
بسم الله الرحمن الرحيم

دانشکده شیمی

پایان نامه کارشناسی ارشد شیمی تجزیه

مطالعه ارتباط کمی ساختار- فعالیت برخی از بازدارنده‌های c-Met سولفون‌دار جدید با فعالیت ضد سرطان

نگارنده:

محدثه لطفی

استاد راهنما:

دکتر منصور عرب چم جنگلی

اسفند ۱۴۰۰

اگر این مختصر را ارزشی است؛

آن ارزش تقدیم به :

دستان آسمانی و شایان ستایش پدرم

دل دریایی و چشمان پر امید مادرم

قلب مهربان و پر عاطفه خواهر و برادرم

و

تمامی پویندگان علم و دانش

سپاسگزاری

این جانب از استاد محترم جناب آقای دکتر منصور عرب چم جنگلی. تشکر می نمایم که مرا در انجام این پروژه یاری نمودند و از خداوند متعال پیروزی و موفقیت روز افزون ایشان را خواستارم. از اساتید گرانقدر م جناب آقای دکتر قدمعلی باقریان دهقی وجناب آقای دکتر ناصر گودرزی که زحمت داوری پایان نامه ی حاضر را داشته و از محضر ایشان در کلاس های خویش بهره ی وافر برده ام تشکر می نمایم.

از خانواده ی عزیزم و همچنین افسانه ی عزیزم به خاطر تمام زحماتشان تشکر می نمایم. از دوست عزیزم سرکار خانم مریم مظفری برای تمام یاری رسانی هایی که در به پایان رسانیدن این پایان نامه داشتند تشکر می نمایم. همچنین از سایر اساتید محترم دانشکده شیمی که در طول دوران تحصیلاتم افتخار شاگردیشان را داشتم، کارمندان محترم دانشکده و تمامی کسانی که مرا به نحوی یاری رساندند نهایت تشکر و قدردانی را به جا آورده و از حضرت تعالی سلامت و توفیق روز افزون آن ها را مسئلت می نمایم.

تعهد نامه

اینجانب **محدثه لطفی** دانشجوی دوره کارشناسی ارشد **شیمی تجزیه** دانشکده **شیمی** دانشگاه صنعتی شاهرود، نویسنده پایان نامه **مطالعه ارتباط کمی ساختار- فعالیت بازدارنده‌های c-Met سولفون‌دار جدید با فعالیت ضد سرطان** تحت راهنمایی **آقای دکتر منصور عرب چم جنگلی** متعهد می‌شوم:

- 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 می‌باشد 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 بوده‌اند در مقالات مستخرج از پایان نامه رعایت می‌گردد.
- 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، برنامه‌های رایانه‌ای، نرم افزارها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می‌باشد. 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 نمی‌باشد.

چکیده

هدف از این مطالعه استفاده از تکنیک غربالگری داده‌ها برای کاهش ابعاد ماتریس داده است. برای این منظور از روش جدید پیش‌پردازش جریمه‌ای Ridge برای کاهش ابعاد داده‌ها استفاده شد. سپس از روش انحراف قدر مطلق به‌طور هموار کوتاه شده (SCAD) به‌عنوان روش انتخاب متغیر جریمه‌ای برای انتخاب مؤثرترین توصیف کننده‌ها و مدل‌سازی استفاده شد. توصیف کننده‌های انتخاب شده به‌عنوان ورودی برای ساخت مدل QSAR با استفاده از روش SCAD استفاده شد. برای این منظور، مدل QSAR با استفاده از اجرای چندگانه Ridge به‌عنوان یک تکنیک پیش‌پردازش برای غربالگری داده‌ها توسعه داده شد. توصیف کننده‌های باقی‌مانده به مدل جریمه شده SCAD وارد شدند. پس از اعمال SCAD به داده‌های مجموعه آموزش، از برخی توصیف کننده‌های مؤثر (۹) در ساخت مدل استفاده شد. مدل بهینه هدف (Ridge-SCAD) با استفاده از محاسبه پارامترهای آماری، دامنه کاربرد و آزمون Y-تصادفی مورد ارزیابی قرار گرفت. آزمون‌های مختلف ذکر شده نتایج رضایت بخشی را نشان می‌دهند. عملکرد مدل Ridge-SCAD برای پیشنهاد برخی از مهارکننده‌های قوی c-Met استفاده شد. کارایی ترکیبات پیشنهادی با استفاده از داکینگ مولکولی و محاسبه قانون پنج لیپینسکی بررسی شد. نتایج حاصل قدرت ترکیبات جدید را اثبات می‌کند. برای هدف مقایسه، مدل QSAR بدون استفاده از روش Ridge به‌عنوان یک تکنیک پیش‌پردازش توسعه داده شد و همه توصیف کننده‌های محاسبه‌شده به‌عنوان ورودی‌های SCAD تعریف شدند. مدل SCAD با استفاده از داده‌های مجموعه آموزش ساخته شد و ۱۲ توصیف کننده به‌عنوان توصیف کننده معنی‌دار به‌دست آمد. فعالیت دارویی داده‌های آزمون با استفاده از مدل SCAD بهینه پیش‌بینی شد. پارامترهای آماری محاسبه شده مدل SCAD با مدل Ridge-SCAD مقایسه شد. نتایج نشان دهنده قدرت پیش‌بینی بالای مدل Ridge-SCAD در برابر مدل SCAD است.

کلمات کلیدی: QSAR، c-Met، پیش‌پردازش، Ridge، SCAD، داکینگ مولکولی

مقاله مستخرج از این پایان نامه تحت عنوان زیر در ژورنال ISI تحت داوری قرار گرفته است:

M. Lotfi, M. Arab Chamjangali and Z. Mozafari. Ridge regression coupled with a new uninformative variable elimination algorithm as a new descriptor screening method: Application of data reduction in QSAR study of some sulfonated derivatives as c-Met inhibitors. Chemometrics and Intelligent Laboratory Systems, under review manuscript on 13 May 2022 .

فهرست مطالب

فهرست شکل‌ها:

فهرست جدول‌ها:

فهرست رابطه‌ها:

فصل اول

مقدمه

مقدمه

از بزرگ‌ترین مشکلاتی که همواره در مسیر توسعه جوامع بشری قرار داشته، مربوط به سلامت و مبارزه با انواع بیماری‌ها از قبیل سرطان، ایدز، دیابت، آلزایمر و ... بوده است. از این‌رو بایستی به‌دنبال داروها و روش‌هایی برای از بین بردن مشکلات بود. سرطان شاید یکی از پیچیده‌ترین مشکلات بشر در زمینه سلامتی باشد که هر ساله هزینه‌های مالی و عاطفی فراوانی به جوامع تحمیل می‌کند. علی‌رغم پیشرفت در پیشگیری، تشخیص زود هنگام، مدیریت جراحی و پیشرفت در شیمی درمانی، توانایی درمان بیماران سرطانی همچنان هدف مبهمی است و نیاز شدیدی به توسعه درمان‌های مؤثرتر دارد. بنابراین، جستجوی گسترده عوامل ضد سرطانی جدید با آشکار سازی اهداف مولکولی نوین به‌دنبال کشف ترکیباتی است که با چنین اهدافی تعامل دارند. از آنجا که طراحی دارو در آزمایشگاه زمان‌بر و پرهزینه بوده، طراحی محاسباتی دارو کمک شایانی، به صرفه‌جویی در زمان، هزینه‌های آزمایشگاهی و پژوهشی کرده است. از این رو نیاز به استفاده از روش‌های تئوری و محاسباتی بدون انجام آزمایش که بتواند ویژگی یا فعالیت ترکیبات را پیش‌بینی کند، ضروری به‌نظر می‌رسد. ظهور علم کمومتریکس توانسته راه حلی برای رفع این محدودیت‌ها باشد. بدین صورت که، با به‌دست آوردن رابطه کمی بین ساختار شیمیایی ترکیبات و فعالیت دارویی آن‌ها می‌توان فعالیت دارویی ترکیبات مشابه را پیش‌بینی کرد.

۱- ۱ کمومتریکس[1]

اصطلاح کمومتریکس اولین بار توسط شیمیدان آلی اسوانت ولد[2] سال ۱۹۷۲ در یک مجله سوئدی ذکر شد. ولد با همکاری بروس آر. کووالسکی[3] که در دانشگاه واشنگتن بر روی الگوشناسی در شیمی تجزیه مطالعه می‌کرد، انجمن بین المللی کمومتریکس[4] (ICS) را در سال ۱۹۷۴ تأسیس کردند [۱]. کمومتریکس، توسط ICS، به‌صورت زیر تعریف شده است:

"کمومتریکس استفاده از روش‌های ریاضی، آماری برای به‌وجود آوردن ارتباط بین سنجش‌های انجام شده روی یک سیستم یا فرآیند شیمیایی برای فهم بهتر اطلاعات شیمیایی است."

اطلاعات شیمیایی که اساس بسیاری از فرآیندهای مهم را تشکیل می‌دهد، به سه ویژگی مهم فرآیند اندازه‌گیری بستگی دارد، از جمله:

۱. خواص شیمیایی: شامل استوکیومتری، تعادل جرم، تعادل شیمیایی، سینتیک و ...

۲. خواص فیزیکی: شامل دما، انتقال انرژی، انتقال فاز و ...

۳. خواص آماری: شامل منابع خطا در فرآیند اندازه‌گیری، کنترل عوامل تداخل، کالیبراسیون سیگنال‌های پاسخ، مدل‌سازی سیگنال‌های پیچیده چند متغیره و ...

اگر یکی از این سه ویژگی کامل نباشد، سیستم اندازه‌گیری نمی‌تواند نتایج قابل اطمینان را ارائه دهد، که گاهی اوقات پیامدهای فاجعه باری به‌همراه خواهد داشت. نقش آمار و کمومتریکس در رسیدگی به سومین ویژگی، بحرانی است. این نقش اساسی، انگیزه اولیه را برای پیشرفت در زمینه کمومتریکس فراهم می‌کند [۲].

[1] Chemometrics
[2] Swante Wold
[3] Bruce R. Kowaski
[4] International Chemometrics Society

از بین کاربردهای مختلف کمومتریکس مهم‌ترین و شاخص‌ترین این کاربردها طراحی دارو است، که ارتباط کمی ساختار- فعالیت[1] (QSAR) را بررسی می‌کند و خواص مولکول‌ها را به ویژگی‌های ساختاری آن‌ها نسبت می‌دهد [۳]. در ادامه به شرح QSAR پرداخته شد.

۱- ۲ ارتباط کمی ساختار- فعالیت (QSAR)

۱- ۲- ۱ تعریف

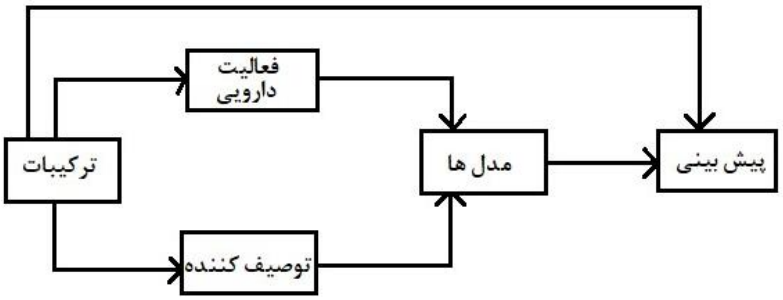
ارتباط کمی ساختار – فعالیت (QSAR)، در ساده‌ترین تعریف، روشی جهت ساختن مدل‌های ریاضی یا محاسباتی است که می‌تواند ارتباط معنا دار آماری را بین ساختار و عملکرد ایجاد کند. تلاش برای برقراری ارتباط بین ساختار شیمیایی و فعالیت زیستی موضوع QSAR است [۴].

۱- ۳ اصول مدل‌سازی QSAR

در مطالعات QSAR ارتباط ساختار مولکول‌ها با فعالیت دارویی آن‌ها طی مراحل زیر بیان می‌شود [۵]:

۱. جمع‌آوری داده‌ها
۲. رسم و بهینه‌سازی ساختار ترکیبات
۳. محاسبه و غربالگری توصیف کننده‌ها
۴. انتخاب مؤثرترین توصیف‌کننده‌ها
۵. ایجاد مدل‌های آماری
۶. ارزیابی مدل

[1] Quantitative structure- activity relationship



شکل ۱-۱ شمای کلی از مراحل QSAR

بنابراین همان‌طور که شکل ۱-۱ نشان می‌دهد، اساس رابطه کمی ساختار – فعالیت (QSAR) ایجاد رابطه بین فعالیت دارویی مشاهده شده و ویژگی‌های ساختاری یک مولکول است. با در نظر گرفتن مجموعه‌ای از مولکول‌ها، یک مدل پیش‌بینی ایجاد می‌شود که می‌تواند برای پیش‌بینی فعالیت دارویی سایر مولکول‌ها مورد استفاده قرار گیرد [۶].

۱- ۳- ۱ جمع‌آوری داده‌ها

اولین مرحله مدل‌سازی جمع آوری داده‌ها است که مقادیر تجربی فعالیت آنها باید با صحت قابل قبول اندازه‌گیری شده باشد. داده‌های تجربی می‌توانند در آزمایشگاه تولید شده یا از مقادیر منتشر در مقالات یا پایگاه داده‌ها استخراج شوند. در مدل‌سازی QSAR مجموعه داده‌ها با استفاده از آنالیزهای متفاوت تقسیم‌بندی از جمله کنارد-استون[1] (KS) انجام می‌شود [۷, ۸]. تقسیم‌بندی KS بر اساس محاسبه فاصله اقلیدسی[2] است. بنابراین مجموعه داده‌ها به دو دسته مجموعه آموزش[3] و مجموعه آزمون[4] تقسیم می‌شوند. الگوریتم KS داده‌ها را طوری تقسیم می‌کند که مجموعه‌های تقسیم‌بندی شده مانند مجموعه آزمون، نماینده مناسبی از مجموعه آموزش باشد. از این‌رو داده‌های مجموعه آموزش و آزمون به‌ترتیب برای ساخت

[1] Kennard-Stone (KS)
[2] Euclidean distance
[3] Training set
[4] Test set

مدل و جهت بررسی قدرت پیش‌بینی و اعتبار مدل استفاده می‌شود [۹]. لازم به ذکر است که مجموعه آزمون در هیچ یک از مراحل انتخاب متغیر و ساخت مدل‌های QSAR شرکت نخواهد داشت.

۱- ۳- ۲ رسم و بهینه سازی ساختار ترکیبات

بعد از جمع آوری مجموعه داده‌ها ساختار شیمیایی ترکیبات با استفاده از نرم افزار هایپرکم[۱] رسم و بهینه‌سازی می‌شود. ساختار ترکیبات به کمک روش‌های شیمی محاسباتی[۲] بهینه می‌شود تا مختصاتی از ساختار به‌دست آید که در آن، انرژی مولکول حداقل باشد. که این روش‌ها عبارتند از:

۱. روش‌های مکانیک کوانتومی[۳] (شامل روش‌های آغازین[۴]، روش‌های نیمه تجربی[۵]، روش‌های تابع چگالی[۶])

۲. روش مکانیک مولکولی[۷]

۳. روش‌های دینامیک مولکولی[۸]

در نرم افزارهایی نظیر هایپرکم اغلب از روش نیمه تجربی (AM1) استفاده می‌شود. بهینه‌سازی ساختار ترکیبات شیمیایی با استفاده از این روش انجام می‌شود و پایدارترین حالت با کم‌ترین انرژی برای هر ساختار به‌دست می‌آید. ساختارهای بهینه شده به‌عنوان ورودی نرم‌افزارهای محاسبه‌ای توصیف کننده-های مولکولی به‌کار گرفته می‌شوند.

[1] Hyperchem
[2] Computational Chemistry
[3] Quantum Mechanic
[4] Ab Initio Methods
[5] Semi-empirial Methods
[6] Density Functional Theory
[7] Molecular Mechanic
[8] Molecular dynamic

۱- ۳- ۳ محاسبه توصیف کننده‌ها

توصیف کننده‌ها به‌عنوان اعداد کمی، نشان دهنده ویژگی مواد شیمیایی هستند و به ایجاد ارتباط ریاضی کمک می‌کنند. بنابراین، انواع مختلف توصیف کننده نقش مهمی در شناسایی و همچنین تجزیه و تحلیل ترکیبات ایفا می‌کنند. اگرچه طیف گسترده‌ای از توصیف کننده‌ها برای استفاده در دسترس هستند، اما، هدف استفاده یا توسعه توصیف کننده‌هایی است که به‌راحتی محاسبه شوند و مقدار مشخصی از اطلاعات شیمیایی را ارائه دهند [۱۰].

۱- ۳- ۴ پیش پردازش[1]

بسیاری از برنامه‌ها قادر به تولید صدها یا هزاران توصیف کننده مولکولی مختلف هستند. به‌طوری که این برنامه‌ها می‌توانند برای تعداد کمی از ترکیبات (n)، هزاران توصیف کننده (p)، محاسبه کنند. به عبارتی $n < p$، که به‌عنوان داده‌های با ابعاد بالا[2] نامیده می‌شوند. ابعاد داده‌ها باید کاهش یابد تا علاوه بر این‌که مدل قدرت پیش‌بینی بالایی داشته باشد، قدرت تفسیر پذیری روشنی هم داشته باشد. مدل‌هایی که شامل مجموعه بزرگی از توصیف کننده‌های مولکولی هستند، تمایل به داشتن واریانس بزرگ دارند که اغلب منجر به عملکرد پیش‌بینی ضعیف می‌شود. بنابراین انتخاب تنها توصیف کننده‌های مولکولی مهم می‌تواند دقت پیش‌بینی را بهبود بخشد [۱۱]. از طرفی دیگر در داده‌هایی با ابعاد بالا مشکل هم‌خطی بین توصیف کننده‌ها وجود دارد، به‌طوری‌که وابستگی خطی منجر به ایجاد پدیده هم‌خطی شدید شده و صحت پیش‌بینی مدل را کم می‌کند. به‌همین دلیل بهتر است ابعاد داده‌ها کاهش یابد. بنابراین انتخاب توصیف کننده‌های مؤثری که بتواند ارتباط معنا دار بین متغیر وابسته و متغیرهای مستقل را نشان دهد، باعث بهبود صحت

[1] preprocessing
[2] High Dimensional Data

پیش‌بینی می‌شود. با توجه به این توضیحات استفاده از روش غربالگری[1] قبل از توسعه مدل حائز اهمیت است.

۱- ۳- ۴- ۱ پیش پردازش اولیه

به‌منظور پیدا کردن توصیف کننده‌های مؤثر، در مرحله‌ی اول، توصیف کننده‌هایی با مقادیر ثابت و نسبتاً ثابت (انحراف استاندارد کم‌تر از ۰/۰۰۱) [۱۲, ۱۳] از مجموعه داده‌ها حذف می‌شود. در مرحله بعد از بین دو توصیف کننده با همبستگی بالای ۰/۹ (R >۰/۹) آن توصیف کننده‌ای که ارتباط بیش‌تری با متغیر وابسته (فعالیت دارویی) دارد، نگه داشته می‌شود و توصیف کننده دیگر حذف می‌گردد تا مجموعه‌ای تنک با توصیف کننده‌های موثرتر توسعه یابد.

۱- ۳- ۴- ۲ استفاده از روش رگرسیون Ridge برای غربالگری توصیف کننده ها

در این پروژه به‌منظور غربالگری بیش‌تر توصیف کننده‌ها و حذف توصیف کننده‌های فاقد اطلاعات مفید، از روش رگرسیون انقباضی Ridge [۱۴] به‌عنوان یک روش غربالگری جدید استفاده گردید. روش رگرسیون Ridge با اعمال محدودیت مجموع مجذور پارامترها کم‌تر از یک مقدار ثابت بر روی تابع کم‌ترین توان‌های دوم موجب انقباض پارامترها و کاهش هم‌خطی و همچنین کاهش واریانس می‌شود و در نهایت دقت پیش‌بینی مطلوب‌تری نسبت به روش برآورد کم‌ترین توان‌های دوم ارائه می‌دهد. اما برآوردگر Ridge نیز در مواقعی که تعداد متغیرها بیش از تعداد ترکیبات است، همانند روش کم‌ترین توان‌های دوم قابل محاسبه نیست. در برآوردگر Ridge با تغییر کوچکی در داده‌ها تغییر زیادی در مدل نهایی صورت نمی‌پذیرد، بنابراین مدل پایداری را ارائه می‌دهد. اما رگرسیون Ridge هیچ یک از پارامترها را حذف نمی‌کند و به این دلیل برای مدل‌هایی با تعداد متغیرهای زیاد مدل ساده‌ای را نخواهیم داشت. با تلفیق یک روش جدید از این روش، برای غربالگری و حذف متغیرهایی با اهمیت کم‌تر استفاده شده است. این روش با استفاده از

[1] screening

تکنیک رد مرحله‌ای تک تک[1]روی توصیف کننده‌های باقی‌مانده اجرا می‌شود. به این منظور برای اجرای روش Ridge به‌عنوان یک روش غربالگری به‌ترتیب سه ماتریس ساخته شد. ماتریس زائد[2] (R) شامل توصیف کننده‌های با کم‌ترین اطلاعات مفید است که ارتباط معنا داری با پاسخ ندارند. ماتریس اصلی (X) متشکل از توصیف کننده‌هایی است، که ارتباط معنا داری با پاسخ دارند و ماتریس XR به‌ترتیب شامل توصیف کننده‌های ماتریس‌های X و R است.

۱- ۳- ۵ انتخاب مؤثرترین توصیف کننده ها و مدل سازی

انتخاب متغیر و مدل‌سازی به روش خطی و غیرخطی انجام می‌شود، که روش‌های خطی به روش‌های کلاسیک و انقباضی یا جریمه‌ای[3] تقسیم می‌شوند. از بین روش‌های کلاسیک می‌توان به روش‌های کمترین توان‌های دوم معمولی (OLS) که شامل انتخاب متغیر پیش رونده[4]، انتخاب متغیر حذفی پس رونده[5] و انتخاب متغیر گام به گام[6] اشاره کرد [۱۵]. هر چند این‌روش‌ها، زیرمجموعه مناسبی را از بین تمام زیرمجموعه‌های ممکن انتخاب می‌نمایند، اما زمانی که ابعاد داده‌ها بزرگ باشد تعداد زیر مجموعه‌های منتخب افزایش می‌یابد و با افزایش تعداد متغیرهای مدل، مقدار ضریب تعیین به‌طور کاذب بالا بوده و قدرت پیش‌بینی مدل پایین می‌آید. علاوه بر این، روش‌های انتخاب متغیر کلاسیک دارای مشکل بی‌ثباتی هستند. به‌طوری‌که با تغییری کوچک در داده‌ها، مدل‌های خیلی متفاوتی به‌وجود می‌آید. بنابراین استفاده از روش-های انتخاب متغیر خطی چندگانه به‌دلیل داشتن معایبی هم‌چون ناپایداری، بایاس بالا، ناکارآمدی در حضور هم‌خطی، برای انتخاب توصیف کننده‌های منتخب توصیه نمی‌شود [۱۶]. روش‌های مبتنی بر جریمه شامل

[1] Leave One Out
[2] Redundant
[3] Shrinkage and Penalized methods
[4] Forward
[5] Backward
[6] Stepwise

روش Ridge [۱۴]، روش انحراف قدر مطلق به‌طور هموار کوتاه شده[1] (SCAD) [۱۷] و ... می‌باشد. در ادامه به شرح روش‌های کلاسیک و روش‌های جریمه‌ای پرداخته شد.

۱- ۳- ۵- ۱ روش مدل‌سازی

۱- ۳- ۵- ۲ روش کم‌ترین توان‌های دوم معمولی

روش کم‌ترین توان‌های دوم معمولی (OLS) در سال ۱۸۰۹ توسط گوس[2] معرفی شد، که به‌طور رایج برای تخمین ضرایب رگرسیون استفاده می‌شود.

مدل رگرسیون خطی چندگانه[3] زیر را در نظر بگیرید رابطه ۱-۱ :

$$y = X\beta + \epsilon$$ رابطه ۱-۱

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 مؤلفه‌های خطای تصادفی می‌باشد.

تخمین کم‌ترین توان‌های دوم معمولی (OLS) با مینیمم کردن مجموع توان‌های دوم خطاهای مدل یعنی $\epsilon^T\epsilon = (y - X\beta)^T(y - X\beta)$ به‌وجود می‌آید. در این‌صورت برآوردگر OLS بردار پارامترها به‌صورت رابطه ۱-۲ محاسبه می‌شود:

$$\hat{\beta} = argmin_{\beta}\, \epsilon^T$$ رابطه ۱-۲

با توجه به این‌که

$$\frac{\partial \epsilon^T \epsilon}{\partial \beta} = -2X^T y + 2X^T X\beta = 0$$ رابطه ۱-۳

می‌توان نتیجه گرفت که :

$$\hat{\beta} = (X^T X)^{-1} X^T y$$ رابطه ۱-۴

[1] Smoothly Clipped Absolute Deviation
[2] Gauss
[3] Multiple Linear Regression (MLR)

برای بهبود مدل‌های خطی، یک پارامتر جریمه[1] به روش برآورد OLS اضافه شده که روش‌های رگرسیون‌های مبتنی بر جریمه[2] نامیده می‌شود، که در ادامه به‌شرح آن‌ها پرداخته شد.

۱- ۳- ۵- ۳ روش های رگرسیون مبتنی بر جریمه

روش‌های رگرسیون مبتنی بر جریمه، همه متغیرهای پیش‌بینی کننده را در مدل نگه می‌دارند، اما، با منقبض کردن ضرایب رگرسیون به سمت صفر و صفر کردن برخی از آن‌ها عمل برآوردیابی و انتخاب متغیر را انجام می‌دهند. این انقباض معمولاً از طریق ایجاد محدودیت روی دامنه تغییرات ضرایب در هنگام برآوردیابی صورت می‌گیرد. هدف از این کار وارد کردن اریبی در مدل با هدف کاهش واریانس و در نهایت کاهش مجموع مربعات خطا و بالا بردن دقت پیش‌بینی است. این روش‌ها باعث افزایش دقت پیش‌بینی، افزایش پایداری و بهبود تفسیر مدل‌های توسعه یافته می‌شوند.

۱- ۳- ۵- ۴ رگرسیون Ridge

برای اولین بار رگرسیون Ridge در سال ۱۹۷۰ توسط هورل و کنارد[3] معرفی شد. پایه و اساس روش برآوردگر OLS یک رگرسیون خطی، این است که $(\mathrm{x}^T\mathrm{x})^{-1}$ وجود داشته باشد و به دو دلیل این معکوس وجود ندارد: ۱. ماتریس طرح پر رتبه ستونی نباشد ۲. مشکل هم‌خطی وجود داشته باشد. یکی از روش‌های حل این مشکل استفاده از روش رگرسیون Ridge می‌باشد.

راهی آسان برای تضمین معکوس پذیری، اضافه کردن ماتریس قطری $\lambda\mathrm{I}$ به $\mathrm{x}^T\mathrm{x}$ است که در آن λ یک مقدار ثابت مثبت و I ماتریس همانی می‌باشد. ازاین‌رو برآوردگر رگرسیون Ridge پارامتر β به‌صورت رابطه ۱-۵ است:

$$\hat{\beta} = (\mathrm{x}^{\mathrm{T}}\mathrm{x} + \lambda\mathrm{I})^{-1}\mathrm{x}^{\mathrm{T}}$$ رابطه ۱-۵

[1] Penalty
[2] penalized regression methods
[3] Hoerl and Kennard

که $\lambda > 0$ است. همان‌گونه که با مینمم کردن $\epsilon^T\epsilon$ نسبت به β، برآوردگر OLS به‌دست آمد، این برآوردگر را نیز می‌توان با مینیمم کردن عبارت مجموع توان‌های دوم خطا $\epsilon^T\epsilon$ نسبت به β تحت شرط $\beta^T\beta = \sum_{j=1}^{p} {\beta_j}^2 \leq t$ محاسبه کرد. در این‌جا $t \geq 0$ یک پارامتر تنظیم کننده است که میزان انقباض ضرایب را کنترل می‌کند. به بیانی هر چه مقدار t کوچک‌تر باشد ضرایب بزرگ‌تر بیش‌تر جریمه می‌شوند. برای حل مسئله بهینه‌سازی بالا از ضرایب لاگرانژ می‌توان استفاده کرد. به بیانی دیگر برآوردگر Ridge به‌صورت رابطه ۱-۶ به‌دست می‌آید:

$$\hat{\beta}_R = argmin_{\beta}\ (\epsilon^T\epsilon + \lambda\beta^T\beta) \qquad \text{رابطه ۱-۶}$$

که در آن λ ضریب لاگرانژ یا همان پارامتر انقباض Ridge می‌باشد. از آنجایی که

$$\frac{\partial(\epsilon^T\epsilon+\lambda\beta^T\beta)}{\partial\beta} = -2x^Ty + 2x^Tx\hat{\beta} + 2\lambda\beta = 0 \qquad \text{رابطه ۱-۷}$$

داریم:

$$(\mathrm{x^Tx} + \lambda\mathrm{I}_p)\beta = \mathrm{x^Ty} \qquad \text{رابطه ۱-۸}$$

و بنابراین

$$\hat{\beta}_R = (\mathrm{x^Tx} + \lambda\mathrm{I}_p)^{-1}\mathrm{x^Ty} \qquad \text{رابطه ۱-۹}$$

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

$$\mathrm{E}(\hat{\beta}_\mathrm{R}) = \beta + \lambda(\mathrm{X^TX} + \lambda\mathrm{I})^{-1}\beta \qquad \text{رابطه ۱-۱۰}$$

$$\mathrm{Var}(\hat{\beta}_\mathrm{R}) = (\mathrm{X^TX} + \lambda\mathrm{I})^{-1}\mathrm{X^TX}(\mathrm{X^TX} + \lambda\mathrm{I})^{-1}\sigma^2 \qquad \text{رابطه ۱-۱۱}$$

در صورتی‌که در خصوص برآوردگر OLS داریم:

$$\in \hat{\beta} = \beta \qquad \text{رابطه ۱-۱۲}$$

$$\mathrm{Var}(\hat{\beta}) = (\mathrm{X^TX})^{-1}\sigma^2 \qquad \text{رابطه ۱-۱۳}$$

هورل و کنارد اثبات کرده‌اند که اگر $\beta^T\beta$ کراندار باشد می‌توان $\lambda > 0$-یی را پیدا کرد به‌گونه‌ای‌که:

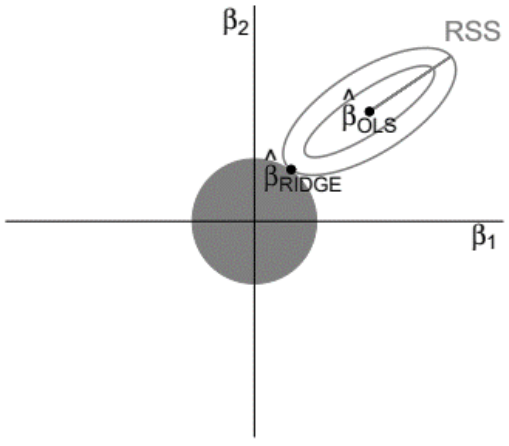
$$\mathrm{MSE}(\hat{\beta}_\mathrm{R}) < \mathrm{MSE}(\hat{\beta}) \qquad \text{رابطه ۱-۱۴}$$

بنابراین رگرسیون Ridge می‌تواند برآورد OLS را بهبود ببخشد [۱۴].

