس اســـت. امروزه پیشـــرفت در علوم کامپیوتر و روش‌های محا سباتی منجر به جایگزینی روش‌های منطقی و محا سباتی به‌جای روش‌های سنتی و تـ صادفی در طراحی دارو شده ا ست، این مو ضوع باعث صرفه‌جویی در زمان و هزینه‌های آزمایـشگاهی و پژوهـشی

می‌شود. هدف از این فصل بیان تعاریف و مفاهیم مقدماتی در فرآیند طراحی محاسباتی دارو است.

اصطلاح کمومتریکس برای اولین بار توسط اسوانت ولد[1] دانشمند جوان سوئدی که در زمینه شیمی فیزیک آلی فعالیت دا شت، مطرح گردید. همکاری ولد با بروس آر. کووالسکی[2] که در دانشگاه وا شنگتن بر روی الگو شنا سی در شیمی تجزیه مطالعه می‌کرد، منجر به تأ سیس انجمن بین المللی کمومتریکس (ICS)[3] در سال۱۹۷۴ گردید[۳۸]. بنا به تعریف ICS، کمومتریکس عبارت است از کاربرد روش‌های ریاضی، آماری برای برقراری ارتباط بین سنجش‌های انجام شده روی یک سیستم یا فرایند شیمیایی به‌منظور درک بهتر اطلاعات شیمیایی است.

برخی از کاربردهای کمومتریکس عبارتند از:

۱. بهینه سازی
۲. تخمین پارامترهای سینتیکی
۳. بهینه‌سازی پارامترهای کنترل کیفی
۴. بررسی رفتار کروماتوگرافی غیر خطی
۵. پردازش سیگنال
۶. طراحی دارو
۷. کالیبراسیون غیر خطی و مدل‌سازی[۳۷].

در میان کاربردهای مختلف کمومتریکس یکی از مهم‌ترین و شاخص‌ترین این کاربردها طراحی دارو می‌باشد که به بررسی ارتباط کمّی ساختار- فعالیت (QSAR[4]) می‌پردازد، که خواص مولکول‌ها را به ویژگی‌های ساختاری آن‌ها نسبت می‌دهد[۳۹]. در ادامه به بحث در مورد QSAR، انواع روش‌های QSAR، کاربرد آن در طراحی دارو و مراحل مختلف مطالعات QSAR پرداخته خواهد شد.

[1] Svante wold
[2] Bruce R. Kowaski
[3] International Chemometrics Society
[4] Quantitative structure- activity relationship

۲-۲ ارتباط کمی ساختار -فعالیت (QSAR)

۲-۲-۱ تعریف

ارتباط کمی ساختار –فعالیت در ساده‌ترین تعریف یک روش برای ساخت مدل‌های محا سباتی و آماری است، که تلاش می‌کند ارتباط آماری معناداری را بین ساختار و عملکرد برقرار کند. مولکول‌های م شابه با تغییر کوچکی در ساختار شان می‌توانند فعالیت بیولوژیکی کاملاً متفاوتی دا شته با شند و در واقع QSAR نیز سعی در پیدا کردن رابطه‌ی هماهنگ میان فعالیت بیولوژیکی و ویژگی‌های مولکولی، به‌منظور کاربرد این قواعد برای ارز یابی فعالیت ترکیبات جد ید دارد[۳۷-۴۲]. بررســـی بیولوژیکی فراورده‌های طبیعی و برخی از مشـــتقات آن‌ها که خواص دارویی دارند، انگیزه ای جهت پیدا کردن وابستگی‌های احتمالی ساختار شیمیایی با فعالیت بیولوژیکی ایجاد نمود و م شخص شد که برخی از واحدهای ســـاختار مولکول‌هایی که از نظر بیولوژیکی فعال هســـتند، در مولکول‌های دیگری که دارای همان ویژگی بیولوژیکی می‌باشـــند، وجود دارد. این امر رهنمودی برای طراحی ســـاختار ترکیباتی با فعالیت بیشتر و در عین حال عملکردی اختصاصی‌تر می‌باشد. در حال حاضر نیز تغییر و اصلاح ساختار یک ترکیب الگو، عمده‌ترین روش در دست‌یابی به داروهای جدید می‌باشد.

تغییرات ســاختاری منجر به پیدایش ویژگی‌های جدید می‌شــود که در نهایت موجب تغییر پخش دارو در ســـلول بافت، تغییر در کیفیت اثر متقابل با جایگاه های فعال آنزیم و گیرنده ها و در نهایت تغییرسرعت واکنش‌ها در جایگاه‌های مورد نظر می‌گردد[۴۳].

روش‌های QSAR را می‌توان بر اســـاس معیارهای گوناگونی طبقه بندی کرد که رایج ترین این طبقه بندی‌ها در بخش بعدی مطرح شده است.

۲-۲-۲ طبقه بندی بر اساس نوع روش‌های کمومتریکس

بعضـــی مواقع روش‌های QSAR بر اســـاس نوع تکنیک به‌کار رفته برای ایجاد رابطه‌ی بین ویژگی‌های ساختاری و فعالیت بیولوژیکی(نوع مدل‌سازی) تقسیم می‌شوند.

- ✓ روش‌های خطی مانند تحلیل مؤلفه‌های ا صلی (PCA) [۱] ، رگر سیون کمترین توان‌های دوم جزئی (PLS)[۲] ، رگرسـیون خطی چندگانه (MLR)[۳] وجنگل‌های تصـادفی (RF)[۴] و...
- ✓ روش‌های غیر خطی مانند شـبکه عصـبی مصـنوعی (ANN)[۵] ، الگوریتم نزدیک ترین همسایه ها (kNN)[۶] و ...[۴۴]

۲-۲-۳ طبقه بندی بر اساس نوع اطلاعات مورد استفاده در مدل‌سازی

QSAR را می‌توان به دو د سته کلی QSAR مبتنی بر لیگاند و QSAR مبتنی بر ساختار پروتئین نیز تقسیم بندی کرد.

۲-۲-۳-۱ QSAR مبتنی بر لیگاند[۷]

اکثر مطالعات QSAR انجام گرفته با استفاده از روش مبتنی بر لیگاند بوده است. این روش به‌طور معمول از اطلاعات مجموعه‌ای از ترکیبات با فعالیت مشخص استفاده می‌کند. اغلب، این روش در مواردی که اطلاعات ساختاری پروتئین هدف وجود نداشته باشد استفاده می‌شود. مفهوم کلیدی در روش‌های مبتنی بر لیگاند این است که ترکیباتی که ساختار مشابه و یا اجزای ساختاری مشابه نسبت به ترکیبات فعال شناخته شده دارند، به احتمال زیاد فعالیتی همانند ترکیبات مشابه خود دارند[۴۵].

[1] Principal component analysis
[2] Partial least squares regression
[3] Multiple Linear Regression
[4] Random forest
[5] Artificial neural network
[6] k-Nearest-neighbors
[7] Structure-based drug design

۲-۲-۳-۲ QSAR مبتنی بر ساختار پروتئین

آگاهی از ساختار هدف بیولوژیکی، به ویژه جایگاه اتصال پروتئین، از دیدگاه طراحی دارو بسیار حائز اهمیت است، زیرا اطلاعات دقیق برهم‌کنش اتصال لیگاند-پروتئین می‌تواند در طراحی داروها استفاده شود. استفاده از این روش با عنوان طراحی داروی منطقی[۱] یا طراحی داروی مبتنی بر ساختار پروتئین شناخته می‌شود. این روش به میزان اطلاعاتی که راجع به دارو و پروتئین در دسترس می‌باشد وابسته است. در حال حاضر، بلور نگاری اشعه ی ایکس[۲] و طیف سنجی NMR[۳] عمده ترین روش‌های تجربی تعیین ساختار ماکرومولکول‌های زیستی در سطح اتمی به شمار می‌رود. ایده‌ی اصلی و بنیان روش NMR این است که سطوح انرژی هسته‌های اتم به وسیله‌ی میدان مغناطیسی شکافته می‌شود و انتقال بین این سطوح انرژی از طریق برانگیختن نمونه با تابش امواج رادیویی امکان پذیر است[۴۶].

بلور نگاری اشعه‌ی ایکس اطلاعات بسیار دقیقی در سطح اتمی فراهم می‌سازد و می‌تواند مکان هر اتم در پروتئین، ا سیدهای نوکلئیک و یا مکان مهارکننده‌ها، یون‌ها و لیگاندهای ات صالی به مولکول هدف را مشخص سازد. بلوری شدن، فرایند دشواری است و تمامی پروتئین‌ها دارای قابلیت ایجاد بلور نیستند؛ حتی اگر بلور نیز ایجاد شود، ممکن است دارای کیفیت مناسب نباشد. بنابراین در نوع پروتئین مورد استفاده برای فرایند بلوری شدن محدودیت وجود دارد. به طور کلی، بلور نگاری اشعه‌ی ایکس برای تعیین ساختار پروتئین‌های کمتر انعطاف پذیرکه بلورهای منظم ایجاد می‌کنند، مناسب است. برای مطالعه‌ی پروتئین‌هایی با انعطاف پذیری بالا مشکلاتی وجود دارد. از آنجا که چگالی الکترونی مناطق انعطاف پذیر پروتئین در منطقه‌ی وسیعی پخش می‌شود غالبا در نقشه‌ی چگالی الکترونی دیده نشده و در نهایت در ساختار نیز لحاظ نمی‌شوند[۴۷].

[1] Rational drug design
[2] X-ray crystallography
[3] Nuclear magnetic resonance

اما ممکن است هیچ نوع اطلاعات تجربی راجع به آن‌ها وجود نداشته باشد که در چنین مواردی باید از روش‌های محاسباتی استفاده کرد. با داشتن این اطلاعات مساله پیش رو عبارت است از طراحی دارویی که به‌طور مناسبی با پروتئین که جایگاه هدف می‌باشد، برهم‌کنش برقرار کند که به این نوع طراحی داکینگ مولکولی[1] می‌گویند. داکینگ دارو و پروتئین می‌تواند از طریق برهمکنش‌های آب‌گریز، الکترواستاتیک، پیوندهای هیدروژنی و واندروالس باشد. امروزه استفاده از ابزارهای محاسباتی چون مطالعات داکینگ مولکولی می‌تواند اطلاعات ارزشمندی در جهت طراحی دارو مفید و موثر در اختیار بگذارند. مطالعات داکینگ مولکولی اطلاعاتی نظیر نحوه برهم‌کنش دارو و پروتئین، جهت گیری دارو در سایت اتصال پروتئین و انرژی آزاد اتصال دارو به پروتئین را فراهم می‌کنند[۴۵]. انرژی اتصال لیگاند و پروتئین، ثابت بازداری تئوری، انرژی الکترواستاتیک و برخی از اطلاعات خروجی داکینگ مولکولی از مجموعه اطلاعات ارزشمندی است که در این پروژه به‌عنوان توصیف‌کننده مولکولی مورد بررسی قرار گرفته است.

۲-۳ شبیه سازی داکینگ مولکولی

طراحی دارو یکی ازچالش بر انگیزترین پژوهش‌ها در دنیای علم کنونی ا ست و پیشرفت در این زمینه نیازمند پیشرفت در سایر حوزه‌های دانش نیز می‌باشد.

اصولا در طراحی دارو از سه محیط برای مطالعه استفاده می شود که عبارت اند از :

۱. In silico: در این محیط مولکول در ابزار های کامپیوتری مورد مطالعه قرار می‌گیرد.
۲. In vitro: مطالعه مولکول در محیط کشت سلولی می‌باشد.
۳. In vivo: مطالعه مولکول در بدن جانوران مانند موش آزمایشگاهی صورت می‌گیرد.

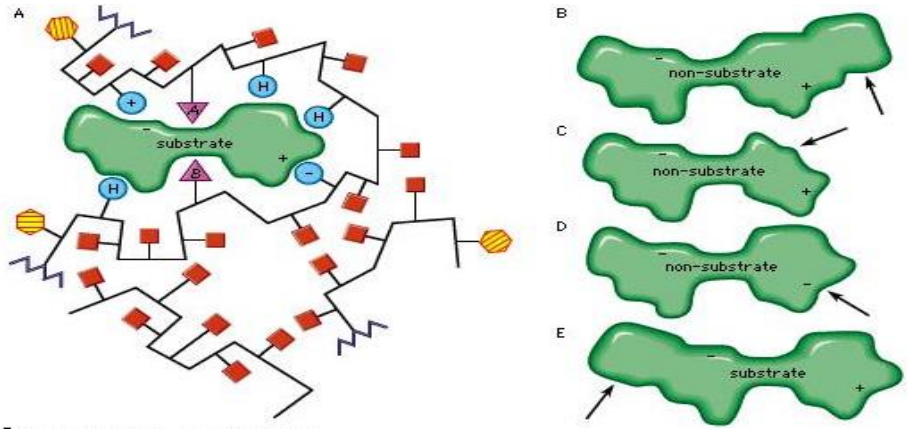
دراین میان شبیه سازی‌ها و پژوهش‌های in silico (انجام شبیه سازی با ا ستفاده از کامپیوتر) به دلیل دسترس پذیری و هزینه پایین، در برابر دیگر آزمایش‌ها جایگاه ویژه خود را دارند. مسیر شبیه

[1] Molecular docking

سازی دقیق و اصولی عملکرد یک دارو شامل مراحل شبیه سازی داکینگ مولکولی[1]، دینامیک مولکولی[2] و ... می‌باشد[۴۹،۴۸].

اتصال یک دارو به گیرنده پروتئینی[3] خود فرایند ساده‌ای نیست و به پارامترهای ترمودینامیکی مختلفی وابسته است. علاوه بر آن در طی این فرایند پدیده‌های گوناگونی نظیر حرکت و جنبش دارو و گیرنده آن به‌منظور یافتن محل پیوند، تغییر در شکل و ساختار[4] سه بعدی گیرنده و یا دارو به منظور آ شکار شدن محل پیوند و تاثیر مولکول‌های آب و یون‌های محلول در آن بر روی گیرنده پروتئینی، بر روند شبیه سازی تاثیر گذار می‌باشند[۴۸].

در بیشتر موارد دارو از طریق مهار رقابتی یعنی اتصال به جایگاه فعال و جلوگیری از ورود سوبسترا به آن جایگاه عمل می‌کند از این موضوع به‌عنوان تئوری قفل و کلید در عملکرد دارو نام برده می‌شود. وجود موقعیت فضایی خاص بر هم‌کنش دارو با گیرنده در سال ۱۸۹۴ توسط فیشر کشف شد[۴۴]. او الگوی مشهور قفل و کلید را بیان کرد و دارو را همانند یک کلید و گیرنده را همانند قفل معرفی نمود. که دارو به عنوان کلیدی می‌تواند فعالیت گیرنده را افزایش یا کاهش دهد[۴۹].

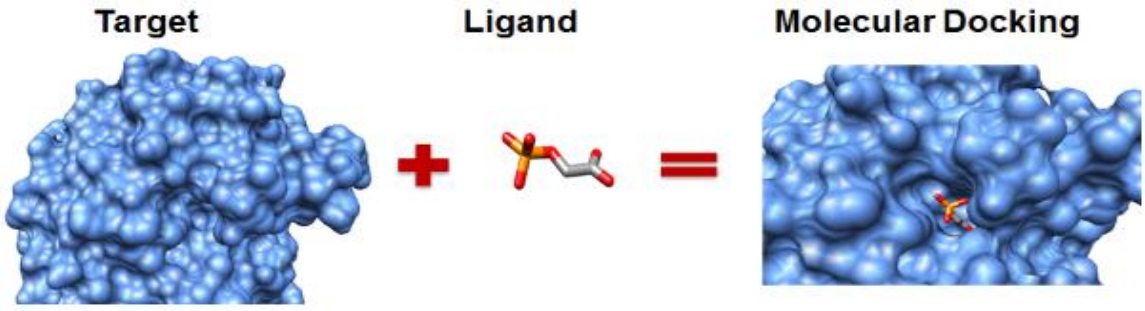


شکل(۲-۲) طرح شماتیک از مدل قفل و کلید

[1] Molecular docking simulation
[2] Molecular Dynamics (MD)
[3] Protein receptor
[4] Conformation

روش محاسباتی مورد استفاده که مولکول‌های کوچک را درون ساختار ماکرو مولکول‌ها قرار داده داک[1] نام دارد و امتیازدهی[2] انرژی حاصل از بر هم کنش این مولکول‌های کوچک در ساختار ماکرو مولکول‌ها داکینگ نام دارد.

داکینگ یک الگوریتم خودکار کامپیوتری است که نحوه‌ی اتصال ترکیب به جایگاه فعال پروتئین را مشخص می‌کند. این روش شامل تعیین جهت‌گیری[3] و موقعیت[4] ترکیب، ساختارهندسی کنفورماسیونی و امتیازدهی می‌باشد. امتیازدهی می‌تواند معیار اندازه‌گیری انرژی اتصال، انرژی آزاد یا یک معیار عددی باشد. هر الگوریتم خودکار داکینگ به نحوی تلاش می‌کند تا ترکیب را در جهت‌گیری‌ها و کنفورماسیون‌های متفاوت در جایگاه فعال قرار دهد و امتیازی را برای هرکدام محاسبه کند. با استفاده از دانش به‌دست آمده از مطالعات داکینگ، نیاز به سنتز و بررسی ترکیبات کمتری است. دلیل اصلی برای استفاده از داکینگ، پیش‌بینی ترکیباتی است که به خوبی به پروتئین متصل می‌شوند وعلاوه بر این مشاهده ساختار هندسی ترکیب متصل شده به جایگاه فعال پروتئین است و شکل(۲-۳) نمایشی از اتصال لیگاند به ماکرومولکول می‌باشد[۵۰]. ذکر این نکته ضروری است که در این پروژه از مجموعه اطلاعات حاصل از اتصال پروتئین و لیگاند به عنوان توصیف‌کننده استفاده شده است.



شکل(۲-۳) اتصال لیگاند به ماکرو مولکول در طی فرایند داکینگ مولکولی [۵۰]

[1] Dock
[2] Scoring
[3] Orientation
[4] Position

۲-۴ مراحل انجام داکینگ مولکولی

داکینگ مولکولی به دو صورت انجام می‌گیرد:

۱. داک لیگاند درون پروتئین با پروتئین جهت اعتبار سنجی فرایند داکینگ مولکولی
۲. داک لیگاندهای مشابه با لیگاند درون پروتئین در شرایط تعیین شده توسط فرایند اعتبار سنجی به جهت مدل‌سازی

در هر دو مورد مذکور انجام مراحل زیر ضروری است.

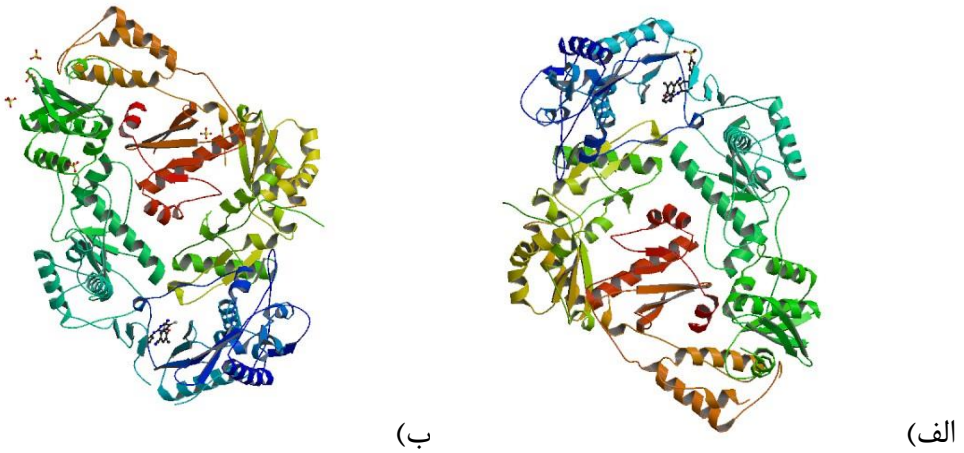
۲-۴-۱ فرایند اعتبار سنجی

قبل از هر عملیات داکینگ گیرنده-لیگاند، اعتبار سنجی برای فرایند داکینگ مولکولی ضروری است. در این مرحله پس از آماده سازی پروتئین که بدون مولکول‌های آب و کوفاکتورهاست و هم‌چنین لیگاند موجود در پروتئین به طور جداگانه در یک فایل جدید ذخیره می‌شود به عبارتی یک بار لیگاند را از گیرنده خارج و دوباره به آن وارد می‌شود و سپس این دو با هم داک می‌شوند که به این عملیات داک-ریداک[1] یا خود-داکینگ[2] هم می‌گویند و به این ترتیب بهترین حالت اتصال محاسبه شده توسط نرم افزار با آنچه که از ساختار کریستالوگرافی به دست آمده است مقایسه می‌شود. به‌طوری‌که باید میانگین ریشه توان‌های دوم انحراف[3] (RMSD) حاصل از فرایند اعتبار سنجی کمتر از مقدار $۲/۰A^o$ به‌د ست آید تا فرایند داکینگ دارای اعتبار با شد، به‌طوری‌که در طی این فرایند ساختار کریستالوگرافی مجددا با لیگاند خودش داک می‌شود و RMSD آن محاسبه می‌گردد. RMSD نشان‌دهنده‌ی تناظر نقطه به نقطه دو ترکیب با هم ا ست و میزان اختلاف دو ساختار را در حالت سه بعدی نشان می‌دهد. RMSD اتصال لیگاند و گیرنده باید از $۲A^o$ کمتر باشد تا بتوان داکینگ را با شرایط تعریف شده ادامه داد و اگر نه باید شرایط را تا جایی‌که به زیر $۲A^o$ بیاید تغییر داد. با توجه به شرایط بهینه حاصل از این فرایند، داکینگ سایر لیگاندها در پروتئین با همین شرایط انجام می‌شود.

[1] Dock-redock
[2] Selfdocking
[3] Root mean square deviation

۲-۴-۲ آماده سازی پروتئین

صحت نتایج داکینگ به‌طور م ستقیم به کیفیت ساختار کریـ ستالوگرافی جایگاه فعال پروتئین وابسته است. ساختار کریستالوگرافی پروتئین‌ها با اشعه ایکس مشخص می‌شوند. اگر ساختار همراه با لیگاند کمپلکس شـــده با پروتئین و یا بدون لیگاند و کوفاکتور و فقط دارای بخش پروتئینی در جایگاه فعال فراهم شـوند سـاختارهای "هولو"[۱] و سـاختار "آپو"[۲] گفته می‌شـود. سـاختارهای هولو مهمترین روش شـــناســـایی جایگاه فعال پروتئین هســـتند. مرکز ثقل لیگاندی که در پروتئین اســت همان مرکز تقریبی جایگاه فعال پروتئین می‌باشـــد. از آن‌جا که حتی بهترین ســـاختارهای کریســـتالوگرافی اغلب ارزش تفکیک یک آنگســـتروم یا بیشـــتر دارند، ابتدا لیگاند و مولکول‌های آب موجود در ســـاختار کریســـتالوگرافی را حذف کرده و با فرمت بانک اطلاعاتی پروتئین[۳] ذخیره می‌کنند و پس از ورود به محیط برنامه، هیدروژن‌هایی را که تو سط کریـ ستالوگرافی پروتئین دیده نمی شوند، ا ضافه می‌نمایند. شکل(۲-۴) ساختارهای کریستالوگرافی دو پروتئین استفاده شده در این پروژه را نشان می‌دهد[۵۱].



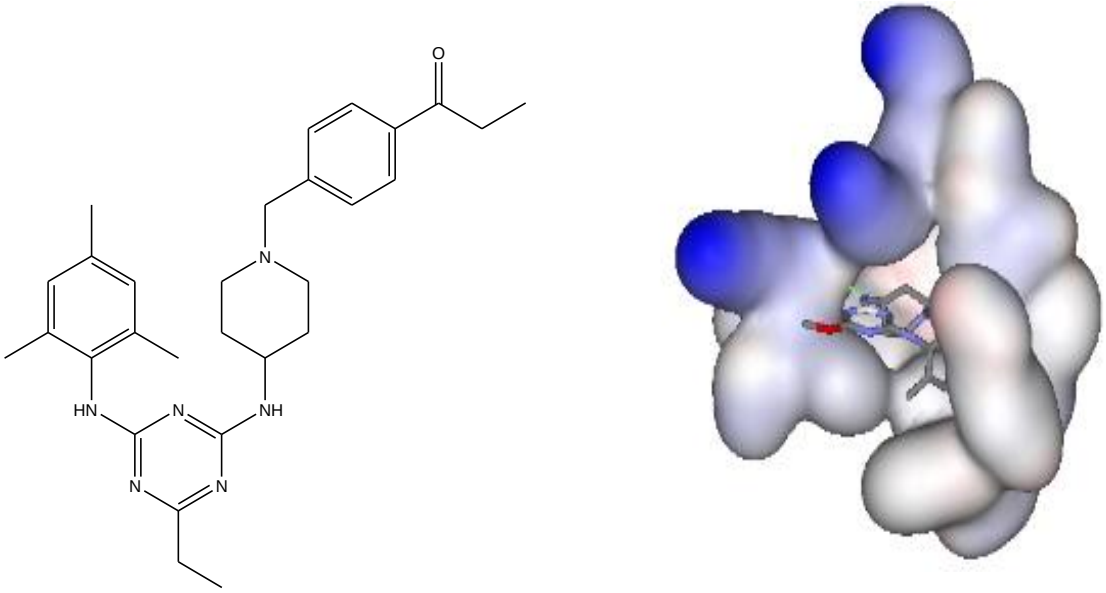
الف) ب)

شکل(۲-۴) ساختار کریستالوگرافی پروتئین الف) کد کریستالوگرافی 3MEC ب) کد کریستالوگرافی 3M8Q [۵۱]

[1] Soaked structure
[2] Apo structure
[3] Protein data bank

۲-۴-۳ ساختن لیگاند

لیگاند در برنامه داکینگ فراخوانی شده و در جایگاه فعال پروتئین هدف قرار می‌گیرد و سپس هر شبیه سازی داکینگ به‌طور جداگانه برای همه لیگاندها توسط کاربر انجام می‌شود. نمایشی از کمپلکس لیگاند با جایگاه فعال در شکل(۲-۵) آورده شده است.



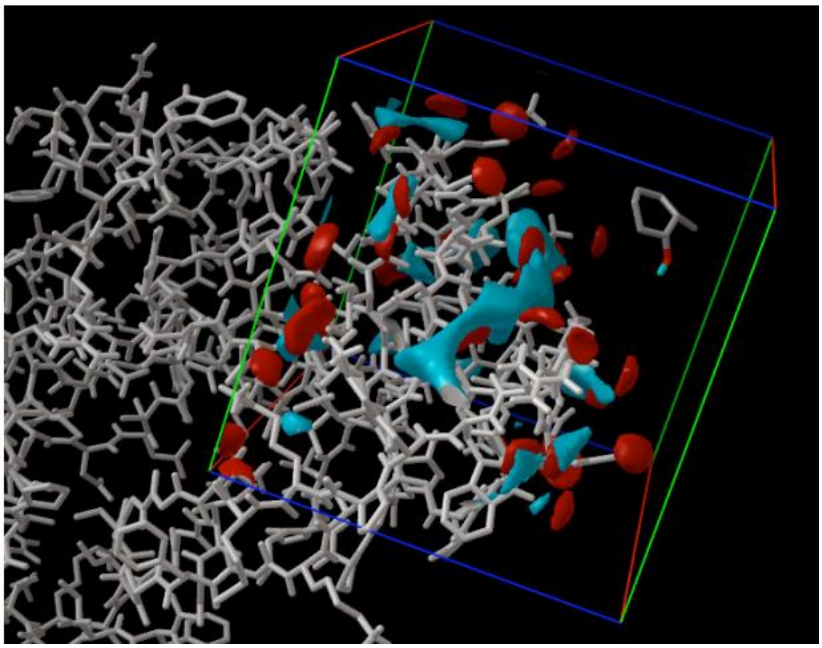
شکل(۲-۵) ساختار کمپلکس لیگاند و پروتئین

۲-۴-۴ تنظیم کردن جعبه شبکه‌ای[1]

در داکینگ مولکولی نمی‌توان کل فضا را در نظر گرفت. باید بخشی از فضا که لیگاند درون جایگاه فعال قرار می گیرد را برای فرایند داکینگ تعریف کرد تا هم دقت بالا باشد و هم زمان انجام فرایند داکینگ به صرفه با شد. به‌منظور سرعت بخ شیدن به محا سبات، یک فا صله‌ی محدود کننده[2] تنظیم می‌شود، این فاصله محدود کننده معمولا یک جعبه مستطیل شکل به نام جعبه شبکه‌ای است. بخش‌هایی از پروتئین که دور از جایگاه فعال هستند، به‌طور معمول هیچ اثر قابل اندازه گیری بر نتایج امتیازدهی ندارند و از این رو ایجاد شبکه با ابعاد یک سان برای پو شش دهی ا سید های آمینه جایگاه فعال پروتئین تو صیه می شود. اگر فایل مربوط به ساختار کریستالوگرافی هولو با شد به این معنی که

[1] Grid box
[2] Cut off

دارای لیگاند کمپلکس شده باشد جایگاه فعال پروتئین مشخص خواهد بود و مختصات جعبه شبکه ای همان مختصات جایگاه فعال و به شکل مکعبی با ابعاد یکسان تعریف می شود. ابعاد جعبه باید طوری انتخاب شود که اسیدهای آمینه جایگاه فعال را در بر بگیرد. شکل(۲-۶) جعبه شبکه‌ای در بر گیرنده اسید های آمینه را نشان می‌دهد[۵۳،۵۲].



شکل(۲-۶) جعبه شبکه ای [۵۲]

۲-۴-۵ گزینه‌های داکینگ

در هنگام تنظیم ورودی ها برای یک محاسبه داکینگ، گزینه‌هایی برای جایگاه فعال انعطاف پذیر، روش‌های نمره دهی، روش‌های جستجو، حلال پوشی، برخورد با جایگاه فعال احاطه شده و غیره موجود هستند.

۲-۴-۶ انجام محاسبه‌ی داکینگ

زمانی‌که ورودی‌ها تنظیم شدند می‌توان محاسبات داکینگ را انجام داد. این محاسبات گاهی توسط همان کامپیوتری که صفحه‌ی رابط گرافیکی از آن استفاده شده، انجام می‌شوند و گاهی می‌توانند به یک سرور دیگر فرستاده شوند.

۲-۴-۷ آنالیز و تحلیل نتایج

مهم‌ترین نتیجه محاسبات داکینگ، انرژی اتصال لیگاند به جایگاه فعال است. این مقداری است که برای تعیین بهترین ترکیب مهار کننده، بین ترکیبات مختلف مقایسه می‌شود. چند حالت و وضعیت از لیگاند در جایگاه فعال که با بهترین انرژی اتصال همراه است، به‌طور چشمی بررسی می‌شود تا از معقول و منا سب بودن آن اطمینان حا صل گردد. گاهی اوقات، و ضعیتی که تو سط یک محا سبه داکینگ تولید می‌شود، به محقق ایده‌ای برای چگونگی تغییر ترکیبات در دوره بعدی محاسبات را می‌دهد.

دو مولفه کلیدی الگوریتم جستجو[۱] و الگوریتم رتبه‌بندی یا امتیازدهی[۲] در برنامه داکینگ وجود دارد. الگوریتم جستجو، مولکول را در موقعیت‌ها و صورت‌بندی‌های متفاوت در جایگاه فعال پروتئین قرار می‌دهد. انتخاب الگوریتم جستجو، مشخص می‌کند که برنامه با چه صحتی در طول فرایند، موقعیت‌های ممکن مولکول را چک می‌کند و چه مدت زمانی برای آن نیاز است. قابل ذکر ا ست که الگوریتم جستجو صحت نتایج به‌د ست آمده از برنامه داکینگ را مشخص نمی‌کند. الگوریتم امتیازدهی م سئول تعیین این ا ست که آیا جهت‌گیری‌های انتخاب شده تو سط الگوریتم ج ستجو، از لحاظ انرژی مناسب‌ترین هستند و مسئول محاسبه انرژی اتصال است.

توصیف‌کننده های حاصل از داکینگ مولکولی را می‌توان به صورت زیر دسته بندی کرد: توصیف‌کننده‌های حاصل از برهم‌کنش لیگاند و پروتئین که به صورت انواع مختلفی از انرژی قابل محاسبه هستند:

۱. انرژی آزاد اتصال محاسبه شده[۳] (BE)
۲. انرژی واندروالس(vdW)+ پیوند هیدروژنی(Hbond)+ حلال‌زدایی (desolv)، (VE)
۳. انرژی بین مولکولی نهایی[۴] (IE)

[1] The search algorithm
[2] The scoring algorithm
[3] Estimated Free Energy of Binding
[4] Final Intermolecular Energy

۴. انرژی الکتروستاتیک[۱] (EE)
۵. مجموع انرژی داخلی نهایی[۲] (TI)
۶. انرژی آزاد چرخشی[۳] (TE)
۷. انرژی سیستم‌های آزاد[۴] (UE)
۸. ثابت بازداری (K_i)
۹. منفی لگاریتم ثابت بازداری (-Log K_i)

توصیف‌کننده‌های حاصل از نوع اسیدهای آمینه که در ارتباط با لیگاند در کمپلکس قرار گرفته‌اند:

۱. هیدروفوبیک [۵]
۲. هیدروفیلیک [۶]
۳. هیدروفیلیک اسیدی[۷]
۴. هیدروفیلیک بازی [۸]

توصیف‌کننده‌های حاصل از نوع اسیدهای آمینه بر اساس ساختار اسید آمینه:

۱. دارای زنجیره جانبی آلیفاتیک [۹]
۲. دارای زنجیره‌های جانبی حاوی گروه‌های هیدرکسیلی [۱۰]
۳. دارای زنجیره‌های جانبی حاوی اتم گوگرد[۱۱]
۴. دارای زنجیره‌های جانبی حاوی گروه‌های اسیدی[۱۲] یا آمیدی
۵. دارای زنجیره جانبی حاوی گروه‌های بازی [۱۳]
۶. دارای حلقه‌های آروماتیک[۱۴]
۷. ایمینواسیدها [۱۵]

[1] Electrostatic Energy
[2] Final Total Internal Energy
[3] Torsional Free Energy
[4] Unbound System's Energy
[5] Hydrophobic
[6] Hydrophilic
[7] Acidic Hydrophilic
[8] Basic Hydrophilic
[9] With aliphatic side chain
[10] With side chains containing Hydroxylic groups
[11] With side chains containing Sulfur atoms
[12] With side chains containing acidic groups or their amides
[13] With side chains containing Basic group
[14] Containing Aromatic Rings
[15] Imino Acids

در پروژه پیش رو، سعی بر استفاده همزمان از توصیف‌کننده‌های ساختاری مولکول‌ها و توصیف‌کننده‌های استخراج شده از داکینگ مشتقات آزابنزن و پروتئین بوده تا بتوان از اطلاعات حا صل از برهم‌کنش لیگاند و پروتئین نیز ا ستفاده نمود. برر سی این برهم کنش‌ها حائز اهمیت ا ست زیرا می‌توان از انرژی اتصال مطلوب و پیوندهای برقرار شده با اسیدهای آمینه به اهمیت آن لیگاند پی برد.

