

دانشگاه صنعتی شاهرود

دانشکده شیمی

پایان نامه کارشناسی ارشد شیمی تجزیه

پیش بینی بازدارندگی (K_i) تعدادی از مشتقات تتراهیدرو پیریدین به
عنوان مهار کننده LIM پروتئین کیناز (LIMK) با استفاده از عملگر
انقباضی و انتخاب حداقل مطلق تطبیقی (ALASSO)

نگارنده:

سلطان نافع سلطان

استاد راهنما:

دکتر منصور عرب چم جنگلی

تیرماه ۱۴۰۵

تصویب نامه

تعهد نامه

اینجانب سلطان نافع سلطان دانشجوی دوره کارشناسی ارشد رشته شیمی گرایش شیمی تجزیه دانشکده شیمی دانشگاه صنعتی شاهرود نویسنده پایان نامه "پیش بینی بازدارندگی (K_i) تعدادی از مشتقات تتراهیدرو پیریدین به عنوان مهار کننده LIM پروتیین کیناز (LIMK) با استفاده از عملگر انقباض و انتخاب حداقل مطلق تطبیقی (ALASSO) " تحت راهنمایی جناب آقای دکتر منصور عرب چم جنگلی متعهد می شوم:

- تحقیقات این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش های محقق های دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان نامه تاکنون توسط خود یا افراد دیگر برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.

• کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا «Shahrood University of Technology» به چاپ خواهد رسید

- حقوق معنوی تمام افرادی که در به دست آمدن نتایج پایان نامه تأثیرگذار بوده اند 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 / امضاء دانشجو:

مالکیت نتایج و حق نشر

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

چکیده

پروتئین کینازها، به‌ویژه کیناز LIM (LIMK) که در تنظیم اسکلت سلولی و متاستاز سرطان نقش دارد، اهداف دارویی مهمی در تحقیقات نوین محسوب می‌شوند. با توجه به نقش LIMK در تهاجم سلول‌های سرطانی، توسعه مهارکننده‌های اختصاصی آن، گامی اساسی در جهت یافتن درمان‌های نوین ضدسرطان خواهد بود. مشتقات تتراهیدروپیریدین به دلیل انعطاف‌پذیری ساختاری و توانایی برقراری برهم‌کنش‌های متنوع، پتانسیل بالایی به عنوان مهارکننده LIMK نشان داده‌اند. با این حال، تعیین تجربی ثابت بازدارندگی (K_i) این ترکیبات فرآیندی زمان‌بر و پرهزینه است. در این پروژه، مدل‌سازی کمی ساختار-فعالیت (QSAR) به همراه داکینگ مولکولی، رویکردی کارآمد برای پیش‌بینی سریع و هدفمند فعالیت ترکیبات تتراهیدروپیریدین گزارش شده است. این رویکرد امکان شناسایی ترکیبات دارای پتانسیل بازدارندگی علیه پروتئین LIMK و همچنین طراحی ترکیبات جدید با ویژگی‌های ساختاری مطلوب را در میان مشتقات تتراهیدروپیریدین فراهم می‌سازد. در این مطالعه از روش انقباضی لاسوی تطبیقی (ALASSO) برای انتخاب متغیر و هم‌چنین ساخت مدل QSAR استفاده شده است. روش ALASSO با انتخاب بهترین متغیرها و کاهش تعداد متغیرها امکان ساخت مدل QSAR با کمترین خط بیش‌برازش و تفسیرپذیری و با قدرت پیش‌بینی بالا را فراهم می‌کند. اعمال روش ALASSO منجر به ایجاد مدلی با تعداد ۷ توصیفگر گردید. کارایی مدل با استفاده از داده‌های آزمون خارجی، روش LOO و پارامترهای آماری ارزیابی گردید. مقادیر MSE و R^2 به ترتیب برابر با ۰/۱۰ و ۰/۸۷ برای سری داده‌های آزمون خارجی و مقدار Q^2 برای پیش‌بینی کل داده‌ها به روش LOO برابر ۰/۷۸ بدست آمد. نتایج حاصل بیانگر مدلی تنک (با تعداد ۷ متغیر) بوده و خطای پیش‌بینی قابل قبولی را نشان می‌دهد. به علاوه مقادیر R^2 و Q^2 مدل به ترتیب از مقادیر پذیرفته شده ۰/۶ و ۰/۵ بیشتر بوده که بیانگر رضایت بخش بودن مدل QSAR است. با استفاده از مدل QSAR فعالیت ترکیبات پیشنهادی پیش‌بینی گردید و درستی ترکیبات پیشنهادی با استفاده از داکینگ مولکولی بررسی شده و به اثبات رسید. بنابراین این مطالعه ضمن غلبه بر محدودیت‌های مدل‌های QSAR سنتی، زمینه را برای طراحی منطقی و بهینه‌سازی مشتقات جدید تتراهیدروپیریدین به عنوان مهارکننده‌های مؤثر LIMK هموار می‌کند.

کلمات کلیدی: پروتئین کیناز LIM (LIMK)، مشتقات تتراهیدروپیریدین، Adaptive LASSO (ALASSO)، سرطان، مهارکننده‌های پروتئین کیناز

فهرست مطالب

فصل اول: مقدمه.....	۱
۱-۱- مقدمه	۲
۲-۱- علم کمومتری کس.....	۲
۳-۱- ارتباط کمی ساختار-فعالیت (QSAR).....	۳
۴-۱- گام‌های کلیدی در توسعه مدل‌های پیش‌گوی QSAR.....	۴
۱-۴-۱- جمع‌آوری و آماده‌سازی داده‌ها در مدل‌سازی QSAR.....	۵
۲-۴-۱- رسم ساختار ترکیبات و بهینه‌سازی.....	۶
۳-۴-۱- تولید توصیف‌گرهای مولکولی.....	۶
۴-۴-۱- پیش‌پردازش و انتخاب توصیف‌گرهای کلیدی.....	۷
۱-۴-۴-۱- روش کمترین مربعات معمولی (OLS).....	۹
۲-۴-۴-۱- لاسوی تطبیقی (ALASSO).....	۱۰
۵-۴-۱- ساخت و توسعه مدل‌های پیش‌بینی‌کننده.....	۱۰
۶-۴-۱- ارزیابی قدرت پیش‌بینی مدل.....	۱۱
۱-۶-۴-۱- ضریب همبستگی خطی (R).....	۱۲
۲-۶-۴-۱- ضریب تعیین (R^2).....	۱۲
۳-۶-۴-۱- مجموع مربعات خطاهای پیش‌بینی (PRESS).....	۱۲
۴-۶-۴-۱- خطای استاندارد پیش‌بینی (SEP).....	۱۳
۵-۶-۴-۱- خطای مطلق میانگین (MAE).....	۱۳
۶-۶-۴-۱- خطای نسبی پیش‌بینی (REP).....	۱۳
۷-۶-۴-۱- میانگین مربع خطا (MSE).....	۱۴
۸-۶-۴-۱- میانگین خطای نسبی (MRE).....	۱۴
۹-۶-۴-۱- ارزیابی مدل با استفاده از مجموعه آزمون.....	۱۴
۱۰-۶-۴-۱- اعتبارسنجی متقابل با روش حذف تک‌تایی (LOO-CV).....	۱۵
۱۱-۶-۴-۱- تحلیل نمودار خطای باقیمانده.....	۱۵
۱۲-۶-۴-۱- آزمون Y-تصادفی.....	۱۶
۵-۱- شبیه‌سازی اتصال مولکولی.....	۱۶
۱-۵-۱- پیش‌پردازش ساختار پروتئین برای داکینگ.....	۱۷
۲-۵-۱- پیش‌پردازش لیگاندها و مولکول‌های کاندید.....	۱۷
۶-۱- سرطان.....	۱۸
۷-۱- ضرورت تحقیق.....	۱۹
۸-۱- پیشینه پژوهش.....	۲۱
فصل دوم: بخش تجربی.....	۲۵
۲-۱- نرم افزارهای به‌کار رفته در این پژوهش.....	۲۶
۲-۲- نرم افزار هایپرکم.....	۲۶

۲۷	۳-۲- نرم افزار دراگون
۲۷	۴-۲- نرم افزار SPSS
۲۸	۵-۲- نرم افزار R/Rstudio
۲۹	۶-۲- نرم افزار MATLAB
۳۲	۱۰-۲- نرم افزار لیگ پلات
۳۲	۲-۱۱- شرح ترکیبات و داده‌های تجربی استفاده‌شده در مدل‌سازی
۳۷	فصل سوم: بحث و نتایج
۳۸	۳-۱- مدل‌سازی فعالیت دارویی مشتقات منتخب با استفاده از مدل Adaptive LASSO (ALASSO)
۳۹	۳-۲- بهینه‌سازی و آماده‌سازی ساختار سه‌بعدی ترکیبات
۴۰	۳-۳- تولید توصیف‌گرهای کمی مولکولی برای مدل‌سازی QSAR
۴۰	۳-۳- انتخاب بهینه‌ی توصیف‌گرهای مؤثر
۴۱	۳-۴- انتخاب متغیرها با مدل رگرسیون انقباضی ALASSO و ارزیابی عملکرد
۴۳	۳-۵- ارزیابی جامع عملکرد و قابلیت اطمینان مدل پیش‌بین فعالیت زیستی مبتنی بر ALASSO
۴۴	۳-۵-۱- ارزیابی عملکرد مدل پیش‌بین ALASSO بر روی مجموعه داده آزمون
	۳-۵-۲- ارزیابی توان پیش‌بینی و تعمیم‌پذیری مدل ALASSO با استفاده از اعتبارسنجی حذف
۴۷	تک‌نمونه‌ای (LOO-CV)
۵۲	۳-۵-۳- ارزیابی عملکرد مدل ALASSO بر پایه شاخص‌های آماری
۵۳	۳-۵-۴- بررسی اعتبار مدل ALASSO با آزمون Y-تصادفی
۵۵	فصل چهارم: بحث و نتیجه‌گیری
۵۶	۴-۱- جمع‌بندی نهایی عملکرد مدل ALASSO و تحلیل اهمیت توصیف‌کننده‌ها
	۴-۲- پیشنهاد ترکیبات جدید با فعالیت دارویی مناسب با استفاده از عملکرد انقباض و انتخاب حداقل مطلق
۵۸	تطبیقی (ALASSO)
۶۲	۴-۳- مطالعه داکینگ مولکولی
۶۳	۴-۳-۱- آماده‌سازی پروتئین
۶۴	۴-۳-۲- آماده‌سازی لیگاند
۶۵	۴-۳-۳- آماده‌سازی ترکیبات پیشنهادی برای فرایند داکینگ مولکولی
۶۵	۴-۳-۴- اجرای فرایند داکینگ مولکولی ترکیبات پیشنهادی با پروتئین 8AAU
۶۶	۴-۴- نتایج داکینگ مولکولی
۶۹	۴-۵- نتایج مقایسه‌ای قانون لیپینسکی
۷۰	۴-۶- نتیجه‌گیری نهایی
۷۲	۴-۷- آینده‌نگری
۷۳	۴-۸- پیوست
۷۵	۴-۹- منابع

فهرست جدول ها

- جدول (۲-۱). ساختار و مقادیر pK_i ترکیبات مورد مطالعه. ۳۴
- جدول (۳-۱). جدول توصیف گرهای انتخاب شده با استفاده از روش ALASSO. ۴۲
- جدول (۳-۲). ارزیابی عملکرد پیش بینی مدل ALASSO روی ترکیبات مجموعه آزمون. ۴۵
- جدول (۳-۳). نتایج حاصل از اعتبارسنجی حذف تک نمونه ای (LOO) برای مدل ALASSO ، شامل مقادیر تجربی و پیش بینی شده pK_i و درصد خطای محاسبه شده برای هر ترکیب. ۴۸
- جدول (۳-۴). شاخص های آماری ارزیابی عملکرد مدل QSAR توسعه یافته، شامل پارامترهای خطا و معیارهای برازش برای مجموعه های آزمون و Q_{LOO}^2 اعتبارسنجی حذف تک نمونه ای (LOO). ۵۲
- جدول (۴-۱) ساختار و فعالیت ترکیب ۱۶ و ۳۰ و ترکیبهای پیشنهادی. ۶۰
- جدول (۴-۲) پارامترهای محاسبه شده قانون لیپینسکی برای ترکیب ۳۰ و ۱۶ و ترکیبهای پیشنهادی. ۷۰
- جدول (۴-۱) ساختار و فعالیت ترکیب ۱۶ و ۳۰ و ترکیبهای پیشنهادی. ۶۰
- جدول (۴-۲) پارامترهای محاسبه شده قانون لیپینسکی برای ترکیب ۳۰ و ۱۶ و ترکیبهای پیشنهادی. ۷۰

فهرست شکل ها

- شکل (۲-۱). اسکلت اصلی ترکیبات تترا هیدرو پیریدین..... ۳۳
- شکل (۳-۱). نمودار مقادیر پیش بینی شده مدل لاسوی تطبیقی بر حسب مقادیر تجربی برای داده های سری آزمون..... ۴۶
- شکل (۳-۲). نمودار باقیمانده های استاندارد شده بر حسب مقادیر تجربی برای داده های سری آزمون..... ۴۷
- شکل (۳-۳). نمودار مقایسه مقادیر پیش بینی شده pK_i توسط مدل ALASSO بر حسب مقادیر تجربی متناظر آن ها بر اساس اعتبارسنجی حذف تک نمونه ای (LOO)..... ۵۰
- شکل (۳-۴). نمودار باقی مانده های استاندارد شده بر حسب مقادیر تجربی pK_i برای مدل ALASSO در چهارچوب اعتبارسنجی حذف تک نمونه ای (LOO)..... ۵۲
- شکل (۳-۵). نتایج Y- تصادفی..... ۵۴
- شکل (۴-۱) نمودار درصد مشارکت توصیف کننده ها در مدل ALASSO..... ۵۶
- شکل (۴-۲) ساختار کریستالوگرافی پروتئین 8AAU را نشان می دهد..... ۶۴
- شکل (۴-۳) برهمکنش ترکیب ۳۰ با اسید آمینه های کلیدی جایگاه فعال پروتئین 8AAU..... ۶۶
- شکل (۴-۴) برهمکنش ترکیب ۱۶ با اسید آمینه های کلیدی جایگاه فعال پروتئین 8AAU..... ۶۷
- شکل (۴-۵) برهمکنش ترکیب G6 با اسید آمینه های کلیدی جایگاه فعال پروتئین 8AAU..... ۶۸
- شکل (۴-۶) برهمکنش ترکیب G2 با اسید آمینه های کلیدی جایگاه فعال پروتئین 8AAU..... ۶۸

فهرست رابطه ها

۹.....	رابطه (۱ - ۱).....
۹.....	رابطه (۲ - ۱).....
۱۲.....	رابطه (۴ - ۱).....
۱۲.....	رابطه (۵ - ۱).....
۱۳.....	رابطه (۶ - ۱).....
۱۳.....	رابطه (۷ - ۱).....
۱۳.....	رابطه (۸ - ۱).....
۱۴.....	رابطه (۹ - ۱).....
۱۴.....	رابطه (۱۰ - ۱).....
۱۴.....	رابطه (۱۱ - ۱).....
۱۵.....	رابطه (۱۲ - ۱).....
۴۳.....	رابطه (۱ - ۳).....
۴۳.....	رابطه (۲ - ۳).....
۴۶.....	رابطه (۳ - ۳).....

فصل اول: مقدمه

۱-۱- مقدمه

سرطان، یکی از چالش‌های اساسی بشر در عصر حاضر، نیازمند دستیابی به راهکارهای درمانی مؤثر و هدفمند است. این بیماری پیچیده و چندعاملی، با گسترش روزافزون، نیاز به درمان‌هایی ایمن و مقرون به صرفه را بیش از پیش نمایان ساخته است. از آنجا که روش‌های سنتی طراحی دارو اغلب مبتنی بر آزمون و خطا بوده و فرآیندی هزینه‌بر و زمان‌گیر محسوب می‌شوند، توجه پژوهشگران به بهره‌گیری از رویکردهای نظری و محاسباتی برای پیش‌بینی اثربخشی ترکیبات پیش از سنتز آن‌ها متمرکز شده‌اند. در این راستا، علم کمومتریکس^۱ با ارائه ابزارهای آماری و تحلیلی، به ویژه از طریق روش‌های کمی ساختار-فعالیت (QSAR)^۲، امکان برقراری ارتباط میان ویژگی‌های ساختاری مولکول‌ها و فعالیت‌های زیستی آن‌ها را فراهم کرده است. این رویکرد با پیش‌بینی فعالیت دارویی ترکیبات، گامی مؤثر در طراحی داروهای نوین محسوب می‌شود. یکی از معیارهای مهم در ارزیابی قدرت یک مهارکننده دارویی، ثابت ترمودینامیکی K_i است که بیانگر قدرت اتصال آن به آنزیم هدف است. هرچه مقدار K_i کوچک‌تر باشد، مهارکنندگی ترکیبات دارویی قوی‌تر است. غالباً تعیین آزمایشگاهی مقادیر K_i معمولاً پرهزینه و زمان‌بر است، لذا توسعه مدل‌های پیشگو بر اساس رابطه کمی ساختار-فعالیت (QSAR) می‌تواند جایگزینی مناسب و کارآمدی برای آزمون‌های آزمایشگاهی در تعیین مقادیر K_i باشد.

۱-۲- علم کمومتریکس

علم کمومتریکس، که ریشه‌های آن به دهه‌های اولیه دهه ۱۹۷۰ میلادی بازمی‌گردد، حاصل یک هم‌افزایی میان دانشمندان برجسته‌ای است که بر ادغام روش‌های آماری و محاسباتی با تحلیل داده‌های شیمیایی تمرکز داشتند. هدف اصلی این علم، ارائه ابزاری قدرتمند برای حل مسائل پیچیده در حوزه‌های مختلف شیمی است. کمومتریکس به‌طور اساسی، کاربرد یکپارچه ریاضیات، آمار و علوم کامپیوتر در علم شیمی برای تحقق اعداف زیر است:

¹ Chemometrics

² Quantitative Structure-Activity Relationship (QSAR)

بهینه‌سازی فرآیند: این بخش بر طراحی آزمایش‌های (DOE)¹ تمرکز دارد تا با کمترین هزینه و زمان، بیشترین بازدهی و اطلاعات حاصل شود. هدف، هدایت آزمایش‌ها به سمتی است که نتایج بهینه و معنادار به دست آید.

استخراج دانش: در دنیای امروز، حجم داده‌های تولید شده توسط دستگاه‌های تجزیه ای بسیار بالاست. کمومتریکس ابزارهایی را برای پردازش و تحلیل این داده‌های پرحجم و چندمتغیره (مانند طیف‌سنجی‌ها) فراهم می‌کند تا حداکثر اطلاعات مفید و قابل تفسیر از آن‌ها استخراج گردد.

مدل‌سازی پیش‌بینانه: در اینجا هدف اصلی، خلق مدل‌هایی است که قادر باشند رفتار سامانه‌های شیمیایی را به دقت توصیف کرده و نتایج آینده را پیش‌بینی نمایند. این مدل‌ها به محققان کمک می‌کنند تا ترکیبات جدید را طراحی کرده، فرآیندها را بهینه‌سازی کنند یا نتایج آزمایش‌های آتی را تخمین بزنند.

از زمان انتشار مقالات اولیه در اواسط دهه ۱۹۷۰، کمومتریکس به سرعت جایگاه خود را به عنوان یک رشته علمی مستقل و تأثیرگذار تثبیت کرد. این علم توانسته است داده‌های خام و پیچیده را به اطلاعات ارزشمند شیمیایی تبدیل کند و به ابزاری کلیدی در دست دانشمندان تبدیل شود. امروزه، کاربردهای کمومتریکس بسیار گسترده و متنوع است و در حوزه‌هایی چون شیمی تجزیه، کشفیات دارویی (از جمله طراحی داروهای ضدسرطان)، کنترل فرآیندهای صنعتی، بیوانفورماتیک و حتی علوم محیطی مورد استفاده قرار می‌گیرد. این علم، پلی میان دنیای داده‌های حجیم و فهم عمیق‌تر پدیده‌های شیمیایی ایجاد کرده است [۴-۱].

۳-۱- ارتباط کمی ساختار-فعالیت (QSAR)

رابطه کمی ساختار-فعالیت (QSAR) یک روش محاسباتی قدرتمند است که به ما امکان می‌دهد تا چگونگی تأثیر ویژگی‌های ساختاری ترکیبات شیمیایی بر فعالیت بیولوژیکی یا دارویی آن‌ها را درک کرده و پیش‌بینی کنیم. این رویکرد که تاریخچه‌ای طولانی در طراحی دارو دارد، به دنبال یافتن ارتباطی نظام‌مند

¹ Design of Experiments

بین خواص مولکولی و عملکرد آنهاست. هسته اصلی QSAR بر ایجاد مدل‌های ریاضی استوار است. این مدل‌ها با استفاده از تکنیک‌های آماری، به‌ویژه روش‌های رگرسیون، یک رابطه قابل سنجش و معنادار را بین شاخص‌های کمی ساختاری مولکول‌ها و معیارهای فعالیت زیستی آنها (مانند قدرت مهار یک آنزیم) برقرار می‌کنند. فرض بنیادی این است که تغییر در ساختار و ویژگی‌های مولکولی، منجر به تغییرات قابل پیش‌بینی در پاسخ بیولوژیکی می‌شود. در این فرایند، توصیف‌گرها^۱ نقش حیاتی ایفا می‌کنند. توصیف‌گرها، معیارهای عددی یا توصیفی هستند که خصوصیات مختلف یک مولکول را کمی‌سازی می‌کنند؛ این خصوصیات می‌توانند شامل ابعاد هندسی، توزیع الکترونی، خواص فیزیکوشیمیایی مانند آب‌گریزی یا قطبیت، و بسیاری ویژگی‌های دیگر باشند. انتخاب توصیف‌گرهای مناسب، گام اول و اساسی در ساخت یک مدل QSAR مناسب است. مدل‌های QSAR با پیش‌بینی درست فعالیت مولکول‌ها از طریق توصیف‌گرها، قادرند تا تعداد زیادی از ترکیبات شیمیایی را بر اساس پتانسیل دارویی مورد انتظارشان رتبه بندی کنند. این فرایند مدل سازی کامپیوتری، با ارائه پیش‌بینی‌های اولیه از میزان فعالیت ترکیبات، نیاز به بررسی‌های آزمایشگاهی گسترده را به شدت کاهش داده و در نتیجه، روند کشف و توسعه دارو را تسریع و بهینه‌سازی می‌کند [۵-۷].

۱-۴- گام‌های کلیدی در توسعه مدل‌های پیش‌گوی QSAR

توسعه مدل QSAR فرآیندی گام‌به‌گام برای یافتن رابطه کمی بین ساختار مولکولی و اثر زیستی است.

گردآوری داده: جمع‌آوری و پالایش اطلاعات فعالیت زیستی ترکیبات.

بهینه‌سازی ساختار: رسم و پایدارسازی هندسه^۲ سه‌بعدی مولکول‌ها.

تولید توصیف‌گر: محاسبه^۳ شاخص‌های عددی نمایانگر ویژگی‌های مولکولی.

گزینش توصیف‌گرهای کلیدی: حذف توصیف‌گرهای نامرتبط برای جلوگیری از بیش‌برازش و افزایش تفسیرپذیری.

¹ Descriptors

ساخت مدل: ایجاد مدل رگرسیونی برای ارتباط بین توصیفگرها و فعالیت زیستی.

اعتبارسنجی: ارزیابی عملکرد مدل با معیارهای آماری برای تضمین تعمیم‌پذیری به داده‌های جدید [۸].

۱-۴-۱- جمع‌آوری و آماده‌سازی داده‌ها در مدل‌سازی QSAR

بنیان هر مدل QSAR موفق، بر پایه‌ی مجموعه‌ای دقیق و سازمان‌یافته از داده‌ها بنا نهاده می‌شود. این مرحله حیاتی شامل انتخاب هوشمندانه‌ی ترکیبات شیمیایی است که فعالیت زیستی آن‌ها به صورت تجربی و از منابع علمی معتبر تعیین شده است. در این گام، اطمینان از وجود همگنی کافی در ساختار ترکیبات و یکنواختی شرایط آزمایشگاهی که فعالیت‌ها در آن اندازه‌گیری شده‌اند، از اهمیت بالایی برخوردار است؛ چرا که هرگونه ناهماهنگی می‌تواند نتایج تحلیل‌های بعدی را تحت تأثیر قرار دهد. پس از گزینش اولیه ترکیبات، فرآیند آماده‌سازی داده‌ها آغاز می‌شود. در این بخش، ترکیبات یا داده‌های فعالیت‌ی که به نظر می‌رسد دارای خطا یا نویز قابل توجهی هستند، شناسایی و حذف می‌شوند تا از هرگونه تأثیر منفی بر فرایند مدل‌سازی جلوگیری شود. در ادامه، مجموعه داده‌های پالایش شده به دو زیرمجموعه مجزا به نام های مجموعه آموزش و مجموعه آزمون تقسیم می‌شوند:

مجموعه آموزشی^۱: عموماً بخش بزرگ‌تری از داده‌ها را به خود اختصاص می‌دهد، به عنوان مجموعه اصلی برای “یادگیری”^۲ و ساخت مدل QSAR استفاده می‌شود و مدل با استفاده از روابط موجود در این مجموعه، پارامترهای خود را تنظیم می‌کند. مجموعه آزمون^۳: شامل داده‌هایی است که کاملاً از مراحل ساخت مدل کنار گذاشته شده و برای سنجش نهایی توانایی مدل در پیش‌بینی فعالیت ترکیبات جدید و ناشناخته مورد استفاده قرار می‌گیرد. این ارزیابی، معیاری قابل اعتماد از عملکرد واقعی و قابلیت تعمیم‌پذیری مدل ارائه می‌دهد. این تقسیم‌بندی داده‌ها به افزایش صحت و اعتبار مدل QSAR توسعه‌یافته کمک زیادی می‌کند.

