دانشکده شیمی

رساله دکتری شیمی تجزیه

به‌کارگیری تکنیک‌های رگرسیونی جریمه‌شده به‌عنوان روش‌های جدید انتخاب متغیر در مطالعات روابط کمی ساختار-فعالیت (QSAR) و ساختار- خاصیت (QSPR)

نگارنده:

زینب مظفری

استاد راهنما:

دکتر منصور عرب چم جنگلی

اساتید مشاور:

دکتر محمد آرشی

دکتر ناصر گودرزی

اسفند ۱۴۰۰

به پاس تعبیر عظیم و انسانی‌شان از کلمه ایثار ، به پاس عاطفه سرشار و گرمای امیدبخش وجودشان که در این سردترین روزگاران بهترین پشتیبان است، به پاس قلب‌های بزرگشان که فریادرس است و به پاس محبت‌های بی دریغشان که هرگز فروکش نمی‌کند ،

این رساله را تقدیم می کنم :

به استوارترین تکیه گاهم ،دستان پرمهر پدرم

به همراه ترین نگاه زندگیم ، چشمان پر امید مادرم

به بهترین هم بازی دوران کودکی و بهترین حامی دوران بزرگسالی، برادر عزیزم، ابراهیم

به انگیزه، امید و چراغ همیشه تابان زندگیم ، خواهر نازنینم ، فروزان

به نو ورود باغ زندگیمان، مژگان

و به تمام کسانی که نیروی حرکت به سوی مقصود را در وجود من زنده کردند

هرچه آموختم در مکتب عشق شما آموختم و هرچه بکوشم قطره‌ای از دریای بی

کران مهربانی‌تان را سپاس توانم بگویم.

امروز هستی‌ام به امید شماست و فردا کلید باغ بهشتم رضای شما

ره آوردی گران سنگ‌تر از این رساله نداشتم تا به خاک پایتان نثار کنم، باشد که

حاصل تلاشم نسیم گونه غبار خستگی‌تان را بزداید.

بوسه بر دستان پرمهرتان، خداپشت و پناهتان

قدر دانی و تشکر:

حمد و ثنای خدای متعال را که به فضلش آدمی را اندیشه و عمل عطا فرمود تا بکاود و بداند و از علم و آگاهیش چراغی برافروزد و هستی خویش را روشنی بخشد. پس از ارادت خاضعانه به درگاه خداوند بی همتا لازم است از انسانهای عاشق همواره در تفکر و شناخت ، آنهائی که زندگی خویش را وقف علم و بشریت کردند قدردانی نمایم . در اینجا لازم میدانم که از زحمات و راهنمائی های خردمندانه وبی دریغ استاد راهنمای گرانقدرم، جناب **آقای دکتر منصور عرب** تقدیر و تشکر نمایم، که موفقیت های پژوهشی و آموزشی بنده حاصل زحمات ایشان در باروری و بروز رسانی اطلاعات اینجانب می باشد. در تدوین این رهاورد خود را مدیون راهنمایی و رهنمودهای ارزنده و مساعدت صمیمانه استاد مشاور ارجمندم، جناب آقای **دکتر محمد آرشی** می دانم ، که نقش بسزایی را در به پایان رساندن و موفقیت این اثر داشته اند. به راستی که انجام این رساله بدون نظرهای صائب، پیگیری های دلسوزانه و تشویق های امید بخش این دو عزیز میسّر نبود. تشکر و قدردانی خود را از جناب آقای **دکتر ناصر گودرزی** استاد و مشاور ارجمند اعلام می دارم که الطاف بی شائبه ایشان در طول مدت تحصیل شامل حال بنده شد. از استاد بزرگوارم **جناب آقای دکتر قدمعلی باقریان**، برای تمام زحمات بی دریغشان و زحمت داوری رساله ، ممنونم. از اساتید بزرگوار، **جناب آقای دکتر فاطمی و جناب آقای دکتر روزبه** که زحمت داوری و تصحیح این رساله را به عهده داشتند کمال سپاس را دارم. از خداوند بزرگ توفیق، طول عمر، سلامتی و موفقیت را برای همه اساتید بزرگوار خواهانم .

از همکلاسی، هم اتاقی، هم دانشگاهی، هم آزمایشگاهی، دوست و خواهر عزیزم خانم **بهاره عربخانی** که در طی این سال ها همراه همیشگی روزهای سخت و خاطره‌ساز ایام خوش در شهر غربت بود ،صمیمانه تقدیر می کنم و برای ایشان از خداوند بزرگ عاقبت به خیری و موفقیت آرزومندم.

از همراهان همیشگی ام **خانم غزاله نجفی عرب، خانم هانیه قنبری، خانم نسرین محمدوست، خانم آرزو افروخه، خانم دکتر اشرفی ، خانم دکتر دوستی ، خانم محدثه لطفی ، آقای مهندس داوود نادعلی ، آقای مهندس مؤمنی و آقای مهندس یزدانی** که اوقات خوشی را در کنار هم سپری کرده‌ایم و در این مدت، همیشه بنده را مورد لطف و مرحمت خود قرار داده‌اند، کمال تقدیر و تشکر را دارم. در نهایت بر خود لازم می‌دانم تا از دوستان عزیزم در آزمایشگاه **شیمی تجزیه**، شیمی آلی و شیمی معدنی، اعضای هیات علمی دانشکده شیمی به خصوص **آقای دکتر بهرامیان، آقای دکتر میرزایی ، سرکار خانم دکتر مصدرالامور و سرکار خانم دکتر کلاتر**، کارشناسان و کارکنان دانشکده شیمی، و کارشناسان تحصیلات تکمیلی به خصوص **آقای مهندس خانعلی زاده**، که در دوره ۷ ساله تحصیلی (کارشناسی ارشد و دکتری) همواره اینجانب را راهنمایی و همراهی نموده‌اند، سپاسگزارم.

تعهد نامه

اینجانب **زینب مظفری** دانشجوی دوره دکتری **شیمی تجزیه** دانشکده **شیمی** دانشگاه صنعتی شاهرود، نویسنده رساله **به‌کارگیری تکنیک‌های رگرسیونی جریمه‌شده به‌عنوان روش‌های جدید انتخاب متغیر در مطالعات روابط کمی ساختار-فعالیت (QSAR) و ساختار- خاصیت (QSPR)** تحت راهنمایی **آقای دکتر منصور عرب چم جنگلی** و مشاوره **آقای دکتر محمد آرشی** و **آقای دکتر ناصر گودرزی** متعهد می‌شوم:

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

تاریخ امضای دانشجو

چکیده

اهداف مهم این رساله، استفاده از روش‌های انتخاب متغیر جریمه‌شده برای انتخاب مؤثرترین توصیف کننده‌ها و جفت کردن روش‌های جریمه‌شده با روش مدل‌سازی غیر خطی شبکه عصبی مصنوعی (ANN) است. در این راستا، از روش‌های مختلفی برای ایجاد ارتباط بین توصیف کننده‌های منتخب و پاسخ هدف استفاده شد. بنابراین، در مطالعه اول از ترکیب انحراف قدر مطلق به‌طور هموار بریده شده (SCAD) و ANN به عنوان یک رویکرد جدید (SCAD-LM-ANN) در مطالعات کمی ساختار-فعالیت (QSAR) استفاده شد. SCAD-LM-ANN از مزایای ذاتی مفید روش انقباضی SCAD در کاهش داده‌هایی با ابعاد بالا قبل از روش مدل‌سازی استفاده می‌کند. عملکرد مدل SCAD-LM-ANN با ایجاد ارتباط بین توصیف کننده‌های به‌دست‌آمده از دراگون و فعالیت‌های دارویی برای مجموعه‌ای از مشتقات تیواستامید/استانیلید به عنوان مهارکننده‌های HIV مورد بررسی قرار گرفت. روش SCAD با اجرای ارزیابی تقاطعی ۱۰-fold بر روی مجموعه داده‌ها در غیاب ترکیبات مجموعه آزمون، اجرا شد. ۱۱ توصیف کننده در λ با کم‌ترین خطای ارزیابی تقاطعی (λ_{min}) انتخاب شدند. توصیف کننده‌های منتخب به‌عنوان ورودی ANN مورد استفاده قرار گرفتند. تمام پارامترهای مؤثر بر عملکرد مدل، بهینه شدند و مدل SCAD-LM-ANN با معماری ۱-۵-۵ به‌عنوان مدل QSAR بهینه انتخاب شد. چندین پارامتر آماری برای ارزیابی مدل در نظر گرفته شد. نتایج حاصل، تعمیم پذیری و قدرت پیش بینی مدل SCAD-LM-ANN پیشنهادی را اثبات می‌کند. با توجه به رابطه ایجاد شده در مدل QSAR برتر، مشتقات جدیدی طراحی و به عنوان مهارکننده‌های فعال HIV برای مطالعات بیش‌تر پیشنهاد شدند. صحت ترکیبات پیشنهادی با تجزیه و تحلیل برهم‌کنش‌های گیرنده- لیگاند (LR) به‌دست آمده از مطالعات داکینگ مولکولی، مورد مطالعه و تأیید قرار گرفت. در بخش دوم این رساله، رویکرد جدیدی به عنوان ترکیبی از حداقل قدر مطلق انقباض و عملگر انتخاب کننده سازگار (ALASSO) و ANN برای ساخت مدل مهارکننده‌های پروتئاز ۳-شبه کیموتریپسین ($3CL^{pro}$) به‌عنوان ترکیبات قوی SARS CoV-2 معرفی شد. با توجه به اهمیت این بیماری از سال ۲۰۱۹، توصیه و طراحی ترکیبات فعال جدید بسیار مهم است. پس از ارزیابی صحت و اعتبار مدل توسعه‌یافته ALASSO-LM-ANN، ترکیبات جدیدی با استفاده از توصیف کننده‌های مؤثر پیشنهاد شدند و فعالیت دارویی ترکیبات جدید پیش‌بینی شد. بررسی برهم‌کنش‌های LR نیز با استفاده از مطالعه داکینگ مولکولی انجام شد. خواص PK و قانون

پنج لیپینسکی برای تمام ترکیبات پیشنهادی محاسبه شد و ترکیبات جدید پیشنهادی دارای خواص دارویی قابل قبولی هستند. در مطالعه سوم، ترکیب حداقل انحراف مطلق - حداقل قدر مطلق انقباض و عملگر انتخاب کننده (LAD-LASSO) به‌عنوان یک روش انتخاب متغیر جدید برای مطالعات QSAR مبتنی بر ANN معرفی شد. مدل ANN همراه با روش انتخاب متغیر کارآمد LAD-LASSO برای پیش‌بینی فعالیت دارویی سه مجموعه داده از ترکیبات شیمیایی، مورد ارزیابی قرار گرفت. از روش انتخاب متغیر LAD-LASSO استفاده شد و توصیف کننده‌هایی که بیشترین ارتباط را با فعالیت‌های دارویی داشتند انتخاب شدند. توصیف کننده‌های منتخب به‌عنوان ورودی ANN تعریف شدند و مدل‌های طراحی شده، بهینه شدند. فعالیت دارویی ترکیبات مجموعه آموزش با استفاده از مدل‌های بهینه ANN پیش‌بینی شد. ضرایب تعیین (R^2) برای داده‌های آزمون در سه مجموعه داده برابر با ۰/۸۷، ۰/۸۴ و ۰/۸۷ بود. دامنه کاربرد و آزمون Y-تصادفی نیز کارایی مدل‌های توسعه یافته را ثابت کردند. در نهایت، مدل‌های QSAR ایجاد شده برای پیشنهاد ترکیبات شیمیایی جدید و قوی با اصلاح ساختاری مولکول‌های ضعیف در هر سه مجموعه داده مورد استفاده قرار گرفتند. مقادیر پاسخ ترکیبات پیشنهادی جدید با استفاده از مدل‌های ANN بهینه پیش‌بینی شدند. بر اساس اطلاعات LR، وجود برهم‌کنش‌های مختلف آب‌دوست و آب‌گریز در جایگاه فعال گیرنده نشان‌دهنده پتانسیل بالای ترکیبات شیمیایی در برقراری اتصال پایدار است. در آخرین کار این مطالعه، ترکیبی از SCAD و ANN در رابطه کمی ویژگی ساختاری- شاخص‌های بازداری (QSRR) (RIs) استفاده شد. روش SCAD پیشنهادی قبل از استفاده از روش قدرتمند مدل‌سازی ANN، ابعاد داده‌ها را کاهش می‌دهد. کارایی روش‌های SCAD-ANN با ساخت یک مدل QSRR بین مؤثرترین توصیف کننده‌های مولکولی و RI برای دو مجموعه از ترکیبات آلی فرار (VOCs) ارزیابی شد. روش SCAD برای داده‌های آموزشی اعمال شد و توصیف کننده‌های مؤثر در λ_{min} انتخاب شدند و به‌عنوان ورودی‌های روش مدل‌سازی ANN تعریف شدند. تمام پارامترهای ANN به‌طور هم‌زمان بهینه شدند. نتایج به‌دست‌آمده نشان می‌دهد که مدل‌های QSRR ساخته‌شده از قدرت پیش بینی قابل قبولی برخوردار است.

کلمات کلیدی: QSAR، QSPR، سرطان، HIV، Coronavirus، SCAD، ALASSO، LAD-LASSO، ANN، اتصال مولکولی

مقالات چاپ شده در مجلات خارجی JCR

1- **Mozafari, Z**., Arab Chamjangali, M., Arashi, M., & Goudarzi, N. (2021). Performance of smoothly clipped absolute deviation as a variable selection method in the artificial neural network- based QSAR studies. Journal of Chemometrics, e3338.

2- **Mozafari, Z**., Chamjangali, M. A., Arashi, M., & Goudarzi, N. (2021). Suggestion of active 3-chymotrypsin like protease ($3CL^{Pro}$) 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.

3- **Mozafari, Z**., Chamjangali, M. A., Arashi, M., & Goudarzi, N. (2021). QSRR models for predicting the retention indices of VOCs in different datasets using an efficient variable selection method coupled with artificial neural network modeling: ANN-based QSPR modeling. journal of iranian chemical society. Accepted manuscript on 18 Dec. 2021.

4- **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, 104510.

مقالات ارائه شده در کنفرانس‌های معتبر داخلی به شکل سخنرانی و پوستر

5. **Mozafari, Z**., Arab Chamjangali, M., Arashi, M., & Goudarzi, N. Hybrid QSPR models for the prediction of the linear retention index of volatile compounds in flour, 7th Iranian Biennial Chemometrics Seminar, 30-31 Oct. 2019, Shahrood University of Technology, Shahrood, Iran, (**Poster**)

6. **Mozafari, Z**., Arab Chamjangali, M., Arashi, M., & Goudarzi, N. Application of a new hybrid of SCAD - artificial neural network in QSAR study of HIV inhibitors, 7th Iranian Biennial Chemometrics Seminar, 30-31 Oct. 2019, Shahrood University of Technology, Shahrood, Iran, (**Oral**)

7. **Mozafari, Z**., Arab Chamjangali, M., Arashi, M., & Goudarzi, N. Application of the LAD-LASSO as a dimensional reduction technique in the ANN-based QSAR study: Discovery of potent inhibitors using molecular docking simulation, 10th Theoretical and Computational Chemistry Workshop, 23 – 25 Nov 2021,Sharif University of Technology, Tehran, Iran, (**Oral**)

فهرست مطالب

فهرست شکل‌ها:

فهرست جدول‌ها:

فهرست رابطه‌ها:

فهرست کلمات اختصاری

QSAR (Quantitative structure- activity relationships)
QSPR (Quantitative structure- properties relationships)
CADD (Computer Aided Drug Design)
RI (Retention index)
OECD (The Organization for Economic Co-operation and Development)
KS (Kennard-Stone)
LASSO (Least absolute deviation- Least absolute shrinkage and selection operator)
LAD (Least Absolute Deviations)
SCAD(Smoothly Clipped Absolute Deviation)
VIF (Variance inflation factor)
PCR (Principal component regression)
PLS (Partial least square)
ANN (Artificial neural network)
LM (Levenberg –Marquardt)
BR (Bayesian regularization)
LOO (Leave-one-out)
AD (Applicability Domain)
CV (Cross validation)
RMSD (Root-mean-square deviation)
SARS (Severe acute respiratory syndrome coronavirus)
MERS (Middle East respiratory syndrome coronavirus)
LAD- LASSO (Least absolute deviation (LAD)- least absolute shrinkage and selection operator (LASSO))
VOCs (Volatile organic compounds)
OLS (Ordinary least squares)
RMSE (Root mean square error)
LASSO (Least Absolute Shrinkage and Selection Operator)
ALASSO (Adaptive least Absolute Shrinkage and Selection Operator)
MCP (Minimax concave penalty)
SPSS (Statistical Package for Social Science)
SMILES (Simplified molecular-input line-entry system)
SIS (Sure Independence Screening)
ISIS (Iterative Sure Independence Screening)

مقدمه

۱- ۱ مقدمه

یکی از مشکلاتی که انسان همیشه با آن مواجه است، مقابله و درمان انواع بیماری‌هایی است که زندگی بشریت را با خطر روبرو می‌کند و همواره یکی از مهم‌ترین مسائل چالش‌برانگیز دانشمندان، توسعه داروهای مؤثر برای رفع و یا کاهش اثرات بیماری‌ها می‌باشد. بیماری‌هایی مانند ایدز، انواع سرطان و مقاومت ویروس‌ها در برابر انواع داروها و جهش‌های متنوعی که ایجاد می‌شود، منجر به تلاش مستمر محققین برای یافتن داروهای مؤثر و کارآمد در مواجهه با بیماری‌ها می‌شود.

توسعه دارو در گذشته به روش‌های آزمون و خطا انجام می‌شده است که به دلایل صرف زمان و هزینه مضاعف همواره دردسر ساز بوده است. علاوه بر این دانشمندان با توجه به مسائلی همچون عدم آگاهی اولیه از امکان سنتز و نتایج آزمایشگاهی دال بر یک سنتز موفق، عدم اطلاع از میزان فعالیت دارویی ترکیبات بالقوه و در دسترس نبودن امکانات انجام آنالیزهای متفاوت، همیشه در مسیر توسعه داروها با مشکلات عدیده‌ای روبرو می‌شوند. از جمله مهم‌ترین اهداف محققین و شیمیدان‌ها می‌توان به پیش‌بینی فعالیت دارویی ترکیبات قبل از سنتز اشاره کرد. زیرا با تحقق این هدف، در بخش عمده‌ای از هزینه‌ها، زمان و به‌کارگیری نیروی انسانی صرفه‌جویی خواهد شد.

علاوه بر مشکلات اشاره شده در توسعه داروها، پیش‌بینی خواص مربوط به ترکیبات شیمیایی نیز همواره مورد توجه محققین و شیمیدان‌ها بوده است. به‌طوری‌که با توجه به نبود امکانات کامل و جامع آزمایشگاهی و به مخاطره نیافتن طبیعت، تعیین مقدار برخی از خواص شیمیایی و فیزیکی ترکیبات نیز دارای اهمیت منحصر به‌فردی می‌باشد. بنابراین استفاده از روش‌های تئوری و شیمی محاسباتی توانسته است دانشمندان را در تحقق این هدف همراهی کند، تا بدون انجام آزمایش‌های متعدد فعالیت دارویی ترکیبات قبل از سنتز پیش‌بینی شود. علاوه بر این با به‌کارگیری شیمی محاسباتی، پیش‌بینی خواص فیزیکوشیمیایی ترکیبات نیز امکان‌پذیر می‌باشد.

شیمی نظری محاسباتی[1] اساساً به محاسبه عددی ساختارهای مولکولی و برهمکنش‌های مولکولی می‌پردازد. اصطلاح شیمی محاسباتی معمولاً زمانی استفاده می‌شود که یک روش ریاضی به اندازه کافی توسعه یافته باشد که بتوان آن را برای پیاده‌سازی بر روی رایانه، خودکار کرد. شیمی محاسباتی استفاده از مهارت‌های شیمی، ریاضی و رایانه برای حل مسائل جالب شیمیایی است. از رایانه‌ها برای تولید اطلاعاتی مانند ویژگی‌های مولکول‌ها یا شبیه‌سازی برهم‌کنش ترکیبات استفاده می‌شود. شیمی محاسباتی، روشی مفید برای بررسی مواد شیمیایی است که یافتن آن‌ها بسیار دشوار یا خرید آن‌ها بسیار گران است. همچنین به شیمیدانان کمک می‌کند قبل از اجرای آزمایش‌های واقعی، برخی از ویژگی‌های ترکیبات مورد مطالعه را پیش‌بینی کنند تا بتوانند به‌طور کارآمدی برای انجام آزمایش‌ها آماده شوند. از طرفی هزینه‌های آزمایشگاهی و به‌کارگیری نیروی انسانی و به پیروی از آن خطرات قرارگیری در معرض مواد شیمیایی نیز کاهش می‌یابد. بدون شک، امروزه بدون شبیه‌سازی مولکولی و بررسی‌های تئوری اولیه، هیچ سنتز آزمایشگاهی و آنالیز دستگاهی هدفمند و به‌صرفه نخواهد بود [۱].

۱- ۲ کمومتریکس

کمومتریکس[2] با به‌کارگیری روش‌های داده محور به استخراج داده‌ها از سیستم‌های شیمیایی می‌پردازد. در سال ۱۹۷۰ با توجه به پیشرفت الکترونیک و به‌کارگیری فزاینده کامپیوترها در علوم مختلف، این علم شناخته شد. اصطلاح کمومتریکس برای اولین بار در سال ۱۹۷۱، توسط سوانت ولد[3] بیان شد و انجمن بین المللی کمومتریکس توسط بروس کوالسکی[4] استاد شیمی تجزیه دانشگاه واشنگتن و ولد، تشکیل داده شد [۲]. علم کمومتریکس یک دانش بین رشته‌ای است، به‌طوری‌که با استفاده از علم شیمی محاسباتی، آمار کاربردی و کامپیوتر به تحلیل سیستم‌های متفاوت در شیمی، بیوفیزیک، بیوشیمی، زیست شناسی و

[1] Computational theoretical chemistry
[2] Chemometrics
[3] Svante Wold
[4] Bruce Kowalski

مهندسی شیمی می‌پردازد. به‌عبارت دیگر، کمومتریکس به معنای استفاده از روش‌های محاسباتی در اندازه‌گیری داده‌های شیمیایی می‌باشد. شیمیدان‌های تجزیه با به‌کارگیری تکنیک‌های متفاوت کمومتریکس، به پیشرفت‌های متنوعی در علوم دستگاهی، طراحی و انتخاب فرایندها و بهبود در طراحی و توسعه داروها به‌کمک روش‌های مدل‌سازی قدرتمند دست یافته‌اند.

بنابراین با توجه به کاربردهای ویژه تکنیک‌های علم کمومتریکس، می‌توان از آن برای توسعه ارتباط کمی ساختار- فعالیت[1](QSAR) و یا ارتباط کمی ساختار- خاصیت[2](QSPR) استفاده کرد.

۱- ۳ رابطه کمی ساختار- فعالیت (QSAR)

رابطه کمی ساختار- فعالیت (QSAR) یک مدل ریاضی و محاسباتی برای آشکار کردن روابط بین فعالیت‌های دارویی و ویژگی‌های ساختاری ترکیبات شیمیایی است. QSAR به‌عنوان یک روش طراحی دارو بیش از ۵۰ سال پیش توسط هانش و فوجیتا[3]توسعه یافت [۳]. از آن زمان تاکنون، QSAR یک روش کارآمد برای ساخت مدل‌های ریاضی است که تلاش می‌کند تا با استفاده از تکنیک‌های رگرسیونی، یک همبستگی آماری معنادار بین ساختار شیمیایی ترکیبات و فعالیت دارویی (pIC_{50}، pEC_{50}، K_i، و غیره) پیدا کند [۴]. اصل اساسی این روش به‌صورتی است که تغییرات در ویژگی‌های ساختاری (خواص فیزیکی و شیمیایی) باعث تغییرات متفاوتی در فعالیت‌های دارویی می‌شود [۵]. مدل‌سازی QSAR تعداد زیادی از مواد شیمیایی را از نظر فعالیت‌های دارویی مورد نظر اولویت دهی می‌کند. این روش یک فرایند شبیه‌سازی رایانه‌ای[4]است که با پیش‌بینی فعالیت دارویی ترکیبات، تعداد مواد شیمیایی کاندید برای انجام آزمایش را به میزان قابل توجهی کاهش می‌دهد [۶, ۷].

[1]Quantitative structure- activity relationships
[2]Quantitative structure- properties relationships
[3]Hansh and Fujita
[4]In silico

در توسعه مدل‌های QSAR، باید از مجموعه‌ای از داده‌های همگن استفاده شود، به‌این معنی که فعالیت دارویی آن‌ها، در شرایط آزمایشگاهی مشابه و با روش اندازه‌گیری یکسان اندازه‌گیری شده باشد، تا مدل نهایی از اعتبار قابل قبولی برخوردار باشد [۸]. استفاده از مجموعه داده‌های همگن منجر به انتخاب توصیف‌کننده‌های کم‌تر می‌شود و از این‌رو مدل‌های QSAR با تفسیرپذیری بالاتری ایجاد می‌شود. توصیف‌کننده‌های مولکولی مقادیر عددی و بیانگر ویژگی‌های فیزیکوشیمیایی ترکیبات مورد مطالعه هستند. توصیف‌کننده‌های مولکولی به دو صورت محاسباتی و یا تجربی استخراج می‌شوند و ساختار مولکولی را کمی می‌کنند و نشان‌دهنده اطلاعات مربوط به ساختار شیمیایی هستند. از ساده‌ترین توصیف کننده‌ها، می‌توان به تعداد و انواع اتم‌ها یا نوع پیوندهای شیمیایی موجود در ساختار مولکول اشاره کرد. توصیف‌کننده‌های مولکولی دقیق‌تر را نیز می‌توان از طریق نظریه‌های مختلف، مانند مکانیک کوانتومی و معادلات پیچیده‌تر ریاضی استخراج کرد [۹].

توسعه مدل‌های QSAR برای ترکیبات شیمیایی سنتز شده این امکان را به شیمیدان‌ها می‌دهد تا ترکیباتی با فعالیت دارویی بهتر پیشنهاد دهند. امروزه شیوع بیماری‌های همه‌گیر، نبود درمان و داروهای مؤثر برای برخی از بیماری‌ها، زندگی بشریت را در مواجهه با این بیماری‌ها با مخاطره همراه نموده است. بنابراین دانشمندان در تلاش هستند تا در راستای مهم‌ترین دغدغه بشریت که حفظ و ارتقا سلامتی و درمان بیماری‌های انسانی است، گام رو به رشدی بردارند. از جمله بیماری‌های چالش برانگیز می‌توان به ایدز، سرطان، آلزایمر، بیماری‌های همه‌گیر جدید همچون کووید-۱۹ اشاره کرد. البته مقاومت ویروس‌ها در برابر داروها و جهش‌های ویروسی همگی منجر به تلاش دانشمندان در پیشبرد تحقیقات بهبود طراحی و توسعه دارو شده است. امروزه با توجه به پیشرفت علوم الکترونیک و وجود کامپیوترها و نرم‌افزارهای محاسباتی قدرتمند، طراحی دارو به روش منطقی و مبتنی بر اصول محاسباتی به‌خوبی و با موفقیت جایگزین مناسبی برای طراحی داروی سنتی می‌باشد. بنابراین با توجه به مزایای یادشده، طراحی دارو به کمک

کامپیوتر[1] به دلیل صرفه‌جویی در زمان و هزینه‌های پژوهشی و کاهش نیروی انسانی لازم، ابزار مناسبی در پیشبرد فرایند توسعه دارو بوده است و ذهن دانشمندان طراحی دارو را به خود معطوف نموده است.

۱- ۴ رابطه کمی ساختار- خاصیت (QSPR)

رابطه کمی ساختار- خاصیت (QSPR) یک رابطه ریاضی بین ویژگی‌های ساختاری و خاصیت مربوط به مجموعه‌ای از مواد شیمیایی را توصیف می‌کند. طبیعتاً استفاده از چنین روابط ریاضی برای پیش‌بینی ویژگی هدف انواع مواد شیمیایی قبل از اندازه‌گیری‌های تجربی پرهزینه، فشرده و زمان‌بر مورد توجه محققین شیمیدان بوده است. استفاده از مدل‌های QSPR برای غربالگری پایگاه‌های اطلاعاتی شیمیایی قبل از سنتز و اندازه‌گیری تجربی آن‌ها، برای تولیدکنندگان مواد شیمیایی، شرکت‌های دارویی و سازمان‌های دولتی، به‌ویژه در زمان‌هایی که با کمبود منابع انسانی، بودجه‌ای و یا امکانات آزمایشگاهی مواجه بوده‌اند، بسیار حائز اهمیت است. با توجه به روند رو به رشد پایگاه‌های ترکیبات شیمیایی سنتز شده از یک‌طرف و فشارهای قانونی و اجتماعی برای ارزیابی به‌موقع خطرات بهداشتی و زیست محیطی مواد شیمیایی از سوی دیگر، تحریم‌های سیاسی موجود و ممنوعیت خرید مواد شیمیایی اولیه و استانداردهای شیمیایی از بازار جهانی و تورم اقتصادی موجود در جوامع جهان سوم، نیاز به ارائه مدل‌های QSPR قابل اعتماد همواره ضروری است [۱۰]. QSPR با استفاده از روش‌های کمومتریکس به بررسی چگونگی تأثیر تغییرات ساختار شیمیایی مولکول بر یک خاصیت شیمیایی معین می‌پردازد. بنابراین با استفاده از شیمی محاسباتی می‌توان بینش جدیدی از نظریه‌های مبتنی بر ارائه مدل‌های QSPR را درک نمود [۱۱, ۱۲]. از مطالعات QSPR برای تخمین خواصی مانند چگالی، نقطه جوش، حلالیت، شاخص بازداری کروماتوگرافی

[1] Computer Aided Drug Design

(RI[1]) و فشار بخار مواد شیمیایی استفاده می‌شود. استراتژی اصلی QSPR یافتن یک رابطه کمی بهینه است که با استفاده از آن بتوان خواص ترکیبات شیمیایی را پیش‌بینی کرد [۱۳].

۱-۵ اهمیت ساخت مدل‌های ارتباط کمی ساختار- فعالیت/ ویژگی (QSAR/QSPR)

کیفیت، ایمنی و کارایی یک داروی راه‌یافته به بازار، با انجام آزمایش‌های مختلف مورد ارزیابی قرار می‌گیرد [۱۴]. تولید دارو فرآیندی زمان‌بر و پرهزینه است و از مرحله اولیه کشف ترکیب بالقوه دارویی تا توسعه و عرضه آن، اغلب به‌طور متوسط ۱۵ سال طول می‌کشد [۱۵] و هزینه عرضه هر دارو به بازار حدود ۹۰۰ میلیون دلار برآورد شده است [۱۶]. هزینه بالا و فرآیند طولانی به دلیل ریسک بالای شکست توسعه دارو است. تخمین زده شده است که تنها ۱۱ درصد از داروهایی که مرحله توسعه را به پایان رساندند، توسط سازمان جهانی غذا و دارو ایالت متحده تأیید شدند [۱۷]. طبق تحقیقات انجام شده، مشخص شد که ۱۰ درصد از شکست توسعه دارو به دلیل ویژگی‌های فارماکوکینتیک ضعیف بوده است، در حالی‌که در مرحله بالینی، ۳۰ درصد حذف ترکیب رهبر، ناشی از عدم کارایی و ۳۰ درصد دیگر ناشی از سمیت یا ایمنی بالینی است [۱۷, ۱۸]. بنابراین، پیش‌بینی این شکست‌ها قبل از مرحله بالینی به‌منظور کاهش هزینه‌های توسعه دارو مفید خواهد بود. ادعا می‌شود که با ۱۰ درصد بهبود پیش‌بینی می‌توان به‌صرفه جویی ۱۰۰ میلیون دلاری در هزینه‌های توسعه هر دارو دست یافت [۱۸]. بنابراین، روش‌های مختلفی مانند روش‌های in vitro، in vivo یا in silico، در اوایل مرحله توسعه دارو برای فیلتر کردن شکست‌های احتمالی استفاده می‌شوند. نمونه‌ای از روش in silico، مدل‌های رابطه کمی ساختار-فعالیت (QSAR) است که می‌تواند برای درک عملکرد دارو، طراحی ترکیبات جدید و نمایش اطلاعات شیمیایی مناسب، مورد استفاده قرار گیرد [۱۹-

[1]Retention index

۲۲]. بنابراین، می‌توان آزمایش‌های پرهزینه یا آنالیز خاصیت فیزیکوشیمیایی مواد خطرناک و سمی یا ترکیبات ناپایدار را با بررسی ارتباط کمی توصیف‌کننده‌های مولکولی – خاصیت شیمیایی جایگزین کرد، که این امر به‌نوبه خود می‌تواند برای پیش‌بینی پاسخ‌های هدف ترکیبات جدید استفاده شود.

۱- ۵- ۱ مراحل ساخت مدل‌های QSAR/QSPR

همان‌طور که گفته شد هدف از انجام مطالعات QSAR/QSPR، یافتن ارتباط بین توصیف‌کننده‌های ساختاری و ویژگی هدف (فعالیت دارویی و یا خاصیت فیزیکوشیمیایی) است. اما پیدا کردن این ارتباط در طی یک روش مستقیم حاصل نمی‌شود. بلکه مراحل متفاوتی برای ساخت مدل‌های QSAR/QSPR انجام می‌شود که شکل ۱-۱ نمایانگر تمامی مراحل مورد نیاز می‌باشد. همان‌طور که شکل ۱-۱ نشان می‌دهد مراحل اصلی ساخت مدل QSAR/QSPR با شماره‌های ۱ تا ۶ مشخص شده است. یک مدل QSAR/QSPR معتبر باید از ۵ اصل اساسی منتشر شده توسط سازمان همکاری اقتصادی و توسعه (OECD[1])، پیروی کند که به ترتیب زیر است [۲۳]:

۱) وجود یک پاسخ هدف تعریف شده

۲) استفاده از یک الگوریتم بدون ابهام

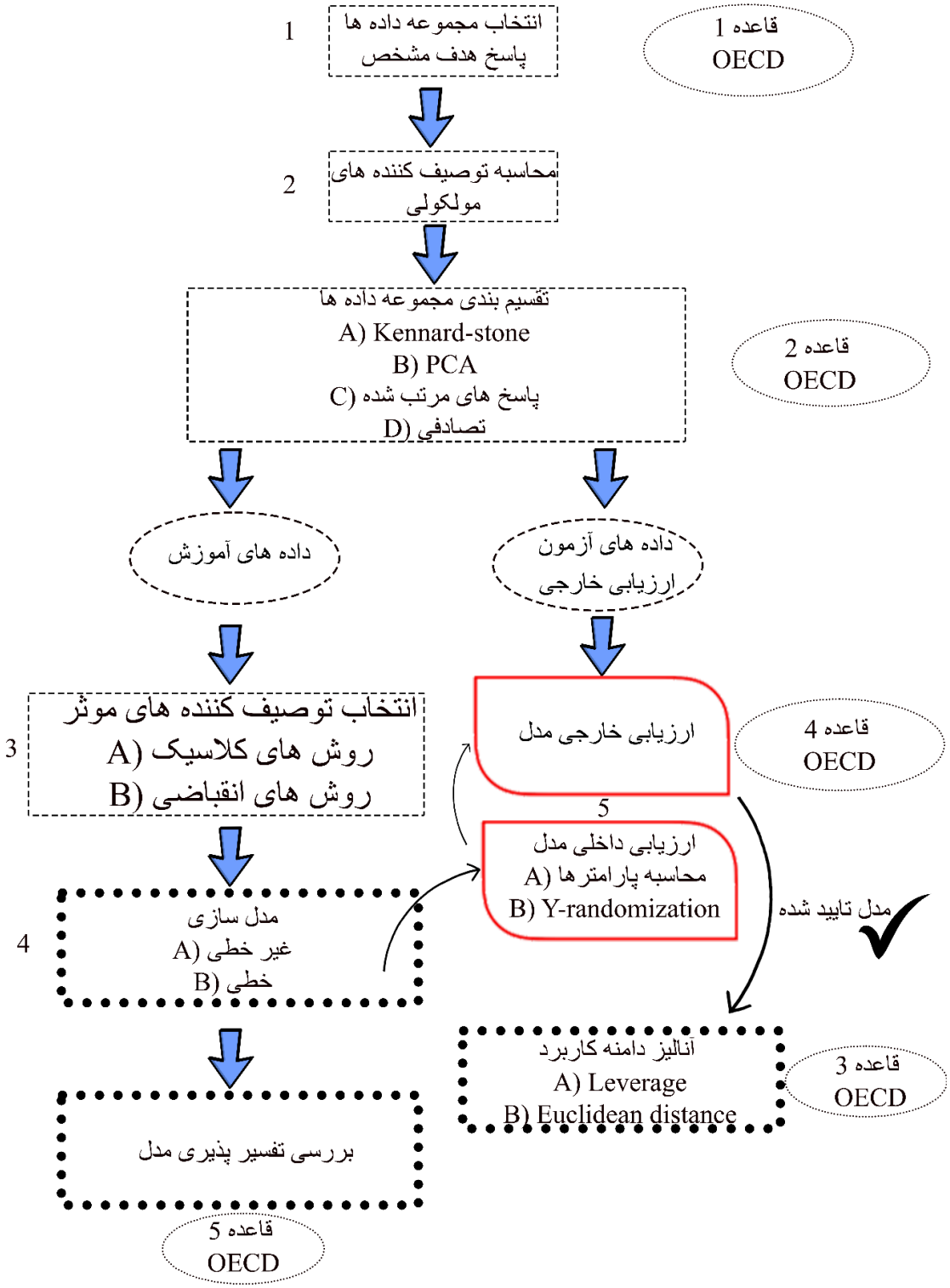
۳) وجود دامنه کاربرد تعریف شده معتبر

۴) تأیید پارامترهای آماری مدل و قدرت پیش‌بینی مناسب مدل

۵) بررسی تفسیرپذیری مدل در صورت امکان

شکل ۱-۳ جزییات هر یک از مراحل ساخت مدل QSAR/QSPR را نشان می‌دهد. ارتباط هر کدام از بندهای قاعده OECD به مراحل ساخت مدل QSAR در شکل ۱-۳ آورده شده است.

[1]The Organiation for Economic Co-operation and Development



شکل ۱-۱ مراحل ساخت، توسعه و ارزیابی مدل‌های QSAR/QSPR

۱- ۵- ۲ جمع‌آوری و انتخاب مجموعه داده‌ها

اولین مرحله مدل‌سازی QSAR/QSPR، جمع‌آوری و انتخاب ترکیبات شیمیایی مورد نظر از منابع قابل اعتماد و در دسترس است. پاسخ هدف مورد نظر برای مدل‌سازی QSAR/QSPR، بایستی در شرایط عملی یکسان اندازه‌گیری شده باشد تا نتیجه قابل قبول‌تر و مناسب‌تری به‌دست آید. در مدل‌سازی QSAR/QSPR مجموعه داده‌ها با استفاده از آنالیزهای متفاوت تقسیم‌بندی از جمله کنارد-استون[۱](KS) انجام می‌شود [۲۴, ۲۵]. تقسیم‌بندی KS بر اساس محاسبه فاصله اقلیدسی (Euclidean distance) است. بنابراین مجموعه داده‌ها به سه دسته آموزشی[۲]، ارزیابی[۳]و آزمون[۴]تقسیم می‌شود. الگوریتم KS داده‌ها را به‌طوری تقسیم می‌کند که مجموعه‌های تقسیم‌بندی شده مانند مجموعه ارزیابی و مجموعه آزمون نماینده مناسبی از مجموعه آموزش باشد. از این‌رو داده‌های مجموعه آموزش، ارزیابی و آزمون به‌ترتیب برای ساخت مدل، ارزیابی داخلی مدل و ارزیابی خارجی مدل به‌کار گرفته می‌شوند. لازم به ذکر است که مجموعه آزمون در هیچ یک از مراحل انتخاب متغیر و ساخت مدل‌های QSAR/QSPR شرکت نخواهد داشت.

۱- ۵- ۳ رسم و بهینه‌سازی ساختارهای شیمیایی

به‌منظور محاسبه صحیح توصیف کننده‌ها، به ساختارهای شیمیایی با حالت پایدار مولکول نیاز می‌باشد. ایجاد پایدارترین حالت ساختار شیمیایی ترکیبات با حداقل انرژی با استفاده از نرم‌افزارهای متفاوتی از جمله هایپرکم[۵]امکان‌پذیر است [۲۶]. بنابراین پس از رسم ساختار شیمیایی ترکیبات مورد نظر، بهینه‌سازی آن‌ها با استفاده از روش‌های متفاوتی همچون روش بهینه‌سازی نیمه تجربی (AM1) انجام

[1]Kennard-Stone
[2]Training
[3]Validation
[4]Test
[5]Hyperchem

می‌شود. بنابراین پایدارترین حالت با کم‌ترین انرژی برای هر ساختار به دست می‌آید. ساختارهای بهینه شده به‌عنوان ورودی نرم‌افزارهای محاسبه‌ای توصیف کننده‌های مولکولی به‌کار گرفته می‌شوند.

۱- ۵- ۴ استخراج توصیف کننده‌ها

توصیف‌کننده‌های مولکولی، مقادیر عددی و بیانگر ویژگی‌های مربوط به ترکیبات شیمیایی هستند. توصیف‌کننده‌های مولکولی به دو دسته اصلی توصیف‌کننده‌های تجربی و محاسباتی تقسیم می‌شوند. توصیف‌کننده‌های تجربی با استفاده از پارامترهای آزمایشگاهی اندازه‌گیری می‌شوند و تنها برای ترکیبات سنتز شده و در دسترس شیمیدان قابل محاسبه هستند. توصیف‌کننده‌های نظری و یا محاسباتی با استفاده از ساختار ترکیبات شیمیایی و بر اساس معادلات ریاضی ساده و یا پیچیده محاسبه می‌شوند، بنابراین برای همه ترکیبات شیمیایی پس از رسم و بهینه‌سازی ساختار قابل محاسبه هستند و هیچ محدودیت محاسباتی برای آن وجود ندارد. با پیشرفت‌های اخیر در علم شیمی محاسباتی و ایجاد نرم‌افزارهای محاسباتی کارآمد مانند دراگون[1] تعداد زیادی توصیف کننده مولکولی برای یک مولکول قابل محاسبه است.

۱- ۵- ۵ پیش‌پردازش[2] توصیف‌کننده‌های محاسبه شده

همان‌طور که گفته شد، پس از محاسبه توصیف کننده‌ها تعداد زیادی از این توصیف‌کننده‌های مولکولی (متغیرهای مستقل) به‌وجود می‌آید، که همه این‌ها دارای اهمیت و ارتباط ویژه با پاسخ هدف نمی‌باشند. بنابراین قبل از انتخاب مؤثرترین توصیف کننده‌ها، عملیات پیش‌پردازش و یا غربالگری[3] انجام می‌شود. به‌طوری‌که، توصیف‌کننده‌های مولکولی زائد و فاقد اطلاعات مفید از فضای ماتریس متغیرهای مستقل خارج می‌شوند. این فرایند علاوه بر این‌که سبب کاهش زمان محاسبات در مراحل انتخاب متغیر و مدل‌سازی می‌شود، بلکه صحت پیش‌بینی مدل‌های توسعه یافته را نیز افزایش می‌دهد. به‌منظور انجام فرایند

[1] Dragon
[2] Pre-processing
[3] Screening

پیش‌پردازش ابتدا توصیف‌کننده‌هایی با مقادیر ثابت و نسبتاً ثابت با استفاده از نرم‌افزار آماری R [۲۷] و با استفاده از بسته نرم‌افزاری caret [۲۸] حذف می‌شوند. سپس همبستگی بین توصیف کننده‌ها نیز با استفاده از دستور corecoeff در نرم‌افزار متلب[1]مورد بررسی قرار خواهد گرفت. از بین دو توصیف کننده با همبستگی بالای ۰/۹، توصیف کننده‌ای که بیش‌ترین همبستگی را با متغیر وابسته (فعالیت دارویی یا خاصیت فیزیکوشیمیایی) داشت حفظ شده و دیگری حذف خواهد شد.

۱- ۵- ۶ انتخاب توصیف‌کننده‌های مؤثر

روش‌های متفاوتی برای انتخاب مؤثرترین توصیف کننده‌ها استفاده می‌شود که می‌توان به روش‌های انتخاب متغیر کلاسیک و روش‌های انقباضی[2]اشاره کرد. روش‌های کلاسیک همچون رگرسیون خطی چندگانه[3]، دارای مشکلاتی چون بی ثباتی، واریانس بالای برآوردگرها، بایاس بالا و تعداد زیاد توصیف‌کننده‌های منتخب و به‌دنبال آن کاهش تفسیرپذیری می‌باشند. از این‌رو روش‌های رگرسیونی انقباضی یا جریمه شده[4] جدید به هدف حل مشکلات موجود معرفی شدند. روش حداقل قدر مطلق انقباض و عملگر انتخاب کننده (LASSO[5]) [۲۹]، لاسوی تطبیقی (ALASSO[6]) [۳۰]، تابع جریمه انحراف قدر مطلق به‌طور هموار بریده شده (SCAD[7]) [۳۱] و حداقل انحراف مطلق- حداقل قدر مطلق انقباض و عملگر انتخاب کننده (LAD-LASSO)[8][۳۲] به‌عنوان روش‌های انقباضی جدید بسیار کارآمد هستند. در ادامه به معرفی انواع روش‌های انتخاب متغیر پرداخته می‌شود.

---

[1]MATLAB
[2]Penalized methods
[3] Multiple Linear Regression (MLR)
[4] Shrinkage or penalized regression methods
[5] Least absolute shrinkage and selection operator (LASSO)
[6] Adaptive LASSO (ALASSO)
[7] Smoothly clipped absolute deviation (SCAD)
[8]Least absolute deviation- Least absolute shrinkage and selection operator

۱- ۵- ۶- ۱ انتخاب متغیر به روش رگرسیون خطی چندگانه

رگرسیون، معادله‌ای برای ایجاد یک رابطه بین یک متغیر وابسته از یک‌طرف و یک یا چند متغیر مستقل از طرف دیگر می‌باشد. اگر تنها یک متغیر مستقل وجود داشته باشد، رگرسیون را ساده و در غیر این صورت، رگرسیون را چندگانه می‌گویند. در بسیاری موارد نمی‌توان تغییرات یک متغیر وابسته را فقط به مقادیر یک متغیر مستقل ارتباط داد. به‌عبارت دیگر برای پیش‌بینی مقادیر یک متغیر، دانستن مقادیر دو یا چند متغیر دیگر لازم است. زمانی‌که تعداد متغیرهای مستقل مورد مطالعه زیاد باشد لازم است که تکنیکی برای انتخاب متغیرهای مهم و اثرگذار مورد استفاده قرار گیرد تا بتوان متغیرهای معنی‌دار را از متغیرهای بی‌اهمیت تمییز داد. روش‌های رگرسیون خطی چندگانه متفاوتی برای انتخاب بهترین زیر مجموعه از تمام مجموعه‌ها وجود دارد. از جمله مهم‌ترین آن‌ها می‌توان به انتخاب پیش‌رونده[1] حذفی پس‌رونده[2] و روش گام‌به‌گام[3] اشاره کرد. هر چند این‌روش‌ها، بهترین زیرمجموعه را از بین تمام زیرمجموعه‌های ممکن انتخاب می‌نمایند، اما زمانی که ابعاد داده‌ها بزرگ باشد تعداد زیر مجموعه‌های منتخب افزایش می‌یابد و با افزایش تعداد متغیرهای مدل، مقدار ضریب تعیین به‌طور کاذب بالا بوده و قدرت پیش‌بینی مدل پایین می‌آید. علاوه بر این، روش‌های انتخاب متغیر کلاسیک دارای مشکل بی ثباتی هستند. به‌طوری‌که با تغییری کوچک در داده‌ها، مدل‌های خیلی متفاوتی به وجود می‌آید. با توجه به مشکلات موجود، محققین رگرسیون جریمه شده را معرفی کردند که با اعمال یک پارامتر جریمهٔ[4] مناسب دو عمل انتخاب متغیر و برآورد ضرایب به‌طور هم‌زمان انجام می‌شود. بنابراین همان‌طور که گفته شد استفاده از روش انتخاب متغیر خطی چندگانه به دلیل داشتن معایبی هم‌چون ناپایداری، بایاس بالا، ناکارآمدی در حضور هم‌خطی، برای انتخاب توصیف کننده‌های منتخب توصیه نمی‌شود [۲۹].

[1] Forward
[2] Backward
[3] Stepwise
[4] Penalty

۱-۵-۶-۲ LASSO

مدل رگرسیون خطی زیر را در نظر بگیرید:

$y_i=x_i^{'}\beta+\epsilon_i \qquad i=1, \ldots,n$

که در آن $x_i=(x_{i1},...,x_{ip})^{'}$ ماتریس توصیف‌کننده‌های مولکولی با ابعاد p و y_i متغیر وابسته و $\beta=\beta_1,...,\beta_p$ ضرایب رگرسیونی و ϵ_i نیز خطای تصادفی با میانه ۰ می‌باشد. تیبشیرانی[1] (۱۹۹۶) روشی جدید برای تخمین مدل‌های خطی ارائه کرد و آن‌را LASSO نامید. این روش به‌صورت هم‌زمان به برآورد و انتخاب متغیر می‌پردازد. انگیزه اصلی تیبشیرانی در تعریف LASSO، برگرفته از پیشنهاد بریمن[2] (۱۹۹۳) است.

به پیشنهاد بریمن

$$\sum_{i=1}^{n}\left\{y_i-\sum_{j=1}^{p} c_j\widehat{\beta_j}\ x_{ij}\right\}^2 \qquad \text{رابطه ۱-۱}$$

تحت شرایط $c_j \geq 0$ و $\sum_{j=1}^{k} c_j \leq t$ ضریب رگرسیونی حداقل می‌شود که در آن t یک مقدار ثابت است. اگر t=0 باشد آن‌گاه متغیری در مدل نخواهد بود و اگر □=t باشد مدل با مجموعه‌ای از متغیرها ساخته می‌شود .

LASSO را در حالت کلی می‌توان با مینیمم کردن عبارت خطای جریمه شده به‌صورت زیر به‌دست آورد:

$$\hat{\beta}^{LASSO}=\arg\min_{\beta\in R^p}\sum_{i=1}^{n}|y_i-x_i^{'}\beta|+\lambda\sum_{i=1}^{p}|\beta_i| \qquad \text{رابطه ۱-۲}$$

که در آن λ یک پارامتر تنظیم‌کننده نامنفی $(\lambda > 0)$، $|\beta| = (|\beta_1|.\cdots.|\beta_P|)^T$ می‌باشد. پارامتر λ سطح تنکی و تعداد ضرایب صفر را در برآوردگر LASSO کنترل می‌کند. به‌عبارت دیگر، هرچه جریمه بزرگ‌تری

[1] Tibshirani

[2] Breiman

به‌کار برده شود، تعداد بیش‌تری از ضرایب به سمت صفر منقبض می‌شوند. در حقیقت λ و t رفتار یا رابطه‌ای معکوس با یک‌دیگر دارند. LASSO با افزایش λ ضرایب را به سمت صفر منقبض می‌کند و زمانی‌که $\lambda \to 0$ ضرایب بیش‌تری در مدل باقی خواهند ماند. در واقع LASSO مجموع توان‌های دوم خطای مدل رگرسیون را تحت این محدودیت که مجموع مقادیر مطلق ضرایب رگرسیونی کم‌تر از یک مقدار ثابت باشند، حداقل می‌سازد. به دلیل اعمال این محدودیت، LASSO تمایل به ایجاد ضرایب دقیقاً برابر با صفر دارد که در نتیجه LASSO می‌تواند به‌عنوان یک روش انتخاب متغیر کارآمد در نظر گرفته شود. اهمیت ذاتی LASSO زمانی بیش‌تر آشکار می‌شود که با داده‌هایی با ابعاد بالا سر و کار داریم. از آنجایی‌که متغیرهای خیلی کمی ارتباط معنادار با متغیر وابسته دارند، بنابراین استفاده از روشی که بتواند ضرایب متغیرهای بی‌اهمیت را صفر کند مورد نیاز است. تیبشیرانی با معرفی LASSO روشی را معرفی کرد که با آن می‌توان ابعاد داده‌ها را کاهش داد. با توجه به این‌که تعداد متغیرها کم می‌شود، مدل‌های ساخته شده تفسیرپذیر، ساده و پایدارتر هستند [۲۹].

LASSO پنجره جدیدی را به روی مبحث انتخاب متغیر گشود، به‌طوری‌که بسیاری از روش‌های انتخاب متغیر که بعد از سال ۱۹۹۶ ارائه شدند به‌نوعی با LASSO در ارتباط هستند و هر یک با هدف اصلاح یکی از نقاط ضعف LASSO ایجاد شدند. در واقع در استفاده از توابع جریمه‌ای، سؤالی که پیش می‌آید، نوع جریمه‌ای است که باید مورد استفاده قرار گیرد. فن و لی[1] (۲۰۰۱) نشان دادند که یک تابع جریمه خوب باید سه ویژگی مطلوب زیر را داشته باشد.

- ❖ تنکی[2]: برآوردگر حاصل باید به‌طور خودکار ضریب برآورد شده‌ای که مقدار کوچک دارد را صفر کند تا متغیرهای مناسب انتخاب شود و به‌این وسیله پیچیدگی مدل کاهش می‌یابد.

[1] Fan and Li

[2] Sparsity

- نا اریبی[1]: برآوردگر به‌دست آمده برای ضرایب رگرسیونی با مقادیر بزرگ، تقریباً نا اریب باشد تا اریبی مدل کاهش یابد.
- پیوستگی[2]: برآوردگر حاصل پیوسته باشد تا باعث پایداری مدل شود.

با این‌که LASSO در بسیاری از مسائل عملکرد خوبی را از خود نشان داده است اما دارای محدودیت‌هایی نیز می‌باشد. به‌عنوان مثال نتیجه جریمه LASSO، برآوردگری با اریبی زیاد می‌باشد، زیرا اندازه انقباضی که به ضرایب رگرسیونی کوچک و بزرگ اختصاص می‌دهد یکسان است و این باعث می‌شود که ضرایب کوچکی که دارای ارتباط مؤثر با متغیر وابسته هستند نیز منقبض و صفر شوند. همچنین LASSO برای یک مدل رگرسیونی خطی با p متغیر پیشگو و n مشاهده، حداکثر n متغیر انتخاب می‌کند. این محدودیت‌ها نشان‌دهنده کارایی پیش‌بینی پایین LASSO در تخمین دقیق ضرایب غیر صفر می‌باشد. منظور از کارایی پیش‌بینی، سازگاری روش در انتخاب متغیر است. به‌این معنی‌که یک روش با احتمال نزدیک به یک بتواند ضرایب غیر صفر را به‌درستی شناسایی کند. از این‌رو در سال ۲۰۰۱ یک جریمه مناسب با ویژگی‌های نا اریب، تنک و پیوسته به نام SCAD معرفی شد [۳۱] و همچنین لاسوی تطبیقی (ALASSO) با کارایی پیش‌بینی مناسبی در سال ۲۰۰۶ برای رفع ناسازگاری LASSO معرفی شد [۳۰]. روش‌های نامبرده در ادامه به اختصار شرح داده می‌شوند.

۱- ۵- ۶- ۳ SCAD

همان‌طور که گفته شد در مقایسه با LASSO، تابع جریمه SCAD، دارای ویژگی‌هایی چون نا اریبی، تنکی و پیوستگی می‌باشد. با توجه به اینکه در روش LASSO ضرایب بزرگ نیز جریمه می‌شوند،

[1] Unbiasedness

[2] Continuity

ضرایب کوچک‌تر فرصت حضور در مدل را پیدا می‌کنند و این سبب بیش برازش[1] می‌گردد. به‌این ترتیب LASSO از ویژگی نا اریبی بی بهره است. تابع جریمه SCAD به‌صورت زیر ارائه شده است:

$$\hat{\beta}^{SCAD}=\arg\min_{\beta\in R^p}\left\{\|Y-X\beta\|^2+\lambda\sum_{i=1}^{p}P(|\beta_i|);a.\lambda\right\}$$ رابطه ۱-۳

که $P(.;a,\lambda)$ تابع جریمه SCAD است و دارای سه آرگومان زیر است:

$$P(t;a.\lambda)=\begin{cases}\lambda t & 0\le t\le\lambda\\ \dfrac{2a\lambda t-t^2-\lambda^2}{2(a-1)} & \lambda<t<a\lambda\\ \dfrac{\lambda^2(a+1)}{2} & t\ge a\lambda\end{cases}$$ رابطه ۱-۴

$a>2$ و $\lambda>0$ پارامترهای جریمه هستند و اندازه انقباض ضرایب را کنترل می‌کنند. با توجه به رابطه ۱-۳ ملاحظه می‌شود که در فاصله $|\beta_j|\le\lambda$ تابع جریمه SCAD بر LASSO منطبق است و پس از آن تا زمانی که $|\beta_j|<a\lambda$ جریمه SCAD یک تابع درجه دوم است. سپس به ازای $|\beta_j|\ge a\lambda$ به یک تابع با مقدار ثابت تبدیل می‌شود و وقتی $\alpha\to\infty$ ، برآوردگر SCAD با برآوردگر LASSO معادل است. فن و لی با مطالعات متعدد نشان دادند که a برای SCAD برابر با ۳/۷ می‌باشد [۳۱]. برنامه SCAD را می‌توان با استفاده از بسته نرم‌افزاری ncvreg در برنامه R اجرا نمود [۳۳].

۱-۵-۶-۴ ALASSO

به‌صورت کلی تابع جریمه ALASSO، مانند LASSO است. اما برخلاف LASSO برای ضرایب مختلف، جریمه‌های متفاوتی را در نظر می‌گیرد.

$$\hat{\beta}_{ALASSO}=\arg\min_{\beta\in R^p}\left\{\|Y-X\beta\|^2+\lambda\sum_{i=1}^{p}\hat{w}_i|\beta_i|\right\}$$ رابطه ۱-۵

[1] Overfitting

$\hat{w}_i$ وزن بردار می‌باشد. با توجه به معادله فوق، اگر وزن‌ها به‌درستی انتخاب شوند این نوع از LASSO دارای کارایی پیش‌بینی بالاتری خواهد بود. وزن‌ها نیز به‌صورت زیر تعریف می‌شوند:

$$\hat{w}_i = \frac{1}{|\hat{\beta}_i|^{\gamma}} \qquad \gamma>0$$ رابطه ۱-۶

که در آن β_i می‌تواند برآوردگر رگرسیون کم‌ترین توان‌های دوم، LASSO و یا Ridge باشد. زو (۲۰۰۶) ALASSO را به هدف رفع ناسازگاری مدل LASSO معرفی کرد. ALASSO با توجه به اعمال جریمه‌های متفاوت روی ضرایب با بزرگی متفاوت، سبب بهبود کارایی پیش‌بینی مدل شد. ALASSO را می‌توان با استفاده از بسته نرم‌افزاری parcore در برنامه R اجرا نمود [۳۴].

۱- ۵- ۶- ۵ LAD-LASSO

در مدل رگرسیون خطی

$$y_i = x_i'\beta + \epsilon_i \qquad i=1, \ldots, n$$ رابطه ۱-۷

برای برآورد پارامترها از حداقل سازی عبارت OLS که برابر است با $\sum_{i=1}^{n}(y_i - x_i'\beta)^2$ استفاده می‌شود. اما عملکرد برآوردهای رگرسیونی OLS، با وجود داده‌های دور افتاده محدود می‌شود. برای مقابله با مشاهده‌های دور افتاده روش‌های رگرسیونی استواری چون برآوردگر حداقل قدر مطلق انحراف[1](LAD) پیشنهاد می‌شود.

$$\hat{\beta}_{LAD} = \operatorname{argmin} \sum_{i=1}^{n} |y_i - x_i\beta|$$ رابطه ۱-۸

در سال ۱۹۹۶ تیبشیرانی برای صفر کردن ضرایب بی‌اهمیت LASSO را معرفی نمود که با نگاهی به رابطه ۱-۲ مشاهده می‌شود که با اضافه شدن عبارت جریمه به معادله حداقل مربعات معمولی ضرایبی با

[1] Least Absolute Deviations

مقدار ناچیز صفر می‌شوند. هر چه مقدار λ بزرگ‌تر باشد میزان انقباض بیش‌تر و ضرایب رگرسیون کم‌تری باقی می‌ماند و در غیر این‌صورت با کوچک شدن مقدار λ تعداد ضرایب غیر صفر بیش‌تری باقی می‌ماند.

با توجه به اینکه LASSO برای همه ضرایب از یک مقدار λ واحد برای انقباض استفاده می‌کند پارامتر تنظیم یکسانی برای همه پارامترها اعمال می‌شود و مدل از بایاس و اریبی رنج می‌برد. با اضافه شدن تابع جریمه نرم L1 به برآوردگر LAD می‌توان یک مدل تنک به‌دست آورد که علاوه بر مواجهه با بایاس، به اریبی در برابر داده‌های دور افتاده، نیز حساسیت ندارد. این روش در مقایسه با LAD می‌تواند عمل انتخاب متغیر و برآورد پارامتر را به‌صورت هم‌زمان انجام دهد و در مقایسه با LASSO در مقابل مشاهده‌های دورافتاده استوار است.

$$\hat{\beta}_{\text{LAD-LASSO}}=\operatorname{argmin}\sum_{i=1}^{n}\left|y_i-x_i'\beta\right|+\lambda_j\sum_{j=1}^{p}\left|\beta_j\right| \qquad \text{رابطه ۱-۹}$$

برآوردگر LAD-LASSO را می‌توان به‌عنوان یک برآوردگر بیزی در نظر گرفت، به‌طوری که هر ضریب رگرسیونی β_j دارای پارامتر مقیاسی $n\lambda_j$ است و بر این اساس مقدار $\widehat{\lambda_j}=\frac{1}{n|\beta_j|}$ است که با استفاده از برآورد LAD معمولی برآورد می‌شود، که با استفاده از تکنیک تخمین LAD معمولی تخمین زده می‌شود. LAD-LASSO را می‌توان با استفاده از بسته نرم‌افزاری MTE در برنامه R بدون برنامه ریزی محاسباتی پیچیده به‌دست آورد [۳۵].

۱- ۵- ۶- ۶ ارزیابی توصیف‌کننده‌های منتخب

توصیف‌کننده‌های منتخب روش انتخاب متغیر مورد نظر باید از نظر وجود همبستگی و هم‌خطی مورد ارزیابی قرار گیرند. به‌عبارت دیگر توصیف‌کننده‌های مولکولی منتخب باید فاقد همبستگی بالای ۰/۹ (معنی‌دار) و فاقد هم‌خطی باشند. پدیده همبستگی با استفاده از محاسبه مربع ضریب همبستگی به‌دست

می‌آید. با رسم نمودار نقشه رنگی[1]نتایج با کیفیت و وضوح بیش‌تری قابل نمایش خواهد بود. مقادیر همبستگی نزدیک به ۱ و ۱- نشاندهنده همبستگی بالا بین دو توصیف کننده می‌باشد. پدیده هم‌خطی با استفاده از محاسبه پارامتر افزایش عامل واریانس[2](VIF) مورد بررسی قرار گرفت. VIF مربوط به هر توصیف کننده منتخب با استفاده از رابطه ۱-۱۰ محاسبه شد.

$$VIF=\frac{1}{1-R_i^2}.\ i=1.2.3\cdots.p$$ رابطه ۱-۱۰

p در رابطه ۱-۱۰ تعداد پارامترهای منتخب روش انتخاب متغیر و R_i^2 نیز مجذور همبستگی چندگانه است که از رگرسیون یک متغیر بر سایر متغیرها به‌دست می‌آید. قرارگیری مقدار VIF در محدوده ۱-۵ دلالت بر عدم وجود پدیده هم‌خطی نگران کننده بین توصیف‌کننده‌های منتخب می‌باشد [۳۸-۳۶].

۱- ۵- ۷ مدل‌سازی

به‌منظور ایجاد ارتباط بین متغیرهای مستقل و وابسته، مدل‌های ریاضی، ساخته می‌شود. با استفاده از مدل‌های توسعه یافته به‌راحتی می‌توان پاسخ هدف مربوط به ترکیبات جدید را تخمین زد. با پیشرفت علم و تکنولوژی، روش‌های مدل‌سازی متفاوتی گسترش یافته‌اند که از آن جمله می‌توان به رگرسیون خطی چندگانه (MLR)، رگرسیون اجزای اصلی (PCR)[3]، کم‌ترین توان‌های دوم جزئی (PLS)[4]، شبکه عصبی مصنوعی (ANN)[5]اشاره کرد. در این پژوهش از روش شبکه عصبی مصنوعی استفاده شده است.

۱- ۵- ۷- ۱ شبکه عصبی مصنوعی

شبکه عصبی یک برنامه نرم‌افزاری است که می‌تواند همانند مغز انسان عمل نماید. در واقع یک شبکه عصبی مصنوعی ایده‌ای است برای پردازش اطلاعات که از سیستم عصبی زیستی الهام گرفته شده

[1]Heatmap
[2]Variance inflation factor
[3]Principal component regression
[4]Partial least square
[5]Artificial neural network

و مانند مغز انسان به پردازش اطلاعات می‌پردازد. این سیستم‌ها از تعداد زیادی عنصر پردازش به نام نورون تشکیل شده‌اند که برای حل یک مسئله به‌صورت هماهنگ با هم عمل می‌کنند. شبکه‌های عصبی مصنوعی نظیر مغز انسان‌ها، با مثال یاد می‌گیرند و با پردازش روی داده‌های تجربی، دانش یا قانون نهفته در ورای داده‌ها را به ساختار شبکه منتقل می‌کنند. به‌همین خاطر به این سیستم‌ها، هوشمند گفته می‌شود، زیرا شبکه‌ها بر اساس محاسبات روی داده‌های عددی یا مثال‌ها، قوانین کلی را فرا می‌گیرند. امروزه از شبکه‌های عصبی به‌عنوان یک ابزار کارآمد در زمینه‌های مختلف علمی از جمله صنایع الکترونیک، پزشکی، اکتشاف نفت وگاز، رباتیک، شیمی و داروسازی استفاده می‌شود. در این پروژه از روش شبکه عصبی به‌عنوان یک مدل قدرتمند غیرخطی برای ایجاد ارتباط بین متغیرهای مستقل و وابسته استفاده شده است.

-آموزش شبکه‌های پیشخور با تکنیک پس انتشار[1]خطا

تکنیک پس انتشار خطا یک روش متداول آموزش با ناظر برای شبکه‌های پیشخور است یعنی برای به‌دست آوردن ارتباط بین متغیرهای ورودی و خروجی در یادگیری به الگوی آموزشی نیاز است. به‌طور کلی آموزش به کمک تکنیک پس انتشار طبق مراحل زیر انجام می‌شود [۳۹].

۱- انتشار ورودی‌ها از نورون‌های ورودی به سمت نورون‌های خروجی

۲-اختصاص ماتریس وزن‌های تصادفی به هریک از اتصالات

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

۴- پس انتشار خطا از نورون‌های خروجی به سمت نورون‌های ورودی و اصلاح وزن‌ها

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

[1]Back propagation

مراحل فوق تا رسیدن به حداکثر تکرار[1](دور آموزش) مجاز انجام می‌شود و یا این‌که مقدار تابع کارآیی از مقداری که تعیین شده کم‌تر باشد. شبکه عصبی انتخاب شده در این پروژه یک شبکه پیشخور با الگوریتم آموزشی پس انتشار خطا می‌باشد[۳۹].

در این مطالعه، یک ANN پیشخور با الگوریتم پس انتشار خطا برای مدل‌سازی QSAR/QSPR در نظر گرفته شده است. در اکثر تحقیقات در زمینه شیمی محاسباتی، ساخت مدل ANN با یک لایه کافی است [۴۰]. بنابراین، مدل‌های ANN توسعه یافته، دارای سه لایه شامل یک لایه ورودی، یک لایه پنهان و یک لایه خروجی است که برای بهینه‌سازی پارامترهای ANN استفاده می‌شود. برای به دست آوردن بهترین مدل شبکه عصبی مصنوعی، تمامی پارامترهای مؤثر بر عملکرد پیش‌بینی مدل شبکه عصبی مصنوعی مانند تعداد ورودی‌ها، گره‌ها در لایه پنهان، دورهای آموزشی و توابع آموزش و انتقال بهینه‌سازی می‌شوند. به‌این منظور، چهار مدل شبکه عصبی مصنوعی مختلف با استفاده از دو الگوریتم آموزشی متفاوت (لونبرگ – مارکوارت (LM)[2]و تنظیم بیزین (BR)[3]و دو تابع انتقال متفاوت برای لایه پنهان (لگاریتم سیگموئیدی و تانژانت هایپربولیک سیگموئیدی) برای همه مطالعات QSAR/QSPR طراحی می‌شود. در تمام مدل‌های شبکه عصبی مصنوعی، تابع خطی به‌عنوان تابع انتقال خروجی استفاده می‌شود و پارامترهای دیگری مانند تعداد ورودی، تعداد گره‌ها و تعداد دورهای آموزشی به‌طور هم‌زمان بهینه می‌شوند. در نهایت از بین همه ساختارهای شبکه عصبی تعیین شده، مدل شبکه عصبی بهینه با استفاده از کم‌ترین مقدار MSE/RMSE مربوط به مجموعه ارزیابی انتخاب می‌شود و از مدل شبکه عصبی بهینه برای پیش‌بینی مقادیر پاسخ ترکیبات مجموعه آزمون استفاده خواهد شد.

[1]Epoch
[2]Levenberg –Marquardt
[3]Bayesian regularization

۱- ۵- ۷- ۲ چیدمان توصیف‌کننده‌های منتخب به‌عنوان ورودی مدل شبکه عصبی مصنوعی

توصیف‌کننده‌های مهم و دارای ضرایب غیر صفر انتخاب شده به‌وسیله روش‌های رگرسیون انقباضی بر اساس ترتیب ورودی ماتریس توصیف کننده‌ها انتخاب می‌شوند. از آن‌جایی‌که برای مدل‌سازی شبکه عصبی مصنوعی و بهینه‌سازی پارامترهای شبکه عصبی ترتیب ورودی توصیف‌کننده‌های منتخب حائز اهمیت است، بنابراین باید برای این مسئله، تدبیری اندیشه شود تا بهترین زیر مجموعه به مدل معرفی شود. با توجه به تعداد زیاد این زیرمجموعه‌ها که از حاصل‌ضرب تعداد حالات توصیف کننده‌ها (دوتایی، سه‌تایی، n-1 تایی که n برابر با تعداد توصیف‌کننده‌های منتخب است) در تعداد گره و تعداد دور آموزشی به‌دست می‌آید، طراحی و استفاده از تمام تولیدات تصادفی به‌عنوان ورودی ANN عملاً غیرممکن است. از این‌رو در این مطالعه از دو روش برای چیدمان توصیف‌کننده‌های منتخب در مدل ANN استفاده می‌شود.

- **چیدمان توصیف کننده‌ها بر اساس بزرگی ضرایب استاندارد شده**

در روش اول ضرایب استاندارد نشده توصیف کننده‌ها با استفاده از اجرای روش‌های رگرسیونی انقباضی به‌دست می‌آید و در نهایت، ضرایب استاندارد نشده در نسبت بین انحراف معیار متغیر مستقل مربوطه و انحراف معیار متغیر وابسته ضرب می‌شود تا ضرایب رگرسیونی از بزرگی مقادیر توصیف کننده‌ها مستقل شود. بنابراین ضرایب استاندارد شده با استفاده از رابطه ۱-۱۱ محاسبه می‌شود [۴۱]:

رابطه ۱-۱۱

$$\beta_{\text{شده استاندارد}} = \beta_{\text{نشده استاندارد}} \times \frac{S_{\text{مستقل متغیر}}}{S_{\text{وابسته متغیر}}}$$

در رابطه ۱-۱۱ صورت کسر نشان‌دهنده انحراف استاندارد محاسبه شده برای مقادیر هر توصیف کننده به‌عنوان متغیر مستقل و مخرج کسر به‌معنی انحراف استاندارد مربوط به متغیر وابسته می‌باشد. پس از محاسبه ضرایب استاندارد شده، توصیف کننده‌ها بر اساس بزرگی قدر مطلق ضرایب استاندارد شده چیده و به‌عنوان ورودی شبکه عصبی مصنوعی تعریف و بهینه‌سازی می‌شوند.

- **پچیدمان توصیف کننده‌ها بر اساس اهمیت در شبکه عصبی مصنوعی**

روش دیگر برای ایجاد چیدمان منطقی، استفاده از روش شبکه عصبی است. به‌طوری‌که ابتدا مدل شبکه عصبی با تعداد کل توصیف‌کننده‌های منتخب آموزش داده می‌شود و پارامترهای دیگر شبکه عصبی از جمله گره، دور آموزشی، تابع انتقال و تابع آموزش بهینه به‌دست می‌آید. سپس برای تخمین اهمیت توصیف کننده i ام، همه توصیف‌کننده‌های منتخب روش انتخاب متغیر انقباضی تصادفی می‌شوند. مدل شبکه عصبی بهینه هر بار با استفاده ماتریس ورودی دست‌کاری شده در حضور توصیف کننده i با مقادیر تصادفی، آموزش داده می‌شود و مقادیر پاسخ مجموعه ارزیابی پیش‌بینی می‌شود. مقدار MSE/RMSE با استفاده از پاسخ‌های پیش‌بینی‌شده و واقعی مجموعه ارزیابی محاسبه می‌شود. این فرآیند تا زمانی‌که مقادیر هر کدام از توصیف کننده‌ها یک‌بار با مقادیر تصادفی جایگزین و مقدار خطای ارزیابی محاسبه شود، تکرار می‌شود. در نتیجه، به تعداد توصیف‌کننده‌های منتخب، مقدار MSE/RMSE به دست خواهد آمد. بالاترین MSE/RMSE به این معنی است که مدل ANN در غیاب آن توصیف کننده با مقادیر واقعی، خطای بیش‌تری را متحمل می‌شود و چنین توصیف کننده‌ای بیش‌ترین اهمیت را در ساخت مدل دارد. بنابراین، توصیف کننده‌ها بر اساس اهمیت محاسبه شده (مقدار MSE/RMSE) مرتب می‌شوند. در نهایت پارامترهای مدل شبکه عصبی با استفاده از توصیف‌کننده‌های چیده شده بر اساس اهمیت در شبکه عصبی مصنوعی بهینه می‌شوند. برای درک بهتر مسئله، لازم به ذکر است که زیرمجموعه اول شامل توصیف‌کننده‌های مهم اول و دوم (دو توصیف کننده با بیش‌ترین MSE/RMSE مجموعه ارزیابی مدل شبکه عصبی) است و زیرمجموعه‌های بعدی با افزودن توصیف‌کننده‌های دیگر بر اساس ترتیب اهمیتشان به زیرمجموعه اول ایجاد می‌شوند. بنابراین، در بهینه‌سازی شرایط شبکه عصبی مصنوعی، تعداد ورودی‌ها با استفاده از زیرمجموعه‌هایی شامل ۲ تا تعداد کل توصیف‌کننده‌های منتخب به‌عنوان ورودی شبکه عصبی مصنوعی، تعریف شده و به‌طور هم‌زمان همراه با سایر پارامترهای شبکه عصبی بهینه می‌شوند. در این رساله مدل با

استفاده از هر دو روش تعیین اهمیت توصیف کننده‌ها، توسعه یافته‌اند و نتیجه مربوط به بهترین روش به‌عنوان اساس تعیین اهمیت برتر گزارش شده است و سایر مراحل QSAR/QSPR ادامه یافته است.

۱- ۵- ۸ ارزیابی مدل

هدف اصلی هر مدل‌سازی QSAR/QSPR قابل اعتماد و دقیق، ایجاد یک مدل قوی با ظرفیت بالای پیش‌بینی فعالیت ترکیبات جدید است. از این‌رو ارزیابی مدل توسعه یافته در مطالعات QSAR/QSPR از مراحل اصلی ساخت مدل به‌شمار می-رود. ارزیابی مدل از طریق شاخص‌های کمی و آنالیزهای آماری متفاوتی انجام پذیر است. از دسته روش‌های ارزیابی مدل می‌توان به محاسبه پارامترهای آماری برای مجموعه ارزیابی خارجی (آزمون) و مجموعه پیش‌بینی شده به‌وسیله تکنیک رد مرحله‌ای تک تک[1](LOO)، نمودار باقی‌مانده‌های استاندارد شده، نمایش دامنه کاربرد[2]مدل و آزمون Y- تصادفی اشاره کرد.

۱- ۵- ۸- ۱ بررسی مدل با استفاده از نتایج پیش‌بینی شده برای مجموعه آزمون

برای ارزیابی مدل توسعه یافته، از پیش‌بینی داده‌های مجموعه آزمون که در هیچ یک از مراحل انتخاب متغیر و مدل‌سازی حضور نداشته‌اند، استفاده می‌شود. به‌طوری‌که ترکیبات مجموعه ارزیابی خارجی (آزمون) با استفاده از مدل‌های شبکه عصبی بهینه، پیش‌بینی می‌شوند. مقادیر خطای حاصل از پیش‌بینی و نتایج پارامترهای آماری قابل قبول، صحت و دقت مدل ساخته شده شبکه عصبی جفت شده با روش‌های انتخاب متغیر انقباضی را اثبات می‌کند.

[1]Leave-one-out
[2]Applicability Domain

۱- ۵- ۸- ۲ ارزیابی مدل با استفاده از تکنیک رد مرحله‌ای تک تک

تکنیک LOO با به‌کارگیری کل مجموعه داده‌ها برای بررسی استواری[1]مدل به‌کار گرفته می‌شود، به این معنی که پایداری مدل را با تغییر در داده‌های مدل اثبات می‌کند. بنابراین شبکه عصبی بهینه برای پیش‌بینی همه داده‌ها، مورد استفاده قرار می‌گیرد. با این تفاوت که در این تکنیک هر داده مورد مطالعه یک‌بار به‌عنوان داده آزمون خارج شده و مدل با مابقی داده‌ها آموزش داده می‌شود و پیش‌بینی پاسخ مربوط به داده خارج شده با استفاده از مدل بهینه به‌دست می‌آید. این عملیات به تعداد کل داده‌ها تکرار شده تا پاسخ همه داده‌ها یک‌بار پیش‌بینی شود. با محاسبه پارامترهای آماری برای نتایج حاصل از تکنیک رد مرحله‌ای تک استواری و استحکام مدل‌های QSAR و QSPR توسعه یافته تأیید می‌شود.

۱- ۵- ۸- ۳ نمودار باقی‌مانده‌های استاندارد شده

علاوه بر روش‌های ارزیابی یاد شده، نمودار باقی‌مانده استاندارد شده نیز، برای بررسی نتایج پیش‌بینی شده با استفاده از مدل‌های QSAR/QSPR بهینه، مورد استفاده قرار می‌گیرد. بنابراین مقادیر باقی‌مانده‌های استاندارد شده (r_i) با استفاده از رابطه ۱-۱۲، برای نتایج پیش‌بینی شده مجموعه آزمون و تکنیک LOO به‌طور مجزا محاسبه شده و نمودار باقی‌مانده‌های استاندارد شده، از رسم r_i بر حسب مقادیر واقعی پاسخ به‌دست می‌آید.

$$r_i = \frac{e_i}{s_{e_i}} = \frac{(y_i - \hat{y}_i)}{s_{e_i}}$$ رابطه ۱-۱۲

که در آن e_i تفاوت بین پاسخ‌های واقعی و پیش‌بینی شده برای هر مشاهده است و $i = ۱،...،n$ و s_{ei} انحراف استاندارد مقادیر باقی‌مانده است. توزیع یکنواخت داده‌ها حول محور صفر نشان‌دهنده عدم وجود خطای سیستماتیک در مدل‌های توسعه یافته است و برازش مناسب مدل‌های توسعه یافته QSAR/QSPR را ثابت می‌کند.

[1]Robustness

۱- ۵- ۸- ۴ پارامترهای آماری

پارامترهای آماری دسته‌ای از شاخص‌های کمی آماری هستند که به‌وسیله آن‌ها صحت نتایج ارائه شده توسط مدل اندازه‌گیری می‌شود. پارامترهای آماری محاسبه شده در این رساله در جدول ۱-۱ آورده شده است. پارامترهای آماری از نوع خطا دارای محدوده پذیرش خاصی نیستند، بلکه هر چقدر پارامترهای خطای مدل‌های توسعه یافته کم‌تر باشد، آن مدل از برتری قابل توجهی نسبت به سایر مدل‌ها برخوردار است. از جمله پارامترهای ارزیابی دیگر می‌توان به ضرایب تعیین[1] R^2 و Q^2_{LOO} اشاره کرد. مقادیر قابل قبول ضرایب رگرسیونی R^2و Q^2_{LOO}، نشان‌دهنده قدرت پیش‌بینی مدل هستند. به‌طوری‌که هر چه این پارامترها به مقدار ۱ نزدیک‌تر باشند، نشان‌دهنده این است که مقادیر پیش‌بینی شده با خطای کمی نزدیک به مقادیر واقعی می‌باشند. مقدار R^2 بزرگ‌تر از ۰/۶ و نزدیک به ۱ از مشخصه‌های لازم برای یک مدل QSAR/QSPR توسعه یافته می‌باشد. Q^2_{LOO} از نظر فرمول محاسباتی با پارامتر R^2 برابری دارد و تنها تفاوت آن در نوع پیش‌بینی مقادیر پاسخ است که با استفاده از روش ارزیابی تقاطعی[2]به روش رد مرحله‌ای تک تک (بخش ۱- ۵- ۸- ۲) پیش‌بینی می‌شود. مقدار Q^2_{LOO} بزرگ‌تر از ۰/۵ و نزدیک‌تر به ۱، برای تأیید اعتمادپذیری و قدرت پیش‌بینی مناسب مدل، ضروری است.

-محاسبه پارامترهای آماری تروپشا[3]

محققین در یافته‌اند که پارامترهای خطا و مقادیر ضرایب رگرسیونی قابل قبول، به‌تنهایی نمی‌توانند برتری و قدرت پیش‌بینی مدل‌های QSAR/QSPR توسعه یافته را اثبات نمایند [۴۲, ۴۳]. از این‌رو تروپشا، پارامترهای ارزیابی دیگری که وابسته به ضریب تعیین (R^2) مدل برتر است را پیشنهاد کرد. همان‌طور که

[1]Determination Coefficients
[2]Cross validation
[3]Tropsha

گفته شد، R^2 به‌عنوان یک پارامتر ارزیابی پر کاربرد، از رسم پاسخ‌های پیش‌بینی شده بر حسب پاسخ‌های تجربی به‌دست می‌آید. پارامتر R_0^2 ضریب تعیین حاصل از رسم پاسخ‌های پیش‌بینی شده بر حسب پاسخ‌های تجربی با عرض از مبدأ صفر است. $R_0'^2$ ضریب تعیین حاصل از رسم پاسخ‌های تجربی بر حسب پاسخ‌های پیش‌بینی شده با عرض از مبدأ صفر است. از طرفی، k شیب معادله رگرسیون برای نمودار پاسخ‌های پیش‌بینی شده بر حسب پاسخ‌های تجربی با عرض از مبدأ صفر و ´k شیب معادله رگرسیون برای نمودار پاسخ‌های تجربی بر حسب مقادیر پاسخ‌های پیش‌بینی شده با عرض از مبدأ صفر است و وجود شیب معادله رگرسیون بین دو حد ۰/۸۵ و ۱/۱۵ حاکی از برتری مدل توسعه یافته QSAR/QSPR است.$R_{0.R}^2$ و$R_{0.R}'^2$ نشان‌دهنده میزان نزدیکی پارامترهای R_0^2،R^2 و یاR^2،$R_0'^2$ به هم هستند. نزدیکی این پارامترها به ضریب تعیین مدل و تفاضل کم‌تر از ۰/۱ ، اعتماد پذیری و تعمیم‌پذیری بالای مدل را اثبات می‌کند. فرمول محاسباتی همه پارامترهای یاد شده، در جدول ۱-۱ آورده شده است.

-محاسبه پارامترهای آماری روی[1]

محقق دیگری به نام روی، سال‌ها بعد پارامترهای آماری تروپشا را گسترش داد و برخی پارامترهای دیگر را از روی پارامترهای تروپشا معرفی نمود [۴۴]. پارامترهای $R_m'^2$ و R_m^2 پارامترهای آماری دیگری هستند که اختلاف بین R^2 و R_0^2 را در مدل نشان می‌دهند. اگر مدل توسعه یافته دارای R_m^2 بیش‌تر از ۰/۵ باشد، مدل از اعتماد پذیری مناسب و قدرت پیش‌بینی خوبی برخوردار است. فرمول محاسباتی این پارامتر در جدول ۱-۱ آورده شده است. R-R پارامتر حاصل از تفاضل دو پارامتر $R_m'^2$ و R_m^2 است. مقدار کوچک‌تر از ۰/۳ این پارامتر قابل قبول است و تأیید دیگری بر مدل QSAR/QSPR توسعه یافته است.

[1]Roy

جدول ۱-۱ پارامترهای آماری

ردیف	پارامتر آماری	فرمول محاسباتی	محدوده قابل قبول
۱	PRESS	$\sum(y_i - \hat{y}_i)^2$	-
۲	SEP	$\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}}$	-
۳	MAE	$\frac{\sum_{i=1}^{n}\lvert y_i - \hat{y}\rvert}{n}$	-
۴	REP(%)	$\frac{100}{\bar{y}} \times \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}}$	-
۵	MSE	$\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}$	-
۶	MRE	$\frac{\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}}{y_i}\right\rvert}{n} \times 100$	-
۷	R^2	$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	R^2>۰/۶
۸	Q^2_{Loo}	R^2 مربوط به داده‌های پیش‌بینی شده بر اساس ارزیابی تقاطعی به تکنیک LOO	$\boldsymbol{Q^2_{Loo}}$> ۰/۵
۹	R_0^2	ضریب تعیین داده‌های پیش‌بینی شده بر حسب تجربی با عرض از مبدأ صفر	نزدیک به R^2
۱۰	R_0^2 نسبی	$R^2_{0.R} = \frac{(R^2 - R_0^2)}{R^2}$	<۰/۱
۱۱	R_m^2	$R^2 \times [1 - (R^2 - R_0^2))^{\frac{1}{2}}]$	>۰/۵
۱۲	$R_0'^2$	ضریب تعیین داده‌های تجربی بر حسب پیش‌بینی شده با عرض از مبدأ صفر	نزدیک به R^2
۱۳	$R_0'^2$ نسبی	$R'^2_{0.R} = \frac{(R^2 - R_0'^2)}{R^2}$	<۰/۱
۱۴	$R_m'^2$	$R^2 \times [1 - ((R^2 - R_0'^2))^{\frac{1}{2}}]$	>۰/۵
۱۵	R-R	$\lvert R_m^2 - R_m'^2\rvert$	<۰/۳
۱۶	k	شیب نمودار داده‌های پیش‌بینی شده بر حسب تجربی با عرض از مبدأ صفر	۰/۸۵ ≤ $\boldsymbol{k}$ ≤ ۱/۱۵
۱۷	k'	شیب نمودار داده‌های تجربی بر حسب پیش‌بینی شده با عرض از مبدأ صفر	۰/۸۵ ≤ $\boldsymbol{k'}$ ≤ ۱/۱۵

۱- ۵- ۸- ۵ دامنه کاربرد مدل

یکی دیگر از روش‌های ارزیابی استحکام و اعتمادپذیری مدل‌های QSAR/QSPR توسعه یافته، آنالیز دامنه کاربرد است. دامنه کاربرد، یک فضای شیمیایی تئوری است که با استفاده از توصیف‌کننده‌های مولکولی مجموعه آموزش و پاسخ تجربی و پیش‌بینی شده مربوطه ایجاد می‌شود. بنابراین اگر داده شیمیایی جدید نیز در این فضای شیمیایی تئوری قرار گیرد، نشان‌دهنده این است که مدل توانسته در مواجهه با داده‌ای که تاکنون در مدل حضور نداشته، آن‌را به‌خوبی پیش‌بینی نماید. دامنه کاربرد با استفاده از محاسبه پارامتر Leverage (H) مجموعه داده‌های آموزش به‌دست می‌آید. مقادیر H هر ترکیب با استفاده از انعکاس پاسخ واقعی بر پاسخ پیش‌بینی شده محاسبه می‌شود. مقادیر H با توجه به رابطه ۱-۱۳ قابل محاسبه است:

$$H= x_i (X^TX)^{-1} x_i^T$$ رابطه ۱-۱۳

که در اینجا X ماتریس توصیف‌کننده‌های مولکولی داده‌های مجموعه آموزش می‌باشد و x_i بردار توصیف‌کننده‌های مربوط به هر داده است. T نیز نشان‌دهنده وارون ماتریس است. برای نمایش فضای شیمیایی دامنه کاربرد از نمودار ویلیام[1] استفاده می‌شود. به‌طوری‌که مقادیر باقی‌مانده‌ها با استفاده از مقادیر پیش بینی شده توسط مدل برتر و مقادیر واقعی محاسبه می‌شود. در نهایت مقادیر باقی‌مانده‌ها با استفاده از رابطه زیر استاندارد می‌گردد.

$$r_i = \frac{(y_i-\hat{y}_i)-\bar{y}}{s_{e_i}}$$ رابطه ۱-۱۴

که در آن y_i و $\hat{y}_i$ به‌ترتیب نشاندهنده پاسخ‌های واقعی و پیش‌بینی شده برای هر مشاهده است و $i=1,...,n$ و s_{ei} انحراف استاندارد مقادیر باقی‌مانده و $\bar{y}$ میانگین مقادیر واقعی پاسخ است. بنابراین، نمودار ویلیام از رسم مقادیر باقی‌مانده‌های استاندارد شده در مقابل مقادیر H رسم می‌شود. برای آنالیز دامنه کاربرد، داده‌های شیمیایی باید در دو محدوده اطمینان قابل قبول نمودار ویلیام قرار گیرند. اولین شرط قابل قبول،

[1] William's plot

قرارگیری و عدم تجاوز مقدار باقی‌مانده‌های استاندارد شده داده‌های شیمیایی در محدوده ۳ برابری بزرگ‌تر/ کوچک‌تری از انحراف استاندارد می‌باشد. علاوه بر این شرط، مقادیر محاسبه شده H نیز نباید از مقدار حد آستانه یا حد هشدار *h برابر با 3p/n بزرگ‌تر باشند. p در این معادله برابر با تعداد توصیف‌کننده‌های مدل به‌علاوه یک و n تعداد داده‌های مجموعه آموزش می‌باشد. بنابراین، قرارگیری داده‌های H، در محدوده‌های قابل قبول، استحکام و قابل اعتماد بودن مدل‌های QSAR/QSPR توسعه یافته را اثبات می‌کند [۴۵, ۴۶].

۱- ۵- ۸- ۶ آزمون Y- تصادفی

آزمون Y- تصادفی برای بررسی استحکام مدل و عدم وجود ارتباط تصادفی بین متغیرهای مستقل و متغیر وابسته به‌کار گرفته می‌شود. از این‌رو مقادیر پاسخ هدف در محدوده تغییرات پاسخ، ۱۰۰۰ بار تصادفی می‌شود. مدل شبکه عصبی بهینه با پاسخ‌های تصادفی توسعه داده می‌شود و پس از ۱۰۰۰ بار اجرا، پاسخ‌های مربوط به هر مجموعه با استفاده از مدل بهینه پیش‌بینی می‌شود. یک مدل QSAR/QSPR مطمئن باید دارای ضرایب تعیین بسیار کوچک‌تر از مقدار قابل قبول ۰/۶ در مدل‌های تصادفی باشد. مقادیر بسیار کوچک‌تر R^2 نشان‌دهنده این است که ارتباط بین توصیف کننده‌ها و مقادیر واقعی پاسخ، شانسی و تصادفی نبوده و یک رابطه منطقی ریاضی بین آن‌ها برقرار است و مدل توسعه یافته بهینه از استحکام و اعتمادپذیری مناسبی برخوردار است.

۱- ۵- ۸- ۷ بررسی مشارکت توصیف‌کننده‌های منتخب در شبکه عصبی

با توجه به مدل غیرخطی توسعه‌یافته، سهم مشارکت توصیف‌کننده‌های منتخب در مدل نهایی با استفاده از مدل بهینه شبکه عصبی پیشنهادی مورد بررسی قرار می‌گیرد. برای محاسبه درصد مشارکت هر توصیف کننده موجود در مدل بهینه، مقادیر هر توصیف کننده در محدوده تغییرات خاص خود تصادفی می‌شود. مدل‌های بهینه هر بار در حضور یک توصیف کننده با مقادیر تصادفی و سایر توصیف کننده‌ها با مقادیر واقعی خود توسعه داده می‌شوند. مدل‌های توسعه‌یافته برای پیش‌بینی مقادیر پاسخ مجموعه ارزیابی

مورد استفاده قرار خواهند گرفت. مقدار پارامتر میانگین قدر مطلق خطا (MAE) مجموعه ارزیابی در حضور توصیف کننده با مقادیر تصادفی، محاسبه می‌شود. این فرآیند برای همه توصیف کننده‌ها تکرار شده و به تعداد توصیف‌کننده‌های مدل MAE محاسبه می‌شود. درصد مشارکت هر توصیف کننده (C_i%)، با استفاده از رابطه ۱-۱۵ محاسبه می‌شود.

$$C_i=(\frac{MAE_i}{\sum MAE_i})\times 100$$ رابطه ۱-۱۵

۱- ۶ شبیه‌سازی داکینگ مولکولی

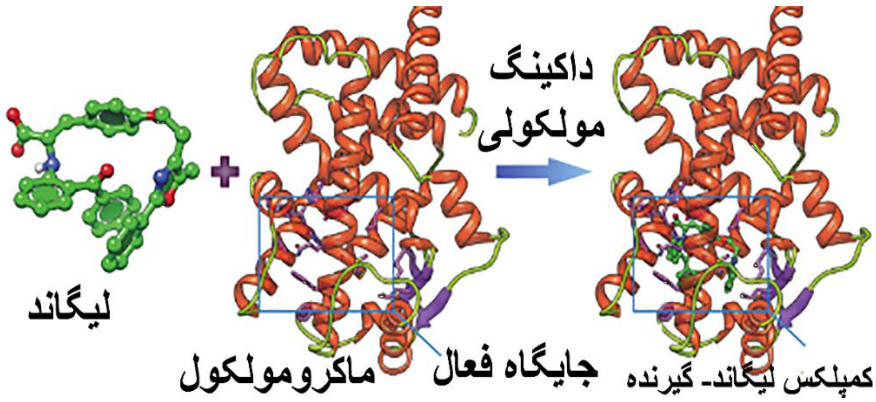
بهبود در فرایند طراحی دارو یکی از بحث‌های چالش برانگیز در دنیای علم و تکنولوژی امروزی به‌شمار می‌رود. اصولاً در طراحی دارو از سه محیط برای بررسی ترکیب پیشنهادی استفاده می‌شود:

- In silico: بررسی ترکیب شیمیایی بالقوه دارویی با استفاده از ابزارهای نرم‌افزاری
- In vitro: بررسی ترکیب شیمیایی بالقوه دارویی در محیط کشت سلولی
- In vivo: بررسی ترکیب شیمیایی بالقوه دارویی در بدن جانوران مانند موش آزمایشگاهی

با توجه به محدودیت‌های اشاره شده در بخش مقدمه، شبیه‌سازی‌های کامپیوتری به دلیل هزینه پایین، زمان و نیروی انسانی مورد نیاز کم‌تر، جایگاه ویژه‌ای را نسبت به آنالیزهای آزمایشگاهی به خود اختصاص داده است. شبیه‌سازی دقیق و اصولی عملکرد یک ترکیب شبه دارویی با استفاده از شبیه‌سازی داکینگ مولکولی[1] دینامیک مولکولی[2] و ... میسر است [۴۷, ۴۸]. از این‌رو برهم‌کنش‌های متفاوت ترکیبات شیمیایی بالقوه و ماکرومولکول (آنزیم‌ها، پروتئین‌ها و DNA) با استفاده از نرم‌افزار داکینگ مولکولی امکان‌پذیر است.

[1]Molecular docking simulation
[2]Molecular Dynamics (MD)

داکینگ یک الگوریتم خودکار کامپیوتری است که نحوه اتصال ترکیب به جایگاه فعال پروتئین را مشخص می‌کند. این روش شامل تعیین جهت‌گیری[1] و موقعیت[2] ترکیب، ساختار هندسی کنفورماسیونی و امتیازدهی می‌باشد. امتیازدهی می‌تواند معیار اندازه‌گیری انرژی اتصال، انرژی آزاد یا یک معیار عددی باشد. هر الگوریتم خودکار داکینگ به نحوی تلاش می‌کند تا ترکیب را در جهت‌گیری‌ها و کنفورماسیون‌های متفاوت در جایگاه فعال قرار دهد و امتیازی را برای هرکدام محاسبه کند. با استفاده از دانش به‌دست آمده از مطالعات داکینگ، به سنتز و بررسی ترکیبات کم‌تری نیاز خواهد شد. دلیل اصلی برای استفاده از داکینگ، پیش‌بینی ترکیباتی است که به‌خوبی به پروتئین متصل می‌شوند و علاوه بر این مشاهده ساختار هندسی ترکیب متصل شده به جایگاه فعال پروتئین است.



شکل (۱-۱) اتصال لیگاند –ماکرومولکول (گیرنده) با استفاده از مطالعه داکینگ مولکولی [۴۹]

۱- ۷ مراحل اجرای داکینگ مولکولی

جهت اجرای داکینگ مولکولی و بررسی برهم‌کنش‌های ترکیبات فعال رعایت دو مرحله اساسی الزامی است:

[1]Orientation
[2]Position

- اجرای فرایند اعتبار سنجی[1]یا داک- ریداک قبل از فرایند داکینگ مولکولی که با استفاده از داکینگ لیگاند کریستالوگرافی در ساختار گیرنده در تعداد اجراهای الگوریتم ژنتیک متفاوت انجام می‌شود و داک-ریداک با کم‌ترین ریشه میانگین مربعات انحراف [2](RMSD کم‌تر از ۱ آنگستروم)، حداقل تعداد خوشه و بیش‌ترین تعداد پیکربندی ترکیب در خوشه اول به‌عنوان شرط بهینه داکینگ انتخاب می‌شود.
- اجرای فرایند داکینگ مولکولی لیگاندهای فعال مجموعه داده‌ها و ترکیبات جدید پیشنهادی در جایگاه فعال ساختار گیرنده، در شرایط بهینه به‌دست آمده از فرایند داک- ریداک

برای اجرای فرایند داکینگ و فرایند اعتبار سنجی انجام مراحل زیر ضروری است.

۱- ۷- ۱ آماده‌سازی پروتئین

صحت نتایج داکینگ به‌طور مستقیم به کیفیت ساختار کریستالوگرافی جایگاه فعال پروتئین وابسته است. ساختار کریستالوگرافی پروتئین‌ها با اشعه ایکس مشخص می‌شوند. اگر ساختار همراه با لیگاند کمپلکس شده با پروتئین و یا بدون لیگاند و کوفاکتور و فقط دارای بخش پروتئینی در جایگاه فعال فراهم شوند، به‌ترتیب ساختارهای "هولو"[3]و ساختار "آپو"[4]نام دارند. ساختارهای هولو مهم‌ترین روش شناسایی جایگاه فعال پروتئین هستند. مرکز ثقل لیگاندی که در پروتئین است همان مرکز تقریبی جایگاه فعال پروتئین می‌باشد. از این‌رو، ابتدا لیگاند و مولکول‌های آب موجود در ساختار کریستالوگرافی را حذف کرده

[1]Validation
[2]Root-mean-square deviation
[3]Soaked structure
[4]Apo structure

و با فرمت بانک اطلاعاتی پروتئین[1]ذخیره می‌کنند و پس از ورود به محیط برنامه، هیدروژن‌هایی را که در ساختار کریستالوگرافی پروتئین دیده نمی‌شوند، اضافه می‌نمایند.

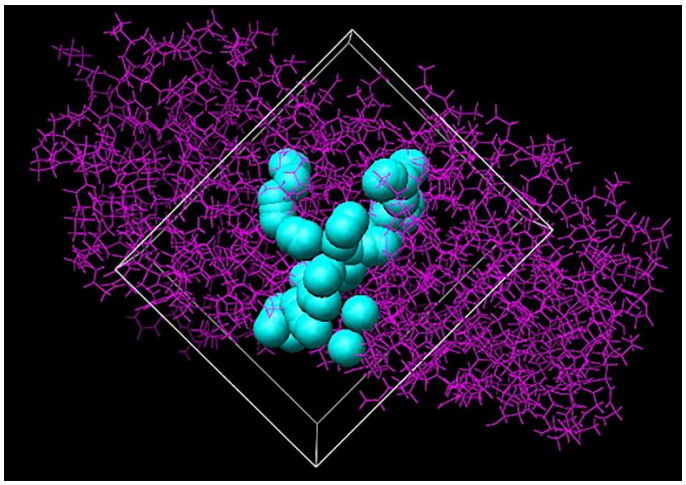
۱- ۷- ۲ ساختن لیگاند

لیگاند در برنامه داکینگ فراخوانی شده و در جایگاه فعال پروتئین هدف قرار می‌گیرد و سپس هر شبیه‌سازی داکینگ به‌طور جداگانه برای همه لیگاندهای مورد نظر توسط کاربر انجام می‌شود.

۱- ۷- ۳ تنظیم کردن جعبه شبکه‌ای[2]

در داکینگ مولکولی نمی‌توان کل فضا را به‌عنوان جایگاه فعال گیرنده در نظر گرفت. باید بخشی از فضا که لیگاند درون جایگاه فعال قرار می‌گیرد را برای فرایند داکینگ تعریف کرد تا هم دقت بالا باشد و هم‌زمان انجام فرایند داکینگ به‌صرفه باشد. به‌منظور سرعت بخشیدن به محاسبات، یک فاصله محدود کننده[3]تنظیم می‌شود، این فاصله محدود کننده معمولاً یک جعبه مستطیل شکل به نام جعبه شبکه‌ای است. بخش‌هایی از پروتئین که دور از جایگاه فعال هستند، به‌طور معمول هیچ اثر قابل اندازه‌گیری بر نتایج امتیازدهی ندارند و از این‌رو ایجاد شبکه با ابعاد یکسان برای پوشش دهی اسیدهای آمینه جایگاه فعال پروتئین توصیه می‌شود. اگر فایل مربوط به ساختار کریستالوگرافی هولو باشد به این معنی که دارای لیگاند کمپلکس شده باشد، جایگاه فعال پروتئین مشخص خواهد بود و مختصات جعبه شبکه‌ای همان مختصات جایگاه فعال و به شکل مکعبی با ابعاد یکسان تعریف می‌شود. ابعاد جعبه باید طوری انتخاب شود که اسیدهای آمینه جایگاه فعال را در بر بگیرد. شکل (۱-۲) جعبه شبکه‌ای در برگیرنده اسیدهای آمینه را نشان می‌دهد [۵۰, ۵۱].

[1]Protein data bank
[2]Grid box
[3]Cut off



شکل (۱-۲) جعبه شبکه‌ای [۵۲]

۱- ۷- ۴ گزینه‌های داکینگ

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

۱- ۷- ۵ انجام محاسبه داکینگ

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

۱- ۷- ۶ آنالیز و تحلیل نتایج

مهم‌ترین نتیجه محاسبات داکینگ، انرژی اتصال لیگاند به جایگاه فعال است. این مقداری است که برای تعیین بهترین ترکیب مهار کننده، بین ترکیبات مختلف مقایسه می‌شود. چند حالت و وضعیت از لیگاند در جایگاه فعال که با بهترین انرژی اتصال همراه است، بررسی می‌شود تا از معقول و مناسب بودن

آن اطمینان حاصل گردد. گاهی اوقات، وضعیتی که توسط یک محاسبه داکینگ تولید می‌شود، به محقق ایده‌ای برای چگونگی تغییر ترکیبات در دوره بعدی محاسبات را می‌دهد.