برای یافتن درک بصری روش Ridge، با فرض وجود تنها دو متغیر در مدل رگرسیونی ناحیه محدودیت به‌صورت رابطه ۱-۱۵ :

$$\beta_1^2 + \beta_2^2 \leq t$$ رابطه ۱-۱۵

همان‌طور که در شکل ۱-۲ نشان داده شده دایره ناحیه تاوان است. محل برخورد مجموع توان‌های دوم باقی‌مانده‌ها[1] (RSS) روش کم‌ترین توان‌های دوم با دایره، مختصات برآوردگر Ridge را نمایش می‌دهد. با عمود کردن این نقطه بر محور افقی، $\hat{\beta}_1$ و با عمود کردن آن بر محور عمودی، $\hat{\beta}_2$ حاصل می‌شود. روش Ridge فرآیندی پیوسته است که ضرایب تولیدی آن پایدار هستند اما برآوردهایی اریب ارائه می‌دهد که پذیرفتن مقداری اریبی در مقابل یافتن پاسخی قابل اطمینان‌تر معقول به‌نظر می‌رسد. روش Ridge ضرایب را به سمت صفر منقبض می‌کند، هرچه مقدار پارامتر Ridge بیش‌تر باشد ضرایب برآورد شده انقباض بیش‌تری داشته و کوچک‌تر می‌شوند ولی هیچ‌گاه حتی در بی‌نهایت نیز به صفر نمی‌رسند این امر به سبب خاصیت دایره (ناحیه تاوان روش Ridge) رخ می‌دهد. روش Ridge یک روش مدل‌سازی است، به‌دلیل این‌که همه متغیرها را در مدل نگه می‌دارد برای انتخاب متغیر استفاده نمی‌شود [۱۶].



شکل ۱-۲ عملکرد رگرسیون Ridge

[1] Residual Sum of Squares

۱- ۳- ۵- ۵ روش SCAD

در سال ۲۰۰۱، فن و لی[1] تابع جریمه انحراف قدر مطلق به‌طور هموار کوتاه شده (SCAD) را به معادله OLS اضافه کردند که ضرایب رگرسیون را تخمین می‌زند:

رابطه ۱-۱۶ $$\hat{\beta}^{ols} = \arg \min_{\beta \in \mathbb{R}^{p+1}} \|Y - X\beta\|^2 = (X^T X)^{-1} X^T$$

رابطه ۱-۱۷ $$\hat{\beta}^{SCAD} = \arg \min_{\beta \in \mathbb{R}^{p+1}} \{\|Y - X\beta\|^2 + \lambda \sum_{i=1}^{p} p(|\beta_i|); \alpha, \lambda\}$$

که $p(.;\alpha,\lambda)$ تابع جریمه SCAD هست و سه آرگومان زیر را دنبال می‌کند:

رابطه ۱-۱۸ $$P(t;\alpha,\lambda) = \begin{cases} \lambda t & 0 \le t \le \lambda \\ \frac{2\alpha\lambda t - t^2 - \lambda^2}{2(\alpha-1)} & \lambda \le t < \alpha\lambda \\ \frac{(\alpha+1)\lambda^2}{2} & t \ge \alpha\lambda \end{cases}$$

جایی که $\alpha > ۲$ و $\lambda > ۰$ پارامترهای جریمه هستند و بزرگی ضرایب را کنترل می‌کنند. فن و لی مقدار بهینه α را به‌طور تجربی ۳/۷ بیان کردند. با یک مطالعه نشان دادند که مقدار ۳/۷ برای α می‌تواند نتایج خوبی داشته باشد.

فن و لی ویژگی‌های یک تابع تاوان خوب را به‌صورت زیر بیان کردند:

۱. نا اریبی: برآوردگر به‌دست آمده برای ضرایب رگرسیونی که در واقع مقادیر آن‌ها بزرگ است، تقریباً نا اریب باشد. این ویژگی اریبی مدل را کاهش می‌دهد.
۲. تنکی[2]: برآوردگر نتیجه شده باید به‌طور خودکار ضرایب برآورد شده‌ای که مقدار کوچکی دارند را برابر صفر قرار دهد تا متغیرهای مناسب را انتخاب کند. این کار پیچیدگی مدل را کاهش می‌دهد.
۳. پیوستگی: برآوردگر نتیجه شده دارای تغییرات پیوسته باشد و ناپایداری در پیش‌بینی مدل را کاهش دهد [۱۷].

[1] Fan & Li
[2] Sparsity

رگرسیون Ridge از ویژگی تنکی بی‌بهره است، زیرا توانایی خارج کردن ضرایب بی اهمیت از مدل را ندارد. تابع جریمه SCAD، دارای این سه ویژگی می‌باشد. در این تحقیق از روش SCAD برای انتخاب متغیر و مدل‌سازی استفاده شده است.

۱- ۴ ارزیابی مدل

از مرحله‌های اساسی ساخت و توسعه مدل QSAR، ارزیابی مدل است. در این پروژه، قدرت پیش‌بینی و تعمیم‌پذیری مدل با استفاده از پیش‌بینی فعالیت دارویی داده‌ها مجموعه آزمون، مجموعه پیش‌بینی شده به‌وسیله تکنیک رد مرحله‌ای تک تک (LOO)، محاسبه پارامترهای آماری، آزمون‌های دامنه کاربرد[۱] و Y-تصادفی[۲] مورد ارزیابی قرار گرفت.

۱- ۴- ۱ محاسبه پارامترهای آماری

بعد از اینکه مدل ساخته شد توسط یک مجموعه پارامتر آماری ارزیابی می‌شود تا از توانایی پیش‌بینی مدل برای نمونه‌های مشابه اطمینان حاصل شود. فرض کنید $x_1, x_2, \dots, x_n$ نشان‌دهنده متغیرهای مستقل (پیشگو) و $y_1, y_2, \dots, y_n$ نشان‌دهنده متغیرهای وابسته در مدل مورد بررسی باشند. برخی از این پارامترها عبارتند از :

ضریب همبستگی[۳]: یکی از معیارهای مورد استفاده در تعیین همبستگی دو متغیر است. این ضریب بین ۱ تا ۱- است. مقدار بزرگ‌تر آن نشان‌دهنده این است که ارتباط خطی بیش‌تری میان متغیر وابسته و متغیرهای مستقل وجود دارد، و در صورت عدم وجود رابطه بین دو متغیر، برابر صفر است. همبستگی بین دو متغیر x,y به‌صورت رابطه ۱-۱۹ تعریف می‌شود:

رابطه ۱-۱۹

$$R = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2\,\Sigma(y_i-\bar{y})^2}}$$

[1] Applicability Domain
[2] Y- Randomization
[3] Correlation coefficient

ضریب تعیین[1]: معیاری برای بیان دقت خط رگرسیون برآورد شده، می‌باشد. نسبتی از واریانس برحسب متغیر وابسته است که از متغیرهای مستقل قابل پیش‌بینی باشد. این ضریب به‌صورت رابطه ۱-۲۰ تعریف می‌شود:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$ رابطه ۱-۲۰

اگر R^2= ۱ باشد آن‌گاه SSR=SST یا به عبارتی SSE = ۰ یعنی زمانی که از متغیرهای مستقل استفاده کنید هیچ خطای وجود ندارد که این بهترین حالت ممکن است. اگر R^2= ۰ باشد آن‌گاه SSR=۰ یا به عبارتی SSE = SSR یعنی استفاده از متغیرهای مستقل هیچ تأثیری بر برآورد خط رگرسیونی ندارد.

- مجموع توان‌های دوم رگرسیون[2] (SSR): بیانگر مجموع توان‌های دوم انحراف مقادیر پیش‌بینی شده متغیر وابسته از میانگین مقادیر آن است. طبق رابطه ۱-۲۱ است:

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$ رابطه ۱-۲۱

- مجموع توان‌های دوم کامل[3] (SST): طبق رابطه ۱-۲۲ 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 ۱-۲۲

- مجموع توان‌های دوم خطا[4] (SSE): مجموع توان‌های دوم انحراف مقادیر واقعی متغیر وابسته از مقادیر پیش‌بینی شده برای آن است رابطه ۱-۲۳ :

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$ رابطه ۱-۲۳

بنابراین با توجه به روابط فوق ضریب تعیین را می‌توان به‌صورت رابطه ۱-۲۴ تعریف کرد:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$ رابطه ۱-۲۴

[1] Determination coefficient
[2] Sum Square Regression
[3] Sum Square Total
[4] Sum Square Error

طبق رابطه ۱-۲۴ اگر به ازای تمام نقاط $y_i = \hat{y}_i$ باشد، یعنی تمام مشاهدات بر روی خط برازش شده قرار گرفته باشند، 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[1]: از آن‌جا که R^2 (ضریب تعیین)، تعداد پارامترهای موجود در مدل را به حساب نمی‌آورد، در ارزیابی و مقایسه مدل‌های مختلف با تعداد متفاوت متغیرهای مستقل ضریب تعیین تصحیح شده استفاده می‌گردد، که طبق رابطه ۱-۲۵ بیان می‌شود:

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1} \cdot \frac{SSE}{SST} = 1 - (1 - R^2)$$ رابطه ۱-۲۵

در این رابطه p نشان‌دهنده تعداد متغیرهای مستقل و n نشان‌دهنده تعداد ترکیبات مورد بررسی می‌باشد.

FIT: به‌عنوان یک پارامتر آماری برای مقایسه مدل‌های مختلف با تعداد متفاوت متغیرهای مستقل به‌کار می‌رود که هر چه مقدار این پارامتر برای یک مدل بیش‌تر باشد، آن مدل مناسب‌تر می‌باشد. مقدار آن از رابطه ۱-۲۶ به‌دست می‌آید [۱۸, ۱۹].

$$FIT = \frac{R^2(n-p-1)}{(n+p^2)(1-R^2)}$$ رابطه ۱-۲۶

که R^2 مجذور ضریب همبستگی مدل، p تعداد توصیف کننده‌های مدل و n تعداد ترکیبات مربوط به مدل است.

مجموع توان دوم باقی‌مانده‌ها[2] **(PRESS)**: مجموع توان دوم تفاوت بین مقدار کمیت مشاهده شده (y_i) و مقدار برآورد شده $(\hat{y}_i)$ می‌باشد رابطه ۱-۲۷ :

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ رابطه ۱-۲۷

[1] Adjusted determination coefficient
[2] Predictive Residual Sum of Squares

خطای استاندارد پیش‌بینی[1] (SEP): طبق رابطه ۱-۲۸ بیان می‌شود:

$$SEP = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}}$$ رابطه ۱-۲۸

خطای مطلق میانگین[2] (MAE): طبق رابطه ۱-۲۹ بیان می‌شود:

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}|}{n}$$ رابطه ۱-۲۹

خطای نسبی پیش‌بینی[3] (REP): طبق رابطه ۱-۳۰ بیان می‌شود:

$$REP(\%) = \frac{100}{\overline{y}} \times \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}}$$ رابطه ۱-۳۰

میانگین توان‌های دوم خطا[4] (MSE): طبق رابطه ۱-۳۱ بیان می‌شود:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}$$ رابطه ۱-۳۱

میانگین خطای نسبی[5] (MRE): طبق رابطه ۱-۳۲ بیان می‌شود:

$$MRE = \frac{\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}}{y_i}\right|}{n} \times 100$$ رابطه ۱-۳۲

تلرانس[6]: همان‌گونه که می‌دانیم وجود یا عدم وجود هم‌خطی در بین متغیرهای پیش‌بینی کننده بسیار مهم 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 پارامتری محاسبه می‌شود که به آن تلرانس گفته می‌شود. برای هر متغیر x_i تلرانس از رابطه ۱-۳۳ به‌دست می‌آید [۲۰, ۲۱]:

[1] Standard Error of Prediction
[2] Mean Absolute Error
[3] Relative Error of Prediction
[4] Mean Square Error
[5] Mean Relative Error
[6] Tolerance

$$T_i = 1 - R_i^2 \qquad i = 1,2,3,\dots,p$$ رابطه ۱-۳۳

که در آن R_i^2 مجذور ضریب همبستگی چندگانه و p تعداد متغیرهای پیشگو است که از رگرسیون x_i بر تمام متغیرهای پیش‌بینی کننده دیگر به‌دست می‌آید. در نظر گرفته شده است که اگر برای هر یک از متغیرهای مستقل، T_i از ۰/۱ بیش‌تر باشد، پدیده هم‌خطی وجود نخواهد داشت.

عامل تورم واریانس[1]: در برخی مواقع به‌جای تلرانس از عامل افزایش واریانس استفاده می‌شود که رابطه آن با تلرانس به‌صورت رابطه ۱-۳۴ است:

$$VIF = \frac{1}{T_i} = \frac{1}{1-R_i^2} \qquad i = 1,2,3,\dots,p$$ رابطه ۱-۳۴

بدیهی 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 تا ۱۰ قابل قبول است و مقادیر بیش‌تر از ۱۰ اکثراً به‌عنوان علامتی از این‌که داده‌ها مشکل هم‌خطی دارند، تلقی می‌شود. اگر هیچ‌گونه رابطه خطی بین متغیرهای پیش‌بینی کننده وجود نداشته باشد، آنگاه R_i^2 صفر و VIF_i برابر یک می‌شود [۲۲-۲۵].

$\boldsymbol{R_0^2}$: پارامتر ضریب تعیین مدل با عرض از مبدأ صفر است، زمانی که نمودار فعالیت پیش‌بینی شده برحسب فعالیت تجربی رسم شده است. در صورتی که مقدار این پارامتر نزدیک به مقدار R^2 باشد، قابل قبول است [۲۶].

$\boldsymbol{R'^2_0}$: ضریب تعیین مدل با عرض از مبدأ صفر است، زمانی که نمودار فعالیت تجربی بر حسب فعالیت پیش‌بینی رسم شده است. در صورتی که مقدار این پارامتر نزدیک به مقدار R^2 باشد، قابل قبول است [۲۶].

[1] Variance Inflation Factor (VIF)

$\boldsymbol{R_m^2}$: این پارامتر اختلاف بین R^2 و R_0^2 را در مدل محدود می‌کند، به‌طوری که مدل با R_m^2 بیش‌تر از ۰/۵ قابل قبول بوده و مدل توسعه یافته از قدرت پیش‌بینی بالاتری برخوردار است [۲۷, ۲۸] و از رابطه ۱-۳۵ محاسبه می‌شود:

$$R_m^2 = R^2 \times [1 - ((R^2 - R_0^2))^{1/2}]$$ رابطه ۱-۳۵

$\boldsymbol{R'^2_m}$: این پارامتر اختلاف بین R^2 و R'^2_0 را در مدل محدود می‌کند، به‌طوری که مدل با R'^2_m بیش‌تر از ۰/۵ قابل قبول بوده و مدل از قدرت پیش‌بینی بالاتری برخوردار است [۲۷, ۲۸] و از رابطه ۱-۳۶ محاسبه می‌شود:

$$R'^2_m = R^2 \times [1 - ((R^2 - R'^2_0))^{1/2}]$$ رابطه ۱-۳۶

$\mathbf{R - R}$: پارامتر حاصل از تفاضل دو پارامتر R'^2_m و R_m^2 است. محدوده قابل قبول این پارامتر کم‌تر از ۰/۲ تعریف شده است و طبق رابطه ۱-۳۷ محاسبه می‌شود [۲۹]:

$$R - R = |R_m^2 - R'^2_m|$$ رابطه ۱-۳۷

$\boldsymbol{R^2_{0,R}}$: این پارامتر نشان‌دهنده نزدیکی دو پارامتر R^2 و R_0^2 می‌باشد. هر چه این مقادیر به‌هم نزدیک‌تر باشند و نسبت تفاضل آن‌ها به R^2 کوچک‌تر از ۰/۱ باشد، مدل دارای قدرت پیش‌بینی بالاتری است [۲۷] از رابطه ۱-۳۸ محاسبه می‌شود:

$$R_{0,R}^2 = \frac{(R^2 - R_0^2)}{R^2}$$ رابطه ۱-۳۸

$\boldsymbol{R'^2_{0,R}}$: این پارامتر نشان‌دهنده نزدیکی دو پارامتر R^2 ، R'^2_0 می‌باشد. هر چه این مقادیر به‌هم نزدیک‌تر باشند و نسبت تفاضل آن‌ها به R^2 کوچک‌تر از ۰/۱ باشد، مدل دارای قدرت پیش‌بینی بالاتری است [۲۷] از رابطه ۱-۳۹ محاسبه می‌شود:

$$R'^2_{0.R} = \frac{(R^2 - R'^2_0)}{R^2}$$ رابطه ۱-۳۹

K: شیب معادله رگرسیون برای نمودار فعالیت پیش‌بینی شده برحسب فعالیت تجربی با عرض از مبدأ صفر می‌باشد. پارامتر K باید در محدوده ۰/۸۵ تا ۱/۱۵ باشد تا مدل توسعه یافته از نظر آماری مورد تأیید واقع گردد [۲۷].

K´: شیب معادله رگرسیون برای نمودار مقادیر تجربی برحسب مقادیر پیش‌بینی شده با عرض از مبدأ صفر می‌باشد و K´ باید در محدوده ۰/۸۵ تا ۱/۱۵ باشد تا مدل قدرت پیش‌بینی بالایی داشته باشد [۲۷].

۱- ۴- ۲ ارزیابی مدل با استفاده از نتایج پیش بینی شده برای مجموعه آزمون

برای ارزیابی مدل توسعه یافته، از پیش‌بینی داده‌های مجموعه آزمون که در هیچ یک از مراحل مدل‌سازی حضور نداشته‌اند، استفاده می‌شود. به‌طوری‌که ترکیبات مجموعه آزمون با استفاده از مدل توسعه یافته، پیش‌بینی می‌شود. مقادیر خطای حاصل از پیش‌بینی و نتایج پارامترهای آماری قابل قبول، صحت و دقت مدل ساخته شده را اثبات می‌کند.

۱- ۴- ۳ ارزیابی مدل با استفاده از تکنیک رد مرحه ای تک تک (LOO)

تکنیک LOO نیز برای بررسی قدرت پیش‌بینی مدل برای کل مجموعه داده‌ها به‌کار گرفته می‌شود. به‌طوری‌که از مدل توسعه یافته برای پیش‌بینی همه داده‌ها، مورد استفاده قرار می‌گیرد. با این تفاوت که در این تکنیک هر داده مورد مطالعه یک‌بار به‌عنوان داده آزمون خارج شده و مدل با مابقی داده‌ها آموزش داده می‌شود و پیش‌بینی پاسخ مربوط به داده خارج شده با استفاده از مدل به‌دست می‌آید. این عملیات به تعداد کل داده‌ها تکرار شده تا پاسخ همه داده‌ها یک‌بار پیش‌بینی شود. با محاسبه پارامترهای آماری برای نتایج حاصل از تکنیک رد مرحله‌ای تک تک قدرت پیش‌بینی و تعمیم پذیری مدل توسعه یافته تأیید می‌شود.

۱- ۴- ۴ استفاده از نمودار خطای باقی مانده ها

عبارت خطای باقی‌مانده، اختلاف بین مقادیر پیش‌بینی شده و مقادیر تجربی می‌باشد، که اگر پراکندگی مقادیر در دو طرف نمودار صفر باشد، این مسئله نشان‌دهنده تصادفی بودن خطا است. اما اگر عمده نقاط در این نمودار، در یک طرف خط صفر باشند، به این معنی است که خطای جهت‌داری اتفاق افتاده است.

۱- ۴- ۵ استفاده از دامنه کاربرد

دامنه کاربرد فضای ویژگی‌های مولکولی را که مدل بر روی آن‌ها آموزش دیده است و باید در مورد آنها استفاده شود، تعیین می‌کند [۳۰]. یک منطقه تئوری در فضای شیمیایی است که به‌وسیله توصیف کننده‌های نهایی و پاسخ مدل تعریف می‌شود. توسط این روش بررسی این‌که آیا یک ماده شیمیایی درون دامنه ساختاری مدل قرار می‌گیرد یا خارج آن امکان می‌یابد. اگر ماده شیمیایی خارج دامنه ساختاری مدل قرار گیرد عدم اطمینان افزایش یافته و مقدار پیش‌بینی شده غیر واقعی است. مقیاس دوری یک ترکیب از دامنه کاربرد مدل با درجه نفوذ[1] (H) آن سنجیده می‌شود که به‌صورت رابطه ۱-۴۰ محاسبه می‌گردد:

$$H_i = x_i\,(X^T X)^{-1} x_i \qquad (i = 1, \dots, n)$$ رابطه ۱-۴۰

در این رابطه H_i درجه نفوذ مربوط به ترکیب i در فضای توصیف کننده‌ها، x_i بردار ردیفی توصیف کننده ترکیب i و X ماتریس توصیف کننده‌ها می‌باشد. بالاوند T نیز به ترانهاده[2] بردار و ماتریس اشاره می‌کند. مقدار بحرانی $H > h^*$ است که $h^* = 3p/n$ و p تعداد توصیف کننده‌های مدل به‌علاوه یک و n تعداد ترکیبات در مجموعه آموزش است. وقتی H یک ترکیب بیش‌تر از مقدار بحرانی است یعنی آن ترکیب از نظر ساختاری دور از داده‌های مجموعه آموزشی است و خارج دامنه کاربرد مدل در نظر گرفته می‌شود.

[1] Leverage
[2] transpose

نمودار ویلیام[1] برای تشخیص ساده گرافیکی مقادیر پیش‌بینی شده پرت (باقی‌مانده‌های استاندارد شده آن‌ها بیش‌تر از ۳± می‌باشد) و همین‌طور ترکیباتی که از لحاظ ساختاری بر مدل تأثیر گذارند (H > h*) به‌کار می‌رود. این نمودار از رسم مقادیر باقی‌مانده‌های استاندارد شده در برابر مقادیر H تمامی ترکیبات به‌دست می‌آید. مقادیر باقی‌مانده‌های استاندارد شده نیز با استفاده از رابطه ۱-۴۱ محاسبه می‌شود:

$$r_i = \frac{e_i}{s_{e_i}} = \frac{(y_i - \hat{y}_i)}{s_{e_i}}$$ رابطه ۱-۴۱

که در آن y_i و $\hat{y}_i$ به‌ترتیب مقادیر واقعی و پیش‌بینی پاسخ هر مشاهده است (i = ۱، ...، n)، e_i تفاوت بین پاسخ‌های واقعی و پیش‌بینی شده برای هر مشاهده است و s_{e_i} انحراف استاندارد مقادیر باقی‌مانده است.

۱- ۴- ۶ استفاده از آزمون Y-تصادفی

برای اثبات قدرت مدل ایجاد شده و این‌که نتایج به‌دست آمده تصادفی نبوده است از آزمون Y-تصادفی استفاده می‌شود. بدین منظور $pIC_{۵۰}$ (فعالیت دارویی به‌عنوان متغیر پاسخ) به‌صورت تصادفی در محدوده داده‌های متغیر پاسخ ایجاد شد.

۱- ۴- ۷ بررسی نحوه تأثیر متغیرهای مستقل بر متغیر وابسته

از آن‌جایی که واحدهای اندازه‌گیری متغیرهای مستقل (ضرایب رگرسیون) یکسان نمی‌باشد، میزان تأثیر و اهمیت یک متغیر مستقل بر روی متغیر وابسته، با استفاده از ضرایب رگرسیون غیر استاندارد مشخص نمی‌شود. برای حل این مسئله، از ضریب استاندارد شده که طبق رابطه ۱-۴۲ محاسبه می‌شود، استفاده گردید [۳۱].

$$\beta'_k = (\frac{S_K}{S_Y})\beta_K$$ رابطه ۱-۴۲

[1] William's Plot

در این رابطه β'_k ضریب رگرسیون استاندارد شده متغیر مستقل مورد نظر (توصیف‌کننده k ام)، S_K انحراف استاندارد توصیف‌کننده k ام، S_Y انحراف استاندارد متغیر وابسته (فعالیت دارویی) و β_K ضریب رگرسیون غیراستاندارد همان توصیف‌کننده است، که به‌صورت رابطه ۱-۴۳ و رابطه ۱-۴۴ زیر بیان می‌شود:

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

رابطه ۱-۴۳

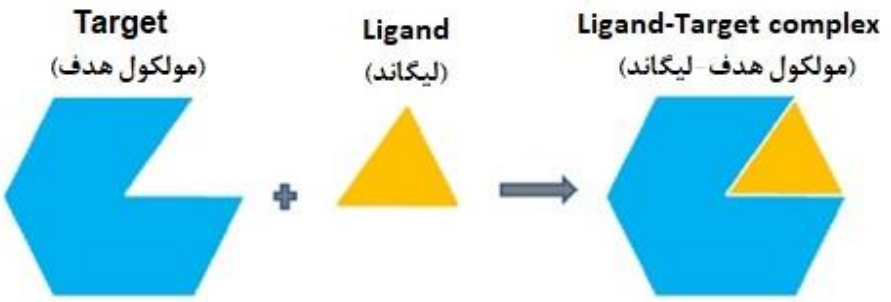
$$S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

رابطه ۱-۴۴

x_{ik} مقدار متغیر مستقل (توصیف‌کننده) k ام برای ترکیب i ام و $\bar{x}_k$ مقدار میانگین توصیف کننده، n تعداد ترکیبات، Y_i مقدار متغیر وابسته برای ترکیب i ام و$\bar{Y}$ مقدار میانگین متغیر وابسته می‌باشد.

۱- ۵ شبیه سازی داکینگ مولکولی

داکینگ مولکولی یک روش شبیه‌سازی رایانه‌ای است که به‌طور گسترده‌ای برای پیش‌بینی ترکیب یک مجموعه گیرنده-لیگاند استفاده می‌شود، جایی که گیرنده معمولاً یک پروتئین است و لیگاند یک مولکول کوچک است [۳۲]. داکینگ مولکولی یکی از ابزارهای کلیدی در زیست شناسی مولکولی ساختاری و طراحی دارو است. هدف از داکینگ پروتئین-لیگاند، پیش‌بینی جایگاه اتصال مناسب لیگاند به پروتئینی با ساختار سه بعدی شناخته شده است [۳۳]. در شکل ۱-۳ ایده روش داکینگ مولکولی به‌صورت اجمالی نشان داده شده است. در این فرآیند لیگاند در حفرات مولکول هدف (پروتئین هدف) قرار می‌گیرد. بنابراین، محاسبه اتصال به‌عنوان یک رویکرد مهم برای درک برهم‌کنش پروتئین-لیگاند در نظر گرفته می‌شود [۳۴, ۳۵].



شکل ۱-۳ اتصال لیگاند به مولکول هدف

داکینگ مولکولی به دو روش انجام می‌شود:

۱. داکینگ لیگاند درون پروتئین با پروتئین برای اعتبار سنجی فرآیند داکینگ مولکولی

۲. داکینگ لیگاندهای همانند با لیگاند درون پروتئین در شرایط مشخص شده توسط فرآیند اعتبارسنجی برای ایجاد مدل‌سازی

۱- ۵- ۱ فرآیند اعتبار سنجی

با توجه به‌این‌که ابزار و روش‌های متفاوتی برای داکینگ مولکولی وجود دارد و همچنین در نظر گرفتن این نکته که برخی از این روش‌ها نتایج بهتری برای بعضی دسته‌بندی مولکول‌های هدف ارائه می‌دهند باید در انتخاب روش مناسب داکینگ مولکولی دقت کرد. به بیان دیگر بایستی روش‌های متفاوت را به‌کار گرفت و در نهایت روشی که بهترین نتیجه را برای مولکول هدف مورد نظر ارائه می‌دهد انتخاب نمود. یکی از معیارهای کمک کننده به انتخاب روش مناسب، اندازه گیری انحراف ریشه میانگین مربع[1] (RMSD) اتم‌های لیگاند موجود در ساختار کریستالوگرافی و داک شده است [۳۲].

موفقیت الگوریتم‌های داکینگ در پیش‌بینی حالت اتصال لیگاند به‌طور معمول بر اساس انحراف ریشه میانگین مربع بین موقعیت‌های اتم سنگین لیگاندها و حالت پیش‌بینی شده توسط الگوریتم، اندازه‌گیری

[1] Root Mean Square Deviation

می‌شود. انعطاف پذیری سیستم یک چالش عمده در جستجوی حالت مناسب است [۳۶]. معمولاً وقتی RMSD کمتر از ۲ آنگستروم باشد عملکرد خوب در نظر گرفته می‌شود.