¹ Training Set

² Learning

³ Test Set

۱-۴-۲- رسم ساختار ترکیبات و بهینه سازی

فرآیند مدل سازی با ترسیم دقیق ساختارهای مولکولی آغاز می شود. برای این منظور، از نرم افزارهای بهینه سازی قدرتمند بهره گرفته شده است. پس از ترسیم اولیه، مرحله ی بهینه سازی ساختاری صورت می گیرد. هدف از این مرحله، رسیدن به پایدارترین حالت ممکن برای هر مولکول با کمترین سطح انرژی است. در میان روش های متعدد بهینه سازی، رویکرد نیمه تجربی AM1 که در دسته بندی روش های مبتنی بر مکانیک کوانتومی قرار می گیرد، برای این تحقیق انتخاب شده است. این روش به ما امکان می دهد تا ساختار مولکول های مورد مطالعه را مدل سازی و انرژی آن ها را محاسبه کنیم. به دست آوردن اطلاعاتی نظیر مختصات سه بعدی مولکول در حالت بهینه و پایدار، پیش نیازی اساسی برای محاسبه ی توصیف گرهای مورد نیاز مدل QSAR است. این مختصات، اساس محاسبات بعدی را تشکیل می دهند [۱۰].

۱-۴-۳- تولید توصیف گرهای مولکولی

برای درک و کمی سازی روابط میان ساختار مولکولی و فعالیت زیستی، از مجموعه ای از شاخص های عددی موسوم به توصیف گر استفاده می شود. این توصیف گرها، در واقع، نمایانگر ویژگی های منحصر به فرد فیزیکوشیمیایی هر مولکول هستند [۱۱].

محاسبه ی دقیق این توصیف گرها، سنگ بنای مطالعات QSAR را تشکیل می دهد؛ زیرا باید این شاخص ها، جنبه های پیچیده ی ساختاری مولکول را به طور قابل قبولی بازتاب دهند. هرگونه خطا در تعیین ساختار مولکولی، به طور مستقیم بر صحت محاسبه ی توصیف گرها تأثیر گذاشته و منجر به ایجاد مجموعه ای از توصیف گرهای ناکارآمد یا نادرست خواهد شد. در نتیجه، مدل های QSAR که بر پایه ی چنین توصیف گرهای معیوب بنا شوند، قادر به ارائه ی پیش بینی های درست و قابل اعتماد نخواهند بود [۱۲]. توصیف گرها عموماً به دو دسته ی اصلی طبقه بندی می شوند:

توصیف گرهای تجربی: این دسته از شاخص ها از طریق اندازه گیری های مستقیم آزمایشگاهی به دست

می آیند.

توصیف‌گرهای نظری (محاسباتی): این توصیف‌گرها با استفاده از روش‌های محاسباتی بدون نیاز به اطلاعات تجربی محاسبه و استخراج می‌گردند.

با توجه به سهولت در محاسبه و همچنین قابلیت اعمال بر روی مولکول‌هایی که هنوز سنتز نشده‌اند، توصیف‌گرهای نظری از کاربرد گسترده‌تری در مطالعات QSAR نسبت به توصیف‌گرهای تجربی برخوردارند [۱۳].

۱-۴-۴- پیش‌پردازش و انتخاب توصیف‌گرهای کلیدی

در عصر حاضر، با پیشرفت‌های چشمگیر در حوزه شیمی محاسباتی و در دسترس بودن نرم‌افزارهای پیشرفته، امکان محاسبه‌ی طیف وسیعی از توصیف‌گرها برای هر مولکول فراهم شده است. این امر منجر به شکل‌گیری مجموعه‌داده‌های حجیم در مطالعات QSAR شده است که اغلب با چالش‌هایی نظیر همبستگی بالای بین توصیف‌گرها و حضور ویژگی‌های نامرتبط با پاسخ مدل مواجه هستند. این پیچیدگی‌ها، فرآیند ساخت مدل را دشوار می‌سازند. برای غلبه بر این موانع، ابتدا مجموعه‌ی داده‌ها تحت فرآیند پیش‌پردازش قرار می‌گیرند. این مرحله شامل مجموعه‌ای از عملیات سازماندهی و پاک‌سازی داده‌های خام است تا آن‌ها را برای تحلیل و مدل‌سازی آماده سازد. در این راستا، توصیف‌گرهایی که دارای مقادیر تکراری (بیش از ۹۰٪ مشابهت) هستند، یا همبستگی شدیدی با سایر توصیف‌گرها نشان می‌دهند، و همچنین آن دسته از ویژگی‌هایی که ارتباط معناداری با متغیر وابسته ندارند، شناسایی و حذف می‌شوند. این اقدام، علاوه بر کاهش ابعاد مجموعه داده، به تمرکز بر ویژگی‌های کلیدی و مؤثر کمک کرده و در نهایت منجر به ارتقاء دقت و قابلیت تعمیم مدل‌های QSAR می‌گردد. با وجود اینکه فرآیندهای کاهش ابعاد، تعداد قابل توجهی از توصیف‌گرها را حذف می‌کنند، همچنان ممکن است تعداد زیادی از آن‌ها باقی بمانند که برای ساخت یک مدل بهینه، فراتر از نیاز است. لذا، مرحله‌ی گزینش توصیف‌گرهای کلیدی از میان موارد باقیمانده، امری ضروری محسوب می‌شود. انتخاب هوشمندانه‌ی توصیف‌گرها دارای مزایای متعددی است: افزایش قابلیت تفسیرپذیری مدل: استفاده از تعداد کمتری توصیف‌گر، درک و تفسیر عملکرد مدل را تسهیل می‌بخشد.

کاهش خطر بیش‌برازش^۱: حذف ویژگی‌های غیرضروری، احتمال ایجاد مدلی که صرفاً بر روی داده‌های آموزشی عملکرد خوبی دارد را کاهش می‌دهد.

بهبود کارایی محاسباتی: مدل‌های ساده‌تر، با سرعت پردازش بالاتر و هزینه‌ی محاسباتی کمتری ساخته می‌شوند [۱۵ و ۱۴].

هدف در این مرحله، کاهش ابعاد فضای ورودی از طریق حذف توصیفگرهای تکراری یا نامرتبط، بدون از دست دادن اطلاعات حیاتی است. فرآیند انتخاب متغیر^۲ به طور کلی در دو دسته‌ی عمده قرار می‌گیرد:

روش‌های خطی: این رویکرد بر فرض وجود رابطه‌ی خطی بین متغیر پاسخ و توصیفگرها استوار است. روش‌های کلاسیک مانند انتخاب گام به گام افزایشی^۳، حذف گام به گام کاهشی^۴ و رگرسیون گام به گام^۵ در این دسته قرار می‌گیرند [۱۶]. همچنین، روش‌های مبتنی بر جریمه^۶ مانند LASSO^۷، SCAD^۸ و ALASSO^۹ که به طور همزمان به انتخاب توصیفگرها و همچنین تخمین ضرایب رگرسیونی می‌پردازند، از اهمیت بالایی برخوردارند [۱۵].

روش‌های غیرخطی: در این رویکرد، فرض بر ماهیت غیرخطی رابطه بین متغیر پاسخ و توصیفگرها است. از جمله مهم‌ترین روش‌های غیرخطی می‌توان به شبکه‌های عصبی مصنوعی (ANN^{۱۰})، ماشین بردار پشتیبان (SVM^{۱۱})، درخت تصمیم (DT^{۱۲}) و جنگل تصادفی (RF^{۱۳}) اشاره کرد [۱۶].

روش‌های مبتنی بر جریمه، به دلیل افزایش دقت پیش‌بینی، پایداری بیشتر و ارائه‌ی مدل‌هایی با قابلیت تفسیرپذیری بهتر نسبت به روش‌های خطی کلاسیک، مزایای قابل توجهی دارند. روش‌های کلاسیک ممکن

¹ Overfitting

² Feature Selection

³ Forward Selection

⁴ Backward Elimination

⁵ Stepwise Regression

⁶ Penalized Methods

⁷ Least Absolute Shrinkage and Selection Operator

⁸ Smoothly Clipped Absolute Deviation

⁹ Adaptive LASSO

¹⁰ Artificial Neural Networks (ANN)

¹¹ Support Vector Machine (SVM)

¹² Decision Tree (DT)

¹³ Random Forest (RF)

است با چالش‌هایی نظیر سوگیری در برآورد ضرایب و عملکرد ضعیف در مواجهه با هم‌خطی^۱ روبرو شوند [۱۵ و ۱۶] که روش‌های مبتنی بر جریمه این چالش‌ها را کاهش می‌دهد.

در این پژوهش، با هدف انتخاب مؤثرترین توصیفگرها در مدل QSAR، از روش (ALASSO^۲) بهره گرفته شده است. این روش که نوعی تکنیک جریمه‌ای محسوب می‌شود، به طور کارآمدی به گزینش ویژگی‌های مرتبط می‌پردازد. برای درک بهتر ALASSO، ابتدا به معرفی روش کمترین مربعات معمولی (OLS^۳) پرداخته می‌شود:

۱-۴-۴-۱- روش کمترین مربعات معمولی (OLS)

مدل رگرسیون خطی چندگانه به صورت زیر تعریف می‌شود:

$$y = \chi\beta + \epsilon \quad \text{رابطه (۱-۱)}$$

مطابق معادله $y = (y_1, y_2, \dots, y_n)^T$ بردار پاسخ، $X = (x_1, \dots, x_n)^T$ ماتریس طرح در اندازه $n \times p$ با رتبه کامل ستونی p شامل متغیرهای مستقل (پیشگو)، $\beta = (\beta_1, \dots, \beta_p)^T$ بردار ضرایب رگرسیونی و $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$ بردار مولفه‌های خطای تصادفی است.

برآوردگر OLS با کمینه کردن مجموع مربعات خطاها به دست می‌آید:

$$\beta^{OLS} = (X^T X)^{-1} X^T y \quad \text{رابطه (۲-۱)}$$

استفاده از برآوردگر OLS با وجود مزایای فراوانی همچون سادگی محاسبات، بدون بایاس بودن ضرایب، تفسیرپذیری مناسب [۲۴، ۱۸، ۱۷]، دارای محدودیت‌هایی نیز هست. نخست، اگرچه OLS یک برآوردگر بدون بایاس است، اما در حضور هم‌خطی شدید بین متغیرهای پیش‌بینی‌کننده یا در مسائل با ابعاد بالا، واریانس ضرایب برآوردشده می‌تواند افزایش یابد و در نتیجه دقت پیش‌بینی روی داده‌های جدید کاهش پیدا کند. در چنین شرایطی، اعمال محدودیت بر ضرایب و کوچک‌سازی آن‌ها به سمت صفر، همانند روش‌های

^۱ Multicollinearity

^۲ Adaptive LASSO (ALASSO)

^۳ Ordinary Least Squares (OLS)

رگرسیون جریمه‌شده، با ایجاد اندکی بایاس و کاهش واریانس، معمولاً موجب بهبود عملکرد پیش‌بینی می‌شود. افزون بر این، از آنجا که OLS فاقد سازوکار داخلی برای انتخاب متغیر است، در مسائلی با تعداد زیاد پیش‌بینی‌کننده‌ها، مدل نهایی ممکن است شامل متغیرهای متعدد باشد که تفسیرپذیری آن را کاهش می‌دهد.

۱-۴-۲- لاسوی تطبیقی (ALASSO)

هرچند شباهت‌های ظاهری زیادی بین تابع جریمه لاسو و لاسوی تطبیقی وجود دارد، اما لاسوی تطبیقی درست برعکس لاسو، برای ضرایب مختلف جریمه‌های یکسانی را در نظر نمی‌گیرد. لاسوی تطبیقی در واقع همان لاسو معمولی است ولی تابع جریمه آن وزن دار بوده و برای ضرایب وزن خاصی در نظر می‌گیرد:

$$\hat{\beta}^{\text{ALASSO}} = \operatorname{argmin} \left\{ \sum_{i=1}^n (y_i - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p \widehat{W}_j |\beta_j| \right\}$$

که در آن $\widehat{W}_j$ وزن بردار بوده که به صورت $\widehat{W}_j = \frac{1}{|\beta_j|^\gamma}$ تعریف شده که اگر به درستی انتخاب شوند دارای کارایی پیشگویی بالایی خواهد بود. در این رابطه، γ یک پارامتر تنظیم مثبت ($\gamma > 0$) است که میزان تأثیر وزن‌ها بر جریمه اعمال‌شده به ضرایب را کنترل می‌کند. هرچه مقدار γ بزرگ‌تر باشد، اختلاف بین جریمه ضرایب کوچک و بزرگ افزایش یافته و انتخاب متغیرها با دقت بیشتری انجام می‌شود. لاسوی تطبیقی در سال ۲۰۰۶ توسط زو^۲ به منظور رفع مشکل ناسازگاری لاسو ارائه شد که با اعمال مقدار جریمه‌های مختلف روی ضرایب با بزرگی متفاوت کارایی مدل پیشگو را بهبود می‌بخشد [۱۹].

۱-۴-۵- ساخت و توسعه مدل‌های پیش‌بینی‌کننده

پس از طی فرآیند دقیق گزینش توصیفگرهای مولکولی و انتخاب مؤثرترین آن‌ها توسط روش ALASSO، مرحله‌ی کنونی به ساخت مدل پیش‌بینی‌کننده اختصاص دارد. این مدل، که بر پایه‌ی مجموعه‌ی منتخب توصیفگرها استوار است، وظیفه‌ی برقراری ارتباط کمی میان ساختار مولکولی و فعالیت زیستی را بر عهده

¹ Adaptive LASSO

² Zou

دارد ALASSO به عنوان یک روش رگرسیون با قابلیت انتخاب ویژگی، مدل نهایی را با ضرایب تخمین زده شده برای ویژگی‌های منتخب ارائه می‌دهد. این ضرایب، نمایانگر اهمیت و جهت تأثیر هر توصیفگر در پیش‌بینی فعالیت زیستی هستند. استفاده از ALASSO در این مرحله تضمین می‌کند که مدل توسعه‌یافته، بر روی مهم‌ترین ویژگی‌ها تمرکز کرده، از پیچیدگی غیرضروری پرهیز نموده و در نتیجه، از قابلیت تعمیم‌پذیری و پایداری بالایی برخوردار باشد. هرچند مدل ALASSO یک مدل خطی می‌باشد و ممکن است برای داده‌های مرتبط با ساختار و فعالیت بیولوژیکی که ذاتا دارای تابعیت غیرخطی هستند دارای محدودیت‌هایی باشد ولی روشن بودن مدل ریاضی با تعدا توصیفگرهای کم و ضرایبی با برآورد درست امکان تفسیر مدل و بررسی اثر هر توصیفگر روی فعالیت را فراهم می‌کند.

۱-۴-۶- ارزیابی قدرت پیش‌بینی مدل

پس از ساخت مدل پیش‌بینی‌کننده بر اساس توصیفگرهای منتخب توسط ALASSO، مرحله‌ی حیاتی بعدی، ارزیابی دقیق عملکرد و سنجش قابلیت اطمینان آن است. هدف اصلی در این مرحله، اطمینان از توانایی مدل در ارائه پیش‌بینی‌های دقیق برای ترکیبات جدید و ناشناخته است، که این امر پیش از هرگونه اقدام تجربی یا سنتز آزمایشگاهی اهمیت بالایی دارد. صرف تطابق مدل با داده‌هایی که برای ساخت آن به کار رفته‌اند (داده‌های آموزش)، معیار کافی برای ارزیابی اعتبار مدل QSAR نیست. در عوض، سنجش توانایی مدل در تعمیم‌پذیری^۱ به داده‌های جدید، که خارج از مجموعه داده‌های آموزشی بوده‌اند، شاخص اصلی موفقیت و کاربردی بودن مدل محسوب می‌شود. این ارزیابی، اعتبار نهایی مدل را برای پیش‌بینی فعالیت زیستی ترکیبات شیمیایی تأیید می‌نماید. برای حصول اطمینان از دقت و سنجش توانایی تعمیم مدل‌های کمی ساختار-فعالیت (QSAR)، استفاده از مجموعه‌ای از شاخص‌های آماری استاندارد امری حیاتی است. این معیارها، هر کدام به طور مجزا، جنبه‌ای خاص از عملکرد و اعتبار مدل را مورد ارزیابی قرار می‌دهند. از طریق تحلیل این شاخص‌های آماری، امکان مقایسه‌ی دقیق مدل‌های مختلف فراهم می‌گردد و در نهایت، پژوهشگر می‌تواند مطمئن‌ترین و کارآمدترین مدل را جهت پیش‌بینی فعالیت زیستی ترکیبات

¹ Generalization

شیمیایی انتخاب نماید.

۱-۴-۶-۱- ضریب همبستگی خطی^۱ (R)

ضریب همبستگی ابزاری برای ارزیابی شدت و جهت رابطه خطی میان متغیرهای مستقل و وابسته در مدل QSAR است. این ضریب که در رابطه (۱-۴) ارائه شده است، از طریق فرمول مشخصی محاسبه می‌گردد، مقادیری بین ۱- تا ۱ را به خود اختصاص می‌دهد. مقدار R نزدیک به ۱ (چه مثبت و چه منفی) حاکی از وجود یک رابطه خطی قوی بین متغیرهاست، در حالی که مقادیر نزدیک به صفر، بیانگر ضعف یا نبود چنین رابطه‌ای است. [۲۰].

$$R = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad \text{رابطه (۱-۳)}$$

۱-۴-۶-۲- ضریب تعیین^۲ (R²)

ضریب تعیین (R²) به عنوان یک شاخص آماری، نسبت واریانس موجود در متغیر وابسته را که مدل QSAR قادر به پیش‌بینی آن است، مشخص می‌کند. این معیار که در رابطه (۱-۵) ارائه شده است، نشان می‌دهد چه کسری از تغییرات کلی در متغیر وابسته، توسط متغیرهای مستقل به‌کار گرفته شده در مدل، قابل تبیین می‌باشد. به طور خلاصه، R² میزان تطابق و تناسب کلی داده‌ها با مدل پیش‌بینی‌کننده را اندازه‌گیری می‌کند [۲۱].

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad \text{رابطه (۱-۴)}$$

۱-۴-۶-۳- مجموع مربعات خطاهای پیش‌بینی (PRESS^۳)

شاخص مجموع مربعات خطای پیش‌بینی (PRESS) که در رابطه (۱-۶) ارائه شده است، از طریق محاسبه مجموع توان دوم تفاضل بین مقادیر واقعی مشاهده شده (y_i) و مقادیر برآورد شده توسط مدل (ŷ_i) به دست

¹ Correlation coefficient

² Coefficient of determination

³ Predictive Residual Sum of Squares (PRESS)

می‌آید و معیاری برای ارزیابی عملکرد مدل QSAR محسوب می‌شود. این شاخص به طور خاص، بر سنجش توانایی مدل در پیش‌بینی مشاهدات جدید و خارج از مجموعه داده آموزشی تمرکز دارد [۲۲].

$$PRESS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \text{رابطه (۱-۵)}$$

۱-۴-۶-۴- خطای استاندارد پیش‌بینی (SEP)^۱

شاخص خطای استاندارد پیش‌بینی (SEP) میانگین انحراف مطلق یا جذر میانگین مربعات خطا بین مقادیر پیش‌بینی شده توسط مدل QSAR و مقادیر واقعی مشاهده شده را طبق رابطه (۷-۱) منعکس می‌کند. این معیار به عنوان یک شاخص مستقیم از دقت کلی پیش‌بینی‌های مدل عمل کرده و مقدار کمتر آن، نشان‌دهنده توانایی مدل در پیش‌بینی مقادیر جدید با صحت بالاتر است [۲۳].

$$SEP = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n}} \quad \text{رابطه (۱-۶)}$$

۱-۴-۶-۵- خطای مطلق میانگین (MAE)^۲

شاخص میانگین قدر مطلق خطا (MAE) با محاسبه میانگین قدر مطلق اختلاف بین مقادیر واقعی (مشاهده شده) و مقادیر پیش‌بینی شده توسط مدل، میزان خطای متوسط مدل در پیش‌بینی‌هایش را طبق رابطه (۸-۱) تعیین می‌کند. این شاخص، برخلاف معیارهای پیچیده‌تر مانند RMSE، به دلیل بیان مستقیم خطا در واحد متغیر وابسته، درک و تفسیر ساده‌تری دارد [۲۴].

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}|}{n} \quad \text{رابطه (۱-۷)}$$

۱-۴-۶-۶- خطای نسبی پیش‌بینی (REP)^۳

شاخص نسبت خطای پیش‌بینی (REP) با سنجش نسبت خطای استاندارد پیش‌بینی (SEP) به میانگین مقادیر واقعی (که معمولاً میانگین متغیر وابسته در مجموعه داده‌های آموزشی یا اعتبارسنجی است) مطابق رابطه (۹-۱) محاسبه می‌شود. این شاخص، دقت پیش‌بینی‌های مدل را در قالب یک مقدار درصدی ارائه

^۱ Standard Error of Prediction (SEP)

^۲ Mean Absolute Error (MAE)

^۳ Relative Error of Prediction (REP)

می‌دهد و قابلیت مقایسه‌ی عملکرد مدل‌های مختلف را، حتی در مجموعه داده‌هایی با مقیاس‌های متفاوت، تسهیل می‌نماید [۲۵].

$$REP (\%) = \frac{100}{\bar{y}} \times \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n}} = \frac{SEP}{\bar{y}} \times 100 \quad \text{رابطه (۸-۱)}$$

۱-۴-۶-۷- میانگین مربع خطا (MSE)^۱

شاخص میانگین مربع خطا (MSE) معیاری برای سنجش میزان دقت مدل از طریق محاسبه میانگین مربعات اختلاف بین مقادیر پیش‌بینی شده توسط مدل و مقادیر واقعی (مشاهده شده) است. این شاخص که در رابطه (۱۰-۱) ارائه شده است، علاوه بر نمایان ساختن اندازه کلی خطاها، به خطاهای بزرگتر وزن بیشتری اختصاص می‌دهد و بدین ترتیب، پراکندگی و شدت خطاهای مدل را در پیش‌بینی‌های خود به خوبی منعکس می‌سازد [۲۶].

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n} \quad \text{رابطه (۹-۱)}$$

۱-۴-۶-۸- میانگین خطای نسبی (MRE)^۲

شاخص میانگین خطای نسبی (MRE) به منظور کمی‌سازی انحراف پیش‌بینی‌های مدل از مقادیر واقعی به کار گرفته می‌شود. این شاخص که در رابطه (۱۱-۱) ارائه شده است، خطاهای پیش‌بینی را به صورت درصد بیان می‌کند و میانگین این خطاها را نشان می‌دهد. به عبارت دقیق‌تر، MRE درصد متوسطی از مقادیر واقعی را که پیش‌بینی‌های مدل از آن منحرف شده‌اند، نمایان می‌سازد [۲۷].

$$MRE = \frac{\sum_{i=1}^n \left| \frac{y_i - \hat{y}}{y_i} \right|}{n} \times 100 \quad \text{رابطه (۱۰-۱)}$$

۱-۴-۶-۹- ارزیابی مدل با استفاده از مجموعه آزمون

برای حصول اطمینان از سنجش دقیق و قابل اعتماد عملکرد مدل QSAR، بخشی از داده‌ها که در

^۱ Mean Squared Error (MSE)

^۲ Mean Relative Error (MRE)

مرحله آموزش مدل به کار گرفته نشده‌اند، به عنوان مجموعه آزمون مجزا در نظر گرفته می‌شوند. انجام فرآیند پیش‌بینی مقادیر برای این مجموعه داده مستقل، امکان ارزیابی قابل اتکایی از توانایی تعمیم‌پذیری مدل و دقت واقعی آن را در برخورد با ترکیبات شیمیایی جدید و ناشناخته فراهم می‌آورد [۲۸ و ۱۵].

۱-۴-۶-۱۰- اعتبارسنجی متقابل با روش حذف تک‌تایی (LOO-CV)^۱

در چهارچوب روش اعتبارسنجی متقابل حذف یک به یک (LOO-CV)، فرآیند ارزیابی مدل به صورت تکراری انجام می‌شود. در هر گام از این تکرارها، تنها یک داده منفرد از مجموعه آموزشی برداشته شده و کنار گذاشته می‌شود. سپس، مدل با اتکا به تمامی داده‌های موجود (به استثنای آن یک نمونه حذف شده) مجدداً آموزش داده می‌شود. پس از تکمیل فرآیند آموزش، مدل برای پیش‌بینی مقدار متغیر وابسته مربوط به همان نمونه‌ای که در آن مرحله کنار گذاشته شده بود، به کار گرفته می‌شود. جهت سنجش توانایی تعمیم‌پذیری و قدرت پیش‌بینی مدل در این روش، از شاخص Q_{LOO}^2 استفاده می‌گردد. این شاخص که در رابطه (۱-۱۲) آورده شده است، معیاری عملکردی را بر روی داده‌هایی ارائه می‌دهد که به طور صریح در فرآیند آموزش مدل دخالتی نداشته‌اند. [۲۹].

$$Q_{LOO}^2 = 1 - \frac{\sum (\hat{y}_i - y_i)^2}{\sum (y_i - \bar{y})^2} \quad \text{رابطه (۱-۱۱)}$$

۱-۴-۶-۱۱- تحلیل نمودار خطای باقیمانده^۲

خطاهای باقیمانده، که معرف اختلاف بین مقادیر پیش‌بینی شده توسط مدل و مقادیر واقعی مشاهده شده هستند، اطلاعات ارزشمندی را در خصوص کیفیت و کارایی مدل ارائه می‌دهند. در نمودار پراکندگی این خطاها، توزیع متقارن و تصادفی نقاط حول محور افقی، بیانگر این است که خطاهای مدل عمدتاً ماهیت تصادفی داشته و هیچ‌گونه سوگیری سیستماتیکی در پیش‌بینی‌های آن وجود ندارد. مشاهده چنین الگوی پراکندگی ایده‌آل، نشان‌دهنده مناسب بودن مدل آماری برای توصیف داده‌های مورد مطالعه و روابط حاکم بر آن‌ها است. [۲۹].

^۱ Leave-One-Out Cross-Validation (LOO-CV)

^۲ Residual plot

۱-۴-۶-۱۲-آزمون Y-تصادفی^۱

به منظور سنجش اعتبار همبستگی‌های استخراج شده در مدل QSAR و اطمینان از اینکه روابط مشاهده شده صرفاً تصادفی نیستند، از آزمون Y-تصادفی بهره گرفته می‌شود. در این رویکرد، مقادیر مشاهده شده برای متغیر وابسته (Y) در مجموعه داده، به صورت تصادفی مورد بازآرایی قرار می‌گیرند. سپس، ضریب تعیین (R^2) با استفاده از این مجموعه داده‌ی «به‌ریخته» محاسبه می‌شود. چنانچه مقدار R^2 حاصل از مدل اصلی، به طور محسوسی بزرگتر از R^2 به دست آمده از داده‌های جابجا شده باشد، این تفاوت معنادار، اعتبار مدل اصلی را تصدیق می‌کند. این نتیجه نشان می‌دهد که همبستگی مشاهده شده، فراتر از یک تصادف آماری بوده و مدل، قابلیت قابل توجهی در توصیف رابطه ساختاری-فعالیتی بین مولکول‌ها و ویژگی‌های آن‌ها دارد. [۳۰].