۲-۴-۸ انواع داکینگ مولکولی

داکینگ مولکولی انواع متفاوتی دارد که شامل موارد زیر می‌باشد.

۱. برهم‌کنش لیگاند-پروتئین: برهم‌کنش یک مولکول کوچک (لیگاند) به یک ماکرو مولکول (پروتئین)
۲. برهم‌کنش پروتئین-پروتئین: معمولا محل اتصال سطحی بزرگتر نسبت به اتصال لیگاند-پروتئین دارد.
۳. برهم‌کنش پروتئین-DNA

۲-۴-۹ کاربردهای داکینگ مولکولی

داکینگ مولکولی کاربردهای زیادی در مطالعات مربوط به طراحی داروها دارد که شامل موارد زیر است.

۱. غربال گری مجازی[1]
۲. کشف دارو
۳. پیش‌بینی انرژی آزاد اتصال
۴. برهم‌کنش پروتئین- پروتئین یا پروتئین- اسید نوکلئیک
۵. مکانیسم‌های واکنش‌های آنزیمی
۶. مهندسی پروتئین

متداول‌ترین کاربرد داکینگ مولکولی در اتصال پروتئین-لیگاند است. هدف نهایی اتصال پروتئین-لیگاند پیش‌بینی فعالیت بیولوژیکی لیگاند و برهم‌کنش لیگاند با پروتئین می‌باشد[۴۵] و در

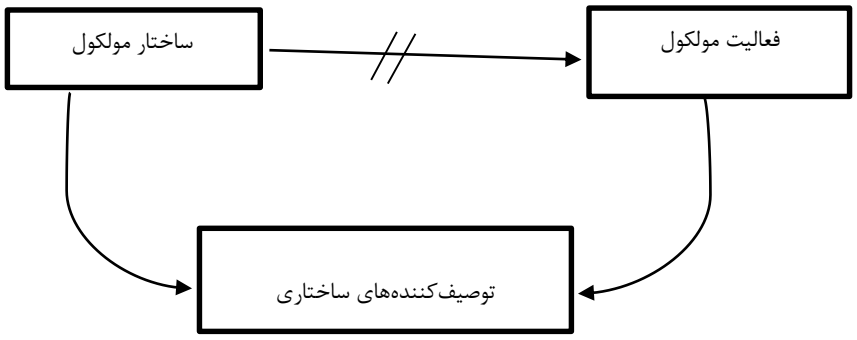
[1] Virtual screening

این پروژه از این هدف (اتصال لیگاند-پروتئین) جهت تولید توصیف‌کننده‌های مهمی اعم از انرژی آزاد اتصـــال، انرژی واندروالس، انرژی بین مولکولی نهایی، انرژی پیوند هیدروژنی، انرژی الکتروســـتاتیک، مجموع انرژی داخلی نهایی، انرژی آزاد چرخشی، ثابت بازداری تئوری استفاده شده است.

۲-۵ اصول مدل‌سازی QSAR

در مطالعات QSAR هرگز نمی‌توان به‌طور مستقیم ارتباط ساختارمولکول با فعالیت بیولوژیکی آن‌را بیان کرد، بلکه باید مراحل زیر به ترتیب دنبال شود [۴۱، ۵۴]:

۱. جمع آوری و انتخاب سری داده‌ها
۲. رسم و بهینه‌سازی ساختار ترکیبات
۳. محاسبه توصیف کننده‌ها
۴. تجزیه و تحلیل و ارزیابی توصیف‌کننده‌ها و انتخاب موثرترین آن‌ها
۵. ایجاد مدل‌های آماری
۶. تجزیه و تحلیل و ارزیابی توصیف‌کننده‌ها و انتخاب مناسب ترین مدل
۷. ارزیابی قدرت پیش‌بینی و اعتبار مدل انتخاب شده



شکل(۲-۷) شمای کلی از مراحل QSAR

دانســتن این‌که چگونه ســاختار مولکولی منجر به اثر خاصــی بر روی یک ســیســتم بیولوژیکی می‌شود، نشان می‌دهد که کلید مهمی برای دست‌یابی به اطلاعات در دسترس وجود دارد. توسعه این روابط منجر به ایجاد یک مدل پیش‌بینی کننده می‌شـود. اگر یک سـری مواد شـیمیایی وجود داشـته باشد و تلاش شود که یک رابطه بین اثرات بیولوژیکی (به عنوان مثال فعالیت) و ویژگی‌های شیمیایی (به عنوان مثال ســاختار) ایجاد گردد، در این صــورت یک ارتباط کمّی ســاختار –فعالیت یا QSAR

تشکیل می‌گردد[۵۱].

۲-۵-۱ جمع آوری و انتخاب سری داده‌ها

اولین مرحله مدل‌سازی جمع‌آوری و انتخاب یک سری ترکیباتی است که مقادیر تجربی خاصیت و یا فعالیت مورد نظر آن‌ها باید با صحت قابل قبول اندازه‌گیری شده باشد. سری ترکیبات باید از نظر ساختاری مشابه باشند و در شرایط عملی یکسانی به‌دست آمده باشند و یا به‌عبارتی دیگر، پیوستگی داده‌ها باید رعایت شود تا نتیجه قابل اعتماد و مناسب‌تری به‌دست آید. بعد از این‌که سری داده‌ها جمع آوری شد لازم است روی آن‌ها پیش پردازش انجام شود. این مرحله وابسته به طبیعت، مقدار و کیفیت داده‌ها است. پیش پردازش داده‌ها شامل حذف داده‌های نامربوط، نامطمئن و با نویز زیاد است که می‌تواند تاثیر منفی روی کل فرآیند مدل‌سازی داشته باشد. این روش‌ها بستگی به اهداف مدل‌سازی و دانش اولیه درباره خصوصیات داده‌ها دارد.

در مدل‌سازی QSAR سری داده‌ها گاهاً به سه دسته سری آموزشی[۱] ، سری ارزیابی[۲] و سری آزمون[۳] تقسیم می‌شود. به مجموعه داده‌ای که برای برازش مدل استفاده شده است، مجموعه داده‌های آموزشی می‌گویند. با استفاده از سری ارزیابی مدل‌های مختلف ارزیابی شده و بر اساس پارامترهای آماری به‌دست آمده برای این سری بهترین مدل انتخاب می‌شود و در نهایت یک روش برای اندازه‌گیری قدرت پیش‌بینی مدل، آزمودن آن بر روی مجموعه‌ای از داده‌ها است که در برازش آن مدل مورد استفاده قرار نگرفته باشد. در یادگیری ماشین و داده‌کاوی به چنین مجموعه‌ای، مجموعه داده‌های آزمون می‌گویند و با استفاده از سری آزمون قدرت پیش‌بینی و اعتبار مدل، مورد سنجش قرار می‌گیرد. سری آزمون به‌نحوی انتخاب می شود که نماینده جمعیت مولکول‌های سری آموزشی باشد.

[1] Training set
[2] Validation set
[3] Test set

۲-۵-۲ رسم و بهینه‌سازی ساختار ترکیبات

برای محاسبه برخی از توصیف کننده‌ها به اطلاعاتی چون مختصات سه بعدی و حالت پایدار با مینیمم انرژی بهینه مولکول نیاز می‌باشد که با استفاده از شیمی محاسباتی قابل بررسی می‌باشد. بنابراین قبل از محاسبه توصیف کننده‌ها ساختار تمام ترکیبات موجود در سری داده‌ها با استفاده از نرم افزار HyperChem8.1 رسم و بهینه می‌شوند تا پایدارترین حالت آن‌ها با مینیمم انرژی حاصل شود. اگر ساختارها به شکل صورت‌بندی با مینیمم انرژی نباشند مقادیر غیر صحیحی برای توصیف‌کننده‌ها ایجاد می شود[۵۵]. پس از رسم و بهینه سازی ساختار ترکیبات اطلاعاتی از قبیل زوایای پیوندی، طول پیوندها، زوایای چرخشی و ... را در اختیار می‌گذارد. داده‌های حاصل از این نرم‌افزار را می‌توان به‌عنوان ورودی به سایر نرم‌افزارها معرفی نمود.

۲-۵-۳ محاسبه توصیف‌کننده‌ها

توصیف‌کننده‌ها مقادیر عددی هستند که ویژگی‌های مختلف ساختاری و الکترونی مولکول‌ها را به‌طور کمی نشان می‌دهند. هر توصیف‌کننده اطلاعات خاصی از مولکول را که بر کمیت مدل‌سازی موثر است، در اختیار می‌گذارد[۵۶]. برخی از ویژگی‌های یک توصیف‌کننده مناسب عبارت است از :

- ساده بودن
- توانایی تفسیر ساختار مولکول
- عدم همبستگی با سایر توصیف کننده‌ها
- قابلیت تمایز بین ایزومرهای مختلف مولکول
- قابل کاربرد برای دامنه وسیعی از ساختارهای مولکولی

۲-۵-۴ تجزیه و تحلیل و ارزیابی توصیف‌کننده‌ها و انتخاب موثرترین آن‌ها

تعداد توصیف کننده‌های حاصل بسیار زیاد است و از آن‌جایی‌که ممکن است همبستگی بالایی بین آن‌ها باشد فرایند مدل‌سازی را دشوار می‌سازد. برای مدل‌سازی بهتر باید متغیرهایی با همبستگی بالا را حذف نمود. در مرحله بعد متغیرهای مناسب باید گزینش گردد تا حاوی اطلاعات مفید و ارتباط

بالا با پاسخ‌های بیولوژیکی باشند. برای حذف متغیرهایی با ضریب همبستگی بالا[1] ابتدا ماتریس مربع همبستگی[2] تشکیل داده می شود و از بین دو زوج با ضریب همبستگی بالا (R>۰/۸) آن‌که ضریب همبستگی کمتری با پاسخ بیولوژیکی دارد حذف می‌شود[۵۷]. برای تشکیل ماتریس مربع همبستگی از نرم افزار SPSS[3] استفاده می‌شود. این مربع اطلاعاتی راجع به میزان همبستگی بین متغیرهای مستقل و هم‌چنین میزان همبستگی متغیرهای مستقل با متغیر وابسته را در اختیار می‌گذارد[۵۹،۵۸].

از روش‌های متفاوتی برای انتخاب موثرترین توصیف‌کننده‌ها استفاده می‌شود که می‌توان به روش انتخاب متغیر پیش‌رونده، انتخاب متغیر گام به گام[4] ، روش منقبض کننده حداقل قدر مطلق و عملگر انتخاب کننده (LASSO)[5] و الگوریتم ژنتیک اشاره کرد که در این تحقیق از روش انتخاب متغیر رگرسیون گام به گام و روش منقبض کننده حداقل قدر مطلق و عملگر انتخاب کننده (لاسو) استفاده شده است.

۲-۵-۴-۱ انتخاب متغیر به روش رگرسیون خطی

رگرسیون معادله‌ای برای ایجاد یک رابطه بین یک متغیر از یک طرف و یک یا چند متغیر از طرف دیگر می‌باشد. اگر تنها یک متغیر مستقل داشته باشد، رگرسیون را ساده و در غیر این صورت، رگرسیون را چندگانه می‌گویند[۶۰]. در بسیاری موارد نمی‌توان تغییرات یک متغیر را فقط به مقادیر یک متغیر دیگر مربوط کرد. به‌عبارت دیگر برای پیش‌بینی مقادیر یک متغیر، دانستن مقادیر دو یا چند متغیر دیگر لازم است.

اما زمانی‌که تعداد متغیرهای مستقل مورد مطالعه زیاد باشد لازم است که تکنیکی برای انتخاب متغیرهای مهم و اثرگذار مورد استفاده قرار گیرد تا بتوان متغیرهای معنی دار را از متغیرهای بی

[1] Correlation matrix
[2] Correlation coefficient
[3] Statistical package for the social sciences
[4] Stepwise Regression
[5] Least Absolute Shrinkage and Selection Operator

اهمیت تمیز داد. روش‌های رگرسیون متفاوتی برای انتخاب متغیرهای مستقل وجود دارد. از جمله مهم‌ترین آن‌ها می‌توان به انتخاب پیش‌رونده[1]، حذفی پس‌رونده[2] و روش گام به گام اشاره کرد..

- **روش انتخاب متغیر پیش‌رونده**

در روش انتخاب متغیر پیش‌رونده متغیرهای مستقل بر اساس این‌که کدام متغیر بیشترین همبستگی را با متغیر وابسته (پاسخ بیولوژیکی) دارد، انتخاب می‌شود. سپس متغیر مستقلی که آماره‌ی فیشر (F) بزرگتر و خطای استاندارد کمتری را برای مدل ایجاد می‌کند وارد معادله مدل می‌شود.

- **روش انتخاب متغیر حذفی پس‌رونده**

در روش حذفی پس‌رونده با ایجاد یک رگرسیون خطی ساده که شامل تمام متغیرهای مستقل ا ست شروع شده و سپس سهم هر عبارت در مدل م شخص می شود. متغیری که کمترین تاثیر و پایین‌ترین سطح معناداری را دارد کنار گذا شته شده و پارامترهای معادله رگر سیون دوباره محا سبه می‌گردند. به‌همین ترتیب متغیر های فاقد سطح معناداری قابل قبول ادامه پیدا می‌کند و این روند زمانی خاتمه می یابد که خروج یک متغیر یاعث کاهش قدرت پیش‌بینی مدل نسبت به مدل قبلی شود.

- **روش انتخاب متغیر گام به گام**

روش رگرسیون گام به گام ترکیبی از روش‌های پیش‌رونده و پس‌رونده است. در این روش متغیرهای مستقل به‌ترتیب و مشابه با روش پیش‌رونده وارد معادله می‌شوند و با ورود هر متغیر جدید کلیه متغیرهای موجود در مدل برر سی شده و اگر هر کدام از آن‌ها سطح معناداری خود را از د ست داده با شند قبل از ورود متغیر جدید از مدل خارج می شود. این روند تا هنگامی که ورود متغیر بعدی

[1] Forward
[2] Backward

تغییر شگرفی در سطح معناداری ایجاد نکند ادامه می‌یابد.

البته لازم به ذکر اســـت که روش‌های انتخاب متغیر فوق خود دارای مشـــکل عدم اســـتواری می‌باشـــند. به عنوان مثال با تغییر کوچک در داده‌ها مدل‌های خیلی متفاوتی را به‌وجود می‌آورند، که این امر درســـتی پیش‌بینی را کاهش می‌دهد. از مزایای لاســـو به عنوان یک روش انتخاب متغیر جدید می‌توان به افزایش دقت پیش‌بینی و بهبود تفسیر مدل‌های ساخته شده اشاره کرد. برآوردگرهای لاسو کاربردهای ساده‌ای از روش‌های برآورد پارامتر، مانند کمترین توان‌های دوم و ضرایب با پیش‌بینی‌های ناچیز را دقیقا به صـــفر کاهش می‌دهند. بنابراین، مدل‌های حاصـــل، روی اثرگذارترین اثرات، متمرکز شـــده و باعث دقت پیش‌بینی می‌گردد. علاوه بر این، برآوردهای لاســـو پایاتر از دیگر روش‌های انتخاب متغیر است.

درادامه برای معرفی روش انتخاب متغیر به کمک رگرسیون لاسو، لازم است مقدمه‌ای از روش کمترین توان‌های دوم معمولی و رگرسیون مرزی آورده شود.

۲-۵-۴-۲ روش کمترین توان‌های دوم معمولی [1]

مدل رگرسیون خطی چندگانه زیر را در نظر بگیرید:

$$y = X\beta + \epsilon$$ رابطه(۲-۱)

که در آن $y = (y_1, \dots, y_n)^T$ بردار ستونی پاسخ‌ها، $X = (x_1, \dots, x_n)^T$ ماتریس طرح در اندازه $n \times p$ با رتبه کامل ستونی p شامل متغیرهای مستقل(پیش‌گو) ، $\beta = (\beta_1, \dots, \beta_p)^T$ بردار ضرایب رگرسیونی و $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$ بردار مولفه‌های خطای تصادفی است.

تخمین کمترین توان‌های دوم معمولی (OLS) با مینیمم کردن مجموع توان‌های دوم خطاهای مدل یعنی $\epsilon^T\epsilon = (y - X\beta)^T(y - X\beta)$ به‌دست می‌آید. در این حالت برآوردگر [2] OLS بردار پارامترها

[1] Ordinary least square
[2] Coefficient

به‌صورت زیر حاصل می‌شود:

$$\hat{\beta} = argmin_{\beta}\, \epsilon^T$$ رابطه(۲-۲)

با توجه به این‌که

$$\frac{\partial \epsilon^T \epsilon}{\partial \beta} = -2X^T y + 2X^T X\beta = 0$$ رابطه(۲-۳)

می‌توان نتیجه گرفت که

$$\hat{\beta} = (\mathrm{X^T X})^{-1}\mathrm{X^T y}$$ رابطه(۲-۴)

به دو دلیل اســـتفاده از این برآوردگر مطلوب نمی‌باشـــد. اول اینکه در مورد صـــحت پیش‌بینی، برآوردگر OLS نا اریب[1] و دارای واریانس زیاد می‌باشد. صحت پیش‌بینی معمولا با انقباض برآوردها به صفر یا صفر کردن برخی از ضرایب بهبود یافته و دلیل دوم نیز تفسیر پذیری است، به صورتی‌که با وجود تعداد زیادی از پیش‌بینی کننده‌ها زیر مجموعه کمتری از پارامترها با تاثیرات بیشتر تعیین شود. معمولا می‌توان درســـتی پیش‌بینی را با انقباض تعدادی از ضـــرایب و یا با صـــفر قرار دادن آن‌ها بهبود بخشید[۵۴، ۶۱]. روش پیشنهادی برای بهبود روش برآورد OLS، رگرسیون‌های انقباضی است، از جمله به رگرسیون مرزی[2] و لاسو می‌توان اشاره کرد.

۲-۵-۴-۳ رگرسیون مرزی

رگرسیون مرزی (ریج) در سال ۱۹۶۲ برای اولین بار توسط هورل و کنارد[3] معرفی شد. لازم به ذکر است اساس و پایه روش برآوردگر OLS یک رگرسیون خطی این است که $(\mathrm{x}^T\mathrm{x})^{-1}$ وجود داشته باشد و دو دلیل وجود دارد که این معکوس وجود نداشته باشد: ماتریس طرح پر رتبه ستونی نباشد و دیگری نیز مشکل هم‌خطی است که ممکن است وجود داشته باشد. روش رگرسیون مرزی یکی از بهترین و محبوب ترین گزینه‌ها برای رفع این مشکل می‌باشد.

[1] Bias
[2] Ridge Regression
[3] Hoerl and Kennard

اضافه کردن ماتریس قطری λI به $x^T x$ راهی آسان برای تضمین معکوس پذیری می‌باشد که در آن λ یک مقدار ثابت مثبت و I ماتریس همانی است. بنابراین برآوردگر رگرسیون ریج پارامتر β به صورت زیر می‌باشد :

$$\hat{\beta} = (x^T x + \lambda I)^{-1} x^T y$$ رابطه(۲-۵)

که $\lambda > 0$ است. همان‌طور که با مینمم کردن $\epsilon^T \epsilon$ نسبت به β، برآوردگر OLS حاصل شد، این برآوردگر را نیز می‌توان با مینیمم کردن عبارت مجموع توان‌های دوم خطا $\epsilon^T \epsilon$ نسبت به β تحت شرط $\beta^T \beta = \sum_{j=1}^{p} {\beta_j}^2 \leq t$ به‌دست آورد. در این‌جا $t \geq 0$ یک پارامتر تنظیم کننده می‌باشد که میزان انقباض ضرایب را کنترل می‌کند. به‌عبارتی هر چه مقدار t کوچکتر باشد ضرایب بزرگتر بیشتر جریمه می‌شوند. برای حل مساله بهینه سازی فوق می‌توان از ضرایب لاگرانژ استفاده کرد. به‌عبارت دیگر برآوردگر ریج به‌صورت زیر حاصل می‌شود:

$$\widehat{\beta_R} = argmin_{\beta} \, (\epsilon^T \epsilon + \lambda \beta^T \beta)$$ رابطه(۲-۶)

که در آن λ ضریب لاگرانژ یا همان پارامتر انقباض ریج است. با توجه به این‌که

$$\frac{\partial(\epsilon^T \epsilon + \lambda \beta^T \beta)}{\partial \beta} = -2x^T y + 2x^T x\hat{\beta} + 2\lambda\beta = 0$$ رابطه(۲-۷)

داریم:

$$(x^T x + \lambda I_p)\beta = x^T y$$ رابطه(۲-۸)

و در نتیجه

$$\widehat{\beta_R} = (x^T x + \lambda I_p)^{-1} x^T y$$ رابطه(۲-۹)

$\hat{\beta}_R$ یک برآوردگر اریب با میانگین و واریانس زیر می‌باشد :

$$E(\hat{\beta}_R) = \beta + \lambda(X^T X + \lambda I)^{-1}\beta$$ رابطه(۲-۱۰)

$$Var(\hat{\beta}_R) = (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1} \sigma^2$$ رابطه(۲-۱۱)

در حالی‌که در خصوص برآوردگر OLS داریم:

$$\in \hat{\beta} = \beta$$

$$\mathrm{Var}(\hat{\beta}) = (\mathrm{X^T X})^{-1}\sigma^2$$ رابطه(۲-۱۲)

هورل و کنارد[1] ثابت کرده‌اند که اگر $\beta^T\beta$ کراندار باشد می توان $\lambda > 0$-یی را پیدا کرد به‌طوری‌که:

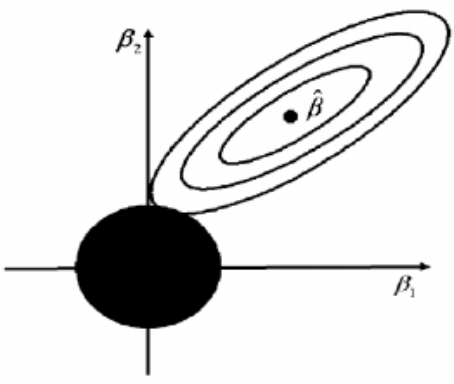
$$\mathrm{MSE}(\hat{\beta}_R) < \mathrm{MSE}(\hat{\beta})$$ رابطه(۲-۱۳)

بنابراین رگرسیون مرزی می‌تواند برآورد OLS را بهبود ببخشد[۶۲].

در شـــکل(۲-۸)، ناحیه‌ی دایره‌ای، ناحیه تاوان$(\beta_1^2 + \beta_2^2 \leq t) = \beta^T\beta$ برای حالت p=۲ و نواحی بیضوی، مجموع توان دوم مانده‌ها با مرکزیت برآوردگر (OLS) را در رابطه:

$$\epsilon^T\epsilon = (\mathrm{y} - \mathrm{X}\beta)^T(y - X\beta)$$ رابطه(۲-۱۴)

نشان می‌دهد. این شکل به وضوح ناتوانی روش مرزی را در صفر برآورد کردن ضرایب نشان می‌دهد، زیرا برخورد دو ناحیه نمی‌تواند در نقطه ای باشد که در آن یکی از ضرایب صفر است. روش‌های جدید انقباضی میل دادن ضرایب به سمت صفر را به‌گونه‌ای انجام می‌دهند که بعضی از ضرایب، مربوط به متغیرهای بی اثر، دقیقا صـــفر برآورد شـــده و به‌این ترتیب متغیر مربوط به آن ضـــرایب از مدل خارج خواهد شد. بنابراین در این روش‌ها برآوردیابی و انتخاب متغیر تواما صورت می‌پذیرد. باید توجه داشت که در تمام روش‌های انقباضی ابتدا به‌ازای مقادیر مختلف، پارامتر کنترل برآورد ضرایب محاسبه شده و ســـپس با اســـتفاده از معیارهای ارزیابی مدل مثل برآورد بهینه از میان مجموعه برآوردهای به‌دســـت آمده انتخاب می‌شود[۶۱].



شکل(۲-۸) عملکرد رگرسیون مرزی

[1] Hoerl and Kennard

۲-۵-۴-۴ روش منقبض کننده حداقل قدر مطلق و عملگر انتخاب کننده (LASSO)

تیبشیرانی[1] روشی جدید برای تخمین مدل‌های خطی ارائه کرد و آن‌را لاسو (روش منقبض کننده حداقل قدر مطلق و عملگر انتخاب کننده) نامید. این روش به‌صورت همزمان به برآورد پارامترها و انتخاب متغیر می‌پردازد. انگیزه اصلی تیبشیرانی در تعریف لاسو، از پیشنهاد فردی به نام بریمن[2] (۱۹۹۳) می آید. به پیشنهاد بریمن عبارت

رابطه(۲-۱۵) $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} c_j \hat{\beta}_j x_{ij})^2$

تحت شرایط $\sum_{j=1}^{k} c_j \le t$ و $c_j \ge 0$ مینیمم می‌شود.

تیبشیرانی این دو شرط را به یک شرط تبدیل کرد و اسم کانادایی "لاسو" را برای آن انتخاب کرد. این روش اساسا شبیه رگرسیون مرزی می‌باشد، با این تفاوت که به‌جای استفاده از تابع تاوان درجه دوم، از تابع تاوان مجموع قدرمطلق ضرایب استفاده می‌شود و عبارت مجموع توان‌های دوم خطا $\epsilon^T\epsilon$ تحت شـــرط $\sum_{j=1}^{k}|\beta_j| \le t$ مینیمم می‌شـــود. در اینجا نیز $t \ge 0$ پارامتر تنظیم کننده بوده و میزان انقباض ضرایب را کنترل می کند. لاسو را در حالت کلی نیز با استفاده از ضرایب لاگرانژ می‌توان با مینیمم کردن عبارت خطای جریمه شده به‌صورت زیر به‌دست آورد :

رابطه(۲-۱۶) $\widehat{\beta_l} = argmin_{\beta}(\epsilon^T\epsilon + \lambda|\beta|^T)$

که در آن λ یک پارامتر تنظیم کننده نامنفی و $|\beta| = (|\beta_1|, \dots, |\beta_p|)^T$ می باشـــد و هم‌چنین $\lambda\sum_{j=1}^{k}|\beta_j|$ را تاوان L_1 می‌نامیم که این تاوان برای موفقیت لاسو حیاتی می‌باشد. لاسو با افزایش λ ضرایب را به سمت صفر منقبض می‌کند و زمانی‌که λ به اندازه کافی بزرگ باشد، بعضی ضرایب را دقیقا صفر برآورد می‌کند.

در حقیقت لاسو مجموع توان‌های دوم خطای مدل رگرسیون را تحت این محدودیت که مجموع

[1] Tibshirani
[2] $Brieman$

مقادیر مطلق ضرایب رگر سیونی کمتر از یک مقدار ثابت با شند، حداقل می سازد. به‌خاطر ماهیت این محدودیت، لاسو تمایل به ایجاد ضرایب دقیقا برابر با صفر دارد که در نتیجه مدل‌های قابل تفسیری را به‌وجود می‌آورد.

تیبشیرانی لاسو را جهت بهبود صحت پیش بینی و تفسیر پذیری مدل رگرسیون معرفی کرد و با تغییر در فرایند برازش مدل، تنها یک زیرمجموعه از متغیرهای کمکی را به‌جای اســـتفاده از همه‌ی متغیرها جهت ا ستفاده در مدل نهایی به‌کار گرفت. این روش ویژگی‌های سادگی تف سیر پذیری روش انتخاب زیرمجموعه ای[1] و پایداری روش رگرسیون ریج را داراست[۶۳].

۲-۵-۴-۵ لاسو در مدل متعامد

بر خلاف روش رگر سیون مرزی که می‌توان صورت ب سته‌ای برای برآوردگر $\widehat{\beta_R}$ یافت، در روش لاسو حل مساله بهینه سازی دارای جواب مشخص بسته ای نیست و باید به ازای λ های مختلف به کمک روش‌های عددی حل شود اما در رگرسیون متعامد، یعنی حالتی‌که $X^TX = I_p$ ، می‌توان صورت بسته‌ای، همانند $\widehat{\beta_R}$ برای روش لاسو به‌دست آورد.

در این حالت فرض کنید که $X^TX = I_p$ و

$$\widehat{\beta_L} = \left(\hat{\beta}_1^{LASSO}. \ldots . \hat{\beta}_p^{LASSO}\right)^T$$ رابطه(۲-۱۷)

در این صورت می‌توان نشان داد:

$$\hat{\beta}_j^{LASSO} = \text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \gamma)^+$$ رابطه(۲-۱۸)

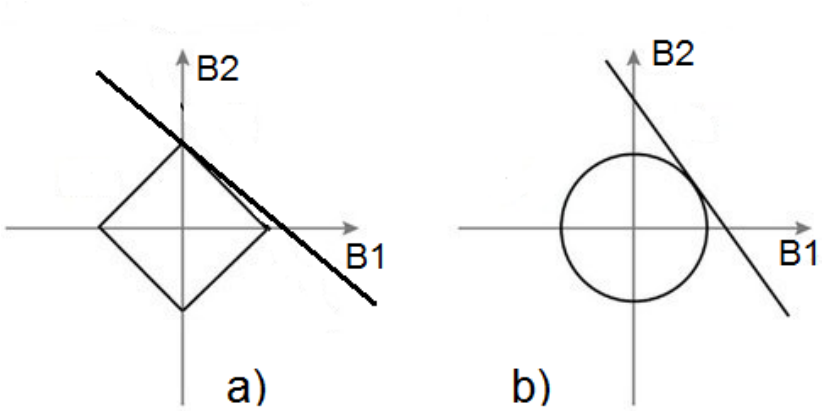
و γ با شرط $\sum|\hat{\beta}_j| = t$ تعیین می‌شود. در اینجا $a^+ = \max(0,a)$ و $\hat{\beta}_j$ همان j مین جز برآوردگر کمترین توان‌های دوم است و

[1] Subset selection

$$\text{sign}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$$

۲-۵-۴-۶ هندسه لاسو

برای این‌که بهتر بتوان روش کار لاسو را از لحاظ مشهودی درک کرد روش انتخاب متغیر را در حالت هندسی و برای یک مساله دو بعدی مورد بررسی قرار گرفته است. در این روش تلاش بر آن بوده که هندسه لاسو به طور مختصر با هندسه رگرسیون مرزی مقایسه شود. در حالت p=۲ دو پارامتر β_1 و β_2 در شکل(۲-۹) به‌عنوان محورهای مختصات نشان داده شده‌اند.



شکل(۲-۹) a) جریمه لاسو b) جریمه ریج

در حقیقت در شکل(۲-۹) جریمه‌های رگرسیون لاسو (a) و رگرسیون مرزی (b) از طریق معادلات $|\beta_1| + |\beta_2| = r$ و $\beta_1^2 + \beta_2^2 = r$ رسم شده‌اند.

به وضوح در در حالت (a) زمانی‌که پاره خط فرضی محور β_1 را قطع می‌کند محور β_2 را روی معادله $|\beta_1| + |\beta_2| = r$ قطع می‌کند و بر آن مماس است که در این حالت $\beta_1 = 0$ و $\beta_2 = r$ می‌باشد. در حالی‌که وقتی پاره خط فرضی در حالت (b) محور β_1 را قطع می‌کند چنان‌چه بخواهد در معادله $\beta_1^2 + \beta_2^2 = r$ صدق کند محور β_2 را در نقطه بالاتر از نقطه معادله دایره قطع می‌کند و به‌عبارت دقیق تر حالت $\beta_1 = 0$ و $\beta_2 = V_r$ رخ نمی‌دهد. بر مبنای این تفسیر، جریمه لاسو انتخاب متغیر انجام می‌دهد و جریمه ریج نمی‌تواند متغیرهای بی معنی را حذف کرده ($\beta_1 = 0$) و متغیرهای

معنی دار را β_2 انتخاب می‌کند.

۲-۵-۵ ایجاد مدل‌های آماری

مدل، در واقع یک رابطه ریاضی است که بیان‌کننده‌ی رابطه بین متغیر مستقل و متغیر وابسته است و به کمک آن می‌توان با داشتن مقادیر متغیر مستقل، متغیر وابسته را تخمین زد. مرحله اصلی در مطالعات QSAR ساختن مدل می‌باشد. مدل‌های QSAR را می‌توان بر اساس پارامترهایی که به صورت تجربی بدست آمده و یا از طریق تئوری محاسبه شده‌اند، به زبان ریاضی بیان نمود و اگر نتایج این نوع مطالعات همبستگی قابل قبولی را ارائه کند، علاوه بر شفاف سازی نحوه‌ی ارتباط بین خواص مولکول‌ها و ویژگی‌های ساختمانی آن‌ها، به پژوهشگران در پیش‌بینی رفتار مولکول‌های جدید بر اساس رفتار مولکول‌های مشابه کمک می‌کند.

چندین روش متفاوت برای ساختن مدل QSAR وجود دارد که می‌توان به رگرسیون خطی چندگانه (MLR)، رگرسیون اجزای اصلی (PCR)[۱]، کمترین توان‌های دوم جزئی(PLS)[۲]، شبکه عصبی مصنوعی (ANN)[۳] و جنگل‌های تصادفی (RF)[۴] اشاره کرد. در این پژوهش از دو روش شبکه عصبی مصنوعی و جنگل‌های تصادفی استفاده شده است.

۲-۵-۶ مقدمه ای بر شبکه عصبی

شبکه عصبی مصنوعی یک سامانه پردازشی داده‌ها است که از مغز انسان ایده گرفته و پردازش داده‌ها را به عهده پردازنده‌های کوچک و بسیار زیادی سپرده که به صورت شبکه‌ای به‌هم پیو سته و موازی با یک‌دیگر رفتار می‌کنند تا یک مسئله را حل نمایند. در این شبکه‌ها به کمک دانش برنامه‌نویسی، ساختار داده‌ای طراحی می‌شود که می‌تواند همانند نرون[۵] عمل کند. به این ساختار داده

[1] Principle Component Regression
[2] Partial Least Square
[3] Artifitial Neuronal Netwotk
[4] Random forests
[5] Neuron

نرون گفته می شود. با ایجاد شبکه‌ای بین این نرون‌ها و اعمال یک الگوریتم آموز شی به آن، شبکه را آموزش می‌دهند. در این شبکه عصبی نرون‌ها دارای دو حالت فعال (روشن یا یک) و غیرفعال (خاموش یا صفر) می‌با شند و هر یال (سیناپس یا ارتباط بین گره‌ها) دارای یک وزن[1] می‌با شد. یال‌های با وزن مثبت موجب تحریک یا فعال نمودن گره فعال بعدی می‌شـــوند و یال‌های با وزن منفی، گره متصـــل بعدی را غیرفعال یا مهار می‌کنند [۶۴].