۱- ۵- ۲ آماده‌سازی پروتئین

ساختار کریستالوگرافی پروتئین‌ها از بانک اطلاعات پروتئین[1] استخراج می‌شود. به ساختار همراه با لیگاند کمپلکس شده با پروتئین، ساختار "هولو[2]" و ساختار بدون لیگاند، کوفاکتور و فقط دارای بخش پروتئینی در جایگاه فعال، ساختار "آپو[3]" نامیده می‌شود [۳۷]. ساختارهای هولو بهترین روش شناسایی جایگاه فعال پروتئین می‌باشند. مرکز ثقل[4] لیگاندی که در پروتئین وجود دارد همان مرکز تقریبی جایگاه فعال پروتئین است. درستی نتایج داکینگ مستقیماً به کیفیت ساختار کریستالوگرافی جایگاه فعال پروتئین بستگی دارد. ازآن‌جایی که حتی بهترین ساختارهای کریستالوگرافی اغلب قدرت تفکیک ۱ آنگستروم یا بیش‌تر دارند، مولکول‌های آب در ساختار را حذف کرده و سپس بعد از ورود به برنامه، هیدروژن‌هایی را که به‌وسیله کریستالوگرافی پروتئین دیده نمی‌شوند، اضافه نموده و با فرمت pdb ذخیره می‌شود [۳۸].

۱- ۵- ۳ ساختن لیگاند

بعد از این‌که، لیگاند در برنامه داکینگ فراخوانی می‌شود، در جایگاه فعال پروتئین هدف قرار می‌گیرد و بعد برای همه لیگاندها، شبیه‌سازی داکینگ به‌طور جداگانه توسط کاربر انجام می‌شود.

۱- ۵- ۴ تنظیم کردن جعبه شبکه‌ای[5]

در داکینگ مولکولی نمی‌توان کل فضا را به‌عنوان جایگاه فعال گیرنده در نظر گرفت. باید بخشی از فضا که لیگاند درون جایگاه فعال قرار می‌گیرد را برای فرآیند داکینگ تعریف کرد تا هم دقت بالا باشد و

[1] Protein Data Bank
[2] Soaked structure
[3] Apo structure
[4] Gravity center
[5] Grid box

هم‌زمان انجام فرآیند داکینگ به‌صرفه باشد. به‌جهت سرعت بخشیدن به محاسبات، یک فاصله محدود کننده[1] تنظیم می‌شود، این فاصله محدود کننده معمولاً یک جعبه مستطیل شکل به‌نام جعبه شبکه‌ای است [۳۸]. بخش‌هایی از پروتئین که دور از جایگاه فعال هستند، به‌طور معمول هیچ اثر قابل اندازه‌گیری بر نتایج امتیازدهی ندارند و از این‌رو ایجاد شبکه با ابعاد یکسان برای پوشش‌دهی اسیدهای آمینه جایگاه فعال پروتئین توصیه می‌شود. اگر فایل مربوط به ساختار دارای لیگاند کمپلکس شده باشد (هولو)، جایگاه فعال پروتئین مشخص خواهد شد و بنابراین مختصات جعبه شبکه‌ای همان مختصات جایگاه فعال است که به شکل مکعبی با ابعاد یکسان تعریف می‌شود. ابعاد جعبه باید به‌گونه‌ای انتخاب شود که اسیدهای آمینه جایگاه فعال را در بر بگیرد [۳۹, ۴۰].

۱- ۵- ۵ گزینه داکینگ

به‌منظور محاسبه داکینگ، ورودی‌ها نظیر روش‌های نمره‌دهی، جایگاه فعال انعطاف پذیر، حلال پوشی، برخورد با جایگاه فعال احاطه شده، روش‌های جستجو و ... تنظیم می‌شوند.

۱- ۵- ۶ انجام محاسبه داکینگ

زمانی‌که ورودی‌ها تنظیم شدند محاسبات داکینگ انجام می‌شود. اغلب برنامه‌ها امکان اجراهای داکینگ به‌صورت تکی و تعداد اندکی قادر به اجرای محاسبات داکینگ در یک حالت مجموعه‌ای می‌باشند. این محاسبات در مواقعی توسط همان کامپیوتری که صفحه رابط گرافیکی از آن استفاده شده، انجام می‌گیرد و در مواقعی هم می‌تواند به یک سرور دیگر فرستاده شود.

[1] Cut off

۱- ۵- ۷ آنالیز و تحلیل نتایج

از مهم‌ترین نتایج محاسبات داکینگ، انرژی اتصال لیگاند به‌جایگاه فعال می‌باشد. این انرژی در بین ترکیبات مختلف برای تعیین بهترین مهار کننده، استفاده می‌شود. چند حالت و وضعیت از لیگاند در جایگاه فعال که با بهترین انرژی اتصال همراه است، به‌طور دقیق بررسی شده تا از مناسب بودن آن اطمینان حاصل شود. در بعضی موارد، وضعیتی که توسط یک محاسبه داکینگ تولید شده، به کاربر طرحی برای چگونگی تغییر ترکیبات در مراحل بعدی محاسبات می‌دهد.

در برنامه داکینگ، دو مؤلفه کلیدی الگوریتم رتبه‌بندی یا امتیازدهی[۱] و الگوریتم جستجو[۲] 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 فرآیند، موقعیت‌های ممکن مولکول را ارزیابی می‌کند و چه مدت زمانی برای آن نیاز است. الگوریتم رتبه‌بندی مسئول تعیین این است که آیا جهت‌گیری‌های انتخاب شده توسط الگوریتم جستجو، از لحاظ انرژی مناسب‌ترین هستند و مسئول محاسبه انرژی اتصال است. قابل ذکر است که الگوریتم جستجو صحت نتایج به‌دست آمده از برنامه داکینگ را مشخص نمی‌کند. بنابراین صحت نتایج در درجه اول به صحت الگوریتم رتبه‌بندی بستگی دارد. [۳۸].

۱- ۵- ۸ کاربردهای داکینگ مولکولی

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

۱. غربالگری مجازی[۳]

[1] The scoring algorithm
[2] The search algorithm
[3] Virtual screening

۲. بهبود طراحی و کشف دارو

۳. پیش‌بینی انرژی آزاد اتصال

۴. بررسی برهم‌کنش لیگاند- گیرنده

متداول‌ترین کاربرد داکینگ مولکولی در اتصال پروتئین-لیگاند است. هدف نهایی اتصال پروتئین-لیگاند پیش‌بینی فعالیت دارویی لیگاند و برهم‌کنش لیگاند با پروتئین می‌باشد و در این تحقیق از این هدف (اتصال لیگاند-پروتئین) جهت بررسی برهم‌کنش‌های ترکیبات پیشنهادی بالقوه در جایگاه فعال گیرنده استفاده می‌شود.

۱- ۶ سرطان

سرطان یک بیماری ژنتیکی است که در آن کنترل رشد سلول‌ها از بین رفته است. بر اساس آمار سازمان جهانی بهداشت[1] (WHO)، سرطان دومین علت مرگ‌ومیر در جهان پس از بیماری‌های قلبی و عروقی به‌شمار می‌آید و تقریباً ۱۰ میلیون مرگ در سال ۲۰۲۰ در سراسر جهان به خود اختصاص داده است [۴۱]. در حالت نرمال بین میزان تقسیم سلول، مرگ سلولی و تمایز سلولی یک تعادل وجود دارد. سلول‌ها به‌طور دقیقی به‌وسیله فاکتورهای رشد، پیام‌رسان‌ها، پیام‌های محیطی و برخی پروتئین‌ها کنترل می‌شوند. یک مجموعه جهش‌های متوالی در ژن‌ها اتفاق می‌افتد که منجر به از دست رفتن مسیر رشد طبیعی سلول‌ها و بروز سلول‌های سرطانی می‌شود. سلول‌های سرطانی کنترل رشد خود بر چرخه سلولی را از دست داده و به‌طور مداوم و بدون توجه به پیام‌های محیطی و فاکتورهای رشد، به تکثیر ادامه می‌دهند. یکی از فاکتورهای رشد، فاکتور رشد هپاتوسیت[2] می‌باشد که به‌همراه گیرنده خود (c-Met)[3] نقش‌های مهمی در اعمال بیولوژیکی بدن و سرطان انجام می‌دهد. گیرنده فاکتور رشد هپاتوسیت یکی از اعضای

[1] World health organization
[2] Hepatocyte Growth Factor
[3] Mesenchymal-epithelial transition factor

خانواده گیرنده تیروزین کیناز (RTK) محسوب می‌شود. تیروزین کینازها گیرنده‌های سطح سلول هستند و اتصال لیگاندهای فاکتور رشد باعث فعال شدن بیش‌تر مسیرهای انتقال سیگنال می‌شود. با این حال، تغییرات ساختاری بیش‌ازحد فاکتورهای رشد و تیروزین کینازهای آن‌ها از طریق چندین مکانیسم با پیشرفت تومور در ارتباط است. مهار این تیروزین کینازهای فاکتور رشد به‌عنوان یک استراتژی درمان سیستماتیک برای سرطان‌های مختلف ایجاد شده است [۴۲, ۴۳]. از دسته ترکیبات بازدارنده c-Met، ترکیبات سولفون دار است. سولفون‌ها ($R-SO_2-R$)، مولکول‌های دارای واحد سولفون، که واسطه‌های چند منظوره در شیمی آلی هستند. کاربردهای مختلفی در زمینه‌هایی مانند شیمی دارویی و پلیمرها دارند. مشتقات سولفون دسته مهمی از ترکیبات فعال زیستی هستند و شامل طیف وسیعی از فعالیت‌ها می‌باشند. این ترکیبات دارای فعالیت‌های متنوعی به‌عنوان عوامل ضد باکتری، حشره‌کش، ضد قارچ، علف‌کش، ضد هپاتیت، ضد تومور، ضد التهاب، ضد سرطان، ضد HIV-1 و ضد سل می‌باشند [۴۴]. در واقع، اولین داروی مبتنی بر سولفون، سولفونال[1]، به سال ۱۸۸۸ باز می‌گردد [۴۵].

۱- ۷ مروری برکارهای انجام شده و ضرورت تحقیق

با توجه به این‌که ترکیبات مورد مطالعه در این تحقیق از دسته ترکیبات c-Met سولفون دار هستند، مروری بر مطالعات منتشر شده در راستای ارائه مدل‌های QSAR برای پیش‌بینی فعالیت دارویی این دسته از ترکیبات انجام شد و اهمیت ارائه مدل QSAR برای این نوع بازدارنده‌ها مورد بررسی قرار گرفت [۴۶-۴۹]. با توجه به جستجوی انجام شده مدل‌های QSAR توسعه یافته در این مقالات از قدرت پیش‌بینی مناسبی برخوردار هستند.

در این تحقیق، از روش مدل‌سازی انقباضی انحراف قدر مطلق به‌طور هموار کوتاه شده (SCAD) به‌عنوان روشی برای انتخاب متغیر و مدل‌سازی استفاده شد. روش انقباضی SCAD در مطالعات QSAR به

[1] sulfonal

دو شکل طبقه‌بندی[1] و رگرسیونی استفاده شده است. موارد مربوط به کاربرد روش‌های انقباضی در مطالعات طبقه‌بندی مد نظر اهداف این پروژه نبوده است و جزئیات آن در بخش مرور بر کارها ذکر نشده است [۱۲, ۵۰-۵۴]. در ادامه به‌مرور مطالعات منتشر شده از به‌کارگیری روش‌های انقباضی SCAD در مدل رگرسیونی در طی سال‌های ۲۰۱۰-۲۰۲۱ پرداخته شد.

الجمال و همکاران[2] در سال ۲۰۱۶، مدل QSAR را برای ۱۱۱ بازدارنده‌های استیل کولین استراز ارائه دادند و از روش غربالگری مستقل مطمئن[3] (SIS) به‌عنوان پیش‌پردازش برای کاهش بعد داده‌ها استفاده شد. در نهایت مدل‌های QSAR با به‌کارگیری روش‌های انقباضی LASSO، ALASSO[4]، SCAD و BRIDGE توسعه یافت. نتایج ضرایب تعیین مجموعه آزمون به‌ترتیب ۰/۸۰، ۰/۸۱، ۰/۸۷ و ۰/۹۱ به‌دست آمد. نتایج گواهی بر قدرت پیش‌بینی مناسب مدل‌های توسعه یافته است [۱۱].

یونگ زیا[5] و همکاران در سال ۲۰۱۸، مدل‌های QSAR را با استفاده از ۴ مجموعه از ترکیبات مختلف و به‌کارگیری روش‌های انقباضی SCAD، LASSO، Elastic Net و BRIDGE ارائه دادند. پارامترهای آماری متفاوتی برای ارزیابی مدل به‌کار گرفته شد که نتایج مدل‌های توسعه یافته از اعتبار قابل قبولی برخوردار بودند [۵۵].

ماجومدار و همکارانش[6] در سال ۲۰۱۸ مدل‌های QSAR را برای مجموعه داده همگن متشکل از ۹۵ آمین، ارائه نمودند. با استفاده از توصیف کننده‌های محاسبه شده، فعالیت جهش‌زایی این ترکیبات را با استفاده از روش‌های مدل‌سازی متفاوتی همچون PCR[7]، PLS[8] ، RF[9]، LASSO و SCAD پیش‌بینی

[1] Classification
[2] Algamal and et al.
[3] Sure Independence Screening
[4] Adaptive Least Absolute Shrinkage and Selection Operator
[5] Yong Xia and et al.
[6] Majumdar and et al.
[7] Principal Component Regression
[8] Partial Least Squares
[9] Random Forest

کردند. میانگین مربعات خطای پیش‌بینی برای مدل‌های QSAR مربوطه به‌دست آمد و به‌ترتیب برابر با ۲۹/۱۱، ۱۸/۸۶، ۱۷/۲۵، ۲۶/۸۵ و ۲۵/۸۱ است. نتایج نشان می‌دهد که مدل‌های ساخته شده از اعتبار مناسبی برخوردار هستند [۵۶].

مظفری و همکاران[1] درسال ۲۰۲۱ برای ۵۷ ترکیب از مشتقات آریلازولیل تیواستامید/ استانیلید به‌عنوان مهارکننده HIV یک مدل QSAR با روش شبکه عصبی مصنوعی[2] (ANN) گزارش دادند که از روش انقباضی SCAD به‌عنوان انتخاب متغیر استفاده شد و نتیجه پارامتر ضریب تعیین برای مجموعه آزمون ۰/۹۲ به‌دست آورده شد [۵۷].

علاوه بر این طبق جستجوی کتابخانه‌ای انجام شده، در سال‌های اخیر گزارشی مبنی بر استفاده از روش Ridge به‌عنوان غربالگری در مطالعات QSAR وجود نداشته است و تنها یک گزارش از به‌کارگیری روش غربالگری مستقل مطمئن (SIS) قبل از مدل‌سازی SCAD توسط الجمال و همکارش در سال ۲۰۱۶ منتشر شده است [۱۱]. بنابراین با توجه به این‌که این روش در مطالعات QSAR استفاده نشده است، در این پروژه برای اولین بار از روش رگرسیون Ridge به‌عنوان روش غربالگری جفت شده با روش انقباضی SCAD برای پیش‌بینی فعالیت دارویی بازدارنده‌های سرطان استفاده شده است. لازم به ذکر است تکنیک غربالگری رد مرحله‌ای تک تک Ridge مورد استفاده در این مطالعه به‌طور ابتکاری و برای اولین بار در مطالعات QSAR پیشنهاد شده است.

۱- ۸ نوآوری تحقیق

جستجوی کتابخانه‌ای نشان می‌دهد که مدل‌های QSAR توسعه یافته، از روش SCAD برای ساخت مدل خطی QSAR استفاده شده است. در داده‌هایی با ابعاد بالا به‌کارگیری روش‌های غربالگری قبل

[1] Mozafari and et al.
[2] Artificial Neural Network

از انتخاب متغیر از اهمیت ویژه‌ای برخوردار است. ابعاد داده‌ها باید کاهش یابد تا علاوه بر این‌که مدل قدرت پیش‌بینی بالایی داشته باشد، قدرت تفسیرپذیری روشنی هم داشته باشد. در مواجهه با داده‌هایی با ابعاد بالا، به‌کارگیری روش غربالگری Ridge قبل از مدل‌سازی SCAD، مدل قدرتمندی با تفسیرپذیری مناسب-تری به‌دست می‌آید. استفاده از آنالیز دو مرحله‌ای Ridge-SCAD می‌تواند از نظر محاسباتی کارآمدتر باشد و برآورد مدل بهتری را نسبت به روش‌های جریمه‌ای یک مرحله‌ای ارائه دهد. بنابراین ابتدا از روش رگرسیون Ridge برای کاهش بعد داده‌ها و سپس از روش انقباضی SCAD برای انتخاب متغیر و مدل‌سازی به‌منظور پیش‌بینی فعالیت دارویی بازدارنده‌های c-Met استفاده شد و گزارشی مبنی بر ترکیب کردن روش رگرسیون Ridge با روش انقباضی SCAD وجود ندارد. در نهایت کارایی مدل توسعه یافته برای پیشنهاد برخی از ترکیبات جدید c-Met با فعالیت دارویی ضد سرطان استفاده شد.

فصل دوم

نتایج تجربی

۲- ۱ نرم افزارهای مورد استفاده

برای انجام کلیه مراحل مدل‌سازی از بسته‌های نرم‌افزاری مختلفی استفاده گردید که در ادامه به معرفی مختصر نرم‌افزارهای مورد استفاده در این تحقیق پرداخته می‌شود.

۲- ۱- ۱ نرم افزار هایپرکم[1]

برای رسم ساختار مولکول‌ها و بهینه‌سازی آن‌ها با استفاده از روش‌های کوانتومی و مکانیکی، از نرم افزار هایپرکم استفاده می‌شود [۵۸]. طول پیوند، زاویه پیوندی و زوایای پیچشی در مولکول با این برنامه بهینه می‌شود. داده‌های حاصل از این نرم افزار به‌عنوان ورودی سایر نرم افزارها از جمله دراگون[2] استفاده می‌شود.

۲- ۱- ۲ نرم افزار دراگون

نرم افزار دراگون به‌گونه‌ای طراحی شده است که انواع توصیف کننده‌های مولکولی مشتق شده از ساختار ترکیبات مختلف را در اختیار کاربر قرار دهد و به کاربر این امکان را می‌دهد تا توصیف کننده‌های مولکولی را انتخاب کند که برای تحقیقات مناسب‌تر هستند. اولین نسخه دراگون در سال ۱۹۹۴ توسعه یافت [۵۹]. نرم افزار دراگون نسخه ۵/۵ [۶۰] امکان محاسبه ۳۲۲۴ توصیف کننده مختلف را برای شیمیدان‌های محاسباتی فراهم می‌کند. جهت محاسبه توصیف کننده‌ها لازم است از ساختارهای هندسی بهینه ترکیبات استفاده شود.

[1] Hyperchem
[2] Dragon

۲- ۱- ۳ نرم افزار SPSS [1]

نرم افزار SPSS [۶۱] امکان تجزیه و تحلیل آماری داده‌ها را فراهم می‌کند. برخی از قابلیت‌های این بسته نرم افزاری عبارتند از : محاسبه میانگین ساده برای داده‌ها، نمایش اطلاعات به‌صورت متنوع در قالب نمودار و انجام رگرسیون تک متغیره و چند متغیره. در این تحقیق از نرم افزار SPSS نسخه ۲۱ برای محاسبه هم‌خطی موجود بین متغیرها استفاده شد.

۲- ۱- ۴ نرم افزار R و R-Studio

نرم افزار R [۶۲] یک زبان برنامه نویسی برای محاسبات آماری و گرافیکی است. طیف گسترده‌ای از مدل‌های آماری (مدل‌سازی خطی و غیرخطی، آزمون‌های آماری کلاسیک، طبقه‌بندی و غیره) و تکنیک‌های گرافیکی را فراهم می‌کند. یکی از نقاط قوت نرم افزار R آسان بودن طراحی نمودارهای با کیفیت است که شامل نمادها و فرمول‌های ریاضی مورد نیاز می‌باشد. نرم افزار R-Studio یک محیط گرافیکی مناسب‌تر برای اجرای برنامه‌های تحت R است. استفاده از R-Studio برای اجرای بسته‌های نرم افزاری آماری موجود در نرم افزار R ضروری می‌باشد. در این تحقیق از نرم افزارهای R و R-Studio، برای پیش پردازش توصیف کننده با استفاده از بسته نرم افزاری Caret، برای غربالگری توصیف کننده‌ها به کمک روش انقباضی Ridge با استفاده از بسته نرم افزاری glmnet و انتخاب مؤثرترین توصیف کننده‌ها با استفاده از بسته نرم افزاری ncvreg استفاده شد.

۲- ۱- ۵ نرم افزار متلب [2]

نرم افزار متلب [۶۳] یکی از جامع‌ترین و کارآمدترین نرم افزارهای علمی و محاسباتی است که طی چند سال گذشته تهیه و به بازار عرضه شد و در طی سال‌های اخیر با تدوین نسخه‌های جدیدتر و کامل‌تر

[1] Statistical Package for the Social Science
[2] MATLAB

روزبه‌روز مورد توجه محققین قرار گرفته است. MATLAB [1] به معنای آزمایشگاه ماتریس است. متلب یک محیط محاسباتی عددی و زبان برنامه‌نویسی است و امکان به‌کارگیری ماتریس، ترسیم توابع و داده‌ها و اجرای الگوریتم‌ها را فراهم می‌کند. در این تحقیق برای محاسبه همبستگی بین توصیف کننده‌ها با استفاده از دستور corecoeff و برخی از ارزیابی‌های مدل برتر از نرم افزار متلب ۲۰۱۷a استفاده شد.

۲- ۱- ۶ نرم افزار اتوداک[2]

نرم افزار اتوداک نسخه ۴/۲ یکی از نرم افزارهای مدل‌سازی مولکولی می‌باشد که قابل اجرا بر روی سیستم عامل ویندوز است [۶۴]. این نرم افزار برای پیش‌بینی چگونگی اتصال مولکول‌ها به‌عنوان سوبسترا یا دارو به یک مولکول گیرنده با ساختار سه‌بعدی طراحی شده است.

۲- ۱- ۷ نرم افزار ویورلایت[3]

نرم افزار ویورلایت نسخه ۵/۰ امکان بررسی جزئی ماکرومولکول و لیگاند را به کاربر می‌دهد و می‌توان با استفاده از این نرم افزار ماکرومولکول را مشاهده و ویرایش کرد. همچنین جایگاه فعال پروتئین را می‌توان مشاهده و تعیین مختصات نمود. همین‌طور می‌توان طول پیوندهای هیدروژنی و هیدروفوبی مربوط به لیگاند و اسیدهای آمینه موجود در جایگاه فعال را نیز بررسی کرد [۶۵]. در این مطالعه جهت آماده‌سازی لیگاند و پروتئین برای انجام داکینگ مولکولی و بررسی بر هم‌کنش‌های لیگاند-پروتئین از نرم افزار ویورلایت نسخه ۵/۰ استفاده گردید.

۲- ۱- ۸ نرم افزار Discovery Studio

این نرم افزار امکان بررسی کمپلکس لیگاند و گیرنده را پس از اجرای داکینگ مولکولی فراهم می‌کند. به‌گونه‌ای که تمامی پیوندهای آب‌گریز و هیدروژنی و همین‌طور طول پیوندهای هیدروژنی و آب‌گریز

[1] Matrix Laboratory
[2] Autodock
[3] ViewerLite

را نشان می‌دهد. بررسی بر هم‌کنش‌های اسیدهای آمینه موجود در جایگاه فعال با اجزای اتمی لیگاند با استفاده از این نرم افزار امکان پذیر است.

۲- ۲ مدل سازی فعالیت دارویی مشتقات سولفون دار به عنوان بازدارنده های c-Met

کمی کردن ویژگی‌های ساختاری و فعالیت دارویی ترکیبات دارویی و پیدا کردن رابطه حاکم بر کمیت‌های اندازه‌گیری شده، همواره موضوع در حال توسعه و مورد توجه بوده و از طریق ساخت مدل‌های QSAR صورت می‌گیرد. در سال‌های اخیر، به‌دلیل توسعه و پیشرفت تکنولوژی، استفاده از شیمی محاسباتی، شبیه‌سازی مولکولی و طراحی و ساخت مدل‌های QSAR متفاوت به‌کمک رایانه به‌طور گسترده تری مورد توجه شیمیدان‌های محاسباتی قرار گرفته است.

روش‌های محاسباتی دارای مزایایی همچون کاهش سنتزهای آزمایشگاهی و صرفه‌جویی در زمان، هزینه و نیروی انسانی می‌شوند. بنابراین محققین شیمی محاسباتی همواره در تلاش تا روش‌های آماری و محاسباتی پیشرفته‌تری را برای بررسی ارتباط بین متغیرهای وابسته و متغیر مستقل به‌کار ببرند که بتوان متغیرهایی که ارتباط بیش‌تری با پاسخ دارند را شناسایی کرد، بنابراین ارائه یک مدل QSAR با قدرت تفسیر پذیری و پیش‌بینی مناسب مورد توجه است. به‌منظور بهبود مدل‌های QSAR، به‌کارگیری روش‌های انتخاب متغیر و مدل‌سازی جدید و کارآمد نظیر روش‌های انتخاب متغیر جریمه‌ای به‌دلیل داشتن مزایایی چون تنکی، پایداری و بایاس کم، دارای اهمیت است. در داده‌های با ابعاد بالا بین توصیف کننده‌ها مشکل همبستگی بالا وجود دارد و همبستگی بالا منجر به ایجاد پدیده هم‌خطی می‌شود. لذا انتظار می‌رود با به کارگیری یک روش غربالگری متغیرها به‌همراه روش و انتخاب متغیر جریمه‌ای بتوان توصیف کننده‌هایی با تعداد محدودتر ولی با اطلاعات مفید را از تعداد بسیار زیادی توصیف کننده‌ها استخراج نمود. به‌منظور ارائه

یک مدل تنک و پایدار از روش رگرسیون Ridge برای غربالگری و کاهش تعداد توصیف کننده‌های محاسبه-ای استفاده شد و روش رگرسیون جریمه‌ای SCAD برای انتخاب مؤثرترین توصیف کننده‌ها بر فعالیت دارویی و ایجاد یک مدل خطی بین آن‌ها و فعالیت دارویی دسته‌ای از مشتقات سولفون دار به‌عنوان بازدارنده‌های سرطان به‌کار گرفته شد. لازم به‌ذکر است که روش رگرسیون Ridge که اساساً یک روش رگرسیون با تمام متغیرها و بدون قابلیت انتخاب متغیر می‌باشد به کمک یک روش ابتکاری به‌عنوان روش غربالگری قبل از توسعه مدل‌های QSAR به‌کار گرفته شد. به‌منظور بررسی عملکرد روش Ridge برای غربالگری توصیف کننده‌ها قبل از مدل‌سازی با SCAD (Ridge-SCAD)، مدل QSAR با استفاده از تمام متغیرها بدون هیچ نوع غربالگری با به‌کارگیری روش SCAD توسعه یافت و نتایج پیش‌بینی مدل، با هم مقایسه شد.

از آنجایی‌که سنتز و ارزیابی تجربی فعالیت دارویی ترکیبات فعال زمان‌بر و پرهزینه است، توسعه روش‌های تئوری از قبیل ساخت مدل QSAR برای ارزیابی و پیش‌بینی فعالیت‌های دارویی ترکیبات مشابه بسیار مهم است و باعث صرفه‌جویی در زمان و هزینه و به‌کارگیری نیروی انسانی می‌شود. با توجه به اهمیت موضوع، در این بخش مدل QSAR برای پیش‌بینی فعالیت ترکیبات مورد مطالعه و پیشنهاد ترکیبات مشابه جدید توسعه یافت. ازاین‌رو مدل Ridge-SCAD برای پیش‌بینی فعالیت دارویی ($pIC_{۵۰}$) برخی از مشتقات سولفون دار به‌عنوان بازدارنده‌های بیماری سرطان ساخته شد. فعالیت دارویی این ترکیبات و پتانسیل بالقوه آن‌ها به‌عنوان کاندیدهای پیشنهادی با فعالیت ضد سرطان با استفاده از داکینگ مولکولی و محاسبه پارامترهای قاعده پنج لیپینسکی[1] مورد بررسی قرار گرفت و نتایج حاصل امکان استفاده از آن‌ها به‌عنوان کاندیدهای جدید فعالیت ضد سرطان را اثبات نمود. در ادامه این بخش با جزئیات بیش‌تری به مراحل

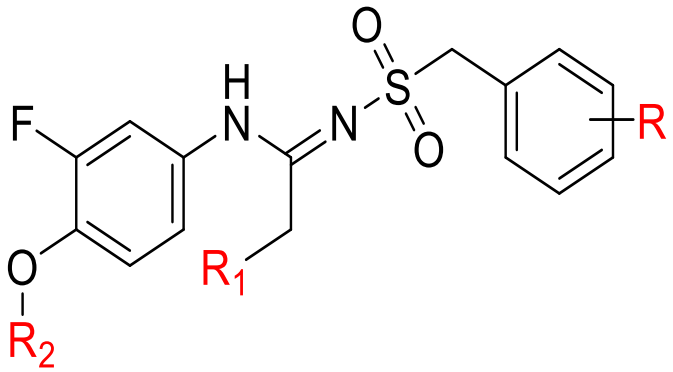
[1] Lipinski's rule of five

ساخت مدل‌های Ridge-SCAD و SCAD (برای بررسی کارایی مدل Ridge-SCAD) این دسته از بازدارنده‌ها پرداخته شده است.