۱-۵- شبیه‌سازی اتصال مولکولی^۲

شبیه‌سازی اتصال مولکولی (داکینگ مولکولی)، روشی محاسباتی در قلمرو شیمی محاسباتی و بیوفیزیک است که به پیش‌بینی نحوه برهم‌کنش و اتصال یک مولکول کوچک (شناخته شده به عنوان لیگاند) با یک مولکول بزرگ (مانند پروتئین) می‌پردازد. این فرآیند با کاوش حالت‌های احتمالی قرارگیری لیگاند در ناحیه فعال پروتئین و محاسبه انرژی حاصل از این برهم‌کنش‌ها، پایدارترین و محتمل‌ترین سناریوی اتصال را تعیین می‌کند. در سطح مولکولی، داکینگ امکان شناسایی انواع نیروهای دخیل در اتصال، از جمله پیوندهای هیدروژنی، نیروهای واندروالس، برهم‌کنش‌های الکترواستاتیکی و برهم‌کنش‌های π - π stacking را فراهم می‌آورد.

در این پژوهش، شبیه‌سازی اتصال مولکولی به عنوان یک رویکرد مکمل در کنار مدل QSAR به کار گرفته شده است تا اعتبار فیزیکی و شیمیایی پیش‌بینی‌های حاصل از مدل آماری را تقویت کند. این روش به عنوان پلی عمل می‌کند که ارتباط بین ساختار مولکولی و فعالیت زیستی استنباط شده از مدل QSAR را

¹ Y-Randomization Test

² Molecular Docking

به سازوکار مولکولی واقعی اتصال، پیوند می‌دهد. داکینگ مولکولی، ارزیابی توانایی اتصال ترکیبات طراحی شده به پروتئین هدف را تسهیل کرده و امکان اولویت‌بندی آن‌ها را بر اساس انرژی اتصال و ماهیت برهم‌کنش‌ها فراهم می‌سازد. علاوه بر این، نتایج حاصل از داکینگ برای درک اهمیت توصیف‌گرهای کلیدی که از طریق مدل ALASSO انتخاب شده‌اند، ارزشمند است. به طور کلی، در این پایان‌نامه، شبیه‌سازی اتصال مولکولی به عنوان ابزاری تحلیلی جهت تأیید و تکمیل یافته‌های مدل‌سازی QSAR و همچنین راهنمایی منطقی برای طراحی مولکول‌های جدید به کار گرفته شده است.

۱-۵-۱- پیش‌پردازش ساختار پروتئین برای داکینگ

مرحله پیش‌پردازش ساختار پروتئین، نقشی اساسی در افزایش دقت و قابلیت اطمینان پیش‌بینی‌های مربوط به برهم‌کنش لیگاند-پروتئین در فرآیند داکینگ مولکولی ایفا می‌کند. در این مرحله، ساختار سه بعدی پروتئین که از پایگاه‌های داده ساختاری معتبر استخراج شده است، مورد بازبینی قرار می‌گیرد. ابتدا، مولکول‌های آب غیرضروری، یون‌ها و سایر مولکول‌های کوچک که در جایگاه اتصال دخالتی ندارند، از ساختار حذف می‌شوند. در ادامه، اتم‌های هیدروژن به مدل اضافه شده و وضعیت یونیزاسیون (پروتونه شدن یا نشدن) باقی‌مانده‌های اسیدی و بازی در ساختار پروتئین به دقت تنظیم می‌گردد. پس از آن، فرآیند بهینه‌سازی هندسی موضعی^۱ اجرا می‌شود تا هرگونه اعوجاج ساختاری و نقایص در پیوندها اصلاح گردد. در نهایت، با شناسایی و تعیین دقیق ناحیه فعال^۲ پروتئین، بستری مناسب و قابل اعتماد برای اجرای فرآیند داکینگ و تفسیر صحیح نتایج آن فراهم می‌شود.

۱-۵-۲- پیش‌پردازش لیگاندها و مولکول‌های کاندید

مرحله پیش‌پردازش لیگاندها و مولکول‌های کاندید، گامی ضروری در فرآیند داکینگ مولکولی به شمار می‌رود و با هدف ایجاد بازنمایی دقیق و واقع‌گرایانه از رفتار شیمیایی این مولکول‌ها در محیط بیولوژیکی طراحی شده است. در ابتدای این فرآیند، ساختار مولکولی هر لیگاند یا ترکیب پیشنهادی ترسیم شده و

¹ Local Geometry Optimization

² Active Site

سپس به یک مدل سه بعدی تبدیل می شود. متعاقب آن، بهینه سازی هندسی^۱ صورت می گیرد تا پایدارترین صورتبندی^۲ برای مولکول حاصل شود.

در ادامه، پارامترهای کوانتومی-شیمیایی مانند حالت پروتونه شدن مناسب، توزیع بارهای جزئی^۳ بر روی اتمها، و نحوه آرایش فضایی پیوندهای قابل چرخش^۴ تعیین می گردند. حذف یا اصلاح کانفورماسیونهای ناپایدار یا غیرطبیعی، دقت نمونه برداری^۵ در مراحل بعدی داکینگ را به طور قابل توجهی بهبود می بخشد. در نهایت، ساختارهای پیش پردازش شده به فرمت استاندارد ذخیره می شوند تا مقایسه دقیق و منصفانه انرژی های اتصال^۶ و الگوهای برهم کنش مولکول ها با پروتئین هدف امکان پذیر گردد [۳۱].

۱-۶- سرطان

رشد سرطان مکانیسم پیچیده ای است که شامل مسیرهای پیام رسان متعدد، تغییرات اسکلت سلولی و تغییرات سلولی است که به سلول های تومور اجازه تکثیر، تهاجم و انتشار متاستاتیک را می دهد [۳۲]. یکی از تنظیم کننده های حیاتی این فرآیندها شامل LIMK، نوعی از سرین/ترئونین کینازها است که کوفلین، که یک پروتئین جداکننده اکتین است، را فسفریله می کند. کینازهای LIM با داشتن یک سازماندهی ساختاری از دامنه ها، نقش مهم در برهم کنش های پروتئین-پروتئین و فعالیت کیناز دارند. LIMK ها از طریق چندین مسیر پیام رسان که باعث توسعه تومور و متاستاز می شوند، نقش دارند. فعالیت غیرطبیعی پروتئین کینازها در بسیاری از بیماری ها از جمله سرطان ها مشاهده می شود و به همین دلیل، این آنزیم ها به عنوان هدف^۷ مهم برای طراحی دارو در نظر گرفته می شوند. مهارکننده های پروتئین کیناز می توانند از فعالیت بیش از حد یا نابجای این آنزیم ها جلوگیری کنند و در نتیجه به عنوان داروهای ضدسرطان، ضدالتهاب یا تعدیل کننده سیستم عصبی به کار روند. ثابت بازدارندگی^۸ K_i یک ثابت ترمودینامیکی است که بیانگر قدرت اتصال یک

¹ Geometry Optimization

² Conformation

³ Partial Charges

⁴ Rotatable Bonds

⁵ Sampling

⁶ Binding Energies

⁷ Targets

⁸ Inhibition

بازدارنده به آنزیم است. هر چه مقدار Ki کوچک‌تر باشد، بازدارنده با قدرت بیشتری به جایگاه فعال آنزیم متصل شده و در نتیجه مهارکنندگی قوی‌تری دارد [۳۴ و ۳۳]. دانستن Ki نه تنها در انتخاب مهارکننده قوی اهمیت دارد، بلکه در بهینه‌سازی ساختارهای شیمیایی و طراحی ترکیبات جدید نیز نقش کلیدی ایفا می‌کند. از آنجایی که تعیین Ki به روش‌های آزمایشگاهی پرهزینه و زمان‌بر است، توسعه مدل‌های محاسباتی برای پیش‌بینی Ki می‌تواند بسیار سودمند باشد. بنابراین توسعه مدل‌های پیشگو بر اساس رابطه کمی ساختار-فعالیت (QSAR) برای پیش‌بینی Ki مشتقات تتراهیدروپیریدین می‌تواند جایگزینی دقیق و کارآمد برای آزمون‌های پرهزینه اولیه باشد. بنابراین استفاده از روش‌های رگرسیون با کارایی بالا مانند روش‌های انقباضی (جریمه ای) لاسوی تطبیقی در این زمینه می‌تواند منجر به ایجاد مدل‌های بهتر با قدرت پیشگویی بالاتر گردد. ایجاد مدل‌های QSAR امکان پیشنهاد ترکیبات جدید با قدرت مهارکنندگی LIMK را فراهم می‌کند که درستی این ترکیبات پیشنهادی با داکینگ مولکولی صحت سنجی خواهد شد. در نهایت داکینگ مولکولی به‌عنوان یک ابزار مکمل قادر است نحوه جای‌گیری و نوع برهم‌کنش‌های غیرکووالانسی (هیدروژنی، واندروالس، هیدروفوبیک) بین مشتقات منتخب و جایگاه فعال LIMK را شبیه‌سازی کند. ترکیب نتایج QSAR و داکینگ نه تنها دقت انتخاب بازدارنده‌های بالقوه را افزایش می‌دهد، بلکه مبنای علمی برای طراحی منطقی دارو^۱ و بهینه‌سازی ساختار مشتقات جدید فراهم می‌آورد. بنابراین ترکیب مطالعات QSAR و داکینگ مولکولی امکان یافتن ترکیبات جدید با فعالیت مناسب مهارکنندگی LIMK که امکان سنتز را دارند فراهم می‌کند.

۷-۱- ضرورت تحقیق

پروتئین کینازها به دلیل نقش محوری در مسیرهای سیگنال‌دهی سلولی و ارتباط مستقیم با بروز سرطان‌ها، بیماری‌های التهابی و اختلالات عصبی، از مهم‌ترین اهداف دارویی در تحقیقات نوین محسوب می‌شوند. مهار اختصاصی این آنزیم‌ها می‌تواند مسیرهای غیرطبیعی سلولی را متوقف کرده و در درمان بیماری‌های صعب‌العلاج نقش تعیین‌کننده داشته باشد. در این میان، پروتئین کیناز (LIMK) به دلیل

¹ Rational Drug Design

دخالت در تنظیم پویایی اسکلت سلولی و نقش اثبات شده در متاستاز سرطان، به عنوان یک هدف دارویی ویژه مورد توجه قرار گرفته است. یافتن مهارکننده های مؤثر و اختصاصی برای LIMK می تواند به توسعه درمان های ضدسرطان جدید منجر شود. اخیراً مشتقات تتراهیدروپیریدین به عنوان مهارکننده های احتمالی پروتئین کیناز LIMK گزارش شده اند. از طرف دیگر، تعیین تجربی ثابت بازدارندگی (K_i) مشتقات تتراهیدروپیریدین فرآیندی زمان بر، پرهزینه و نیازمند زیرساخت های آزمایشگاهی پیشرفته است. به همین دلیل، استفاده از روش های محاسباتی مانند مدل سازی QSAR و داکینگ مولکولی ضرورتی اجتناب ناپذیر دارد. این روش ها با پیش بینی صحیح K_i و بررسی نحوه برهم کنش سه بعدی بازدارنده ها با جایگاه فعال آنزیم، می توانند غربالگری سریع و هدفمند ترکیبات را امکان پذیر کنند. به طور خاص، مشتقات تتراهیدروپیریدین به دلیل قابلیت برقراری برهم کنش های متنوع و ساختار شیمیایی انعطاف پذیر، پتانسیل بالایی برای توسعه مهارکننده های LIMK دارند. به کارگیری روش های آماری نوین مانند ALASSO در مدل سازی QSAR، نوآوری این تحقیق را دوچندان کرده و امکان شناسایی دقیق متغیرهای کلیدی و ترکیبات امیدبخش را فراهم می آورد. در این پژوهش، به دلیل نبود مطالعات پیشین در زمینه مهارکننده های مشتقات تتراهیدروپیریدین برای پروتئین کیناز LIM، رویکردی نوآورانه مبتنی بر مدل سازی QSAR اتخاذ شده است. یکی از چالش های اساسی QSAR، وجود تعداد زیاد توصیف کننده های مولکولی در برابر ترکیبات محدود است که خطر بیش برآزش و کاهش قابلیت تعمیم مدل را به همراه دارد. برای رفع این مشکل، استفاده از رگرسیون جریمه ای به ویژه ALASSO ضروری است. ALASSO انتخاب هوشمندانه متغیرهای کلیدی، علاوه بر افزایش صحت پیش بینی ثابت بازدارندگی (K_i)، امکان تفسیر زیست شیمیایی نتایج را نیز فراهم می سازد. بنابراین، به کارگیری این روش موجب توسعه مدلی صحیح، پایدار و کاربردی برای غربالگری و طراحی منطقی مهارکننده های جدید پروتئین کیناز LIM خواهد شد. یکی از محدودیت های اصلی مدل های آماری و QSAR این است که صرفاً بر اساس داده های عددی و روابط ریاضی فعالیت زیستی ترکیبات را پیش بینی می کنند، اما اطلاعات مستقیم در مورد نحوه جای گیری بازدارنده در جایگاه فعال پروتئین و نوع برهم کنش های مولکولی به دست نمی دهند. در حالی که در طراحی منطقی دارو، درک دقیق مکانیسم اتصال

و تعامل‌های سه‌بعدی بین لیگاند و پروتئین هدف اهمیت اساسی دارد. در این زمینه، داکینگ مولکولی به‌عنوان یک ابزار محاسباتی قدرتمند، امکان شبیه‌سازی اتصال مشتقات شیمیایی به پروتئین هدف را فراهم می‌کند. ضرورت به‌کارگیری داکینگ مولکولی در این تحقیق از آن جهت است که نتایج مدل QSAR و پیش‌بینی Ki را تکمیل می‌کند. ترکیب این دو رویکرد QSAR و داکینگ مولکولی موجب خواهد شد که انتخاب ترکیبات امیدبخش دقیق‌تر و هدفمندتر انجام گیرد، درک بهتری از بازدارندگی LIMK به دست آید و نهایتاً مسیر برای طراحی منطقی و بهینه‌سازی مشتقات جدید تتراهیدرو پیریدین هموار شود. در مجموع، ضرورت این تحقیق در سه محور اصلی قابل توجیه است:

۱. **اهمیت زیستی و بالینی LIMK:** یک هدف دارویی مهم در سرطان و بیماری‌های وابسته است.

۲. **صرفه‌جویی علمی و اقتصادی:** روش‌های محاسباتی هزینه و زمان کشف دارو را به‌طور چشمگیری کاهش می‌دهند.

۳. **نوآوری علمی:** ترکیب QSAR مبتنی بر ALASSO با داکینگ مولکولی، رویکردی جدید و کارآمد برای طراحی منطقی دارو ارائه می‌دهد.

۸-۱- پیشینه پژوهش

به‌منظور بررسی جایگاه روش لاسوی تطبیقی در مطالعات مدل‌سازی QSAR/QSPR^۱، مقالات منتشرشده طی ده سال اخیر مورد بررسی قرار گرفت. نتایج این بررسی نشان داد که اگرچه روش ALASSO و نسخه‌های توسعه‌یافته آن در سال‌های اخیر به‌عنوان یکی از روش‌های مؤثر انتخاب متغیر در داده‌های با ابعاد بالا مورد توجه قرار گرفته‌اند، اما کاربرد آن در مطالعات QSAR همچنان محدود بوده و بیشتر پژوهش‌ها بر توسعه الگوریتم یا ارزیابی عملکرد آن در مجموعه‌داده‌های خاص متمرکز بوده‌اند.

آناند^۲ و همکارانش در سال ۲۰۱۵ تحقیقاتی در مورد مشتقات تتراهیدروپیریدین و بررسی رابطه کمی

^۱ Quantitative Structure-Property Relationship (QSPR)

^۲ Anand

ساختار-فعالیت سه بعدی (3D-QSAR) بعنوان مهار کننده های VEGFR-2 kinase انجام دادند. در این تحقیق مطالعه رابطه 3D-QSAR بر روی مجموعه‌ای از مشتقات تتراهیدرو- H^3 -ایمیدازو[۴،۵]پیریدین انجام شده است تا اساس ساختاری فعالیت مهاری VEGFR-2 درک شود. روش‌های تحلیل میدان مولکولی مقایسه‌ای (CoMFA) و تحلیل شاخص‌های تشابه مولکولی مقایسه‌ای (CoMSIA) برای استخراج مدل‌های QSAR به کار گرفته شده‌اند. نتایج نشان می‌دهد مدل‌های CoMFA عملکرد بهتری نسبت به مدل‌های CoMSIA داشتند. استحکام مدل با اعتبارسنجی متقابل leave-five-out بررسی شد. مشخص شد که این مدل از نظر آماری قوی است و انتظار می‌رود در طراحی ترکیبات جدید با فعالیت مهاری VEGFR-2 افزایش یافته کمک کند [۳۵].

در سال ۲۰۱۵، الگمال^۱ و همکاران با معرفی روش لاسوی تطبیقی اصلاح شده (Adjusted Adaptive LASSO)، یک مدل QSAR با ابعاد بالا برای پیش‌بینی فعالیت ضدسرطانی مشتقات ایمیدازو[۴،۵]-bپیریدین^۲ ارائه کردند. در این پژوهش، نسخه اصلاح شده ALASSO به منظور انتخاب بهینه توصیف‌کننده‌های مولکولی، کاهش اثر هم‌خطی و افزایش قدرت پیش‌بینی توسعه داده شد. نتایج نشان داد که روش پیشنهادی در مقایسه با روش‌های لاسو LASSO^۳ و ALASSO متداول، توصیف‌کننده‌های مؤثرتری را انتخاب کرده، دقت پیش‌بینی بالاتری ارائه داده و مدل حاصل از پایداری و تفسیرپذیری بیشتری برخوردار است. این مطالعه نشان داد که استفاده از نسخه‌های توسعه یافته ALASSO می‌تواند کارایی مدل‌های QSAR با ابعاد بالا را به‌طور محسوسی افزایش دهد [۳۶].

در ادامه، الگمال و لی^۴ در سال ۲۰۱۷ نسخه جدیدی از ALASSO برای توسعه یک مدل رابطه کمی ساختار-فعالیت طبقه‌بندی شده به منظور پیش‌بینی فعالیت ضد ویروس هپاتیت C مشتقات تیاوره^۵ معرفی کردند. هدف اصلی این پژوهش، انتخاب دقیق توصیف‌کننده‌های مؤثر و افزایش دقت طبقه‌بندی در

¹ Algamil

² Imidazo[4,5-b]pyridine

³ Least Absolute Shrinkage and Selection Operator

⁴ Algamil & Lee

⁵ Thiourea

مجموعه داده‌های با ابعاد بالا بود. نتایج نشان داد که روش پیشنهادی نسبت به روش‌های متداولی نظیر لاسو، لاسوی تطبیقی، Elastic Net، اسکد (SCAD)^۱ و رگرسیون بریج^۲ عملکرد بهتری داشته و علاوه بر افزایش دقت طبقه‌بندی، از نظر پایداری مدل و دامنه کاربرد نیز نتایج مطلوب‌تری ارائه کرده است [۳۷].

از سوی دیگر، داتا^۳ و همکاران در سال‌های ۲۰۱۷ و ۲۰۱۹ با توسعه الگوریتم لاسوی تطبیقی مبتنی بر همبستگی (CorrLASSO)^۴، از این روش در مدل‌سازی رابطه کمی ساختار-ویژگی (QSPR) برای پیش‌بینی تمایل اتصال دارو به DNA^۵ مشتقات ۹-آنیلینوآکریدین^۶ به عنوان عوامل ضدتومور استفاده کردند. در این پژوهش، CorrLASSO در کنار الگوریتم ژنتیک (GA)^۷ به کار گرفته شد تا انتخاب توصیف‌کننده‌ها و ساخت مدل به صورت هم‌زمان انجام شود. نتایج نشان داد که CorrLASSO نسبت به مدل‌های مبتنی بر لاسو کلاسیک و رگرسیون خطی چندگانه مبتنی بر الگوریتم ژنتیک (GA-MLR)^۸، مدل‌هایی با دقت پیش‌بینی بیشتر، تعداد توصیف‌کننده‌های کمتر و قابلیت تفسیر بالاتر تولید می‌کند و برای داده‌های شیمیایی با ابعاد بالا گزینه‌ای کارآمد محسوب می‌شود [۳۸، ۳۹].

در مجموع، بررسی مقالات ده سال اخیر نشان می‌دهد که ALASSO و نسخه‌های توسعه‌یافته آن عمدتاً در مطالعات QSAR و QSPR مرتبط با ترکیبات ضدسرطان، ضدویروس و عوامل ضدتومور، با هدف انتخاب توصیف‌کننده‌های مولکولی، کاهش ابعاد داده‌ها، کنترل هم‌خطی و افزایش قدرت پیش‌بینی مدل‌ها مورد استفاده قرار گرفته‌اند. با این حال، بر اساس منابع بررسی‌شده در این پژوهش، مطالعه‌ای که به طور هم‌زمان از روش لاسوی تطبیقی (ALASSO) برای انتخاب توصیفگر و مدل‌سازی QSAR با هدف پیش‌بینی ثابت بازدارندگی (Ki) مشتقات تتراهیدروپیریدین به عنوان مهارکننده‌های پروتئین کیناز LIMK استفاده کرده باشد، یافت نشد. از این رو، پژوهش حاضر از نظر نوع ترکیبات مورد مطالعه، هدف زیستی، متغیر پاسخ و

¹ Smoothly Clipped Absolute Deviation (SCAD)

² Bridge Regression

³ Datta

⁴ Correlation-based Adaptive LASSO

⁵ DNA Drug Binding Affinity

⁶ 9-anilinoacridine

⁷ Genetic Algorithm

⁸ Genetic Algorithm–Multiple Linear Regression

به‌کارگیری روش ALASSO در انتخاب توصیف‌کننده‌های مولکولی، دارای جنبه نوآورانه بوده و می‌تواند زمینه‌ساز توسعه مدل‌های صحیح‌تر در طراحی منطقی مهارکننده‌های پروتئین‌کیناز LIMK و کشف داروهای جدید باشد.

فصل دوم: بخش تجربی

۱-۲- نرم افزار های به کار رفته در این پژوهش

در این بخش، خلاصه‌ای از نرم‌افزارهای مورد استفاده در این تحقیق آورده شده است

۲-۲- نرم افزار هایپرکم^۱

نرم‌افزار هایپرکم به‌عنوان یکی از ابزارهای معتبر در شیمی محاسباتی، قابلیت ایجاد، ترسیم و بازسازی سه‌بعدی ساختارهای مولکولی را فراهم می‌سازد. این برنامه با تکیه بر الگوریتم‌های محاسبات کوانتومی و رویکردهای نیمه‌تجربی، هندسه مولکول‌ها را به‌گونه‌ای بهینه می‌کند که ساختار نهایی در کمینه انرژی پتانسیل قرار گیرد؛ وضعیتی که نشان‌دهنده پایدارترین حالت ترمودینامیکی است. افزون بر این، هایپرکم امکان استخراج دقیق پارامترهای هندسی از جمله طول پیوندها، زاویه‌های پیوندی و زوایای دی‌هدرال را فراهم می‌کند؛ داده‌هایی که در تحلیل عمیق ویژگی‌های ساختاری و بررسی روابط ساختار-فعالیت (SAR) نقش محوری دارند [۴۰]. در این پژوهش از نسخه ۰٫۸ نرم‌افزار هایپرکم استفاده گردید و هندسه‌های بهینه‌شده حاصل از آن، به‌عنوان داده‌های ورودی ضروری برای نرم‌افزار دراگون در محاسبه انواع توصیف‌گرهای مولکولی مورد بهره‌برداری قرار گرفتند. به این ترتیب، ساختارهای مینیمم‌سازی‌شده فراهم‌شده توسط هایپرکم، مبنای تولید پارامترهای کمی مورد نیاز برای توسعه و تحلیل مدل‌های کیوسار را تشکیل دادند [۴۱].

¹ Hyperchem

۲-۳- نرم افزار دراگون^۱

نرم افزار دراگون که توسط گروه کمومتریक्स دانشگاه میلان توسعه یافته است، یکی از ابزارهای معتبر و گسترده در حوزه شیمی اطلاعات به شمار می رود. کارکرد اصلی این نرم افزار، محاسبه و استخراج مجموعه ای وسیع از توصیف گرهای مولکولی بر مبنای ساختارهای شیمیایی ورودی است. توصیف گرهای مولکولی در واقع مقادیر عددی اند که ویژگی های ساختاری، توپولوژیکی، هندسی و الکترونیکی یک مولکول را به طور کمی بیان کرده و نقش بنیادین در توسعه مدل های کیوسار ایفا می کنند. در این پژوهش از نسخه ۵/۵ این برنامه استفاده شد؛ نسخه ای که امکان محاسبه بیش از سه هزار توصیف گر را فراهم کرده و آن ها را در قالب ۴۴ دسته بندی مستقل سازمان دهی می کند [۴۲].

۲-۴- نرم افزار SPSS

نرم افزار IBM SPSS Statistics که نخستین نسخه های آن در دهه ۱۹۷۰ توسط پژوهشگران دانشگاه استنفورد طراحی شد و از سال ۲۰۰۹ تحت مالکیت و پشتیبانی شرکت IBM قرار گرفت، یکی از کاربردی ترین سامانه ها در حوزه تحلیل داده های کمی به شمار می رود. این نرم افزار با فراهم کردن مجموعه گسترده ای از ابزارهای آماری، امکان مدیریت، پردازش و تفسیر داده ها را در مطالعات علمی فراهم می سازد. SPSS قابلیت اجرای انواع روش های رگرسیونی، از جمله رگرسیون های ساده و چندمتغیره، انجام آزمون های آماری متنوع، اعمال محاسبات ریاضی پایه (نظیر لگاریتم گیری و محاسبه مقدار مطلق)، و تولید نمودارها و جداول توصیفی را داراست. ترکیب سهولت کاربری با تنوع بالای روش های تحلیلی موجب شده است که این نرم افزار در رشته های گوناگون، به ویژه علوم پزشکی و حوزه های تجربی، به طور گسترده مورد استفاده قرار گیرد.