۲-۵-۶-۱ ساختمان و عملکرد نرون مصنوعی

یک نرون مصنوعی در حقیقت مدل محاسباتی است که از نرون‌های عصبی واقعی انسان، الهام گرفته اسـت. در شـبکه‌های عصـبی مصـنوعی نیز بلوک‌های سـاختاری و یا نرون‌ها، دسـتگاه‌های محاسـباتی خیلـی سـاده‌ای هستند و ارتباط بین نرون‌ها عملکرد شبکه را تعیین می‌کند و هدف از آموزش شـبکه‌هـای عصـبی مصـنوعی تعیین ارتباط مناسب، جهت حل مسائل مختلف است. در واقع شبکه‌های عصبی مصنوعی نیز با ایده گـرفتن از رفتار شبکه عصبی بیولوژیکی شکل یافته است. برای ایجاد یک شبکه عصبی مصنوعی باید سیستمی طراحی شود که دارای تعدادی ورودی باشد و با توجه به اهمیت هر یک، آن‌ها را با یک‌دیگر جمع جبری کند و توسط یک تابع موسوم به تابع انتقال[2]، آن‌ها را به نرون‌های دیگر ارسال نماید.

۲-۵-۶-۲ تابع انتقال

زمانی که ورودی‌ها با توجه به اهمیت آن‌ها با یک‌دیگر جمع جبری شوند تو سط تابع انتقال به نرون بعدی منتقل می‌شـوند. در واقع تابع انتقال یکی از اجزای شـبکه عصـبی می‌باشـد که برحسـب کاربرد می‌تواند توابع متفاوتی با شد. دو تابع انتقال رایج در شبکه‌های عصبی که در این پژوهش برای بهینه‌سازی شبکه به کار گرفته شده است عبارتند از:

[1] Weight
[2] Transfer function

۲,۵,۶,۲,۱ تابع انتقال لگاریتم سیگموئید[1](log sig)

از این تابع انتقال در شـــبکه‌های پس انتشـــار اســـتفاده می‌شـــود این تابع مقادیر ورودی را در محدوده منفی بی‌نهایت تا مثبت بی‌نهایت دریافت کرده و خروجی بین ۰ و ۱ تولید می‌نماید.

۲,۵,۶,۲,۲ تابع انتقال تانژانت سیگموئید[2] (tansig)

این تابع انتقال مقادیر ورودی را در محدوده منفی بی‌نهایت تا مثبت بی‌نهایت دریافت کرده و خروجی بین ۱+ و۱- تولید می‌کند.

۲-۵-۷ آموزش شبکه‌های جلو سو با تکنیک پس‌انتشار[3]

پس انتشـــار خطا یک روش متداول آموزش با ناظر برای شـــبکه‌های جلوســـو اســـت یعنی برای به‌دســـت آوردن ارتباط بین متغیرهای ورودی و خروجی در یادگیری به الگوی آموزشـــی نیاز اســـت. به‌طور کلی آموزش به کمک تکنیک پس‌انتشار طبق مراحل زیر انجام می‌شود[۶۴].

۱- انتشار ورودی‌ها از نرون‌های ورودی به سمت نرون‌های خروجی

۲-اختصاص ماتریس وزن‌های تصادفی به هریک از اتصالات

۳- مقایسه خروجی‌های شبکه با مقادیر واقعی (مقادیر هدف) و محاسبه‌ی خطای شبکه

۴- پس‌انتشار خطا از نرون‌های خروجی به سمت نرون‌های ورودی و اصلاح وزن‌ها

۵- ارزیابی عملکرد شبکه با توجه به تابع کارآیی تعیین شده

مراحل فوق تا رســـیدن به حداکثر تکرار[4] مجاز تکرار می‌شـــود و یا این‌که مقدار تابع کارآیی از مقداری که تعیین شده کم‌تر با شد. شبکه ع صبی انتخاب شده در این پروژه یک شبکه جلو سو با الگوریتم آموزشی پس انتشار می‌باشد[۶۴] .

[1] Logarithm sigmoid transfer function
[2] Hyperbolic tangent transfer function
[3] Back propagation
[4] Epoch

۲-۶ جنگل‌های تصادفی [1](RF)

۲-۶-۱ مقدمه

مدل‌ســـازی با اســـتفاده از روش‌ها و مدل‌های آماری، کاربرد فراوانی در علوم مختلف دارد که می‌توان از جمله زمینه های کاربرد آن به علوم مهندســی، کامپیوتر، پزشــکی و ... اشــاره کرد. از جمله روش‌های مرسوم مدل‌سازی می‌توان به رگرسیون خطی و غیر خطی، رگرسیون لجستیک، شبکه های عصبی مصنوعی و ... اشاره کرد که هر کدام بسته به ساختار الگوریتم‌شان قابلیت استفاده برای داده های کمی یا کیفی (رده بندی یا رگرسیون) و یا هر دو را دارا می‌باشند.

یکی از روش های جدید مدل سازی، روش جنگل تصادفی است که از این پس به اختصار RF نامیده می‌شود. جنگل تصادفی جز تکنیک های یادگیری ماشین[2] می باشد. در سال ۲۰۰۱ توسط لئو بریمن[3] و آدله کاتلر[4] ارائه شـــد، به‌طوری که همه محققان، جنگل تصـــادفی را با نام این دو محقق می‌شناسند. نام این روش از نام جنگل‌های تصمیم تصادفی[5] گرفته شده که اولین بار توسط تین کام هو[6] پیشنهاد شده است. اهمیت مسئله وقتی زیاد می‌شود که تعداد متغیرها زیاد شود. این روش، یک روش پیشــرفته ماشــین یادگیری اســت که رده بندی را با اســتفاده از درختان تصــمیم گیری[7] انجام می‌دهد[۶۵].

اساس کار مدل جنگل تصادفی ریشه در مدل درخت تصمیم است، به همین دلیل لازم است که برای بیان مدل جنگل تصادفی، ابتدا روش رده بندی درخت تصمیم بیان شود.

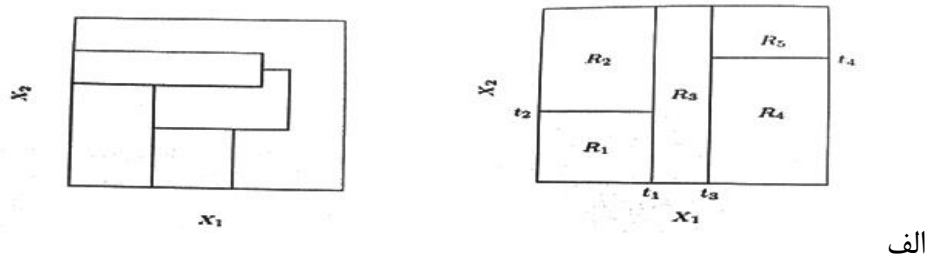
---

[1] Random Forests
[2] Machine learning
[3] Leo Breiman
[4] Adele Cutler
[5] Random decision forests
[6] Tin Kam Ho
[7] CART(classification and regression tree)

۲-۶-۲ روش درخت رگرسیونی[1] (تصمیم)

این روش بر پایه تقسیم کردن مجموعه داده‌ها به قسمت‌های کوچکتر است. در روش رگرسیون خطی، پیش بینی‌کننده‌های رگر سیونی، مدل‌هایی هستند که در آن‌ها یک مدل پیش‌بینی واحد روی ف ضای کل داده‌ها انجام می‌دهد. اما ممکن ا ست به دلیل تفاوت رفتار متغیر پا سخ در نواحی مختلف، برقراری یک مدل واحد، کارایی لازم را نداشته باشد. لذا یک روش جایگزین، تقسیم بندی فضای داده‌ها به بخش‌های کوچکتر است تا بتوان رفتار متغیر پاسخ را به‌طور موضعی مدل‌سازی نمود. در روش درخت رگرسیونی هدف این است که مقادیر متغیرهای توصیفی (ابر مکعب در فضای m بعدی) در هر ناحیه، به‌گونه‌ای تقسیم بندی شود که داده‌های واقع در هر بخش تا حد ممکن هم‌گون با شند به‌طوری‌که بتوان آن‌ها را توسط ساده ترین مدل پیش‌بینی نمود.

به‌طور کلی، افراز[2] فضای متغیرهای توضیحی را می‌توان به هر شکلی انجام داد. اما در روش درخت رگرسیونی، تنها افراز قابل قبول است که تمامی نواحی ساخته شده به شکل مربع یا مستطیل باشند. در شکل(۲-۱۰)، نمونه‌ای از افراز قابل قبول و غیرقابل قبول در فضای دومتغیره با روش درخت رگرسیونی نشان داده شده است.



ب الف

شکل(۲-۱۰) (الف) افراز قابل قبول و شکل (ب) افراز غیرقابل قبول

[1] Regression Tree
[2] Partition

۲-۶-۳ اساس روش جنگل‌های تصادفی

اساس روش RF وابسته به ماهیت روش درخت رگرسیونی است و از آن‌جایی که ساختار درخت رگرسیونی به‌طور کامل معرفی گردید، اینک می‌توان به معرفی روش RF پرداخت.

در روش RF مجموعه‌ای از درخت‌های رگرسیونی تشکیل می شوند و هر درخت مدلی را تولید می‌کند که مدل نهایی، برآیند یا ترکیبی از همه این مدل‌ها است. به‌عبارت دیگر هر یک از این درخت‌ها سهمی در مدل نهایی دارد. یکی از تفاوت‌های اساسی RF با درخت رگرسیونی آن است که در درخت رگرسیونی برای افراز فضای متغیرها در هر مرحله، از کلیه متغیرها استفاده می‌گردد در حالی‌که در RF فقط از زیر مجموعه‌ای از متغیرها استفاده می شود. دیگر تفاوت اساسی این دو روش داده‌هایی است که، در ساخت مدل شرکت دارند. به این معنا که در درخت رگرسیونی، همه داده‌ها در ساخت مدل شرکت دارند، اما در روش RF تنها بخشی از داده‌ها در ساخت مدل استفاده می شوند. به‌طور کلی می‌توان گفت روش RF، ترکیبی از چندین درخت رگرسیونی است که در ساخت آن چندین نمونه با جای‌گذاری از داده‌ها شرکت دارند و در هر درخت برای ساخت هر گره، تنها یک زیرمجموعه تصادفی از متغیرهای توضیحی شرکت می‌کنند.

۲-۶-۳-۱ الگوریتم روش جنگل‌های تصادفی

فرض کنید $i = 1 , \dots , N$ و (X_i , Y_i) مجموعه داده‌های مدل ساز (آموزش) با شند که در آن، $X_i = (X_{1i,..,}X_{iM})$ برداری از M متغیر توضیحی و Y_i متغیر پاسخ متناظر آن است. اگر تعداد کل درخت‌های مدل با ntree نشان داده شود، مراحل پنج‌گانه زیر بیانگر الگوریتم ساخت درخت iام $(i=1,\dots,ntree)$ است.

۱) یک نمونه N تایی به روش جای‌گذاری از مجموعه داده‌های مدل‌ساز گرفته می‌شود. زیر مجموعه‌ای از داده‌های اصلی (حدود یک‌سوم از داده‌های آموزشی) که در این نمونه حضور ندارند

را OOB[1] نامیده که برای هر درخت نقش داده‌های آزمون را ایفا می‌کند. روش جداسازی داده‌های OOB روش نمونه گیری خودگردان اســـت. با تکرار عملیات نمونه گیری، تعدادی مجموعه داده OOB از مجموعه آموز شی به‌وجود می‌آید که می‌توان برای هر مجموعه داده آموز شی یک درخت تصمیم محاسبه کرد.

۲) به‌طور تصــادفی از بین M متغیر توضــیحی، m متغیر انتخاب می‌شــود ($m \ll M$). برای تقســیم فضای متغیرهای توضیحی به دو قسمت و براساس اصول درخت رگرسیونی، فقط از این m متغیر و نمونه N تایی انتخاب شــده در گام ۱ اســتفاده می‌شــود تا بهترین متغیر و بهترین نقطه افراز در اولین مرحله از مراحل بازگ شتی بد ست آید. محدودیت این مرحله آن ا ست که تعداد م شاهدات موجود در هر یک از دو ناحیه باید بیشتر از n_r باشد و n_r پارامتری است که در اختیار کاربر بوده و بیانگر حداقل تعداد مشـــاهدات موجود در هر ناحیه اســـت. معمولا در مدل رگرســـیونی، $m=\frac{M}{3}$ پیشنهاد می‌شود و در مدل رده‌بندی $m = \sqrt{M}$ در نظر گرفته می‌شود.

۳) برای هریک از دو ناحیه تولید شده در گام ۲، مجددا به‌طور تصادفی از بین M متغیر توضیحی، m متغیر انتخاب می‌گردد و با اســـتفاده از همان نمونه N تایی انتخاب شـــده در گام ۱، روش درخت رگر سیون اعمال می‌گردد. این عمل منجر به افراز هر یک از نواحی موجود به دو قسمت می شود. لازم به ذکر است که افراز نواحی در صورتی انجام می‌شود که تعداد مشاهدات موجود در کل ناحیه بیشتر از $2n_r$ باشد. یعنی اگر هریک از نواحی دارای تعداد مشاهداتی کم‌تر از $2n_r$ باشد نباید افراز دیگری روی هیچ یک از نواحی صورت گیرد.

۴) گام ۳ برای تمام نواحی افراز شده تا زمانی تکرار می‌شود که تعداد مشاهدات در تمامی این نواحی کمتر از $2n_r$ باشد.

۵) حال درخت رگرسیونی i ام طوری تشکیل شده است که فضای متغیرهای توضیحی به r_i ناحیه‌ی R_{1i} , R_{2i}, …, R_{r_i} تقسیم گردیده ا ست و تعداد م شاهدات هر ناحیه کمتر از $2n_r$ می‌با شد، مدل

[1] Out of bag

درخت رگرسیونی به دست آمده، به صورت زیر است.

رابطه(۲-۱۹) $\hat{f}_i(X) = \sum_{j=1}^{r_i} \hat{c}_j I_{R_i j}(X)$

که در آن، $\hat{c}_j = \hat{y}_j$ و $I_{R_{ij}}(X) = \begin{cases} 1. X \in R_{ij} \\ 0. X \notin R_{ij} \end{cases}$.

از آن‌جایی که در روش RF تعداد ntree درخت رگرسیونی وجود دارد، می‌توان گفت که تعداد ntree مدل به‌صورت رابطه(۲۳-۲) خواهیم داشت. اگر برای مقدار مشاهده شده‌ی x مدل خروجی درخت iام را به صورت $\hat{f}_i(X)$ نشان دهیم، آن‌گاه برآورد متغیر پاسخ در این نقطه، با میانگین‌گیری از مقادیر $\hat{f}_i(X): i = 1، \dots، ntree$ یعنی رابطه(۲-۲۰) به‌دست می‌آید.

رابطه(۲-۲۰) $\hat{y}(X) = \frac{1}{ntree} \sum_{i=1}^{ntree} \hat{f}_i(X)$

لازم به ذکر است که در روش RF، تعداد درخت‌ها (پارامترntree)،تعداد توصیف‌کننده‌های انتخاب شده در هر مرحله افراز (پارامترmtry) و تعداد حداقل مشاهدات در هر افراز (پارامتر n_r)، پارامترهایی هستند که قابل تغییر بوده و توسط کاربر تعیین می‌گردد [۶۶].

۲-۶-۳-۲ مزیت‌های روش جنگل‌های تصادفی

۱) این روش برای داده‌های بزرگ به‌خوبی اجرا می‌شود.
۲) این روش هزارها متغیر ورودی را می‌تواند در بر بگیرد بدون این‌که متغیری را حذف کند.
۳) این روش، برآوردی برای متغیرهایی که مهم هستند در رده بندی ارائه می‌دهد.
۴) یادگیری این روش آسان است.
۵) در صورت در اختیار دا شتن تعداد م شاهدات کم، با توجه به نقش داده‌های OOB، می‌توان از آن به‌عنوان داده‌های آزمون استفاده کرد.
۶) این روش توانایی انتخاب متغیر را داراست و بر خلاف سایر روش‌های یادگیری ماشین، نیازمند روشی دیگر برای انتخاب متغیر نیست[۶۷، ۶۸].

۲-۷ ارزیابی مدل

۲-۷-۱ استفاده از پارامترهای آماری

برای اطمینان از این‌که مدل نهایی، مدل مناسبی است که توانایی پیش‌بینی نمونه‌های مختلفی از یک جمعیت را داراســـت، باید مدل را ارزیابی کرد. این ارزیابی از طریق شـــاخص‌های کمی انجام می‌شود که به‌وسیله آن‌ها صحت نتایج ارائه شده توسط مدل مورد سنجش قرار می‌گیرند. فرض کنید $x_1, x_2, \ldots, x_n$ نشان‌دهنده متغیرهای مستقل(پیش‌گو) و $y_1, y_2, \ldots, y_n$ نشان‌دهنده متغیرهای وابسته در مدل مورد بررسی می‌باشند. برخی از این شاخص‌های کمی عبارتند از:

ضریب همبستگی[1]**:** ساده‌ترین راه برای بررسی میزان همبستگی دو یا چند متغیر، محاسبه آماره‌ی ضریب همبستگی آن‌هاست. ضریب همبستگی دو متغیر $x.y$ با رابطه(۲-۲۱) تعریف می‌شود. مقدار این آماره بین ۱ تا ۱- متغیر اســت. مقدار بزرگتر آن نشـــان‌دهنده‌ی این اســت که ارتباط خطی بیش‌تری میان متغیر وابسته و متغیرهای مستقل وجود دارد.

$$R = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}}$$ رابطه(۲-۲۱)

ضریب تعیین[2]: به‌عنوان یک شاخص برای بیان دقت خط رگرسیون برآورد شده، به‌کار می‌رود و نشان‌دهنده‌ی نسبت تغییرات متغیر وابسته توضیح داده شده توسط متغیر مستقل است. به‌عنوان مثال R^2 برابر با ۰/۹۲۸۶نشـــان می‌دهد که ۹۲/۸۶ درصـــد تغییرات در متغیر وابســـته می‌تواند توســـط متغیر مستقل توضیح داده شود. رابطه ریاضی مربوط به ضریب تعیین به صورت زیر است:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$ رابطه(۲-۲۲)

مجموع توان‌های دوم رگرسیون(SSR)[3] طبق رابطه(۲-۲۳) بیانگر مجموع توان‌های دوم انحراف مقادیر پیش‌بینی شده‌ی متغیر وابسته از میانگین مقادیر آن است.

[1] Correlation coefficient
[2] Determination coefficient
[3] Sum Square Regression

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$ رابطه(۲-۲۳)

مجموع توان‌های دوم کامل(SST)[1] طبق رابطه(۲-۲۴) نشان‌گر مجموع توان‌های دوم انحراف مقادیر واقعی متغیر وابسته از میانگین مقادیر آن است.

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$ رابطه(۲-۲۴)

مجموع توان‌های دوم خطا (SSE)[2] نیز مبین مجموع توان‌های دوم انحراف مقادیر واقعی متغیر وابسته از مقادیر پیش‌بینی شده برای آن است:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$ رابطه(۲-۲۵)

بنابراین با توجه به روابط فوق می‌توان نوشت:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$ رابطه(۲-۲۶)

طبق رابطه(۲-۲۶) اگر تمام مشاهدات بر روی خط برازش شده قرار گرفته باشند، یعنی به ازای تمام نقاط $y_i = \hat{y}_i$ باشد، مقدار R^2 برابر یک می‌شود و هرگونه انحرافی از این حالت باعث می‌شود که مقدار R^2 از یک کوچکتر شود.

ضریب تعیین تصحیح شده[3]: ضریب تعیین تصحیح شده یکی از پارامترهایی است که در ارزیابی و مقایسه مدل‌های مختلف با تعداد متفاوت متغیرهای پیش‌گو استفاده می‌شود و از آن‌جایی که R^2 (ضریب تعیین)، تعداد پارامترهای موجود در مدل را به حساب نمی‌آورد، به‌این منظور از ضریب تعیین تصحیح شده طبق رابطه زیر برای انتخاب مدل برتر استفاده می‌شود:

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1} \cdot \frac{SSE}{SST} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$ رابطه(۲-۲۷)

که در این رابطه p تعداد متغیرهای مستقل و n تعداد ترکیبات مورد بررسی می باشد.

آماره F: آزمون F یا آزمون فیشر در واقع آزمون معنی‌دار بودن آماری در تحلیل رگرسیون ساده و چند متغیره است و برابر با نسبت میانگین توان‌های دوم رگرسیون (MSR) به میانگین توان‌های دوم باقی‌مانده‌ها (MSE) است. بیان ریاضی آن به‌صورت زیر می‌باشد:

[1] Sum Square Total
[2] Sum Square Error
[3] Adjusted determination coefficient

$$F = \frac{MSR}{MSE} = \frac{SSR/df_m}{SSE/df_{res}} \quad , \quad df_m = p \quad , \quad df_{res} = n - p - 1$$ رابطه(۲-۲۸)

df_m درجه آزادی مدل، p تعداد متغیرهای مستقل مدل، n تعداد کل ترکیبات مربوط به مدل، df_{res} درجه آزادی باقی‌مانده‌ها است. و SSE، SSR در روابط قبل توضیح داده شد.

آماره T: برای بررسی معنی دار بودن ضرایب متغیرهای مستقل در هر مدل از آماره T استفاده می‌شود. به وسیله آماره T صحت فرضیه صفر مورد بررسی قرار می گیرد: هرچه قدر مطلق آماره T بزرگتر باشد فرضیه صفر با قدرت بیشتری رد می‌شود. بیان ریاضی آن به‌صورت زیر می‌باشد:

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{s.e.(\hat{\beta})}$$ رابطه(۲-۲۹)

$\hat{\beta}$ به عنوان برآوردگر پارامتر β و $s.e(\hat{\beta})$ انحراف استاندارد برآوردگر β می‌باشد.

مجموع توان دوم باقی ما نده ها[1] (PRESS): برابر مجموع توان دوم تفاوت بین مقدار کمیت مشاهده شده(y_i) و مقدار برآورد شده $(\hat{y}_i)$ است.

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ رابطه(۲-۳۰)

خطای استاندارد پیش‌بینی[2] (SEP)

$$SEP = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n}}$$ رابطه(۲-۳۱)

خطای مطلق میانگین[3] (MAE)

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}|}{n}$$ رابطه(۲-۳۲)

[1] Predictive Residual Sum of Squares
[2] Standard Error of Prediction
[3] Mean Absolute Error

خطای نسبی پیش‌بینی [1] (REP)

$$REP(\%) = \frac{100}{\bar{y}} \times \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}}$$ رابطه(۲-۳۳)

میانگین توان‌های دوم خطا [2] (MSE)

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}$$ رابطه(۲-۳۴)

میانگین خطای نسبی [3] (MRE)

$$MRE = \frac{\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}}{y_i}\right|}{n} \times 100$$ رابطه(۲-۳۵)

تلرانس [4]: همان‌طور که می‌دانیم وجود یا عدم وجود هم‌خطی [5] در بین متغیرهای پیش‌بینی کننده مسئله مهمی است که باید بررسی شود. پدیده هم‌خطی می‌تواند به یکی از دو دلیل زیر باشد:

- ❖ حداقل دو تا از متغیرهای مستقل از لحاظ خطی خیلی به‌هم وابسته باشند. یعنی قدر مطلق ضریب همبستگی بین آن دو خیلی به یک نزدیک باشد.
- ❖ حداقل یکی از متغیرهای مستقل وابستگی شدیدی به مجموعه سایر متغیرهای مستقل داشته باشد که معمولاً این مورد مهم‌تر است.

لذا برای کشف پدیده هم‌خطی برای هر متغیر مستقل مقداری محاسبه می‌شود که به آن تلرانس گفته می‌شود. تلرانس برای هر متغیر x_i از رابطه زیر به دست می‌آید[۶۰]:

$$T_i = 1 - R_i^2 \qquad i = 1.2.3 \cdots . p$$ رابطه(۲-۳۶)

که در آن p تعداد متغیرهای پیش‌گو و R_i^2 مجذور ضریب همبستگی چندگانه است که از رگرسیون x_i بر تمام متغیرهای پیش بینی کننده ی دیگر به‌دست می‌آید. پیشنهاد شده است که اگر برای هر یک از

[1] Relative Error of Prediction
[2] Mean Square Error
[3] Mean Relative Error
[4] Tolerance
[5] Multicollinearity

متغیرهای مستقل، T_i از ۰/۱ بیش‌تر باشد، پدیده هم‌خطی وجود ندارد.

عامل تورم واریانس[1]: در برخی مواقع از عامل افزایش واریانس به‌جای تلرانس استفاده می‌شود که رابطه آن با تلرانس به صورت زیر است:

$$VIF = \frac{1}{T_i} = \frac{1}{1-R_i^2} \qquad i = 1.2.3.\cdots.p$$ رابطه(۲-۳۷)

واضح است که اگر x_i رابطه خطی با سایر متغیرها داشته باشد، آن‌گاه R_i^2 نزدیک به یک است و عامل افزایش واریانس بزرگ می‌شود. مقادیر VIF_i بین محدوده ۱ تا۱۰ قابل قبول بوده و مقادیر بیشتر از ۱۰ اغلب به عنوان علامتی از این‌که داده‌ها مشکل هم‌خطی دارند، تلقی می‌شود. اگر هیچ‌گونه رابطه خطی بین متغیرهای پیش بینی کننده نباشد، آن‌گاه R_i^2 صفر و VIF_i برابر یک است. [۷۰،۶۹].

۲-۷-۲ استفاده از نمودار برگشتی

در نمودار برگشتی مقادیر پیش بینی شده‌ی کمیت مورد نظر برحسب مقادیر تجربی رسم می‌شود و به کمک ضریب تعیین (R^2) به‌دست آمده از نمودار، پراکندگی نقاط اطراف خط برگشت تعیین می‌شود. محدوده‌ی تغییرات ضریب تعیین به‌دست آمده بین ۰ و ۱ است. هرچه قدر به ۱ نزدیک‌تر با شد، نتایج پیش‌بینی شده به مقادیر واقعی نزدیک‌ترند. اما اگر این مقدار برابر صفر با شد، بین داده‌ها هیچ گونه همبستگی وجود ندارد.

۲-۷-۳ استفاده از نمودار خطای باقی‌مانده‌ها

منظور از عبارت خطای باقی‌مانده، اختلاف بین مقادیر پیش‌بینی شده و مقادیر تجربی است. اگر پراکندگی مقادیر در دو طرف نمودار صفر باشد، این امر نشان‌دهنده‌ی تصادفی بودن خطاست. ولی اگر عمده‌ی نقاط در این نمودار، در یک طرف خط صفر باشند، این بدان معناست که خطای جهت‌داری رخ داده است.

[1] Variance Inflation Factor (VIF)

۲-۷-۴ استفاده از سری آزمون

اعتبار و اهمیت مدل‌های پیش‌بینی کننده وقتی مشخص می‌شود که بتوانند خاصیت مورد نظر را برای مولکول‌های جدیدی که در دسته‌ی آموزش موجود نیستند پیش‌بینی کنند. بدین منظور از ابتدای کار تعدادی از مولکول‌ها به‌عنوان سری آزمون خارجی کنار گذاشته می‌شود و در نهایت از مدل منتخب بهینه شده، جهت پیش‌بینی خاصیت مورد نظر مولکول‌های موجود در سری تست که در مدل‌سازی استفاده نشده‌اند، استفاده می‌شود.

۳ فصل سوم

مطالعه ارتباط کمی ساختار- فعالیت مشتقات آزابنزن‌ها با استفاده از شبکه عصبی مصنوعی و جنگل‌های تصادفی

۳-۱ مدل‌سازی فعالیت دارویی مشتقات آزابنزن به عنوان بازدارنده‌های غیرنوکلئوزیدی ویروس HIV

در راستای مطالعه ارتباط کمی ساختار فعالیت ضد ایدز گروهی از مشتقات آزابنزن، از روش شبیه‌سازی داکینگ مولکولی برای تجزیه و تحلیل بر هم کنش های موثر بین پروتئین و مهارکننده‌های آن استفاده شد و توصیف‌کننده‌های جدید به دست آمد و با روش انتخاب متغیر گام به گام و روش نوین لاسو توصیف‌کننده‌های مناسب انتخاب شدند. این توصیف‌کننده‌های مربوط به برهم کنش‌های موثر، برای مدل سازی QSAR استفاده شدند. جهت بهبود در مدل سازی، در بخش تولید متغیرهای مناسب، توصیف‌کننده‌های ساختاری با توجه به ساختار پایه مشتقات آزابنزن نیز پیشنهاد داده شدند. بهترین توصیف‌کننده‌های ساختاری و داکینگ مولکولی با استفاده از روش‌های رگرسیون گام به گام و لاسو انتخاب شده و مدل‌سازی شبکه عصبی مصنوعی[1] و جنگل‌های تصادفی[2] و ارزیابی مدل برتر در این بخش مورد بررسی قرار گرفته است.

۳-۱-۱ نرم افزارهای مورداستفاده

در علم کمومتریکس استفاده از بسته‌های نرم‌افزاری متفاوت برای انجام تمام مراحل مدل‌سازی به‌کار گرفته می‌شود که در ادامه به اختصار بسته‌های نرم‌افزاری استفاده شده در این تحقیق معرفی می‌شوند.

۳-۱-۱-۱ بسته نرم‌افزاری Hyperchem

از بسته نرم افزاری Hyperchem8.1 [۷۱] برای رسم شکل مولکول‌ها و بهینه‌سازی ساختار با استفاده از روش‌های کوانتومی و مکانیکی، استفاده می‌شود. به‌کمک این برنامه می‌توان طول پیوند، زاویه پیوندی و زوایای پیچشی را در مولکول تعیین کرد. داده‌های حاصل از این نرم افزار را می‌توان به‌عنوان ورودی به سایر نرم افزارها معرفی نمود. هم‌چنین به کمک این نرم افزار می‌توان تعدادی از توصیف‌کننده‌ها از

[1] Artificial Neural Network (ANN)
[2] Random forest(RF)

جمله حجم مولی و قطبش‌پذیری را محاسبه کرد.

۳-۱-۱-۲ نرم افزار AutoDock 4.2

نرم افزار AutoDock 4.2 یک مجموعه از نرم افزار مدل‌سازی مولکولی است که قابل اجرا بر روی سیستم عامل ویندوز می‌باشد[۷۲]. این نرم افزار رایگان و برای عموم در دسترس می‌باشد. این نرم افزار برای پیش‌بینی چگونگی اتصال مولکول‌هایی به‌عنوان سوبسترا یا دارو به یک مولکول گیرنده با ساختار سه بعدی معلوم طراحی شده است و از انواع الگوریتم‌ها برای ارزیابی انرژی و یافتن موقعیت‌های مناسب اتصال استفاده می‌کند.

نرم افزارAutodock4.2 شامل یک رابط کاربر است که ساختار گرافیکی مولکول‌ها را نمایش می‌دهد. این نرم افزار قادر به ساخت مولکول‌ها با استفاده از الگوهای ساخت، تغییر نوع اتم و پیوند و ایجاد ساختارهای جدید از ساختارهای موجود است. کار با این نرم افزار آسان است و در ضمن دارای قابلیت طراحی مولکول‌ها و به حداقل رساندن انرژی آن‌ها می‌باشد که باعث بالا رفتن سرعت کار می‌شود.

۳-۱-۱-۳ نرم افزار ViewerLite5.0

این نرم افزار امکان بررسی جزیی ماکرومولکول و لیگاند را به کاربر می‌دهد با استفاده از این نرم افزار می توان ماکرومولکول را مشاهده، ویرایش و ... کرد. همین‌طور می‌توان جایگاه فعال پروتئین را مشاهده و تعیین مختصات نمود. علاوه بر این می‌توان طول پیوندهای هیدروژنی و هیدروفوبی مربوط به لیگاند و اسیدهای آمینه موجود در جایگاه فعال را نیز مورد بررسی قرار داد[۷۳]. در این پایان نامه برای آماده سازی لیگاند و پروتئین جهت انجام داکینگ، بررسی حضور پیوندهای هیدروژنی و محاسبه طول پیوندهای هیدروژنی، بررسی اسیدهای آمینه و جدا کردن آن‌ها بر اساس انواع هیدروفیلی و هیدروفوبی و ... از نرم افزار ViewerLite5.0 استفاده شد. از این نرم افزار هم قبل از انجام داکینگ برای آماده سازی لیگاند و پروتئین و هم بعد از داکینگ برای بررسی برهم کنش‌های لیگاند-پروتئین

استفاده می‌شود.

۳-۱-۱-۴ نرم افزار Notepad++

++Notepad برنامه‌ای رایگان برای ویرایش کد اصـــلی برنامه ها و جایگزینی مناســـب برای Notepad ویندوز ا ست. این نرم افزار، زبان‌های برنامه نویسی متعددی را پشتیبانی می‌نماید. امکانات این برنامه شامل تفکیک و قالب بندی کدها، تکمیل خودکار کدها، ویرایش همزمان چند فایل، قابلیت جستجوی کامل و تبدیل فرمت‌های متفاوت عکس به مختصات عددی و ... می‌باشد. در این پایان نامه خروجی داکینگ که به فرمت dlg[1] می‌باشد با نرم افزار Notepad++v.5.3.1 نمایش داده شد و تمامی توصیف‌کننده ها با استفاده از این نرم افزار بازخوانی و استخراج شد. به علاوه در این پایان نامه از این نرم افزار برای انتخاب لیگاند با بهترین انرژی و تبدیل آن به فرمت بانک اطلاعاتی پروتئین[2] اســـتفاده شـــد تا بتوان از این لیگاند در نرم افزار ViewerLite5.0 جهت نمایش پیوند هیدروژنی با پروتئین استفاده کرد.

۳-۱-۱-۵ نرم افزار LigPlot2012

این نرم افزار امکان بررســـی کمپلکس لیگاند و گیرنده را پس از اجرای داکینگ فراهم می‌کند به‌طوری که تمامی پیوندهای هیدروفوبی و هیدروژنی و هم‌چنین طول پیوندهای هیدروژنی را به نمایش می‌گذارد. با ا ستفاده از این نرم افزار برر سی برهم‌کنش‌های ا سیدهای آمینه موجود در جایگاه فعال با اجزای اتمی لیگاند قابل بررسی می‌باشد[۷۴]. در این پایان‌نامه نیز وجود پیوندهای هیدروژنی، طول پیوندهای هیدروژنی و نوع اتم‌های درگیر در پیوند هیدروژنی کمپلکس‌های لیگاند-پروتئین با این نرم افزار مورد بررسی قرار گرفته است.