دو مؤلفه کلیدی الگوریتم جستجو[1] و الگوریتم رتبه‌بندی یا امتیازدهی[2] در برنامه داکینگ وجود دارد. الگوریتم جستجو، مولکول را در موقعیت‌ها و صورت‌بندی‌های متفاوت در جایگاه فعال پروتئین قرار می‌دهد. انتخاب الگوریتم جستجو، مشخص می‌کند که برنامه با چه صحتی در طول فرایند، موقعیت‌های ممکن مولکول را ارزیابی می‌کند و چه مدت زمانی برای آن نیاز است. قابل ذکر است که الگوریتم جستجو صحت نتایج به‌دست آمده از برنامه داکینگ را مشخص نمی‌کند. الگوریتم امتیازدهی مسئول تعیین این است که آیا جهت‌گیری‌های انتخاب شده توسط الگوریتم جستجو، از لحاظ انرژی مناسب‌ترین هستند و مسئول محاسبه انرژی اتصال است.

۱- ۷- ۷ کاربردهای داکینگ مولکولی

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

۱. غربال‌گری مجازی[3]

۲. بهبود طراحی و کشف دارو

۳. پیش‌بینی انرژی آزاد اتصال

۴. بررسی برهم‌کنش لیگاند- گیرنده

متداول‌ترین کاربرد داکینگ مولکولی در اتصال پروتئین-لیگاند است. هدف نهایی اتصال پروتئین-لیگاند پیش‌بینی فعالیت بیولوژیکی لیگاند و برهم‌کنش لیگاند با پروتئین می‌باشد و در این پروژه از این

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

هدف (اتصال لیگاند-پروتئین) جهت بررسی برهم‌کنش‌های ترکیبات پیشنهادی بالقوه در جایگاه فعال گیرنده استفاده می‌شود.

۱- ۸ ضرورت و اهمیت مدل‌سازی QS(A/P)R ترکیبات شیمیایی با استفاده از مدل‌های شبکه عصبی توسعه یافته با توصیف کننده‌های منتخب روش‌های جریمه‌ای و مروری بر کارهای انجام شده

با توجه به توضیحاتی که تا این بخش در مورد مراحل مدل‌سازی QSPR/QSAR داده شد، چالش‌هایی که یک شیمیدان محاسباتی با آن مواجه است به‌خوبی مشخص شد. به‌طور معمول، در مراحل اولیه مطالعات QSPR/QSAR، تعداد زیادی توصیف کننده مولکولی تولید می‌شود. به‌منظور ساخت یک مدل QSPR/QSAR با قابلیت پیش‌بینی و تفسیرپذیری بالا، یک روش انتخاب متغیر برای انتخاب زیر مجموعه بهینه از توصیف کننده‌ها الزامی است، به شرطی که مدل توسعه یافته بالاترین همبستگی بین مقادیر واقعی و پیش‌بینی‌شده ویژگی هدف را برای ترکیبات مورد استفاده در مدل فراهم سازد. انتخاب مؤثرترین توصیف کننده‌ها از انبوه زیادی از توصیف‌کننده‌های مولکولی محاسبه شده، همواره یک موضوع چالش برانگیز در ساخت مدل‌های QSPR/QSAR بوده است. در به‌کارگیری روش‌های انتخاب متغیر دو جنبه اصلی از جمله بهبود عملکرد پیش‌بینی برآوردگرها و ارائه تفسیر مناسبی از ارتباط بین متغیرهای مستقل و وابسته دنبال می‌شود. از این‌رو هدف از استفاده از روش‌های انتخاب متغیر در ساخت مدل‌های QSPR/QSAR، به‌دست آوردن مؤثرترین توصیف کننده‌ها با بیش‌ترین ارتباط با متغیر وابسته است. سؤال اساسی در ارائه مدل‌های QSPR/QSAR این است که چگونه می‌توان توصیف‌کننده‌های مؤثر را تعیین کرد تا به‌طور رضایت بخشی وابستگی بین ویژگی هدف و توصیف کننده‌ها را نشان دهد. یک مدل

QSPR/QSAR با کارایی بالا، باید به ازای استفاده از حداقل پارامترهای فیزیکوشیمیایی قدرت پیش‌بینی مناسبی را در مقابل ترکیبات شیمیایی جدید نشان دهد. از این‌رو استفاده از روش‌های انتخاب متغیر کارآمد برای کاهش تعداد توصیف کننده‌ها توصیه می‌شود. روش‌های انتخاب متغیر با حذف توصیف کننده‌های زائد باعث کاهش توصیف کننده‌ها و انتخاب توصیف‌کننده‌های مؤثر می‌شود و در نتیجه سبب افزایش تفسیرپذیری مدل می‌شود. روش‌های کاهش ابعاد داده‌های متفاوتی در ساخت مدل‌های QSPR/QSAR کاربرد دارد. از بین روش‌های توسعه یافته، روش‌های کلاسیک به‌دلیل داشتن معایبی از جمله بایاس بالا در تخمین ضرایب رگرسیونی، عملکرد پایین در حضور هم‌خطی، ناپایداری در برابر تغییر نمونه‌های آزمایشی و غیره کارآمد نیستند. بنابراین استفاده از روش‌های جریمه‌شده، به دلیل رفع محدودیت‌های یاد شده در بحث مدل‌سازی QSPR/QSAR بسیار مورد توجه می‌باشند. روش‌های انتخاب متغیر جریمه شده، با انقباض ضرایب متغیرهای کم اهمیت به صفر منجر به تولید مدل‌های QSPR/QSAR تنک با قابلیت تفسیرپذیری بالا می‌شود. بنابراین استفاده از روش‌های انتخاب متغیر انقباضی به دلیل داشتن برخی مزایای ذاتی، از جمله داشتن برآوردگرهایی با بایاس و خطای پیش‌بینی کم و پایداری و تنکی قابل توجه در توسعه مدل‌های QSPR/QSAR شدیداً توصیه می‌شود. از این‌رو در این رساله، کارایی روش‌های انتخاب متغیر انقباضی متفاوت در مجموعه داده‌های متفاوتی مورد ارزیابی قرار گرفت.

در ادامه بحث، اهمیت ساخت مدل‌های QSPR/QSAR و تحقیقات انجام شده در سال‌های اخیر پیرامون به‌کارگیری روش‌های انتخاب متغیر انقباضی در ساخت مدل‌های QSPR/QSAR، تحت بررسی قرار گرفت.

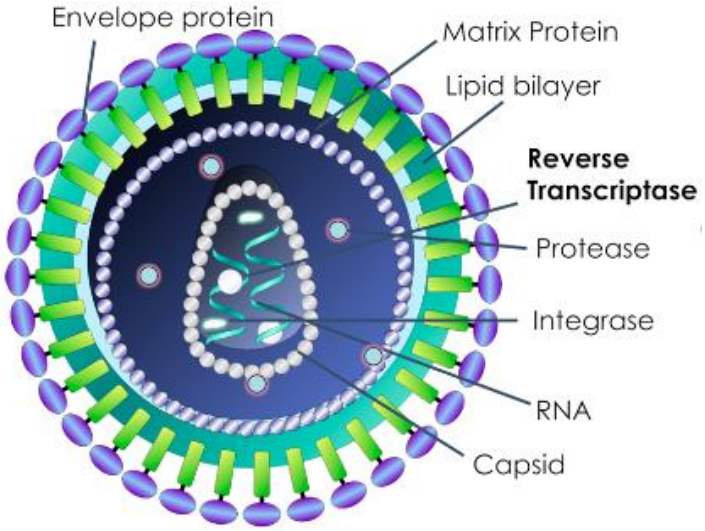
۱-۸-۱ اهمیت پیش‌بینی فعالیت دارویی بازدارنده‌های ایدز با استفاده از مدل SCAD-ANN

ویروس نقص ایمنی اکتسابی انسانی[1](HIV) باعث ایجاد بیماری ایدز (AIDS)[2]می‌شود و عامل مرگ و میر در سراسر جهان است. بر اساس داده‌های اپیدمی ایدز سازمان جهانی بهداشت[3]در سال ۲۰۲۰، حدود ۱ میلیون مرگ و میر به‌اضافه ۳۷/۷ میلیون نفر مبتلا به ایدز شناخته شده است [۵۵-۵۳]. طبق نتایج منتشر شده، اگرچه مرگ و میر ناشی از ایدز در سال ۲۰۰۵ (۱/۹ میلیون) به‌طور قابل توجهی در مقایسه با سال ۲۰۱۶ (۱ میلیون) کاهش یافته است، اما متأسفانه تعداد عفونت‌های جدید در هر سال ثابت می‌ماند. یکی از دلایل اصلی کاهش مرگ و میر ناشی از ایدز، معرفی درمان ترکیبی ضد رتروویروسی (cART)[4]است. با این‌حال، اکثر داروهای موجود در حال حاضر برای استفاده بالینی مستعد ایجاد مقاومت دارویی هستند. به‌طور خلاصه، cART چندین نقطه از چرخه همانندسازی را هدف قرار می‌دهد و معمولاً شامل بازدارنده‌های نوکلئوزیدی آنزیم نسخه‌بردار معکوس (NRTIs)، بازدارنده‌های غیر نوکلئوزیدی آنزیم نسخه‌بردار معکوس (NNRTIs) و بازدارنده‌های پروتئاز (PIs) می‌شود [۵۶].

تکثیر HIV یک فرآیند چند مرحله‌ای است و هر مرحله برای تکثیر موفقیت آمیز ویروس بسیار مهم است. به‌طور خلاصه، چرخه زندگی ویروس زمانی آغاز می‌شود که HIV به گیرنده $CD4^+$، موجود در سطح لنفوسیت T، متصل می‌شود. علاوه بر اتصال به گیرنده CD_4^+، HIV باید به یکی از گیرنده‌های کموکاین مانند CCR5 یا CXCR4 متصل شود تا وارد سلول شود (شکل ۱-۲). این گیرنده‌ها با پروتئین‌های پوششی ویروسی مانند گلیکوپروتئین‌های gp120 و پروتئین گذرنده gp41 تعامل دارند. هنگامی‌که HIV به سطح سلول میزبان نزدیک می‌شود، gp120 به گیرنده CD_4^+ متصل می‌شود، و این امر باعث تقویت

[1]Human immunodeficiency virus
[2]Acquired immunodeficiency syndrome
[3]World Health Organization AIDS Epidemic
[4]Combination antiretroviral therapy

بیش‌تر اتصال با گیرنده مرکزی می‌شود و منجر به تغییر ساختاری در gp120 می‌شود و به gp120 اجازه می‌دهد تا در غشای سلول میزبان باز شود و روی خود تا شود. نوکلئوکپسید ویروس وارد سلول میزبان می‌شود و رشته‌های RNA و آنزیم‌های کلیدی مانند نسخه بردار معکوس (RT)، اینتگراز (IN) و پروتئاز (PR) را آزاد می‌کند. برای تکثیر در داخل سلول میزبان، DNA پلیمراز ویروسی از پرایمر RNA میزبان، به‌ویژه tRNA، برای سنتز DNA ویروسی تک رشته‌ای استفاده می‌کند، سپس با یک الگوی مشخص ویروس هیبرید RNA:DNA تشکیل می‌شود. نسخه بردار معکوس (RT)، رشته RNA را از هیبرید حذف می‌کند. سپس DNA باقی‌مانده ویروس به DNA ویروسی دو رشته‌ای تبدیل می‌شود. متعاقباً، اینتگراز، DNA ویروس دو رشته‌ای را به هسته می‌برد و DNA ویروس را در ژنوم سلول میزبان ادغام می‌کند. فعال شدن سلول باعث رونویسی DNA به mRNA می‌شود. سپس mRNA ویروسی به سیتوپلاسم منتقل می‌شود، جایی که به‌عنوان طرحی برای ساخت بلوک‌های طولانی‌تر از ویروس جدید استفاده می‌شود [۵۴, ۵۷].



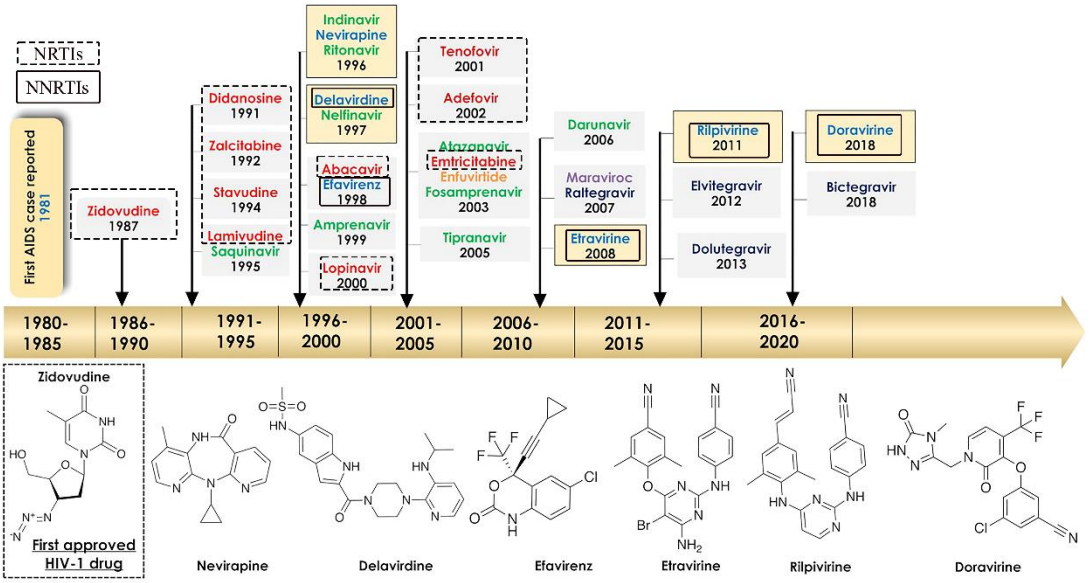
شکل ۱-۲ ساختار کلی HIV [۵۴]

داروهایی که در مراحل کلیدی تکثیر تداخل ایجاد می‌کنند، باعث توقف آن‌ها در مراحل زیر می‌شوند [۵۸].

- مسدود کردن ورود ویروس به سلول میزبان با مهارکننده‌های ورودی، از جمله مهارکننده‌های همجوشی[1](FIs)
- متوقف نمودن تکثیر با مهارکننده‌های نوکلئوزیدی نسخه بردار معکوس (NRTIs) و یا مهارکننده‌های نسخه بردار معکوس غیر نوکلئوزیدی (NNRTIs)
- مهار ادغام DNA ویروسی با ماده ژنتیکی میزبان با مسدود کردن فعالیت اینتگراز (InSTI)،
- انسداد بلوغ بیش‌تر ویروس با مهارکننده پروتئاز ((PI)

رونویسی معکوس یک DNA پلیمراز وابسته به RNA است که از یک رشته RNA برای سنتز DNA ویروسی دو رشته‌ای استفاده می‌کند [۵۹]. بر اساس محل اتصال و طبقه شیمیایی ترکیبات، مهارکننده‌های RT به NRTIs و NNRTIs طبقه‌بندی می‌شوند [۶۰, ۶۱]. شکل ۱-۳ نقطه عطف کوتاهی در توسعه داروهای HIV را نشان می‌دهد و همان‌طور که مشخص است، مهارکننده‌های مهم RT دارای برجستگی قابل توجهی می‌باشند. NNRTIs (nevirapine، delavirdine، efavirenz، etravirine، rilpivirine و doravirine) از دسته داروهای RT مورد تأیید سازمان غذا و داروی ایالات متحده[2]است. ترکیبات NNRTIs با قدرت بیش‌تر و سمیت کم‌تر همواره برای مقابله با HIV بسیار کارآمد هستند. از طرفی، با توجه به مقاومت دارویی سریع سویه‌های ویروسی جدید، پیشنهاد NNRTI ها با اثربخشی بهتر بسیار حائز اهمیت است [۵۴].

[1]Fusion Inhibitor
[2]U.S. Food and Drug Administration



شکل ۱-۳ جدول زمانی کوتاهی از توسعه داروهای HIV، داروهای شناسایی شده برای RT با خط چین (NRTIs) و خط (NNRTIs) جدا شده‌اند[۵۴]

بنابراین با توجه به اهمیت بازدارنده‌های غیرنوکلئوزیدی نسخه بردار معکوس در مقابله با HIV، پیشنهاد ترکیبات جدید NNRTIs با قدرت بیش‌تر، سمیت کم‌تر در امر طراحی دارو حائز اهمیت است. بنابراین در اولین بخش مطالعات تجربی این رساله تلاش شد و یک مدل QSAR برای پیش‌بینی فعالیت دارویی بازدارنده‌های NNRTIs توسعه یافت. همان‌طور که در بخش ۱- ۸ توضیح داده شد، انتخاب مؤثرترین توصیف کننده‌ها موضوع چالشی در ساخت مدل‌های QSAR است و استفاده از روش‌های انتخاب متغیر انقباضی به دلیل داشتن مزایای ذاتی متفاوت، حائز اهمیت است. بنابراین کارایی روش انتخاب متغیر انقباضی SCAD در ساخت مدل QSAR مورد بررسی قرار گرفت. بنابراین در ادامه و در بخش مروری بر کارها، به بررسی مطالعات و پروژه‌هایی که در سال‌های اخیر در مورد به‌کارگیری روش انتخاب متغیر SCAD در ساخت مدل‌های QSAR بحث شده است پرداخته خواهد شد.

۱- ۸- ۲ اهمیت پیش‌بینی فعالیت دارویی بازدارنده‌های SARS-COV-2 با استفاده از مدل ALASSO-ANN

سندرم حاد تنفسی ویروس کرونا[1](SARS-CoV-2)، به‌عنوان هفتمین ویروس کرونای انسانی، در ووهان، استان هوبی، چین، در طی اپیدمی اخیر ذات‌الریه در ژانویه ۲۰۲۰ کشف شد [۶۲, ۶۳]. از آن زمان، این ویروس در سراسر جهان گسترش یافته است و تا به الان، حدود ۲۵۸ میلیون عفونت و حدود ۵ میلیون قربانی داشته است [۶۴]. طبق آمار منتشر شده، SARS-CoV-2،SARS-CoV و کروناویروس سندرم تنفسی خاورمیانه (MERS-CoV)[2]به‌ترتیب باعث ذات‌الریه شدید با میزان مرگ و میر ۲/۹٪، ۹/۶٪ و ~۳۶٪ می‌شوند [۶۷-۶۵].

SARS-CoV-2 متعلق به طبقه بتا-کرونا ویروس از خانواده کروناویردایی[3]و از راسته نیدوویرالز[4] است. SARS-CoV-2 بسیار شبیه به کروناویروس‌های SARS مانند است [۶۲] و تقریباً ۹۶٪ دارای اشتراک ساختاری هستند. پروتئین‌های SARS-CoV-2 دو سوم ژنوم ویروس را اشغال می‌کنند و اهداف برجسته‌ای برای داروهای تعدیل کننده تکثیر ویروس هستند. این گروه شامل آنزیم‌های تکثیر و رونویسی ویروسی RNA پلیمراز، هلیکاز و پروتئاز شبه پاپائین ((PL^{pro})[5]همچون پروتئاز ۳C-مانند ($3CL^{pro}$ و یا M^{pro}) می‌شود [۶۸]. SARS-CoV-2 به‌عنوان یک ویروس RNA تک رشته‌ای، مستقیماً پس از عفونت، سلول از سیستم‌های ترجمه سلولی برای تولید پلی‌پروتئین‌های ویروسی استفاده می‌کند [۶۹]. غربالگری کتابخانه‌ای ترکیبات و مطالعات داکینگ مولکولی بر اساس ساختار کریستالوگرافی SARS-CoV و MERS، مهارکننده‌های $3CL^{Pro}$ متفاوت، عملکرد بالقوه‌ای را علیه SARS-CoV-2 نشان داده است [۷۰, ۷۱].

[1]Severe acute respiratory syndrome coronavirus
[2]Middle East respiratory syndrome coronavirus
[3]Coronaviridae
[4]Nidovirales
[5]Papain-like protease

بنابراین پیشنهاد ترکیبات $3CL^{pro}$ جدید با فعالیت دارویی بالقوه از اهمیت بالایی برخوردار است. از این‌رو ارائه یک مدل QSAR قابل اعتماد برای پیش‌بینی فعالیت دارویی ترکیبات $3CL^{pro}$ جدید توصیه می‌شود. در دومین بخش این رساله، یک مدل QSAR با استفاده از توصیف‌کننده‌های منتخب روش انتخاب متغیر انقباضی لاسو تطبیقی (ALASSO) برای پیش‌بینی فعالیت دارویی بازدارنده‌های $3CL^{pro}$ توسعه یافت.

۱- ۸- ۳ اهمیت پیش‌بینی فعالیت دارویی برخی از بازدارنده‌های ایدز و سرطان با استفاده از مدل LAD-LASSO-ANN

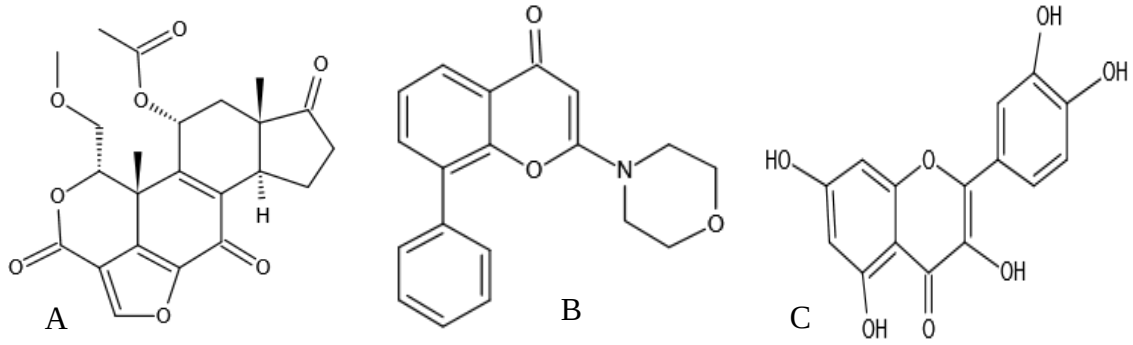
عادات غذایی نامناسب و سبک زندگی وابسته به میز، باعث بروز بیماری‌های مختلفی مانند سرطان، عوارض قلبی عروقی، دیابت و اختلالات ایمنی می‌شود [۷۲]. سرطان یک مشکل عمده بهداشت عمومی در سراسر جهان است و دومین عامل مرگ و میر در دنیا است. متأسفانه نرخ رشد آن در کشورهای توسعه یافته‌تر و در حال توسعه دیده می‌شود. در سال ۲۰۲۰، تشخیص و درمان سرطان به دلیل همه‌گیری بیماری کروناویروس ۲۰۱۹ با مشکل مواجه شد. به‌عنوان مثال، کاهش دسترسی به مراقبت به دلیل بسته شدن مراکز مراقبت‌های بهداشتی منجر به تأخیر در تشخیص و درمان می‌شود، که ممکن است منجر به کاهش کوتاه مدت در بروز سرطان و به دنبال آن افزایش در مرحله پیشرفته بیماری و در نهایت افزایش مرگ و میر شود. از این‌رو، ارائه دقیقی از آمار مبتلایان و مرگ و میر دقیق زمان‌بر است [۷۳]. طبق اطلاعیه سازمان جهانی بهداشت در سال ۲۰۲۰، سرطان حدود ۱۰ میلیون مرگ و میر را به دنبال داشته است و طبق تقسیم‌بندی انجام شده، سرطان سینه با ۲/۲۶ میلیون جمعیت، در رتبه اول مرگ و میر می‌باشد. سرطان ریه و سرطان روده به ترتیب با ۲/۲۱ و ۱/۹۳ میلیون مرگ و میر رتبه دوم و سوم سرطان پرخطر را به خود اختصاص داده‌اند. در ادامه سرطان پروستات، پوست و معده مجموعاً ۳/۵ میلیون مرگ و میر را به همراه داشته‌اند [۶۴]. با توجه به آمار اشاره شده حدود ۲۲ درصد و ۱۹ درصد از کل مرگ و میر مربوط به بیماری

سرطان به‌ترتیب به سرطان‌های ریه و روده تعلق دارد. نرخ بقای بیماران مبتلا به سرطان ریه و روده، علی‌رغم پیشرفت‌هایی که در درمان دارویی و تشخیص آن صورت گرفته است، هنوز بسیار پایین است بنابراین، تحقیق برای دستیابی به یک رویکرد درمانی مناسب ضروری است [۷۴].

فسفوئینوزیتید-۳-کینازها[1](PI3Ks)، کینازهایی لیپیدی هستند که در پاسخ به فاکتورهای رشد توسط تعداد معینی از پروتئین گیرنده تیروزین کیناز (RTKs)[2] یعنی گیرنده فاکتور رشد اپیدرمی (EGFR)[3] گیرنده فاکتور رشد اپیدرمی انسانی (HER2)[4]حساس به فعال شدن هستند. PI3K ها به سه طبقه اصلی تقسیم می‌شوند [۷۵]. شواهد نشان می‌دهد جهش فعال کننده PI3Kα معمولاً در بسیاری از سرطان‌ها مشاهده می‌شود. بنابراین، مهارکننده‌های مولکولی کوچک علیه PI3Kα برای درمان سرطان حائز اهمیت است [۷۶]. تحقیقات زیادی در مورد مهار کننده‌های PI3Kα وجود دارد و برخی از مهارکننده‌ها وارد آزمایش‌های بالینی نیز شده‌اند. چندین ماده شیمیایی، از جمله بازدارنده‌های نسل اول PI3K مانند ورتمانین[5](شکل ۱-۴ A)، به‌عنوان مهارکننده‌های PI3K به‌طور گسترده شناخته شده‌اند. مهارکننده‌های نسل دوم PI3K مولکول‌های کوچکی هستند که از جمله می‌توان به LY294002 (شکل ۱-۴ B) به‌عنوان آنالوگ کوئرستین[6](شکل ۲-۱ C) اشاره کرد [۷۷]. مطالعات سرطان شناسی، LY294002 و آنالوگ‌های آن را به‌عنوان قوی‌ترین مهارکننده‌های PI3K معرفی نموده است [۷۸, ۷۹]. طبق تحقیقات انجام شده مشتقات کرومنو [c-۴،۳] پیرازول۴-(۲H)-اون[7]حاوی زنجیره بنزیل و آلکیل به‌عنوان مهار کننده‌های بالقوه PI3Kα و از دسته آنالوگ‌های B شناخته شده هستند [۸۰]. این مشتقات در برابر سلول‌های سرطانی ریه

[1]Phosphoinositide-3-kinases
[2]Receptor protein tyrosine kinases
[3]Epidermal growth factor receptor
[4]Human epidermal growth factor receptor 2
[5]Wortmannin
[6]Quercetin
[7]Chromeno[4,3-c]pyrazol-4(2H)-one

(A549) و کولورکتال (HCT-116) مورد آزمایش قرار گرفته و فعالیت دارویی مناسب و اثرات ضد تکثیری قابل توجهی را نشان دادند [۸۱].



شکل ۱-۴ شکل بازدارنده‌های بالقوه PI3K.. A ورتمانین، B: LY294002 و C کوئرستین

بنابراین طبق توضیحات ارائه شده در مورد اهمیت بازدارنده‌های PI3K، پیش‌بینی فعالیت ترکیبات آنالوگ و پیشنهاد ترکیبات جدید سنتز نشده از اهمیت بسیاری برخوردار است. بنابراین در این مطالعه از اطلاعات ساختاری و فعالیت دارویی ضد سرطانی روده و ریه این مشتقات استفاده شد تا مدل‌های QSAR با قدرت پیش‌بینی قابل اعتماد معرفی شود. به‌منظور افزایش قدرت پیش‌بینی و تفسیرپذیری مدل، استفاده از یک روش انتخاب متغیر انقباضی کارآمد توصیه می‌شود. از این‌رو در سومین بخش این رساله از روش انتخاب متغیر انقباضی حداقل انحراف مطلق - حداقل قدر مطلق انقباض و عملگر انتخاب کننده (LAD-LASSO)[1] به دلیل مزایایی چون تنکی، مقاوم به وجود داده‌های دور افتاده و نزدیک بودن به یک مدل حقیقی (اوراکل) [۳۲] برای کاهش ابعاد داده‌ها قبل از مدل‌سازی با ANN استفاده شد. در ادامه مطالعات انجام شده در زمینه QSAR برای ساخت مدل‌های طبقه‌بندی و رگرسیونی مورد بررسی قرار گرفته است.

[1]Least absolute deviation (LAD)- least absolute shrinkage and selection operator (LASSO)

۱- ۸- ۴ اهمیت پیش‌بینی شاخص بازداری ترکیبات آلی فرار با استفاده از مدل SCAD-ANN

ترکیبات آلی فرار (VOCs)[1]مواد شیمیایی آلی هستند که به‌طور کلی دارای فشار بخار بالا و حلالیت پایین در آب هستند. ترکیبات مختلف زیادی وجود دارد که ممکن است به‌عنوان ترکیبات آلی فرار طبقه‌بندی شوند. بسیاری از VOC ها مواد شیمیایی ساخته شده توسط انسان هستند که در ساخت رنگ‌ها، مواد دارویی و خنک کننده‌ها استفاده و تولید می‌شوند. VOC ها آلاینده‌های رایج آب‌های زیرزمینی یا آلاینده‌های خطرناک هوا هستند [۸۲]. VOC ها همچنین نقش عمده‌ای در تشکیل آلاینده‌های ثانویه مختلف از طریق واکنش‌های فیتوشیمیایی در حضور نور خورشید و اکسیدهای نیتروژن دارند. علاوه بر این، برخی از VOC ها می‌توانند در تخریب لایه اوزون و ایجاد آلودگی‌های پایدار در مناطق نقش داشته باشد. بنابراین، این ترکیبات در طول دو دهه اخیر یک موضوع مهم زیست محیطی بوده و توجه گروه‌های مختلف تحقیقاتی را به خود جلب کرده است [۸۸-۸۳]. روش‌های تجزیه‌ای ویژه با حساسیت و گزینش پذیری بالا برای شناسایی و اندازه‌گیری VOC ها در اختیار دانشمندان قرار دارد که از جمله آن‌ها می‌توان به استفاده از کروماتوگرافی گازی در شناسایی و اندازه‌گیری شاخص بازداری این ترکیبات اشاره کرد [۸۸]. محدودیت‌هایی برای استفاده از این روش آزمایشگاهی وجود دارد. به‌طوری‌که در برخی موارد استانداردهای موجود با درجه خلوص بالا وجود ندارد یا در برخی موارد ممکن است این استانداردهای مورد نیاز در دسترس نباشند. عدم دسترسی به ستون‌های متفاوت جهت انجام آنالیز و یا عدم تکرارپذیری داده‌های اندازه‌گیری شده، همگی، از جمله محدودیت‌هایی است که به‌کارگیری روش‌های تئوری در پیش‌بینی شاخص بازداری این ترکیبات را حائز اهمیت ویژه‌ای می‌سازد [۹۱-۸۹].

[1]Volatile organic compounds

یک جایگزین معتبر برای مقابله با این محدودیت‌ها، ادغام اطلاعات ساختاری VOC های شناسایی‌شده و شاخص‌های بازداری تجربی آن‌ها به‌منظور ارائه یک مدل نظری برای تخمین RI مربوط به VOC های جدید است. بنابراین ارتباط کمی ساختار – ویژگی (QSPR)، یک روش امیدوارکننده‌ای را برای تخمین شاخص بازداری بر اساس توصیف‌کننده‌های ساختار مولکولی فراهم می‌سازد [۸۳, ۸۴, ۹۴-۹۲].

۱- ۹ مروری بر تحقیقات انجام شده در مورد به‌کارگیری روش‌های انتخاب متغیر جریمه‌ای در ساخت مدل‌های QSAR/QSPR

در این بخش جستجوی کتابخانه در مورد به‌کارگیری روش‌های انقباضی در ساخت مدل‌های QSAR/QSPR انجام شده است که به‌ترتیب سال در ادامه مرتب شده‌اند.

با توجه به جستجوی انجام شده در پایگاه داده‌های متفاوت، از روش‌های انقباضی SCAD، ALASSO و LAD-LASSO در مطالعات QSAR/QSPR به دو شکل طبقه‌بندی[1] و رگرسیونی استفاده شده است. موارد مربوط به کاربرد روش‌های انقباضی در مطالعات طبقه‌بندی QSAR/QSPR[2] مد نظر اهداف این رساله نبوده است و جزییات آن در بخش مرور بر کارها ذکر نشده است [۱۰۶-۹۵]. لازم به ذکر است که در برخی از این تحقیقات، روش‌های انقباضی SCAD، ALASSO و LAD-LASSO تنها به‌منظور مقایسه با روش پیشنهادی انقباضی مورد نظر نویسندگان، به کار گرفته شده است و لزوماً به‌عنوان روش برتر مطالعه برگزیده نشده است. در ادامه به‌مرور مطالعات منتشر شده از به‌کارگیری روش‌های انقباضی SCAD، ALASSO و LAD-LASSO در مدل‌های رگرسیونی QSAR/QSPR پرداخته خواهد شد.

[1] Classification
[2] Classification

پنگ[1] و همکارانش در سال ۲۰۰۶ نقطه جوش ۵۳۰ هیدروکربن را با استفاده از جفت مدل SCAD و مدل Kriging پیش‌بینی نمودند. مدل SCAD-Kriging با ۱۵ توصیف کننده با مدل حداقل مربعات خطی معمولی (OLS)[2] مقایسه شد. مقدار ریشه میانگین مربعات خطا (RMSE)[3] برای مدل پیشنهادی نویسنده ۱۷٪ کاهش یافت [۱۰۷].

الجمال و همکارانش در سال ۲۰۱۵ مدل QSAR را برای پیش‌بینی فعالیت دارویی (pIC_{50}) ۶۵ مشتق ایمیدازو [b-۴،۵] پیریدین به‌عنوان بازدارنده‌های سرطان توسعه دادند. در این مطالعه از فرم اصلاح شده‌ای از روش انقباضی ALASSO به‌عنوان مدل استفاده شد و کارایی روش با مدل‌های LASSO و ALASSO مقایسه شد. به‌طوری‌که روش‌های انتخاب متغیر انقباضی بر روی کل توصیف‌کننده‌های محاسبه شده، ۲۵ بار (با تقسیم‌بندی مجموعه آموزش و آزمون متفاوت) اجرا شد و توصیف کننده‌ها بر اساس فراوانی ٪۵۰ به بالای تکرار توصیف کننده‌ها، برای مدل‌سازی مورد استفاده قرار گرفتند. پارامتر Q^2 برای داده‌های مجموعه آزمون به‌ترتیب برای مدل‌های QSAR توسعه یافته با مدل‌های انقباضی ALASSO اصلاح شده (۲۶ متغیر)، LASSO (۱۶ متغیر) و ALASSO (۱۲ متغیر) برابر با ۰/۸۷، ۰/۸۰ و ۰/۷۶ به‌دست آمد. نتایج نشان می‌دهد که مدل توسعه یافته انقباضی دارای قدرت پیش‌بینی مناسبی هستند [۱۰۸].

الجمال[4] و همکارانش در سال ۲۰۱۶، روش انقباضی Bridge با نُرم $L_{1/2}$ را برای ساخت مدل QSAR ۱۱۱ بازدارنده‌های استیل کولین استراز[5] ارائه کردند. به‌منظور مقایسه روش پیشنهادی از روش‌های انتخاب متغیر دیگری همچون LASSO[6]، ALASSO[7] و SCAD نیز استفاده شد. پارامترهای آماری متفاوتی برای این مدل‌ها محاسبه شد. پارامتر ضرایب تعیین مجموعه آزمون به ترتیب برای مدل‌های Bridge،

[1] Peng
[2] Ordinary least squares
[3] Root mean square error
[4] Algamal
[5] Acetylcholinesterase
[6] Least Absolute Shrinkage and Selection Operator
[7] Adaptive least Absolute Shrinkage and Selection Operator

LASSO، ALASSO و SCAD به ترتیب برابر با ۰/۹۱، ۰/۸۰، ۰/۸۱ و ۰/۸۷ به‌دست آمد. نتایج نشان می‌دهد که همه روش‌های انتخاب متغیر به کار گرفته شده در این مطالعه دارای قدرت پیش‌بینی قابل قبولی هستند [۱۰۹].

الجمال و همکارانش در سال ۲۰۱۷، مدل‌های اصلاح شده تنک مبتنی بر ضرایب تنظیم‌کننده متفاوت (RS-QSPR[1]) و هم‌چنین مدل SCAD را برای پیش‌بینی شاخص بازداری ۱۶۹ ترکیب اسانس ارائه کردند. مدل‌های RS-QSPR و SCAD به‌ترتیب ۴ و ۹ توصیف کننده مؤثر را انتخاب کردند. ضریب تعیین مجموعه آزمون برای این دو مدل به‌ترتیب برابر با ۰/۸۷ و ۰/۹۴ به‌دست آمد. نتایج نشان‌دهنده قدرت پیش‌بینی هر دو مدل QSPR توسعه یافته می‌باشد [۱۱۰].

الجمال و همکارانش در سال ۲۰۱۸ مدل‌های انقباضی QSAR (Bridge، LASSO و PBridge) را برای پیش‌بینی نقطه ذوب ۶۰ ترکیب نیتروآروماتیک ارائه دادند. توصیف‌کننده‌های منتخب هر روش به‌ترتیب برابر با ۶، ۷ و ۴ برای روش‌های BRIDGE، LASSO و PBridge به‌دست آمد. پارامترهای آماری متفاوتی برای مدل‌های توسعه یافته محاسبه شد. پارامتر ضریب تعیین مجموعه داده‌های خارجی برای مدل‌های انقباضی یاد شده به‌ترتیب برابر با ۰/۹۱، ۰/۸۱ و ۰/۹۴ به‌دست آمد. نتایج حاکی از اعتبار قابل قبول مدل‌های انقباضی توسعه یافته است [۱۱۰].

ماجومدار[2] و همکارانش در سال ۲۰۱۸ مدل‌های QSAR را برای مجموعه داده همگن متشکل از ۹۵ آمین، ارائه نمودند. با استفاده از توصیف‌کننده‌های محاسبه‌شده، فعالیت جهش‌زایی این ترکیبات را با استفاده از روش‌های مدل‌سازی متفاوتی همچون PCR[3]، PLS[4]، RF[5]، LASSO و SCAD پیش‌بینی کردند. میانگین مربعات خطای پیش‌بینی برای مدل‌های QSAR مربوطه به‌دست آمد و به‌ترتیب برابر با ۲۹/۱۱،

[1] robust sparse QSPR
[2] Majumdar
[3] Principal component regression
[4] Partial least squares regression
[5] Random forest

۱۸/۸۶، ۱۷/۲۵، ۲۶/۸۵، ۲۵/۸۱ است. نتایج نشان می‌دهد که مدل‌های ساخته شده از اعتبار مناسبی برخوردار هستند [۱۱۱].

یونگ زیا[1] و همکارانش در سال ۲۰۱۸، چندین روش انقباضی از جمله SCAD، MCP[2]، EN[3] را برای پیش‌بینی فعالیت دارویی ۴ دسته داده عمومی به‌کار گرفتند. پارامترهای آماری متفاوتی برای ارزیابی مدل‌های انقباضی توسعه یافته محاسبه شدند. نتایج مربوط به هر ۴ دسته داده از اعتبار قابل قبولی برخوردار بودند [۱۱۲].

علاوه بر این طبق جستجوی کتابخانه‌ای انجام شده، در سال‌های اخیر گزارشی مبنی بر استفاده از روش LAD-LASSO در پیش‌بینی فعالیت دارویی ترکیبات متفاوت در قالب ارائه یک مدل QSAR وجود نداشته است. تنها یک گزارش از به‌کارگیری روش‌های مدل‌سازی انقباضی LAD-Bridge، LAD-SCAD و LAD-LASSO در طبقه‌بندی ترکیبات فعال و غیرفعال ۱۰۸ بازدارنده آنفلوانزا توسط الجمال و همکارش در سال ۲۰۱۹ منتشر شده است [۱۱۳]. بنابراین با توجه به این‌که این روش در مطالعات QSAR استفاده نشده است، در این رساله برای اولین بار از روش انتخاب متغیر LAD-LASSO جفت شده با ANN برای پیش‌بینی فعالیت دارویی بازدارنده‌های سرطان ریه، روده و بازدارنده‌های ایدز استفاده شده است.

[1] Yong Xia
[2] Minimax concave penalty
[3] Elastic-net

۱- ۱۰ نوآوری تحقیق

با توجه به موارد اشاره شده در بخش مروری بر کارها، به‌خوبی مشخص شد که به‌کارگیری روش‌های انقباضی به‌دلیل داشتن مزایای ذاتی چون تنکی، پایداری، بایاس کم و کارایی بالا در انتخاب توصیف کننده‌های مؤثر در سال‌های اخیر مورد توجه محققین بوده است. به‌طوری که مطالعات متفاوتی مبنی بر به‌کارگیری روش‌های انقباضی برای ساخت مدل QSAR/QSPR ارائه شده است. نتایج جستجوی کتابخانه‌ای نشان می‌دهد که مطالعات اندکی به بررسی کارایی روش‌های انقباضی به‌عنوان روش انتخاب متغیر مناسب و ترکیب آن با روش‌های مدل‌سازی خطی و غیر خطی پرداخته‌اند. به‌این منظور، در این رساله، با توجه به کارایی و قدرت پیش بینی مدل غیر خطی شبکه عصبی مصنوعی (ANN)[1]، مدل‌های هیبریدی مبتنی بر روش‌های انقباضی به‌عنوان روش انتخاب متغیر تنک و پایدار و مدل شبکه عصبی به‌عنوان مدل غیر خطی قدرتمند در ساخت مدل‌های QSAR/QSPR پیشگو برای پیش بینی فعالیت دارویی/ ویژگی ترکیبات شیمیایی ارائه شد [۱۱۴, ۱۱۵].

بنابراین هدف اصلی این رساله ایجاد مدل‌های QSAR تنک و تفسیرپذیر با حداقل تعداد توصیف کننده مناسب برای پیش بینی فعالیت دارویی/ ویژگی ترکیبات شیمیایی است. به‌این منظور با توجه به کارایی و پراکندگی روش‌های انقباضی هم‌چون SCAD، ALASSO و LAD-LASSO، در انتخاب توصیف کننده‌هایی با بیشترین تأثیر بر متغیر وابسته، تلاش شد تا به‌صورت هدفمندی از این روش‌ها در ساخت مدل‌های QSAR/QSPR مبتنی بر شبکه عصبی مصنوعی استفاده شود. لازم به‌ذکر است پس از ارزیابی موفق مدل‌های QSAR/QSPR توسعه یافته، از مدل‌های شبکه عصبی توسعه یافته با توصیف کننده‌های منتخب روش‌های انقباضی متفاوت، برای پیشنهاد ترکیبات جدید بالقوه استفاده شده است.

[1] Artificial Neural Network

ش تجربی

۲- ۱ معرفی نرم‌افزارهای مورد استفاده برای مدل‌سازی QSAR

در این رساله از نرم‌افزارهای متفاوتی برای مدل‌سازی استفاده شده است که در ادامه به‌طور مختصر به توضیح و بررسی هر یک از آن‌ها پرداخته شده است.

۲- ۱- ۱ نرم‌افزار هایپرکم

نرم‌افزار هایپرکم [۱۱۶] به‌عنوان یک ابزار پیشرفته در شیمی محاسباتی، برای رسم و بهینه‌سازی مولکول‌های ساده و پیچیده با استفاده از روش‌های مکانیکی و کوانتومی شناخته می‌شود. با استفاده از این نرم‌افزار می‌توان، طول پیوند، زاویه پیوند، زاویه پیچشی و غیره ترکیبات شیمیایی را بهینه کرد. هرچند برخی از توصیف‌کننده‌های ترکیبات شیمیایی از قبیل ضریب شکست، قطبش پذیری، چربی‌دوستی و غیره با نرم‌افزار هایپرکم قابل محاسبه است، ولی غالباً برای محاسبه توصیف کننده‌های ترکیبات استفاده نمی‌شود و در مطالعات QSAR، خروجی این نرم‌افزار به‌عنوان ورودی برای نرم‌افزار دراگون به کار گرفته می‌شود.

۲- ۱- ۲ نرم‌افزار دراگون[1]

از نرم‌افزار دراگون جهت استخراج توصیف‌کننده‌های مولکولی استفاده می‌شود. نرم‌افزار دراگون توسط گروه کمومتریکس دانشگاه میلانو طراحی شده است [۱۱۷]. نرم‌افزار دراگون نسخه ۵/۵، قادر به محاسبه ۳۲۲۴ توصیف کننده می‌باشد که به ۲۲ دسته اصلی تقسیم می‌شوند. توصیف‌کننده‌هایی از جمله نوع و تعداد اتم‌ها، گروه‌های عاملی، توصیف‌کننده‌های توپولوژیکی و هندسی و غیره با استفاده از این نرم‌افزار قابل محاسبه است. اطلاعات کاملی از جمله ماهیت توصیف کننده‌ها، فرمول محاسبه، مراجع و جزییات توصیف کننده‌ها در کتاب مرجع توصیف‌کننده‌های مولکولی قابل دسترس است [۹].

[1] Dragon

۲- ۱- ۳ نرم‌افزار SPSS[1]

SPSS یکی از توانمندترین و جامع‌ترین نرم‌افزارهای آماری برای تحلیل داده‌ها است. نرم‌افزار آماری SPSS ساخته دانشگاه استنفورد[2]در سال ۱۹۷۰ است. از جمله کاربردهای نرم‌افزار آماری SPSS می‌توان به موارد زیر اشاره کرد:

- تحلیل و آنالیز داده‌های ورودی
- تهیه جداول و نمودارهای آماری
- به‌دست آوردن توزیع و رفتار داده‌های ورودی
- ایجاد داده‌های تصادفی
- پردازش انواع رگرسیون
- محاسبه انواع آزمون‌های آماری

در این رساله برای انجام آنالیزهای آماری متفاوت روی توصیف‌کننده‌های استخراج شده نرم‌افزار دراگون، اجرای روش رگرسیون خطی چندگانه برای انتخاب مؤثرترین توصیف کننده‌ها، محاسبه هم‌خطی[3] بین توصیف کننده‌ها از نرم‌افزار SPSS 25 استفاده شد [۱۱۸].

۲- ۱- ۴ نرم افزارهای R و R-studio

نرم‌افزار R، یک زبان برنامه‌نویسی و محیط نرم‌افزاری رایگان برای محاسبات آماری و علم داده‌ها است. نرم‌افزار R دارای بسته‌های[4]آماری و تحلیلی مختلف بر اساس موضوع مورد نیاز است، در نتیجه بسیار مورد توجه محققین قرار گرفته است. از جمله کاربردهای نرم‌افزار R می‌توان به موارد زیر اشاره کرد:

- برنامه‌نویسی و محیط نرم‌افزاری برای محاسبات آماری و علم داده‌ها.

[1]Statistical Package for Social Science
[2]Stanford University
[3]Collinearity
[4]Packages

- برنامه‌نویسی ساده و پیشرفته شامل عبارت‌های شرطی، حلقه و غیره.
- انجام عملیات داده‌کاوی و یادگیری ماشین مانند دسته‌بندی، خوشه‌بندی، انتخاب متغیر، مدل‌سازی، تحلیل شبکه و غیره با استفاده از کتابخانه‌های نرم افزار.
- امکان ذخیره، بازیابی و دست‌کاری داده‌ها.

R-Studio نیز یک ابزار رایگان برای استفاده از نرم‌افزار R است و امکان توسعه و به اشتراک‌گذاری آنالیزهای آماری را برای محققین فراهم می‌کند. نرم‌افزار R-Studio دارای محیط گرافیکی و نمایشی بهتری نسبت به نرم‌افزار R است و R-Studio، نرم‌افزار کاربرپسند[1]بوده و استفاده از آن، برای اجرای بسته‌های نرم‌افزاری آماری موجود در نرم افزار R ضروری می‌باشد.

در این رساله از نرم‌افزارهای R و R-studio برای پیش‌پردازش داده‌ها، تقسیم‌بندی داده‌ها، اجرای روش‌های رگرسیون انقباضی برای انتخاب مؤثرترین توصیف کننده‌ها و هم‌چنین، اجرای آزمون‌های متفاوت ارزیابی مدل‌های برتر استفاده شد.

۲- ۱- ۵ نرم‌افزار متلب

متلب[2]مخفف کتابخانه ماتریکس[3]است و یکی از پرکاربردترین زبان‌های برنامه‌نویسی شناخته شده است [۱۱۹]. متلب امکان محاسبات فنی و عددی، رسم داده‌ها، پیاده‌سازی الگوریتم‌ها، دست‌کاری ماتریس، یادگیری ماشین، شبیه‌سازی، پردازش تصویر و غیره را فراهم می‌سازد. در واقع یک محیط آسان برای ادغام برنامه‌نویسی، ترسیم و محاسبه است. متلب دارای کتابخانه‌های کاربردی فراوانی است که یادگیری و اجرای روش‌های متفاوت را برای کاربر آسان می‌کند. در این رساله به‌منظور پیش‌پردازش داده‌ها، مدل‌سازی شبکه عصبی مصنوعی و ارزیابی مدل‌های توسعه یافته از نرم‌افزار متلب ۲۰۱۷a استفاده شد.

[1]User-friendly
[2]MATLAB
[3]Matrix Laboratory

۲- ۱- ۶ نرم افزار Origin Lab

Origin Lab نرم‌افزاری قدرتمند است که برای تجزیه و تحلیل داده‌های آماری، پردازش توابع ریاضی و ویژوال‌بیسیک و رسم نمودارهای گرافیکی توسعه یافته است [۱۲۰]. از کاربردهای این نرم‌افزار می‌توان به مواردی همچون تجزیه و تحلیل داده‌ها و رسم نمودارهای آماری، قابلیت به اشتراک‌گذاری داده‌ها، نتایج و نمودارها با سایر نرم‌افزارها همچون اکسل[1] رسم نمودارهای دو بعدی و سه بعدی از توابع و داده‌های آماری، ذخیره نمودار با کیفیت بالا اشاره کرد. در این رساله برای نمایش برخی از نمودارها از نرم‌افزار Origin Lab استفاده شد.

۲- ۱- ۷ نرم‌افزار اتوداک[2]

Autodock4.2 یک نرم‌افزار رایگان شبیه‌سازی و مدل‌سازی مولکولی است. این نرم‌افزار بر روی سیستم عامل به‌آسانی قابل اجرا می‌باشد. برنامه Autodock4.2 یک روش خودکار توسعه یافته برای پیش‌بینی برهم‌کنش[3]لیگاندها با گیرنده‌های بیوماکرومولکولی است. هدف اصلی استفاده از این نرم‌افزار طراحی ترکیبات فعال زیستی در زمینه طراحی دارو به کمک رایانه می‌باشد که در آن با بررسی اتصال گیرنده – لیگاند، بهترین موقعیت لیگاند در جایگاه فعال گیرنده به‌دست می‌آید [۱۲۱].

۲- ۱- ۸ نرم افزار ویورلایت[4]

ViewerLite یک برنامه رایگان برای مشاهده، ویرایش، بررسی وجود پیوندهای هیدروژنی و هیدروفوبی گیرنده- لیگاند، تخمین طول پیوند و زوایای پیوند مولکول‌های شیمیایی و هم‌چنین تجزیه و تحلیل ساختارهای کریستالوگرافی پروتئین‌ها است [۱۲۲]. در این رساله، از نرم‌افزار ViewerLite نسخه ۵،

[1]Excel
[2]AutoDock
[3]Interaction
[4]ViewerLite

برای آماده‌سازی پروتئین و لیگاند و تعیین مختصات جایگاه فعال پروتئین قبل از اجرای فرایند داکینگ مولکولی استفاده شد.

۲- ۱- ۹ نرم افزار بیوویا دی اس[1]

BIOVIA Discovery Studio برای مدل‌سازی و شبیه‌سازی در علوم دارویی بسیار کارآمد است. این نرم‌افزار برای مطالعه پروتئین, پپتید, لیگاندهای کوچک, لیپیدها و کربوهیدرات‌ها ایجاد شده است. نرم‌افزار مدل‌سازی و شبیه‌سازی BIOVIA Discovery Studio برای محققین امکان شبیه‌سازی، تجسم و تجزیه و تحلیل سیستم‌های شیمیایی و بیولوژیکی و آنالیز نتایج داکینگ مولکولی را فراهم می‌سازد. در این رساله برای نمایش برهم‌کنش‌های گیرنده- لیگاند در جایگاه فعال پروتئین از نرم افزار بیوویا دی اس استفاده شده است [۱۲۳].

۲- ۱- ۱۰ نرم افزار VMD[2]

VMD محیط گرافیکی مناسبی برای شبیه‌سازی، نمایش و تجزیه و تحلیل سیستم‌های بیولوژیکی است. این نرم‌افزار قادر به فراخوانی فایل‌هایی با پسوند PDB می‌باشد و با استفاده از این نرم‌افزار می‌توان کمپلکس گیرنده-لیگاند را مورد بررسی قرار داد. به‌طوری‌که برهم‌کنش‌های متفاوت هیدروفیلی و هیدروفوبی با رنگ‌آمیزی‌های مختلف قابل مشاهده است.

[1]BIOVIA Discovery Studio
[2]Visual Molecular Dynamics

۲- ۲ پیش‌بینی فعالیت دارویی برخی از مشتقات استانیلید/ استامید به‌عنوان بازدارنده‌های ایدز با استفاده از مدل SCAD-ANN

۲- ۲- ۱ مقدمه

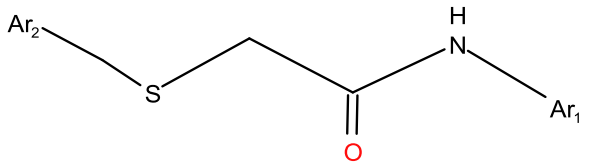
سندرم نقص ایمنی اکتسابی (ایدز) یک بیماری مزمن و تهدید کننده برای زندگی بشریت است که توسط ویروس نقص ایمنی انسانی (HIV) ایجاد می‌شود. HIV با آسیب رساندن به سیستم ایمنی بدن، توانایی بدن را برای مبارزه با عفونت و بیماری مختل می‌کند. HIV یک عفونت مقاربتی[1]است و همچنین می‌تواند از طریق تماس با خون آلوده یا از مادر به کودک در دوران بارداری، زایمان یا شیردهی سرایت کند. هیچ درمان قطعی برای ایدز وجود ندارد، اما داروها می‌توانند به‌طور چشم‌گیری پیشرفت بیماری را کاهش دهند. این داروها مرگ و میر ناشی از ایدز را در بسیاری از کشورهای توسعه یافته کاهش داده است [۱۲۴]. ترکیبات فعال دارویی متفاوتی به‌عنوان بازدارنده‌های ایدز، مورد استفاده محققین دارویی قرار گرفته‌اند که از بین بازدارنده‌های بالقوه، NNRTIs دارای برخی ویژگی‌های مطلوب مانند سمیت کم، فعالیت بالا و انتخاب پذیری مناسب هستند [۱۲۵, ۱۲۶]. مشتقات Arylazolylthioacetamide/acetanilide به‌عنوان گروه جدیدی از NNRTI ها، با خواص ساختاری مناسب، فعالیت ضد ایدز ویژه‌ای را در برابر سویه‌های جهش یافته نشان می‌دهند. در نتیجه طراحی و سنتز این ترکیبات حائز اهمیت بوده و مورد توجه محققین قرار گرفته است. این بازدارنده‌ها معمولاً از طریق اصلاح ساختاری پنج ترکیب رهبر طراحی و سنتز می‌شوند [۱۲۷-۱۳۳]. از آنجایی‌که سنتز و ارزیابی تجربی فعالیت دارویی ترکیبات فعال زمان‌بر و پرهزینه است، توسعه روش‌های تئوری از قبیل ساخت مدل QSAR برای ارزیابی و پیش‌بینی فعالیت‌های دارویی ترکیبات مشابه بسیار مهم است و باعث صرفه‌جویی در زمان، هزینه و به‌کارگیری نیروی انسانی می‌شود. با توجه به

[1]Sexually transmitted infection (STI)

اهمیت موضوع، در این بخش مدل QSAR برای پیش‌بینی فعالیت ترکیبات مورد مطالعه و پیشنهاد ترکیبات مشابه جدید بالقوه به‌کمک داکینگ مولکولی توسعه یافت. ازاین‌رو مدل شبکه عصبی جفت شده با روش انتخاب متغیر انقباضی SCAD برای پیش‌بینی فعالیت دارویی (pEC_{50}) برخی از مشتقات استانیلید/ استامید به‌عنوان بازدارنده‌های ایدز ساخته شد. مدل توسعه یافته با توصیف‌کننده‌های منتخب روش SCAD کمک شایانی را در راستای پیشنهاد ترکیبات فعال مشابه جدید نموده است. در ادامه این بخش با جزییات بیش‌تری به مراحل ساخت مدل SCAD-ANN برای این دسته از بازدارنده‌ها پرداخته شده است.

۲- ۲- ۲ مجموعه داده‌ها

در این تحقیق مجموعه‌ای ۵۷ تایی از بازدارنده‌های ایدز از مجموعه مشتقات Arylazolylthioacetamide/acetanilide جمع‌آوری شد و برای مدل‌سازی QSAR مورد بررسی قرار گرفت [۱۳۴-۱۳۶]. ساختار پایه ترکیبات مورد مطالعه در شکل ۲-۱ نشان داده شده است. ساختارهای شیمیایی ترکیبات و فعالیت‌های دارویی مربوطه در مقیاس لگاریتمی ($pEC_{50}=\log\frac{1}{EC_{50}}$) محاسبه و در جدول ۲-۱ خلاصه شده است. EC_{50} غلظت مؤثر (بر حسب مولار) مورد نیاز برای مهار تکثیر HIV است [۱۳۷]. در این مطالعه از pEC_{50} با محدوده از ۴/۶۲ تا ۷/۷۴ به‌عنوان متغیر وابسته در ساخت مدل QSAR استفاده شده است. تقسیم‌بندی مجموعه داده‌ها با استفاده از الگوریتم KS انجام شد و داده‌ها به سه دسته آموزش (۳۵ ترکیب)، ارزیابی (۱۱ ترکیب) و آزمون (۱۱ ترکیب) تقسیم شدند و در مراحل مختلف مدل‌سازی QSAR مورد استفاده قرار گرفتند.



شکل ۲-۱ ساختار پایه ترکیبات Arylazolylthioacetamide/acetanilide مورد مطالعه

جدول ۲-۱ ساختار ترکیبات شیمیایی به همراه مقادیر واقعی و پیش‌بینی شده pEC_{50}

ردیف	Ar_1	Ar_2	pE_{50} واقعی	pE_{50} پیش‌بینی شده
۱	2-fluorophenyl	2-(3-(2-Chlorophenyl) pyrazin-2-yl	۵/۲۸	۵/۴
۲[t]	2-Chlorophenyl	2-(3-(2-chlorophenyl) pyrazin-2-yl)	۵/۳۴	۵/۳۷
۳[v]	2,4-dichlorophenyl	2-(3-(2-Chlorophenyl) pyrazin-2-yl)	۵/۲۸	۵/۱۶
۴	2-Bromophenyl	2-(3-(2-chlorophenyl) pyrazin-2-yl)	۵/۳۸	۵/۳۴
۵	2-Bromo-4-methylphenyl	2-(3-(2-chlorophenyl) pyrazin-2-yl)	۵/۳۷	۵/۲۹
۶[t]	2-Bromo-4-chlorophenyl	2-(3-(2-chlorophenyl) pyrazin-2-yl)	۵/۲۹	۵/۱۷
۷[v]	4-Acetyl-2-bromophenyl	2-(3-(2-chlorophenyl) pyrazin-2-yl)	۵/۶	۵/۳
۸[t]	Methyl 3-bromo-4-benzoat	(2-(3-(2-chlorophenyl) pyrazin-2-yl	۵/۴۳	۵/۱۴
۹[v]	2-nitrophenyl	2-(3-(2-Chlorophenyl) pyrazin-2-yl)	۵/۰۵	۵/۲
۱۰	4-methyl-2-nitrophenyl	2-(3-(2-Chlorophenyl) pyrazin-2-yl	۵/۷۷	۵/۷۴
۱۱[t]	2-chloropyridin-3-yl	2-(3-(2-Chlorophenyl) pyrazin-2-yl)	۵/۵۴	۵/۳۵
۱۲	2-chloropyridin-3-yl	2-(3-(2-Chlorophenyl) pyrazin-2-yl)	۵/۶۲	۵/۷
۱۳	(3,4-dihydroisoquinolin-2(1H)-yl) ethenone	2-(3-(2-Chlorophenyl) pyrazin-2-yl)	۵/۳۹	۵/۲۸
۱۴[v]	Phenyl	2-(5-(2-chlorophenyl)-1,2,4-triazin-6-yl)	۶/۳۸	۶/۲۷
۱۵	2-chloro Phenyl	2-(5-(2-chlorophenyl)-1,2,4-triazin-6-yl)	۶/۰۹	۶/۵۵
۱۶	2-fluoro Phenyl	2-(5-(2-chlorophenyl)-1,2,4-triazin-6-yl)	۶/۶۸	۶/۳۷
۱۷	2-bromo Phenyl	2-(5-(2-chlorophenyl)-1,2,4-triazin-6-yl)	۷/۰۴	۶/۹۶
۱۸[v]	2-bromo 4-methylPhenyl	2-(5-(2-chlorophenyl)-1,2,4-triazin-6-yl)	۶/۷۵	۶/۶۱
۱۹[t]	4-methyl-2-nitrophenyl	2-(5-(2-chlorophenyl)-1,2,4-triazin-6-yl)	۷/۳۲	۷/۶۹
۲۰	2-nitrophenyl	2-(5-(2-chlorophenyl)-1,2,4-triazin-6-yl)	۷/۶۴	۷/۵۸
۲۱	4-acetyl-2-bromophenyl	2-(5-(2-chlorophenyl)-1,2,4-triazin-6-yl)	۷/۰۷	۶/۸۸
۲۲	2-chloropyridin-3-yl	2-(5-(2-chlorophenyl)-1,2,4-triazin-6-yl)	۷/۱۳	۷/۱۹
۲۳[v]	2,4-dichlorophenyl	2-(5-(2-chlorophenyl)-1,2,4-triazin-6-yl)	۶/۵۹	۶/۵۸
۲۴[t]	o-tolyl	2-(5-(2-chlorophenyl)-1,2,4-triazin-6-yl)	۶/۵۳	۶/۱۳
۲۵	ethyl 3-bromo-4-benzoate	2-(5-(2-chlorophenyl)-1,2,4-triazin-6-yl)	۶/۷۲	۶/۷۱
۲۶	3-bromo-5-methylpyridin-2-yl	2-(5-(2-chlorophenyl)-1,2,4-triazin-6-yl)	۶/۶۳	۶/۷۸
۲۷	pyridin-2-yl	2-(5-(2-chlorophenyl)-1,2,4-triazin-6-yl)	۵/۴۸	۶/۰۲
۲۸[v]	Methyl-3-thiophene-2-carboxylate	2-(5-(2-chlorophenyl)-1,2,4-triazin-6-yl)	۵/۵۴	۵/۴۱
۲۹[t]	1-(3,4-dichlorophenyl)ethan-1-one	2-(5-(2-chlorophenyl)-1,2,4-triazin-6-yl)	۵/۲۱	۴/۵۸
۳۰	1-(3,4-dihydroisoquinolin-2(1H)-yl)ethan-1-one	2-(5-(2-chlorophenyl)-1,2,4-triazin-6-yl)	۶/۱۴	۵/۸۵

ادامه جدول ۲-۱

ردیف	Ar_1	Ar_2	pE_{50} واقعی	pE_{50} پیش‌بینی شده
۳۱	phenyl	2-(5-(naphthalene-1-yl)-1,2,4-triazin-6-yl)	۵/۸۲	۵/۸۹
۳۲	2-chlorophenyl	2-(5-(naphthalene-1-yl)-1,2,4-triazin-6-yl)	۶/۸۷	۶/۶۶
۳۳	2-fluoro Phenyl	2-(5-(naphthalene-1-yl)-1,2,4-triazin-6-yl)	۶/۴۴	۶/۹
۳۴[v]	2-bromo Phenyl	2-(5-(naphthalene-1-yl)-1,2,4-triazin-6-yl)	۶/۷۶	۶/۹۶
۳۵	2-bromo 4-methylPhenyl	2-(5-(naphthalene-1-yl)-1,2,4-triazin-6-yl)	۶/۷۱	۶/۶۴
۳۶	4-methyl-2-nitrophenyl	2-(5-(naphthalene-1-yl)-1,2,4-triazin-6-yl)	۷/۳۴	۷/۲۲
۳۷	2-nitrophenyl	2-(5-(naphthalene-1-yl)-1,2,4-triazin-6-yl)	۷/۳۷	۷/۳۷
۳۸[t]	4-acetyl-2-bromophenyl	2-(5-(naphthalene-1-yl)-1,2,4-triazin-6-yl)	۷/۲۹	۶/۷۸
۳۹	2-chloropyridin-3-yl	2-(5-(naphthalene-1-yl)-1,2,4-triazin-6-yl)	۷/۲۳	۶/۷۹
۴۰v	o-tolyl	2-(5-(naphthalene-1-yl)-1,2,4-triazin-6-yl)	۶/۰۸	۶/۰۲
۴۱[t]	methyl 3-bromo-4-benzoate	2-(5-(naphthalene-1-yl)-1,2,4-triazin-6-yl)	۷/۱۶	۶/۸۶
۴۲[v]	ethyl 3-bromo-4-benzoate	2-(5-(naphthalene-1-yl)-1,2,4-triazin-6-yl)	۶/۶۸	۶/۹۷
۴۳	ethyl 3-chloro-4-benzoate	2-(5-(naphthalene-1-yl)-1,2,4-triazin-6-yl)	۶/۶۹	۶/۷۸
۴۴	2-bromo-4-sulfamoylphenyl	2-(5-(naphthalene-1-yl)-1,2,4-triazin-6-yl)	۷/۷۴	۷/۷۶
۴۵[t]	naphthalen-1-yl	2-(5-(naphthalene-1-yl)-1,2,4-triazin-6-yl)	۶/۱۲	۵/۷۸
۴۶	2-bromo-4-sulfamoylphenyl	2-(5-(Naphthalen-1-yl)pyrimidin-4-yl)	۶/۷۴	۶/۷۷
۴۷[v]	Methyl 3-bromo-benzoate	2-(5-(Naphthalen-1-yl)pyrimidin-4-yl)	۵/۶۲	۵/۳۴
۴۸	Ethyl 3-bromo-benzoate	2-(5-(Naphthalen-1-yl)pyrimidin-4-yl)	۵/۳۵	۵/۳۶
۴۹[t]	Methyl 3-chloro-benzoate	2-(5-(Naphthalen-1-yl)pyrimidin-4-yl)	۵/۵۵	۵/۳
۵۰	Ethyl 3-chloro-benzoate	2-(5-(Naphthalen-1-yl)pyrimidin-4-yl)	۵/۳۴	۵/۲۳
۵۱	3-Chloro-benzoic acid	2-(5-(Naphthalen-1-yl)pyrimidin-4-yl)	۵/۰۹	۵/۲۴
۵۲	2-nitrophenyl	2-(5-(Naphthalen-1-yl)pyrimidin-4-yl)	۶/۰۴	۶/۰۴
۵۳	3-Chloro-N-methoxy-benzamide	2-(5-(Naphthalen-1-yl)pyrimidin-4-yl)	۶/۱۲	۶/۱
۵۴	3-Chloro-N-hydroxy-benzamide	2-(5-(Naphthalen-1-yl)pyrimidin-4-yl)	۶/۲	۶/۰۲
۵۵	Ethyl 2-(3-chloro-benzamide)acetate	2-(5-(Naphthalen-1-yl)pyrimidin-4-yl)	۶/۸۲	۶/۵۱
۵۶	6-Chloro-5- picolinic acid	2-(5-(Naphthalen-1-yl)pyrimidin-4-yl)	۴/۶۲	۴/۶۱
۵۷	Ethyl 2-(3-chloro-benzamide)propanoate	2-(5-(Naphthalen-1-yl)pyrimidin-4-yl)	۵/۵۶	۵/۸۶

* v و t به ترتیب نمایانگر داده‌های مجموعه ارزیابی و آزمون هستند و سایر ترکیبات در مجموعه آموزش قرار دارند.

۲- ۲- ۳ رسم و بهینه‌سازی ساختار Arylazolylthioacetamide/acetanilide ها

ساختار شیمیایی مشتقات مورد مطالعه با استفاده از نرم‌افزار هایپرکم رسم شد و بهینه‌سازی ساختارها مطابق روش کار بخش (۱- ۵- ۳) تا رسیدن به حداقل انرژی بهینه شدند.

۲- ۲- ۴ استخراج توصیف کننده‌ها

ساختارهای بهینه ترکیبات مورد مطالعه در نرم‌افزار دراگون فراخوانی شدند و سپس برای هر ترکیب به تعداد ۳۲۲۴ توصیف کننده در ۲۲ دسته متفاوت محاسبه شدند.

۲- ۲- ۵ پیش‌پردازش و انتخاب توصیف‌کننده‌های مؤثر

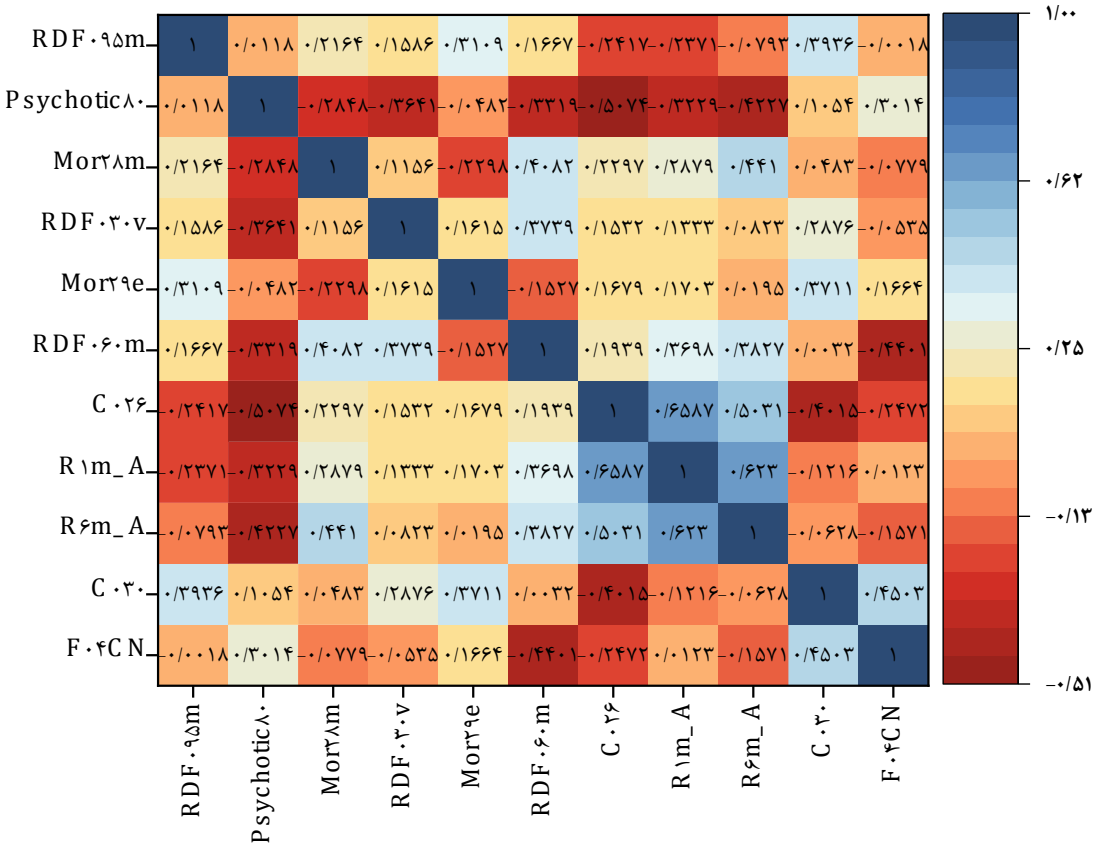
با توجه به تعداد زیاد توصیف‌کننده‌های محاسبه شده و احتمال وجود توصیف‌کننده‌های اضافی و فاقد اطلاعات مفید، فرآیند پیش‌پردازش و غربالگری توصیف کننده‌ها به‌منظور کاهش تعداد توصیف‌کننده‌ها و در نتیجه بهبود صحت پیش‌بینی، کاهش زمان مدل‌سازی و افزایش تفسیرپذیری مدل بر روی کل توصیف‌کننده‌های محاسبه شده اعمال شد. در این فرآیند، توصیف‌کننده‌هایی با مقادیر ثابت و تقریباً ثابت (توصیف‌کننده‌هایی با واریانس کم‌تر از ۰/۰۰۱) از مجموع توصیف‌کننده‌های محاسبه شده حذف شدند و در نهایت از بین دو توصیف کننده همبسته ($R^2 >$ ۰/۹۰) که دارای اطلاعات تقریباً یکسانی هستند، توصیف کننده با بیش‌ترین همبستگی با پاسخ، نگه داشته شد و توصیف‌کننده بعدی حذف شد. پس از کاهش اولیه توصیف کننده‌ها، فرآیند انتخاب متغیر انقباضی با روش ارزیابی تقاطعی ده فولد[۱] موجود در بسته ncvreg در نرم‌افزار R، روی داده‌های مجموعه آموزش و ارزیابی اجرا شد تا مؤثرترین توصیف کننده‌ها انتخاب شود. با اجرای روش SCAD، تعداد ۱۱ توصیف کننده مربوط به λ_{min} (λ دارای کم‌ترین خطای ارزیابی تقاطعی) انتخاب شدند. نام و نوع توصیف‌کننده‌های منتخب روش SCAD در جدول ۲-۲ خلاصه شده‌اند. به‌منظور

[۱] Ten fold cross validation (10-fold-CV-SCAD)

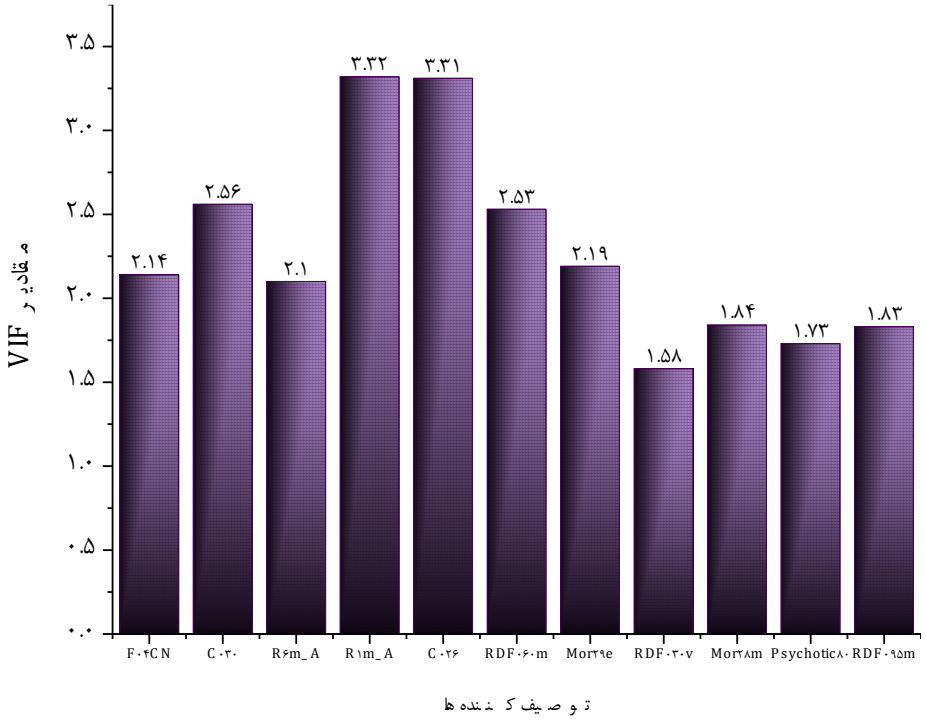
بررسی دقیق‌تر توصیف‌کننده‌های منتخب با روش SCAD، احتمال وجود همبستگی و هم‌خطی بین توصیف کننده‌ها، به‌ترتیب با محاسبه مقادیر ضریب همبستگی بین دو توصیف کننده و مقادیر افزایش تورم واریانس (VIF) توصیف کننده (مطابق رابطه ۱-۱۰) مطالعه گردید. به‌این منظور نمودارهای نقشه رنگی و VIF برای توصیف‌کننده‌های منتخب رسم شد و نتایج حاصل در شکل ۲-۲ و شکل ۲-۳ نمایش داده شده‌اند. نتایج مربوط به نمودار نقشه رنگی (شکل ۲-۲) نشان می‌دهد که همبستگی معناداری بین توصیف‌کننده‌های انتخاب شده با روش SCAD وجود ندارد. علاوه بر این نمودار VIF (شکل ۲-۳) نیز نشان می‌دهد که همه توصیف‌کننده‌های منتخب با روش SCAD مقادیر VIF کم‌تر از ۱۰ دارند، این شواهد بیانگر این است که بین توصیف‌کننده‌های منتخب با روش SCAD هم خطی قابل ملاحظه‌ای وجود ندارد [۱۳۸, ۱۳۹].

جدول ۲-۲ توصیف‌کننده‌های منتخب SCAD

ردیف	نماد	طبقه‌بندی	معنا	اهمیت
۱	F04[C-N]	2D frequency fingerprints	Frequency of C-N at topological distance 04	۰/۳۲۴
۲	C030	Atom centered fragments	X—CH—X	۰/۲۵۶
۳	R6m_A	GETAWAY	R autocorrelation of lag 6 / weighted by atomic masses	۰/۲۰۲
۴	R1m_A	GETAWAY	R autocorrelation of lag 1 / weighted by atomic masses	۰/۱۷
۵	C026	Atom centered fragments	R—CX—R	۰/۱۲۳
۶	RDF060m	RDF	Radial destribution function -6.0 / weighted by atomic masses	۰/۱۰
۷	Mor29e	3D-MoRSE	3D- MoRSE- signal 29/ weighted by atomic van der waals volumes	۰/۰۹۳
۸	RDF030v	RDF	Radial destribution function -3.0/ weighted by atomic masses	۰/۰۸۹
۹	Mor28m	3D-MORSE	3D- MoRSE- signal 29/ weighted by atomic masses	۰/۰۸۷
۱۰	Psychotic80	Molecular properties	Ghose- viswandhan- Wendoloski antipsychotic – like index at 80%	۰/۰۸۶
۱۱	RDF095m	RDF	Radial destribution function -9.5/ weighted by atomic masses	۰/۰۸۴



شکل ۲-۲ نمودار نقشه رنگی جهت نمایش همبستگی بین توصیف‌کننده‌های منتخب SCAD



شکل ۲-۳ نمودار مقادیر VIF توصیف‌کننده‌های منتخب SCAD

۲- ۲- ۶ مدل‌سازی شبکه عصبی با استفاده از توصیف‌کننده‌های منتخب SCAD

به‌منظور ساخت مدل QSAR غیرخطی از روش شبکه عصبی مصنوعی استفاده شد. در این مطالعه، یک مدل شبکه عصبی مصنوعی (ANN) پیشخور با الگوریتم آموزشی پس انتشار خطا مورد استفاده قرار گرفت. برنامه آن در یک m-file در نرم‌افزار MATLAB نوشته شد. با توجه به توصیه مقالات منتشر شده، استفاده از یک لایه پنهان در مدل ANN، در مطالعات شیمی محاسباتی کفایت می‌کند [۱۴۰, ۱۴۱]. بنابراین، مدل‌های ANN سه لایه‌ای شامل یک لایه ورودی، یک لایه پنهان و یک لایه خروجی برای بهینه‌سازی پارامترهای ANN و نهایتاً ساخت مدل نهایی استفاده شد. برای به دست آوردن بهترین مدل ANN، تمامی پارامترهای مؤثر بر عملکرد پیش‌بینی مدل، مانند تعداد ورودی‌ها، گره‌های لایه پنهان، دور آموزش و توابع آموزش و انتقال بهینه شدند. برای این منظور، چهار مدل شبکه عصبی مصنوعی مختلف با استفاده از دو الگوریتم آموزشی متفاوت لونبرگ – مارکوارت (LM) (تابع آموزشی trainlm در جعبه ابزار متلب) و تنظیم بایزین (BR) (تابع آموزشی trainbr در جعبه ابزار متلب) و دو تابع انتقال لگاریتم سیگموئیدی و تانژانت هایپربولیک سیگموئیدی (در جعبه ابزار برنامه متلب به ترتیب با توابع logsig و tansig شناخته می‌شوند) طراحی شدند. در تمام مدل‌های ANN طراحی شده، تابع خطی (purlin) به‌عنوان تابع انتقال خروجی استفاده شد. سایر پارامترهای مدل‌های ANN، مانند تعداد ورودی، تعداد گره‌ها و تعداد دورهای آموزش به‌طور هم‌زمان بهینه شدند. در بهینه‌سازی تعداد ورودی‌های ANN، زیرمجموعه‌هایی با تعداد توصیف‌کننده در بازه ۲ تا ۱۱ از میان توصیف کننده‌های انتخاب شده با SCAD به‌عنوان ورودی ANN در نظر گرفته شدند. با توجه به تعداد زیاد این زیرمجموعه‌ها (حدود $۱۰^{۱۰}$ زیرمجموعه دوتایی تا یازده‌تایی)، طراحی و استفاده از تمام توصیف‌کننده‌های تصادفی ایجاد شده به‌عنوان ورودی ANN عملاً غیرممکن است. بنابراین طبق توضیحات مندرج در بخش ۱- ۵- ۷- ۲ از چیدمان شبکه عصبی برای تعیین اهمیت توصیف کننده‌های منتخب روش SCAD به‌عنوان ورودی شبکه عصبی مورد استفاده قرار گرفت.

به‌طوری‌که ابتدا شبکه عصبی با ۱۱ زیرمجموعه به‌عنوان ورودی شبکه عصبی مصنوعی تعریف شد. برای این منظور، ابتدا مدل‌های ANN (۴ معماری ANN با توجه به دو تابع آموزش و انتقال متفاوت) با کل توصیف‌کننده‌های منتخب روش SCAD بهینه شدند. مدل SCAD-LM-ANN با معماری ۱-۴-۱۱ و با دور آموزشی ۵، تابع آموزش LM و تابع انتقال tansig دارای حداقل خطای مجموعه ارزیابی است. پس از شناسایی مدل بهینه شبکه عصبی توسعه یافته با ۱۱ توصیف کننده، از مدل بهینه برای تعیین اهمیت توصیف کننده‌های منتخب روش SCAD استفاده شد. برای تخمین اهمیت توصیف کننده i ام، همه ۱۱ توصیف کننده وارد مدل ANN بهینه شدند، درحالی‌که مقادیر توصیف کننده i ام با مقادیر تصادفی در محدوده تغییرات مقادیر واقعی توصیف کننده i جایگزین شدند. مدل ANN با ماتریس ورودی دست‌کاری شده آموزش داده شد و مقادیر pEC_{50} داده‌های مجموعه ارزیابی پیش‌بینی شد و مقدار MSE مجموعه ارزیابی محاسبه شد. این فرآیند برای همه توصیف کننده‌ها تکرار شد، به‌طوری‌که مقادیر همه توصیف کننده‌ها یک‌بار با مقادیر تصادفی جایگزین شدند و MSE مجموعه ارزیابی محاسبه شد. در نتیجه، ۱۱ MSE برای ۱۱ توصیف کننده به‌دست آمد. بالاترین MSE نشان می‌دهد که مدل ANN در غیاب آن توصیف کننده، خطای بیش‌تری را در مدل بهینه متحمل می‌شود و چنین توصیف کننده‌ای بیش‌ترین اهمیت را در توسعه مدل ANN دارد. توصیف کننده‌ها بر اساس اهمیت آن‌ها (مقادیر MSE از زیاد به کم) در جدول ۲-۲ مرتب شده‌اند. از مقادیر اهمیت محاسبه شده ۱۱ توصیف کننده، برای طراحی ۱۰ زیرمجموعه استفاده شد. به‌طوری‌که زیرمجموعه اول شامل توصیف‌کننده‌های مهم اول و دوم بود و زیرمجموعه‌های بعدی به‌ترتیب با افزودن یک توصیف‌کننده دیگر بر اساس ترتیب اهمیت توصیف کننده‌ها به زیرمجموعه‌های قبلی ایجاد شدند. پس از ایجاد زیر مجموعه‌های ورودی با استفاده از توصیف کننده‌های منتخب روش SCAD، تعداد ورودی‌ها، تعداد نورون‌های لایه پنهان و دور آموزشی به‌طور هم‌زمان بهینه شدند. در بهینه‌سازی پارامترهای ANN، تعداد ورودی‌ها با استفاده از زیرمجموعه‌های ایجاد شده با ۲ تا ۱۱ توصیف

کننده به‌عنوان ورودی‌های ANN در بازه ۲ تا ۱۱ تغییر یافت. علاوه بر این، تعداد نورون‌ها در لایه پنهان و دور آموزشی به‌ترتیب در محدوده ۲ تا ۱۰ (با گام ۱) و ۵ تا ۵۰ (با گام ۵) به‌طور هم‌زمان تغییر یافت. تمام مدل‌های ممکن و ایجاد شده ANN، با استفاده از مجموعه داده‌های آموزش (شامل ۳۵ ترکیب) آموزش داده شدند و برای پیش‌بینی مقادیر pEC_{50} مربوط به ۱۱ ترکیب موجود در مجموعه ارزیابی استفاده شدند. مقادیر MSE مجموعه ارزیابی برای تمام مدل‌های آموزش دیده محاسبه و به‌عنوان معیاری برای انتخاب بهترین مدل ANN مورد استفاده قرار گرفت. با توجه به نتایج، معماری شبکه و مقادیر MSE برای چهار مدل ANN با توابع آموزش و انتقال متفاوت که دارای کم‌ترین مقادیر MSE بودند در جدول ۲-۳ خلاصه شده‌اند. نتایج نشان می‌دهند که مدل ANN با ۵ توصیف کننده پر اهمیت و منتخب روش SCAD به‌عنوان ورودی، ۵ گره در لایه پنهان و ۵ دور آموزش با تابع انتقال لگاریتم سیگموئیدی و الگوریتم آموزشی LM دارای کم‌ترین مقدار MSE برای پیش بینی داده‌های مجموعه ارزیابی می‌باشد. بنابراین این مدل با نماد LM-SCAD-ANN نشان داده شد و با معماری ۱-۵-۵ به‌عنوان مدل برتر برای پیش‌بینی مقادیر pEC_{50} ترکیبات مورد مطالعه انتخاب شد.

جدول ۲-۳ ساختارهای شبکه‌های توسعه یافته با توصیف‌کننده‌های منتخب SCAD با کمترین MSE مجموعه ارزیابی

تعداد توصیف کننده	تابع آموزش	تابع انتقال	تعداد گره	تعداد دور آموزش	$MSE_{validation}$	$R^2_{validation}$
۱۰	تنظیم بایزین	لگاریتم-سیگموئید	۲	۱۵	۰/۰۶	۰/۸۹
۵	**لونبرگ- مارکوارت**	**لگاریتم-سیگموئید**	**۵**	**۵**	**۰/۰۳**	**۰/۹۷**
۶	تنظیم بایزین	تانژانت-سیگموئید	۴	۵	۰/۰۶	۰/۸۹
۹	لونبرگ- مارکوارت	تانژانت-سیگموئید	۳	۲۰	۰/۰۴	۰/۹۱

برای مقایسه عملکرد روش SCAD در انتخاب مؤثرترین توصیف کننده‌ها، از روش متداول انتخاب متغیر رگرسیون گام‌به‌گام (SR) استفاده شد. SR بر روی مجموعه داده‌های ارزیابی و آموزش اعمال شد. تعداد ۱۱ توصیف کننده شامل (F04CN, R1m_A, RDF095m, HATS5v, F09CN, C030, C005,

Mor12v, Mor22u, RDF080m, DISPv انتخاب شدند. زیرمجموعه‌هایی شامل ۲ تا ۱۱ توصیف کننده بر اساس اهمیت آن‌ها در مدل ANN (مطابق با بخش ۱- ۵- ۷- ۲) ایجاد شدند و به‌عنوان ورودی برای طراحی و بهینه سازی مدل‌های شبکه عصبی استفاده شدند. نتایج نشان داد مدل ANN با تابع آموزش LM و با استفاده از زیر مجموعه‌های ۱۱ تایی از توصیف کننده‌های SR با تعداد ۲ گره در لایه پنهان و ۵ دور آموزش حداقل مقدار MSE برای مجموعه ارزیابی ایجاد نمود. مدل بهینه SR-LM-ANN برای پیش‌بینی داده‌های مجموعه آزمون (pEC_{50}) استفاده شد.

۲- ۲- ۷ ارزیابی مدل SCAD-LM-ANN

یکی از اساسی‌ترین مراحل مدل‌سازی، بررسی قدرت پیش‌بینی مدل با استفاده از تکنیک‌های مختلف آماری است. در این تحقیق، قدرت پیش بینی، اعتبار و تعمیم پذیری مدل بهینه توسعه یافته SCAD-LM-ANN با استفاده از پیش‌بینی داده‌های مجموعه آزمون، پیش بینی پاسخ کل ترکیبات با تکنیک LOO، نمودار باقی‌مانده‌ها، محاسبه پارامترهای آماری متفاوت، آنالیز دامنه کاربرد و آزمون Y-تصادفی مورد ارزیابی قرار گرفت.

۲- ۲- ۷- ۱ ارزیابی مدل SCAD-LM-ANN با استفاده از پیش‌بینی داده‌های مجموعه آزمون

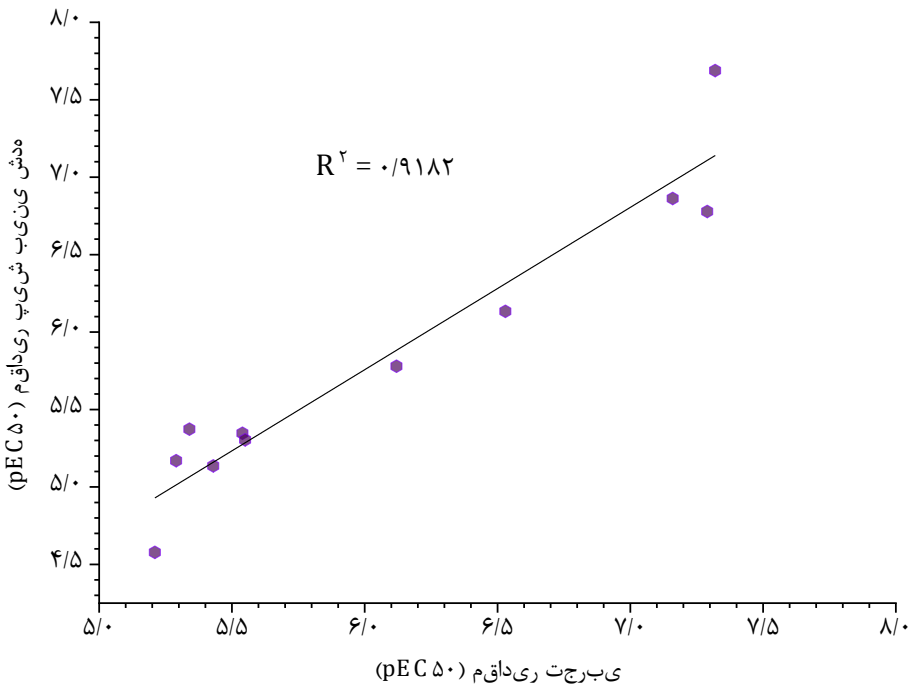
اعتبار و اهمیت مدل با استفاده از داده‌های مجموعه آزمون که در حین انتخاب متغیر و مدل‌سازی حضور نداشتند، سنجیده شد. به‌این منظور فعالیت دارویی داده‌های مجموعه آزمون با استفاده از مدل توسعه یافته بهینه SCAD-LM-ANN آموزش دیده در شرایط بهینه و با معماری ۱-۵-۵ پیش‌بینی شد. مقادیر پیش‌بینی شده فعالیت ترکیبات موجود در مجموعه آزمون در جدول ۲-۴ آورده شده است. مقادیر خطای کم اغلب ترکیبات موجود در مجموعه آزمون، نشان‌دهنده قدرت پیش‌بینی مناسب مدل توسعه یافته است. برای بررسی بیش‌تر قدرت پیش بینی مدل در مجموعه آزمون، مقادیر pEC_{50} پیش‌بینی شده به‌وسیله مدل بر حسب مقادیر واقعی آن‌ها رسم شد. شکل ۲-۴ نشان‌دهنده مقدار R^2 قابل قبول مدل بهینه ANN است.

با توجه به این‌که R^2 مربوط به مجموعه آموزش و ارزیابی برای مدل SCAD-LM-ANN به‌ترتیب برابر با ۰/۹۱ و ۰/۹۴ است بنابراین با مقایسه مقادیر مذکور با مقدار R^2 مجموعه آزمون، قدرت پیش‌بینی و تعمیم‌پذیری مدل SCAD-LM-ANN اثبات می‌شود.

همان‌طور که در بخش ۲- ۲- ۶ گفته شد، به‌منظور مقایسه عملکرد SCAD-LM-ANN، از روش SR-LM-ANN استفاده شد. مقادیر MSE و R^2_{test} برای مجموعه آزمون به‌ترتیب برابر با ۰/۴۲ و ۰/۴۸ به‌دست آمد. نتایج نشان می‌دهد که قدرت پیش‌بینی مدل SR-ANN رضایت‌بخش نیست و SCAD-LM-ANN توانایی پیش‌بینی بسیار بهتری را نسبت به SR-LM-ANN دارد.

جدول ۲-۴ نتایج حاصل از ارزیابی مدل SCAD-ANN با استفاده از مجموعه آزمون

شماره ترکیب در مجموعه داده‌ها	pEC_{50}		درصد خطا
	مقدار واقعی	مقدار پیش‌بینی شده	
۲	۵/۳۴	۵/۳۷	۰/۶۱
۶	۵/۲۹	۵/۱۷	-۲/۲۸
۸	۵/۴۳	۵/۱۴	-۵/۴۲
۱۱	۵/۵۴	۵/۳۵	-۳/۴۹
۱۹	۷/۳۲	۷/۶۹	۵/۰۴
۲۴	۶/۵۳	۶/۱۳	-۶/۰۸
۲۹	۵/۲۱	۴/۵۸	-۱۲/۱۵
۳۸	۷/۲۹	۶/۷۸	-۷/۰۱
۴۱	۷/۱۶	۶/۸۶	-۴/۱۶
۴۵	۶/۱۲	۵/۷۸	-۵/۵۷
۴۹	۵/۵۵	۵/۳	-۴/۴۵



شکل ۲-۴ نمودار تغییرات مقادیر پیش‌بینی شده pEC_{50} به‌وسیله مدل SCAD-LM-ANN در شرایط بهینه در مقابل مقادیر تجربی برای داده‌های مجموعه آزمون

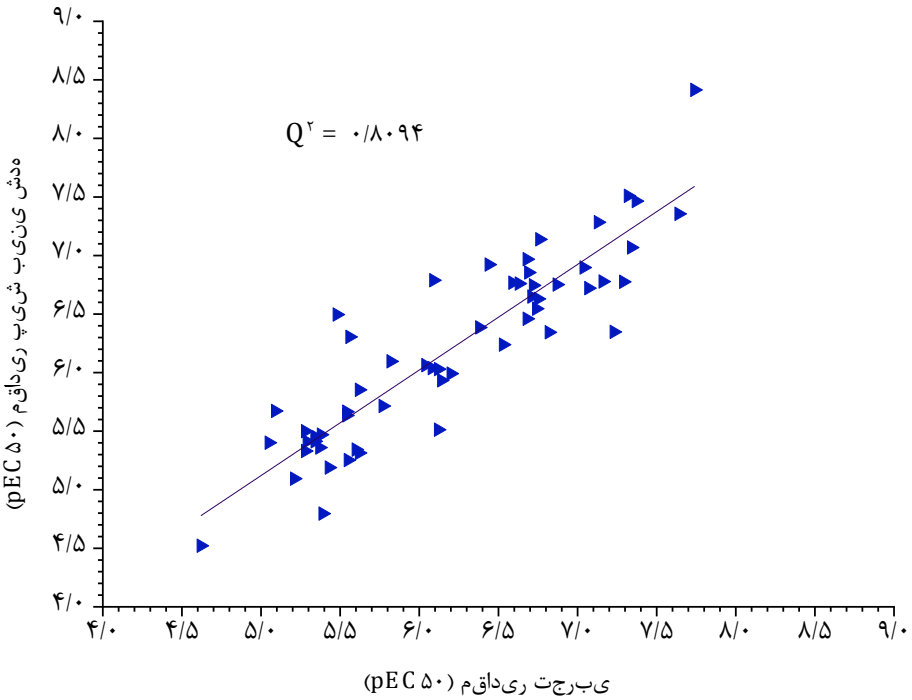
۲- ۲- ۷- ۲ ارزیابی مدل SCAD-ANN با استفاده از روش رد مرحله‌ای تک تک

در این بخش، یکی از تکنیک‌های قدرتمند برای ارزیابی مدل بهینه و بررسی قدرت پیش‌بینی مدل به‌نام تکنیک رد مرحله‌ای تک تک (LOO) برای پیش‌بینی همه داده‌های مورد مطالعه به‌عنوان داده آزمون مورد استفاده قرار گرفت. در این مطالعه هر بار یکی از داده‌های مورد مطالعه کنار گذاشته شد و مدل SCAD-LM-ANN با سایر داده‌ها آموزش داده شد و فعالیت دارویی ترکیب کنار گذاشته با مدل آموزش داده شده، پیش‌بینی گردید. این فرایند برای همه ترکیبات انجام شد و همه داده‌ها یک‌بار به‌عنوان داده آزمون در نظر گرفته شدند. ازاین‌رو فعالیت دارویی همه ترکیبات تکرار شد به‌گونه‌ای که تکنیک LOO توسط مدل SCAD-LM-ANN پیش‌بینی شدند و نتایج حاصله در جدول ۲-۵ آورده شده‌اند. برای مطالعه بیش‌تر قدرت پیش بینی مدل، نمودار مقادیر پیش‌بینی شده $pEC_{۵۰}$ تمام ترکیبات بر حسب مقادیر تجربی آن‌ها رسم شد. مقدار Q^2_{LOO} که در شکل ۲-۵ نشان داده شده است بیش‌تر از مقدار قابل قبول (Q^2> ۰/۵)

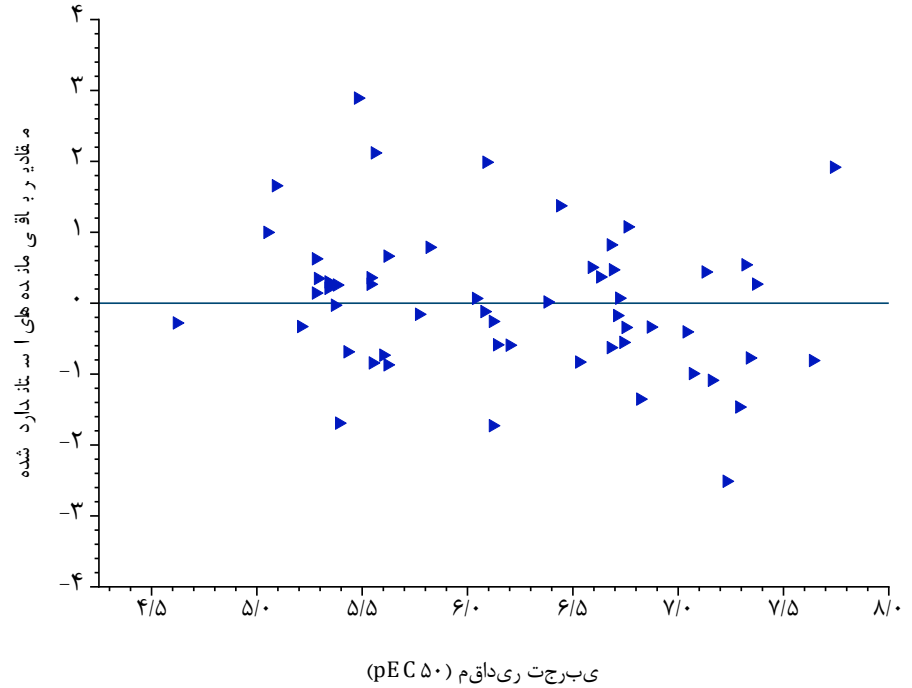
بوده که نشان‌دهنده این است که مدل توسعه یافته از پایداری مناسبی برخوردار است. علاوه بر این، به‌منظور بررسی عدم وجود خطای سیستماتیک در مدل توسعه یافته، مقادیر باقی‌مانده بر حسب مقادیر تجربی رسم شد. نتایج حاصله شکل ۲-۶ نشان‌دهنده توزیع یکنواخت و تصادفی داده‌ها حول محور صفر می‌باشد که بیانگر عدم وجود خطای سیستماتیک در مدل شبکه عصبی توسعه یافته با استفاده از روش SCAD به‌عنوان روش انتخاب متغیر (SCAD-LM-ANN) است.