در این پژوهش، به منظور غربالگری اولیه توصیف گرهای مولکولی، از روش انتخاب متغیر گام به گام در محیط SPSS استفاده شد. متغیرهای انتخاب شده در این مرحله، به عنوان مجموعه نهایی ورودی، در مدل

¹ Dragon

رگرسیون خطی چندگانه (MLR) به کار گرفته شدند تا ارتباط میان توصیف‌گرهای مولکولی و میزان فعالیت زیستی ترکیبات به صورت کمی مدل‌سازی و تحلیل شود [۴۳].

۲-۵- نرم افزار R/Rstudio

در این پژوهش، تمامی تحلیل‌های آماری، پردازش داده‌ها و مراحل مرتبط با آماده‌سازی و پالایش توصیف‌گرهای مولکولی در محیط برنامه‌نویسی R و از طریق نرم‌افزار RStudio انجام می‌گیرد. زبان R به عنوان یکی از ابزارهای قدرتمند در تحلیل‌های عددی و آماری، امکان پیاده‌سازی الگوریتم‌های پیچیده، بررسی داده‌های با ابعاد بالا و توسعه مدل‌های پیش‌بینی‌کننده را فراهم می‌سازد. افزون بر این، RStudio به عنوان یک محیط توسعه یکپارچه، امکانات مدیریت داده، اجرای منظم کدها و سازمان‌دهی فرایندهای تحلیلی را ارائه می‌کند.

با توجه به انعطاف‌پذیری گسترده R، وجود بسته‌های تخصصی در حوزه شیمی محاسباتی و یادگیری ماشین، و توانایی آن در رسیدگی به داده‌های پیچیده، این محیط به عنوان بستر اصلی برای تحلیل‌های ریاضی و آماری در این تحقیق انتخاب شده است. در بخش پیش‌پردازش داده‌ها - که شرح دقیق آن در فصل سوم ارائه خواهد شد. از قابلیت‌های این نرم‌افزار برای پالایش و آماده‌سازی مجموعه توصیف‌گرهای مولکولی استفاده شده است.

در گام نخست، جهت جداسازی اصولی داده‌ها به سه زیرمجموعه آموزش، اعتبارسنجی و آزمون، از الگوریتم کنارد-استون بهره گرفته می‌شود. این روش با انتخاب نمونه‌ها بر اساس پراکندگی آن‌ها در فضای توصیف‌گرها، تضمین می‌کند که هر زیرمجموعه بازتاب مناسبی از ساختار کلی داده‌ها داشته باشد و از شکل‌گیری هرگونه سوگیری در فرآیند مدل‌سازی جلوگیری شود. انجام چنین تفکیکی نقش کلیدی در افزایش توان تعمیم‌پذیری مدل‌های QSAR دارد.

به منظور کاهش ابعاد داده‌ها و حذف متغیرهای فاقد ارزش اطلاعاتی، از قابلیت‌های موجود در بسته نرم‌افزاری caret استفاده شده است. در این مرحله، توصیف‌گرهایی که تنوع بسیار کمی میان نمونه‌ها از خود نشان می‌دهند، شناسایی و کنار گذاشته می‌شوند؛ زیرا چنین متغیرهایی معمولاً توانایی تفکیک‌کنندگی ندارند و سهم قابل‌توجهی در توضیح رفتار یا فعالیت زیستی مولکول‌ها ایفا نمی‌کنند. حضور این نوع توصیف‌گرها می‌تواند موجب افزایش نویز، پیچیدگی غیرضروری و در نهایت کاهش کارایی مدل‌های پیش‌بینی شود. حذف آن‌ها کمک می‌کند مجموعه‌ای کارآمدتر و مناسب‌تر از ویژگی‌ها برای مراحل بعدی مدل‌سازی فراهم گردد.

به طور کلی، بهره‌گیری از محیط R و RStudio در فرآیند پیش‌پردازش داده‌ها نقش بنیادینی در مهیا کردن ورودی‌های بهینه، کاهش سربار محاسباتی، بهبود کیفیت متغیرهای مورد استفاده و افزایش دقت و پایداری مدل‌های QSAR دارد. این مرحله، به‌عنوان نخستین گام در فرایند مدل‌سازی، شرایط لازم را برای انجام روش‌های انتخاب ویژگی، توسعه مدل‌های آماری و غیرخطی، و اجرای تحلیل‌های پیش‌بینی در ادامه پژوهش فراهم می‌آورد [۴۴].

۲-۶- نرم افزار MATLAB

در پژوهش حاضر، محیط محاسباتی متلب (MATLAB)، توسعه‌یافته توسط کمپانی MathWorks، به عنوان بستر اصلی اجرای آنالیزهای کمومتریکس و پیاده‌سازی الگوریتم‌های ریاضی انتخاب شده است. این پلتفرم به واسطه معماری ماتریس‌محور خود، ابزاری بی‌بدیل در شیمی محاسباتی و به‌ویژه مطالعات کمی ارتباط ساختار-فعالیت (QSAR) به شمار می‌رود. قابلیت‌های برجسته این نرم‌افزار در مدیریت ماتریس‌های داده‌ای حجیم، توسعه مدل‌های آماری و انجام محاسبات پیچیده خطی و غیرخطی، آن را به استاندارد پذیرفته‌شده جهت استخراج الگوهای معنادار از فضا‌های شیمیایی چندبعدی مبدل ساخته است.

در مسیر انجام این تحقیق، از نسخه R2014b نرم‌افزار مذکور جهت اجرای فاز حیاتی پیش‌پردازش ماتریس توصیف‌گرهای فیزیکوشیمیایی و ساختاری، و همچنین تدوین ساختار مدل‌های پیش‌بین بهره گرفته

شد (رویه‌های اجرایی و الگوریتم‌های مربوطه در فصل سوم این رساله به تفصیل تبیین خواهند شد). از منظر متدولوژی، گام پیش‌پردازش در این پلتفرم با هدف پالایش فضای متغیرها و ارتقای کیفیت داده‌های ورودی انجام می‌پذیرد. در این راستا، ماتریس همبستگی توصیف‌گرها به دقت آنالیز می‌گردد تا پدیده هم‌خطی چندگانه در میان متغیرهای مستقل شناسایی شود. از آنجا که حضور توصیف‌گرهای دارای همبستگی و وابستگی خطی شدید موجب تورم واریانس، ناپایداری ضرایب رگرسیونی و در نهایت افت قدرت تفسیرپذیری و توان پیش‌بینی مدل QSAR می‌گردد، غربالگری این متغیرهای حشو^۱ امری الزامی است. این فیلترینگ آماری در محیط متلب، با حذف ویژگی‌های تکراری، تضمین‌کننده استخراج یک زیرمجموعه بهینه، مستقل و غنی از اطلاعات جهت ورود به فاز نهایی برازش مدل می‌باشد [۴۵].

۲-۷- نرم‌افزار اتوداک^۲

بسته نرم‌افزاری اتوداک امروزه به مثابه یکی از ارکان بنیادین در حوزه شیمی محاسباتی و رهیافت‌های طراحی داروی ساختارمحور تثبیت شده است. رسالت اصلی این پلتفرم محاسباتی، شبیه‌سازی دقیق و پیش‌بینی نحوه جهت‌گیری فضایی و اتصال مولکول‌های کوچک (لیگاندها) درون پاکت اتصالی^۳ یا جایگاه فعال ماکرومولکول‌های زیستی می‌باشد. این الگوریتم با بهره‌گیری از توابع امتیازدهی پیشرفته و متدهای جستجوی کنفورماسیونی، بهینه‌ترین آرایش فضایی کمپلکس لیگاند-گیرنده را جستجو نموده و مقادیر انرژی آزاد اتصال ($\Delta G_{\text{binding}}$) را با دقت بالایی برآورد می‌نماید.

در پیکربندی متدولوژیک رساله حاضر، اجرای عملیات داکینگ مولکولی با استعانت از نسخه ۲,۴,۶ نرم‌افزار AutoDock صورت پذیرفت. غایت اصلی از انجام این شبیه‌سازی‌های این‌سلیکو (In silico)، پایش و آنالیز دقیق نحوه استقرار سری ترکیبات مورد بررسی (اعم از مشتقات با اکتیویته بالا، ضعیف و ساختارهای پیشنهادی نوین) درون حفره کاتالیتیکی آنزیم/پروتئین هدف بوده است. در گام نهایی، نتایج استخراج‌شده از

¹ Redundant

² Autodock

³ Binding Pocket

این آنالیزهای ساختاری، به عنوان شالوده‌ای جهت بررسی شبکه برهم‌کنش‌های غیرکوالانسی، شناسایی آمینواسیدهای کلیدی در تثبیت کمپلکس کاندیداهای دارویی مورد استناد قرار گرفتند [۴۶].

۲-۸- نرم افزار ویورلایت^۱

بسته گرافیکی ویورلایت به عنوان یک رابط کاربری توانمند در راستای تصویرسازی سه‌بعدی شناخته می‌شود. این پلتفرم از قابلیت پردازش و رندرینگ^۲ طیف وسیعی از فرمت‌های استاندارد ذخیره‌سازی مختصات اتمی، از جمله PDB، MOL و MOL2 برخوردار است. به واسطه این ابزار، امکان رصد دقیق ماکرومولکول‌های پروتئینی و لیگاندهای آلی تحت الگوهای نمایشی متنوعی نظیر مدل‌های گلوله-میله^۳، فضاپرکن^۴ و سایر نماهای گرافیکی تعاملی فراهم می‌گردد. در چارچوب متدولوژی رساله پیش‌رو، نرم‌افزار ویورلایت در گام‌های آغازین پژوهش، به منظور پیش‌پردازش هندسی، بازبینی و آماده‌سازی فایل‌های مختصاتی گیرنده هدف (تارگت) و مولکول‌های لیگاند، پیش از ورود به فاز اصلی شبیه‌سازی‌ها و آنالیزهای محاسباتی مورد بهره‌برداری قرار گرفته است.

۲-۹- نرم افزار دیسکاوری استودیو^۵

پلتفرم محاسباتی دیسکاوری استودیو به مثابه یک بسته نرم‌افزاری جامع و یکپارچه در عرصه مدل‌سازی مولکولی و انفورماتیک زیست‌شیمیایی ایفای نقش می‌کند. این محیط شبیه‌سازی، ماژول‌های پیشرفته‌ای را جهت واکاوی سیستم‌های بیوماکرومولکولی در بر داشته و اجرای آنالیزهای پیچیده‌ای نظیر الگوریتم‌های داکینگ، شبیه‌سازی‌های دینامیک مولکولی (MD)، تکوین مدل‌های QSAR و پیاده‌سازی استراتژی‌های طراحی داروی ساختارمحور را تسهیل می‌نماید. برخورداری از یک رابط کاربری گرافیکی قدرتمند در این نرم‌افزار، امکان رندرینگ سه‌بعدی و پایش تعاملی شبکه‌های اتصالی میان لیگاند و تارگت پروتئینی، و همچنین ردیابی مسیرهای سیگنالینگ سلولی را با وضوح بالا فراهم می‌آورد.

¹ Viewer light

² Rendering

³ Ball-and-Stick

⁴ Space-filling/CPK

⁵ Discovery Studio

در پیکربندی متدولوژیک رساله حاضر، قابلیت‌های پردازشی این نرم‌افزار به‌طور ویژه جهت استخراج و واکاوی شبکه پیوندهای هیدروژنی، ترسیم دیاگرام‌های برهم‌کنش لیگاند-گیرنده در قالب نماهای توپولوژیک دوبعدی و فضایی سه‌بعدی، و نهایتاً شناسایی آمینواسیدهای کلیدی مستقر در پاکت اتصالی^۱ گیرنده هدف مورد بهره‌برداری قرار گرفت. اتخاذ این استراتژی بصری‌سازی، بستر مناسبی را برای ارزیابی جهت‌گیری فضایی poses کاندیداهای دارویی مهیا ساخته و امکان تفسیر گرافیکی و کمی خروجی‌های حاصل از فرآیند داکینگ را با دقتی بالا میسر نمود [۴۷].

۱۰-۲- نرم افزار لیگ پلات^۲

نرم‌افزار LigPlot+ یک ابزار بیوانفورماتیکی برای نمایش برهم‌کنش‌های بین پروتئین و لیگاند است. این نرم‌افزار به‌صورت خودکار نمودارهای دوبعدی از پیوندهای هیدروژنی و تماس‌های آب‌گریز تولید می‌کند. LigPlot+ در مطالعات داکینگ مولکولی و طراحی دارو کاربرد گسترده‌ای دارد. این برنامه امکان مقایسه برهم‌کنش‌های چند لیگاند با یک پروتئین یا چند پروتئین مشابه را فراهم می‌کند. خروجی‌های آن به پژوهشگران کمک می‌کند تا جایگاه‌های اتصال و نوع تعاملات مولکولی را بهتر تحلیل کنند. به دلیل سادگی نمایش و دقت مناسب، LigPlot+ یکی از نرم‌افزارهای پرکاربرد در زیست‌شناسی ساختاری و کشف دارو محسوب می‌شود.

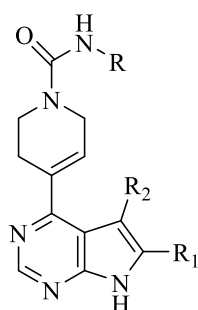
۲-۱۱- شرح ترکیبات و داده‌های تجربی استفاده‌شده در مدل‌سازی

مجموعه داده‌های مورد مطالعه در پژوهش حاضر، متشکل از ۴۴ ساختار شیمیایی متعلق به تعدادی از تتراهیدرو پیریدین به عنوان مهار کننده LIM پروتئین کیناز (LIMK) می‌باشد [۴۸]. از منظر بیومولکولی، پروتئین کیناز LIMK از طریق فسفریلاسیون و متعاقباً غیرفعال‌سازی پروتئین، نقش محوری در تنظیم دینامیک اسکلت سلولی ایفا نموده است. در عرصه شیمی دارویی، مشتقات حاوی داربست هتروسیکلی تتراهیدروپیریدین به دلیل نفوذپذیری بافتی مطلوب و فعالیت‌های زیستی گسترده، به عنوان فارماکوفراهای

¹ Binding Pocket

² Lig plot

ارزشمندی جهت توسعه بازدارنده‌های انتخابی این کیناز شناخته می‌شوند. این ترکیبات دارویی با استقرار در جایگاه فعال آنزیم و اتصال رقابتی به پاکت ATP، مسیر سیگنالینگ LIMK/Rho/ROCK/LIMK را مسدود می‌کنند. در نهایت، مهار هدفمند این پروتئین توسط لیگندهای تتراهیدروپیریدینی، از بازآرایی پاتولوژیک سیتواسکلتون جلوگیری کرده و یک استراتژی درمانی نوین و کارآمد برای مهاجرت سلول‌های سرطانی در شرایط سلول زنده^۱ مهیا می‌سازد. داده‌های فعالیت بیولوژیکی این ترکیبات در مطالعه اولیه، در قالب شاخص pKi LIMK2 گزارش گردیده است. جهت کالیراسیون داده‌ها و ورود آن‌ها به فاز مدل‌سازی کمی ارتباط ساختار-فعالیت (QSAR)، مقادیر مذکور از طریق تبدیلات ریاضی به شاخص مولار تبدیل شدند تا توزیع نرمال‌تری یافته و به عنوان متغیر وابسته یا بردار پاسخ هدف، در الگوریتم‌های رگرسیونی ایفای نقش نمایند. تنوع ساختاری تعبیه‌شده در این کتابخانه لیگاندی، بستر ایده‌آلی را جهت رصد تاثیرات ناشی از تنوع‌بخشی به استخلاف‌های پیرامونی حلقه آروماتیک و ممانعت فضایی و الکترونیکی بر میزان تمایل اتصالی فراهم می‌آورد. همان‌گونه که در شکل (۱-۲) و جدول (۱-۲) به تصویر کشیده شده، داربست ساختاری مشترک و مقادیر لگاریتمی فعالیت بیولوژیکی (موسوم به pKi LIMK2) برای تک‌تک اعضای اصلی گزارش شده است. هدف اصلی از سنتز این سری از مشتقات در مطالعات مرجع، دستیابی به بازدارندگی (Ki) تعدادی از مشتقات تتراهیدرو پیریدین به عنوان مهار کننده LIM پروتئین کیناز (LIMK) بوده که در این تحقیق، پتانسیل مهاري آن‌ها با بهره‌گیری از عملگر انقباض و انتخاب حداقل مطلق تطبیقی (ALASSO) مدل‌سازی پیش‌بینی می‌گردد.



شکل (۱-۲). اسکلت اصلی ترکیبات تتراهیدرو پیریدین.

¹ In-vivo

جدول (۲- ۱). ساختار و مقادیر pKi ترکیبات مورد مطالعه.

Number of compound	R	R ₁	R ₂	p Ki LIMK2
1 _{test}	3-chlorobenzene	H	H	6.67
2	phenyl -3,3-dimethylcarbamate	H	H	7.41
3	benzene	H	CH ₃	6.63
4	3-fluorobenzene	H	CH ₃	6.50
5	3-chlorobenzene	H	CH ₃	6.98
6	3-bromobenzene	H	CH ₃	6.97
7	3-methoxybenzene	H	CH ₃	6.58
8	3-trifluoromethoxybenzene	H	CH ₃	6.82
9	benzo [1,3]dioxole	H	CH ₃	6.16
10	N,N-dimethylaniline	H	CH ₃	6.37
11	N-phenylacetamide	H	CH ₃	6.74
12	phenyl -3,3-dimethylcarbamate	H	CH ₃	7.92
13	3-trifluorocarbonbenzene	H	CH ₃	6.79
14	3-benzocyanide	H	CH ₃	6.57
15	4-chloro-3-trifluorocarbonbenzene	H	CH ₃	6.94
16	3-pyridine	H	CH ₃	5.55
17	2-thiophene	H	CH ₃	6.73
18	naphthalene	H	CH ₃	7.04
19	1,1'-biphenyl	H	CH ₃	7.60
20	4-Phenylpyridine	H	CH ₃	6.89
21 _{test}	1-phenyl-1H-pyrrole	H	CH ₃	7.00
22 _{test}	1-phenyl-1H-imidazole	H	CH ₃	6.74
23 _{test}	2-methoxy-1,1'-biphenyl	H	CH ₃	7.85
24	3-methoxy-1,1'-biphenyl	H	CH ₃	8.10
25 _{test}	4-methoxy-1,1'-biphenyl	H	CH ₃	7.60
26	benzene	H	CycPr	6.87

ادامه جدول (۱-۲)

Number of compound	R	R ₁	R ₂	p K _i LIMK2
27	3-fluorobenzene	H	CycPr	7.07
28	3-chlorobenzene	H	CycPr	7.49
29	3-bromobenzene	H	CycPr	7.49
30	phenyl -3,3-dimethylcarbamate	H	CycPr	8.52
31	1,1'-biphenyl	H	CycPr	8.00
32	1-phenyl-1H-pyrrole	H	CycPr	8.00
33 _{test}	2-methoxy-1,1'-biphenyl	H	CycPr	8.15
34	3-methoxy-1,1'-biphenyl	H	CycPr	8.30
35	4-methoxy-1,1'-biphenyl	H	CycPr	7.68
36	3-chlorobenzene	H	Cl	6.98
37 _{test}	3-bromobenzene	H	Cl	7.12
38	phenyl -3,3-dimethylcarbamate	H	Cl	8.05
39	benzene	CH ₃	CH ₃	6.97
40	3-methoxybenzene	CH ₃	CH ₃	7.1
41	3-bromobenzene	CH ₃	CH ₃	7.64
42	phenyl -3,3-dimethylcarbamate	CH ₃	CH ₃	7.74
43	1,1'-biphenyl	CH ₃	CH ₃	7.77
44	4-methoxy-1,1'-biphenyl	CH ₃	CH ₃	7.57

Cycpr=cyclopropyl

Test: داده های سری آزمون

فصل سوم: بحث و نتایج

۳-۱- مدل سازی فعالیت دارویی مشتقات منتخب با استفاده از مدل Adaptive

LASSO (ALASSO)

مدل سازی کمی رابطه میان ساختار مولکولی و رفتار زیستی آن ها از اصول کلیدی در طراحی داروهای نوین محسوب می شود. هدف اصلی این رویکرد، برقراری پیوندی قابل اعتماد میان ویژگی های شیمیایی یک ترکیب و میزان فعالیت دارویی آن است تا بتوان عملکرد ترکیبات ناشناخته را پیش از سنتز یا آزمایش تجربی پیش بینی کرد. با توجه به حجم بالای توصیف کننده های مولکولی و محدود بودن تعداد نمونه ها، یکی از مهم ترین چالش ها در ساخت مدل های QSAR، مواجهه با داده هایی است که ابعاد آنها نسبت به اندازه مجموعه داده بسیار بزرگ است. چنین شرایطی معمولاً منجر به بیش برآزش، افت پایداری و کاهش توان پیش بینی مدل می شود. به همین دلیل، استفاده از روش هایی که بتوانند انتخاب ویژگی و کنترل پیچیدگی مدل را به طور هم زمان انجام دهند، ضرورت دارد. در میان روش های منظم سازی^۱، تکنیک های انقباضی نقش ویژه ای در بهبود کیفیت مدل های رگرسیونی دارند؛ زیرا ضمن اعمال جریمه بر ضرایب، ویژگی های غیرضروری را حذف کرده و تمرکز مدل را بر توصیف کننده های تأثیرگذار افزایش می دهند [۴۹-۵۱]. روش هایی مانند ریدج^۲، لاسو و الاستیک نت^۳ در این زمینه شناخته شده اند [۵۲-۵۵]. اما نسخه کلاسیک LASSO با وجود کاربردهای گسترده، محدودیت هایی نیز دارد. اعمال وزن یکسان بر تمامی ضرایب باعث می شود متغیرهای مهم نیز بیش از اندازه کوچک شوند و در حضور هم بستگی شدید میان توصیف کننده ها، انتخاب ویژگی توسط مدل از پایداری کافی برخوردار نباشد [۱۳]. برای رفع این نواقص، در این پژوهش از روش ALASSO استفاده شده است؛ رویکردی که با تخصیص وزن های متفاوت و تطبیقی به ضرایب، قادر است متغیرهای واقعاً مؤثر را با دقت بالاتری شناسایی کند و هم زمان سوگیری موجود در مدل های لاسو را کاهش دهد. مزیت دیگر ALASSO این است که می تواند در داده هایی با تعداد بسیار زیاد ویژگی عملکردی پایدار ارائه دهد و انتخاب متغیرها را به شکلی قابل اعتماد انجام دهد [۱۹]. در این مطالعه، ALASSO نه تنها

¹ Regularization

² Ridge

³ Elastic Net

برای انتخاب مجموعه‌ای بهینه از توصیف‌کننده‌ها به کار گرفته شده، بلکه نقش مدل رگرسیونی اصلی را نیز ایفا کرده است. به این ترتیب، فرآیند انتخاب ویژگی و مدل‌سازی در یک چارچوب واحد انجام شده و از پیچیدگی ناشی از ترکیب چندین روش مختلف جلوگیری شده است. پس از تعیین ضرایب و ساخت مدل نهایی، عملکرد آن با شاخص‌های آماری معتبر مورد ارزیابی قرار گرفت تا میزان صحت پیش‌بینی و توانایی مدل در بازتاب رابطه ساختار-فعالیت مشخص شود. نتایج ارزیابی‌ها نشان دادند که ALASSO قادر است با کارایی بالا، الگوهای موجود در داده‌ها را آشکار کرده و رفتار زیستی ترکیبات جدید را با صحت مناسبی پیش‌بینی کند.

۳-۲- بهینه‌سازی و آماده‌سازی ساختار سه‌بعدی ترکیبات

پس از تعیین ترکیبات هدف در این پژوهش، ساختارهای سه‌بعدی آن‌ها با بهره‌گیری از نرم‌افزار HyperChem (نسخه ۸/۰) ایجاد گردید. این مرحله به‌منزله گام اولیه برای تمام تحلیل‌های بعدی در حوزه محاسبات ساختاری محسوب می‌شود. جهت دستیابی به آرایش پایدار و انرژی حداقلی، بر روی هر مولکول به‌صورت جداگانه فرآیند بهینه‌سازی هندسی انجام شد تا شکل فضایی واقع‌بینانه‌ای از ترکیب حاصل شود. در این مطالعه از روش نیمه‌تجربی AM1^۱ به عنوان پایه تئوری محاسبات استفاده شد؛ انتخاب این مدل مبتنی بر توانایی آن در فراهم‌سازی توازن مناسب میان سرعت محاسبات و دقت قابل قبول در تعیین انرژی و خصوصیات هندسی مولکول‌های آلی متوسط حجم است. روند بهینه‌سازی تا همگرایی انرژی ادامه یافت و به محض آن‌که مقدار ریشه میانگین مربعات گرادیان نیرو به کمتر از ۰/۰۰۱ کیلوکالری بر مول رسید، محاسبات متوقف گردید. این مقدار، نشانه رسیدن مولکول به حداقل انرژی پایدار و ساختار تعادلی آن بود. در نهایت، ساختارهای پایدار به‌دست‌آمده به عنوان داده پایه برای محاسبه و استخراج توصیف‌کننده‌های مولکولی به کار رفتند تا مراحل بعدی مدل‌سازی و تحلیل کمی ساختار-فعالیت بر اساس آن‌ها انجام گیرد.

¹ Austin Model 1

۳-۳- تولید توصیف‌گرهای کمی مولکولی برای مدل سازی QSAR

پس از نهایی‌سازی ساختارهای مولکولی از طریق بهینه‌سازی هندسی، گام بعدی در این پژوهش، استخراج معیارهای عددی (توصیف‌گرها) بود که بتوانند ویژگی‌های ساختاری ترکیبات را به‌طور کمی نشان دهند. برای تحقق این هدف، از نرم‌افزار Dragon (نسخه ۵/۵) استفاده شد. این نرم‌افزار با تحلیل دقیق ساختار سه‌بعدی مولکول‌ها، مجموعه‌ای غنی از پارامترهای توصیفی را محاسبه می‌کند که جنبه‌های مختلف مولکولی مانند توزیع الکترونی، توپولوژی اتمی، و خواص فیزیکوشیمیایی را منعکس می‌نمایند. در طول فرآیند استخراج توصیف‌گرها، در مجموع ۳۲۲۴ پارامتر کمی برای هر ترکیب محاسبه شد. این پارامترها در دسته‌بندی‌های متنوعی از جمله شاخص‌های هندسی، آماری، و فیزیکوشیمیایی جای می‌گیرند. هدف از این استخراج گسترده، ایجاد یک پایگاه داده جامع از ویژگی‌های مولکولی بود که بتواند اساس تحلیل‌های بعدی در راستای مدل‌سازی رابطه کمی ساختار-فعالیت (QSAR) را تشکیل دهد. این مجموعه داده‌ها، امکان شناسایی و کمی‌سازی روابط میان ساختار مولکولی و فعالیت زیستی را در مراحل بعدی فراهم می‌سازد.