[1] Docking log file
[2] Protein data bank

۳-۱-۱-۶ نرم‌افزار SPSS[1]

SPSS [۷۵] که نخستین نسخه آن در سال ۱۹۷۰ توسط جمعی از فارغ التحصیلان دانشگاه استانفورد آمریکا ارائه شده، امکان تجزیه و تحلیل آماری داده‌ها را فراهم می‌آورد. برخی از قابلیت‌های این بسته نرم افزاری عبارتند از:

- ❖ محاسبه میانگین ساده برای داده‌ها
- ❖ نمایش اطلاعات به‌صورت متنوع در قالب نمودار و جدول
- ❖ انجام رگرسیون تک متغیره و چند متغیره

در این پایان نامه برای محاسبه همبستگی و هم‌خطی موجود بین متغیرها و برای انجام عملیات انتخاب متغیر به روش گام به گام از SPSS .v.21 استفاده شد.

۳-۱-۱-۷ نرم افزار MATLAB

نرم افزار MATLAB [۷۶] یکی از جامع‌ترین و کارآمد ترین نرم افزارهای علمی و محاسباتی است که طی چند سال گذشته تهیه و به بازار عرضه شد و در سال‌های اخیر با تدوین نسخه‌های جدیدتر و کامل‌تر روز به روز بر غنای آن افزوده می‌شود.

MATLAB[2]به معنای آزمایشگاه ماتریس است. ورودی‌ها اساساً به صورت ماتریس در نظر گرفته می‌شوند و هیچ نیازی به مشخص کردن ابعاد ماتریس نمی‌باشد. در MATLAB حتی اعداد اسکالر، ماتریس‌های (x) به‌حساب می‌آیند و بردارها، حالت خاصی از ماتریس‌های سطری یا ستونی در نظر گرفته می‌شوند. از جمله کاربردهای جالب آن که مورد توجه شیمیدانان قرارگرفته است شبکه عصبی مصنوعی و جنگل‌های تصادفی است که در این پژوهش از آن استفاده شده است. اطلاعات شیمیایی به‌عنوان ورودی و به‌صورت یک ماتریس به نرم افزار داده می‌شود. در محیط این برنامه با استفاده از آرایه‌ها و فرامین موجود امکان مدل‌سازی غیرخطی فعالیت

[1] Statistical Package for the Social Science
[2] Matrix Laboratory

بیولوژیکی با ساختار ترکیبات فراهم می‌گردد.

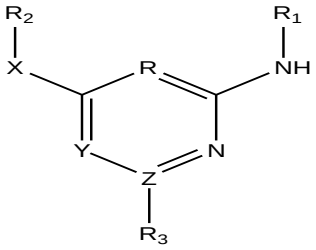
۳-۱-۱-۸ نرم افزار R

R یک زبان و محیط برای محاسبات آماری و گرافیکی می‌باشد. R طیف گسترده‌ای از مدل‌های آماری (مدل‌سازی خطی و غیرخطی، آزمون‌های آماری کلاسیک، تجزیه و تحلیل سری های زمانی، طبقه بندی، خوشه بندی، ...) و تکنیک های گرافیکی را فراهم می‌کند. یکی از نقاط قوت R سهولت در طراحی نمودارهای با کیفیت است که شامل نمادها و فرمول های ریاضی مورد نیاز است. هم‌چنین R یک نرم افزار در دسترس و رایگان است [۷۷]. در این پروژه روش انتخاب متغیر لاسو با استفاده از این نرم افزار انجام شد.

۳-۱-۲ معرفی سری داده‌های مورد استفاده در این مطالعه

سری داده‌ها شامل فعالیت دارویی ۷۳ ترکیب از مشتقات آزابنزن می‌باشد که توسط لیو و همکارانش[1] گزارش شده است [۷۸-۸۲]. اسکلت اصلی لیگاندها (از این پس مشتقات آزابنزن تحت عنوان لیگاندها نام برده می‌شود) در شکل (۳-۱) آورده شده است. جزئیات استخلاف‌ها و مقادیر عددی EC_{50} (بر حسب میکرو مولار) هر لیگاند که مستخرج از مراجع [۷۸-۸۲] می‌باشد در جدول(۳-۱) آورده شده است. EC_{50} غلظتی از دارو می‌باشد که تا ٪۵۰ تکثیر ویروس HIV-1 را مهار می‌کند. ذکر این نکته ضروری است که هرچه دارو قوی‌تر باشد، غلظت کمتری از آن، برای داشتن ٪۵۰ اثر مهار کنندگی، نیاز است. pEC_{50} برابر با منفی لگاریتم مقادیر مربوط به EC_{50} بر حسب مولار می‌باشد که به‌عنوان متغیر وابسته در نظر گرفته شده است.

[1] Liu & et al



شکل(۳-۱) اسکلت اصلی ترکیبات مورد مطالعه

جدول(۳-۱) ساختار و مقادیر EC_{50} ترکیبات مورد مطالعه

No.	R	Y	Z	X	R_1	R_2	R_3	$EC_{50}(\mu M)$	pEC_{50}(M)
1	C	N	C	O	piperidin	2,4,6-TriMe-Ph	-NH-4-CN-Ph	0.14 ± 0.05	6.85
2	C	N	C	O	piperidin	2,6-DiMe-4-CN-Ph	-NH-4-CN-Ph	0.038 ± 0.002	7.42
3	C	N	C	O	piperidin-1-yl)methyl)- 4-SO_2NH_2-Ph	2,4,6-TriMe-Ph	-NH-4-CN-Ph	0.19 ± 0.04	6.72
4	C	N	C	O	piperidin-1-yl)methyl)- 4-SO_2CH_3-Ph	2,4,6-TriMe-Ph	-NH-4-CN-Ph	0.13 ± 0.08	6.89
5	C	N	C	O	piperidin-1-yl)methyl)- 4-Pyridyl	2,4,6-TriMe-Ph	-NH-4-CN-Ph	0.24 ± 0.12	6.62
6	C	N	C	O	piperidin-1-yl)methyl)- 4-SO_2NH_2-Ph	2,6-DiMe-4-CN-Ph	-NH-4-CN-Ph	0.13 ± 0.05	6.89
7	C	N	C	O	piperidin-1-yl)methyl)- 4-SO_2CH_3-Ph	2,6-DiMe-4-CN-Ph	-NH-4-CN-Ph	0.058 ± 0.028	7.24
8	C	N	C	O	piperidin-1-yl)methyl)- 4-Pyridyl	2,6-DiMe-4-CN-Ph	-NH-4-CN-Ph	0.047 ± 0.011	7.33
9	C	N	C	NH	piperidin-1-yl)methyl)- 4-SO_2CH_3-Ph	2,4,6-TriMe-Ph	-NH-4-CN-Ph	0.077 ± 0.065	7.11
10	N	N	N	NH	piperidin-1-yl)methyl)- Pyridin-4-yl	2,4,6-TriMe-Ph	NH_2	0.0049 ± 0.002	8.31
11	N	N	N	NH	piperidin-1-yl)methyl)- 4-COOEt-Ph	2,4,6-TriMe-Ph	NH_2	0.01155 ± 0.0005	7.94
12	N	N	N	NH	piperidin-1-yl)methyl)- 4-$CONH_2$-Ph	2,4,6-TriMe-Ph	NH_2	0.02043 ± 0.019	7.69
13	N	N	N	NH	piperidin-1-yl)methyl)- 4-SO_2NH_2-Ph	2,4,6-TriMe-Ph	NH_2	0.00575 ± 0.002	8.24
14	N	N	N	NH	piperidin-1-yl)methyl)- 4-SO_2Me-Ph	2,4,6-TriMe-Ph	NH_2	0.00555 ± 0.001	8.26
15	N	N	N	NH	piperidin-1-yl)methyl)- Phenyl	2,4,6-TriMe-Ph	NH_2	0.01025 ± 0.0009	7.99
16	N	N	N	NH	piperidin-1-yl)methyl)- 4-NO_2-Ph	2,4,6-TriMe-Ph	NH_2	0.00787 ± 0.001	8.10

No.	R	Y	Z	X	R_1	R_2	R_3	$EC_{50}(\mu M)$	pEC_{50}(M)
17	N	N	N	NH	piperidin-1-yl)methyl)- 4-CN-Ph	2,4,6-TriMe-Ph	NH_2	0.00566 ± 0.003	8.25
18	N	N	N	NH	piperidin-1-yl)methyl)- Pyridin-4-yl	2,4,6-TriMe-Ph	NHMe	0.00725 ± 0.001	8.14
19	N	N	N	NH	piperidin-1-yl)methyl)- 4-COOEt-Ph	2,4,6-TriMe-Ph	NHMe	0.01666 ± 0.007	7.78
20	N	N	N	NH	piperidin-1-yl)methyl)- 4-$CONH_2$-Ph	2,4,6-TriMe-Ph	NHMe	0.00461 ± 0.001	8.34
21	N	N	N	NH	piperidin-1-yl)methyl)- 4-SO_2NH_2-Ph	2,4,6-TriMe-Ph	NHMe	0.00699 ± 0.001	8.16
22	N	N	N	NH	piperidin-1-yl)methyl)- 4-SO_2Me-Ph	2,4,6-TriMe-Ph	NHMe	0.00606 ± 0.006	8.22
23	N	N	N	NH	piperidin-1-yl)methyl)- Pyridin-4-yl	2,4,6-TriMe-Ph	OMe	0.01332 ± 0.002	7.88
24	N	N	N	NH	piperidin-1-yl)methyl)- 4-COOEt-Ph	2,4,6-TriMe-Ph	OMe	0.01471 ± 0.007	7.83
25	N	N	N	NH	piperidin-1-yl)methyl)- 4-$CONH_2$-Ph	2,4,6-TriMe-Ph	OMe	0.00864 ± 0.001	8.06
26	N	N	N	NH	piperidin-1-yl)methyl)- 4-SO_2NH_2-Ph	2,4,6-TriMe-Ph	OMe	0.01062 ± 0.002	7.97
27	N	N	N	NH	piperidin-1-yl)methyl)- 4-SO_2Me-Ph	2,4,6-TriMe-Ph	OMe	0.0119 ± 0.004	7.92
28	C	-C-NO_2	C	O	4-CN-Ph	2,6-DiMe-Ph	H	0.58	6.24
29	C	-C-NO_2	C	O	4-CN-Ph	2,4,6-TriMe-Ph	H	0.056	7.25
30	C	-C-NO_2	C	O	4-CN-Ph	2,6-DiMe-4-CN-Ph	H	0.11	6.96
31	C	-C-NO_2	C	O	4-CN-Ph	2,6-DiMe-4-Br-Ph	H	0.27	6.57
32	C	-C-NO_2	C	O	4-CN-Ph	2,6-DiMe-4-Cl-Ph	H	0.16	6.80
33	C	-C-NO_2	C	O	4-CN-Ph	2,6-Di-OMe-Ph	H	4.28	5.37
34	C	-C-NO_2	C	O	4-CN-Ph	2,6-DiCl-Ph	H	0.62	6.21
35	C	-C-NO_2	C	O	4-CN-Ph	2,4,6-TriCl-Ph	H	0.17	6.77
36	C	-C-NO_2	C	O	4-CN-Ph	2,6-DiCl-4-NO_2-Ph	H	16.8	4.77
37	C	-C-NO_2	C	O	4-CN-Ph	2,4,6-TriBr-Ph	H	0.11	6.96
38	C	-C-NO_2	C	O	4-CN-Ph	2,6-DiBr-4-Me-Ph	H	0.034	7.47
39	C	-C-NO_2	C	O	4-CN-Ph	2,4,6-TriF-Ph	H	0.72	6.14
40	C	-C-NO_2	C	NH	4-CN-Ph	2,4,6-TriMe-Ph	H	30.17	4.52
41	C	-C-NO_2	C	NH	4-CN-Ph	2,6-DiBr-4-Me-Ph	H	3.85	5.41
42	C	-C-NO_2	C	NH	4-CH_3-Ph	2,4-DiBr-6-F-Ph	H	3.05	5.52
43	C	-C-NO_2	C	O	4-NO_2-Ph	2,6-DiMe-4-CN-Ph	H	0.72	6.14
44	C	-C-Cl	N	O	p-Cl-Ph	2,4,6-TriMe-Ph	-	0.139 ± 0.050	6.86
45	C	-C-Cl	N	O	p-Me-Ph	2,4,6-TriMe-Ph	-	0.201 ± 0.043	6.70
46	C	-C-Cl	N	O	p-CN-Ph	2,4,6-TriMe-Ph	-	0.079 ± 0.015	7.10
47	C	-C-Cl	N	O	p-CN-Ph	2,4,6-TriCl-Ph	-	0.155 ± 0.042	6.81

No.	R	Y	Z	X	R_1	R_2	R_3	$EC_{50}(\mu M)$	pEC_{50}(M)
48	C	-C-Cl	N	O	p-CN-Ph	2,4,6-TriBr-Ph	-	0.18 ± 0.06	6.74
49	C	-C-Cl	N	O	p-CN-Ph	2,6-DiBr-4-Me-Ph	-	0.034 ± 0.012	7.47
50	C	-C-Cl	N	O	p-CN-Ph	2,6-DiBr-4-Br-Ph	-	0.40 ± 0.15	6.40
51	C	-C-Cl	N	O	p-CN-Ph	2,6-DiCl-Ph	-	0.22 ± 0.051	6.66
52	C	-C-Cl	N	O	p-CN-Ph	4-CN-2,6-DiMe-Ph	-	0.144 ± 0.029	6.84
53	C	-C-Cl	N	O	p-CN-Ph	2,6-DiMe-Ph	-	0.54 ± 0.21	6.27
54	C	-C=O	N	O	p-Cl-Ph	2,4,6-TriMe-Ph	H	0.78 ± 0.17	6.11
55	C	-C=O	N	O	p-Me-Ph	2,4,6-TriMe-Ph	H	1.55 ± 0.32	5.81
56	C	-C=O	N	O	p-NO_2-Ph	2,4,6-TriMe-Ph	H	5.08 ± 1.00	5.29
57	C	-C=O	N	O	p-OMe-Ph	2,4,6-TriMe-Ph	H	1.59 ± 0.21	5.80
58	C	-C=O	N	O	p-CN-Ph	2,4,6-TriMe-Ph	H	0.26 ± 0.04	6.59
59	C	-C=O	N	O	p-CN-Ph	2,4,6-TriBr-Ph	H	0.63 ± 0.2	6.20
60	C	-C=O	N	O	p-CN-Ph	2,6-DiBr-4-Me-Ph	H	0.21 ± 0.03	6.68
61	C	-C=O	N	O	p-CN-Ph	2,6-DiMe-4-Br-Ph	H	1.19 ± 0.24	5.92
62	C	-C=O	N	O	p-CN-Ph	2,4,6-TriMe-Ph	H	14.4 ± 4.21	4.84
63	C	-C=O	N	O	p-CN-Ph	2,6-DiMe-Ph	H	1.90 ± 0.31	5.72
64	C	-C=O	N	O	p-CN-Ph	4-CN-2,6-DiMe-Ph	H	2.83 ± 0.54	5.55
65	C	-C=O	N	O	p-CN-Ph	2,4,6-TriMe-Ph	Me	1.16 ± 0.25	5.94
66*	C	-C=O	N	O	p-CN-Ph	2,4,6-TriMe-Ph	Me	3.36 ± 0.99	5.47
67	C	C	C	O	2,4,6-TriMe-Ph	4-CN-Ph	-OMe	0.84 ± 0.43	6.08
68	C	C	C	O	2,6-DiMe-4-CN	4-CN-Ph	-OMe	0.70 ± 0.50	6.15
69	C	C	C	O	2,4,6-TriMe-Ph	4-NO_2-Ph	=O	1.40 ± 0.3	5.85
70	C	C	C	O	2,4,6-TriMe-Ph	4-Me-Ph	=O	1.40 ± 0.4	5.85
71	C	C	C	O	2,4,6-TriMe-Ph	4-CN-Ph	=O	0.37 ± 0.10	6.43
72	C	C	C	O	2,6-DiMe-4-CN-Ph	4-CN-Ph	=O	0.15 ± 0.06	6.82
73	C	C	C	O	2,6-DiMe-Ph	4-CN-Ph	=O	1.40 ± 0.17	5.85

*** لیگاند ۶۶ به جای NH موجود در اسکلت اصلی دارای NMe می‌باشد.**

علامت - موجود در جدول، به معنای عدم وجود استخلاف R_3 در این لیگاندهاست.

۳-۱-۳ مدل‌سازی فعالیت دارویی ترکیبات مورد مطالعه با استفاده از توصیف‌کننده‌های حاصل از داکینگ مولکولی

بعد از انتخاب سری داده‌ها، توصیف‌کننده‌های حاصل از داکینگ مولکولی برای این ترکیبات استخراج گردید و سپس این توصیف‌کننده‌ها برای ساخت مدل‌های QSAR با استفاده از الگوریتم‌های

شـبکه عصـبی(ANN) و جنگل‌های تصـادفی (RF) مورد اسـتفاده قرار گرفتند. در ادامه به چگونگی اسـتخراج توصـیف‌کننده‌های حاصـل از داکینگ مولکولی و انتخاب آن‌ها و سـپس به معرفی مدل‌های برتر QSAR و ارائه نتایج حاصل از ارزیابی آن‌ها پرداخته خواهد شد.

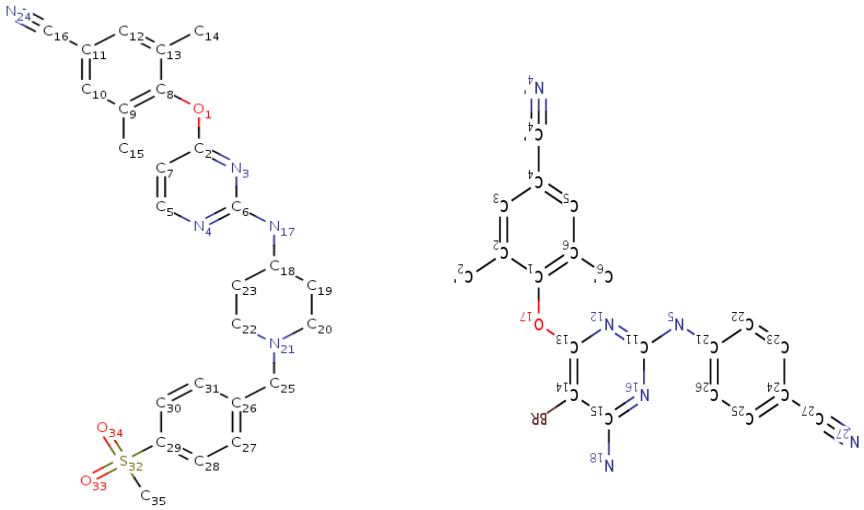
۳-۱-۳-۱ استخراج توصیف‌کننده‌های حاصل از داکینگ مولکولی

تو صیف‌کننده‌های داکینگ مولکولی از ساختار کمپلکس حا صل از وارد کردن لیگاند (ترکیبات مورد مطالعه) در جایگاه فعال پروتئین ا ستخراج گردید. برای پیدا کردن چنین ساختار کمپلک سی در بهترین شـــرایط، ابتدا مراحل آماده ســـازی لیگاندها (لیگاند پروتئین و ترکیبات مورد مطالعه)، آماده سـازی پروتئین، تعیین بهترین شـرایط داکینگ مولکولی به روش اعتبار سـنجی و در نهایت داکینگ لیگاندها در سـاختار پروتئین اجرا گردید. در ادامه به اختصـار روش کار در هر مرحله و نتایج حاصـله آورده خواهد شد.

– استخراج و آماده سازی لیگاند موجود در پروتئین

همانطور که قبلا گفته شـد، توصـیف‌کننده‌های مولکولی از سـاختار یک کمپلکس حاصـل از داکینگ لیگاندها (سـری داده‌ها) در پروتئین مناسـب، اسـتخراج می‌شـوند. شـرایط بهینه داکینگ مولکولی به روش اعتبار ســـنجی و با ورود و خروج‌های متوالی لیگاند موجود در ســـاختار پروتئین به جایگاه فعال پروتئین حا صل می‌گردد. لذا لازم ا ست لیگاند و پروتئین جهت انجام اعتبار سنجی آماده شوند. برای ا ستخراج و آماده سازی لیگاند موجود در جایگاه فعال پروتئین به صورت زیر عمل شد: ابتدا با توجه به شـــباهت نســـبی بین لیگاندهای موجود در جایگاه فعال پروتئین‌ها و لیگاندهای مورد مطالعه و با توجه به نتایج و گزارش های موجود در مورد مناســـب‌ترین آنزیم های مهارکننده ی غیرنوکلئوزیدی نسخه‌بردار معکوس که حاوی لیگاندهای درونی نیز هستند[۵۱]، ساختار اولیه این نوع آنزیم‌ها انتخاب گردید این پروتئین‌ها با شناسه ساختار 3M8Q و 3MEC که حاوی لیگاندهای درونی هســـتند از بانک اطلاعاتی پروتئین [۵۱] دریافت و به فرمت بانک اطلاعاتی پروتئین (PDB) ذخیره

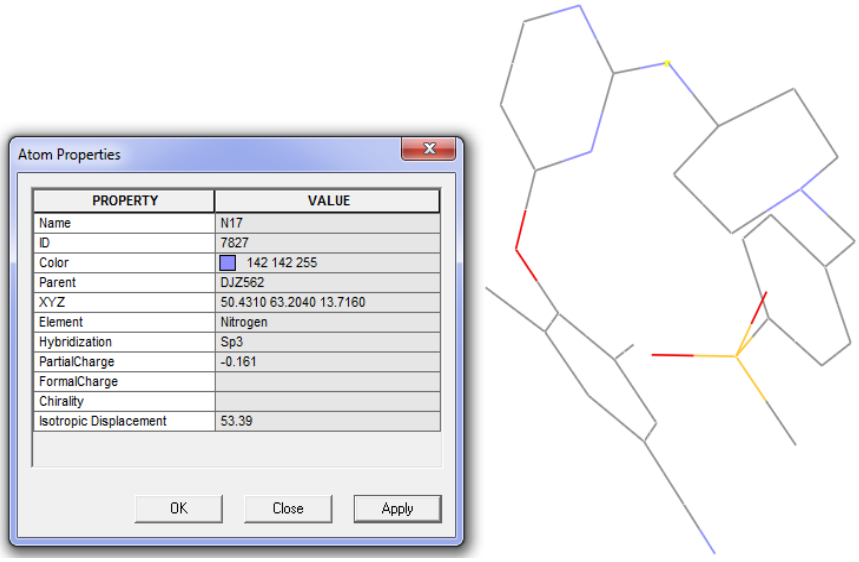
شدند. فایل دانلود شده در نرم افزار ViewerLite5.0 باز شده و سپس تمامی اسیدهای آمینه (ساختار آنزیم)، مولکول‌های آب و کوفاکتورهای همراه از این فایل حذف شـــدند. باقی‌مانده ســـاختار که همان لیگاند موجود درجایگاه فعال پروتئین می‌باشد به فرمت PDB ذخیره گردید. لیگاندهای استخراج شده از این پروتئین‌ها درشکل(۳-۲) نشان داده شده است. این لیگاندها بعد از تعیین مختصات مرکز ثقل در مرحله اعتبار سنجی مورد استفاده قرار خواهند گرفت.



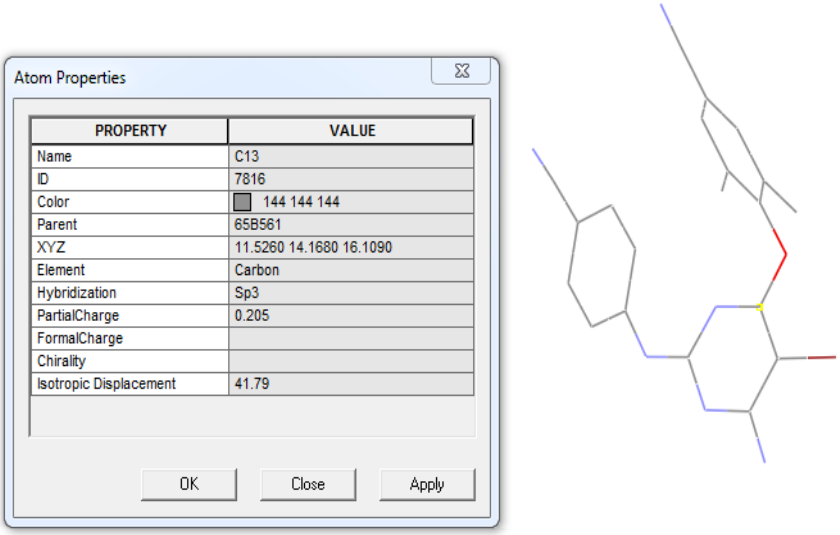
الف) ب)

شکل(۳-۲) ساختار لیگاندهای موجود در پروتئین الف)موجود در پروتئین 3MEC ب) موجود در پروتئین 3M8Q

از آنجایی‌که در ســـاختار کریســـتالوگرافی پروتئین‌ها، این لیگاندها با جهت‌گیری مشـــخص در جایگاه فعال پروتئین قرار دارند، لذا مختصـــات مرکز ثقل لیگاندها نشـــان‌دهنده جایگاه فعال پروتئین مربوطه می‌باشد و جهت تعیین مختصات جایگاه فعال در عملیات داکینگ مورد استفاده قرار می‌گیرد. ا ستخراج مخت صات مرکز ثقل لیگاندها در نرم افزار ViewerLite5.0 طبق مراحل ن شان داده شده در شکل(۳-۳) و شکل(۳-۴) انجام گردید. نتایج به د ست آمده مختصات ۵۰/۴۳۱ ، ۶۳/۲۰۴، ۱۳/۷۱۶ برای مرکز ثقل لیگاند موجود در پروتئین 3M8Q و مختصـــات ۱۱/۵۲۸، ۱۴/۱۶۸ و ۱۶/۱۰۹ برای مرکز ثقل لیگاند موجود در پروتئین 3MEC را نشان داد.



شکل(۳-۳) مختصات جایگاه فعال به‌دست آمده از جایگاه لیگاند در پروتئین 3M8Q



شکل(۳-۴) مختصات جایگاه فعال به‌دست آمده از جایگاه لیگاند در پروتئین 3MEC

– آماده سازی پروتئین جهت استفاده در فرایند اعتبار سنجی درونی و داکینگ مولکولی

همان‌طور که در بخش قبل گفته شد فایل کریستالوگرافی کمپلکس‌های پروتئین‌ها با لیگاندهای درونی از بانک اطلاعاتی پروتئین با شناسه‌های ساختار کریستالوگرافی 3M8Q و 3MEC دریافت شدند. ارزش تفکیک[1] مربوط به ساختارهای کریستالوگرافی برای پروتئین با شناسه کریستالوگرافی 3M8Q برابر با ۲/۷ Aº و برای پروتئین با شناسه کریستالوگرافی 3MEC برابر با Aº

[1] Resolution value

۲/۳ می‌باشد که این مقادیر ارزش تفکیک برای ساختارهای کریستالوگرافی پروتئین‌ها که در مطالعات داکینگ مورد ا ستفاده قرار می‌گیرند قابل قبول ه ستند[۵۱]. برای آماده سازی پروتئین جهت فرایند اعتبار ســـنجی و داکینگ لیگاندهای مورد مطالعه (ترکیبات ســـری داده‌ها) در پروتئین‌های مربوطه، مراحل زیر انجام شـد. ابتدا فایل پروتئین دانلود شـده و در نرم افزار ViewerLite5.0 فراخوانی شـد و مولکول‌های آب، کوفاکتورها و سایر مولکول‌هایی که در جایگاه فعال پروتئین (گیرنده) حضور ندارند و لیگاند موجود در ساختار کریستالوگرافی حذف شدند. ساختار ا صلاح شده به فرمت بانک اطلاعاتی پروتئین (PDB) ذخیره گردید. دسـتکاری و آماده سـازی بیشـتر پروتئین در نرم افزار Autodock4.2 انجام شد.

– فرایند اعتبار سنجی (داک-ریداک) جهت تعیین بهترین شرایط داکینگ

یکی از مهم‌ترین مراحل در مطالعات داکینگ مولکولی تعیین بهترین شـــرایط داک لیگاند در جایگاه فعال پروتئین می‌باشـد. معمولا این شـرایط از طریق فرایند اعتبار سـنجی تعیین می‌شـود. در فرایند اعتبارســـنجی لیگاند موجود در جایگاه فعال پروتئین به طور متوالی و در شـــرایط مختلف در همان پروتئین داک می‌گردد تا شـــرایطی که در آن بهترین داکینگ با توجه به میزان انرژی اتصـــال، تعداد خوشه‌ها، تعداد کنفورماسیون‌های موجود در برترین خوشه و زمان اجرا انجام می شود به د ست آید. فرایند اعتبار سنجی در نرم افزار Autodock4.2 طی مراحل زیر انجام گرفت.

۱. فراخوانی لیگاند و پروتئین

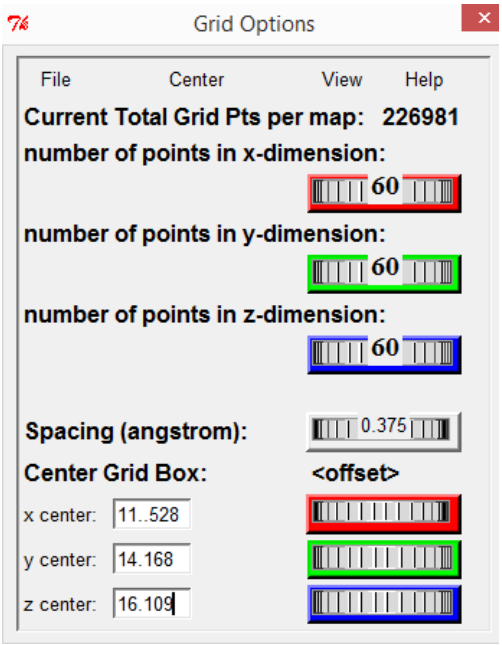
در این مرحله لیگاند آماده شده با توجه به مراحل ذکر شده در بخش آماده سازی لیگاند در نرم افزار Autodock4.2 فراخوانی شد. به منظور آماده سازی بی شتر پروتئین فایل آماده شده در مرحله قبل در نرم افزار Autodock4.2 فراخوانی شد. از آنجایی‌که فایل‌های کریستالوگرافی کمبود هیدروژن دارند در این نرم افزار، کلیه اتم‌های هیدروژن به سـاختار کریسـتالوگرافی افزوده شـد[۵۲]. در مرحله بعد، اتم‌های هیدروژن غیر قطبی (متصـــل به اتم‌های کربن) در اتم کربن مربوطه ادغام شـــد و بار

الکتریکی کلمن نیز به ساختار کریستالوگرافی افزوده شد.

۲. تنظیم فاصله بین نقاط شبکه، ابعاد شبکه، مختصات مرکز ثقل

با توجه به بخش(۲-۴-۴) لازم به ذکر است که فاصله بین نقاط شبکه[1] بر روی ۰/۳۷۵ A^o تنظیم شده است که در حدود یک چهارم طول پیوند یگانه کربن-کربن (کربن با هیبریداسیون $sp3$) می‌باشد و این فاصله در بخش اتوگرید تنظیم شد. بر اساس حجم مولکولی لیگاند طراحی شده، شبکه‌ای با ابعاد مشخص در راستای محورهای سه‌گانه مختصات که در برگیرنده جایگاه فعال گیرنده مورد نظر باشد، در نظر گرفته شد، در این پروژه با توجه به دستورالعمل‌های اجرای داکینگ مولکولی موجود در مقالات متفاوت، شبکه‌ای با ابعاد (A^{o3}) ۶۰×۶۰×۶۰ در نظر گرفته شد[۸۳-۸۶]. مختصات مرکز شبکه با استفاده از مختصات مرکز ثقل[2] لیگاند کریستالوگرافی اولیه که در بخش آماده سازی لیگاند توضیح داده شد در تنظیمات مربوط به مختصات شبکه[3] نرم افزار تعریف شد، مختصات x، y و z مرکز ثقل برای پروتئین با شناسه‌های ساختار کریستالوگرافی 3M8Q و 3MEC همانطور که دربخش آماده سازی لیگاند ذکر شد در این مرحله از نرم افزار تعریف شد. شکل(۳-۵) روند تعریف شبکه و مختصات آن را نشان می‌دهد.

[1] Spacing
[2] Center box
[3] Autogrid



شکل(۳-۵) تعیین مختصات شبکه در بخش اتوگرید نرم افزار

۳. انتخاب الگوریتم جستجو در داکینگ مولکولی

فرایند داکینگ مولکولی، با اســتفاده از الگوریتم‌های متفاوت ژنتیک الگوریتم لامارکین[۱]، ژنتیک الگوریتم[۲]، جســـتجو محلی[۳] و ارزیابی انرژی[۴] اجرا می‌شـــود که از این بین الگوریتم ژنتیک لامارکین روش اجرایی کارآمدتر با نتایج داکینگ تکرار پذیرتر نسبت به سایر الگوریتم‌ها می‌با شد[۵۲]. در این پروژه نیز از الگوریتم ژنتیک لامارکین استفاده شده است.

۴. تنظیم پارامترهای الگوریتم ژنتیک لامارکین

در این بخش باید مقادیر مربوط به پارامترهای جســـتجو الگوریتم ژنتیک تنظیم شـــود. تعداد دفعات اجرای الگوریتم[۵] فاکتور مهمی اســـت که باید ارزیابی و بهینه گردد. افزایش ســـطح این فاکتور منجر به افزایش دقت و صحت محاسبات خواهد گردید اما باید بررسی شود که آیا میزان افزایش دقت و صحت روش، در مقایسه با میزان زمان لازم برای اجرای شبیه سازی داکینگ مولکولی به‌صرفه بوده

[1] Lamarckian Genetic Algorithm(LGA)
[2] Genetic Algorithm(GA)
[3] Local Search
[4] Evaluate energy(EPDB)
[5] Number of genetic algorithm runs

و تعادلی میان عوامل برقرار باشـــد. در نرم افزار Autodock4.2 برای حداکثر تعداد ارزیابی‌های انرژی ســه حد کوتاه[1]، متوســط[2] و طولانی[3] تعریف شــده اســت که اغلب در پروژه‌های محاســباتی متوســط محاسبه انرژی به تعداد متوسط یعنی ۲۵۰۰۰۰۰ تعریف می‌گردد[۸۷-۸۹] و در این بررسی نیز برای تمامی داک‌های پروتئین و لیگاندهای درون پروتئین حداکثر تعداد ارزیابی‌های انرژی در سطح متوسط قرار داده شــد. ســایر تنظیمات پارامترهای جســتجو الگوریتم ژنتیک در حالت پیش فرض برنامه قرار داشـــته و تغییری در آن‌ها ایجاد نگردید و فقط تعداد اجرای الگوریتم ژنتیک[4] برای لیگاندهای موجود در ساختار پروتئین در فرایند اعتبار سنجی (داک-ریداک) در مقادیر ۱۰۰، ۱۵۰ و ۲۰۰ قرار داده شد.