۲- ۲- ۱ معرفی مجموعه داده های مورد استفاده در این مطالعه

مجموعه داده‌ها شامل فعالیت دارویی ۶۳ ترکیب از مشتقات سولفون دار می‌باشد که توسط نان و همکارانش[1] در سال‌های ۲۰۱۹ و ۲۰۲۰ ارائه شده است [۶۶, ۶۷]. اسکلت اصلی لیگاندها (از این پس ترکیبات مورد مطالعه که همان مشتقات سولفون دار می‌باشند به‌نام لیگاندها بیان می‌شوند) در شکل ۲-۱ و شکل ۲-۲ نشان داده شده است. جزئیات استخلاف‌ها و مقادیر منهای لگاریتم $IC_{۵۰}$ ($pIC_{۵۰}$) هر لیگاند در جدول ۲-۱ و جدول ۲-۲ آورده شده است. $IC_{۵۰}$ غلظتی از دارو برحسب مولار است که تا ۵۰٪ تکثیر پروتئین c-Met را مهار می‌کند. لازم به‌ذکر است که هر چقدر دارو قوی‌تر باشد، غلظت کم‌تری از آن، برای داشتن ۵۰٪ اثر مهارکنندگی، لازم بوده و در نتیجه مقدار $pIC_{۵۰}$ بزرگ‌تری دارد. در کلیه مراحل ساخت مدل QSAR از $pIC_{۵۰}$ به‌عنوان متغیر وابسته و توصیف کننده‌های مولکولی به‌عنوان متغیر مستقل در نظر گرفته شدند. تقسیم‌بندی مجموعه داده‌ها با استفاده از الگوریتم KS انجام شد و در کلیه مراحل ساخت مدل QSAR مجموعه آموزش (متشکل از ۵۳ ترکیب) برای ساخت مدل‌های خطی Ridge-SCAD و SCAD و مجموعه آزمون (متشکل از ۱۰ ترکیب) برای ارزیابی نهایی مدل‌ها به‌کار گرفته شده‌اند.

[1] Nan et al.



شکل ۲-۱ اسکلت اصلی ساختار ترکیبات مورد مطالعه مربوط به ترکیبات ۱ تا ۳۹

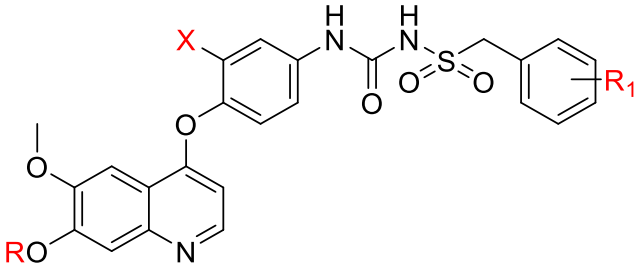
جدول ۲-۱ جزئیات ساختار ۱ تا ۳۹ به همراه مقادیر pIC_{50} ترکیبات مورد مطالعه

ردیف	R_2	R	R_1	فعالیت‌دارویی (pIC_{50})
۱	thieno[3,2-b]pyridine	H	n-butyl	۵/۰۲
۲[t]	thieno[3,2-b]pyridine	4-F	n-butyl	۵/۱۵
۳	thieno[3,2-d]pyrimidine	H	n-butyl	۴/۷۳
۴[t]	thieno[3,2-d]pyrimidine	4-F	n-butyl	۴/۹۵
۵	6,7-dimethoxyquinazoline	H	n-butyl	۴/۵۲
۶	6,7-dimethoxyquinazoline	4-F	n-butyl	۴/۶۶
۷	6,7-dimethoxyquinoline	H	n-butyl	۵/۰۸
۸	6,7-dimethoxyquinoline	4-F	n-butyl	۵/۲۹
۹	6,7-dimethoxyquinoline	H	Phenyl	۴/۴۹
۱۰[t]	6,7-dimethoxyquinoline	4-F	Phenyl	۴/۶۳
۱۱	6,7-dimethoxyquinoline	H	Methoxymethyl	۵/۴۵
۱۲	6,7-dimethoxyquinoline	4-F	Methoxymethyl	۵/۷۸
۱۳	6,7-dimethoxyquinoline	H	H	۵/۲۳
۱۴	6,7-dimethoxyquinoline	4-F	H	۵/۳۹
۱۵	6,7-dimethoxyquinoline	H	1-Cyclohexenyl	۴/۸۴
۱۶	6,7-dimethoxyquinoline	4-F	1-Cyclohexenyl	۴/۹۹
۱۷	6,7-dimethoxyquinoline	H	3-Pyridyl	۴/۶۹
۱۸	6,7-dimethoxyquinoline	4-F	3-Pyridyl	۴/۸۱
۱۹[t]	6,7-dimethoxyquinoline	H	3-Thienyl	۴/۷۸
۲۰	6,7-dimethoxyquinoline	4-F	3-Thienyl	۴/۹۵
۲۱	6,7-dimethoxyquinoline	2-Methyl	Methoxymethyl	۵/۳۱
۲۲	6,7-dimethoxyquinoline	3-Methyl	Methoxymethyl	۵/۱۰
۲۳	6,7-dimethoxyquinoline	4-Methyl	Methoxymethyl	۵/۰۶

ادامه جدول ۲-۱

ردیف	R_2	R	R_1	فعالیت‌دارویی (pIC_{50})
۲۴	6,7-dimethoxyquinoline	2-F	Methoxymethyl	۵/۵۰
۲۵	6,7-dimethoxyquinoline	3-F	Methoxymethyl	۵/۶۱
۲۶[t]	6,7-dimethoxyquinoline	4-Cl	Methoxymethyl	۶/۰۴
۲۷	6,7-dimethoxyquinoline	4-Br	Methoxymethyl	۵/۱۰
۲۸[t]	6,7-dimethoxyquinoline	3,4-di-Cl	Methoxymethyl	۶/۰۶
۲۹	6,7-dimethoxyquinoline	4-CF_3	Methoxymethyl	۴/۸۸
۳۰	7-butoxy-6-methoxyquinoline	4-Cl	Methoxymethyl	۵/۵۵
۳۱	7-butoxy-6-methoxyquinoline	3,4-di-Cl	Methoxymethyl	۵/۳۹
۳۲[t]	4-(3-((6-methoxyquinolin-7-yl)oxy)propyl)morpholine	4-Cl	Methoxymethyl	۶/۵۰
۳۳	4-(3-((6-methoxyquinolin-7-yl)oxy)propyl)morpholine	3,4-di-Cl	Methoxymethyl	۶/۲۳
۳۴	6-methoxy-7-(3-(piperidin-1-yl)propoxy)quinoline	4-Cl	Methoxymethyl	۶/۳۹
۳۵	6-methoxy-7-(3-(piperidin-1-yl)propoxy)quinoline	3,4-di-Cl	Methoxymethyl	۶/۱۱
۳۶	6-methoxy-7-(3-(4-methylpiperazin-1-yl)propoxy)quinoline	4-Cl	Methoxymethyl	۶/۴۴
۳۷	6-methoxy-7-(3-(4-methylpiperazin-1-yl)propoxy)quinoline	3,4-di-Cl	Methoxymethyl	۶/۱۳
۳۸	6-methoxy-7-(3-(4-methylpiperidin-1-yl)propoxy)quinoline	4-Cl	Methoxymethyl	۶/۲۶
۳۹	6-methoxy-7-(3-(4-methylpiperidin-1-yl)propoxy)quinoline	3,4-di-Cl	Methoxymethyl	۶/۰۶

[t] نمایانگر داده‌های مجموعه آزمون می‌باشد. سایر داده‌ها مربوط به مجموعه آموزش می‌باشند.



شکل ۲-۲ اسکلت اصلی ساختار ترکیبات مورد مطالعه مربوط به ترکیبات ۴۰ تا ۶۳

جدول ۲-۲ جزئیات ساختار ۴۰ تا ۶۳ به همراه مقادیر pIC_{50} ترکیبات مورد مطالعه

ردیف	X	R	R_1	فعالیت‌دارویی (pIC_{50})
۴۰	H	Methyl	H	۵/۳۶
۴۱	H	Methyl	4-F	۵/۴۹
۴۲	F	Methyl	H	۵/۵۲
۴۳	F	Methyl	4-F	۵/۵۷
۴۴	F	Methyl	3-F	۵/۵۱
۴۵	F	Methyl	2-F	۵/۵۱
۴۶[t]	F	Methyl	4-Me	۵/۱۵
۴۷	F	Methyl	3-Me	۵/۲۲
۴۸	F	Methyl	2-Me	۵/۲۱
۴۹[t]	F	Methyl	4-Cl	۵/۵۴
۵۰	F	Methyl	4-Br	۵/۵۱
۵۱	F	Methyl	4-CF_3	۵/۰۹
۵۲	F	Methyl	2,4-Cl	۵/۵۳
۵۳	F	n-butyl	H	۵/۵۱
۵۴	F	n-butyl	4-F	۵/۵۴
۵۵[t]	F	n-butyl	4-Me	۵/۰۳
۵۶	F	3-morpholinopropyl	H	۵/۸۹
۵۷	F	3-morpholinopropyl	4-F	۶/۴۷
۵۸	F	3-morpholinopropyl	4-Me	۵/۴۳
۵۹	F	3-(piperidin-1-yl)propyl	H	۵/۹۱
۶۰	F	3-(piperidin-1-yl)propyl	4-F	۶/۵۰
۶۱	F	3-(piperidin-1-yl)propyl	4-Me	۵/۶۶
۶۲	F	3-(pyrrolidin-1-yl) propyl	H	۶/۰۳
۶۳	F	3-(pyrrolidin-1-yl) propyl	4-F	۶/۸۰

[t] نمایانگر داده‌های مجموعه آزمون می‌باشد. سایر داده‌ها مربوط به مجموعه آموزش می‌باشند.

۲- ۲- ۲ رسم و بهینه سازی ساختار هندسی ترکیبات

ساختار شیمیایی ترکیبات مورد مطالعه با استفاده از نرم افزار هایپرکم رسم شد و بهینه‌سازی ساختارها مطابق با روش کار (بخش ۱- ۳- ۲)، با استفاده از روش نیمه تجربی کوانتومی AM1، به‌عنوان یکی از کارآمدترین روش‌های نیمه‌تجربی، بهینه شد و بهینه‌سازی تا زمانی ادامه یافت که جذر میانگین مربعات گرادیان[1] انرژی به ۰/۰۰۱ کیلو کالری بر مول برسد.

۲- ۲- ۳ محاسبه توصیف کننده ها

به‌منظور کمی کردن ساختارهای شیمیایی بهینه شده و استخراج توصیف کننده‌ها، از نرم افزار دراگون استفاده شد. به این منظور ترکیبات مورد مطالعه در نرم افزار دراگون فراخوانی شدند و تعداد ۳۲۲۴ توصیف کننده برای ساختار هر ترکیب در ۲۲ دسته از جمله نوع و تعداد اتم‌ها، گروه‌های عاملی، توصیف‌کننده‌های مکانی و غیره محاسبه شد.

۲- ۲- ۴ پیش پردازش اولیه

با توجه به تعداد زیاد توصیف‌کننده‌های محاسبه شده و احتمال وجود توصیف‌کننده‌های اضافی و فاقد اطلاعات مفید، فرآیند پیش‌پردازش توصیف کننده‌ها به‌منظور کاهش تعداد توصیف‌کننده‌ها و در نتیجه بهبود صحت پیش‌بینی، کاهش زمان مدل‌سازی و افزایش تفسیرپذیری مدل بر روی کل توصیف‌کننده‌های محاسبه شده اعمال شد. به این منظور توصیف کننده‌هایی که مقادیر ثابت برای تمام ترکیبات داشتند حذف گردید و ۱۵۹۱ توصیف کننده به‌دست آمد. سپس، ضرایب همبستگی بین تمام توصیف کننده‌ها محاسبه شد. از بین دو توصیف کننده همبسته ($R > ۰/۹۰$) که دارای اطلاعات تقریباً یکسانی هستند، توصیف کننده با بیش‌ترین همبستگی با پاسخ، نگه داشته شده و توصیف‌کننده بعدی حذف شد و در نتیجه تعداد ۴۰۱

[1] Root Mean Square Gradient

توصیف کننده باقی ماند. در مرحله بعد توصیف کننده‌های نسبتاً ثابت (توصیف کننده‌هایی که انحراف استاندارد کمتر از ۰/۰۰۱ داشتند) حذف شدند و ۳۹۲ توصیف کننده باقی ماند. هر چند تعداد توصیف کننده‌ها به ۳۹۲ کاهش یافت، ولی اگر بتوان به‌گونه‌ای با حذف توصیف کننده‌های با تأثیر کم‌تر، تعداد توصیف کننده‌ها را کاهش بیش‌تری داد، در این‌صورت می‌توان به مدل‌های خطی SCAD تنک‌تر دست یافت. بنابراین، روش رگرسیون Ridge به‌همراه یک فرآیند چند مرحله‌ای جدید به‌عنوان روش غربالگری روی توصیف کننده‌های باقی‌مانده اعمال شد.

۲- ۲- ۵ استفاده از روش رگرسیون Ridge برای غربالگری توصیف کننده ها

روش رگرسیون Ridge یک روش مدل‌سازی است و به‌دلیل این‌که همه‌ی متغیرها را در مدل نگه می‌دارد به‌عنوان روش انتخاب متغیر استفاده نمی‌شود. در این پروژه با استفاده از تکنیکی که در ادامه به شرح آن پرداخته می‌شود، از روش Ridge به‌عنوان روش غربالگری استفاده شد. روش Ridge با استفاده از تکنیک رد مرحله‌ای تک تک روی ۳۹۲ توصیف کننده باقی‌مانده اجرا شد. با به‌کارگیری روش رد مرحله‌ای تک تک به کمک Ridge، هر بار یکی از داده‌های مورد مطالعه (۱ ترکیب از ۶۳ ترکیب) کنار گذاشته شد و مدل Ridge با استفاده از باقی‌مانده داده‌ها (۶۲ ترکیب) توسعه داده شد. این فرآیند برای همه داده‌ها تکرار شد، به‌طوری که در نهایت ۶۳ مدل Ridge با ۳۹۲ توصیف کننده با ضرایب رگرسیونی متفاوت به‌دست آمد. به‌منظور انجام فرآیند غربالگری، ابتدا مقادیر میانگین ضرایب توصیف کننده j ام ($\overline{\beta_j}$) و انحراف استاندارد ضرایب توصیف کننده j ام (σ_{β_j}) مطابق با رابطه ۲-۱ و رابطه ۲-۲ محاسبه شدند. در این روابط β_{ij} و N به‌ترتیب نمایانگر ضرایب توصیف کننده j ام در مدل i ام و تعداد ضرایب توصیف کننده j ام در مدل i ام است.

رابطه ۲-۱

$$\bar{\beta}_j = \frac{\sum_{i=1}^{N} \beta_{ij}}{N}$$

رابطه ۲-۲

$$\sigma_{\beta_j} = \sqrt{\frac{\sum_{i=1}^{N} (\beta_{ij} - \bar{\bar{\beta}}_j)^2}{N-1}}$$

برای استفاده از تکنیک غربالگری Ridge، نیاز به یک معیار مناسب جهت تشخیص حد آستانه[1] برای شناسایی توصیف کننده‌های زائد (R) و اصلی (X) از یک‌دیگر و ایجاد ماتریس‌های جدید X، R و XR می‌باشد (طبق تعریف بخش ۱- ۳- ۴- ۲)، تا توصیف کننده‌های معنا دار از بی معنا متمایز گردد. معیار C (قدر مطلق نسبت میانگین ضرایب توصیف کننده j ام به انحراف استاندارد همان توصیف کننده) به‌عنوان معیار شناسایی و ایجاد ماتریس‌های R، X و XR مطابق با رابطه ۲-۳ مورد استفاده قرار گرفت. مقدار ماکسیمم C برای ماتریس R به‌عنوان حد آستانه برای شناسایی و ایجاد ماتریس‌های جدید X، R و XR تعریف شد. به‌طوری‌که پس از محاسبه پارامتر C مربوط به توصیف کننده‌های X موجود در ماتریس XR، توصیف کننده‌هایی که مقدار C کم‌تری نسبت به ماکسیمم C ماتریس R دارند، به‌عنوان توصیف کننده‌های زائد (R) جدید شناخته می‌شوند و از ماتریس X حذف می‌گردند و به‌عنوان ماتریس R جدید شناخته می‌شوند و سایر توصیف کننده‌ها با C بزرگتر از C ماکسیمم ماتریس R به‌عنوان ماتریس X جدید شناخته می‌شوند. بدین ترتیب ماتریس X، R و XR جدید برای هر مرحله به این روش ایجاد شد.

رابطه ۲-۳

$$C = \left| \frac{\bar{\beta}_j}{\sigma_{\beta_j}} \right|$$

لازم به ذکر است در مرحله اول که روی داده‌های اصلی (۳۹۲) روش Ridge اجرا می‌شود، ماتریس زائدی (R) وجود ندارد که برای آن پارامتر C محاسبه گردد، از این رو توصیف کننده‌هایی که میانگین ضرایب رگرسیونی آن‌ها با انحراف استانداردشان برابر بود ($C_0 : \bar{\beta}_j = \sigma_{\beta_j}$)، یا به‌عبارتی نسبت میانگین به انحراف استاندارد برابر یک یا کوچک‌تر از یک داشتند ($C_0 \leq 1$)، به‌عنوان توصیف کننده‌های بی معنای اولیه (R) شناخته شدند و به این صورت اولین ماتریس R ایجاد شد. توصیف کننده‌های باقی‌مانده به ماتریس X

[1] Threshold

تعلق دارند و همچنین ماتریس XR به‌ترتیب از مجموع توصیف کننده‌های X و R تشکیل شد. طبق فلوچارت (شکل ۲-۳)، مدل Ridge با ماتریس‌های جدید X (ماتریس X اول در غیاب توصیف کننده‌های R جدید)، R (متشکل از توصیف کننده‌هایی با پارامتر C ماتریس X کوچک‌تر از C ماکسیمم ماتریس R) و XR (متشکل از توصیف کننده‌های به‌ترتیب ماتریس X جدید و توصیف کننده‌های R جدید) با روند اشاره شده در شکل ۲-۳ تشکیل شدند. ساخت مدل‌های Ridge و ایجاد ماتریس‌های R، X و XR جدید تا جایی ادامه می‌یابد که دیگر توصیف کننده‌ای C ماتریس X ای (C_x) کوچک‌تر از C ماکسیمم محاسبه شده از ماتریس R جدید (C_{maxR}) وجود نداشته باشد و در این مرحله غربالگری به اتمام می‌رسد، زیرا ساخت ماتریس R جدید دیگر امکان پذیر نیست. در ادامه بحث جزئیات مربوط به فلوچارت شکل ۲-۳ به‌طور مرحله‌ای توضیح داده شده است.

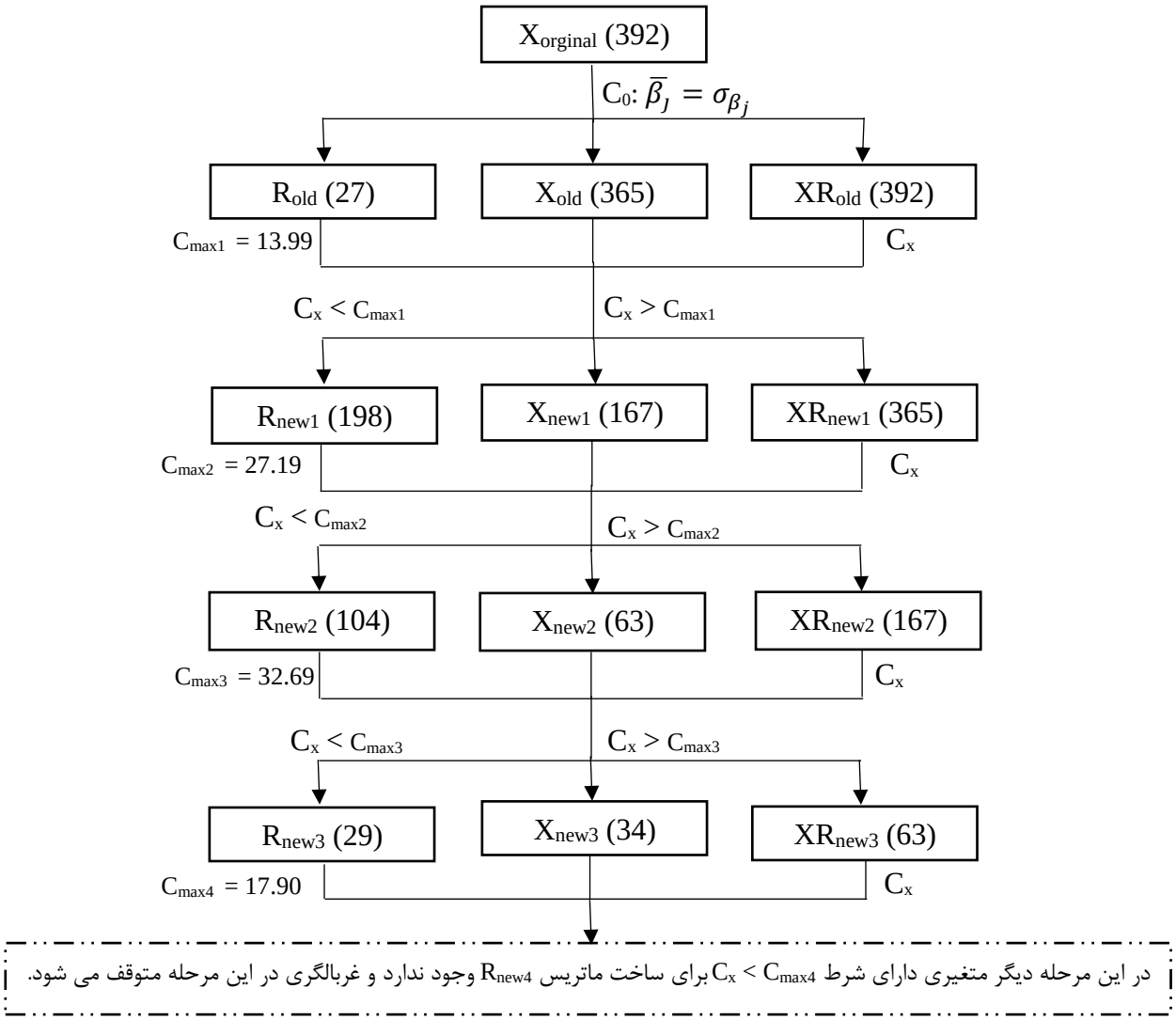
همان‌طور که در شکل ۲-۳ نشان داده شده است، در مرحله اول که ایجاد ماتریس R غیر ممکن است. روش Ridge با تکنیک رد مرحله‌ای تک تک با j توصیف کننده شامل ۳۹۲ توصیف کننده ($X_{orginal}$) اجرا شد. پس از این‌که ضرایب رگرسیونی Ridge مربوط به ۳۹۲ توصیف کننده به‌دست آمد، از j توصیف کننده، ۲۷ توصیف کننده (با شرط $C_0 \leq 1$)، جزء ماتریس R (R_{old})، ۳۶۵ توصیف کننده جزء ماتریس X (X_{old}) و ۳۹۲ توصیف کننده متعلق به ماتریس XR هستند (XR_{old}).

پس از ایجاد ماتریس‌های X، R و XR مرحله قبل، Ridge با تکنیک رد مرحله‌ای تک تک با j توصیف کننده مرحله قبل (XR_{old}) اجرا شد، ضرایب رگرسیونی برای j توصیف کننده محاسبه شد و پارامتر C (رابطه ۲-۳) برای ضرایب ۳۶۵ توصیف کننده X موجود در ماتریس XR و ۲۷ توصیف کننده ماتریس R به‌دست آمد. ماکسیمم پارامتر C ماتریس R به‌عنوان معیار غربالگری انتخاب شد. بنابراین از ۳۹۲ توصیف کننده مرحله قبل ۱۶۷ توصیف کننده متعلق به توصیف کننده‌های ماتریس X موجود در ماتریس XR، دارای پارامتر C بزرگتر از ماکسیمم پارامتر C ($C_x > C_{max1}$) ماتریس R است و به‌عنوان ماتریس X جدید

شناخته می‌شود (X_{new1}). در این مرحله، ماتریس R (R_{new1}) متشکل از ۱۹۸ توصیف کننده می‌باشد. در نهایت ماتریس XR جدید (XR_{new1}) این مرحله نیز با j توصیف کننده شامل ۳۶۵ توصیف کننده ایجاد شد.

در مرحله سوم، Ridge با تکنیک رد مرحله‌ای تک تک با استفاده از j توصیف کننده (۳۶۵ توصیف کننده ماتریس XR_{new1}) اجرا شد و ضرایب رگرسیونی برای j توصیف کننده محاسبه شد و پارامتر C (رابطه ۲-۳) برای ضرایب ۱۶۷ توصیف کننده X موجود در ماتریس XR (شامل ۳۶۵ توصیف کننده ماتریس XR_{new1}) و ۱۹۸ توصیف کننده ماتریس R (R_{new1}) به‌دست آمد. از ۳۶۵ توصیف کننده مرحله قبل ۶۳ توصیف کننده دارای پارامتر C بزرگتر از ماکسیمم پارامتر C ($C_x > C_{max2}$) ماتریس R (R_{new2}) است و به‌عنوان ماتریس X جدید شناخته می‌شود (X_{new2}). در این مرحله، ماتریس R (R_{new2}) متشکل از ۱۰۴ توصیف کننده می‌باشد. ماتریس XR جدید این مرحله (XR_{new2}) نیز با j توصیف کننده شامل ۱۶۷ توصیف کننده به‌دست آمد.

در مرحله بعد مجدداً Ridge با تکنیک رد مرحله‌ای تک تک با j توصیف کننده مرحله چهارم(شامل ۶۳ توصیف کننده (XR_{new3})) اجرا شد و ضرایب رگرسیونی برای j توصیف کننده محاسبه شد و پارامتر C (رابطه ۲-۳) برای ضرایب ۳۴ توصیف کننده X موجود در ماتریس XR و ۲۹ توصیف کننده ماتریس R (R_{new3}) به‌دست آمد. در این مرحله هیچ توصیف کننده‌ای دارای پارامتر C کوچک‌تر از ماکسیمم پارامتر C ($C_x < C_{max3}$) ماتریس R (R_{new3}) نبوده و تشکیل ماتریس R جدید غیر ممکن است. بنابراین اجرای Ridge با تکنیک رد مرحله‌ای تک تک پس از اجرای ۴ مرحله متوالی متوقف شد و در نهایت ۳۴ توصیف کننده (توصیف کننده‌های موجود در ماتریس X_{new3}) به‌عنوان متغیرهای مستقل باقی‌مانده با بیش‌ترین تأثیر با پاسخ (فعالیت دارویی) شناسایی شد. این توصیف کننده‌ها به‌عنوان ورودی مدل‌سازی انقباضی SCAD به‌کار گرفته شدند.



شکل ۲-۳ فلوچارت نمایش روند غربالگری Ridge (زیروند Orginal مربوط به توصیف کننده های مرحله پیش پردازش، زیروند old مربوط به توصیف کننده های اولین مرحله غربالگری Ridge و زیروند new مربوط به توصیف کننده های حاصل از هر مرحله اجرای روند غربالگری Ridge است.)

۲- ۲- ۶ انتخاب متغیر و مدل سازی SCAD

در این بخش به‌منظور یافتن مؤثرترین توصیف کننده‌ها و ساختار مدل QSAR از روش انقباضی SCAD استفاده شد. ۳۴ توصیف کننده‌ی باقی‌مانده از مرحله غربالگری Ridge به‌عنوان ورودی مدل در نظر گرفته شد. لازم به‌ذکر است که در مراحل پیش‌پردازش اولیه و غربالگری از ۶۳ داده برای انتخاب بهترین توصیف کننده‌ها استفاده شد و ضرایب استخراج شد. در این مرحله ۶۳ داده به دو مجموعه تقسیم می‌شود. پس از تقسیم بندی داده‌ها به دو مجموعه آموزش و آزمون (مطابق با روش بخش ۱- ۳- ۱) از مجموعه آموزش برای انتخاب مؤثرترین توصیف کننده‌ها و مدل‌سازی استفاده شد. بنابراین روش SCAD با به-کارگیری بسته نرم افزاری ncvreg در نرم افزار R-Studio با تکنیک ۱۰ فولد- ارزیابی تقاطعی[1]، به‌طوری که برای یک λ داده شده، مجموعه داده‌های آموزش (۵۳) به‌طور تصادفی به ۱۰ فولد اختصاص می‌یابد. سپس یک فولد (فولد اول) به‌عنوان مجموعه آزمون و سایر فولدها (۹ فولد دیگر) به‌عنوان مجموعه آموزش در نظر گرفته می‌شود. پس از برازش مدل با استفاده از مجموعه آموزش (۹ فولد)، مقدار متغیر پاسخ (فعالیت دارویی) به‌ازای داده‌های آزمون (فولد اول) پیش‌بینی می‌شود و خطای پیش‌بینی (PE) مجموعه آزمون (فولد اول) طبق رابطه ۲-۴ به‌دست می‌آید.