۳-۳- انتخاب بهینه‌ی توصیف‌گرهای مؤثر

پس از محاسبه‌ی کامل مجموعه اولیه توصیف‌گرهای مولکولی، مرحله بعد شامل غربالگری و انتخاب زیرمجموعه‌ای بهینه از آن‌ها برای توسعه مدل بود. هدف از این روند، جلوگیری از افزایش غیرضروری ابعاد داده، حذف اطلاعات تکراری ناشی از همبستگی بالا، و دستیابی به مدلی ساده، پایدار و با توان پیش‌بینی مناسب بود.

مراحل پیش‌پردازش و غربالگری اولیه به ترتیب زیر انجام شد:

الف) حذف توصیف‌گرهای همسان:

در گام نخست، تمام توصیف‌گرهایی که در بین تمامی ترکیبات مقدار ثابت داشتند از مجموعه حذف شدند. این عملیات تعداد توصیف‌گرهای قابل استفاده را به ۱۵۷۵ مورد کاهش داد.

ب) حذف توصیف‌گرهای با واریانس ناچیز:

در ادامه، توصیف‌گرهایی با تغییرپذیری بسیار کم (کمتر از ۰/۰۰۱) از مجموعه داده کنار گذاشته شدند. پس از این مرحله، ۱۴۶۳ توصیف‌گر برای انجام فرآیند انتخاب متغیر باقی ماند.

ج) کاهش همبستگی:

برای توصیف‌گرهای باقیمانده، ماتریس ضرایب همبستگی دوجانبه محاسبه شد. در مواردی که همبستگی میان دو توصیف‌گر بیشتر از ۰/۹ بود، آن توصیف‌گری حذف شد که ارتباط ضعیف‌تری با متغیر پاسخ (فعالیت زیستی) نشان می‌داد. این مرحله تعداد متغیرها را به ۳۸۶ مورد کاهش داد. با وجود کاهش قابل توجه تعداد توصیف‌گرها از ۳۲۲۴ به ۳۸۶ مورد، همچنان انتخاب مؤثرترین زیرمجموعه ضروری بود. برای این منظور، روش ALASSO مورد استفاده قرار گرفت. در این روش، متغیر وابسته شامل مقادیر فعالیت دارویی و متغیرهای مستقل شامل مجموعه ۳۸۶ تایی توصیف‌گرها در نظر گرفته شدند.

پیش از اجرای ALASSO، داده‌ها با استفاده از الگوریتم KS^1 و بر اساس فاصله اقلیدسی^۲ به دو بخش داده‌های آموزش و داده‌های آزمون تقسیم شدند. در این پژوهش کلیه مراحل انتخاب متغیر تنها بر روی مجموعه آموزش انجام گرفت و داده‌های آزمون برای ارزیابی نهایی مدل کنار گذاشته شدند.

۳-۴- انتخاب متغیرها با مدل رگرسیون انقباضی ALASSO و ارزیابی عملکرد

در این پژوهش، ابتدا از مدل رگرسیون انقباضی LASSO استاندارد (با $\alpha=1$) برای دستیابی به یک برآورد اولیه از ضرایب متغیرها و شناسایی توصیف‌گرهای بالقوه مؤثر استفاده شد. فرآیند انتخاب پارامتر تنظیم (λ) برای این مدل از طریق اعتبارسنجی متقابل ۱۰ لایه^۳ انجام گرفت و مقدار بهینه λ (که منجر به کمترین خطا در مجموعه اعتبارسنجی شد) برای محاسبه ضرایب نهایی ضرایب LASSO مورد استفاده قرار گرفت. ضرایب غیر-صفر LASSO اولیه، پایه و اساس محاسبه وزن‌های تطبیقی^۴ برای مدل ALASSO را تشکیل

¹ Kennard Stone

² Euclidean distance

³ 10-fold cross-validation

⁴ Adaptive weights

دادند.

سپس، با استفاده از معکوس مقادیر قدر مطلق ضرایب LASSO اولیه (و اعمال یک بایاس کوچک برای جلوگیری از تقسیم بر صفر)، وزن‌های تطبیقی محاسبه شدند. این وزن‌ها به گونه‌ای طراحی شده‌اند که توصیف‌گرهای مهم‌تر (با ضرایب LASSO بزرگتر) وزن کمتر و توصیف‌گرهای کمتر مهم (با ضرایب LASSO کوچکتر) وزن بیشتری دریافت کنند. این رویکرد اطمینان می‌دهد که در مرحله ALASSO، جریمه اعمال شده برای متغیرهای مهم کمتر و برای متغیرهای کمتر مهم بیشتر خواهد بود، که این خود به انتخاب صحیح تر و حفظ متغیرهای مهم کمک می‌کند.

در مرحله نهایی انتخاب متغیر، مدل ALASSO با استفاده از وزن‌های محاسبه شده اجرا گردید. مجدداً، اعتبارسنجی متقابل ۱۰-لایه برای تعیین پارامتر تنظیم بهینه ($\lambda_{\min}$) به کار گرفته شد. در نهایت، توصیف‌گرهایی که پس از اجرای مدل ALASSO، ضرایب نهایی غیرصفر دریافت کردند، به عنوان مجموعه بهینه متغیرهای پیش‌بین انتخاب شدند. طبق نتایج حاصل از اجرای این مدل، ۷ توصیف‌گر شناسایی گردیدند که جزئیات آن‌ها در جدول (۱-۳) آورده شده است.

جدول (۱-۳). جدول توصیف‌گرهای انتخاب‌شده با استفاده از روش ALASSO

ضرایب استاندارد شده	معنا	طبقه بندی	نماد	ردیف
+۰/۴۹	Eigenvalue 05 from edge adj. matrix weighted by edge degrees	edge adjacency indices	EEig05x	۰۱
-۰/۲۰	3D-MoRSE - signal 27 / weighted by atomic masses	3D-MoRSE descriptors	Mor27m	۰۲
+۰/۱۷	3D-MoRSE - signal 28 / weighted by atomic masses	3D-MoRSE descriptors	Mor28m	۰۳
-۰/۰۹	1st component symmetry directional WHIM index / weighted by atomic Sanderson electronegativities	WHIM descriptors	G1e	۰۴
+۰/۲۵	1st component accessibility directional WHIM index / weighted by atomic polarizabilities	WHIM descriptors	E1p	۰۵
+۰/۰۱	H autocorrelation of lag 5 / weighted by atomic masses	GETAWAY descriptors	H5m	۰۶
+۰/۲۰	H autocorrelation of lag 7 / weighted by atomic polarizabilities	GETAWAY descriptors	H6p	۰۷

مدل نهایی بر پایه توصیف‌گرهای منتخب توسط ALASSO ساخته شد. چنانچه در رابطه (۱-۳) به

نگارش درآمده است.

$$pK_i = -1.41 + (+2.20 \times \text{EEig05x}) + (-1.12 \times \text{Mor27m}) + (+1.02 \times \text{Mor28m}) + (-10.03 \times \text{G1e}) + (+3.01 \times \text{E1p}) + (+0.15 \times \text{H5m}) + (+2.05 \times \text{H6p})$$

رابطه (۳-۱)

در این معادله، همان ضرایب بهینه به دست آمده از مرحله انتخاب متغیر، ابتدا مطابق با رابطه (۳-۲) استاندارد سازی شده و سپس به عنوان ضرایب مدل نهایی در نظر گرفته شدند.

$$\beta_j^{std} = \beta_j \times \frac{SD(X_j)}{SD(Y)}$$

رابطه (۳-۲)

که در آن $SD(X_j)$ انحراف معیار مقادیر توصیف گر (X_j) در مجموعه داده برای ۴۴ ترکیب که بیانگر میزان پراکندگی مقادیر آن توصیف گر بوده و $SD(Y)$ انحراف معیار مقادیر متغیر پاسخ Y (فعالیت بیولوژیکی) در مجموعه داده که میزان پراکندگی مقادیر پاسخ را نشان می دهد.

۳-۵- ارزیابی جامع عملکرد و قابلیت اطمینان مدل پیش بین فعالیت زیستی

مبتنی بر ALASSO

عملکرد مدل نهایی ALASSO (رابطه (۳-۱))، به عنوان یک مدل خطی، با دقت و از جنبه های مختلف مورد ارزیابی قرار گرفت تا از استحکام و قابلیت تعمیم آن به ترکیبات جدید اطمینان حاصل شود. در ابتدای فرآیند مدل سازی، مجموعه داده ها با استفاده از الگوریتم کنارد-استون به دو بخش آموزش و آزمون تقسیم شدند. این تقسیم بندی به منظور حفظ گستره و پراکندگی فضای شیمیایی داده ها انجام گرفت و در تمامی مراحل بعدی مدل سازی، ارزیابی و محاسبات آماری، از همین مجموعه های آموزش و آزمون ثابت استفاده شد. برای سنجش قدرت پیش بینی و پایداری مدل در برابر تغییرات احتمالی در مجموعه داده ها، از رویکرد اعتبارسنجی متقابل استفاده گردید. به طور مشخص، از روش اعتبارسنجی متقابل ۱۰-لایه بهره گرفته شد. در این روش، کل مجموعه داده ها به ۱۰ بخش (لایه) مساوی تقسیم شده و در هر مرحله، یکی از این بخش ها به عنوان مجموعه آزمون موقت در نظر گرفته می شود، در حالی که ۹ بخش باقی مانده برای آموزش

مدل به کار گرفته می‌شدند. عملکرد مدل در هر مرحله از اعتبارسنجی متقابل با محاسبه معیارهایی چون ضریب تعیین (R^2)، میانگین قدر مطلق خطا (MAE) و میانگین مربعات خطا (MSE) بر روی مجموعه آزمون موقت، سنجیده شد. در نهایت، میانگین این معیارها بر روی هر ۱۰ مرحله، به عنوان شاخص نهایی عملکرد مدل در ادامه گزارش گردید. این نتایج، ارزیابی بیانی از توانایی مدل در پیش‌بینی فعالیت زیستی ترکیبات سری آزمون به روش ارزیابی متقاطع را نشان می‌دهد. علاوه بر این، به منظور اطمینان از اینکه همبستگی‌های مشاهده شده بین توصیف‌گرهای منتخب و فعالیت زیستی، تصادفی نبوده و از اعتبار آماری برخوردار هستند، آزمون Y- تصادفی اجرا شد.

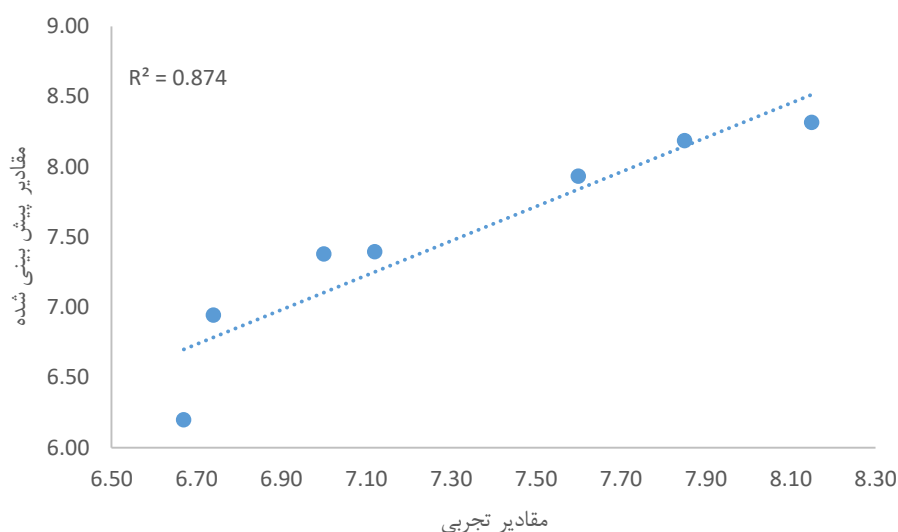
۳-۵-۱- ارزیابی عملکرد مدل پیش‌بین ALASSO بر روی مجموعه داده آزمون

به منظور سنجش قابلیت اطمینان و قدرت تعمیم مدل نهایی QSAR، که بر اساس توصیف‌گرهای منتخب روش ALASSO ساخته شده است، عملکرد آن بر روی یک مجموعه داده آزمون مستقل مورد ارزیابی قرار گرفت. این مجموعه داده، شامل ترکیبات مولکولی بود که در هیچ مرحله‌ای از فرآیند آموزش مدل یا انتخاب متغیرها نقش ایفا نکرده بودند. استفاده از چنین مجموعه‌ای، امکان ارزیابی واقعی از توانایی مدل در پیش‌بینی فعالیت زیستی ترکیبات جدید را فراهم می‌آورد، که این امر ماهیت کاربرد عملیاتی مدل را منعکس می‌سازد. در این ارزیابی، تنها توصیف‌گرهایی که پیشتر توسط الگوریتم ALASSO به عنوان متغیرهای کلیدی و دارای ضرایب غیر-صفر شناسایی شده بودند، به عنوان ورودی مدل مورد استفاده قرار گرفتند. این انتخاب دقیق، تضمین می‌کند که مدل صرفاً بر اساس مؤثرترین و مرتبط‌ترین ویژگی‌های مولکولی عمل پیش‌بینی را انجام دهد. فعالیت زیستی ترکیبات موجود در مجموعه داده آزمون، با استفاده از مدل ALASSO نهایی پیش‌بینی گردید. سپس، مقادیر پیش‌بینی شده به دقت با مقادیر واقعی و تجربی گزارش شده برای این ترکیبات مقایسه شد. این مقایسه، میزان صحت پیش‌بینی‌های مدل را در شرایطی کاملاً مستقل از داده‌های آموزشی، مشخص می‌سازد (جزئیات نتایج در جدول (۲-۳) ارائه شده است).

جدول (۳-۲). ارزیابی عملکرد پیش‌بینی مدل ALASSO روی ترکیبات مجموعه آزمون.

درصد خطا %	pK_i پیش‌بینی	pK_i واقعی	شماره ترکیب در مجموعه داده‌ها
-۷/۰۸	۶/۲۰	۶/۶۷	۱
+۵/۴۰	۷/۳۸	۷/۰۰	۲۱
+۳/۰۰	۶/۹۴	۶/۷۴	۲۲
+۴/۲۷	۸/۱۹	۷/۸۵	۲۳
+۴/۳۶	۷/۹۳	۷/۶۰	۲۵
+۲/۰۱	۸/۳۱	۸/۱۵	۳۳
+۳/۸۶	۷/۳۹	۷/۱۲	۳۷

هماهنگی قابل توجه میان مقادیر پیش‌بینی شده توسط مدل ALASSO و مقادیر واقعی مشاهده شده، گواهی بر صحت قابل قبول مدل و توانایی آن در بازنمایی صحیح روابط ساختار-فعالیت در مجموعه داده‌های آزمون بود. نتایج حاصل از این مقایسه به صورت خطا، که نشان‌دهنده نزدیکی پیش‌بینی‌های مدل به داده‌های واقعی است، در جدول (۳-۲) نمایش داده شده است. به منظور ارزیابی کمی قدرت پیش‌بینی مدل، نموداری از مقادیر پیش‌بینی شده بر حسب مقادیر واقعی ترسیم گردید (شکل (۳-۱)). در این نمودار، ضریب تعیین (R^2) برای مجموعه داده آزمون مقدار ۰/۷۸۴ محاسبه شد. مقدار R^2 به دست آمده بالاتر از ۰/۶ بوده که بیانگر توانایی مدل در پیش‌بینی دقیق فعالیت زیستی ترکیبات در مجموعه داده آزمون است. علاوه بر این، جهت ارزیابی عمیق‌تر عملکرد مدل و بررسی هرگونه سوگیری احتمالی، مقادیر باقی‌مانده (تفاوت میان مقادیر واقعی و پیش‌بینی شده) محاسبه شده و نمودار آن‌ها بر حسب مقادیر واقعی رسم گردید (شکل (۳-۲)). توزیع متوازن و یکنواخت نقاط داده در اطراف خط افقی صفر، نشان‌دهنده عدم وجود بایاس سیستماتیک در پیش‌بینی‌های مدل ALASSO است. این پراکندگی متعادل، همچنین قابلیت تعمیم‌پذیری مطلوب مدل به داده‌های آزمون که مستقل از فرآیند آموزش بوده‌اند را تأیید می‌کند.



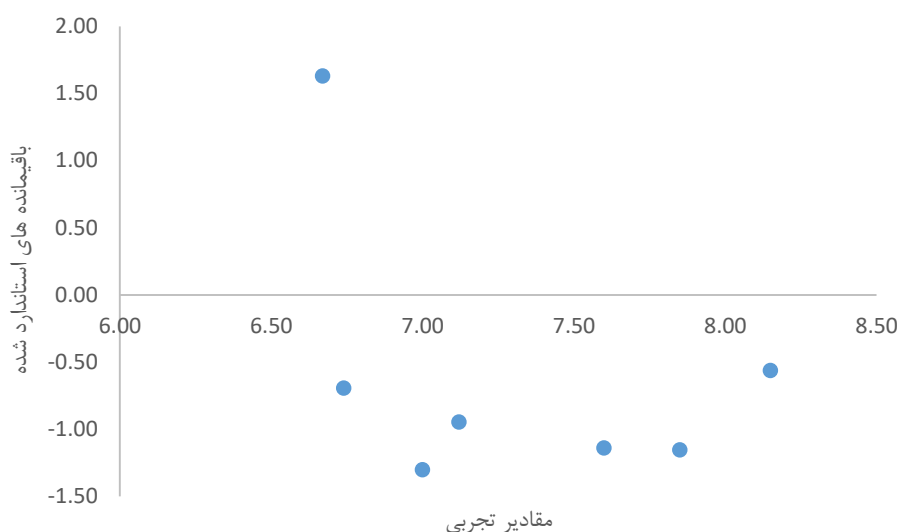
شکل (۳-۱). نمودار مقادیر پیش بینی شده مدل لاسوی تطبیقی بر حسب مقادیر تجربی برای داده های سری آزمون.

علاوه بر این، جهت ارزیابی عمیق تر عملکرد مدل و بررسی هرگونه سوگیری احتمالی، مقادیر باقی مانده (تفاوت میان مقادیر واقعی و پیش بینی شده) محاسبه شده و بر اساس معادله (۳-۳) استانداردسازی شده و در نهایت نمودار آن ها بر حسب مقادیر واقعی رسم گردید (شکل (۳-۲)).

$$\text{Standard residuals} = \left(\frac{R_{\text{siduals}}}{SD_{\text{residuals}}} \right) \quad \begin{matrix} \text{رابطه (۳-۳)} \\ \text{(۳)} \end{matrix}$$

که در آن $SD_{\text{residuals}}$ انحراف استاندارد کل باقیمانده هاست.

با توجه به محدود بودن تعداد نمونه های آزمون، تغییرات جزئی در پراکندگی باقیمانده های مجموعه آزمون چنانچه در شکل (۳-۲) مشاهده می شود، می تواند قابل انتظار باشد. به منظور بررسی پایداری مدل و اطمینان از مستقل بودن عملکرد آن از نحوه تقسیم بندی داده ها، در ادامه آزمون اعتبارسنجی حذف تک به تک (LOO) انجام شد که نتایج آن قابلیت اطمینان مدل توسعه یافته را تأیید کرد.



شکل (۳-۲). نمودار باقیمانده های استاندارد شده بر حسب مقادیر تجربی برای داده های سری آزمون.

۳-۵-۲- ارزیابی توان پیش‌بینی و تعمیم‌پذیری مدل ALASSO با استفاده از

اعتبارسنجی حذف تک‌نمونه‌ای (LOO-CV)

به‌منظور ارزیابی میزان پایداری و قابلیت تعمیم‌پذیری مدل ALASSO، از روش اعتبارسنجی حذف تک‌نمونه‌ای (LOO-CV) بهره گرفته شد. این رویکرد که به‌عنوان یکی از شیوه‌های سنجش دقیق در مطالعات QSAR شناخته می‌شود، شامل کنار گذاشتن متوالی تنها یک نمونه از مجموعه داده‌ها در هر مرحله است. مدل ALASSO با بهره‌گیری از سایر نمونه‌های موجود، مجدداً آموزش داده می‌شود و سپس مقدار فعالیت زیستی نمونه‌ی حذف‌شده توسط مدل پیش‌بینی می‌گردد. این فرآیند تا زمانی تکرار می‌شود که هر ترکیب منحصربه‌فرد، یک‌بار به‌عنوان داده‌ی آزمون مورد استفاده قرار گیرد. نتایج به‌دست‌آمده از اجرای این روش برای تمامی ترکیبات، در جدول (۳-۳) منعکس شده است.

جدول (۳-۳). نتایج حاصل از اعتبارسنجی حذف تک‌نمونه‌ای (LOO) برای مدل ALASSO، شامل مقادیر تجربی و پیش‌بینی‌شده pK_i و درصد خطای محاسبه‌شده برای هر ترکیب.

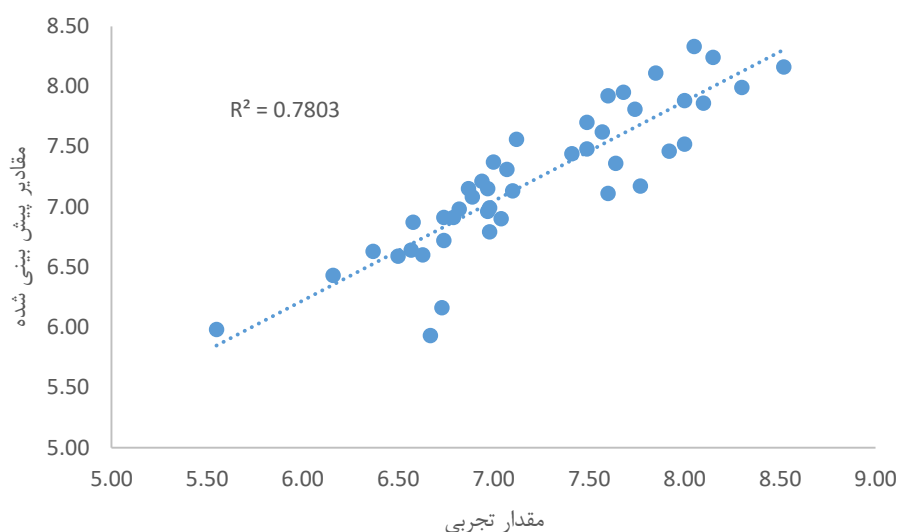
درصد خطا	مقدار پیش‌بینی شده pK_i	مقدار واقعی pK_i	شماره ترکیب	درصد خطا	مقدار پیش‌بینی شده pK_i	مقدار واقعی pK_i	شماره ترکیب
+۳/۳۱	۸/۱۱	۷/۸۵	۲۳	-۱۱/۰۹	۵/۹۳	۶/۶۷	۱
-۲/۹۶	۷/۸۶	۸/۱۰	۲۴	+۰/۴۰	۷/۴۴	۷/۴۱	۲
+۴/۲۱	۷/۹۲	۷/۶۰	۲۵	-۰/۴۵	۶/۶۰	۶/۶۳	۳
+۴/۰۸	۷/۱۵	۶/۸۷	۲۶	+۱/۳۸	۶/۵۹	۶/۵۰	۴
+۳/۳۹	۷/۳۱	۷/۰۷	۲۷	-۲/۷۲	۶/۷۹	۶/۹۸	۵
-۰/۱۳	۷/۴۸	۷/۴۹	۲۸	+۲/۵۸	۷/۱۵	۶/۹۷	۶
+۲/۸۰	۷/۷۰	۷/۴۹	۲۹	+۴/۴۱	۶/۸۷	۶/۵۸	۷
-۴/۲۳	۸/۱۶	۸/۵۲	۳۰	+۲/۳۵	۶/۹۸	۶/۸۲	۸
-۱/۵۰	۷/۸۸	۸/۰۰	۳۱	+۴/۳۸	۶/۴۳	۶/۱۶	۹
-۶/۰۰	۷/۵۲	۸/۰۰	۳۲	+۴/۰۸	۶/۶۳	۶/۳۷	۱۰
+۱/۱۰	۸/۲۴	۸/۱۵	۳۳	-۰/۳۰	۶/۷۲	۶/۷۴	۱۱
-۳/۷۳	۷/۹۹	۸/۳۰	۳۴	-۵/۸۱	۷/۴۶	۷/۹۲	۱۲
+۳/۵۲	۷/۹۵	۷/۶۸	۳۵	+۱/۷۷	۶/۹۱	۶/۷۹	۱۳
+۰/۱۴	۶/۹۹	۶/۹۸	۳۶	+۱/۰۷	۶/۶۴	۶/۵۷	۱۴
+۶/۱۸	۷/۵۶	۷/۱۲	۳۷	+۳/۸۹	۷/۲۱	۶/۹۴	۱۵
+۳/۴۸	۸/۳۳	۸/۰۵	۳۸	+۷/۷۵	۵/۹۸	۵/۵۵	۱۶
-۰/۱۴	۶/۹۶	۶/۹۷	۳۹	-۸/۴۷	۶/۱۶	۶/۷۳	۱۷
+۰/۴۲	۷/۱۳	۷/۱۰	۴۰	-۱/۹۹	۶/۹۰	۷/۰۴	۱۸
-۳/۶۶	۷/۳۶	۷/۶۴	۴۱	-۶/۴۵	۷/۱۱	۷/۶۰	۱۹
+۰/۹۰	۷/۸۱	۷/۷۴	۴۲	+۲/۷۶	۷/۰۸	۶/۸۹	۲۰
-۷/۷۲	۷/۱۷	۷/۷۷	۴۳	+۵/۲۹	۷/۳۷	۷/۰۰	۲۱
+۰/۶۶	۷/۶۲	۷/۵۷	۴۴	+۲/۵۲	۶/۹۱	۶/۷۴	۲۲

همان‌گونه که نتایج جدول (۳-۳) نشان می‌دهد، مدل توسعه‌یافته مبتنی بر روش لاسوی تطبیقی (ALASSO) از توانایی مناسبی در پیش‌بینی مقادیر (pK_i) برخوردار است. مقایسه مقادیر واقعی و پیش‌بینی‌شده نشان می‌دهد که برای اکثر ترکیبات، اختلاف بین این دو مقدار اندک بوده و درصد خطای پیش‌بینی در محدوده قابل قبولی قرار دارد. به‌گونه‌ای که حدود ۶۵ درصد از ترکیبات دارای درصد خطای کمتر از ۵ درصد هستند که در مطالعات QSAR، بیانگر دقت مناسب مدل در پیش‌بینی فعالیت زیستی ترکیبات جدید است.