جدول ۲-۵ نتایج حاصل از ارزیابی مدل SCAD-LM-ANN به روش رد مرحله‌ای تک تک برای کل داده‌ها

pEC_{50}			pEC_{50}					
درصد خطا	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب	درصد خطا	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
-۳/۳۶	۵/۹۳	۶/۱۴	۳۰	۴/۱۶	۵/۵	۵/۲۸	۱
۴/۷۵	۶/۱	۵/۸۲	۳۱	۱/۳۸	۵/۴۱	۵/۳۴	۲
-۱/۷۲	۶/۷۵	۶/۸۷	۳۲	۰/۹۵	۵/۳۳	۵/۲۸	۳
۷/۴۹	۶/۹۲	۶/۴۴	۳۳	۱/۶۷	۵/۴۷	۵/۳۸	۴
۵/۵۹	۷/۱۴	۶/۷۶	۳۴	-۰/۱۹	۵/۳۶	۵/۳۷	۵
-۰/۹۱	۶/۶۵	۶/۷۱	۳۵	۲/۳۱	۵/۴۱	۵/۲۹	۶
-۳/۷	۷/۰۷	۷/۳۴	۳۶	-۴/۶۱	۵/۳۴	۵/۶	۷
۱/۲۸	۷/۴۶	۷/۳۷	۳۷	-۴/۴۴	۵/۱۹	۵/۴۳	۸
-۷/۰۶	۶/۷۸	۷/۲۹	۳۸	۶/۹۵	۵/۴	۵/۰۵	۹
-۱۲/۲۱	۶/۳۵	۷/۲۳	۳۹	-۰/۹۶	۵/۷۱	۵/۷۷	۱۰
-۰/۶۸	۶/۰۴	۶/۰۸	۴۰	۲/۲۸	۵/۶۷	۵/۵۴	۱۱
-۵/۳۴	۶/۷۸	۷/۱۶	۴۱	۴/۱۴	۵/۸۵	۵/۶۲	۱۲
۴/۳۲	۶/۹۷	۶/۶۸	۴۲	-۱۱/۰۲	۴/۸	۵/۳۹	۱۳
۲/۴۷	۶/۸۶	۶/۶۹	۴۳	۰/۰۹	۶/۳۹	۶/۳۸	۱۴
۸/۷۱	۸/۴۱	۷/۷۴	۴۴	۱۱/۴۷	۶/۷۹	۶/۰۹	۱۵
-۹/۹۲	۵/۵۱	۶/۱۲	۴۵	-۳/۲۹	۶/۴۶	۶/۶۸	۱۶
-۲/۸۸	۶/۵۵	۶/۷۴	۴۶	-۲/۰۲	۶/۹	۷/۰۴	۱۷
-۵/۴۴	۵/۳۱	۵/۶۲	۴۷	-۱/۷۸	۶/۶۳	۶/۷۵	۱۸
۱/۸۲	۵/۴۵	۵/۳۵	۴۸	۲/۵۹	۷/۵۱	۷/۳۲	۱۹
-۵/۳۴	۵/۲۵	۵/۵۵	۴۹	-۳/۷۲	۷/۳۶	۷/۶۴	۲۰
۱/۹۶	۵/۴۴	۵/۳۴	۵۰	-۴/۹۴	۶/۷۲	۷/۰۷	۲۱
۱۱/۴۳	۵/۶۷	۵/۰۹	۵۱	۲/۱۷	۷/۲۸	۷/۱۳	۲۲
۰/۳۹	۶/۰۶	۶/۰۴	۵۲	۲/۶۹	۶/۷۷	۶/۵۹	۲۳
-۱/۴۸	۶/۰۳	۶/۱۲	۵۳	-۴/۴۷	۶/۲۴	۶/۵۳	۲۴
-۳/۳۶	۵/۹۹	۶/۲	۵۴	۰/۳۶	۶/۷۴	۶/۷۲	۲۵
-۶/۹۷	۶/۳۴	۶/۸۲	۵۵	۱/۹۶	۶/۷۶	۶/۶۳	۲۶
-۲/۱۳	۴/۵۲	۴/۶۲	۵۶	۱۸/۵۴	۶/۵	۵/۴۸	۲۷
۰/۷۴	۶/۳	۵/۵۶	۵۷	۱/۶۹	۵/۶۳	۵/۵۴	۲۸
				-۲/۲۲	۵/۰۹	۵/۲۱	۲۹



شکل ۲-۵ نمودار تغییرات مقادیر پیش‌بینی شده همه داده‌ها بر اساس تکنیک LOO در مقابل مقادیر تجربی



شکل ۲-۶ نمودار باقی‌مانده‌های حاصل از پیش‌بینی فعالیت دارویی ترکیبات با استفاده از تکنیک LOO و مقادیر تجربی برحسب مقادیر تجربی

۲- ۲- ۷- ۳ ارزیابی مدل SCAD-LM-ANN با استفاده از پارامترهای آماری

یکی از روش‌های ارزیابی مدل توسعه یافته، محاسبه پارامترهای آماری مربوط به مدل می‌باشد. بنابراین پارامترهای آماری معرفی شده در بخش ۱- ۵- ۸- ۴ برای فعالیت دارویی پیش‌بینی شده ترکیبات مجموعه آزمون و فعالیت‌های دارویی پیش‌بینی شده برای کل ترکیبات به روش رد مرحله‌ای تک تک محاسبه و در جدول ۲-۶ خلاصه شدند. نتایج حاصله در جدول ۲-۶ نشان می‌دهد که پارامترهای آماری در محدوده قابل قبول قرار دارند. بنابراین مطابق با بخش ۱- ۵- ۸- ۴، بزرگ‌تر بودن مقادیر پارامترهای تروپشا و روی همچون R_0^2، R_0^2 نسبی، R_m^2 و غیره از مقدار قابل قبول ۰/۵ و نزدیک بودن این پارامترها به مقدار R^2 قدرت پیش بینی و تعمیم پذیری مدل توسعه یافته SCAD-LM-ANN اثبات می‌شود.

جدول ۲-۶ پارامترهای آماری محاسبه شده برای مجموعه آزمون و داده‌های پیش‌بینی شده با تکنیک LOO برای مدل SCAD-LM-ANN

ردیف	پارامتر آماری	مقادیر pEC_{50} پیش بینی شده با SCAD-LM-ANN		محدوده قابل قبول پارامتر آماری
		ترکیبات مجموعه آزمون	کل ترکیبات با تکنیک LOO	
۱	PRESS	۱/۳۶	۶/۹	-
۲	SEP	۰/۳۵	۰/۳۵	-
۳	MAE	۰/۳۱	۰/۲۶	-
۴	REP(%)	۵/۷۹	۵/۶۳	-
۵	MSE	۰/۱۲	۰/۱۲	-
۶	MRE	۵/۱	۴/۳	-
۷	R^2	۰/۹۲	-	R^2>۰/۶
۸	Q^2_{LOO}	-	۰/۸۱	$\boldsymbol{Q^2_{Loo}}$> ۰/۵
۹	R^2_0	۰/۹۱	۰/۸۰	نزدیک به R^2
۱۰	R^2_0 نسبی	۰/۰۱	۰/۰۱	<۰/۱
۱۱	R^2_m	۰/۹۲	۰/۸۱	>۰/۵
۱۲	R'^2_0	۰/۸۹	۰/۸	نزدیک به R^2
۱۳	R'^2_0 نسبی	۰/۰۳	۰/۰۱	<۰/۱
۱۴	R'^2_m	۰/۹۲	۰/۸۱	>۰/۵
۱۵	R-R	۰/۰۲	۰/۰۰	<۰/۳
۱۶	k	۰/۹۶	۱/۰۰	$۰/۸۵ \leq \boldsymbol{k} \leq ۱/۱۵$
۱۷	k′	۱/۰۴	۱/۰۰	$۰/۸۵ \leq \mathbf{k'} \leq ۱/۱۵$

۲- ۲- ۷- ۴ ارزیابی مدل SCAD-LM-ANN با استفاده از دامنه کاربرد

دامنه کاربرد (AD) یک معیار مفید برای ارزیابی اعتمادپذیری[1] مدل‌های QSAR است. دامنه کاربرد مدل توسعه یافته SCAD-LM-ANN با استفاده از رسم نمودار ویلیام مورد تجزیه و تحلیل قرار

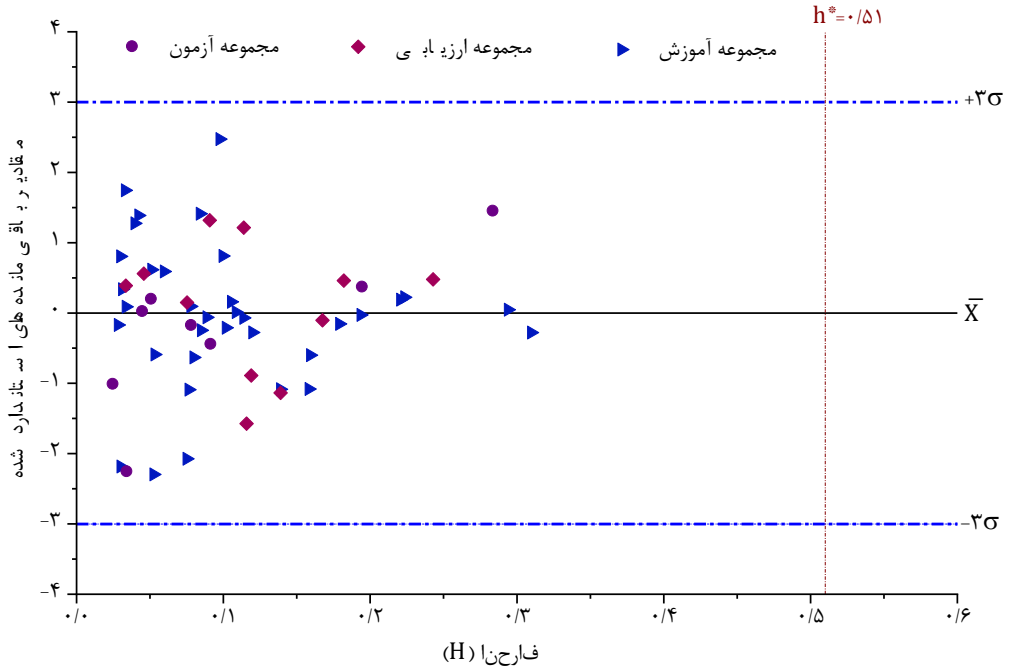
[1] Reliability

گرفت. طبق تعریف دامنه کاربرد در بخش ۱- ۵- ۸- ۵، تعیین فضای ساختارهای شیمیایی برای بررسی پیش بینی قابل اعتماد مدل، ضروری است. به عبارت دیگر، درجه تعمیم پذیری یک مدل QSAR توسعه یافته به میزان گستردگی دامنه کاربرد وابسته است. در یک فضای شیمیایی با p توصیف‌کننده معین، تخمین‌هایی برای ساختارهای شیمیایی جدید با استفاده از داده‌های آموزشی به دست می‌آید [۱۴۲]. در این بخش به‌منظور بررسی دامنه کاربرد، محاسبه مقدار انحراف (H) ضروری است. مقدار H یک ماده شیمیایی جدید، با اندازه‌گیری فاصله ماهالونبیس[1]آن ترکیب از مرکز مجموعه آموزشی متناسب است. بنابراین ماتریس H با ابعاد ۵۷×۵۷ با استفاده از رابطه ۱-۱۳ محاسبه شد. به‌این منظور ابتدا مقدار H مطابق با رابطه ۱-۱۳ محاسبه گردید و سپس بر اساس مقادیر pEC_{50} پیش بینی شده توسط مدل SCAD-LM-ANN، مقادیر باقی‌مانده‌های استاندارد شده محاسبه شد. از رسم نمودار باقی‌مانده‌های استاندارد شده بر حسب مقادیر H، نمودار ویلیام استخراج شد و در شکل ۲-۷ نشان داده شد.

صحت آنالیز دامنه کاربرد، با قرارگیری داده‌های شیمیایی در دو محدوده اطمینان قابل قبول نمودار ویلیام مشخص می‌شود. اولین شرط قابل قبول، قرارگیری و عدم تجاوز مقدار باقی‌مانده‌های استاندارد شده داده‌های شیمیایی در محدوده ۳ برابری بزرگ‌تر/ کوچک‌تری از انحراف استاندارد (σ) می‌باشد. علاوه بر این شرط، مقادیر محاسبه شده H نیز نباید از مقدار حد آستانه یا حد هشدار *h برابر با 3p/n بزرگ‌تر باشند. p در این معادله برابر با تعداد توصیف‌کننده‌های مدل به‌علاوه یک و n تعداد داده‌های مجموعه آموزش می‌باشد. بنابراین، قرارگیری داده‌های H، در محدوده‌های قابل قبول، استحکام و قابل اعتماد بودن مدل‌های QSAR/QSPR توسعه یافته را اثبات می‌کند. با توجه به نتایج ارزیابی دامنه کاربرد نشان داده شده در نمودار ویلیام (شکل ۲-۷)، مشاهده می‌شود که مقادیر محاسبه شده برای H محاسبه شده کم‌تر از مقدار هشدار انحراف است و مقادیر باقی‌مانده‌های استاندارد محاسبه شده نیز در محدوده ±۳σ می‌باشد. در نتیجه،

[1]Mahalanobis

هیچ‌کدام از داده‌های به کار رفته در مدل SCAD-LM-ANN به‌عنوان داده دور افتاده شناخته نشده است و همه داده‌ها در دامنه کاربرد قابل قبول قرار گرفته‌اند.

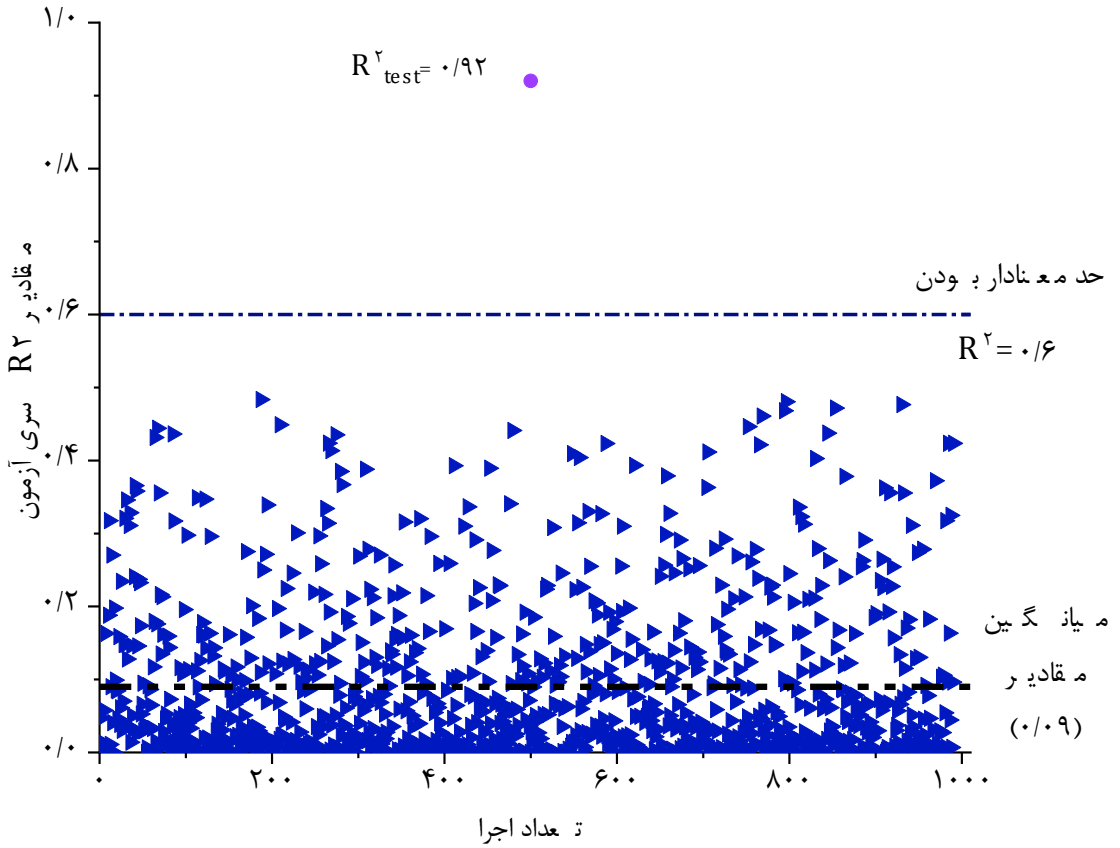


شکل ۲-۷ دامنه کاربرد مدل SCAD-LM-ANN، خطوط نقطه چین افقی و عمودی در دو انتهای نمودار به ترتیب نمایانگر مقادیر $\pm 3\sigma$ و h^* است.

۲- ۲- ۷- ۵ ارزیابی مدل SCAD-LM-ANN با استفاده از آزمون Y-تصادفی

به‌منظور بررسی عدم وجود ارتباط تصادفی ایجاد شده توسط مدل SCAD-LM-ANN بین ویژگی‌های ساختار ترکیبات شیمیایی و فعالیت دارویی مربوطه از آزمون Y- تصادفی استفاده شد. برای انجام آزمون Y- تصادفی ابتدا مقادیر ازاین‌رو متغیر وابسته (pEC_{50}) ترکیبات در مجموعه آموزش مورد مطالعه در محدوده حداقل و حداکثر مقادیر (۴/۶۲ تا ۷/۷۴) ۱۰۰۰ بار به‌صورت کاملاً تصادفی تغییر داده شدند و مدل شبکه عصبی با استفاده از داده‌های تصادفی متغیر وابسته توسعه داده شده و برای پیش بینی مقادیر pEC_{50} مجموعه آزمون به‌کار گرفته شدند. نتایج R^2 حاصل از پیش بینی pEC_{50} ترکیبات مجموعه آزمون با ۱۰۰۰ مدل توسعه یافته با متغیر وابسته تصادفی،

در شکل ۲-۸ آورده شده است. میانگین مقادیر R^2 برابر با ۰/۰۹ شد و مقادیر R^2 حاصل برای هر بار اجرا از مقدار قابل قبول ۰/۶ کم‌تر است (R^2<۰/۶). علاوه بر مقدار R^2 مربوط به مجموعه آزمون که به‌وسیله مدل پیشنهادی SCAD-LM-ANN با استفاده از متغیر اصلی وابسته برابر با ۰/۹۲ می‌باشد که به‌طور معناداری با مقادیر R^2 مجموعه آزمون پیش بینی شده در آزمون Y-تصادفی تفاوت دارد. می‌توان نتیجه‌گیری کرد که مدل پیشنهادی بر اساس ارتباط شانسی بین متغیرهای مستقل و متغیر وابسته پایه‌گذاری نشده است و ارتباط بین ساختار شیمیایی (توصیف کننده‌های منتخب روش SCAD) و فعالیت ترکیبات مورد مطالعه در مدل توسعه یافته SCAD-LM-ANN به‌طور منطقی و معنادار است.



شکل ۲-۸ نمودار مقادیر R^2 به‌دست آمده در آزمون Y-تصادفی بر حسب تعداد اجرا برای ۱۰۰۰ اجرای Y-تصادفی و پیش‌بینی فعالیت ترکیبات مجموعه آزمون به‌وسیله مدل SCAD-LM-ANN با استفاده از پاسخ تصادفی شده در شرایط بهینه

۲- ۳ پیش‌بینی فعالیت دارویی برخی از مشتقات ۳- chymotrypsin like protease ($3CL^{Pro}$) به‌عنوان بازدارنده‌های SARS-COV-2 با استفاده از مدل ALASSO-ANN

۲- ۳- ۱ مقدمه

ویروس سندرم حاد تنفسی (SARS) می‌تواند منجر به سندرم حاد تنفسی شود و ویروس SARS Coronavirus (CoV) در سال ۲۰۰۳ در آسیا برای اولین بار شناسایی شد و بیش از ۸۰۰۰ نفر را دچار عفونت و ۸۰۰ نفر قربانی داشت [۱۴۳]. در پایان دسامبر ۲۰۱۹، SARS-CoV-2، به‌عنوان یک کرونا ویروس جدید، منجر به سندرم حاد تنفسی به‌نام کووید-۱۹ شد. همه‌گیری کووید-۱۹ به‌سرعت در سراسر جهان گسترش یافت و به یک خطر و چالش جهانی برای سلامت عمومی بشریت و ثبات اقتصادی همه جوامع تبدیل شد [۱۴۴, ۱۴۵]. سازمان جهانی بهداشت (WHO) این بیماری را در مارس ۲۰۲۰ به‌عنوان یک اپیدمی اعلام کرد و این بیماری تا به الان حدود ۲۵۸ میلیون عفونت و حدود ۵ میلیون قربانی داشته است. با توجه به همه‌گیری بیماری کووید-۱۹ و این واقعیت که تعداد بیماران و مرگ و میر روزانه در حال افزایش است، پیشنهاد ترکیبات جدید مشابه دارو برای پیشگیری یا درمان این بیماری یکی از مهم‌ترین چالش‌های محققان است. یک راه مناسب برای دستیابی به این ترکیبات، طراحی ترکیبات قوی جدید با مطالعه ساختار داروهایی است که بر بیماری‌های ناشی از ویروس‌هایی با ساختار مشابه ویروس کووید-۱۹ تأثیر می‌گذارند. در بین ویروس‌های خانواده SARS، ویروس‌های SARS-CoV دارای توالی ساختاری مشابه با SARS-CoV-2 هستند [۱۴۶]. کرونا ویروس جدید SARS-CoV-2 دارای ژنومی با حدود ۸۰ درصد شباهت به ویروس بتا درگیر در سندرم SARS-CoV است [۱۴۷]. مسدود کردن پروتئازهای ویروسی کدگذاری شده توسط ژنوم ویروس، نقش حیاتی در جلوگیری از تکثیر ویروس و پیشرفت بیماری ایفا می‌کند. در میان

پروتئازهای کرونا ویروس، پروتئاز papain-like (PL^{pro}) و پروتئاز 3-chymotrypsin-like ($3CL^{pro}$) برای تکثیر ویروس حیاتی هستند. آنزیم $3CL^{pro}$ به‌عنوان یک هدف دارویی مناسب در مطالعات کووید-۱۹ توصیه می‌شود [۱۴۸]. در مطالعات اخیر SARS-CoV، مهارکننده‌های $3CL^{pro}$ در دسته‌های متفاوتی، از جمله ایزاتین [۱۴۹, ۱۵۰]، پلی فنول‌ها [۱۵۱]، سینانسرین [۱۵۲]، پیریدین [۱۵۳]، سولفونیل دی بنزن [۱۵۴]، استامید [۱۵۵] و تانشینون [۱۵۶] و سایر مولکول‌های کوچک [۱۵۵, ۱۵۷, ۱۵۸] گنجانده شده‌اند. اگرچه مهارکننده‌های سنتز شده $3CL^{pro}$ به دلیل گسترش بیش‌تر این بیماری در سطح جهان، مقاومت خوبی در برابر SARS-CoV نشان می‌دهند، اما همچنان برای مبارزه با عفونت این ویروس به ترکیباتی جدید با فعالیت ضد ویروسی نیاز است. یک راه سریع برای دستیابی به این ترکیبات، طراحی و کشف ترکیبات مشابه با مشتقات گزارش شده قبلی با فعالیت ضد ویروسی مناسب است. در این راستا، برخی از مشتقات مانند ایزاتین [۱۴۹, ۱۵۰]، پیریدین [۱۵۳]، سولفونیل دی بنزن [۱۵۴]، استامید [۱۵۵] و تانشینون [۱۵۶] با اثر بازدارندگی مناسب نسبت به SARS-CoV می‌توانند به‌عنوان کاندیدای مناسب در نظر گرفته شوند. روش‌های آزمون و خطای رایج برای طراحی دارو بسیار زمان‌بر و پرهزینه هستند. در نتیجه، رویکردهای طراحی دارویی به کمک کامپیوتر (CADD) برای طراحی ترکیبات جدید و قوی به‌شدت توصیه می‌شود. ابزارهای CADD و بیوانفورماتیک، به‌عنوان روش‌های سریع، ساده و اقتصادی [۱۵۹]، نقش مهمی را در طراحی ترکیبات بالقوه جدید دارند. یک مدل قدرتمند و معتبر QSAR دانش درستی را در مورد الگوهای ساختاری که همبستگی مناسبی با فعالیت دارویی ترکیبات شبه دارویی دارند، ارائه می‌کند [۴, ۱۶۲-۱۶۰]. مدل‌های QSAR دقت و صحت مطلوبی دارند و برای تخمین فعالیت دارویی ترکیبات جدید با ساختارهای مشابه با ساختار اصلی مجموعه داده‌ها مناسب هستند. پیش‌بینی فعالیت دارویی مهارکننده‌های جدید $3CL^{pro}$ قبل از سنتز به محققان کمک می‌کند تا از سنتز ترکیبات با فعالیت دارویی کم و نامناسب اجتناب کنند و با استفاده از این رویکرد در طراحی دارو در زمان و هزینه صرفه‌جویی کنند. یکی از کاربردی‌ترین

جنبه‌های مطالعات QSAR که اخیراً مورد توجه بسیاری از محققان قرار گرفته است، استفاده از مدل‌های QSAR برای پیشنهاد ترکیبات جدید با فعالیت دارویی مناسب، بر اساس رابطه توسعه یافته بین فعالیت دارویی و توصیف‌کننده‌های مؤثر است. بنابراین این مفهوم مستلزم آن است که مدل QSAR با استفاده از تعداد کمی از توصیف‌کننده‌های مؤثر ساخته شده باشد. ازاین‌رو، استفاده از روش‌های انتخاب متغیر کارآمد و ترکیب آن‌ها با روش‌های مدل‌سازی قدرتمند برای پیشنهاد ترکیبات بالقوه ضروری است. روش‌های انتخاب متغیر با کارایی بالا، توصیف‌کننده‌های اضافی را حذف می‌کنند، درحالی‌که توصیف‌کننده‌های مؤثر مرتبط با فعالیت دارویی را تا حد امکان حفظ می‌کنند. در مطالعات QSAR، پرداختن به داده‌های با ابعاد بالا و حذف توصیف‌کننده‌های اضافی و بی‌اهمیت همیشه موضوع چالش برانگیز محققان بوده است [۱۶۲]. از بین روش‌های انتخاب متغیر، روش‌های جریمه‌ای به دلیل داشتن مزایایی چون، بایاس کم، تنکی مناسب، پایداری و عملکرد بالا در حضور هم‌خطی بسیار مورد توجه قرار گرفته‌اند. از جمله روش‌های رگرسیون جریمه‌شده، می‌توان به لاسو تطبیقی (ALASSO) اشاره کرد که برای غلبه بر محدودیت‌های ذاتی روش‌های کلاسیک مناسب است [۳۰]. بنابراین، در این بخش از مطالعه، برای ساخت مدل QSAR از روش انتخاب متغیر ALASSO برای انتخاب مؤثرترین توصیف کننده‌ها استفاده شد، که در ادامه به آن پرداخته می‌شود. در این مطالعه، یک مدل QSAR تنک با قدرت پیش‌بینی و تفسیرپذیری بالا برای پیش‌بینی فعالیت دارویی ترکیبات ضد CoV-2 ارائه شد. به‌این منظور، ترکیبی از ALASSO و روش مدل‌سازی شبکه عصبی برای مدل‌سازی غیرخطی فعالیت دارویی برخی از بازدارنده‌های $3CL^{pro}$ ساخته شد. بنابراین، مدل QSAR پیشنهادی در این مطالعه، مهارکننده‌های جدید $3CL^{pro}$ با خواص فارماکوکینتیک مناسب را پیشنهاد می‌کند.

۲- ۳- ۲ مجموعه داده‌ها

مجموعه داده‌ها شامل ۹۰ مهارکننده $3CL^{pro}$ می‌باشد که ساختار ترکیبات و فعالیت دارویی ($IC_{۵۰}$) آن‌ها از مقالات منتشر شده استخراج شد [۱۴۹, ۱۵۰, ۱۵۳-۱۵۶]. لگاریتم مقادیر ($1/IC_{۵۰}$) محاسبه شد و به‌عنوان پاسخ در مراحل مدل‌سازی QSAR مورد استفاده قرار گرفت. $IC_{۵۰}$ غلظت مؤثر (بر حسب مولار) مورد نیاز برای بازداری ساختارهای شیمیایی ۹۰ ترکیب به یک فایل فرمت سیستم ورودی خطی ورودی مولکولی ساده شده (SMILES)[1] تبدیل شد و همراه با مقادیر $pIC_{۵۰}$ در جدول ۲-۷ خلاصه شد. تقسیم‌بندی مجموعه داده‌ها با استفاده از الگوریتم KS انجام شد و داده‌ها به سه دسته آموزش (۶۴ ترکیب)، ارزیابی (۱۳ ترکیب) و آزمون (۱۳ ترکیب) تقسیم شدند و در مراحل مختلف مدل‌سازی QSAR مورد استفاده قرار گرفتند.

[1] Simplified molecular-input line-entry system

جدول ۲-۷ مجموعه داده‌ها به همراه مقادیر واقعی و پیش‌بینی شده pIC_{50}

ردیف	SMILES	$pIC_{۵۰}$ واقعی	$pIC_{۵۰}$ پیش‌بینی شده
۱[v]	c1(ccc2c(c1)C(=O)C(=O)N2Cc1c(onc1C)C)C#N	۵/۱۴	۵/۵۸
۲	c1(ccc2c(c1)C(=O)C(=O)N2Cc1c(cc(cc1)F)Cl)I	۵/۰۳	۵/۳
۳[t]	c1(ccc2c(c1)C(=O)C(=O)N2C[C@@H]1COc2c(O1)cccc2)I	۴/۸۷	۴/۷۹
۴	c1ccc2c(c1)C(=O)C(=O)N2Cc1cc2c(s1)cccc2	۴/۸۸	۵/۱۲
۵[v]	c1cc(c2c(c1)C(=O)C(=O)N2Cc1cc2c(s1)cccc2)N(=O)=O	۵/۷	۵/۵۸
۶[t]	c1cc(c2c(c1)C(=O)C(=O)N2Cc1cc2c(s1)cccc2)Br	۶/۰۱	۶/۷۸
۷	c1(ccc2c(c1)C(=O)C(=O)N2Cc1cc2c(s1)cccc2)F	۵/۳۲	۵/۱۹
۸	c1(ccc2c(c1)C(=O)C(=O)N2Cc1cc2c(s1)cccc2)I	۶/۰۲	۵/۱۹
۹	c1ccc2c(c1Cl)C(=O)C(=O)N2Cc1cc2c(s1)cccc2	۴/۹۵	۵/۶۳
۱۰	c1(ccc2c(c1)C(=O)C(=O)N2C/C=C/c1cc2c(s1)cccc2)I	۴/۶۳	۴/۶۴
۱۱	c1(ccc2c(c1)C(=O)C(=O)N2Cc1ccc(s1)C(=O)Nc1ccc(cc1)Cl)I	۴/۹	۴/۹۷
۱۲	c1(ccc2c(c1)C(=O)C(=O)N2Cc1ccc(s1)C(=O)N1CCCCC1)I	۴/۷۶	۴/۸۱
۱۳	c1c(c(cc(c1Cl)S(=O)(=O)c1c(cc(cc1N(=O)=O)C(F)(F)F)N(=O)=O)C)Cl	۶/۵۲	۶/۲۶
۱۴	c1c(ccc(c1)S(=[O-])(=[O-])c1ccc(cc1)OC(=O)C(=C(Cl)Cl)Cl)OC(=O)C(=C(Cl)Cl)Cl	۶/۰۵	۵/۹۶
۱۵	c1c(ccc(c1)[S+3](=[O-])(=[O-])[C-]1=CN=C(N=C1N)N)Cl	۵/۲۲	۵/۳۴
۱۶	c1c(ccc(c1)[S@@](=[O-])(=O)c1ccc(cc1N(=O)=O)C(F)(F)F)Cl	۴/۹۲	۴/۷۶
۱۷	c1cccc(c1)[S@@](=[O-])(=O)c1nc(c(c(c1C#N)C)N(=O)=O)C	۴/۸۹	۵/۱۷
۱۸	c1c(ccc(c1)S(=[O-])(=[O-])c1nc(c(c(c1C#N)C)N(=O)=O)C)C	۴/۸۹	۴/۸۱
۱۹[t]	c1c(ccc(c1)[S@@](=O)(=[O-])c1n(=O)cc(cc1)N(=O)=O)Cl	۴/۸۲	۵/۰۶
۲۰	c1(ccc(cc1)N/C=C(/C(=O)OCC)\C#N)S(=[O-])(=[O-])c1ccc(cc1)N/C=C(/C(=O)OCC)\C#N	۴/۸	۴/۷۷
۲۱[v]	c1(cc(c(c(c1)Br)O)Br)S(=[O-])(=[O-])c1ccc(cc1)C(=O)O	۴/۸	۴/۹۶
۲۲	c1(ccccc1C(=O)C)S(=[O-])(=[O-])c1c(cccc1)C(=O)O	۴/۸	۴/۵۸
۲۳[t]	c1(ccc(cc1)N(=O)=O)S(=[O-])(=[O-])c1ccc(cc1)N(=O)=O	۴/۶	۴/۴۴
۲۴	c1(ccc(cc1)NC=C(C(=O)OCC)C(=O)OCC)S(=[O-])(=[O-])c1ccc(cc1)NC=C(C(=O)OCC)C(=O)OCC	۴/۴۹	۴/۶۵
۲۵[v]	c1cccc(c1)C#Cc1oc(cc1)C(=O)Sc1nnc([nH]1)C(F)(F)F	۵/۵۲	۵/۰۳
۲۶	c1(cc(nn1C)C(F)(F)F)c1ccc(s1)C(=O)Nc1c(c(no1)C)N(=O)=O	۵/۳	۴/۸۱
۲۷[v]	c1ccsc1[S+3](=[O-])(=[O-])Nc1[nH]nc(c1)c1cc(oc1C)C(C)(C)C	۵/۰۰	۴/۸
۲۸	c1ccsc1C(=O)Nc1[nH]nc([c-]1[S+3](=[O-])(=[O-])c1ccc(cc1)Cl)SC	۴/۸۲	۴/۷۶
۲۹[t]	c1coc(c1)CNc1sc(c2c1C(=O)CC(C2)(C)C)C#N	۴/۸	۴/۶۷
۳۰	c1ccc(s1)c1n(nc(c1C#N)SC)c1c(c(nn1C)C)N(=O)=O	۴/۷۴	۴/۷۲
۳۱	c1csc(c1)c1ccc(s1)CNc1n(nc(c1N(=O)=O)C)C	۴/۷	۴/۶۶
۳۲	c1csc(n1)C(=O)CC1=NCCS1	۴/۴	۴/۶۹
۳۳	c1(ccc2c(c1)C(=O)C(=O)N2)S(=[O-])(=[O-])N1CCN(CC1)C	۴/۱۱	۴/۳۹
۳۴[v]	c1(ccc2c(c1)C(=O)C(=O)N2)S(=[O-])(=[O-])N1CCN(CC1)Cc1cc(ccc1)Cl	۴/۵۰	۴/۴۳
۳۵	c1(ccc2c(c1)C(=O)C(=O)N2)S(=[O-])(=[O-])N1CCN(CC1)Cc1cc(c(c(c1)OC)OC)OC	۴/۴۹	۴/۴
۳۶	c1(ccc2c(c1)C(=O)C(=O)N2)S(=[O-])(=[O-])N1CCN(CC1)CCc1ccccc1	۴/۴۶	۴/۳۸
۳۷	c1(ccc2c(c1)C(=O)C(=O)N2)S(=[O-])(=[O-])N1CCN(CC1)C(=O)c1ccco1	۵/۰۰	۴/۹۵
۳۸	c1(ccc2c(c1)C(=O)C(=O)N2C)S(=[O-])(=[O-])N1CCN(CC1)c1ccccn1	۴/۲۹	۴/۷۱

ادامه جدول ۲-۷

ردیف	SMILES	$pIC_{۵۰}$ واقعی	$pIC_{۵۰}$ پیش‌بینی شده
۳۹ v	c1(ccc2c(c1)C(=O)C(=O)N2)S(=[O-])(=[O-])N1CCCCC1	۵/۳۵	۵
۴۰	c1(ccc2c(c1)C(=O)C(=O)N2)S(=[O-])(=[O-])N1CCOCC1	۴/۹۰	۴/۸۳
۴۱	c1(ccc2c(c1)C(=O)C(=O)N2)S(=[O-])(=[O-])N1CCC(CC1)C	۵/۹۳	۵/۵۹
۴۲ v	c1(ccc2c(c1)C(=O)C(=O)N2)S(=[O-])(=[O-])N1CCCC[C@@H]1C	۵/۶۵	۴/۹۱
۴۳	c1(ccc2c(c1)C(=O)C(=O)N2)S(=[O-])(=[O-])N1C[C@H](C[C@H](C1)C)C	۵/۳۷	۵/۵۲
۴۴	c1(ccc2c(c1)C(=O)C(=O)N2C)S(=[O-])(=[O-])N1CCN(CC1)C	۴/۹۳	۴/۳۹
۴۵	c1(ccc2c(c1)C(=O)C(=O)N2Cc1ccccc1)S(=[O-])(=[O-])N1CCN(CC1)C	۴/۱۷	۴/۴۵
۴۶	c12c(N(C(=O)C1=O)Cc1ccc3c(c1)cccc3)ccc(c2)S(=[O-])(=[O-])N1CCN(CC1)C	۴/۰۸	۴/۴۱
۴۷	c12c(N(C(=O)C1=O)C)ccc(c2)S(=[O-])(=[O-])N1CCN(CC1)CCc1ccccc1	۴/۸۶	۴/۳۸
۴۸	c12c(N(C(=O)C1=O)C)ccc(c2)S(=[O-])(=[O-])N1CCN(CC1)c1ccccn1	۵/۲۶	۴/۷۱
۴۹ v	c12c(N(C(=O)C1=O)Cc1ccccc1)ccc(c2)S(=[O-])(=[O-])N1CCCCC1	۴/۸۵	۴/۹۸
۵۰	c12c(N(C(=O)C1=O)C)ccc(c2)S(=[O-])(=[O-])N1CCOCC1	۵/۰۰	۵/۰۵
۵۱ t	c12c(N(C(=O)C1=O)Cc1ccccc1)ccc(c2)S(=[O-])(=[O-])N1CCOCC1	۴/۸۶	۴/۷۵
۵۲	c12c(N(C(=O)C1=O)Cc1ccc3c(c1)cccc3)ccc(c2)S(=[O-])(=[O-])N1CCOCC1	۴/۴۰	۴/۸۶
۵۳	c12c(N(C(=O)C1=O)C)ccc(c2)S(=[O-])(=[O-])N1CCC(CC1)C	۵/۹۸	۵/۵۷
۵۴ t	c12c(N(C(=O)C1=O)Cc1ccccc1)ccc(c2)S(=[O-])(=[O-])N1CCC(CC1)C	۵/۷۷	۴/۹۳
۵۵	c12c(N(C(=O)C1=O)Cc1ccc3c(c1)cccc3)ccc(c2)S(=[O-])(=[O-])N1CCC(CC1)C	۴/۷۵	۵/۰۵
۵۶	c12c(N(C(=O)C1=O)C)ccc(c2)S(=[O-])(=[O-])N1C[C@H](C[C@H](C1)C)C	۵/۵۵	۵/۵۵
۵۷	c12c(N(C(=O)C1=O)Cc1ccccc1)ccc(c2)S(=[O-])(=[O-])N1C[C@H](C[C@H](C1)C)C	۵/۳۳	۵/۳۹
۵۸	c1c(cc(cc1Cl)NC(=O)CSc1nccc(n1)c1nc(sc1)c1ccccc1)Cl	۵/۵۲	۵/۶۷
۵۹	c1c(ccc(c1)NC(=O)CSc1nc(ccn1)c1ccc(s1)c1cc(n(n1)C)C(F)(F)F)Cl	۵/۰۰	۴/۶۷
۶۰ v	c1cc(ccc1NC(=O)CSc1nc(ccn1)c1c(c(sc1SC)c1nc(cs1)C)C)Cl	۴/۹۶	۴/۶۹
۶۱ t	c1(ccccc1NC(=O)CSc1nc(ccn1)c1c(c(sc1SC)c1nc(cs1)C)C)Cl	۴/۹۲	۴/۸۶
۶۲	c1c(ccc(c1)NC(=O)CSc1nccc(n1)c1nc(sc1)c1nc(sc1)C)Cl	۴/۸۵	۵/۰۱
۶۳ t	c1c(ccc(c1)NC(=O)CSc1nccc(n1)c1cc(no1)c1ccccc1)Cl	۴/۸۲	۴/۵
۶۴	c1cc(ccc1NC(=O)CSc1nc(ccn1)c1cc(no1)c1ccc(cc1Cl)Cl)C(F)(F)F	۴/۸۲	۴/۸۱
۶۵ t	c1cc(ccc1NC(=O)CSc1nccc(n1)c1cc(no1)c1c(cccc1)Cl)Cl	۴/۸۲	۴/۵۱
۶۶	c1cc(ccc1NC(=O)CSc1nc(cc(n1)O)CCC)Cl	۴/۵۲	۴/۴۹
۶۷	c1c(ccc(c1)NC(=O)CSc1nc(c(c(=O)[nH]1)C#N)c1cccc(c1)OC)S(=[O-])(=[O-])[NH2-]	۴/۴۰	۴/۵۳
۶۸	c1c(ccc(c1)NC(=O)CSc1nc(ccn1)c1cccs1)C(C)C	۴/۴۰	۴/۴۹
۶۹	n1c(nc(cc1c1ccccc1)O)SCC(=O)Nc1ccc(c(c1)OC)OC	۴/۳۵	۴/۶۱
۷۰ t	c1cc(ccc1NC(=O)CSc1[nH]c(=O)c(c(n1)c1cc(ccc1)OC)C#N)C(=O)C	۴/۲۲	۴/۴۱

ادامه جدول ۲-۷

ردیف	SMILES	$pIC_{۵۰}$ واقعی	$pIC_{۵۰}$ پیش‌بینی شده
۷۱	c1(cccc(c1)NC(=O)[C@@H](CC)Sc1nc(c(c(=O)[nH]1)C#N)c1ccccc(c1)OC)C(=O)C	۴/۲۲	۴/۴۱
۷۲	c1(ccc(c(c1)NC(=O)CSc1[nH]c(=O)cc(n1)C)Oc1ccccc1)Cl	۴/۰۰	۴/۴۶
۷۳	c1c(ccc(c1)c1oc(cc1)C(=O)Oc1cncc(c1)Cl)Cl	۷/۲۰	۶/۵۲
۷۴ [t]	c1c(ccc(c1)c1oc(cc1)C(=O)Oc1cncc(c1)Cl)N(=O)=O	۷/۲۲	۶/۶۷
۷۵ [v]	c1c(cc(c(c1)c1oc(cc1)C(=O)Oc1cncc(c1)Cl)N(=O)=O)Cl	۶/۹۱	۶/۷۴
۷۶	c1ccc(c(c1)c1oc(cc1)C(=O)Oc1cncc(c1)Cl)N(=O)=O	۶/۶۸	۶/۷۸
۷۷	c1cc(cc(c1)c1oc(cc1)C(=O)Oc1cncc(c1)Cl)N(=O)=O	۶/۳۰	۶/۷۷
۷۸	c1(C(=O)Oc2cncc(c2)Cl)ccncc1	۶/۷۹	۶/۶۳
۷۹	c1(C(=O)Oc2cncc(c2)Cl)cnccc1	۶/۱۶	۶/۳۹
۸۰	c1(C(=O)Oc2cncc(c2)Cl)ccc(cc1)Cl	۶/۳۶	۶/۴۱
۸۱ [v]	c1(C(=O)Oc2cncc(c2)Cl)c(cccc1)N(=O)=O	۶/۴۸	۶/۱۳
۸۲	c1(C(=O)Oc2cncc(c2)Cl)cc(ccc1)N(=O)=O	۶/۱۶	۶/۰۱
۸۳ [t]	c1(C(=O)Oc2cncc(c2)Cl)c2c(ccc1)cccc2	۶/۹۱	۶/۶۶
۸۴ [v]	c1(C(=O)Oc2cncc(c2)Cl)c(=O)oc2c(c1)cccc2	۶/۹۷	۶/۷۸
۸۵	C1CC(c2c(C1)c1c(cc2)c2c(C(=O)C1=O)c(co2)C)(C)C	۴/۰۵	۴/۳۹
۸۶	C1C[C@](c2c(C1)c1c(cc2)c2c(C(=O)C1=O)c(co2)C)(C)CO	۴/۶۱	۴/۴۲
۸۷	C1C[C@](c2c(C1)c1c(cc2)c2c(C(=O)C1=O)c(co2)C)(C)C(=O)OC	۴/۶۸	۴/۴۲
۸۸	c1cc(c2c(c1)c1c(cc2)c2c(C(=O)C1=O)c(co2)C)C	۴/۴۱	۴/۵۸
۸۹	c1cc(c2c(c1)c1c(cc2)C2=C(C(=O)C1=O)[C@H](CO2)C)C	۴/۸۴	۴/۸۷
۹۰	C1CC(c2c(C1)c1c(cc2)C=C(C(=O)C1=O)C(C)C)(C)C	۴/۶۸	۴/۴۴

۲- ۳- ۳ رسم و بهینه‌سازی ساختار بازدارنده‌های $3CL^{pro}$

ساختارهای سه بعدی ۹۰ بازدارنده $3CL^{pro}$ با استفاده از نرم‌افزار هایپرکم و مطابق با بخش ۱- ۵- ۳ رسم شد. ساختارهای شیمیایی مطابق با روش کار بخش ۱- ۵- ۳ تا رسیدن به حداقل انرژی بهینه شدند و در نهایت با فرمت hin.* ذخیره شدند.

۲- ۳- ۴ استخراج توصیف کننده‌ها

ساختارهای بهینه ۹۰ بازدارنده $3CL^{pro}$ مورد مطالعه در نرم‌افزار دراگون فراخوانی شدند و سپس برای هر ترکیب شیمیایی به تعداد ۳۲۲۴ توصیف کننده محاسبه شدند.

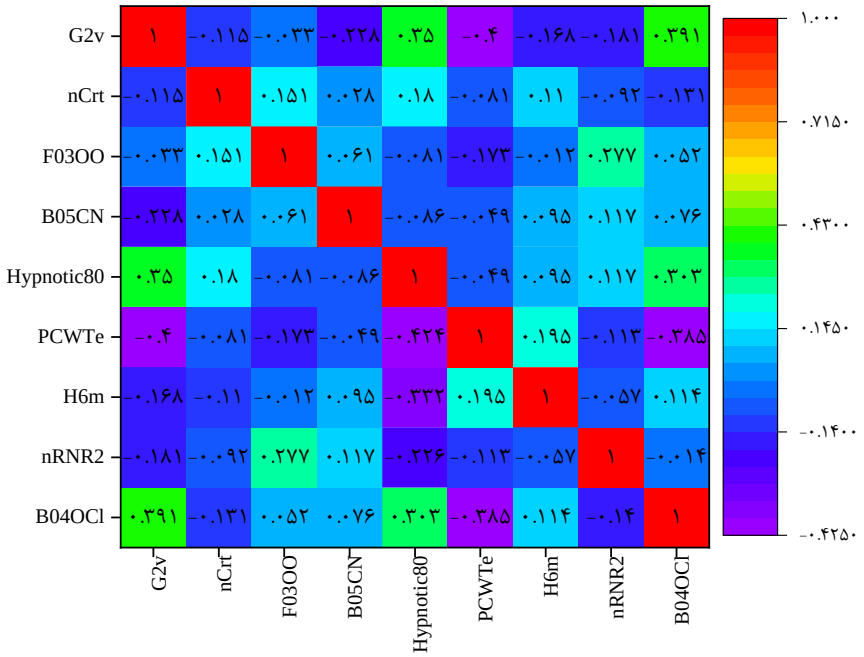
۲- ۳- ۵ پیش‌پردازش و انتخاب توصیف‌کننده‌های مؤثر

توصیف‌کننده‌های مولکولی با استفاده از ساختارهای بهینه و با استفاده از نرم‌افزار دراگون محاسبه شدند. از بین توصیف‌کننده‌های محاسبه شده، توصیف کننده‌هایی با مقادیر ثابت و نسبتاً ثابت (توصیف کننده‌هایی با واریانس کم‌تر از ۰/۰۰۱) با اجرای بسته نرم‌افزاری caret در برنامه R حذف شدند [۱۶۳]. بنابراین حدود ۱۳۸۶ توصیف کننده که اطلاعات خاصی به مدل اضافه نمی‌کنند، از مجموعه داده‌ها حذف شدند. از بین دو توصیف کننده با همبستگی بالاتر از ۰/۹، توصیف کننده‌ای که همبستگی کم‌تری با پاسخ داشت حذف شد. پس از انجام مراحل پیش‌پردازش انجام شده، ماتریسی شامل ۶۴۵ توصیف کننده برای مطالعات بیش‌تر ذخیره شد. سپس روش انتخاب متغیر ALASSO به‌عنوان روش انتخاب متغیر جریمه‌شده بر روی ماتریس داده‌ها (شامل ۶۴۵ متغیر مستقل و $pIC_{۵۰}$ به‌عنوان متغیر وابسته) اجرا شد. رگرسیون ALASSO با روش ارزیابی تقاطعی ۱۰ فولد موجود در بسته نرم‌افزاری parcor در برنامه R بر روی مجموعه داده‌های ارزیابی و آموزش اجرا شد تا موثرترین توصیف کننده‌ها انتخاب شود [۳۴]. با اجرای روش ALASSO به تعداد ۹ توصیف‌کننده مربوط به λ_{min} (λ دارای کم‌ترین خطای ارزیابی تقاطعی) به‌عنوان مؤثرترین توصیف کننده‌ها انتخاب شدند. مقادیر مربوط به ضرایب رگرسیونی ۹ توصیف کننده منتخب روش

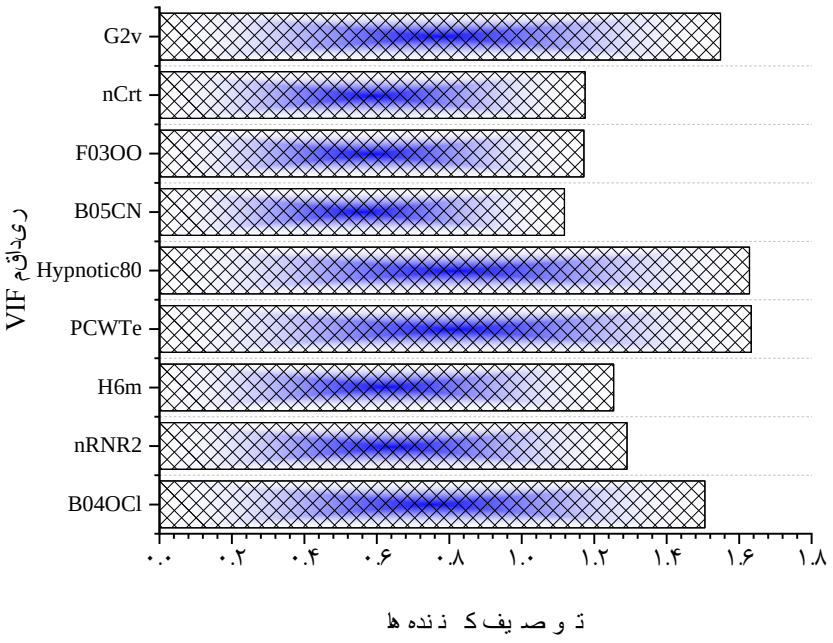
ALASSO در جدول ۲-۸ آورده شده است. ارزیابی‌های مربوط به کیفیت آماری توصیف‌کننده‌های منتخب روش ALASSO، از جمله عدم وجود همبستگی و هم‌خطی بین توصیف کننده‌ها با محاسبه مقادیر ضرایب همبستگی بین دو توصیف کننده و مقادیر افزایش تورم واریانس (VIF) توصیف کننده (مطابق با رابطه ۱-۱۰) مطالعه شد. نتایج مربوط به آن‌ها در نمودارهای نقشه رنگی و VIF (شکل ۲-۹ و شکل ۲-۱۰) آورده شده است. شکل ۲-۹ نشان می‌دهد که همبستگی معناداری بین توصیف‌کننده‌های انتخاب شده با روش ALASSO وجود ندارد. علاوه بر این، در شکل ۲-۱۰ نیز، مقادیر کم‌تر از ۱۰ مربوط به آنالیز افزایش تورم واریانس مشاهده می‌شود. نتایج حاکی از آن است که بین توصیف کننده‌های منتخب روش ALASSO هم خطی شدیدی وجود ندارد [۱۳۸, ۱۳۹].

جدول ۲-۸ توصیف‌کننده‌های منتخب ALASSO

ردیف	نماد	طبقه‌بندی	معنا	ضرایب استاندارد شده ALASSO
۱	B04[O-Cl]	2D binary fingerprints	Presence/absence of O - Cl at topological distance 4	۰/۶۶
۲	nRNR2	Functional group counts	number of tertiary amines (aliphatic)	۰/۱۲
۳	H6m	GETAWAY descriptors	H autocorrelation of lag 6 / weighted by mass	۰/۱
۴	PCWTe	Charge descriptors	partial charge weighted topological electronic index	۰/۰۹
۵	Hypnotic-80	Drug-like indices	Ghose-Viswanadhan-Wendoloski hypnotic-like index at 80%	۰/۰۸
۶	B05[C-N]	2D binary fingerprints	Presence/absence of C - N at topological distance 5	۰/۰۷
۷	F03[O-O]	2D frequency fingerprints	Frequency of O - O at topological distance 3	۰/۰۴
۸	nCrt	Functional group counts	number of ring tertiary C(sp3)	۰/۰۳
۹	G2v	WHIM descriptors	nd component symmetry directional ۲ WHIM index / weighted by van der Waals volume	۰/۰۰۲



شکل ۲-۹ نمودار نقشه رنگی جهت نمایش همبستگی بین توصیف‌کننده‌های منتخب روش ALASSO



شکل ۲-۱۰ نمودار مقادیر VIF توصیف‌کننده‌های منتخب روش ALASSO

۲- ۳- ۶ مدل‌سازی شبکه عصبی با استفاده از توصیف‌کننده‌های منتخب ALASSO

رابطه بین توصیف‌کننده‌های ساختاری و pIC_{50} با استفاده از یک مدل شبکه عصبی مصنوعی پیشخور با الگوریتم آموزشی پس انتشار خطا ایجاد شد. همان‌طور که در بخش ۲- ۳- ۶ اشاره شد، مجموعه داده‌ها به سه دسته آموزش، ارزیابی و آزمون تقسیم شد. گروه اول (۶۴ ترکیب) برای آموزش مدل شبکه عصبی و از گروه دوم (۱۳ ترکیب) برای ارزیابی و انتخاب مدل بهینه استفاده شد. باقی‌مانده ترکیبات در مدل‌سازی شبکه عصبی شرکت نداشتند و به‌عنوان یک مجموعه آزمون خارجی مستقل، برای بررسی تعمیم‌پذیری مدل در شبکه عصبی تعریف شدند. به‌منظور یافتن شبکه عصبی بهینه، پارامترهای شبکه عصبی از جمله تعداد ورودی‌ها، تعداد گره‌های لایه پنهان، دورهای آموزشی و توابع انتقال (لگاریتم سیگموئیدی و تانژانت هایپربولیک سیگموئیدی به ترتیب با توابع logsig و tansig در جعبه ابزار متلب) و توابع آموزش لونبرگ مارکورات و تنظیم بایزین (به ترتیب با توابع آموزشی trainlm و trainbr در جعبه ابزار متلب شناخته می‌شوند) به‌طور هم‌زمان بهینه شد. در تمام مدل‌های شبکه عصبی مصنوعی، تابع خطی به‌عنوان لایه خروجی استفاده شد. توصیف‌کننده‌های منتخب روش ALASSO بر اساس بزرگی ضرایب استاندارد شده ALASSO (جدول ۲-۸) چیده شدند و به‌عنوان ورودی در ساخت مدل شبکه عصبی مورد استفاده قرار گرفتند. در بهینه‌سازی هم‌زمان پارامترهای شبکه عصبی مصنوعی، تعداد توصیف کننده‌ها، گره‌های لایه پنهان و دورهای آموزشی به‌طور هم‌زمان از ۲ تا ۹ (با گام ۱)، ۲ تا ۱۰ (با گام ۱) و ۵ تا ۵۰ (با گام ۵) تغییر یافت. مدل‌های ممکن ایجاد شده ANN، با استفاده از مجموعه داده‌های آموزش (شامل ۶۴ ترکیب) ساخته شدند و برای پیش بینی مقادیر pIC_{50} مربوط به ترکیبات مجموعه ارزیابی استفاده شدند. مقادیر MSE مجموعه ارزیابی برای تمام مدل‌های آموزش دیده محاسبه شد و به‌عنوان معیاری برای انتخاب بهترین مدل ANN مورد استفاده قرار گرفت. نتایج معماری‌های شبکه و مقادیر MSE برای چهار مدل ANN با توابع آموزش و انتقال متفاوت که دارای کم‌ترین مقدار MSE هستند، در جدول ۲-۹ آورده

شده است. نتایج نشان می‌دهند که مدل ANN با ۹ توصیف کننده منتخب روش ALASSO به‌عنوان ورودی، ۲ گره در لایه پنهان و ۵ دور آموزش با تابع انتقال تانژانت هایپربولیک سیگموئیدی و الگوریتم آموزشی LM دارای کم‌ترین مقدار MSE برای پیش بینی داده‌های مجموعه ارزیابی می‌باشد. بنابراین مدل با نماد ALASSO-LM-ANN نشان داده شده با معماری ۱-۲-۹ به‌عنوان مدل برتر برای پیش‌بینی مقادیر $pIC_{۵}$. ترکیبات مورد مطالعه انتخاب شد.

جدول ۲-۹ ساختارهای شبکه‌های توسعه یافته با توصیف‌کننده‌های منتخب ALASSO با کمترین MSE مجموعه ارزیابی

تعداد توصیف کننده	تابع آموزش	تابع انتقال	تعداد گره	تعداد دور آموزش	$MSE_{validation}$	$R^2_{validation}$
۹	تنظیم بایزین	لگاریتم-سیگموئید	۵	۵	۰/۱۴	۰/۸۲
۹	لونبرگ- مارکوارت	لگاریتم-سیگموئید	۲	۴۰	۰/۱۲	۰/۸۵
۹	تنظیم بایزین	تانژانت-سیگموئید	۹	۵	۰/۱۶	۰/۸۱
۹	**لونبرگ- مارکوارت**	**تانژانت-سیگموئید**	**۲**	**۵**	**۰/۱۱**	**۰/۸۶**

با هدف مقایسه عملکرد ALASSO در انتخاب متغیر برای مدل‌سازی ANN، از روش کلاسیک متداول SR برای انتخاب موثرترین توصیف کننده‌ها استفاده شد. پس از اجرای SR روی مجموعه داده‌های ارزیابی، ۱۸ توصیف کننده به‌عنوان بهترین زیرمجموعه از توصیف کننده‌های SR انتخاب شد. به منظور مقایسه کارایی SR و SCAD، به تعداد ۹ توصیف کننده روش SR (چیده شده بر اساس بزرگی ضرایب رگرسیونی) برای طراحی و بهینه سازی مدل ANN مورد استفاده قرار گرفتند. نتایج نشان داد مدل ANN با تابع آموزش LM و با استفاده از زیر مجموعه‌های ۲ تا ۹ تایی از توصیف کننده‌های منتخب روش SR با تعداد ۳ گره در لایه پنهان و ۵ دور آموزش حداقل MSE را برای مجموعه ارزیابی ایجاد نمود. بنابراین مدل بهینه SR-LM-ANN برای پیش بینی فعالیت دارویی بازدارنده‌های $3CL^{pro}$ موجود در مجموعه آزمون استفاده شد.

۲- ۳- ۷ ارزیابی مدل ALASSO-LM-ANN

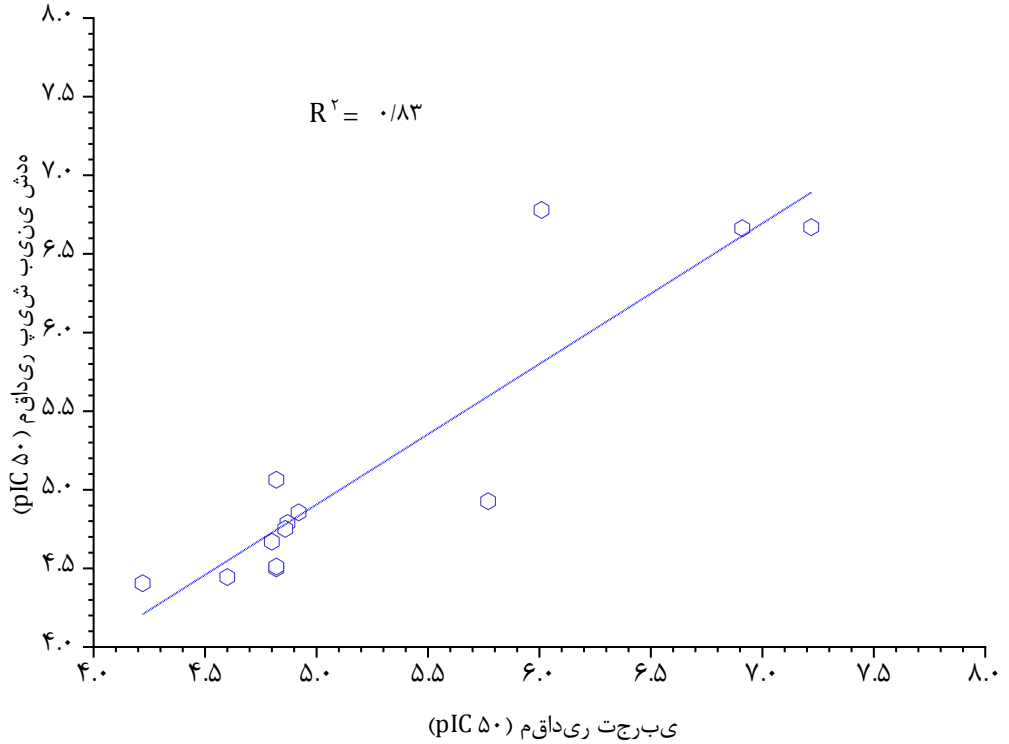
ارزیابی قدرت پیش‌بینی مدل شبکه عصبی پیشنهادی (ALASSO-LM-ANN) گامی مهم در مطالعات QSAR است. همان‌طور که در بخش ۲- ۲- ۷ اشاره شد، قدرت پیش بینی، اعتبار و تعمیم پذیری مدل‌های توسعه یافته در این رساله، با استفاده از پیش بینی پاسخ ($pIC_{۵۰}$) داده‌های مجموعه آزمون با مدل بهینه (ALASSO-LM-ANN)، پیش بینی پاسخ کل ترکیبات با تکنیک رد مرحله‌ای تک تک (LOO)، محاسبه پارامترهای آماری، دامنه کاربرد و آزمون Y-تصادفی نیز مورد ارزیابی بیش‌تر قرار گرفت که در ادامه نتایج روش‌های ارزیابی متفاوت آورده شده است.

۲- ۳- ۷- ۱ ارزیابی مدل ALASSO-LM-ANN با استفاده از مجموعه آزمون

عدم حضور داده‌های مجموعه آزمون در هر دو مرحله انتخاب توصیف‌کننده‌های مؤثر به روش ALASSO و مدل‌سازی با ANN، معیار مناسبی برای ارزیابی مدل ALASSO-LM-ANN به‌حساب می‌آیند. به‌این منظور فعالیت دارویی ($pIC_{۵۰}$) مربوط به ۱۳ ترکیب موجود در مجموعه آزمون با استفاده از مدل ANN در شرایط بهینه ALASSO-LM-ANN با معماری ۱-۲-۹ پیش‌بینی شدند و نتایج در جدول ۲-۱۰ آورده شده است. مقادیر خطای کم اغلب ترکیبات موجود در مجموعه آزمون (جدول ۲-۱۰) نشان‌دهنده قدرت پیش‌بینی مناسب مدل توسعه‌یافته ALASSO-LM-ANN است. شکل ۲-۱۰ از رسم مقادیر پیش‌بینی شده $pIC_{۵۰}$ به‌وسیله مدل بهینه بر حسب مقادیر واقعی $pIC_{۵۰}$ به دست آمد. با توجه به این‌که R^2 مربوط به مجموعه آموزش و ارزیابی برای مدل ALASSO-LM-ANN به ترتیب برابر با ۰/۸۲ و ۰/۸۶ است بنابراین با مقایسه مقادیر مذکور با مقدار R^2 مجموعه آزمون (شکل ۲-۱۱) قدرت پیش‌بینی و تعمیم‌پذیری مدل ALASSO-LM-ANN اثبات می‌شود.

جدول ۲-۱۰ نتایج حاصل از ارزیابی مدل ALASSO-LM-ANN با استفاده از مجموعه آزمون

شماره ترکیب	$pIC_{۵۰}$		درصد خطا
	مقدار واقعی	مقدار پیش‌بینی شده	
۳	۴/۸۷	۴/۷۹	-۱/۶۷
۶	۶/۰۱	۶/۷۸	۱۲/۸
۱۹	۴/۸۲	۵/۰۶	۵/۰۵
۲۳	۴/۶	۴/۴۴	-۳/۳۷
۲۹	۴/۸	۴/۶۷	-۲/۷۲
۵۱	۴/۸۶	۴/۷۵	-۲/۲۵
۵۴	۵/۷۷	۴/۹۳	-۱۴/۶
۶۱	۴/۹۲	۴/۸۶	-۱/۳
۶۳	۴/۸۲	۴/۵	-۶/۶۳
۶۵	۴/۸۲	۴/۵۱	-۶/۴
۷۰	۴/۲۲	۴/۴۱	۴/۴
۷۴	۷/۲۲	۶/۶۷	-۷/۶۲
۸۳	۶/۹۱	۶/۶۶	-۳/۵۷



شکل ۲-۱۱ نمودار تغییرات مقادیر پیش‌بینی شده pIC_{50} به‌وسیله مدل ALASSO-LM-ANN در شرایط بهینه در مقابل مقادیر تجربی برای داده‌های مجموعه آزمون

۲- ۳- ۷- ۲ ارزیابی مدل ALASSO-LM-ANN با پیش بینی pIC_{50} تمام ترکیبات مجموعه داده با استفاده از روش رد مرحله‌ای تک تک

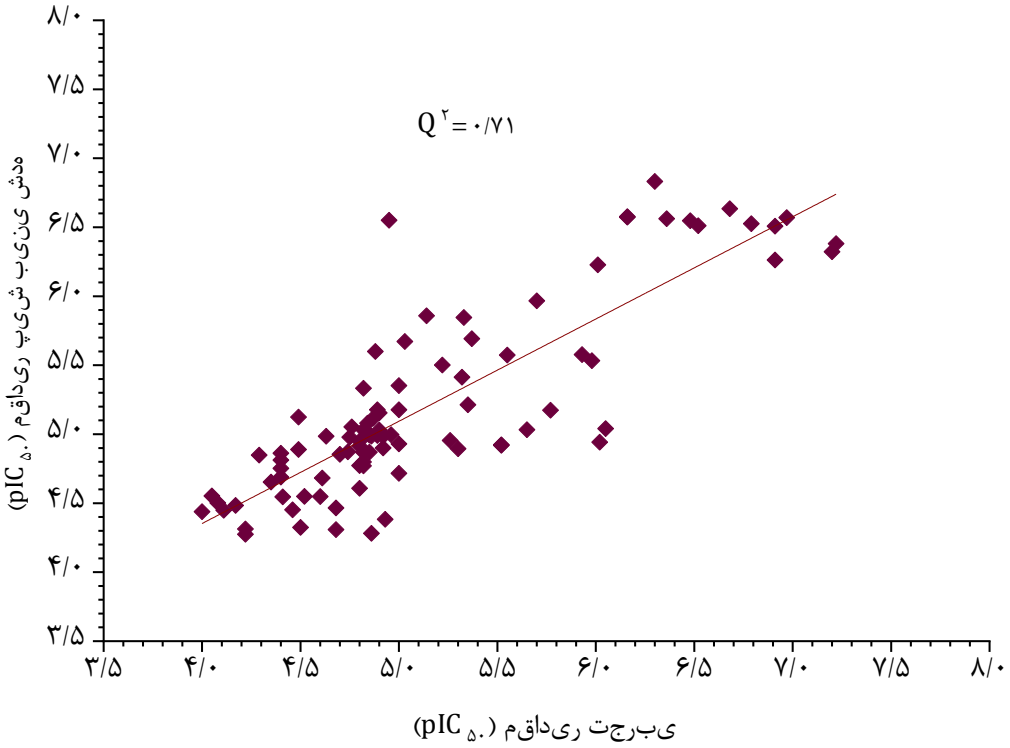
در این بخش، قدرت پیش‌بینی مدل ALASSO-LM-ANN، با استفاده از تکنیک رد مرحله‌ای تک تک (LOO) برای پیش‌بینی فعالیت دارویی همه ترکیبات مورد بررسی قرار گرفت. در این تکنیک، هر داده یک‌بار به‌عنوان داده آزمون در نظر گرفته شد و مدل ALASSO-LM-ANN با معماری ۱-۲-۹ با ترکیبات باقی‌مانده در مجموعه داده‌ها آموزش داده شد و فعالیت دارویی داده آزمون کنار گذاشته شده، پیش‌بینی شد. این روش ۹۰ بار برای ۹۰ بازدارنده $3CL^{pro}$ تکرار شد تا زمانی که فعالیت دارویی همه ترکیبات توسط مدل ALASSO-LM-ANN پیش‌بینی شد و نتایج حاصله در جدول ۲-۱۱ خلاصه شد. برای مطالعه بیش‌تر قدرت پیش بینی مدل، مقادیر پیش‌بینی‌شده pIC_{50} همه ترکیبات بر حسب مقادیر تجربی آن‌ها رسم شد (شکل ۲-۱۲). همان‌طور که شکل ۲-۱۲ نشان می‌دهد، مقدار Q^2_{LOO} از حد قابل قبول (Q^2=۰/۵) بالاتر است. بنابراین این نتیجه حاکی از استحکام مناسب مدل می‌باشد. علاوه بر این، مقادیر باقی‌مانده استاندارد شده با استفاده از رابطه ۱-۱۲ محاسبه شد و نتایج آن در مقابل مقادیر واقعی pIC_{50}، رسم شد (شکل ۲-۱۳). با توجه به نمودار باقی‌مانده‌ها، توزیع تصادفی مقادیر باقی‌مانده حول محور صفر، نشان‌دهنده عدم وجود خطای سیستماتیک در مدل پیشنهادی شبکه عصبی توسعه یافته با استفاده از روش ALASSO به‌عنوان روش انتخاب متغیر (ALASSO-LM-ANN) است.

جدول ۲-۱۱ نتایج حاصل از ارزیابی مدل ALASSO-LM-ANN به روش رد مرحله‌ای تک تک برای کل داده‌ها

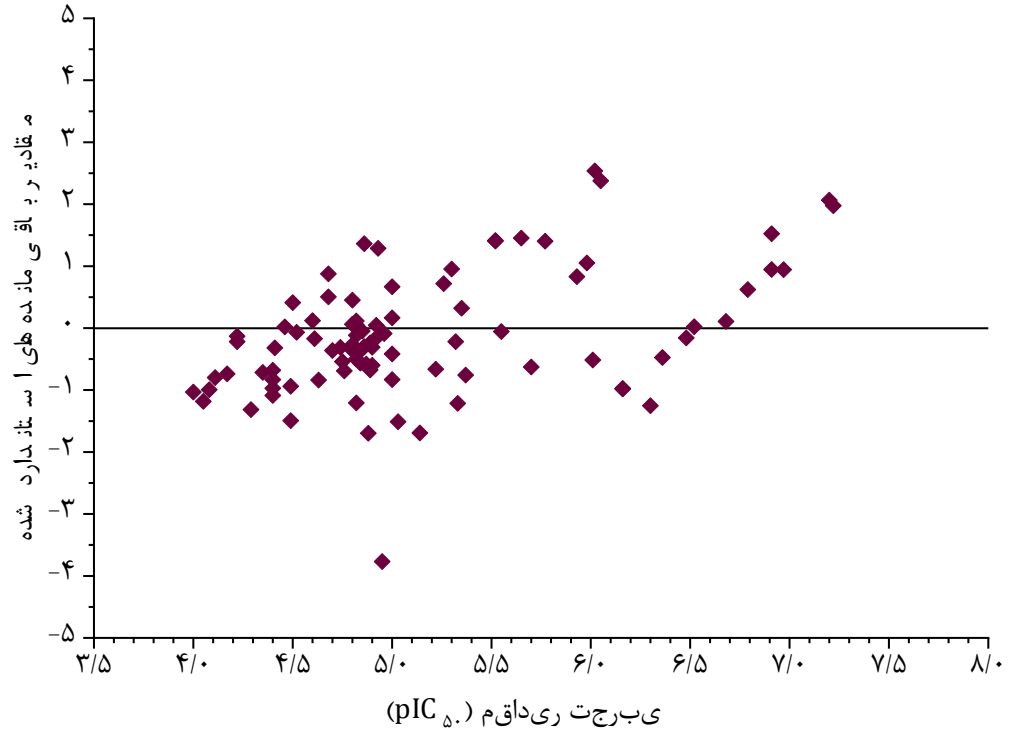
شماره ترکیب	pIC_{50}		درصد خطا	شماره ترکیب	pIC_{50}		درصد خطا
	مقدار واقعی	مقدار پیش‌بینی شده			مقدار واقعی	مقدار پیش‌بینی شده	
۱	۵/۱۴	۵/۸۶	۱۳/۹۸	۲۹	۴/۸	۴/۹۱	۲/۳۸
۲	۵/۰۳	۵/۶۷	۱۲/۷۶	۳۰	۴/۷۴	۴/۸۷	۲/۷۹
۳	۴/۸۷	۵/۱۲	۵/۰۵	۳۱	۴/۷	۴/۸۵	۳/۲۷
۴	۴/۸۸	۵/۶	۱۴/۷۶	۳۲	۴/۴	۴/۷۵	۸/۰۲
۵	۵/۷	۵/۹۷	۴/۶۸	۳۳	۴/۱۱	۴/۴۵	۸/۲۶
۶	۶/۰۱	۶/۲۳	۳/۶۳	۳۴	۴/۵	۴/۳۲	-۳/۹
۷	۵/۳۲	۵/۴۱	۱/۷۵	۳۵	۴/۴۹	۴/۸۹	۸/۸۸
۸	۶/۰۲	۴/۹۴	-۱۷/۹۱	۳۶	۴/۴۶	۴/۴۵	-۰/۱۹
۹	۴/۹۵	۶/۵۵	۳۲/۳۳	۳۷	۵/۰۰	۵/۳۵	۷/۰۶
۱۰	۴/۶۳	۴/۹۹	۷/۶۸	۳۸	۴/۲۹	۴/۸۵	۱۳/۰۳
۱۱	۴/۹	۵/۱۵	۵/۱۹	۳۹	۵/۳۵	۵/۲۱	-۲/۵۵
۱۲	۴/۷۶	۵/۰۵	۶/۱۷	۴۰	۴/۹	۵/۰۳	۲/۶۷
۱۳	۶/۵۲	۶/۵۱	-۰/۱۴	۴۱	۵/۹۳	۵/۵۸	-۵/۹۴
۱۴	۶/۰۵	۵/۰۴	-۱۶/۶۹	۴۲	۵/۶۵	۵/۰۳	-۱۰/۹۳
۱۵	۵/۲۲	۵/۵	۵/۴	۴۳	۵/۳۷	۵/۶۹	۵/۹۶
۱۶	۴/۹۲	۴/۹	-۰/۳۹	۴۴	۴/۹۳	۴/۳۸	-۱۱/۰۸
۱۷	۴/۸۹	۵/۱۸	۵/۸۷	۴۵	۴/۱۷	۴/۴۸	۷/۵۲
۱۸	۴/۸۹	۴/۹۹	۲/۰۷	۴۶	۴/۰۸	۴/۵	۱۰/۳۴
۱۹	۴/۸۲	۵/۳۳	۱۰/۶۴	۴۷	۴/۸۶	۴/۲۸	-۱۱/۹۱
۲۰	۴/۸	۴/۹۵	۳/۱۴	۴۸	۵/۲۶	۴/۹۶	-۵/۷۹
۲۱	۴/۸	۴/۷۷	-۰/۵۶	۴۹	۴/۸۵	۴/۹۹	۲/۹۴
۲۲	۴/۸	۴/۶۱	۴/۰۰	۵۰	۵/۰۰	۵/۱۸	۳/۵۵
۲۳	۴/۶	۴/۵۵	-۱/۱۲	۵۱	۴/۸۶	۴/۹۹	۲/۶۱
۲۴	۴/۴۹	۵/۱۲	۱۴/۱۳	۵۲	۴/۴	۴/۸۶	۱۰/۵
۲۵	۵/۵۲	۴/۹۲	-۱۰/۸۷	۵۳	۵/۹۸	۵/۵۳	-۷/۴۷
۲۶	۵/۳	۴/۸۹	-۷/۶۴	۵۴	۵/۷۷	۵/۱۷	-۱۰/۳۳
۲۷	۵/۰۰	۴/۹۳	-۱/۴	۵۵	۴/۷۵	۴/۹۸	۴/۸۲
۲۸	۴/۸۲	۴/۸۷	۱/۰۳	۵۶	۵/۵۵	۵/۵۷	۰/۴۲

ادامه جدول ۲-۱۱

درصد خطا	pIC۵۰		شماره ترکیب	درصد خطا	pIC۵۰		شماره ترکیب
	مقدار پیش‌بینی شده	مقدار واقعی			مقدار پیش‌بینی شده	مقدار واقعی	
-۵/۷۴	۶/۵۷	۶/۹۷	۸۴	۹/۶۸	۵/۸۵	۵/۳۳	۵۷
۱۲/۴	۴/۵۵	۴/۰۵	۸۵	-۱۰/۸۱	۴/۹۲	۵/۵۲	۵۸
۱/۵۷	۴/۶۸	۴/۶۱	۸۶	-۵/۶۷	۴/۷۲	۵/۰۰	۵۹
-۴/۵۷	۴/۴۷	۴/۶۸	۸۷	۰/۷۵	۵/۰۰	۴/۹۶	۶۰
۳/۰۸	۴/۵۵	۴/۴۱	۸۸	۱/۳۶	۴/۹۹	۴/۹۲	۶۱
۴/۹۶	۵/۰۸	۴/۸۴	۸۹	۰/۴۱	۴/۸۷	۴/۸۵	۶۲
-۷/۹۴	۴/۳۱	۴/۶۸	۹۰	-۱/۰۲	۴/۷۷	۴/۸۲	۶۳
				۴/۳۹	۵/۰۳	۴/۸۲	۶۴
				-۰/۴۴	۴/۸	۴/۸۲	۶۵
				۰/۶۵	۴/۵۵	۴/۵۲	۶۶
				۹/۳۸	۴/۸۱	۴/۴۰	۶۷
				۶/۵۷	۴/۶۹	۴/۴	۶۸
				۶/۹۸	۴/۶۵	۴/۳۵	۶۹
				۱/۳۱	۴/۲۸	۴/۲۲	۷۰
				۲/۲۱	۴/۳۱	۴/۲۲	۷۱
				۱۰/۹۶	۴/۴۴	۴/۰۰	۷۲
				-۱۲/۱۹	۶/۳۲	۷/۲	۷۳
				-۱۱/۶۱	۶/۳۸	۷/۲۲	۷۴
				-۵/۸۲	۶/۵۱	۶/۹۱	۷۵
				-۰/۶۸	۶/۶۳	۶/۶۸	۷۶
				۸/۴۴	۶/۸۳	۶/۳۰	۷۷
				-۳/۸۹	۶/۵۳	۶/۷۹	۷۸
				۶/۷۷	۶/۵۸	۶/۱۶	۷۹
				۳/۱۷	۶/۵۶	۶/۳۶	۸۰
				۱/۰۴	۶/۵۵	۶/۴۸	۸۱
				۶/۷۱	۶/۵۷	۶/۱۶	۸۲
				-۹/۳۷	۶/۲۶	۶/۹۱	۸۳



شکل ۲-۱۲ نمودار تغییرات مقادیر پیش‌بینی شده همه داده‌ها بر اساس تکنیک LOO در مقابل مقادیر تجربی



شکل ۲-۱۳ نمودار باقی‌مانده‌های حاصل از پیش‌بینی فعالیت دارویی ترکیبات با استفاده از تکنیک LOO و مقادیر تجربی برحسب مقادیر تجربی

همان‌طور که در بخش ۲- ۳- ۶ اشاره شد، به منظور مقایسه عملکرد روش انتخاب متغیر ALASSO از روش SR استفاده شد. پس از مقایسه مدل SR-LM-ANN با مدل ALASSO-LM-ANN مشاهده می‌شود که R^2_{test} و MSE برای مجموعه آزمون مربوط به مدل SR-LM-ANN به‌ترتیب برابر با ۰/۵۳ و ۰/۴۰ به‌دست آمد. علاوه بر این Q^2_{LOO} و MSE مربوط به کل داده‌های پیش‌بینی شده با تکنیک LOO به‌ترتیب برابر با ۰/۶۷ و ۰/۲۲ به‌دست آمد. نتایج به‌دست‌آمده تائید می‌کند که مدل ALASSO-LM-ANN دارای قابلیت پیش‌بینی قابل توجهی در مقایسه با مدل SR-LM-ANN است، که این موضوع نشان‌دهنده عملکرد بالای ALASSO در انتخاب توصیف‌کننده‌های مهم در مطالعات QSAR مبتنی بر ANN است

۲- ۳- ۷- ۳ ارزیابی مدل ALASSO-LM-ANN با استفاده از پارامترهای آماری

برای بررسی‌های بیش‌تر قدرت پیش‌بینی مدل پیشنهادی ALASSO-LM-ANN، پارامترهای آماری ذکر شده در بخش ۱- ۵- ۸- ۴ برای فعالیت دارویی پیش‌بینی شده ترکیبات مجموعه آزمون و فعالیت دارویی پیش بینی شده برای کل ترکیبات به روش رد مرحله‌ای تک تک، محاسبه و در جدول ۲-۱۲ خلاصه شدند. نتایج حاصله (جدول ۲-۱۲) نشان می‌دهد که پارامترهای آماری برای مدل ALASSO-LM-ANN دارای مقادیری در محدوده قابل قبول هستند. به‌طوری‌که پارامترهای مربوط به آزمون تروپشا و روی همگی از مقدار هشدار ۰/۵ بزرگ‌تر و به مقدار R^2 نزدیک هستند. شیب نمودار حاصل از مقادیر پیش بینی شده بر حسب مقادیر تجربی (و بالعکس) در عرض از مبدأ صفر نیز در محدوده ۰/۸۵ تا ۱/۱۵ قرار دارند که این نتیجه نیز حاکی از صحت مدل توسعه یافته ANN با توصیف کننده‌های منتخب روش ALASSO می‌باشد.

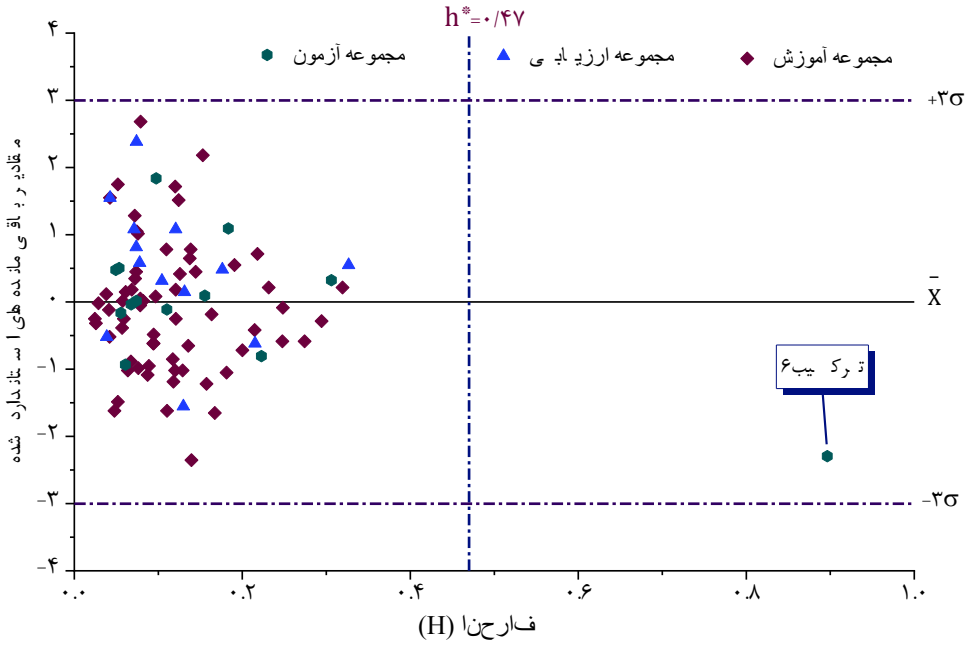
جدول ۲-۱۲ پارامترهای آماری محاسبه شده برای مجموعه آزمون و داده‌های پیش‌بینی شده با تکنیک LOO برای مدل ALASSO-LM-ANN

ردیف	پارامتر آماری	نتایج pIC50 پیش بینی شده با ALASSO-LM-ANN		محدوده قابل قبول
		ترکیبات مجموعه آزمون	کل ترکیبات به روش LOO	
۱	PRESS	۱/۴۷	۶/۳۳	-
۲	SEP	۰/۳۴	۰/۴۳	-
۳	MAE	۰/۲۸	۰/۳۳	-
۴	REP(%)	۶/۰۸	۸/۲۷	-
۵	MSE	۰/۱۱	۰/۱۸	-
۶	MRE	۵/۱۴	۶/۳۲	-
۷	R^2	۰/۸۳	-	R^2>۰/۶
۸	Q^2_{Loo}	-	۰/۷۱	$\boldsymbol{Q^2_{Loo}}$> ۰/۵
۹	R^2_0	۰/۸۵	۰/۶۳	نزدیک به R^2
۱۰	R^2_0 نسبی	۰/۰۱	۰/۱۱	<۰/۱
۱۱	R^2_m	۰/۷۷	۰/۵۱	>۰/۵
۱۲	R'^2_0	۰/۸۶	۰/۷۱	نزدیک به R^2
۱۳	R'^2_0 نسبی	۰/۰۰	۰/۰۰	<۰/۱
۱۴	R'^2_m	۰/۷۷	۰/۶۳	>۰/۵
۱۵	R-R	۰/۰۱	۰/۰۸	<۰/۳
۱۶	k	۰/۹۸	۱	$۰/۸۵ \leq \boldsymbol{k} \leq ۱/۱۵$
۱۷	k′	۱/۰۳	۰/۹۹	$۰/۸۵ \leq \mathbf{k'} \leq ۱/۱۵$

۲- ۳- ۷- ۴ ارزیابی ALASSO-LM-ANN با استفاده از دامنه کاربرد

دامنه کاربرد (AD) یک استراتژی ارزشمند برای تائید اعتبار مدل‌های QSAR توسعه یافته است. به‌منظور نمایش دامنه کاربرد مدل ALASSO-LM-ANN، نمودار ویلیام (شکل ۲-۱۴) مورد استفاده قرار گرفت. بنابراین مطابق با روش کار بخش ۱- ۵- ۸- ۵ و ۲- ۲- ۷- ۴ ابتدا مقدار انحراف (H) با استفاده از رابطه ۱-۱۳ به‌دست آمد. سپس بر اساس مقادیر پیش بینی شده pIC_{50} توسط مدل ALASSO-LM-

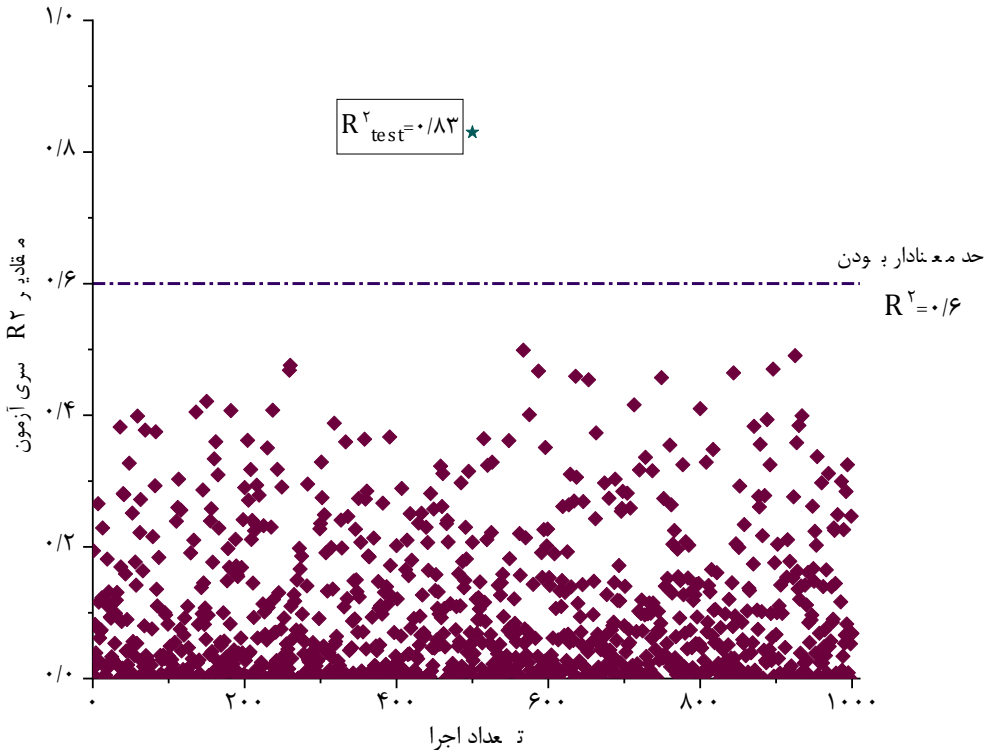
ANN، مقادیر باقی‌مانده‌ها برای همه ترکیبات محاسبه شد و مقادیر مربوط به باقی‌مانده‌های استاندارد شده با استفاده از رابطه ۱-۱۴ به‌دست آمد. نمودار ویلیام از رسم باقی‌مانده‌های استاندارد شده بر حسب مقادیر انحراف به‌دست آمد (شکل ۲-۱۴). نمودار ویلیام نشان می‌دهد که مقادیر H محاسبه شده کمتر از حد آستانه‌ای h* است. خط عمودی، نشان‌دهنده مقدار آستانه h* است که با استفاده از رابطه ۳p/n (p برابر با تعداد توصیف‌کننده‌ها به‌علاوه ۱ (۱۰) و n نیز برابر با داده‌های مجموعه آموزش و برابر با ۶۴ هست) محاسبه شد و برابر با ۰/۴۷ به‌دست آمد. علاوه بر این مقادیر باقی‌مانده‌های استاندارد شده در محدوده ±۳σ قرار دارند. خطوط افقی، نشان‌دهنده حدود ±۳σ برای مقادیر باقی‌مانده‌های استاندارد شده است. نتایج شکل ۲-۱۴ نشان می‌دهد که ۹۹٪ از کل داده‌های به کار رفته در ساخت، ارزیابی و آزمون مدل ALASSO-LM-ANN در فواصل آستانه‌ای دامنه کاربرد قرار دارند. در نتیجه، مجموعه داده‌ها، از جمله مجموعه آموزش، ارزیابی و آزمون در محدوده قابل اعتماد وجود دارند و مدل ALASSO-LM-ANN از اعتبار مناسبی برخوردار است.



شکل ۲-۱۴ دامنه کاربرد مدل ALASSO-LM-ANN، خطوط نقطه چین افقی و عمودی در دو انتهای نمودار به ترتیب نمایانگر مقادیر ۳σ ± و h* است.

۲- ۳- ۷- ۵ ارزیابی مدل ALASSO-LM-ANN با استفاده از آزمون Y-تصادفی

اعتبار ارتباط ساختار-فعالیت دارویی ایجاد شده با مدل برتر ALASSO-LM-ANN و عدم وجود ارتباط تصادفی بین ویژگی‌های ساختار ترکیبات شیمیایی و فعالیت دارویی مربوطه با استفاده از آزمون Y-تصادفی مورد بررسی قرار گرفت. برای انجام آزمون Y-تصادفی، ابتدا مقادیر فعالیت دارویی در محدوده ۷/۲۲-۴/۰۰ به تعداد ۱۰۰۰ بار تصادفی شدند. مدل‌های شبکه عصبی ALASSO-LM-ANN با استفاده از پاسخ‌های تصادفی متغیر وابسته مجموعه آموزش ساخته شدند و برای پیش بینی مقادیر فعالیت دارویی ترکیبات مجموعه آزمون به‌کار گرفته شدند. مقادیر R^2 حاصل از پیش بینی فعالیت دارویی ترکیبات مجموعه آزمون با ۱۰۰۰ مدل توسعه یافته با متغیر وابسته تصادفی، به‌دست آمد و بر حسب تعداد اجرا ترسیم شدند (شکل ۲-۱۵). نتایج به‌دست‌آمده در شکل ۲-۱۵ نشان می‌دهد که مقادیر R^2 حاصل از هر بار اجرای مدل ALASSO-LM-ANN با داده‌های تصادفی از مقدار قابل قبول ۰/۶ نیز کوچک‌تر هستند [۴۳]. شکل ۲-۱۵ نشان می‌دهد که مقادیر R^2 آزمون Y- تصادفی، در محدوده ۰/۰۰ تا ۰/۴۹ قرار دارند و تنها ۶/۴ درصد از R^2 ها دارای مقادیر بیش‌تر از ۰/۳ هستند. بنابراین مطابق با نتایج شکل ۲-۱۵، مقادیر R^2 داده‌های تصادفی به‌طور قابل توجهی کوچک‌تر از R^2 مجموعه آزمون (R^2=۰/۸۳) حاصل از مدل پیشنهادی ALASSO-LM-ANN توسعه یافته با متغیر اصلی است. نتایج به‌دست‌آمده ثابت می‌کند که ارتباط بین ۹ توصیف کننده منتخب روش ALASSO و فعالیت دارویی (pIC_{50}.) تصادفی نیست و مدل ALASSO-LM-ANN بر اساس یک رابطه منطقی و معنادار بین متغیرهای مستقل و وابسته ایجاد شده است.



شکل ۲-۱۵ نمودار مقادیر R^2 به‌دست آمده در آزمون Y-تصادفی بر حسب تعداد اجرا برای ۱۰۰۰ اجرای Y-تصادفی و پیش‌بینی فعالیت ترکیبات مجموعه آزمون به‌وسیله مدل ALASSO-LM-ANN با استفاده از پاسخ تصادفی شده در شرایط بهینه

۲- ۴ پیش‌بینی فعالیت دارویی برخی از بازدارنده‌های ایدز و سرطان با استفاده از مدل LAD-LASSO-ANN

۲- ۴- ۱ مقدمه

کمی کردن خواص فیزیکوشیمیایی و فعالیت‌های دارویی ترکیبات دارویی طبیعی و سنتزی و یافتن الگو و روابط حاکم بر کمیت‌های اندازه‌گیری شده همواره موضوع جالب و در حال پیشرفتی است که در روش QSAR به آن پرداخته می‌شود. در سال‌های اخیر استفاده از شیمی محاسباتی و شبیه‌سازی مولکولی جهت طراحی دارو به کمک رایانه مورد توجه محققین دارویی قرار گرفته است. از مزایای روش‌های محاسباتی می‌توان به کاهش تعداد ترکیبات ساخته شده برای پیدا کردن ترکیب رهبر، بالا بردن سرعت محاسبات و آزمایش‌ها با پیش‌بینی قابل اعتماد از طریق ویژگی‌های دارویی ساختار مولکول و هم‌چنین کاهش استفاده از آزمایش‌های حیوانات و بالینی اشاره کرد. ازاین‌رو محققان شیمی، همواره در تلاش هستند تا روش‌های آماری پیشرفته‌تری را برای ایجاد معادلات ریاضی بین داده‌های دارویی و ویژگی‌های ساختاری به‌کار ببرند. بنابراین ارائه یک مدل QSAR با قابلیت پیش‌بینی و تفسیرپذیری بالا همواره مورد توجه بوده است. یک مدل QSAR با کارایی بالا، باید به ازای استفاده از حداقل متغیرهای وابسته، قدرت پیش‌بینی مناسبی را در مقابل ترکیبات شیمیایی جدید نشان دهد. ازاین‌رو استفاده از روش‌های انتخاب متغیر کارآمد برای کاهش تعداد توصیف کننده‌ها توصیه می‌شود. روش‌های انتخاب متغیر با حذف توصیف‌کننده‌های زائد باعث کاهش توصیف کننده‌ها و انتخاب توصیف‌کننده‌های مؤثر می‌شود و تفسیرپذیری مدل افزایش می‌یابد. در طی سال‌های اخیر روش‌های انتخاب متغیر متفاوتی برای گزینش توصیف‌کننده‌هایی با بیش‌ترین ارتباط در مطالعات QSAR استفاده شده است که از این دسته می‌توان به روش‌های انتخاب متغیر کلاسیک و انقباضی اشاره کرد. روش‌های انتخاب متغیر کلاسیک همچون انتخاب پیش‌رو، پس‌رو و روش رگرسیون

گام‌به‌گام در مطالعات QSAR مورد توجه بوده است که همه این‌روش‌ها مبتنی بر روش حداقل مربعات معمولی (OLS) هستند. همان‌طور که واضح است روش OLS در مواجهه با داده‌هایی با تعداد متغیر زیاد و نمونه کم (داده‌هایی با ابعاد بالا و نمونه‌های تنک) به‌دلیل وجود همبستگی بین متغیرها یا هم خطی چندگانه موفق عمل نمی‌کنند. علاوه بر این، وجود مشاهدات دور افتاده از جمله عواملی است که برآوردگر OLS را محدود می‌نماید. برای حل این مشکل می‌توان از برآوردگرهای استوار چون برآوردگر کم‌ترین قدر مطلق انحرافات (LAD) بهره گرفت که علاوه بر مقاوم بودن در برابر مشاهدات دور افتاده، برآوردگرهایی با کارایی مطلوب را ایجاد می‌نمایند. بنابراین ونگ[1] و همکارانش (۲۰۰۷) با افزودن پارامتر جریمه به تابع هدف LAD مدلی را پیشنهاد دادند که علاوه بر مقابله در برابر مشاهدات دور افتاده، قابلیت انتخاب متغیرهای مؤثر را نیز دارد [۳۲]. از جمله روش‌های انتخاب متغیر رگرسیونی انقباضی با تابع جریمه L1 می‌توان به روش انتخاب متغیر انقباضی LASSO [۲۹] اشاره کرد. LASSO به دلیل داشتن تابع جریمه $\lambda \sum_{i=1}^{p}|\beta_i|$ ضرایب پارامترهایی بی‌اهمیت را صفر می‌کند و با توجه به این ماهیت ذاتی، توانایی به‌کارگیری به‌عنوان یک روش انتخاب متغیر کارآمد را دارد [۲۹]. اخیراً محققین از روش LASSO برای انتخاب توصیف‌کننده‌هایی با بیش‌ترین ارتباط با پاسخ دارویی را در مطالعات QSAR استفاده نموده‌اند و کارایی روش را مورد بررسی قرار داده‌اند [۱۶۸-۱۶۴]. در این مطالعه سعی شده است تا به‌طور هم‌زمان از مزایای ذاتی LASSO و LAD در مواجهه با داده‌های با ابعاد بالا استفاده شود تا مدل، علاوه بر مقابله با مشاهدات دور افتاده بتواند در مواجهه با هم خطی چندگانه نیز به‌طور قابل قبولی عمل نماید.

در راستای تحقیقات پیشین، تلاش شد که با جفت کردن روش انتخاب متغیر LAD-LASSO با روش مدل‌سازی شبکه عصبی، مقادیر فعالیت دارویی برخی از ترکیبات دارویی پیش‌بینی شود. بنابراین در این بخش از مطالعه، سعی شد با به‌کارگیری یک روش انتخاب متغیر انقباضی جفت شده با شبکه عصبی،

[1] Wang

مدل‌های QSAR با قدرت پیش‌بینی قابل قبول و تفسیرپذیری مناسب ارائه شود. ازاین‌رو روش LAD-LASSO با قابلیت بالایی در انتخاب متغیر بر روی سه نمونه داده‌های دارویی متشکل از بازدارنده‌های ایدز و سرطان کارسینوم کولورکتال انسان و سرطان ریه اجرا شد. توصیف‌کننده‌های منتخب به‌عنوان ورودی روش مدل‌سازی شبکه عصبی مورد استفاده قرار گرفتند. مدل‌های LAD-LASSO-LM-ANN در هر سه نمونه داده از قدرت پیش‌بینی مناسبی برخوردار هستند. بنابراین از مدل‌های توسعه یافته برای پیشنهاد ترکیبات جدید بالقوه استفاده شد.

۲- ۴- ۲ مجموعه داده‌ها

به‌منظور بررسی کارایی روش انتخاب متغیر LAD-LASSO، از سه مجموعه داده استفاده شد. داده‌ها به‌ترتیب متشکل از ۷۳ ترکیب (بازدارنده‌های ایدز) [۱۲۷, ۱۶۹-۱۷۲]، ۷۲ ترکیب (بازدارنده‌های سرطان از سلول کارسینوم کولورکتال انسان[1]) [۸۰, ۸۱, ۱۷۳] و ۷۰ ترکیب (سرطان بافت ریه انسان[2]) هستند [۸۰, ۸۱, ۱۷۳]. فرمت سیستم ورودی خطی ورودی مولکولی ساده شده (SMILES) و فعالیت‌های دارویی مربوط به هر دسته داده در جدول ۲-۱۳ تا جدول ۲-۱۵ خلاصه شده است. فعالیت‌های دارویی مربوط به هر مجموعه داده بر اساس منفی لگاریتم فعالیت دارویی (EC_{50} برای داده‌های ضد ایدز و IC_{50} برای هر دو دسته داده ضد سرطان) به‌دست آمد. تقسیم بندی مجموعه داده‌ها با استفاده از الگوریتم KS انجام شد و داده‌ها به سه دسته آموزش، ارزیابی و آزمون تقسیم بندی شدند و در مراحل مختلف مدل‌سازی QSAR مورد استفاده قرار گرفتند.