۵. تنظیم گزینه‌های اجرای داکینگ

تمامی گزینه‌های اجرای داکینگ در مقادیر پیش فرض برنامه Autodock4.2 قرار داده شـــد. برخی از این گزینه‌ها عبارتند از میزان جابجایی لیگاند در فضـــا در راســـتای محور y،x و z (انتقال[5])، دوران در فضای سه بعدی حول محورهای y،x و z (چهارگان [6]) و میزان چرخش[7].

پس از تنظیم شـــرایط فرایند اعتبار ســـنجی، داکینگ لیگاند و پروتئین در جایگاه فعال همان پروتئین در شــرایط گفته شــده انجام گرفت و تمامی پارامترهای تعداد اجرا الگوریتم ژنتیک، RMSD اتصال لیگاند و پروتئین باید از $2A^o$ کمتر باشد تا بتوان داکینگ را با شرایط تعریف شده ادامه داد و اگر نه باید شــرایط را تا جایی‌که به زیر $2A^o$ بیاید تغییر داد، میانگین انرژی اتصــال که هرچه منفی تر باشد از لحاظ انرژی در بهترین حالت خود است، این انرژی زمانی مطلوب خواهد بود که خوشه مورد نظر حاوی تعداد قابل توجهی کنفورماســیون باشــد (تعداد کل کنفورماســیون‌ها با تعداد دفعات اجرای الگوریتم ژنتیک برابر است)[۹۰]. تعداد خوشه‌های های چند کنفورماسیونه، تعداد کنفورماسیون‌های

[1] Short
[2] Medium
[3] Long
[4] Number of GA runs
[5] Translation
[6] Quaternion
[7] Torsion

موجود در خو شه اول و زمان اجرا نیز حائز اهمیت ا ست. داکینگی موفق ا ست که تعداد کمی خو شه داشته باشد و ایده آل ترین حالت وجود یک خوشه با حداکثر کنفورماسیون‌های مجتمع شده می‌باشد. هرچه زمان اجرا کمتر باشـد داکینگ لیگاندها با پروتئین در زمان کوتاه تری انجام گرفته و برای کاربر به صـــرفه خواهد بود. تمامی پارامترهای ذکر شـــده برای تعداد اجراهای متفاوت الگوریتم ژنتیک لامارکین به دست آمد. نتایج حاصله در جدول(۳-۲) و جدول(۳-۳) آورده شده است.

جدول(۳-۲) نتایج حالات متفاوت اجرا و انتخاب حالت بهینه برای پروتئین با شناسه کریستالوگرافی 3MQ8

تعداد اجرا الگوریتم ژنتیک	RMSD (Å)	**میانگین انرژی اتصال(kcal/mol)**	**تعداد خوشه‌های های چند کنفورماسیونه**	**تعداد کنفورماسیون‌های موجود در خوشه اول**	**زمان اجرا**
۱۰۰	۰/۴۴	-۱۱/۶۹	۱۳	۲۸	۱ ساعت و ۱۴ دقیقه
۱۵۰	**۰/۶۹**	**-۱۲/۲۰**	۷	۷۵	**۱ ساعت و ۵۶ دقیقه**
۲۰۰	۰/۵۱	-۱۱/۷۶	۱۷	۵۴	۲ ساعت و ۲۲ دقیقه

جدول(۳-۳) نتایج حالات متفاوت اجرا و انتخاب حالت بهینه برای پروتئین با شناسه کریستالوگرافی 3MEC

تعداد اجرا الگوریتم ژنتیک	RMSD (Å)	میانگین انرژی اتصال(kcal/mol)	تعداد خوشه‌های چند کنفورماسیونه	تعداد کنفورماسیون‌های موجود در خوشه اول	زمان اجرا
۱۰۰	۰/۶۲	-۱۱/۶۴	۱۰	۴۳	۳۴ دقیقه
۱۵۰	**۰/۶۲**	**-۱۱/۹۳**	**۵**	**۱۳۴**	**۵۵ دقیقه**
۲۰۰	۰/۶۳	-۱۱/۸۳	۱۲	۱۰۴	۲ ساعت و۲۲ دقیقه

با توجه به نتایج هر دو جدول حالت بهینه دارای اعتبار برای ادامه فرایند داکینگ، انتخاب و مشخص گردید. داکینگ مولکولی با اجرای ژنتیک الگوریتم ۱۵۰، دارای RMSD معتبر (ارزش تفکیک < ۲ A^o)، کمترین تعداد خوشه، بیشترین تعداد کنفورماسیون در خوشه برتر و زمان قابل قبول از نظر کاربر می‌باشد. لذا شرایط مشخص شده در جداول فوق برای تمامی لیگاندهای مورد مطالعه با توجه به پروتئین‌های مربوطه انتخاب گردید.

– داکینگ لیگاند های مورد مطالعه در پروتئین‌های مربوط

بعد از آماده ســـازی لیگاندها و انجام عملیات داکینگ مولکولی لیگاندها با پروتئین مربوطه در شرایط بهینه حا صل از فرایند اعتبار سنجی، تو صیف‌کننده‌های حا صل از داکینگ مولکولی برای این ترکیبات اســـتخراج گردید. در ادامه به نوع توصـــیف‌کننده‌های حاصـــل از داکینگ مولکولی و انتخاب بهترین آن‌ها و ســـپس به معرفی مدل‌های برتر QSAR و ارائه نتایج حاصـــل از ارزیابی آن‌ها پرداخته خواهد شد.

– آماده سازی لیگاندهای مورد مطالعه قبل از استفاده در فرایند داکینگ مولکولی

در این مرحله از مطالعه، ساختار مولکولی هر ترکیب ابتدا در نرم افزار Hyperchem8.1 تر سیم شد. سپس با احتساب اتم‌های هیدروژن، ساختار سه بعدی ترکیبات با ا ستفاده از روش نیمه تجربی کوانتومیAM1 بهینه شـــد و بهینه‌ســـازی تا زمانی‌که جذر میانگین توان‌های دوم گرادیان انرژی به ۰/۰۰۱ کیلو کالری بر مول بر سد ادامه یافت و در نهایت لیگاندها بهینه شدند و با فرمت مولی[1] ذخیره گردیدند. توجه به این نکته که فایل ورودی نرم افزار داکینگ باید با فرمت بانک اطلاعاتی پروتئین با شد ضروری ا ست. از این‌رو تمامی لیگاندها در نرم افزار ViewrLite5.0 باز شد و علاوه بر برر سی تمامی اســـتخلاف‌ها و تایید درســـتی رســـم ســـاختار، لیگاندها با فرمت بانک اطلاعاتی پروتئین ذخیره شدند و از آن‌ها در فرایند داکینگ مولکولی استفاده شد.

– اجرای عملیات داکینگ مولکولی لیگاندها و پروتئین

با توجه به کلیه نکات اشـــاره شـــده دربخش آماده ســـازی پروتئین، پروتئین آماده شـــد و در نرم افزار داکینگ فراخوانی شـــد و تمامی تنظیمات مربوط به افزایش بارکلمن، افزایش هیدروژن، تعریف مختصات جایگاه فعال و شبکه روی آن اعمال گردید و در مرحله نهایی الگوریتم ژنتیک لامارکین برای اجرای فرایند داکینگ با استفاده از شرایط بهینه مشخص شده در جدول(۳-۲) و جدول(۳-۳) مورد

[1] Molfile format

استفاده قرار گرفت. تمامی لیگاندها به صورت جداگانه با پروتئین مربوطه (با توجه به پیشنهاد مقالات سری داده‌ها لیگاندهایی با شماره ترکیب (۱-۴۳) با پروتئین 3M8Q و لیگاندهایی با شماره ترکیب (۴۴-۷۳) با پروتئین 3MEC داک شدند) در نرم افزار Autodock4.2 فراخوانی شده و با شرایط یکسان تعیین شده توسط فرایند اعتبار سنجی داک شدند و نتایج خروجی آن‌ها با فرمت پیش‌فرض dlg[1] ذخیره گردید. خروجی داکینگ با هدف ا ستخراج تو صیف‌کننده‌هایی با اطلاعات ارز شمند مورد آنالیز قرار گرفت.

۳-۱-۳-۲ استخراج توصیف‌کننده‌های حاصل از داکینگ مولکولی

برای ایجاد ارتباط ریاضی بین ساختار– فعالیت، به مقادیر عددی نیازمند است، بر هم کنش‌های بین لیگاندها با پروتئین با استفاده از داکینگ مولکولی آنالیز شد و مجموعه‌ای از توصیف‌کننده‌های عددی که اغلب از نوع انرژی هستند با استفاده از نرم افزار Notpad++ استخراج گردید و این توصیف‌کننده‌ها عبارتند از:

۱. انرژی آزاد اتصال محاسبه شده[2] (BE)
۲. انرژی واندروالس(vdW)+ پیوند هیدروژنی(Hbond)+ حلال‌زدایی (desolv)، (VE)
۳. انرژی بین مولکولی نهایی[3] (IE)
۴. انرژی الکتروستاتیک[4] (EE)
۵. مجموع انرژی داخلی نهایی[5] (TI)
۶. انرژی آزاد چرخشی[6] (TE)
۷. انرژی سیستم‌های آزاد[7] (UE)
۸. ثابت بازداری (K_i)
۹. منفی لگاریتم ثابت بازداری (-Log K_i)

این توصیف‌کننده‌ها حاوی اطلاعات مفیدی بوده و بسیار قابل توجه می‌باشند زیرا اطلاعات

[1] Docking log file
[2] Estimated Free Energy of Binding
[3] Final Intermolecular Energy
[4] Electrostatic Energy
[5] Final Total Internal Energy
[6] Torsional Free Energy
[7] Unbound System's Energy

مربوط به بر هم کنش لیگاندها-پروتئین را در بر گرفته‌اند. همه این توصیف‌کننده‌ها برای ۷۳ لیگاند مورد مطالعه محاسبه و سپس به عنوان ورودی در مدل‌سازی‌های QSAR مورد استفاده قرار گرفت.

۳-۱-۳-۳ انتخاب توصیف‌کننده‌های مناسب داکینگ مولکولی

انتخاب توصیف‌کننده‌های مناسب، مهم‌ترین مرحله در مطالعات QSAR است تا ارتباط بین ساختار مولکول و فعالیت دارویی تف سیر پذیر با شد. تو صیف‌کننده‌های حا صل از داکینگ لیگاندها و پروتئین، مجموعا ۹ توصیف‌کننده ذکر شده در بخش قبل می‌باشد. داده‌های مربوط ذخیره شدند. مجموعه ارزیابی و مجموعه آزمون که هرکدام شامل ۱۱ ترکیب که به طور تصادفی انتخاب شده‌اند می‌باشد. داده‌های مربوط به سری ارزیابی و آزمون از مجموع داده‌ها حذف گردید. انتخاب بهترین توصیف‌کننده‌ها با استفاده از داده‌های مربوط به مجموعه آموزش انجام شد. برای انتخاب بهترین توصیف‌کننده‌ها، از روش‌های انتخاب متغیر رگرسیون گام به گام و لاسو استفاده گردید.

با به‌کارگیری نرم افزار SPSS و با اجرای روش رگرسیون گام به گام برای ۹ توصیف‌کننده استخراج شده از داکینگ مولکولی به‌عنوان متغیرهای مستقل، و منفی لگاریتم فعالیت دارویی (pEC_{50}) به عنوان متغیر واب سته، تعداد ۲ تو صیف‌کننده انتخاب شدند که نام و نوع آن‌ها در جدول(۳-۴) نشان داده شده است.

جدول(۳-۴) توصیف کننده‌های انتخاب شده از داکینگ مولکولی با روشSR

No	Symbol	Meaning
۱	TI	Final Total Internal Energy
۲	pK_i	-Log(Estimated Inhibition Constant, Ki)

بار دیگر برای انتخاب متغیرهای حاصل از داکینگ مولکولی، با به‌کارگیری نرم افزار R و با اجرای روش لاسو روی داده‌های مجموعه آموزش رابطه (۱۶-۲) مینمم شد و برای ۹ توصیف‌کننده به‌عنوان متغیرهای مستقل، و منفی لگاریتم فعالیت بازدارندگی (pEC_{50}) به عنوان متغیر وابسته، تعداد ۴

توصیف‌کننده انتخاب شدند که نام و نوع آن‌ها در جدول(۳-۵) نشان داده شده است.

جدول(۳-۵) توصیف کننده‌های انتخاب شده از داکینگ مولکولی با روش LASSO

No	Symbol	Meaning
۱	TI	Final Total Internal Energy
۲	pK_i	-Log(Estimated Inhibition Constant, Ki)
۳	EE	Electrostatic Energy
۴	VE	(vdW + Hbond + desolv) Energy

پارامتر های موثر بر محاسبه توصیف‌کننده VE در جدول(۳-۵) مربوط به انرژی واندروالس(vdW)، انرژی پیوند هیدروژنی(Hbond) و انرژی حلال‌زدایی (desolv) می‌باشد.

۳-۱-۳-۴ بررسی توصیف‌کننده‌های داکینگ مولکولی انتخاب شده با استفاده از روش‌های رگرسیون گام به گام و لاسو

تعدادی از توصیف‌کننده‌ها دارای اثر یکسان بر روی فعالیت مهارکنندگی هستند، لذا پس از بررسی همبستگی بین متغیرها، از میان توصیف‌کننده‌هایی با ضریب همبستگی بزرگتر از ۰/۹، آنکه همبستگی کمتری با متغیر وابسته داشته باشد، حذف می شود. ضریب همبستگی بین دو متغیر با استفاده از رابطه(۲-۲۹) محاسبه شد. پدیده‌ی همبستگی بین توصیف‌کننده‌های منتخب با نرم افزار SPSS مورد بررسی قرار گرفت و ماتریس همبستگی آن‌ها به دست آمد. جدول (۳-۶) و جدول (۳-۷) ماتریس همبستگی بین توصیف‌کننده‌ها را نشان می‌دهد، نتایج حاصله عدم وجود همبستگی معنادار بین این توصیف‌کننده‌ها را تایید می‌نماید.

جدول(۳-۶) ماتریس همبستگی کل توصیف کننده‌های انتخاب شده از داکینگ مولکولی با روش SR

	TI	pK_i
TI	۱	
pK_i	-۰/۳۵۴	۱

جدول(۳-۷) ماتریس همبستگی کل توصیف کننده‌های انتخاب شده از داکینگ مولکولی با روش LASSO

	TI	pK_i	EE	VE
TI	۱			
pK_i	-۰/۳۵۴	۱		
EE	۰/۷۷۹	-۰/۰۱۷	۱	
VE	-۰/۲۰۱	-۰/۶۹۹	-۰/۵۶۲	۱

وجود هم‌خطی در بین متغیرهای پیش‌بینی کننده مسئله مهمی است که باید بررسی شود. مقادیر VIF بیشتر از ۱۰ اغلب به عنوان علامتی از این‌که داده‌ها مشکل هم‌خطی دارند، تلقی می شود. VIF طبق رابطه(۲-۳۷) محاسبه شد. به منظور بررسی وجود پدیده‌ی هم‌خطی در بین توصیف‌کننده‌های انتخاب شده، عامل افزایش واریانس مربوط به هر توصیف‌کننده محاسبه شد که نتایج آن در جدول(۳-۸) و جدول(۳-۹) آورده شده است و با توجه به این‌که برای کلیه متغیرهای مستقل، عامل افزایش واریانس از ۱۰کمتر است، عدم وجود هم‌خطی در بین توصیف‌کننده‌ها اثبات می‌گردد.

جدول(۳-۸) مقادیر عامل افزایش واریانس کل توصیف کننده‌های انتخاب شده از داکینگ مولکولی با روشSR

توصیف‌کننده	TI	pK_i
VIF	۱/۱۴۴	۱/۱۴۴

جدول(۳-۹) مقادیر عامل افزایش واریانس کل توصیف کننده‌های انتخاب شده از داکینگ مولکولی با روشLASSO

توصیف‌کننده	TI	pK_i	EE	VE
VIF	۵/۵۱۰	۴/۸۸۴	۴/۲۳۹	۳/۶۰۹

۳-۱-۴ مدل سازی به روش های غیر خطی و خطی با استفاده از توصیف‌کننده‌های داکینگ مولکولی

در این بخش از روش‌های مدل‌سازی شبکه عصبی مصنوعی و جنگل‌های تصادفی با استفاده از توصیف‌کننده‌های داکینگ مولکولی انتخاب شده با هر دو روش انتخاب متغیر برای ایجاد مدل‌های

QSAR به کار گرفته شـــدند که در ادامه به چگونگی مدل‌ســـازی با هر کدام از روش‌های مدل‌ســـازی پرداخته خواهد شد.

۳-۱-۴-۱ مدل‌سازی با استفاده از روش شبکه عصبی مصنوعی (ANN)

یکی از راه‌های یافتن رابطه غیرخطی بین متغیرهای مستقل و متغیر وابسته، استفاده از شبکه عصبی مصنوعی برای مدل‌سازی می‌باشد. شبکه عصبی انتخاب شده در این پروژه یک شبکه پیشرو با الگوریتم آموزشی پس انتشار می‌باشد که الگوریتم آن با برنامه متلب نوشته شده است. در فرآیند بهینه سازی پارامترهای شبکه، سری داده‌ها به سه بخش سری آزمون (۱۱ ترکیب) و سری ارزیابی (۱۱ ترکیب) و همچنین سری آموزش (۵۱ ترکیب) تقسیم شدند. توصیف‌کننده‌های انتخاب شده به روش‌های رگرسیون گام به گام و لاسو به‌عنوان ورودی به شبکه عصبی مصنوعی داده شدند. لازم به ذکر است آماره T طبق رابطه(۲-۲۹) برای توصیف‌کننده‌های انتخاب شده توسط روش انتخاب متغیر لاسو با استفاده از نرم افزار SPSS محاسبه شد و توصیف‌کننده‌ها بر اساس بزرگی آماره T مرتب شدند و سپس این توصیف‌کننده‌ها به عنوان ورودی در شبکه عصبی استفاده شدند. خروجی شبکه نیز فعالیت ترکیبات مورد مطالعه می‌باشد. برای یافتن شرایط بهینه، آموزش شبکه با دو الگوریتم آموزشی لونبرگ-مارکوات و تنظیم بایزین و دوتابع انتقال تانژانت سیگموئید و لگاریتم سیگموئید انجام شد. از ترکیب این دو الگوریتم و دو تابع انتقال، چهار نوع شبکه عصبی مختلف طراحی گردید. در هر یک از این شبکه‌ها تعداد ورودی، تعداد نرون‌های لایه پنهان (گره) و تعداد دور آموز شی بهینه شدند. مقدار بهینه هریک از این پارامترها زمانی اســـت که میانگین توان‌های دوم خطا[1] برای ســـری ارزیابی دارای کمترین مقدار باشد. برای بهینه سازی تعداد ورودی‌های شبکه، تعداد نرون‌های لایه‌ی پنهان و تعداد دور آموز شی از روش بهینه سازی همزمان استفاده گردید. برای این منظور هر شبکه با تعداد ورودی توصیف‌کننده‌ها از ۱ تا ۴ و تعداد گره از ۲ تا ۱۰ با گام ۱ و تعداد دور آموز شی از ۲ تا ۵۰ با گام ۲ به

[1] MSE

طور همزمان آموزش داده شد. در روند بهینه‌سازی فوق، به حداقل رساندن میانگین توان‌های دوم خطا برای سری ارزیابی به‌عنوان معیار انتخاب شد، این کار موجب می شود بتوان نقش تمام پارامترها را به طور همزمان و یک‌جا بررسی کرد. با توجه به نتایج حاصل برای تعداد مختلف توصیف‌کننده‌ها و گره‌ها و دورهای آموزش و هم‌چنین توابع متفاوت آموزش و انتقال، بهترین شبکه‌های به‌دست آمده براساس کمترین مقدار MSE با استفاده از توصیف‌کننده‌های انتخاب شده به روش‌های رگرسیون گام به گام و لاسو در جدول(۳-۱۰) و جدول(۳-۱۱ خلاصه شده است.

جدول(۳-۱۰) توابع و پارامترهای شبکه‌های بهینه توصیف کننده‌های انتخاب شده از داکینگ مولکولی با روش SR

MSE	تعداد دور آموزش	تعداد گره	تابع انتقال	تابع آموزش	تعداد توصیف کننده
۰/۱۶۸۰	۸	۴	لگاریتم-سیگموئید	تنظیم بایزین	۲
۰/۱۵۰۸	۱۸	۳	لگاریتم-سیگموئید	لونبرگ- مارکوات	۲
۰/۱۷۰۳	۴۸	۳	تانژانت-سیگموئید	تنظیم بایزین	۲
۰/۱۴۵۱	**۱۲**	**۴**	**تانژانت-سیگموئید**	**لونبرگ- مارکوات**	**۲**

جدول(۳-۱۱)توابع و پارامترهای شبکه‌های بهینه توصیف کننده‌های انتخاب شده از داکینگ مولکولی با روش LASSO

MSE	تعداد دور آموزش	تعداد گره	تابع انتقال	تابع آموزش	تعداد توصیف کننده
۰/۱۵۷۸	۲	۱۰	لگاریتم-سیگموئید	تنظیم بایزین	۳
۰/۱۲۷۵	۲۸	۳	لگاریتم-سیگموئید	لونبرگ- مارکوات	۴
۰/۱۶۱۰	۱۴	۵	تانژانت-سیگموئید	تنظیم بایزین	۳
۰/۱۰۹۳	**۸**	**۷**	**تانژانت-سیگموئید**	**لونبرگ- مارکوات**	**۳**

با توجه به نتایج به دست آمده در جدول(۳-۱۰) و جدول(۳-۱۱)، الگوریتم آموزشی لونبرگ مارکوات نسبت به الگوریتم آموزشی تنظیم بایزین دارای میانگین توان‌های دوم خطا (MSE) کمتری است. همچنین تابع تانژانت سیگموئیدی دارای خطای کمتری نسبت به لگاریتم سیگموئیدی است. بنابراین تابع انتقال تانژانت سیگموئیدی به عنوان تابع انتقال لایه پنهان استفاده شد. طبق نتایج به دست آمده ۲ توصیف‌کننده انتخاب شده با استفاده از روش رگرسیون گام به گام و ۳ توصیف‌کننده انتخاب شده با استفاده از روش رگرسیون لاسو به عنوان توصیف‌کننده‌هایی که بیشترین ارتباط را با فعالیت بازدارندگی دارند، توسط مدل سازی شبکه عصبی انتخاب شدند. با توجه به نتایج ذکر شده

مدل شبکه عصبی برای توصیف‌کننده‌های انتخاب شده توسط روش لاسو MSE کمتری را نسبت به مدل با توصیف‌کننده‌های منتخب روش رگرسیون گام به گام دارد و این موضوع توانمندی روش انتخاب متغیر لاسو را در انتخاب موثرترین توصیف‌کننده‌ها به خوبی نشان می‌دهد.

۳-۱-۴-۲ مدل‌سازی با استفاده از روش جنگل‌های تصادفی (RF)

ابتدا سری داده‌ها به طور تصادفی به دو مجموعه آموزش (۶۲ ترکیب) و سری آزمون (۱۱ ترکیب) تقسیم شدند. برای ایجاد مدل جنگل‌های تصادفی از داده‌های سری آموزش استفاده گردید. همانطور که در بخش (۲-۶-۳-۱) گفته شد، $\frac{1}{3}$ داده‌ها در ساخت درخت شرکت ندارند، این داده‌ها OOB[1] هستند که برای بهینه‌سازی پارامترهای مؤثر در نظرگرفته شدند. برای دستیابی به بهترین مدل ابتدا سه پارامتر مهم و تأثیرگذار یعنی تعداد درختان[2]، تعداد توصیف کننده‌های انتخاب شده در هر مرحله افراز[3] و مقدار مشاهدات باقی‌مانده در هر گره[4] بهینه شدند. برای بهینه نمودن، تعداد درخت از ۱۰ تا ۵۰۰ با گام ۱۰، تعداد توصیف‌کننده‌های انتخاب شده در هر مرحله افراز با توجه به تعداد توصیف‌کننده‌های منتخب برای هر دو روش رگرسیون گام به گام و لاسو به‌ترتیب از ۱ تا ۲ و ۱ تا ۴ با گام ۱ و مقدار مشاهدات باقی‌مانده در هر گره از ۲ تا ۱۰ با گام ۱ تغییر داده شد و در هر مرحله مقدار خطای مربوط به مجموعه OOB محا سبه گردید. مقادیر MSE مربوط به داده‌های OOB در شرایط مختلف برای مدل با متغیرهای منتخب رگر سیون گام به گام در جدول(۳-۱۲) و برای مدل با متغیرهای منتخب لاسو در جدول(۳-۱۳) برای چندین حالت که کمترین MSE را دارد، آورده شده است.

[1] Out of bag
[2] n_{tree}
[3] Mtry
[4] Node size

جدول(۳-۱۲) کمترین مقادیر MSE همراه با Mtry و Ntree متناظر با آن‌ها برای مدل با متغیرهای منتخب SR

N_{tree}	Mtry	Node Size	MSE OOB
۴۹۰	**۲**	**۱**	**۰/۱۱۰۸**
۱۴۰	۲	۲	۰/۱۱۸۶
۳۶۰	۲	۳	۰/۱۳۲۲
۳۶۰	۲	۴	۰/۱۴۴۹
۳۱۰	۲	۵	۰/۱۵۵۴
۳۲۰	۲	۶	۰/۱۷۰۵
۱۳۰	۲	۷	۰/۱۷۷۲
۲۳۰	۲	۸	۰/۱۹۰۶
۳۴۰	۲	۹	۰/۲۰۰۲
۴۴۰	۲	۱۰	۰/۲۱۲۳

با توجه به جدول(۳-۱۲) تعداد درختان برابر با۴۹۰ و تعداد توصیف کننده‌های انتخاب شده در هر مرحله افراز برابر با ۲ و تعداد گره برابر با ۱ به‌عنوان مقادیر بهینه برای مدل با متغیرهای منتخب روش رگرسیون گام به گام انتخاب شدند.

جدول(۳-۱۳)کمترین مقادیر MSE همراه با Mtry و Ntree متناظر برای مدل با متغیرهای منتخب روش LASSO

N_{tree}	M_{try}	Node Size	MSE OOB
۴۶۰	**۴**	**۱**	**۰/۰۹۲۸**
۱۰۰	۴	۲	۰/۰۹۷۴
۴۱۰	۴	۳	۰/۱۰۵۶
۱۵۰	۳	۴	۰/۱۰۷۱
۲۵۰	۳	۵	۰/۱۱۴۱
۴۳۰	۳	۶	۰/۱۳۱۶
۲۴۰	۴	۷	۰/۱۳۴۹
۱۶۰	۴	۸	۰/۱۴۲۹
۱۰۰	۳	۹	۰/۱۵۶۴
۲۰۰	۳	۱۰	۰/۱۷۱۹

با توجه به جدول(۳-۱۳) تعداد درختان برابر با۴۶۰ و تعداد توصیف کننده‌های انتخاب شده در هر مرحله افراز برابر با ۴ و تعداد گره برابر با ۱ به‌عنوان مقادیر بهینه برای مدل با متغیرهای منتخب روش لاسو انتخاب شدند.

۳-۱-۴-۳ ارزیابی مدل

در بخش‌های پیش رو، با توجه به دو روش مدل سازی غیر خطی و خطی مورد ا ستفاده در این پروژه ارزیابی داده‌ها با اســتفاده از داده‌های ســری آزمون و به وســیله رســم نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی و نمودار خطای باقی‌مانده برحسب مقادیر تجربی مورد بررسی قرار می‌گیرد.

- ارزیابی مدل شبکه عصبی

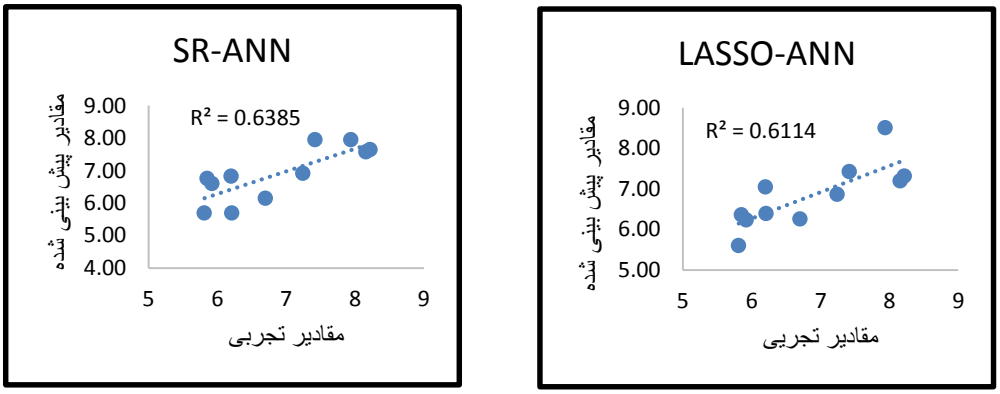
- ارزیابی مدل با استفاده از داده‌های سری آزمون

اعتبار و اهمیت مدل وقتی مشــخص می‌گردد که خواص مولکول‌های مجهولی که در دســته‌ی آموزش موجود نیستند را پیش‌بینی کند. بدین منظور مدل‌های منتخب برای هر دو نوع توصیف‌کننده انتخاب شده توسط روش رگرسیون گام به گام و لاسو جهت پیش‌بینی فعالیت بازدارندگی ۱۱ ترکیب که در فرایند مدل سازی ا ستفاده نشده‌اند، به کارگرفته شد و با ا ستفاده از شبکه عصبی بهینه شده، مقادیر pEC_{50} این ترکیبات پیش‌بینی شد. نتایج مربوط به پیش‌بینی فعالیت بازدارندگی این ترکیبات در جدول(۳-۱۴) نشان داده شده است و خطای کم پیش بینی فعالیت داده‌های سری آزمون در مدل با توصــیف‌کننده‌های منتخب لاســو، قدرت پیش‌بینی مدل را تایید می‌نماید. شــکل(۳-۶) نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی را برای داده‌های سری آزمون ن شان می‌دهد. ضــریب تعیین بیانگر این اســت که روش به چه میزان در پیش‌بینی فعالیت دارویی ترکیبات مورد برر سی، توانمند ا ست. نمودار باقی‌مانده‌ها معیاری برای شایستگی مدل به د ست آمده، می‌با شد و اگر باقی‌مانده‌ها به طور یکنواخت حول محور افقی پراکنده باشند، نشان می‌دهد که مدل مناسبی به دست آمده و هیچ خطای سی ستماتیکی وجود ندارد. شکل(۳-۷) نمودار باقی‌مانده‌ها را که از تفا ضل مقدار تجربی فعالیت دارویی با مقدار محا سبه شده تو سط مدل غیرخطی به د ست می‌آید، بر ح سب مقدار تجربی نشان می‌دهد. تقارن پراکندگی نقاط در دو طرف محور افقی عدم وجود خطای سیستماتیک را

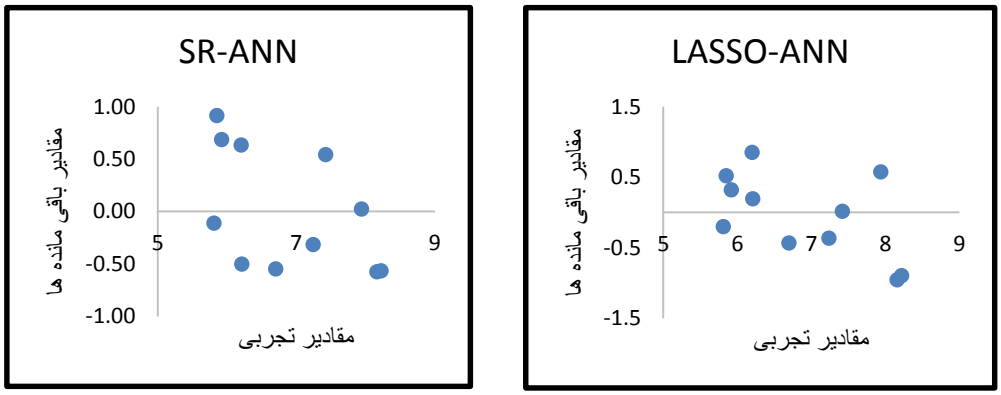
نشان می‌دهد.

جدول(۳-۱۴) نتایج حاصل از ارزیابی مدل شبکه عصبی مصنوعی با استفاده از داده‌های سری آزمون

شماره ترکیب	مقدار تجربی	مقدار پیش‌بینی		درصد خطا	
		SR-ANN	LASSO -ANN	SR-ANN	LASSO –ANN
۲	۷/۴۲	۷/۹۶	۷/۴۳	۷/۳۰	۰/۱۷
۷	۷/۲۴	۶/۹۲	۶/۸۷	-۴/۳۸	-۵/۰۶
۱۱	۷/۹۴	۷/۹۶	۸/۵۱	۰/۲۹	۷/۲۳
۲۱	۸/۱۶	۷/۵۸	۷/۲۰	-۷/۰۷	-۱۱/۷۵
۲۲	۸/۲۲	۷/۶۵	۷/۳۲	-۶/۸۹	-۱۰/۹۰
۳۴	۶/۲۱	۵/۷۱	۶/۴۰	-۸/۱۱	۳/۰۶
۴۵	۶/۷۰	۶/۱۵	۶/۲۶	-۸/۱۹	-۶/۵۱
۵۵	۵/۸۱	۵/۷۰	۵/۶۱	-۱/۹۲	-۳/۴۹
۵۹	۶/۲۰	۶/۸۳	۷/۰۵	۱۰/۱۹	۱۳/۷۱
۶۱	۵/۹۲	۶/۶۱	۶/۲۴	۱۱/۵۹	۵/۳۸
۷۳	۵/۸۵	۶/۷۶	۶/۳۷	۱۵/۶۱	۸/۸۹



شکل(۳-۶) نمودار تغییرات مقادیر پیش‌بینی شده‌در مقابل مقادیر تجربی برای داده‌های سری آزمون



شکل(۳-۷) نمودار باقی‌مانده برحسب مقادیر تجربی برای داده‌های سری آزمون

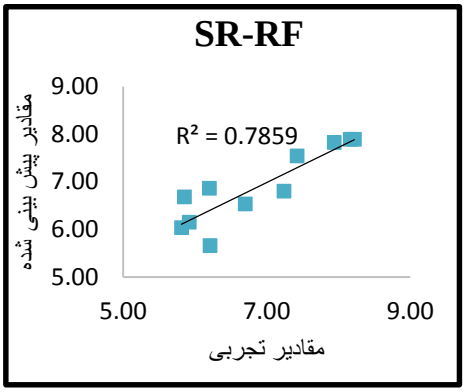
– ارزیابی مدل جنگل‌های تصادفی

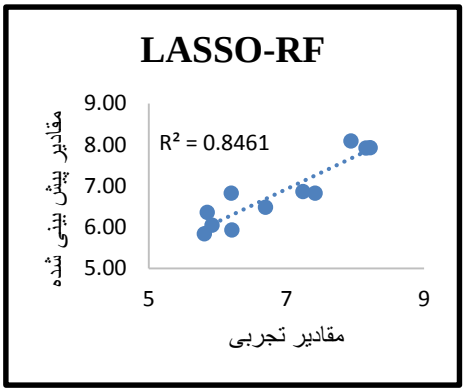
– ارزیابی مدل با استفاده از داده‌های سری آزمون

اهمیت مدل‌های بهینه منتخب با پیش‌بینی فعالیت داده‌های سری آزمون مشخص می‌گردد. به این منظور مدل‌های منتخب با شرایط بهینه برای پیش‌بینی فعالیت داده‌های سری آزمون، به‌کار برده شدند. جدول(۳-۱۵) نتایج پیش‌بینی حاصل از ارزیابی مدل جنگل‌های تصادفی با استفاده از داده‌های سری آزمون را نشان می‌دهد. شکل(۳-۸) نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی را برای داده‌های سری آزمون نشان می‌دهد. ضریب تعیین مدل جنگل‌های تصادفی نشان دهنده توانمندی روش در پیش‌بینی فعالیت دارویی ترکیبات مورد برر سی ا ست. توزیع متقارن داده‌ها حول محور افقی (خطای صفر) حاکی از عدم وجود خطای سیستماتیک است. نمودار خطای باقی‌مانده برحسب مقادیر تجربی، برای مدل‌های ذکر شده در شکل(۳-۹) نشان داده شده است.