رابطه ۲-۴
$$PE_1 = \sum_{i \in I_1} (y_i - \hat{y}_i)^2$$

در رابطه ۲-۴، $i \in I_1$ است، که I_1 اندیس ترکیبات گروه اول است. در مرحله‌ی بعد، یک فولد دیگر به‌عنوان مجموعه آزمون کنار گذاشته می‌شود و مابقی فولدها به‌عنوان مجموعه آموزش استفاده می‌شوند و خطای پیش‌بینی (PE) محاسبه می‌شود. این روند k بار تکرار می‌شود و در هر تکرار یکی از فولدها به‌عنوان داده‌های مجموعه آزمون و بقیه فولدها به‌عنوان داده‌های مجموعه آموزش در نظر گرفته می‌شوند. در نتیجه

[1] 10- fold Cross-validation

k برآورد خطای $PE_1, \dots, PE_k$ آزمون به‌دست آمده و در نهایت، امتیاز ارزیابی تقاطعی (CV) به‌صورت رابطه ۲-۵ محاسبه می‌شود:

$$CV(\lambda) = \frac{1}{k}\sum_{i=1}^{k} PE_i$$ رابطه ۲-۵

این کار برای λ های مختلف تکرار شده و سپس λ ای انتخاب می‌شود که مقدار خطای پیش‌بینی را مینیمم کند. با توجه به λ_{min} مدل با حداقل خطای پیش‌بینی ارزیابی تقاطعی انتخاب شد [۶۸]. در نهایت با اجرای روش SCAD، تعداد ۹ توصیف کننده انتخاب شد. مدل SCAD (رابطه ۲-۶) با استفاده از توصیف کننده‌های منتخب توسعه یافت. نماد، طبقه بندی و معنای هر یک از این توصیف کننده‌ها در جدول ۲-۳ آورده شده است.

$pIC_{۵۰} = ۷/۴ + ۰/۵۷\ DECC + ۰/۲۵\ MATS4e - ۰/۰۴\ C025 - ۰/۹۶\ Mor26p - ۰/۰۳\ H046 - ۰/۰۲$ رابطه ۲-۶

$N075 - ۰/۰۰۱\ SEigZ + ۰/۵۲\ Mor23m - ۲/۴\ GATS4e$

به‌منظور بررسی دقیق‌تر توصیف‌کننده‌های منتخب مدل Ridge-SCAD، احتمال وجود همبستگی با محاسبه مربع ضریب همبستگی بین دو توصیف کننده در جدول ۲-۴ مورد بررسی قرار گرفت. نتایج نشان می‌دهد که بین توصیف‌کننده‌های منتخب همبستگی معنا داری وجود ندارد. علاوه بر این احتمال وجود هم‌خطی بین توصیف کننده‌های منتخب با محاسبه پارامتر مقادیر افزایش تورم واریانس (VIF) (مطابق رابطه ۱-۳۴) مورد بررسی قرار گرفت و نتایج آن در ستون آخر جدول ۲-۴ گزارش شده است. مقادیر VIF کمتر از ۱۰ توصیف کننده‌های مدل Ridge-SCAD بیانگر عدم وجود هم‌خطی بین توصیف‌کننده‌ها می‌باشد [۲۵-۲۲].

جدول ۲-۳ توصیف کننده های انتخاب شده توسط روش Ridge-SCAD

ردیف	نماد	طبقه‌بندی	معنا
۱	DECC	topological	Eccentric
۲	C-025	Atom-centred fragment	R-CR-R
۳	GATS4e	2D autocorrelation	Geary autocorrelation – lag 4
۴	Mor26p	3D-MoRSE	3D-MoRSE – signal 26
۵	Mor23m	3D-MoRSE	3D-MoRSE – signal 23
۶	H-046	Atom-centred fragment	H attached to C (sp^3) no X attached to next C
۷	MATS4e	2D autocorrelation	Moran autocorrelation – lag 4
۸	N-075	Atom-centred fragment	R-N-R / R-N-X
۹	SEigZ	Eigenvalue-based indices	Eigenvalue sum from Z

جدول ۲-۴ ماتریس همبستگی و مقادیر عامل افزایش واریانس توصیف کننده های انتخاب شده مدل Ridge-SCAD

توصیف کننده‌ها	DECC	C-025	GATS4e	Mor26p	Mor23m	H-046	MATS4e	N-075	SEigZ	VIF
DECC	۱									۳/۷۷۹
C-025	-۰/۰۰۶	۱								۱/۲۵۳
GATS4e	-۰/۳۰۶	-۰/۱۳۶	۱							۱/۵۷۵
Mor26p	-۰/۵۳۵	-۰/۱۲۱	۰/۳۱۵	۱						۱/۷۱۸
Mor23m	۰/۵۳۲	-۰/۱۵۸	-۰/۰۵	-۰/۳۵۶	۱					۱/۹۴۷
H-046	-۰/۲۱۷	-۰/۳۱	۰/۴۴۳	۰/۲۹۳	-۰/۰۵۷	۱				۱/۵۴۹
MATS4e	۰/۶۹۱	۰/۰۷۲	-۰/۱۴۴	-۰/۵۶۳	۰/۶۴۹	-۰/۲۲۵	۱			۴/۶۱۳
N-075	-۰/۳	-۰/۱۲۴	۰/۳۸۳	۰/۳۶۷	-۰/۲۶۶	۰/۳۴۵	-۰/۴۸۴	۱		۱/۶۳
SEigZ	۰/۵۳	-۰/۱۲۲	-۰/۴۲۷	-۰/۳۰۹	۰/۱۸۵	-۰/۳۱۹	۰/۰۷۶	-۰/۱۹۴	۱	۲/۴۲۳

به‌منظور بررسی عملکرد روش Ridge برای غربالگری توصیف کننده‌ها، مدل SCAD بدون استفاده از روش غربالگری Ridge ساخته شد. پس از پیش‌پردازش‌های اولیه توصیف کننده‌ها، روش SCAD با به‌کارگیری بسته نرم افزاری ncvreg در نرم افزار R-Studio با تکنیک ۱۰ فولد- ارزیابی تقاطعی اجرا شد. به‌طوری که برای یک λ داده شده، مجموعه داده‌های آموزش (۵۳) به‌طور تصادفی به ۱۰ فولد اختصاص می‌یابد. سپس فولد اول به‌عنوان مجموعه آزمون و سایر فولدها (۹) به‌عنوان مجموعه آموزش در نظر گرفته

می‌شود. پس از برازش مدل با استفاده از مجموعه آموزش، مقدار متغیر پاسخ (فعالیت دارویی) به‌ازای داده‌های آزمون پیش‌بینی می‌شود و خطای پیش‌بینی طبق رابطه ۲-۴ به‌دست می‌آید. این روند k بار تکرار می‌شود و در هر تکرار یکی از فولدها به‌عنوان داده‌های آزمون و بقیه فولدها به‌عنوان داده‌های آموزش در نظر گرفته می‌شوند. در نتیجه k برآورد خطای $PE_1, \dots, PE_k$ آزمون به‌دست آمده و در نهایت، امتیاز ارزیابی تقاطعی (CV) به‌صورت رابطه ۲-۵ محاسبه می‌شود. این کار برای λ های مختلف تکرار شده و سپس λ ای انتخاب می‌شود که مقدار خطای پیش‌بینی را مینیمم کند. با توجه به λ_{min} مدل با حداقل خطای پیش‌بینی ارزیابی تقاطعی انتخاب شد. با اجرای روش SCAD، تعداد ۱۲ توصیف کننده انتخاب شدند. نام و نوع توصیف‌کننده‌های منتخب روش SCAD در جدول ۲-۵ خلاصه شده است. مدل SCAD در رابطه ۲-۷ آورده شده است.

رابطه ۲-۷ $pIC_{50} = 4/2 + 1/1\ DECC - 2/3\ MATS2p - 0/55\ Mor22u + 0/27\ Mor09m + 0/05\ Mor23m - 0/02$

$Mor30m - 0/09\ Mor32v + 0/43\ G3m + 0/04\ E1m - 0/27\ G3p - 0/43\ C025 - 0/24\ LAI$

جدول ۲-۵ توصیف کننده های انتخاب شده توسط روش SCAD

ردیف	نماد	طبقه‌بندی	معنا
۱	DECC	topological	eccentric
۲	C-025	Atom-centred fragment	R-CR-R
۳	Mor22u	3D-MoRSE	3D-MoRSE – signal 22
۴	Mor09m	3D-MoRSE	3D-MoRSE – signal 09
۵	MATS2p	2D autocorrelation	Moran autocorrelation – lag 2
۶	LAI	molecular properties	Lipiniski Alert index.
۷	Mor23m	3D-MoRSE	3D-MoRSE – signal 23
۸	Mor32v	3D-MoRSE	3D-MoRSE – signal 32
۹	G3m	WHIM	3st component symmetry directional WHIM index
۱۰	Mor30m	3D-MoRSE	3D-MoRSE – signal 30
۱۱	E1m	WHIM	1st component accessibility directional WHIM index
۱۲	G3p	WHIM	3st component symmetry directional WHIM index

۲- ۲- ۷ ارزیابی مدل Ridge-SCAD

از مرحله‌های اساسی ساخت مدل QSAR، ارزیابی مدل است. در این پروژه، قدرت پیش‌بینی و تعمیم پذیری مدل توسعه یافته Ridge-SCAD با استفاده از پیش‌بینی فعالیت دارویی داده‌های مجموعه آزمون، فعالیت دارویی کل داده‌های پیش‌بینی شده به‌وسیله تکنیک رد مرحله‌ای تک تک (LOO)، محاسبه پارامترهای آماری، آزمون‌های دامنه کاربرد و Y-تصادفی مورد ارزیابی قرار گرفت.

۲- ۲- ۷- ۱ ارزیابی مدل Ridge-SCAD و مدل SCAD با استفاده از پیش بینی داده های مجموعه آزمون

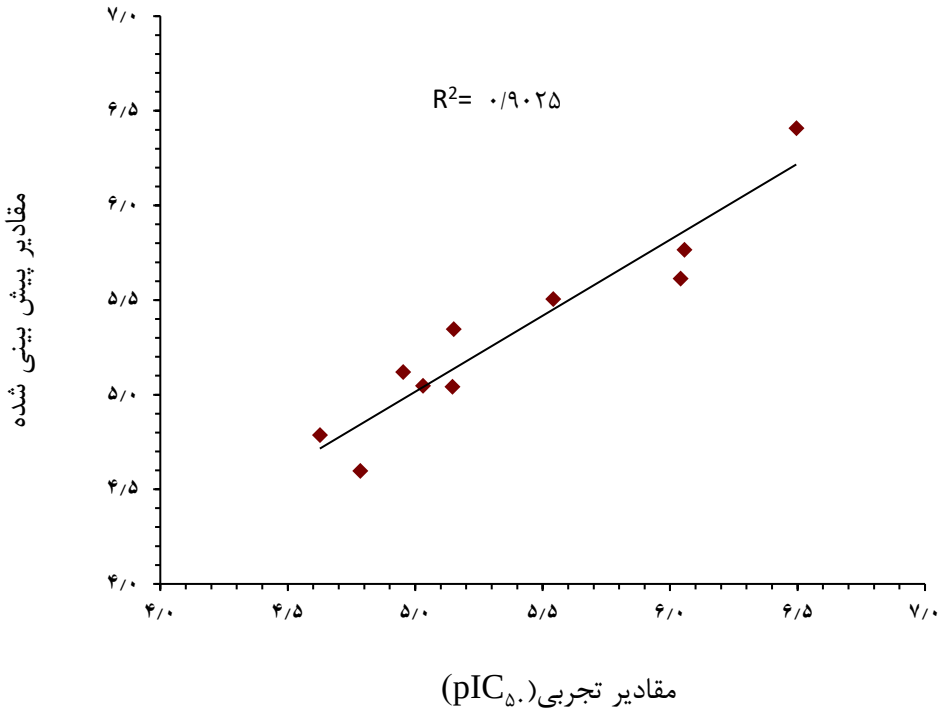
اعتبار و اعتمادپذیری[1] مدل توسعه یافته Ridge-SCAD با استفاده از داده‌های مجموعه آزمون که در ساخت مدل شرکت نداشته‌اند، سنجیده شد. بنابراین فعالیت دارویی ($pIC_{۵۰}$) ترکیبات مجموعه آزمون (۱۰ ترکیب) مشخص شده در جدول ۲-۶ با استفاده از مدل Ridge-SCAD (رابطه ۲-۶) پیش‌بینی شدند. مقادیر پیش‌بینی شده در جدول ۲-۶ خلاصه شده است. خطای کم پیش‌بینی شده، دلالت بر قدرت پیش‌بینی مدل Ridge-SCAD دارد. علاوه بر این مقادیر پیش‌بینی شده فعالیت دارویی برحسب مقادیر واقعی متناظر آن‌ها رسم شد (شکل ۲-۴). R^2 مربوط به داده‌های مجموعه آزمون برابر با ۰/۹۰۲۵ به دست آمد. مقدار R^2 بزرگ‌تر از ۰/۶ نشان‌دهنده قدرت پیش‌بینی مدل Ridge-SCAD است. نزدیکی R^2 مربوط به مجموعه آزمون (R^2_{test}= ۰/۹۰۲۵) به R^2 مربوط به مجموعه آموزش (R^2_{train}= ۰/۹۳۱۷) تعمیم پذیری مدل توسعه یافته Ridge-SCAD را اثبات می‌کند. علاوه بر این مقادیر باقی مانده‌های پیش‌بینی شده از محاسبه تفاضل مقادیر پیش‌بینی شده و تجربی به‌دست آمد. نمودار باقی مانده‌ها بر حسب مقادیر تجربی رسم شد (شکل ۲-۵). توزیع مقادیر باقی‌مانده‌ها حول محور صفر، نشان‌دهنده عدم وجود خطای سیستماتیک در مدل توسعه یافته Ridge-SCAD است.

[1] Reliability

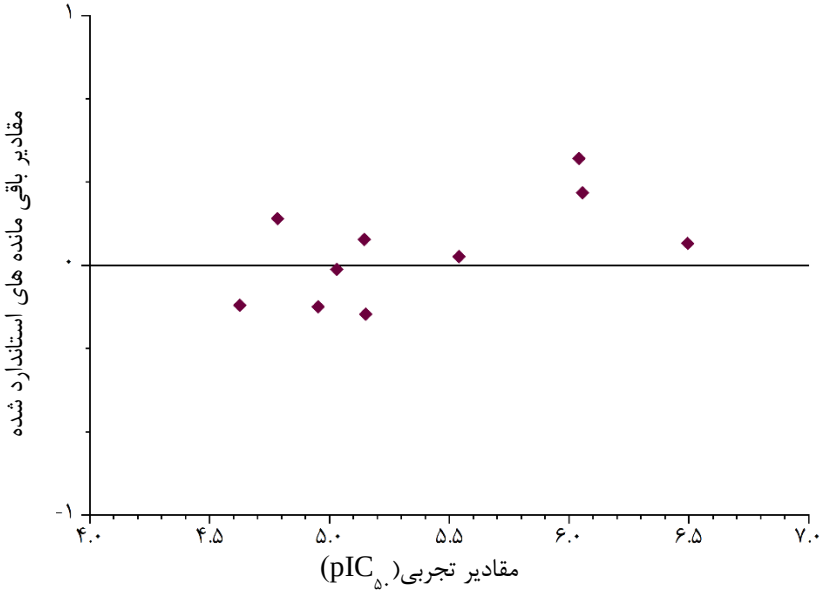
همچنین، به‌منظور بررسی عملکرد روش Ridge برای غربالگری توصیف کننده‌ها، ارزیابی مدل SCAD بدون استفاده از روش غربالگری Ridge با استفاده از ترکیبات مجموعه آزمون (۱۰ ترکیب) مشخص شده در جدول ۲-۶ و با مدل SCAD (رابطه ۲-۷) پیش‌بینی شدند. مقادیر خطای پیش‌بینی مدل SCAD نسبت به مدل Ridge-SCAD در برخی از ترکیبات مجموعه آزمون افزایش یافته است که این امر دلالت بر قدرت پیش‌بینی کم‌تر مدل SCAD دارد. علاوه بر این مقادیر پیش‌بینی شده فعالیت دارویی برحسب مقادیر واقعی متناظر آن‌ها رسم شد (شکل ۲-۶). R^2 مربوط به داده‌های مجموعه آزمون (R^2_{test}=۰/۸۰۴۲) و مجموعه آموزش (R^2_{train}= ۰/۹۴۹۱) به‌دست آمد که هر دو دارای مقادیر قابل قبول هستند، اما نزدیکی مقادیر R^2 دو مجموعه کم‌تر از روش Ridge-SCAD است. با مقایسه نتایج قابلیت بالای Ridge در کاهش ابعاد داده که منجر به انتخاب متغیرهای مؤثرتر و بهتر می‌شود، مشهود است.

جدول ۲-۶ نتایج حاصل از پیش بینی فعالیت دارویی داده های مجموعه آزمون توسط مدل Ridge-SCAD و مدل SCAD در مجموعه داده ها

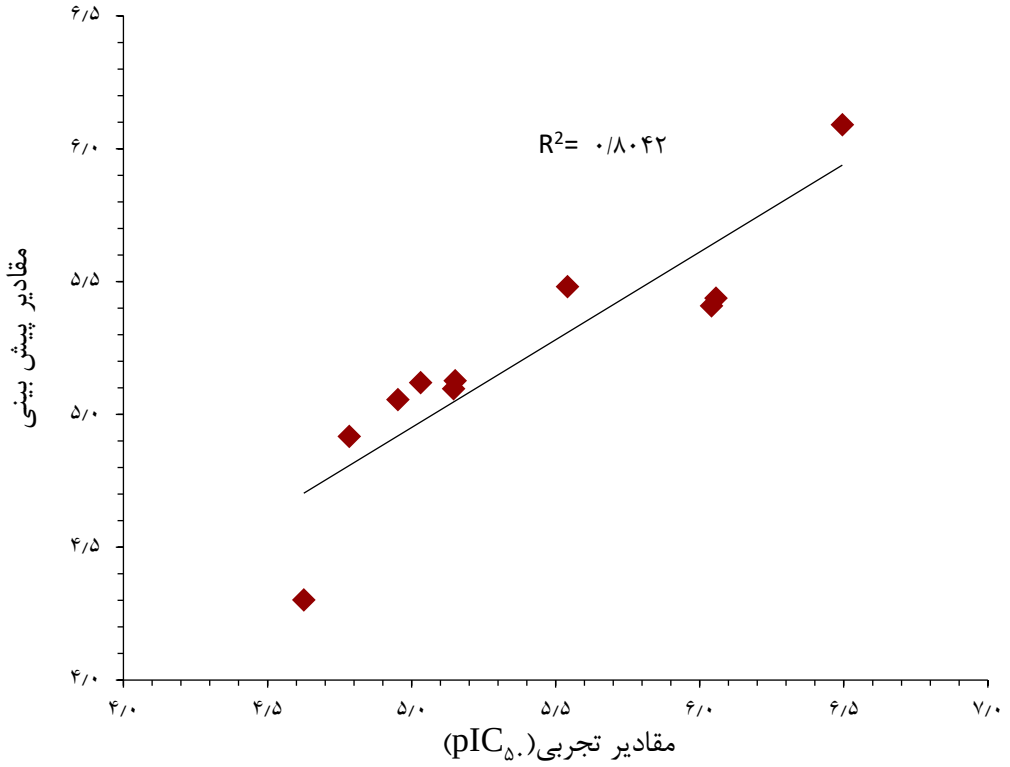
شماره ترکیب در مجموعه داده‌ها	مقدار پیش‌بینی شده ($pIC_{۵۰}$)		درصد خطا	
	مقدار تجربی ($pIC_{۵۰}$)	Ridge-SCAD	SCAD	Ridge-SCAD	SCAD
۲	۵/۱۵	۵/۳۵	۵/۱۳	-۳/۸	۰/۴۸
۴	۴/۹۵	۵/۱۲	۵/۰۶	-۳/۳	-۲/۱
۱۰	۴/۶۳	۴/۷۹	۴/۳۰	-۳/۵	۷/۰
۱۹	۴/۷۸	۴/۶۰	۴/۹۲	۳/۹	-۲/۸
۲۶	۶/۰۴	۵/۶۱	۵/۴۱	۷/۱	۱۰
۲۸	۶/۰۶	۵/۷۶	۵/۴۴	۴/۸	۱۰
۳۲	۶/۴۹	۶/۴۱	۶/۰۹	۱/۴	۶/۲
۴۶	۵/۱۵	۵/۰۴	۵/۱۰	۲/۰	۰/۹۷
۴۹	۵/۵۴	۵/۵۱	۵/۴۸	۰/۶۳	۱/۱
۵۵	۵/۰۳	۵/۰۵	۵/۱۲	-۰/۳۲	-۱/۸



شکل ۲-۴ نمودار تغییرات مقادیر پیش بینی شده ($pIC_{۵۰}$) مدل Ridge-SCAD در مقابل مقادیر تجربی برای مجموعه آزمون



شکل ۲-۵ نمودار مقادیر باقی مانده های پیش بینی فعالیت دارویی با استفاده از مدل Ridge-SCAD و مقادیر تجربی بر حسب مقادیر تجربی



شکل ۲-۶ نمودار تغییرات مقادیر پیش بینی شده ($pIC_{۵۰}$) مدل SCAD در مقابل مقادیر تجربی برای مجموعه آزمون

۲- ۲- ۷- ۲ ارزیابی مدل Ridge-SCAD با استفاده از تکنیک رد مرحله ای تک تک

در این بخش، یکی از تکنیک‌های قدرتمند برای ارزیابی مدل بهینه و بررسی قدرت پیش‌بینی مدل به‌نام تکنیک رد مرحله‌ای تک تک (LOO) برای پیش‌بینی همه داده‌های مورد مطالعه به‌عنوان داده آزمون مورد استفاده قرار گرفت. در این مطالعه، هر بار یکی از داده‌های مورد مطالعه کنار گذاشته شد و مدل Ridge-SCAD با سایر داده‌ها آموزش داده شد و فعالیت دارویی ترکیب کنار گذاشته با مدل آموزش داده شده، پیش‌بینی گردید. این فرآیند برای همه ترکیبات انجام شد و همه داده‌ها یک‌بار به‌عنوان داده آزمون در نظر گرفته شدند. ازاین‌رو فعالیت دارویی همه ترکیبات تکرار شد، به‌گونه‌ای که تکنیک LOO توسط مدل Ridge-SCAD پیش‌بینی شدند و نتایج حاصل در جدول ۲-۷ آورده شده‌اند. برای مطالعه بیش‌تر پایداری قدرت پیش‌بینی مدل، نمودار مقادیر پیش‌بینی شده $pIC_{۵۰}$ تمام ترکیبات بر حسب مقادیر تجربی

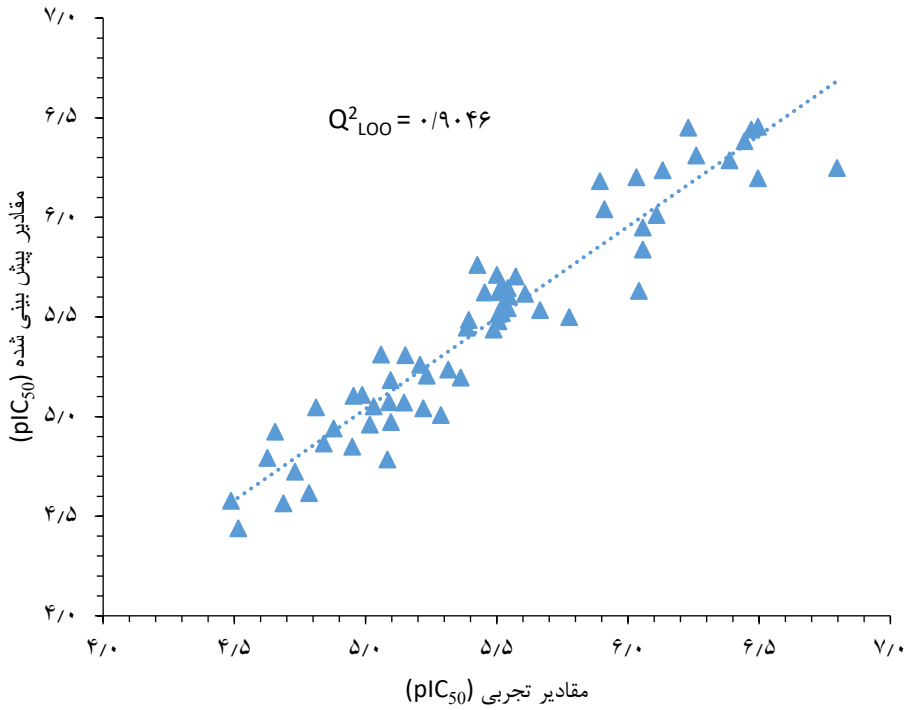
آن‌ها رسم شد. مقدار Q^2_{LOO} که در شکل ۲-۷ نشان داده شده است، بیش‌تر از مقدار قابل قبول ($Q^2 >$ ۰/۵) بوده که نشان‌دهنده این است که مدل توسعه یافته از پایداری قدرت پیش‌بینی و تعمیم پذیری مناسبی برخوردار است. علاوه بر این، به‌منظور بررسی عدم وجود خطای سیستماتیک در مدل توسعه یافته، مقادیر باقی‌مانده بر حسب مقادیر تجربی رسم شد (شکل ۲-۸). نتایج حاصل نشان‌دهنده توزیع یکنواخت و تصادفی داده‌ها حول محور صفر می‌باشد که بیانگر عدم وجود خطای سیستماتیک در مدل Ridge-SCAD است.

جدول ۲-۷ نتایج حاصل از ارزیابی مدل Ridge-SCAD به روش رد مرحله ای تک تک برای کل داده ها

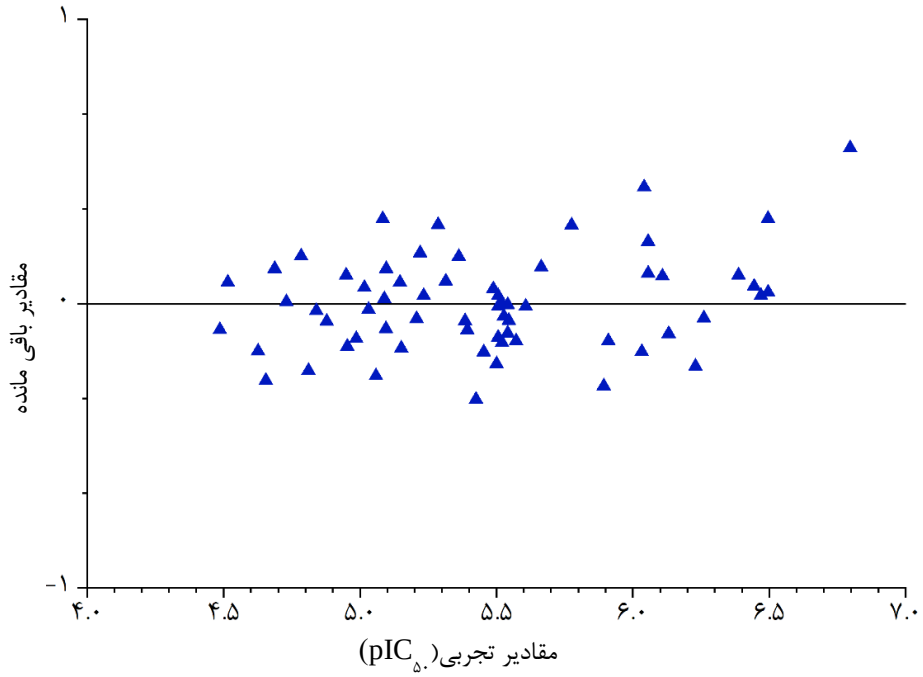
pIC۵۰			شماره ترکیب	pIC۵۰			شماره ترکیب
درصد خطا	مقدار پیش‌بینی شده	مقدار واقعی		درصد خطا	مقدار پیش‌بینی شده	مقدار واقعی	
۳/۵۱	۴/۶۲	۴/۷۸	۱۹	۱/۱۵	۴/۹۶	۵/۰۲	۱
۲/۰۳	۴/۸۵	۴/۹۵	۲۰	-۳/۰۳	۵/۳۱	۵/۱۵	۲
۱/۴۹	۵/۲۳	۵/۳۱	۲۱	۰/۱۵	۴/۷۲	۴/۷۳	۳
۲/۴۲	۴/۹۷	۵/۱۰	۲۲	-۳/۰۳	۵/۱۰	۴/۹۵	۴
-۵/۰۲	۵/۳۱	۵/۰۶	۲۳	۱/۶۸	۴/۴۴	۴/۵۲	۵
-۳/۸۴	۵/۷۱	۵/۵۰	۲۴	-۵/۷۹	۴/۹۲	۴/۶۶	۶
-۰/۱۶	۵/۶۲	۵/۶۱	۲۵	۵/۸۸	۴/۷۸	۵/۰۸	۷
۶/۸۰	۵/۶۳	۶/۰۴	۲۶	۵/۲۷	۵/۰۱	۵/۲۹	۸
-۱/۷۲	۵/۱۸	۵/۱۰	۲۷	-۲/۰۴	۴/۵۸	۴/۴۹	۹
۳/۶۱	۵/۸۴	۶/۰۶	۲۸	-۳/۶۱	۴/۷۹	۴/۶۳	۱۰
-۱/۲۶	۴/۹۴	۴/۸۸	۲۹	-۳/۱۱	۵/۶۲	۵/۴۵	۱۱
-۱/۰۶	۵/۶۰	۵/۵۵	۳۰	۴/۷۹	۵/۵۰	۵/۷۸	۱۲
-۱/۷۰	۵/۴۸	۵/۳۹	۳۱	۰/۵۵	۵/۲۰	۵/۲۳	۱۳
۰/۶۳	۶/۴۵	۶/۵۰	۳۲	-۱/۱۲	۵/۴۵	۵/۳۹	۱۴
-۳/۵۳	۶/۴۵	۶/۲۳	۳۳	-۰/۵۰	۴/۸۶	۴/۸۴	۱۵
۱/۵۸	۶/۲۹	۶/۳۹	۳۴	-۲/۴۶	۵/۱۱	۴/۹۹	۱۶
۱/۵۹	۶/۰۱	۶/۱۱	۳۵	۲/۶۳	۴/۵۶	۴/۶۹	۱۷
۰/۹۵	۶/۳۸	۶/۴۴	۳۶	-۴/۸۸	۵/۰۵	۴/۸۱	۱۸

ادامه جدول ۲-۷

درصد خطا	pIC_{50}		شماره ترکیب	درصد خطا	pIC_{50}		شماره ترکیب
	مقدار پیش‌بینی شده	مقدار واقعی			مقدار پیش‌بینی شده	مقدار واقعی	
-۰/۰۶	۵/۵۴	۵/۵۴	۴۹	-۱/۷۲	۶/۲۴	۶/۱۳	۳۷
-۰/۱۷	۵/۵۲	۵/۵۱	۵۰	-۰/۸۰	۶/۳۱	۶/۲۶	۳۸
۰/۳۴	۵/۰۷	۵/۰۹	۵۱	۱/۷۷	۵/۹۵	۶/۰۶	۳۹
-۰/۸۰	۵/۵۷	۵/۵۳	۵۲	۳/۱۰	۵/۲۰	۵/۳۶	۴۰
۰/۵۱	۵/۴۸	۵/۵۱	۵۳	۰/۹۵	۵/۴۴	۵/۴۹	۴۱
-۱/۸۷	۵/۶۴	۵/۵۴	۵۴	۰/۰۰	۵/۵۲	۵/۵۲	۴۲
-۰/۳۹	۵/۰۵	۵/۰۳	۵۵	-۲/۳۵	۵/۷۰	۵/۵۷	۴۳
-۴/۹۱	۶/۱۸	۵/۸۹	۵۶	-۲/۴۷	۵/۶۶	۵/۵۲	۴۴
۲/۲۸	۵/۲۳	۵/۶۶	۶۱	-۲/۱۴	۵/۶۲	۵/۵۱	۴۵
-۲/۸۰	۶/۲۰	۶/۰۳	۶۲	۱/۴۶	۵/۰۷	۵/۱۵	۴۶
۸/۰۷	۶/۲۵	۶/۸۰	۶۳	۳/۴۳	۵/۰۴	۵/۲۲	۴۷
				-۱/۰۱	۵/۲۶	۵/۲۱	۴۸



شکل ۲-۷ نمودار تغییرات مقادیر پیش بینی شده همه داده ها بر اساس تکنیک LOO در مقابل مقادیر تجربی



شکل ۲-۸ نمودار باقی مانده های حاصل از پیش بینی فعالیت دارویی ترکیبات با استفاده از تکنیک LOO و مقادیر تجربی برحسب مقادیر تجربی

۲- ۲- ۷- ۳ ارزیابی مدل Ridge-SCAD و مدل SCAD با استفاده از پارامترهای آماری

به‌منظور بررسی بیش‌تر کارایی مدل Ridge-SCAD پارامترهای آماری معرفی شده در بخش (۱-۴- ۱) محاسبه شد. کارایی مدل توسعه یافته Ridge-SCAD با محاسبه این پارامترهای آماری برای مجموعه آزمون و آموزش مورد بررسی قرار گرفت و نتایج حاصل از محاسبه پارامترهای آماری به‌همراه محدوده یا مقدار قابل قبول هر پارامتر در جدول ۲-۸ خلاصه شده اند. نتایج جدول ۲-۸ نشان می‌دهد که پارامترهای (R_0^2، R'^2_0، R_m^2 و R'^2_m) از مقدار محدوده قابل قبول ۰/۵ بزرگ‌تر و به مقدار R^2 نزدیک هستند. شیب نمودار حاصل از مقادیر پیش‌بینی شده بر حسب مقادیر تجربی (و بالعکس) در عرض از مبدأ صفر نیز در محدوده ۰/۸۵ تا ۱/۱۵ قرار دارند که این نتیجه نیز حاکی از صحت مدل توسعه یافته Ridge-SCAD می‌باشد. از آن‌جا که R^2 (ضریب تعیین)، تعداد توصیف کننده‌های موجود در مدل را به حساب نمی‌آورد، در ارزیابی و مقایسه مدل‌های مختلف با تعداد متفاوت متغیرهای مستقل (توصیف کننده)، پارامترهای آماری ضریب تعیین تصحیح شده (R^2_{adj}) و FIT استفاده شد.