[1] Human colorectal carcinoma
[2] Human lung cancer tissue

جدول ۲-۱۳ مجموعه داده‌های ضد ایدز به همراه مقادیر واقعی و پیش‌بینی شده pEC_{50}

ردیف	SMILES	pEC_{50} واقعی	pEC_{50} پیش‌بینی شده
۱	c1(nc(nc(c1)NC1CCNCC1)Nc1ccc(cc1)C#N)Oc1c(cc(cc1C)C)C	۶/۸۵	۶/۶۸
۲[v]	c1(nc(nc(c1)NC1CCNCC1)Nc1ccc(cc1)C#N)Oc1c(cc(cc1C)C#N)C	۷/۴۲	۶/۸۳
۳[t]	c1(nc(nc(c1)NC1CCN(CC1)Cc1ccc(cc1)S(=O)(=O)N)Nc1ccc(cc1)C#N)Oc1c(cc(cc1C)C)C	۶/۷۲	۶/۰۹
۴	c1(nc(nc(c1)NC1CCN(CC1)Cc1ccc(cc1)S(=O)(=O)C)Nc1ccc(cc1)C#N)Oc1c(cc(cc1C)C)C	۶/۸۹	۷/۰۵
۵	c1(nc(nc(c1)NC1CCN(CC1)Cc1ccncc1)Nc1ccc(cc1)C#N)Oc1c(cc(cc1C)C)C	۶/۶۲	۷/۳۲
۶	c1(nc(nc(c1)NC1CCN(CC1)Cc1ccc(cc1)S(=O)(=O)N)Nc1ccc(cc1)C#N)Oc1c(cc(cc1C)C#N)C	۶/۸۹	۶/۶۸
۷[v]	c1(nc(nc(c1)NC1CCN(CC1)Cc1ccc(cc1)S(=O)(=O)C)Nc1ccc(cc1)C#N)Oc1c(cc(cc1C)C#N)C	۷/۲۴	۷/۰۰
۸[t]	c1(nc(nc(c1)NC1CCN(CC1)Cc1ccncc1)Nc1ccc(cc1)C#N)Oc1c(cc(cc1C)C#N)C	۷/۳۳	۷/۲۶
۹	c1(nc(nc(c1)NC1CCN(CC1)Cc1ccc(cc1)S(=O)(=O)N)Nc1ccc(cc1)C#N)Nc1c(cc(cc1C)C)C	۷/۱۱	۶/۷۷
۱۰	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccncc1)N)Nc1c(cc(cc1C)C)C	۸/۳۱	۸/۲۳
۱۱[v]	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccc(cc1)C(=O)OCC)N)Nc1c(cc(cc1C)C)C	۷/۹۴	۷/۵۷
۱۲	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccc(cc1)C(=O)N)N)Nc1c(cc(cc1C)C)C	۷/۶۹	۸/۰۳
۱۳	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccc(cc1)S(=O)(=O)N)N)Nc1c(cc(cc1C)C)C	۸/۲۴	۷/۶۶
۱۴	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccc(cc1)S(=O)(=O)C)N)Nc1c(cc(cc1C)C)C	۸/۲۶	۷/۷
۱۵	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccccc1)N)Nc1c(cc(cc1C)C)C	۷/۹۹	۸/۹۳
۱۶	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccc(cc1)[N](=O)O)N)Nc1c(cc(cc1C)C)C	۸/۱	۷/۹۲
۱۷	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccc(cc1)C#N)N)Nc1c(cc(cc1C)C)C	۸/۲۵	۸/۴۸
۱۸[t]	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccncc1)NC)Nc1c(cc(cc1C)C)C	۸/۱۴	۸/۳
۱۹	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccc(cc1)C(=O)OCC)NC)Nc1c(cc(cc1C)C)C	۷/۷۸	۷/۸۵
۲۰	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccc(cc1)C(=O)N)NC)Nc1c(cc(cc1C)C)C	۸/۳۴	۸/۲۵
۲۱[v]	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccc(cc1)S(=O)(=O)N)NC)Nc1c(cc(cc1C)C)C	۸/۱۶	۷/۸۹
۲۲[v]	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccc(cc1)S(=O)(=O)C)NC)Nc1c(cc(cc1C)C)C	۸/۲۲	۷/۸۸
۲۳	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccncc1)OC)Nc1c(cc(cc1C)C)C	۷/۸۸	۸/۱۳
۲۴[t]	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccc(cc1)C(=O)OCC)OC)Nc1c(cc(cc1C)C)C	۷/۸۳	۷/۹۱
۲۵	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccc(cc1)C(=O)N)OC)Nc1c(cc(cc1C)C)C	۸/۰۶	۸/۰۱
۲۶	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccc(cc1)S(=O)(=O)N)OC)Nc1c(cc(cc1C)C)C	۷/۹۷	۷/۹۶
۲۷	c1(nc(nc(n1)NC1CCN(CC1)Cc1ccc(cc1)S(=O)(=O)C)OC)Nc1c(cc(cc1C)C)C	۷/۹۲	۷/۸۵
۲۸	c1(c(cnc(c1)Nc1ccc(cc1)C#N)N(=O)=O)Oc1c(cccc1C)C	۶/۲۴	۶/۲۸
۲۹	c1(c(cnc(c1)Nc1ccc(cc1)C#N)N(=O)=O)Oc1c(cc(cc1C)C)C	۷/۲۵	۶/۲۴
۳۰	c1(c(cnc(c1)Nc1ccc(cc1)C#N)N(=O)=O)Oc1c(cc(cc1C)C#N)C	۶/۹۶	۶/۲
۳۱[t]	c1(c(cnc(c1)Nc1ccc(cc1)C#N)N(=O)=O)Oc1c(cc(cc1C)Br)C	۶/۵۷	۶/۴۹
۳۲	c1(c(cnc(c1)Nc1ccc(cc1)C#N)N(=O)=O)Oc1c(cc(cc1C)Cl)C	۶/۸	۵/۹۷
۳۳	c1(c(cnc(c1)Nc1ccc(cc1)C#N)N(=O)=O)Oc1c(cccc1OC)OC	۵/۳۷	۵/۳۴
۳۴[v]	c1(c(cnc(c1)Nc1ccc(cc1)C#N)[N](=O)O)Oc1c(cccc1Cl)Cl	۶/۲۱	۵/۹۵
۳۵[t]	c1(c(cnc(c1)Nc1ccc(cc1)C#N)N(=O)=O)Oc1c(cc(cc1Cl)Cl)Cl	۶/۷۷	۶/۱
۳۶	c1(c(cnc(c1)Nc1ccc(cc1)C#N)N(=O)=O)Oc1c(cc(cc1Cl)N(=O)=O)Cl	۴/۷۷	۵/۰۰
۳۷	c1(c(cnc(c1)Nc1ccc(cc1)C#N)N(=O)=O)Oc1c(cc(cc1Br)Br)Br	۶/۹۶	۶/۱۲
۳۸	c1(c(cnc(c1)Nc1ccc(cc1)C#N)N(=O)=O)Oc1c(cc(cc1Br)C)Br	۷/۴۷	۶/۶۹
۳۹	c1(c(cnc(c1)Nc1ccc(cc1)C#N)N(=O)=O)Oc1c(cc(cc1F)F)F	۶/۱۴	۶/۴۷
۴۰	c1(c(cnc(c1)Nc1ccc(cc1)C#N)N(=O)=O)Nc1c(cc(cc1C)C)C	۴/۵۲	۵/۲۶
۴۱	c1(c(cnc(c1)Nc1ccc(cc1)C#N)N(=O)=O)Nc1c(cc(cc1Br)C)Br	۵/۴۱	۵/۹۳
۴۲	c1(c(cnc(c1)Nc1ccc(cc1)C#N)[N](=O)O)Nc1c(cc(cc1Br)Br)F	۵/۵۲	۵/۸۳
۴۳	c1(c(cnc(c1)Nc1ccc(cc1)N(=O)=O)N(=O)=O)Oc1c(cc(cc1C)C#N)C	۶/۱۴	۶/۰۳

ادامه جدول ۲-۱۳

ردیف	SMILES	pEC_{50} واقعی	pEC_{50} پیش‌بینی شده
۴۴[t]	c1(c(nnc(c1)Nc1ccc(cc1)Cl)Cl)Oc1c(cc(cc1C) C)C	۶/۸۶	۶/۳۲
۴۵[v]	c1(c(nnc(c1)Nc1ccc(cc1)C)Cl)Oc1c(cc(cc1C)C)C	۶/۷	۶/۲۶
۴۶	c1(c(nnc(c1)Nc1ccc(cc1)C#N)Cl)Oc1c(cc(cc1C)C)C	۷/۱	۶/۳۴
۴۷	c1(c(nnc(c1)Nc1ccc(cc1)C#N)Cl)Oc1c(cc(cc1Cl)Cl)Cl	۶/۸۱	۵/۹۳
۴۸	c1(c(nnc(c1)Nc1ccc(cc1)C#N)Cl)Oc1c(cc(cc1Br)Br)Br	۶/۷۴	۶/۰۸
۴۹	c1(c(nnc(c1)Nc1ccc(cc1)C#N)Cl)Oc1c(cc(cc1Br)C)Br	۷/۴۷	۶/۶۳
۵۰	c1(c(nnc(c1)Nc1ccc(cc1)C#N)Cl)Oc1c(cc(cc1C)Br)C	۶/۴	۶/۶۰
۵۱	c1(c(nnc(c1)Nc1ccc(cc1)C#N)Cl)Oc1c(cccc1Cl)Cl	۶/۶۶	۵/۹۹
۵۲	c1(c(nnc(c1)Nc1ccc(cc1)C#N)Cl)Oc1c(cc(cc1C)C#N)C	۶/۸۴	۶/۶۶
۵۳	c1(c(nnc(c1)Nc1ccc(cc1)C#N)Cl)Oc1c(cccc1C)C	۶/۲۷	۶/۵۲
۵۴	c1(c(=O)[nH]nc(c1)Nc1ccc(cc1)Cl)Oc1c(cc(cc1C)C)C	۶/۱۱	۵/۹۶
۵۵[v]	c1(c(=O)[nH]nc(c1)Nc1ccc(cc1)C)Oc1c(cc(cc1C)C)C	۵/۸۱	۶/۱۹
۵۶	c1(c(=O)[nH]nc(c1)Nc1ccc(cc1)N(=O)=O)Oc1c(cc(cc1C)C)C	۵/۲۹	۵/۵۹
۵۷	c1(c(=O)[nH]nc(c1)Nc1ccc(cc1)OC)Oc1c(cc(cc1C)C)C	۵/۸	۶/۴۴
۵۸[t]	c1(c(=O)[nH]nc(c1)Nc1ccc(cc1)C#N)Oc1c(cc(cc1C)C)C	۶/۵۹	۵/۹۳
۵۹[v]	c1(c(=O)[nH]nc(c1)Nc1ccc(cc1)C#N)Oc1c(cc(cc1Br)Br)Br	۶/۲۰	۶/۰۰
۶۰	c1(c(=O)[nH]nc(c1)Nc1ccc(cc1)C#N)Oc1c(cc(cc1Br)C)Br	۶/۶۸	۶/۲۶
۶۱[v]	c1(c(=O)[nH]nc(c1)Nc1ccc(cc1)C#N)Oc1c(cc(cc1C)Br)C	۵/۹۲	۶/۳
۶۲	c1(c(=O)[nH]nc(c1)Nc1cc(ccc1)Cl)Oc1c(cc(cc1C)C)C	۴/۸۴	۴/۹۳
۶۳[t]	c1(c(=O)[nH]nc(c1)Nc1ccc(cc1)C#N)Oc1c(cccc1C)C	۵/۷۲	۵/۹۵
۶۴	c1(c(=O)[nH]nc(c1)Nc1ccc(cc1)C#N)Oc1c(cc(cc1C)C#N)C	۵/۵۵	۵/۴۵
۶۵	c1(c(=O)n(nc(c1)Nc1ccc(cc1)C#N)C)Oc1c(cc(cc1C)C)C	۵/۹۴	۵/۹۵
۶۶	c1(c(=O)n(nc(c1)N(c1ccc(cc1)C#N)C)C)Oc1c(cc(cc1C)C)C	۵/۴۷	۶/۰۲
۶۷	c1(cc(nc(c1)Nc1ccc(cc1)C#N)OC)Oc1c(cc(cc1C)C)C	۶/۰۸	۶/۱۳
۶۸	c1(cc(nc(c1)Nc1ccc(cc1)C#N)OC)Oc1c(cc(cc1C)C#N)C	۶/۱۵	۶/۰۰
۶۹[t]	c1(cc(=O)[nH]c(c1)Nc1ccc(cc1)[N](=O)O)Oc1c(cc(cc1C)C)C	۵/۸۵	۵/۷۵
۷۰	c1(cc(=O)[nH]c(c1)Nc1ccc(cc1)C)Oc1c(cc(cc1C)C)C	۵/۸۵	۶/۵۶
۷۱	c1(cc(=O)[nH]c(c1)Nc1ccc(cc1)C#N)Oc1c(cc(cc1C)C)C	۶/۴۳	۶/۰۸
۷۲	c1(cc(=O)[nH]c(c1)Nc1ccc(cc1)C#N)Oc1c(cc(cc1C)C#N)C	۶/۸۲	۵/۹۴
۷۳[v]	c1(cc(=O)[nH]c(c1)Nc1ccc(cc1)C#N)Oc1c(cccc1C)C	۵/۸۵	۶/۰۲

t و v به‌ترتیب نمایانگر داده‌های مجموعه آزمون و ارزیابی می‌باشد. سایر داده‌ها مربوط به مجموعه آموزش می‌باشد.

جدول ۲-۱۴ مجموعه داده‌های ضد سرطان کارسینوم کولورکتال به همراه مقادیر واقعی و پیش‌بینی شده pIC_{50}

ردیف	SMILES	pIC_{50} واقعی	pIC_{50} پیش‌بینی شده
۱	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1ccccc1	۵/۶۷	۵/۶۲
۲ t	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1cc(ccc1)C	۶/۳۶	۶/۰۳
۳	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1ccc(cc1)C	۶/۳۲	۵/۸۵
۴	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1ccc(cc1)OC	۶/۲۴	۵/۶۱
۵ t	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1c(cccc1)F	۵/۴۷	۵/۴۶
۶	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1cc(ccc1)F	۵/۶۳	۶/۰۰
۷ v	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1c(cccc1)Cl	۵/۸۷	۵/۵۶
۸	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1cc(ccc1)Cl	۵/۳۶	۵/۷۳
۹ t	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1cc(ccc1)Br	۵/۳۶	۴/۹۰
۱۰ t	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1ccc(cc1)Br	۵/۲۰	۵/۰۰
۱۱ t	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1c(cccc1F)F	۵/۶۳	۵/۳۵
۱۲	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1cc(c(cc1)Cl)Cl	۷/۰۹	۶/۷۵
۱۳	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1cc(c(cc1)F)Cl	۶/۷۲	۷/۱۳
۱۴	n1(nc2c(c1)c(=O)oc1c2cccc1)C	۴/۲۳	۴/۵۲
۱۵	n1(nc2c(c1)c(=O)oc1c2cccc1)CC	۴/۲۸	۴/۲۵
۱۶ t	n1(nc2c(c1)c(=O)oc1c2cccc1)CCC	۴/۲۱	۴/۲۳
۱۷	n1(nc2c(c1)c(=O)oc1c2cccc1)C(C)C	۴/۱۸	۴/۴۹
۱۸	n1(nc2c(c1)c(=O)oc1c2cccc1)CCCC	۴/۳۷	۴/۲۲
۱۹ t	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(C)C	۴/۲۰	۴/۴۱
۲۰ v	n1(nc2c(c1)c(=O)oc1c2cccc1)CCCCC	۴/۴۵	۴/۲۷
۲۱ v	n1(nc2c(c1)c(=O)oc1c2cccc1)C1CCCC1	۴/۵۴	۴/۶۵
۲۲	n1(nc2c(c1)c(=O)oc1c2cccc1)C1CCCCC1	۴/۶۵	۴/۵
۲۳	n1(nc2c(c1)c(=O)oc1c2cccc1)CC=C	۴/۳۰	۴/۴۸
۲۴	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=C)C	۴/۳۵	۴/۴۴
۲۵	n1(nc2c(c1)c(=O)oc1c2cccc1)CC=C(C)C	۴/۵۴	۴/۷۴
۲۶	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccccc1	۴/۲۰	۴/۴۹
۲۷	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1)C	۴/۱۷	۴/۳۴
۲۸	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1)OC	۴/۲۱	۴/۱۳
۲۹ t	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1)c1ccccc1	۵/۴۷	۵/۵۹
۳۰	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1)F	۴/۳۰	۴/۳۳
۳۱ v	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1)Br	۴/۲۸	۴/۱
۳۲ v	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccccc1Cl	۴/۷۴	۴/۹۸
۳۳	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1cccc(c1)Cl	۴/۵۹	۴/۲۳
۳۴	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1)Cl	۴/۷۱	۴/۶۴
۳۵	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1F)F	۵/۱۰	۵/۳۲
۳۶	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1Cl)Cl	۴/۴۹	۴/۶۷
۳۷	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(c(c1)Cl)Cl	۵/۳۴	۵/۵۵

ادامه جدول ۲-۱۴

ردیف	SMILES	pIC_{50} واقعی	pIC_{50} پیش‌بینی شده
۳۸[t]	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1cccc(c1)N(=O)=O	۴/۷۶	۵/۱۲
۳۹	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1)C(F)(F)F	۴/۷۴	۴/۹۳
۴۰[v]	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccccc1	۴/۶۶	۴/۵۶
۴۱	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1)C	۴/۷۴	۴/۶۴
۴۲	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1)OC	۴/۷۲	۵/۰۶
۴۳[v]	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1)c1ccccc1	۶/۶۷	۶/۷۷
۴۴	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1)F	۵/۷۶	۵/۶۲
۴۵[v]	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1)Br	۴/۷۹	۴/۹۲
۴۶	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccccc1Cl	۵/۲۹	۵/۵
۴۷	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1cccc(c1)Cl	۵/۳۹	۵/۱۴
۴۸[t]	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1)Cl	۵/۶۵	۵/۴۲
۴۹	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1F)F	۶/۴۴	۶/۲۹
۵۰	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1Cl)Cl	۵/۶۵	۵/۴۷
۵۱[v]	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(c(c1)Cl)Cl	۷/۰۰	۶/۸
۵۲	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1cccc(c1)N(=O)=O	۶/۲۰	۶/۳۱
۵۳	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1)C(F)(F)F	۶/۰۰	۶/۱۷
۵۴	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccccc1	۴/۳۷	۴/۴۹
۵۵	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccc(C)cc1	۵/۰۹	۵/۳۲
۵۶[t]	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccc(cc1)OC	۵/۲	۴/۸۶
۵۷	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccc(cc1)C(C)(C)C	۵/۱۷	۵/۲۷
۵۸	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccc(cc1)F	۴/۴۸	۴/۷۱
۵۹	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccc(cc1)Cl	۴/۵۶	۵/۰۵
۶۰[v]	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccc(cc1)Br	۴/۷۱	۴/۴۳
۶۱	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1c(cccc1)N(=O)=O	۴/۹۳	۵/۱۴
۶۲	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1cc(ccc1)N(=O)=O	۵/۳۸	۵/۸۴
۶۳	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccc(cc1)N(=O)=O	۵/۹۸	۵/۷۹
۶۴	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccc(cc1)C(F)(F)F	۵/۵۷	۴/۹۹
۶۵	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1c(ccc(c1)C)C	۶/۴۴	۶/۴۲
۶۶	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1cc(c(cc1)Cl)N(=O)=O	۶/۹۲	۶/۸۱
۶۷	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1c(C)cc(C)cc1C	۷/۰۹	۶/۷
۶۸	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1c(cc(cc1C(C)C)C(C)C)C(C)C	۷/۱۵	۷/۳
۶۹[v]	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)CC	۴/۰۸	۴/۱۵
۷۰	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)CCC	۴/۱۷	۴/۱۳
۷۱	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)C1CC1	۴/۰۱	۴/۴۷
۷۲	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)CCCCl	۴/۰۷	۴/۲۸

t و v به‌ترتیب نمایانگر داده‌های مجموعه آزمون و ارزیابی را می‌باشد. سایر داده‌ها مربوط به مجموعه آموزش می‌باشد.

جدول ۲-۱۵ مجموعه داده‌های ضد سرطان ریه به همراه مقادیر واقعی و پیش‌بینی شده pIC_{50}

ردیف	SMILES	pIC_{50} واقعی	pIC_{50} پیش‌بینی شده
۱[t]	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1ccccc1	۵/۱۴	۴/۸۷
۲	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1cc(ccc1)C	۶/۰۷	۶/۱۴
۳	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1ccc(cc1)C	۶/۱۱	۵/۷۲
۴	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1ccc(cc1)OC	۵/۷۵	۵/۰۴
۵	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1c(cccc1)F	۴/۹۹	۵/۴۹
۶	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1cc(ccc1)F	۵/۶۵	۶/۱۸
۷[t]	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1c(cccc1)Cl	۵/۴۸	۴/۸۵
۸[v]	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1cc(ccc1)Cl	۵/۰۳	۵/۱۲
۹	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1cc(ccc1)Br	۵/۲	۴/۹۴
۱۰	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1ccc(cc1)Br	۴/۹۱	۵/۱
۱۱[v]	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1c(cccc1F)F	۵/۳۷	۵/۵۴
۱۲	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1cc(c(cc1)Cl)Cl	۶/۵۵	۶/۲۴
۱۳	n1(nc2c(c1)c(=O)oc1c2cccc1)Cc1cc(c(cc1)F)Cl	۶/۵۱	۶/۴۶
۱۴	n1(nc2c(c1)c(=O)oc1c2cccc1)C	۴/۰۴	۴/۱۷
۱۵	n1(nc2c(c1)c(=O)oc1c2cccc1)CC	۴/۰۷	۴/۱۳
۱۶[v]	n1(nc2c(c1)c(=O)oc1c2cccc1)CCC	۴/۰۵	۴/۰۹
۱۷[v]	n1(nc2c(c1)c(=O)oc1c2cccc1)C(C)C	۴/۱۷	۴/۵۷
۱۸	n1(nc2c(c1)c(=O)oc1c2cccc1)CCCC	۴/۱۴	۳/۹۷
۱۹[t]	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(C)C	۳/۹۴	۴/۰۰
۲۰	n1(nc2c(c1)c(=O)oc1c2cccc1)CCCCC	۴/۲۳	۴/۰۱
۲۱	n1(nc2c(c1)c(=O)oc1c2cccc1)C1CCCC1	۴/۶۱	۴/۶۱
۲۲[t]	n1(nc2c(c1)c(=O)oc1c2cccc1)C1CCCCC1	۴/۷۱	۴/۳۲
۲۳	n1(nc2c(c1)c(=O)oc1c2cccc1)CC=C	۴/۱۱	۴/۳۴
۲۴[t]	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=C)C	۴/۴۶	۴/۳
۲۵	n1(nc2c(c1)c(=O)oc1c2cccc1)CC=C(C)C	۴/۵	۴/۹۵
۲۶[t]	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccccc1	۴/۰۱	۴/۳۵
۲۷	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1)C	۴/۱	۴/۱۱
۲۸[t]	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1)OC	۴/۱۶	۳/۷۵
۲۹	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1)c1ccccc1	۵/۱۸	۵/۲۴
۳۰	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1)F	۴/۰۷	۴/۰۳
۳۱	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1)Br	۴/۲۱	۴/۶۲
۳۲	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccccc1Cl	۴/۷۵	۴/۶۶
۳۳[v]	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1cccc(c1)Cl	۴/۴۹	۴/۵۸
۳۴	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1)Cl	۴/۳۴	۴/۶۷
۳۵	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1F)F	۵/۰۳	۴/۹۹
۳۶[v]	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1Cl)Cl	۴/۵۶	۴/۵۸
۳۷	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(c(c1)Cl)Cl	۴/۸۸	۴/۵۶
۳۸	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1cccc(c1)N(=O)=O	۴/۴۵	۴/۲۲
۳۹	n1(nc2c(c1)c(=O)oc1c2cccc1)CC(=O)c1ccc(cc1)C(F)(F)F	۴/۵	۴/۹۱
۴۰	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccccc1	۴/۴۷	۴/۵۸
۴۱	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1)C	۴/۵۵	۴/۷۸
۴۲[t]	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1)OC	۴/۴۱	۴/۸۷
۴۳[v]	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1)c1ccccc1	۶/۲۱	۶/۱۵
۴۴	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1)F	۵/۶۱	۵/۶۴
۴۵[v]	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1)Br	۴/۴۳	۴/۷۷
۴۶	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccccc1Cl	۵/۰۷	۴/۶۱
۴۷	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1cccc(c1)Cl	۴/۹۸	۵/۳۷
۴۸	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1)Cl	۵/۲	۴/۶۶
۴۹	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1F)F	۶/۱۱	۵/۶۳

ادامه جدول ۲-۱۵

ردیف	SMILES	pIC_{50} واقعی	pIC_{50} پیش‌بینی شده
۵۰	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1Cl)Cl	۵/۲	۴/۸
۵۱	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(c(c1)Cl)Cl	۶/۵۱	۵/۹۳
۵۲	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1cccc(c1)N(=O)=O	۵/۶۵	۶/۰۱
۵۳[v]	n1(nc2c(c1)c(=O)oc1c2cccc1)C/C(=N/O)/c1ccc(cc1)C(F)(F)F	۵/۴۹	۶/۰۹
۵۴	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccccc1	۴/۲۲	۴/۶
۵۵	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccc(C)cc1	۵/۰۱	۴/۶۶
۵۶	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccc(cc1)OC	۵/۰۳	۴/۵۸
۵۷	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccc(cc1)C(C)(C)C	۵/۱۱	۵/۶
۵۸	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccc(cc1)F	۴/۳	۴/۷۴
۵۹	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccc(cc1)Cl	۴/۴۹	۴/۸۵
۶۰[v]	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccc(cc1)Br	۴/۶۶	۴/۵۸
۶۱	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1c(cccc1)N(=O)=O	۴/۸۶	۴/۶۸
۶۲[t]	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1cc(ccc1)N(=O)=O	۵/۳۵	۵/۲
۶۳	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccc(cc1)N(=O)=O	۶/۰۱	۵/۹
۶۴	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1ccc(cc1)C(F)(F)F	۵/۷۲	۵/۴۴
۶۵	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1c(ccc(c1)C)C	۶/۴۶	۶/۰۲
۶۶[t]	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1cc(c(cc1)Cl)N(=O)=O	۶/۳۲	۵/۹۳
۶۷[t]	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1c(C)cc(C)cc1C	۶/۷۷	۶/۳
۶۸	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)c1c(cc(cc1C(C)C)C(C)C)C(C)C	۶/۹۲	۶/۱۶
۶۹	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)CC	۴/۰۵	۴/۵۲
۷۰	n1(nc2c(c1)c(=O)oc1c2cccc1)S(=O)(=O)CCC	۴/۱	۴/۵۴

t و v به‌ترتیب نمایانگر داده‌های مجموعه آزمون و ارزیابی می‌باشد. سایر داده‌ها مربوط به مجموعه آموزش می‌باشد.

۲- ۴- ۳ رسم و بهینه‌سازی ساختار ترکیبات شیمیایی مجموعه داده‌های متفاوت

برای محاسبه توصیف‌کننده‌های مولکولی با مقادیر صحیح، ساختار ترکیبات مورد مطالعه با استفاده از نرم‌افزار هایپرکم و مطابق با روش کار بخش ۱- ۵- ۳ بهینه شدند. ساختارهای بهینه شده با پسوند hin.* ذخیره شدند و به‌عنوان ورودی نرم‌افزار محاسباتی دراگون استفاده شدند.

۲- ۴- ۴ استخراج توصیف‌کننده‌های ساختاری

ساختارهای بهینه شده ترکیبات مورد مطالعه برای هر سه مجموعه داده‌ها به‌طور مجزا در نرم افزار دراگون فراخوانی شدند و سپس برای هر ترکیب تعداد ۳۲۲۴ توصیف کننده در ۲۲ دسته متفاوت محاسبه شدند.

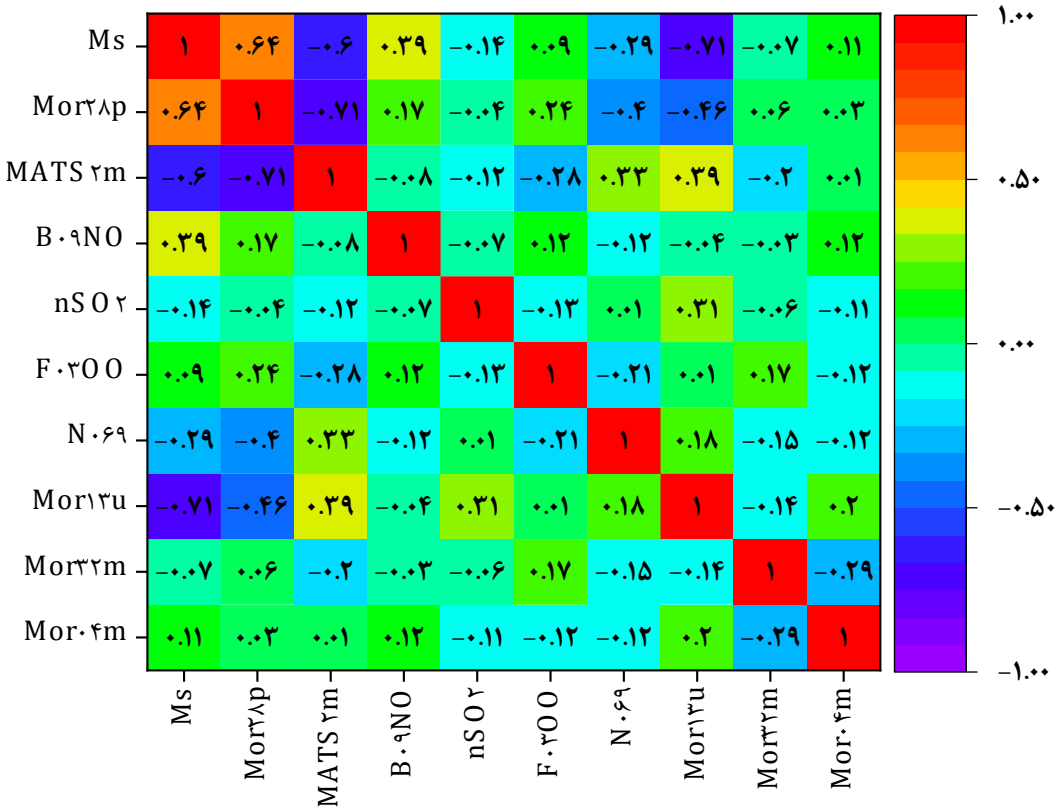
۲- ۴- ۵ پیش‌پردازش و انتخاب توصیف‌کننده‌های مؤثر

با توجه به این‌که وجود توصیف کننده‌هایی با مقادیر ثابت و نسبتاً ثابت اطلاعات مفیدی را به مدل نمی‌افزاید، بنابراین این نوع از توصیف کننده‌ها (توصیف کننده‌هایی با واریانس کمتر از ۰/۰۰۱) با استفاده از بسته نرم‌افزاری caret در نرم‌افزار R حذف شدند. علاوه بر این توصیف‌کننده‌هایی با همبستگی بالای ۰/۹ نیز در نرم‌افزار متلب مورد بررسی قرار گرفتند و از بین دو توصیف کننده با همبستگی بالای ۰/۹، توصیف کننده‌ای با بیش‌ترین همبستگی با پاسخ، نگه داشته شد و توصیف کننده بعدی حذف شد. پس از انجام مراحل پیش‌پردازش داده‌ها، تعداد توصیف کننده‌ها برای مجموعه داده‌های ۷۳، ۷۲ و ۷۰ تایی به ترتیب از ۳۲۲۴ توصیف کننده به ۳۳۵، ۴۳۸ و ۴۲۹ توصیف کننده کاهش یافت. این توصیف کننده‌ها به‌طور مستقل به‌عنوان ورودی روش انتخاب متغیر LAD-LASSO مورد استفاده قرار گرفتند. به‌طوری‌که توصیف‌کننده‌های حاصل از مرحله پیش‌پردازش به‌عنوان متغیرهای مستقل و فعالیت‌های دارویی مربوطه به‌عنوان متغیر وابسته در روش انتخاب متغیر LAD-LASSO به‌کار گرفته شدند. به‌منظور اجرای روش انتخاب متغیر LAD-LASSO، مجموعه داده‌ها با استفاده از الگوریتم KS به سه بخش آموزش، ارزیابی و آزمون تقسیم شدند (شماره ترکیبات مربوط به هر مجموعه در جدول مجموعه داده‌ها مشخص شده است). در مرحله انتخاب متغیر، داده‌های مجموعه آزمون از ابتدا بیرون گذاشته شدند و از مجموعه داده‌های مجموعه آموزش و ارزیابی برای انتخاب توصیف‌کننده‌های مؤثر، استفاده شد. روش انتخاب متغیر انقباضی با روش ارزیابی تقاطعی ده فولد موجود در بسته quantreg در نرم افزار R روی مجموعه داده‌های ارزیابی و آموزش اجرا شد تا موثرترین توصیف کننده‌ها انتخاب شود. با اجرای روش LAD-LASSO در نهایت توصیف‌کننده‌هایی با ضرایب غیر صفر (۱۰، ۱۴ و ۹ توصیف کننده به‌ترتیب برای مجموعه داده‌های ۷۳، ۷۲ و ۷۰ تایی) در λ_{min} مربوطه استخراج شدند. نام و نوع این توصیف کننده‌ها به همراه ضرایب رگرسیونی استاندارد شده مربوط به هر توصیف کننده در جدول ۲-۱۶ خلاصه شده‌اند. به‌منظور بررسی آماری دقیق‌تر

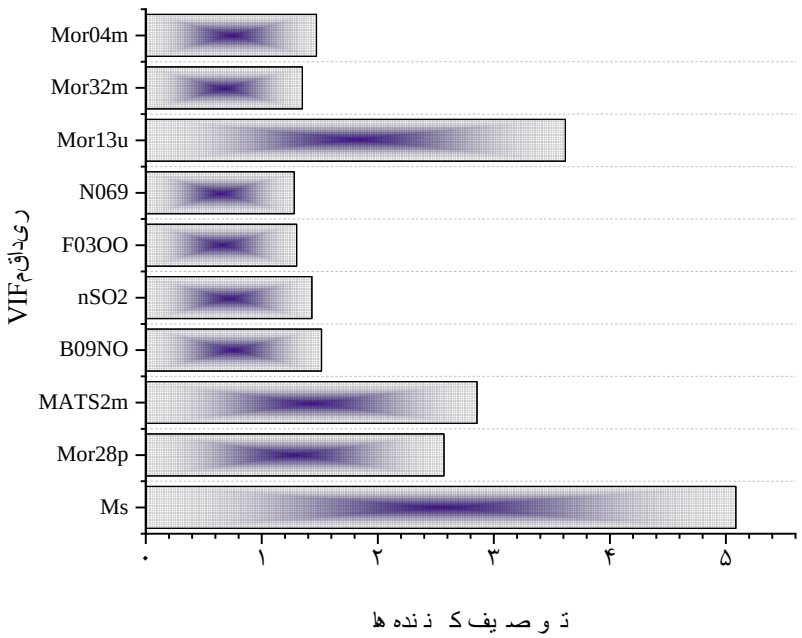
توصیف‌کننده‌های منتخب روش LAD-LASSO هر سه مجموعه داده‌ها، احتمال وجود همبستگی و هم‌خطی بین توصیف کننده‌ها به‌ترتیب با محاسبه مقادیر ضرایب همبستگی بین دو توصیف کننده و مقادیر افزایش تورم واریانس (VIF) توصیف کننده (مطابق با رابطه ۱-۱۰) مطالعه شد. به‌این منظور نمودارهای نقشه رنگی و VIF برای توصیف کننده‌های منتخب روش LAD-LASSO برای هر سه مجموعه داده رسم شد و نتایج حاصل در شکل ۲-۱۶ تا شکل ۲-۲۱ نمایش داده شده‌اند. نتایج نمودار نقشه رنگی (شکل ۲-۱۶ و شکل ۲-۱۸ و شکل ۲-۲۰) نشان می‌دهد که بین توصیف‌کننده‌های منتخب هر سه مجموعه داده‌ها به روش LAD-LASSO همبستگی معناداری وجود ندارد. علاوه بر این، نمودارهای VIF هر سه مجموعه داده‌ها (شکل ۲-۱۷ و شکل ۲-۱۹ و شکل ۲-۲۱) نیز نشان می‌دهد که مقادیر VIF محاسبه شده برای متغیرهای منتخب هر سه مجموعه داده‌ها، در محدوده ۱-۱۰ قرار دارند. این شواهد بیانگر این است که بین توصیف کننده‌های منتخب روش LAD-LASSO هم خطی شدیدی وجود ندارد.

جدول ۲-۱۶ توصیف‌کننده‌های منتخب LAD-LASSO برای هر سه مجموعه داده‌ها

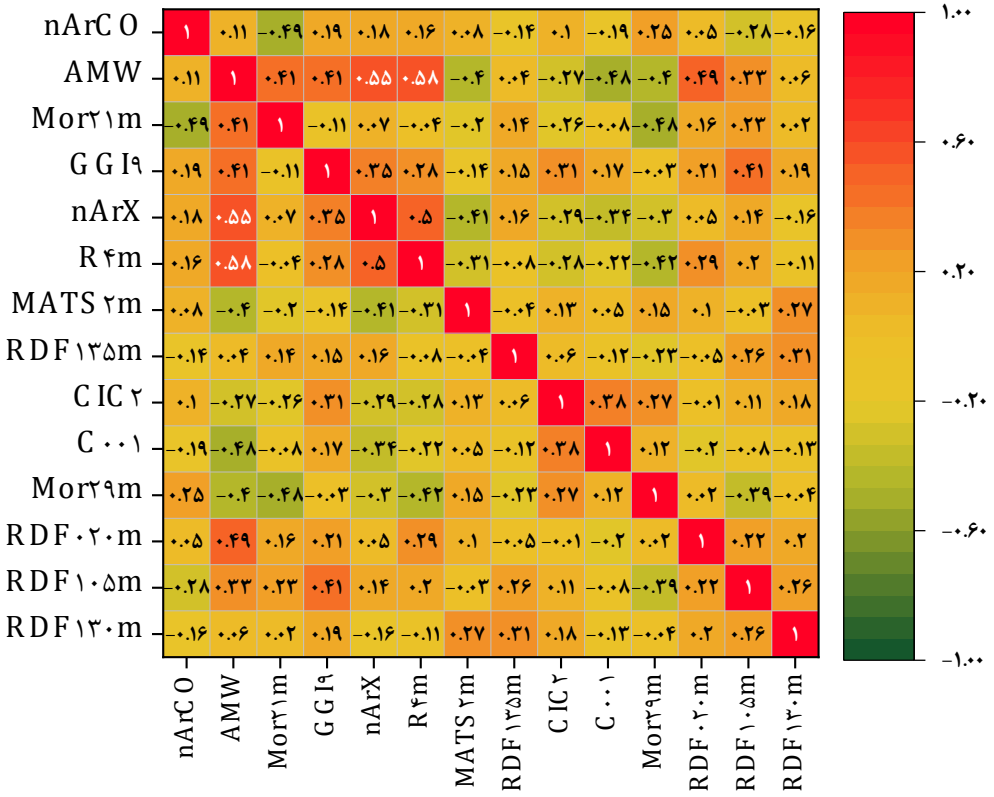
مجموعه داده‌ها	ردیف	نماد	معنا	طبقه‌بندی	ضرایب استاندارد شده
بازدارنده‌های ایدز	۱	Ms	Electro-topological State	Constitutional indices	۰/۶۵
	۲	Mor28p	signal 28 / weighted by polarizability	3D-MoRSE descriptors	-۰/۵۱
	۳	MATS2m	Moran autocorrelation of lag 2 weighted by mass	2D autocorrelations	۰/۳۸
	۴	B09NO	Presence/absence of N - O at topological distance 9	2D binary fingerprints	-۰/۳۷
	۵	nSO2	number of sulfites (thio-/dithio-)	Functional group counts	۰/۲۷
	۶	F03OO	Frequency of C - O at topological distance 3	2D frequency fingerprints	-۰/۲۴
	۷	N069	Ar-NH2 / X-NH2	Atom-centered fragments	۰/۲۲
	۸	Mor13u	signal 13 / unweighted	3D-MoRSE descriptors	۰/۱۹
	۹	Mor32m	signal 32 / weighted by mass	3D-MoRSE descriptors	۰/۱۷
	۱۰	Mor04m	signal 04 / weighted by mass	3D-MoRSE descriptors	-۰/۱۴
بازدارنده‌های سرطان کارسینوم کولورکتال	۱	nArCO	number of ketones (aromatic)	Functional group counts	-۰/۶۵
	۲	AMW	average molecular weight	Constitutional indices	۰/۵۲
	۳	Mor21m	signal 21 / weighted by mass	3D-MoRSE descriptors	-۰/۴۲
	۴	GGI9	topological charge index of order 9	2D autocorrelations	۰/۳۹
	۵	nArX	number of X on aromatic ring	Functional group counts	۰/۳۵
	۶	R4m	R autocorrelation of lag 4 / weighted by mass	GETAWAY descriptors	-۰/۲۷
	۷	MATS2m	Moran autocorrelation of lag 2 weighted by mass	2D autocorrelations	۰/۱۷
	۸	RDF135m	Radial Distribution Function - 135 / weighted by mass	RDF descriptors	۰/۱۶
	۹	CIC2	Complementary Information Content index (neighborhood symmetry of 2-order)	Information indices	۰/۱۴
	۱۰	C001	CH3R / CH4	Atom-centered fragments	۰/۱۴
	۱۱	Mor29m	signal 29 / weighted by mass	3D-MoRSE descriptors	۰/۱
	۱۲	RDF020m	Radial Distribution Function - 020 / weighted by mass	RDF descriptors	-۰/۰۹
	۱۳	RDF105m	Radial Distribution Function - 105 / weighted by mass	RDF descriptors	-۰/۰۸
	۱۴	RDF130m	Radial Distribution Function - 130 / weighted by mass	RDF descriptors	۰/۰۶
بازدارنده‌های سرطان ریه	۱	AMW	average molecular weight	Constitutional indices	۰/۶۱
	۲	RDF110m	Radial Distribution Function - 110 / weighted by mass	RDF descriptors	۰/۵۶
	۳	Mor12e	signal 12 / weighted by Sanderson electronegativity	3D-MoRSE descriptors	-۰/۵۴
	۴	nCp	number of terminal primary C(sp3)	Functional group counts	۰/۴۴
	۵	Mor29m	signal 29 / weighted by mass	3D-MoRSE descriptors	۰/۳۳
	۶	F084	F attached to C1(sp2)	Atom-centered fragments	۰/۳۱
	۷	RDF130m	Radial Distribution Function - 130 / weighted by mass	RDF descriptors	۰/۱۵
	۸	nArOR	number of ethers (aromatic)	Functional group counts	-۰/۱۳
	۹	RDF070m	Radial Distribution Function - 070 / weighted by mass	RDF descriptors	-۰/۰۸



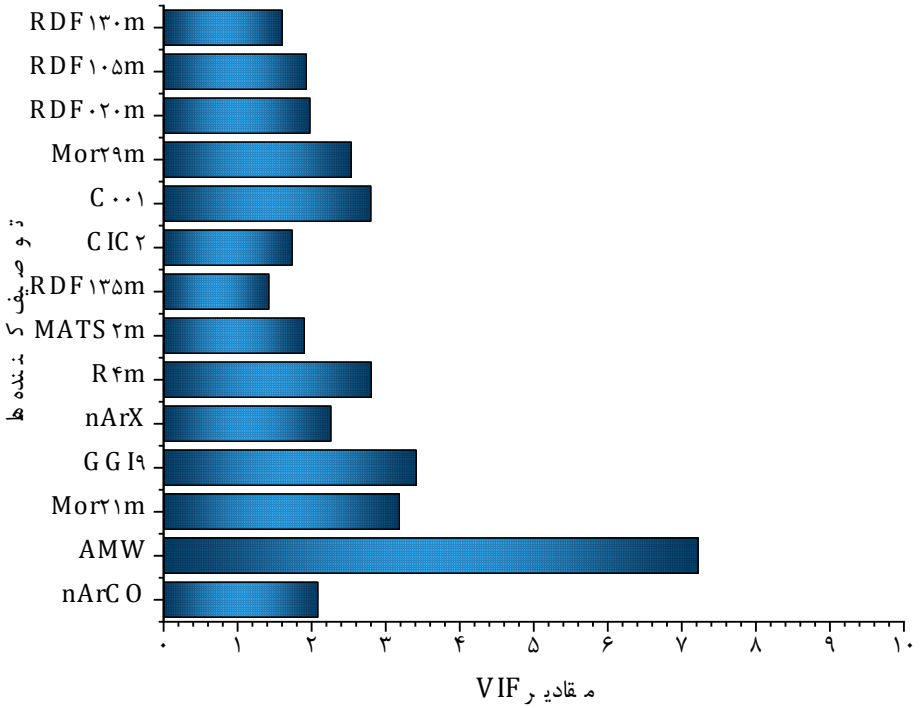
شکل ۲-۱۶ نمودار نقشه رنگی جهت نمایش همبستگی بین توصیف‌کننده‌های منتخب LAD-LASSO برای بازدارنده‌های ایدز



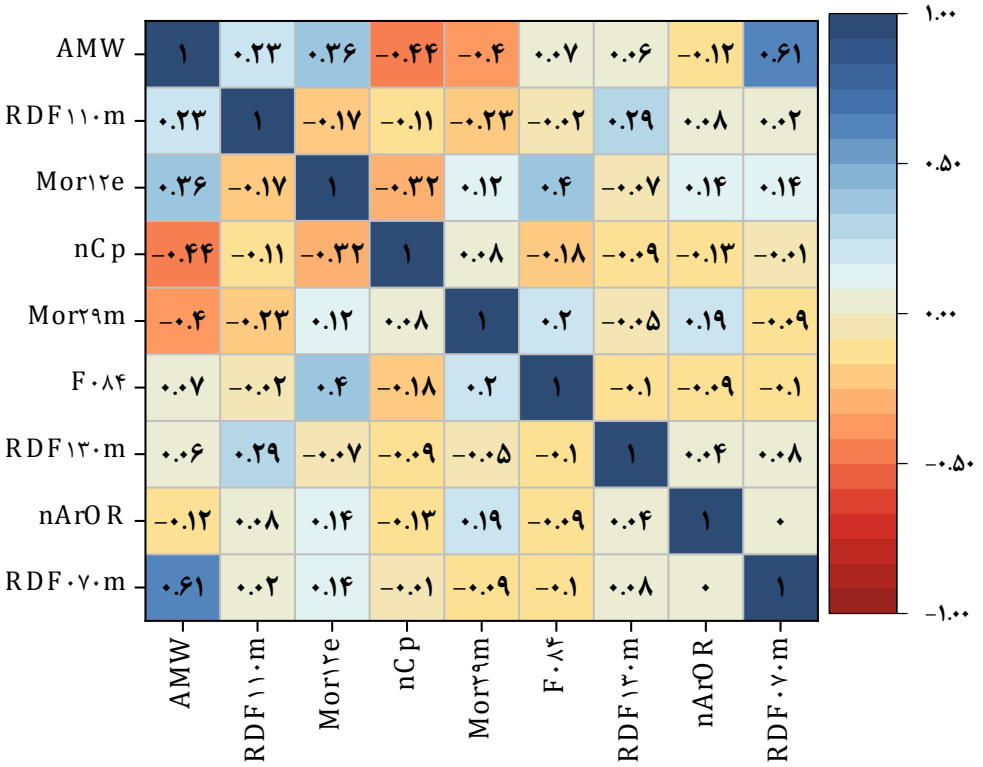
شکل ۲-۱۷ نمودار مقادیر VIF توصیف‌کننده‌های منتخب LAD-LASSO برای بازدارنده‌های ایدز



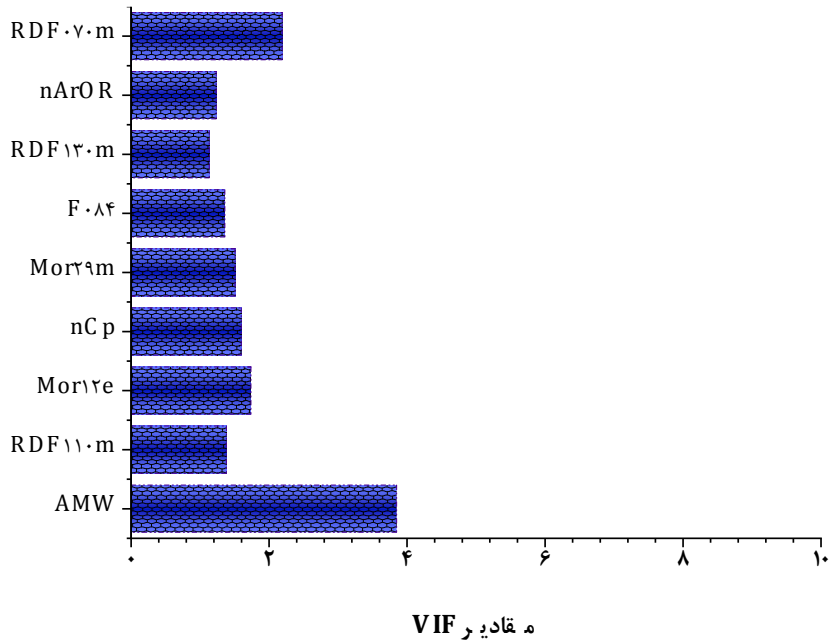
شکل ۲-۱۸ نمودار نقشه رنگی جهت نمایش همبستگی بین توصیف‌کننده‌های منتخب LAD-LASSO برای بازدارنده‌های سرطان کارسینوم کولورکتال



شکل ۲-۱۹ نمودار مقادیر VIF توصیف‌کننده‌های منتخب LAD-LASSO برای بازدارنده‌های سرطان کارسینوم کولورکتال



شکل ۲-۲۰ نمودار نقشه رنگی جهت نمایش همبستگی بین توصیف‌کننده‌های منتخب LAD-LASSO برای بازدارنده‌های سرطان ریه



شکل ۲-۲۱ نمودار مقادیر VIF توصیف‌کننده‌های منتخب LAD-LASSO برای بازدارنده‌های سرطان ریه

۲- ۴- ۶ مدل‌سازی شبکه عصبی با استفاده از توصیف‌کننده‌های منتخب LAD-LASSO

برای برقراری ارتباط بین توصیف‌کننده‌های منتخب و متغیر وابسته مربوطه از مدل شبکه عصبی مصنوعی سه لایه‌ای با الگوریتم پس انتشار خطا استفاده شد. به این منظور، یک شبکه پرسپترون، با یک لایه ورودی، یک لایه پنهان و یک لایه خروجی طراحی شد [۱۴۰, ۱۴۱]. برای انتخاب بهترین ساختار شبکه عصبی، بهینه‌سازی تعداد نورون‌های لایه ورودی، تعداد نورون‌های لایه پنهان، تعداد دور آموزش، توابع انتقال (تانژانت هایپربولیک سیگموئیدی و لگاریتم سیگموئیدی که به ترتیب با توابع tansig و logsig در جعبه ابزار متلب مشخص می‌شوند) و توابع آموزش (تنظیم بایزین و لونبرگ- مارکوارت که به ترتیب با توابع trainbr و trainlm در جعبه ابزار متلب مشخص می‌شوند) به‌طور هم‌زمان انجام شد. علاوه بر این از تابع خطی purlin نیز به‌عنوان لایه خروجی استفاده شد. بنابراین، توصیف‌کننده‌های منتخب بر اساس بزرگی ضرایب استاندارد شده LAD-LASSO (جدول ۲-۱۶) چیده شدند و به‌عنوان ورودی در ساخت مدل شبکه عصبی مورد استفاده قرار گرفتند. معیار بهینه‌سازی حداقل نمودن پارامتر MSE مجموعه ارزیابی است. از این‌رو برای بهینه‌سازی پارامترهای شبکه عصبی مصنوعی، تعداد توصیف کننده‌ها، گره‌ها و دورهای آموزشی به‌طور هم‌زمان از ۲ تا تعداد کل توصیف‌کننده‌های منتخب (با گام ۱)، ۲ تا ۱۰ (با گام ۱) و ۵ تا ۵۰ (با گام ۵) تغییر یافت. ساختارهای شبکه عصبی با ورودی‌های متفاوت (۳۲۴۰ حالت، ۴۶۸۰ حالت و ۲۸۸۸ حالت برای مجموعه داده‌های متشکل از ۷۳، ۷۲ و ۷۰ ترکیب دارویی) برای هر سه مجموعه داده‌ها طراحی شد. آموزش شبکه عصبی با استفاده از داده‌های مجموعه آموزش و با توابع آموزش تنظیم بایزین و لونبرگ- مارکوارت انجام شد. پس از بهینه‌سازی پارامترهای شبکه عصبی، ساختار بهینه برای هر مجموعه مورد مطالعه، با توجه به حداقل مقدار MSE مجموعه ارزیابی انتخاب شد. شرایط بهینه ساختار شبکه عصبی بهینه برای هر سه مجموعه داده، در جدول ۲-۱۷ خلاصه شده است. مقادیر فعالیت‌های دارویی داده‌های

مجموعه آزمون با استفاده از ساختار شبکه عصبی بهینه مربوط به هر مجموعه داده پیش‌بینی شد. بنابراین مدل شبکه عصبی با نماد LAD-LASSO-LM-ANN (با معماری‌های متفاوت مندرج در جدول ۲-۱۷ برای هر مجموعه داده) به‌عنوان مدل برتر برای پیش بینی مقادیر فعالیت دارویی ترکیبات مورد مطالعه در هر سه مجموعه داده انتخاب شد.

جدول ۲-۱۷ ساختارهای شبکه‌های توسعه یافته با توصیف‌کننده‌های منتخب LAD_LASSO با کمترین MSE مجموعه ارزیابی

مجموعه داده‌ها	تعداد توصیف کننده	تابع آموزش	تابع انتقال	تعداد گره	تعداد دور آموزش	$MSE_{validation}$	$R^2_{validation}$
بازدارنده‌های ایدز	۸	تنظیم بایزین	لگاریتم-سیگموئید	۶	۵	۰/۱۶	۰/۸۳
	۵	**لونبرگ-مارکوارت**	**لگاریتم-سیگموئید**	**۶**	**۵**	**۰/۱۲**	**۰/۹۱**
	۸	تنظیم بایزین	تانژانت-سیگموئید	۲	۵	۰/۱۶	۰/۸۳
	۵	لونبرگ- مارکوارت	تانژانت-سیگموئید	۷	۲۰	۰/۱۴	۰/۸۶
بازدارنده‌های سرطان کارسینوم کولورکتال	۱۴	تنظیم بایزین	لگاریتم-سیگموئید	۸	۵۰	۰/۱۹	۰/۸۱
	۱۴	**لونبرگ-مارکوارت**	**لگاریتم-سیگموئید**	**۲**	**۵**	**۰/۰۹**	**۰/۹۰**
	۱۰	تنظیم بایزین	تانژانت-سیگموئید	۲	۵۰	۰/۱۹	۰/۸۲
	۱۳	لونبرگ- مارکوارت	تانژانت-سیگموئید	۲	۵	۰/۱۰	۰/۹۱
بازدارنده‌های سرطان ریه	۹	تنظیم بایزین	لگاریتم-سیگموئید	۲	۱۰	۰/۱۰	۰/۸۵
	۷	لونبرگ- مارکوارت	لگاریتم-سیگموئید	۳	۱۰	۰/۰۹	۰/۸۵
	۷	تنظیم بایزین	تانژانت-سیگموئید	۴	۴۰	۰/۰۹	۰/۸۶
	۷	**لونبرگ-مارکوارت**	**تانژانت-سیگموئید**	**۴**	**۴۰**	**۰/۰۸**	**۰/۸۶**

به‌منظور تائید کارایی روش پیشنهادی LAD-LASSO-LM-ANN در این مطالعه، روش انتخاب متغیر کلاسیک SR نیز بر روی داده‌های مجموعه آموزش و ارزیابی هر سه مجموعه داده‌ها اجرا شد. تعداد توصیف‌کننده‌های منتخب SR برای هر سه مجموعه داده‌ها به‌ترتیب برابر با تعداد ۱۲، ۱۴ و ۱۱ به‌دست آمد. از این توصیف کننده‌ها برای طراحی و بهینه‌سازی مدل‌های شبکه عصبی استفاده شد و بهینه‌سازی پارامترهای ANN به‌طور هم‌زمان انجام شد. نتایج نشان داد مدل ANN با توابع آموزش LM برای هر سه مجموعه داده و با استفاده از ۱۲ توصیف کننده، ۳ گره در لایه پنهان و ۵ دور آموزشی برای مجموعه داده‌های ضد ایدز، ۱۴ توصیف کننده، ۲ گره در لایه پنهان و ۱۰ دور آموزشی برای مجموعه داده‌های ضد سرطان کارسینوم کولورکتال و ۱۱ توصیف کننده، ۴ گره در لایه پنهان و ۲۵ دور آموزشی برای مجموعه داده‌های ضد سرطان ریه، کمترین MSE را برای مجموعه ارزیابی ایجاد نمود. مدل‌های بهینه SR-LM-ANN برای پیش بینی فعالیت دارویی مجموعه آزمون استفاده شد.

۲- ۴- ۷ ارزیابی مدل LAD-LASSO-LM-ANN هر سه مجموعه داده

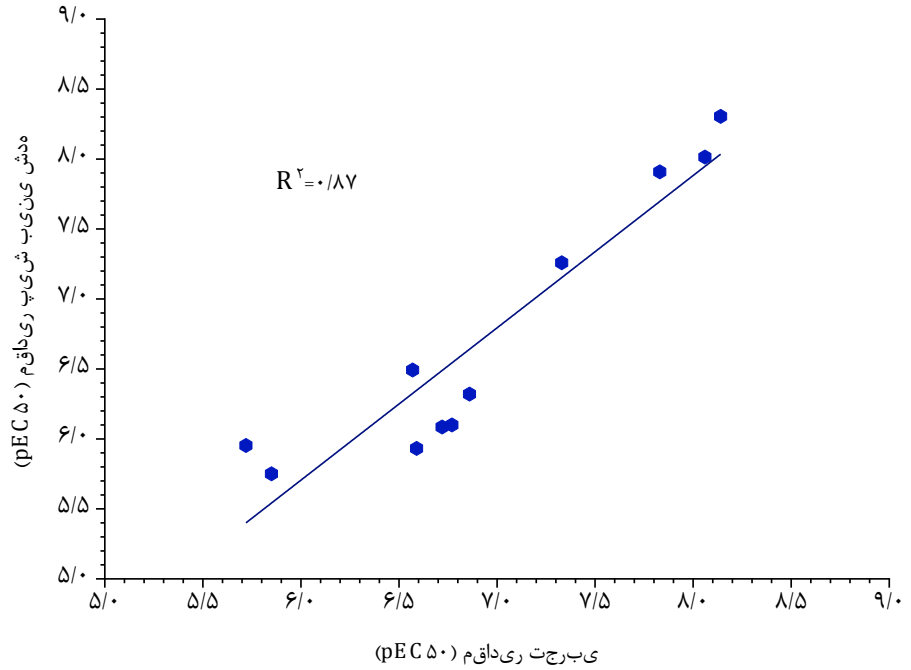
هدف اصلی از ساخت مدل QSAR، ایجاد یک مدل با قدرت پیش‌بینی مناسب، قابل اعتماد و دقیق برای پیش‌بینی فعالیت دارویی ترکیبات جدید است. در این تحقیق، قدرت پیش بینی، اعتبار و تعمیم پذیری مدل توسعه یافته LM-ANN با استفاده از توصیف کننده‌های منتخب روش انتخاب متغیر LAD-LASSO با به‌کارگیری پیش بینی داده‌های مجموعه آزمون و پیش‌بینی پاسخ کل ترکیبات با استفاده از تکنیک LOO، آزمون‌های پرکاربردی هم‌چون Y-تصادفی، دامنه کاربرد و محاسبه پارامترهای آماری مورد ارزیابی قرار گرفت.

۲- ۴- ۷- ۱ ارزیابی مدل LAD-LASSO-LM-ANN با استفاده از پیش بینی داده‌های مجموعه آزمون

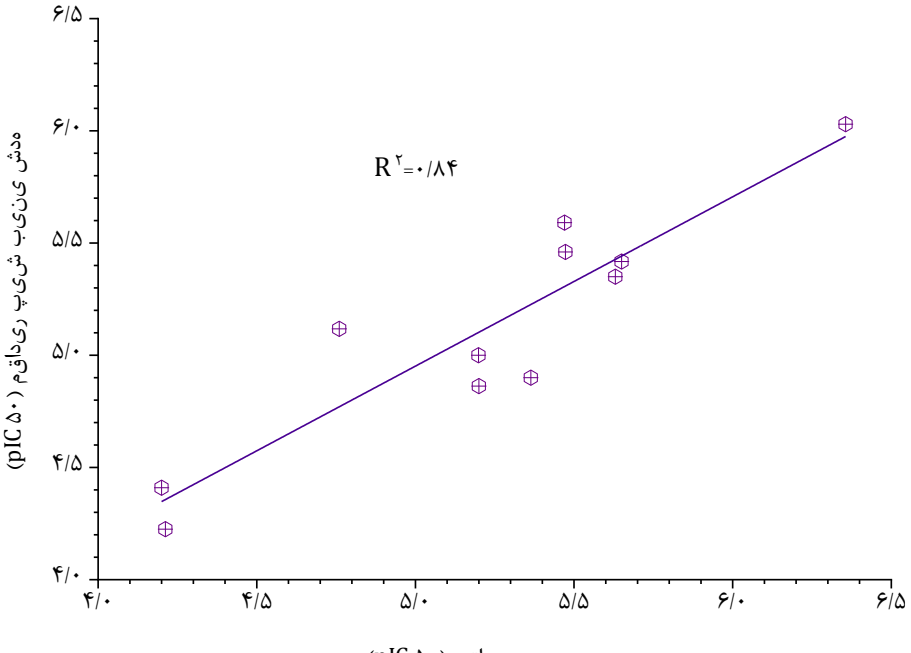
به‌منظور بررسی اعتبار و قدرت پیش‌بینی مدل بهینه LAD-LASSO-LM-ANN توسعه یافته برای هر مجموعه داده، فعالیت دارویی داده‌های مجموعه آزمون با استفاده از مدل آموزش دیده در شرایط بهینه پیش‌بینی شد. لازم به ذکر است که مجموعه داده‌های آزمون در هیچ یک از مراحل انتخاب توصیف کننده‌های مؤثر و مدل‌سازی حضور نداشته‌اند. مقادیر فعالیت پیش‌بینی شده به‌وسیله مدل LAD-LASSO-LM-ANN برای ترکیبات موجود در مجموعه آزمون هر سه مجموعه داده در جدول ۲-۱۸ آورده شده است. نتایج حاصل حاکی از خطای کم و قابل قبول اغلب ترکیبات مجموعه آزمون است. برای بررسی بیش‌تر قدرت پیش بینی مدل‌های توسعه یافته LAD-LASSO-LM-ANN مقادیر فعالیت پیش‌بینی شده به‌وسیله مدل LAD-LASSO-LM-ANN بر حسب فعالیت دارویی تجربی ترکیبات موجود در مجموعه آزمون رسم شد. مقادیر R^2 بزرگ‌تر از ۰/۶ نشان‌دهنده قدرت پیش‌بینی قابل قبول مدل LAD-LASSO-LM-ANN در هر سه مجموعه داده مورد مطالعه است.

جدول ۲-۱۸ نتایج حاصل از ارزیابی مدل LAD-LASSO-LM-ANN با استفاده از مجموعه آزمون

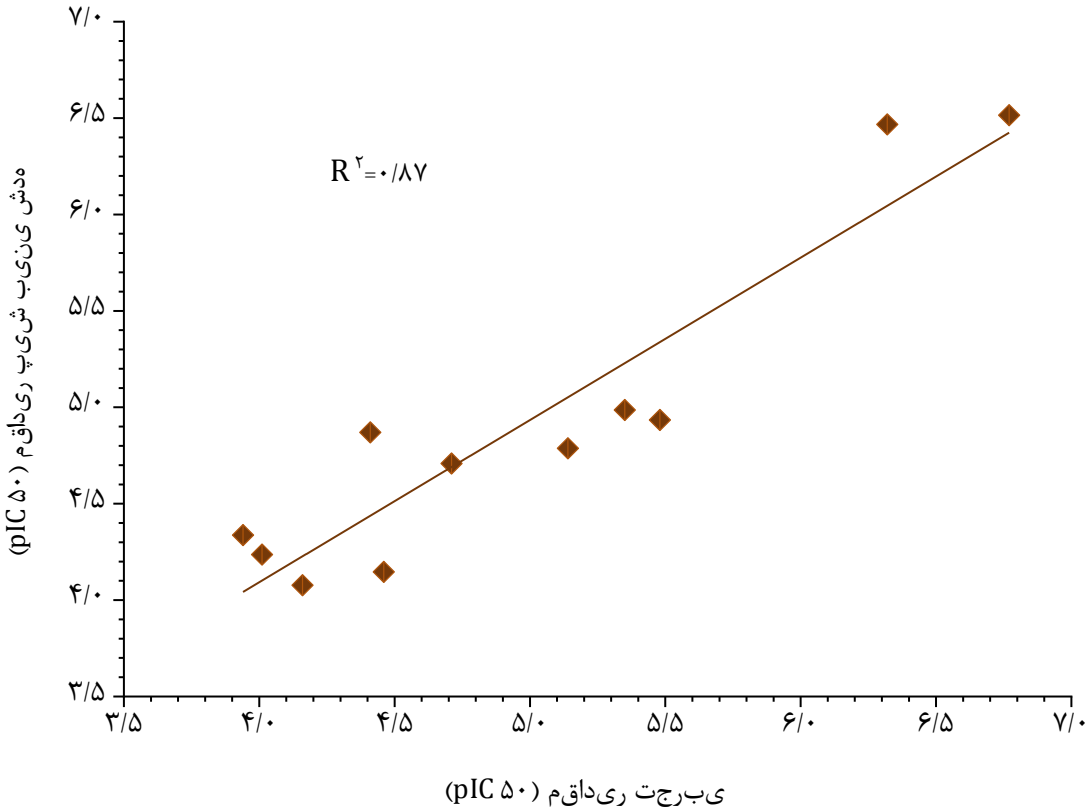
مجموعه داده‌ها	شماره ترکیب	فعالیت دارویی ($pIC_{۵۰}/pEC_{۵۰}$)		درصد خطا
		مقدار واقعی	مقدار پیش‌بینی شده	
بازدارنده‌های ایدز	۳	۶/۷۲	۶/۰۹	-۹/۴۴
	۸	۷/۳۳	۷/۲۶	-۰/۹۷
	۱۸	۸/۱۴	۸/۳	۲/۰۳
	۲۴	۷/۸۳	۷/۹۱	۱/۰۰
	۲۵	۸/۰۶	۸/۰۱	-۰/۵۷
	۳۱	۶/۵۷	۶/۴۹	-۱/۱۸
	۳۵	۶/۷۷	۶/۱	-۹/۹
	۴۴	۶/۸۶	۶/۳۲	-۷/۸۷
	۵۸	۶/۵۹	۵/۹۳	-۹/۹۸
	۶۳	۵/۷۲	۵/۹۵	۴/۰۷
	۶۹	۵/۸۵	۵/۷۵	-۱/۶۹
بازدارنده‌های سرطان کارسینوم کولورکتال	۲	۶/۳۶	۶/۰۳	-۵/۱۴
	۵	۵/۴۷	۵/۴۶	-۰/۲۴
	۹	۵/۳۶	۴/۹	-۸/۶۵
	۱۰	۵/۲	۵	-۳/۸۳
	۱۱	۵/۶۳	۵/۳۵	-۴/۹۷
	۱۶	۴/۲۱	۴/۲۳	۰/۳۱
	۱۹	۴/۲	۴/۴۱	۴/۹۹
	۲۹	۵/۴۷	۵/۵۹	۲/۲۱
	۳۸	۴/۷۶	۵/۱۲	۷/۵۱
	۴۸	۵/۶۵	۵/۴۲	-۴/۱۱
	۵۶	۵/۲	۴/۸۶	-۶/۴۹
بازدارنده‌های سرطان ریه	۱	۵/۱۴	۴/۷۹	-۶/۸۷
	۷	۵/۴۸	۴/۹۳	-۹/۹۶
	۱۹	۳/۹۴	۴/۳۴	۱۰/۰۹
	۲۲	۴/۷۱	۴/۷۱	-۰/۰۴
	۲۴	۴/۴۶	۴/۱۵	-۷/۰۴
	۲۶	۴/۰۱	۴/۲۴	۵/۶۵
	۲۸	۴/۱۶	۴/۰۸	-۱/۹۸
	۴۲	۴/۴۱	۴/۸۷	۱۰/۴۳
	۶۲	۵/۳۵	۴/۹۹	-۶/۸
	۶۶	۶/۳۲	۶/۴۷	۲/۳۳
	۶۷	۶/۷۷	۶/۵۲	-۳/۷۶



شکل ۲-۲۲ نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی برای داده‌های ضد ایدز مجموعه آزمون



شکل ۲-۲۳ نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی برای داده‌های ضد سرطان کارسینوم کولورکتال مجموعه آزمون



شکل ۲-۲۴ نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی برای داده‌های ضد سرطان ریه مجموعه آزمون

همانطور که در بخش ۲- ۴- ۷ اشاره شد، از روش SR-LM-ANN برای بررسی عملکرد مدل LAD-LASSO-LM-ANN استفاده شد. پاسخ ترکیبات مجموعه آزمون با استفاده از مدل‌های شبکه عصبی بهینه با معماری‌های اشاره شده در بخش ۲- ۴- ۷ پیش بینی شد. مقادیر MSE مجموعه داده آزمون به‌ترتیب برابر با ۰/۳۲، ۰/۲۶ و ۰/۴۶ برای مجموعه داده‌های ایدز، سرطان کارسینوم کولورکتال و سرطان ریه به‌دست آمد. مقادیر R^2 مربوط به مجموعه آزمون نیز به ترتیب برابر با ۰/۷۶، ۰/۷۴ و ۰/۷۶ برای مجموعه داده‌های ایدز، سرطان کارسینوم کولورکتال و سرطان ریه به‌دست آمد. نتایج نشان می‌دهد که مدل LAD-LASSO-LM-ANN در پیش‌بینی فعالیت‌های دارویی ترکیبات مورد نظر با قدرت بیش‌تری نسبت به مدل SR-LM-ANN عمل کرده است.

۲- ۴- ۷- ۲ ارزیابی مدل LAD-LASSO-LM-ANN با پیش بینی فعالیت دارویی تمام ترکیبات مجموعه داده‌ها با استفاده از روش رد مرحله‌ای تک تک

در این بخش برای ارزیابی مدل بهینه و بررسی قدرت پیش بینی مدل از تکنیک قدرتمند دیگری به‌عنوان روش رد مرحله‌ای تک تک (LOO) استفاده شد. با به‌کارگیری این تکنیک هر داده مورد مطالعه یک‌بار به‌عنوان داده آزمون خارج شد و مدل LAD-LASSO-LM-ANN با باقی‌مانده داده‌ها آموزش داده شد و فعالیت دارویی داده خارج شده با استفاده از مدل آموزش دیده، پیش‌بینی شد. این عملیات برای همه ترکیبات موجود در مجموعه داده‌ها تکرار شد و فعالیت دارویی همه داده‌ها توسط مدل LAD-LASSO-LM-ANN پیش‌بینی شدند. نتایج حاصل از پیش‌بینی همه داده‌ها بر اساس تکنیک LOO، برای هر سه مجموعه داده‌ها در جدول ۲-۱۹، جدول ۲-۲۰ و جدول ۲-۲۱ آورده شده‌اند. برای مطالعه بیشتر قدرت پیش بینی مدل، نمودار مقادیر پیش‌بینی شده فعالیت دارویی تمام ترکیبات بر حسب مقادیر تجربی فعالیت دارویی آن‌ها رسم شد. مقادیر Q^2_{LOO} برای هر سه مجموعه داده‌ها، از مقدار قابل قبول ۰/۵ بزرگ‌تر است، که نشان‌دهنده پایداری و استحکام مناسب مدل‌های بهینه توسعه یافته با توصیف‌کننده‌های منتخب روش LAD-LASSO می‌باشد. علاوه بر این نمودار باقی‌مانده‌ها برای هر سه مجموعه داده‌ها، از رسم مقادیر پیش‌بینی شده فعالیت دارویی بر حسب مقادیر واقعی آن‌ها به‌دست آمد (شکل ۲-۲۶، شکل ۲-۲۸ و شکل ۲-۳۰). توزیع یکنواخت داده‌ها حول محور صفر، نشان‌دهنده عدم وجود خطای سیستماتیک در مدل‌های شبکه عصبی توسعه یافته با استفاده از روش LAD-LASSO به‌عنوان روش انتخاب متغیر (LAD-LASSO-LM-ANN) است.

جدول ۲-۱۹ نتایج حاصل از ارزیابی مدل LAD-LASSO-LM-ANN با تکنیک LOO برای کل مجموعه داده‌های ضد ایدز

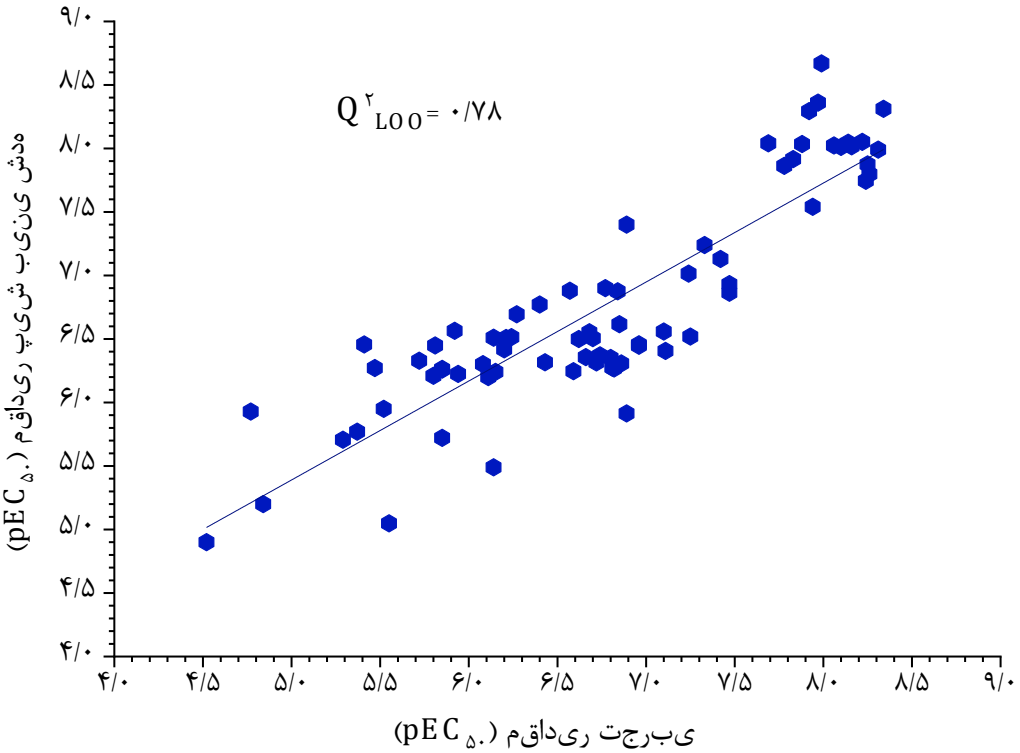
درصد خطا	pEC_{50}		شماره ترکیب	درصد خطا	pEC_{50}		شماره ترکیب
	مقدار پیش‌بینی شده	مقدار واقعی			مقدار پیش‌بینی شده	مقدار واقعی	
-۸/۱۱	۶/۸۶	۷/۴۷	۳۸	-۳/۴۱	۶/۶۲	۶/۸۵	۱
۶/۰۳	۶/۵۱	۶/۱۴	۳۹	-۳/۹۱	۷/۱۳	۷/۴۲	۲
۸/۴۱	۴/۹	۴/۵۲	۴۰	-۶/۰۲	۶/۳۲	۶/۷۲	۳
۸/۴۱	۴/۹	۴/۵۲	۴۱	۷/۴	۷/۴	۶/۸۹	۴
۱۹/۳۶	۶/۴۶	۵/۴۱	۴۲	-۱/۸۱	۶/۵	۶/۶۲	۵
۷/۷۹	۵/۹۵	۵/۵۲	۴۳	-۱۴/۱۷	۵/۹۱	۶/۸۹	۶
-۱۰/۵۹	۵/۴۹	۶/۱۴	۴۴	-۳/۱	۷/۰۲	۷/۲۴	۷
-۸/۰۶	۶/۳۱	۶/۸۶	۴۵	-۱/۲۳	۷/۲۴	۷/۳۳	۸
-۲/۸۸	۶/۵۱	۶/۷	۴۶	-۹/۹	۶/۴۱	۷/۱۱	۹
-۷/۶۳	۶/۵۶	۷/۱	۴۷	-۳/۸۴	۷/۹۹	۸/۳۱	۱۰
-۷/۸۶	۶/۲۷	۶/۸۱	۴۸	-۵/۰۳	۷/۵۴	۷/۹۴	۱۱
-۵/۴۷	۶/۳۷	۶/۷۴	۴۹	۴/۵۷	۸/۰۴	۷/۶۹	۱۲
-۷/۲۱	۶/۹۳	۷/۴۷	۵۰	۶/۰۰	۷/۷۵	۸/۲۴	۱۳
۵/۸۱	۶/۷۷	۶/۴	۵۱	-۵/۵۹	۷/۸	۸/۲۶	۱۴
-۴/۵۳	۶/۳۶	۶/۶۶	۵۲	۸/۵	۸/۶۷	۷/۹۹	۱۵
۰/۵۲	۶/۸۸	۶/۸۴	۵۳	-۱/۰۷	۸/۰۱	۸/۱	۱۶
۶/۷۷	۶/۶۹	۶/۲۷	۵۴	-۴/۵۵	۷/۸۷	۸/۲۵	۱۷
۱/۴۷	۶/۲	۶/۱۱	۵۵	-۱/۱۶	۸/۰۵	۸/۱۴	۱۸
۱۱	۶/۴۵	۵/۸۱	۵۶	۱/۰۷	۷/۸۶	۷/۷۸	۱۹
۷/۸۷	۵/۷۱	۵/۲۹	۵۷	-۰/۳۳	۸/۳۱	۸/۳۴	۲۰
۷/۰۷	۶/۲۱	۵/۸	۵۸	-۱/۷۵	۸/۰۲	۸/۱۶	۲۱
-۵/۲	۶/۲۵	۶/۵۹	۵۹	-۲/۰۴	۸/۰۵	۸/۲۲	۲۲
۳/۵۲	۶/۴۲	۶/۲	۶۰	۱/۹۷	۸/۰۴	۷/۸۸	۲۳
-۱/۹	۶/۵۵	۶/۶۸	۶۱	۱/۱	۷/۹۲	۷/۸۳	۲۴
۱۰/۸۸	۶/۵۶	۵/۹۲	۶۲	-۰/۴۵	۸/۰۲	۸/۰۶	۲۵
۷/۴۴	۵/۲	۴/۸۴	۶۳	۴/۹	۸/۳۶	۷/۹۷	۲۶
۱۰/۶۴	۶/۳۳	۵/۷۲	۶۴	۴/۷۳	۸/۲۹	۷/۹۲	۲۷
-۹/۰۳	۵/۰۵	۵/۵۵	۶۵	۴/۴۱	۶/۵۲	۶/۲۴	۲۸
۴/۸	۶/۲۳	۵/۹۴	۶۶	-۱۰/۰۹	۶/۵۲	۷/۲۵	۲۹
۱۴/۶۵	۶/۲۷	۵/۴۷	۶۷	-۷/۳۸	۶/۴۵	۶/۹۶	۳۰
۳/۶۶	۶/۳	۶/۰۸	۶۸	۴/۷۳	۶/۸۸	۶/۵۷	۳۱
۱/۴۹	۶/۲۴	۶/۱۵	۶۹	-۶/۶۱	۶/۳۵	۶/۸	۳۲
-۲/۱۹	۵/۷۲	۵/۸۵	۷۰	۷/۴۵	۵/۷۷	۵/۳۷	۳۳
۶/۹۹	۶/۲۶	۵/۸۵	۷۱	۴/۷۹	۶/۵۱	۶/۲۱	۳۴
-۱/۷۵	۶/۳۲	۶/۴۳	۷۲	۱/۹۲	۶/۹	۶/۷۷	۳۵
-۸/۰۹	۶/۲۷	۶/۸۲	۷۳	۲۴/۳	۵/۹۳	۴/۷۷	۳۶
				-۷/۲۲	۶/۴۶	۶/۹۶	۳۷

جدول ۲-۲۰ نتایج حاصل از ارزیابی مدل LAD-LASSO-LM-ANN به روش رد مرحله‌ای تک تک برای کل مجموعه داده‌های ضد سرطان کارسینوم کولورکتال

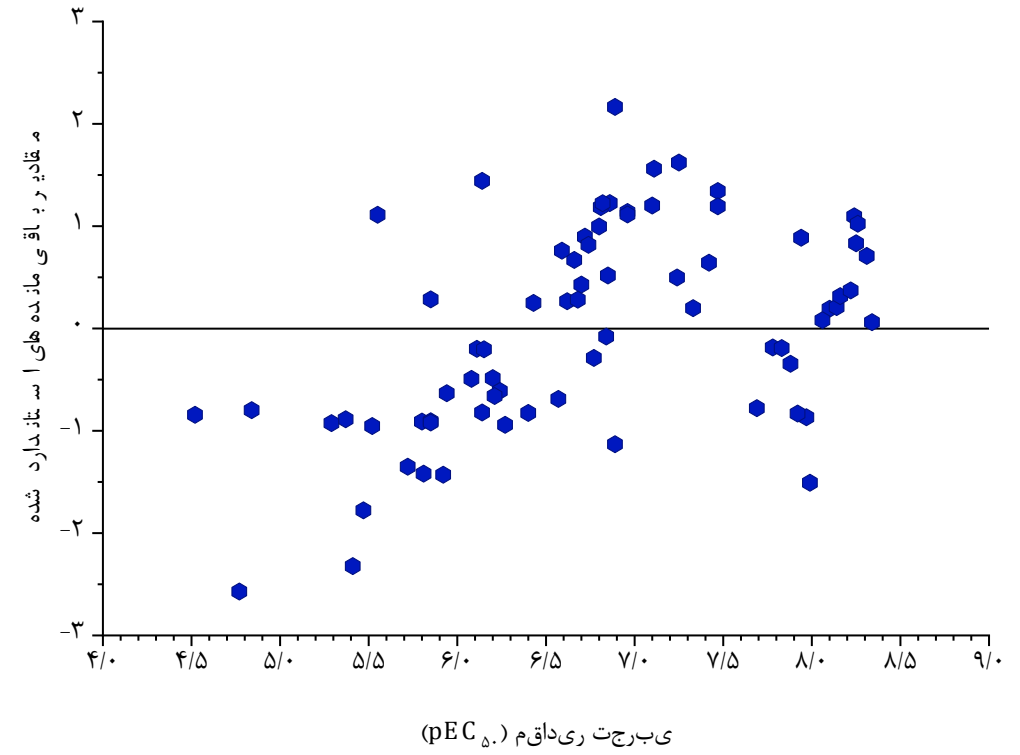
شماره ترکیب	pIC۵۰		درصد خطا	شماره ترکیب	pIC۵۰		درصد خطا
	مقدار واقعی	مقدار پیش‌بینی شده			مقدار واقعی	مقدار پیش‌بینی شده	
۱	۵/۶۷	۵/۴۷	-۳/۵۵	۴۱	۴/۷۴	۴/۸۹	۳/۱۶
۲	۶/۳۶	۵/۹۴	-۶/۵۷	۴۲	۴/۷۲	۴/۷۹	۱/۴۸
۳	۶/۳۲	۵/۹۳	-۶/۱۵	۴۳	۶/۶۷	۶/۱۷	-۷/۴۹
۴	۶/۲۴	۵/۶۹	-۸/۸۴	۴۴	۵/۷۶	۴/۸۴	-۱۶/۰۶
۵	۵/۴۷	۵/۹۶	۸/۹۳	۴۵	۴/۷۹	۴/۷۲	-۱/۵۵
۶	۵/۶۳	۵/۹۱	۴/۹	۴۶	۵/۲۹	۵/۶۹	۷/۵
۷	۵/۸۷	۴/۸۱	۱۸/۰۰	۴۷	۵/۳۹	۴/۸۵	-۱۰/۰۵
۸	۵/۳۶	۵/۸	۸/۱۷	۴۸	۵/۶۵	۵/۴۱	-۴/۲۶
۹	۵/۳۶	۵/۰۳	-۶/۲۳	۴۹	۶/۴۴	۶/۵۵	۱/۷۷
۱۰	۵/۲	۴/۹۴	-۴/۹۳	۵۰	۵/۶۵	۵/۴۸	-۳/۰۵
۱۱	۵/۶۳	۶/۲۷	۱۱/۲۹	۵۱	۷/۰۰	۶/۹۶	-۰/۶۲
۱۲	۷/۰۹	۶/۶۴	-۶/۳۸	۵۲	۶/۲	۶/۴۲	۳/۵۹
۱۳	۶/۷۲	۷/۳۲	۸/۸۹	۵۳	۶/۰۰	۶/۵۲	۸/۶۲
۱۴	۴/۲۳	۴/۴۵	۵/۳۲	۵۴	۴/۳۷	۴/۵	۲/۹۲
۱۵	۴/۲۸	۴/۱۸	-۲/۲۹	۵۵	۵/۰۹	۵/۶۱	۱۰/۲۱
۱۶	۴/۲۱	۴/۱۹	-۰/۵۶	۵۶	۵/۲۰	۵/۱۵	-۱/۰۳
۱۷	۴/۱۸	۴/۴	۵/۱۷	۵۷	۵/۱۷	۶/۰۰	۱۶/۰۵
۱۸	۴/۳۷	۴/۲۲	-۳/۳۳	۵۸	۴/۴۸	۴/۷۲	۵/۲۸
۱۹	۴/۲	۴/۳۶	۳/۹۲	۵۹	۴/۵۶	۵/۲۶	۱۵/۳۵
۲۰	۴/۴۵	۴/۲۴	-۴/۶۱	۶۰	۴/۷۱	۴/۹۲	۴/۵۶
۲۱	۴/۵۴	۵/۰۱	۱۰/۲۶	۶۱	۴/۹۳	۵/۳۹	۹/۲۷
۲۲	۴/۶۵	۴/۷۳	۱/۶۳	۶۲	۵/۳۸	۶/۱۷	۱۴/۷۲
۲۳	۴/۳	۴/۳۴	۰/۹۵	۶۳	۵/۹۸	۴/۷۴	-۲۰/۷۹
۲۴	۴/۳۵	۴/۳۵	۰/۰۳	۶۴	۵/۵۷	۴/۹۲	-۱۱/۷
۲۵	۴/۵۴	۴/۶	۱/۳	۶۵	۶/۴۴	۵/۲۴	-۱۸/۷۱
۲۶	۴/۲	۴/۳۵	۳/۶۹	۶۶	۶/۹۲	۷/۶۷	۱۰/۸۵
۲۷	۴/۱۷	۴/۳	۳/۰۹	۶۷	۷/۰۹	۵/۷۴	-۱۹/۰۹
۲۸	۴/۲۱	۴/۲۳	۰/۳۷	۶۸	۷/۱۵	۶/۴۴	-۹/۹۷
۲۹	۵/۴۷	۴/۹۸	-۹/۰۲	۶۹	۴/۰۸	۴/۱۵	۱/۶۵
۳۰	۴/۳	۴/۴۳	۲/۹۱	۷۰	۴/۱۷	۴/۱۲	-۱/۱۲
۳۱	۴/۲۸	۴/۸۲	۱۲/۶۲	۷۱	۴/۰۱	۴/۴۶	۱۱/۰۸
۳۲	۴/۷۴	۴/۹	۳/۳۴	۷۲	۴/۰۷	۴/۳۳	۶/۴۳
۳۳	۴/۵۹	۴/۱۹	-۸/۷۵				
۳۴	۴/۷۱	۴/۴۷	-۵/۰۹				
۳۵	۵/۱	۵/۱۸	۱/۵۴				
۳۶	۴/۴۹	۵/۲	۱۵/۸۴				
۳۷	۵/۳۴	۴/۵۱	-۱۵/۴۹				
۳۸	۴/۷۶	۴/۶۴	-۲/۴۴				
۳۹	۴/۷۴	۴/۹۴	۴/۲۲				
۴۰	۴/۶۶	۴/۷۵	۲/۰۳				

جدول ۲-۲۱ نتایج حاصل از ارزیابی مدل LAD-LASSO-LM-ANN به روش رد مرحله‌ای تک تک برای کل مجموعه داده‌های ضد سرطان ریه

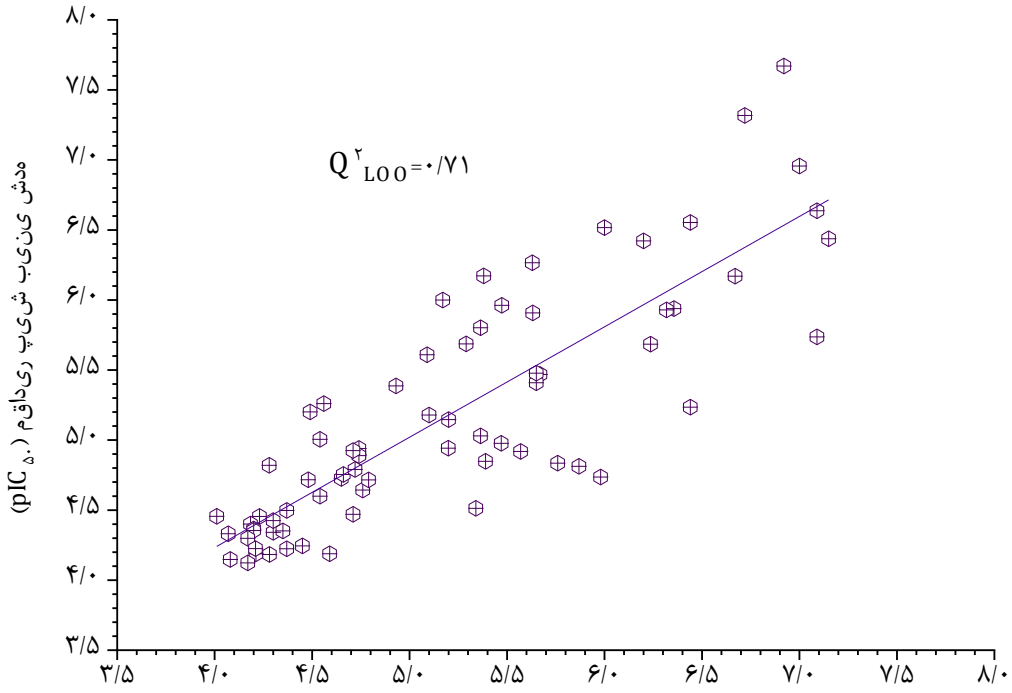
شماره ترکیب	pIC_{50}		درصد خطا	شماره ترکیب	pIC_{50}		درصد خطا
	مقدار واقعی	مقدار پیش‌بینی شده			مقدار واقعی	مقدار پیش‌بینی شده	
۱	۵/۱۴	۵/۰۸	-۱/۱۲	۴۱	۴/۵۵	۴/۶۲	۱/۴۷
۲	۶/۰۷	۶/۰۷	-۰/۰۲	۴۲	۴/۴۱	۴/۳	-۲/۴۹
۳	۶/۱۱	۶/۳۹	۴/۵۶	۴۳	۶/۲۱	۶/۲۴	۰/۴۸
۴	۵/۷۵	۴/۸۳	-۱۵/۹۸	۴۴	۵/۶۱	۵/۵۹	-۰/۳۶
۵	۴/۹۹	۴/۲۴	-۱۵/۰۶	۴۵	۴/۴۳	۵/۲۳	۱۸/۰۶
۶	۵/۶۵	۵/۷۷	۲/۱۲	۴۶	۵/۰۷	۴/۴۵	-۱۲/۲۱
۷	۵/۴۸	۵/۰۷	-۷/۵۲	۴۷	۴/۹۸	۴/۷۴	-۴/۷۳
۸	۵/۰۳	۵/۵۲	۹/۷۷	۴۸	۵/۲	۵/۳۱	۲/۱۴
۹	۵/۲	۴/۷۷	-۸/۳۴	۴۹	۶/۱۱	۵/۶۸	-۷/۰۴
۱۰	۴/۹۱	۴/۲۸	-۱۲/۸۳	۵۰	۵/۲	۵/۴۶	۴/۹۳
۱۱	۵/۳۷	۵/۹	۹/۹۳	۵۱	۶/۵۱	۶/۷۹	۴/۳
۱۲	۶/۵۵	۶/۸۹	۵/۱۹	۵۲	۵/۶۵	۵/۴۳	-۳/۸۹
۱۳	۶/۵۱	۶/۵۵	۰/۶۱	۵۳	۵/۴۹	۵/۳۱	-۳/۲۸
۱۴	۴/۰۴	۴/۱۱	۱/۶۱	۵۴	۴/۲۲	۴/۵۴	۷/۵۲
۱۵	۴/۰۷	۴/۱	۰/۷۷	۵۵	۵/۰۱	۴/۳۳	-۱۳/۶۳
۱۶	۴/۰۵	۴/۱۳	۱/۹۱	۵۶	۵/۰۳	۴/۴۴	-۱۱/۸۳
۱۷	۴/۱۷	۴/۳۸	۴/۹۸	۵۷	۵/۱۱	۴/۷	-۸/۰۶
۱۸	۴/۱۴	۴/۱۲	-۰/۵۵	۵۸	۴/۳	۴/۸	۱۱/۶۳
۱۹	۳/۹۴	۴/۴۲	۱۲/۱۹	۵۹	۴/۴۹	۴/۸۴	۷/۸
۲۰	۴/۲۳	۴/۱۳	-۲/۴۳	۶۰	۴/۶۶	۴/۶۲	-۰/۸۶
۲۱	۴/۶۱	۴/۷۱	۲/۱۹	۶۱	۴/۸۶	۴/۸۷	۰/۱۶
۲۲	۴/۷۱	۴/۷۴	۰/۷۱	۶۲	۵/۳۵	۴/۷۴	-۱۱/۳۵
۲۳	۴/۱۱	۴/۲۸	۴/۰۵	۶۳	۶/۰۱	۵/۴۸	-۸/۷۹
۲۴	۴/۴۶	۴/۲۵	-۴/۶۱	۶۴	۵/۷۲	۴/۸۶	-۱۵/۰۹
۲۵	۴/۵	۴/۷۴	۵/۳۱	۶۵	۶/۴۶	۵/۸۴	-۹/۵۹
۲۶	۴/۰۱	۴/۳۷	۸/۹۷	۶۶	۶/۳۲	۶/۳۲	۰/۰۰
۲۷	۴/۱	۴/۲۱	۲/۷	۶۷	۶/۷۷	۶/۶۵	-۱/۷۸
۲۸	۴/۱۶	۴/۱	-۱/۵۶	۶۸	۶/۹۲	۶/۰۱	-۱۳/۱۳
۲۹	۵/۱۸	۵/۳۳	۲/۹۵	۶۹	۴/۰۵	۴/۱۸	۳/۱۸
۳۰	۴/۰۷	۴/۲۳	۴/۰۱	۷۰	۴/۱	۴/۲۱	۲/۵۶
۳۱	۴/۲۱	۴/۶۷	۱۰/۹۹				
۳۲	۴/۷۵	۴/۶۵	-۲/۲				
۳۳	۴/۴۹	۴/۳۸	-۲/۳۹				
۳۴	۴/۳۴	۴/۶۷	۷/۵۲				
۳۵	۵/۰۳	۵/۷۱	۱۳/۵۴				
۳۶	۴/۵۶	۴/۶	۰/۸۵				
۳۷	۴/۸۸	۴/۳۵	-۱۰/۸۸				
۳۸	۴/۴۵	۴/۶۹	۵/۳۹				
۳۹	۴/۵	۴/۹۸	۱۰/۶۸				
۴۰	۴/۴۷	۴/۲۹	-۳/۹۹				



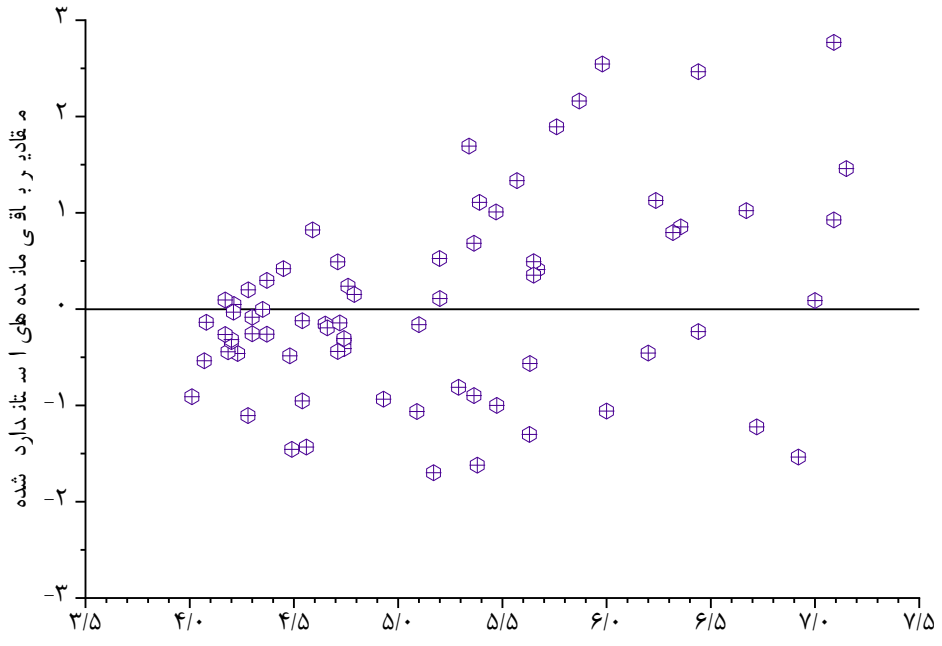
شکل ۲-۲۵ نمودار تغییرات مقادیر پیش‌بینی شده همه داده‌های ضد ایدز بر اساس تکنیک LOO در مقابل مقادیر تجربی



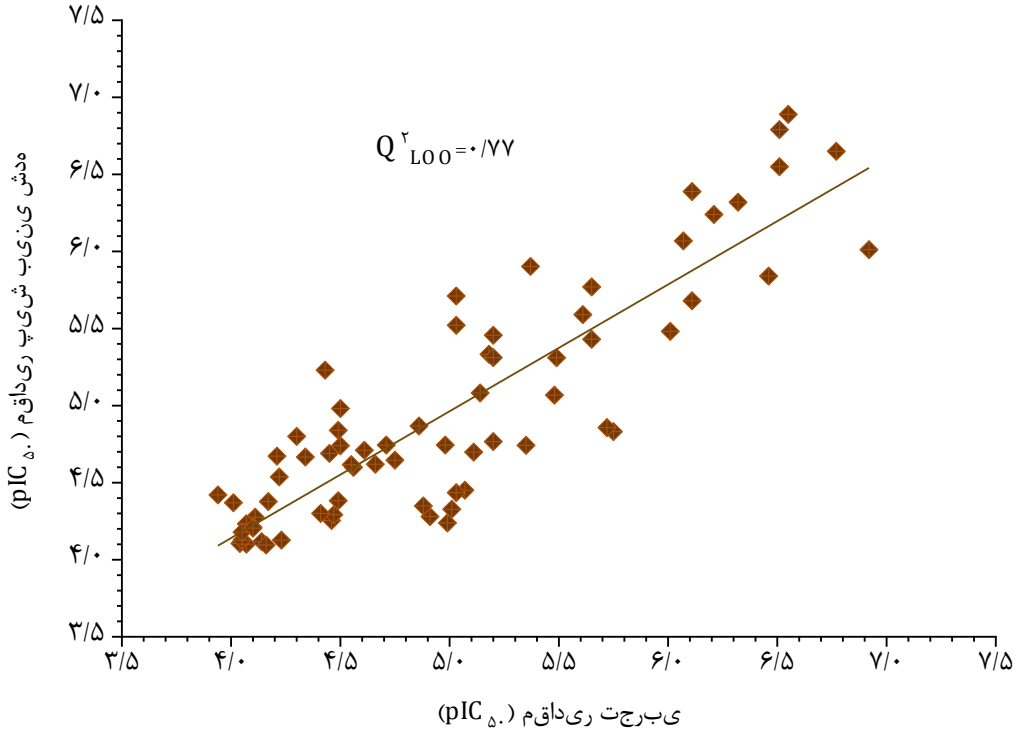
شکل ۲-۲۶ نمودار باقی‌مانده‌های پیش‌بینی شده داده‌های ضد ایدز با استفاده از تکنیک LOO برحسب مقادیر تجربی



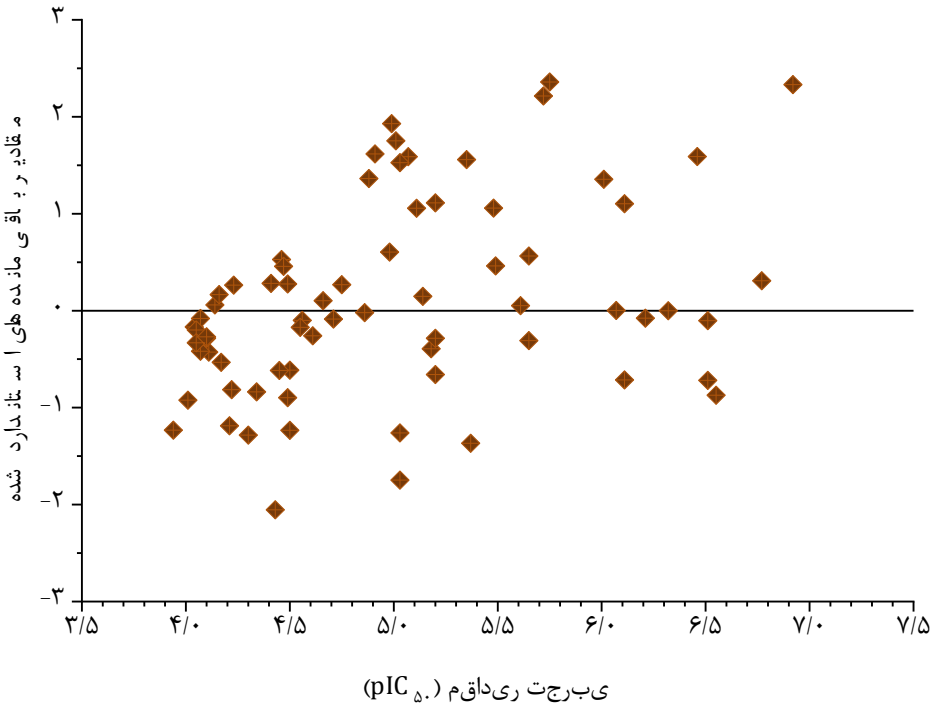
شکل ۲-۲۷ نمودار تغییرات مقادیر پیش‌بینی شده همه داده‌های ضد سرطان کارسینوم کولورکتال بر اساس تکنیک LOO در مقابل مقادیر تجربی



شکل ۲-۲۸ نمودار باقی‌مانده‌های پیش‌بینی شده داده‌های ضد سرطان کارسینوم کولورکتال با استفاده از تکنیک LOO برحسب مقادیر تجربی



شکل ۲-۲۹ نمودار تغییرات مقادیر پیش‌بینی شده همه داده‌های ضد سرطان ریه بر اساس تکنیک LOO در مقابل مقادیر تجربی



شکل ۲-۳۰ نمودار باقی‌مانده‌های پیش‌بینی شده داده‌های ضد سرطان ریه با استفاده از تکنیک LOO برحسب مقادیر تجربی

۲- ۴- ۷- ۳ ارزیابی مدل LAD-LASSO-LM-ANN با استفاده از پارامترهای آماری

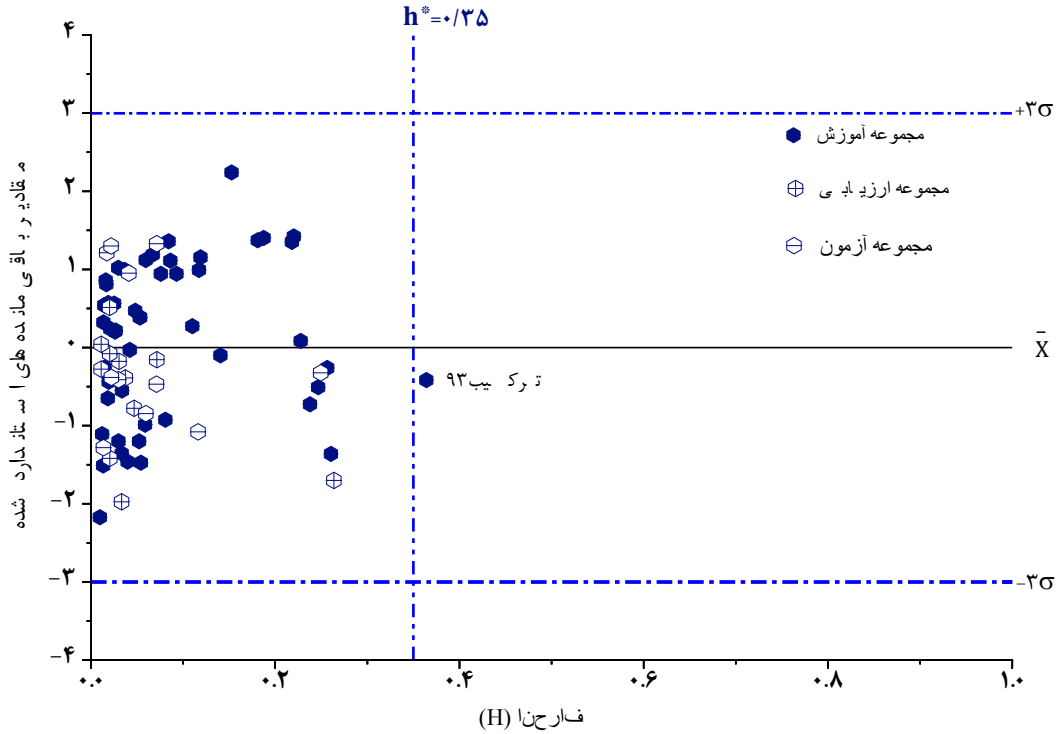
علاوه بر تکنیک‌های ذکر شده، یکی از روش‌های ارزیابی مدل LAD-LASSO-ANN توسعه یافته، محاسبه پارامترهای آماری متفاوت می‌باشد. از جمله پارامترهای آماری پر کاربرد می‌توان به پارامترهای متفاوت بررسی خطای مدل، پارامترهای آماری تروپشا و روی اشاره کرد. بنابراین پارامترهای آماری معرفی شده در بخش ۱- ۵- ۸- ۴ برای فعالیت دارویی پیش‌بینی شده ترکیبات مجموعه آزمون و فعالیت‌های دارویی پیش‌بینی شده برای کل ترکیبات به روش رد مرحله‌ای تک تک محاسبه و در جدول ۲-۲۲ خلاصه شدند. نتایج حاصل از محاسبات آماری جدول ۲-۲۲ نشان می‌دهد که پارامترهای آماری در محدوده قابل قبول قرار دارند. بنابراین مطابق با بخش ۱- ۵- ۸- ۴، با توجه به بزرگتر بودن مقادیر پارامترهای تروپشا و روی همچون R_0^2، R_0^2 نسبی، R_m^2 و غیره از مقدار قابل قبول ۰/۵ و نزدیک بودن این پارامترها به مقدار R^2 قدرت پیش بینی و تعمیم پذیری مدل توسعه یافته LAD-LASSO-ANN اثبات می‌شود. علاوه بر این شیب نمودار حاصل از مقادیر پیش بینی شده بر حسب مقادیر تجربی (و بالعکس) در عرض از مبدأ صفر نیز در محدوده ۰/۸۵ تا ۱/۱۵ قرار دارند که این نتیجه نیز حاکی از صحت مدل توسعه یافته ANN با توصیف کننده‌های منتخب روش LAD-LASSO می‌باشد.