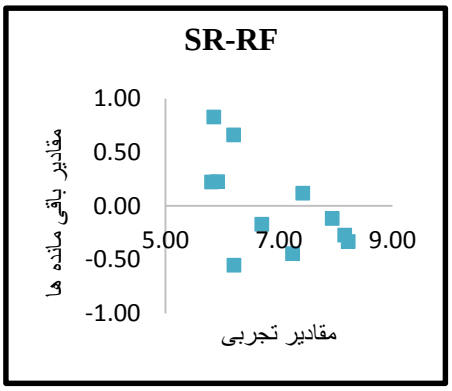
جدول(۳-۱۵) نتایج حاصل از ارزیابی مدل‌های جنگل تصادفی با استفاده از داده‌های سری آزمون

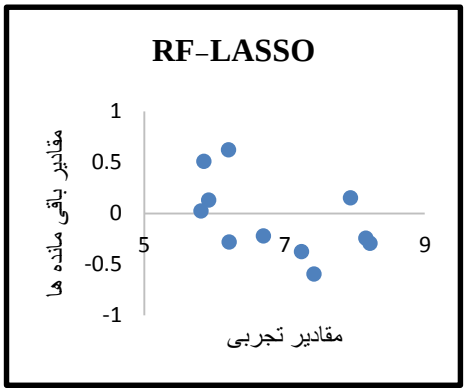
شماره ترکیب	مقدار تجربی	مقدار پیش‌بینی		درصد خطا	
		SR-RF	LASSO - RF	SR-RF	LASSO – RF
۲	۷/۴۲	۷/۵۴	۶/۸۳	۱/۶۱	-۸/۰۱
۷	۷/۲۴	۶/۸۰	۶/۸۷	-۶/۱۱	-۵/۱۶
۱۱	۷/۹۴	۷/۸۲	۸/۰۹	-۱/۴۵	۱/۹۱
۲۱	۸/۱۶	۷/۸۹	۷/۹۲	-۳/۳۱	-۲/۹۶
۲۲	۸/۲۲	۷/۸۹	۷/۹۳	-۴/۰۱	-۳/۵۶
۳۴	۶/۲۱	۵/۶۶	۵/۹۳	-۸/۸۹	-۴/۵۰
۴۵	۶/۷۰	۶/۵۳	۶/۴۸	-۲/۵۳	-۳/۲۹
۵۵	۵/۸۱	۶/۰۳	۵/۸۳	۳/۸۶	۰/۴۳
۵۹	۶/۲۰	۶/۸۶	۶/۸۲	۱۰/۶۹	۱۰/۰۳
۶۱	۵/۹۲	۶/۱۵	۶/۰۵	۳/۸۲	۲/۱۹
۷۳	۵/۸۵	۶/۶۸	۶/۳۶	۱۴/۱۷	۸/۷۲





شکل(۳-۸) نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی برای داده‌های سری آزمون





شکل(۳-۹) نمودار باقی‌مانده‌ها برحسب مقادیر تجربی برای داده‌های سری آزمون

۳-۱-۵ نتیجه گیری مدل‌سازی با توصیف‌کننده‌های داکینگ مولکولی

با توجه به مدل‌سازی انجام شده مشاهده می‌شود که مدل خطی جنگل تصادفی با توصیف‌کننده‌های داکینگ منتخب به وسیله روش انتخاب متغیر لاسو برتری مناسبی (R^2=۰/۸۴) را نسبت به مدل غیر خطی شبکه عصبی (R^2=۰/۶۱)نشان داده است. مشاهده می‌شود که مدل حاصل، قابل قبول می‌باشد اما مورد دلخواه نبوده و از این‌رو مجموعه‌ای از توصیف‌کننده‌های ساختاری از سری داده‌های لیگاندها استخراج گردید و کلیه روش‌های انتخاب متغیر و مدل سازی‌های خطی و غیر خطی روی مجموعه‌ی توصیف‌کننده‌های داکینگ مولکولی و ساختاری تکرار شد. در بخش‌های بعدی شاهد بهبود مدل خواهید بود.

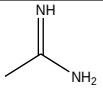
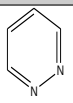
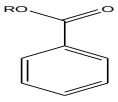
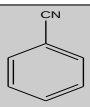
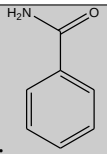
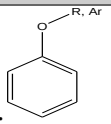
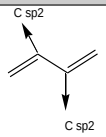
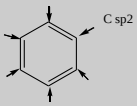
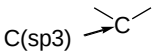
۳-۱-۶ مدل‌سازی با استفاده همزمان از مجموعه توصیف‌کننده‌های داکینگ مولکولی و ساختاری

مجموعه توصیف‌کننده های استخراج شده توسط داکینگ مولکولی به مجموعه توصیف‌کننده‌های ساختاری استخراج شده از لیگاندهای سری داده‌ها افزوده شد. روش‌های انتخاب متغیر رگرسیون گام به گام و لاسو روی مجموعه توصیف‌کننده‌ها اعمال شد و مناسب‌ترین توصیف‌کننده‌ها انتخاب و به عنوان ورودی به مدل‌های شبکه عصبی و جنگل‌های تصادفی داده شدند. در ادامه به چگونگی مدل‌سازی با هر کدام از روش‌های مدل‌سازی پرداخته خواهد شد.

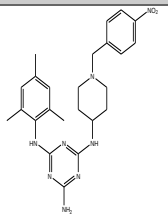
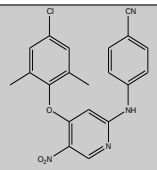
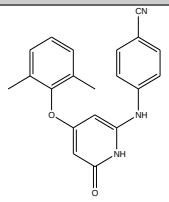
۳-۱-۶-۱ توصیف‌کننده‌های ساختاری

توصیف‌کننده‌هایی که در عین سادگی و قابل تفسیر بودن، بتوانند کل ویژگی‌های ترکیبات مورد بررسی را نشان دهند قابل توجه هستند از این رو با توجه به لیگاندهای سری داده‌ها توصیف‌کننده‌های ساده‌ای چون حضور حلقه تری آزین، حضور حلقه پیریمیدین، تعداد گروه‌های نیترو، تعداد پیوندهای هیدروژنی و ... در جدول(۳-۱۶) تعریف گردید. جهت محاسبه مقادیر این توصیف‌کننده‌ها برخی بر حسب حضور گروه مربوطه همانند حضور حلقه تری آزین مقدار یک و بر حسب عدم حضور آن در لیگاند عدد صفر را خواهند داشت و برخی نیز بر اساس تعداد حضور گروه مربوطه همچون تعداد نیتروژن‌ها و یا تعداد پیوندهای هیدروژنی با شمارش تعداد حضورشان محاسبه می‌شوند. جدول(۳-۱۷) مقادیر چهارده توصیف‌کننده قطعه مولکولی محاسبه شده برای چند ترکیب را نشان می‌دهد.

جدول(۳-۱۶) توصیف‌کننده‌های قطعه مولکولی مربوط به کل لیگاندها

No.	Symbol	Meaning
۱	A_1	No. 
۲	A_2	No. 
۳	A_3	No. of Nitro group
۴	A_4	No. 
۵	A_5	No. 
۶	A_6	No. 
۷	A_7	No. 
۸	A_8	No. 
۹	A_9	No. 
۱۰	A_{10}	No. 
۱۱	A_{11}	No. Of H donor atoms
۱۲	A_{12}	No.non-AromaticConjugated C (sp^2) 
۱۳	A_{13}	 No. C sp^2 (Aromatic)
۱۴	A_{14}	No.of secondary C (sp^3) 

جدول(۳-۱۷) مقادیر توصیف‌کننده‌های قطعه مولکولی برای چند ترکیب

		توصیف‌کننده													
No.	ترکیب	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}
۱۶		۱	۰	۱	۰	۱	۰	۰	۰	۰	۰	۴	۰	۱۵	۳
۳۲		۰	۰	۱	۰	۰	۰	۱	۰	۰	۱	۱	۰	۱۷	۰
۷۲		۰	۰	۰	۰	۰	۰	۲	۰	۰	۱	۲	۵	۱۲	۰

۳-۱-۷ انتخاب توصیف‌کننده‌های مناسب از مجموع توصیف‌کننده‌های داکینگ مولکولی و ساختاری

انتخاب توصیف‌کننده‌های مناسب، مهم‌ترین مرحله در مطالعات QSAR است تا ارتباط بین ساختار مولکول و فعالیت دارویی تفسیر پذیر باشد. پس از محاسبه توصیف‌کننده‌های ساختاری که در بخش قبل ذکر شد، مجموعه توصیف‌کننده‌های ساختاری و داکینگ مولکولی در یک جا ذخیره گردید و پس از ذخیره سازی داده‌های مربوط، مجموعه داده‌های ارزیابی و آزمون از مجموع داده‌ها حذف گردید و انتخاب بهترین توصیف‌کننده‌ها با استفاده از داده‌های مربوط به مجموعه آموزش با استفاده از هر دو روش انتخاب متغیر رگرسیون گام به گام و لاسو انجام شد. ۱۴ توصیف‌کننده مربوط به توصیف‌کننده‌های ساختاری به ۹ توصیف‌کننده مربوط به توصیف‌کننده‌های داکینگ مولکولی اضافه شد و مجموعا ۲۳ توصیف‌کننده حاصل گردید. داده‌های مربوط به مجموعه ارزیابی و مجموعه آزمون

از مجموع داده‌ها حذف گردید و مدل‌ســـازی با داده‌های مربوط به مجموعه آموزش انجام شـــد. با به‌کارگیری نرم افزار SPSS و با اجرای رگرسیون گام به گام برای ۲۳ توصیف‌کننده به‌عنوان متغیرهای م ستقل، و منفی لگاریتم فعالیت دارویی به‌عنوان متغیر واب سته، تعداد ۳ تو صیف‌کننده انتخاب شدند که نام و نوع آن‌ها درجدول(۳-۱۸) نشان داده شده است.

جدول(۳-۱۸)توصیف کننده‌های انتخاب شده از مجموع متغیرهای ساختاری و داکینگ مولکولی با روش SR

No	Symbol	Class	Meaning
۱	A_5	Structure Derived	Number of 1-3-5 Triazines
۲	BE	Docking Derived	Binding Energy Of ligands and protein
۳	A_{12}	Structure Derived	Number of non-conjugated aromatic C (sp^2)

بار دیگر برای انتخاب متغیرهای حاصـل از مجموع متغیرهای سـاختاری و داکینگ مولکولی، با به‌کارگیری نرم افزار R و با اجرای روش لاســـو روی داده‌های مجموعه آموزش رابطه (۱۶-۲) مینیمم شـد و برای ۹ توصـیف‌کننده به‌عنوان متغیرهای مسـتقل، و منفی لگاریتم فعالیت دارویی (pEC_{50}) به عنوان متغیر وابسته، تعداد ۴ توصیف‌کننده انتخاب شدند که نام و نوع آن‌ها در جدول(۳-۱۹)نشان داده شده است.

جدول(۳-۱۹) توصیف کننده‌های انتخاب شده با روش LASSO

No	Symbol	Class	Meaning
۱	A_5	Structure Derived	Number of 1-3-5 Triazines
۲	pK_i	Docking Derived	(-Log) of Estimated Inhibition Constant, K_i
۳	A_{12}	Structure Derived	Number of non-conjugated aromatic C (sp^2)
۴	A3	Structure Derived	Number of NO_2 group
۵	A_1	Structure Derived	Number of primary amines (aromatic)
۶	VE	Docking Derived	vdW + Hbond + desolv Energy
۷	A_2	Structure Derived	Number of amidines derivatives
۸	A_{14}	Structure Derived	Number of total secondary C (sp^3)
۹	A_8	Structure Derived	Number of Pyridazines
۱۰	TI	Docking Derived	Final Total Internal Energy

پارامتر های موثر بر محاسبه توصیف‌کننده VE در جدول(۳-۱۹) مربوط به انرژی واندروالس(vdW)، انرژی پیوند هیدروژنی(Hbond) و انرژی حلال‌زدایی (desolv) می‌باشد.

۳-۱-۸ بررسی مجموع توصیف‌کننده‌های داکینگ مولکولی و ساختاری انتخاب شده با استفاده از روش‌های رگرسیون گام به گام و لاسو

همانطور که گفته شد تعدادی از توصیف‌کننده‌ها حاوی نتیجه اثر یکسان بر روی فعالیت مهارکنندگی هستند، لذا پس از بررسی همبستگی بین متغیرها، از میان توصیف‌کننده‌هایی با ضریب همبستگی بزرگتر از ۰/۹، آنکه همبستگی کمتری با متغیر وابسته داشته باشد، حذف می‌شود. ضریب همبستگی بین دو متغیر با استفاده از رابطه(۲-۲۱) محاسبه شد. پدیده ی همبستگی بین توصیف‌کننده‌های منتخب با نرم افزار SPSS مورد بررسی قرار گرفت و ماتریس همبستگی آن‌ها به دست آمد. جدول(۳-۲۰) و جدول(۳-۲۱) ماتریس همبستگی بین توصیف‌کننده‌ها را برای روش انتخاب متغیر رگرسیون گام به گام و لاسو نشان می‌دهد، نتایج حاصله عدم وجود همبستگی معنادار بین این توصیف‌کننده‌ها را تایید می‌نماید.

جدول(۳-۲۰) ماتریس همبستگی کل توصیف کننده‌های انتخاب شده توسط SR

	A_5	BE	A12
A_5	۱		
BE	۰/۲۸۷	۱	
A_{12}	-۰/۲۵۱	-۰/۱۱۲	۱

جدول(۳-۲۱) ماتریس همبستگی کل توصیف کننده‌های انتخاب شده توسط LASSO

	A_5	pK_i	A_{12}	A_3	A_1	VE	A_2	A_{14}	A_8	TI
A_5	۱									
pK_i	-۰/۲۸۸	۱								
A_{12}	-۰/۲۵۱	۰/۱۹۹	۱							
A_3	-۰/۲۶۱	-۰/۴۰۰	-۰/۲۱۸	۱						
A_1	۰/۶۱۳	-۰/۲۲۰	-۰/۱۷۰	-۰/۱۱۲	۱					
VE	-۰/۲۴۵	-۰/۶۹۹	۰/۱۳۳	۰/۴۸۴	-۰/۰۹۱	۱				
A_2	-۰/۲۶۶	-./۰۱۲	۰/۷۲۸	-۰/۱۹۲	-۰/۱۶۳	۰/۲۳۱	۱			
A_{14}	۰/۷۴۷	-۰/۰۱۴	-۰/۳۸۳	-۰/۳۷۷	۰/۴۵۸	-۰/۶۵۶	-۰/۳۵۷	۱		
A8	-۰/۲۲۸	۰/۲۵۵	-۰/۲۴۷	-۰/۲۲۵	-۰/۱۴۰	-۰/۰۵۳	-۰/۱۸۵	-۰/۳۰۵	۱	
TI	۰/۸۳۳	۰/۳۵۴	-۰/۲۲۷	-۰/۲۰۳	۰/۳۸۴	-۰/۲۰۱	-۰/۲۸۳	۰/۷۲۹	-۰/۳۰۲	۱

به منظور بررسی وجود پدیده‌ی هم‌خطی در بین توصیف‌کننده‌های انتخاب شده، عامل افزایش واریانس مربوط به هر توصیف‌کننده، محاسبه شد که نتایج آن در جدول(۳-۲۲) و جدول(۳-۲۳) آورده شده است و با توجه به این‌که برای کلیه متغیرهای مستقل، عامل افزایش واریانس از ۱۰کمتر است، عدم وجود هم‌خطی در بین توصیف‌کننده‌ها اثبات می‌گردد.

جدول(۳-۲۲) مقادیر عامل افزایش واریانس مجموع توصیف کننده‌های انتخاب شده از داکینگ مولکولی و توصیف‌کننده‌های ساختاری با روش SR

توصیف‌کننده	**A_5**	**BE**	**A_{12}**
VIF	۱/۱۵۱	۱/۰۹۲	۱/۰۶۹

جدول(۳-۲۳) مقادیر عامل افزایش واریانس مجموع توصیف کننده‌های انتخاب شده از داکینگ مولکولی و توصیف‌کننده‌های ساختاری با روش LASSO

توصیف‌کننده	**A_5**	**pK_i**	**A_{12}**	**A_3**	**A_1**	**VE**	**A_2**	**A_{14}**	**A_8**	**TI**
VIF	۵/۲۲۷	۷/۰۷۹	۳/۶۳۲	۲/۴۵۳	۱/۷۹۷	۶/۵۱۴	۲/۸۰۹	۷/۳۴۴	۵/۲۷۷	۴/۷۲۵

۳-۱-۹ مدل‌ســـازی به روش‌های غیر خطی و خطی با اســـتفاده از مجموع توصیف‌کننده‌های داکینگ مولکولی و ساختاری

در این بخش مجموع توصیف‌کننده‌های داکینگ مولکولی و توصیف‌کننده‌های ساختاری انتخاب شده با روش‌های انتخاب متغیر رگرسیون گام به گام و لاسو، به عنوان ورودی مدل‌های شبکه عصبی و جنگل تصادفی در نظر گرفته شدند. در ادامه به چگونگی مدل‌سازی با هر کدام از روش‌های مدل‌سازی پرداخته خواهد شد.

۳-۱-۹-۱ مدل‌سازی با استفاده از روش شبکه عصبی مصنوعی (ANN)

در این بخش نیز شبکه عصبی انتخاب شده یک شبکه پیشرو با الگوریتم آموزشی پس انتشار می‌باشد. دراین فرایند نیز مجددا فرآیند بهینه‌سازی پارامترهای شبکه، سری داده‌ها به سه بخش سری آزمون (۱۱ ترکیب) و سری ارزیابی (۱۱ ترکیب) و هم‌چنین سری آموزش (۵۱ ترکیب) تقسیم شدند. توصیف کننده‌های انتخاب شده به روش‌های رگرسیون گام به گام و لاسو به‌عنوان ورودی به شبکه عصبی مصنوعی داده شدند. خروجی شبکه نیز فعالیت ترکیبات مورد مطالعه می‌باشد. لازم به ذکر است آماره T طبق رابطه(۲-۲۹) برای توصیف‌کننده‌های انتخاب شده توسط روش انتخاب متغیر لاسو با استفاده از نرم افزار SPSS محاسبه شد و توصیف‌کننده‌ها بر اساس بزرگی آماره T مرتب شدند و سپس این توصیف‌کننده‌ها به عنوان ورودی در شبکه عصبی استفاده شدند. برای آموزش شبکه با دو الگوریتم آموزشـــی لونبرگ-مارکوات و تنظیم بایزین و دوتابع انتقال تانژانت ســـیگموئید و لگاریتم ســـیگموئید انجام شـــد. از ترکیب این دو الگوریتم و دو تابع انتقال، چهار نوع شـــبکه عصـــبی مختلف طراحی گردید. تعداد ورودی‌ها، تعداد نرون‌های لایه پنهان (گره) و تعداد دور آموزشـــی بهینه شـــدند. مقدار بهینه هریک از این پارامترها زمانی اســـت که میانگین توان‌های دوم خطا[1] برای ســـری ارزیابی دارای کمترین مقدار باشد. برای بهینه سازی تعداد ورودی‌های شبکه، تعداد نرون‌های لایه‌ی پنهان و

[1] MSE

تعداد دور آموزشی از روش بهینه‌سازی همزمان استفاده گردید. برای این منظور هر شبکه با تعداد ورودی توصیف‌کننده‌ها و تعداد گره از ۲ تا ۱۰ با گام ۱ و تعداد دور آموزشی از ۲ تا ۵۰ با گام ۲ به طور همزمان آموزش داده شد. در روند بهینه سازی فوق، به حداقل ر ساندن میانگین خطاهای سری آموزش به‌عنوان معیار انتخاب شد. با توجه به نتایج حاصل برای تعداد مختلف توصیف‌کننده‌ها و گره‌ها و دورهای آموزش و هم‌چنین توابع متفاوت آموزش و انتقال، بهترین شبکه‌های به‌د ست آمده برا ساس کمترین مقدار MSE با استفاده از توصیف‌کننده‌های انتخاب شده به روش‌های رگرسیون گام به گام و لاسو در جدول(۳-۲۴) و جدول(۳-۲۵) خلاصه شده است.

جدول(۳-۲۴)توابع و پارامترهای شبکه‌های بهینه توصیف کننده‌های منتخب از مجموع متغیرهای ساختاری و داکینگ مولکولی با روش SR

MSE	تعداد دور آموزش	تعداد گره	تابع انتقال	تابع آموزش	تعداد توصیف کننده
۰/۱۲۲۶	۱۰	۲	لگاریتم-سیگموئید	تنظیم بایزین	۳
۰/۱۴۱۱	۴	۲	لگاریتم-سیگموئید	لونبرگ- مارکوات	۳
۰/۱۱۴۸	**۲**	**۲**	**تانژانت-سیگموئید**	**تنظیم بایزین**	**۳**
۰/۱۲۶۷	۲	۷	تانژانت-سیگموئید	لونبرگ- مارکوات	۳

جدول(۳-۲۵)توابع و پارامترهای شبکه‌های بهینه توصیف کننده‌های انتخاب شده از مجموع متغیرهای ساختاری و داکینگ مولکولی با روش LASSO

MSE	تعداد دور آموزش	تعداد گره	تابع انتقال	تابع آموزش	تعداد توصیف کننده
۰/۰۷۵۴	۲	۸	لگاریتم-سیگموئید	تنظیم بایزین	۳
۰/۰۶۴۵	**۶**	**۳**	**لگاریتم-سیگموئید**	**لونبرگ- مارکوات**	**۴**
۰/۰۸۴۲	۵۰	۹	تانژانت-سیگموئید	تنظیم بایزین	۱۰
۰/۰۵۲	۸	۴	تانژانت-سیگموئید	لونبرگ- مارکوات	۱۰

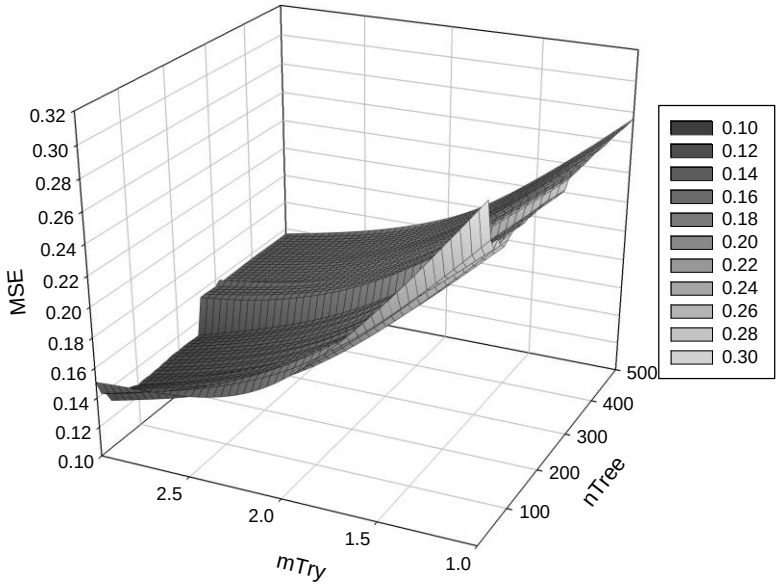
با توجه به نتایج به د ست آمده در جدول(۳-۲۴) برای مدل با تو صیف‌کننده‌های انتخاب شده توسط رگرسیون گام به گام، الگوریتم آموزشی تنظیم بایزین نسبت به الگوریتم آموزشی لونبرگ مارکوات دارای میانگین توان‌های دوم خطا (MSE) کمتری ا ست. همچنین تابع تانژانت سیگموئیدی دارای خطای کمتری نسبت به لگاریتم سیگموئیدی ا ست. بنابراین تابع انتقال تانژانت سیگموئیدی به عنوان تابع انتقال لایه پنهان ا ستفاده شد. طبق نتایج به د ست آمده ۳ تو صیف‌کننده انتخاب شده با

استفاده از روش رگرسیون گام به گام به عنوان توصیف‌کننده‌هایی که بیشترین ارتباط را با فعالیت بازدارندگی دارند، توسط مدل سازی شبکه عصبی انتخاب شدند. همچنین با توجه به نتایج به دست آمده در جدول(۳-۲۵) برای مدل با توصیف‌کننده‌های انتخاب شده توسط رگرسیون لاسو، الگوریتم آموزشی لونبرگ مارکوات نسبت به الگوریتم آموزشی تنظیم بایزین دارای میانگین توان‌های دوم خطا (MSE) کمتری است. همچنین تابع لگاریتم سیگموئیدی دارای خطای کمتری نسبت به تانژانت سیگموئیدی است. بنابراین تابع انتقال لگاریتم سیگموئیدی به عنوان تابع انتقال لایه پنهان استفاده شد. طبق نتایج به دست آمده ۴ توصیف‌کننده اول از جدول(۳-۱۹) انتخاب شده با استفاده از روش رگرسیون لاسو به عنوان توصیف‌کننده‌هایی که بیشترین ارتباط را با فعالیت بازدارندگی دارند، توسط مدل‌سازی شبکه عصبی انتخاب شدند.

۳-۱-۹-۲ مدل‌سازی با استفاده از روش جنگل‌های تصادفی(RF)

ابتدا سری داده‌ها به طور تصادفی به دو مجموعه آموزش (۶۲ ترکیب) و سری آزمون (۱۱ ترکیب) تقسیم شدند. برای ایجاد مدل جنگل‌های تصادفی از داده‌های سری آموزش استفاده گردید. همانطور که در بخش(۲-۶-۳-۱) گفته شد، $\frac{1}{3}$ داده‌ها در ساخت درخت شرکت ندارند، این داده‌ها OOB[1] هستند که برای بهینه‌سازی پارامترهای مؤثر در نظرگرفته شدند. برای دستیابی به بهترین مدل ابتدا سه پارامتر مهم و تأثیرگذار یعنی تعداد درختان[2]، تعداد توصیف کننده‌های انتخاب شده در هر مرحله افراز [3]و مقدار مشاهدات باقی‌مانده در هر گره[4] بهینه شدند. برای بهینه نمودن، تعداد درخت از ۱۰ تا ۵۰۰ با گام ۱۰، تعداد توصیف‌کننده‌های انتخاب شده در هر مرحله افراز با توجه به تعداد توصیف‌کننده‌های منتخب برای هر دو روش رگرسیون گام به گام و لاسو به‌ترتیب از ۱ تا ۳ و ۱ تا ۱۰ با گام ۱ و مقدار مشاهدات باقی‌مانده در هر گره از ۲ تا ۱۰ با گام ۱ تغییر داده شد و در هر مرحله مقدار خطای مربوط به مجموعه OOB محاسبه گردید. برای مدل با توصیف‌کننده‌های انتخاب شده به وسیله رگرسیون گام به گام، نتایج بهینه سازی پارامترهای تعداد درختان، تعداد توصیف‌کننده‌های انتخاب شده در هر مرحله افراز در گره بهینه در شکل(۳-۱۰) آمده است و هم‌چنین برای متغیر حاصل از روش رگرسیون SR جدول(۳-۲۶) چندین حالت که کمترین MSE را برای OOB دارد برای متغیرهای انتخاب شده توسط روش رگرسیون گام به گام را نشان می‌دهد. برای مدل با توصیف‌کننده‌های انتخاب شده به وسیله لاسو شکل(۳-۱۱) نتایج بهینه سازی پارامترها در گره بهینه و جدول(۳-۲۷) نیز، چندین حالت که کمترین MSE را برای OOB دارد نشان می‌دهد.

[1] Out of bag
[2] n_{tree}
[3] Mtry
[4] Node size

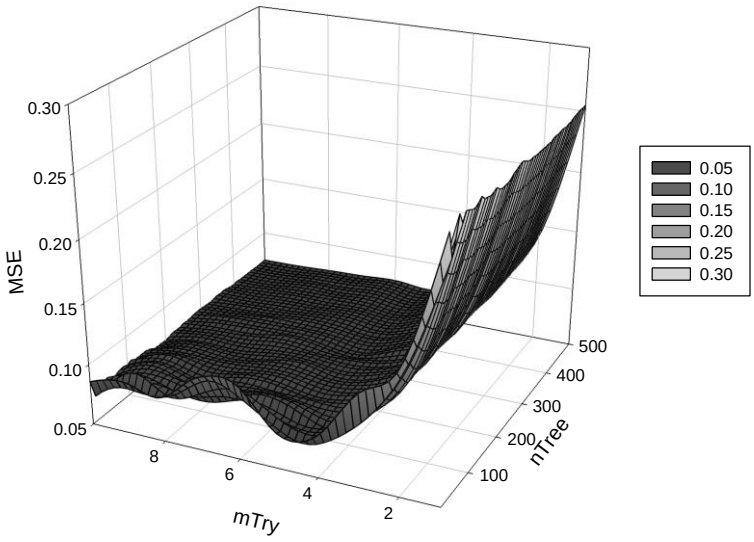


شکل(۳-۱۰)بهینه سازی پارمترهای Mtry و ntree در Node size بهینه برای متغیر حاصل از روش رگرسیون SR

جدول(۳-۲۶) کمترین مقادیرMSE همراه با Mtry و ntree متناظر آن‌ها برای متغیر حاصل از روش رگرسیون SR

N_{tree}	M_{try}	Node Size	MSE OOB
۱۰۰	**۳**	**۱**	**۰/۰۹۷۹**
۱۰۰	۳	۲	۰/۱۰۲۷
۳۱۰	۳	۳	۰/۱۲۶۰
۱۲۰	۳	۴	۰/۱۳۶۲
۳۱۰	۳	۵	۰/۱۴۸۸
۵۰۰	۳	۶	۰/۱۶۲۷
۲۴۰	۳	۷	۰/۱۷۵۱
۱۴۰	۳	۸	۰/۱۸۶۳
۱۷۰	۳	۹	۰/۱۹۱۲
۵۰۰	۳	۱۰	۰/۱۹۷۸

با توجه به جدول (۳-۲۶) برای متغیر حاصل از روش رگرسیون SR تعداد درختان برابر با ۱۰۰ و تعداد توصیف کننده‌های انتخاب شده در هر مرحله افراز برابر با ۳ و تعداد گره برابر ۱ به‌عنوان مقادیر بهینه انتخاب شدند.



شکل(۳-۱۱) بهینه سازی پارمترهای Mtryو ntreeدر Node sizeبهینه برای متغیرهای حاصل از روش LASSO

جدول(۳-۲۷) کمترین مقادیرMSE همراه با Mtryو ntreeمتناظر آن‌ها برای متغیرهای حاصل از روش LASSO

N_{tree}	M_{try}	Node Size	MSE OOB
۲۰۰	۱۰	۱	۰/۰۷۱۴
۲۱۰	**۸**	**۲**	**۰/۰۶۹۸**
۱۱۰	۹	۳	۰/۰۷۵۴
۲۶۰	۹	۴	۰/۰۹۱۸
۱۲۰	۹	۵	۰/۰۹۷۷
۳۸۰	۹	۶	۰/۱۰۹۳
۱۴۰	۸	۷	۰/۱۱۴۵
۴۸۰	۱۰	۸	۰/۱۲۳۱
۴۶۰	۱۰	۹	۰/۱۳۴۹
۱۲۰	۱۰	۱۰	۰/۱۴۰۴

با توجه به جدول(۳-۲۷) تعداد درختان برابر با ۲۱۰ و تعداد توصیف کننده‌های انتخاب شده در هر مرحله افراز برابر با ۸ و تعداد گره برابر با ۲ به‌عنوان مقادیر بهینه انتخاب شدند.

با توجه به مدل‌سازی جنگل‌های تصادفی مربوط به دو حالت مختلف، توصیف‌کننده‌های منتخب حاصل از روش‌های انتخاب متغیر رگرسیون گام به گام و لاسو، مدل جنگل‌های تصادفی مربوط به مجموع توصیف‌کننده‌های داکینگ مولکولی و توصیف‌کننده‌های ساختاری منتخب روش

لاسو، بهبود در مقدار MSE را نشان می‌دهد. مدل جنگل‌های تصادفی با شرایط بهینه ۲۱۰ =N_{tree} و۸ = M_{try} و Node Size=۲ به‌عنوان مدل برتر معرفی می‌گردد و ارزیابی‌های پیش رو نیز این برتری را به‌درستی نشان می‌دهد.

۳-۱-۹-۳ ارزیابی مدل

با توجه به دو روش مدل‌سازی مورد استفاده در بخش‌های قبلی، ارزیابی داده‌ها با استفاده از داده‌های سری آزمون و به وسیله رسم نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی و نمودار خطای باقی‌مانده برحسب مقادیر تجربی مورد بررسی قرار گرفت.