جدول ۲-۸ پارامترهای آماری برای مجموعه آموزش و آزمون داده‌های پیش بینی شده مدل Ridge-SCAD

ردیف	پارامتر	مجموعه آزمون	مجموعه آموزش	محدوده قابل قبول
۱	PRESS	۰/۱۵	۱/۱	–
۲	SEP	۰/۲۰	۰/۱۴	–
۳	MAE	۰/۱۷	۰/۱۱	–
۴	REP	۳/۸	۲/۶	–
۵	MSE	۰/۰۴	۰/۰۲	–
۶	MRE	۳/۱	۱/۱	–
۷	R^2	۰/۹۰	۰/۹۳	>۰/۶
۸	R_{adj}^2	-	۰/۹۲	–
۹	FIT	-	۴/۳۸	–
۱۰	R_0^2	۰/۸۵	۰/۹۳	نزدیک به R^2
۱۱	R نسبی	۰/۰۶	۰/۰۰	<۰/۱
۱۲	R^2_m	۰/۷۰	۰/۹۳	>۰/۵
۱۳	R'^2_0	۰/۸۹	۰/۹۳	نزدیک به R^2
۱۴	R' نسبی	۰/۰۱	۰/۰۰	<۰/۱
۱۵	R'^2_m	۰/۸۰	۰/۹۳	>۰/۵
۱۶	R-R	۰/۰۴	۰/۰۰	<۰/۲
۱۷	K	۰/۹۹	۱/۰۰	۰/۸۵ < K < ۱/۱۵
۱۸	K'	۱/۰۱	۱/۰۰	۰/۸۵ < K < ۱/۱۵

برای بررسی کارایی Ridge به‌عنوان روش غربالگری توصیف کننده قبل از مدل‌سازی با SCAD، مدل SCAD ساخته شده با تمام توصیف کننده‌ها بدون غربالگری Ridge با محاسبه برخی از پارامترهای آماری برای مجموعه آزمون و آموزش مورد ارزیابی و مقایسه با مدل Ridge-SCAD قرار گرفت و نتایج حاصل در جدول ۲-۹ خلاصه شده‌اند.

جدول ۲-۹ مقایسه پارامترهای آماری محاسبه شده برای مجموعه آزمون داده های پیش بینی شده مدل Ridge-SCAD و SCAD

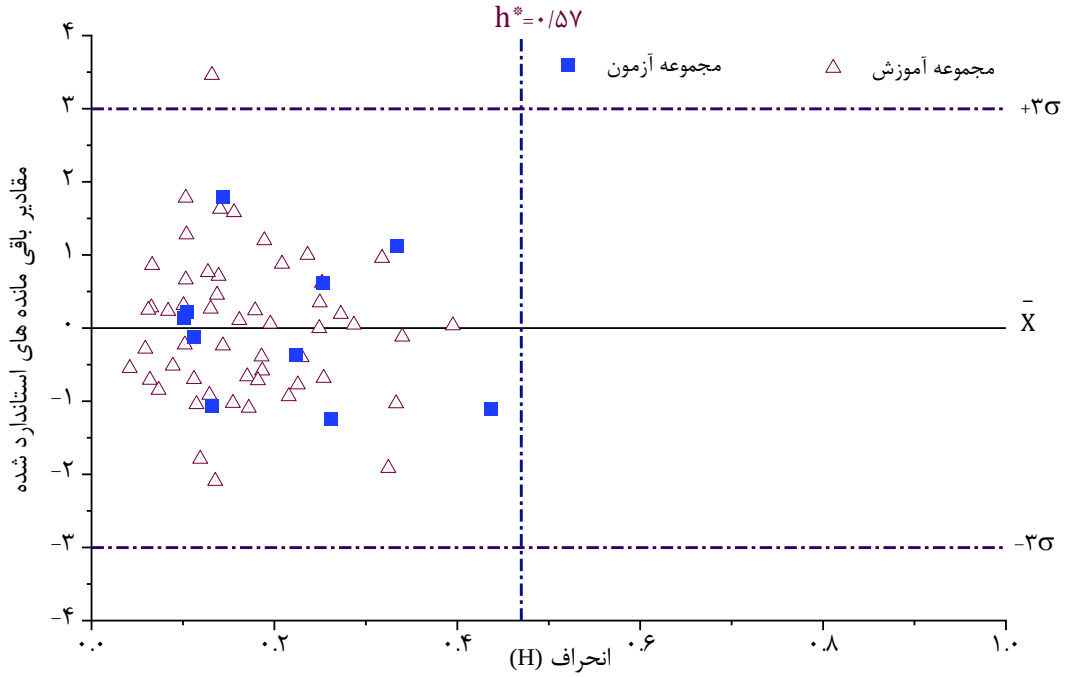
ردیف	پارامتر	Ridge-SCAD		SCAD		محدوده قابل قبول
		مجموعه آزمون	مجموعه آموزش	مجموعه آزمون	مجموعه آموزش	
۱	R^2	۰/۹۰	۰/۹۳	۰/۸	۰/۹۵	>۰/۶
۲	R_{adj}^2	–	۰/۹۲	–	۰/۹۱	–
۳	FIT	–	۴/۳۸	-	۳/۷۸	–

مطابق با نتایج جدول ۲-۹ مقدار R_{adj}^2 مدل Ridge-SCAD و مدل SCAD بزرگ‌تر از مقدار قابل قبول ۰/۵ به‌دست آمد. مقدار R_{adj}^2 مدل Ridge-SCAD با تعداد توصیف کننده‌های کم‌تر نسبت به مدل SCAD با توصیف کننده‌های بیش‌تر، به مقدار R^2 خود نزدیک‌تر است. پارامتر FIT برای هر دو مدل Ridge-SCAD و SCAD محاسبه شد، نتایج نشان می‌دهد که FIT مدل Ridge-SCAD بیش‌تر از مدل SCAD می‌باشد، بنابراین مدل Ridge-SCAD با تعداد توصیف کننده کم‌تر (تنکی بیش‌تر) دارای قدرت پیش‌بینی بیش‌تری نسبت به مدل SCAD است. مقایسه نتایج پارامترهای آماری در مدل Ridge-SCAD و SCAD (جدول ۲-۹) نشان می‌دهد که مدل Ridge-SCAD ساده‌تر، تفسیرپذیرتر (توسعه یافته با توصیف کننده‌های کم‌تری) است، در حالی‌که نتایج آماری بهتری در پیش‌بینی فعالیت دارویی ترکیبات نسبت به مدل SCAD دارد و قدرت پیش‌بینی و تعمیم‌پذیری مدل Ridge-SCAD از مدل SCAD بیش‌تر است.

۲- ۲- ۷- ۴ ارزیابی مدل با استفاده از دامنه کاربرد

از آزمون دامنه کاربرد به‌عنوان یک معیار مناسب برای بررسی اعتمادپذیری و استحکام مدل توسعه یافته Ridge-SCAD استفاده شد. دامنه کاربرد یک مدل QSAR با استفاده از نمودار ویلیام نشان داده می‌شود. به‌منظور استخراج نمودار ویلیام، مقادیر انحراف (H) مجموعه داده‌های آموزش و آزمون (مطابق با رابطه ۱-۴۰) و توضیحات (بخش ۱- ۴- ۵) به‌دست آمد. مقادیر باقی‌مانده‌های استاندارد شده نیز با استفاده

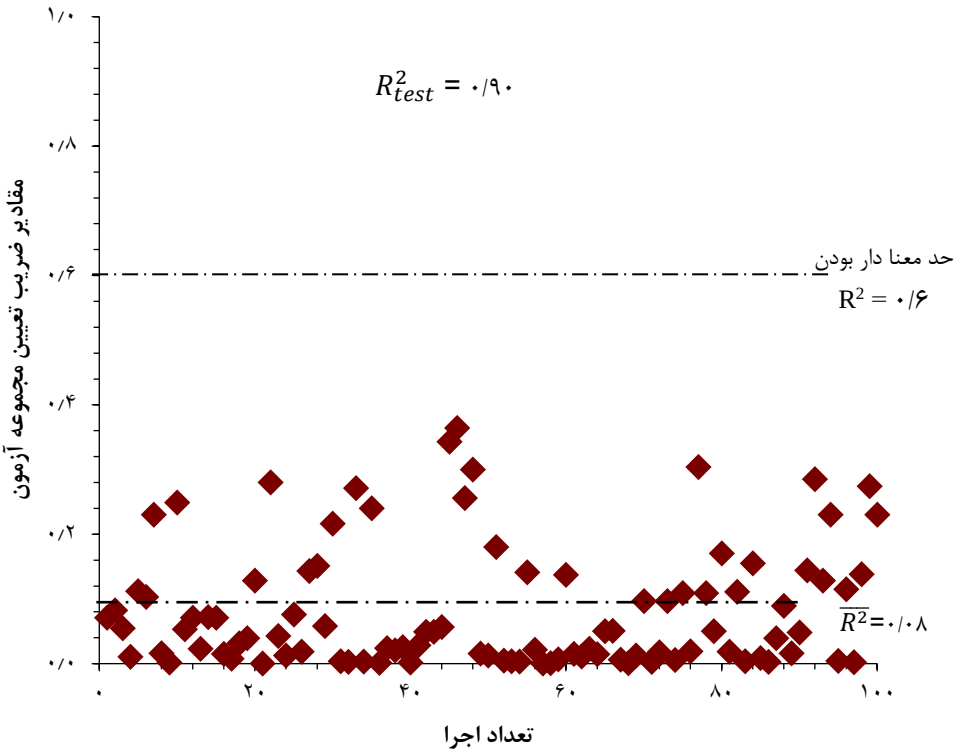
از رابطه ۱-۴۱ محاسبه شد. از رسم مقادیر باقی‌مانده‌های استاندارد شده بر حسب مقادیر انحراف (H) نمودار ویلیام به‌دست آمد. برای تفسیر و تحلیل نمودار دامنه کاربرد رعایت دو شرط الزامی است. ابتدا این‌که مقادیر باقی‌مانده‌های استاندارد شده (محور عمودی) باید در محدوده ۳ برابر بزرگ‌تر یا کوچک‌تر از انحراف استاندارد ($\pm 3\sigma$) قرار داشته باشند، تا نشان دهد باقی‌مانده به‌صورت کاملاً تصادفی توزیع شده‌اند. علاوه‌براین باید مقادیر انحراف (H) در محدوده‌ای کوچک‌تر از h^* که برابر با $3p/n$ است، قرار گیرند. همان‌طور که شکل ۲-۹ نشان می‌دهد، همه داده‌ها دارای H کوچک‌تر از h^* هستند و تنها یک داده دارای باقی‌مانده‌ای است که از محدوده $+3\sigma$ تجاوز کرده است و ۹۹٪ کل داده‌های مورد مطالعه در محدوده‌های فوق قرار دارند و داده پرت نمی‌باشند.



شکل ۲-۹ نمودار ویلیام (دامنه کاربرد) مدل Ridge-SCAD، خطوط نقطه چین افقی و عمودی در دو انتها نمودار به ترتیب نمایانگر مقادیر $\pm 3\sigma$ و h^* می باشد.

۲- ۲- ۷- ۵ ارزیابی مدل Ridge-SCAD با استفاده از Y-تصادفی

برای حصول اطمینان از تصادفی نبودن ارتباط ایجاد شده بین توصیف کننده‌های نهایی و فعالیت دارویی به‌وسیله مدل Ridge-SCAD، از آزمون Y-تصادفی استفاده شد. برای انجام این آزمون، مقادیر فعالیت دارویی ترکیبات مورد مطالعه در محدوده ۴/۴۹ تا ۶/۸۰ به تعداد ۱۰۰ بار به‌طور کاملاً تصادفی تغییر داده شدند. مدل Ridge-SCAD با استفاده از پاسخ‌های تصادفی فعالیت دارویی ترکیبات آموزش ساخته شد و مدل توسعه یافته با استفاده از پاسخ‌های تصادفی شده برای پیش‌بینی فعالیت دارویی ترکیبات مجموعه آزمون مورد استفاده قرار گرفت و مقادیر R^2 برای نتایج حاصل محاسبه گردید. نتایج R^2 حاصل از پیش‌بینی $pIC_{۵۰}$ ترکیبات مجموعه آزمون با ۱۰۰ مدل توسعه یافته با متغیر وابسته تصادفی، درشکل ۲-۱۰ آورده شده است. همان‌طور که نتایج شکل ۲-۱۰ نشان می‌دهد، مقادیر R^2 حاصل از پیش‌بینی فعالیت دارویی ترکیبات توسط مدل توسعه یافته تصادفی مجموعه آزمون کوچک‌تر از مقدار قابل قبول ۰/۶ است و میانگین ۱۰۰ تکرار نیز برابر با ۰/۰۸ است که در نمودار مشخص شده است. بنابراین مقادیر بسیار کم R^2، عدم تصادفی بودن ارتباط ایجاد شده بین توصیف کننده‌های مدل و فعالیت دارویی را اثبات می‌کند و ارتباط ایجاد شده توسط مدل Ridge-SCAD کاملاً منطقی و معنا دار است.



شکل ۲-۱۰ نمودار مقادیر R^2 به دست آمده در آزمون Y-تصادفی بر حسب تعداد اجرا، برای ۱۰۰ بار اجرای آزمون Y-تصادفی و پیش بینی فعالیت ترکیبات مجموعه آزمون به وسیله مدل Ridge-SCAD توسعه یافته با استفاده از پاسخ تصادفی شده

فصل سوم

نتیجه‌گیری و آینده‌نگری

۳- ۱ تجزیه و تحلیل توصیف کننده های مدل Ridge-SCAD و تأثیر آن‌ها بر فعالیت دارویی با استفاده از ضرایب رگرسیون استاندارد شده

قابلیت تفسیر پذیری مدل QSAR توسعه یافته یکی از مهم‌ترین معیارها در مطالعات QSAR است. در این بخش از مطالعه، اثر توصیف کننده‌های منتخب مدل توسعه یافته Ridge-SCAD بر فعالیت دارویی، با توجه به تأثیر ضرایب رگرسیونی استاندارد شده توصیف کننده‌های مدل Ridge-SCAD و نمودار سهم مشارکت توصیف کننده‌ها مورد بررسی قرار گرفت. بنابراین ضرایب استاندارد شده ۹ توصیف‌کننده مدل Ridge-SCAD با استفاده از ضرایب غیر استاندارد و رابطه ۱-۴۲ محاسبه شد و در جدول ۳-۱ آورده شده است. علامت مثبت ضرایب توصیف کننده، نشان‌دهنده اثر افزایشی توصیف کننده بر فعالیت دارویی می‌باشد، در حالی‌که ضرایب با علامت منفی نشان‌دهنده اثر کاهشی توصیف کننده بر فعالیت دارویی می‌باشد. در ادامه رابطه بین این توصیف کننده‌ها و فعالیت دارویی ترکیبات بررسی خواهد شد.

جدول ۳-۱ اثر متوسط توصیف‌کننده های انتخاب شده توسط روش Ridge-SCAD

توصیف کننده‌ها	DECC	MATS4e	C-025	Mor26p	H-046	N-075	SEigZ	Mor23m	GATS4e
ضریب تأثیر	+۰/۴۵۵۵	+۰/۰۲۴۰	-۰/۳۹۸۰	-۰/۲۰۲۴	-۰/۱۷۱۰	-۰/۰۱۱۸	-۰/۰۰۱۳	+۰/۱۷۵۶	-۰/۲۶۴۳

۳- ۱-۱ بررسی میزان و چگونگی تأثیر توصیف کننده ها بر فعالیت دارویی مجموعه بازدارنده های سرطان و کاربرد مدل ارائه شده در پیشنهاد ترکیبات جدید

۳- ۱-۱-۱ محاسبه سهم مشارکت هر توصیف کننده در مدل Ridge-SCAD

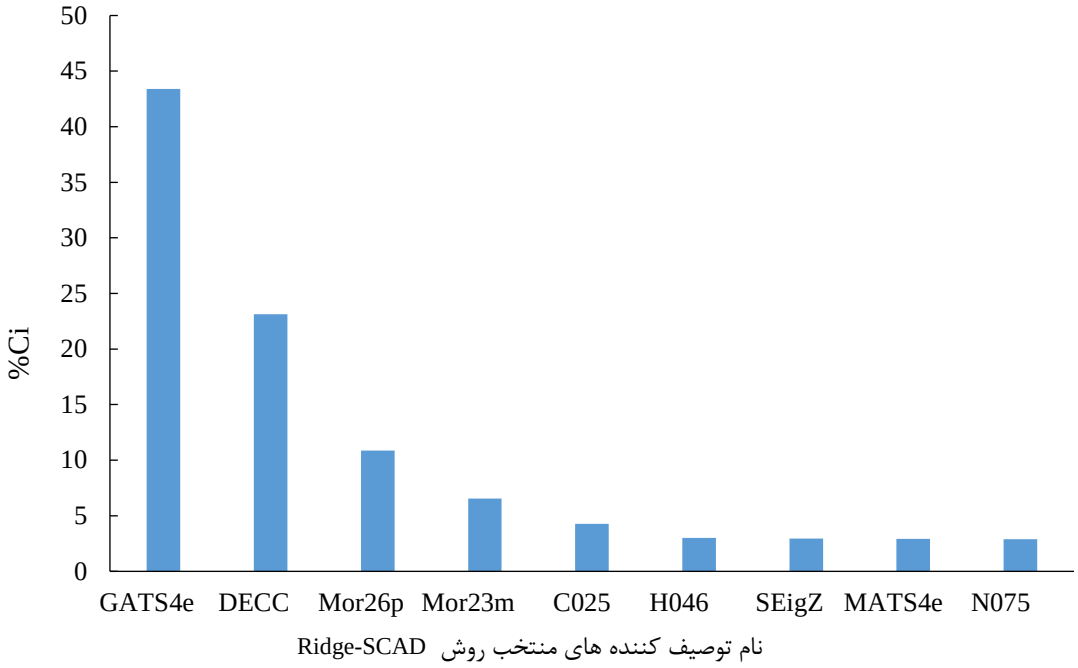
به‌منظور بررسی اهمیت و میزان مشارکت هر توصیف کننده در مدل Ridge-SCAD درصد سهم مشارکت هر توصیف کننده (C_i٪) برآورد شد. روش کار به این صورت است که برای محاسبه سهم مشارکت

توصیف کننده i، مقادیر آن توصیف کننده به صفر تغییر یافت. پس از پیش‌بینی فعالیت دارویی ترکیبات مجموعه آموزش، مقدار MSE مجموعه آموزش در حضور توصیف کننده i با مقادیر تصادفی شده، محاسبه شد (MSE_i). این فرآیند برای هر ۹ توصیف کننده تکرار شد و ۹ مقدار (MSE_i) برای مجموعه آموزش به‌دست آمد. سپس، مقادیر C_i٪ (۹،...،۲،۱)=i برای همه توصیف کننده‌ها با استفاده از رابطه ۳-۱ به‌دست آمد.

رابطه ۳-۱

$$C_i=(\frac{MSE_i}{\sum MSE_i})\times 100$$

نتایج به‌دست‌آمده برای سهم مشارکت هر توصیف کننده در مدل Ridge-SCAD در شکل ۳-۱ آورده شده است. نتایج نشان می‌دهد که ۳ توصیف کننده GATS4e، DECC و Mor26p حدوداً ۸۰ ٪ در توسعه مدل Ridge-SCAD مشارکت دارند. بنابراین با توجه به نمودار سهم مشارکت (شکل ۳-۱) و تأثیر ضرایب این توصیف کننده‌ها در مدل Ridge-SCAD (جدول ۳-۱) چگونگی اثر توصیف کننده‌هایی با بیش‌ترین سهم مشارکت در مدل Ridge-SCAD بر فعالیت دارویی ترکیبات مورد بررسی بیش‌تر قرار گرفت، تا از نتایج آن برای پیشنهاد ترکیبات جدید با فعالیت دارویی مناسب استفاده گردد.



شکل ۳-۱ نمودار سهم مشارکت توصیف کننده ها در مدل Ridge-SCAD

۳- ۱- ۱- ۲ توصیف کننده دو بعدی همبسته[۱]

• توصیف کننده های گروه Geary Autocorrelation:

توصیف کننده‌های دو بعدی خود همبسته از گراف مولکول و از طریق محاسبه مجموع وزن‌های اتم‌های انتهای کل مسیرها با طول مسیر[۲] مورد نظر، به‌دست می‌آیند. نوعی از این توصیف کننده‌ها گروه Geary Autocorrelation است که ضریب گری[۳] نام دارد و طبق (رابطه ۳-۲) محاسبه می‌شوند:

رابطه ۳-۲

$$c(d) = \frac{\frac{1}{2\triangle}\sum_{i=1}^{A}\sum_{j=1}^{A}\delta_{ij}(W_i - W_j)^2}{\frac{1}{A-1}\sum_{i=1}^{A}(W_i - \bar{W})^2}$$

[۱۲] D autocorrelation
[2] Lag
[3] Geary Autocorrelation

$c(d)$ ضریب گری می‌باشد، که W یک ویژگی‌اتم، $\overline{W}$ میانگین مقدار آن ویژگی روی مولکول، A تعداد اتم‌ها و d فاصله مکانی است، δ_{ij} نیز که به تابع کرونکر[1] معروف است، در حالتی که $d = d_{ij}$ باشد، δ_{ij} برابر یک است، و در غیر این صورت صفر است و $\triangle$ هم مجموعه δ هاست. مقدار این توصیف کننده که از جنس فاصله است، از صفر تا بی‌نهایت متغیر است [۶۹]. از میان این توصیف کننده‌ها GATS4e در مرحله انتخاب توصیف کننده‌های مؤثر در مدل Ridge-SCAD پیشنهادی برگزیده شده است. این توصیف کننده براساس الکترونگاتیویته ساندرسون، وزن‌دار شده است. با توجه به منفی بودن ضریب رگرسیون برای این توصیف کننده (جدول ۳-۱) کاهش مقدار آن سبب افزایش مقدار فعالیت دارویی ($pIC_{۵۰}$) می‌شود.

۳- ۱- ۱- ۳ توصیف کننده مکانی[2]

توصیف کننده‌های مکانی که از گراف مولکولی به‌دست می‌آیند، به‌عنوان توصیف کننده‌های عمومی مطرح هستند، چون به‌طور مستقیم از ساختار مولکول محاسبه می‌شوند و مستقل از صورت‌بندی مولکول هستند. DECC جزء این دسته از توصیف کننده‌ها می‌باشند که در مدل Ridge-SCAD ظاهر شده‌اند. DECC به شاخه دار بودن ساختار ترکیب ارتباط دارد و با توجه به علامت ضریب رگرسیونی (جدول ۳-۱)، هرچه انشعابات و استخلاف‌های ترکیب شیمیایی بیش‌تر باشد، فعالیت دارویی نیز افزایش می‌باشد [۷۰].

۳- ۱- ۱- ۴ توصیف کننده نمایش سه بعدی ساختار مولکول براساس تفرق الکترون[3]

توصیف کننده‌ها ۳D-MORSE (نمایش سه بعدی ساختار مولکول براساس تفرق الکترون) از طریق معادله تبدیلی که در پراش الکترون استفاده می‌شود، محاسبه می‌گردند:

رابطه ۳-۳

$$I(s) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} A_i A_j \frac{\sin(sr_{ij})}{sr_{ij}}$$

[1] Kroncker
[2] Topological
[3] ۳ D- Molecule Representation of Structure based on Electron diffraction

I شدت الکترون پراکنده شده، A_t و A_j خاصیت اتمی‌اتم‌های i و j، s زاویه پراکندگی، r_{ij} فاصله بین اتم‌های i و j و N تعداد کل اتم‌ها را نشان می‌دهد. این روش باعث می‌شود که ساختار سه بعدی مولکول به یک کد ثابت تبدیل شود [۶۹]. توصیف کننده‌های سه بعدی که در مدل Ridge-SCAD دیده می‌شوند، عبارتند از Mor26p که با استفاده از قطبش پذیری وزن دار شده است، و دارای ضریب تأثیر منفی می‌باشد (جدول ۳-۱). با توجه به این‌که مقادیر توصیف کننده Mor26p منفی است، بنابراین با منفی‌تر شدن مقدار این توصیف کننده مقدار فعالیت دارویی ($pIC_{۵۰}$) ترکیب مورد نظر افزایش می‌یابد.

۳- ۲ پیشنهاد ترکیبات جدید دارای فعالیت مناسب ضد سرطان با استفاده از مدل Ridge-SCAD و مطالعه داکینگ مولکولی ترکیبات پیشنهادی

یکی از اهداف مطالعات QSAR که اخیراً مورد توجه زیادی قرار گرفته است، پیشنهاد ترکیبات جدید و در نتیجه تسهیل فرآیند طراحی دارو است. مدل پیشنهادی نشان می‌دهد که چگونه تغییرات در توصیف کننده‌های ظاهر شده در مدل QSAR می‌تواند فعالیت دارویی را افزایش دهد. از این‌رو با توجه به اثر هر توصیف کننده در مدل Ridge-SCAD بر فعالیت دارویی مربوطه، می‌توان به چگونگی ایجاد یک اصلاح ساختاری مؤثر که منجر به افزایش فعالیت دارویی شود، پی برد. قابل توجه است که برای مثال، توصیف کننده‌های C-025 و H-046 با ضریب اثر منفی در مدل نشان‌دهنده این است که هرچه تعداد کربن‌های R-CR-R و تعداد هیدروژن‌های متصل به کربن (Sp^3)، که کربن بعدی به هتروواتمی متصل نباشد، کم‌تر باشد، فعالیت دارویی ترکیبات بیش‌تر می‌شود. بنابراین تلاش شد با در نظر گرفتن اثر توصیف کننده‌های مدل Ridge-SCAD بر فعالیت دارویی ترکیبات مورد مطالعه، ترکیبات جدید سنتز نشده با فعالیت دارویی مناسب پیشنهاد شود. به این منظور، مطابق با اثر توصیف کننده‌ها بر فعالیت دارویی، اصلاحات ساختاری روی ترکیب ضعیف (ترکیب شماره ۵) ایجاد شد. علاوه براین تلاش شد تا با ایجاد تغییرات بر روی ترکیب فعال (ترکیب شماره ۶۳) نیز ترکیباتی با پتانسیل دارویی بیش‌تر پیشنهاد شود.