جدول ۲-۲۲ پارامترهای آماری محاسبه شده برای مجموعه آزمون و داده‌های پیش‌بینی شده با تکنیک LOO برای مدل LAD-LASSO-LM-ANN هر سه مجموعه از داده‌ها

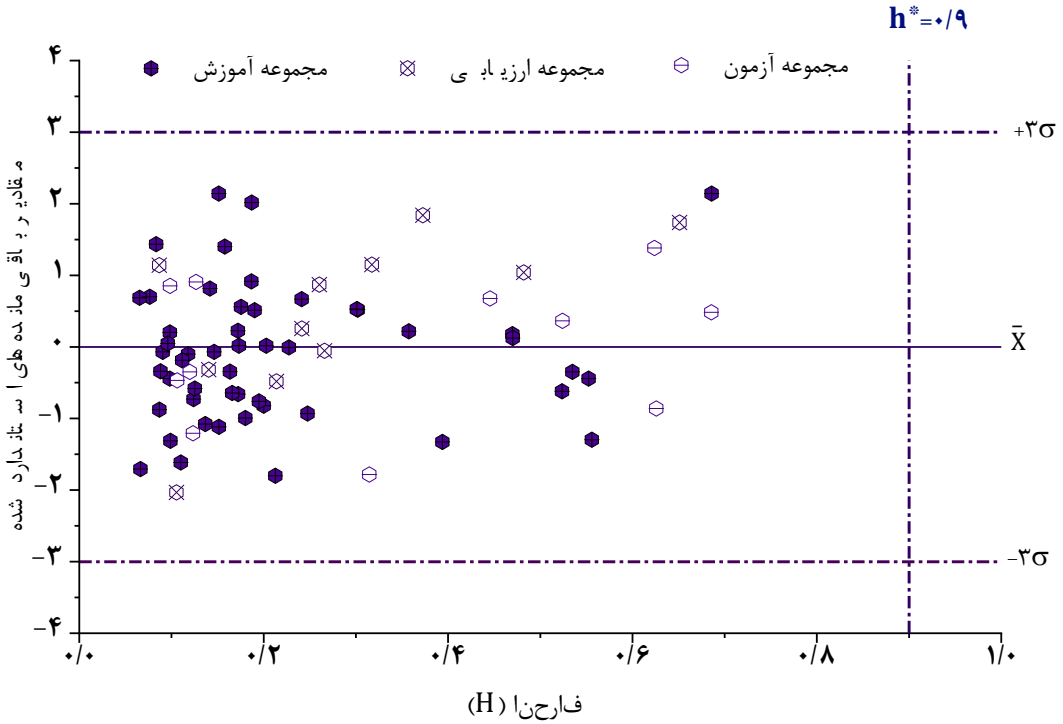
ردیف	پارامتر آماری	مجموعه داده‌های ضد ایدز		مجموعه داده‌های ضد سرطان کارسینوم کولورکتال		مجموعه داده‌های ضد سرطان ریه		محدوده قابل قبول
		نتایج فعالیت دارویی پیش بینی شده با LAD-LASSO-ANN						
		کل ترکیبات مجموعه آزمون	کل ترکیبات به روش LOO	ترکیبات مجموعه آزمون	کل ترکیبات به روش LOO	ترکیبات مجموعه آزمون	کل ترکیبات به روش LOO	
۱	PRESS	۱/۴۸	۱۴/۶۲	۰/۸۰	۱۶/۹۶	۱/۱۷	۱۰/۵۷	-
۲	SEP	۰/۳۷	۰/۴۵	۰/۲۷	۰/۴۹	۰/۳۳	۰/۳۹	-
۳	MAE	۰/۲۹	۰/۳۶	۰/۲۳	۰/۳۰	۰/۲۸	۰/۲۸	-
۴	REP(%)	۵/۶۲	۶/۶۱	۵/۱۴	۹/۳۷	۶/۵۴	۷/۷۶	-
۵	MSE	۰/۱۳	۰/۲۰	۰/۰۷	۰/۲۴	۰/۱۱	۰/۱۵	-
۶	MRE	۵/۲۴	۵/۹۲	۴/۴۲	۶/۸۴	۵/۸۹	۵/۹۳	-
۷	R^2	۰/۸۷	-	۰/۸۴	-	۰/۸۷	-	R^2>۰/۶
۸	Q^2_{Loo}	-	۰/۷۸	-	۰/۷۱	-	۰/۷۷	$\boldsymbol{Q^2_{Loo}}$> ۰/۵
۹	R_0^2	۰/۸۶	۰/۷۲	۰/۷۷	۰/۶۶	۰/۸۵	۰/۷۳	نزدیک به R^2
۱۰	R_0^2 نسبی	۰/۰۱	۰/۰۸	۰/۰۸	۰/۰۷	۰/۰۲	۰/۰۵	<۰/۱
۱۱	R_m^2	۰/۷۸	۰/۵۹	۰/۶۲	۰/۵۵	۰/۷۵	۰/۶۲	>۰/۵
۱۲	$R_0'^2$	۰/۸	۰/۷۸	۰/۸۴	۰/۷	۰/۸۷	۰/۷۶	نزدیک به R^2
۱۳	$R_0'^2$ نسبی	۰/۰۸	۰/۰۰	۰/۰۰	۰/۰۱	۰/۰۰	۰/۰۱	<۰/۱
۱۴	$R_m'^2$	۰/۶۵	۰/۷۲	۰/۵۷	۰/۵۹	۰/۷۳	۰/۶۶	>۰/۵
۱۵	R-R	۰/۰۶	۰/۰۶	۰/۰۷	۰/۰۱	۰/۰۲	۰/۰۳	<۰/۳
۱۶	k	۰/۹۷	۱	۰/۹۸	۰/۹۹	۰/۹۸	۱	۰/۸۵ ≤ $\boldsymbol{k}$ ≤ ۱/۱۵
۱۷	k′	۱/۰۳	۱	۱/۰۲	۱	۱/۰۱	۱	۰/۸۵ ≤ $\mathbf{k'}$ ≤ ۱/۱۵

۲- ۴- ۷- ۴ ارزیابی مدل LAD-LASSO-LM-ANN با استفاده از دامنه کاربرد

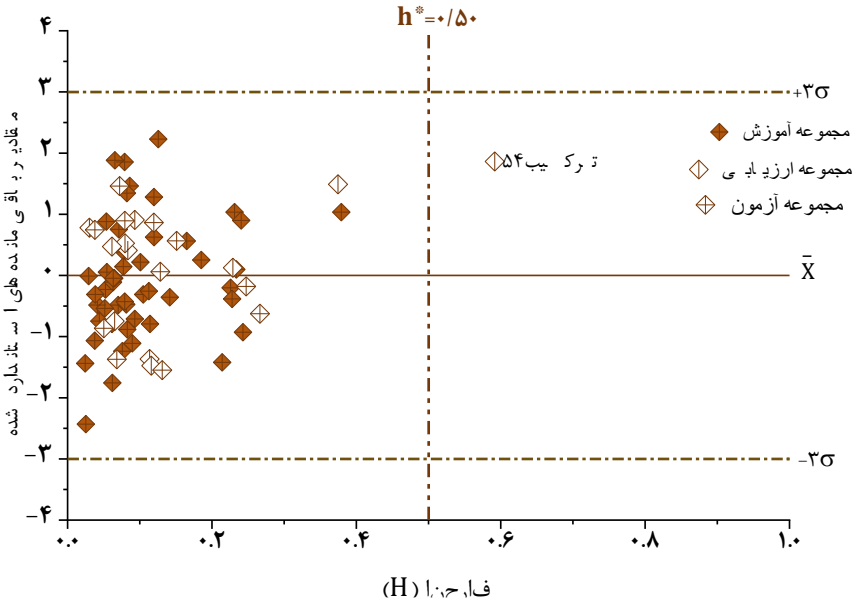
از جمله روش‌های ارزیابی اعتبار مدل QSAR توسعه یافته می‌توان به آزمون دامنه کاربرد اشاره کرد. در واقع با استفاده از این‌روش، یک فضای شیمیایی تئوری با استفاده از توصیف‌کننده‌های مولکولی مجموعه آموزش و فعالیت دارویی مربوطه ایجاد می‌شود. آزمون دامنه کاربرد مدل توسعه یافته LAD-LASSO-LM-ANN با استفاده از رسم نمودار ویلیام مورد تجزیه و تحلیل قرار گرفت. بنابراین مطابق با توضیحات ارائه شده در بخش‌های ۱- ۵- ۸- ۵ و ۲- ۲- ۷- ۴ مقادیر H هر ترکیب با استفاده از رابطه ۱-۱۳ محاسبه شد. سپس مقادیر باقی‌مانده‌های استاندارد شده طبق رابطه ۱-۱۴ محاسبه شد. نمودار ویلیام از رسم مقادیر باقی‌مانده‌های استاندارد شده بر حسب مقابل مقادیر H به‌دست آمد. همان‌طور که شکل ۲-۳۱ و شکل ۲-۳۲ نشان می‌دهند، مقادیر محاسبه شده H و باقی‌مانده‌های استاندارد شده برای مجموعه داده‌های ضد ایدز و ضد سرطان کارسینوم کولورکتال، در محدوده قابل قبول قرار دارند و هیچ کدام از داده‌های به‌کار گرفته شده در ساخت، ارزیابی و آزمون مدل LAD-LASSO-LM-ANN به‌عنوان داده دور افتاده نیستند. با توجه به نمودار ویلیام (شکل ۲-۳۳) مربوط به مجموعه داده‌های ضد سرطان ریه، مشاهده می‌شود که تنها یک داده بزرگ‌تر از مقدار هشدار *h است و سایر داده‌ها در محدوده قابل قبول قرار گرفته‌اند. بنابراین مدل‌های LAD-LASSO-LM-ANN توسعه یافته برای هر سه مجموعه داده‌ها از استحکام و اطمینان‌پذیری قابل قبولی برخوردار هستند.



شکل ۲-۳۱ دامنه کاربرد مدل LAD-LASSO-LM-ANN برای مجموعه داده‌های ضد ایدز، خطوط نقطه چین افقی و عمودی در دو انتهای نمودار به ترتیب نمایانگر مقادیر ۳σ ± و h* است.



شکل ۲-۳۲ دامنه کاربرد مدل LAD-LASSO-LM-ANN برای مجموعه داده‌های ضد سرطان کارسینوم کولورکتال، خطوط نقطه چین افقی و عمودی در دو انتهای نمودار به ترتیب نمایانگر مقادیر ۳σ ± و h* است.

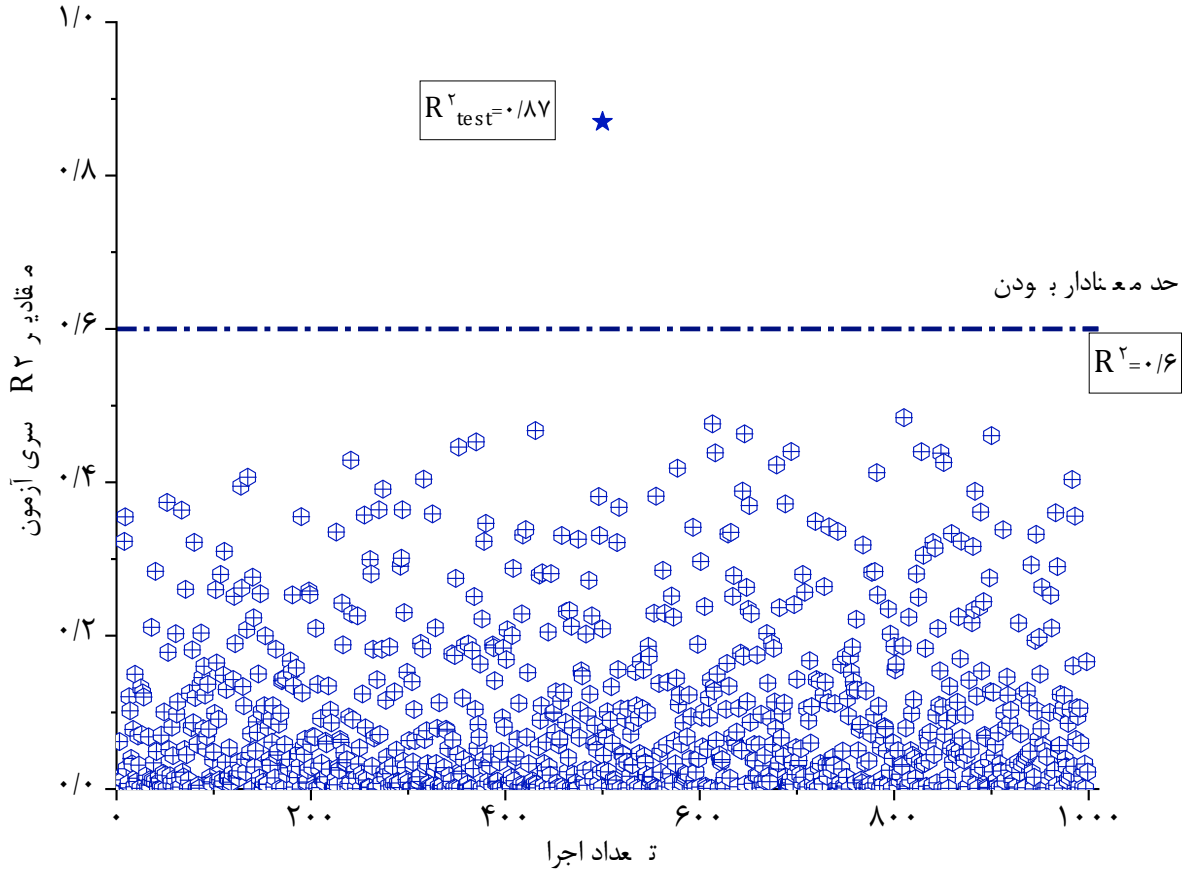


شکل ۲-۳۳ دامنه کاربرد مدل LAD-LASSO-LM-ANN برای مجموعه داده‌های ضد سرطان ریه، خطوط نقطه چین افقی و عمودی در دو انتهای نمودار به ترتیب نمایانگر مقادیر ۳σ ± و h* است.

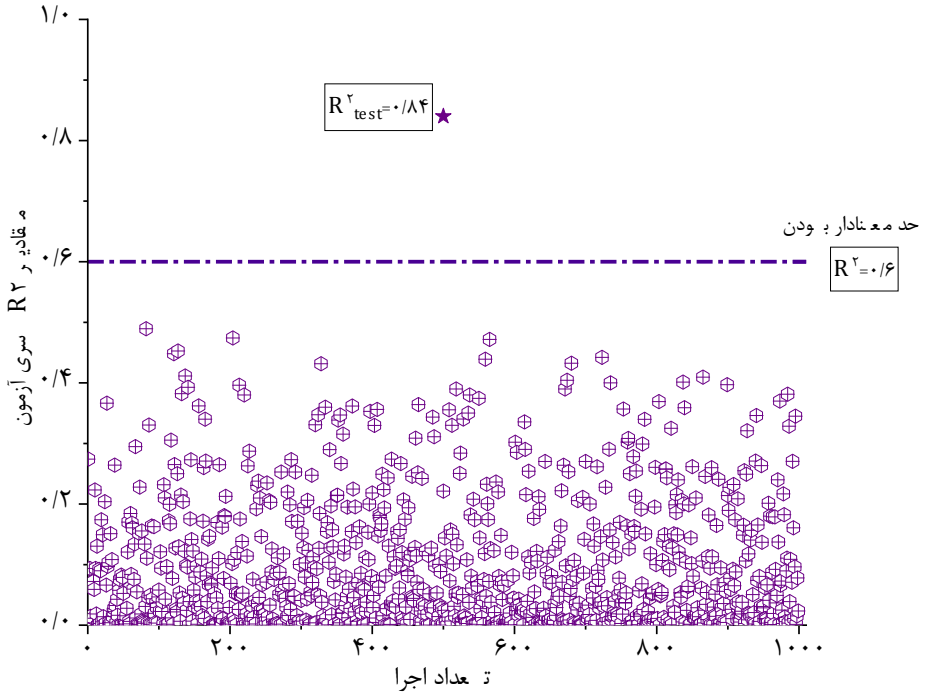
۲- ۴- ۷- ۵ ارزیابی مدل LAD-LASSO-LM-ANN با استفاده از آزمون Y-تصادفی

آزمون Y-تصادفی برای بررسی عدم وجود ارتباط تصادفی ایجاد شده توسط مدل LAD-LASSO-LM-ANN بین متغیرهای مستقل و متغیر وابسته به‌کار گرفته شد. بنابراین برای انجام آزمون Y-تصادفی، ابتدا مقادیر فعالیت دارویی مربوط به هر مجموعه داده، در محدوده تغییرات خودش (حداقل و حداکثر مقدار فعالیت دارویی) ۱۰۰۰ بار به طور کاملاً تصادفی تغییر داده شدند. مدل شبکه عصبی بهینه مربوط به هر مجموعه داده با استفاده از فعالیت‌های دارویی تصادفی آموزش داده شد و مقادیر فعالیت‌های دارویی مجموعه آزمون با استفاده از مدل بهینه LAD-LASSO-LM-ANN توسعه یافته با متغیر وابسته تصادفی، پیش‌بینی شدند. این فرایند برای هر ۳ مجموعه داده انجام شد. نمودار مقادیر پیش‌بینی شده بر حسب مقادیر واقعی برای هر مجموعه داده رسم شد و مقادیر R^2 مربوطه به‌دست آمد. نتایج R^2 حاصل از پیش بینی فعالیت دارویی ترکیبات مجموعه آزمون برای هر ۱۰۰۰ مدل توسعه یافته با متغیر وابسته به‌دست آمد. مقادیر R^2

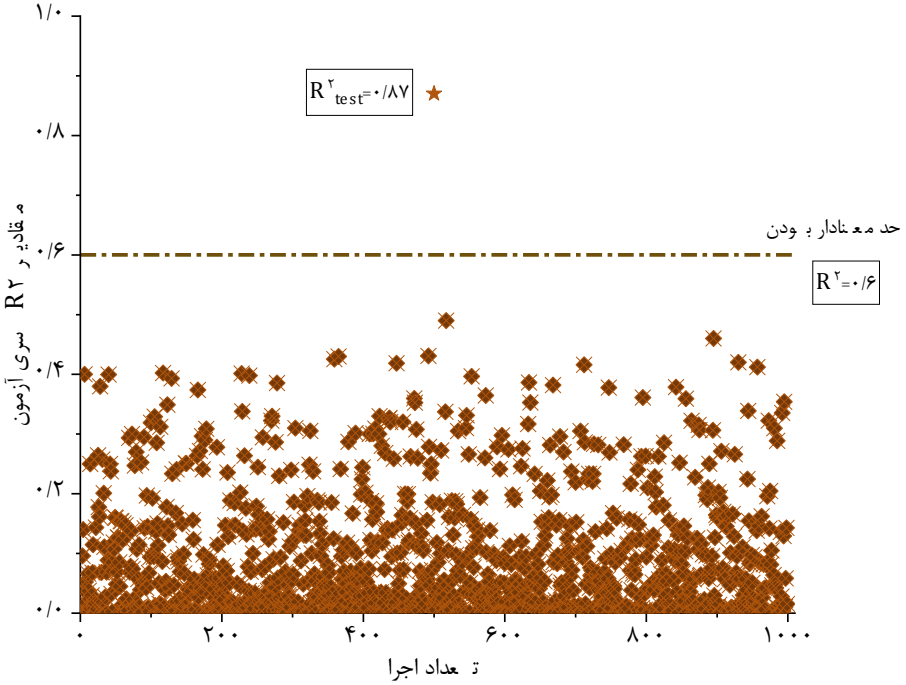
مربوط به ۱۰۰۰ اجرا در هر سه مجموعه داده در شکل ۲-۳۴ تا شکل ۲-۳۵ نشان داده شد. همان‌طور که مشاهده می‌شود مقادیر R^2 حاصل از مدل LAD-LASSO-LM-ANN توسعه یافته با متغیر وابسته تصادفی از مقدار قابل قبول ۰/۶ کوچک‌تر هستند، در نتیجه استنباط می‌شود که مدل‌های توسعه یافته LAD-LASSO-LM-ANN بر اساس ارتباط منطقی بین توصیف کننده‌های منتخب روش LAD-LASSO و فعالیت دارویی مربوطه ساخته شده است و ارتباط ایجاد شده به‌طور تصادفی و شانسی به وجود نیامده است.



شکل ۲-۳۴ نمودار مقادیر R^2 به‌دست آمده در آزمون Y-تصادفی بر حسب تعداد اجرا برای ۱۰۰۰ اجرای Y-تصادفی و پیش‌بینی فعالیت ترکیبات ضد ایدز مجموعه آزمون به‌وسیله مدل LAD-LASSO-LM-ANN با استفاده از پاسخ تصادفی شده در شرایط بهینه



شکل ۲-۳۵ نمودار مقادیر R^2 به‌دست آمده در آزمون Y-تصادفی بر حسب تعداد اجرا برای ۱۰۰۰ اجرای Y-تصادفی و پیش‌بینی فعالیت ترکیبات ضد سرطان کارسینوم کولورکتال مجموعه آزمون به‌وسیله مدل LAD-LASSO-LM-ANN با استفاده از پاسخ تصادفی شده در شرایط بهینه



شکل ۲-۳۶ نمودار مقادیر R^2 به‌دست آمده در آزمون -Yتصادفی بر حسب تعداد اجرا برای ۱۰۰۰ اجرای -Yتصادفی و پیش‌بینی فعالیت ترکیبات ضد سرطان ریه مجموعه آزمون به‌وسیله مدل LAD-LASSO-LM-ANN با استفاده از پاسخ تصادفی شده در شرایط بهینه

۲- ۵ پیش‌بینی شاخص بازداری برخی از ترکیبات آلی فرار با استفاده از مدل SCAD-ANN

۲- ۵- ۱ مقدمه

با توجه به مشکلات و محدودیت‌های کارهای آزمایشگاهی، استفاده از روش‌های تئوری برای محاسبه خواص و ویژگی‌های ترکیبات بسیار حائز اهمیت می‌باشد. امروزه مطالعات کمومتریکس راهی برای استخراج حداکثر اطلاعات از نتایج تجربی با استفاده از انجام یک سری محاسبات آماری و ریاضی می‌باشد. مطالعات ارتباط کمی ساختار - خاصیت (QSPR)، یکی از روش‌های کمومتریکس است که با استفاده از آن می‌توان روابط خطی و یا غیر خطی بین ویژگی‌های ساختاری و خواص ترکیبات را پیدا نمود و از روی آن خاصیت مورد نظر را برای ترکیبات دیگر پیش‌بینی نمود [۱۸۰-۱۷۴]. بنابراین، در شیمی محاسباتی، یافتن رابطه بین خصوصیات ساختاری ترکیبات و خواص آن‌ها مورد توجه محققین است. در میان رویکردهای مختلف شیمی محاسباتی، مطالعه QSPR ابزار مفیدی برای یافتن یک رابطه منطقی مناسب است و به‌طور گسترده در زمینه‌های مختلف مانند کمومتریکس و شیمی دارویی استفاده شده است. QSPR به‌عنوان یک ابزار قوی برای تجزیه و تحلیل خواص کروماتوگرافی نیز شناخته شده است. اخیراً محاسبه شاخص بازداری در مقالات متفاوت گزارش شده است [۸۹, ۹۱, ۱۸۶-۱۸۱]. با توجه به این‌که شناسایی ترکیبات اغلب با تطبیق پیک‌های کروماتوگرافی گازی (GC) آنالیت با پیک‌های استاندارد انجام می‌شود، درحالی‌که نمونه‌های استاندارد خالص گاهی اوقات در دسترس نیستند، امر تطبیق پیک با مشکل مواجه می‌شود. تکنیک GC همچنین به آماده‌سازی نمونه و ستون و بهینه‌سازی پارامتر نیاز دارد که زمان‌بر و پرهزینه است [۹۱, ۱۸۷]. از طرفی، به دلیل پیچیدگی روش‌های تجزیه‌ای و تحلیل کمی، تعیین شاخص بازداری (RI)[1] مواد شیمیایی

[1] Retention index

همیشه برای محققین امری زمان بر است. بنابراین، معرفی روش‌های کارآمد برای پیش‌بینی مقادیر RI برای ترکیبات ناشناخته بدون هیچ گونه اندازه‌گیری یا آزمایش مورد توجه است [۸۹]. مدل‌های مطالعه ارتباط کمی ساختار- شاخص بازداری (QSRR) (RI) می‌توانند برای پیش‌بینی شاخص‌های بازداری ساختارهای شیمیایی مورد استفاده قرار گیرند [۸۹, ۹۱, ۱۸۶]. همان‌طور که گفته شد، هدف اصلی مطالعه QSRR، ایجاد یک رابطه ریاضی بین RI به‌عنوان پاسخ یک سیستم کروماتوگرافی و خواص مولکولی به‌عنوان توصیف‌کننده‌های ساختار آنالیت، می‌باشد. بنابراین انتخاب مؤثرترین توصیف کننده‌ها با بیش‌ترین ارتباط با متغیر پاسخ، امر مهمی در ساخت مدل QSRR است. از این‌رو، معرفی تکنیک‌های انتخاب متغیر جدید و کارآمد برای افزایش عملکرد مدل‌سازی، بهبود تفسیرپذیری مدل، و کاهش پیچیدگی محاسبات و ذخیره‌سازی ضروری است [۱۸۸, ۱۸۹]. بنابراین، به‌عنوان یک روش انتخاب متغیر قدرتمند، از SCAD برای ساخت مدل QSRR با نتایج رضایت‌بخش استفاده شد [۱۰۷, ۱۰۹, ۱۹۰]. با توجه به جنبه‌های هیجان انگیز SCAD، مانند پراکندگی و ثبات بالا، این‌روش می‌تواند جایگزین مناسبی برای روش‌های کلاسیک انتخاب متغیر و یا LASSO باشد. با توجه به جستجوی انجام شده در مقالات منتشر شده، هیچ گزارشی مبنی بر استفاده از SCAD-ANN در مطالعات QSRR وجود ندارد. بنابراین، در این مطالعه از روش SCAD به‌عنوان تکنیک کاهش متغیر در مطالعه QSRR مقادیر RI برای برخی از ترکیبات آلی فرار استفاده شد.

طبق توضیحات ذکر شده، در این بخش از مطالعه، روش SCAD به‌عنوان یک روش جدید انتخاب متغیر جفت شده با روش ANN به‌عنوان مدل‌سازی غیرخطی قدرتمند (SCAD-ANN) برای پیش‌بینی مقادیر شاخص‌های بازداری (RI) ترکیبات آلی فرار (VOCs) متفاوت استفاده شد. VOC ها ترکیبات شیمیایی آلیفاتیک و معطر با وزن مولکولی و نقطه جوش کم هستند [۱۹۱]. منابع VOC شامل حلال‌ها، سوخت‌ها، رنگ‌ها، مواد شوینده، سیگار و مواد غذایی است. بیش‌تر VOC های تولید شده توسط صنایع

اغلب حاوی بنزن، تولوئن، زایلن، فوران و کلروفرم بوده که برای سلامت انسان مضر هستند [۱۹۲]. در این پژوهش، پس از محاسبه توصیف‌کننده‌های ساختاری برای مجموعه‌های مختلف VOCs و غربالگری متغیرها، توصیف‌کننده‌های معنی‌دار مطابق با λ_{min} استخراج شدند. با توجه به رابطه پیچیده و غیرخطی بین متغیرهای وابسته و مستقل، از ANN برای ایجاد مدل QSRR استفاده شد. عملکرد مدل‌های SCAD-ANN توسعه‌یافته با استفاده از پارامترهای آماری برای پیش‌بینی RI های مجموعه‌های آزمون و کل مجموعه داده‌ها مورد ارزیابی قرار گرفت. با به‌کارگیری SCAD-ANN پیشنهادی مشخص شد که روش مدل‌سازی ANN جفت شده با تکنیک انتخاب متغیر جریمه‌شده SCAD، مدل‌های QSRR دقیقی را برای پیش‌بینی شاخص‌های بازداری ترکیبات VOC ایجاد می‌کند.

۲- ۵- ۲ مجموعه داده‌ها

به‌منظور توسعه مدل‌های QSRR و بررسی کارایی روش انتخاب متغیر SCAD، از دو مجموعه داده متفاوت استفاده شد. اولین مجموعه شامل ۱۳۲ ترکیب آلی فرار متشکل از آلکان‌ها، آلکن‌ها، آمین‌ها، اترها، الکل‌ها، آلکیل بنزن و آلکیل هالیدها بود. شاخص بازداری ترکیبات مورد مطالعه مجموعه داده A با استفاده از دستگاه GC مدل ۴۳۹ Packard Becker (Delft, Netherlands) مجهز به دو آشکارساز هدایت حرارتی اندازه‌گیری شد. فاز ساکن C67 با نام شیمیایی ۱۹،۱۹-Diethyl-14,24 - ditridecylheptatricontane بود و داده‌های GC برای حدود ۱۳۲ ترکیب آلی فرار در C67 در دمای □ ۱۳۰ اندازه‌گیری شد [۱۹۳]. مجموعه داده B شامل ۵۲ ترکیب آلی فرار متشکل از پیرازین‌ها، پیریدین‌ها، فوران‌ها و غیره است. RI های ترکیبات مورد مطالعه مجموعه داده دوم با استفاده از دستگاه Varian CP-3800 GC و ستون DB-5ms (30m×0.25mm×0.25mm) استخراج شد [۱۹۴]. ساختار ترکیبات به یک فایل با فرمت سیستم ورودی خطی ورودی مولکولی ساده شده (SMILES) تبدیل شد. نام شیمیایی و فرمت SMILES ساختارهای مورد مطالعه همراه با مقادیر واقعی و پیش بینی شده RI مربوط به هر دو مجموعه داده به‌ترتیب در جدول ۲-۲۳

و جدول ۲-۲۴ خلاصه شده‌اند. مقادیر RI برای همه ترکیبات در مجموعه خود، تحت شرایط یکسان اندازه‌گیری شده است. لازم به ذکر است که دو مجموعه داده، همگنی لازم برای ترکیب شدن به‌عنوان یک مجموعه داده را نداشتند، بنابراین به همین دلیل، به‌عنوان دو مجموعه داده مجزا برای مطالعات بیش‌تر مورد استفاده قرار گرفتند. مقادیر RI به‌عنوان متغیر وابسته در کل مطالعه در نظر گرفته شد. مجموعه داده‌ها با استفاده از نرم‌افزار R و اجرای الگوریتم KS به مجموعه‌های آموزش، ارزیابی و آزمون تقسیم شدند. در ادامه توضیحات مربوط به ساخت مدل QSRR، مجموعه داده اول شامل ۱۳۲ ترکیب آلی فرار به نام مجموعه داده A و مجموعه داده دوم شامل ۵۲ ترکیب آلی فرار با نام مجموعه داده B نامگذاری شده است.

جدول ۲-۲۳ ساختار شیمیایی و SMILES مربوط به ترکیبات مورد مطالعه مجموعه داده‌های A به همراه شاخص بازداری

ردیف	ساختار شیمیایی	SMILES	RI واقعی	RI پیش بینی شده
۱[t]	1-Butanol	C(CCC)O	۰/۴۲۲۱	۰/۴۱۴۳
۲	2-Methyl-2-propanol	CC(C)(O)C	۰/۳۳۴۱	۰/۳۳۲۳
۳[v]	1-Pentanol	C(CCCC)O	۰/۴۹۷۲	۰/۴۹۰۱
۴[t]	2-Methyl-2-butanol	CC(CC)(O)C	۰/۴۲۵۹	۰/۴۱۸۳
۵	1-Hexanol	C(CCCCC)O	۰/۵۷۰۳	۰/۵۶۷۹
۶[t]	Cyclohexanol	C1(CCCCC1)O	۰/۶۰۶۲	۰/۶۲۴۱
۷	2-Methyl-2-pentanol	CC(CCC)(O)C	۰/۴۹۲۳	۰/۴۹۳۳
۸	1-Heptanol	C(CCCCCC)O	۰/۶۴۳	۰/۶۵۱۸
۹[v]	2-Methyl-2-hexanol	CC(CCCC)(O)C	۰/۵۶۱۱	۰/۵۶۹۵
۱۰	2-Butanol	C[C@@H](CC)O	۰/۳۸۸۵	۰/۳۸۵۵
۱۱	2-Pentanol	C[C@@H](CCC)O	۰/۴۵۹۴	۰/۴۶۳۹
۱۲	2-Hexanol	C[C@@H](CCCC)O	۰/۵۳۱۳	۰/۵۳۵۲
۱۳	2-Heptanol	C[C@@H](CCCCC)O	۰/۶۰۲۷	۰/۶۰۳۷
۱۴[t]	2-Phenylethanol	C(Cc1ccccc1)O	۰/۷۵۰۴	۰/۷۴۰۴
۱۵[v]	Benzyl alcohol	C(c1ccccc1)O	۰/۶۹۲۸	۰/۶۵۵۵
۱۶	Pentanal	C(=O)CCCC	۰/۴۶۰۵	۰/۴۶۸۷
۱۷	Hexanal	C(=O)CCCCC	۰/۵۳۳۶	۰/۵۳۳۵
۱۸[v]	2-Butanone	CC(=O)CC	۰/۳۸۰۸	۰/۳۷۵۴
۱۹[v]	2-Pentanone	CC(=O)CCC	۰/۴۴۸۱	۰/۴۴۹۴
۲۰	Cyclopentanone	C1(=O)CCCC1	۰/۵۲۷	۰/۵۰۶۵
۲۱	2-Hexanone	CC(=O)CCCC	۰/۵۲۰۶	۰/۵۰۳۵
۲۲[t]	Cyclohexanone	C1(=O)CCCCC1	۰/۶۰۸	۰/۵۸۵
۲۳	2-Heptanone	CC(=O)CCCCC	۰/۵۹۲۶	۰/۵۶۳۳
۲۴	Dipropylether	O(CCC)CCC	۰/۴۶۸۱	۰/۴۵۴۱
۲۵[v]	Dibutyl ether	O(CCCC)CCCC	۰/۶۱	۰/۶۱۰۱
۲۶	Tetrahydrofuran	O1CCCC1	۰/۴۳۴۱	۰/۳۹۹۱
۲۷[t]	1,4-Dioxane	O1CCOCC1	۰/۴۷۶۶	۰/۴۷۳۵
۲۸[t]	Methyl phenyl ether	O(c1ccccc1)C	۰/۶۴۶۶	۰/۵۷۷۵
۲۹	Phenetole	CCOc1ccccc1	۰/۶۹۶۲	۰/۶۶۱۸
۳۰	Nitroethane	CCN(=O)=O	۰/۴۰۲۵	۰/۴۱۵۲
۳۱[t]	1-Nitropropane	C(CC)N(=O)=O	۰/۴۷۱۱	۰/۴۸۴۳
۳۲[v]	1-Nitrobutane	C(CCC)N(=O)=O	۰/۵۴۵۱	۰/۵۵۹۵
۳۳	1-Nitropentane	C(CCCC)N(=O)=O	۰/۶۱۷۷	۰/۵۹۷۵
۳۴	1-Nitrobenzene	c1(ccccc1)N(=O)=O	۰/۷۴۶۶	۰/۷۵۶۳
۳۵	1-Cyanoethane	C(C)C#N	۰/۳۴۵۹	۰/۳۲۴۵
۳۶	1-Cyanopropane	C(CC)C#N	۰/۴۱۷۴	۰/۴۰۶۷
۳۷	1-Cyanobutane	C(CCC)C#N	۰/۴۹۴	۰/۴۶۵۸
۳۸	1-Cyanopentane	C(CCCC)C#N	۰/۵۶۶۹	۰/۵۲۷۵
۳۹	Pyridine	c1ccccn1	۰/۵۱۳۸	۰/۵۱۳۹
۴۰	2-Picoline	c1(ccccn1)C	۰/۵۷۰۵	۰/۵۶۵۷
۴۱[v]	3-Picoline	c1c(cccn1)C	۰/۶۰۰۳	۰/۵۷۵۸
۴۲	4-Picoline	c1cc(ccn1)C	۰/۵۹۸۹	۰/۵۷۴۶
۴۳	2,3-Lutidine	c1(c(cccn1)C)C	۰/۶۶۴۷	۰/۶۲۲۵
۴۴	2,4-Lutidine	c1(cc(ccn1)C)C	۰/۶۵۳۹	۰/۶۱۴۸
۴۵[v]	2,5-Lutidine	c1(ccc(cn1)C)C	۰/۶۵۳۹	۰/۶۳۱۴
۴۶	2,6-Lutidine	c1(cccc(n1)C)C	۰/۶۲۰۹	۰/۶۰۸۴
۴۷[t]	3,4-Lutidine	c1c(c(ccn1)C)C	۰/۷۰۰۷	۰/۶۴۰۶
۴۸	3,5-Lutidine	c1c(cc(cn1)C)C	۰/۶۸۵۴	۰/۶۲۸۲
۴۹	3-Chloropyridine	c1c(cccn1)Cl	۰/۶۲۹۳	۰/۶۲۳۹

ادامه جدول ۲-۲۳

ردیف	ساختار شیمیایی	SMILES	RI واقعی	RI پیش بینی شده
۵۰	1-Acetoxypropane	C(CC)OC(=O)C	۰/۴۵۴۲	۰/۴۷۹
۵۱	1-Acetoxybutane (butyl acetate)	C(CCC)OC(=O)C	۰/۵۲۷۳	۰/۵۴۱
۵۲	1-Acetoxypentane (pentyl acetate)	C(CCCC)OC(=O)C	۰/۵۹۹۴	۰/۶۱۷۱
۵۳	1,1,1-Trifluorooctane	C(CCCCCCC)(F)(F)F	۰/۵۱۸۹	۰/۵۳۴۹
۵۴[v]	Fluorobenzene	c1(ccccc1)F	۰/۴۷۴۲	۰/۴۸۴۴
۵۵	Hexafluorobenzene	c1(c(c(c(c(c1F)F)F)F)F)F	۰/۳۹۲۸	۰/۳۸۹۸
۵۶	Trifluoromethylbenzene	c1(ccccc1)C(F)(F)F	۰/۴۶۹۱	۰/۴۷۳۴
۵۷	Dichloromethane	C(Cl)Cl	۰/۳۵۵	۰/۳۵۴۹
۵۸	Trichloromethane	C(Cl)(Cl)Cl	۰/۴۲۹۷	۰/۴۵۹۵
۵۹[t]	Tetrachloromethane	C(Cl)(Cl)(Cl)Cl	۰/۴۸۲۳	۰/۵۶۲۳
۶۰	1-Chlorobutane	C(CCC)Cl	۰/۴۵۵۹	۰/۴۴۲۶
۶۱[t]	1-Chloropentane	C(CCCC)Cl	۰/۵۲۹۱	۰/۵۲۰۱
۶۲	1-Chlorohexane	C(CCCCC)Cl	۰/۶۰۱۷	۰/۵۷۱۴
۶۳[v]	Chlorobenzene	c1(ccccc1)Cl	۰/۶۱۵۳	۰/۵۹۲
۶۴	1-Bromopropane	c1(ccccc1)Cl	۰/۴۴۷۵	۰/۵۸۰۸
۶۵	1-Bromobutane	C(CCC)Br	۰/۵۲۱۳	۰/۴۵۷۵
۶۶[v]	1-Bromopentane	C(CCCC)Br	۰/۵۹۴۶	۰/۵۴۷۲
۶۷[v]	Bromobenzene	c1(ccccc1)Br	۰/۶۸۲۸	۰/۶۶۲۹
۶۸[t]	1-Butanethiol	C(CCC)S	۰/۵۱۰۶	۰/۴۷۵
۶۹[v]	1-Pentanethiol	C(CCCC)S	۰/۵۸۳۷	۰/۵۶۳۳
۷۰	n-Hexanethiol	C(CCCCC)S	۰/۶۵۶۴	۰/۶۴۰۲
۷۱[t]	Thiophene	c1cccs1	۰/۴۸۵۹	۰/۵۱۱
۷۲[v]	1-Hexene	C=CCCCC	۰/۴۱۷۴	۰/۴۴۵
۷۳[t]	Cyclohexene	C1=CCCCC1	۰/۵۰۳	۰/۵۰۹۴
۷۴	1,4-Cyclohexadiene	C1=CCC=CC1	۰/۵۱۶۱	۰/۵۰۵۴
۷۵	1,3-Cyclohexadiene	C1=CC=CCC1	۰/۴۹۱۳	۰/۵۰۳۸
۷۶	1-Heptene	C=CCCCCC	۰/۴۸۹۱	۰/۵۰۲۲
۷۶[v]	1-Octene	C=CCCCCCC	۰/۵۵۹۹	۰/۵۷۳۹
۷۸[t]	1-Nonene	C=CCCCCCCC	۰/۶۳۱۹	۰/۶۴۵۱
۷۹[v]	1-Decene	C=CCCCCCCCC	۰/۷۰۳۱	۰/۷۰۹
۸۰	1-Pentyne	C#CCCC	۰/۳۴۴	۰/۳۹
۸۱	1-Hexyne	C#CCCCC	۰/۴۱۹۱	۰/۴۴۱۸
۸۲[t]	2-Hexyne	CC#CCCC	۰/۴۵۹۶	۰/۴۶۰۹
۸۳[v]	3-Hexyne	CCC#CCC	۰/۴۴۳۱	۰/۴۴۵۱
۸۴[v]	1-Heptyne	C#CCCCCC	۰/۴۹۱۳	۰/۵۰۶۳
۸۵	1-Octyne	C#CCCCCCC	۰/۵۶۲۴	۰/۵۷۱۷
۸۶[v]	4-Octyne	CCCC#CCCC	۰/۵۷۸۷	۰/۵۹۲۳
۸۷	1-Nonyne	C#CCCCCCCC	۰/۶۳۴۱	۰/۶۴۵۵
۸۸	1-Decyne	C#CCCCCCCCC	۰/۷۰۵۵	۰/۷۱۰۵
۸۹[v]	Benzene	c1ccccc1	۰/۴۸۱۱	۰/۴۷۷۳
۹۰	Toluene	c1(ccccc1)C	۰/۵۵۸۶	۰/۵۳۶۸
۹۱[v]	Ethylbenzene	c1(ccccc1)CC	۰/۶۲۲۹	۰/۶۱۳
۹۲[v]	Naphthalene	c1cccc2ccccc12	۰/۸۶۳	۰/۷۷۲۵

ادامه جدول ۲-۲۳

ردیف	ساختار شیمیایی	SMILES	RI واقعی	RI پیش بینی شده
۹۳[t]	Azulene	c1ccc2ccccccc12	۰/۹۴۲۸	۰/۷۸۹۷
۹۴	Pentane	CCCCC	۰/۳۵۷۱	۰/۳۷۴۳
۹۵	Cyclopentane	C1CCCC1	۰/۴۱۹۹	۰/۴۳
۹۶	2,2-Dimethylbutane	CC(CC)(C)C	۰/۳۸۶۵	۰/۳۹۱۳
۹۷	2,3-Dimethylbutane	CC(C(C)C)C	۰/۴۰۹	۰/۴۰۰۲
۹۸[v]	Hexane	CCCCCC	۰/۴۲۸۶	۰/۴۵۱۱
۹۹[t]	Cyclohexane	C1CCCCC1	۰/۴۹۵۱	۰/۵۱۱۷
۱۰۰[v]	2,2-Dimethylpentane	CC(CCC)(C)C	۰/۴۴۸۹	۰/۴۵۳۸
۱۰۱	2,3-Dimethylpentane	CC([C@@H](CC)C)C	۰/۴۸۵۲	۰/۴۷۹۳
۱۰۲	2,4-Dimethylpentane	CC(CC(C)C)C	۰/۴۴۸۹	۰/۴۷۰۶
۱۰۳	2,2,3-Trimethylbutane	CC(C(C)C)(C)C	۰/۴۶۴۶	۰/۴۳۳۷
۱۰۴[t]	Heptane	CCCCCCC	۰/۵	۰/۵۱۶۸
۱۰۵	Cycloheptane	C1CCCCCC1	۰/۵۹۷	۰/۵۹۵۵
۱۰۶	Methylcyclohexane	C1(CCCCC1)C	۰/۵۴۱۱	۰/۵۷۰۵
۱۰۷	2,3-Dimethylhexane	CC([C@@H](CCC)C)C	۰/۵۴۷	۰/۵۳۴
۱۰۸[t]	2,4-Dimetylhexane	CC(C[C@@H](CC)C)C	۰/۵۲۳۷	۰/۵۲۲۹
۱۰۹[t]	3,4-Dimetylhexane	CC[C@@H]([C@@H](CC)C)C	۰/۵۵۶۶	۰/۵۳۴۸
۱۱۰	2,2,4-Trimethylpentane	CC(CC(C)C)(C)C	۰/۴۹۷۵	۰/۴۸۷۲
۱۱۱[v]	2,3,4-Trimethylpentane	CC(C(C(C)C)C)C	۰/۵۴۷	۰/۴۹۸۶
۱۱۲	cis-1,2-Dimethylcyclohexane	[C@@H]1([C@H](CCCC1)C)C	۰/۶۱۸۹	۰/۶۲۲۶
۱۱۳	trans-1,2-Dimethylcyclohexane	[C@@H]1([C@@H](CCCC1)C)C	۰/۵۹۷۶	۰/۶۲۲۵
۱۱۴	cis-1,4-Dimethylcyclohexane	[C@@H]1(CC[C@H](CC1)C)C	۰/۵۹۷۴	۰/۶۲۱۴
۱۱۵[t]	trans-1,4-Dimethylcyclohexane	[C@@H]1(CC[C@@H](CC1)C)C	۰/۵۸۲	۰/۶۱۸۹
۱۱۶	Octane	CCCCCCCC	۰/۵۷۱۴	۰/۵۸۰۹
۱۱۷	Cyclooctane	C1CCCCCCC1	۰/۶۸۷۹	۰/۶۶۶۷
۱۱۸[t]	Nonane	CCCCCCCCC	۰/۶۴۲۹	۰/۶۵۶۳
۱۱۹	Decane	CCCCCCCCCC	۰/۷۱۴۳	۰/۷۱۱۹
۱۲۰	Cyclodecane	C1CCCCCCCCC1	۰/۸۴۰۹	۰/۸۲۱۱
۱۲۱[t]	cis-Hydrindane	C1CCc2ccccc12	۰/۷۳۷۱	۰/۷۲۰۱
۱۲۲	trans-Hydrindane	C1CCc2ccccc12	۰/۷۱۳۶	۰/۷۱۰۵
۱۲۳	cis-Decalin	C1CCC[C@H]2CCCC[C@@H]12	۰/۸۲۱۸	۰/۸۲۳۴
۱۲۴	trans-Decalin	C1CCC[C@H]2CCCC[C@H]12	۰/۷۹۲۲	۰/۸۱۹۶
۱۲۵[t]	Adamantane	[C@@H]12C[C@@H]3C[C@@H](C[C@H](C1)C3)C2	۰/۸۰۹۷	۰/۸۱۶۶
۱۲۶[t]	Undecane	CCCCCCCCCCC	۰/۷۸۵۷	۰/۷۹۴۶
۱۲۷	Dodecane	CCCCCCCCCCCC	۰/۸۵۷۱	۰/۸۴۸۶
۱۲۸	Tridecane	CCCCCCCCCCCCC	۰/۹۲۸۶	۰/۹۲۷۹
۱۲۹	Tetradecane	CCCCCCCCCCCCCC	۱/۰۰۰	۰/۹۹۵۹
۱۳۰[t]	Tetramethylsilane	[Si](C)(C)(C)C	۰/۳۰۳۹	۰/۳۳۰۹
۱۳۱	Hexamethyldisilane	[Si]([Si](C)(C)C)(C)(C)C	۰/۴۹۱۵	۰/۴۹۰۴
۱۳۲	Tetramethylthin	[Sn](C)(C)(C)C	۰/۴۳۰۴	۰/۴۲۹۹

جدول ۲-۲۴ ساختار شیمیایی و SMILES مربوط به ترکیبات مورد مطالعه مجموعه داده‌های B به همراه شاخص بازداری

ردیف	ساختار شیمیایی	SMILES	RI واقعی	RI پیش بینی شده
۱[v]	2-Methyl-1H-pyrrole	c1(c(cc(cc1C(C)(C)C)C)C(C)(C)C)O	۸۲۱	۸۵۶
۲[v]	2-Ethyl-2,5-dimethylpyrazine	CCCCCCCCCCCCCC	۱۰۹۱	۱۰۸۷
۳	Vinylpyrazine	C([C@H](C(=O)OC[C@H](CC(C)(C)C)C)C)O	۹۶۵	۹۴۷
۴	2-Furanmethanol	C(=O)/C=C/CCCCCCCC	۹۱۰	۸۵۳
۵[t]	Pyrazine	C(=O)/C=C/C=C/CCCCC	۷۵۵	۷۴۰
۶	Benzaldehyde	C(=O)/C=C\C=C/CCCCC	۱۰۱۴	۹۸۲
۷	Menthol	CCCCCCCCCCCCC	۱۱۹۴	۱۱۹۰
۸[v]	Methylvinylpyrazine	c1cccc2ccccc12	۱۰۲۹	۱۰۳۸
۹	2-Pentylthiophene	CCCCCCCCCCCC	۱۲۲۴	۱۱۷۰
۱۰	2-Ethyl-5-methylpyrazine	[C@H]1([C@H](CC[C@H](C1)C)C(C)C)O	۱۰۰۱	۱۰۱۵
۱۱[t]	Nonanal	C(c1ccco1)n1cccc1	۱۰۸۵	۱۱۱۳
۱۲	(Z,Z)2,4-decadienal	c1(cccs1)CCCCC	۱۲۲۶	۱۳۰۹
۱۳	Tridecane	c1(ccccc1)CCCCC	۱۲۵۸	۱۲۹۹
۱۴[t]	2-Pentylfuran	C(=O)CCCCCCCC	۱۱۲۹	۹۹۶
۱۵	2,5-Diethylpyrazine	CCCCCCCCCCC	۱۱۰۴	۱۰۹۱
۱۶[v]	1-(2-furanmethyl)-1H-pyrrole	c1(cnc(cn1)CC)CC	۱۲۰۷	۱۱۸۲
۱۷[v]	Dodecane	c1(c(nc(cn1)C)CC)C	۱۱۹۲	۱۱۹۹
۱۸	3-Methylphenol	c1(c(nc(cn1)C)CC)C	۱۰۱۶	۱۰۸۰
۱۹[t]	3-Ethyl-2,5-dimethylpyrazine	Cc1cc(O)ccc1	۱۰۸۴	۱۰۸۵
۲۰[t]	Tetradecane	c1(ccccc1)CCCC	۱۳۳۳	۱۳۹۹
۲۱	2,4-decadienal (E,E)	CC(=O)/C=C/CCCC	۱۲۴۹	۱۳۱۹
۲۲	2-Undecenal	CC1=CC[C@@H](C(=C)C)CC1	۱۲۷۶	۱۳۶۵
۲۳	2-Butenal	c1(c(nccn1)C)C=C	۷۰۵	۶۴۸
۲۴	Butylated hydroxytoluene	C([C@H](CCCC)CC)O	۱۴۹۰	۱۵۱۵
۲۵[t]	Propanoic acid, 2-methyl-, 3-hydroxy-2,4,4-trimethylpentyl esters	c1(cncc(n1)C=C)C	۱۴۸۰	۱۳۸۴
۲۶[v]	Naphthalene	c1(cnc(cn1)C)CC	۱۲۰۴	۱۲۱۸
۲۷	2-Methylpyrazine	c1(cncc(n1)C)CC	۸۵۵	۸۲۸
۲۸[t]	Pentyl-benzene	CCCCCCCCCC	۱۳۳۳	۱۱۶۸
۲۹	Pentanol	o1c(ccc1)CCCCC	۸۰۱	۷۷۰
۳۰[v]	Furfural	C(=C\C(=O)CCCC)/C	۸۶۷	۸۳۸
۳۱	Methyl heptenone	S(SC)SC	۱۰۱۶	۹۹۱
۳۲[t]	2,3-Pentanedione	C=C[C@H](CCCCC)O	۸۰۶	۶۹۱
۳۳[t]	Decane	C(=O)c1ccccc1	۱۰۳۴	۱۰۰۰
۳۴	Ethylpyrazine	C(=O)/C=C/CCCC	۹۳۷	۹۲۹
۳۵	Undecane	c1(cnccn1)C=C	۱۱۱۵	۱۰۹۹
۳۶[t]	Butyl-benzene	c1(cnccn1)CC	۱۱۳۰	۱۰۶۵
۳۷	Limonene	c1(cnc(cn1)C)C	۱۰۴۰	۱۰۴۱
۳۸	3-Octen-2-one	C(=O)CCCCCC	۹۸۰	۱۰۴۳
۳۹	2-Ethyl-6-methyl pyrazine	CC(=O)CCCCC	۹۹۱	۱۰۰۹
۴۰[t]	2,5-Dimethylpyrazine	Cc1ccc(cc1)C	۹۰۸	۹۲۶
۴۱	Pyrrole	C(CCCCC)O	۷۴۷	۷۵۲
۴۲	2-Ethyl-1-hexanol	o1c(ccc1)CO	۱۰۱۴	۱۰۳۲
۴۳	Hexanal	[nH]1c(ccc1)C	۸۴۲	۸۰۱
۴۴	2-Heptanone	C(=O)c1ccco1	۸۸۳	۸۸۹
۴۵[v]	1-Hexanol	c1(cnccn1)C	۸۷۳	۸۷۲
۴۶	1-Octen-3-ol	s1cnc(c1)C	۱۰۰۸	۹۸۲
۴۷	2-Methyl-6-vinylpyrazine	C(=O)CCCCC	۱۰۴۸	۱۰۳۲
۴۸	Heptanal	C(CCCC)O	۹۲۵	۹۰۷
۴۹	2-Methyl-1H-pyrrole	[nH]1cccc1	۸۵۸	۸۳۶
۵۰[v]	Xylene	c1cnccn1	۹۲۵	۸۸۱
۵۱[v]	2(E)-Heptenal	CC(=O)C(=O)CC	۹۶۸	۹۶۸
۵۲	Dimethyl trisulfide	C(=O)/C=C/C	۹۶۷	۹۹۰

۲- ۵- ۳ رسم و بهینه‌سازی ساختار ترکیبات آلی فرار مجموعه داده‌های متفاوت

ساختارهای شیمیایی همه ترکیبات آلی فرار با استفاده از نرم‌افزار هایپرکم ترسیم شد. فرآیند بهینه‌سازی مطابق با روش کار بخش ۱- ۵- ۳ و با به‌کارگیری روش نیمه تجربی AM1 انجام شد و تا رسیدن به حداقل انرژی به ۰/۰۰۱ ادامه یافت. ساختارهای بهینه شده با پسوند hin.* ذخیره شدند.

۲- ۵- ۴ استخراج توصیف‌کننده‌های ساختاری

ساختارهای بهینه ترکیبات آلی فرار هر دو مجموعه داده به‌طور مجزا در نرم افزار دراگون فراخوانی شدند و سپس برای هر ترکیب به تعداد ۳۲۲۴ توصیف‌کننده مولکولی در ۲۲ دسته متفاوت محاسبه شدند.

۲- ۵- ۵ پیش‌پردازش و انتخاب توصیف‌کننده‌های مؤثر

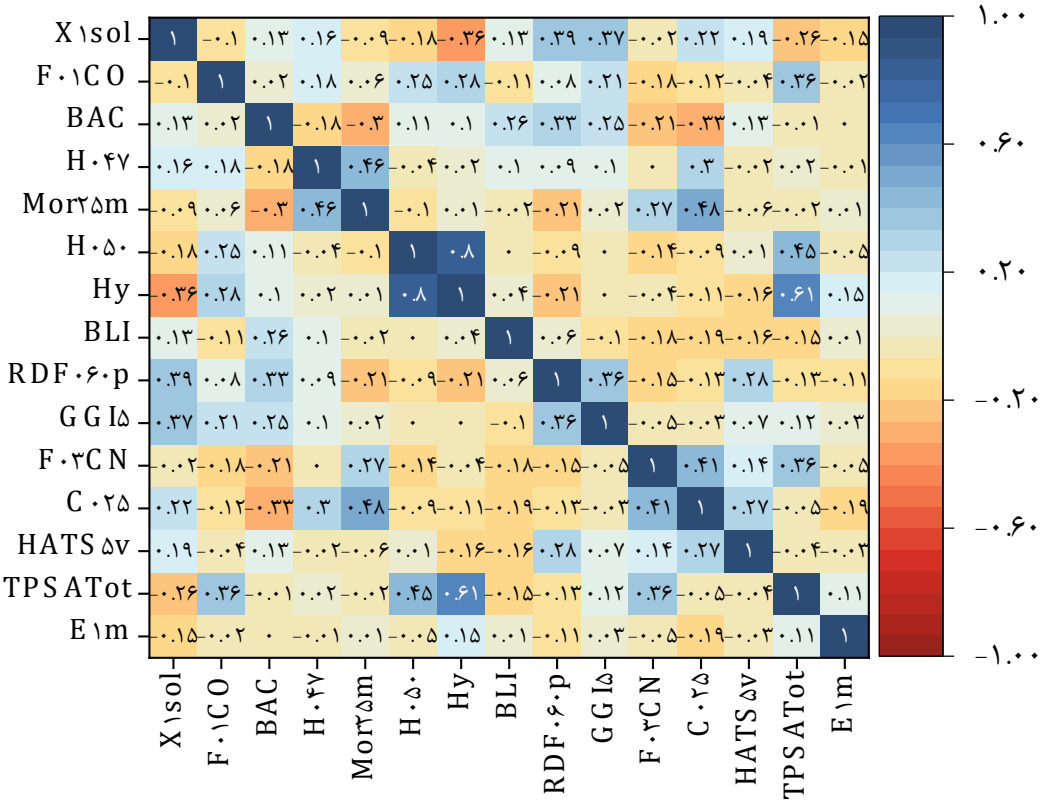
به‌منظور کاهش ابعاد داده‌ها و افزایش تفسیرپذیری مدل QSRR، باید تعداد توصیف کننده‌ها کاهش یابد. برای تحقق این هدف، یک استراتژی دو مرحله‌ای متشکل از مرحله غربالگری و انتخاب توصیف کننده انجام شد. از این‌رو، مقادیر ثابت و نسبتاً ثابت (متغیرهای با واریانس کم‌تر از ۰/۰۰۱) در مرحله غربالگری و با استفاده از اجرای بسته نرم افزاری caret در نرم افزار R حذف شدند [۱۶۳, ۱۹۵]. همچنین از بین دو توصیف کننده با همبستگی بالای ۰/۹، توصیف کننده‌ای که کم‌ترین ارتباط را با پاسخ داشت نیز حذف شد و توصیف کننده با بیش‌ترین همبستگی با متغیر پاسخ نگه داشته شد. پس از غربالگری، توصیف‌کننده‌های باقی‌مانده در یک ماتریس مرتب شدند، به‌طوری‌که توصیف کننده‌ها و RI به‌ترتیب به‌عنوان متغیرهای مستقل و وابسته در نظر گرفته شدند. داده‌های مجموعه آزمون از مجموعه داده‌ها حذف شدند. سپس، روش SCAD با استفاده از روش ارزیابی تقاطعی ده فولد (۱۰-fold-CV) موجود در بسته نرم افزاری ncvreg در نرم افزار R، روی مجموعه ارزیابی و آموزش اجرا شد [۱۶۳, ۱۹۵]. توصیف‌کننده‌های مربوط به λ_{min} (λ با حداقل خطای ارزیابی تقاطعی (CV)) شناسایی شد و توصیف‌کننده‌های مربوط به λ_{min} به‌عنوان مؤثرترین توصیف کننده‌ها در نظر گرفته شدند. پارامتر α در محاسبات SCAD، بر اساس مطالعه

بیزی فن و لی روی ۳/۷ تنظیم شد [۳۱]. بنابراین به تعداد ۱۵ و ۸ توصیف کننده، به‌عنوان مؤثرترین توصیف کننده‌ها برای مجموعه داده‌های A و B به‌ترتیب، به‌دست آمد و نام و نوع توصیف کننده‌های منتخب روش SCAD برای هر دو مجموعه داده‌ها در جدول ۲-۲۵ خلاصه شدند.

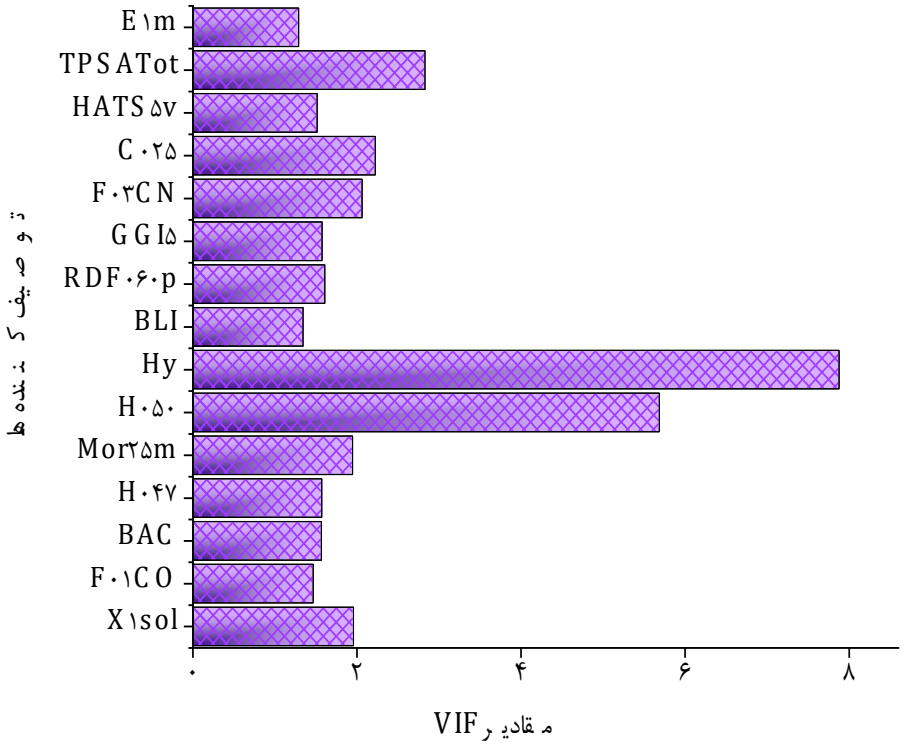
به‌منظور ارزیابی دقیق آماری توصیف‌کننده‌های منتخب روش SCAD، احتمال وجود همبستگی و هم‌خطی چندگانه نیز با محاسبه ضریب همبستگی بین دو توصیف کننده و مقادیر افزایش تورم واریانس (VIF) مورد بررسی قرار گرفتند. به‌این منظور نمودار نقشه رنگی و VIF برای توصیف کننده‌های منتخب روش SCAD رسم شد و نتایج حاصل در شکل ۲-۳۷ تا شکل ۲-۴۰ نمایش داده شده‌اند. نمودار نقشه رنگی هر دو مجموعه داده A و B (شکل ۲-۳۷ و شکل ۲-۳۹) نشاندهنده عدم وجود همبستگی معنادار بین توصیف کننده‌های منتخب روش SCAD است. با هدف بررسی پدیده هم‌خطی، پارامتر VIF محاسبه شده برای توصیف‌کننده‌های منتخب (شکل ۲-۳۸ و شکل ۲-۴۰) نیز نشان می‌دهد که مقادیر VIF، از مقدار هشدار ۱۰ کم‌تر هستند. بنابراین هم‌خطی شدید نیز بین توصیف‌کننده‌های منتخب روش SCAD هر دو مجموعه داده A و B وجود ندارد [۱۳۸, ۱۳۹].

جدول ۲-۲۵ توصیف‌کننده‌های منتخب SCAD برای مجموعه داده‌های A و B

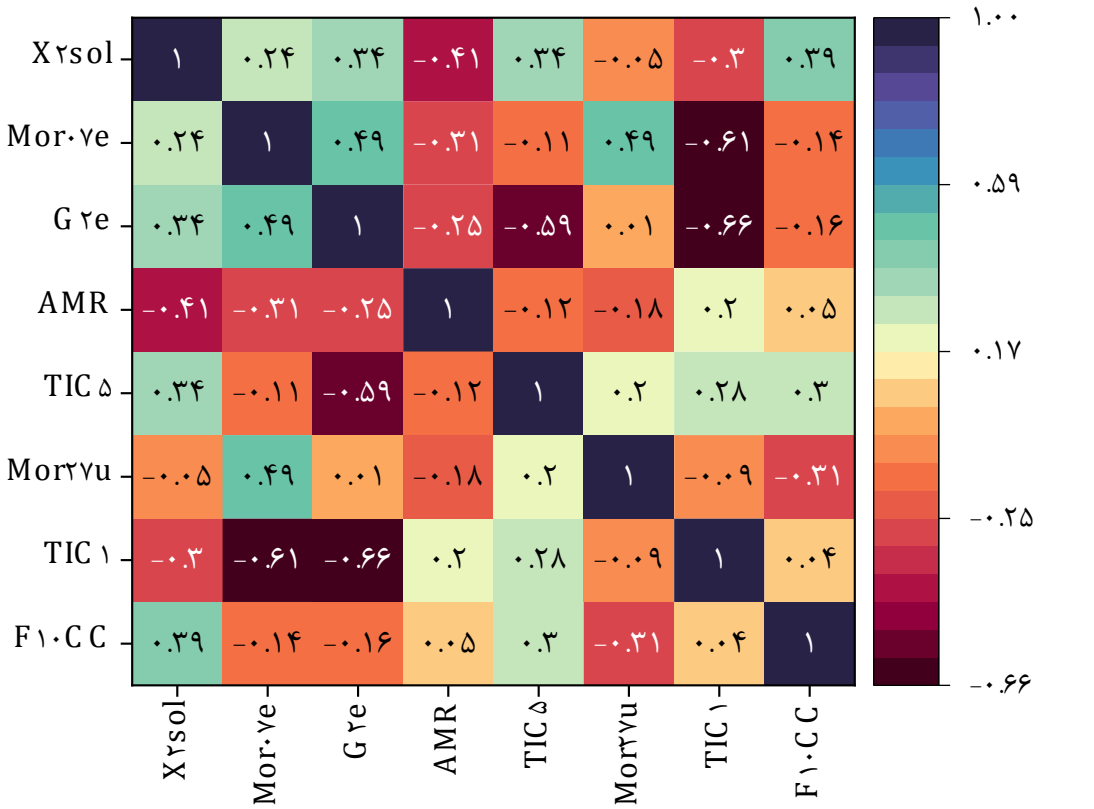
مجموعه داده‌ها	ردیف	نماد	معنا	طبقه‌بندی	اهمیت
A	۱	X1sol	solvation connectivity index of order 1	Connectivity indices	۱۶/۵
	۲	F01CO	Frequency of C - O at topological distance 1	2D Frequency finferprints	۹/۷۶
	۳	BAC	Balaban centric index	Topological indices	۸/۴۶
	۴	H-047	H attached to C1(sp3)/C0(sp2)	Atom-centred fragments	۸/۱
	۵	Mor25m	signal 25 / weighted by mass	3D-MoRSE descriptors	۷/۰۳
	۶	H-050	H attached to a heteroatom	Atom-centred fragments	۶/۷۱
	۷	Hy	hydrophilic factor	Molecular properties	۶/۶۲
	۸	BLI	Kier benzene-likeliness index Topological indice	Topological indices	۶/۶۱
	۹	RDF060p	Radial Distribution Function - 060 / weighted by polarizability	RDF descriptors	۶/۰۱
	۱۰	GGI5	topological charge index of order 5	2D autocorrelations	۵/۸۷
	۱۱	F03CN	Radial Distribution Function - 030 / unweighted	2D Frequency finferprints	۵/۳۹
	۱۲	C025	R--CR--R	Atom-centred fragments	۴/۹۴
	۱۳	HATS5v	leverage-weighted autocorrelation of lag 5 / weighted by van der Waals volume	GETAWAY descriptors	۴/۹
	۱۴	TPSA(Tot)	topological polar surface area using N,O,S,P polar contributions	Molecular properties	۴/۲۶
	۱۵	E1m	1st component accessibility directional WHIM index / weighted by mass	WHIM descriptors	۴/۲۲
B	۱	X2sol	solvation connectivity index of order 2	Connectivity indices	۱۲/۵۶
	۲	Mor07e	signal 07 / weighted by Sanderson electronegativity	3D-MoRSE descriptors	۵/۰۲
	۳	G2e	2nd component symmetry directional WHIM index / weighted by Sanderson electronegativity	WHIM descriptors	۴/۸۷
	۴	AMR	Ghose-Crippen molar refractivity	Molecular properties	۴/۵۵
	۵	TIC5	Total Information Content index (neighborhood symmetry of 5-order)	Information indices	۳/۷۱
	۶	Mor27u	signal 27 / unweighted	3D-MoRSE descriptors	۳/۱۳
	۷	TIC1	Total Information Content index (neighborhood symmetry of 1-order)	Information indices	۳/۱۱
	۸	F10CC	Frequency of C - C at topological distance 10	2D Frequency finferprints	۱/۵۵



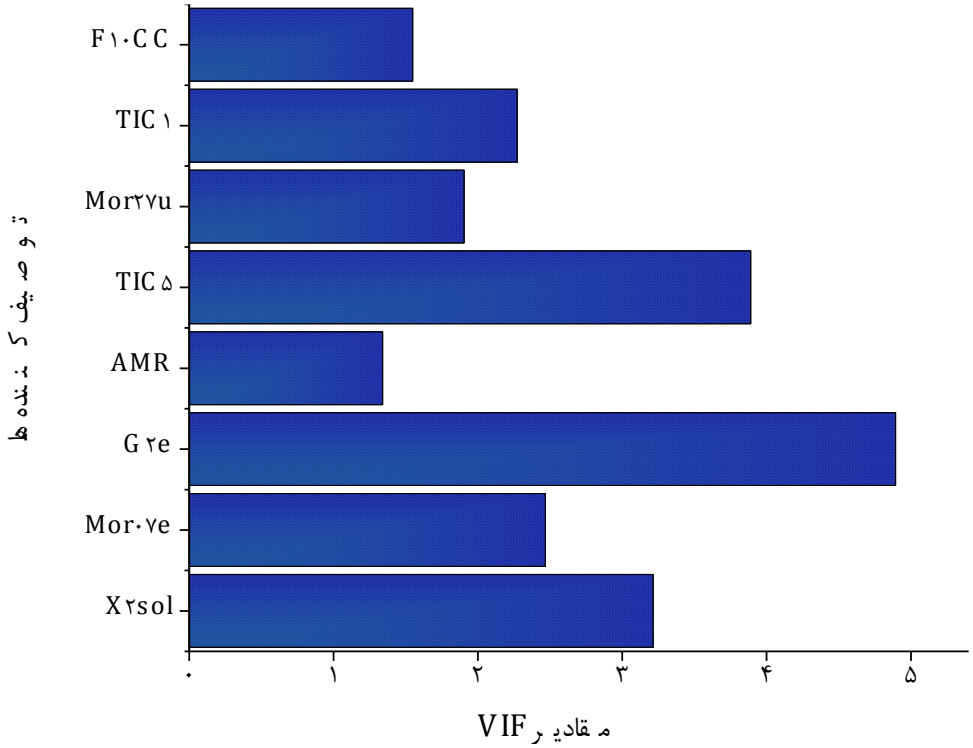
شکل ۲-۳۷ نمودار نقشه رنگی جهت نمایش همبستگی بین توصیف‌کننده‌های منتخب SCAD برای مجموعه داده A



شکل ۲-۳۸ نمودار مقادیر VIF توصیف‌کننده‌های منتخب SCAD برای مجموعه داده A



شکل ۲-۳۹ نمودار نقشه رنگی جهت نمایش همبستگی بین توصیف‌کننده‌های منتخب SCAD برای مجموعه داده B



شکل ۲-۴۰ نمودار مقادیر VIF توصیف‌کننده‌های منتخبSCAD برای مجموعه داده B

۲- ۵- ۶ مدل‌سازی شبکه عصبی با استفاده از توصیف‌کننده‌های منتخب SCAD

مدل‌سازی ANN برای ایجاد یک رابطه غیرخطی بین توصیف‌کننده‌های منتخب و RI استفاده شد. در این مطالعه از یک مدل شبکه عصبی پیشخور با الگوریتم آموزشی پس انتشار خطا استفاده شد. برای آموزش مدل ANN، بهینه‌سازی هم‌زمان تمام پارامترهای مهم مانند تعداد ورودی‌ها، تعداد گره‌ها در لایه پنهان، دورهای آموزشی و توابع آموزش و انتقال ضروری است [۴۰]. مدل‌های ANN حاوی یک لایه ورودی، یک لایه پنهان و یک لایه خروجی برای بهینه‌سازی پارامترهای ANN استفاده شدند. در بهینه‌سازی هم‌زمان پارامترهای ANN، مدل‌های ANN با معماری‌های مختلف با استفاده از دو الگوریتم آموزشی متنوع لونبرگ مارکوارت و تنظیم بایزین (توابع آموزشی trainlm و trainbr در جعبه ابزار متلب) و دو تابع انتقال متفاوت لگاریتم سیگموئیدی و تانژانت هایپربولیک سیگموئیدی (در جعبه ابزار برنامه متلب به ترتیب با توابع logsig و tansig شناخته می‌شوند) برای هردو مجموعه داده A و B به‌کار گرفته شد. برای بهینه‌سازی تعداد ورودی‌های شبکه عصبی مصنوعی، ابتدا تعداد نورون‌های لایه ورودی از ۲ تا کل توصیف‌کننده‌های منتخب SCAD در هر مجموعه داده تعریف شد. همان‌طور که در مطالعات قبلی گفته شد، تعداد زیر گروه‌هایی که می‌توان از تعداد توصیف‌کننده‌های منتخب تشکیل داد بسیار زیاد است و استفاده از همه این زیرگروه‌ها برای آموزش ANN، بسیار وقت‌گیر است. بنابراین از روش ANN برای یافتن ترکیب بهینه توصیف‌کننده‌های منتخب SCAD استفاده شد. مدل‌های ANN تولید شده با توجه به تعداد کل توصیف‌کننده‌های منتخب ساخته شدند و توابع انتقال و آموزش، گره و دور آموزشی بهینه به‌دست آمد. پس از آن، مقادیر همه توصیف کننده‌ها در محدوده تغییرات مربوط به خودشان، تصادفی شدند. هر بار مجموعه جدیدی از توصیف کننده‌ها با یک توصیف کننده دست‌کاری شده، در حضور سایر توصیف‌کننده‌های واقعی برای ساخت مدل‌های ANN با شرایط بهینه استفاده شد. مقادیر RI مربوط به مجموعه ارزیابی با استفاده از مدل بهینه SCAD-ANN پیش‌بینی شد. مقدار $RMSE_i$ مجموعه ارزیابی با استفاده از همه توصیف‌کننده‌های منتخب و توصیف کننده

i ام دست‌کاری شده، محاسبه شد. این فرآیند برای همه توصیف کننده‌ها تکرار شد تا در نهایت ۱۵ و ۸ مقدار RMSE با استفاده از توصیف‌کننده‌های هر دو مجموعه داده A و B به‌دست آمد. مقدار $RMSE_i$ بالاتر نشان می‌دهد که مدل ANN در حضور توصیف کننده i ام دست‌کاری شده (غیاب توصیف کننده با مقادیر واقعی خودش) دچار خطای بیش‌تری می‌شود. بنابراین توصیف کننده i ام با خطای بیش‌تر نسبت به سایر توصیف کننده‌ها از اهمیت بیش‌تری برخوردار است. مقادیر $RMSE_i$ توصیف کننده‌ها به‌طور کاهشی (از نظر مقادیر RMSE) مرتب شدند و به‌عنوان ورودی ANN تعریف شدند (پارامتر اهمیت در جدول ۲-۲۵). به‌طوری‌که شبکه عصبی مصنوعی ابتدا با دو توصیف کننده اول با اهمیت بیش‌تر آموزش دید و تا تعداد کل توصیف‌کننده‌های مهم بهینه‌سازی ادامه یافت. به‌این معنی‌که، در بهینه‌سازی شرایط ANN، تعداد ۱۴ زیر مجموعه برای مجموعه داده A و ۷ زیر مجموعه برای مجموعه داده B با توجه به اهمیت توصیف کننده‌ها به‌عنوان ورودی استفاده شدند. تعداد نورون‌های لایه پنهان در محدوده ۲ تا ۱۰ با گام ۱ و تعداد دورهای آموزشی از ۵ تا ۵۰ (با گام ۵) تغییر یافت. مدل‌های ANN طراحی شده، با استفاده از داده‌های مجموعه آموزشی، آموزش داده شدند. در نتیجه مدل‌های توسعه یافته برای پیش‌بینی مقادیر RI مجموعه ارزیابی استفاده شد. بهترین مدل ANN با توجه به کم‌ترین مقدار RMSE مجموعه ارزیابی انتخاب شد. بهترین معماری از مدل‌های شبکه عصبی مصنوعی با توابع آموزش و انتقال متفاوت با توجه به حداقل مقادیر RMSE در جدول ۲-۲۶ خلاصه شده است. نتایج به دست آمده نشان می‌دهد که مدل‌های شبکه عصبی مصنوعی SCAD-LM-ANN با معماری ۱-۲-۱۰ برای مجموعه داده A و SCAD-BR-ANN با معماری ۱-۴-۷ برای مجموعه داده B دارای RMSE مجموعه ارزیابی حداقل بوده و به‌عنوان مدل‌های برتر برای پیش بینی RI ترکیبات مورد مطالعه برگزیده شدند.

جدول ۲-۲۶ ساختارهای شبکه‌های توسعه یافته با توصیف‌کننده‌های منتخب SCAD با کمترین MSE مجموعه ارزیابی هردو مجموعه داده A و B

مجموعه داده‌ها	تعداد توصیف کننده	تابع آموزش	تابع انتقال	تعداد گره	تعداد دور آموزش	$RMSE_{validation}$	$R^2_{validation}$
مجموعه A	۱۵	تنظیم بایزین	لگاریتم-سیگموئید	۷	۱۵	۰/۰۳	۰/۹۴
	۱۰	**لونبرگ- مارکوارت**	**لگاریتم-سیگموئید**	**۲**	**۲۰**	**۰/۰۲**	**۰/۹۵**
	۱۵	تنظیم بایزین	تانژانت-سیگموئید	۵	۱۰	۰/۰۳	۰/۹۵
	۱۰	لونبرگ- مارکوارت	تانژانت-سیگموئید	۲	۲۰	۰/۰۳	۰/۹۶
مجموعه B	**۷**	**تنظیم بایزین**	**لگاریتم-سیگموئید**	**۴**	**۱۰**	**۲۲/۳۹**	**۰/۹۸**
	۸	لونبرگ- مارکوارت	لگاریتم-سیگموئید	۲	۱۵	۲۶/۱۵	۰/۹۷
	۶	تنظیم بایزین	تانژانت-سیگموئید	۸	۱۵	۲۴/۷۶	۰/۹۷
	۶	لونبرگ- مارکوارت	تانژانت-سیگموئید	۲	۲۰	۳۰/۴	۰/۹۷

علاوه بر این، کارایی روش SCAD به‌عنوان روش انتخاب متغیر جریمه‌ای با روش انتخاب متغیر کلاسیک SR مقایسه شد. در این راستا، متغیرهای انتخاب شده با روش‌های SR برای هر دو مجموعه داده به‌طور مجزا و با چینش شبکه عصبی به‌عنوان ورودی ANN تعریف شدند. پس از آموزش و بهینه‌سازی پارامترهای شبکه عصبی، مقادیر RI داده‌های مجموعه آزمون با استفاده از مدل ANN بهینه (SR-LM-ANN ۱-۲-۱۰ برای مجموعه داده A و SR-BR-ANN ۱-۸-۷ برای مجموعه داده B) پیش‌بینی شدند.

۲- ۵- ۷ ارزیابی مدل SCAD-ANN

قدرت پیش‌بینی مدل‌های پیشنهادی SCAD-ANN با استفاده از رویکردهای مختلف ارزیابی شد. از جمله می‌توان به پیش‌بینی مقادیر RI مجموعه آزمون با استفاده از مدل‌های بهینه، پیش‌بینی RI کل

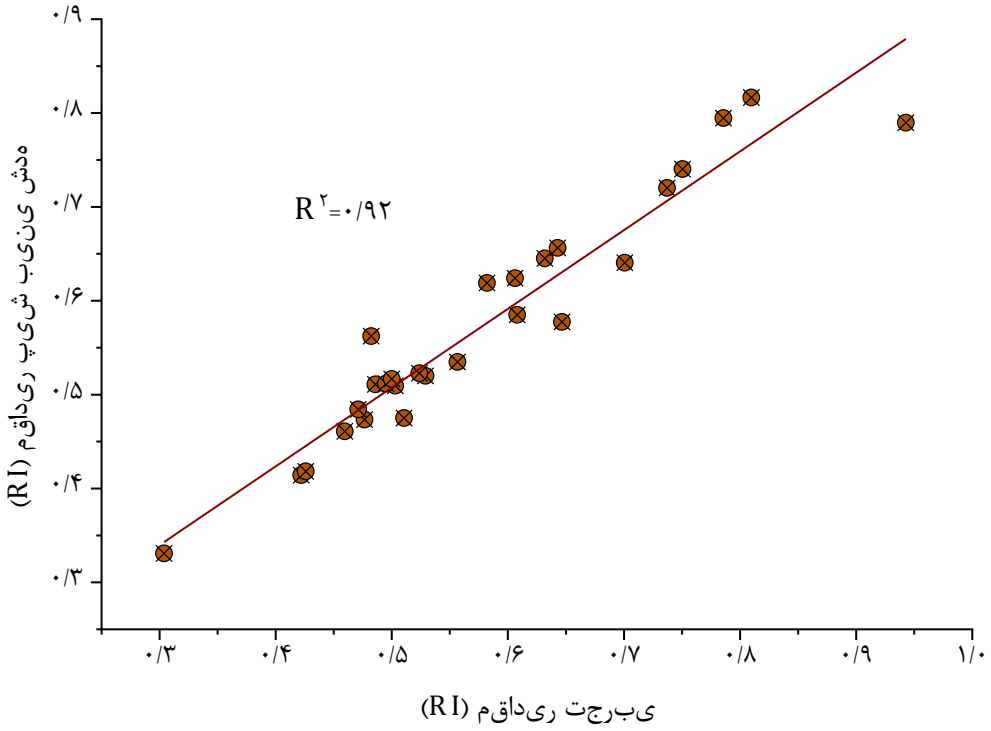
داده‌ها با استفاده از تکنیک LOO، محاسبه پارامترهای آماری، آزمون‌های پر کاربرد دامنه کاربرد و Y-تصادفی اشاره کرد.

۲- ۵- ۷- ۱ ارزیابی مدل SCAD-ANN با استفاده از پیش بینی مجموعه داده‌های آزمون

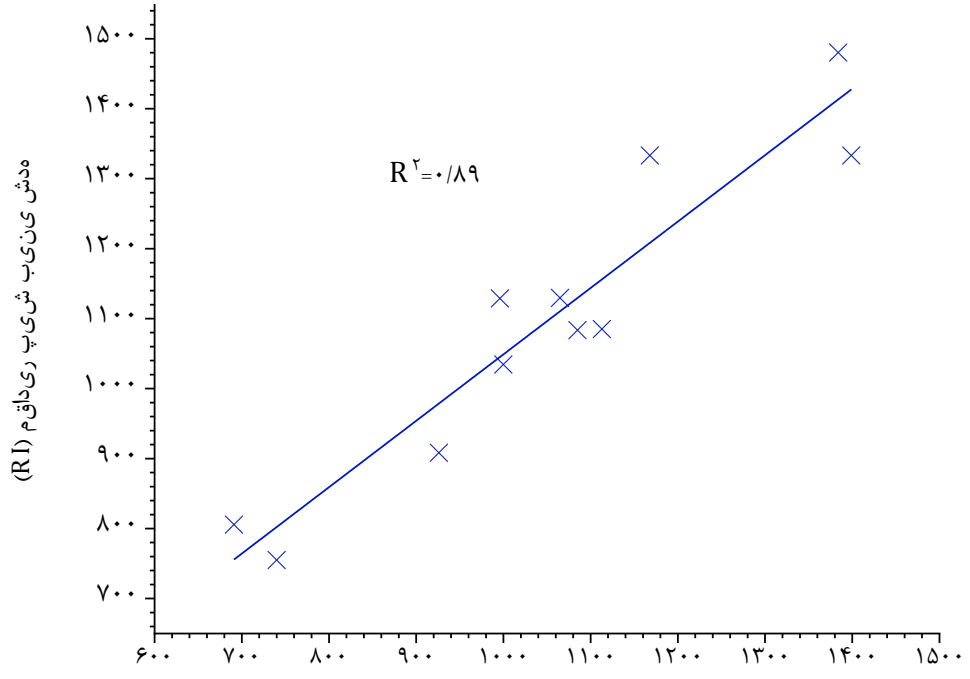
به‌منظور بررسی اعتبار و قدرت پیش بینی مدل، مقادیر RI مجموعه آزمون با استفاده از مدل‌های بهینه (SCAD-LM_ANN با معماری ۱-۲-۱۰ برای مجموعه داده A و مدل SCAD-BR-ANN با معماری ۱-۴-۷ برای مجموعه داده B) پیش‌بینی شد و نتایج هر کدام در جدول ۲-۲۷ آورده شد. همان‌طور که مشاهده می‌شود مقادیر خطای پیش بینی اغلب ترکیبات مجموعه آزمون کم است. نتایج نشان دهنده قدرت پیش بینی مدل توسعه یافته است. علاوه بر این نمودار مقادیر RI پیش‌بینی‌شده در برابر مقادیر واقعی متناظر آن‌ها رسم شد. مقدار ضریب تعیین بالاتر از حد هشدار (R^2=۰/۶) نشان‌دهنده قدرت پیش‌بینی قابل قبول مدل SCAD-ANN در هر دو مجموعه داده مورد مطالعه می‌باشد.

جدول ۲-۲۷ نتایج حاصل از ارزیابی مدل SCAD-ANN با استفاده از مجموعه آزمون

مجموعه داده‌ها	شماره ترکیبات مجموعه آزمون	شاخص بازداری (RI)		درصد خطا
		مقدار واقعی	مقدار پیش‌بینی شده	
مجموعه A	۱	۰/۴۲۲۱	۰/۴۱۴۳	-۱/۸۵
	۴	۰/۴۲۵۹	۰/۴۱۸۳	-۱/۷۹
	۶	۰/۶۰۶۲	۰/۶۲۴۱	۲/۹۶
	۱۴	۰/۷۵۰۴	۰/۷۴۰۴	-۱/۳۴
	۲۲	۰/۶۰۸	۰/۵۸۵	-۳/۷۸
	۲۷	۰/۴۷۶۶	۰/۴۷۳۵	-۰/۶۶
	۲۸	۰/۶۴۶۶	۰/۵۷۷۵	-۱۰/۶۹
	۳۱	۰/۴۷۱۱	۰/۴۸۴۳	۲/۸۱
	۴۷	۰/۷۰۰۷	۰/۶۴۰۶	-۸/۵۸
	۵۹	۰/۴۸۲۳	۰/۵۶۲۳	۱۶/۶
	۶۱	۰/۵۲۹۱	۰/۵۲۰۱	-۱/۷
	۶۸	۰/۵۱۰۶	۰/۴۷۵	-۶/۹۷
	۷۱	۰/۴۸۵۹	۰/۵۱۱	۵/۱۶
	۷۳	۰/۵۰۳	۰/۵۰۹۴	۱/۲۷
	۷۸	۰/۶۳۱۹	۰/۶۴۵۱	۲/۰۸
	۸۲	۰/۴۵۹۶	۰/۴۶۰۹	۰/۲۸
	۹۳	۰/۹۴۲۸	۰/۷۸۹۷	-۱۶/۲۴
	۹۹	۰/۴۹۵۱	۰/۵۱۱۷	۳/۳۶
	۱۰۴	۰/۵	۰/۵۱۶۸	۳/۳۶
	۱۰۸	۰/۵۲۳۷	۰/۵۲۲۹	-۰/۱۶
	۱۰۹	۰/۵۵۶۶	۰/۵۳۴۸	-۳/۹۲
	۱۱۵	۰/۵۸۲	۰/۶۱۸۹	۶/۳۵
	۱۱۸	۰/۶۴۲۹	۰/۶۵۶۳	۲/۰۹
	۱۲۱	۰/۷۳۷۱	۰/۷۲۰۱	-۲/۳
	۱۲۵	۰/۸۰۹۷	۰/۸۱۶۶	۰/۸۵
	۱۲۶	۰/۷۸۵۷	۰/۷۹۴۶	۱/۱۳
	۱۳۰	۰/۳۰۳۹	۰/۳۳۰۹	۸/۸۸
مجموعه B	۵	۷۴۰	۷۵۴/۸	۲/۰۰
	۱۱	۱۱۱۳	۱۰۸۵	-۲/۵۲
	۱۴	۹۹۶	۱۱۲۸/۸	۱۳/۳۳
	۱۹	۱۰۸۵	۱۰۸۳/۵	-۰/۱۴
	۲۰	۱۳۹۹	۱۳۳۳/۱	-۴/۷۱
	۲۵	۱۳۸۴	۱۴۸۰/۲	۶/۹۵
	۲۸	۱۱۶۸	۱۳۳۳/۱	۱۴/۱۳
	۳۲	۶۹۱	۸۰۵/۶	۱۶/۵۸
	۳۳	۱۰۰۰	۱۰۳۴/۳	۳/۴۳
	۳۶	۱۰۶۵	۱۱۲۹/۶	۶/۰۶



شکل ۲-۴۱ نمودار تغییرات مقادیر پیش‌بینی شده RI به‌وسیله مدل SCAD-LM-ANN در شرایط بهینه در مقابل مقادیر تجربی برای داده‌های مجموعه آزمون مجموعه A



شکل ۲-۴۲ نمودار تغییرات مقادیر پیش‌بینی شده RI به‌وسیله مدل SCAD-BR-ANN در شرایط بهینه در مقابل مقادیر تجربی برای داده‌های مجموعه آزمون مجموعه B

۲- ۵- ۷- ۲ ارزیابی مدل SCAD-ANN با پیش بینی شاخص‌های بازداری تمام ترکیبات مجموعه داده‌های A و B با استفاده از روش رد مرحله‌ای تک تک

در این بخش، یکی از تکنیک‌های کارآمد برای ارزیابی مدل بهینه و بررسی قدرت پیش بینی مدل از تکنیک رد مرحله‌ای تک تک استفاده شد. در این روش هر ترکیب یک‌بار به‌عنوان داده آزمون حذف شد و مدل بهینه SCAD-ANN با استفاده از داده‌های باقی‌مانده آموزش داده شد. پس از اجرای مدل‌های بهینه برای هر دو مجموعه A و B، مقادیر RI همه ترکیبات پیش‌بینی و در جدول ۲-۲۸ و جدول ۲-۲۹ آورده شد. نتایج خطای پیش بینی اغلب ترکیبات کم بوده و این امر نشان دهنده قدرت پیش بینی مناسب مدل SCAD-ANN است. علاوه بر این، مقادیر پیش‌بینی‌شده بر حسب مقادیر واقعی RI برای هر مجموعه داده ترسیم شدند (شکل ۲-۴۵ و شکل ۲-۴۶ به‌ترتیب برای مجموعه A و B). مقادیر Q^2_{LOO} مربوط به مجموعه داده‌های A و B به‌ترتیب برابر با ۰/۹۴ و ۰/۸۸ به‌دست آمد. مقادیر Q^2_{LOO} بزرگ‌تر از ۰/۵ حاکی از اعتبار و پایداری بالای مدل توسعه یافته SCAD-ANN است.

علاوه بر این، نتایج حاصل برای هر دو مجموعه داده با استفاده از نمودارهای باقی‌مانده‌ها نیز ارزیابی شد. به‌این منظور، نمودارهای باقی‌مانده‌های استاندارد شده (r_i) بر حسب مقادیر واقعی رسم شد (شکل ۲-۴۴ و شکل ۲-۴۶). الگوی تصادفی و پراکندگی داده‌ها حول محور صفر در نمودار باقی‌مانده‌ها نشان می‌دهد که هیچ خطای سیستماتیکی در مدل‌های ANN توسعه یافته با توصیف کننده‌های منتخب روش SCAD وجود ندارد.