بررسی جزئی‌تر نتایج نشان می‌دهد که بعضی ترکیبات دارای کمترین میزان خطای پیش‌بینی (کمتر از ۱ درصد) هستند که بیانگر توانایی مدل در برآورد دقیق فعالیت زیستی این دسته از ترکیبات است. در مقابل، تعداد محدودی از ترکیبات نیز دارای خطاهای بزرگ‌تری هستند. بروز چنین اختلافاتی در مدل‌های QSAR دور از انتظار نبوده و می‌تواند ناشی از عواملی نظیر پراکندگی ذاتی داده‌های زیستی، تفاوت‌های ساختاری قابل توجه برخی ترکیبات نسبت به سایر اعضای مجموعه داده و یا محدود بودن تعداد نمونه‌های آموزشی باشد.

از سوی دیگر، مقدار ضریب تعیین پیش‌بینی حذف تک‌نمونه‌ای ($Q_{Loo}^2 = 0.78$) بیانگر پایداری و قدرت پیش‌بینی مطلوب مدل است. با توجه به اینکه در مطالعات QSAR معمولاً مقادیر بزرگ‌تر از ۰/۵۰ به‌عنوان حداقل معیار پذیرش مدل و مقادیر بالاتر از ۰/۷۰ نشان‌دهنده مدل‌های با قابلیت پیش‌بینی بالا در نظر گرفته می‌شوند، مقدار به‌دست‌آمده در این پژوهش حاکی از استحکام و قابلیت تعمیم مناسب مدل توسعه‌یافته است [۵۰]. همچنین نزدیکی مقادیر پیش‌بینی‌شده به مقادیر تجربی برای بخش عمده‌ای از ترکیبات نشان می‌دهد که مدل ALASSO توانسته است بدون بیش‌برازش، رابطه بین توصیف‌گرهای مولکولی و فعالیت زیستی را به‌خوبی استخراج کند.

به‌طور کلی، نتایج اعتبارسنجی حذف تک‌نمونه‌ای (LOO) نشان می‌دهد که مدل ALASSO علاوه بر برخورداری از صحت مناسب، از پایداری و قابلیت تعمیم مطلوبی نیز برخوردار است. بنابراین، این مدل می‌تواند به‌عنوان ابزاری قابل اعتماد برای پیش‌بینی فعالیت زیستی مشتقات جدید تتراهیدروپیریدین و غربالگری اولیه ترکیبات بالقوه مهارکننده پروتئین‌کیناز LIMK مورد استفاده قرار گیرد.



شکل (۳-۳). نمودار مقایسه مقادیر پیش‌بینی شده pK_i توسط مدل ALASSO بر حسب مقادیر تجربی متناظر آن‌ها بر اساس اعتبارسنجی حذف تک‌نمونه‌ای (LOO).

شکل (۳-۳) نمودار پراکنش مقادیر پیش‌بینی شده در برابر مقادیر واقعی پاسخ حاصل از اعتبارسنجی حذف تک‌نمونه‌ای را برای مدل مبتنی بر روش ALASSO نشان می‌دهد. قرارگیری بخش عمده نقاط در امتداد خط روند بیانگر وجود یک رابطه خطی مناسب بین مقادیر واقعی و مقادیر پیش‌بینی شده است. این موضوع نشان می‌دهد که مدل توانسته است الگوی کلی تغییرات فعالیت زیستی ترکیبات را به درستی شناسایی کرده و رفتار کلی داده‌ها را بازسازی کند.

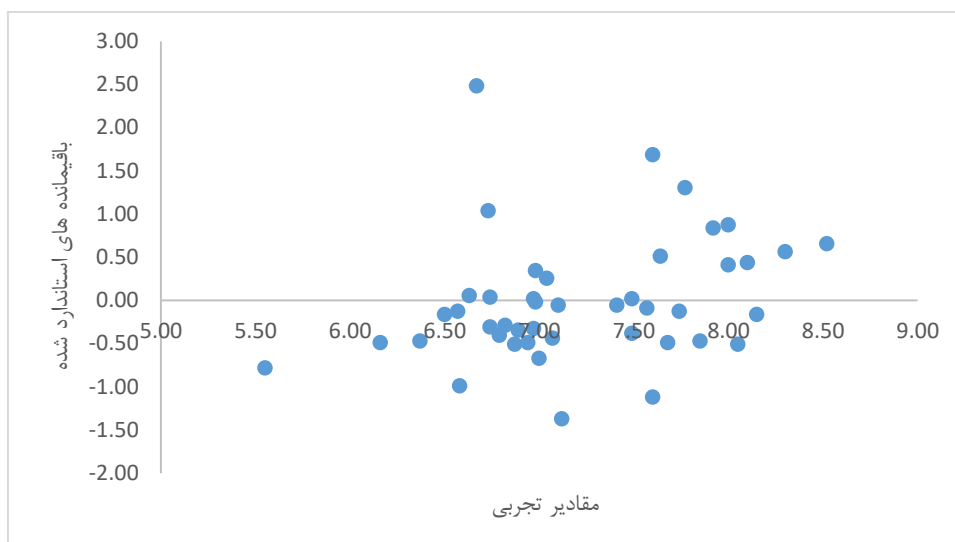
از سوی دیگر، توزیع نسبتاً یکنواخت نقاط در سراسر دامنه تغییرات پاسخ بیانگر آن است که عملکرد مدل تنها به یک بازه خاص از داده‌ها محدود نشده و در مقادیر پایین، متوسط و بالای فعالیت زیستی، روند پیش‌بینی مشابهی را حفظ کرده است. همچنین، عدم مشاهده خوشه‌بندی غیرعادی یا تمرکز نقاط در ناحیه‌ای خاص نشان می‌دهد که مدل نسبت به بخش‌های مختلف فضای شیمیایی مجموعه داده، عملکرد متعادلی داشته است.

اگرچه در تعداد بسیار محدودی از ترکیبات فاصله بیشتری از خط روند مشاهده می‌شود، اما این نقاط به‌صورت کاملاً پراکنده در نمودار توزیع شده‌اند و الگوی مشخص یا جهت‌داری را تشکیل نمی‌دهند. چنین رفتاری بیانگر آن است که خطاهای مدل غالباً ماهیتی تصادفی داشته و ناشی از پیچیدگی ذاتی داده‌های زیستی، تفاوت‌های ساختاری میان برخی ترکیبات یا خطاهای اجتناب‌ناپذیر آزمایشگاهی هستند، نه وجود ضعف سیستماتیک در ساختار مدل.

همچنین مقدار $Q_{Loo}^2 = 0.78$ نشان می‌دهد که همبستگی مشاهده‌شده در نمودار صرفاً یک برازش ظاهری نیست، بلکه مدل از توانایی مناسبی در پیش‌بینی نمونه‌های جدید نیز برخوردار است. این مقدار بیانگر آن است که مدل توسعه‌یافته علاوه بر برازش مناسب داده‌های موجود، از پایداری داخلی مطلوبی برخوردار بوده و قابلیت تعمیم قابل قبولی برای کاربردهای پیش‌بینی در مطالعات QSAR دارد.

همچنین شکل (۳-۴) نمودار باقیمانده‌های حاصل از اعتبارسنجی حذف تک‌نمونه‌ای را برای مدل مبتنی بر روش ALASSO نشان می‌دهد. همان‌گونه که مشاهده می‌شود، باقیمانده‌ها به‌صورت تصادفی در اطراف خط صفر پراکنده شده‌اند و الگوی منظم یا روند مشخصی در توزیع آن‌ها دیده نمی‌شود. این موضوع نشان می‌دهد که مدل توانسته است رابطه بین توصیفگرهای مولکولی و فعالیت زیستی را به‌خوبی استخراج کند و عدم مشاهده الگوهایی نظیر روند افزایشی یا کاهشی، شکل قیفی (ناهمسانی واریانس) یا ساختارهای منظم در پراکندگی باقیمانده‌ها، بیانگر آن است که فرضیات اصلی مدل نقض نشده و مدل از سوگیری سیستماتیک و بیش‌برازش قابل توجه برخوردار نیست.

در مجموع، شکل (۳-۳) و شکل (۳-۴) تأیید می‌کنند که مدل ALASSO علاوه بر دستیابی به یک رابطه خطی مناسب بین مقادیر واقعی و پیش‌بینی‌شده، از نظر بصری نیز عملکرد قابل قبولی ارائه می‌دهد و عدم وجود انحرافات سیستماتیک در پراکندگی نقاط، قابلیت اعتماد مدل را برای پیش‌بینی فعالیت زیستی مشتقات جدید تتراهیدروپیریدین به‌عنوان مهارکننده‌های بالقوه پروتئین کیناز LIMK تقویت می‌کند.



شکل (۳-۴). نمودار باقی مانده های استاندارد شده بر حسب مقادیر تجربی pK_i برای مدل ALASSO در چهارچوب اعتبارسنجی حذف تک نمونه ای (LOO).

۳-۵-۳- ارزیابی عملکرد مدل ALASSO بر پایه شاخص های آماری

به منظور سنجش کامل عملکرد مدل ALASSO، مجموعه جامعی از شاخص های آماری محاسبه شد. این ارزیابی با هدف تعیین میزان خطا، سنجش دقت و بررسی قابلیت تعمیم پذیری مدل صورت گرفت. شاخص های مذکور هم برای مجموعه داده آزمون خارجی و هم در چارچوب اعتبارسنجی حذف تک نمونه ای (LOO) برای پیش بینی فعالیت زیستی ترکیبات، مورد تحلیل قرار گرفتند. نتایج حاصل، همراه با حدود توصیه شده برای هر پارامتر، در جدول (۳-۴) به تفصیل نمایش داده شده است.

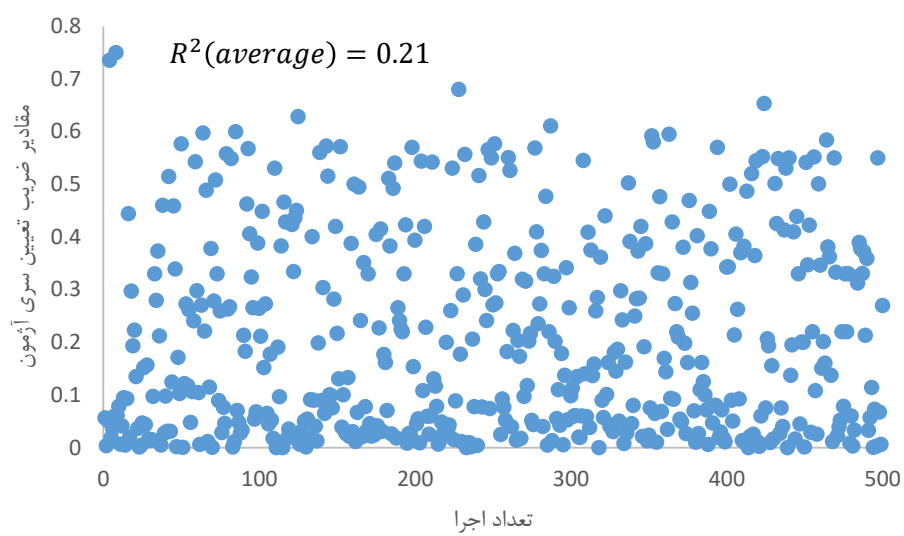
جدول (۳-۴). شاخص های آماری ارزیابی عملکرد مدل QSAR توسعه یافته، شامل پارامترهای خطا و معیارهای برازش برای مجموعه های آزمون و Q_{LOO}^2 اعتبارسنجی حذف تک نمونه ای (LOO)

ردیف	پارامتر آماری	مجموعه آزمون	LOO	محدوده قابل قبول
۱	PRESS	۰/۷۳	۱۳/۰۷	-
۲	SEP	۰/۳۲	۰/۵۵	-
۳	MAE	۰/۳۱	۰/۳۸	-
۴	REP	۴/۴۲	۷/۵۵	-
۵	MSE	۰/۱۰	۰/۳۰	-
۶	MRE	۴/۲۷	۵/۳۷	-
۷	R^2	۰/۸۷	-	$> ۰/۶$
۸	Q^2	-	۰/۷۸	$> ۰/۵$

بررسی همزمان شاخص‌های جدول (۳-۴) نشان می‌دهد که مدل ALASSO عملکردی نسبتاً قابل قبول دارد. از یک سو، ضریب تعیین بسیار بالا ($R^2=0.87$) و مقادیر نسبتاً پایین خطاها در مجموعه آزمون، حاکی از توانایی قوی مدل در یادگیری روابط ساختار-فعالیت و تبیین واریانس داده‌ها است. از سوی دیگر، مقدار $Q_{LoO}^2 = 0.78$ که به‌طور قابل ملاحظه‌ای بالاتر از حداقل مقدار پذیرفته‌شده در مطالعات QSAR است، نشان‌دهنده پایداری، قابلیت تعمیم مناسب و قدرت پیش‌بینی مطلوب مدل در اعتبارسنجی داخلی است. همچنین، نزدیک بودن مقادیر R^2 و Q^2 بیانگر آن است که مدل علاوه بر برازش مناسب داده‌های آموزشی، در پیش‌بینی نمونه‌های جدید نیز عملکرد قابل قبولی داشت.

۳-۵-۴- بررسی اعتبار مدل ALASSO با آزمون Y-تصادفی

برای اطمینان از عدم وجود همبستگی صرفاً تصادفی بین ویژگی‌های مولکولی و متغیر پاسخ، آزمون Y-تصادفی بر روی مدل ALASSO توسعه‌یافته اجرا گردید. در این فرایند، مقادیر متغیر پاسخ (Y) به صورت تصادفی و مستقل از توصیفگرهای ورودی، ۵۰۰ بار جایگزین شد. در هر تکرار، مدل ALASSO با همان معماری اصلی آموزش داده شد و ضریب تعیین (R^2) برای مجموعه آزمون خارجی محاسبه گردید. نمودار پراکنش مقادیر R^2 حاصل از این تکرارهای تصادفی، در برابر تعداد تکرارها، در شکل (۳-۵) نمایش داده شده است. همانطور که مشهود است، اکثر مقادیر R^2 در نزدیکی صفر متمرکز شده‌اند و توزیع آن‌ها فاقد هرگونه روند یا الگوی مشخص است. اگرچه در برخی تکرارهای محدود، مقادیر R^2 بالاتری مشاهده می‌شود، اما این مقادیر به صورت پراکنده، نامنظم و کاملاً تصادفی ظاهر شده‌اند و هیچ‌گونه تداوم یا ساختار معناداری ندارند. این پراکندگی نامنظم و تمرکز غالب مقادیر R^2 در نواحی پایین، نشان می‌دهد که پس از برهم زدن تصادفی پاسخ‌ها، مدل قادر به استخراج یک رابطه پایدار بین توصیفگرها و فعالیت زیستی نیست. این رفتار تأیید می‌کند که عملکرد قابل قبول مدل نهایی ALASSO ناشی از تصادف یا برازش تصادفی داده‌ها نبوده، بلکه مدل به طور واقعی بر اساس رابطه ساختار-فعالیت مولکولی بنا شده است. در نتیجه، نتایج آزمون Y-تصادفی، صحت، پایداری و قابلیت اعتماد مدل ALASSO را تأیید می‌نماید.



شکل (۳-۵). نتایج Y - تصادفی

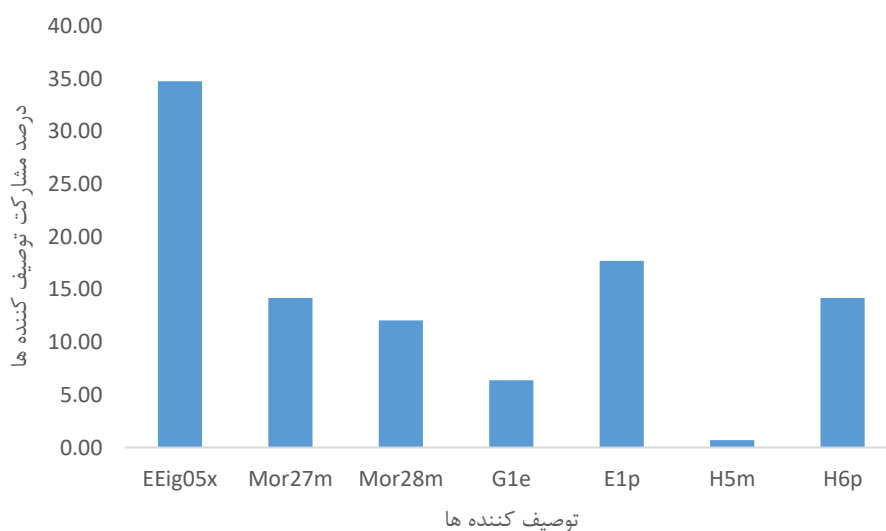
فصل چہارم: بحث و نتیجہ گیری

۴-۱- جمع‌بندی نهایی عملکرد مدل ALASSO و تحلیل اهمیت توصیف‌کننده‌ها

از آنجا که روش ALASSO یکی از روش‌های منظم‌سازی مبتنی بر جریمه است که فرآیند انتخاب متغیر و برآورد ضرایب مدل را به صورت هم‌زمان انجام می‌دهد. توصیف‌کننده‌های باقی‌مانده در مدل نهایی دارای ضرایب متفاوتی هستند که بیانگر میزان تأثیر نسبی آن‌ها بر فعالیت زیستی ترکیبات مورد مطالعه است. با این حال، به دلیل تفاوت در مقیاس اندازه‌گیری توصیف‌کننده‌های مولکولی، مقایسه مستقیم ضرایب خام نمی‌تواند معیار مناسبی برای ارزیابی اهمیت واقعی متغیرها باشد. به همین دلیل، از ضرایب رگرسیونی استاندارد شده در بخش سوم (با استفاده از معادله (۲-۳) استفاده شد تا امکان مقایسه مستقیم اثر توصیف‌کننده‌ها فراهم شود.

درصد سهم مشارکت هر توصیف‌کننده مطابق رابطه زیر تعیین گردید:

$$\text{رابطه (۴-۱)} \quad \text{درصد مشارکت توصیف‌گر } i = \frac{|\text{ضریب استاندارد توصیف‌گر } i|}{\sum |\text{ضرایب استاندارد شده}|}$$



شکل (۴-۱) نمودار درصد مشارکت توصیف‌کننده‌ها در مدل ALASSO

نتایج حاصل از شکل (۴-۱) نشان داد که توصیف‌کننده EEig05x بیشترین سهم را در مدل ALASSO به خود اختصاص داده است. توصیف‌کننده EEig05x یکی از توصیف‌کننده‌های توپولوژیکی موجود در نرم‌افزار دراگون است. این شاخص از مقادیر ویژه با وزن‌دهی بر اساس نوع و ویژگی پیوندها محاسبه می‌شود و اطلاعاتی درباره نحوه آرایش و اتصال پیوندهای موجود در ساختار مولکول ارائه می‌دهد. مقدار این توصیف‌کننده تحت تأثیر عواملی مانند الگوی اتصال اتم‌ها، میزان شاخه‌دار بودن، حضور حلقه‌ها و پیچیدگی ساختاری مولکول قرار دارد؛ بنابراین می‌تواند تغییرات ظریف در توپولوژی مولکول را منعکس کند. در مدل‌های QSAR/QSPR، EEig05x به عنوان شاخصی از ویژگی‌های ساختاری مولکول در نظر گرفته می‌شود و نقش آن بر اساس ضریب رگرسیونی مدل تفسیر می‌شود. چنانچه ضریب این توصیف‌کننده در معادله مدل مثبت باشد، افزایش مقدار EEig05x با افزایش مقدار متغیر پاسخ همراه است؛ در مقابل، ضریب منفی بیانگر رابطه معکوس میان این توصیف‌کننده و پاسخ مدل خواهد بود. بنابراین، علامت مثبت یا منفی به خود توصیف‌کننده مربوط نیست، بلکه نشان‌دهنده جهت اثر آن در مدل آماری است. برتری قابل توجه این توصیف‌کننده نسبت به سایر متغیرها نشان می‌دهد که بخش عمده اطلاعات مورد نیاز برای پیش‌بینی فعالیت زیستی ترکیبات از طریق این ویژگی مولکولی تأمین می‌شود. بنابراین EEig05x را می‌توان مهم‌ترین عامل ساختاری در مدل نهایی دانست. پس از آن، توصیف‌کننده E1p در رتبه دوم قرار گرفت که بیانگر نقش مؤثر این متغیر در توضیح تغییرات فعالیت زیستی است. همچنین توصیف‌کننده‌های Mor27m و H6p با مقادیر تقریباً مشابه سهم قابل توجهی در عملکرد مدل نشان دادند. مشارکت نزدیک این دو توصیف‌کننده بیانگر آن است که هر دو ویژگی اطلاعات ساختاری مکملی را برای مدل فراهم کرده‌اند. در مرتبه بعد، توصیف‌کننده Mor28m قرار دارد که اگرچه اهمیت آن کمتر از متغیرهای اصلی مدل است، اما همچنان نقش معناداری در قدرت پیش‌بینی مدل ایفا می‌کند. همچنین توصیف‌کننده Gle با مقدار حدود ۱۰ درصد دارای سهم کمتری در مقایسه با سایر متغیرها بوده است. کمترین میزان مشارکت مربوط به توصیف‌کننده H5m است که مقدار آن نزدیک به ۱ مشاهده شد. با وجود سهم اندک این متغیر، انتخاب آن توسط الگوریتم

ALASSO نشان می‌دهد که این ویژگی همچنان دارای اطلاعات مفیدی بوده و در کنار سایر توصیف‌کننده‌ها به بهبود عملکرد کلی مدل کمک کرده است.

به طور کلی، نتایج نشان می‌دهد که مدل ALASSO بیشترین اتکا را به توصیف‌کننده EEig05x دارد، در حالی که سایر توصیف‌کننده‌ها نقش مکمل و پشتیبان را در ساختار مدل ایفا می‌کنند. این الگوی مشارکت بیانگر آن است که اگرچه یک ویژگی مولکولی نقش غالبی در پیش‌بینی فعالیت زیستی دارد، اما دستیابی به بهترین عملکرد مدل تنها از طریق ترکیب چندین توصیف‌کننده منتخب امکان‌پذیر بوده است. از دیدگاه QSAR، این موضوع نشان می‌دهد که فعالیت زیستی ترکیبات مورد مطالعه تحت تأثیر مجموعه‌ای از ویژگی‌های ساختاری مختلف قرار دارد و روش ALASSO توانسته است با حذف متغیرهای غیرضروری، مدل نهایی را به مجموعه‌ای محدود از توصیف‌کننده‌های اطلاعاتی و قابل تفسیر تقلیل دهد. در نتیجه، مدل حاصل علاوه بر برخورداری از تفسیرپذیری مناسب، دارای ساختار تنک^۱ بوده و خطر بیش‌برازش در آن کاهش یافته است.

۴-۲- پیشنهاد ترکیبات جدید با فعالیت دارویی مناسب با استفاده از عملگر

انقباض و انتخاب حداقل مطلق تطبیقی (ALASSO)

در حوزه کمومتریکس و طراحی دارو، بهره‌گیری از رویکردهای مدل‌سازی کمی جهت غربالگری و معرفی ساختارهای شیمیایی نوین با پاسخ بیولوژیکی هدف، امری ضروری ناپذیر است. در این راستا، الگوریتم لاسوی تطبیقی به عنوان یک تکنیک پیشرفته و هوشمند انتخاب متغیر، امکان استخراج و شناسایی دقیق توصیف‌گرهای ساختاری تعیین‌کننده در فعالیت زیستی را میسر می‌سازد. بر مبنای توصیف‌گرهای کلیدی استخراج‌شده توسط الگوریتم لاسوی تطبیقی، ساختار شیمیایی ترکیب شماره (۱۶) که دارای پاسخ بیولوژیکی ضعیفی با مقدار فعالیت (۵/۵۵) بود، به عنوان اسکلت پایه جهت اعمال تغییرات ساختاری هدفمند انتخاب گردید. برای پیش‌بینی فعالیت ترکیبات پیشنهادی به صورت زیر عمل می‌شود. در گام

¹ Sparse

نخست، ساختارهای دوبعدی ترکیبات پیشنهادی ترسیم گشت و به منظور دستیابی به پایدارترین کنفورماسیون‌ها، عملیات مینیمم‌سازی انرژی و بهینه‌سازی هندسی با استفاده از نرم افزار هایپرکم بر روی آن‌ها اعمال شد. متعاقباً، ماتریس توصیف‌گرهای فیزیکوشیمیایی و ساختاری برای مولکول‌های بهینه‌شده توسط نرم‌افزار دراگون محاسبه گردید. با اعمال تغییرات هدفمند در ساختار مشتقات تتراهیدروپیریدین، گروه‌های الکترون‌کشنده و حلقه‌های مختلف به اسکلت اصلی مولکول افزوده شدند. این اصلاحات ساختاری با هدف افزایش تنوع ساختاری انجام گرفت. هدف از این اصلاحات، طراحی شبیه سازی مجموعه‌ای از مشتقات تتراهیدروپیریدین با پتانسیل دارویی ارتقایافته بود. ساختار شیمیایی و مقادیر کمی فعالیت پیش‌بینی‌شده برای این ترکیبات تتراهیدروپیریدین در جدول (۱-۴) گزارش شده است. در مرحله بعد، توانمندی بیولوژیکی این ساختارهای جدید توسط مدل تلفیقی شبکه عصبی-لاسوی تطبیقی پیش‌بینی شد. بررسی خروجی‌ها اثبات نمود که همبستگی منطقی و معناداری میان پروفایل ساختاری و پاسخ‌های زیستی پیش‌بینی‌شده وجود دارد که کاملاً با مبانی تئوری کیوسار همخوانی دارد. جهت سنجش قابلیت‌های فارماکولوژیک ترکیبات طراحی‌شده، شبیه‌سازی داکینگ مولکولی پروتئین هدف با شناسه 8AAU به اجرا درآمد. در این آنالیز شبیه سازی، مقادیر انرژی آزاد اتصال^۱ در کنار شبکه‌ای از برهم‌کنش‌های غیرکوالانسی شامل پیوندهای هیدروژنی، به‌دقت بررسی گردید. خروجی‌های داکینگ مولکولی، درک عمیقی از جهت‌گیری فضایی^۲ لیگاندها در سایت فعال اتصال آن‌ها، ارائه نمود. روی‌هم‌رفته، پژوهش حاضر یک پلتفرم منطقی و مبتنی بر ساختار جهت تفسیر مکانیسم اثر ترکیبات ارائه داده و مسیر را برای طراحی استراتژیک لیگاندهای بهینه‌تر هموار می‌سازد. از بین ترکیبات پیشنهادی ترکیباتی قابل قبول هستند که رنج عددی آنها بین ۸-۹/۳۰ باشد که pKi آن نزدیک به ترکیب فعال (۳۰) باشد و از ترکیب کم فعال (۱۶) بیشتر باشد. G2 و G6 عملکرد بهتری را نشان دادند و داکینگ مولکولی آنها مورد بررسی قرار گرفت.