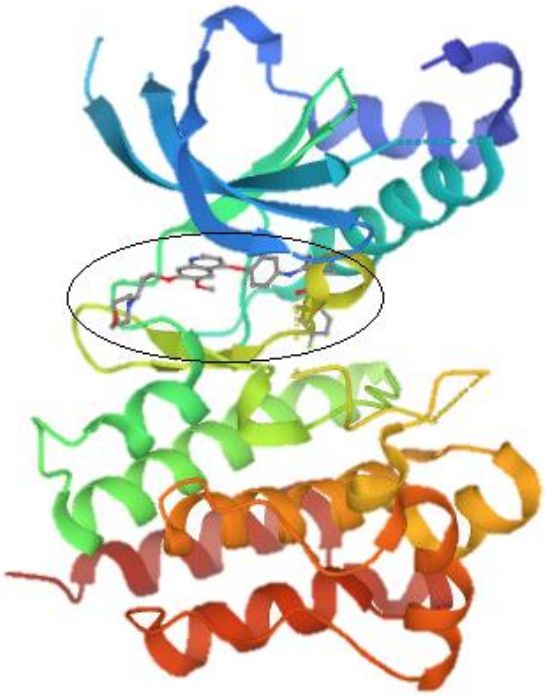
ازاین‌رو برای طراحی ترکیبات جدید با فعالیت دارویی مناسب، از معنای شیمیایی توصیف‌کننده‌های منتخب برای اصلاح ساختاری ترکیبات در راستای ایجاد ترکیبات با فعالیت دارویی مناسب، استفاده شد. ساختار ترکیبات پیشنهادی و مقادیر pIC_{50} پیش‌بینی‌شده مرتبط به آن‌ها در جدول ۳-۲ آورده شده است.

ساختار این ترکیبات در هایپرکم رسم و بهینه‌سازی شدند. توصیف کننده‌های مدل برتر با استفاده از نرم افزار دراگون محاسبه شد و همچنین فعالیت این ترکیبات با استفاده از مدل برتر Ridge-SCAD پیش‌بینی شد، که فعالیت برخی از آن‌ها بیش‌تر از فعال‌ترین ترکیب مورد مطالعه (ترکیب ۶۳) بوده و فعالیت برخی ترکیبات در محدوده ترکیب فعال به‌دست آمد و نتایج در جدول ۳-۲ آورده شده است.

در ادامه برای اثبات درستی فعالیت پیش‌بینی شده برای ترکیبات فعال پیشنهادی، برهم‌کنش ترکیبات در جایگاه فعال گیرنده آنزیمی با استفاده از نرم‌افزار داکینگ مولکولی مورد بررسی قرار گرفت که در ادامه چگونگی انجام کار آورده شده است.

۳- ۲- ۱ مطالعه داکینگ مولکولی برای بازدارنده های جدید پیشنهادی

به‌منظور انجام عملیات داکینگ مولکولی باید مراحل متفاوتی را انجام داد که در فصل دوم (بخش‌های ۱- ۵- ۲ تا ۱- ۵- ۶) به آن‌ها اشاره شده است. مطابق با مراحل گفته شده در قبل، ابتدا ساختار کریستالوگرافی گیرنده مورد استفاده برای ترکیبات مورد مطالعه به پیشنهاد مقالات شناسایی شد [۷۱, ۷۲]. با مراجعه به مقالات، گیرنده با کد ۳LQ8 انتخاب و ساختار کریستالوگرافی آن از سایت بانک اطلاعاتی پروتئین با پسوند pdb دانلود شد. ارزش تفکیک مربوط به ساختار کریستالوگرافی گیرنده ۳LQ8 برابر ۲/۰۲ آنگستروم می‌باشد که این مقدار ارزش تفکیک برای انجام مطالعات داکینگ مولکولی مناسب است [۷۳]. ساختار کریستالوگرافی گیرنده ۳LQ8 و لیگاند کریستالوگرافی موجود در زنجیره A در شکل ۳-۲ نشان داده شده است.



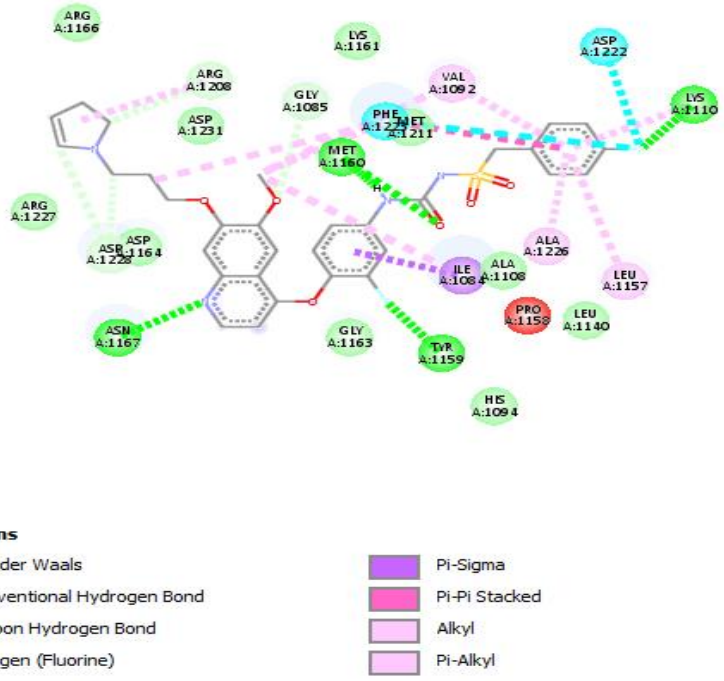
شکل ۳-۲ ساختار کریستالوگرافی 3LQ8 (منطقه انتخاب شده نشان دهنده لیگاند کریستالوگرافی و مابقی زنجیره های اسید آمینه ای است)

در مرحله بعد ساختار کریستالوگرافی گیرنده 3LQ8 در نرم افزار ویورلایت فراخوانی شد و عملیات آماده‌سازی متفاوتی از جمله حذف مولکول‌های آب، حذف کوفاکتورها و حذف زنجیره اسید آمینه‌ای فاقد لیگاند کریستالوگرافی انجام شد و باقی‌مانده ساختار که شامل زنجیره اسید آمینه‌ای A و لیگاند کریستالوگرافی است، با پسوند pdb ذخیره شد. سپس زنجیره اسید آمینه‌ای به‌عنوان فایل گیرنده و ترکیب موجود در ساختار کریستالوگرافی به‌عنوان لیگاند به‌طور مجزا با پسوند pdb ذخیره شدند و به‌عنوان ورودی‌های نرم‌افزار داکینگ برای انجام مرحله اعتبار سنجی داکینگ مورد استفاده قرار گرفتند. در مرحله بعد، مختصات جایگاه فعال گیرنده با توجه به مختصات مرکز ثقل لیگاند کریستالوگرافی و با استفاده از نرم‌افزار ویورلایت با مختصات X، Y و Z برابر با ۰/۸۲-، ۴/۸۶ و ۳۰/۶۱۲ به‌دست آمد.

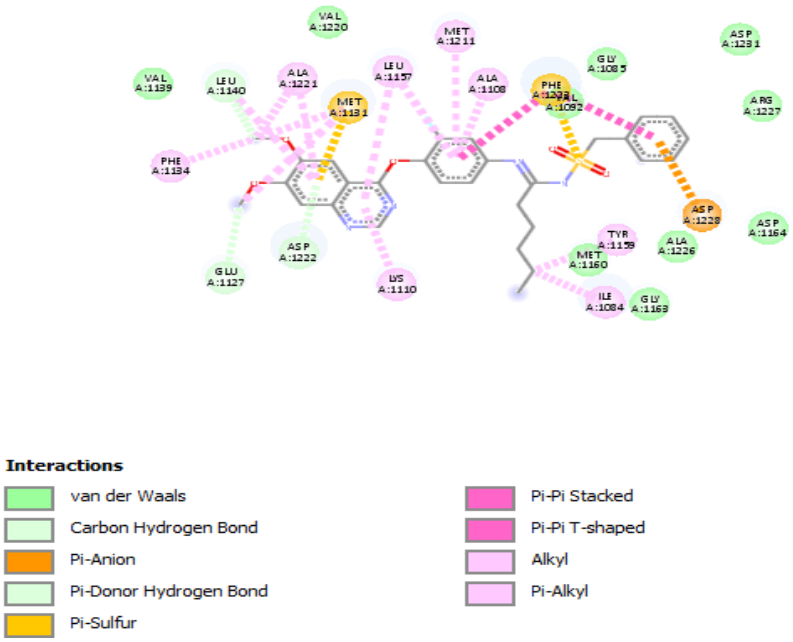
پس از آماده‌سازی فایل‌های ورودی (زنجیره اسید آمینه حاوی ساختار لیگاند کریستالوگرافی به‌عنوان گیرنده و لیگاند کریستالوگرافی به‌عنوان لیگاند)، فرآیند اعتبار سنجی داکینگ مولکولی با استفاده از نرم‌افزار اتوداک ۴/۲ انجام شد. به این منظور ابتدا فایل گیرنده در محیط نرم‌افزار وارد شد. آماده‌سازی‌های بیش‌تر روی ساختار گیرنده از جمله افزودن هیدروژن برای جبران کمبود هیدروژن ساختار کریستالوگرافی، ادغام هیدروژن‌های غیر قطبی متصل به هر اتم کربن و افزایش بار الکتریکی کلمن[1] انجام شد [۳۹]. در مرحله بعد به‌منظور انجام فرآیند اعتبار سنجی، ساختار لیگاند کریستالوگرافی به‌عنوان لیگاند در نرم‌افزار اتوداک ۴/۲ فراخوانی شد و با پسوند pdbqt ذخیره شد. در مرحله بعد خروجی pdbqt ساختار گیرنده نیز ذخیره شد و مختصات یک شبکه مشتمل بر جایگاه فعال گیرنده، با پیشنهاد مقالات منتشر شده در حوزه مطالعات داکینگ مولکولی، با ابعاد ۶۰ ×۶۰×۶۰ آنگستروم و فاصله بین نقاط ۰/۳۷۵ آنگستروم (در حدود یک چهارم طول پیوند یگانه کربن-کربن) ایجاد شد و بر هم‌کنش‌های گیرنده-لیگاند در این شبکه مورد بررسی قرار گرفت [۴۰, ۷۴, ۷۵]. داکینگ مولکولی با استفاده از اجرای الگوریتم ژنتیک لامارکین (LGA)، در تعداد اجراهای متفاوت ۱۰۰، ۱۵۰ و ۲۰۰ انجام شد. لازم به ذکر است که سایر پارامترهای الگوریتم ژنتیک در حالات پیش‌فرض پیشنهادی نرم افزار در نظر گرفته شدند. عملیات اعتبارسنجی در شرایط ذکر شده، انجام شد و فایل‌های خروجی با پسوند dlg استخراج و ذخیره شدند. با توجه به نتایج داکینگ مولکولی در مرحله اعتبار سنجی در ۳ حالت با تعداد اجراهای متفاوت الگوریتم LG، تعداد اجرای الگوریتم LG برابر با ۱۰۰، کم‌ترین مقدار ریشه میانگین مربعات انحراف (RMSD) و بیش‌ترین تعداد کنفورماسیون در خوشه اول را دارا بوده و در نتیجه، برای انجام فرآیند داکینگ مولکولی ترکیبات مورد نظر از تعداد اجرای الگوریتم LG برابر با ۱۰۰ استفاده شد. در فرآیند انجام داکینگ مولکولی ترکیبات مورد مطالعه، همانند فرآیند اعتبار سنجی، تمامی شرایط (مختصات جایگاه فعال، ابعاد شبکه و پارامترهای پیش‌فرض الگوریتم ژنتیک) در

[1] Kollman

شرایط داکینگ مولکولی ترکیبات جدید نیز تعریف شد. بنابراین داکینگ مولکولی در شرایط بهینه برای ترکیب ۶۳ با بیش‌ترین فعالیت ($pIC_{۵۰}$= ۶/۸۰)، ترکیب ۵ با کم‌ترین فعالیت ($pIC_{۵۰}$= ۴/۵۲) و تمامی ترکیبات پیشنهادی در جایگاه فعال گیرنده انجام شدند. بهترین پیکربندی‌ها از ترکیبات مورد مطالعه و پیشنهادی با توجه به حداقل انرژی آزاد اتصال و بیش‌ترین تعداد پیکربندی در خوشه اول، شناسایی شدند و برهم‌کنش آن‌ها با اسید آمینه‌های کلیدی با استفاده از نرم‌افزار Discovery Studio مورد بررسی قرار گرفت. شکل‌های دوبعدی برهمکنش‌های ترکیب فعال ۶۳ (شکل ۳-۳) و ترکیب ضعیف ۵ (شکل ۳-۴) آورده شد.

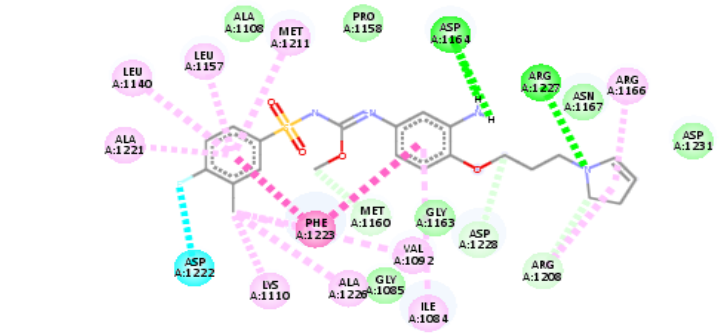


شکل ۳-۳ بر همکنش ترکیب ۶۳ با اسید آمینه های کلیدی

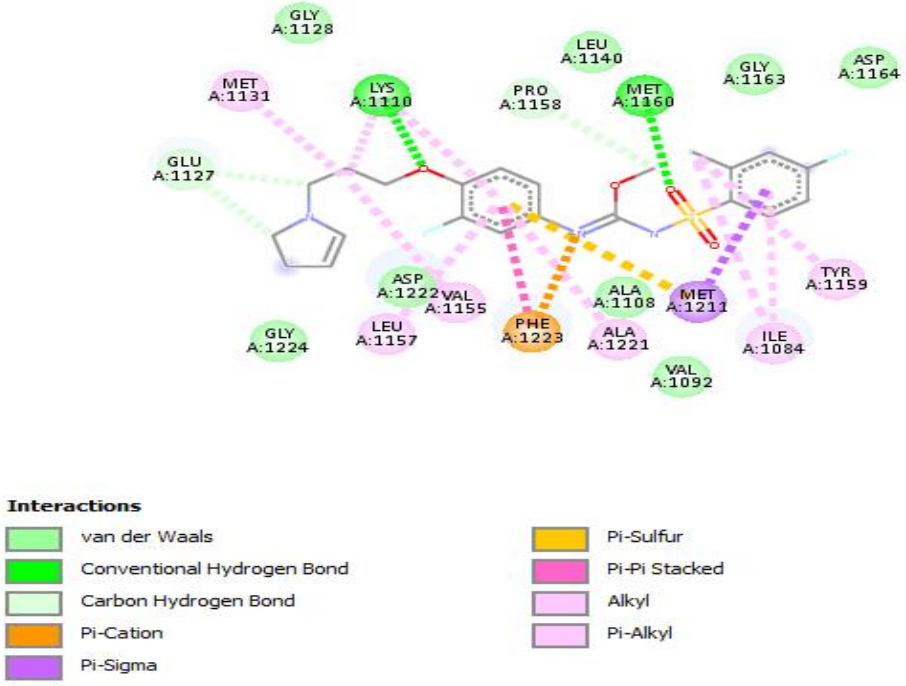


شکل ۳-۴ بر همکنش ترکیب ۵ با اسید آمینه های کلیدی

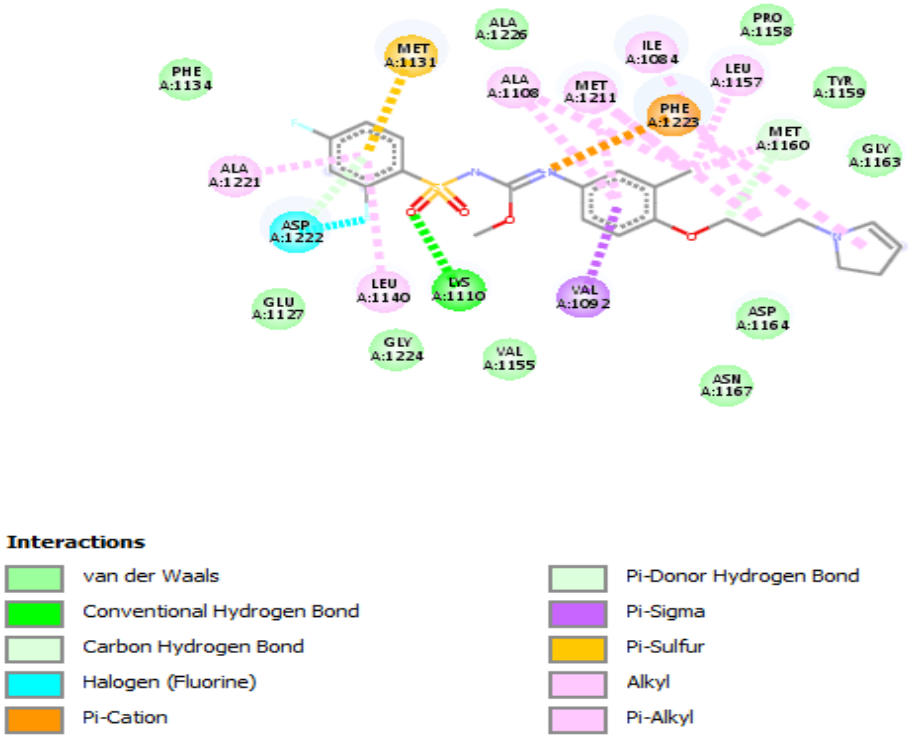
همان‌طور که شکل‌ها و نتایج جدول ۳-۲ نشان می‌دهد، ترکیب فعال ۶۳ برهم‌کنش‌های هیدروژنی H-N Asn1167 (2.29Å)، H-F Tyr1159 (2.58Å)، O-H Met1160 (2.49Å) H-0 Met1160 (2.05Å) و H-F Lys1110 (1.96Å) (موارد داخل پرانتز طول پیوند اسیدآمینه کلیدی با یک اتم در ساختار مورد مطالعه را نشان می‌دهد) را با اسید آمینه داشته و ترکیب ضعیف هیچ پیوند هیدروژنی با اسید آمینه برقرار نکرده و همه پیوندها از نوع هیدروفوبی بودند. بنابراین ترکیب فعال (ترکیب شماره ۶۳) پیوندهای هیدروژنی مناسبی را با اسید آمینه‌ها برقرار کرده است، که به‌طور معنا داری از نظر بر هم‌کنش متفاوت از ترکیب ضعیف‌تر (ترکیبات شماره ۵) است. این به آن معنا است که بر هم‌کنش‌های حاصل از داکینگ مولکولی معیار مناسبی برای ارزیابی میزان فعالیت دارویی ترکیبات مورد مطالعه هستند. بنابراین برهم‌کنش‌های لیگاند-گیرنده برای تمام ترکیبات پیشنهادی استخراج و مورد تجزیه و تحلیل قرار گرفت و نتایج برخی از ترکیبات پیشنهادی (۱، ۲ و ۳) در شکل ۳-۵ تا شکل ۳-۷ آورده شده است.



شکل ۳-۵ بر هم‌کنش ترکیب جدید ۱ با اسید آمینه های کلیدی

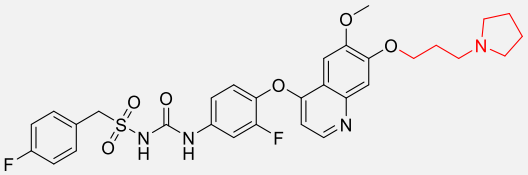
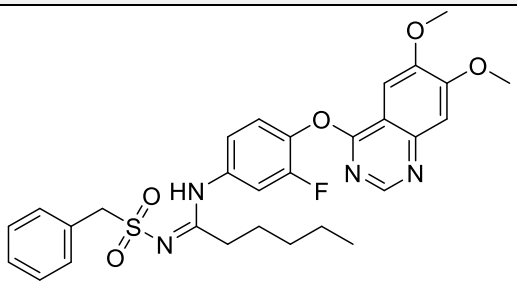
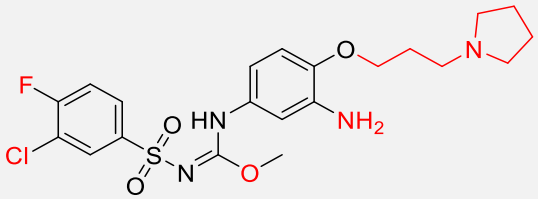
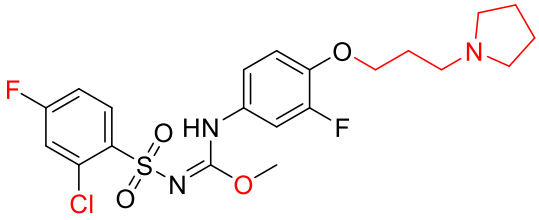
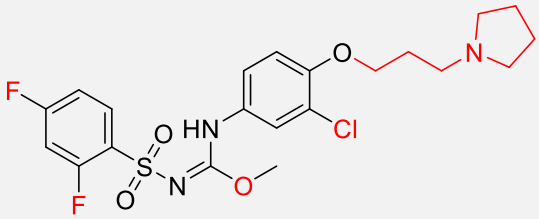
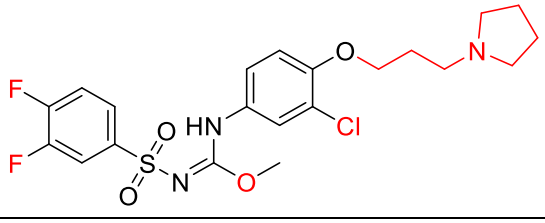


شکل ۳-۶ بر همکنش ترکیب جدید ۲ با اسید آمینه های کلیدی

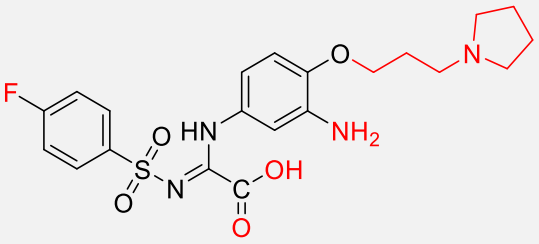
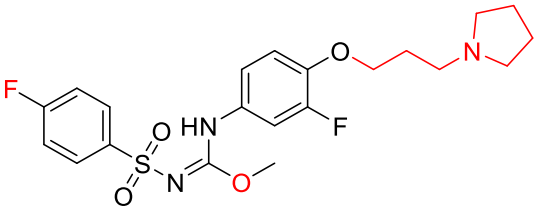
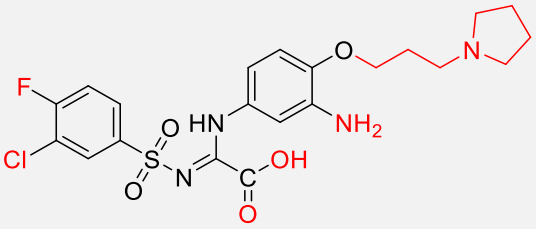
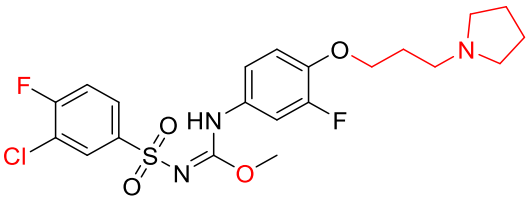
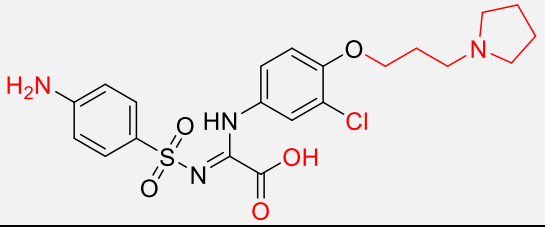


شکل ۳-۷ بر همکنش ترکیب جدید ۳ با اسید آمینه های کلیدی

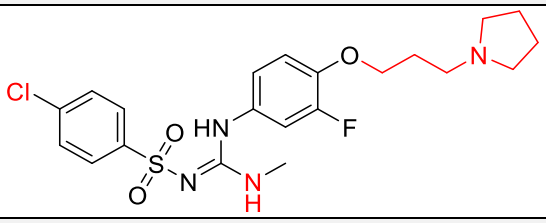
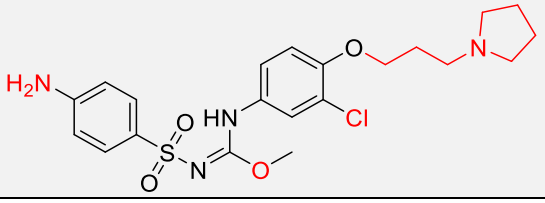
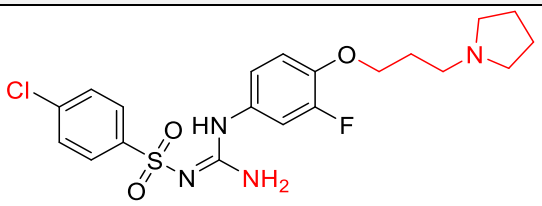
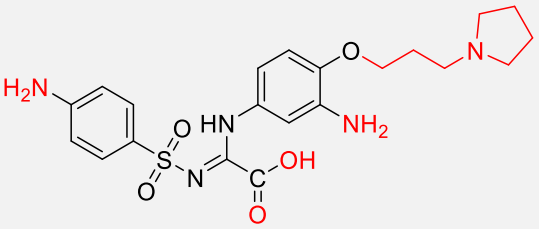
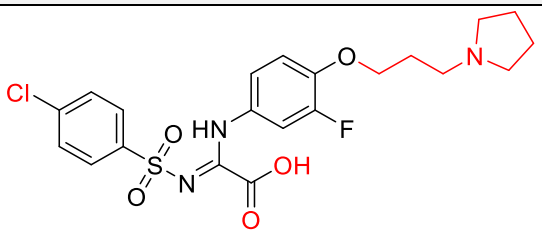
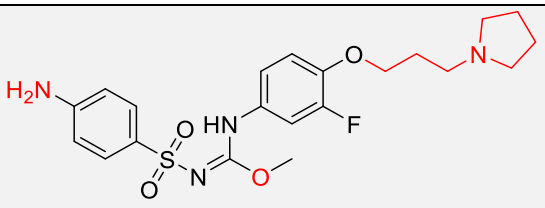
جدول ۳-۲ پارامترهای محاسبه شده برای ترکیبات مورد مطالعه و ترکیبات پیشنهادی

شماره ترکیب	ساختار شیمیایی	پیوند هیدروژنی	طول پیوند هیدروژنی (Å)	پیوند هیدروفوبی	pIC_{50}
۶۳		H Asn1167- N H Tyr1159- F O Met1160- H H Met1160- O H Lys1110- F	۲/۲۹ ۲/۵۸ ۲/۴۹ ۲/۰۵ ۱/۹۶	Ala1108 Asp1164 Asp1231 Arg1166 Arg1227 Gly1163 Lys1161 Leu1140 Met1211 Phe1223	۶/۷۹۶
۵		-	-	Ala1226 Asp1231 Asp1164 Arg1227 Gly1163 Gly1085 Met1160 Val1092 Val1139 Val1220 Phe1223	۴/۵۱۵
NC۱		O Asp1164- H O Asp1164- H H Arg1227- N	۲/۱۴ ۲/۵۱ ۳/۰۳	Ala1108 Asn1167 Asp1231 Gly1163 Gly1085 Pro1158 Phe1223	۶/۹۴۹
NC۲		H Lys1110- O H Met1160- O	۲/۳۵ ۲/۰۴	Ala1108 Asp1164 Asp1222 Gly1163 Gly1224 Gly1128 Leu1140 Val1092 Phe1223	۶/۹۳۶
NC۳		H Lys1110- O	۱/۶۳	Ala1226 Asn1167 Asp1164 Gly1163 Gly1224 Glu1127 Pro1158 Phe1134 Tyr1159 Val1155	۶/۹۱۶
NC۴		H Lys1110- O	۱/۷۸	Ala1226 Gly1163 His1094 Pro1158 Phe1134 Val1155 Val1220	۶/۷۵۴

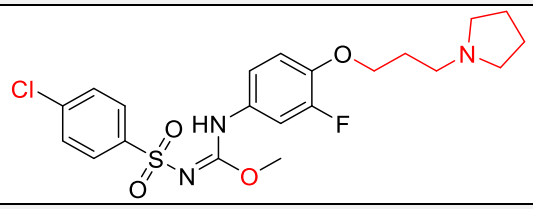
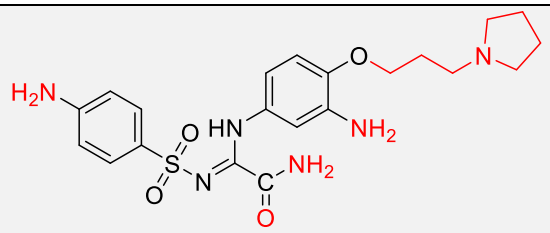
ادامه جدول ۳-۲

شماره ترکیب	ساختار شیمیایی	پیوند هیدروژنی	طول پیوند هیدروژنی (Å)	پیوند هیدروفوبی	pIC_{50}
NC۵		O Asp1222- H H Lys1110- O H Lys1110- O O Phe1223- H	۲/۱۰ ۱/۹۶ ۲/۸۵ ۲/۲۳	Ala1226 Gly1163 Gly1224 His1202 Ile130 Leu1195 Leu1140 Met1211 Met1160 Phe1200 Tyr1159 Val1139 Val1155 Phe1134	۶/۷۱۵
NC۶		H Lys1110- O	۱/۸۹	Ala1226 Gly1163 Gly1224 Glu1127 Lys1161 Leu1140 Met1160 Pro1158 Phe1134 Val1155 Phe1223	۶/۷۰۴
NC۷		O Asp1224- H H Lys1110- O S Met1131- H	۲/۳۵ ۲/۹۹ ۲/۸۳	Ala1226 Asn1209 Gly1163 Gly1224 Gly1128 Pro1158 Tyr1159 Val1139 Val1092 Val1155 Phe1223	۶/۵۹۸
NC۸		H Asp1164- O H Met1160- O	۲/۳۳ ۲/۳۵	Ala1226 Asn1167 Asp1222 Asp1228 Arg1227 Gly1163 Gly1085 Lys1110 Leu1140 Pro1158 Phe1223	۶/۵۹۵
NC۹		O Lys1161- H O Met1160- H O Pro1158- H	۲/۳۴ ۲/۱۵ ۲/۱۴	Ala1108 Asp1222 Gly1163 His1162 Lys1110 Phe1200 Val1092 Val1220 Tyr1159	۶/۵۸۹

ادامه جدول ۳-۲

شماره ترکیب	ساختار شیمیایی	پیوند هیدروژنی	طول پیوند هیدروژنی (Å)	پیوند هیدروفوبی	pIC_{50}
NC۱۰		H Lys1110- O	۱/۶۵	Ala1226 Gly1163 Gly1224 Glu1127 Leu1140 Met1160 Val1155	۶/۵۵۱
NC۱۱		H Asp1222- O	۱/۹۹	Asp1164 Gly1163 Met1160 Pro1158 Tyr1159 Val1092 Val1109	۶/۵۴۳
NC۱۲		H Lys1110- O H Met1160- O	۱/۷۹ ۲/۲۶	Gly1163 Lys1161 Pro1158 Val1092 Val1155	۶/۴۷۹
NC۱۳		O Asp1228- H H Asp1164- O H Tyr1159- O O Met1160- H O Pro1158- H O Pro1158- H	۲/۰۸ ۱/۸۱ ۲/۴۲ ۲/۲۵ ۲/۲۲ ۲/۲۲	Ala1221 Asp1222 Asn1167 Arg1086 Gly1163 Gly1085 His1094 Val1092 Tyr1159	۶/۴۷۷
NC۱۴		O Phe1223- H H Lys1110- O S Met1131- H	۲/۰۲ ۲/۶۲ ۲/۷۲	Ala1221 Ala1226 Asp1222 Gly1163 Gly1224 Val1155 Val1092 Phe1223	۶/۴۴۶
NC۱۵		O Asp1222- N H Asp1222- O O Glu1127- H	۲/۷۱ ۲/۰۰ ۲/۰۸	ALA1221 ASN1209 GLY1163 GLY1228 LEU1140 LYS1161 PHE1124	۶/۳۸۳

ادامه جدول ۳-۲

شماره ترکیب	ساختار شیمیایی	پیوند هیدروژنی	طول پیوند هیدروژنی (Å)	پیوند هیدروفوبی	pIC_{50}
NC۱۶		H Lys1110- O H Met1160- O	۱/۹۷ ۲/۴۰	Gly1163 Glu1127 His1094 Leu1140 Pro1158 Val1155	۶/۳۶۲
NC۱۷		H Asp1222-N O Pro1158- H H Met1160- O	۲/۲۲ ۲/۱۴ ۱/۸۴	Ala1108 Asn1167 Asn1209 Asp1164 Gly1163 Lys1110 Tyr1159 Val1220 Val1139 Val1092 Phe1223	۶/۳۳۵

NC نماد ترکیبات جدید پیشنهادی است.