جدول ۲-۲۸ نتایج حاصل از ارزیابی مدل SCAD-ANN با تکنیک LOO برای کل داده‌های مجموعه A

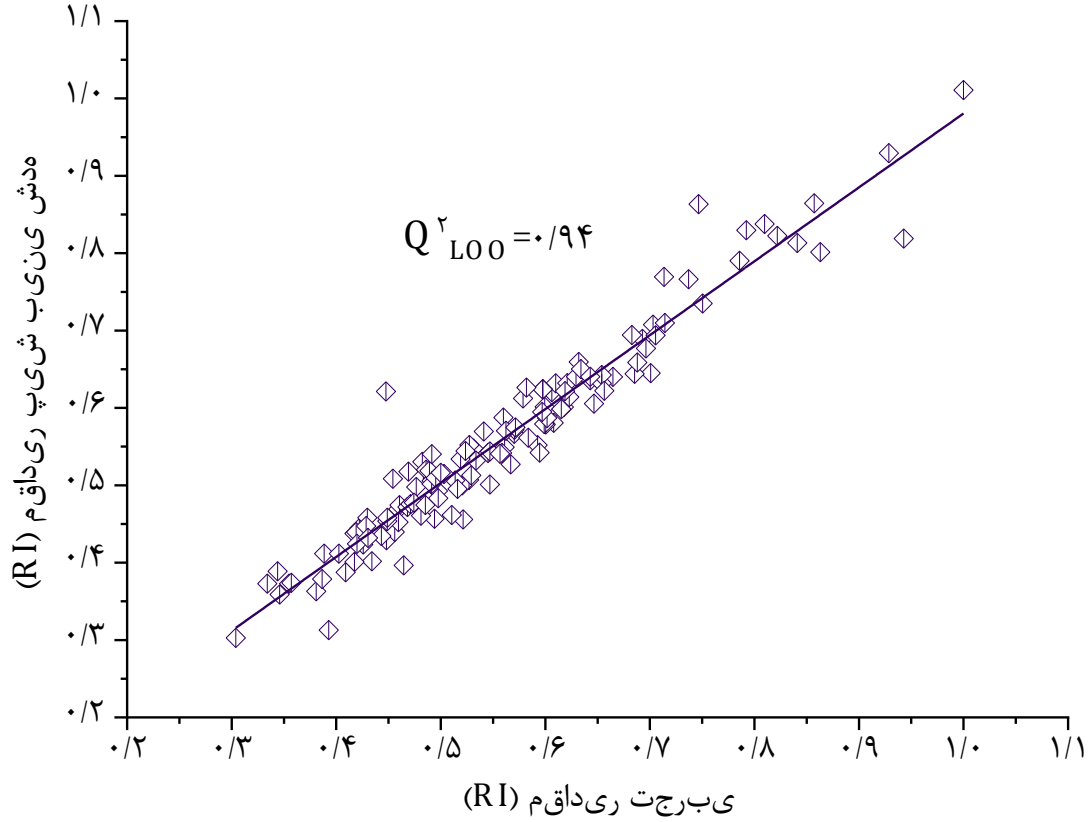
درصد خطا	RI		شماره ترکیب	RI		درصد خطا	RI		شماره ترکیب
	مقدار پیش‌بینی شده	مقدار واقعی		مقدار پیش‌بینی شده	مقدار واقعی		مقدار پیش‌بینی شده	مقدار واقعی	
-۰/۰۴	۰/۵۷	۰/۵۷	۴۰	۰/۴۴	۰/۴۲	-۰/۶۶	۰/۴۴	۰/۴۲	۱
۰/۶۵	۰/۵۸	۰/۶	۴۱	۰/۳۷	۰/۳۳	-۱/۱۵	۰/۳۷	۰/۳۳	۲
۰/۵۸	۰/۵۸	۰/۶	۴۲	۰/۴۹	۰/۵	۰/۱۳	۰/۴۹	۰/۵	۳
۰/۷۳	۰/۶۴	۰/۶۶	۴۳	۰/۴۲	۰/۴۳	۰/۰۷	۰/۴۲	۰/۴۳	۴
۰/۶۶	۰/۶۳	۰/۶۵	۴۴	۰/۵۷	۰/۵۷	۰/۱۱	۰/۵۷	۰/۵۷	۵
۰/۳۴	۰/۶۴	۰/۶۵	۴۵	۰/۶۲	۰/۶۱	-۰/۴۳	۰/۶۲	۰/۶۱	۶
-۰/۳۳	۰/۶۳	۰/۶۲	۴۶	۰/۵۱	۰/۴۹	-۰/۵۱	۰/۵۱	۰/۴۹	۷
۱/۶۵	۰/۶۵	۰/۷	۴۷	۰/۶۴	۰/۶۴	۰/۲۱	۰/۶۴	۰/۶۴	۸
۱/۲۳	۰/۶۴	۰/۶۹	۴۸	۰/۵۵	۰/۵۶	۰/۳۵	۰/۵۵	۰/۵۶	۹
-۰/۳۳	۰/۶۴	۰/۶۳	۴۹	۰/۴۱	۰/۳۹	-۰/۶۹	۰/۴۱	۰/۳۹	۱۰
-۱/۶۱	۰/۵۱	۰/۴۵	۵۰	۰/۴۷	۰/۴۶	-۰/۲۹	۰/۴۷	۰/۴۶	۱۱
-۰/۷۴	۰/۵۵	۰/۵۳	۵۱	۰/۵۵	۰/۵۳	-۰/۴۸	۰/۵۵	۰/۵۳	۱۲
-۰/۰۶	۰/۶	۰/۶	۵۲	۰/۵۹	۰/۶	۰/۳۳	۰/۵۹	۰/۶	۱۳
-۰/۴۵	۰/۵۳	۰/۵۲	۵۳	۰/۷۴	۰/۷۵	۰/۴۵	۰/۷۴	۰/۷۵	۱۴
-۰/۱	۰/۴۸	۰/۴۷	۵۴	۰/۶۹	۰/۶۹	۰/۱۱	۰/۶۹	۰/۶۹	۱۵
۲/۳۸	۰/۳۱	۰/۳۹	۵۵	۰/۴۷	۰/۴۶	-۰/۴۱	۰/۴۷	۰/۴۶	۱۶
-۱/۴۴	۰/۵۲	۰/۴۷	۵۶	۰/۵۳	۰/۵۳	۰/۰۵	۰/۵۳	۰/۵۳	۱۷
-۰/۵۳	۰/۳۷	۰/۳۶	۵۷	۰/۳۶	۰/۳۸	۰/۵۳	۰/۳۶	۰/۳۸	۱۸
-۰/۸۵	۰/۴۶	۰/۴۳	۵۸	۰/۴۳	۰/۴۵	۰/۵۷	۰/۴۳	۰/۴۵	۱۹
-۱/۴۳	۰/۵۳	۰/۴۸	۵۹	۰/۵۱	۰/۵۳	۰/۵۹	۰/۵۱	۰/۵۳	۲۰
۰/۴۸	۰/۴۴	۰/۴۶	۶۰	۰/۵	۰/۵۲	۰/۶۱	۰/۵	۰/۵۲	۲۱
۰/۴۸	۰/۵۱	۰/۵۳	۶۱	۰/۵۸	۰/۶۱	۰/۸۱	۰/۵۸	۰/۶۱	۲۲
۰/۴۴	۰/۵۹	۰/۶	۶۲	۰/۵۵	۰/۵۹	۱/۲۱	۰/۵۵	۰/۵۹	۲۳
۰/۵	۰/۶	۰/۶۲	۶۳	۰/۴۷	۰/۴۷	-۰/۱۱	۰/۴۷	۰/۴۷	۲۴
-۳/۱۶	۰/۶۲	۰/۴۵	۶۴	۰/۶۳	۰/۶۱	-۰/۶۵	۰/۶۳	۰/۶۱	۲۵
۱/۹۴	۰/۴۶	۰/۵۲	۶۵	۰/۴	۰/۴۳	۰/۹۶	۰/۴	۰/۴۳	۲۶
۱/۵۵	۰/۵۴	۰/۵۹	۶۶	۰/۵	۰/۴۸	-۰/۶۳	۰/۵	۰/۴۸	۲۷
-۰/۳۴	۰/۶۹	۰/۶۸	۶۷	۰/۶۱	۰/۶۵	۱/۲۲	۰/۶۱	۰/۶۵	۲۸
۱/۴۶	۰/۴۶	۰/۵۱	۶۸	۰/۶۸	۰/۷	۰/۵۷	۰/۶۸	۰/۷	۲۹
۰/۶۶	۰/۵۶	۰/۵۸	۶۹	۰/۴۱	۰/۴	-۰/۲۷	۰/۴۱	۰/۴	۳۰
۱/۰۲	۰/۶۲	۰/۶۶	۷۰	۰/۴۸	۰/۴۷	-۰/۱۳	۰/۴۸	۰/۴۷	۳۱
-۱/۰۱	۰/۵۲	۰/۴۹	۷۱	۰/۵۴	۰/۵۵	۰/۱	۰/۵۴	۰/۵۵	۳۲
-۰/۶۱	۰/۴۴	۰/۴۲	۷۲	۰/۶	۰/۶۲	۰/۴۹	۰/۶	۰/۶۲	۳۳
-۰/۳۴	۰/۵۱	۰/۵	۷۳	۰/۸۶	۰/۷۵	-۳/۴۷	۰/۸۶	۰/۷۵	۳۴
۰/۶۲	۰/۵	۰/۵۲	۷۴	۰/۳۶	۰/۳۵	-۰/۳۸	۰/۳۶	۰/۳۵	۳۵
-۰/۴۳	۰/۵۱	۰/۴۹	۷۵	۰/۴	۰/۴۲	۰/۴۹	۰/۴	۰/۴۲	۳۶
-۰/۸۸	۰/۵۲	۰/۴۹	۷۶	۰/۴۶	۰/۴۹	۱/۱۱	۰/۴۶	۰/۴۹	۳۷
-۰/۸۲	۰/۵۹	۰/۵۶	۷۷	۰/۵۳	۰/۵۷	۱/۱۹	۰/۵۳	۰/۵۷	۳۸
-۰/۸۲	۰/۶۶	۰/۶۳	۷۸	۰/۵۱	۰/۵۱	۰/۲۳	۰/۵۱	۰/۵۱	۳۹

ادامه جدول ۲-۲۸

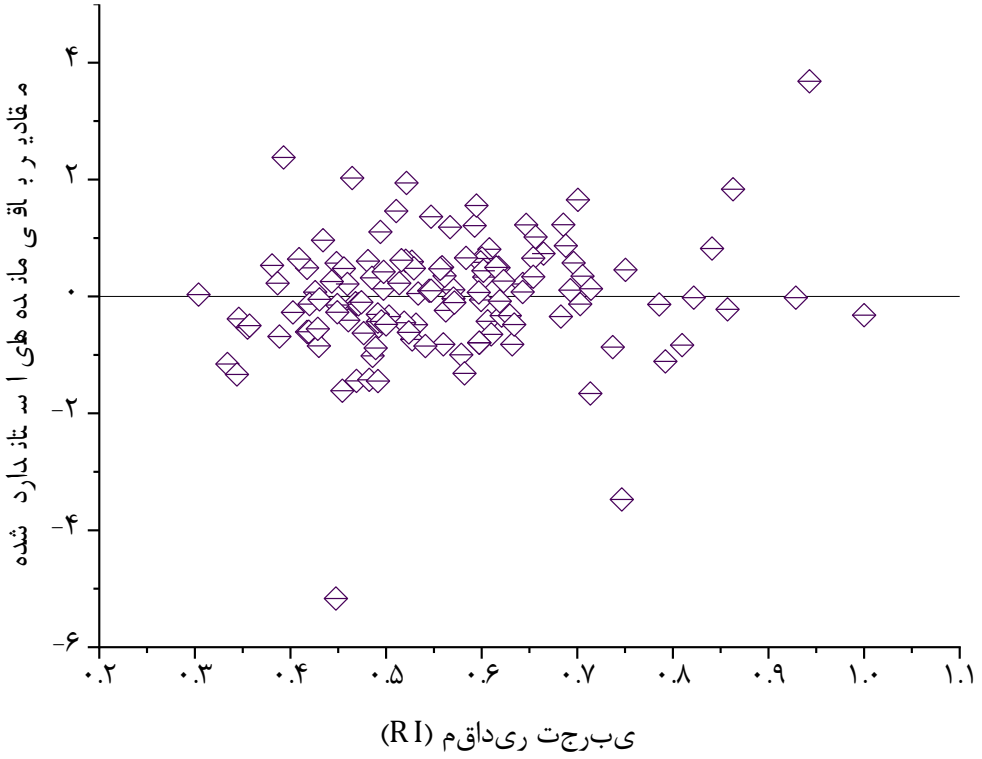
درصد خطا	RI		شماره ترکیب	درصد خطا	RI		شماره ترکیب
	مقدار پیش‌بینی شده	مقدار پیش‌بینی شده			مقدار پیش‌بینی شده	مقدار پیش‌بینی شده	
۰/۱	۰/۵۴	۰/۵۵	۱۰۷	-۰/۱۳	۰/۷۱	۰/۷	۸۱
-۰/۶۲	۰/۵۴	۰/۵۲	۱۰۸	-۱/۳۳	۰/۳۹	۰/۳۴	۸۲
۰/۴۷	۰/۵۴	۰/۵۶	۱۰۹	-۰/۶	۰/۴۴	۰/۴۲	۸۳
۰/۴۲	۰/۴۸	۰/۵	۱۱۰	۰/۲۱	۰/۴۵	۰/۴۶	۸۴
۱/۳۶	۰/۵	۰/۵۵	۱۱۱	۰/۲۶	۰/۴۳	۰/۴۴	۸۵
-۰/۰۸	۰/۶۲	۰/۶۲	۱۱۲	-۰/۳۱	۰/۵	۰/۴۹	۸۶
-۰/۷۹	۰/۶۲	۰/۶	۱۱۳	-۰/۲۴	۰/۵۷	۰/۵۶	۸۷
-۰/۷۹	۰/۶۲	۰/۶	۱۱۴	۱	۰/۶۱	۰/۵۸	۸۸
-۱/۳۲	۰/۶۳	۰/۵۸	۱۱۵	-۰/۴۸	۰/۶۵	۰/۶۳	۸۹
-۰/۱۱	۰/۵۷	۰/۵۷	۱۱۶	۰/۳۵	۰/۶۹	۰/۷۱	۹۰
۰/۸۷	۰/۶۶	۰/۶۹	۱۱۷	۰/۶۱	۰/۴۶	۰/۴۸	۹۱
۰/۰۸	۰/۶۴	۰/۶۴	۱۱۸	۰/۵	۰/۵۴	۰/۵۶	۹۲
۰/۱۳	۰/۷۱	۰/۷۱	۱۱۹	۰/۲۶	۰/۶۱	۰/۶۲	۹۳
۰/۸۲	۰/۸۱	۰/۸۴	۱۲۰	۱/۸۳	۰/۸	۰/۸۶	۹۴
-۰/۸۷	۰/۷۷	۰/۷۴	۱۲۱	۳/۶۸	۰/۸۲	۰/۹۴	۹۵
-۱/۶۶	۰/۷۷	۰/۷۱	۱۲۲	-۰/۵	۰/۳۷	۰/۳۶	۹۶
-۰/۰۲	۰/۸۲	۰/۸۲	۱۲۳	-۰/۱۳	۰/۴۲	۰/۴۲	۹۷
-۱/۱۱	۰/۸۳	۰/۷۹	۱۲۴	۰/۲۳	۰/۳۸	۰/۳۹	۹۸
-۰/۸۳	۰/۸۴	۰/۸۱	۱۲۵	۰/۶۴	۰/۳۹	۰/۴۱	۹۹
-۰/۱۴	۰/۷۹	۰/۷۹	۱۲۶	-۰/۵۶	۰/۴۵	۰/۴۳	۱۰۰
-۰/۲۲	۰/۸۶	۰/۸۶	۱۲۷	-۰/۴۳	۰/۵۱	۰/۵	۱۰۱
-۰/۰۲	۰/۹۳	۰/۹۳	۱۲۸	-۰/۱۵	۰/۴۵	۰/۴۵	۱۰۲
-۰/۳۲	۱/۰۱	۱	۱۲۹	۰/۳۲	۰/۴۷	۰/۴۹	۱۰۳
۰/۰۳	۰/۳	۰/۳	۱۳۰	-۰/۲۷	۰/۴۶	۰/۴۵	۱۰۴
-۱/۴۵	۰/۵۴	۰/۴۹	۱۳۱	۲/۰۳	۰/۴	۰/۴۶	۱۰۵
-۰/۰۵	۰/۴۳	۰/۴۳	۱۳۲	-۰/۴۸	۰/۵۲	۰/۵	۱۰۶

جدول ۲-۲۹ نتایج حاصل از ارزیابی مدل SCAD-ANN با تکنیک LOO برای کل داده‌های مجموعه B

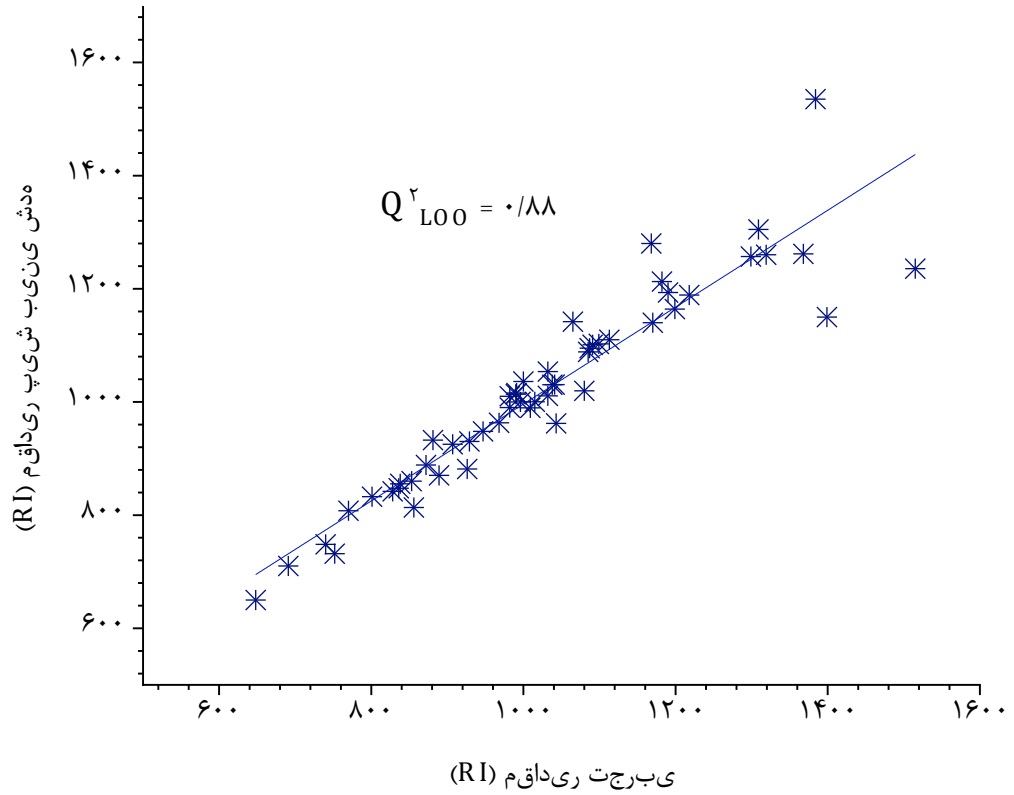
RI				RI			
درصد خطا	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب	درصد خطا	مقدار پیش‌بینی شده	مقدار واقعی	شماره ترکیب
-۱/۶۸	۸۴۱/۹۱	۸۲۸	۲۷	۴/۹۸	۸۱۳/۳۵	۸۵۶	۱
-۹/۵۹	۱۲۸۰	۱۱۶۸	۲۸	-۰/۷۵	۱۰۹۵/۱۴	۱۰۸۷	۲
-۴/۹	۸۰۷/۷۴	۷۷۰	۲۹	-۰/۱۱	۹۴۸	۹۴۷	۳
-۲/۰۷	۸۵۵/۳۶	۸۳۸	۳۰	-۰/۸۲	۸۶۰	۸۵۳	۴
-۲/۴۹	۱۰۱۵/۶۳	۹۹۱	۳۱	-۱/۱۴	۷۴۸/۴۲	۷۴۰	۵
-۲/۷۵	۷۱۰	۶۹۱	۳۲	-۰/۸۱	۹۹۰	۹۸۲	۶
-۳/۶۲	۱۰۳۶/۲	۱۰۰۰	۳۳	-۰/۲۸	۱۱۹۳/۳۸	۱۱۹۰	۷
-۰/۱۴	۹۳۰/۲۹	۹۲۹	۳۴	۰/۷۲	۱۰۳۰/۴۹	۱۰۳۸	۸
-۰/۳۳	۱۱۰۲/۶۶	۱۰۹۹	۳۵	۲/۵۶	۱۱۴۰	۱۱۷۰	۹
-۷/۲۱	۱۱۴۱/۷۷	۱۰۶۵	۳۶	۱/۴۴	۱۰۰۰/۳۷	۱۰۱۵	۱۰
۰/۹۹	۱۰۳۰/۶۹	۱۰۴۱	۳۷	۰/۲۷	۱۱۱۰	۱۱۱۳	۱۱
۷/۷۶	۹۶۲/۰۵	۱۰۴۳	۳۸	۰/۳۱	۱۳۰۵	۱۳۰۹	۱۲
۱/۹۴	۹۸۹/۴۶	۱۰۰۹	۳۹	۳/۲۳	۱۲۵۶/۹۸	۱۲۹۹	۱۳
۴/۸۲	۸۸۱/۳۹	۹۲۶	۴۰	-۰/۴	۱۰۰۰	۹۹۶	۱۴
۲/۶۷	۷۳۱/۹۱	۷۵۲	۴۱	-۰/۹۳	۱۱۰۱/۱۳	۱۰۹۱	۱۵
۲/۰۷	۱۰۱۰/۶۵	۱۰۳۲	۴۲	-۲/۶۱	۱۲۱۲/۸	۱۱۸۲	۱۶
-۳/۹۵	۸۳۲/۶۶	۸۰۱	۴۳	۲/۹۱	۱۱۶۴/۰۸	۱۱۹۹	۱۷
۲/۱۲	۸۷۰/۱۷	۸۸۹	۴۴	۵/۵۹	۱۰۱۹/۶۷	۱۰۸۰	۱۸
-۱/۹	۸۸۸/۵۷	۸۷۲	۴۵	-۰/۳۱	۱۰۸۸/۳۹	۱۰۸۵	۱۹
-۲/۸۶	۱۰۱۰/۱	۹۸۲	۴۶	۱۷/۸	۱۱۵۰	۱۳۹۹	۲۰
-۲/۰۹	۱۰۵۳/۵۶	۱۰۳۲	۴۷	۴/۴۷	۱۲۶۰/۱	۱۳۱۹	۲۱
-۲/۰۱	۹۲۵/۲	۹۰۷	۴۸	۷/۷۶	۱۲۶۱/۸۱	۱۳۶۸	۲۲
-۱/۳۷	۸۴۷/۴۳	۸۳۶	۴۹	-۰/۳۱	۶۵۰	۶۴۸	۲۳
-۵/۸۵	۹۳۲/۵۸	۸۸۱	۵۰	۱۸/۴۵	۱۲۳۵/۴۱	۱۵۱۵	۲۴
۰/۴۹	۹۶۳/۲۳	۹۶۸	۵۱	-۱۰/۹۳	۱۵۳۵/۲۲	۱۳۸۴	۲۵
-۲/۳۸	۱۰۱۳/۵۴	۹۹۰	۵۲	۲/۳۸	۱۱۸۹/۰۱	۱۲۱۸	۲۶



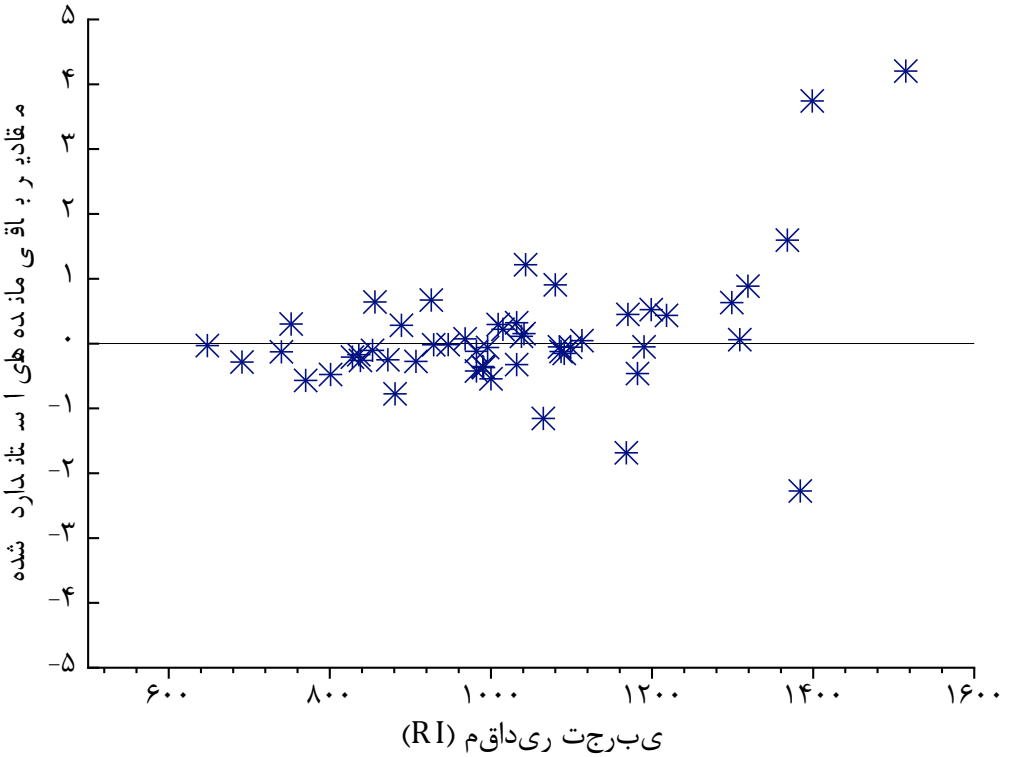
شکل ۲-۴۳ نمودار تغییرات مقادیر پیش‌بینی شده RI همه داده‌های مجموعه A بر اساس تکنیک LOO در مقابل مقادیر تجربی



شکل ۲-۴۴ نمودار باقی‌مانده‌های پیش‌بینی شده RI همه داده‌های مجموعه A با استفاده از تکنیک LOO برحسب مقادیر تجربی



شکل ۲-۴۵ نمودار تغییرات مقادیر پیش‌بینی شده RI همه داده‌های مجموعه B بر اساس تکنیک LOO در مقابل مقادیر تجربی



شکل ۲-۴۶ نمودار باقی‌مانده‌های پیش‌بینی شده RI همه داده‌های مجموعه B با استفاده از تکنیک LOO برحسب مقادیر تجربی

پارامترهای آماری RMSE و R^2 مربوط به مجموعه آزمون و کل داده‌ها (تکنیک LOO) برای هر دو مجموعه داده A و B محاسبه شدند. RMSE مربوط به مجموعه آزمون برای مجموعه داده A و B به‌ترتیب برابر با ۰/۰۵ و ۱۰۴/۰۲ و علاوه بر این برای کل داده‌ها به‌ترتیب برابر با ۰/۰۵ و ۸۸/۵۰ به‌دست آمد. علاوه بر این ضرایب تعیین مربوط به مجموعه آزمون نیز به ترتیب برابر با ۰/۹۰ و ۰/۸۷ و ضرایب تعیین مربوط به کل داده‌ها (تکنیک LOO) به ترتیب برابر با ۰/۸۴ و ۰/۸۴ برای مجموعه داده A و B به دست آمد. با مقایسه نتایج مدل‌های برتر SCAD-ANN (جدول ۲-۲۶) برای هر دو مجموعه داده مشخص می‌شود که مدل‌های پیشنهادی در پیش‌بینی RI ترکیبات مورد نظر با قدرت بیش‌تری نسبت به مدل‌های SR-ANN عمل کرده‌اند.

۲- ۵- ۷- ۳ ارزیابی مدل SCAD-ANN با استفاده از پارامترهای آماری

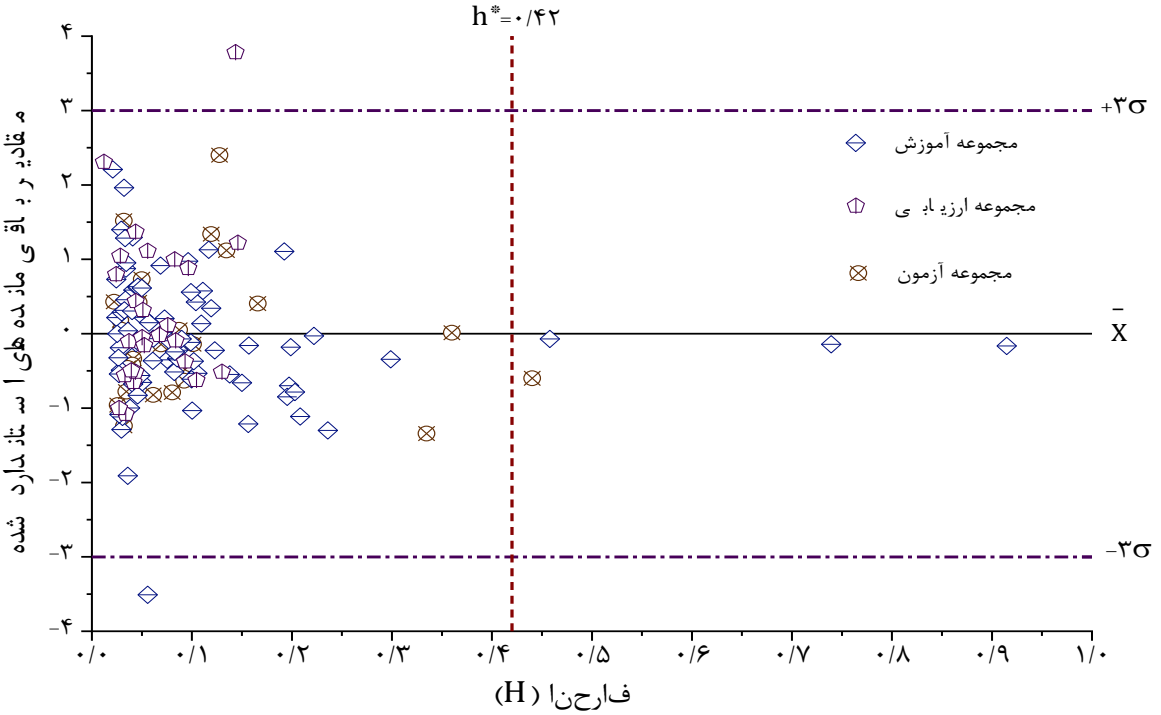
پارامترهای آماری مختلف نیز برای ارزیابی بیش‌تر توانایی پیش‌بینی مدل‌های SCAD-ANN پیشنهادی محاسبه شد. بنابراین پارامترهای آماری معرفی شده در بخش ۱- ۵- ۸- ۴ برای RI پیش‌بینی شده ترکیبات مجموعه آزمون و RI پیش‌بینی شده برای کل ترکیبات به روش رد مرحله‌ای تک تک محاسبه و در جدول ۲-۳۰ خلاصه شدند. نتایج حاصل از محاسبات آماری جدول ۲-۳۰ نشان می‌دهد که پارامترهای آماری در محدوده قابل قبول قرار دارند. پارامترهای آماری تروپشا و روی (پارامترهای R_0^2،R_0^2 نسبی،R_m^2 و غیره) از جمله پارامترهای آماری پر کاربرد وابسته به R^2 هستند. بزرگ‌تر بودن این پارامترها از مقدار هشدار ۰/۶ و نزدیک بودن آن‌ها به R^2 نشاندهنده تعمیم پذیری و قدرت پیش بینی رضایت بخش مدل‌های توسعه یافته QSAR مبتنی بر SCAD-ANN است. علاوه بر این شیب نمودار حاصل از مقادیر پیش بینی شده بر حسب مقادیر تجربی RI (و بالعکس) در عرض از مبدأ صفر نیز در محدوده ۰/۸۵ تا ۱/۱۵ قرار دارند که این نتیجه نیز حاکی از صحت مدل توسعه یافته ANN با توصیف کننده‌های منتخب روش SCAD می‌باشد.

جدول ۲-۳۰ پارامترهای آماری محاسبه شده برای مجموعه آزمون و داده‌های پیش‌بینی شده با تکنیک LOO برای مدل SCAD-ANN هر دو مجموعه از داده‌ها

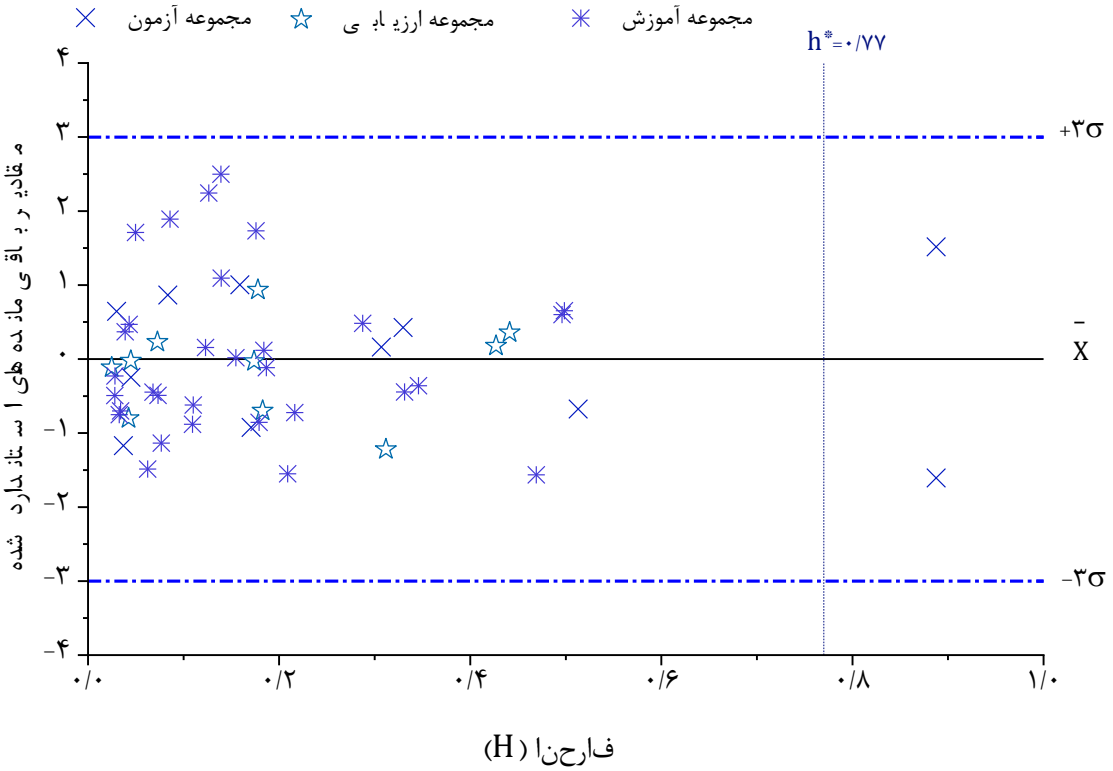
ردیف	پارامتر آماری	مجموعه داده A		مجموعه داده B		محدوده قابل قبول
		مقادیر RI پیش بینی شده با SCAD-ANN				
		ترکیبات مجموعه آزمون	کل ترکیبات به روش LOO	ترکیبات مجموعه آزمون	کل ترکیبات به روش LOO	
۱	PRESS	۰/۰۵	۰/۱۵	۷۸۴۲۶	۲۲۹۰۰۰	-
۲	SEP	۰/۰۴	۰/۰۳	۸۴/۴۴	۶۶/۳۲	-
۳	MAE	۰/۰۳	۰/۰۲	۶۶/۹۱	۳۷/۷۸	-
۴	REP(%)	۵/۳۹	۵/۹۸	۸/۰۲	۶/۴۵	-
۵	RMSE	۰/۰۴	۰/۰۳	۸۴/۴۳	۶۶/۳۲	-
۶	MRE	۴/۳۴	۴/۱۴	۶/۵۳	۳/۳۲	-
۷	R^2	۰/۹۲	-	۰/۸۹	-	R^2>۰/۶
۸	Q^2_{LOO}	-	۰/۹۴	-	۰/۸۸	Q^2_{Loo}> ۰/۵
۹	R^2_0	۰/۹۱	۰/۹۳	۰/۸۹	۰/۸۶	نزدیک به R^2
۱۰	R^2_0 نسبی	۰/۰۱	۰/۰۱	۰/۰۱	۰/۰۲	<۰/۱
۱۱	R^2_m	۰/۸۳	۰/۸۵	۰/۸	۰/۷۶	>۰/۵
۱۲	R'^2_0	۰/۹۲	۰/۹۳	۰/۸۸	۰/۸۸	نزدیک به R^2
۱۳	R'^2_0 نسبی	۰	۰/۰۱	۰	۰/۰۲	<۰/۱
۱۴	R'^2_m	۰/۸۲	۰/۹۳	۰/۷۹	۰/۷۴	>۰/۵
۱۵	R-R	۰/۰۱	۰	۰/۰۱	۰/۰۲	<۰/۳
۱۶	k	۰/۹۸	۱/۰۰	۱/۰۴	۰/۹۹	۰/۸۵ ≥k≤ ۱/۱۵
۱۷	k′	۱/۰۱	۱/۰۰	۰/۹۵	۱	۰/۸۵ ≥k'≤ ۱/۱۵

۲- ۵- ۷- ۴ ارزیابی مدل SCAD-ANN با استفاده از دامنه کاربرد

هدف از دامنه کاربرد، اثبات اعتمادپذیری مدل پیشنهادی SCAD-ANN توصیه شده برای ترکیبات جدید (مجموعه آزمون)، است. بنابراین، وجود مقادیر RI پیش‌بینی‌شده در محدوده فضای شیمیایی، صحت نتایج را اثبات می‌کند. نمودار ویلیام به‌عنوان بهترین نمایش از دامنه کاربرد، مطابق با روش کار بخش‌های ۱- ۵- ۸- ۵ و ۲- ۲- ۷- ۴ برای مدل‌های QSRR مبتنی بر شبکه عصبی و توسعه یافته با توصیف کننده‌های منتخب روش SCAD به‌دست آمد. بنابراین مقادیر انحراف بر اساس رابطه ۱-۱۳ محاسبه شد. به‌این ترتیب مقادیر استاندارد شده باقی‌مانده‌ها بر حسب مقادیر پیش بینی شده و واقعی RI و مطابق با رابطه ۱-۱۴ محاسبه شد. از رسم مقادیر باقی‌مانده‌های استاندارد شده بر حسب مقادیر انحراف نمودار ویلیام به‌دست آمد (شکل ۲-۴۷ و شکل ۲-۴۸). علاوه بر این مقادیر h^* برای هر دو مجموعه داده A و B به‌ترتیب برابر با ۰/۴۲ و ۰/۷۷ به‌دست آمد. نمودار ویلیام نشان می‌دهد که بیش از ۹۵ درصد داده‌ها در محدوده اطمینان (کوچک‌تر از h^* و قرارگیری در محدوده $\pm 3\sigma$) قرار دارند، به این معنی که نتایج پیش‌بینی شده مدل‌های SCAD-ANN هر دو مجموعه داده دارای اعتماد و استحکام قابل قبولی هستند.



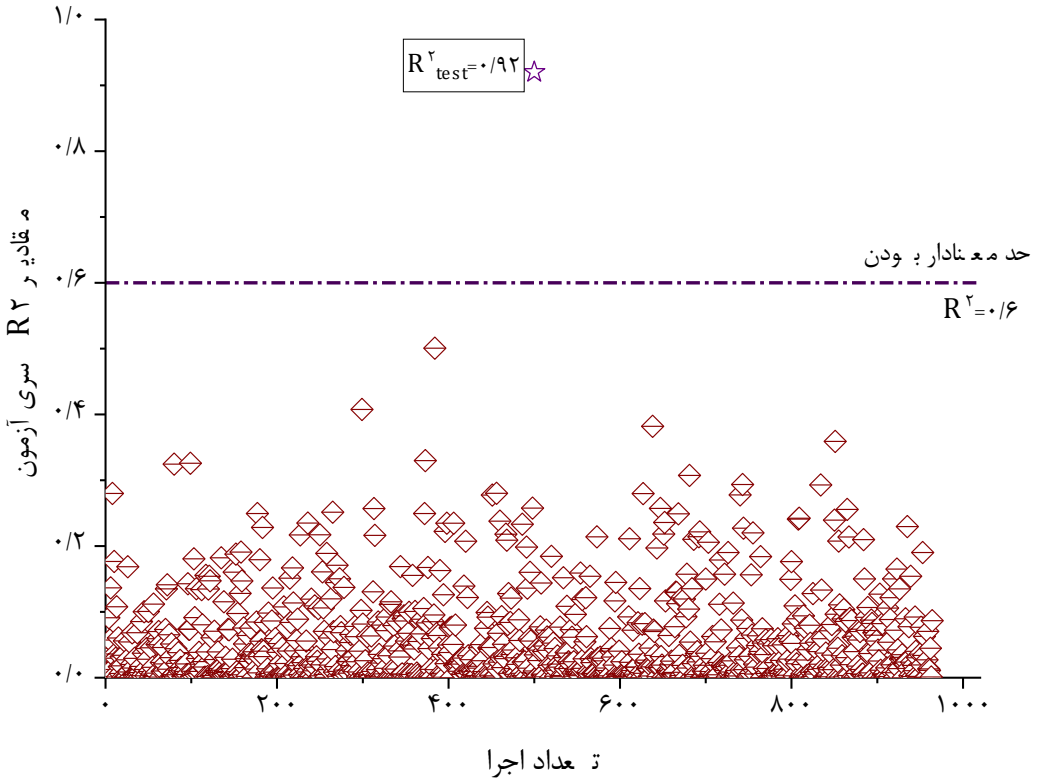
شکل ۲-۴۷ دامنه کاربرد مدل SCAD-LM-ANN برای مجموعه A، خطوط نقطه چین افقی و عمودی در دو انتهای نمودار به ترتیب نمایانگر مقادیر ۳σ ± و h* است.



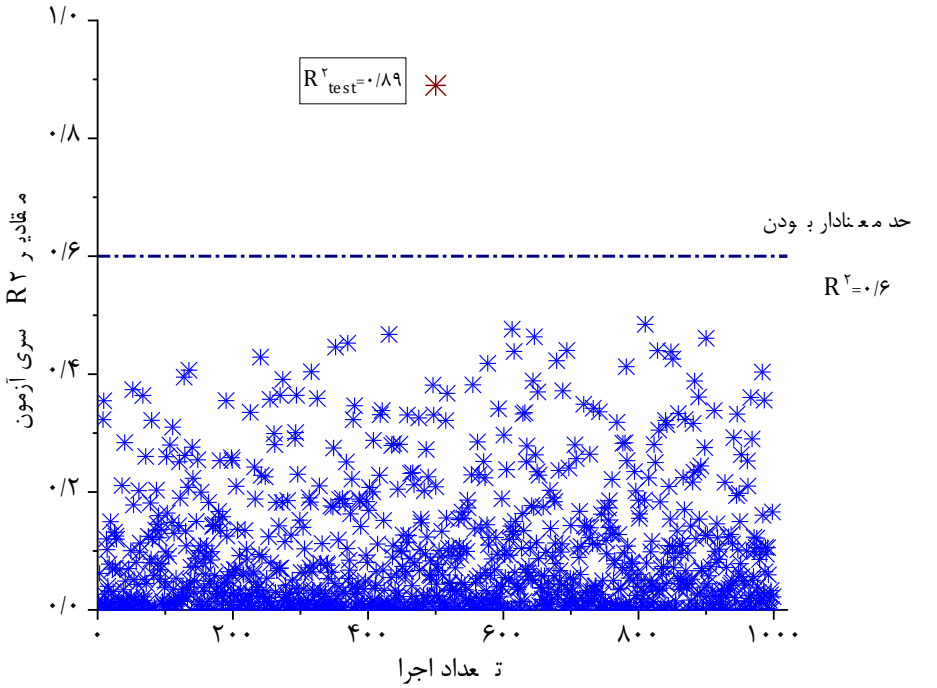
شکل ۲-۴۸ دامنه کاربرد مدل SCAD-BR-ANN برای مجموعه B، خطوط نقطه چین افقی و عمودی در دو انتهای نمودار به ترتیب نمایانگر مقادیر ۳σ ± و h* است.

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

به‌منظور بررسی اعتبار و عدم وجود ارتباط شانسی ایجاد شده توسط مدل‌های SCAD-ANN بین توصیف کننده‌های منتخب و شاخص بازداری مربوطه از آنالیز Y-تصادفی استفاده شد. برای اجرای آزمون Y- تصادفی، ابتدا مقادیر متغیر وابسته (RI) مجموعه آموزش مجموعه داده‌های A و B، در محدوده حداقل و حداکثر مقادیر RI، ۱۰۰۰ بار تصادفی شدند. مدل‌های SCAD-ANN پیشنهادی با استفاده از پاسخ‌های دست‌کاری شده، آموزش داده شد و برای پیش بینی مقادیر RI مجموعه آزمون به‌کار گرفته شد. نتایج R^2 حاصل از پیش بینی RI ترکیبات مجموعه آزمون با ۱۰۰۰ مدل توسعه یافته با متغیر وابسته تصادفی، به‌دست آمد و در شکل ۲-۴۹ و شکل ۲-۵۰ نشان داده شد. نتایج به‌دست‌آمده نشان می‌دهد که مقادیر R^2 با مدل‌های SCAD-ANN توسعه یافته با داده‌های دست‌کاری شده، به‌طور قابل‌توجهی کوچک‌تر از ضریب تعیین مجموعه آزمون مدل برتر هر دو مجموعه داده (R^2= ۰/۹۲ برای مجموعه A و R^2= ۰/۸۹ برای مجموعه B) و حتی کوچک‌تر از مقدار قابل قبول ۰/۶ است. بنابراین، نتایج به‌دست‌آمده ثابت می‌کند که رابطه QSRR ایجاد شده تصادفی نیست و بر اساس یک رابطه دقیق و منطقی بین توصیف‌کننده‌های منتخب روش SCAD و RI ایجاد شده است.



شکل ۲-۴۹ نمودار مقادیر R^2 به‌دست آمده در آزمون Y- تصادفی بر حسب تعداد اجرا برای ۱۰۰۰ اجرای Y-تصادفی و پیش‌بینی پیش‌بینی RI ترکیبات آزمون مجموعه A به‌وسیله مدل SCAD-ANN با استفاده از پاسخ تصادفی شده در شرایط بهینه



شکل ۲-۵۰ نمودار مقادیر R۲ به‌دست آمده در آزمون -Yتصادفی بر حسب تعداد اجرا برای ۱۰۰۰ اجرای Y-تصادفی و پیش‌بینی پیش‌بینی RI ترکیبات آزمون مجموعه B به‌وسیله مدل SCAD-ANN با استفاده از پاسخ تصادفی شده در شرایط بهینه

نتیجه گیری و آینده نگری

۳- ۱ نتیجه گیری نهایی مدل‌های توسعه یافته QSAR/QSPR

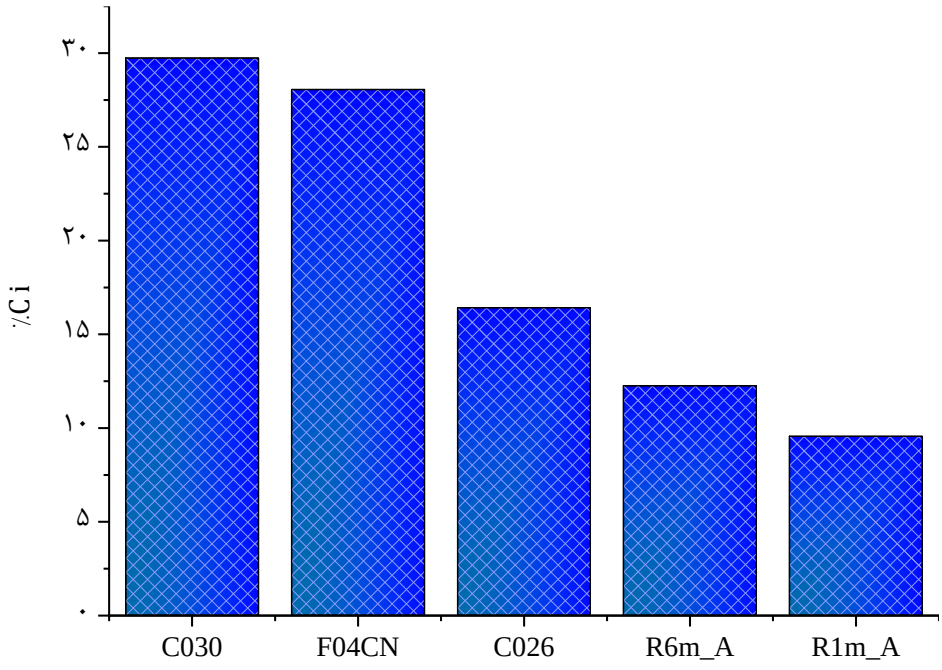
در این رساله تلاش شد تا مدل‌های QSAR/QSPR با اعتبار قابل قبول جهت پیش‌بینی فعالیت دارویی و یا شاخص بازداری ترکیبات شیمیایی ارائه شود. ارزیابی مدل‌های توسعه یافته نشان داد که این مدل‌ها از اعتبار و تعمیم پذیری مناسبی برخوردار هستند. بنابراین در این فصل سعی شده است ارتباط بین توصیف‌کننده‌های مورد استفاده در مدل‌های بهینه شبکه عصبی با متغیر پاسخ هدف (فعالیت دارویی و یا شاخص بازداری) تجزیه و تحلیل گردد. به‌طوری‌که با فهم چگونگی تأثیر هر توصیف کننده بر متغیر پاسخ و میزان سهم مشارکت هر توصیف کننده در مدل برتر، ترکیبات جدیدی با فعالیت دارویی مناسب به‌عنوان ترکیبات کاندید برای سنتز و بررسی بیشتر به‌طور تجربی پیشنهاد گردد. لازم به ذکر است که برای اثبات صحت و درستی فعالیت دارویی ترکیبات پیشنهادی جدید از مطالعه داکینگ مولکولی و بررسی برهم کنش لیگاند – گیرنده استفاده شده است. پتانسیل بالقوه فعالیت دارویی ترکیبات پیشنهادی با مقایسه برهم کنش‌های مؤثر با اسیدآمینه‌های کلیدی در ترکیبات فعال مجموعه داده‌ها با ترکیبات پیشنهادی به‌صورت نظری تأیید گردید. علاوه بر این موارد، ویژگی‌های فارماکوکینتیکی ترکیبات پیشنهادی نیز با استفاده از ابزار وب رایگان Swiss-ADME محاسبه و مورد بررسی و تحلیل قرار گرفته است. در ادامه به جزییات مطالعات ذکر شده QSAR و QSRR برای هر کدام از مجموعه داده‌ها با استفاده از مدل‌های برتر معرفی شده در فصل‌های قبل پرداخته خواهد شد.

۳- ۱- ۱ تجزیه و تحلیل توصیف‌کننده‌های مدل SCAD-LM-ANN برای مجموعه بازدارنده‌های ایدز

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

به‌منظور بررسی اهمیت و میزان مشارکت هر توصیف کننده در مدل SCAD-LM-ANN درصد سهم هر توصیف کننده (C_i%) در مدل SCAD-ANN در شرایط بهینه، برآورد شد. روش کار به‌این صورت بود که برای محاسبه سهم مشارکت توصیف کننده i، مقادیر آن توصیف کننده به‌طور تصادفی در محدوده مقادیر واقعی آن‌ها تغییر داده شد. پارامترهای شبکه عصبی (گره، دور آموزش، توابع آموزش و انتقال) مدل شبکه عصبی با استفاده از یک زیر مجموعه داده با ۵ توصیف کننده به‌عنوان ورودی، در مقادیر بهینه (گزارش شده در بخش ۲- ۲- ۶) قرار گرفتند و این مدل شبکه عصبی برای محاسبه سهم مشارکت هر توصیف کننده مورد استفاده قرار گرفت. مدل شبکه عصبی در شرایط بهینه هر بار با استفاده از یک زیر مجموعه با ۵ توصیف کننده، آموزش داده شد، درحالی‌که مقادیر توصیف کننده i با مقادیر تصادفی جایگزین شده بودند. پس از آموزش شبکه و پیش بینی فعالیت دارویی ترکیبات مجموعه ارزیابی مقدار MAE مجموعه ارزیابی در حالی‌که مقادیر توصیف کننده i تصادفی شده بودند، محاسبه شد (MAE_i). این فرآیند برای هر ۵ دسته توصیف کننده تکرار شد و ۵ مقدار (MAE_i) برای مجموعه ارزیابی به دست آمد. سپس، مقادیر Ci% (i=۱،۲،...،۵) برای همه توصیف کننده‌ها با استفاده از رابطه ۱-۱۵ به دست آمد.

نتایج به‌دست‌آمده برای سهم مشارکت هر توصیف کننده در مدل SCAD-ANN در شکل ۳-۱ آورده شده است. نتایج نشان می‌دهد که سه توصیف کننده F04CN، C030 و C026 درصد مشارکت بالاتری در مدل بهینه SCAD-ANN دارند. بنابراین چگونگی اثر این توصیف کننده‌ها بر فعالیت دارویی ترکیبات مورد بررسی بیشتر قرار گرفت، تا از نتایج آن برای پیشنهاد ترکیبات جدید با فعالیت دارویی مناسب استفاده گردد.



شکل ۳-۱ نمودار سهم مشارکت توصیف کننده‌ها در مدل SCAD-LM-ANN

۳-۱-۱-۲ بررسی رابطه بین توصیف کننده‌های استفاده شده در مدل نهایی (SCAD-LM-ANN) و فعالیت دارویی ترکیبات مورد مطالعه

بررسی میزان اثر توصیف کننده‌ها بر پاسخ (فعالیت دارویی) با مقایسه ضرایب استاندارد شده مدل خطی SCAD مورد بررسی قرار گرفت. به‌این منظور ضرایب استاندارد شده توصیف کننده‌های منتخب برای مدل SCAD مطابق با رابطه ۳-۱ استخراج شدند. علامت مثبت ضرایب رگرسیون (رابطه ۳-۱) نشان می‌دهد که همه توصیف کننده‌ها اثر افزایشی بر فعالیت دارویی دارند، به این معنی که افزایش مقدار توصیف کننده مورد مطالعه باعث افزایش نسبی فعالیت دارویی ترکیب می‌شود.

$$pEC_{50} = ۰/۹۵ + ۰/۳۸F04CN + ۰/۰۹\ C\text{-}030 + ۰/۰۶C\text{-}026 + ۰/۴۹\ R6m\_A + ۰/۵۴\ R1m\_A$$ رابطه ۳-۱

با توجه به نمودار سهم مشارکت (شکل ۳-۱)، توصیف‌کننده‌های F04CN، C030 و C026 که بیش‌ترین مشارکت را در مدل ANN دارند و ضرایب این توصیف کننده‌ها در مدل SCAD (رابطه ۳-۱) مثبت است. بنابراین با افزایش مقادیر همه این توصیف کننده‌ها، فعالیت دارویی (pEC_{50}) افزایش می‌یابد. F04CN به طبقه توصیف‌کننده‌های اثر انگشت فرکانس دو بعدی[1]تعلق دارد و به تعداد C-N در فاصله هندسی ۴Å مربوط است و تعداد کربن موجود در فاصله هندسی ۴Å از اتم نیتروژن را نشان می‌دهد. بررسی عمیق‌تر ساختار ترکیبات نشان می‌دهد که در ساختارهایی با گروه‌های عاملی تقریباً یکسان، با وجود یک حلقه‌تری آزین فعالیت دارویی افزایش یابد و اغلب باعث می‌شود که مقدار توصیف کننده F04CN به‌طور قابل‌توجهی افزایش یابد. بنابراین می‌توان گفت، F04CN به حضور حلقه‌تری آزین در ساختار ترکیبات مورد مطالعه وابسته است و حضور آن در مدل QSAR با تأثیر مثبت، تأثیر وجود حلقه‌تری آزین به فعالیت دارویی را وارد مدل نهایی می‌کند.

دو توصیف کننده دیگری که بیش‌ترین سهم مشارکت را در مدل ANN توسعه یافته دارند، توصیف‌کننده‌های C-030 و C-026 با اثر مثبت بر فعالیت دارویی هستند. این دو توصیف کننده، متعلق به دسته اجزای اتم محور[2]هستند که در ادامه بیش‌تر به توضیح آن‌ها پرداخته می‌شود.

C-030 با نماد (X-CH-X) نشان‌دهنده‌اتم کربن آروماتیک متصل به گروه الکترونگاتیو (X = F، O، N، S، P و هالوژن‌ها) است. اگر ساختار دارای یک گروه الکترونگاتیو متصل به اتم کربن آروماتیک باشد، مقدار C-030 برابر با ۱ و در غیر این صورت ۰ است. در واقع، C-030 نشان‌دهنده وجود یا عدم وجود یک گروه الکترونگاتیو متصل به کربن با هیبریداسیون sp^2 است و تأثیر مثبت حضور گروه‌های الکترونگاتیو در ساختار ترکیبات مورد مطالعه بر فعالیت دارویی را در مدل وارد می‌کند.

[1]Frequency binary fingerprints descriptors
[2]Atom-centered fragments

توصیف کننده C-026 به معنی فراوانی R-CX-R در ترکیبات مورد مطالعه است. R و X به‌ترتیب نشان‌دهنده گروه آلکیل و گروه الکترونگاتیو هستند. C-026 مربوط به تعداد هتروواتم‌های الکترونگاتیو متصل به دو گروه آلکیل است. مقدار C-026 با افزایش تعداد اتم‌های الکترونگاتیو به گروه‌های آلکیل افزایش می‌یابد. ظهور C-026 در مدل نهایی با اثر مثبت نشان می‌دهد که فعالیت دارویی با افزایش تعداد گروه‌های الکترونگاتیو افزایش می‌یابد.

۳- ۱- ۱- ۳ پیشنهاد ترکیبات جدید با فعالیت دارویی مناسب با استفاده از مدل SCAD-LM-ANN ارائه شده

با استفاده از توصیف‌کننده‌های موجود در مدل برتر SCAD-LM-ANN و با توجه به رابطه ۳-۱، می‌توان دریافت که تغییر در مقادیر توصیف‌کننده‌های مهم مدل QSAR می‌تواند فعالیت دارویی را افزایش دهد. بنابراین با دانستن رابطه بین مقادیر توصیف کننده‌ها و جزئیات ساختاری ترکیبات مورد مطالعه می‌توان ترکیبات جدید با فعالیت دارویی قابل قبول پیش بینی کرد پیدا کرد. با توجه به توضیحات بخش ۳- ۱- ۱- ۲ مشاهده شد که فرکانس کربن در فاصله هندسی ۴Å از اتم نیتروژن، وجود اتم‌های کربن متصل به گروه‌های الکترونگاتیو و فراوانی هتروواتم‌های متصل به گروه‌های آلکیل، فعالیت دارویی این دسته از ترکیبات مورد مطالعه را افزایش می‌دهد. در نتیجه، با استفاده از این رویکرد، تغییرات هدفمندی در ویژگی‌های ساختاری ترکیبات مورد مطالعه ایجاد شد و ترکیبات جدیدی با فعالیت دارویی مناسب ایجاد گردید و به‌عنوان ترکیباتی با فعالیت دارویی بالقوه پیشنهاد شدند. ساختار ترکیبات پیشنهادی با استفاده از نرم‌افزار هایپرکم رسم و بهینه‌سازی شد و توصیف‌کننده‌های مدل با استفاده از نرم‌افزار دراگون محاسبه شدند. ۵ توصیف‌کننده مهم محاسبه‌شده برای ترکیبات پیشنهادی به‌صورت یک زیر مجموعه با ۵ توصیف کننده به‌عنوان ورودی در شرایط بهینه وارد مدل بهینه SCAD-LM-ANN شدند و فعالیت‌های دارویی معادل آن‌ها (pEC_{50}) پیش‌بینی شد. ساختار ترکیبات پیشنهادی و مقادیر pEC_{50} پیش‌بینی‌شده مربوط به

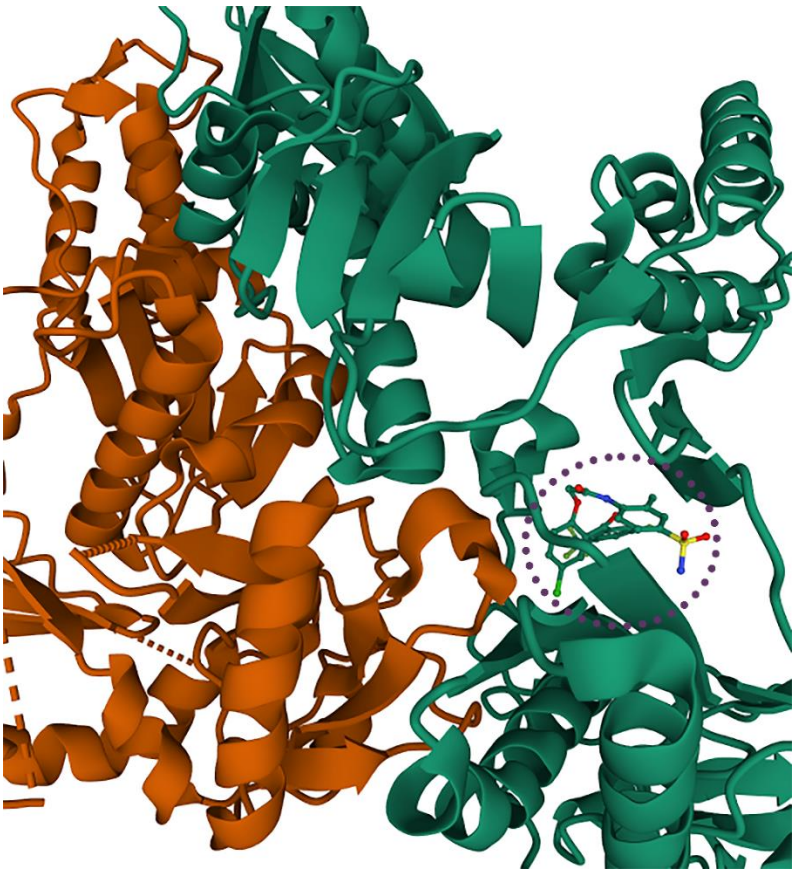
آن‌ها در جدول ۳-۱ خلاصه شده‌اند. نتایج نشان می‌دهد که مقادیر pEC_{50} پیش‌بینی‌شده در محدوده فعالیت ترکیب فعال موجود در مجموعه داده‌ها (ترکیب ۴۴ با فعالیت دارویی برابر با ۷/۷۴) می‌باشد. اگرچه مقادیر pEC_{50} پیش‌بینی‌شده برای ترکیبات پیشنهادی به اندازه‌ای بزرگ هستند که به‌عنوان ترکیبات با فعالیت بالا در نظر گرفته شوند، اما ممکن است عملاً با مولکول‌های هدف بر هم کنش مناسبی نداشته باشند و فعالیت‌های مورد انتظار را نشان ندهند. بنابراین درستی پیشنهاد این ترکیبات با به‌کارگیری یک روش مناسب اثبات شد. عملاً بهترین روش برای اثبات فعالیت بالای ترکیبات پیشنهادی، سنتز آن‌ها و تعیین فعالیت تجربی فعالیت دارویی آن‌هاست. بااین‌حال، این فرآیند زمان‌بر و پرهزینه است و به امکانات آزمایشگاهی نیازمند است که انجام آن در مطالعه حاضر میسر نبوده، بنابراین یک روش نظری برای تائید فعالیت ترکیبات پیشنهادی از طریق، مقایسه برهم‌کنش‌های گیرنده-لیگاند ترکیبات پیشنهادی با ترکیبات فعال مجموعه داده‌های مورد مطالعه با استفاده از نرم‌افزار داکینگ مولکولی مورد استفاده قرار گرفت. مراحل انجام داکینگ مولکولی، استخراج برهم کنش‌های گیرنده- لیگاند و مقایسه این بر هم کنش‌ها برای ترکیبات مختلف در ادامه آورده شده است.

۳- ۱- ۱- ۴ مطالعه داکینگ مولکولی

به‌منظور انجام عملیات داکینگ مولکولی باید مراحل متفاوتی را انجام داد که در فصل اول (بخش‌های ۱- ۷- ۱ تا۱- ۷- ۵) به آن‌ها اشاره شده است. ابتدا ساختار کریستالوگرافی گیرنده مورد استفاده برای ترکیبات مورد مطالعه به پیشنهاد مقالات شناسایی شد. گیرنده با کد 3DLG شناسایی و از سایت بانک اطلاعاتی پروتئین با پسوند pdb دانلود شد [۱۹۶]. ارزش تفکیک[1] برای ساختار کریستالوگرافی گیرنده 3DLG برابر با ۲/۲۰A° است که برای انجام مطالعات داکینگ مولکولی مناسب است [۱۹۷]. ساختار کریستالوگرافی گیرنده 3DLG و لیگاند کریستالوگرافی موجود در زنجیره A در شکل ۳-۲ نشان داده شده

[1]Resolution

است. در مرحله بعد ساختار کریستالوگرافی گیرنده 3DLG در نرم‌افزار ویورلایت ورژن ۵ فراخوانی شد و عملیات آماده‌سازی متفاوتی از جمله حذف مولکول‌های آب، حذف کوفاکتورها و حذف زنجیره اسید آمینه‌ای فاقد لیگاند کریستالوگرافی انجام شد و باقی‌مانده ساختار که شامل زنجیره اسید آمینه‌ای A و لیگاند کریستالوگرافی است که با پسوند pdb ذخیره شد. سپس زنجیره اسید آمینه‌ای به‌عنوان فایل گیرنده و ترکیب موجود در ساختار کریستالوگرافی به‌عنوان لیگاند به‌طور مجزا با پسوند pdb ذخیره شدند و به‌عنوان ورودی‌های نرم‌افزار داکینگ برای انجام مرحله اعتبار سنجی داکینگ مورد استفاده قرار گرفتند. در مرحله بعد، مختصات جایگاه فعال گیرنده با توجه به مختصات مرکز ثقل لیگاند کریستالوگرافی و با استفاده از نرم‌افزار ویورلایت با مختصات x= -۲/۲۲۰ ،y= -۳۴/۵۶۸ ،z= ۲۲/۶۷۵ A^o استخراج گردید.



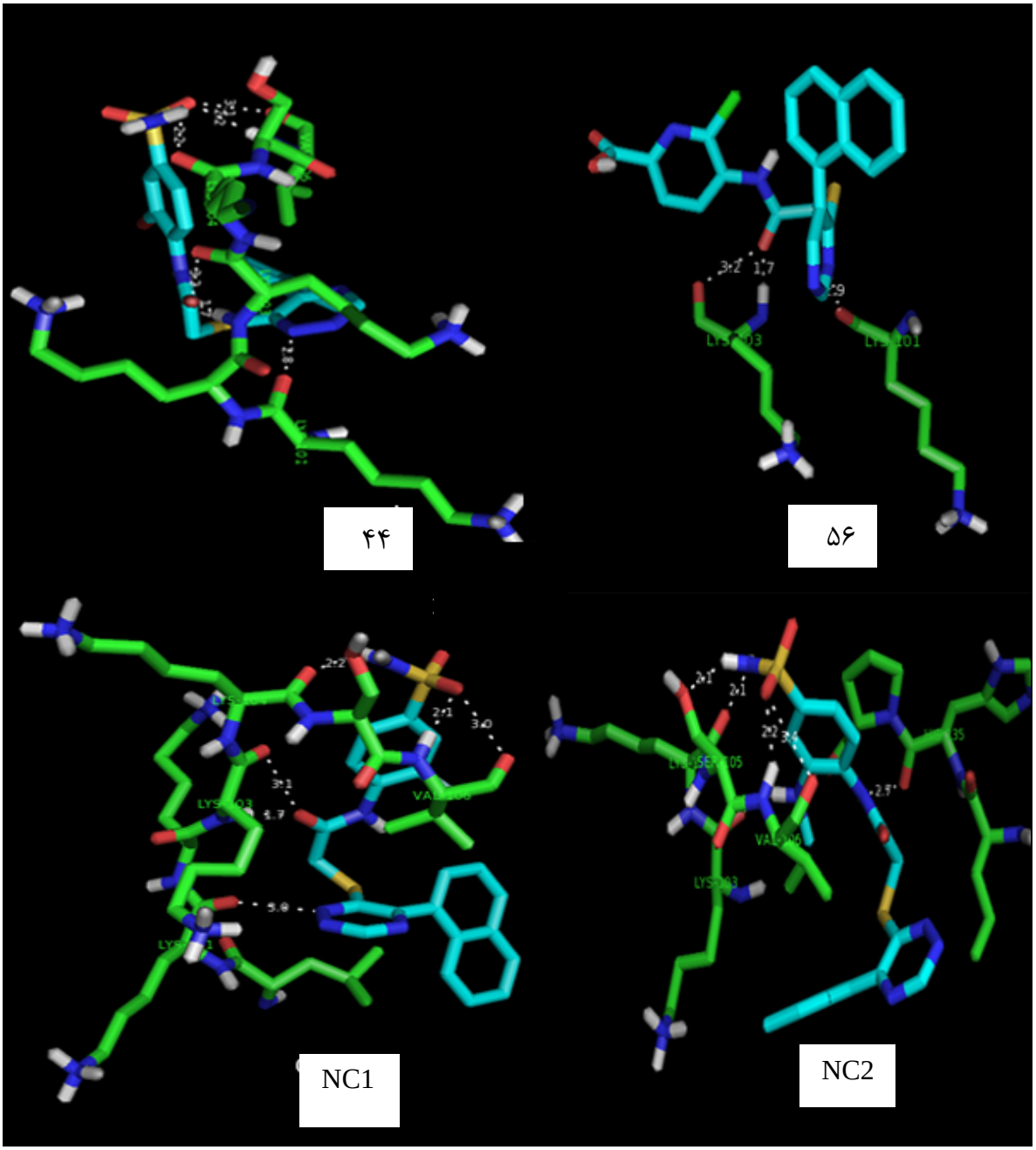
شکل ۳-۲ ساختار کریستالوگرافی 3DLG [۱۹۸] (منطقه نقطه چین نشان‌دهنده لیگاند کریستالوگرافی و مابقی زنجیره‌های اسید آمینه‌ای است)

پس از آماده‌سازی فایل‌های ورودی (زنجیره اسید آمینه حاوی ساختار لیگاند کریستالوگرافی به‌عنوان گیرنده و لیگاند کریستالوگرافی به‌عنوان لیگاند) فرایند اعتبار سنجی داکینگ مولکولی با استفاده از نرم‌افزار Autodock 4.2 انجام شد. به‌این منظور ابتدا فایل گیرنده در محیط نرم‌افزار وارد شد. آماده‌سازی‌های بیش‌تر روی ساختار گیرنده از جمله افزودن هیدروژن برای جبران کمبود هیدروژن ساختار کریستالوگرافی، ادغام هیدروژن‌های غیر قطبی متصل به هر اتم کربن و افزایش بار الکتریکی کلمن[1] انجام شد [۴۴, ۵۰, ۵۱]. در مرحله بعد به‌منظور انجام فرایند اعتبار سنجی، ساختار لیگاند کریستالوگرافی به‌عنوان لیگاند در نرم‌افزار Autodock 4.2 فراخوانی شد و با پسوند pdbqt ذخیره شد. در مرحله بعد خروجی pdbqt ساختار گیرنده نیز ذخیره شد و مختصات یک شبکه در بر گیرنده جایگاه فعال گیرنده با پیشنهاد مقالات منتشر شده در حوزه مطالعات داکینگ مولکولی، با ابعاد ۶۰ ×۶۰×۶۰ A^o و فاصله بین نقاط ۰/۳۷۵ A^o (در حدود یک چهارم طول پیوند یگانه کربن-کربن) ایجاد شد و بر هم کنش‌های گیرنده- لیگاند در این شبکه مورد بررسی قرار گرفت [۱۹۹]. داکینگ مولکولی با استفاده از اجرای الگوریتم ژنتیک لامارکین (LGA)، در تعداد اجراهای متفاوت ۱۰۰، ۱۵۰ و ۲۰۰ انجام شد. لازم به ذکر است که سایر پارامترهای الگوریتم ژنتیک در حالات پیش‌فرض پیشنهادی نرم افزار در نظر گرفته شدند. عملیات اعتبارسنجی در شرایط ذکر شده، انجام شد و فایل‌های خروجی با پسوند dlg استخراج و ذخیره شدند. با توجه به نتایج داکینگ مولکولی در مرحله اعتبار سنجی در ۳ حالت با تعداد اجراهای متفاوت الگوریتم LGA، تعداد اجرای الگئوریتم LGA برابر با ۱۵۰ کم‌ترین مقدار ریشه میانگین مربعات انحراف (RMSD) و بیش‌ترین تعداد کنفورماسیون در خوشه اول را دارا بوده و در نتیجه، برای انجام فرایند داکینگ مولکولی ترکیبات مورد نظر از تعداد اجرای الگوریتم LGA برابر با ۱۵۰ استفاده شد. در فرایند انجام داکینگ مولکولی ترکیبات مورد مطالعه، همانند فرایند اعتبار سنجی، تمامی شرایط (مختصات جایگاه فعال، ابعاد شبکه و پارامترهای پیش‌فرض الگوریتم

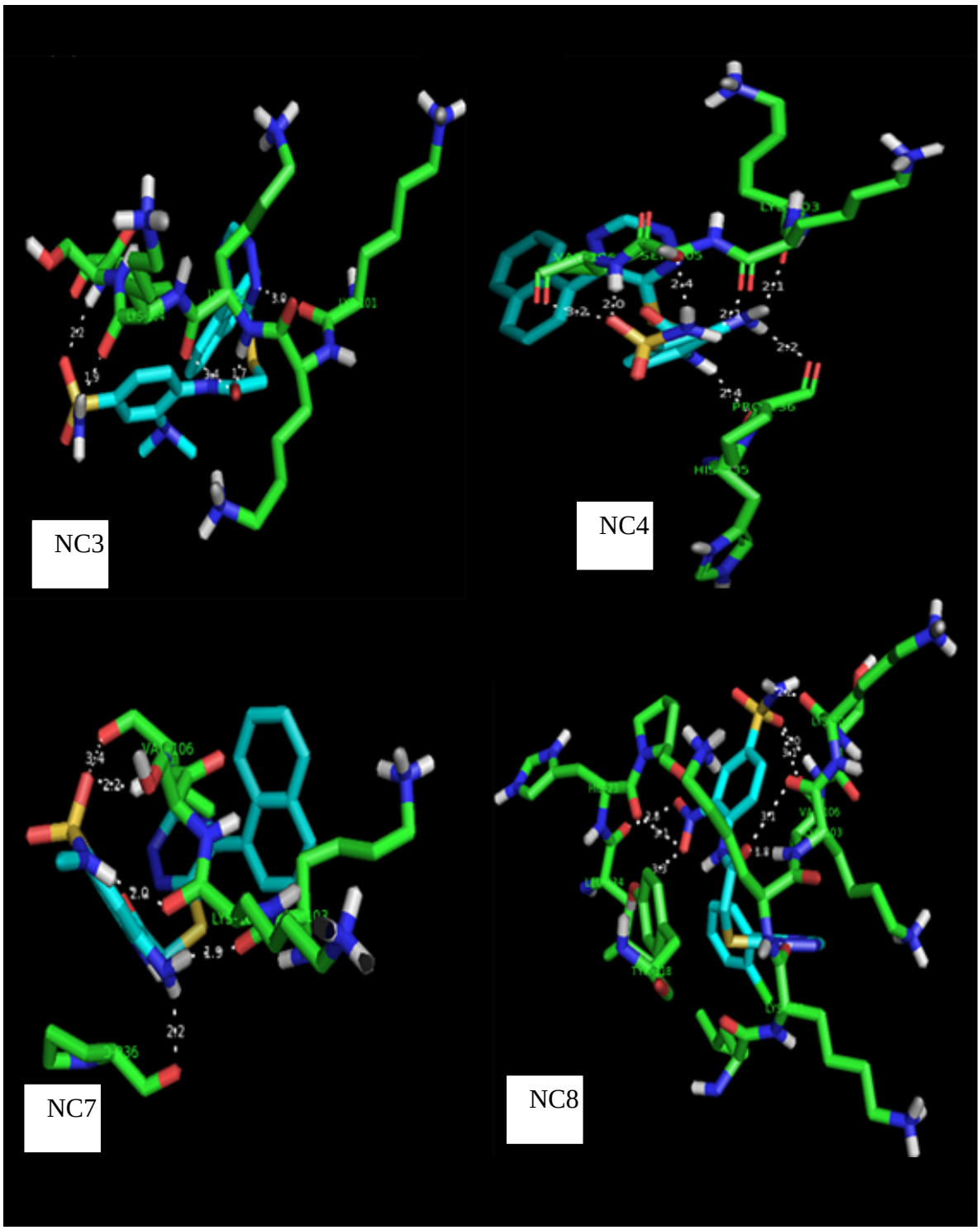
[1] Kollman

ژنتیک) در شرایط داکینگ مولکولی ترکیبات جدید نیز تعریف شد. بنابراین داکینگ مولکولی در شرایط بهینه برای ترکیبات ۴۴ (با بیش‌ترین فعالیت (pEC_{50}= ۷/۷۴)، ترکیب ۵۶ با کم‌ترین فعالیت (۴/۶۲ =pEC_{50}) و تمامی ترکیبات پیشنهادی در جایگاه فعال گیرنده انجام شدند. بهترین پیکربندی‌ها از ترکیبات مورد مطالعه و پیشنهادی با توجه به حداقل انرژی آزاد اتصال و بیش‌ترین تعداد پیکربندی در خوشه اول، شناسایی شدند و برهم‌کنش آن‌ها با اسید آمینه‌های کلیدی با استفاده از نرم‌افزار vmd مورد بررسی قرار گرفت. مطابق با شکل ۳-۳ و شکل ۳-۴ و داده‌های جدول ۳-۱، فعال‌ترین ترکیب (ترکیب شماره ۴۴) دارای برهم‌کنش‌های آب‌دوست با اسیدهای آمینه کلیدی Val106 (2.2 Å)، Val106 (3.1 Å)، Lys104 (2.2 Å، A)، Lys103 (3.2 Å)، Lys103 (1.7 Å) و Lys101 (2.8 Å) می‌باشد. درحالی‌که ضعیف‌ترین ترکیب (ترکیب شماره ۵۶) دارای تعداد برهم‌کنش‌های کم‌تر با تعداد محدودتری اسید آمینه Lys101 (2.9 Å)، Lys103 (3.2 Å) و Lys103 (1.7 Å) است (موارد داخل پرانتز طول پیوند اسیدآمینه کلیدی با یک اتم در ساختار مورد مطالعه را نشان می‌دهد). بنابراین ترکیب فعال (ترکیب شماره ۴۴) پیوندهای هیدروژنی مناسبی را با اسیدآمینه‌ها برقرار کرده است، که به‌طور معناداری از نظر بر هم کنش متفاوت از ترکیب ضعیف‌تر (ترکیب شماره ۵۶) است. این به آن معناست که بر هم کنش‌های حاصل از داکینگ مولکولی معیار مناسبی برای ارزیابی میزان فعالیت دارویی ترکیبات مورد مطالعه هستند. بنابراین برهم‌کنش‌های لیگاند- گیرنده برای تمام ترکیبات پیشنهادی استخراج و مورد تجزیه و تحلیل قرار گرفت و نتایج در شکل ۳-۳ و شکل ۳-۴ آورده شده است. نتایج به‌وضوح نشان می‌دهد که ترکیبات پیشنهادی پیوندهای هیدروژنی مناسبی را با اسیدآمینه‌های کلیدی برقرار نموده‌اند. در نتیجه فعالیت دارویی پیش بینی شده برای آن‌ها به‌وسیله مدل به‌خوبی برآورد شده‌اند. علاوه بر مطالعه داکینگ مولکولی، پارامترهای قاعده لیپینسکی نیز برای ترکیبات پیشنهادی محاسبه شدند و نتایج در جدول ۳-۱ خلاصه شده‌اند. نتایج نشان می‌دهد که تمامی پارامترهای مورد نظر از جمله وزن مولکولی (MW<۵۰۰)، چربی‌دوستی (MLOGP < ۴/۱۵)، تعداد

گیرنده‌های پیوند هیدروژنی (#H-B-acc <۱۰)، تعداد دهنده‌های پیوند هیدروژنی (#H-B-don <۵) و تعداد پیوندهای قابل چرخش (#Rot-B <۱۰) برای ترکیبات پیشنهادی دارای مقادیر قابل قبولی هستند [۲۰۰]. بنابراین ترکیبات پیشنهادی معیارهای شباهت به ترکیبات دارویی را داشته‌اند و به‌عنوان ترکیبات بالقوه فعالی که شباهت به ترکیب رهبر دارند، می‌توانند به‌عنوان کاندیدهای فعال مناسبی برای سنتز و آزمایش‌های بیولوژیکی مورد توجه قرار گیرند. علاوه بر این فاکتورها، پارامتر سهولت سنتز (Syn-Acc) نیز با استفاده از سایت Swiss-ADME [۲۰۰] استخراج شد و در جدول ۳-۱ آورده شد. فاکتور سهولت سنتز می‌تواند دارای مقداری در محدوده ۱ (سنتز بسیار آسان) تا ۱۰ (ترکیب پیچیده و چالش‌برانگیز) باشد [۲۰۱]. فاکتور سهولت سنتز برای ترکیبات پیشنهادی دارای مقدار قابل قبولی می‌باشد و در نتیجه سنتز آن‌ها در محیط آزمایشگاهی امکان‌پذیر خواهد بود.

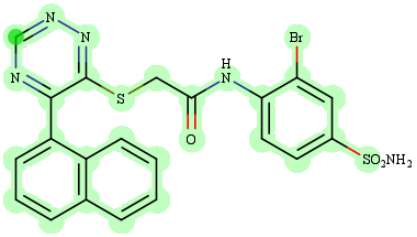
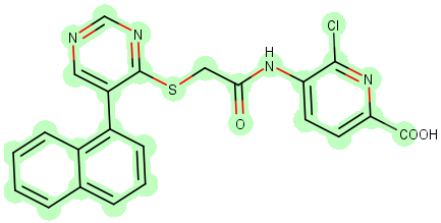
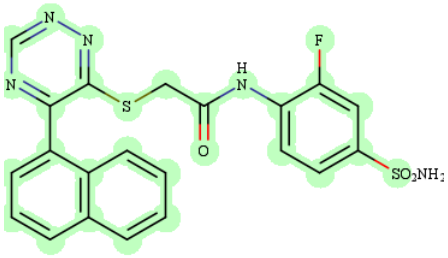
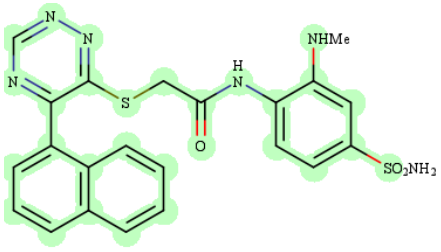
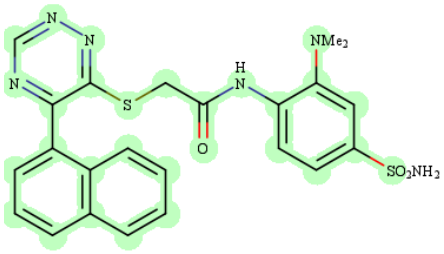


شکل ۳-۳ بررسی برهم‌کنش ترکیبات ۴۴، ۵۶،NC1 و NC2 با گیرنده

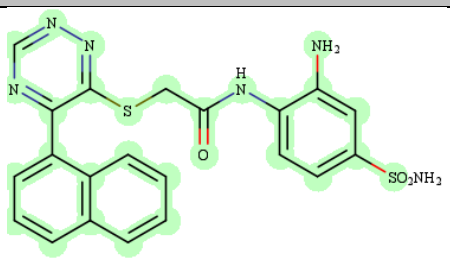
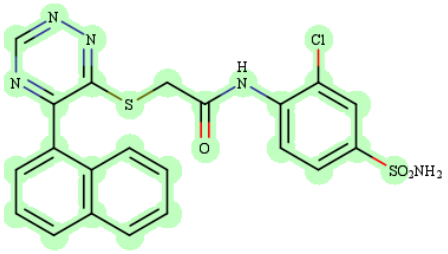
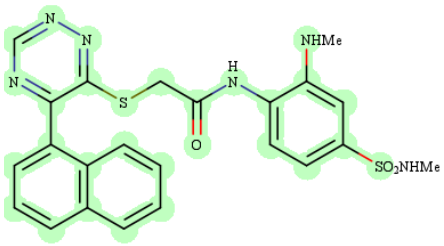
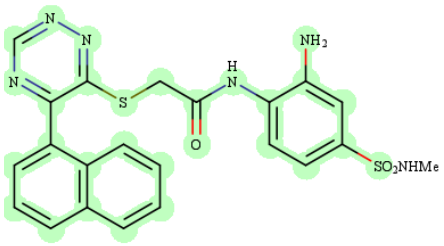
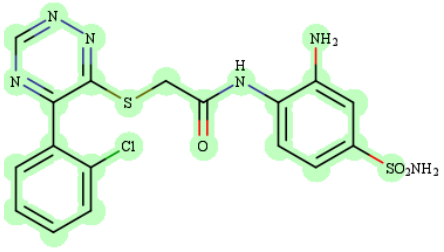


شکل ۳-۴ بررسی برهم‌کنش ترکیبات NC3، NC4، NC7 و NC8 با گیرنده

جدول ۳-۱ پارامترهای PK محاسبه شده برای ترکیبات مورد مطالعه و ترکیبات پیشنهادی

شماره ترکیب	ساختار شیمیایی	MW	MLOGP	#Rot-B	#H-B-don	#H-B-acc	Syn-Acc	pEC_{50}
۴۴		۵۳۰/۴	۱/۹۴	۷	۷	۲	۳/۳۹	۷/۷۴
۵۶		۴۵۰/۹	۰/۹۸	۷	۶	۲	۳/۲۵	۴/۶۲
NC1		۴۶۶/۵	۰/۸۴	۷	۷	۳	۳/۴۸	۷/۸۲
NC2		۴۸۶/۰	۱/۸۳	۷	۷	۲	۳/۴۲	۷/۸۱
NC3		۴۸۰/۶	۱/۰۵	۸	۷	۳	۳/۶۶	۷/۸۱

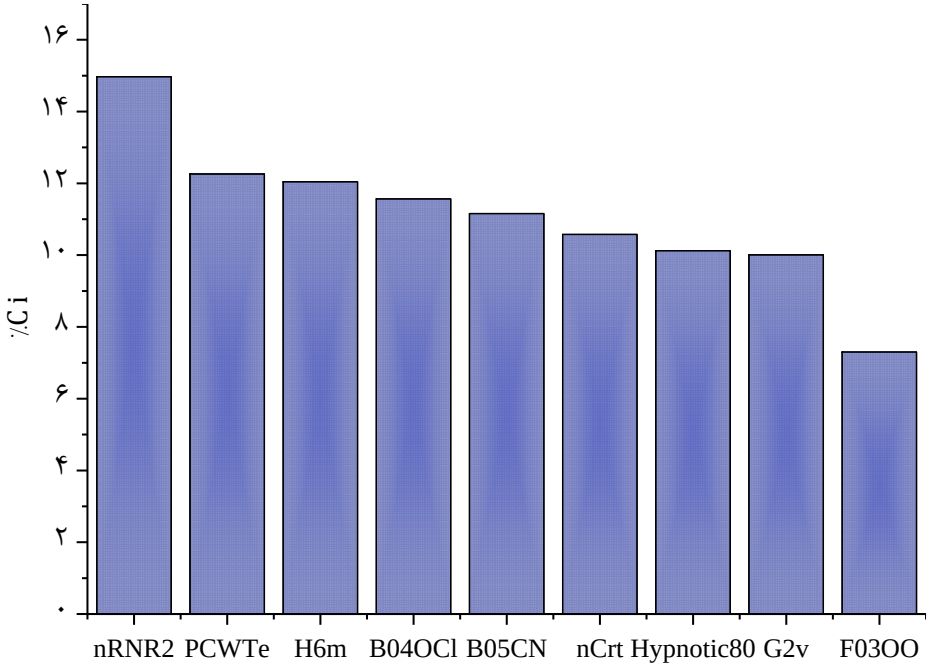
ادامه جدول ۳-۱

شماره ترکیب	ساختار شیمیایی	MW	MLOGP	#Rot-B	#H-B-don	#H-B-acc	Syn-Acc	pEC_{50}
NC4		۴۹۴/۶	۱/۲۷	۹	۷	۳	۳/۷۷	۷/۸
NC5		۴۸۰/۶	۱/۰۵	۸	۷	۳	۳/۵۹	۷/۸
NC6		۴۶۹/۵	۱/۷۲	۷	۸	۲	۳/۳۷	۷/۸
NC7		۴۹۴/۶	۱/۲۷	۸	۷	۲	۳/۷۷	۷/۷۹
NC8		۴۵۰/۹	۰/۵۹	۷	۷	۳	۳/۳۲	۷/۷۸

۳- ۱- ۲ تجزیه و تحلیل توصیف‌کننده‌های مدل ALASSO-ANN برای بازدارنده‌های SARS-COV-2

۳- ۱- ۲- ۱ محاسبه سهم مشارکت هر توصیف کننده در مدل ALASSO-ANN

مدل ALASSO-LM-ANN ساخته شده با استفاده از ۹ توصیف کننده مؤثر شامل (B04[O-Cl]، H6m، nRNR2، PCWTe، Hypnotic80، B05[C-N]، F03[O-O]، nCrt، G2v به‌عنوان مدل برتر در نظر گرفته شد. سهم هر توصیف کننده با محاسبه درصد مشارکت توصیف کننده i (C_i٪) در مدل بهینه ALASSO-LM-ANN با معماری ۱-۲-۹ در شرایط بهینه برآورد شد. به‌این منظور مقادیر توصیف کننده i در محدوده مقادیر واقعی به‌صورت تصادفی تغییر داده شد و مقادیر بقیه توصیف کننده‌ها دارای مقادیر اصلی خود بودند. مدل ALASSO-LM-ANN در شرایط بهینه با استفاده از زیر مجموعه داده‌های ایجاد شده شامل توصیف کننده i با مقادیر تصادفی و ۸ توصیف کننده دیگر با مقادیر واقعی خود آموزش داده شد و برای پیش بینی فعالت دارویی مجموعه ارزیابی به‌کار گرفته شد. با استفاده از نتایج حاصله، مشابه همین فرایند با تصادفی‌سازی تمامی توصیف کننده‌ها تکرار شد. در نهایت ۹ مقدار MAE_i به‌دست آمد. درصد سهم مشارکت هر توصیف کننده (C_i٪) طبق رابطه ۱-۱۵ محاسبه شد. نتایج به‌دست آمده برای درصد مشارکت توصیف کننده‌ها در مدل ALASSO-LM-ANN در شکل ۳-۵ آورده شده است. نتایج نشان می‌دهد که، مدل ALASSO-LM-ANN در حضور توصیف کننده‌هایی که دارای مقادیر C_i بیش‌تر می‌باشند، خطای پیش بینی بیش‌تری را در مدل بهینه ایجاد می‌کند. بنابراین این توصیف کننده بیش‌ترین مشارکت را در مدل ALASSO-LM-ANN به‌عنوان مدل QSAR برتر پیشنهادی دارند. در ادامه به بررسی میزان چگونگی تأثیر این توصیف کننده‌ها بر فعالیت دارویی ترکیبات و رابطه آن‌ها با ساختار ترکیبات مورد مطالعه پرداخته خواهد شد.



شکل ۳-۵ نمودار سهم مشارکت توصیف کننده‌ها در مدل ALASSO-LM-ANN

۳- ۱- ۲- ۲ بررسی رابطه بین توصیف کننده‌های استفاده شده در مدل نهایی (ALASSO-LM-ANN) و فعالیت دارویی ترکیبات مورد مطالعه

برای بررسی بیش‌تر چگونگی تأثیر هر توصیف کننده بر فعالیت دارویی، علامت تأثیر هر توصیف کننده بر فعالیت دارویی تعیین شد. برای این کار، ضرایب استاندارد شده مدل ALASSO با رگرسیون پاسخ (pIC_{50}) بر حسب مقادیر توصیف کننده‌های منتخب مطابق با معادله زیر به‌دست آمد:

$$pIC_{50} = ۰/۴۹ + ۰/۶۶\ B04[O\text{-}Cl] - ۰/۱۲ nRNR2 + ۰/۱ H6m - ۰/۰۹ PCWTe + ۰/۰۸ Hypnotic80$$
$$+ ۰/۰۷\ B05[C\text{-}N] + ۰/۰۴\ F03[O\text{-}O] + ۰/۰۳\ nCrt + ۰/۰۰۲\ G2V$$

رابطه ۳-۲

با توجه به رابطه ۳-۲، علامت ضرایب هفت توصیف کننده ([B04[O-Cl]، H6m، Hypnotic80، B05[CN]، F03[OO]، nCrt و G2v) مثبت بوده و اثر افزایشی بر فعالیت دارویی دارند، در حالی‌که ضرایب دو مورد دیگر (nRNR2 و PCWTe) دارای ضرایب اثر با علامت منفی بوده و تأثیر منفی بر فعالیت دارویی

دارند. با توجه به این‌که توصیف کننده‌هایی مانند B04[O-Cl]، nRNR2، F03[OO] و B05[CN] تأثیر بیش‌تری بر فعالیت دارویی دارند (شکل ۳-۵)، و به‌علاوه رابطه آن‌ها با ساختار شیمیایی ترکیبات تفسیر پذیرتر می‌باشند، این توصیف کننده‌ها برای اصلاح ساختارهای ترکیبات ضعیف به‌منظور یافتن ترکیباتی با فعالیت دارویی بیش‌تر (قوی‌تر) استفاده شدند و مهارکننده‌های جدید $3CL^{pro}$ با فعالیت دارویی (pIC_{50}) مناسب پیشنهاد داده شدند. در ادامه شرح مفصل‌تری از این توصیف کننده‌ها و چگونگی وابستگی فعالیت دارویی به آن‌ها ارائه شده است.

توصیف کننده B04[O-Cl] به طبقه توصیف‌کننده‌های اثرات انگشت دو بعدی[1] تعلق دارد و فرکانس O-Cl در فاصله مکانی ۴Å از مرکز است. مقادیر توصیف کننده B04[O-Cl] اعداد باینری (۰ و ۱) است و در ترکیبات مورد مطالعه به وجود یا عدم وجود گروه کلروپیریدین در مجاورت گروه استر بستگی دارد. در حضور کلروپیریدین، B04[O-Cl] برابر با مقدار ۱ و در غیر این صورت برابر با صفر است. توصیف کننده B04[O-Cl] با علامت تأثیر مثبت در مدل نهایی نشان می‌دهد که وجود گروه کلروپیریدین در مجاورت گروه استر در ساختار ترکیبات سبب بهبود فعالیت دارویی می‌شود. به‌عنوان مثال، ترکیب ۷۴ دارای B04[O-Cl] برابر با ۱ ($pIC_{۵۰}$ = ۷/۲۲) است، درحالی‌که ترکیب ۱ دارای B04[O-Cl] برابر با ۰ ($pIC_{۵۰}$ = ۵/۱۴) است. بنابراین با اصلاح ساختاری و ایجاد ترکیبی با ساختاری که توصیف کننده B04[O-Cl] آن دارای مقدار ۱ باشد، ترکیب پیشنهادی فعالیت دارویی مناسبی خواهد داشت. به‌این ترتیب به‌کمک تغییر در مقدار این توصیف‌کننده می‌توان بازدارنده‌های جدید $3CL^{pro}$ با ویژگی‌های فارموکوکینتیک مناسب را پیشنهاد داد.

توصیف کننده nRNR2 به طبقه شمارش گروه عاملی تعلق دارد و تعداد آمین‌های آلیفاتیک نوع سوم را شمارش می‌کند. در ترکیبات مورد مطالعه، مقدار nRNR2 با حضور بخش پیپرازین کنترل می‌شود.

[1] 2D binary fingerprints

حضور گروه پیپرازین به دلیل اتصال اتم‌های N به گروه‌های آلیفاتیک باعث افزایش مقدار این توصیف کننده می‌شود، در حالی‌که فعالیت دارویی کاهش می‌یابد (به‌عنوان مثال ترکیبات ۳۳-۳۶ و ۴۴-۴۷ با pIC_{50} زیر ۵/۰۰). از طرف دیگر، nRNR2 به دلیل داشتن تأثیر با علامت منفی حضور آن در مدل باعث کاهش فعالیت دارویی می‌شود. می‌توان نتیجه‌گیری کرد که فعالیت دارویی ترکیبات پیشنهادی در غیاب گروه عاملی nRNR2، مانند گروه پیپرازین، به‌اندازه کافی بزرگ است و از این نتیجه‌گیری می‌توان به‌عنوان یک مفهوم اساسی در طراحی بازدارنده‌های جدید $3CL^{pro}$ استفاده کرد.

توصیف کننده F03[O-O] نیز همانند توصیف کننده B04[O-Cl] به طبقه توصیف‌کننده‌های اثرات انگشت دو بعدی تعلق دارد و انعکاس دهنده حضور و فرکانس گروه O-O در فاصله مکانی ۳Å از مرکز است. با توجه به علامت مثبت ضریب تأثیر توصیف کننده F03[O-O] می‌توان گفت برای یک ترکیب هر چه مقدار فراوانی این توصیف کننده بیش‌تر باشد، فعالیت دارویی آن نیز بیش‌تر خواهد شد. با بررسی اجمالی کل مجموعه داده‌ها، می‌توان گفت که در ترکیبات با گروه فوران در مجاورت گروه کربونیل در موقعیت -β استر توصیف کننده F03[O-O] دارای مقدار ۲ (حداکثر مقدار آن) می‌باشند و این ترکیبات از دسته فعال‌ترین ترکیبات موجود در مجموعه داده‌ها هستند. به‌طور مشابه، ترکیبات با دی-اون روی حلقه و در مجاورت گروه کربونیل دارای F03[O-O] برابر با ۱ بوده و از نظر فعالیت دارویی فعال‌تر از ترکیباتی هستند که مقدار توصیف کننده F03[O-O] برای آن‌ها صفر است. بنابراین، این توصیف کننده تأثیر حضور گروه‌های دی-اون و حلقه‌های فوران مجاور گروه کربونیل را با علامت مثبت وارد مدل نهایی QSAR می‌کند.

توصیف کننده B05[C-N] توصیف کننده مهم دیگری با تأثیر مثبت بر فعالیت دارویی است. B05[CN] مربوط به حضور یا عدم حضور کربن و نیتروژن در فاصله مکانی ۵Å است. این توصیف کننده اهمیت وجود پیوند CN مانند سیانید، گروه‌های نیترو و آمین در فاصله مناسب در اطراف حلقه فنیل را نشان می‌دهد.

۳- ۱- ۲- ۳ کاربرد مدل ALASSO-LM-ANN در طراحی و پیشنهاد ترکیبات فعال با اثر ضد کووید-۱۹

یکی از کلیدی‌ترین جنبه‌های کاربردی مطالعات QSAR که اخیراً مورد توجه برخی از محققین قرار گرفته است، استفاده از مدل‌های توسعه یافته برای پیشنهاد ترکیبات جدید و در نتیجه تسهیل فرآیند طراحی ترکیبات دارویی بالقوه است. توصیف‌کننده‌های محدود موجود در مدل نهایی ALASSO-ANN می‌توانند به‌طور مؤثر نشان‌دهنده رابطه بین فعالیت دارویی ترکیبات مورد مطالعه و ویژگی‌های ساختاری آن‌ها باشد. با توجه به این‌که توصیف کننده‌های موجود در مدل ارائه شده در عین قابل تفسیرپذیری و ساده بودن، قدرت پیش‌بینی قابل قبولی نیز دارد، از مدل ارائه شده به‌همراه نوع رابطه بین ساختار و مقدار توصیف کننده‌ها برای پیشنهاد بازدارنده‌های جدید $3CL^{pro}$ با فعالیت‌های دارویی مطلوب استفاده گردید. قابل توجه است که در بین همه توصیف‌کننده‌های مؤثر، توصیف کننده‌های B04[O-Cl]، nRNR2، F03[OO] و B05[CN] دارای تفسیر و معنای شیمیایی واضح و وابسته به ساختار بوده و همچنین رابطه فعالیت ترکیبات با این توصیف کننده در مدل نهایی نمایان است. بنابراین، در طراحی ترکیبات جدید این توصیف کننده‌ها چگونگی ارتباط اصلاحات ساختارهای مورد مطالعه و تأثیر تغییرات ایجاد شده ساختاری بر افزایش فعالیت دارویی را در اختیار قرار می‌دهند. ازاین‌رو برای طراحی ترکیبات جدید با فعالیت دارویی مناسب، از معنای شیمیایی این توصیف‌کننده‌ها برای اصلاح ساختاری ترکیبات در راستای ایجاد ترکیبات با فعالیت دارویی مناسب، استفاده شد. ساختار ترکیبات پیشنهادی و مقادیر $pIC_{۵۰}$ پیش‌بینی‌شده مرتبط به آن‌ها در جدول ۳-۲ آورده شده است. سپس، ساختارهای سه بعدی ترکیبات پیشنهادی با استفاده از برنامه هایپرکم رسم و بهینه شدند و توصیف‌کننده‌های مؤثر با استفاده از نرم‌افزار دراگون محاسبه شدند. مقادیر توصیف‌کننده‌های ترکیبات پیشنهادی (به‌عنوان یک مجموعه آزمون) در مدل ALASSO-LM-ANN در شرایط بهینه به‌عنوان ورودی تعریف شدند و مقادیر فعالیت دارویی ($pIC_{۵۰}$) آن‌ها پیش‌بینی شد. با توجه به مقادیر $pIC_{۵۰}$

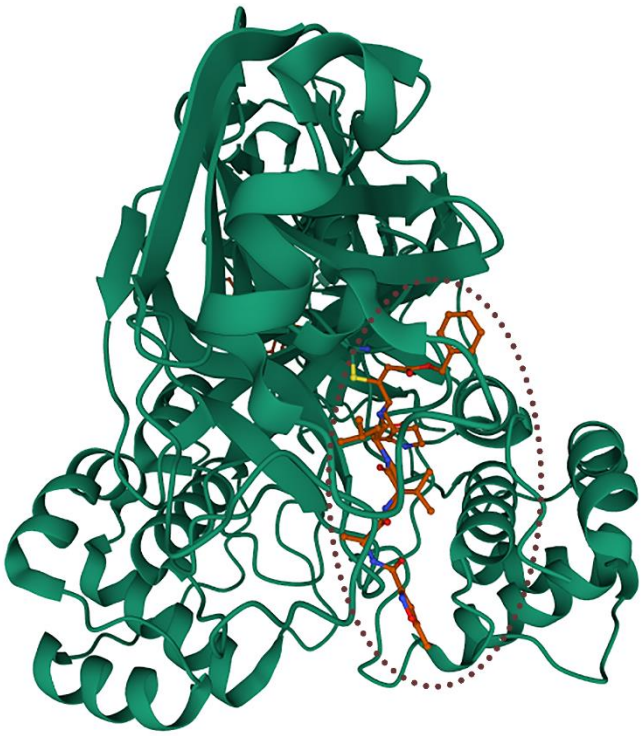
پیش‌بینی‌شده، ترکیبات پیشنهادی (NC۱ -NC۱۷) به‌عنوان بازدارنده‌های پیشنهادی بسیار فعال و ترکیبات پیشنهادی (NC۱۸ -NC۳۰) به‌عنوان بازدارنده‌های پیشنهادی نسبتاً فعال این مطالعه معرفی شدند. در ادامه برای اثبات درستی فعالیت پیش بینی شده برای ترکیبات فعال پیشنهادی، برهم‌کنش ترکیبات در جایگاه فعال گیرنده آنزیمی با استفاده از نرم‌افزار داکینگ مولکولی مورد بررسی قرار گرفت که در ادامه چگونگی انجام کار آورده شده است.

۳- ۱- ۲- ۴ مطالعه داکینگ مولکولی بازدارنده‌های $3CL^{pro}$

صحت فعالیت دارویی ترکیبات پیشنهادی با استفاده از مطالعه داکینگ مولکولی و محاسبه خواص فارماکوکینتیک (PK) با استفاده از ابزار وب رایگان Swiss-ADME مورد بررسی قرار گرفت. در این راستا، مطالعه داکینگ همه ترکیبات پیشنهادی (NC۱ -NC۳۰) و ترکیب فعال (ترکیب ۷۴) و غیرفعال (ترکیب ۷۲) مجموعه داده‌های مورد مطالعه به‌صورت جداگانه با اسید آمینه‌های کلیدی در جایگاه فعال گیرنده پروتئینی انجام شد. برای اجرای داکینگ مولکولی ابتدا، با توجه به توصیه مقالات منتشر شده، گیرنده پروتئینی با کد کریستالوگرافی با کد 6LU7 از سایت بانک اطلاعاتی پروتئین استخراج شد [۲۰۲]. ساختار کریستالوگرافی 6LU7 با ارزش تفکیک برابر با ۲/۱۶ A^o برای انجام محاسبات داکینگ مولکولی از کیفیت مناسبی برخوردار است. شکل ۳-۶ زنجیره اسید آمینه‌ای ساختار کریستالوگرافی 6LU7 و لیگاند کریستالوگرافی موجود در زنجیره آن نشان داده شده است. اجرای داکینگ مولکولی، طبق مراحل ذکر شده در بخش‌های ۱- ۷- ۱ تا ۱- ۷- ۵ انجام شد. قبل از داکینگ ترکیبات پیشنهادی در جایگاه فعال گیرنده، فرایند اعتبار سنجی داکینگ انجام شد. همانند قبل در انجام این فرایند ابتدا داکینگ لیگاند کریستالوگرافی در جایگاه فعال گیرنده اجرا شد و شرایط بهینه داکینگ استخراج گردید. به‌این منظور برای آماده‌سازی گیرنده، فایل دانلود شده 6LU7 با پسوند pdb در نرم‌افزار ویورلایت نسخه ۵ فراخوانی شد. مولکول‌های آب، کوفاکتورها، زنجیره‌های جانبی حذف شدند. ساختار باقی‌مانده (شامل زنجیره اسید آمینه اصلی و لیگاند

کریستالوگرافی)، به دو شکل ذخیره شد. یک‌بار فقط زنجیره اسید آمینه‌ای به‌عنوان ماکرومولکول ورودی نرم‌افزار Autodock4.2 و با پسوند pdb ذخیره شد و بار دیگر لیگاند کریستالوگرافی به‌طور مجزا با پسوند pdb ذخیره شد. در ادامه به‌منظور انجام فرایند اعتبار سنجی داکینگ، گیرنده در نرم‌افزار Autodock4.2 فراخوانی شد. برای آماده‌سازی بیش‌تر گیرنده، اتم‌های هیدروژن و اتم‌های هیدروژن غیر قطبی به ترتیب اضافه و ادغام شدند. سپس بار کولمن نیز برای موازنه بار سیستم اضافه شد. در مرحله بعد لیگاند کریستالوگرافی به جایگاه فعال گیرنده وارد و خروجی آن با پسوند pdbqt، ذخیره شد. سپس ساختار گیرنده اصلاح شده نیز با فرمت pdbqt ذخیره شد. در مرحله بعد، یک جعبه شبکه‌ای با ابعاد ۶۰ ×۶۰×۶۰ A^o و فاصله بین نقاط ۰/۳۷۵ A^o ایجاد شد. مرکز مختصات (x ،y ،z) جایگاه فعال با توجه به مختصات مرکز ثقل لیگاند برابر با x= -۱۰/۷۲۹ A^o, y= ۱۲/۴۱۸ A^o, z=۶۸/۸۱۶ A^o ، در نرم‌افزار Autodock4.2 تنظیم شد. سایر پارامترهای داکینگ مولکولی در مقادیر پیش‌فرض نرم افزار تنظیم شدند و فرایند داکینگ مولکولی با استفاده از الگوریتم ژنتیک لامارکین (LGA) در تعداد اجراهای متفاوت الگوریتم (۲۰۰، ۱۵۰، ۱۰۰) اجرا شد. با توجه به خروجی‌های داکینگ در ۳ اجرای متفاوت الگوریتم LGA، داکینگ با ۱۵۰ اجرا دارای کم‌ترین مقدار RMSD بود. در نتیجه این تعداد اجرا به‌عنوان تعداد اجرای بهینه برای اجرای فرایند داکینگ سایر ترکیبات در جایگاه فعال گیرنده ۶UL۷ در نظر گرفته شد. سپس به‌طور مجزا داکینگ مولکولی برای ساختار شیمیایی ترکیب شماره ۷۴ (۷/۲۲ = $pIC_{۵۰}$.)، به‌عنوان فعال‌ترین ترکیب، ترکیب شماره ۷۲ (۴/۰۰ = $pIC_{۵۰}$.)، به‌عنوان یک ترکیب با فعالیت دارویی کم و ترکیبات پیشنهادی، به‌عنوان بازدارنده‌های جدید $3CL^{pro}$، در شرایط بهینه و با پارامترهای مشابه فرایند اعتبارسنجی داکینگ انجام شد. خروجی فرآیند داکینگ مولکولی به‌صورت یک فایل با فرمت dlg ذخیره شد. اطلاعات مربوط به بهترین پیکربندی لیگاند موردنظر با توجه به کم‌ترین مقدار RMSD، کم‌ترین انرژی آزاد اتصال، کمترین تعداد خوشه و بیشترین تعداد کنفورماسیون در خوشه اول از خروجی dlg داکینگ مولکولی استخراج شد. برهم‌کنش مربوط به

بهترین پیکربندی در جایگاه فعال گیرنده برای همه ترکیبات مورد مطالعه در داکینگ مولکولی با استفاده از نرم‌افزار vmd به‌دست آمد. نتایج در شکل ۳-۷ تا شکل ۳-۱۱ آورده شده است.