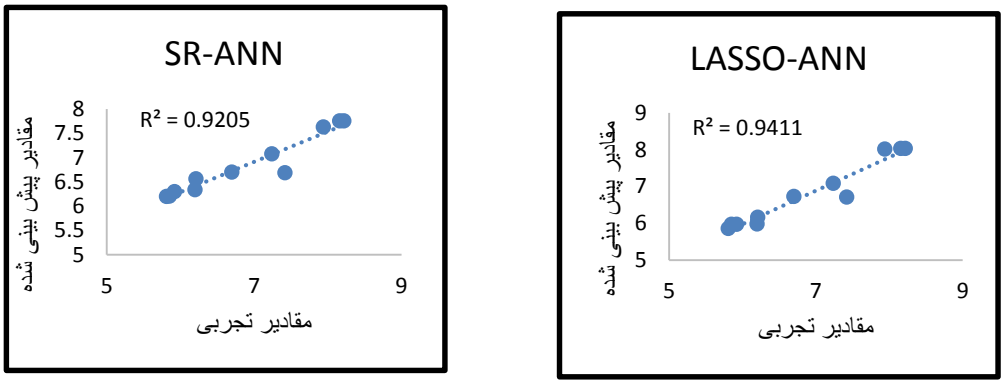
- ارزیابی شبکه عصبی مصنوعی

- ارزیابی مدل با استفاده از داده‌های سری آزمون

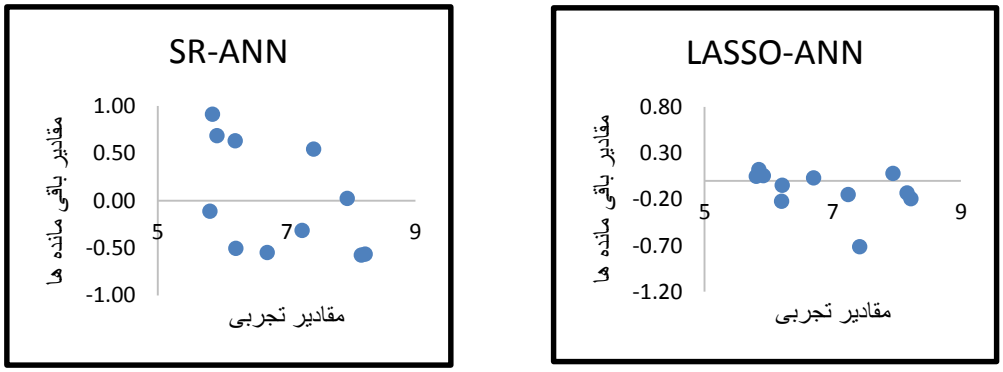
اعتبار و اهمیت مدل با پیش‌بینی فعالیت بازدارندگی داده‌هایی که در مدل‌سازی حضور نداشته‌اند مشخص می شود. بدین منظور مدل‌های انتخاب شده برای هر دو نوع توصیف‌کننده انتخاب شده توسط روش رگرسیون گام به گام و لاسو جهت پیش‌بینی فعالیت بازدارندگی ۱۱ ترکیب که در فرایند مدل‌سازی استفاده نشده‌اند، به کارگرفته شد و با استفاده از شبکه عصبی بهینه شده، مقادیر pEC_{50} این ترکیبات پیش‌بینی شد. جدول (۳-۲۸) نتایج پیش‌بینی حاصل از ارزیابی مدل‌های شبکه عصبی مصنوعی با استفاده از داده‌های سری آزمون را نشان می‌دهد. خطای کم در پیش‌بینی فعالیت دارویی، قدرت پیش‌بینی مدل را نشان می‌دهد. جدول(۳-۱۲) نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی را برای داده‌های سری آزمون نشان می‌دهد. شکل(۳-۱۳) نمودار خطای باقی‌مانده برحسب مقادیر تجربی، داده‌های سری آزمون را نشان می‌دهد. توزیع متقارن داده‌ها حول محور افقی (خطای صفر) حاکی از عدم وجود خطای سیستماتیک است.

جدول(۳-۲۸) نتایج حاصل از ارزیابی مدل‌های شبکه عصبی مصنوعی با استفاده از داده‌های سری آزمون

شماره ترکیب	مقدار تجربی	مقدار پیش‌بینی		درصد خطا	
		SR-ANN	LASSO-ANN	SR-ANN	LASSO-ANN
۲	۷/۴۲	۶/۶۹	۶/۷۱	-۹/۸۶	-۹/۶۳
۷	۷/۲۴	۷/۰۸	۷/۰۹	-۲/۱۹	-۲/۰۹
۱۱	۷/۹۴	۷/۶۳	۸/۰۲	-۳/۸۶	۱/۰۰
۲۱	۸/۱۶	۷/۷۶	۸/۰۳	-۴/۸۸	-۱/۶۳
۲۲	۸/۲۲	۷/۷۵	۸/۰۳	-۵/۶۲	-۲/۳۷
۳۴	۶/۲۱	۶/۵۶	۶/۱۶	۲/۷۱	-۰/۸۰
۴۵	۶/۷۰	۶/۷۰	۶/۷۳	۰/۰۱	۰/۴۷
۵۵	۵/۸۱	۶/۲۰	۵/۸۶	۶/۷۲	۰/۸۵
۵۹	۶/۲۰	۶/۳۴	۵/۹۸	۲/۳۰	-۳/۵۶
۶۱	۵/۹۲	۶/۳۰	۵/۹۷	۶/۴۰	۰/۹۱
۷۳	۵/۸۵	۶/۲۱	۵/۹۷	۶/۱۲	۲/۰۶



شکل(۳-۱۲) نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی برای داده‌های سری آزمون



شکل(۳-۱۳) نمودار باقی‌مانده برحسب مقادیر تجربی برای داده‌های سری آزمون

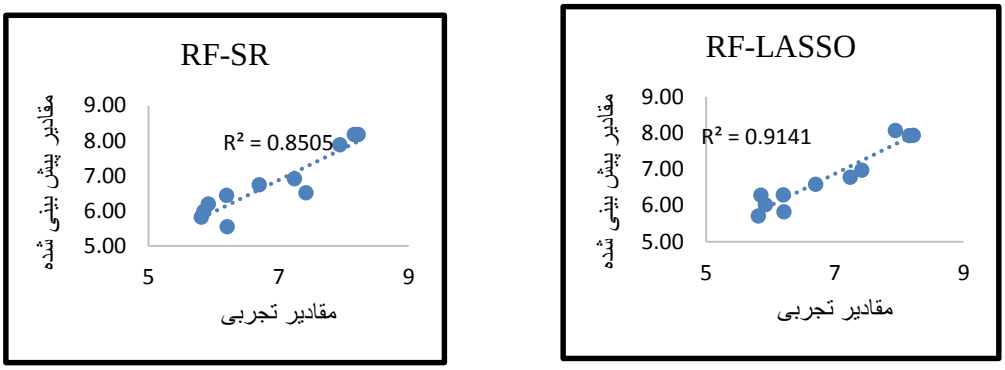
-ارزیابی مدل جنگل‌های تصادفی

– ارزیابی مدل با استفاده از داده‌های سری آزمون

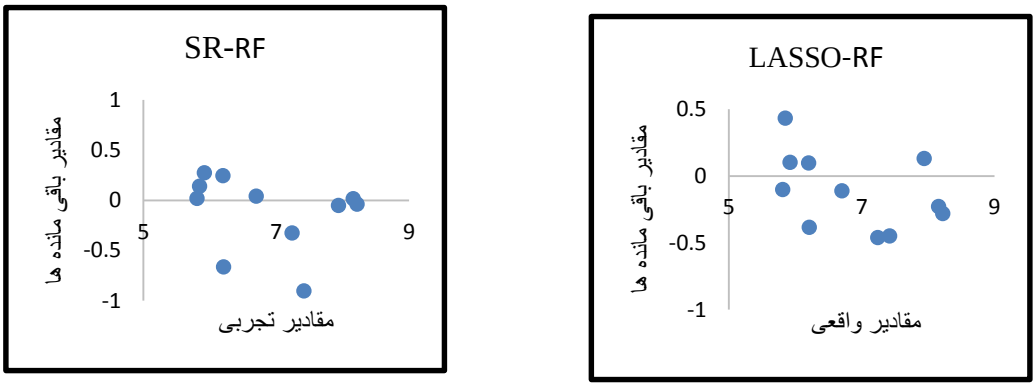
در این مرحله الگوریتم جنگل های تصـــادفی با اســـتفاده از مقادیر بهینه تعداد درخت، تعداد توصـیف‌کننده‌های انتخاب شـده در هر مرحله افراز و تعداد مشـاهدات باقی‌مانده در هر گره اجرا شـد ســـپس مقادیر(pEC_{50}) مربوط به ۱۱ ترکیب ســـری آزمون که در مدل‌ســـازی از آن‌ها اســـتفاده نشده بود، پیش‌بینی گردید. جدول(۳-۲۹) نتایج این پیش‌بینی را برای هر دو مدل جنگل تصادفی با متغیرهای منتخب رگرسیون گام به گام و لاسو نشان می‌دهد. نتایج بیانگر این است که مدل‌های ارائه شـده خطای پیش‌بینی کمی دارند. شـکل(۳-۱۴) مقادیر پیش‌بینی شـده بر حسـب مقادیر واقعی را برای داده‌های ســـری آزمون نشـــان می‌دهد. نتایج بیانگر این اســـت که مدل های ارائه شـــده خطای پیش‌بینی کمی دارند. ضریب تعیین بالای مدل با توصیف کننده‌های انتخاب شده توسط روش لاسو برتری قابل ملاحظه‌ای را در پیش بینی مقادیر فعالیت دارویی مربوط به داده‌های سری آزمون نسبت به مدل با توصیف‌کننده‌های انتخاب شده توسط روش رگرسیون گام به گام نشان می‌دهد. با توجه به شکل(۳-۱۵) توزیع نسبتاً یکنواخت و اتفاقی باقی‌مانده‌ها در دو طرف خط صاف نشان دهنده‌ی عدم وجود خطای سیستماتیک در مدل می‌باشد.

جدول(۳-۲۹) نتایج حاصل از ارزیابی مدل‌های جنگل‌های تصادفی با استفاده از داده‌های سری آزمون

شماره ترکیب	مقدار تجربی	مقدار پیش‌بینی		درصد خطا	
		SR-RF	LASSO-RF	SR-RF	LASSO–RF
۲	۷/۴۲	۶/۵۲	۶/۹۷	-۱۲/۱۹	-۶/۰۳
۷	۷/۲۴	۶/۹۱	۶/۷۸	-۴/۵۲	-۶/۳۶
۱۱	۷/۹۴	۷/۸۹	۸/۰۷	-۰/۶۵	۱/۶۵
۲۱	۸/۱۶	۸/۱۸	۷/۹۳	۰/۲۰	-۲/۸۱
۲۲	۸/۲۲	۸/۱۸	۷/۹۴	-۰/۵۰	-۳/۴۱
۳۴	۶/۲۱	۵/۵۵	۵/۸۳	-۱۰/۷۰	-۶/۱۸
۴۵	۶/۷۰	۶/۷۴	۶/۵۹	۰/۶۲	-۱/۶۷
۵۵	۵/۸۱	۵/۸۳	۵/۷۱	۰/۲۹	-۱/۷۵
۵۹	۶/۲۰	۶/۴۵	۶/۳۰	۳/۹۷	۱/۵۸
۶۱	۵/۹۲	۶/۱۹	۶/۰۲	۴/۶۴	۱/۷۰
۷۳	۵/۸۵	۵/۹۹	۶/۲۸	۲/۳۶	۷/۴۰



شکل(۳-۱۴) نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی برای داده‌های سری آزمون



شکل(۳-۱۵) نمودار باقی‌مانده‌ها برحسب مقادیر تجربی برای داده‌های سری آزمون

۳-۱-۱۰ نتیجه گیری مدل‌سازی با مجموع توصیف‌کننده های داکینگ مولکولی و ساختاری

با توجه به مدل سازی انجام شده و بررسی نتایج موجود در شکل (۳-۱۲) مشاهده می شود که بین مدل‌های غیر خطی شبکه عصبی با توجه به نوع ورودی توصیف‌کننده‌ها و روش انتخاب متغیر، مجموع توصیف‌کننده‌های داکینگ مولکولی و ساختاری انتخاب شده توسط روش لاسو برتری مناسبی (R^2=۰/۹۴) را نسبت به مدل شبکه عصبی با مجموع توصیف‌کننده‌های داکینگ مولکولی و ساختاری انتخاب شده توسط روش انتخاب متغیر رگرسیون گام به گام (R^2=۰/۹۲) نشان داده است. لازم به ذکر است با توجه به نتیجه گیری بخش (۳-۱-۵) توصیف‌کننده‌های ساختاری جهت ایجاد بهبود در نتایج مدل افزوده شدند. با مقایسه نتایج شکل (۳-۶) و شکل (۳-۱۲) بهبود در ضریب تعیین مدل ایجاد شده با مجموع توصیف‌کننده‌های داکینگ مولکولی و ساختاری انتخاب شده توسط روش لاسو (R^2=۰/۹۴) در مقایسه با مدل شبکه عصبی با توصیف‌کننده‌های داکینگ مولکولی انتخاب شده توسط روش لاسو (R^2=۰/۶۱) مشاهده می‌شود. از بین مدل‌های شبکه عصبی ذکر شده، مدل ایجاد شده با مجموع توصیف‌کننده‌های داکینگ مولکولی و ساختاری انتخاب شده توسط روش لاسو به‌عنوان مدل برتر معرفی می‌گردد.

با توجه به مدل‌های جنگل‌های تصادفی و بررسی نتایج موجود در شکل(۳-۱۴) که بین مدل‌های خطی جنگل‌های تصادفی با توجه به نوع ورودی توصیف‌کننده‌ها و روش انتخاب متغیر، مدل با مجموع توصیف‌کننده‌های داکینگ مولکولی و ساختاری انتخاب شده توسط روش لاسو برتری مناسبی (R^2=۰/۹۱) را نسبت به مدل جنگل‌های تصادفی با مجموع توصیف‌کننده‌های داکینگ مولکولی و ساختاری انتخاب شده توسط روش انتخاب متغیر رگرسیون گام به گام (R^2=۰/۸۵) نشان داده است. لازم به ذکر است با توجه به نتیجه گیری بخش (۳-۱-۵) توصیف‌کننده‌های ساختاری جهت ایجاد بهبود در نتایج مدل افزوده شدند. با مقایسه نتایج شکل(۳-۸) و شکل(۳-۱۴) بهبود در

ضریب تعیین مدل ایجاد شده با مجموع توصیف‌کننده‌های داکینگ مولکولی و ساختاری انتخاب شده تو سط روش لا سو (R^2=۰/۹۱) در مقایسه با مدل جنگل‌های تصادفی با تو صیف‌کننده‌های داکینگ مولکولی انتخاب شده تو سط روش لا سو (R^2=۰/۸۴) م شاهده می شود. از بین مدل‌های جنگل‌های تصادفی ذکر شده، مدل ایجاد شده با مجموع توصیف‌کننده‌های داکینگ مولکولی و ساختاری انتخاب شده توسط روش لاسو به‌عنوان مدل برتر معرفی می‌گردد.

از بین دو مدل برتر معرفی شـــده از بین مدل‌های شـــبکه عصـــبی (R^2=۰/۹۴) و جنگل‌های تصادفی(R^2=۰/۹۲) ، مدل شبکه عصبی ایجاد شده با مجموع توصیف‌کننده‌های داکینگ مولکولی و ساختاری انتخاب شده توسط روش لاسو به‌عنوان مدل برتر این پروژه معرفی شد.

۳-۱-۱۰-۱ ارزیابی مدل‌های برتر با استفاده از پارامترهای آماری

علاوه بر نتایجی که در بخش قبل به آن‌ها اشـــاره شـــد، هفت پارامتر آماری نیز، جهت ارزیابی توانایی پیش‌گویی مدل‌های ساخته شده به روش (RF) و (ANN) با استفاده از دو روش انتخاب متغیر رگر سیون SR و LASSO به کار گرفته شد. نتایج در جدول(۳-۳۰) آورده شده ا ست. برتری روش انتخاب متغیر لا سو درهر دو مدل م شاهده می شود و همانطور که در بخش قبل گفته شد از بین این دو مدل روش شبکه عصبی با مجموع توصیف‌کننده‌های داکینگ مولکولی-ساختاری پارامترهای آماری بهتری را داراست. از این‌رو مدل منتخب این پروژه مدل LASSO-ANN خواهد بود.

جدول(۳-۳۰) پارامترهای آماری برای مدل RF و ANN با متغیرهای حاصل از دو روش SR و LASSO

پارامترهای آماری		R^2	PRESS	MSE	MAE	MRE	SEP	REP (%)
SR-ANN	سری ارزیابی	۰/۸۱۰۱	۱/۲۶۴۲	۰/۱۱۴۹	۰/۳۰۳۶	۴/۴۳۱۲	۰/۳۳۹۰	۴/۳۶۹۴
	سری تست	۰/۸۲۳۸	۱/۵۹۴۴	۰/۱۴۴۹	۰/۳۳۴۵	۴/۸۷۵۵	۰/۳۸۰۷	۴/۸۶۳۲
LASSO-ANN	سری ارزیابی	۰/۸۹۳۶	۰/۷۰۸۲	۰/۰۶۴۴	۰/۱۸۳۶	۲/۶۵۲۳	۰/۲۵۳۷	۲/۶۴۲۶
	سری تست	**۰/۹۲۷۴**	**۰/۶۵۷۲**	**۰/۰۵۹۷**	**۰/۱۶۱۸**	**۲/۲۸۲۸**	**۰/۲۴۴۴**	**۲/۳۵۲۳**
SR-RF	سری تست	۰/۸۳۱۱	۱/۵۲۸۵	۰/۱۳۹۰	۰/۲۴۷۵	۳/۶۹۴۶	۰/۳۷۲۸	۳/۵۹۷۹
LASSO-RF	**سری تست**	**۰/۸۹۶۸**	**۰/۹۳۳۷**	**۰/۰۸۴۹**	**۰/۲۵۱۸**	**۳/۶۷۴۱**	**۰/۲۹۱۳**	**۳/۶۶۰۶**

۳-۱-۱۰-۲ ارزیابی مدل ارائه شده توسط شبکه عصبی و جنگل‌های تصادفی با استفاده از آزمون Y- تصادفی

این تکنیک ارزیابی مدل با هدف بررسی هر گونه ارتباط تصادفی بین داده‌ها انجام شد. در این آزمون مقادیر تصادفی از متغیر وابسته تولید گردید. مدل QSAR جدید با استفاده از ماتریس متغیرهای مستقل اصلی و مقادیر تصادفی از متغیر وابسته توسعه یافت. اگر مدل اصلی هیچ گونه ارتباط تصادفی نداشته باشد، تفاوت قابل توجهی بین مقدار ضریب تعیین مدل اصلی و مدل QSAR که با پاسخ‌های تصادفی توسعه یافته، وجود دارد. نتایج حاصل از چندین بار اجرای آزمون Y- تصادفی در جدول(۳-۳۱) نشان داده شده است. مقادیر کوچک ضریب تعیین (R^2) بیانگر عدم وجود ارتباط تصادفی یا وابستگی ساختاری به سری آموزش در مدل توسعه یافته توسط شبکه می‌باشد.

جدول(۳-۳۱) مقادیر R^2 برای سری آزمون با استفاده از آزمونY- تصادفی

تکرار	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
R^2 (LASSO-ANN)	۰/۰۰۲۵	۰/۰۷۶۸	۰/۰۲۶۳	۰/۰۰۱۲	۰/۰۷۶۹	۰/۱۰۰۰	۰/۰۲۶	۰/۰۰۲۶	۰/۰۵۹۸	۰/۰۴۷۵
R^2 (LASSO-RF)	۰/۰۶۶۲	۰/۱۹۳۳	۰/۰۰۱۰	۰/۱۸۰۸	۰/۰۹۴۱	۰/۲۱۶۸	۰/۰۳۱۲	۰/۰۰۱	۰/۱۷۶۰	۰/۱۳۰۷

۳-۱-۱۰-۳ بررسی ارتباط توصیف‌کننده‌های وارد شده در مدل با فعالیت بازدارندگی

با توجه به اینکه مدل‌های ایجاد شده با استفاده از مجموع توصیف‌کننده‌های داکینگ مولکولی و ساختاری انتخاب شده توسط روش انتخاب متغیر لاسو، به‌عنوان مدل‌های برتر تعیین شدند و در ادامه بخش، بررسی اجمالی روی اثرات این متغیرها بر فعالیت دارویی ترکیبات مورد مطالعه صورت خواهد گرفت.

با توجه به اینکه واحدهای اندازه‌گیری متغیرهای مستقل (ضرایب رگرسیون) یکسان نیست، هرگز نمی‌توان از روی ضرایب رگرسیون غیر استاندارد، به میزان اهمیت و تأثیر یک متغیر مستقل بر روی متغیر وابسته پی برد. برای حل این مشکل، از ضریب استاندارد شده که طبق رابطه (۳-۱) محاسبه می‌شود[۶۰]، استفاده شد.

$$\beta'_k = (\frac{S_K}{S_Y})\beta_K$$ رابطه (۳-۱)

در این رابطه β'_k ضریب رگرسیون استاندارد شده توصیف‌کننده k ام ، β_K ضریب رگرسیون غیراستاندارد همان توصیف‌کننده، S_K انحراف استاندارد متغیر مستقل مورد نظر (توصیف‌کننده k ام) و S_Y انحراف استاندارد متغیر وابسته (فعالیت دارویی) می‌باشد که از روابط زیر به دست می‌آیند:

$$S_K = \sqrt{\frac{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}{n-1}}$$ رابطه (۳-۲)

$$S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$ رابطه (۳-۳)

x_{ik} مقدار توصیف‌کننده k ام برای ترکیب i ام و $\bar{x}_k$ مقدار میانگین متغیر مستقل (توصیف‌کننده)، n تعداد ترکیبات، Y_i مقدار متغیر وابسته برای ترکیب i ام و $\bar{Y}$ مقدار میانگین متغیر وابسته است[۶۰]. ضرایب استاندارد شده ۱۰ توصیف‌کننده انتخاب شده توسط لاسو در جدول(۳-۳۲) آورده شده است و در ادامه به بررسی رابطه بین این توصیف‌کننده‌ها و فعالیت دارویی ترکیبات پرداخته خواهد شد.

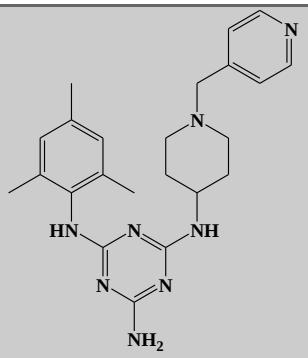
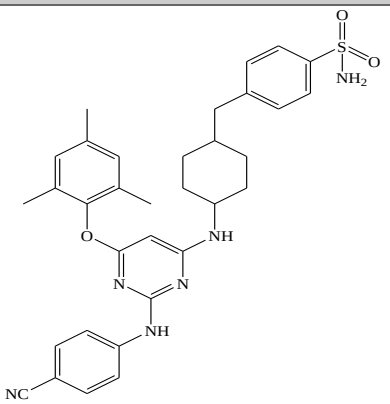
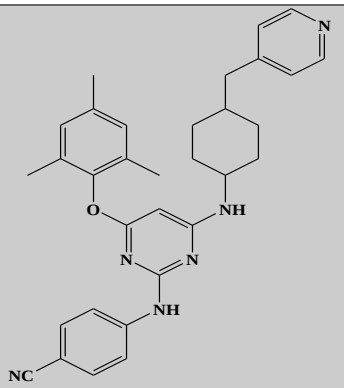
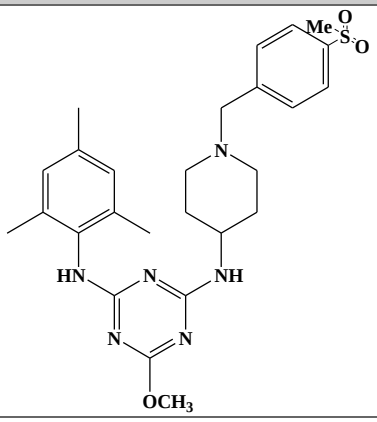
جدول(۳-۳۲) اثر متوسط توصیف‌کننده‌های انتخاب شده توسط روش لاسو موجود در مدل

Variable	A_5	pK_i	A_{12}	A_3	A_1	VE	A_2	A_{14}	A_8	TI
Mean effect	۰/۴۴۴۸	۰/۱۶۹۳	-۰/۱۴	-۰/۰۹۶	۰/۰۵۵۴	-۰/۰۰۲	-۰/۱۱۵	۰/۰۰۴۱	۰/۰۱۴۵	۰/۲۵۰۰

۳-۱-۱۰-۴ توصیف‌کننده A_5

اولین توصیف‌کننده موجود در مدل می‌باشد، این توصیف‌کننده حضور حلقه تری آزین موجود در ساختار مولکول‌ها را نشان می‌دهد که دارای ضریب مثبت در مدل برتر است و نشان می‌دهد که حضور حلقه تری آزین منجر به افزایش(pEC_{50}) می‌شود. به‌عبارت دیگر بازداری آنزیم نسخه‌بردار معکوس در فرایند رونویسی RNA، توسط مولکول‌هایی با حضور حلقه تری آزین زیاد می شود. چند مثال از اثر این توصیف‌کننده در جدول(۳-۳۳) ارائه شده است.

جدول(۳-۳۳) مثال‌هایی از اثر گروه تری آزین‌ها بر فعالیت دارویی

No.	ترکیب	حضور/عدم حضور توصیف‌کننده	pEC_{50}
۱		۱	۸/۳۱
۲		۰	۶/۷۲
۳		۰	۶/۸۹
۴		۱	۷/۹۲

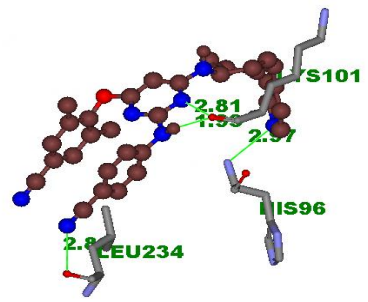
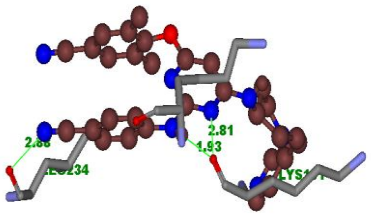
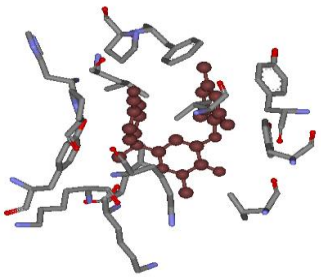
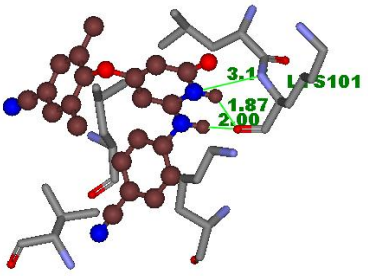
۳-۱-۱۰-۵ توصیف‌کننده pK_i

این توصیف‌کننده مربوط به منفی لگاریتم ضریب بازداری تئوری حاصل از شبیه سازی داکینگ مولکولی(K_i) است. K_i طبق رابطه زیر به‌دست می‌آید:

$$K_i = \exp (\Delta G \times 1000) / (R_{cal} \times T_K)$$

ΔG انرژی آزاد اتصال داکینگ است، R_{cal} ثابت گازها برابر با (Cal/mol. k) ۱/۹۸۷۱۹ و T_K نیز برابر با ۲۹۸ کلوین می‌باشد. این توصیف‌کننده دارای ضریب اثر مثبت در مدل می‌باشد. مقدار pEC_{50} برای ترکیباتی که pK_i بیشتری دارند، زیاد است و باعث افزایش خاصیت دارویی می شود. ترکیباتی با ضریب بازداری تئوری بزرگتر با احتمال بیشتری توانایی برقراری بر هم‌کنش های هیدروژنی با اسیدهای آمینه موجود در جایگاه فعال را دارند. در حالت کلی لیگاندهایی که قادر باشند با اسیدهای آمینه جایگاه فعال گیرنده در تشکیل پیوندهای هیدروژنی و هیدروفوبی مشارکت نمایند، از نظر طراحی مولکول‌های زیست فعال حائز اهمیت هستند، زیرا از طریق روش‌های اصلاح مولکولی، بهینه سازی جایگاه‌های فارماکوفوری احتمالی و ارتقا به مولکول‌های زیست فعال قوی‌تر امکان پذیر خواهد بود. در جدول(۳-۳۴) اسیدهای آمینه‌ای که با لیگاندهای موجود پیوند هیدروژنی برقرار کرده‌اند با رنگ سبز نمایش داده شده است و علاوه بر این طول پیوند نیز مشاهده می شود. با توجه به جدول(۳-۳۴) مشاهده می‌شود که لیگاند اول با اسید آمینه لیزین دو پیوند هیدروژنی و با هر کدام از اسیدهای آمینه هیستیدین و لوسین نیز یک پیوند هیدروژنی برقرارکرده است. لیگاند دوم با اسیدهای آمینه لیزین و لوسین به‌ترتیب دو و یک پیوند هیدروژنی برقرار کرده است. لیگاند سوم با مقدار عددی کمتر توصیف‌کننده نسبت به سایرین، هیچ پیوند هیدروژنی برقرار نکرده است و لیگاند چهارم نیز سه پیوند هیدروژنی با لیزین برقرار کرده است (نمایش پیوندهای هیدروژنی و محاسبه طول پیوند اشکال موجود در جدول با نرم افزار ViewerLite5.0 میسر شده است).

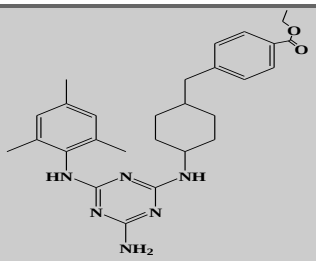
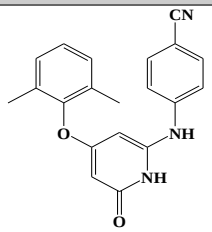
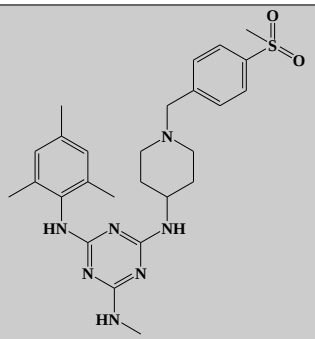
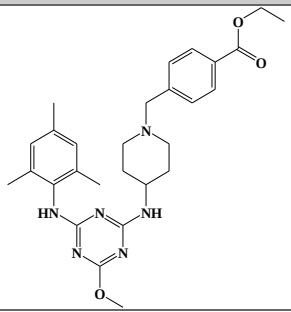
جدول(۳-۳۴) اثر توصیف‌کننده pK_i بر فعالیت دارویی

No.	ترکیب	تعداد پیوند هیدروژنی	طول پیوند هیدروژنی	pK_i (Docking Derived)	pEC_{50}
۱		۴	۱/۹۹ ۲/۸۰ ۲/۸۱ ۲/۵۷	۹/۴۵	۷/۳۳
۲		۳	۲/۸۱ ۱/۹۳ ۲/۸۸	۸/۸۶	۸/۳۴
۳		۰	-	۷/۵۸	۵/۴۷
۴		۳	۳/۱ ۱/۸۷ ۲/۰۰	۹/۶۹	۶/۸۲

۳-۱-۱۰-۶ توصیف‌کننده A_{12}

این توصیف‌کننده تعداد کربن‌های غیر آروماتیک با هیبریداسیون sp^2 است که دارای ضریب اثر منفی در مدل است و نشان می‌دهد که کاهش تعداد کربن‌های غیر آروماتیک با هیبریداسیون sp^2 منجر به افزایش pEC_{50} می شود. به‌عبارت دیگر بازداری آنزیم نسخه‌بردار معکوس در فرایند رونویسی RNA، توسط مولکول‌هایی با تعداد کربن‌های غیر آروماتیک با هیبریداسیون sp^2 کمتر، زیاد می‌شود. چند مثال از اثر این توصیف‌کننده در جدول(۳-۳۵) ارائه شده است.

جدول(۳-۳۵)اثر توصیف‌کننده A_{12} بر فعالیت دارویی

No.	ترکیب	A_{12}	pEC_{50}
۱		۱	۷/۶۹
۲		۵	۵/۸۵
۳		۰	۸/۲۲
۴		۱	۷/۸۳

۳-۱-۱۰-۷ توصیف‌کننده A_3

این توصیف‌کننده مربوط به تعداد گروه‌های نیترو است. این گروه‌ها سبب افزایش توزیع الکترواستاتیک شده و مقدار pEC_{50} را افزایش می‌دهند[۸۰].

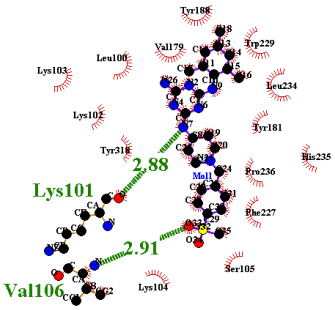
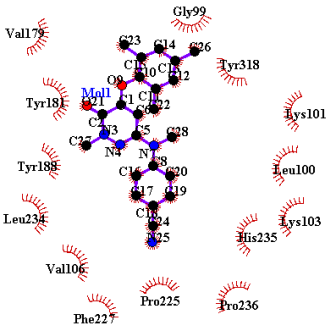
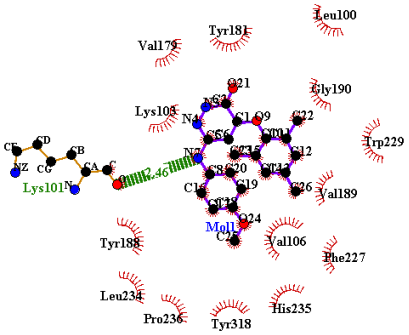
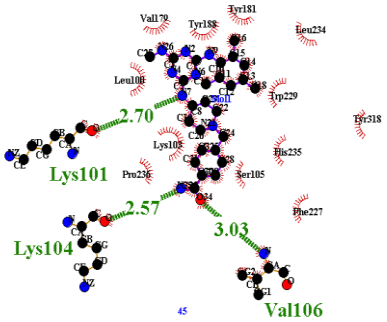
۳-۱-۱۰-۸ توصیف‌کننده A_1

این توصیف‌کننده مربوط به تعداد گروه‌های آمین متصل به حلقه آروماتیک است و دارای ضریب اثر مثبت در مدل است و با افزایش مقدار توصیف‌کننده افزایش در مقدار pEC_{50} مشاهده می‌شود. اتصال گروه آمین به مشتقات موجود، سبب بروز مقاومت مناسبی در برابر نوع ویروس جهش یافته تغییر پذیر HIV-1 می‌شود. از طرفی وجود این گروه حلالیت در آب را افزایش داده و به‌طور بالقوه سبب بهبود مشخصات فارموکینتیکی می‌شود. اتم نیتروژن گروه آمین به‌عنوان یک دهنده پیوند هیدروژنی از نقاط بالقوه یک فارماکوفور است[۷۹،۴].