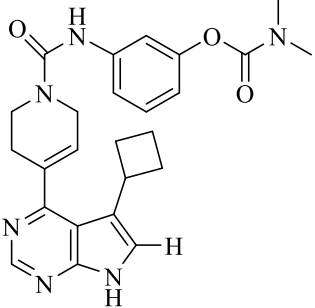
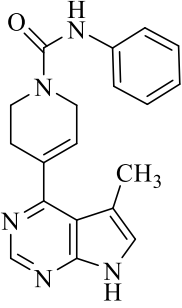
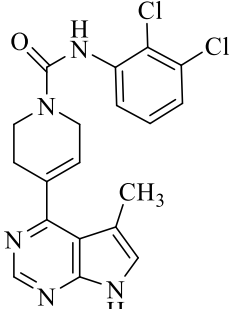
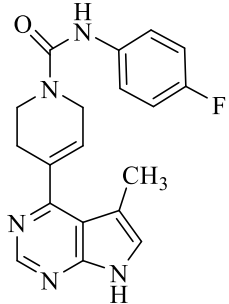
¹ Binding Free Energy

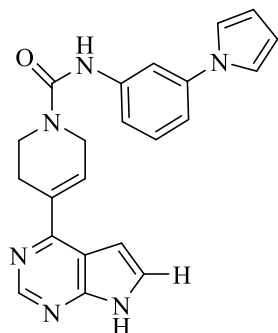
² Binding Pose

جدول (۴-۱) ساختار و فعالیت ترکیب ۱۶ و ۳۰ و ترکیب‌های پیشنهادی

شماره ترکیبات	ساختار ترکیبات	pKi	تعداد پیوند هیدروژنی
۳۰		۸/۵۲	۲
۱۶		۵/۵۵	۱
G6		۹/۲۶	۳
G2		۸/۴۳	۵

ادامه جدول (۱-۴)

شماره ترکیبات	ساختار ترکیبات	pKi	تعداد پیوند هیدروژنی
G1		۷/۵۴	-
G3		۶/۱۳	-
G4		۶/۳۱	-
G5		۶/۲۵	-



۴-۳- مطالعه داکینگ مولکولی

در این پژوهش، به منظور ارزیابی کمی و کیفی برهمکنش‌های بین مولکولی، از شبیه‌سازی داکینگ مولکولی با بهره‌گیری از مجموعه نرم‌افزاری^۱ بهره گرفته شد. پروتکل محاسباتی با آماده‌سازی ساختار سه‌بعدی رسپتور هدف آغاز گردید. این مرحله شامل افزودن هیدروژن‌های قطبی، ادغام هیدروژن‌های غیرقطبی با اتم‌های کربن متصل به آن‌ها و تخصیص بارهای جزئی اتمی به روش کولمن^۲ بود. این اقدامات جهت مدل‌سازی دقیق میدان نیروی الکترواستاتیک ماکرومولکول ضروری است. متعاقباً، ساختار لیگاند کریستالوگرافی نیز تحت فرآیند آماده‌سازی مشابهی قرار گرفت. در نهایت، هر دو مولکول رسپتور و لیگاند در فرمت PDBQT ذخیره شدند که حاوی اطلاعات مختصات اتمی، نوع اتم و بار جزئی برای اجرای محاسبات داکینگ مولکولی است.

فضای جستجوی کانفورماسیونی^۳ از طریق تعریف یک جعبه شبکه^۴ سه‌بعدی با ابعاد ۵۰×۸۰×۷۰ آنگستروم، که به طور کامل جایگاه فعال رسپتور را در بر می‌گرفت، مشخص گردید. تفکیک‌پذیری^۵ این شبکه با تعیین فاصله بین نقاط به مقدار ۱,۰۰ آنگستروم تنظیم شد تا توازن مناسبی بین دقت محاسباتی و هزینه زمانی شبیه‌سازی برقرار گردد. الگوریتم مورد استفاده برای جستجوی کانفورماسیون‌های پایدار، الگوریتم ژنتیک لامارکین^۶ بود. به منظور تضمین قابلیت اعتماد و تکرارپذیری آماری نتایج، برای هر لیگاند ۱۰۰ اجرای مستقل در نظر گرفته شد و سایر پارامترهای محاسباتی بر اساس مقادیر پیش‌فرض نرم‌افزار

¹ AutoDock

² Kollman

³ Conformational Search Space

⁴ Grid Box

⁵ Resolution

⁶ Lamarckian Genetic Algorithm

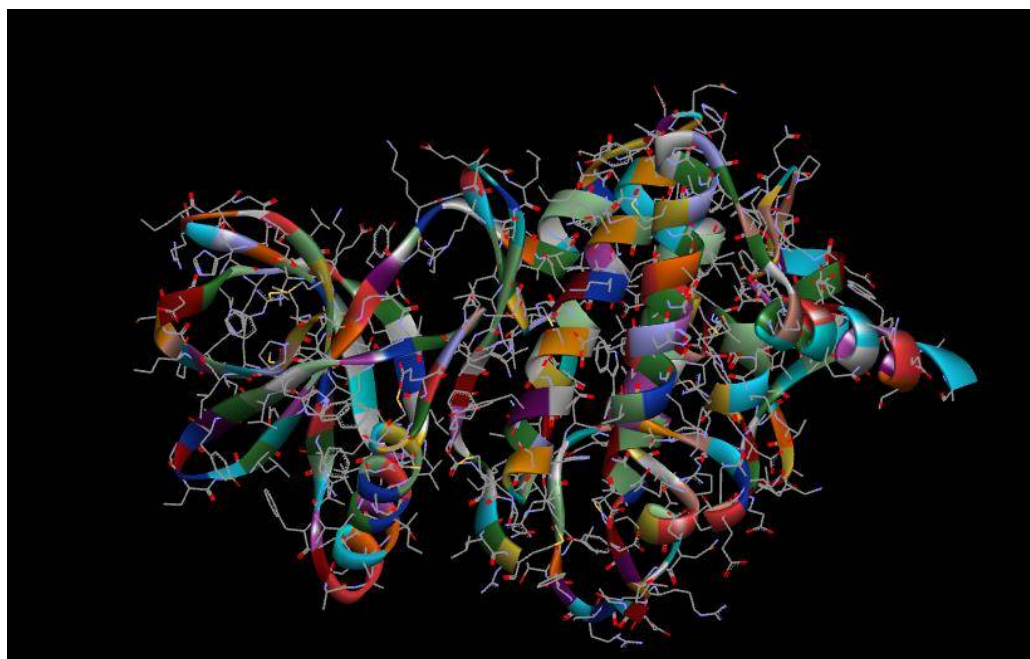
تنظیم گردید. پس از اتمام شبیه‌سازی‌ها، خروجی‌ها که در فایل‌های گزارش داکینگ (DLG) ثبت شده بودند، مورد تجزیه و تحلیل قرار گرفتند. معیار اصلی برای انتخاب بهینه‌ترین حالت اتصال، گزینشی بود که به پایین‌ترین انرژی آزاد اتصال (ΔG_{bind}) تعلق داشته است.

در مرحله پیش‌پردازش لیگاندها جهت اجرای پروتکل داکینگ مولکولی بر روی گیرنده هدف با کد شناسه 8AAU، ابتدا مدل‌سازی سه‌بعدی ترکیبات انجام پذیرفت. از آنجا که کانفورماسیون فضایی لیگاند تأثیر مستقیمی بر دقت محاسبات ترمودینامیکی دارد، فرآیند حداقل‌سازی انرژی و بهینه‌سازی هندسی بر روی تک‌تک مولکول‌ها اعمال گردید تا پایدارترین آرایش فضایی با کمترین سطح انرژی پتانسیل به دست آید. خروجی ساختارهای بهینه‌شده در گام نخست با فرمت mol. استخراج گردید. متعاقباً، به منظور سازگاری فایل‌های ورودی با پلتفرم‌های محاسباتی داکینگ، ساختارهای مذکور به محیط نرم‌افزار ویورلایت^۱ منتقل شدند. در این مرحله، تبدیل فرمت مختصاتی به قالب استاندارد pdb صورت گرفت تا مولکول‌ها جهت ارزیابی و شبیه‌سازی دقیق برهم‌کنش‌های لیگاند-پروتئین آماده گردند.

۴-۳-۱- آماده سازی پروتئین

در این پژوهش، پس از ارزیابی جامع متون علمی، مختصات بلورشناسی گیرنده هدف با شناسه مستند 8AAU از پایگاه داده‌های ساختاری (PDB) استخراج گردید. در گام آماده‌سازی ماکرومولکول از ابزار پردازشی Viewer ویورلایت بهره گرفته شد. در این مرحله از آماده‌سازی ساختاری، کلیه مولکول‌های حلال (آب‌های کریستالوگرافی)، کوفاکتورها، از شبکه مختصاتی حذف گردیدند. این عملیات پالایشی جهت جلوگیری از بروز خطاهای محاسباتی ناشی از ممانعت‌های فضایی و برهم‌کنش‌های کاذب الکترواستاتیک در حین شبیه‌سازی ضروری است. در نهایت، ساختار خالص‌سازی‌شده گیرنده مجدداً با پسوند استاندارد PDB جهت ورود به فاز محاسبات داکینگ ذخیره‌سازی شد (شکل ۴-۲).

¹ Viewer Lite



شکل (۴-۲) ساختار کریستالوگرافی پروتئین 8AAU را نشان می دهد.

۴-۳-۲- آماده سازی لیگاند

گام بعدی در پیش پردازش داده ها، خالص سازی ساختار لیگاند مرجع با بهره گیری از ابزار ویور لایت بود. بدین منظور، ماتریس دربرگیرنده لیگاند از طریق پاک سازی ماکرومولکول پروتئینی، آب های شبکه ای و سایر هترواتم ها اصلاح شد تا از بروز خطاهای محاسباتی و ممانعت های فضایی در طول شبیه سازی ممانعت به عمل آید. خروجی این مرحله، یک فایل مستقل از لیگاند با پسوند PDB بود. از آنجا که موقعیت یابی صحیح جایگاه فعال^۱ برای هدایت الگوریتم داکینگ مولکولی ضروری است، بر اساس محاسبات ساختاری، مرکز این فضای سه بعدی در مختصات $50 \times 80 \times 70$ آنگستروم با گام شبکه ای 1.00 تنظیم گردید. اعمال این پارامترهای فضایی مشخص موجب تمرکز فضای جستجو بر روی پاکت اتصالی^۲ شده و در نتیجه، وضوح برهم کنش های بین مولکولی را به میزان چشمگیری ارتقا می بخشد.

¹ Active site

² Binding pocket

۴-۳-۳- آماده سازی ترکیبات پیشنهادی برای فرایند داکینگ مولکولی

در پژوهش حاضر، نخست توپولوژی سه بعدی لیگاندهای کاندید، طراحی گردید و به منظور دستیابی به پایدارترین کانفورماسیون با حداقل سطح انرژی پتانسیل، تحت عملیات بهینه سازی هندسی قرار گرفتند. مختصات اتمی بهینه شده ترکیبات در وهله اول با فرمت استاندارد MOL استخراج گردید. در گام بعدی پیش پردازش، داده های ساختاری به محیط نرم افزار ویورلایت انتقال یافتند تا با تغییر قالب مختصاتی به PDB، پیش نیازهای ساختاری جهت ورود به فاز شبیه سازی مهیا گردد. در نهایت، این ساختارهای پردازش شده و استاندارد، به عنوان فایل های ورودی به موتور محاسباتی نرم افزار اتوداک معرفی شدند تا برهم کنش های ترمودینامیکی و میل ترکیبی آنها با ماکرومولکول هدف، از طریق پروتکل های داکینگ مولکولی مورد ارزیابی قرار گیرد.

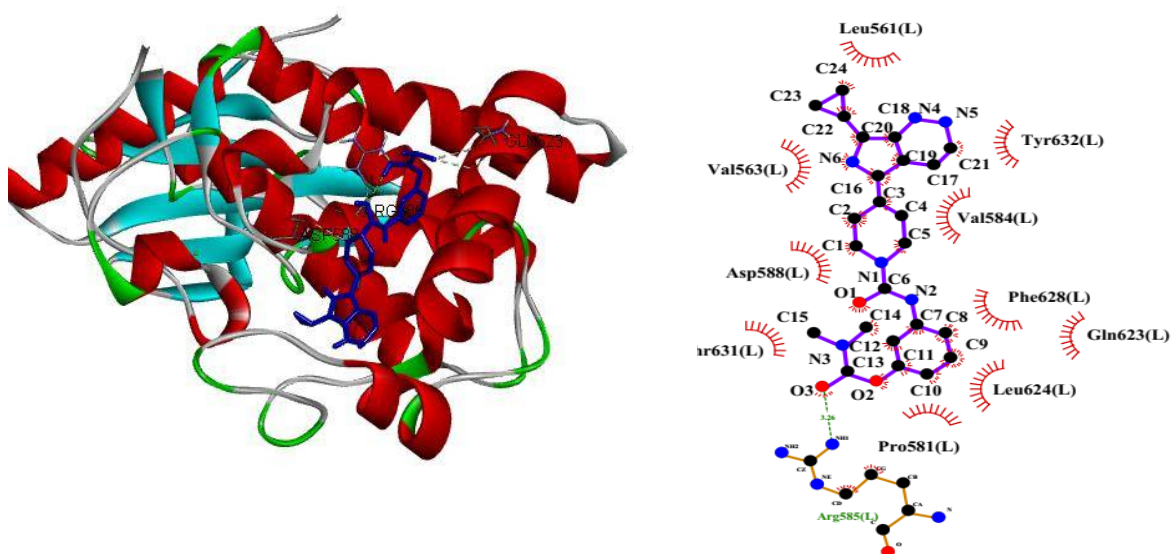
۴-۳-۴- اجرای فرایند داکینگ مولکولی ترکیبات پیشنهادی با پروتئین 8AAU

در راستای اجرای شبیه سازی داکینگ مولکولی، نخست موقعیت فضایی دقیق پاکت اتصالی در ماکرومولکول هدف معین گردید. کاوش در فضای کانفورماسیونی با بهره گیری از الگوریتم جستجوی ژنتیک لامارکین و تحت پارامترهای پیش فرض موتور محاسباتی صورت پذیرفت. مختصات و ابعاد جعبه شبکه نیز در تطابق کامل با مقادیر بهینه شده در فاز پیشین تنظیم شد. این پروتکل محاسباتی جهت ارزیابی تمایل اتصالی دو ساختار نوین پیشنهادی، در کنار دو ترکیب شامل ترکیب شماره ۳۰ به عنوان نمونه بسیار فعال با شاخص pKi معادل ۸/۵۲ میکرومولار و ترکیب شماره ۱۶ به عنوان نمونه با حداقل فعالیت زیستی و pKi برابر با ۵/۵۵ میکرومولار در جایگاه فعال گیرنده 8AAU به اجرا درآمد. در نهایت، با استخراج و تحلیل داده های ثبت شده در فایل های خروجی (dlg)، کانفورماسیونی که دارای کمترین (منفی ترین) مقدار انرژی آزاد اتصال ترمودینامیکی بود، به عنوان پایدارترین مُد اتصالی^۱ برای هر لیگاند برگزیده شد [۵۶-۵۸].

¹ Binding pose

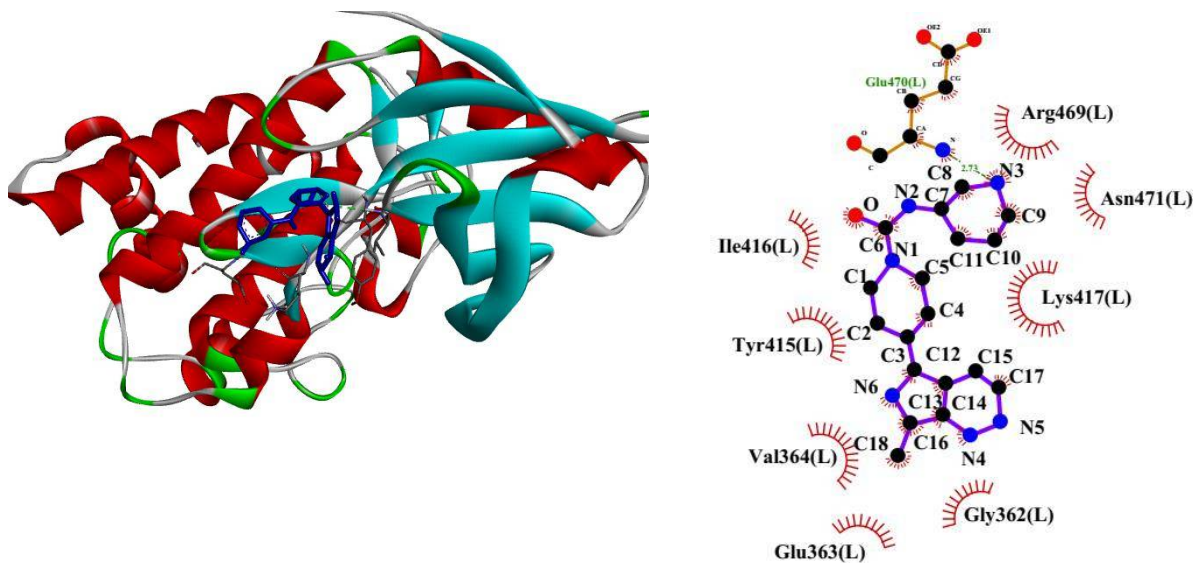
۴-۴- نتایج داکینگ مولکولی

الگوی برهم کنش‌های بین‌مولکولی میان لیگاندها و ماکرومولکول هدف، بر پایه داده‌های استخراج‌شده از شبیه‌سازی در شکل‌های ۳-۴ و ۴-۴ رسم گردید. بررسی‌های ساختاری حاکی از آن بود که هر دو دسته ترکیبات، فارغ از میزان فعالیت زیستی (قوی یا ضعیف)، ظرفیت محدودی در برقراری شبکه‌های پیوند هیدروژنی با ریشه‌های آمینواسیدی کلیدی مستقر در جایگاه فعال از خود بروز دادند. جهت آنالیز دقیق نیروهای غیرکوالانسی، شامل اتصالات هیدروژنی و سایر نیروهای بین‌مولکولی ضعیف در سطح تماس لیگاند-گیرنده، از پلتفرم محاسباتی دیسکآوری استودیو و لیگ پلات^۱ بهره‌گیری شد. در نهایت، پارامترهای ترمودینامیکی از جمله مقادیر انرژی آزاد اتصال، در کنار ماهیت فیزیکوشیمیایی و فراوانی پیوندهای شکل‌گرفته برای هر ترکیب، ارزیابی گردید.



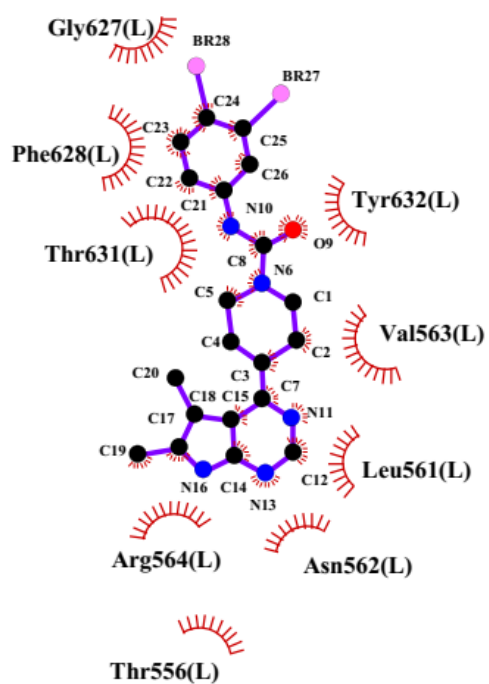
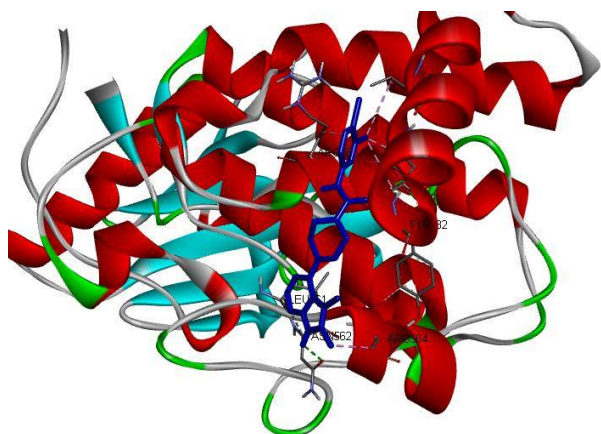
شکل (۳-۴) برهم‌کنش ترکیب ۳۰ با اسید آمینه‌های کلیدی جایگاه فعال پروتئین 8AAU

¹ Lig plot

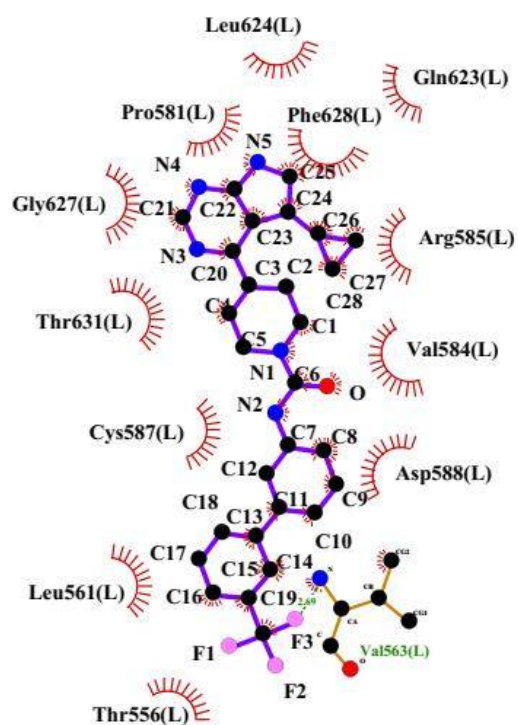
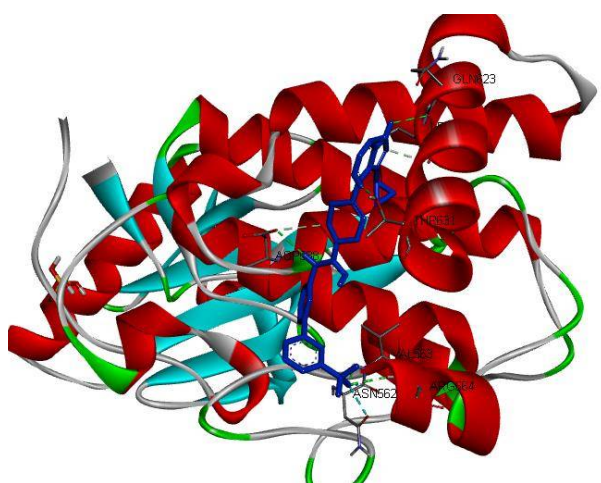


شکل (۴-۴) برهمکنش ترکیب ۱۶ با اسید آمینه های کلیدی جایگاه فعال پروتئین 8AAU

در گام بعدی از این مطالعه، شبکه‌ی تعاملات بین‌مولکولی میان لیگاندهای طراحی‌شده و پاکت اتصالی ماکرومولکول هدف با جزئیات کامل بررسی شد. جهت تصویرسازی توپولوژی این ترکیبات، نگاره‌های فضایی سه بعدی و دوبعدی برهم‌کنش‌ها از طریق پلتفرم دیسکآوری استودیو و لیگ پلات تولید و در شکل‌های ۴-۵ و ۴-۶ به تصویر کشیده شدند. در این فاز، ماهیت فیزیکوشیمیایی و موقعیت مکانی دقیق نیروهای غیرکووالانسی برقرار شده میان ترکیبات و اسید آمینه‌های مستقر در سایت فعال، با دقت ثبت گردید. ارزیابی داده‌های خروجی حاکی از آن بود که تمامی ساختارهای کاندید، موفق به تشکیل شبکه‌ای از پیوندهای هیدروژنی با اسیدهای آمینه جایگاه فعال جدول ۴-۱، ASP588, ARG564, VAL63, ASN562, PHE628, THR631, GLN623, TYR632, می‌کنید تعداد پیوند هیدروژنی در ترکیبات پیشنهادی بیشتر است. در نهایت، الگوهای اتصالات هیدروژنی به عنوان یکی از شاخص‌های بنیادین در سنجش کیفیت و پایداری کمپلکس لیگاند-پروتئین، استخراج و مورد استناد قرار گرفت.



شکل (۴-۵) برهمکنش ترکیب G6 با اسید آمینه های کلیدی جایگاه فعال پروتئین 8AAU



شکل (۴-۶) برهمکنش ترکیب G2 با اسید آمینه های کلیدی جایگاه فعال پروتئین 8AAU

۴-۵- نتایج مقایسه ای قانون لیپینسکی

در این پژوهش، شاخص‌های قانون لیپینسکی به‌عنوان یکی از مراحل کلیدی در طراحی و بهینه‌سازی دارو مورد ارزیابی قرار گرفت تا پیش از انجام آزمون‌های آزمایشگاهی، رفتار فارماکوکینتیکی و ایمنی ترکیبات بررسی شود. در این مرحله، ویژگی‌های مرتبط با جذب، توزیع، متابولیسم، دفع و همچنین سمیت مولکول‌ها تحلیل شدند. انجام این بررسی‌ها امکان تشخیص ترکیبات دارای ویژگی‌های مناسب دارویی و حذف گزینه‌های نامطلوب را فراهم کرد. یافته‌ها نشان دادند که ترکیبات انتخاب‌شده از نظر معیارهای قانون لیپینسکی در بازه قابل پذیرشی قرار دارند و از پتانسیل لازم برای ادامه مسیر توسعه دارویی برخوردارند. به‌طور کلی، به‌کارگیری تحلیل قانون لیپینسکی موجب افزایش دقت پیش‌بینی و بهبود کارایی فرایند طراحی دارو گردید. علاوه بر این، میزان تطابق ترکیبات با پنج قانون لیپینسکی با استفاده از پایگاه Swiss-ADME بررسی شد و نتایج مربوطه در جدول ۲-۴ گزارش شده است.