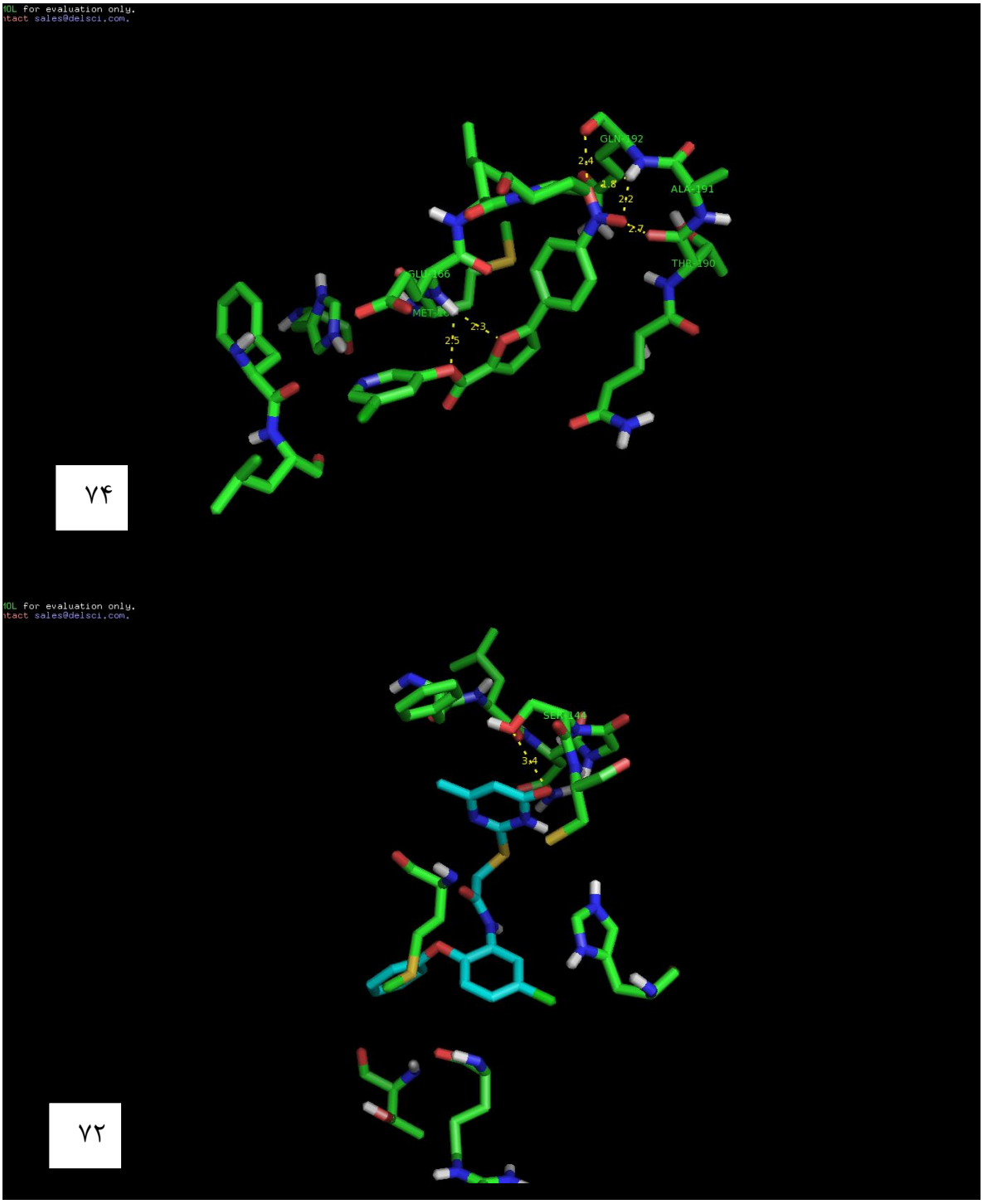
شکل ۳-۶ ساختار کریستالوگرافی 6LU7 [۱۹۸] (منطقه نقطه چین نشان‌دهنده لیگاند کریستالوگرافی و مابقی زنجیره‌های اسید آمینه‌ای است)

با توجه به مقالات منتشر شده، مشاهده شده است که بیش‌تر برهم‌کنش‌های (آب‌دوست، آب‌گریز، واندروالسی و ...) مربوط به بازدارنده‌های $3CL^{pro}$ - گیرنده، با اسید آمینه‌های Thr24، Thr26، Phe140، Asn142، Gly143، Cys145، His163، His164، Glu166، Thr24، Thr19، His19، His163، His164، Glu166، Thr19,Hisn19 برقرار شده است [۲۰۳] و نتایج همچنین با مشاهده متون اخیر مطابقت دارد. علاوه بر این، از بین اسیدآمینه‌های کلیدی نام‌برده شده، اسید آمینه‌های Ala191، Asn142، Glu166، Leu141، Phe140 پیوندهای هیدروژنی مناسبی را با بازدارنده‌های $3CL^{pro}$ برقرار می‌کنند [۲۰۴, ۲۰۵]. نتایج برهم‌کنش ترکیبات موجود در مجموعه داده‌ها و ترکیبات پیشنهادی با گیرنده برای همه ترکیبات با استفاده از نتایج داکینگ مولکولی و به‌کمک نرم افزار PyMOL استخراج شد. نتایج برهم‌کنش‌ها در شکل

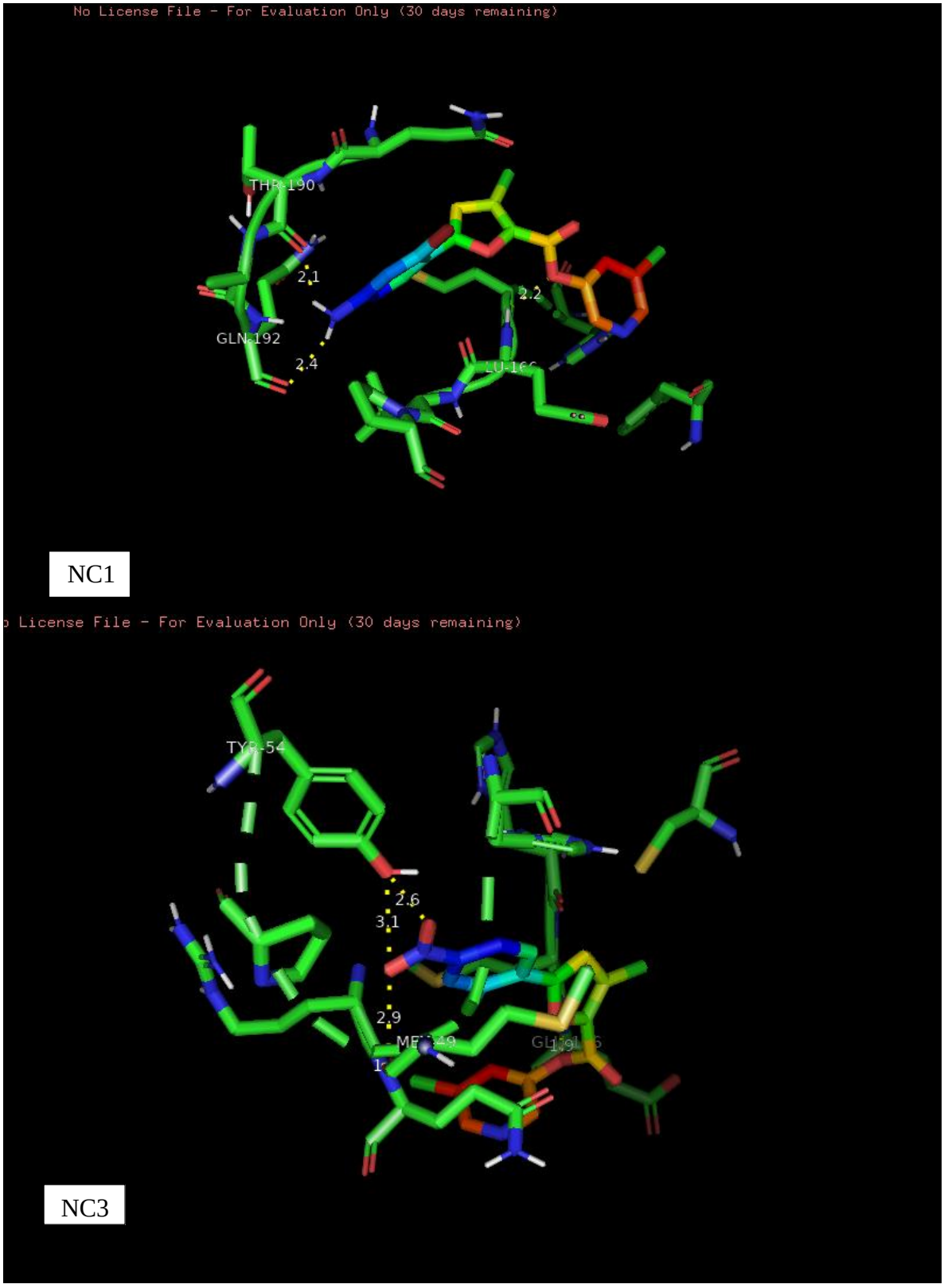
۳-۷ تا شکل ۳-۱۱ آورده شد. شکل ۳-۷ نشان می‌دهد که ترکیب ۷۴ به‌عنوان فعال‌ترین ترکیب به‌خوبی قادر به تشکیل پیوندهای هیدروژنی مناسب با Glu166، Gln192 Ala191 و Thr190 شده است. از طرفی، ترکیب ضعیف این مطالعه (ترکیب ۷۲) با کم‌ترین فعالیت دارویی تنها یک پیوند هیدروژنی با Ser144 برقرار کرده است. علاوه بر بررسی برهم‌کنش ترکیبات فعال و غیر فعال (۷۴ و ۷۲) با گیرنده، ترکیبات پیشنهادی نیز با استفاده از مطالعه داکینگ مولکولی مورد بررسی قرار گرفتند. نتایج برهم‌کنش ترکیبات فعال پیشنهادی با گیرنده (شکل ۳-۸ تا شکل ۳-۱۱) نشان می‌دهد که این ترکیبات پیوندهای هیدروژنی مناسبی را با طول پیوند قابل قبول با اسید آمینه‌های کلیدی مانند Gln192، Thr190، Glu166، His164 برقرار نموده‌اند. علاوه بر این، این ترکیبات با اسید آمینه‌های کلیدی همچون، Ala191، Leu167، Met165، Leu141، Phe140 Ser144 و Gln189 پیوندهای هیدروفوبی برقرار کرده‌اند. نتایج به‌دست آمده در خصوص برهم‌کنش‌های هیدروژنی و هیدروفوبی ترکیبات پیشنهادی با گیرنده، با نتایج گزارش شده در مقالات اخیری که داکینگ مولکولی بازدارنده‌های $3CL^{pro}$ با گیرنده 6LU7 را گزارش داده‌اند، مطابقت دارد [۲۰۳-۲۰۵]. علاوه بر این، برخی از داروهای مرجع ضد ویروسی، مانند Lopinavir و Nelfinavir، که در برابر کووید-۱۹ مؤثر هستند، نیز در مطالعات داکینگ مولکولی مورد مطالعه قرار گرفته‌اند و برهم‌کنش‌های هیدروفوبی و هیدروژنی مشابهی را نشان داده‌اند[۲۰۵]. به‌طوری‌که پیوندهای هیدروژنی با اسیدآمینه‌های کلیدی همچون Glu166، Arg188، His164، و Gln189 برقرار شده است و به‌علاوه پیوندهای هیدروفوبی با اسیدآمینه‌های کلیدی همچون His163، Asp167، Ala191، Leu167، Met165، Leu141، Phe140 Ser114، و Gln189 گزارش شده است [۲۰۵, ۲۰۶]. علاوه بر این دو داروی مرجع، در اواسط اپیدمی کووید-۱۹، داروی Remdesivir نیز به‌عنوان یک مهارکننده مناسب در برابر SARS-CoV-2 اعلام عمومی شد. برهم‌کنش‌های هیدروژنی Remdesivir با اسیدآمینه‌های کلیدی مانند Gly143، Thr24، His164، Glu166 و برهم‌کنش‌های هیدروفوبی نیز با اسیدآمینه‌های کلیدی از جمله His41 و Met49 برقرار شده

است [۲۰۷-۲۱۱]. با توجه به نتایج برهم کنش‌های مذکور در کمپلکس ترکیبات پیشنهادی – گیرنده نیز مشاهده می‌شود و فعال بودن ترکیبات پیشنهادی از حیث مطالعات داکینگ مولکولی نیز قابل توجه می‌باشد.

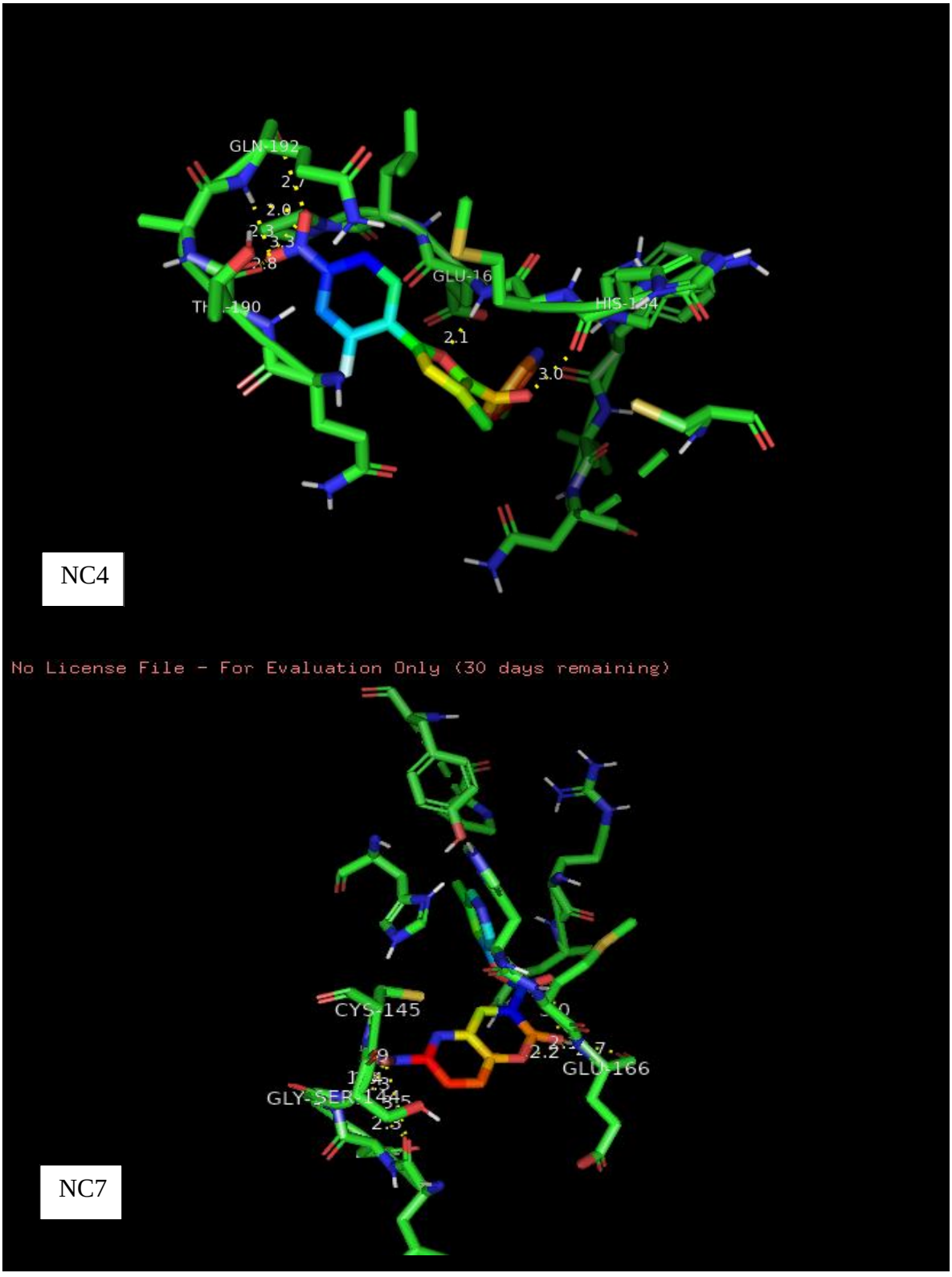
علاوه بر مطالعه داکینگ مولکولی، پارامترهای فارماکوکینتیکی (PK) ترکیبات پیشنهادی و پارامتر سهولت سنتز این ترکیبات نیز با استفاده از ابزار وب رایگان Swiss-ADME محاسبه شد. پارامترهای PK محاسبه شده برای همه ترکیبات پیشنهادی در جدول ۳-۲ خلاصه شده است. نتایج نشان می‌دهد که همه ترکیبات پیشنهادی با پارامترهای قاعده پنج لیپینسکی (وزن مولکولی (۵۰۰> MW)، چربی‌دوستی (۴/۱۵ > MLOGP)، تعداد گیرنده‌های پیوند هیدروژنی (۱۰> #H-B-acc)، تعداد دهنده‌های پیوند هیدروژنی (۵> #H-B-don) و تعداد پیوندهای قابل چرخش (۱۰> #Rot-B) مطابقت دارند. علاوه بر این، فاکتور سهولت سنتز (Syn-Acc) هر ترکیب محاسبه و در جدول ۳-۲ آورده شده است. نتایج حاصله نشان می‌دهد که همه ترکیبات پیشنهادی دارای مقادیر فاکتور سهولت سنتز کم‌تر از ۱۰ هستند. در نتیجه، امکان سنتز آزمایشگاهی این ترکیبات وجود دارد و سنتز آن‌ها در آزمایشگاه با پیچیدگی خاصی همراه نیست. به‌طور خلاصه، نتایج حاصل از فعالیت مناسب پیش‌بینی شده مدل ALASSO-ANN، بر هم‌کنش‌های بالقوه به‌دست‌آمده از داکینگ مولکولی، پارامترهای قابل قبول قاعده پنج لیپینسکی و مقادیر سهولت سنتز مناسب، اهمیت سنتز ترکیبات پیشنهادی را به‌عنوان کاندیدهای مناسبی برای بازدارنده‌های جدید $3CL^{pro}$ را تائید می‌نمایند.



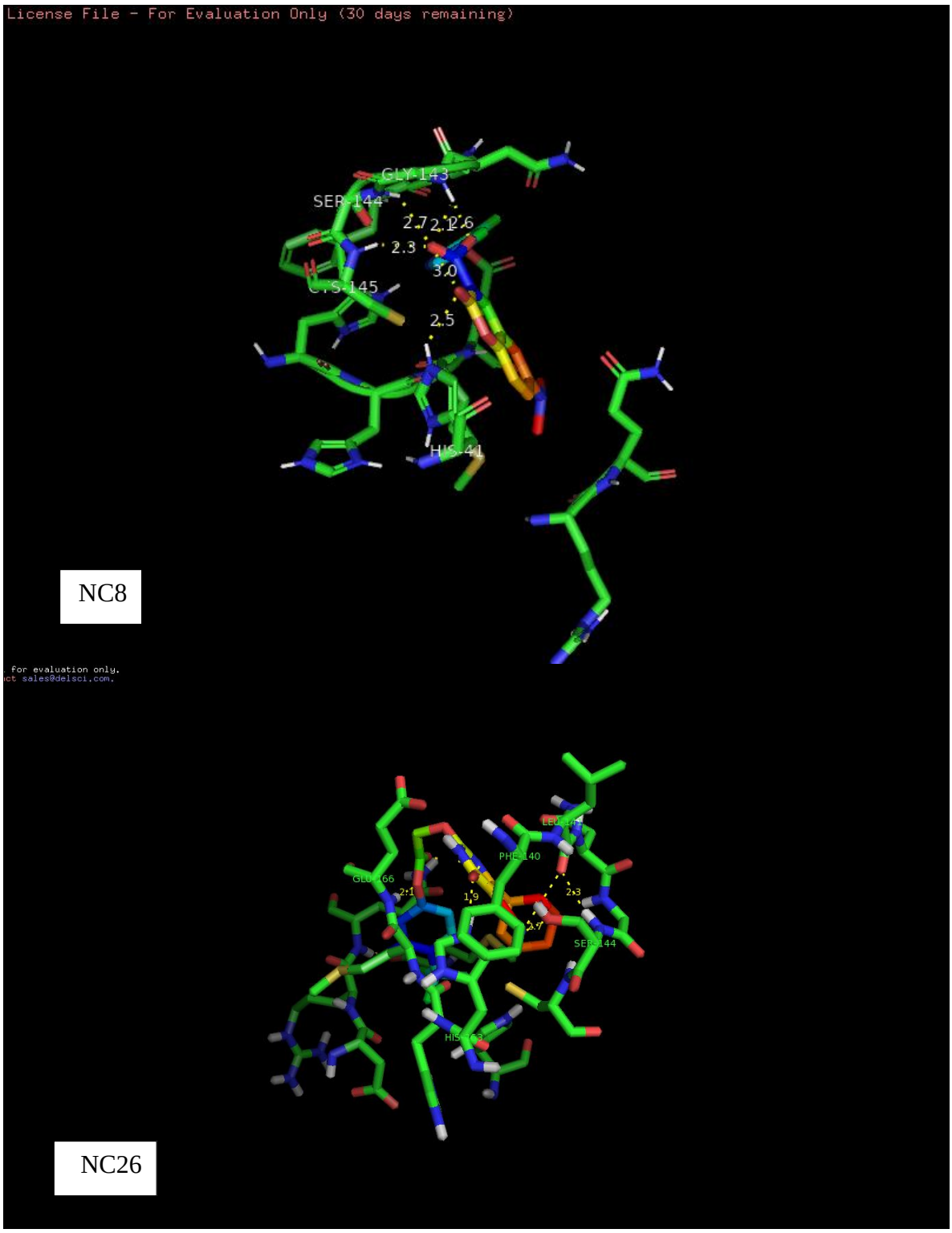
شکل ۳-۷ برهم‌کنش ترکیبات فعال (۷۴) و کم فعال (۷۲) موجود در مجموعه داده‌ها با اسید آمینه‌های کلیدی



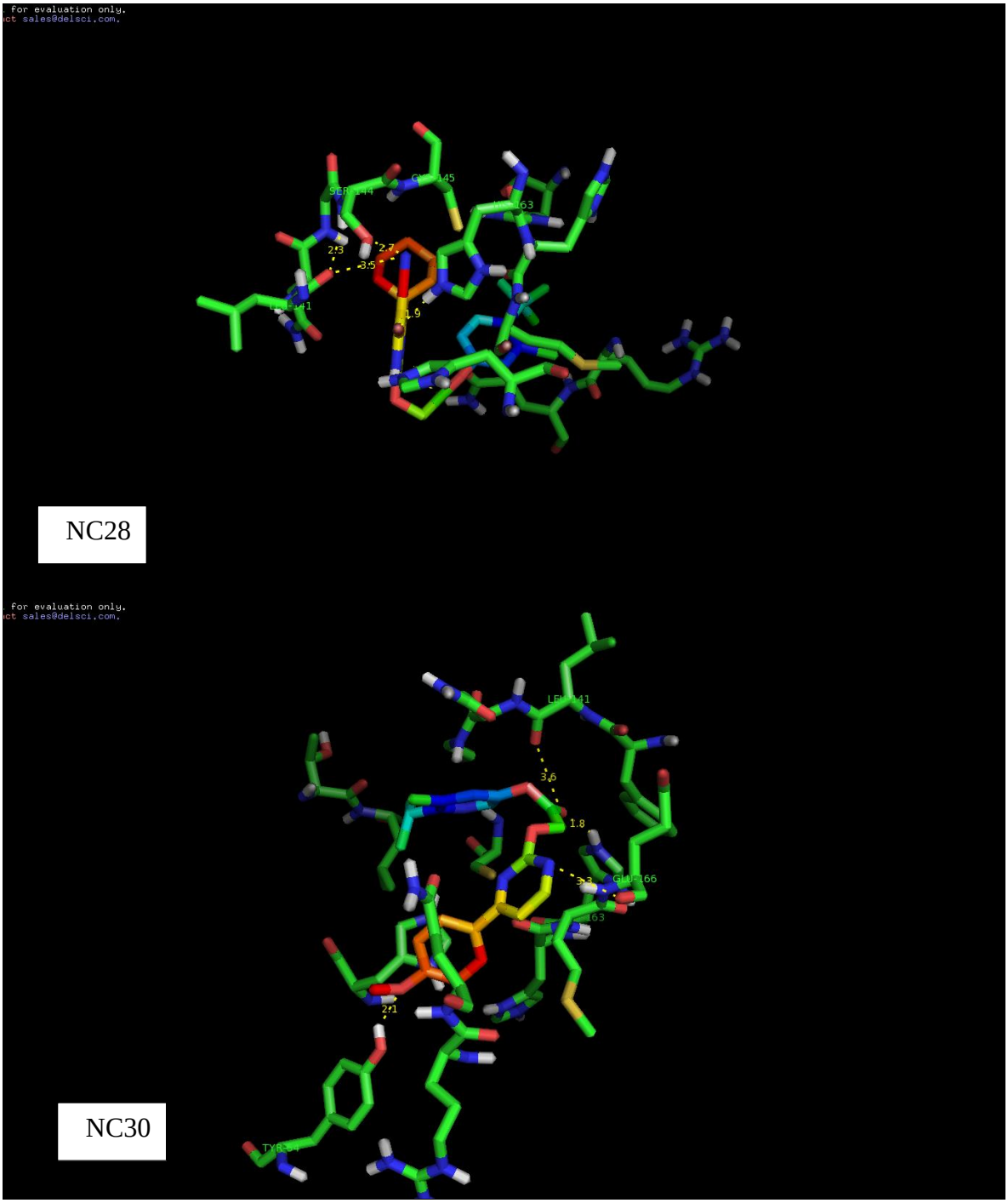
شکل ۳-۸ برهم‌کنش ترکیبات پیشنهادی (NC1 و NC3) با اسید آمینه‌های کلیدی



شکل ۳-۹ برهم‌کنش ترکیبات پیشنهادی (NC4 و NC7) با اسید آمینه‌های کلیدی

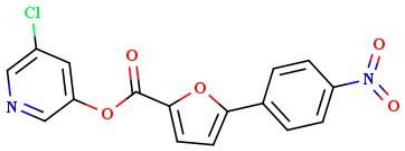
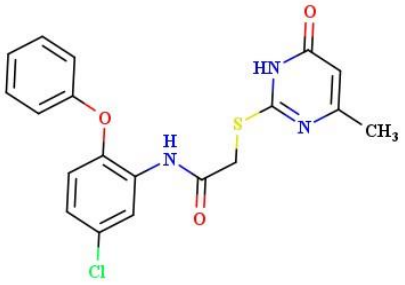
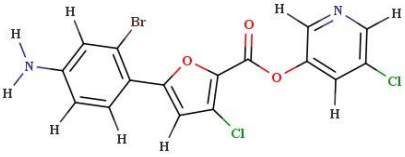
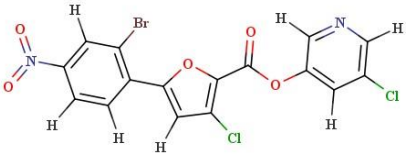
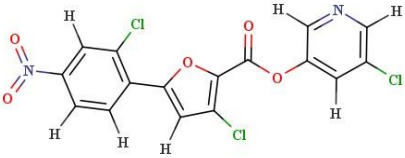
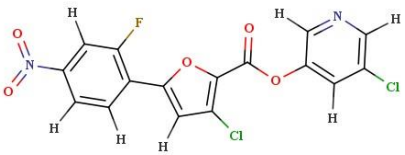


شکل ۳-۱۰ برهم‌کنش ترکیبات پیشنهادی (NC8 و NC26) با اسید آمینه‌های کلیدی

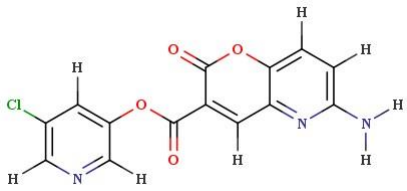
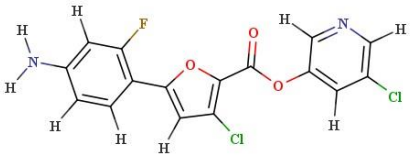
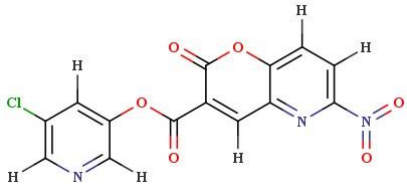
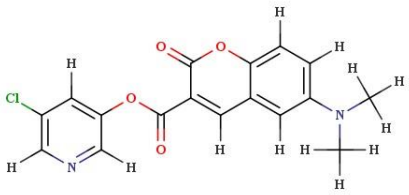
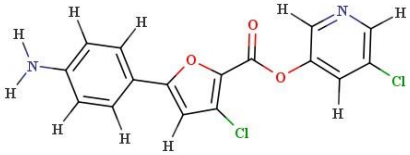


شکل ۳-۱۱ برهم‌کنش ترکیبات پیشنهادی (NC28 و NC30) با اسید آمینه‌های کلیدی

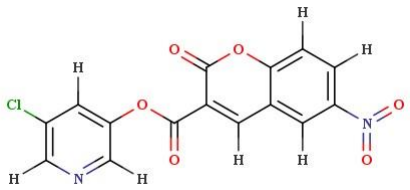
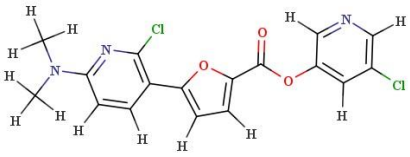
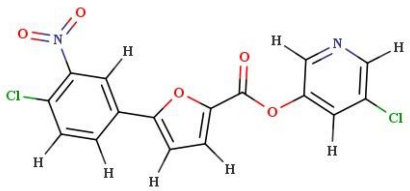
جدول ۳-۲ پارامترهای PK محاسبه شده برای ترکیبات مورد مطالعه و ترکیبات پیشنهادی

شماره ترکیب	ساختار شیمیایی	MW	MLOGP	#Rot-B	#H-B-don	#H-B-acc	Syn-Acc	pEC_{50}
۷۴		۳۴۴/۷۱	۱/۳۵	۵	۰	۶	۳/۰۳	۷/۲۲
۷۲		۴۰۱/۸۷	۲/۶۸	۷	۲	۴	۲/۸۵	۴/۰۰
NC1		۴۲۸/۰۶	۲/۹۴	۴	۱	۴	۳/۲	۶/۷۹
NC2		۴۵۸/۰۵	۲/۴۷	۵	۰	۶	۳/۲۳	۶/۷۹
NC3		۴۱۳/۶	۲/۳۶	۵	۰	۶	۳/۱۴	۶/۷۹
NC4		۴۱۵/۱۶	۱/۴۵	۵	۱	۸	۳/۲۱	۶/۷۹

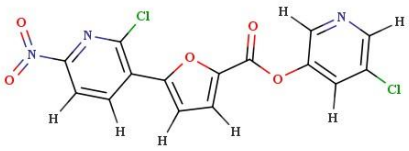
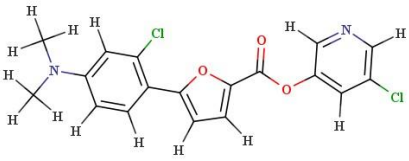
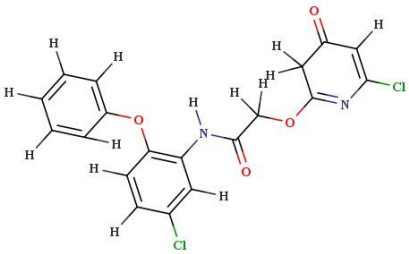
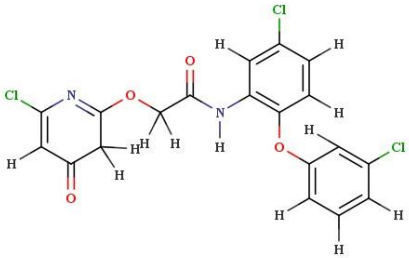
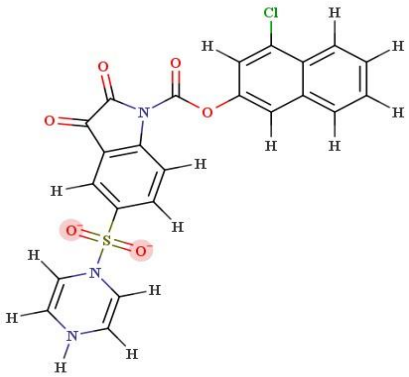
ادامه جدول ۳-۲

شماره ترکیب	ساختار شیمیایی	MW	MLOGP	#Rot-B	#H-B-don	#H-B-acc	Syn-Acc	pEC_{50}
NC5		۳۱۷/۶۸	۱/۴	۳	۱	۶	۳/۰۱	۶/۷۹
NC6		۳۶۷/۱۶	۲/۷۱	۴	۱	۵	۳/۰۹	۶/۷۷
NC7		۳۴۷/۶۷	۱/۳۸	۴	۰	۸	۳/۰۲	۶/۷۲
NC8		۳۴۴/۷۵	۲/۲۵	۴	۰	۵	۳/۰۹	۶/۷
NC9		۳۴۹/۱۷	۲/۳۳	۴	۱	۴	۳/۰۹	۶/۶۸

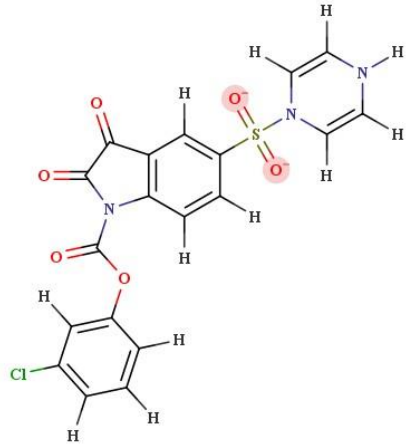
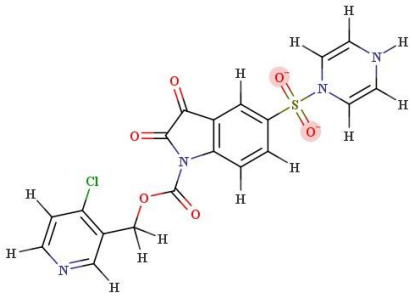
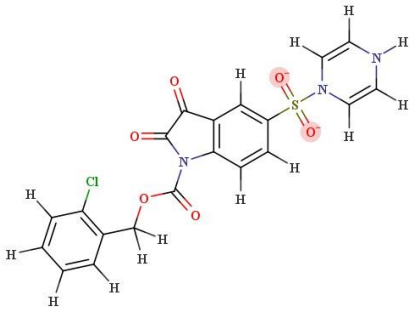
ادامه جدول ۳-۲

شماره ترکیب	ساختار شیمیایی	MW	MLOGP	#Rot-B	#H-B-don	#H-B-acc	Syn-Acc	pEC_{50}
NC10		۳۱۴/۷۲	۱/۸۲	۴	۱	۴	۳/۰۱	۶/۶۶
NC11		۴۲۵/۶۱	۳/۱۲	۴	۰	۴	۳/۱۱	۶/۶۶
NC12		۳۷۸/۶	۳/۰۱	۴	۰	۴	۲/۹۹	۶/۶۵
NC13		۳۴۶/۶۸	۱/۳۳	۴	۰	۷	۲/۹۲	۶/۶۴
NC14		۳۷۸/۲۱	۲/۱۵	۵	۰	۵	۳/۳۶	۶/۶۴
NC15		۳۷۹/۱۵	۱/۸۶	۵	۰	۶	۳/۱	۶/۶۳

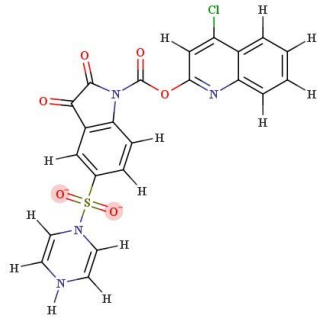
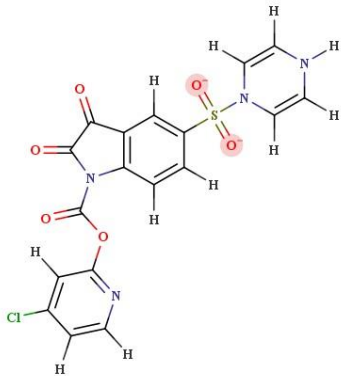
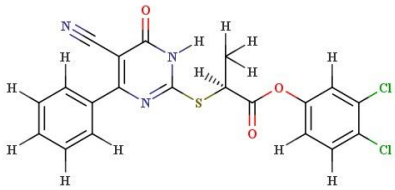
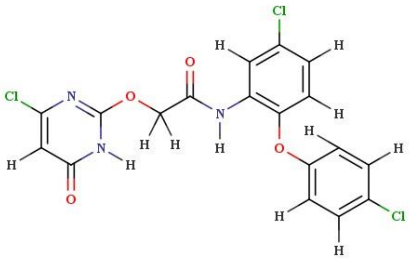
ادامه جدول ۳-۲

شماره ترکیب	ساختار شیمیایی	MW	MLOGP	#Rot-B	#H-B-don	#H-B-acc	Syn-Acc	pEC_{50}
NC16		۳۸۰/۱۴	۱/۶۴	۵	۰	۷	۳/۲۱	۶/۵۷
NC17		۳۷۷/۲۲	۲/۷۹	۵	۰	۴	۳/۲۴	۶/۵۴
NC18		۴۰۵/۲۳	۲/۳	۷	۱	۵	۳/۵۹	۵/۹۲
NC19		۴۳۹/۶۸	۲/۷۹	۷	۱	۵	۳/۶	۵/۸۱
NC20		۴۹۵/۸۹	۲/۰۲	۵	۱	۶	۳/۳۳	۵/۷۶

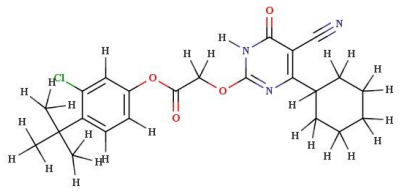
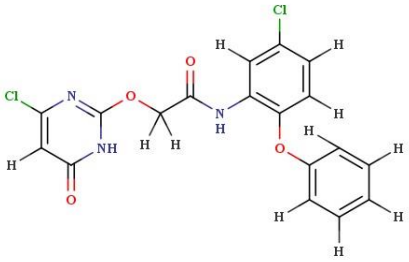
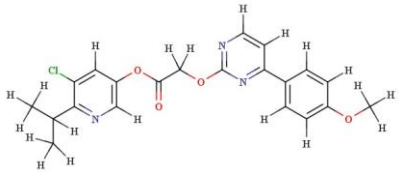
ادامه جدول ۳-۲

شماره ترکیب	ساختار شیمیایی	MW	MLOGP	#Rot-B	#H-B-don	#H-B-acc	Syn-Acc	pEC_{50}
NC21		۴۴۵/۸۳	۱/۷۳	۵	۱	۶	۳/۱۱	۵/۷۱
NC22		۴۶۰/۸۵	۰/۶۹	۶	۱	۷	۳/۱۷	۵/۶۴
NC23		۴۵۹/۸۶	۱/۶۸	۶	۱	۶	۳/۲۲	۵/۵۳

ادامه جدول ۳-۲

شماره ترکیب	ساختار شیمیایی	MW	MLOGP	#Rot-B	#H-B-don	#H-B-acc	Syn-Acc	pEC_{50}
NC24		۴۹۶/۸۸	۱/۷۱	۵	۱	۷	۳/۳۶	۵/۴۵
NC25		۴۴۶/۸۲	۱/۱۴	۵	۱	۷	۳/۱۵	۵/۴۳
NC26		۴۴۶/۳۱	۲/۷۸	۶	۱	۵	۳/۴۹	۵/۴۰
NC27		۴۴۰/۶۶	۲/۶۴	۷	۲	۵	۲/۸	۵/۳۶

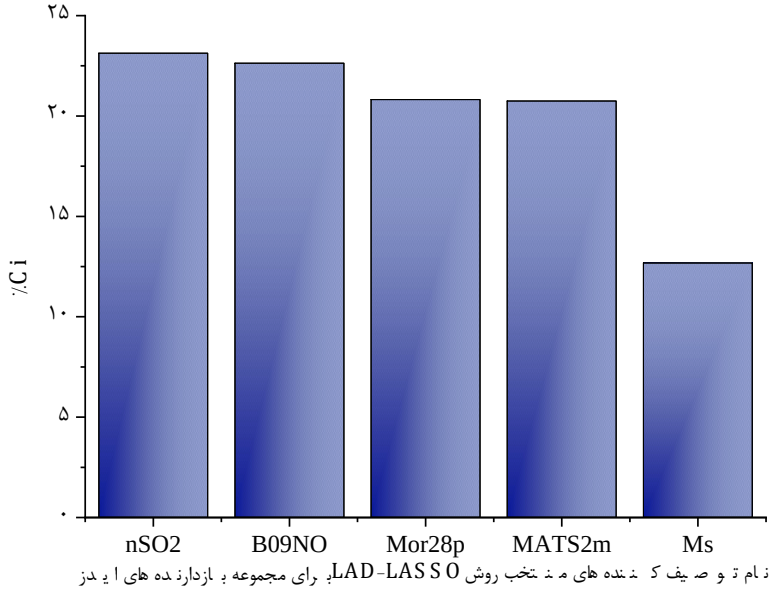
ادامه جدول ۳-۲

شماره ترکیب	ساختار شیمیایی	MW	MLOGP	#Rot-B	#H-B-don	#H-B-acc	Syn-Acc	pEC_{50}
NC28		۴۴۳/۹۲	۲/۷۶	۷	۱	۶	۳/۶۸	۵/۳۳
NC29		۴۰۶/۲۲	۲/۱۵	۷	۲	۵	۲/۸	۵/۲۹
NC30		۴۱۳/۸۵	۲/۱۳	۸	۰	۷	۳/۲۷	۵/۲۶

۳- ۱- ۳ تجزیه و تحلیل توصیف‌کننده‌های مدل LAD-LASSO-LM-ANN برای مجموعه داده‌های ضد ایدز و ضد سرطان

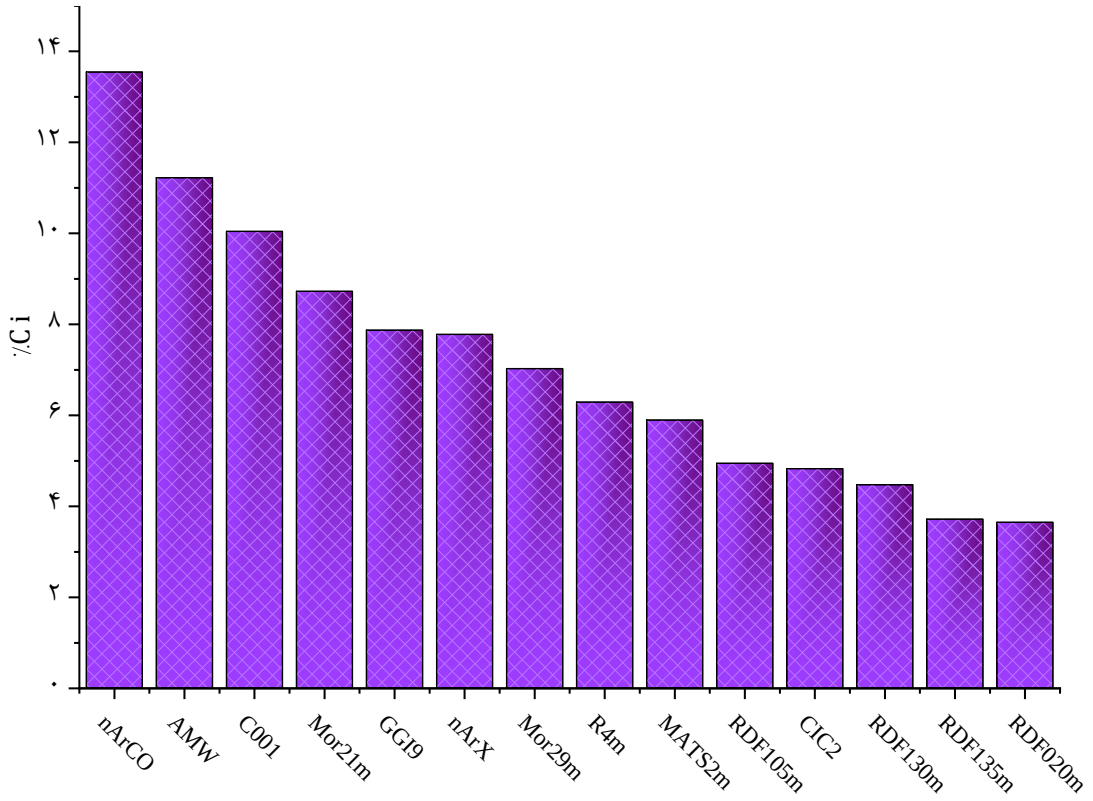
۳- ۱- ۳- ۱ محاسبه سهم مشارکت هر توصیف کننده در مدل LAD-LASSO-LM-ANN

مدل‌های QSAR برتر LAD-LASSO-LM-ANN برای مجموعه داده‌های مختلف با استفاده از توصیف‌کننده‌های منتخب (جدول ۲-۱۶) در فصل قبل ساخت و ارزیابی مدل بیان شد. میزان مشارکت هر توصیف کننده در مدل‌های LAD-LASSO-LM-ANN در شرایط بهینه برای مجموعه داده‌های مختلف، با محاسبه درصد سهم مشارکت هر توصیف کننده (C_i٪) مطابق با توضیحات مندرج در بخش۱- ۵- ۸- ۷ و بر اساس رابطه ۱-۱۵ برآورد شد. نتایج به‌دست آمده برای سهم مشارکت توصیف‌کننده‌های منتخب هر کدام از مجموعه داده‌ها در مدل LAD-LASSO-LM-ANN مربوطه در شکل ۳-۱۲ تا شکل ۳-۱۴ آورده شده است. در ادامه، با توجه به درصد سهم مشارکت توصیف‌کننده‌های منتخب در مدل‌های بهینه، به توضیح برخی از مهم‌ترین توصیف کننده‌ها پرداخته خواهد شد و کاربرد مدل‌های ارائه شده در پیشنهاد ترکیبات جدید با فعالیت دارویی مناسب بیان می‌گردد.



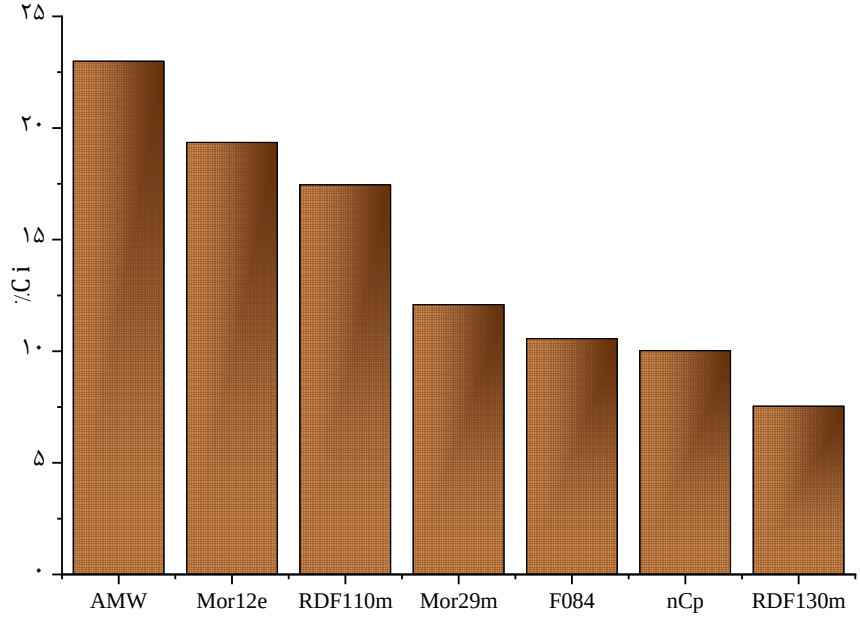
نام توصیف کننده های منتخب روش LAD-LASSO برای مجموعه بازدارنده های ایدز

شکل ۳-۱۲ نمودار سهم مشارکت توصیف کننده‌ها در مدل LAD-LASSO-LM-ANN برای بازدارنده‌های ضد ایدز



نام توصیف کننده های منتخب روش LAD-LASSO برای مجموعه بازدارنده های سرطان کولورکتال

شکل ۳-۱۳ نمودار سهم مشارکت توصیف کننده‌ها در مدل LAD-LASSO-LM-ANN برای بازدارنده‌های سرطان کارسینوم کولورکتال



نام توصیف کننده های منتخب روش LAD-LASSO برای مجموعه بازدارنده های ریه

شکل ۳-۱۴ نمودار سهم مشارکت توصیف کننده‌ها در مدل LAD-LASSO-LM-ANN برای بازدارنده‌های سرطان ریه

۳- ۱- ۳- ۲ بررسی میزان و چگونگی تأثیر توصیف کننده‌ها بر فعالیت دارویی ترکیبات مجموعه داده‌های ضد ایدز و کاربرد مدل LAD-LASSO-LM-ANN ارائه شده در پیشنهاد ترکیبات جدید

برای بررسی تأثیر هر توصیف کننده بر معادله نهایی، مدل LAD-LASSO-LM-ANN برای مجموعه داده‌های ضد ایدز استخراج شد. تأثیر افزایش و کاهش مقادیر هر توصیف کننده بر فعالیت دارویی ترکیبات با توجه به ضرایب استاندارد شده توصیف کننده‌ها (جدول ۲-۱۶) مورد بررسی قرار گرفت. در ادامه شرح مختصری از برخی از توصیف کننده‌ها، با توجه به نتایج سهم مشارکت و تأثیر ضرایب توصیف‌کننده‌های منتخب مدل‌های LAD-LASSO، آورده شده است. در نهایت برهم کنش ترکیبات پیشنهادی با گیرنده با استفاده از مطالعه داکینگ مولکولی مورد ارزیابی قرار گرفت.

از بین توصیف‌کننده‌های مدل LAD-LASSO-LM-ANN برای بازدارنده‌های ایدز، توصیف کننده nSO2 مربوط به وجود گروه‌های SO2 در ترکیبات مورد مطالعه است. ضریب اثر مثبت این توصیف کننده در مدل LAD-LASSO-LM-ANN (جدول ۲-۱۶) نشان‌دهنده این است که حضور این گروه در ترکیب موجب افزایش فعالیت دارویی ترکیب می‌شود. برای مثال دو ترکیب شماره ۱۴ و ۴۰ از نظر اثر وجود و عدم وجود گروه SO_2 مورد مقایسه قرار گرفتند. ترکیب (ساختار شماره ۱۴) با داشتن گروه SO_2 دارای فعالیت دارویی بیش‌تری (pIC_{50} = ۸/۲۶) نسبت به ترکیب فاقد گروه SO_2 (ساختار شماره ۴۰ با pIC_{50} = ۴/۵۲) دارد. توصیف کننده بعدی B09[N-O] است. این توصیف کننده مربوط به وجود/عدم وجود N–O- در فاصله مکانی ۹Å از مرکز است. ضریب اثر منفی این توصیف کننده در مدل LAD-LASSO-LM-ANN نشان‌دهنده این است که در نبود N-O- در فاصله مکانی ۹Å، ترکیبات مورد مطالعه فعالیت دارویی بیش‌تری دارند. این توصیف کننده بیان ارتباط ساختار- فعالیت ترکیباتی با گروه‌های نیترو و سیانو در مجاورت نیتروپیریدین، بنزونیتریل، پیریدازین و پیریدین در مدل ظاهر شده است. به‌عنوان مثال ترکیبات شماره

۳۶، ۵۶، ۶۴ و ۶۹ به‌دلیل مجاورت ۹Å با گروه نیترو و سیانو با حلقه‌های مذکور دارای فعالیت دارویی به‌ترتیب برابر با ۴/۷۹، ۵/۲۹، ۵/۵۵ و ۵/۸۵ هستند و از دسته ترکیباتی با فعالیت دارویی ضعیف و متوسط به‌شمار می‌آیند. توصیف کننده Mor28p مربوط به طبقه 3D-MoRSE، است [۹, ۲۱۲]. این دسته از این توصیف کننده‌ها به‌گونه‌ای تعریف می‌شوند که سهم ویژگی اتمی قطبش پذیری (p)، را در یک زاویه پراکندگی تعیین شده به ویژگی هدف منعکس می‌کنند و امکان تمایز بین ماهیت اتم را فراهم می‌کند. Mor28p، با زاویه پراکندگی $۲۸ Å^{-۱}$، توسط قطبش پذیری اتمی وزن دهی می‌شود. این توصیف کننده دارای مقادیر منفی می‌باشد و با توجه به ضریب منفی Mor28p، نشان می‌دهد که وجود برخی از توده‌های اتمی مطلوب سبب بهبود فعالیت بازدارنده‌های ایدز می‌شود [۲۱۳]. برای مثال دو ترکیب شماره ۲۰ و ۳۶ مورد مقایسه قرار گرفت. همان‌طور که گفته شد، مطابق با ضریب اثر منفی توصیف کننده بر فعالیت دارویی، ترکیب (ساختاره شماره ۲۰) با منفی‌ترین مقدار Mor28p دارای فعالیت دارویی بیش‌تری نسبت به ترکیب (ساختار شماره ۳۶) با مثبت‌ترین مقدار Mor28p است. توصیف کننده مهم دیگر MATS2m است. این توصیف کننده مربوط به طبقه‌بندی خودهمبستگی دوبعدی وزن‌دهی شده با جرم مولکولی است و اطلاعاتی در مورد توزیع جرم مولکولی در امتداد ساختار توپولوژیکی ارائه می‌دهد [۲۱۴]. این توصیف کننده با ضریب اثر مثبت، باعث افزایش فعالیت دارویی می‌شود.

Ms توصیف کننده نشان‌دهنده میانگین حالت الکتروتوپولوژیکی و از دسته توصیف‌کننده‌های شاخص‌های اساسی است. Ms دارای علامت مثبت در مدل LAD-LASSO است. بنابراین ترکیباتی که دارای مقدار بزرگ‌تری از این توصیف کننده هستند فعالیت دارویی بیش‌تری را به خود اختصاص می‌دهند. در مجموعه داده‌های بازدارنده‌های ایدز، ترکیباتی که بر روی حلقه فنیل بال راست دارای گروه‌های فلوئور هستند دارای فعالیت دارویی بیشتری نسبت به ترکیبات مشابه ولی فاقد گروه فلوئور هستند. برای مثال ترکیب شماره ۳۹ (فعالیت دارویی برابر با ۶/۱۴) دارای سه گروه فلوئور بر روی حلقه فنیل می‌باشد و در

مقابل ترکیب شماره ۳۳ (فعالیت دارویی برابر با ۵/۳۷) درای دو گروه متوکسی بر روی حلقه فنیل می‌باشد. از این‌رو بازدارندگی ترکیب ۳۵ کم‌تر از ترکیب ۳۹ می‌باشد.

یکی از مهم‌ترین کاربردهای مدل QSAR، پیشنهاد ترکیبات جدید با استفاده از توصیف‌کننده‌های مدل برتر پیشنهادی است. از این‌رو با توجه به اثر هر توصیف کننده بر فعالیت دارویی مربوطه، می‌توان به چگونگی ایجاد یک اصلاح ساختاری مؤثر که منجر به افزایش فعالیت دارویی شود، پی برد.

بنابراین با توجه به توضیح‌های ارائه شده در بخش ۳- ۱- ۳- ۱ برای مدل QSAR بازدارنده‌های ایدز، تلاش شد تا در راستای تأثیر توصیف کننده‌ها بر فعالیت دارویی و انجام برخی از اصلاحات ساختاری بر روی ترکیبات ضعیف مورد مطالعه (ترکیبات ۴۰ و ۳۶ با pIC_{50} به‌ترتیب برابر با ۴/۵۲ و ۴/۷۷)، ترکیباتی با فعالیت دارویی مناسب پیشنهاد شود. از این‌رو با توجه به تأثیر وجود گروه SO_2 و نبود گروه N - O در فاصله توپولوژیکی ۹Å، ترکیبات جدیدی با فعالیت دارویی مناسب در محدوده ترکیبات فعال، پیشنهاد شد و جزییات ساختاری آن‌ها در جدول ۳-۳ آورده شده است.

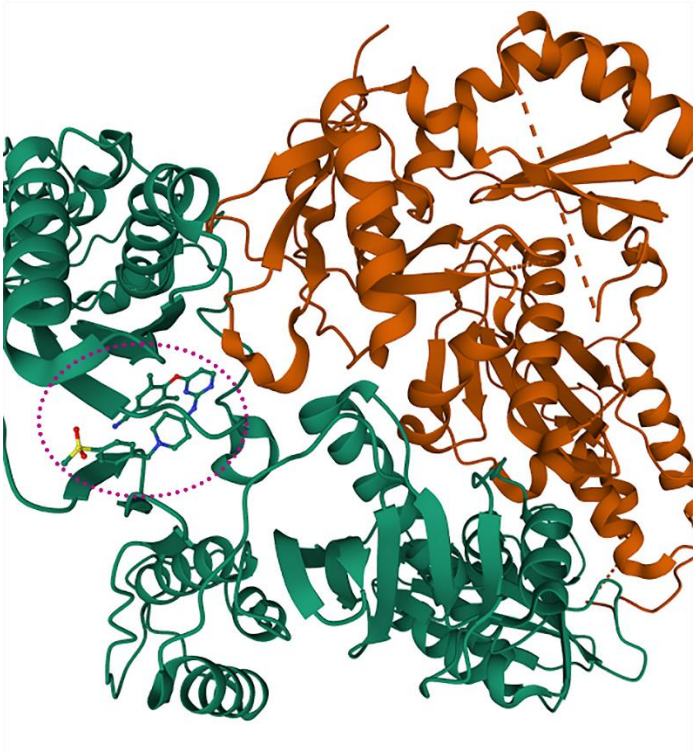
علاوه بر این ترکیبات جدیدی، با توجه به اثر توصیف‌کننده‌های مدل بهینه LAD-LASSO-LM-ANN بر فعالیت دارویی بازدارنده‌های سرطان کارسینوم کولورکتال، پیشنهاد شد. از این‌رو ترکیباتی با تعداد کتون‌های آروماتیک کم‌تر، تعداد گروه‌های هالوژنه متصل به حلقه آروماتیک بیش‌تر و تعداد اتم‌هایی با وزن اتمی بالاتر ایجاد شدند. جزییات ساختاری این ترکیبات پیشنهادی با فعالیت دارویی پیش بینی شده در جدول ۳-۴ آورده شده است. در نهایت با توجه به شرح مختصر توصیف‌کننده‌های منتخب مدل LAD-LASSO-LM-ANN برای بازدارنده‌های سرطان ریه، استنباط می‌شود که وزن مولکولی متوسط (وزن مولکولی/تعداد کل اتم‌ها) بالای ترکیبات، حضور گروه الکترونگاتیو از جمله هالوژن‌ها در ترکیبات، وجود گروه کربن Sp^3 انتهایی و عدم وجود گروه F متصل به کربن Sp^2 در ساختار ترکیب باعث افزایش فعالیت دارویی می‌شود. بنابراین با انجام برخی اصلاحات، ترکیبات جدیدی با فعالیت دارویی مناسب پیشنهاد شد

که جزییات ساختاری این ترکیبات پیشنهادی با فعالیت دارویی قابل قبول در جدول ۳-۴ آورده شده است. پس از اینکه ترکیبات مورد نظر پیشنهاد شدند ساختار هریک با استفاده از برنامه هایپرکم رسم و بهینه‌سازی شد. توصیف‌کننده‌های منتخب هر کدام از مدل‌های LAD-LASSO-LM-LM -ANN برای مجموعه داده‌های مربوطه با استفاده از نرم‌افزار دراگون محاسبه شدند. مقادیر فعالیت دارویی ترکیبات جدید با استفاده از مدل در شرایط بهینه پیش‌بینی شد و نتایج مربوط به فعالیت‌های دارویی پیش‌بینی شده ترکیبات جدید هر کدام از مجموعه داده‌ها، در جدول ۳-۴ آورده شد. در ادامه بحث، برای اثبات فعالیت ترکیبات پیشنهاد شده، از مطالعه داکینگ مولکولی استفاده شد که در بخش ۳- ۱- ۳- ۳ به‌طور کامل به آن پرداخته خواهد شد.

۳- ۱- ۳- ۳ مطالعه داکینگ مولکولی بازدارنده‌های ایدز

پس از ایجاد ترکیبات فعال پیشنهادی با فعالیت دارویی قابل قبول، برهم‌کنش‌های ترکیبات پیشنهادی با اسیدآمینه‌های کلیدی جایگاه فعال گیرنده مورد نظر نیز مورد بررسی قرار گرفت. به‌طوری‌که ابتدا ساختار کریستالوگرافی (3M8Q) به پیشنهاد مقالات منتشر شده از سایت بانک اطلاعاتی پروتئین با فرمت pdb دانلود شد[۱۷۰, ۱۷۲]. ساختار کریستالوگرافی 3M8Q با ارزش تفکیک برابر با ۲/۷۰ A^o برای انجام محاسبات داکینگ مولکولی از کیفیت مناسبی برخوردار است [۲۱۵]. شکل ۳-۱۵ زنجیره اسید آمینه‌ای ساختار کریستالوگرافی 3M8Q و لیگاند کریستالوگرافی موجود در زنجیره را نشان می‌دهد. جایگاه فعال مربوط به ساختار کریستالوگرافی بر اساس مختصات مرکز ثقل لیگاند کریستالوگرافی تعریف شد. فرایند اعتبار سنجی داکینگ مولکولی با استفاده از زنجیره اسید آمینه‌ای A و لیگاند کریستالوگرافی انجام شد. نتایج داکینگ، تعداد اجرای الگوریتم ژنتیک برابر با ۱۵۰ را به‌عنوان داکینگ بهینه مشخص کرد. بنابراین داکینگ ترکیبات جدید پیشنهادی با ساختار کریستالوگرافی مربوط به هر مجموعه داده در تعداد اجرای الگوریتم ژنتیک برابر با ۱۵۰ انجام شد. از این‌رو، ساختارهای بهینه شده ترکیبات مورد مطالعه و

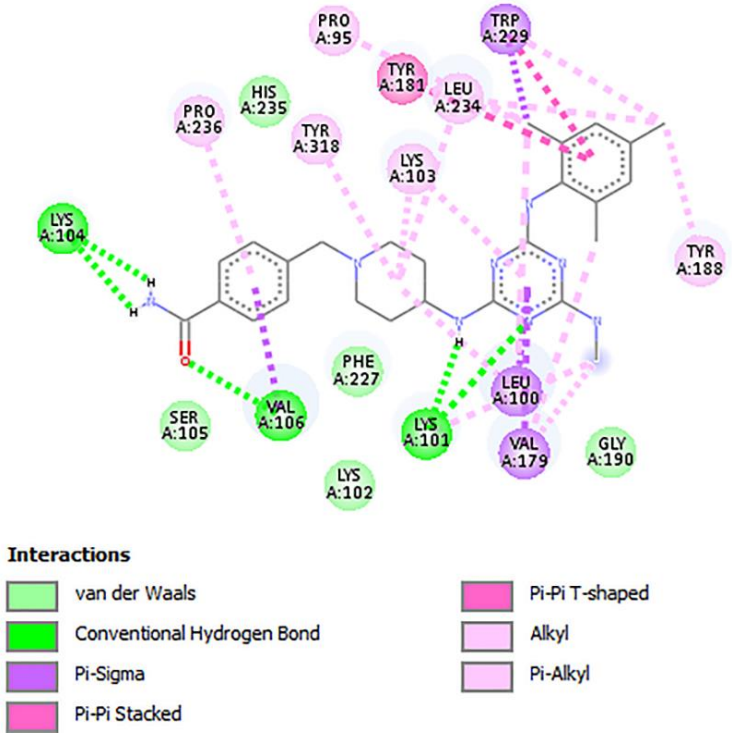
پیشنهادی با پسوند pdb ذخیره شدند و به‌عنوان ورودی لیگاند نرم‌افزار داکینگ مولکولی تعریف شدند. سپس داکینگ ترکیبات فعال و غیر فعال پیشنهادی با ساختار کریستالوگرافی مربوطه در شرایط بهینه داکینگ انجام شد. در نهایت بهترین پیکربندی مربوط به ترکیبات با کم‌ترین انرژی اتصال از نرم‌افزار داکینگ استخراج شدند و برهم‌کنش‌های لیگاند- گیرنده با استفاده از نرم‌افزار Discovery Studio Visualizer به‌دست آمد.



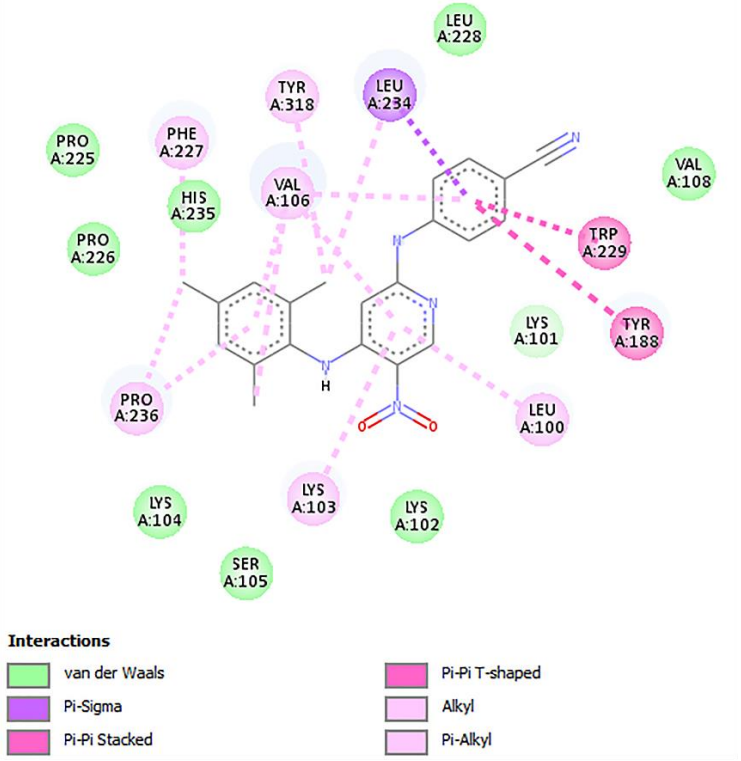
شکل ۳-۱۵ ساختار کریستالوگرافی 3M8Q [۲۱۵] (منطقه نقطه چین نشان‌دهنده لیگاند کریستالوگرافی و مابقی زنجیره‌های اسید آمینه‌ای است)

برهم‌کنش‌های متفاوت ترکیبات فعال و کم‌فعال با گیرنده در شکل ۳-۱۶ تا شکل ۳-۱۸ آورده شده است. با توجه به نتایج مشاهده می‌شود که ترکیب فعال شماره ۲۰ با pIC_{50} برابر با ۸/۳۴ با اسیدآمینه‌های متفاوتی از جمله Lys104، Val106 و Lys101 برهم کنش‌های هیدروژنی برقرار کرده است، بنابراین با توجه به توانایی ترکیب در برقراری پیوندهای هیدروژنی و هیدروفوبی متفاوت، فعالیت این ترکیب نیز اثبات می‌شود. علاوه بر این به‌منظور مقایسه این ترکیب با ترکیبات ضعیف مورد مطالعه، برهم‌کنش‌های

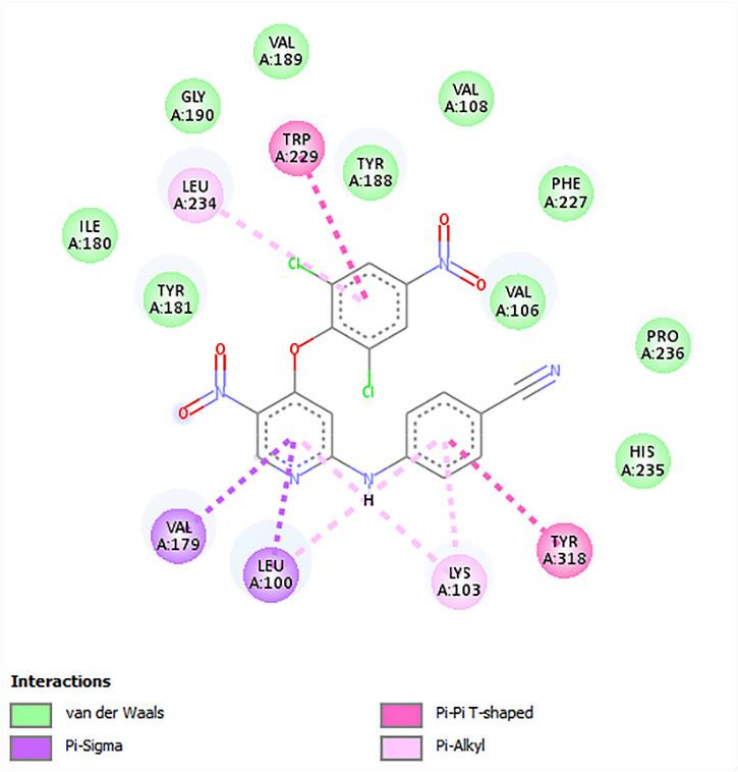
ترکیبات ضعیف ۴۰ و ۳۶ نیز مورد بررسی قرار گرفت و به ترتیب در شکل ۳-۱۷ و شکل ۳-۱۸ آورده شد. نتایج نشان می‌دهد که این دو ترکیب در برقراری پیوند هیدروژنی موفق نبوده‌اند. الگوی مربوط به برهم کنش ترکیبات مورد مطالعه برای مقایسه فعالیت ترکیبات جدید مد نظر قرار گرفتند. برهم کنش برخی از ترکیبات از جمله NC1، NC2، NC4 و NC6 در شکل ۳-۱۹ تا شکل ۳-۲۴ نشان داده شده است. با توجه به نتایج مشاهده می‌شود که این ترکیبات دارای غلظت مؤثر کم‌تر از ۱ میکرومولار هستند و در محدوده ترکیبات فعال قرار دارند. علاوه بر این‌که فعالیت این ترکیبات با استفاده از پیش‌بینی مدل LAD-LASSO-ANN LM مربوطه تأیید شد، در ادامه برهم کنش‌های این ترکیبات نیز مورد بررسی قرار گرفت. به‌طوری‌که ترکیب NC1 به‌عنوان یکی از فعال‌ترین ترکیبات پیشنهادی، برهم کنش‌های هیدروژنی مناسبی با اسید آمینه‌های کلیدی Lys101، VaL106، His235، Pro236، Tyr188 و Val179 برقرار کرده است که این نتیجه گواه مناسبی بر فعالیت چشم‌گیر این ترکیب می‌باشد. ترکیبات NC2، NC4 و NC6 نیز به ترتیب پیوندهای هیدروژنی مناسبی را با اسید آمینه‌های Leu234، Pro236، Lys103، Lys101 و Val106 برقرار کرده‌اند.



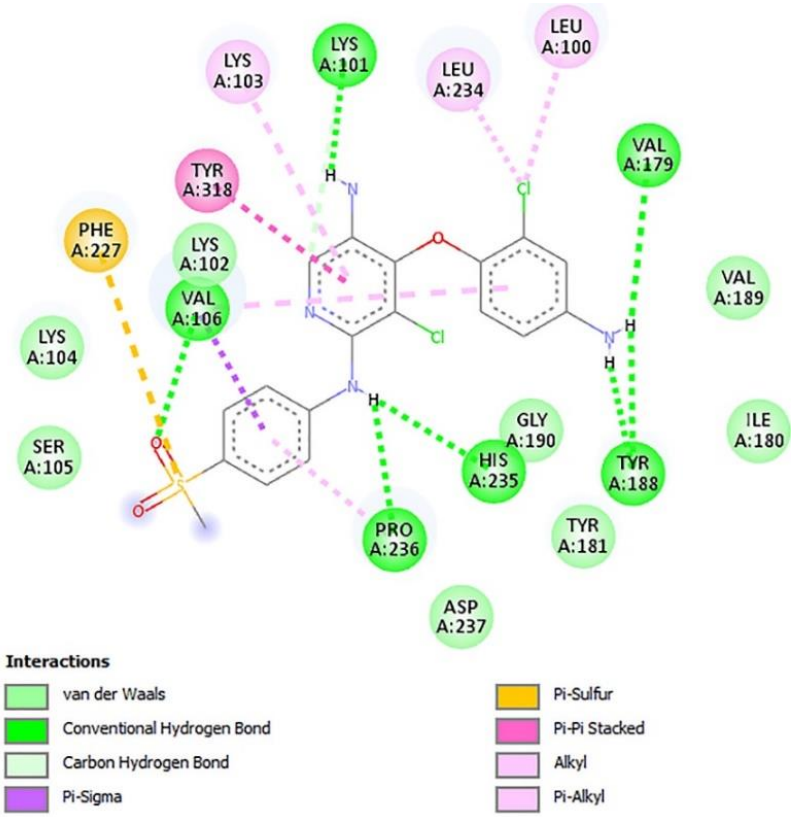
شکل ۳-۱۶ برهم‌کنش ترکیب فعال (۲۰) موجود در مجموعه داده‌های ضد ایدز با اسید آمینه‌های کلیدی



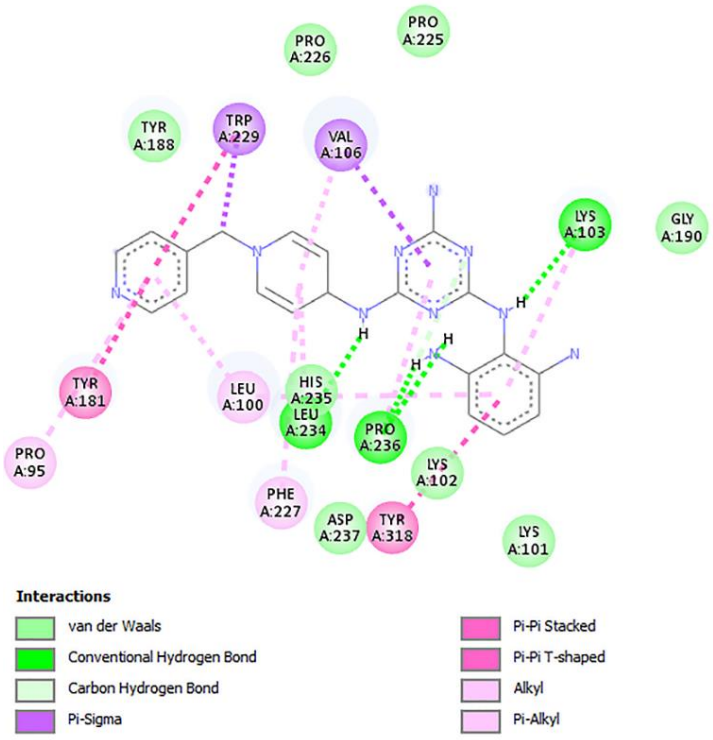
شکل ۳-۱۷ برهم‌کنش ترکیب کم فعال (۴۰) موجود در مجموعه داده‌های ضد ایدز با اسید آمینه‌های کلیدی



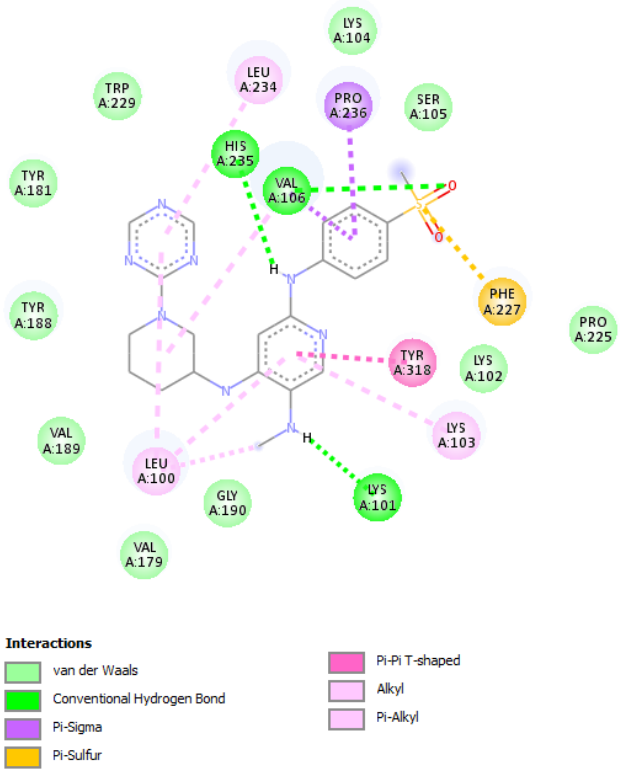
شکل ۳-۱۸ برهم‌کنش ترکیب کم فعال (۳۶) موجود در مجموعه داده‌های ضد ایدز با اسید آمینه‌های کلیدی



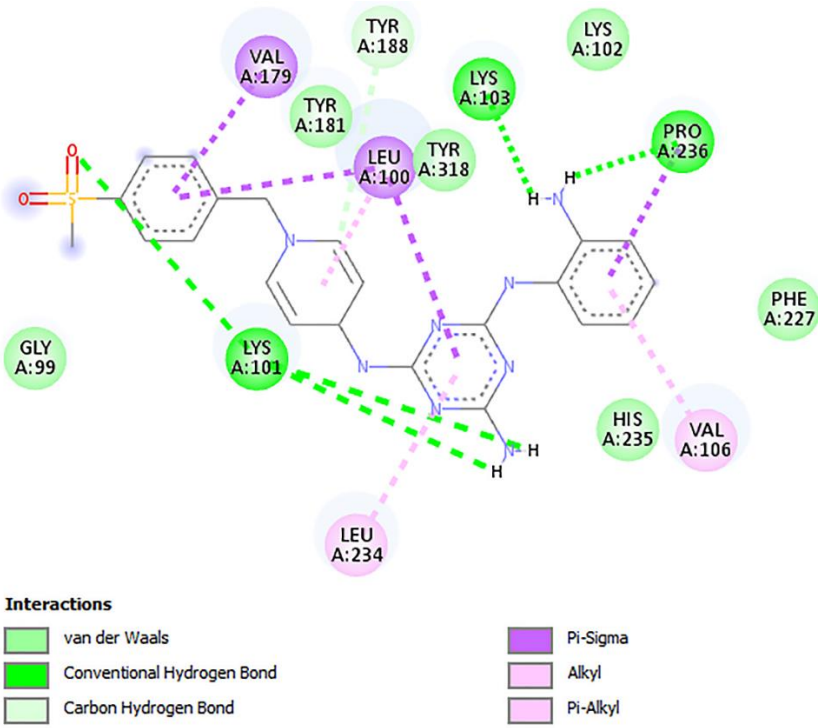
شکل ۳-۱۹ برهم‌کنش ترکیب پیشنهادی NC1 با اسید آمینه‌های کلیدی



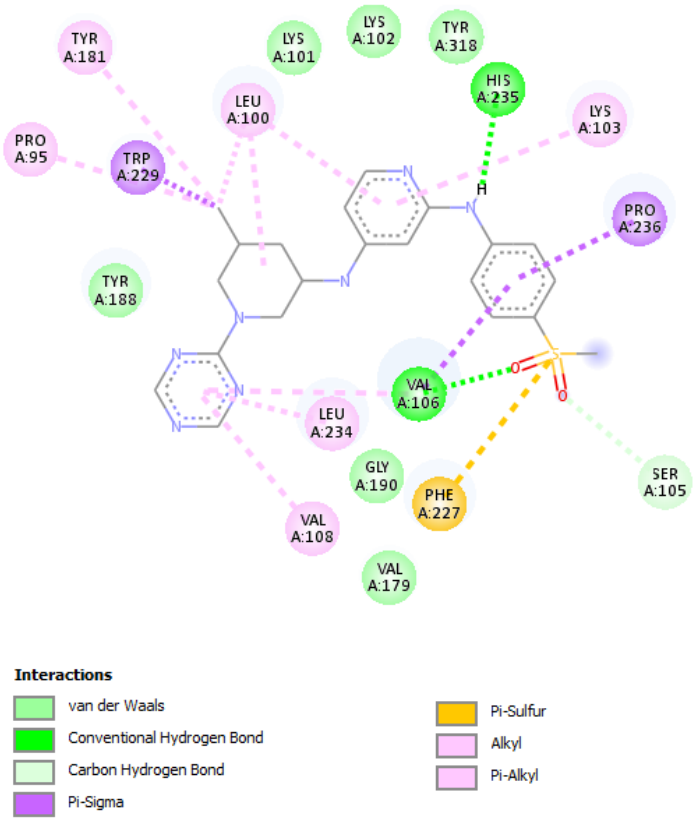
شکل ۳-۲۰ برهم‌کنش ترکیب پیشنهادی NC2 با اسید آمینه‌های کلیدی



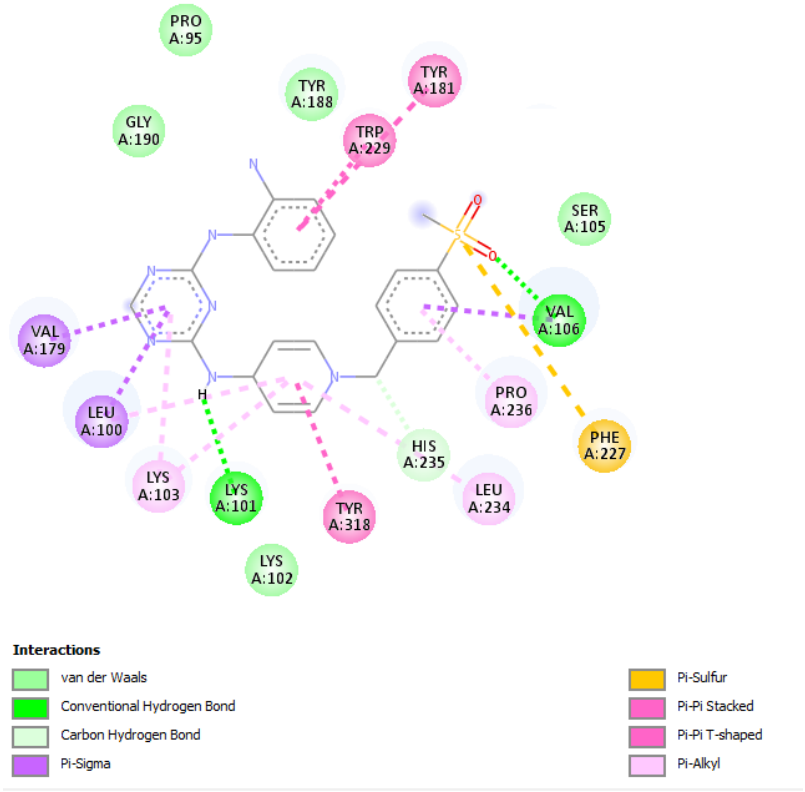
شکل ۳-۲۱ برهم‌کنش ترکیب پیشنهادی NC3 با اسید آمینه‌های کلیدی



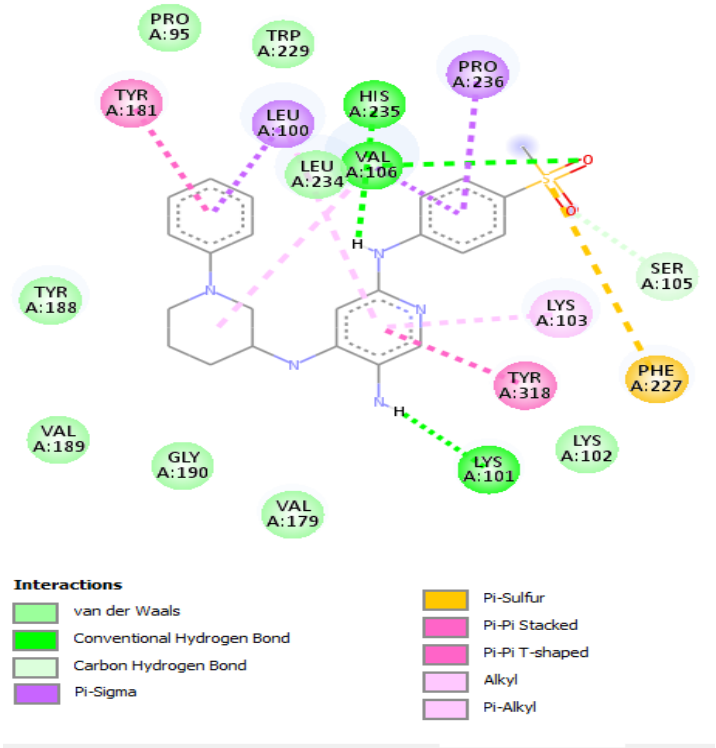
شکل ۳-۲۲ برهم‌کنش ترکیب پیشنهادی NC4 با اسید آمینه‌های کلیدی



شکل ۳-۲۳ برهم‌کنش ترکیب پیشنهادی NC5 با اسید آمینه‌های کلیدی

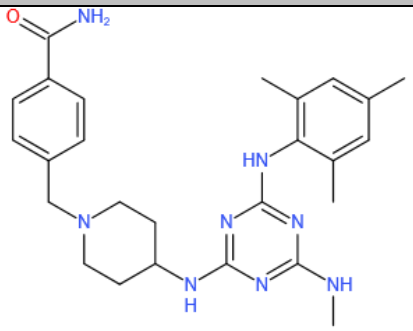
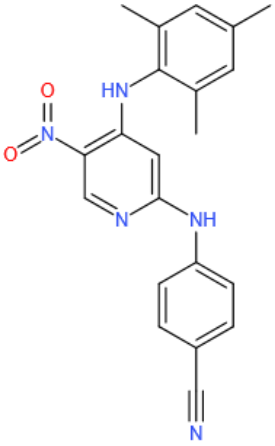
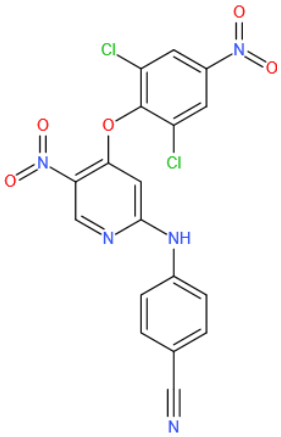


شکل ۳-۲۴ برهم‌کنش ترکیب پیشنهادی NC6 با اسید آمینه‌های کلیدی

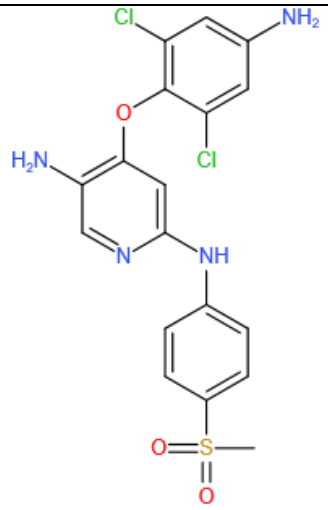
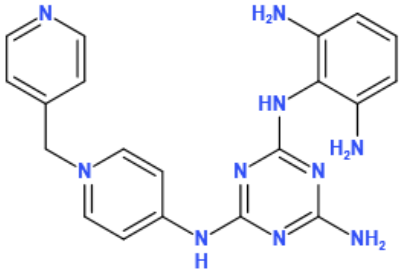
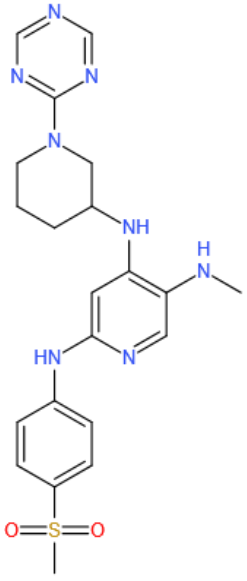


شکل ۳-۲۵ برهم‌کنش ترکیب پیشنهادی NC7 با اسید آمینه‌های کلیدی

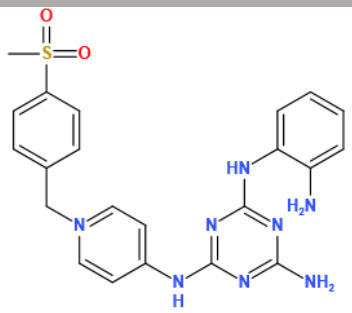
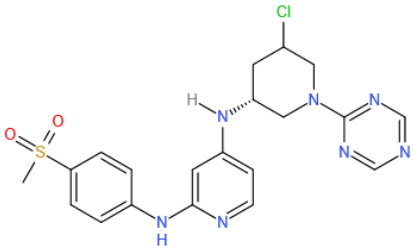
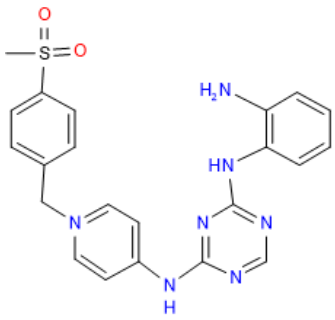
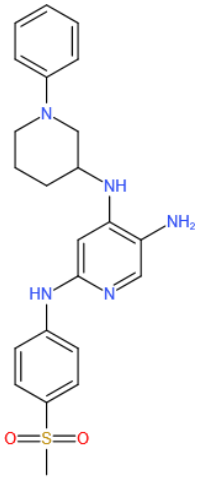
جدول ۳-۳ پارامترهای PK محاسبه شده برای ترکیبات مورد مطالعه ضد ایدز و ترکیبات پیشنهادی

شماره ترکیب	ساختار شیمیایی	MW	MLOGP	#Rot-B	#H-B-don	#H-B-acc	Syn-Acc	pEC_{50}
۲۰		۴۷۴/۶	۱/۶۶	۸	۴	۵	۳/۷۸	۸/۳۴
۴۰		۳۷۳/۴۱	۲/۱	۵	۲	۴	۳/۳۵	۴/۵۲
۳۶		۴۴۶/۲	۱/۴۹	۶	۱	۷	۳/۰۱	۴/۷۷

ادامه جدول ۳-۳

شماره ترکیب	ساختار شیمیایی	MW	MLOGP	#Rot-B	#H-B-don	#H-B-acc	Syn-Acc	pEC_{50}
NC1		۴۰۴/۸۷	۲/۰۵	۵	۳	۴	۳/۱	۸/۴۵
NC2		۴۰۶/۴۹	۰/۴۵	۶	۵	۵	۳/۴۷	۷/۶۳
NC3		۴۵۴/۵۵	۰/۸۵	۷	۳	۶	۳/۹۸	۷/۵

ادامه جدول ۳-۳

شماره ترکیب	ساختار شیمیایی	MW	MLOGP	#Rot-B	#H-B-don	#H-B-acc	Syn-Acc	pEC_{50}
NC4		۴۶۴/۵۴	۰/۴۵	۷	۴	۵	۴/۳۴	۷/۲۳
NC5		۴۵۹/۹۵	۱/۳۵	۶	۲	۶	۴/۱۷	۷/۱۱
NC6		۴۵۳/۵۶	۱/۱۶	۷	۳	۶	۳/۵۴	۶/۹۴
NC7		۴۳۷/۵۶	۲/۴	۶	۳	۳	۳/۸۹	۶/۶۳

علاوه بر بررسی برهم کنش‌های متفاوت هیدروژنی و هیدروفوبی با استفاده از داکینگ مولکولی، پارامترهای قاعده لیپینسکی نیز با استفاده از ابزار وب رایگان Swiss-ADME مورد ارزیابی قرار گرفت. همان‌طور که نتایج جدول ۳-۳ نشان می‌دهد همه ترکیبات پیشنهادی دارای مقادیر قابل قبول می‌باشند و از این نظر خواص فارماکوکینتیکی نیز دارای معیارهای مناسب هستند. پارامتر Syn-Acc نیز نشان می‌دهد که سنتز این ترکیبات در آزمایشگاه امکان‌پذیر است و از درجه آسانی مناسبی برخوردار است.

۳- ۱- ۳- ۴ بررسی میزان و چگونگی تأثیر توصیف کننده‌ها بر فعالیت دارویی ترکیبات مجموعه داده‌های ضد سرطان کارسینوم کولورکتال و ریه و کاربرد مدل LAD-LASSO-LM-ANN ارائه شده در پیشنهاد ترکیبات جدید

-اثر توصیف کننده‌ها بر فعالیت دارویی با استفاده از مدل LAD-LASSO-LM-ANN برای بازدارنده‌های سرطان کارسینوم کولورکتال

با توجه به نتایج جدول ۲-۱۷، مدل LAD-LASSO-LM-ANN برای بازدارنده‌های سرطان کارسینوم کولورکتال، با ۱۴ توصیف کننده در مدل بهینه مشارکت دارند. در ادامه، به شرح مختصری از برخی از توصیف‌کننده‌های منتخب با تفسیرپذیری بیش‌تر پرداخته خواهد شد. توصیف کننده nArCO در طبقه گروه‌های عاملی قرار دارد و مربوط به تعداد کتون‌های آروماتیک می‌باشد. علامت منفی این توصیف کننده در مدل LAD-LASSO-LM-ANN (جدول ۲-۱۶) نشان می‌دهد که حضور این توصیف کننده سبب کاهش مقدار فعالیت دارویی می‌شود. با توجه به ساختارهای مورد مطالعه در مجموعه بازدارنده‌های سرطان کارسینوم کولورکتال، دیده می‌شود که گروه nArCO در ترکیبات شماره ۲۶ تا ۳۳ (فعالیت دارویی ضعیف‌تر با $pIC_{۵}$ کوچک‌تر از ۵/۰۰) وجود دارد که این موضوع نشان‌دهنده اثر منفی وجود گروه‌های کتون آروماتیک بر فعالیت دارویی بوده که به‌وسیله توصیف کننده nArCO وارد مدل شده است. توصیف کننده

nArX (X گروه هالوژن است) از طبقه شمارش گروه‌های عاملی می‌باشد. این توصیف کننده تعداد گروه‌های عاملی هالوژنه متصل به حلقه آروماتیک را نشان می‌دهد. ضریب تأثیر مثبت این توصیف کننده نشان می‌دهد که هرچه تعداد گروه‌های هالوژنه متصل به حلقه آروماتیک بیش‌تر باشد، فعالیت دارویی ترکیب بیش‌تر می‌شود. توصیف کننده AMW از دسته توصیف‌کننده‌های ساختاری می‌باشد. این توصیف کننده نشان‌دهنده وزن مولکولی متوسط (وزن مولکولی/تعداد کل اتم‌ها) است. این توصیف کننده مستقل از اتصال و ساختار مولکولی بوده و اطلاعاتی از ترکیب مولکولی یک ترکیب را منعکس می‌کند [۲۱۶]. با توجه به ضریب اثر مثبت این توصیف کننده (جدول ۲-۱۶)، مقدار بالاتر این توصیف کننده نشان‌دهنده حضور اتم‌هایی با وزن اتمی بالاتر در مولکول است که منجر به افزایش فعالیت دارویی ترکیباتی با چنین شرایطی می‌شوند. با بررسی بیش‌تر ترکیبات مورد مطالعه مشاهده می‌شود که ترکیبات شماره ۱۲(فعالیت دارویی برابر با ۷/۰۹)، ۵۱ (فعالیت دارویی برابر با ۷/۰۰) و ۶۶ (فعالیت دارویی برابر با ۶/۹۲) از دسته ترکیباتی هستند که به‌دلیل داشتن گروه X (۲ گروه Cl در ترکیبات شماره ۱۲ و ۵۱ و یک گروه Cl در ترکیب شماره ۶۶) دارای AMW بالایی نیز می‌باشند و در نتیجه فعالیت دارویی بالاتری نیز دارند. علاوه بر اثر مثبت توصیف کننده AMW بر متغیر وابسته، توصیف‌کننده‌های تابع توزیع شعاعی (RDF) وزن‌دهی شده با جرم (m) در واقع بیان می‌کنند که این اتم‌ها با وزن بالا باید در محل‌های خاصی حضور داشته باشند و وجود اتم‌هایی با وزن بالا به تنهایی باعث افزایش فعالیت دارویی ترکیب نمی‌شود. در واقع این دسته از توصیف کننده‌های RDF، چگونگی توزیع شعاعی این اتم‌ها را در فضای شیمیایی ترکیب نشان می‌دهد. همان‌طور که گفته شد، توصیف‌کننده‌های تابع توزیع شعاعی (RDF)، حاوی اطلاعات اساسی در مورد فاصله‌های بین اتمی در طول پیوند، نوع حلقه و غیره می‌باشند. توصیف‌کننده‌های RDF مبتنی بر توزیع فواصل در نمایش هندسی مولکول و تشکیل یک کد تابع توزیع شعاعی است. RDF020m، RDF105m، RDF130m و RDF135m توصیف‌کننده‌های نرمال شده بر اساس جرم اتمی هستند. این توصیف کننده‌ها حضور اتم‌ها را

در کره‌های مجازی با قطرهای ۲۰، ۱۰۵، ۱۳۰ و ۱۳۵ آنگستروم نشان می‌دهند. این ۴ توصیف کننده دارای ضرایب تأثیر متفاوتی در مدل LAD-LASSO-LM-ANN هستند که در جدول ۲-۱۶ آورده شده است. توصیف کننده‌هایی با ضرایب مثبت نشان می‌دهد که وجود اتم‌ها با وزن بالا در ناحیه خاصی باعث افزایش فعالیت دارویی می‌شود و توصیف کننده RDF با ضریب منفی نشان می‌دهد که قرار گیری برخی از اتم‌ها با وزن بالا در ناحیه مشخصی از فضای شیمیایی ترکیب باعث کاهش فعالیت دارویی ترکیبات می‌شود.

-اثر توصیف کننده‌ها بر فعالیت دارویی با استفاده از مدل LAD-LASSO-LM-ANN برای بازدارنده‌های سرطان ریه

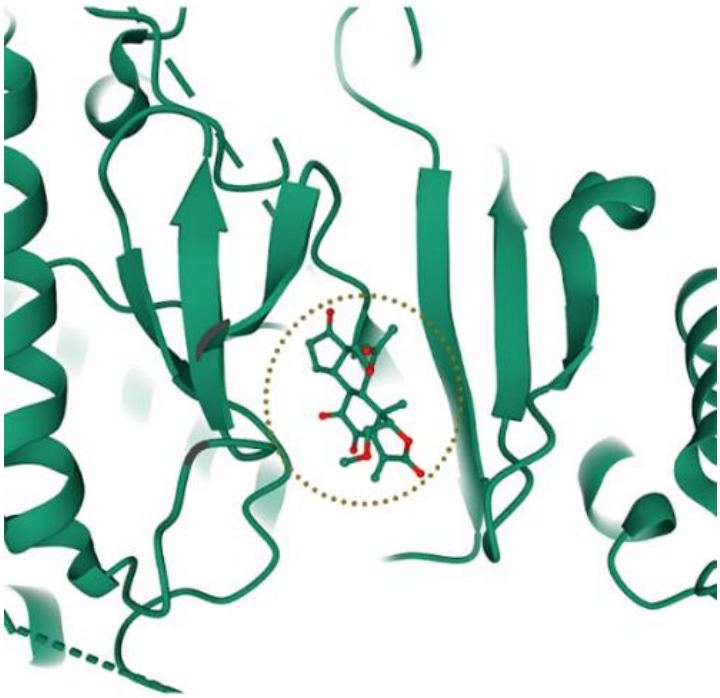
مدل LAD-LASSO-ANN ارائه شده برای بازدارنده‌های سرطان ریه (نتایج جدول ۲-۱۷)، مدل با ۷ توصیف کننده دارای شرایط بهینه است. اولین توصیف کننده با بیش‌ترین سهم مشارکت (شکل ۳-۱۴) و بیش‌ترین ضریب تأثیر (جدول ۲-۱۶) مربوط به AMW از دسته توصیف‌کننده‌های ساختاری می‌باشد. همان‌طور که گفته شد، این توصیف کننده نشان‌دهنده وزن مولکولی متوسط (وزن مولکولی/تعداد کل اتم‌ها) است. با توجه به ضریب اثر مثبت این توصیف کننده (جدول ۲-۱۶)، ترکیباتی با AMW بزرگ‌تر دارای فعالیت دارویی بیش‌تر هستند. توصیف‌کننده‌های RDF110m و RDF130m مربوط به توصیف‌کننده‌های تابع توزیع شعاعی هستند که بیانگر اطلاعات اساسی در توزیع فواصل در نمایش هندسی مولکول و تشکیل یک کد تابع توزیع شعاعی است. همان‌طور که گفته شد وجود گروه‌های سنگین در ناحیه‌ای مشخص از مرکز ساختار شیمیایی باعث افزایش فعالیت دارویی ترکیب می‌شود. این توصیف کننده‌ها بر اساس جرم اتمی وزن دهی شده‌اند. بنابراین با توجه به ضریب تأثیر مثبت این دو توصیف کننده (جدول ۲-۱۶)، ترکیباتی با وزن اتمی بیش‌تر، دارای فعالیت دارویی بیش‌تری نیز هستند. MoR21m و MoR12e از دسته توصیف‌کننده‌های 3D-MoRSE است. MoR21m توصیف کننده سه بعدی وزن دهی شده بر اساس جرم اتمی است. ضریب مثبت این توصیف کننده نیز گواهی بر این موضوع است که ترکیبات با وزن اتمی بیش‌تر

دارای فعالیت دارویی بیش‌تری می‌باشند. توصیف کننده Mor12e بر اساس الکترونگاتیوی وزن دهی می‌شود. Mor12e دارای مقادیر منفی در توصیف کننده و ضریب تأثیر منفی می‌باشد. بنابراین این توصیف کننده، نشان می‌دهد که حضور ترکیبات الکترونگاتیو از جمله هالوژن‌ها در ترکیبات، آن‌ها را فعال‌تر می‌کند [۲۱۷]. بنابراین با بررسی بیش‌تر ترکیبات مورد مطالعه مشاهده می‌شود که با توجه به ضریب تأثیر مثبت توصیف کننده‌های از جنس وزن مولکولی (AMW، RDF110m ، RDF130m و MoR21m) و تأثیر مثبت آن‌ها بر فعالیت دارویی و با توجه به تأثیر مثبت توصیف کننده Mor12e وجود گروه‌های هالوژنه الکترونگاتیو سنگین هم‌چون F و Cl در ناحیه مشخص (با توجه به فاصله مکانی مذکور توصیف کننده‌های RDF و MoRSE) فعالیت دارویی ترکیبات افزایش می‌یابد.از دسته ترکیبات فعال دارای چند گروه هالوژنه می‌توان به ترکیبات شماره ۱۲، ۱۳، ۴۹ و ۵۱ اشاره کرد که با داشتن چندین گروه فلوئور و کربن دارای فعالیت دارویی مناسبی نیز هستند (جدول ۲-۱۳ تا جدول ۲-۱۵). nCp توصیف کننده‌ای از گروه شمارش گروه عاملی می‌باشد و نشان‌دهنده کربن Sp^3 انتهایی می‌باشد. ضریب اثر مثبت این توصیف کننده نشان می‌دهد که ترکیبات پیشنهادی دارای این گروه کربنی فعالیت دارویی بیش‌تری را به خود اختصاص داده‌اند. ترکیبات شماره ۶۸، ۶۷ و ۶۵ با داشتن به‌ترتیب ۶، ۳ و ۲ گروه متیل از دسته ترکیبات با فعالیت دارویی در محدوده ۶/۹۲-۶/۶ هستند. بنابراین هر چه تعداد گروه متیل روی حلقه فنیل (بال راست) بیش‌تر باشد فعالیت دارویی آن ترکیب بیش‌تر است. F084 از گروه قطعات اتم محور[1]است و مربوط به تعداد گروه F متصل به کربن Sp^2 می‌باشد. با توجه به ضریب اثر مثبت این توصیف کننده، وجود گروه F بر حلقه فنیل (بال راست) ساختار ترکیب باعث افزایش فعالیت دارویی بازدارنده‌ها می‌شود. به‌طور مثال ترکیبات شماره ۱۳ و ۴۹ به‌ترتیب با داشتن ۱ و ۲ گروه فلوئور بر روی حلقه فنیل از دسته ترکیباتی با فعالیت مناسب دارویی (بزرگ‌تر از ۶/۰۰) می‌باشند.

[1]Atom centered fragments

- مطالعه داکینگ مولکولی بازدارنده‌های سرطان کارسینوم کولورکتال و ریه

پس از ایجاد ترکیبات فعال پیشنهادی با فعالیت دارویی مناسب، برهم‌کنش‌های ترکیبات پیشنهادی با اسیدآمینه‌های کلیدی جایگاه فعال گیرنده مورد نظر نیز مورد بررسی قرار گرفت. از این‌رو برهم کنش‌های ترکیبات مورد مطالعه و ترکیبات پیشنهادی با استفاده از مطالعه داکینگ مولکولی مورد ارزیابی قرار گرفت. بنابراین، ابتدا ساختار کریستالوگرافی 3HHM به پیشنهاد مقالات منتشر شده از سایت بانک اطلاعاتی پروتئین با فرمت pdb دانلود شد [۸۰, ۸۱, ۱۷۳]. این ساختار دارای ارزش تفکیک ۲/۸۰ A^o است و به هدف انجام فرایند داکینگ از کیفیت مناسبی برخوردار است [۲۱۸]. به‌منظور استخراج مختصات جایگاه فعال گیرنده، مختصات مرکز ثقل لیگاند کریستالوگرافی به کار گرفته شد. ساختار کریستالوگرافی گیرنده و لیگاند مربوطه در فرایند اعتبار سنجی داکینگ مورد استفاده قرار گرفت. تعداد اجراهای الگوریتم ۱۰۰، ۱۵۰ و ۲۰۰ برای داکینگ لیگاند کریستالوگرافی- گیرنده تعریف شد. پس از انجام فرایند اعتبار سنجی، شرایط بهینه داکینگ، با تعداد اجرای ژنتیک الگوریتم برابر با ۱۵۰ دارای کم‌ترین مقدار RMSD شد. بنابراین در ادامه، داکینگ ترکیبات مورد مطالعه و ترکیبات پیشنهادی در جایگاه فعال گیرنده مورد نظر با ۱۵۰ اجرای ژنتیک الگوریتم داک شدند. فایل خروجی dlg برای همه ترکیبات داک شده به‌دست آمد. با استفاده از اطلاعات مربوط به خروجی داکینگ، بهترین پیکربندی با توجه به کم‌ترین انرژی آزاد اتصال استخراج شدند و برهم کنش‌های متفاوت مربوط به لیگاند- گیرنده با استفاده از نرم‌افزار Discovery Studio Visualizer مورد بررسی قرار گرفت.