۳-۱-۱۰-۹ توصیف‌کننده VE

این توصیف‌کننده مربوط به مجموع انرژی واندرالسی، هیدروژنی و حلال‌پوشی است. هر چقدر این انرژی منفی تر باشد نشان‌دهنده پایداری بیشتر کمپلکس لیگاند-پروتئین می‌باشد. با توجه به ضریب اثر متوسط منفی مربوط به این توصیف‌کننده قابل ذکر است که بازداری آنزیم و مقدار pEC_{50} افزایش می‌یابد. از شواهد موجود در جدول(۳-۳۶) نتیجه می‌شود که هر چه مقدار انرژی مربوطه منفی تر باشد علاوه بر افزایش فعالیت دارویی، احتمال برقراری پیوند هیدروژنی با اسیدهای آمینه نیز بیشتر می‌شود. اسیدهای آمینه سبز رنگ موجود در جدول(۳-۳۶) با لیگاند مربوطه پیوند هیدروژنی برقرار کرده‌اند که نام و نوع آن‌ها در جدول(۳-۳۶) ذکر شده است (اشکال موجود در جدول با استفاده از نرم افزار Ligplot2012 به‌دست آمده است).

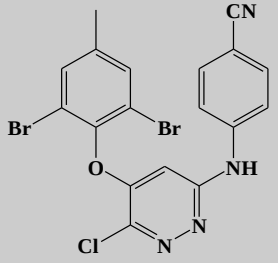
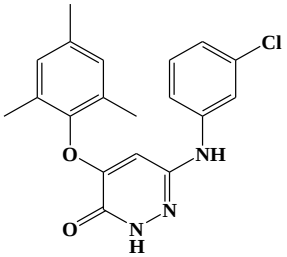
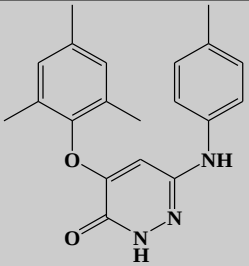
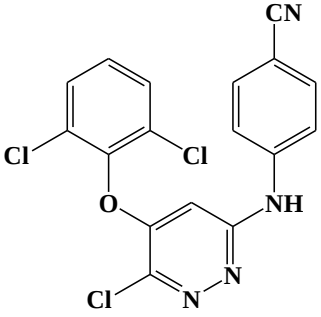
جدول(۳-۳۶) اثر توصیف‌کننده VE

No.	ترکیب	VE(kcal/mol) (Docking Derived)	pEC_{50}
۱		-۱۴/۳۰	۸/۲۶
۲		-۱۱/۵۲	۵/۴۷
۳		-۱۲/۴۹	۵/۸۰
۴		-۱۴/۸۵	۸/۳۴

۳-۱-۱۰-۱۰ توصیف‌کننده A_2

این توصیف‌کننده مربوط به حضور گروه آمیدین است که دارای ضریب اثر منفی در مدل است. با توجه به مشتقاتی که دارای این گروه هستند مشاهده می‌شود که مقدار pEC_{50} در حضور این گروه کم شده است. جدول(۳-۳۷) این موضوع را تایید می‌نماید.

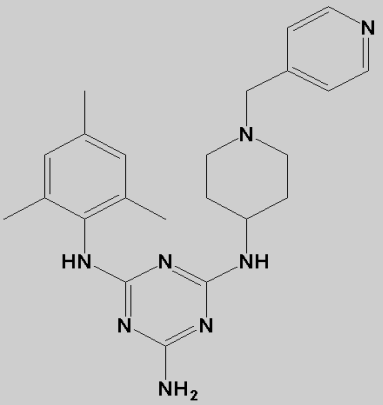
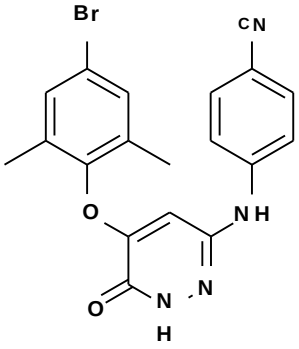
جدول(۳-۳۷) مثال‌هایی از اثر توصیف‌کننده A_2 بر فعالیت دارویی

No.	ترکیب	A_2	pEC_{50}
۱		۰	۷/۴۷
۲		۱	۴/۸۴
۳		۱	۵/۸۱
۴		۰	۶/۶۶

۳-۱-۱۰-۱۱ توصیف‌کننده A_{14}

این تو صیف‌کننده مربوط به کربن‌هایی با هیبریدا سیون sp^3 ا ست که به دو کربن دیگر مت صل شـده‌اند و از نوع دوم[1] هسـتند. با توجه به جدول(۳-۳۸) احتمال افزایش در مقدار pEC_{50} با افزایش در تعداد کربن‌های نوع دوم وجود دارد.

جدول(۳-۳۸) مثال‌هایی از اثر توصیف‌کننده A_{14} بر فعالیت دارویی

No.	ترکیب	A_{14}	pEC_{50}
۱		۳	۸/۳۱
۲		۰	۵/۹۲

۳-۱-۱۰-۱۲ توصیف‌کننده A_4

این توصیف‌کننده مربوط به حضور حلقه پیریدازین در ترکیبات است. این ترکیبات که مشتقات جدید پیشنهاد شده برای بیماری ایدز هستند قادرند که با غلظت‌های نسبتا کم (۳۴نانومولار تا ۵/۰۸

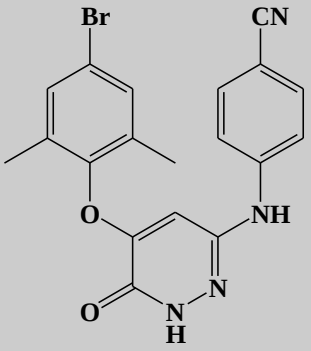
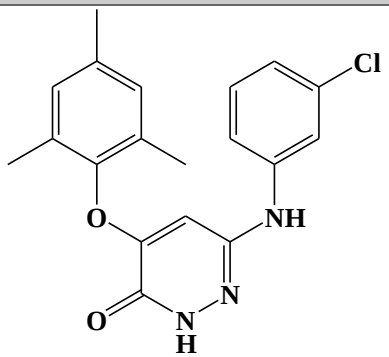
[1] - Secondary carbon

میکرومولار) بازداری خوبی را در برابر داروهای مرجع چون نویراپین[1] و دلاویریدین[2] از خود نشـــان دهند[۸۱].

۳-۱-۱۰-۱۳ توصیف‌کننده TI

این توصـیف‌کننده مربوط به انرژی نهایی درونی کل حاصـل از شـبیه سـازی داکینگ مولکولی ا ست. این نوع از انرژی مربوط به مجموع تغییرات ترم‌های انرژی درگیر در فرایند ت شکیل کمپلکس و لیگاند اســت. این توصــیف‌کننده دارای ضــریب اثر مثبت در مدل اســت به این معنا که هرچه انرژی درونی کل بزرگتر با شد سبب افزایش pEC_{50} می شود. جدول(۳-۳۹) مثال‌هایی از این تو صیف‌کننده را نشان می‌دهد.

جدول(۳-۳۹) مثال‌هایی از اثر توصیف‌کننده TI بر فعالیت دارویی

No.	ترکیب	TI(kcal/mol) (Docking Derived)	pEC_{50}
۱		-۱/۳	۵/۹۲
۲		-۱/۵	۴/۸۴

[1] - Nevirapine
[2] - Delaviridine

۳-۱-۱۱ نتیجه‌گیری نهایی

روش‌های QSAR کلاسیک سریع‌تر، ساده تر و راحت‌تر از روش‌های نوین هستند. آن‌ها شامل توصیف‌کننده‌های فیزیکوشیمیایی هستند که به‌طور واضح تعریف شده‌اند و برای آنالیز تعداد زیادی از ترکیبات و بررسی مجموعه داده‌های مولکولی بسیار مناسب می‌باشند. اما با وجود استفاده گسترده دارای محدودیت‌هایی چون در دسترس نبودن پارامترهای فیزیکوشیمیایی مناسب، استفاده از پارامترهای نامطلوب برای توصیف کردن بر هم‌کنش‌های گیرنده با دارو و نبود خروجی گرافیکی به‌منظور تفسیر راحت‌تر و جامع‌تر و ... می‌باشد. از این‌رو شبیه سازی داکینگ مولکولی به‌جهت در نظر گرفتن برهم‌کنش‌های گیرنده و دارو به‌منظور بهبود در طراحی منطقی دارو پیشنهاد می‌شود[۹۱]. توصیف‌کننده‌های حاصل از شبیه سازی داکینگ مولکولی در تفسیر پذیری بهتر مدل نقش مهمی را ایفا نموده و مدل ایجاد شده در بردارنده ی اطلاعات ارزشمندی است که برهم کنش لیگاند-پروتئین را به خوبی تفسیر می‌کند.

در این پروژه روش لاسو به‌عنوان یک روش انتخاب متغیر نوین پیشنهاد داده شد و در مقایسه با روش رگرسیون گام به گام بهتر عمل کرده است. روش رگرسیون گام به گام علاوه بر مزایایی چون تفسیر ساده، محاسبات ساده و یک نرم افزار پر کاربرد دارای محدودیت‌های زیر می‌باشد :

- متغیرها در یک لحظه وارد و خارج می‌گردند و این دسترسی به یک مدل بهینه را تقریبا غیر ممکن می‌سازد.
- مقدار R^2 به‌طور غیر واقعی بالاست.
- ضرایب رگرسیون نیاز به انقباض دارند
- مقادیر P معنی واقعی صحیحی ندارند.
- در حضور متغیرهای هم‌خط با مشکل روبه‌رو می‌شود.
- خطاهای استاندارد پارامترهای تخمین شده بسیار کوچک است.

از این‌رو از روش انتخاب متغیر انقباضی لاسو استفاده شد و نتایج مدل‌سازی، برتری این روش را نسبت به روش رگرسیون گام به گام نشان می‌دهد.

نتایج نشـان می‌دهند که مدل‌های خطی و غیرخطی توسـعه یافته می‌توانند به عنوان روش‌های موفق برای مدل‌ســازی و پیش‌بینی فعالیت ضــد ایدز ترکیبات مورد مطالعه باشــند و این مدل‌ها می توانند به محققان در طراحی داروهایی با فعالیت ضــد ایدز قوی‌تر کمک کنند. بنابراین بر اســاس این مدل‌ها می‌توان فعالیت ضـدHIV را برای همین طبقه از ترکیبات که هنوز سـنتز نشـده‌اند پیش بینی کرد و در واقع این روش به طراحان دارو این امکان را می‌دهد که قبل از این‌که بخواهند یک ترکیب دارویی را ســنتز کنند، ابتدا با اســتفاده از مدل معتبر خاصــیت دارویی آن را پیش‌گویی کنند و در صورتی‌که دارای فعالیت دارویی مناسبی باشند، ترکیب سنتز شود و بدین ترتیب با صرف هزینه و وقت کم‌تر در مسیر تهیه داروهایی با اثر بخشی بهتر گام بردارند.

آینده نگری

- ✓ می‌توان از روش‌های انتخاب متغیر انقبا ضی جدید چون لارس[1]، ا سکار[2] و ا سکد[3] و ... استفاده نمود.
- ✓ روش شـــبیه ســـازی داکینگ مولکولی و دینامیک مولکولی را برای بررســـی برهم‌کنش سایر مشتقات دارویی و گیرنده می‌توان استفاده کرد.
- ✓ روش شبیه سازی دینامیک مولکولی را نیز برای بررسی برهم‌کنش مشتقات دارویی و گیرنده می‌توان استفاده کرد
- ✓ این تحقیقات را می‌توان با پژوهش روی خواص دارویی ترکیبات جدید ادامه داد تا نهایتاً منجر به ساخت داروهای جدیدتر با اثربخشی بهتر گردد.

[1] (Least angle regression) LARS
[2] (Otagonal shrinkage and clustering algorithm for regression) OSCAR
[3] SCAD

۳-۲ فهرست منابع

[1]. Brocchieri, L., & Karlin, S. (2005). “Protein length in eukaryotic and prokaryotic proteomes”. J. Nucleic acids res, 33(10), 3390-3400.

[۲]. سید مکاری م (۱۳۹۲)،" داروشناسی"، چاپ اول، نشر فرهیختگان

[۳]. صبوری ع، موسوی موحدی ع، (۱۳۷۵)، "سینتیک آنزیمی"، انتشارات دانشگاه تهران، چاپ اول،۲۲۰-۲۰۲

[۴]. ا شنایدر گ، برینگهاس ک، (۲۰۰۸)، "طراحی مولکولی مفاهیم و کاربردها"، قا سمی ج، چاپ اول، ن شر پژوه شی نوآوران شریف

[5]. Finzi, D., Blankson, J., Siliciano, J. D., Margolick, J. B., Chadwick, K., Pierson, T., & Quinn, T. C. (1999). “Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy”. J. Nat. Med, 5(5), 512-517.

[6]. Richman, D. D. (2001). “HIV chemotherapy” .J. Nature, 410(6831), 995-1001.

[7]. Sarafianos, S. G., Marchand, B., Das, K., Himmel, D. M., Parniak, M. A., Hughes, S. H., & Arnold, E. (2009). “Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition”. J. Mol. Biol., 385(3), 693-713.

[8]. Fabian, J., & Lewars, E. (2004). “Azabenzenes (azines) The nitrogen derivatives of benzene with one to six N atoms: Stability, homodesmotic stabilization energy, electron distribution, and magnetic ring current; a computational study”. J. Chem., 82(1), 50-69.

[9]. Najafi, A., & Sobhan Ardakani, S. (2011). “2D autocorrelation modelling of the anti-HIV HEPT analogues using multiple linear regression approaches”. J. Mol Simul, 37(01), 72-83.

[10]. Cichero, E., Buffa, L., & Fossa, P. (2011). “3, 4, 5-Trisubstituted-1, 2, 4-4H-triazoles as WT and Y188L mutant HIV-1 non-nucleoside reverse transcriptase inhibitors: docking-based CoMFA and CoMSIA analyses”. J. Mol model, 17(7), 1537-1550.

[11]. Ruixin, Z., Fei, W., Qi, L., & Tingguo, K. (2011). “Quantitative Structure-Activity Relationship of IOPY/ISPY Analogues as HIV-1 Non-Nucleoside Reverse Transcriptase Inhibitors”. J. ACTA CHIM SIN, 69(15). 1731-1736.

[12]. Goodarzi, M., Freitas, M. P., & Vander Heyden, Y. (2011). “Linear and nonlinear quantitative structure–activity relationship modeling of the HIV-1 reverse transcriptase inhibiting activities of thiocarbamates”. J. Anal chim acta, 705(1), 166-173.

[13]. Kumar, S., Singh, V., & Tiwari, M. (2011). “QSAR modeling of the inhibition of reverse transcriptase enzyme with benzimidazolone analogs”. J. Med Chem Res, 20(9), 1530-1541.

[14]. Putz, M. V., Lazea, M., Putz, A. M., & Duda-Seiman, C. (2011). “Introducing catastrophe-QSAR. Application on modeling molecular mechanisms of pyridinone derivative-type HIV non-nucleoside reverse transcriptase inhibitors”. J. Mol. Sci, 12(12), 9533-9569.

[15]. Karpov, P. V., Baskin, I. I., Zhokhova, N. I., Nawrozkij, M. B., Zefirov, A. N., Yablokov, A. S., ... & Zefirov, N. S. (2011). “One-class approach: models for virtual screening of non-nucleoside HIV-1 reverse transcriptase inhibitors based on the concept of continuous molecular fields”. J. RUSS CHEM

B, 60(11), 2418-2424.

[16]. Cichero, E., & Fossa, P. (2012). “Docking-based 3D-QSAR analyses of pyrazole derivatives as HIV-1 non-nucleoside reverse transcriptase inhibitors”. J. Mol model, 18(4), 1573-1582.

[17]. Mao, Y., Li, Y., Hao, M., Zhang, S., & Ai, C. (2012). “Docking, molecular dynamics and quantitative structure-activity relationship studies for HEPTs and DABOs as HIV-1 reverse transcriptase inhibitors”. J. Mol model, 18(5), 2185-2198.

[18]. de Brito, M. A., Rodrigues, C. R., Cirino, J. J. V., Araújo, J. Q., Honório, T., Cabral, L. M., ... & Albuquerque, M. G. (2012). “Residue-ligand interaction energy (ReLIE) on a receptor-dependent 3D-QSAR analysis of S-and NH-DABOs as non-nucleoside reverse transcriptase inhibitors”. J. Molecules, 17(7), 7666-7694.

[19]. Ashok, P., Babu, K. K., Murugesan, S., & Ganguly, S. (2012). “Design and Docking Studies of Some Novel 1-Phenyl-2, 3, 4, 9-tetrahydro-1H-pyrido [3, 4-b] indole-3-carboxylic acids as Inhibitors of Human Immunodeficiency Virus Type-1 Reverse Transcriptase”. Asian J. Chem, 24(12), 5857-5860.

[20]. Pirhadi, S., & Ghasemi, J. B. (2012). “Pharmacophore identification, molecular docking, virtual screening, and in silico ADME studies of non-nucleoside reverse transcriptase inhibitors”. J. Molec Inf, 31(11-12), 856-866.

[21]. Sapre, N. S., Jain, N., Gupta, S., & Sapre, N. (2013). “Ligand based 3D-QSAR modelling studies on 2-amino-6-aryl sulfonylbenzonitriles (AASBNs) as non-nucleoside reverse transcriptase inhibitors of HIV-1”. J. RSC Adv, 3(26), 10442-10451.

[22]. Akyüz, L., & Sarıpınar, E. (2013). “Conformation depends on 4D-QSAR analysis using EC–GA method: pharmacophore identification and bioactivity prediction of TIBOs as non-nucleoside reverse transcriptase inhibitors”. J. Enzyme Inhib. Med., 28(4), 776-791.

[23]. Debnath, U., Verma, S., Jain, S., Katti, S. B., & Prabhakar, Y. S. (2013). “Pyridones as NNRTIs against HIV-1 mutants: 3D-QSAR and protein informatics”. J. Comput Aided Mol Des, 27(7), 637-654.

[24]. Wu, H. Q., Yao, J., He, Q. Q., & Chen, F. E. (2014). “Docking-based CoMFA and CoMSIA studies on naphthyl-substituted diarylpyrimidines as NNRTIs”. J. SAR QSAR Environ Res, 25(10), 761-775.

[25]. Patil, V. M., Masand, N., & Anand, P. (2014). “QSAR Study on Isatin Analogues Acting as HIV-Reverse Transcriptase Inhibitors”. J. Int Drug Des, 1(5), 1249-1254.

[26]. Li, L., Yan, F., Xu, X., Jia, Q., Wang, Q., & Ma, P. (2016). “A QSAR Model for Predicting the Anti HIV-1 Activity of TIBO Derivatives Using the Norm Indexes”. J. Bull. Korean Chem. Soc, 37(6), 923-930.

[27]. Masand, V. H., Mahajan, D. T., Hadda, T. B., Jawarkar, R. D., Chavan, H., Bandgar, B. P., & Chauhan, H. (2014).“Molecular docking and quantitative structure–activity relationship (QSAR) analyses of indolylarylsulfones as HIV-1 non-nucleoside reverse transcriptase inhibitors”. J. Med Chem Resh, 23(1), 417-425.

[28]. Kumar, S., & Tiwari, M. (2015). “Topomer-CoMFA-based predictive modelling on 2, 3-diaryl-substituted-1, 3-thiazolidin-4-ones as non-nucleoside reverse transcriptase inhibitors”. J. Med Chem, 24(1), 245-257.

[29]. Liu, X., Chen, X., Zhang, L., Zhan, P., & Liu, X. (2015). “3D-QSAR and docking studies on piperidine-substituted diarylpyrimidine analogues as HIV-1 reverse transcriptase inhibitors”. J. Med Chem Res, 24(8), 3314-3326.

[30]. Xu, H. R., Fu, L., Zhan, P., & Liu, X. Y. (2016). “3D-QSAR analysis of a series of S-DABO derivatives as anti-HIV agents by CoMFA and CoMSIA”. J. SAR QSAR Environ Res, 27(12), 999-1014.

[31]. Shaik, B., Agrawal, V., P Gupta, S., & Menon, U. (2017). “Quantitative Structure-Activity Relationship and Docking Studies on a Series of Oxadiazole and Triazole Substituted Naphthyridines as HIV-1 Integrase Inhibitors”. J. Lett Drug Des Discov, 14(1), 10-27.

[32]. Darnag, R., Mazouz, E. M., Schmitzer, A., Villemin, D., Jarid, A., & Cherqaoui, D. (2010). “Support vector machines: development of QSAR models for predicting anti-HIV-1 activity of TIBO derivatives”. Eur J Med Chem, 45(4), 1590-1597.

[33]. Akyüz, L., Sarıpınar, E., Kaya, E., & Yanmaz, E. (2012). “4D-QSAR study of HEPT derivatives by electron conformational–genetic algorithm method”. J. SAR QSAR Environ Res, 23(5-6), 409-433.

[34]. Toropova, A. P., Toropov, A. A., Veselinović, J. B., Miljković, F. N., & Veselinović, A. M. (2014). “QSAR models for HEPT derivates as NNRTI inhibitors based on Monte Carlo method”. Eur J Med Chem, 77, 298-305.

[35]. Nizami, B., Tetko, I. V., Koorbanally, N. A., & Honarparvar, B. (2015). “QSAR models and scaffold-based analysis of non-nucleoside HIV RT inhibitors”. J. Chemometr Intell Lab Syst, 148, 134-144.

[36]. Chamjangali, M. A., Bagherian, G., Ashrafi, M., & Amin, A. H. (2015). “Prediction of the anti-HIV activities of PETT analogs as non-nucleoside HIV-1 reverse transcriptase inhibitors by linear and non-linear QSAR models”. J. Appl Chem, Vol, 9(32).

[37]. Brereton, R. G. (2003). “Chemometrics: data analysis for the laboratory and chemical plant”. John Wiley & Sons.

[38]. Frank, I. E., & Kowalski, B. R. (1982). “Chemometrics”. J.Anal Chem, 54(5), 232R-243R.

[39]. Kowalski, B. R. (1980). “Chemometrics”. J. Anal Chem, 52(5), 112R-122R.

[40]. Wold, S. (1995). “Chemometrics; what do we mean with it, and what do we want from it?”. J. Chemometr. Intell. Lab., 30(1), 109-115.

[41]. He, L., & Jurs, P. C. (2005). “Assessing the reliability of a QSAR model's prediction”. J. Mol Graph Model, 23(6), 503-523.

[42].] Guha, R. (2005). “Methods to improve the reliability, validity and interpretability of QSAR models (Doctoral dissertation The Pennsylvania State University)”.

[۴۳]. شعیری ف، (۱۳۸۵)، پایان نامه کارشناسی ارشد،"پیشگویی فعالیت پیرازولوکوئینولین‌ها با استفاده از شبکه عصبی موجک و کاربرد روشRAFA در اندازه‌گیری همزمان آمپی سیلین و آموکسی سیلین "، دانشگاه علوم پایه دامغان

[۴۴]. یانگ د، (۲۰۰۹)، "طراحی محاسباتی دارو"، رسول امینی س ، چاپ اول، انتشارات جامعه نگر تهران

[۴۵]. قرقانی س، (۱۳۹۱)، رساله دکترا، " روابط کمی ساختار-فعالیت (QSAR) بر مبنای ساختار پروتئین و استفاده از مدل سازی همسانی، داکینگ مولکولی و شبیه سازی دینامیک مولکولی، طراحی داروهای جدید بر مبنای قطعات مولکولی (Fragment) با تاثیر همزمان بر پروتئین‌های استیل کولین استراز و آمین اکسیداز حساس به سمی‌کاربازید"، دانشگاه صنعتی اصفهان.

[۴۶]. مراد ا، گلیایی ب، (۱۳۹۰)، " بررسی محاسباتی برهم کنش پروتئین لیگاند"، مجله تحصیلات تکمیلی بیوفیزیک در مرکزتحقیقات بیوشیمی و بیوفیزیک دانشگاه تهران، شماره۳، صفحه۲۴

[۴۷]. نیکنام ن، (۱۳۸۹)، پایان نامه کارشناسی ارشد، " تعیین ساختار آنزیم لیزوزیم در pH=۷ توسط روش 2D-HNMR"، گروه بیوفیزیک دانشگاه تربیت مدرس

[48]. Alonso, H., Bliznyuk, A. A., & Gready, J. E. (2006). “Combining docking and molecular dynamic simulations in drug design”. J. Med Res Rev, 26(5), 531-568.

[49]. Wermuth, C. G., Ganellin, C. R., Lindberg, P., & Mitscher, L. A. (1998). “Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998)”. J. Pure Appl. Chem, 70(5), 1129-1143.

[50]. https://en-lifesci.tau.ac.il/bioinformatics-unit/Autodock_2016

[51]. www.RCSB.org

[52]. Goodsell, D. S., Morris, G. M., & Olson, A. J. (1996). “Automated docking of flexible ligands: applications of AutoDock”. J. Mol. Recogn, 9(1), 1-5.

[53]. Feinstein, W. P., & Brylinski, M. (2015). “Calculating an optimal box size for ligand docking and virtual screening against experimental and predicted binding pockets”. J. Cheminform, 7(1), 1.

[54]. Yasri A., Hartsough D., (2001) “Toward an optimal procedure for variable selection and QSAR model building”. J. Chem. Inf. Comput. Sci., 41, pp 1218-1227

[55]. Todeschini, R., & Consonni, V. (2008). “Handbook of molecular descriptors” (Vol. 11). John Wiley & Sons.

[56]. www.iasbs .ac.ir/chemistry/chemometrics/…/8th/qsar_introduction.ppt

[57]. Lapin, L. L. (1983). “Probability and statistics for modern engineering”. Thomson Brooks/Cole, 2nd Ed, pp188

[58]. SPSS for windows, the statistical package for IBMPC, SPSS Inc.,http://www.spss.com, 2011.

[59]. [Xu, L., & Zhang, W. J. (2001). “Comparison of different methods for variable selection”. J. Anal. Chim. Acta, 446(1), 475-481.

[۶۰]. فرشادفر ع، (۱۳۸۰)، "اصول و روش‌های پیشرفته آماری (تجزیه رگرسیون)"، چاپ دوم، انتشارات طاق بستان.

[61]. Fan, J., & Li, R. (2001). “Variable selection via nonconcave penalized likelihood and its oracle properties”, J. ASA, 96(456), 1348-1360.

[62]. Tibshirani, R. (1996). “Regression shrinkage and selection via the lasso”. J. R Stat Soc, Series B (Methodological), 267-288.

[۶۳]. شرفی م، (۱۳۹۲)، پایان نامه کار شنا سی ار شد،" ا ستنباط آماری مدل رگر سیونی با خطاهای خودبازگشتی به روش لاسو"، دانشکده علوم ریاضی دانشگاه شیراز

[۶۴]. اشرفی م، (۱۳۸۹)،پایان نامه کارشناسی ارشد ،" مطالعه ارتباط کمی ساختار- فعالیت مشتقات تیوکربامات‌ها به عنوان دسته‌ی جدیدی از بازدارنده‌های غیر نوکلئوزیدی HIV" دانشکده شیمی دانشگاه صنعتی شاهرود

[65]. Breiman, L. (2001). "Random forests". J. Mach. Learn. Res., 45(1), 5-32.

[66]. [Hastie T, Tibshirani R, Friedman J, (2009), "The Elements of Statistical Learning Data Mining Inferences, and Prediction", 2nd Ed, Springer, New York.

[۶۷]. عجم ز، (۱۳۹۲)، پایان نامه کارشناسی ارشد،" پیش بینی فعالیت ضد ایدز مشتقات غیر نوکلئوزیدی تیوکربومات به روش جنگل‌های تصادفی"، دانشکده شیمی، دانشگاه صنعتی شاهرود.

[۶۸]. صفری ف، (۱۳۹۲)،پایان نامه کار شنا سی ار شد ،"مقای سه روش‌های یادگیری ما شین با روش‌های رده بندی متداول جهت رده بندی تصاویر ماهواره ای ابرطیف" دانشکده آمار و علوم ریاضی دانشگاه صنعتی شاهرود

[۶۹]. چترچی س.، هادی ع.، پرایس ب.، (۱۳۸۴)، "تحلیل رگرســـیون با مثال"، نیرومند ح. ع.، چاپ اول، انتشـــارات دانشگاه فردوسی مشهد.

[70]. Goodarzi M., Deshpande S., Murugesan V., Katti S.B., Prabhakar Y.S., (2009), "Is feature selection essentioal for ANN modeling?", J. QSAR Com. Sci., 28, pp 1487-1499.

[71]. http://www.hyper.com/?tabid=360

[72]. Joy, S., Nair, P.S., Hariharan, R., Pillai, M.R. "Detailed comparison of the protein ligand docking efficiencies of GOLD, a commercial package and arguslab, a licensable freeware", J. In Silico Biol., Vol. 6, No. 6, pp. 601-605, 2006.

[73]. http://accelrys.com/products/collaborative-science/biovia-discovery studio/visualization. Html

[74]. Wallace, A. C., Laskowski, R. A., & Thornton, J. M. (1995). "LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions". J. Protein Eng, 8(2), 127-134.

[75]. SPSS for windows Statistical package for IBM PC, SPSS Inc, http: //www. spss.com.

[76]. MATLAB 7.8, the Math Work, Inc., Natick, MA, USA.

[77]. [https://www.r-project.org/

[78]. Chen, X., Liu, X., Meng, Q., Wang, D., Liu, H., De Clercq, E., & Liu, X. (2013). "Novel piperidinylamino-diarylpyrimidine derivatives with dual structural conformations as potent HIV-1 non-nucleoside reverse transcriptase inhibitors". J. Bioorg. Med. Chem., 23(24), 6593-6597.

[79]. Chen, X., Zhan, P., Liu, X., Cheng, Z., Meng, C., Shao, S., & Liu, X. (2012). "Design, synthesis, anti-HIV evaluation and molecular modeling of piperidine-linked amino-triazine derivatives as potent non-nucleoside reverse transcriptase inhibitors". J. Bioorg. Med. Chem, 20(12), 3856-3864.

[80]. Wang, J., Zhan, P., Li, Z., Liu, H., De Clercq, E., Pannecouque, C., & Liu, X. (2014). "Discovery of nitropyridine derivatives as potent HIV-1 non-nucleoside reverse transcriptase inhibitors via a structure-based core refining approach". Eur. J. Med. Chem. 76, 531-538.

[81]. Li, D., Zhan, P., Liu, H., Pannecouque, C., Balzarini, J., De Clercq, E., & Liu, X. (2013). “Synthesis and biological evaluation of pyridazine derivatives as novel HIV-1 NNRTIs” J. Bioorg. Med. Chem, 21(7), 2128-2134.

[82]. Chen, W., Zhan, P., Rai, D., De Clercq, E., Pannecouque, C., Balzarini, J., & Liu, X. (2014). “Discovery of 2-pyridone derivatives as potent HIV-1 NNRTIs using molecular hybridization based on crystallographic overlays”. J. Bioorg. Med. Chem, 22(6), 1863-1872.

[83]. Ahmadzadeh, S., Emamzadeh, R., Nazari, M., (2016). “Molecular Docking Study for the Identification of Role of a critical Residue on Binding Site Properties of HER2 mAb Inhibitors”. Iran Gen Soc.

[84]. Feinstein, W. P., & Brylinski, M. (2015). Calculating an optimal box size for ligand docking and virtual screening against experimental and predicted binding pockets”. J. Cheminform., 7(1), 1.

[85]. Zhou, Q., Zhang, N., Zhang, C., Huang, L., Niu, Y., Zhang, Y., & Liu, W. (2010). “Molecular mechanism of enantioselective inhibition of acetolactate synthase by imazethapyr enantiomers”. J. Agric. Food Chem., 58(7), 4202-4206.

[86]. Parida, P., Deka, P., Shankar, B., & Yadav, R. N. S. (2014). “In silico antigenic site evaluation and antiviral therapy against dengue serotypes”. Bangladesh J. Pharmacol., 9(1), 83-95.

[87]. Awale, M., Kumar, V., Saravanan, P., & Mohan, C. G. (2010). "Homology modeling and atomic level binding study of Leishmania MAPK with inhibitors”. J. Mol. Model., 16(3), 475-488.

[88]. Toor, J. S., Sharma, A., Kumar, R., Gupta, P., Garg, P., & Arora, S. K. (2011). “Prediction of drug-resistance in HIV-1 subtype C based on protease sequences from ART naive and first-line treatment failures in North India using genotypic and docking analysis”. J. Antiviral Res, 92(2), 213-218.

[89]. MacDougall, I. J., & Griffith, R. (2006). “Selective pharmacophore design for α 1-adrenoceptor subtypes”. J Mol Graph Model, 25(1), 146-157.

[90].] http://www.weare.ir/interesting/scientific/molecular-docking

[91]. Bultinck, P., De Winter, H., Langenaeker, W., & Tollenare, J. P. (Eds.). (2003). “Computational medicinal chemistry for drug discovery”. CRC Press

Abstract

For model construction, two kinds of descriptors namely molecular docking descriptors and calculated structural descriptors were used. The new molecular docking descriptors were derived from molecular docking by considering the interaction between compounds (as ligands) and protein (as receptor). The structural descriptors were calculated from structure of compounds. After generation of descriptos, the most important descriptors were selected by applying stepwise regression (SR) and least absolute shrinlage and selector operator (LASSO) as variable selection methods. The selected descriptors were used as inputs for construction QSAR models using Artificial Neural Network (ANN) and Random Forest (RF) as a non-linear and linear methods. For ANN model, the data set was divided into training set, valid set and test set containing 51, 11 and 11 compounds, respectively. The data set was divided into training set and test set including 62 and 11 compounds, respectively for RF modeling. All effective parameters of ANN and RF models were optimized. The validation study of the ANN and RF models was performed by prediction of the activities of the test set compounds. The obtained results from models validation showed correlation coefficient of 0.9274 and 0.8968 for prediction the activity of test set by ANN and RF models, respectively. The mean square error of 0.0597 and 0.0849 were also found for prediction of test set activity by ANN and RF models, respectively.

Keyword: QSAR, Anti -HIV, Artificial neural network, Random forest, Molecular Docking, LASSO

Faculty of Chemistry

M.Sc. Thesis in Analytical Chemistry

Study of Quantitative structure-activity relationship of pyridine & pyrimidine derivatives as inhibitors of HIV

By:**Zeinab Mozafari**

Superviser:

Dr. Mansour Arab Chamjangali

Advisor:

Dr. Mohammad Arashi

Jaunary 2017