در این پژوهش، ویژگی‌های داروشناختی مبتنی بر قوانین لیپینسکی برای ترکیب فعال شماره ۳۰، ترکیب با فعالیت ضعیف شماره ۱۶ و همچنین مجموعه‌ای از ترکیبات جدید طراحی‌شده محاسبه و به‌صورت مقایسه‌ای تحلیل شد. نتایج به‌دست‌آمده نشان داد که تمامی ترکیبات پیشنهادی در محدوده قابل قبول معیارهای قانون پنج‌گانه لیپینسکی قرار دارند. این امر بیانگر آن است که این ترکیبات از نظر خصوصیات فیزیکوشیمیایی دارای شرایط مناسبی برای سنتز و بررسی بیشتر به‌عنوان نامزدهای دارویی هستند. به بیان دیگر، ترکیبات طراحی‌شده از دیدگاه ویژگی‌های داروماند، مشخصات مطلوبی داشته و احتمال می‌رود در محیط‌های زیستی نیز عملکرد مناسبی از خود نشان دهند [۵۹].

در ادامه ارزیابی‌ها، فعالیت زیستی این ترکیبات با بهره‌گیری از مدل پیش‌بینی لاسو برآورد شد. بر اساس داده‌های ارائه‌شده در جدول ۴-۲، مقادیر پیش‌بینی‌شده نشان می‌دهد که ترکیبات پیشنهادی از سطح فعالیت قابل قبول و رضایت‌بخشی برخوردارند. علاوه بر این، میزان سهولت سنتز ترکیبات طراحی‌شده نیز با استفاده از پایگاه داده Swiss-ADME مورد محاسبه قرار گرفت. این شاخص در بازه عددی ۱ تا ۱۰ تعریف می‌شود؛ به‌گونه‌ای که مقادیر نزدیک‌تر به ۱ بیانگر سادگی بیشتر فرآیند سنتز و امکان اجرای عملی بهتر آن

است. با توجه به مقادیر مناسب به دست آمده برای فعالیت زیستی پیش‌بینی شده و همچنین شاخص مطلوب سهولت سنتز، می‌توان نتیجه گرفت که ترکیبات پیشنهادی ظرفیت مناسبی برای ورود به مراحل آزمایشگاهی و ارزیابی‌های زیستی دارند. در گام پایانی، به منظور بررسی دقیق‌تر نوع و نحوه برهم‌کنش‌های مولکولی و همچنین تأیید نتایج پیش‌بینی شده، مطالعات داکینگ مولکولی میان این ترکیبات و پروتئین هدف انجام گرفت.

جدول (۴-۲) پارامترهای محاسبه شده قانون لیپینسکی برای ترکیب ۳۰ و ۱۶ و ترکیب‌های پیشنهادی

سهولت سنتز	تعداد دهنده-های پیوند هیدروژنی	تعداد گیرنده-های پیوند هیدروژنی	Mlo gP	وزن مولکولی g/mol	پیوندهای قابل چرخش	شماره ترکیب
۳/۹۲	۲	۵	۵۴/۱	۴۴۶/۵۰	۸	۳۰
۳/۳۰	۲	۴	۳۶/۱	۳۳۴/۳۸	۴	۱۶
۳/۴۵	۲	۳	۵۵/۱	۵۰۵/۲۱	۴	G6
۳/۹۸	۲	۶	۶۵/۱	۵۰۳/۵۲	۷	G2

۴-۶- نتیجه گیری نهایی

در این فصل، فرآیند توسعه و ارزیابی یک مدل QSAR مبتنی بر روش رگرسیون انقباضی لاسو تطبیقی (ALASSO) برای پیش‌بینی فعالیت زیستی ترکیبات مورد مطالعه ارائه شد. بدین منظور، پس از بهینه‌سازی ساختارهای سب‌بعدی مولکول‌ها و محاسبه ۳۲۲۴ توصیف‌گر مولکولی توسط نرم‌افزار Dragon، مجموعه‌ای از مراحل پیش‌پردازش شامل حذف توصیف‌گرهای ثابت، کم‌واریانس و دارای همبستگی بالا انجام شد که در

نهایت تعداد متغیرها به ۳۸۶ توصیف‌گر کاهش یافت. سپس با استفاده از الگوریتم ALASSO، مؤثرترین ویژگی‌های مولکولی شناسایی شده و هفت توصیف‌گر به عنوان متغیرهای نهایی مدل انتخاب گردیدند. نتایج حاصل نشان داد که مدل ALASSO قادر است با بهره‌گیری از تعداد محدودی از توصیف‌گرهای منتخب، بخش عمده‌ای از تغییرات فعالیت زیستی ترکیبات را تبیین نماید. ارزیابی مدل بر روی مجموعه آزمون خارجی بیانگر دقت پیش‌بینی مطلوب آن بود، به‌گونه‌ای که ضریب تعیین بسیار بالای ($R^2=0.87$) و مقادیر پایین شاخص‌های خطا از جمله MAE و MSE نشان‌دهنده توانایی مناسب مدل در بازتولید داده‌های تجربی است. همچنین بررسی نمودارهای مقادیر پیش‌بینی‌شده در برابر مقادیر واقعی و تحلیل باقی‌مانده‌ها، عدم وجود خطای سیستماتیک و توزیع مناسب خطاها را تأیید نمود که بیانگر پایداری و اعتبار ساختاری مدل توسعه‌یافته است.

به‌منظور ارزیابی توان تعمیم‌پذیری مدل، اعتبارسنجی حذف تک‌نمونه‌ای (LOO-CV) نیز انجام شد. اگرچه مقدار Q_{LOO}^2 برابر ۰/۴۰ اندکی کمتر از مقدار ایده‌آل گزارش‌شده در برخی منابع بود، اما نتایج به‌دست‌آمده همچنان وجود یک رابطه ساختار-فعالیت معنادار را تأیید می‌کنند. با توجه به تعداد نسبتاً محدود نمونه‌ها و ماهیت پیچیده داده‌های زیستی، چنین عملکردی در مطالعات اکتشافی QSAR قابل قبول ارزیابی می‌شود. افزون بر این، نتایج آزمون Y-Randomization نشان داد که پس از تصادفی‌سازی مقادیر متغیر پاسخ، مدل قادر به ایجاد همبستگی معنادار نبوده و مقادیر ضریب تعیین در اغلب تکرارها به سمت صفر میل کرده‌اند. این موضوع نشان می‌دهد که عملکرد مدل نهایی حاصل روابط واقعی میان ساختار مولکولی و فعالیت زیستی بوده و ناشی از برازش تصادفی داده‌ها نیست.

در این تحقیق، مجموعه‌ای از ترکیبات نوین با هدف ارزیابی قابلیت بالقوه آن‌ها در بروز فعالیت زیستی طراحی شد. روند طراحی با توجه به اصول داروشناختی و با در نظر گرفتن الزامات قانون لیپینسکی انجام گرفت تا ترکیبات حاصل از نظر ویژگی‌های اولیه داروماند شرایط مناسبی داشته باشند. برای بررسی خصوصیات فیزیکوشیمیایی و برآورد میزان زیست‌دسترسی احتمالی هر ترکیب، از پایگاه SwissADME بهره گرفته شد و تمامی پارامترهای استخراج‌شده ثبت و مورد تحلیل قرار گرفتند. در گام بعد، به‌منظور

تخمین فعالیت زیستی ترکیبات طراحی شده، مدل پیش‌بینی لاسو به کار گرفته شد. پس از آن، مطالعات داکینگ مولکولی جهت ارزیابی نحوه اتصال این ترکیبات به پروتئین هدف انجام شد. در این بخش، کیفیت برهم‌کنش میان لیگاند و گیرنده و همچنین انرژی‌های اتصال محاسبه و برای ترکیبات مختلف مستند شد تا امکان مقایسه و تحلیل دقیق‌تر فراهم شود.

در مجموع، نتایج این فصل نشان داد که روش ALASSO ابزاری کارآمد برای انتخاب ویژگی و توسعه مدل‌های QSAR در داده‌های با ابعاد بالا است. مدل حاصل ضمن برخورداری از ساختاری نسبتاً ساده و قابل تفسیر، از توان پیش‌بینی مناسبی نیز برخوردار بوده و می‌تواند به عنوان یک ابزار کمکی ارزشمند در غربالگری اولیه ترکیبات و هدایت فرآیند طراحی دارو مورد استفاده قرار گیرد. با این حال، افزایش حجم داده‌های آموزشی، بررسی دقیق‌تر ترکیبات پرت و استفاده از مجموعه‌های داده گسترده‌تر می‌تواند در مطالعات آینده موجب بهبود بیشتر قدرت تعمیم و دقت پیش‌بینی مدل گردد.

۷-۴- آینده نگری

با توجه به نتایج به‌دست‌آمده در این پژوهش، می‌توان انتظار داشت که توسعه مدل‌های QSAR مبتنی بر روش‌های انتخاب متغیر انقباضی، نقش مهمی در تسریع فرآیند طراحی دارو و غربالگری مجازی ترکیبات ایفا کند. در مطالعات آینده، افزایش تعداد ترکیبات موجود در مجموعه داده و گسترش دامنه تنوع ساختاری مولکول‌ها می‌تواند به بهبود قابلیت تعمیم و پایداری مدل منجر شود. همچنین بررسی و مقایسه عملکرد ALASSO با سایر روش‌های پیشرفته انتخاب ویژگی نظیر Elastic Net، SCAD و غیره می‌تواند در شناسایی ساختارهای مؤثرتر و افزایش دقت پیش‌بینی مفید باشد. از سوی دیگر، ترکیب روش‌های انتخاب ویژگی مبتنی بر ALASSO با مدل‌های غیرخطی نظیر شبکه‌های عصبی مصنوعی، ماشین بردار پشتیبان و روش‌های یادگیری عمیق می‌تواند روابط پیچیده‌تر ساختار-فعالیت را آشکار سازد. علاوه بر این، انجام مطالعات داکینگ مولکولی، دینامیک مولکولی و بررسی ناحیه کاربرد در کنار مدل‌های QSAR، می‌تواند اعتبار زیستی نتایج را افزایش داده و زمینه را برای طراحی منطقی مشتقات جدید با فعالیت دارویی بالاتر فراهم نماید. در نهایت، توسعه مدل‌های ترکیبی و استفاده از داده‌های بزرگ مولکولی، چشم‌انداز

امیدوارکننده‌ای برای دستیابی به سامانه‌های پیش‌بینی دقیق‌تر و قابل اعتمادتر در حوزه طراحی دارو ترسیم می‌کند.

۸-۴- پیوست

شماره ترکیب	EEig05x	Mor27m	Mor28m	G1e	E1p	H5m	H6p
۱	3.269	0.015	-0.097	0.157	0.555	0.293	0.199
۲	3.744	-0.167	-0.042	0.149	0.5	0.259	0.23
۳	3.386	0.161	0.074	0.155	0.534	0.222	0.29
۴	3.386	0.147	0.037	0.155	0.531	0.237	0.289
۵	3.386	0.215	0.21	0.155	0.563	0.254	0.307
۶	3.386	-0.087	0.141	0.155	0.582	0.296	0.316
۷	3.386	0.133	0.108	0.152	0.584	0.2	0.301
۸	3.74	0.14	-0.171	0.152	0.503	0.222	0.224
۹	3.479	0.152	-0.124	0.153	0.489	0.2	0.241
۱۰	3.448	0.161	0.157	0.167	0.51	0.198	0.275
۱۱	3.542	0.229	-0.01	0.15	0.517	0.197	0.265
۱۲	3.723	0.023	0.248	0.164	0.508	0.28	0.324
۱۳	3.642	0.022	0.004	0.163	0.493	0.2	0.239
۱۴	3.517	0.056	0.072	0.165	0.495	0.183	0.229
۱۵	3.684	0.142	0.09	0.163	0.551	0.249	0.244
۱۶	3.386	0.202	-0.129	0.173	0.457	0.195	0.206
۱۷	3.386	0.195	0.051	0.157	0.493	0.31	0.184
۱۸	3.531	0.073	0.126	0.16	0.532	0.205	0.26
۱۹	3.651	0.081	0.045	0.157	0.54	0.206	0.282
۲۰	3.651	0.119	-0.06	0.149	0.535	0.207	0.275
۲۱	3.656	0.093	-0.112	0.159	0.6	0.207	0.36
۲۲	3.656	0.088	-0.115	0.16	0.525	0.182	0.263
۲۳	3.719	-0.047	0.134	0.146	0.624	0.317	0.381
۲۴	3.744	-0.03	-0.053	0.154	0.625	0.347	0.403
۲۵	3.757	0.017	-0.017	0.146	0.549	0.291	0.438
۲۶	3.659	0.17	0.061	0.162	0.525	0.263	0.311
۲۷	3.684	0.162	0.044	0.152	0.523	0.281	0.311
۲۸	3.684	0.223	0.172	0.152	0.552	0.3	0.326
۲۹	3.684	0.044	0.12	0.152	0.57	0.348	0.335
۳۰	3.771	-0.009	0.157	0.153	0.626	0.318	0.435
۳۱	3.728	0.143	0.143	0.15	0.619	0.278	0.41
۳۲	3.731	0.158	0.003	0.156	0.608	0.256	0.387

0.456	0.416	0.648	0.165	0.099	-0.029	3.78	۳۳
0.418	0.353	0.611	0.151	-0.036	0	3.797	۳۴
0.428	0.367	0.507	0.151	-0.015	0.035	3.807	۳۵
0.322	0.288	0.57	0.157	0.296	0.194	3.386	۳۶
0.24	0.334	0.599	0.157	0.343	-0.226	3.386	۳۷

ادامه جدول پیوست

H6p	H5m	E1p	G1e	Mor28m	Mor27m	EEig05x	شماره ترکیب
0.3	0.442	0.683	0.149	0.204	0.028	3.723	۳۸
0.233	0.2	0.437	0.153	0.111	0.081	3.653	۳۹
0.255	0.199	0.467	0.15	0.084	0.016	3.654	۴۰
0.246	0.2	0.457	0.153	0.201	-0.261	3.654	۴۱
0.314	0.254	0.518	0.146	0.148	-0.127	3.727	۴۲
0.305	0.237	0.506	0.155	0.003	0.026	3.661	۴۳
0.372	0.309	0.474	0.144	-0.063	-0.011	3.762	۴۴
0.332	0.357	0.498	0.144	-0.188	-0.083	3.763	G1
0.408	0.566	0.533	0.166	0.018	-0.336	3.909	G2
0.193	0.212	0.481	0.177	-0.098	-0.177	3.345	G3
0.199	0.176	0.478	0.155	0.053	-0.009	3.345	G4
0.193	0.249	0.469	0.166	-0.118	-0.179	3.371	G5
0.303	0.687	0.696	0.153	0.641	-0.905	3.517	G6
0.226	0.207	0.583	0.151	-0.264	0.068	3.671	G7

- [1] Wold, S. (1995). Chemometrics; what do we mean with it, and what do we want from it?. *Chemometrics and intelligent laboratory systems*, 30(1), 109-115.
- [2] Brereton, R. G. (2009). *Chemometrics for pattern recognition*. John Wiley & Sons.
- [3] Otto, M. (2023). *Chemometrics: statistics and computer application in analytical chemistry*. John Wiley & Sons.
- [4] Todeschini, R., & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics*. Wiley-VCH.
- [5] Hitaoka, S., & Chuman, H. (2013). Revisiting the Hansch–Fujita approach and development of a fundamental QSAR. *Journal of Pesticide Science*, 38(2), 60-67.
- [6] Cherkasov, A., Muratov, E.N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y.C. and Todeschini, R. (2014). QSAR modeling: where have you been? Where are you going to? *Journal of medicinal chemistry*, 57(12): 4977-5010.
- [7] Golbraikh, A., Wang, X.S., Zhu, H. and Tropsha, A. (2016). Predictive QSAR modeling: methods and applications in drug discovery and chemical risk assessment. *Handbook of computational chemistry*: 1-48.
- [8] 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): 1218-1227.
- [9] 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): 231-243.
- [10] لواين. ايران (۱۳۸۷)، "شیمی کوانتومی" اسلامپور غ، پارسا فر غ، مقاری ع، نجفی ب، جلد سوم، انتشارات فاطمی.
- [11] Abraham MH., Ibrahim A., Zissimos AM. (2004) "Determination of sets of solute descriptors from chromatographic measurements", *J Chromatogr A*. 1037, (1-2), pp 29-47
- [12] Tropsha A. (2010) "Best Practices for QSAR Model Development, Validation, and Exploitation", *Mol. Inf.* 29, pp 476 – 488
- [13] Mozafari, Z., Chamjangali, M. A., Arashi, M., & Goudarzi, N. (2021). Suggestion of active 3-chymotrypsin like protease (3CLPro) inhibitors as potential anti-SARS-CoV-2 agents using predictive QSAR model based on the combination of ALASSO with an ANN model. *SAR and QSAR in Environmental Research*, 32(11), 863-888.
- [14] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
- [15] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.
- [16] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R* (Vol. 103). New York: springer.
- [17] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). Linear model selection and regularization. In *An introduction to statistical learning: with applications in R* (pp. 225-288). New York, NY: Springer US.
- [18] Zhao, Q., & Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1), 272-281.

- [19] Zou H (2006) The adaptive lasso and its oracle properties. *Journal of the American statistical association* 101:1418-1429.
- [20] Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. routledge.
- [21] Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (Vol. 326). John Wiley & Sons.
- [22] Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 109–130.
- [23] Wold, S., Eriksson, L., & Kettaneh, N. (2009). PLS in data mining and data integration. In *Handbook of partial least squares: Concepts, methods and applications* (pp. 327-357). Berlin, Heidelberg: Springer Berlin Heidelberg.
- [24] Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, 30(1), 79-82.
- [25] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geoscientific model development*, 7(3), 1247-1250.
- [26] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.
- [27] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4), 679-688.
- [28] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2), 111-133.
- [29] Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (Vol. 326). John Wiley & Sons.
- [30] Rücker, C., Rücker, G., & Meringer, M. (2007). y-Randomization and its variants in QSPR/QSAR. *Journal of chemical information and modeling*, 47(6), 2345-2357.
- [31] Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry*, 30(16), 2785-2791.
- [32] Mathur, A., Meena, A., & Luqman, S. (2025). LIM Kinase's role in cancer cytoskeleton remodelling and metastasis. *International Journal of Biological Macromolecules*, 146677.
- [33] M.-H. Lee, J.K. Kundu, J.-I. Chae, J.-H. Shim, Targeting ROCK/LIMK/cofilin signaling pathway in cancer, *Arch. Pharm. Res.* 42 (6) (2019) 481–491.
- [34] A. Mathur, A. Meena, S. Luqman, Monoterpenoids: An upcoming class of therapeutic agents for modulating cancer metastasis, *Phytother. Res.* 38 (2) (2023) 939–969.
- [35] Balupuri, A., Balasubramanian, P. K., & Cho, S. J. (2015). 3D-QSAR study of tetrahydro-3 H-imidazo [4, 5-c] pyridine derivatives as VEGFR-2 kinase inhibitors using various charge schemes. *Archives of pharmacal research*, 38(8), 1434-1442.
- [36] Algamil, Z. Y., Lee, M. H., & Al-Fakih, A. M. (2016). High-dimensional quantitative structure–activity relationship modeling of influenza neuraminidase a/PR/8/34 (H1N1) inhibitors based on a two-stage adaptive penalized rank regression. *Journal of Chemometrics*, 30(2), 50-57.
- [37] Algamil, Z. Y., & Lee, M. H. (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), 75-90.
- [38] Datta, S., Dev, V. A., & Eden, M. R. (2017). Developing QSPR for predicting DNA drug binding affinity of 9-Anilinoacridine derivatives using correlation-based adaptive LASSO algorithm. In *Computer Aided Chemical Engineering* (Vol. 40, pp. 2767-2772).

Elsevier.

[39] Datta, S., Dev, V. A., & Eden, M. R. (2019). Using correlation based adaptive LASSO algorithm to develop QSPR of antitumour agents for DNA–drug binding prediction. *Computers & Chemical Engineering*, 122, 258-264.

[40] Froimowitz, M. (1993). HyperChem: a software package for computational chemistry and molecular modeling. *Biotechniques*, 14(6), 1010-1013.

[41] Wolinski, K., Hinton, J. F., Wishart, D. S., Sykes, B. D., Richards, F. M., Pastone, A. ... & Hayes, D. M. (2007). HyperChem Computational Chemistry Software Package, Version 8.0.

[42] Mauri, A., Consonni, V., Pavan, M., & Todeschini, R. (2006). Dragon software: An easy approach to molecular descriptor calculations. *Match*, 56(2), 237-248.

[43] Sen, S., & Yildirim, I. (2022). A Tutorial on How to Conduct Meta-Analysis with IBM SPSS Statistics. *Psych* 2022, 4, 640–667.

[44] Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3): 299-314.

[45] Valentine, D.T. and Hahn, B.H. 2022. *Essential MATLAB for engineers and scientists*: Academic Press.

[46] Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., & Olson, A. J. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of computational chemistry*, 19(14), 1639-1662.

[47] Todeschini, R., & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics* (Vols. 1–2). Wiley-VCH, Weinheim, Germany.

[48] Champiré, A., Berabez, R., Braka, A., Cosson, A., Corret, J., Girardin, C., ... & Routier, S. (2024). Tetrahydropyridine LIMK inhibitors: Structure activity studies and biological characterization. *European Journal of Medicinal Chemistry*, 271, 116391.

[49] Mozafari, Z., Chamjangali, M. A., Arashi, M., & Goudarzi, N. (2022). Application of the LAD-LASSO as a dimensional reduction technique in the ANN-based QSAR study: Discovery of potent inhibitors using molecular docking simulation. *Chemometrics and Intelligent Laboratory Systems*, 222, 104510.

[50] Mozafari, Z., Chamjangali, M. A., & Arashi, M. (2020). Combination of least absolute shrinkage and selection operator with Bayesian Regularization artificial neural network (LASSO-BR-ANN) for QSAR studies using functional group and molecular docking mixed descriptors. *Chemometrics and Intelligent Laboratory Systems*, 200, 103998.

[51] Lotfi, M., Chamjangali, M. A., & Mozafari, Z. (2023). 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*, 232, 104714.

[52] Lotfi, M., Chamjangali, M. A., & Mozafari, Z. (2023). 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*, 232, 104714.

[53] Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Application of Biased Estimation to Regression Problems. *Technometrics*, 12(1), 69–82.

[54] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.

[55] Netzeva, T. I., Worth, A. P., Aldenberg, T., Rusyn, I., & Jeliaskova, N. (2005). Use of statistical cross-validation in the development of QSAR models. European Chemicals Bureau (ECB).

[56] Sepehri, B., Kohnehpoushi, M., & Ghavami, R. (2022). High predictive QSAR models for predicting the SARS coronavirus main protease inhibition activity of ketone-based covalent inhibitors. *Journal of the Iranian Chemical Society*, 19(5), 1865-1876.

[57] Garro Martinez, J. C., Duchowicz, P. R., Estrada, M. R., & Castro, E. A. (2012). A multivariate QSAR study on the anticonvulsant activity of acetamido-N-benzylacetamide derivatives. Influence of different molecular descriptors.

[58] Aminaee, H., Hosseini, S., Davood, A., & Askarizadeh, E. (2022). QSAR study of indole derivatives as active agents against *Candida albicans*: a DFT calculation. *Journal of Receptors and Signal Transduction*, 42(6), 614-622.

[59] Ajala, A., Uzairu, A., Shallangwa, G. A., & Abechi, S. E. (2023). QSAR, simulation techniques, and ADMET/pharmacokinetics assessment of a set of compounds that target MAO-B as anti-Alzheimer agent. *Future Journal of Pharmaceutical Sciences*, 9(1), 4.

Abstract

Protein kinases, particularly LIM kinase (LIMK), which plays a crucial role in cytoskeletal regulation and cancer metastasis, have emerged as important therapeutic targets in modern drug discovery. Considering the significant role of LIMK in cancer cell invasion and migration, the development of selective LIMK inhibitors represents a promising strategy for the discovery of novel anticancer therapies. Tetrahydropyridine derivatives, owing to their structural flexibility and ability to establish diverse molecular interactions, have demonstrated considerable potential as LIMK inhibitors. However, the experimental determination of the inhibition constant (K_i) of these compounds is both time-consuming and costly. In the present study, Quantitative Structure–Activity Relationship (QSAR) modeling combined with molecular docking was employed as an efficient computational approach for the rapid prediction of the inhibitory activity of tetrahydropyridine derivatives and for the identification of potential LIMK inhibitors. The Adaptive Least Absolute Shrinkage and Selection Operator (ALASSO) was utilized for both molecular descriptor selection and QSAR model development. By selecting the most informative descriptors and eliminating redundant variables, ALASSO enabled the construction of a sparse, interpretable, and highly predictive QSAR model while reducing the risk of overfitting. Application of the ALASSO algorithm resulted in a final QSAR model containing seven molecular descriptors. The predictive performance of the model was evaluated using an external test set, leave-one-out (LOO) cross-validation, and several statistical validation criteria. The model achieved an MSE of 0.10 and an R^2 value of 0.87 for the external test set, while the Q^2 value obtained from LOO cross-validation confirmed the robustness and predictive capability of the developed model. Furthermore, the resulting sparse model, consisting of only seven descriptors, demonstrated satisfactory predictive accuracy. The obtained R^2 and Q^2 values exceeded the generally accepted threshold values of 0.60 and 0.50, respectively, indicating that the proposed QSAR model is statistically reliable and possesses good predictive ability. The developed QSAR model was subsequently applied to predict the biological activity of newly designed compounds. The predicted compounds were further evaluated by molecular docking, and the docking results supported the reliability of the QSAR predictions. Overall, this study overcomes several limitations of conventional QSAR models and provides an effective computational framework for the rational design and optimization of novel tetrahydropyridine derivatives as potent LIMK inhibitors.

Keywords: LIM-kinase (LIMK), Tetrahydropyridine derivatives, Adaptive LASSO (ALASSO), Cancer, Protein kinase inhibitors



Shahrood University of Technology
Faculty of Chemistry
Master Dissertation in Analytical Chemistry

**Prediction of the inhibitory constant (K_i) of some
novel tetrahydropyridine derivatives as LIM kinase
(LIMK) inhibitors using Adaptive Least Absolute
Shrinkage and Selection Operator (ALASSO)**

By:
Sultan Nafe Sultan

Supervisor:
Dr. Mansour Arab Chamjangali