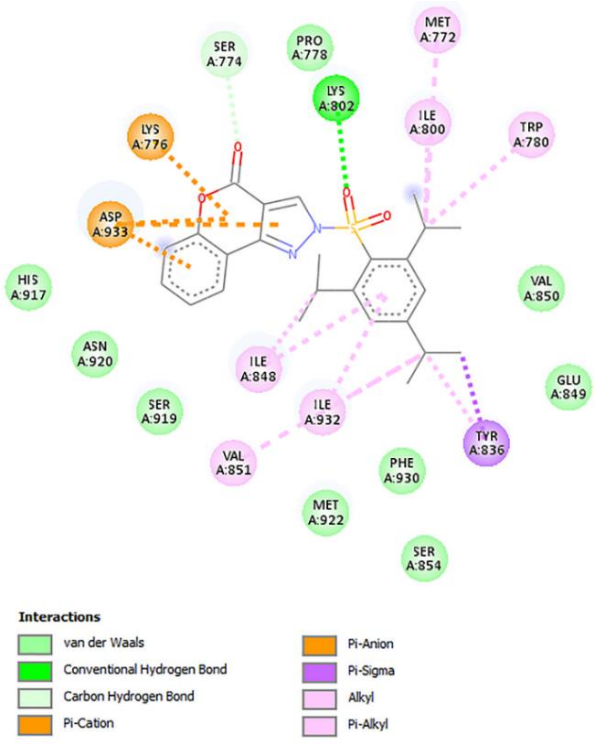


شکل ۳-۲۶ ساختار کریستالوگرافی 3HHM [۲۱۸] (منطقه نقطه چین نشان‌دهنده لیگاند کریستالوگرافی و مابقی زنجیره‌های اسید آمینه‌ای است)

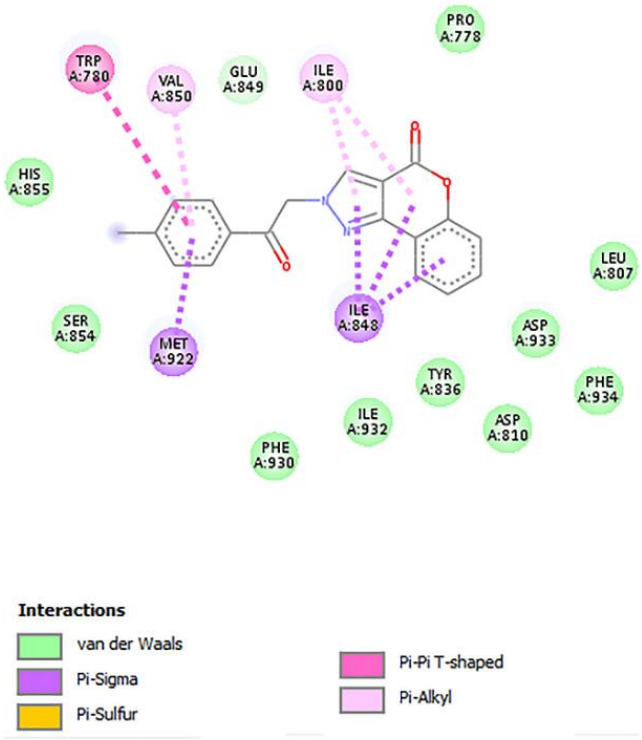
برهم‌کنش‌های متفاوت مربوط به لیگاندهای فعال و کم فعال بازدارنده‌های سرطان با گیرنده مربوطه در شکل ۳-۲۷ و شکل ۳-۲۸ آورده شده است. همان‌طور که در شکل ۳-۲۷ مشاهده می‌شود، ترکیب فعال برهم کنش‌های متفاوت هیدروژنی و هیدروفوبی با اسید آمینه‌های کلیدی برقرار نموده است. پیوند هیدروژنی ترکیب فعال با Lys802 به‌درستی گواهی بر فعالیت مناسب ترکیب در جایگاه فعال گیرنده می‌باشد. ترکیب کم فعال مورد مطالعه توانایی تشکیل پیوندهای هیدروژنی را از خود نشان نداده است. با توجه به نتایج برهم کنش‌های این دو ترکیب، فعالیت ترکیبات پیشنهادی نیز مورد قیاس قرار گرفت. به‌طوری‌که نتایج برهم کنش ترکیبات پیشنهادی با گیرنده در شکل ۳-۲۹ و شکل ۳-۳۰ نشان داده شده است. با توجه به نتایج مشاهده می‌شود که ترکیب پیشنهادی NC1 علاوه بر این‌که فعالیت پیش‌بینی شده مناسبی دارد (جدول ۳-۴)، بلکه برهم کنش‌های هیدروژنی مناسبی را با اسید آمینه‌های کلیدی از جمله Glu849، Val851 و Gln859 برقرار کرده است. ترکیب پیشنهادی NC2 نیز به‌خوبی توانسته است علاوه

بر پیوندهای هیدروفوبی، با اسید آمینه‌های کلیدی Val851 و Gln859 پیوندهای هیدروژنی برقرار کند. ترکیبات پیشنهادی NC3 و NC4 نیز علاوه بر پیوندهای هیدروفوبی متفاوت، با اسید آمینه Glu859 پیوند هیدروژنی برقرار کنند.

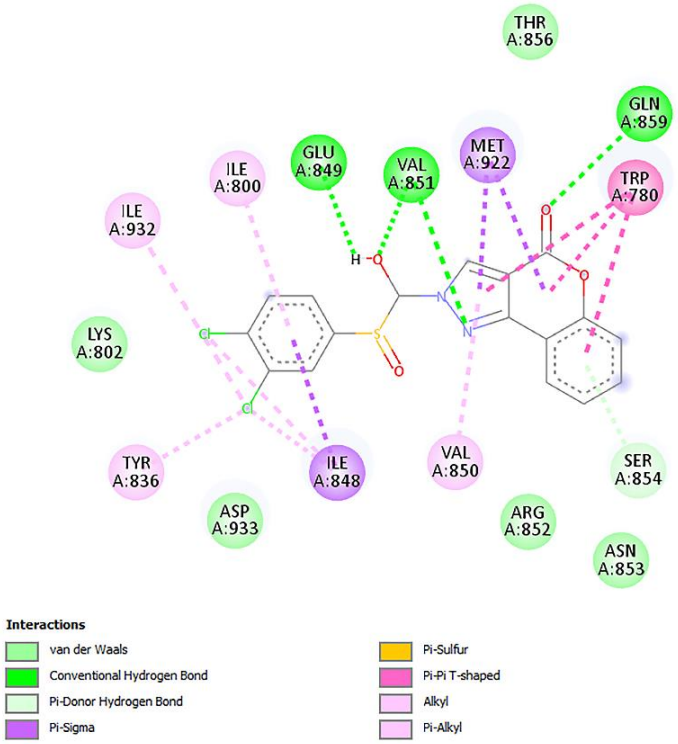
همان‌طور که نتایج نشان داد، ترکیبات پیشنهادی از ارتباط هیدروژنی و هیدروفوبی مناسبی با گیرنده برخوردار هستند و پایداری این ترکیبات در جایگاه فعال گیرنده مورد نظر، مورد تأیید است. علاوه بر بررسی برهم کنش‌های ترکیبات، ویژگی‌های فارماکوکینتیکی ترکیبات نیز مورد محاسبه قرار گرفت و نتایج جدول ۳-۴ نشان می‌دهد که ترکیبات پیشنهادی از نظر قاعده لیپینسکی نیز مورد تأیید هستند و علاوه بر این این درجه آسانی سنتز آزمایشگاهی کم‌تر از ۱۰ این ترکیبات بالقوه نیز محاسبه شد و با توجه به نتایج جدول ۳-۴ مشخص است که سنتز این ترکیبات با پیچیدگی همراه نیست.



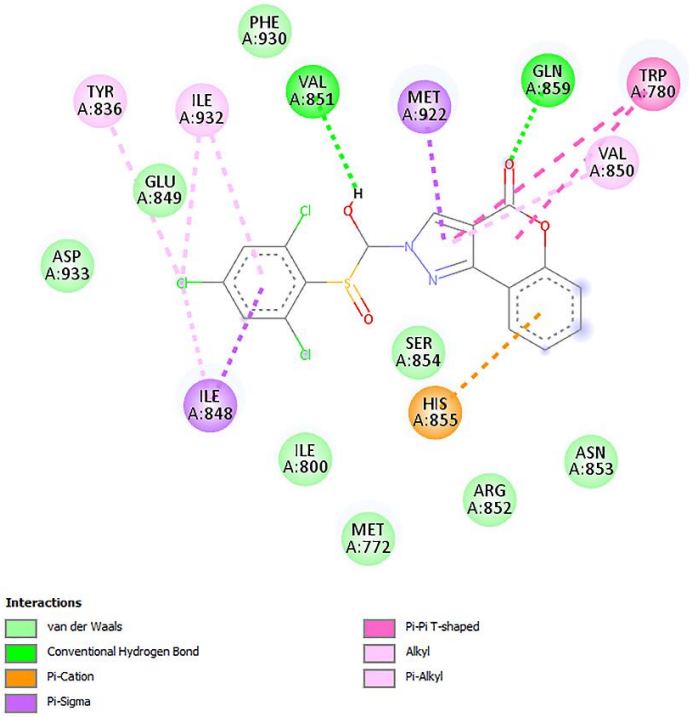
شکل ۳-۲۷ برهم‌کنش ترکیب نسبتاً فعال (۶۸) موجود در مجموعه داده‌های ضد سرطان کاسینوم کولورکتال با اسید آمینه‌های کلیدی



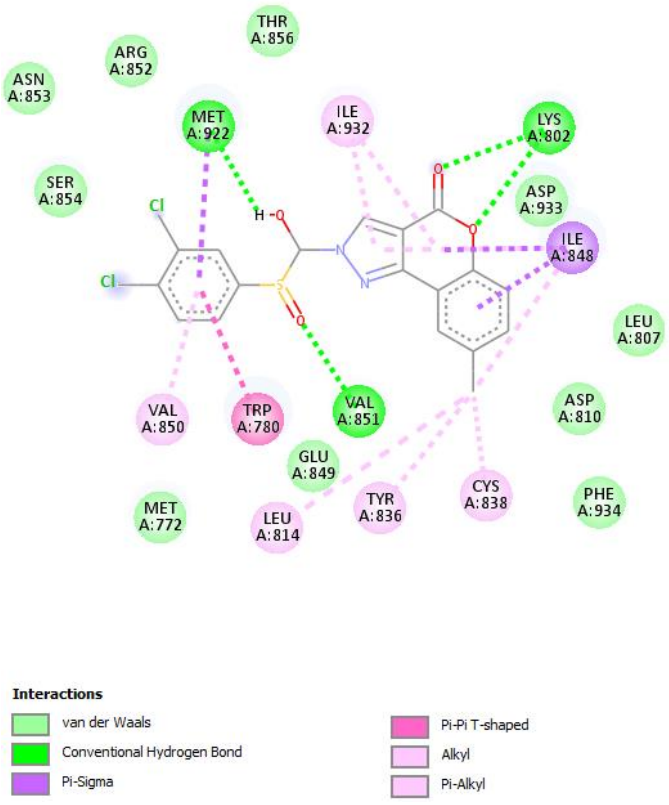
شکل ۳-۲۸ برهم‌کنش ترکیب کم فعال (۲۷) موجود در مجموعه داده‌های ضد سرطان کاسینوم کولورکتال با اسید آمینه‌های کلیدی



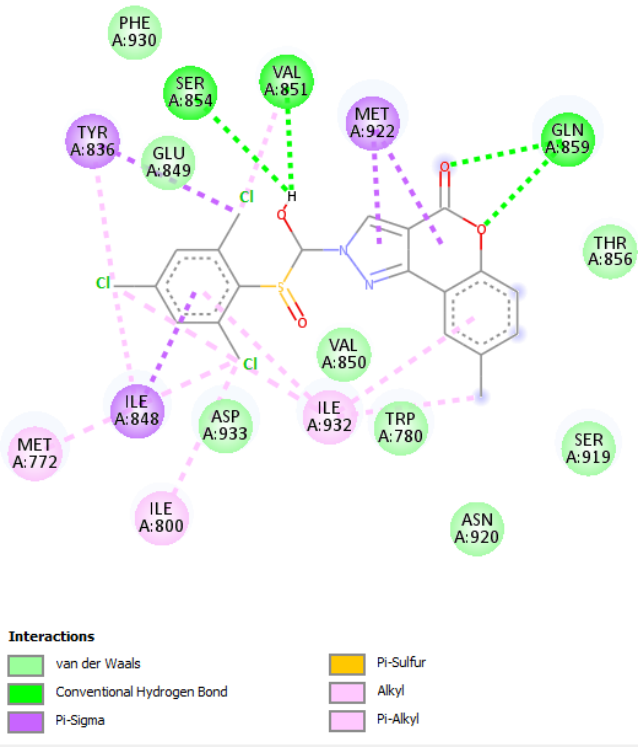
شکل ۳-۲۹ برهم‌کنش ترکیب پیشنهادی NC1 با اسید آمینه‌های کلیدی



شکل ۳-۳۰ برهم‌کنش ترکیب پیشنهادی NC2 با اسید آمینه‌های کلیدی

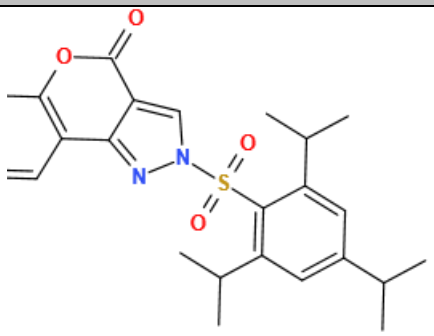
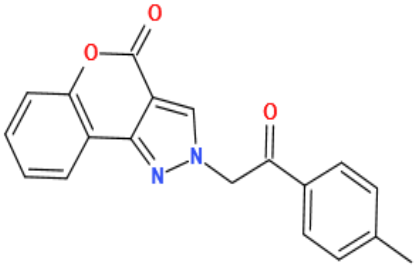
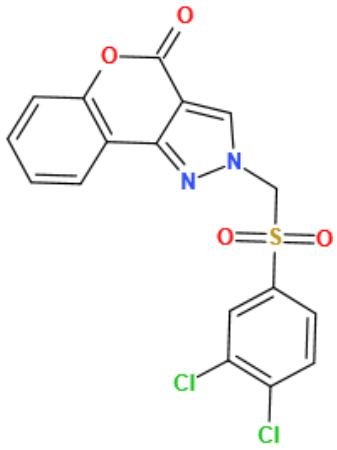


شکل ۳-۳۱ برهم‌کنش ترکیب پیشنهادی NC3 با اسید آمینه‌های کلیدی

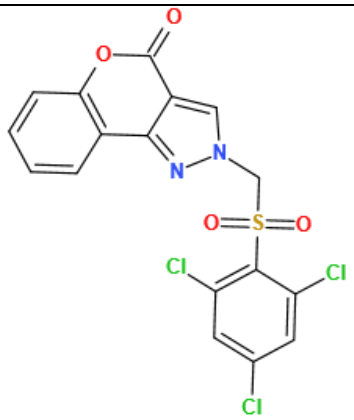
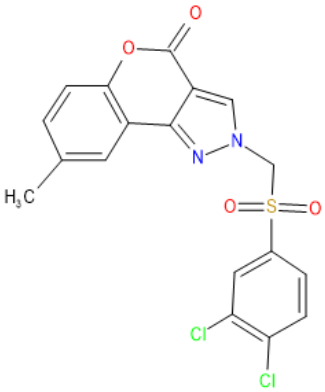
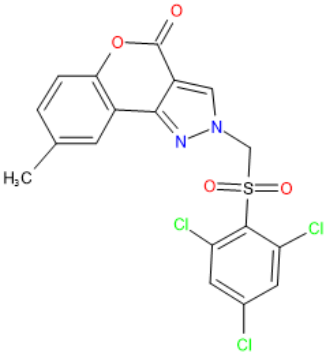


شکل ۳-۳۲ برهم‌کنش ترکیب پیشنهادی NC4 با اسید آمینه‌های کلیدی

جدول ۳-۴ پارامترهای PK محاسبه شده برای ترکیبات مورد مطالعه ضد سرطان و ترکیبات پیشنهادی

شماره ترکیب	ساختار شیمیایی	MW	MLOGP	#Rot-B	#H-B-don	#H-B-acc	Syn-Acc	pIC_{50}
۶۸		۳۱۸/۳۳	۲/۶۲	۳	۰	۴	۲/۸۸	۷/۱۵
۲۶		۴۵۲/۵۷	۵/۲	۵	۰	۵	۴/۶۸	۴/۲۰
NC1		۴۰۹/۲۴	۳/۳۵	۳	۰	۵	۳/۵۱	۸/۹۲

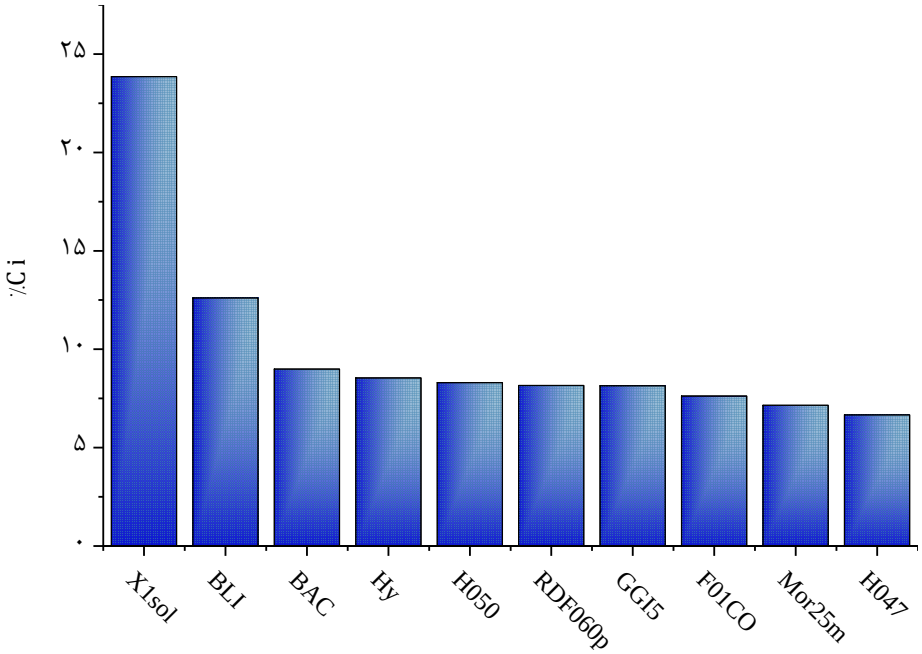
ادامه جدول ۳-۴

شماره ترکیب	ساختار شیمیایی	MW	MLOGP	#Rot-B	#H-B-don	#H-B-acc	Syn-Acc	pIC_{50}
NC2		۴۴۳/۶۹	۳/۸۵	۳	۰	۵	۳/۵۸	۸/۷۸
NC3		۴۵۷/۷۰	۴/۰۳	۳	۰	۵	۳/۵۸	۷/۵۲
NC4		۴۲۳/۲۰	۴/۱۳	۳	۰	۵	۳/۵۱	۶/۴۳

۳- ۱- ۴ تحلیل توصیف‌کننده‌های مدل SCAD-ANN برای ترکیبات آلی فرار

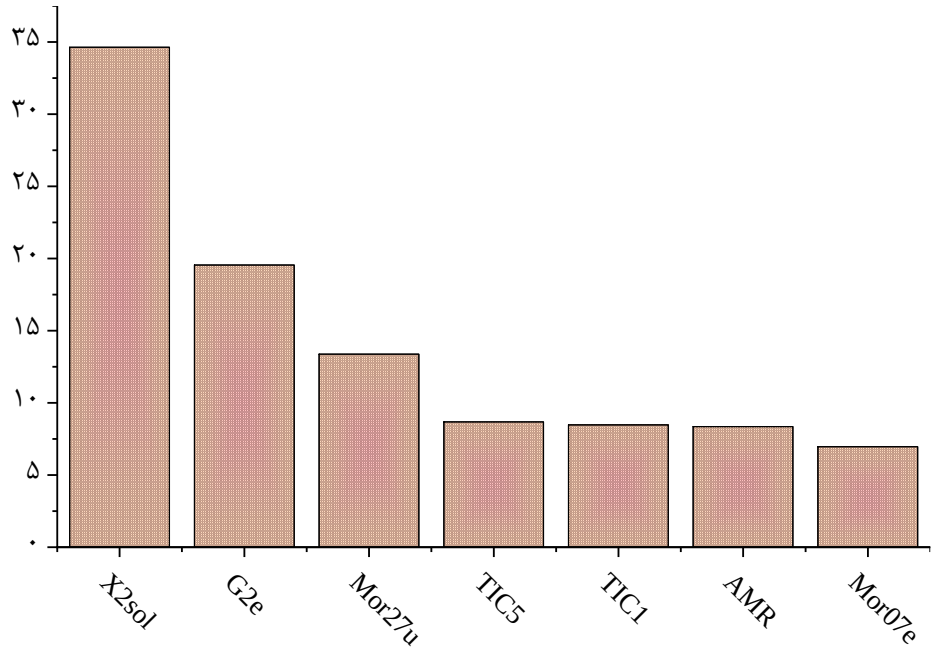
۳- ۱- ۴- ۱ محاسبه سهم مشارکت هر توصیف کننده در مدل SCAD-ANN

سهم مشارکت توصیف‌کننده‌های منتخب SCAD در مدل غیرخطی توسعه‌یافته بهینه مورد بررسی قرار گرفت. برای محاسبه میزان درصد مشارکت هر توصیف کننده، مقادیر هر توصیف کننده در محدوده تغییرات مقادیر واقعی هر توصیف کننده تصادفی شد. این‌بار مدل‌های بهینه SCAD-ANN هر بار در حضور توصیف کننده i با مقادیر تصادفی و سایر توصیف کننده‌ها با مقادیر واقعی خود توسعه داده شدند و مقادیر RI با این شرایط پیش‌بینی شد. مقادیر $RMSE_i$ مربوط به مجموعه ارزیابی در حضور توصیف کننده i_{th} (با مقادیر تصادفی) به‌دست آمد و این فرایند برای همه توصیف کننده‌ها تکرار شد و در نهایت درصد مشارکت هر توصیف کننده (C_i٪) بر اساس رابطه ۱-۱۵ محاسبه شد و نتایج در شکل ۳-۳۳ و شکل ۳-۳۴ نمایش داده شد در ادامه به بررسی میزان و چگونگی تأثیر این توصیف کننده‌ها بر شاخص بازداری ترکیبات و رابطه آن‌ها با ساختار ترکیبات مورد مطالعه پرداخته خواهد شد.



ت و ص یف ک نده های م نتخب روش SCAD برای مجموعه داده های A

شکل ۳-۳۳ نمودار سهم مشارکت توصیف کننده‌ها در مدل SCAD-ANN برای مجموعه A



توصیف کننده های منتخب روش SCAD برای مجموعه داده های B

شکل ۳-۳۴ نمودار سهم مشارکت توصیف کننده‌ها در مدل SCAD-ANN برای مجموعه B

۳- ۱- ۴- ۲ بررسی رابطه بین توصیف کننده‌های استفاده شده در مدل SCAD-ANN و شاخص بازداری ترکیبات مورد مطالعه

برای بررسی بیشتر چگونگی تأثیر هر توصیف کننده بر شاخص بازداری، علامت تأثیر هر توصیف کننده بر RI تعیین شد. برای این کار، ضرایب مدل SCAD مطابق با معادله‌های زیر برای مجموعه داده‌های A و B (رابطه ۳-۳) و B (رابطه ۳-۴) به‌دست آمد.

$$RI=0.35+ 0.97X1sol-0.002F01CO-0.22\ BAC-\ 0.01\ H047-0.03Mor25m+0.03H050+0.03Hy -0.25BLI-0.02RDF060p-0.01\ GGI5+0.015\ F03CN+0.016C025+0.02HATS5V+0.03TPSATot+ 0.001E1m$$

رابطه ۳-۳

$$RI=0.42+ 1.01X2sol - 0.005G2e -\ 0.13Mor27u + 0.027\ TIC5 + 0.006TIC1 + 0.005\ AMR - 0.005\ Mor07e$$

رابطه ۳-۴

با توجه به رابطه ۳-۳ و رابطه ۳-۴، اثر افزایشی و یا کاهشی تأثیر توصیف کننده‌ها بر شاخص بازداری نشان داده شد. با توجه به تفسیرپذیری بهتر و تأثیر بیش‌تر برخی از توصیف کننده‌ها به شرح آن‌ها

پرداخته شد و در ادامه شرحی از این توصیف کننده‌ها و چگونگی وابستگی شاخص بازداری به آن‌ها ارائه شده است.

-اثر توصیف کننده‌ها بر شاخص بازداری با استفاده از مدل SCAD برای مجموعه A

با توجه به نمودار سهم مشارکت توصیف کننده‌ها مؤثرترین توصیف کننده‌ها عبارتند از X1sol، BLI و BAC و این توصیف کننده‌ها به‌ترتیب اثر افزایشی، کاهشی و کاهشی بر RI دارند.

بحرانی‌ترین توصیف کننده با بیش‌ترین سهم در مدل پیشنهادی مربوط به X1sol است. این توصیف کننده از دسته توصیف‌کننده‌های مکانی است و شاخص اتصال حلالیت chi-1 نام دارد [۲۱۹]. توصیف‌کننده‌های شاخص اتصال انحلال پذیری، بیانی از آنتروپی هستند. این توصیف کننده بیانی از پارامترهای اتمی است و با رابطه ۳-۵ نشان داده می‌شود:

$$m_{\chi^s} = (\frac{1}{2^{m+1}}).\sum_{k+1}^{k}\left(\frac{(\prod_{a=1}^{k} L_a)_{\,k}}{(\prod_{a=1}^{n}\delta_a)^{\frac{1}{2}}_{k}}\right)$$

رابطه ۳-۵

که در آن L_a عدد کوانتومی اصلی (۲ برای اتم‌های C، N، O و ۳ برای Si، S، Cl، ...) مربوط به a_{th} اتم در زیرگراف k_{th} است. δa درجه رأس مربوطه است. k تعداد کل زیرگراف‌های مرتبه m ام و n تعداد رئوس در زیرگراف است. $1/2^{\,m+1}$ ضریب نرمال سازی است و m نشان دهنده درجه شاخص اتصال حلالیت است که برای X1sol برابر با ۱ است [۹]. با توجه به علامت مثبت ضریب X1sol در معادله SCAD، ترکیبات با مقادیر X1sol زیاد دارای مقادیر RI بالاتری نیز هستند.

BLI توصیف کننده مورد بحث بعدی است و با توجه به رابطه ۳-۵ دارای ضریب اثر منفی بر شاخص بازداری است. شاخص شباهت بنزن کایر (Keir) یک شاخص آروماتیکی است که از توپولوژی مولکولی محاسبه می‌شود. توصیف‌کننده‌های BLI از جمله شاخص‌های رزونانسی هستند که پایداری بنزن را از نظر تئوری توضیح می‌دهند و میزان جابجایی سیستم‌های مزدوج را پیش‌بینی می‌کنند.

شاخص BAC از دسته توصیف‌کننده‌های مکانی و مربوط به شاخص مرکزی balaban می‌باشد. نام شاخص مرکزی به‌این دلیل است که توپولوژی نمودار را از مرکز مشاهده می‌کند .این توصیف کننده پیچیدگی و انشعاب مولکولی را رمزگذاری می‌کند. پارامتر BAC می‌تواند به یک یا چند ویژگی ساختاری مولکول مانند اندازه، شکل، تقارن، انشعاب و چرخه‌ای بودن حساس باشد. همچنین می‌تواند اطلاعات شیمیایی مربوط به نوع اتم و تعداد پیوند را نیز رمزگذاری کند.

-اثر توصیف کننده‌ها بر شاخص بازداری با استفاده از مدل SCAD برای مجموعه B

از بین توصیف‌کننده‌های مدل بهینه برای مجموعه B، X2sol، G2e و Mor27u بیش‌ترین سهم مشارکت را در ساخت مدل ANN دارند. از این‌رو در ادامه به شرح مفصلی از هر یک از این‌ها پرداخته می‌شود.

بحرانی‌ترین توصیف کننده با بیش‌ترین سهم مشارکت در مدل پیشنهادی X2sol است که نشان‌دهنده شاخص اتصال حلالیت با درجه ۲ است. این توصیف کننده از دسته توصیف‌کننده‌های توپولوژیکی دوبعدی است [۲۱۹]. توضیحات مربوط به این شاخص در بخش قبلی و با رابطه ۳-۵ شرح داده شد. با توجه به علامت مثبت ضریب X2sol در معادله SCAD (رابطه ۳-۴)، ترکیبات با مقادیر X2sol زیاد دارای مقادیر RI بالاتری هستند.

دومین توصیف کننده مهمی (شکل ۳-۳۴) که در مدل ظاهر شده است G2e است که متعلق به شاخص WHIM است. این توصیف کننده، مؤلفه‌ای است که توسط الکترونگاتیوی ساندرسون اتمی وزن شده است. G2e مربوط به شاخص‌های آماری است و بر روی پیش‌بینی‌اتم ها در امتداد محورهای اصلی محاسبه می‌شود. الکترونگاتیوی ساندرسون اتمی یکی از طرح‌های وزن دهی است که برای محاسبه ماتریس کوواریانس وزنی در توصیف کننده G2e استفاده می‌شود و به شکل مولکولی و اندازه کل ساختار مربوط می‌شود [۲۲۰]. ترکیب با مقدار G2e کم‌تر، استخلاف بیش‌تری دارد. بنابراین، اندازه ساختار بزرگ‌تر است.

با توجه به تأثیر منفی G2e بر روی RI، ترکیب با مقدار G2e کم‌تر دارای مقدار RI بیش‌تری است. این مفهوم نشان می‌دهد که افزایش تعداد استخلاف‌ها و حجم کل ساختار باعث می‌شود فاز ساکن با ترکیب مورد مطالعه بیش‌تر برهمکنش داشته باشد و مقدار RI افزایش یابد.

سومین توصیف کننده (Mor27u) به دسته توصیف‌کننده‌های نمایش ساختارهای سه بعدی مولکولی بر اساس پراش الکترونی (3D-MoRSE) تعلق دارد. پارامترهای عددی 3D-MoRSE ساختار مولکولی سه بعدی را به‌طور قابل توجهی نشان می‌دهند و آرایش سه بعدی اتم‌ها را منعکس می‌کنند و پیوندهای شیمیایی در نظر گرفته نمی‌شوند [۲۲۱]. Mor27u (3D-MoRSE-signal 27/unweighted) مبتنی بر ایده به‌دست آوردن اطلاعات از مختصات اتمی سه بعدی با تبدیل مورد استفاده در مطالعات پراش الکترون برای تهیه منحنی‌های پراکندگی نظری است [۲۲۲]. به‌طور کلی توصیف‌کننده‌های 3D-MoRSE فواصل بین اتمی را اندازه‌گیری می‌کنند و دیدگاه‌های متفاوتی از کل ساختار مولکول را نشان می‌دهند، اگرچه معنای آن‌ها هنوز غیر قابل مشخص است [۲۲۳]. با توجه به تأثیر منفی این توصیف کننده بر مقادیر RI، انتظار می‌رود که ترکیبات با مقادیر منفی Mor27u دارای مقادیر RI نسبتاً بالایی باشند. در نتیجه، ظهور Mor27u در مدل نهایی، ویژگی‌های ساختاری مولکول‌های با اندازه بزرگ را در مدل در نظر می‌گیرد.

۳- ۲ نتیجه‌گیری نهایی

در این مطالعه، مدل‌های QSAR\QSRR مبتنی بر به‌کارگیری ترکیب روش‌های انقباضی و مدل غیر خطی شبکه عصبی مصنوعی برای پیش بینی فعالیت دارویی/ شاخص بازداری مجموعه‌داده‌های متفاوتی مورد استفاده قرار گرفت. در ادامه نتیجه‌گیری کلی از نتایج حاصل از هر یک از مدل‌های توسعه یافته QSAR و QSRR ارائه خواهد شد.

در مطالعه اول این رساله، برای اولین بار، روش انقباضی SCAD (فن و لی، ۲۰۰۱) به عنوان روش انتخاب متغیر انقباضی با مدل غیر خطی LM-ANN ترکیب شد و در مطالعات QSAR مورد استفاده قرار گرفت. روش پیشنهادی SCAD-LM-ANN از مزایای ذاتی SCAD در کاهش ابعاد داده‌ها و کارایی LM-ANN در مدل سازی روابط غیر خطی بهره برده است. در نتیجه، مدل ساخته شده علاوه بر قدرت پیش بینی رضایت بخش از تفسیر پذیری و تنکی مناسبی نیز برخوردار است. مدل توسعه‌یافته SCAD-LM-ANN با استفاده از ارزیابی نتایج آماری مجموعه آزمون، تکنیک LOO، دامنه کاربرد، آزمون Y- تصادفی و محاسبه پارامترهای آماری مورد ارزیابی بیش‌تر قرار گرفت. همان‌طور که نتایج (جدول ۲-۶) نشان می‌دهد، مدل از قدرت پیش بینی و تعمیم پذیری قابل قبولی برخوردار است. توانایی ذاتی SCAD در انتخاب متغیر منجر به ساخت یک مدل تفسیرپذیر با مجموعه کوچکی از توصیف‌کننده‌ها می‌شود. توصیف کننده‌های مناسب و صحیح انتخاب شده توسط روش SCAD، امکان طراحی ترکیبات فعال جدید را با توجه به تأثیر مقادیر ضرایب رگرسیونی توصیف‌کننده‌ها بر فعالیت دارویی فراهم می‌کند. صحت فعال بودن ترکیبات پیشنهادی، با استفاده از بررسی برهم‌کنش لیگاند-گیرنده از طریق مطالعه داکینگ مولکولی تأیید شد. مقادیر پیش بینی شده فعالیت دارویی (جدول ۳-۱)، نتایج برهم کنش داکینگ مولکولی ترکیبات پیشنهادی با گیرنده (شکل ۳-۳ و شکل ۳-۴) و نتایج ویژگی‌های فارموکوکینتیکی ترکیبات پیشنهادی (جدول ۳-۱) نشان می‌دهد که ترکیبات جدید فعالیت بالقوه‌ای را به‌عنوان بازدارنده‌های جدید ایدز دارند [۱۱۴].

در مطالعه دوم این رساله، با توجه به اهمیت بررسی مهارکننده‌های $3CL^{pro}$ برای مقابله با بیماری کووید-۱۹، مدل شبکه عصبی با روش انتخاب متغیر انقباضی قدرتمند ALASSO (زو، ۲۰۰۶)، برای پیش‌بینی فعالیت بازدارنده‌های ضد کووید-۱۹، توسعه داده شد. روش انقباضی ALASSO تعداد توصیف کننده‌ها را به ۹ توصیف کننده مؤثر با بیش‌ترین تأثیر بر متغیر وابسته کاهش داد. مدل ALASSO-LM-ANN با قابلیت تفسیرپذیری و قدرت پیش بینی مناسب تولید شد. ارزیابی مدل ALASSO-LM-ANN با استفاده از روش‌های مختلف برآورد شد. نتایج خلاصه شده در جدول ۲-۱۲ نشان می‌دهد که مدل ALASOO-LM-ANN دارای پارامترهای آماری قابل قبولی است. همچنین، نمودار دامنه کاربرد (شکل ۲-۱۴) به وضوح اعتبار مدل توسعه یافته ALASSO-LM-ANN را اثبات می‌کند. بنابراین، کارایی مدل تفسیرپذیر و پیشگو توسعه یافته با توصیف کننده‌های منتخب روش قدرتمند ALASSO برای پیشنهاد ترکیبات جدید مورد استفاده قرار گرفت. صحت ترکیبات پیشنهادی با استفاده از بررسی بر هم کنش ترکیبات پیشنهادی – گیرنده به‌کمک داکینگ مولکولی و محاسبه خواص PK برای تأیید قانون پنج لیپینسکی مورد ارزیابی قرار گرفت. مقایسه ترکیبات فعال مورد مطالعه و ترکیبات پیشنهادی به درستی نشان دهنده قدرت بالای ترکیبات پیشنهادی به‌عنوان بازدارنده‌های بالقوه کووید-۱۹ هستند [۱۱۵].

در مطالعه سوم این رساله، این مدل‌های LAD-LASSO-LM-ANN برای پیش‌بینی فعالیت دارویی سه مجموعه داده متفاوت ضد ایدز و ضد سرطان مورد استفاده قرار گرفت. موثرترین توصیف کننده‌های هر مجموعه داده با استفاده از روش LAD-LASSO (وانگ، ۲۰۰۷) به عنوان ورودی ANN تعریف شدند. پس از ساخت و توسعه مدل‌های ANN، فعالیت‌های دارویی ترکیبات مجموعه آزمون بر اساس مدل بهینه LAD-LASSO-LM-ANN پیش بینی شدند. پارامترهای آماری متفاوتی برای مدل‌های توسعه یافته محاسبه شد (جدول ۲-۲۲) و نتایج همه پارامترهای آماری در محدوده قابل قبول قرار داشتند. در این مطالعه، چگونگی وابستگی فعالیت دارویی به توصیف کننده‌های منتخب هر مجموعه داده مورد بررسی قرار

گرفت. با توجه به ارتباط بین توصیف کننده‌های منتخب و فعالیت دارویی ترکیبات، چندین بازدارنده جدید با فعالیت دارویی مناسب پیشنهاد شد. صحت ترکیبات پیشنهادی با استفاده از بررسی بر هم کنش ترکیبات پیشنهادی – گیرنده (شکل ۳-۱۶ تا شکل ۳-۳۲) و محاسبه خواص PK برای تأیید قانون پنج لیپینسکی مورد ارزیابی قرار گرفت (جدول ۳-۳ و جدول ۳-۴). مقایسه ترکیبات فعال مورد مطالعه و ترکیبات پیشنهادی به درستی نشان دهنده فعالیت دارویی مناسب ترکیبات پیشنهادی است. علاوه بر این، فاکتور سهولت سنتز (کم‌تر از ۱۰) نشان می‌دهد که سنتز ترکیبات در مقیاس آزمایشگاهی نیز امکان پذیر است.

در مطالعه آخر این رساله تحقیقاتی، از روش SCAD به عنوان یک روش انقباضی کارآمد برای انتخاب توصیف کننده‌های مؤثر و پیش بینی مقادیر RI ترکیبات آلی فرار متفاوت با استفاده از مدل غیر خطی ANN برای ساخت مدل QSRR استفاده شد. مدل‌های متناظر SCAD-ANN پیشنهادی به طور هم‌زمان از مزیت SCAD مانند تنکی و پایداری مناسب توام با توانایی پیش‌بینی بسیار بالای روش شبکه عصبی مصنوعی استفاده شد. ارزیابی مدل‌های SCAD-ANN با استفاده از داده‌های مجموعه آزمون، تکنیک LOO، دامنه کاربرد و آزمون Y – تصادفی انجام شد و نتایج نشان دهنده قدرت پیش‌بینی و تعمیم پذیری رضایت‌بخش مدل ارائه شده است.

در نهایت، به منظور بررسی کارایی روش‌های انقباضی به‌عنوان روش‌های مدل‌سازی QSAR و QSPR، هر کدام از روش‌های مورد مطالعه در شرایط یکسان و مشابه با مراحل ارائه شده در هر بخش، بر روی داده‌های آموزش مجموعه داده‌های متفاوت اعمال شد. فعالیت‌های دارویی/ شاخص‌های بازداری ترکیبات آزمون با استفاده از مدل‌های توسعه یافته با توصیف کننده‌های مولکولی منتخب پیش بینی شد. پارامترهای آماری متفاوت (جدول ۱-۱) برای ترکیبات مجموعه آزمون پیش‌بینی شده، محاسبه شد. با مقایسه نتایج مربوط به به‌کارگیری روش‌های انقباضی جفت شده با مدل شبکه عصبی مصنوعی و روش‌های انقباضی به‌عنوان روش‌های انتخاب متغیر و مدل‌سازی هم‌زمان، مشاهده می‌شود که مدل‌های غیرخطی

ANN جفت شده با تکنیک‌های جریمه‌ای (SCAD، ALASSO و LAD-LASSO) از قدرت پیش‌بینی و تعمیم پذیری بهتری برخوردار هستند و در برخی از موارد به‌کارگیری مدل‌های انقباضی به‌تنهایی نتوانسته است نتایج قابل قبولی را ارائه دهد. بنابراین هدف این رساله در توسعه مدل‌های غیرخطی ANN جفت شده با تکنیک‌های جریمه‌ای به‌عنوان روش‌های انتخاب متغیر، به‌طور کارآمد و موفقی عمل نموده است.

در نهایت می‌توان به این نتیجه رسید که برای ساخت مدل‌های QSAR/QSPR با قدرت پیش‌بینی بالا، مدل SCAD و LAD-LASSO عملکرد بالاتری را از خود نشان داده‌اند و نتایج رضایت‌بخشی دارند. در حالی‌که برای ساخت مدل‌های QSAR/QSPR تنک و تفسیرپذیر روش ALASSO برای انتخاب موثرترین توصیف کننده‌ها پیشنهاد می‌شود ولی از آنجایی‌که ALASSO به دنبال برآورد ضرایب حقیقی و درست‌تری است، یافتن مدل QSAR/QSPR با بیش‌ترین ارتباط بین متغیرهای مستقل و وابسته کمی دشوار است.

جدول ۳-۵ مقایسه پارامترهای آماری محاسبه شده برای مجموعه آزمون مدل برتر SCAD-LM-ANN با مدل SCAD برای مشتقات استانیلید/ استامید به‌عنوان بازدارنده‌های ایدز

ردیف	پارامترهای آماری	مقادیر $pEC_{۵۰}$ پیش بینی شده ترکیبات مجموعه آزمون با مدل SCAD	مقادیر $pEC_{۵۰}$ پیش بینی شده ترکیبات مجموعه آزمون با مدل SCAD-LM-ANN
۱	PRESS	۲/۶۳	۱/۳۶
۲	SEP	۰/۴۹	۰/۳۵
۳	MAE	۰/۴۲	۰/۳۱
۴	REP(%)	۸/۰۵	۵/۷۹
۵	MSE	۰/۲۴	۰/۱۲
۶	MRE	۷/۰۷	۵/۱
۷	R^2	۰/۶۵	۰/۹۲
۸	R_0^2	۰/۴۸	۰/۹۱
۹	R_0^2 نسبی	۰/۲۶	۰/۰۱
۱۰	R_m^2	۰/۳۸	۰/۹۲
۱۱	$R_0'^2$	۰/۶۵	۰/۸۹
۱۲	$R_0'^2$ نسبی	۰/۰۰۲	۰/۰۳
۱۳	$R_m'^2$	۰/۶۳	۰/۹۲
۱۴	R-R	۰/۱۷	۰/۰۲
۱۵	k	۱/۰۰	۰/۹۶
۱۶	k′	۰/۹۹	۱/۰۴

جدول ۳-۶ مقایسه پارامترهای آماری محاسبه شده برای مجموعه آزمون مدل برتر ALASSO-LM-ANN با مدل ALASSO برای مشتقات 3-chymotrypsin like protease (3CLPro) به‌عنوان بازدارنده‌های SARS-COV-2

ردیف	پارامترهای آماری	مقادیر pIC_{50} پیش بینی شده ترکیبات مجموعه آزمون با مدل ALASSO	مقادیر pIC_{50} پیش بینی شده ترکیبات مجموعه آزمون با مدلALASSO-LM-ANN
۱	PRESS	۱۴/۱۱	۱/۴۷
۲	SEP	۰/۴۰	۰/۳۴
۳	MAE	۰/۳۰	۰/۲۸
۴	REP(%)	۷/۶۹	۶/۰۸
۵	MSE	۰/۱۶	۰/۱۱
۶	MRE	۵/۸۳	۵/۱۴
۷	R^2	۰/۷۵	۰/۸۳
۸	R_0^2	۰/۵۸	۰/۸۵
۹	نسبی R_0^2	۰/۲۳	۰/۰۱
۱۰	R_m^2	۰/۴۴	۰/۷۷
۱۱	$R_0'^2$	۰/۷۴	۰/۸۶
۱۲	نسبی $R_0'^2$	۰/۰۱	۰/۰۰
۱۳	$R_m'^2$	۰/۳۵	۰/۷۷
۱۴	R-R	۰/۱۶	۰/۰۱
۱۵	k	۰/۹۹	۰/۹۸
۱۶	k′	۱/۰۰	۱/۰۳

جدول ۳-۷ مقایسه پارامترهای آماری محاسبه شده برای مجموعه آزمون مدل برتر LAD-LASSO-ANN با مدل LAD-LASSO برای هر سه مجموعه داده‌های متفاوت

ردیف	پارامتر آماری	نتایج فعالیت دارویی پیش بینی شده ترکیبات مجموعه آزمون با مدل					
		مجموعه داده‌های ضد ایدز		مجموعه داده‌های ضد سرطان کارسینوم کولورکتال		مجموعه داده‌های ضد سرطان ریه	
		LAD-LASSO	LAD-LASSO-ANN	LAD-LASSO	LAD-LASSO-ANN	LAD-LASSO	LAD-LASSO-ANN
۱	PRESS	۱/۷۸	۱/۴۸	۱/۳۷	۰/۸۰	۴/۲۱	۱/۱۷
۲	SEP	۰/۴۰	۰/۳۷	۰/۳۵	۰/۲۷	۰/۶۲	۰/۳۳
۳	MAE	۰/۳۱	۰/۲۹	۰/۲۷	۰/۲۳	۰/۴۵	۰/۲۸
۴	REP(%)	۵/۷۸	۵/۶۲	۶/۷۶	۵/۱۴	۱۲/۴۳	۶/۵۴
۵	MSE	۰/۱۶	۰/۱۳	۰/۱۲	۰/۰۷	۰/۳۸	۰/۱۱
۶	MRE	۴/۵۶	۵/۲۴	۴/۹۰	۴/۴۲	۹/۱۷	۵/۸۹
۷	R^2	۰/۸۲	۰/۸۷	۰/۶۷	۰/۸۴	۰/۵۲	۰/۸۷
۸	R_0^2	۰/۷۹	۰/۸۶	۰/۵۹	۰/۷۷	۰/۴۲	۰/۸۵
۹	R_0^2 نسبی	۰/۰۴	۰/۰۱	۰/۱۲	۰/۰۸	۰/۱۹	۰/۰۲
۱۰	R_m^2	۰/۶۸	۰/۷۸	۰/۴۸	۰/۶۲	۰/۳۶	۰/۷۵
۱۱	$R_0'^2$	۰/۸۰	۰/۸۰	۰/۶۶	۰/۸۴	۰/۵۱	۰/۸۷
۱۲	$R_0'^2$ نسبی	۰/۰۲	۰/۰۸	۰/۰۱	۰/۰۰	۰/۰۲	۰/۰۰
۱۳	$R_m'^2$	۰/۷۱	۰/۶۵	۰/۴۸	۰/۵۷	۰/۲۹	۰/۷۳
۱۴	R-R	۰/۰۱	۰/۰۶	۰/۰۷	۰/۰۷	۰/۰۹	۰/۰۲
۱۵	k	۰/۹۷	۰/۹۷	۰/۹۹	۰/۹۸	۰/۹۸	۰/۹۸
۱۶	k′	۱/۰۳	۱/۰۳	۱/۰۰	۱/۰۲	۱/۰۰	۱/۰۱

جدول ۳-۸ مقایسه پارامترهای آماری محاسبه شده برای مجموعه آزمون مدل برتر SCAD -ANN با مدل SCAD برای مجموعه داده A و B

ردیف	پارامترهای آماری	مجموعه داده A		مجموعه داده B	
		SCAD	SCAD-ANN	SCAD	**SCAD-ANN**
۱	PRESS	۰/۱۰	۰/۰۵	۲۸۴۰۶۰	۷۸۴۲۶
۲	SEP	۰/۰۶	۰/۰۴	۱۶۰/۷۰	۸۴/۴۴
۳	MAE	۰/۰۴	۰/۰۳	۱۲۶/۸۴	۶۶/۹۱
۴	REP(%)	۱۰/۴۱	۵/۳۹	۱۵/۲۸	۸/۰۲
۵	MSE	۰/۰۶	۰/۰۴	۱۶۰/۷۰	۸۴/۴۳
۶	MRE	۶/۷۴	۴/۳۴	۱۲/۱۳	۶/۵۳
۷	R^2	۰/۸۳	۰/۹۲	۰/۷۳	۰/۸۹
۸	R_0^2	۰/۸۲	۰/۹۱	۰/۷۳	۰/۸۹
۹	R_0^2 نسبی	۰/۰۱	۰/۰۱	۰/۰۰	۰/۰۱
۱۰	R_m^2	۰/۷۵	۰/۸۳	۰/۷۳	۰/۸
۱۱	$R_0'^2$	۰/۸۱	۰/۹۲	۰/۶۸	۰/۸۸
۱۲	$R_0'^2$ نسبی	۰/۰۲	۰/۰۰	۰/۰۷	۰/۰۰
۱۳	$R_m'^2$	۰/۷۴	۰/۸۲	۰/۵۷	۰/۷۹
۱۴	R-R	۰/۰۱	۰/۰۱	۰/۰۵	۰/۰۱
۱۵	k	۱/۰۰	۰/۹۸	۱/۰۸	۱/۰۴
۱۶	k'	۰/۹۸	۱/۰۱	۰/۹۱	۰/۹۵

۳- ۳ آینده نگری

- ✓ جفت کردن روش‌های انقباضی جدید هم‌چون LASSO بیزی، LASSO گروهی، LAD-SCAD با مدل غیر خطی شبکه عصبی مصنوعی و بررسی کارایی آن‌ها در انتخاب متغیر برای مطالعات QSAR
- ✓ استفاده از روش‌های انقباضی مورد استفاده در رساله حاضر با روش یادگیری عمیق به‌عنوان روش مدل‌سازی پر کاربرد و کارامد
- ✓ به‌کارگیری روش‌های پیش‌پردازش همچون روش غربالگری مستقل مطمئن (SIS)[1] و الگوریتم تکرار شونده آن (ISIS)[2] قبل از انتخاب متغیر با روش‌های انقباضی
- ✓ استفاده از روش‌های انقباضی برای انتخاب متغیر در مطالعات QSAR مربوط به بیماری‌های آلزایمر، سرطان‌های پروستات و سینه و بررسی میزان کارایی این روش‌ها در مطالعات QSAR این دسته ترکیبات

[1] Sure Independence Screening
[2] Iterative Sure Independence Screening

۳- ۴ منابع

[1] Young D.C. (**2009**), "Computational drug design: a guide for computational and medicinal chemists", John Wiley & Sons,

[2] Wold S. (**1995**) "Chemometrics; what do we mean with it, and what do we want from it?", Chemometrics and Intelligent Laboratory Systems. **30**, 1, pp 109-115

[3] Hansch C., and Fujita T. (**1964**) "p-σ-π Analysis. A method for the correlation of biological activity and chemical structure", Journal of the American Chemical Society. **86**, 8, pp 1616-1626

[4] Cherkasov A., Muratov E.N., Fourches D., Varnek A., et al. (**2014**) "QSAR modeling: where have you been? Where are you going to?", Journal of medicinal chemistry. **57**, 12, pp 4977-5010

[5] Verma J., Khedkar V.M., and Coutinho E.C. (**2010**) "3D-QSAR in drug design-a review", Current topics in medicinal chemistry. **10**, 1, pp 95-115

[6] Golbraikh A., Wang X.S., Zhu H., and Tropsha A. (**2012**) "Predictive QSAR modeling: methods and applications in drug discovery and chemical risk assessment", Handbook of computational chemistry. pp 1309-1342

[7] Ma J., Sheridan R.P., Liaw A., Dahl G.E., et al. (**2015**) "Deep neural nets as a method for quantitative structure–activity relationships", Journal of chemical information and modeling. **55**, 2, pp 263-274

[8] Andrada M.F., Vega-Hissi E.G., Estrada M.R., and Garro Martinez J.C. (**2017**) "Impact assessment of the rational selection of training and test sets on the predictive ability of QSAR models", SAR and QSAR in Environmental Research. **28**, 12, pp 1011-1023

[9] Todeschini R., and Consonni V. (**2008**), "Handbook of molecular descriptors", John Wiley & Sons,

[10] Tropsha A., Gramatica P., and Gombar V.K. (**2003**) "The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models", QSAR & Combinatorial Science. **22**, 1, pp 69-77

[11] Katritzky A.R., Dobchev D.A., Slavov S., and Karelson M. (**2008**) "Legitimate utilization of large descriptor pools for QSPR/QSAR models", Journal of chemical information and modeling. **48**, 11, pp 2207-2213

[12] Katritzky A.R., Petrukhin R., Tatham D., Basak S., et al. (**2001**) "Interpretation of quantitative structure− property and− activity relationships", Journal of chemical information and computer sciences. **41**, 3, pp 679-685

[13] Jiang J., Duan W., Wei Q., Zhao X., et al. (**2020**) "Development of quantitative structure-property relationship (QSPR) models for predicting the thermal hazard of ionic liquids: A review of methods and models", Journal of Molecular Liquids. **301**, pp 112471

[14] Snodin D.J. (**2002**) "An EU perspective on the use of in vitro methods in regulatory pharmaceutical toxicology", Toxicology letters. **127**, 1-3, pp 161-168

[15] Kraljevic S., Stambrook P.J., and Pavelic K. (**2004**) "Accelerating drug discovery: Although the evolution of '- omics' methodologies is still in its infancy, both the pharmaceutical industry and patients could benefit from their implementation in the drug development process", EMBO reports. **5**, 9, pp 837-842

[16] Adams C.P., and Brantner V.V. (**2006**) "Estimating the cost of new drug development:

is it really $802 million?", Health affairs. **25**, 2, pp 420-428
[17] Kola I., and Landis J. (**2004**) "Can the pharmaceutical industry reduce attrition rates?", Nature reviews Drug discovery. **3**, 8, pp 711-716
[18] Lionberger R.A. (**2008**) "FDA critical path initiatives: opportunities for generic drug development", The AAPS journal. **10**, 1, pp 103-109
[19] Andricopulo A.D., Guido R.V., and Oliva G. (**2008**) "Virtual screening and its integration with modern drug design technologies", Current medicinal chemistry. **15**, 1, pp 37-46
[20] Schwaighofer A., Schroeter T., Mika S., and Blanchard G. (**2009**) "How wrong can we get? A review of machine learning approaches and error bars", Combinatorial chemistry & high throughput screening. **12**, 5, pp 453-468
[21] Valerio Jr L.G. (**2009**) "In silico toxicology for the pharmaceutical sciences", Toxicology and applied pharmacology. **241**, 3, pp 356-370
[22] Yap C., Xue Y., and Chen Y. (**2006**) "Application of support vector machines to in silico prediction of cytochrome P450 enzyme substrates and inhibitors", Current topics in medicinal chemistry. **6**, 15, pp 1593-1607
[23] Meeting of the Chemicals Committee OECD principles, For the validation of (quantitative) structure–activity relationship models, **2019**
[24] Hdoufane I., Ounaissi D., Dermoune A., and Cherqaoui D. (**2021**) "Development of QSAR Models Using Singular Value Decomposition Method: A Case Study for Predicting Anti-HIV-1 and Anti-HCV Biological Activities", Biointerface Research in Applied Chemistry. **12**, 3, pp 3090-3105
[25] Puzyn T., Mostrag-Szlichtyng A., Gajewicz A., Skrzyński M., et al. (**2011**) "Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models", Structural Chemistry. **22**, 4, pp 795-804
[26] HyperChem 8 (by Hypercube, Inc.), Coleman W.F., and Arumainayagam C.R., **2012**
[27] R: A language and environment for statistical computing, Team R.C., Vienna, Austria, **2020**
[28] Package ‘caret’, Kuhn M., Wing J., Weston S., Williams A., et al., The R Journal, **2020**
[29] Tibshirani R. (**1996**) "Regression shrinkage and selection via the lasso", Journal of the Royal Statistical Society: Series B (Methodological). **58**, 1, pp 267-288
[30] Zou H. (**2006**) "The adaptive lasso and its oracle properties", Journal of the American statistical association. **101**, 476, pp 1418-1429
[31] Fan J., and Li R. (**2001**) "Variable selection via nonconcave penalized likelihood and its oracle properties", Journal of the American statistical Association. **96**, 456, pp 1348-1360
[32] Wang H., Li G., and Jiang G. (**2007**) "Robust regression shrinkage and consistent variable selection through the LAD-Lasso", Journal of Business & Economic Statistics. **25**, 3, pp 347-355
[33] Package ‘ncvreg’, Breheny P., and Breheny M.P., **2021**
[34] Package ‘parcor’, Kraemer N., Schaefer J., and Kraemer M.N., **2014**
[35] Package ‘quantreg’, Koenker R., Portnoy S., Ng P.T., Zeileis A., et al., Cran R-project. org, **2018**
[36]Jammalamadaka S.R., Introduction to linear regression analysis, in, Taylor & Francis, 2003.
[37] Saleh A.M.E., Arashi M., and Kibria B.G. (**2019**), "Theory of ridge regression estimation with applications", John Wiley & Sons,

[38] Sadeghi F., Afkhami A., Madrakian T., and Ghavami R. (**2021**) "A new approach for simultaneous calculation of pIC 50 and logP through QSAR/QSPR modeling on anthracycline derivatives: a comparable study", Journal of the Iranian Chemical Society. pp 1-16
[39] Chamjangali M.A., Beglari M., and Bagherian G. (**2007**) "Prediction of cytotoxicity data (CC50) of anti-HIV 5-pheny-l-phenylamino-1H-imidazole derivatives by artificial neural network trained with Levenberg–Marquardt algorithm", Journal of Molecular Graphics and Modelling. **26**, 1, pp 360-367
[40] Zupan J., and Gasteiger J. (**1993**), "Neural networks for chemists; an introduction", VCH publishers,
[41] Bring J. (**1994**) "How to standardize regression coefficients", The American Statistician. **48**, 3, pp 209-213
[42] Tropsha A. (**2010**) "Best practices for QSAR model development, validation, and exploitation", Molecular informatics. **29**, 6- 7, pp 476-488
[43] Golbraikh A., and Tropsha A. (**2002**) "Beware of q2!", Journal of molecular graphics and modelling. **20**, 4, pp 269-276
[44] Roy P.P., and Roy K. (**2008**) "On some aspects of variable selection for partial least squares regression models", QSAR & Combinatorial Science. **27**, 3, pp 302-313
[45] Consonni V., Ballabio D., and Todeschini R. (**2009**) "Comments on the definition of the Q 2 parameter for QSAR validation", Journal of chemical information and modeling. **49**, 7, pp 1669-1678
[46] Consonni V., Ballabio D., and Todeschini R. (**2010**) "Evaluation of model predictive ability by external validation techniques", Journal of chemometrics. **24**, 3- 4, pp 194-201
[47] Alonso H., Bliznyuk A.A., and Gready J.E. (**2006**) "Combining docking and molecular dynamic simulations in drug design", Medicinal research reviews. **26**, 5, pp 531-568
[48] Wermuth C., Ganellin C., Lindberg P., and Mitscher L. (**1998**) "Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998)", Pure and Applied Chemistry. **70**, 5, pp 1129-1143
[49] Liao C., Peach M.L., Yao R., and Nicklaus M.C., Molecular docking and structure-based virtual screening, in, Future Medicine, **2013**
[50] Feinstein W.P., and Brylinski M. (**2015**) "Calculating an optimal box size for ligand docking and virtual screening against experimental and predicted binding pockets", Journal of cheminformatics. **7**, 1, pp 1-10
[51] Goodsell D.S., Morris G.M., and Olson A.J. (**1996**) "Automated docking of flexible ligands: applications of AutoDock", Journal of molecular recognition. **9**, 1, pp 1-5
[52] Lang P.T., Aynechi T., Moustakas D., Shoichet B., et al. (**2007**) "Molecular docking and structure-based design", Drug Discovery Research: New Frontiers in the Post-Genomic Era. pp 3-23
[53] Mehellou Y., and De Clercq E. (**2010**) "Twenty-six years of anti-HIV drug discovery: where do we stand and where do we go?", Journal of medicinal chemistry. **53**, 2, pp 521-538
[54] Namasivayam V., Vanangamudi M., Kramer V.G., Kurup S., et al. (**2018**) "The journey of HIV-1 non-nucleoside reverse transcriptase inhibitors (NNRTIs) from lab to clinic", Journal of medicinal chemistry. **62**, 10, pp 4851-4883
[55] Zhan P., Pannecouque C., De Clercq E., and Liu X. (**2016**) "Anti-HIV drug discovery and development: current innovations and future trends: miniperspective", Journal of

medicinal chemistry. **59**, 7, pp 2849-2878
[56] Collaboration A.T.C. (**2008**) "Life expectancy of individuals on combination antiretroviral therapy in high-income countries: a collaborative analysis of 14 cohort studies", The Lancet. **372**, 9635, pp 293-299
[57] Brechtl J.R., Breitbart W., Galietta M., Krivo S., et al. (**2001**) "The use of highly active antiretroviral therapy (HAART) in patients with advanced HIV infection: impact on medical, palliative care, and quality of life outcomes", Journal of pain and symptom management. **21**, 1, pp 41-51
[58]Meanwell N.A., Krystal M.R., Nowicka-Sans B., Langley D.R., et al., Inhibitors of HIV-1 attachment: the discovery and development of temsavir and its prodrug fostemsavir, in, ACS Publications, 2018.
[59] De Clercq E. (**1998**) "The role of non-nucleoside reverse transcriptase inhibitors (NNRTIs) in the therapy of HIV-1 infection", Antiviral research. **38**, 3, pp 153-179
[60] De Clercq E. (**2004**) "Non- nucleoside reverse transcriptase inhibitors (NNRTIs): past, present, and future", Chemistry & biodiversity. **1**, 1, pp 44-64
[61] De Clercq E., and Li G. (**2016**) "Approved antiviral drugs over the past 50 years", Clinical microbiology reviews. **29**, 3, pp 695-747
[62] Wu F., Zhao S., Yu B., Chen Y.-M., et al. (**2020**) "A new coronavirus associated with human respiratory disease in China", Nature. **579**, 7798, pp 265-269
[63] Zhou P., Yang X.-L., Wang X.-G., Hu B., et al. (**2020**) "A pneumonia outbreak associated with a new coronavirus of probable bat origin", nature. **579**, 7798, pp 270-273
[64] Fahmi I. (**2019**) "World Health Organization coronavirus disease 2019 (Covid-19) situation report", DroneEmprit. pp 1-9
[65] Azhar E.I., Hui D.S., Memish Z.A., Drosten C., et al. (**2019**) "The middle east respiratory syndrome (MERS)", Infectious Disease Clinics. **33**, 4, pp 891-905
[66] Hui D.S., and Zumla A. (**2019**) "Severe acute respiratory syndrome: historical, epidemiologic, and clinical features", Infectious Disease Clinics. **33**, 4, pp 869-889
[67] Wang C., Horby P.W., Hayden F.G., and Gao G.F. (**2020**) "A novel coronavirus outbreak of global health concern", The lancet. **395**, 10223, pp 470-473
[68] Ciotti M., Ciccozzi M., Terrinoni A., Jiang W.-C., et al. (**2020**) "The COVID-19 pandemic", Critical reviews in clinical laboratory sciences. **57**, 6, pp 365-388
[69] Zhang L., Lin D., Sun X., Curth U., et al. (**2020**) "Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors", Science. **368**, 6489, pp 409-412
[70] Lee C.-C., Kuo C.-J., Hsu M.-F., Liang P.-H., et al. (**2007**) "Structural basis of mercury- and zinc-conjugated complexes as SARS-CoV 3C-like protease inhibitors", FEBS letters. **581**, 28, pp 5454-5458
[71] Ramajayam R., Tan K.-P., and Liang P.-H. (**2011**) "Recent development of 3C and 3CL protease inhibitors for anti-coronavirus and anti-picornavirus drug discovery", Biochemical Society Transactions. **39**, 5, pp 1371-1375
[72] Barta I., Smerak P., Polivkova Z., Sestakova H., et al. (**2006**) "Current trends and perspectives in nutrition and cancer prevention", Neoplasma. **53**, 1, pp 19-25
[73] Siegel R.L., Miller K.D., and Jemal A. (**2020**) "Cancer statistics, 2020", CA: a cancer journal for clinicians. **70**, 1, pp 7-30
[74] Bade B.C., and Cruz C.S.D. (**2020**) "Lung cancer 2020: epidemiology, etiology, and prevention", Clinics in chest medicine. **41**, 1, pp 1-24

[75] Garcia-Echeverria C., and Sellers W. (**2008**) "Drug discovery approaches targeting the PI3K/Akt pathway in cancer", Oncogene. **27**, 41, pp 5511-5526
[76] Ran T., Lu T., Yuan H., Liu H., et al. (**2012**) "A selectivity study on mTOR/PI3Kα inhibitors by homology modeling and 3D-QSAR", Journal of molecular modeling. **18**, 1, pp 171-186
[77] Walker E.H., Pacold M.E., Perisic O., Stephens L., et al. (**2000**) "Structural determinants of phosphoinositide 3-kinase inhibition by wortmannin, LY294002, quercetin, myricetin, and staurosporine", Molecular cell. **6**, 4, pp 909-919
[78] Arcaro A., Volinia S., Zvelebil M.J., Stein R., et al. (**1998**) "Human phosphoinositide 3-kinase C2β, the role of calcium and the C2 domain in enzyme activity", Journal of Biological Chemistry. **273**, 49, pp 33082-33090
[79] Sadhu C., Masinovsky B., Dick K., Sowell C.G., et al. (**2003**) "Essential role of phosphoinositide 3-kinase δ in neutrophil directional movement", The Journal of Immunology. **170**, 5, pp 2647-2654
[80] Yin Y., Wu X., Han H.-W., Sha S., et al. (**2014**) "Discovery and synthesis of a novel series of potent, selective inhibitors of the PI3Kα: 2-alkyl-chromeno [4, 3-c] pyrazol-4 (2 H)-one derivatives", Organic & biomolecular chemistry. **12**, 45, pp 9157-9165
[81] Lu L., Sha S., Wang K., Zhang Y.-H., et al. (**2016**) "Discovery of chromeno [4, 3-c] pyrazol-4 (2H)-one containing carbonyl or oxime derivatives as potential, selective inhibitors PI3Kα", Chemical and Pharmaceutical Bulletin. pp c16-00388
[82] Sin D.W.-m., Wong Y.-c., Sham W.-c., and Wang D. (**2001**) "Development of an analytical technique and stability evaluation of 143 C3–C12 volatile organic compounds in Summa® canisters by gas chromatography–mass spectrometry", Analyst. **126**, 3, pp 310-321
[83] Luan F., Xue C., Zhang R., Zhao C., et al. (**2005**) "Prediction of retention time of a variety of volatile organic compounds based on the heuristic method and support vector machine", Analytica Chimica Acta. **537**, 1-2, pp 101-110
[84] Jalali-Heravi M., and Kyani A. (**2004**) "Use of computer-assisted methods for the modeling of the retention time of a variety of volatile organic compounds: a PCA-MLR-ANN approach", Journal of chemical information and computer sciences. **44**, 4, pp 1328-1335
[85] Sepehri B., Ghavami R., Farahbakhsh S., and Ahmadi R. (**2021**) "Machine learning-based quantitative structure–retention relationship models for predicting the retention indices of volatile organic pollutants", International Journal of Environmental Science and Technology. pp 1-10
[86] Zhu T., and Tao C. (**2022**) "Prediction models with multiple machine learning algorithms for POPs: The calculation of PDMS-air partition coefficient from molecular descriptor", Journal of Hazardous Materials. **423**, pp 127037
[87] Paskaleva V., and Kochev N. (**2019**) "GROUP CONTRIBUTION MODELING OF RETENTION INDICES OF VOLATILE ORGANIC COMPOUNDS IN PEPPERS", Scientific Work of the. pp 243
[88] He M., Yan P., Yang Z., Ye Y., et al. (**2018**) "Multi-analytical strategy for unassigned peaks using physical/mathematical separation, fragmental rules and retention index prediction: An example of sesquiterpene metabolites characterization in Cyperus rotundus", Journal of pharmaceutical and biomedical analysis. **154**, pp 476-485
[89] Kaliszan R. (**2007**) "QSRR: quantitative structure-(chromatographic) retention

relationships", Chemical reviews. **107**, 7, pp 3212-3246
[90] Kaliszan R., and Bączek T., QSAR in chromatography: quantitative structure–retention relationships (QSRRs), in: Recent advances in QSAR studies, Springer, **2010**
[91] Marrero-Ponce Y., Barigye S.J., Jorge-Rodríguez M.E., and Tran-Thi-Thu T. (**2018**) "QSRR prediction of gas chromatography retention indices of essential oil components", Chemical Papers. **72**, 1, pp 57-69
[92] Fragkaki A., Tsantili-Kakoulidou A., Angelis Y., Koupparis M., et al. (**2009**) "Gas chromatographic quantitative structure–retention relationships of trimethylsilylated anabolic androgenic steroids by multiple linear regression and partial least squares", Journal of Chromatography A. **1216**, 47, pp 8404-8420
[93] Jalali-Heravi M., and Fatemi M.H. (**2001**) "Artificial neural network modeling of Kovats retention indices for noncyclic and monocyclic terpenes", Journal of Chromatography A. **915**, 1-2, pp 177-183
[94] Golmohammadi H., and Fatemi M.H. (**2005**) "Artificial neural network prediction of retention factors of some benzene derivatives and heterocyclic compounds in micellar electrokinetic chromatography", Electrophoresis. **26**, 18, pp 3438-3444
[95] Algamal Z., and Lee M. (**2017**) "A new adaptive L1-norm for optimal descriptor selection of high-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives", SAR and QSAR in Environmental Research. **28**, 1, pp 75-90
[96] Algamal Z., Qasim M., and Ali H. (**2017**) "A QSAR classification model for neuraminidase inhibitors of influenza A viruses (H1N1) based on weighted penalized support vector machine", SAR and QSAR in Environmental Research. **28**, 5, pp 415-426
[97] Algamal Z.Y., Alhamzawi R., and Ali H.T.M. (**2018**) "Gene selection for microarray gene expression classification using Bayesian Lasso quantile regression", Computers in biology and medicine. **97**, pp 145-152
[98] Algamal Z.Y., and Lee M.H. (**2017**) "A novel molecular descriptor selection method in QSAR classification model based on weighted penalized logistic regression", Journal of Chemometrics. **31**, 10, pp e2915
[99] Al-Thanoon N.A., Qasim O.S., and Algamal Z.Y. (**2019**) "A new hybrid firefly algorithm and particle swarm optimization for tuning parameter estimation in penalized support vector machine with application in chemometrics", Chemometrics and Intelligent Laboratory Systems. **184**, pp 142-152
[100] Qasim M., Algamal Z., and Ali H.M. (**2018**) "A binary QSAR model for classifying neuraminidase inhibitors of influenza A viruses (H1N1) using the combined minimum redundancy maximum relevancy criterion with the sparse support vector machine", SAR and QSAR in Environmental Research. **29**, 7, pp 517-527
[101] Qasim O.S., Al-Thanoon N.A., and Algamal Z.Y. (**2020**) "Feature selection based on chaotic binary black hole algorithm for data classification", Chemometrics and Intelligent Laboratory Systems. **204**, pp 104104
[102] Wu S., Jiang H., Shen H., and Yang Z. (**2018**) "Gene selection in cancer classification using sparse logistic regression with L1/2 regularization", Applied Sciences. **8**, 9, pp 1569
[103] Algamal Z.Y., Lee M.H., Al- Fakih A.M., and Aziz M. (**2017**) "High- dimensional QSAR classification model for anti- hepatitis C virus activity of thiourea derivatives based on the sparse logistic regression model with a bridge penalty", Journal of Chemometrics. **31**, 6, pp e2889
[104] Alharthi A., Lee M., Algamal Z., and Al-Fakih A. (**2020**) "Quantitative structure-

activity relationship model for classifying the diverse series of antifungal agents using ratio weighted penalized logistic regression", SAR and QSAR in Environmental Research. **31**, 8, pp 571-583
[105] Alharthi A.M., Lee M.H., and Algamal Z.Y. (**2021**) "Gene selection and classification of microarray gene expression data based on a new adaptive L1-norm elastic net penalty", Informatics in Medicine Unlocked. pp 100622
[106] Algamal Z. (**2017**) "An efficient gene selection method for high-dimensional microarray data based on sparse logistic regression", Electronic Journal of Applied Statistical Analysis. **10**, 1, pp 242-256
[107] Peng X.-L., Yin H., Li R., and Fang K.-T. (**2006**) "The application of Kriging and empirical Kriging based on the variables selected by SCAD", Analytica chimica acta. **578**, 2, pp 178-185
[108] Algamal Z.Y., Lee M.H., Al- Fakih A.M., and Aziz M. (**2015**) "High- dimensional QSAR prediction of anticancer potency of imidazo [4, 5- b] pyridine derivatives using adjusted adaptive LASSO", Journal of Chemometrics. **29**, 10, pp 547-556
[109] Algamal Z., Lee M., Al-Fakih A., and Aziz M. (**2016**) "High-dimensional QSAR modelling using penalized linear regression model with L 1/2-norm", SAR and QSAR in Environmental Research. **27**, 9, pp 703-719
[110] Al-Fakih A., Algamal Z., Lee M., and Aziz M. (**2018**) "A penalized quantitative structure–property relationship study on melting point of energetic carbocyclic nitroaromatic compounds using adaptive bridge penalty", SAR and QSAR in Environmental Research. **29**, 5, pp 339-353
[111] Majumdar S., Basak S.C., Lungu C., Diudea M., et al. (**2018**) "Mathematical structural descriptors and mutagenicity assessment: a study with congeneric and diverse datasets$", SAR and QSAR in Environmental Research. **29**, 8, pp 579-590
[112] Xia L.-Y., Wang Y.-W., Meng D.-Y., Yao X.-J., et al. (**2018**) "Descriptor selection via log-sum regularization for the biological activities of chemical structure", International journal of molecular sciences. **19**, 1, pp 30
[113] Al- Dabbagh Z.T., and Algamal Z.Y. (**2019**) "Least absolute deviation estimator-bridge variable selection and estimation for quantitative structure–activity relationship model", Journal of Chemometrics. **33**, 7, pp e3139
[114] Mozafari Z., Arab Chamjangali M., Arashi M., and Goudarzi N. (**2021**) "Performance of smoothly clipped absolute deviation as a variable selection method in the artificial neural network- based QSAR studies", Journal of Chemometrics. **35**, 5, pp e3338
[115] Mozafari Z., Chamjangali M.A., Arashi M., and 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, pp 863-888
[116] HyperChem(TM) Professional 8.0.5, Hypercube I., 1115 NW 4th Street, Gainesville, Florida 32601, USA., **2008**
[117] Dragon software: An easy approach to molecular descriptor calculations, Mauri A., Consonni V., Pavan M., and Todeschini R., University of Milano-Bicocca, **2006**
[118] IBM SPSS statistics 25 for Windows student, George D., and Mallery P., **2017**
[119] Matlab Version R2017a, The MathWorks Inc., Massachusetts, **2017**
[120] Origin Lab Version 2021 , OriginLab Corporation, Northampton, MA, USA., **2021**

[121] Morris G.M., Huey R., Lindstrom W., Sanner M.F., et al. (**2009**) "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility", Journal of computational chemistry. **30**, 16, pp 2785-2791
[122] ViewerLite v 5.0, Accelrys Inc, San Diego, CA, **2001**
[123] BIOVIA discovery studio, Biovia D.S., San Diego, CA, USA, **2021**
[124] Moss J.A. (**2013**) "HIV/AIDS Review", Radiologic technology. **84**, 3, pp 247-267
[125] de Béthune M.-P. (**2010**) "Non-nucleoside reverse transcriptase inhibitors (NNRTIs), their discovery, development, and use in the treatment of HIV-1 infection: a review of the last 20 years (1989–2009)", Antiviral research. **85**, 1, pp 75-90
[126] Reust C.E. (**2011**) "Common adverse effects of antiretroviral therapy for HIV disease", American family physician. **83**, 12, pp 1443-1451
[127] Chen X., Liu X., Meng Q., Wang D., et al. (**2013**) "Novel piperidinylamino-diarylpyrimidine derivatives with dual structural conformations as potent HIV-1 non-nucleoside reverse transcriptase inhibitors", Bioorganic & medicinal chemistry letters. **23**, 24, pp 6593-6597
[128] De La Rosa M., Kim H.W., Gunic E., Jenket C., et al. (**2006**) "Tri-substituted triazoles as potent non-nucleoside inhibitors of the HIV-1 reverse transcriptase", Bioorganic & medicinal chemistry letters. **16**, 17, pp 4444-4449
[129] Gagnon A., Landry S., Coulombe R., Jakalian A., et al. (**2009**) "Investigation on the role of the tetrazole in the binding of thiotetrazolylacetanilides with HIV-1 wild type and K103N/Y181C double mutant reverse transcriptases", Bioorganic & medicinal chemistry letters. **19**, 4, pp 1199-1205
[130] Moyle G., Boffito M., Stoehr A., Rieger A., et al. (**2010**) "Phase 2a randomized controlled trial of short-term activity, safety, and pharmacokinetics of a novel nonnucleoside reverse transcriptase inhibitor, RDEA806, in HIV-1-positive, antiretroviral-naive subjects", Antimicrobial agents and chemotherapy. **54**, 8, pp 3170-3178
[131] O'Meara J.A., Jakalian A., LaPlante S., Bonneau P.R., et al. (**2007**) "Scaffold hopping in the rational design of novel HIV-1 non-nucleoside reverse transcriptase inhibitors", Bioorganic & medicinal chemistry letters. **17**, 12, pp 3362-3366
[132] Tzoupis H., Leonis G., Durdagi S., Mouchlis V., et al. (**2011**) "Binding of novel fullerene inhibitors to HIV-1 protease: insight through molecular dynamics and molecular mechanics Poisson–Boltzmann surface area calculations", Journal of computer-aided molecular design. **25**, 10, pp 959-976
[133] Wang Z., Wu B., Kuhen K.L., Bursulaya B., et al. (**2006**) "Synthesis and biological evaluations of sulfanyltriazoles as novel HIV-1 non-nucleoside reverse transcriptase inhibitors", Bioorganic & medicinal chemistry letters. **16**, 16, pp 4174-4177
[134] Li X., Lu X., Chen W., Liu H., et al. (**2014**) "Arylazolyl (azinyl) thioacetanilides. Part 16: Structure-based bioisosterism design, synthesis and biological evaluation of novel pyrimidinylthioacetanilides as potent HIV-1 inhibitors", Bioorganic & medicinal chemistry. **22**, 19, pp 5290-5297
[135] Zhan P., Chen W., Li Z., Li X., et al. (**2012**) "Discovery of novel 2-(3-(2-chlorophenyl) pyrazin-2-ylthio)-N-arylacetamides as potent HIV-1 inhibitors using a structure-based bioisosterism approach", Bioorganic & medicinal chemistry. **20**, 23, pp 6795-6802
[136] Zhan P., Li X., Li Z., Chen X., et al. (**2012**) "Structure-based bioisosterism design, synthesis and biological evaluation of novel 1, 2, 4-triazin-6-ylthioacetamides as potent HIV-1 NNRTIs", Bioorganic & medicinal chemistry letters. **22**, 23, pp 7155-7162

[137] Armand-Ugón M., Moncunill G., Gonzalez E., Mena M., et al. (**2010**) "Different selection patterns of resistance and cross-resistance to HIV-1 agents targeting CCR5", Journal of antimicrobial chemotherapy. **65**, 3, pp 417-424
[138] Beiknejad D., Chaichi M.J., and Fatemi M.H. (**2021**) "Prediction of ozonolytic decolorization half-lives of azo dyes in a continuous-flow system using QSPR", Dyes and Pigments. **185**, pp 108915
[139] Montgomery D.C., Peck E.A., and Vining G.G. (**2021**), "Introduction to linear regression analysis", John Wiley & Sons,
[140] Coulibaly P., Anctil F., and Bobée B. (**1999**) "Prévision hydrologique par réseaux de neurones artificiels: état de l'art", Canadian Journal of civil engineering. **26**, 3, pp 293-304
[141] Othman F., and Naseri M. (**2011**) "Reservoir inflow forecasting using artificial neural network", International journal of physical sciences. **6**, 3, pp 434-440
[142] Sahigara F., Mansouri K., Ballabio D., Mauri A., et al. (**2012**) "Comparison of different approaches to define the applicability domain of QSAR models", Molecules. **17**, 5, pp 4791-4810
[143] Tsang K.W., Ho P.L., Ooi G.C., Yee W.K., et al. (**2003**) "A cluster of cases of severe acute respiratory syndrome in Hong Kong", New England Journal of Medicine. **348**, 20, pp 1977-1985
[144] Lu H., Stratton C.W., and Tang Y.W. (**2020**) "Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle", Journal of medical virology. **92**, 4, pp 401
[145] Zhu N., Zhang D., Wang W., Li X., et al. (**2020**) "A novel coronavirus from patients with pneumonia in China, 2019", New England journal of medicine. pp 1-10
[146] Sohrabi C., Alsafi Z., O'neill N., Khan M., et al. (**2020**) "World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19)", International journal of surgery. **76**, pp 71-76
[147] Chen Y.W., Yiu C.-P.B., and Wong K.-Y. (**2020**) "Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CL pro) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates", F1000Research. **9**, pp 10-16
[148] Zhavoronkov A., Zagribelnyy B., Zhebrak A., Aladinskiy V., et al. (**2020**) "Potential non-covalent SARS-CoV-2 3C-like protease inhibitors designed using generative deep learning approaches and reviewed by human medicinal chemist in virtual reality", pp 13-19
[149] Chen L.-R., Wang Y.-C., Lin Y.W., Chou S.-Y., et al. (**2005**) "Synthesis and evaluation of isatin derivatives as effective SARS coronavirus 3CL protease inhibitors", Bioorganic & Medicinal Chemistry Letters. **15**, 12, pp 3058-3062
[150] Liu W., Zhu H.-M., Niu G.-J., Shi E.-Z., et al. (**2014**) "Synthesis, modification and docking studies of 5-sulfonyl isatin derivatives as SARS-CoV 3C-like protease inhibitors", Bioorganic & medicinal chemistry. **22**, 1, pp 292-302
[151] Chen C.-N., Lin C.P., Huang K.-K., Chen W.-C., et al. (**2005**) "Inhibition of SARS-CoV 3C-like protease activity by theaflavin-3, 3'-digallate (TF3)", Evidence-Based Complementary and Alternative Medicine. **2**, 2, pp 209-215
[152] Chen L., Gui C., Luo X., Yang Q., et al. (**2005**) "Cinanserin is an inhibitor of the 3C-like proteinase of severe acute respiratory syndrome coronavirus and strongly reduces virus replication in vitro", Journal of virology. **79**, 11, pp 7095-7103
[153] Niu C., Yin J., Zhang J., Vederas J.C., et al. (**2008**) "Molecular docking identifies the

binding of 3-chloropyridine moieties specifically to the S1 pocket of SARS-CoV Mpro", Bioorganic & medicinal chemistry. **16**, 1, pp 293-302
[154] Lu I.-L., Mahindroo N., Liang P.-H., Peng Y.-H., et al. (**2006**) "Structure-based drug design and structural biology study of novel nonpeptide inhibitors of severe acute respiratory syndrome coronavirus main protease", Journal of medicinal chemistry. **49**, 17, pp 5154-5161
[155] Tsai K.-C., Chen S.-Y., Liang P.-H., Lu I.-L., et al. (**2006**) "Discovery of a novel family of SARS-CoV protease inhibitors by virtual screening and 3D-QSAR studies", Journal of medicinal chemistry. **49**, 12, pp 3485-3495
[156] Park J.-Y., Kim J.H., Kim Y.M., Jeong H.J., et al. (**2012**) "Tanshinones as selective and slow-binding inhibitors for SARS-CoV cysteine proteases", Bioorganic & medicinal chemistry. **20**, 19, pp 5928-5935
[157] Dooley A.J., Shindo N., Taggart B., Park J.-G., et al. (**2006**) "From genome to drug lead: identification of a small-molecule inhibitor of the SARS virus", Bioorganic & medicinal chemistry letters. **16**, 4, pp 830-833
[158] Kao R.Y., Tsui W.H., Lee T.S., Tanner J.A., et al. (**2004**) "Identification of novel small-molecule inhibitors of severe acute respiratory syndrome-associated coronavirus by chemical genetics", Chemistry & biology. **11**, 9, pp 1293-1299
[159] Shehroz M., Zaheer T., and Hussain T. (**2020**) "Computer-aided drug design against spike glycoprotein of SARS-CoV-2 to aid COVID-19 treatment", Heliyon. **6**, 10, pp e05278
[160] Bacha U., Barrila J., Velazquez-Campoy A., Leavitt S.A., et al. (**2004**) "Identification of novel inhibitors of the SARS coronavirus main protease 3CLpro", Biochemistry. **43**, 17, pp 4906-4912
[161] Fujita T., and Winkler D.A. (**2016**) "Understanding the roles of the “two QSARs”", Journal of chemical information and modeling. **56**, 2, pp 269-274
[162] Gramatica P. (**2020**) "Principles of QSAR modeling: comments and suggestions from personal experience", International Journal of Quantitative Structure-Property Relationships (IJQSPR). **5**, 3, pp 61-97
[163] The caret package. R Foundation for Statistical Computing, Vienna, Austria, Kuhn M., URL https://cran. r-project. org/package= caret, **2012**
[164] Ebrahimi M., Khayamian T., and Gharaghani S. (**2012**) "Interactions between Activin-Like Kinase 5 (ALK5) receptor and its inhibitors and the construction of a Docking Descriptor-Based QSAR model", Journal of the Brazilian Chemical Society. **23**, 11, pp 2043-2092
[165] Eklund M., Norinder U., Boyer S., and Carlsson L. (**2012**) "Benchmarking variable selection in QSAR", Molecular informatics. **31**, 2, pp 173-179
[166] Ghasemi G., Arshadi S., Rashtehroodi A.N., Nirouei M., et al. (**2013**) "QSAR investigation on quinolizidinyl derivatives in Alzheimer’s disease", Journal of Computational Medicine. **2013**, pp 13-18
[167] Wacker S., and Noskov S.Y. (**2018**) "Performance of machine learning algorithms for qualitative and quantitative prediction drug blockade of hERG1 channel", Computational Toxicology. **6**, pp 55-63
[168] Zhu X.-W., Xin Y.-J., and Ge H.-L. (**2015**) "Recursive random forests enable better predictive performance and model interpretation than variable selection by LASSO", Journal of chemical information and modeling. **55**, 4, pp 736-746
[169] Chen W., Zhan P., Rai D., De Clercq E., et al. (**2014**) "Discovery of 2-pyridone derivatives as potent HIV-1 NNRTIs using molecular hybridization based on

crystallographic overlays", Bioorganic & medicinal chemistry. **22**, 6, pp 1863-1872
[170] Chen X., Zhan P., Liu X., Cheng Z., et al. (**2012**) "Design, synthesis, anti-HIV evaluation and molecular modeling of piperidine-linked amino-triazine derivatives as potent non-nucleoside reverse transcriptase inhibitors", Bioorganic & medicinal chemistry. **20**, 12, pp 3856-3864
[171] Li D., Zhan P., Liu H., Pannecouque C., et al. (**2013**) "Synthesis and biological evaluation of pyridazine derivatives as novel HIV-1 NNRTIs", Bioorganic & medicinal chemistry. **21**, 7, pp 2128-2134
[172] Wang J., Zhan P., Li Z., Liu H., et al. (**2014**) "Discovery of nitropyridine derivatives as potent HIV-1 non-nucleoside reverse transcriptase inhibitors via a structure-based core refining approach", European journal of medicinal chemistry. **76**, pp 531-538
[173] Yin Y., Hu J.-Q., Wu X., Sha S., et al. (**2019**) "Design, synthesis and biological evaluation of novel chromeno [4, 3-c] pyrazol-4 (2H)-one derivates containing sulfonamido as potential PI3Kα inhibitors", Bioorganic & medicinal chemistry. **27**, 11, pp 2261-2267
[174] Borhani T.N., García-Muñoz S., Luciani C.V., Galindo A., et al. (**2019**) "Hybrid QSPR models for the prediction of the free energy of solvation of organic solute/solvent pairs", Physical Chemistry Chemical Physics. **21**, 25, pp 13706-13720
[175] Li M., Yu H., Wang Y., Li J., et al. (**2020**) "QSPR models for predicting the adsorption capacity for microplastics of polyethylene, polypropylene and polystyrene", Scientific reports. **10**, 1, pp 1-11
[176] Raffetti E., Donato F., Speziani F., Scarcella C., et al. (**2018**) "Polychlorinated biphenyls (PCBs) exposure and cardiovascular, endocrine and metabolic diseases: a population-based cohort study in a North Italian highly polluted area", Environment international. **120**, pp 215-222
[177] Sepehri B. (**2020**) "A review on created QSPR models for predicting ionic liquids properties and their reliability from chemometric point of view", Journal of Molecular Liquids. **297**, pp 112013
[178] Villaverde J.J., Sevilla-Morán B., López-Goti C., Alonso-Prados J.L., et al. (**2018**) "Considerations of nano-QSAR/QSPR models for nanopesticide risk assessment within the European legislative framework", Science of the Total Environment. **634**, pp 1530-1539
[179] Yan F., Shi Y., Wang Y., Jia Q., et al. (**2020**) "QSPR models for the properties of ionic liquids at variable temperatures based on norm descriptors", Chemical Engineering Science. **217**, pp 115540
[180] Zhu T., Gu L., Chen M., and Sun F. (**2021**) "Exploring QSPR models for predicting PUF-air partition coefficients of organic compounds with linear and nonlinear approaches", Chemosphere. **266**, pp 128962
[181] Acimovic M., Pezo L., Jeremic J.S., Cvetkovic M., et al. (**2020**) "QSRR model for predicting retention indices of geraniol chemotype of Thymus serpyllum essential oil", Journal of Essential Oil Bearing Plants. **23**, 3, pp 464-473
[182] Kaliszan R. (**2020**) "Recent advances in quantitative structure–retention relationships", Handbook of Analytical Separations. **8**, pp 587-632
[183] Matyushin D.D., Sholokhova A.Y., Karnaeva A.E., and Buryak A.K. (**2020**) "Various aspects of retention index usage for GC-MS library search: A statistical investigation using a diverse data set", Chemometrics and Intelligent Laboratory Systems. **202**, pp 104042
[184] Naylor B.C., Catrow J.L., Maschek J.A., and Cox J.E. (**2020**) "QSRR automator: a

tool for automating retention time prediction in lipidomics and metabolomics", Metabolites. **10**, 6, pp 237
[185] Pavlić B., Teslić N., Kojić P., and Pezo L. (**2020**) "Prediction of the GC-MS retention time for terpenoids detected in sage (Salvia officinalis L.) essential oil using QSRR approach", Journal of the Serbian Chemical Society. **85**, 1, pp 9-23
[186] Wu L., Gong P., Wu Y., Liao K., et al. (**2013**) "An integral strategy toward the rapid identification of analogous nontarget compounds from complex mixtures", Journal of Chromatography A. **1303**, pp 39-47
[187] Kalhor P., and Yarivand O. (**2016**) "Quantitative Structure-Property Relationship Study to Predict the Retention Times of Some Volatile Compounds in Rosé Wines", Analytical Chemistry Letters. **6**, 4, pp 371-383
[188] Andries J.P., Goodarzi M., and Vander Heyden Y. (**2020**) "Improvement of quantitative structure–retention relationship models for chromatographic retention prediction of peptides applying individual local partial least squares models", Talanta. **219**, pp 121266
[189] Zheng W., and Jin M. (**2020**) "Comparing multiple categories of feature selection methods for text classification", Digital Scholarship in the Humanities. **35**, 1, pp 208-224
[190] Al-Fakih A., Algamal Z., Lee M., and Aziz M. (**2017**) "A sparse QSRR model for predicting retention indices of essential oils based on robust screening approach", SAR and QSAR in Environmental Research. **28**, 8, pp 691-703
[191] Fleming-Jones M.E., and Smith R.E. (**2003**) "Volatile organic compounds in foods: a five year study", Journal of agricultural and food chemistry. **51**, 27, pp 8120-8127
[192] Vinci R.M., Jacxsens L., De Meulenaer B., Deconink E., et al. (**2015**) "Occurrence of volatile organic compounds in foods from the Belgian market and dietary exposure assessment", Food Control. **52**, pp 1-8
[193] Ghavami R., and Faham S. (**2010**) "QSRR models for Kováts' retention indices of a variety of volatile organic compounds on polar and apolar GC stationary phases using molecular connectivity indexes", Chromatographia. **72**, 9, pp 893-903
[194] Xu J., Zhang W., Adhikari K., and Shi Y.-C. (**2017**) "Determination of volatile compounds in heat-treated straight-grade flours from normal and waxy wheats", Journal of Cereal Science. **75**, pp 77-83
[195] Häppölä P., Havulinna A.S., Tasa T., Mars N.J., et al. (**2020**) "A data-driven medication score predicts 10-year mortality among aging adults", Scientific reports. **10**, 1, pp 1-10
[196] Ren J., Chamberlain P.P., Stamp A., Short S.A., et al. (**2008**) "Structural basis for the improved drug resistance profile of new generation benzophenone non-nucleoside HIV-1 reverse transcriptase inhibitors", Journal of medicinal chemistry. **51**, 16, pp 5000-5008
[197] Hevener K.E., Zhao W., Ball D.M., Babaoglu K., et al. (**2009**) "Validation of molecular docking programs for virtual screening against dihydropteroate synthase", Journal of chemical information and modeling. **49**, 2, pp 444-460
[198] Burley S.K., Berman H.M., Bhikadiya C., Bi C., et al. (**2019**) "RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy", Nucleic acids research. **47**, D1, pp D464-D474
[199] Zhou Q., Zhang N., Zhang C., Huang L., et al. (**2010**) "Molecular mechanism of enantioselective inhibition of acetolactate synthase by imazethapyr enantiomers", Journal of agricultural and food chemistry. **58**, 7, pp 4202-4206

[200] Daina A., Michielin O., and Zoete V. (**2017**) "SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules", Scientific reports. **7**, 1, pp 1-13
[201] Alam S., and Khan F. (**2014**) "QSAR and docking studies on xanthone derivatives for anticancer activity targeting DNA topoisomerase IIα", Drug design, development and therapy. **8**, pp 183
[202] Jin Z., Du X., Xu Y., Deng Y., et al. (**2020**) "Structure of M pro from SARS-CoV-2 and discovery of its inhibitors", Nature. **582**, 7811, pp 289-293
[203] De P., Bhayye S., Kumar V., and Roy K. (**2020**) "In silico modeling for quick prediction of inhibitory activity against 3CLpro enzyme in SARS CoV diseases", Journal of Biomolecular Structure and Dynamics. pp 1-27
[204] Toropov A.A., Toropova A.P., Veselinović A.M., Leszczynska D., et al. (**2020**) "SARS-CoV Mpro inhibitory activity of aromatic disulfide compounds: QSAR model", Journal of Biomolecular Structure and Dynamics. pp 1-7
[205] Khaerunnisa S., Kurniawan H., Awaluddin R., Suhartati S., et al. (**2020**) "Potential inhibitor of COVID-19 main protease (Mpro) from several medicinal plant compounds by molecular docking study", Preprints. **2020**, pp 2020030226
[206] Cao B., Wang Y., Wen D., Liu W., et al. (**2020**) "A trial of lopinavir–ritonavir in adults hospitalized with severe Covid-19", New England Journal of Medicine. pp 29-39
[207] Abdelrheem D.A., Ahmed S.A., Abd El-Mageed H., Mohamed H.S., et al. (**2020**) "The inhibitory effect of some natural bioactive compounds against SARS-CoV-2 main protease: insights from molecular docking analysis and molecular dynamic simulation", Journal of Environmental Science and Health, Part A. **55**, 11, pp 1373-1386
[208] Hasan A., Paray B.A., Hussain A., Qadir F.A., et al. (**2021**) "A review on the cleavage priming of the spike protein on coronavirus by angiotensin-converting enzyme-2 and furin", Journal of Biomolecular Structure and Dynamics. **39**, 8, pp 3025-3033
[209] Murugan N.A., Kumar S., Jeyakanthan J., and Srivastava V. (**2020**) "Searching for target-specific and multi-targeting organics for Covid-19 in the Drugbank database with a double scoring approach", Scientific reports. **10**, 1, pp 1-16
[210] Rehman M., AlAjmi M.F., and Hussain A. (**2021**) "Natural compounds as inhibitors of SARS-CoV-2 main protease (3CLpro): A molecular docking and simulation approach to combat COVID-19", Current Pharmaceutical Design. pp 10-19
[211] Yu R., Chen L., Lan R., Shen R., et al. (**2020**) "Computational screening of antagonists against the SARS-CoV-2 (COVID-19) coronavirus by molecular docking", International Journal of Antimicrobial Agents. **56**, 2, pp 106012
[212] Saíz-Urra L., González M.P., and Teijeira M. (**2007**) "2D-autocorrelation descriptors for predicting cytotoxicity of naphthoquinone ester derivatives against oral human epidermoid carcinoma", Bioorganic & medicinal chemistry. **15**, 10, pp 3565-3571
[213] Sun M., Chen J., Cai J., Cao M., et al. (**2010**) "Simultaneously Optimized Support Vector Regression Combined With Genetic Algorithm for QSAR Analysis of KDR/VEGFR- 2 Inhibitors", Chemical biology & drug design. **75**, 5, pp 494-505
[214] Jin X., Peldszus S., and Huck P.M. (**2015**) "Predicting the reaction rate constants of micropollutants with hydroxyl radicals in water using QSPR modeling", Chemosphere. **138**, pp 1-9
[215] Kertesz D.J., Brotherton-Pleiss C., Yang M., Wang Z., et al. (**2010**) "Discovery of

piperidin-4-yl-aminopyrimidines as HIV-1 reverse transcriptase inhibitors. N-benzyl derivatives with broad potency against resistant mutant viruses", Bioorganic & medicinal chemistry letters. **20**, 14, pp 4215-4218
[216] Wong K.Y., Mercader A.G., Saavedra L.M., Honarparvar B., et al. (**2014**) "QSAR analysis on tacrine-related acetylcholinesterase inhibitors", Journal of biomedical science. **21**, 1, pp 1-8
[217] Chiari L.P., da Silva A.P., de Oliveira A.A., Lipinski C.F., et al. (**2021**) "Drug design of new sigma-1 antagonists against neuropathic pain: A QSAR study using partial least squares and artificial neural networks", Journal of Molecular Structure. **1223**, pp 129156
[218] Mandelker D., Gabelli S.B., Schmidt-Kittler O., Zhu J., et al. (**2009**) "A frequent kinase domain mutation that changes the interaction between PI3Kα and the membrane", Proceedings of the National Academy of Sciences. **106**, 40, pp 16996-17001
[219] Zefirov N.S., and Palyulin V.A. (**2001**) "QSAR for boiling points of "small" sulfides. Are the "high-quality structure-property-activity regressions" the real high quality QSAR models?", Journal of chemical information and computer sciences. **41**, 4, pp 1022-1027
[220] Pourbasheer E., Beheshti A., Khajehsharifi H., Ganjali M.R., et al. (**2013**) "QSAR study on hERG inhibitory effect of kappa opioid receptor antagonists by linear and non-linear methods", Medicinal Chemistry Research. **22**, 9, pp 4047-4058
[221] Goudarzi N., and Goodarzi M. (**2010**) "Application of successive projections algorithm (SPA) as a variable selection in a QSPR study to predict the octanol/water partition coefficients (K ow) of some halogenated organic compounds", Analytical Methods. **2**, 6, pp 758-764
[222] Schuur J.H., Selzer P., and Gasteiger J. (**1996**) "The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity", Journal of Chemical Information and Computer Sciences. **36**, 2, pp 334-344
[223] Saíz-Urra L., González M.P., and Teijeira M. (**2006**) "QSAR studies about cytotoxicity of benzophenazines with dual inhibition toward both topoisomerases I and II: 3D-MoRSE descriptors and statistical considerations about variable selection", Bioorganic & medicinal chemistry. **14**, 21, pp 7347-7358

Lipinski's rule of five were calculated for all proposed compounds and the suggested new compounds have acceptable pharmacological properties. In the third study, the combination of the least absolute deviation-least absolute shrinkage and selection operator (LAD-LASSO) was introduced as a new variable selection method for the ANN-based QSAR studies. The ANN model coupled with the efficient LAD-LASSO variable selection method was evaluated by predicting the biological activity of three datasets of chemical compounds. The LAD-LASSO variable selection method was applied, and the descriptors with the most relevance to the biological activities were selected. The selected descriptors were defined as the ANN inputs, and the designed models were optimized. The biological activity of the test set compounds was predicted using the corresponding optimum ANN models. The coefficients of determination (R^2) for the test data in the three datasets were equal to 0.87, 0.84, and 0.87. The applicability domain and Y-randomization test also proved the efficiency of the developed models. Finally, the created QSAR models were utilized to suggest numerous novel, highly potent chemical compounds by structurally modifying the weak molecules in all three datasets. The response value of the new suggested compounds was predicted using the optimum ANN models. According to the LR information, and the presence of different hydrophilic and hydrophobic interactions in the active site of the respective receptor indicates the high potential of chemical compounds. In the last work of this study, a combination of the SCAD and the ANN was utilized in the quantitative structure-retention indices(RIs) relationship (QSRR). The proposed SCAD method reduces the dimension of data before using the robust ANN modeling method. The efficiency of the SCAD-ANN methods was evaluated by the construction of a QSRR model between the most effective molecular descriptors and RIs for two sets of volatile organic compounds (VOCs). The SCAD method was applied to training data, and effective descriptors were selected in λ_{min} and were defined as the inputs to the ANN modeling method. All ANN parameters were optimized simultaneously. Some statistical parameters were computed, and the obtained results indicate that the constructed QSRR models have acceptable values. Also, the applicability domain analysis reveals that more than 95% of the data are in the confidence range. So, the prediction results of the SCAD-ANN models are reliable.

Keywords: QSAR, QSPR, Cancer, HIV, Coronavirus, SCAD, ALASSO, LAD-LASSO, ANN, Molecular docking

Abstract

The important goal of this dissertation is to use the penalized variable selection methods to select the most effective descriptors and coupling the penalized methods with the nonlinear artificial neural network (ANN) modeling. In this regard, different methods were used to make a relationship between the selected descriptors and the target response. So, in the first study, a hybrid of smoothly clipped absolute deviation (SCAD) and ANN was used as a new approach (SCAD-LM-ANN) in the quantitative structure-activity relationship (QSAR) studies. SCAD-LM-ANN exploits the useful shrinkage nature of penalized SCAD method in the reduction of high dimensional data prior to the modeling method. The performance of the SCAD-LM-ANN model was examined by establishing a QSAR model between Dragon derived descriptors and biological activities for a set of thioacetamide/acetanilide derivatives as HIV inhibitors. SCAD method with the 10-fold cross validation was applied to the dataset in the absence of test set compounds. A number of 11 descriptors were selected at a λ with the lowest cross-validation error (λmin). The selected descriptors were used as inputs of the ANN. All parameters affecting model performance were optimized, and the SCAD-LM-ANN model with the architecture of 5-5-1 was selected as the optimal QSAR model. Several statistical parameters were considered for the model evaluation. The obtained results prove the generalizability and predictability of the proposed SCAD-LM-ANN model. According to the established relationship in the recommended QSAR model, new derivatives were designed and suggested as new active HIV inhibitors for further studies. The accuracy of the suggested compounds was studied and confirmed by analyzing the ligand-receptor (LR) interactions derived from the molecular docking studies. In the second part of this dissertation, a new approach as a hybrid of adaptive least absolute shrinkage and selection operator (ALASSO) and ANN was introduced for the 3-chymotrypsin-like protease ($3CL^{pro}$) inhibitors as SARS CoV-2 potent compounds. Given the importance of this disease since 2019, the recommendation and design of new active compounds are crucial. After evaluating the accuracy and validity of the developed ALASSO-LM-ANN model, new compounds were proposed using effective descriptors, and the biological activity of the new compounds was predicted. The investigation of LR interactions were also performed using molecular docking study. The PK properties and

Faculty of Chemistry

Ph.D. Thesis in Analytical Chemistry

Application of penalized regression techniques as new methods of variable selection in the quantitative structure-activity (QSAR) and structure-property (QSPR) studies

By:

Zeinab Mozafari

Supervisor**:**

Dr. Mansour Arab Chamjangali

Advisors**:**

Dr. Mohammad Arashi

Dr. Nasser Goudarzi

February 2022