نتایج به وضوح نشان می‌دهد که ترکیبات پیشنهادی پیوندهای هیدروژنی مناسبی را با اسید آمینه‌های کلیدی برقرار نموده‌اند و درستی فعالیت دارویی پیش‌بینی شده توسط مدل Ridge-SCAD با مطالعات داکینگ و نتایج بر هم‌کنش لیگاند- گیرنده همخوانی دارد. علاوه بر مطالعه داکینگ مولکولی، پارامترهای قاعده لیپینسکی نیز برای ترکیبات پیشنهادی محاسبه شدند و نتایج در جدول ۳-۳ خلاصه شده‌اند. نتایج نشان می‌دهد که تمامی پارامترهای مورد نظر از جمله وزن مولکولی (MW < ۵۰۰)، چربی‌دوستی (MLOGP < ۴/۱۵)، تعداد گیرنده‌های پیوند هیدروژنی (کمتر از ۱۰)، تعداد دهنده‌های پیوند هیدروژنی (کمتر از ۵) و تعداد پیوندهای قابل چرخش (کمتر از ۱۰) برای ترکیبات پیشنهادی دارای مقادیر قابل قبولی هستند [۷۶]. بنابراین ترکیبات پیشنهادی معیارهای شباهت به ترکیبات دارویی را داشته‌اند و

به‌عنوان ترکیبات بالقوه فعالی که شباهت به ترکیب رهبر[1] دارند، می‌توانند به‌عنوان کاندیدهای فعال مناسبی برای سنتز و آزمایش‌های دارویی مورد توجه قرار گیرند.

علاوه بر این فاکتورها، پارامتر سهولت سنتز[2] نیز با استفاده از سایت Swiss-ADME [۷۶] استخراج شد و در جدول ۳-۳ آورده شد. فاکتور سهولت سنتز می‌تواند دارای مقداری در محدوده ۱ (سنتز بسیار آسان) تا ۱۰ (ترکیب پیچیده و چالش‌برانگیز) باشد [۷۷]. فاکتور سهولت سنتز برای ترکیبات پیشنهادی دارای مقدار قابل قبولی می‌باشد و در نتیجه سنتز آن‌ها در محیط آزمایشگاهی امکان‌پذیر خواهد بود.

[1] Lead compound
[2] Synthetic Accessibility

جدول ۳-۳ پارامترهای محاسبه شده قاعده ۵ لیپینسکی

شماره ترکیبات	پیوندهای قابل چرخش	وزن مولکولی g/mol	MlogP	تعداد گیرنده‌های پیوند هیدروژنی	تعداد دهنده‌های پیوند هیدروژنی	سهولت سنتز
۶۳	۱۴	۶۲۶/۷	۲/۵۵	۱۰	۲	۴/۲۸
۵	۱۳	۵۸۵/۷	۴/۰۶	۹	۱	۴/۲۹
NC۱	۱۰	۴۸۴/۱	۲/۸۷	۷	۲	۳/۹۵
NC۲	۱۰	۴۸۷/۱	۳/۷۶	۸	۱	۳/۷۲
NC۳	۱۰	۴۸۷/۱	۳/۷۶	۸	۱	۳/۷۹
NC۴	۱۰	۴۸۷/۱	۳/۷۶	۸	۱	۳/۷۶
NC۵	۱۰	۴۶۴/۵	-۰/۶۹	۸	۳	۳/۸۰
NC۶	۱۰	۴۵۳/۵	۳/۲۸	۸	۱	۳/۷۰
NC۷	۱۰	۴۹۸/۱	-۰/۲	۸	۳	۳/۸۳
NC۸	۱۰	۴۸۷/۱	۳/۷۶	۸	۱	۳/۷۵
NC۹	۱۰	۴۸۰/۱	-۰/۱۷	۷	۳	۳/۷۱
NC۱۰	۱۰	۴۶۸/۱	۲/۹۸	۶	۲	۳/۵۲
NC۱۱	۱۰	۴۶۶/۱	۲/۹	۶	۲	۳/۸۰
NC۱۲	۹	۴۵۴/۱	۲/۷۶	۶	۲	۳/۳۹
NC۱۳	۱۰	۴۶۱/۵	-۱/۱۶	۷	۴	۳/۸۷
NC۱۴	۱۰	۴۸۳/۱	۰/۷۲	۸	۲	۳/۵۸
NC۱۵	۱۰	۴۵۰/۵	۲/۷۹	۷	۲	۳/۷۶
NC۱۶	۱۰	۴۶۹/۱	۳/۳۹	۷	۱	۳/۶۸
NC۱۷	۱۰	۴۶۰/۶	۰/۶۵	۶	۴	۳/۸۸

NC نماد ترکیبات جدید پیشنهادی است.

۳- ۳ نتیجه گیری نهایی

در این تحقیق یک مدل QSAR معتبر جهت پیش‌بینی فعالیت دارویی ترکیبات شیمیایی ضد سرطان ارائه شد. ارزیابی‌های متفاوت مدل Ridge-SCAD توسعه یافته نشان داد که مدل از اعتبار و تعمیم پذیری مناسبی برخوردار است. در این تحقیق از روش Ridge به‌عنوان یک روش غربالگری نوین برای کاهش ابعاد داده‌ها استفاده شد. سپس با اعمال روش انقباضی SCAD روی مجموعه آموزش مؤثرترین توصیف کننده‌ها انتخاب شدند و مدل SCAD ساخته شد. مدل خطی توسعه یافته Ridge-SCAD برای پیش‌بینی فعالیت دارویی ترکیبات مجموعه آزمون مورد استفاده قرار گرفت.

علاوه براین تأثیر توصیف کننده‌های مدل Ridge-SCAD بر فعالیت دارویی با تجزیه‌وتحلیل ضرایب رگرسیونی استاندارد شده و نمودار سهم مشارکت توصیف کننده‌ها استخراج شد. در نهایت با توجه به میزان تأثیر توصیف کننده‌ها بر فعالیت دارویی، ترکیبات جدیدی با فعالیت دارویی مناسب ارائه شد. به‌منظور بررسی بیش‌تر ترکیبات جدید ضد سرطان و اثبات صحت پیش‌بینی فعالیت دارویی آن‌ها توسط مدل Ridge-SCAD از مطالعه داکینگ مولکولی برای استخراج نوع بر هم‌کنش آن‌ها با گیرنده پروتئینی مربوطه استفاده شد. نتایج مربوط به برهم‌کنش ترکیبات پیشنهادی با گیرنده، صحت فعالیت پیش‌بینی شده با استفاده از مدل Ridge-SCAD را اثبات نمود. از طرفی ویژگی‌های فارماکوکینتیک این ترکیبات با قاعده ۵ لیپینسکی نیز بررسی شد. نتایج نشان‌دهنده پتانسیل بالای این ترکیبات به‌عنوان بازدارنده‌های سرطان است. بنابراین مدل توسعه یافته Ridge-SCAD به‌عنوان یک مدل معتبر می‌تواند برای پیش‌بینی فعالیت دارویی ترکیبات ضد سرطان c-Met کاربرد داشته باشد. همچنین، به‌منظور بررسی عملکرد روش Ridge برای غربالگری توصیف کننده‌ها، ارزیابی مدل SCAD بدون استفاده از روش غربالگری انجام شد. پارامترهای محاسباتی دلالت بر قدرت پیش‌بینی بیش‌تر مدل Ridge-SCAD دارد.

۳- ۴ آینده نگری

- ✓ به‌کارگیری روش‌های غربالگری دیگری برای کاهش ابعاد داده‌ها
- ✓ اجرای روش‌های مدل‌سازی انقباضی جدید برای پیش‌بینی فعالیت دارویی متفاوت
- ✓ استفاده از مجموعه داده‌های متنوع دیگر برای ساخت مدل QSAR توسعه یافته با روش‌های انقباضی

۳- ۵ فهرست منابع

[۱] Massart D.L., Vandeginste B.G., Buydens L., De Jong S., et al. (**1998**) "Handbook of chemometrics and qualimetrics: Part A", pp

[۲] Gemperline P. (**2006**), "Practical guide to chemometrics", CRC press ,

[۳] Kowalski B.R. (**1980**) "Chemometrics", Analytical Chemistry. **52**, 5, pp 112-122

[٤] Tropsha A. (**2010**) "Best practices for QSAR model development, validation, and exploitation", Molecular informatics. **29**, 6-7, pp 476-488

[٥] Yasri A., and Hartsough D. (**2001**) "Toward an optimal procedure for variable selection and QSAR model building", Journal of chemical information and computer sciences. **41**, 5, pp 1218-1227

[٦] Leonard J.T., and Roy K. (**2006**) "On selection of training and test sets for the development of predictive QSAR models", QSAR & Combinatorial Science. **25**, 3, pp 235-251

[۷] Puzyn T., Mostrag-Szlichtyng A., Gajewicz A., SkrzzDski M., et al. (**2011**) "Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models", Structural Chemistry. **22**, pp 795-804

[۸] Hdoufane I ,.Ounaissi D., Dermoune A., and Cherqaoui D. (**2021**) "Development of QSAR Models Using Singular Value Decomposition Method: A Case Study for Predicting Anti-HIV-1 and Anti-HCV Biological Activities", Biointerface Research in Applied Chemistry. pp

[۹] Golbraikh A., and Tropsha A. (**2000**) "Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection", Molecular diversity. **5**, 4, pp 231-243

[۱۰] Roy K., Kar S., and Das R.N. (**2015**), "A primer on QSAR/QSPR modeling: fundamental concepts", Springer ,

[۱۱] Algamal Z., Lee M., Al-Fakih A., and Aziz M. (**2016**) "High-dimensional QSAR modelling using penalized linear regression model with L 1/2-norm", SAR and QSAR in Environmental Research. **27**, 9, pp 703-719

[۱۲] Algamal Z., and Lee M. (**2017**) "A new adaptive L1-norm for optimal descriptor selection of high-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives", SAR and QSAR in Environmental Research. **28**, 1, pp 75-9۰

[۱۳] Saber S., Mohamad H., and Aziz M., Development a QSAR Model of 1, 3, 4-Triazole Derivatives for Antioxidant Activity Prediction, in: 2018 International Conference on Advanced Science and Engineering (ICOASE), IEEE, 2018, pp. 380-383.

[۱٤] Hoerl A.E., and Kennard R.W. (**1970**) "Ridge regression: Biased estimation for nonorthogonal problems", Technometrics. **12**, 1, pp 55-67

[۱٥] Xu L., and Zhang W.-J. (**2001**) "Comparison of different methods for variable selection", Analytica Chimica Acta. **446**, 1-2, pp 475-481

[۱٦] Tibshirani R. (**1996**) "Regression shrinkage and selection via the lasso", Journal of the Royal Statistical Society: Series B (Methodological). **58**, 1, pp 267-288

[۱۷] Fan J., and Li R. (**2001**) "Variable selection via nonconcave penalized likelihood and its oracle properties", Journal of the American statistical Association. **96**, 456, pp 1348-1360
[۱۸] Mercader A.G., and Pomilio A.B. (**2010**) "QSAR study of flavonoids and biflavonoids as influenza H1N1 virus neuraminidase inhibitors", European journal of medicinal chemistry. **45**, 5, pp 1724-1730
[۱۹] Todeschini R., and Consonni V. (**2008**), "Handbook of molecular descriptors", John Wiley & Sons ,
[۲۰] Asadi L., Gholivand K., and Zare K. (**2016**) "Phosphorhydrazides as urease and acetylcholinesterase inhibitors: biological evaluation and QSAR study", Journal of the Iranian Chemical Society. **13**, 7, pp 1213-1223
[۲۱] Sadeghi F., Afkhami A., Madrakian T., and Ghavami R. (**2021**) "A new approach for simultaneous calculation of pIC 50 and logP through QSAR/QSPR modeling on anthracycline derivatives: a comparable study", Journal of the Iranian Chemical Society. pp 1-16
[۲۲] Goodarzi M., Deshpande S., Murugesan V., Katti S.B., et al. (**2009**) "Is feature selection essential for ANN modeling?", QSAR & Combinatorial Science. **28**, 11-12, pp 1487-1499
[۲۳] Gholivand K., Valmoozi A.A.E., Salahi M., Taghipour F., et al. (**2017**) "Bisphosphoramidate derivatives: synthesis, crystal structure, anti-cholinesterase activity, insecticide potency and QSAR analysis", Journal of the Iranian Chemical Society. **14**, 2, pp 427-442
[۲٤] Saleh A.M.E., Arashi M., and Kibria B.G. (**2019**), "Theory of ridge regression estimation with applications", John Wiley & Sons ,
[۲٥] Montgomery D.C., Peck E.A., and Vining G.G. (**2021**), "Introduction to linear regression analysis", John Wiley & Sons ,
[۲٦] Golbraikh A., and Tropsha A. (**2002**) "Beware of q2!", Journal of molecular graphics and modelling. **20**, 4, pp 269-276
[۲۷] Veerasamy R., Rajak H., Jain A., Sivadasan S., et al. (**2011**) "Validation of QSAR models-strategies and importance", Int. J. Drug Des. Discov. **3**, pp 511-519
[۲۸] Roy P.P., and Roy K. (**2008**) "On some aspects of variable selection for partial least squares regression models", QSAR & Combinatorial Science. **27**, 3, pp 302-313
[۲۹] Gramatica P., and Sangion A. (**2016**) "A historical excursus on the statistical validation parameters for QSAR models: a clarification concerning metrics and terminology", Journal of chemical information and modeling. **56**, 6, pp 1127-1131
[۳۰] Fatemi M.H., and Izadiyan P. (**2011**) "Cytotoxicity estimation of ionic liquids based on their effective structural features", Chemosphere. **84**, 5, pp 553-563
[۳۱] Bring J. (**1994**) "How to Standardize Regression Coefficients", The American Statistician. **48**, pp 209-213
[۳۲] Morris G.M., and Lim-Wilby M., Molecular docking, in: Molecular modeling of proteins, Springer, **2008**
[۳۳] Huang S.-Y., and Zou X. (**2010**) "Advances and challenges in protein-ligand docking", International journal of molecular sciences. **11**, 8, pp 3034-3016

[٣٤] Halperin I., Ma B., Wolfson H., and Nussinov R. (**2002**) "Principles of docking: An overview of search algorithms and a guide to scoring functions", Proteins: Structure, Function, and Bioinformatics. **47**, 4, pp 409-443

[٣٥] Hernández-Santoyo A., Tenorio-Barajas A.Y., Altuzar V., Vivanco-Cid H., et al. (**2013**) "Protein-protein and protein-ligand docking", Protein engineering-technology and application. pp 63-81

[٣٦] Sousa S.F., Fernandes P.A., and Ramos M.J. (**2006**) "Protein–ligand docking: current status and future challenges", Proteins: Structure, Function, and Bioinformatics. **65**, 1, pp 15-26

[٣٧] Seeliger D., and De Groot B.L. (**2010**) "Conformational transitions upon ligand binding: holo-structure prediction from apo conformations", PLoS computational biology. **6**, 1, pp e1000634

[٣٨] McGovern S.L., and Shoichet B.K. (**2003**) "Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes", Journal of medicinal chemistry. **46**, 14, pp 2895-2907

[٣٩] Goodsell D.S., Morris G.M., and Olson A.J. (**1996**) "Automated docking of flexible ligands: applications of AutoDock", Journal of molecular recognition. **9**, 1, pp 1-5

[٤٠] Feinstein W.P., and Brylinski M. (**2015**) "Calculating an optimal box size for ligand docking and virtual screening against experimental and predicted binding pockets", Journal of cheminformatics. **7**, 1, pp 1-10

[٤١] https://www.who.int/news-room/fact-sheets/detail/cancer

[٤٢] Traxler P. (**2003**) "Tyrosine kinases as targets in cancer therapy–successes and failures", Expert opinion on therapeutic targets. **7**, 2, pp 215-234

[٤٣] Witsch E., Sela M., and Yarden Y. (**2010**) "Roles for growth factors in cancer progression", Physiology.

[٤٤] Xu W.-M., Han F.-F., He M., Hu D.-Y., et al. (**2012**) "Inhibition of tobacco bacterial wilt with sulfone derivatives containing an 1, 3, 4-oxadiazole moiety", Journal of agricultural and food chemistry. **60**, 4, pp 1036-1041

[٤٥] Liu N.-W., Liang S., and Manolikakes G. (**2016**) "Recent advances in the synthesis of sulfones", Synthesis. **48**, 13, pp 1939-1973

[٤٦] Caballero J., Quiliano M., Alzate-Morales J.H., Zimic M., et al. (**2011**) "Docking and quantitative structure–activity relationship studies for 3-fluoro-4-(pyrrolo [2, 1-f][1, 2, 4] triazin-4-yloxy) aniline-٣ ,fluoro-4-(1H-pyrrolo [2, 3-b] pyridin-4-yloxy) aniline, and 4-(4-amino-2-fluorophenoxy)-2-pyridinylamine derivatives as c-Met kinase inhibitors", Journal of Computer-Aided Molecular Design. **25**, 4, pp 349-369

[٤٧] Huang D., Zhu X., Tang C., Mei Y., et al. (**2012**) "3D QSAR pharmacophore modeling for c-Met kinase inhibitors", Medicinal chemistry. **8**, 6, pp 1117-1125

[٤٨] Lee J.Y., Lee K., Kim H.R., and Chae C.H. (**2013**) "3D-QSAR Studies on Chemical Features of 3-(benzo [d] oxazol-2-yl) pyridine-2-amines in the External Region of c-Met Active Site", Bulletin of the Korean Chemical Society. **34**, 12, pp 3553-3558

[٤٩] Balasubramanian P.K., Balupuri A., Bhujbal S.P., and Cho S.J. (**2019**) "3D-QSAR-aided design of potent c-Met inhibitors using molecular dynamics simulation and binding free energy calculation", Journal of Biomolecular Structure and Dynamics. **37**, 8, pp 2165-

2178
[٥٠] Algamal Z.Y., and Lee M.H. (**2017**) "A novel molecular descriptor selection method in QSAR classification model based on weighted penalized logistic regression", Journal of Chemometrics. **31**, 10, pp e2915
[٥١] Algamal Z., Qasim M., and Ali H. (**2017**) "A QSAR classification model for neuraminidase inhibitors of influenza A viruses (H1N1) based on weighted penalized support vector machine ,"SAR and QSAR in Environmental Research. **28**, 5, pp 415-426
[٥٢] Algamal Z.Y., Lee M.H., Al-Fakih A.M., and Aziz M. (**2017**) "High-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives based on the sparse logistic regression model with a bridge penalty", Journal of Chemometrics. **31**, 6, pp e2889
[٥٣] Qasim M., Algamal Z., and Ali H.M. (**2018**) "A binary QSAR model for classifying neuraminidase inhibitors of influenza A viruses (H1N1) using the combined minimum redundancy maximum relevancy criterion with the sparse support vector machine", SAR and QSAR in Environmental Research. **29**, 7, pp 517-527
[٥٤] Al-Thanoon N.A., Qasim O.S., and Algamal Z.Y. (**2019**) "A new hybrid firefly algorithm and particle swarm optimization for tuning parameter estimation in penalized support vector machine with application in chemometrics", Chemometrics and Intelligent Laboratory Systems. **184**, pp 142-152
[٥٥] Xia L.-Y., Wang Y.-W., Meng D.-Y., Yao X.-J., et al. (**2018**) "Descriptor selection via log-sum regularization for the biological activities of chemical structure", International journal of molecular sciences. **19**, 1, pp 30
[٥٦] Majumdar S., Basak S.C., Lungu C., Diudea M., et al. (**2018**) "Mathematical structural descriptors and mutagenicity assessment: a study with congeneric and diverse datasets$", SAR and QSAR in Environmental Research. **29**, 8, pp 579-590
[٥٧] Mozafari Z., Arab Chamjangali M., Arashi M., and Goudarzi N. (**2021**) "Performance of smoothly clipped absolute deviation as a variable selection method in the artificial neural network-based QSAR studies", Journal of Chemometrics. **35**, 5, pp e3338
[٥٨] HyperChem(TM) Professional 8.0.5 H.I., 1115 NW 4th Street, Gainesville, Florida 32601, USA., 2008 pp
[٥٩] Todeschini R., Lasagni M., and Marengo E. (**1994**) "New molecular descriptors for 2D and 3D structures. Theory", Journal of chemometrics. **8**, 4, pp 263-272
[٦٠] Todeschini R., Consonni V., Mauri A., and Pavan M., DRAGONs software for the calculation of molecular descriptors ,version 5.5 for Windows. Milan, Italy, in, 2007.
[٦١] IBM SPSS statistics 25 for Windows student G.D., and Mallery P., 2017 pp
[٦٢] Dessau R.B., and Pipper C.B. (**2008**) ""R"--project for statistical computing", Ugeskrift for laeger. **170**, 5, pp 328-330
[٦٣] Matlab Version R2017a T.M.I., Massachusetts, 2017 pp
[٦٤] Joy S., Nair P.S., Hariharan R., and Pillai R.M. (**2006**) "Detailed Comparison of the Protein-Ligand Docking Efficiencies of GOLD, a Commercial Package and ArgusLab, a Licensable Freeware", In silico biology. **6 6**, pp 601-605
[٦٥] ViewerLite v 5.0 A.I., San Diego, CA, 2001 pp

[٦٦] Nan X., Jiang Y.-F., Li H.-J., Wang J.-H., et al. (**2019**) "Design, synthesis and evaluation of sulfonylurea-containing 4-phenoxyquinolines as highly selective c-Met kinase inhibitors", Bioorganic & medicinal chemistry. **27**, 13, pp 2801-2812

[٦٧] Nan X., Zhang J., Li H.-J., Wu R., et al. (**2020**) "Design, synthesis and biological evaluation of novel N-sulfonylamidine-based derivatives as c-Met inhibitors via Cu-catalyzed three-component reaction", European Journal of Medicinal Chemistry. **200**, pp 112470

[٦٨] کاظمی م، (۱۳۹۷)، "انتخاب متغیر و تشخیص ساختار در مدل های نیمه پارامتری"، دانشگاه صنعتی شاهرود

[٦٩] Todeschini R., and Consonni V., Handbook of Molecular Descriptors, in, 2002.

[٧٠] Sun M., Chen J., Wei H., Yin S., et al. (**2009**) "Quantitative Structure–Activity Relationship and Classification Analysis of Diaryl Ureas Against Vascular Endothelial Growth Factor Receptor-2 Kinase Using Linear and Non-Linear Models", Chemical biology & drug design. **73**, 6, pp 644-654

[٧١] Qian F., Engst S ,.Yamaguchi K., Yu P., et al. (**2009**) "Inhibition of tumor cell growth, invasion, and metastasis by EXEL-2880 (XL880, GSK1363089), a novel inhibitor of HGF and VEGF receptor tyrosine kinases", Cancer research. **69**, 20, pp 8009-8016

[٧٢] Rose P.W., Beran B ,.Bi C., Bluhm W.F., et al. (**2010**) "The RCSB Protein Data Bank: redesigned web site and web services", Nucleic acids research. **39**, suppl_1, pp D392-D401

[٧٣] Chiari L.P.A., da Silva A.P., de Oliveira A.A., Lipinski C.F., et al. (**2021**) "Drug design of new sigma-1 antagonists against neuropathic pain: A QSAR study using partial least squares and artificial neural networks", Journal of Molecular Structure. **1223**, pp 129156

[٧٤] Zhou Q., Zhang N., Zhang C., Huang L., et al. (**2010**) "Molecular mechanism of enantioselective inhibition of acetolactate synthase by imazethapyr enantiomers", Journal of agricultural and food chemistry. **58 7**, pp 4202-4206

[٧٥] Parida P.K., Deka P., Shankar B., and Yadav R.N.S. (**2014**) "In silico antigenic site evaluation and antiviral therapy against dengue serotypes", Bangladesh Journal of Pharmacology. **9**, pp 83-95

[٧٦] Daina A., Michielin O., and Zoete V. (**2017**) "SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules", Scientific Reports. **7**, pp

[٧٧] Ertl P., and Schuffenhauer A. (**2009**) "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions", Journal of cheminformatics. **1**, 1, pp 1-11

Abstract

This study aims to use a screening technique to reduce the dimension of the data matrix. For this purpose, the new penalized Ridge preprocessing technique is used to reduce the data dimensions. Then, the Smoothly Clipped Absolute Deviation (SCAD) as a penalized variable selection method is used to select the most effective descriptors and modeling. Selected descriptors were used as input to construct the QSAR model using the SCAD method. For this purpose, the QSAR model was developed using multiple Ridge implementation as a preprocessing technique for data screening. The remained descriptors were entered into SCAD penalized model. After implementing SCAD to the train set data, some effective descriptors (9) were used in the model construction. The proposed optimum model (Ridge-SCAD) was evaluated using calculation of statistical parameter, applicability domain, and Y- Randomization test. The mentioned different tests show satisfactory results. The performance of the Ridge-SCAD model was applied for the suggestion of some potent c-Met inhibitors. The efficiency of the recommended compounds was investigated using molecular docking and calculation of Lipinski's rule of five. The obtained results prove the potency of the new compounds. For the comparision aim, the QSAR model was developed without using the Ridge method as a preprocessing technique, and all the calculated descriptors were defined as SCAD inputs. The SCAD model was constructed using the train set data, and 12 descriptors were obtained as significant descriptors. The biological activity of the test data was predicted using the optimum SCAD model. The statistical parameters of SCAD model were calculated and compared with the Ridge-SCAD model. The results proved that the high predictability of the Ridge-SCAD model against to the SCAD model.

Keywords: QSAR, c-Met, Preprocessing, Ridge, SCAD, Molecular Docking

Faculty of Chemistry

M.Sc. Thesis in Analytical Chemistry

Quantitative structure-activity relationship study of some new sulfonated c-Met inhibitors with anti-cancer activity

By:

Mohadeseh Lotfi

Supervisor**:**

Dr. Mansour Arab Chamjangali

March 2022