



استفاده از مدل های خطی و غیر خطی مطالعه ارتباط کمی ساختار-فعالیت (QSAR) جهت پیش بینی فعالیت دارویی مشتقات دی آریل پیریمیدین ها

مهدی نکویی^{۱*}، سید محمدرضا مطلبی طبائی^۲، نادیا نکویی^۳

^۱گروه شیمی، واحد شاهرود، دانشگاه آزاد اسلامی، شاهرود، ایران

^۲گروه بیوشیمی، واحد شاهرود، دانشگاه آزاد اسلامی، شاهرود، ایران

^۳گروه علوم آزمایشگاهی، دانشگاه علوم پزشکی شاهرود، شاهرود، ایران

تاریخ ثبت اولیه: ۱۴۰۳/۰۹/۲۵، تاریخ دریافت نسخه اصلاح شده: ۱۴۰۳/۱۲/۱۱، تاریخ پذیرش قطعی: ۱۴۰۳/۱۲/۲۰

چکیده

مطالعه QSAR (ارتباط کمی ساختار-فعالیت) برای پیش بینی فعالیت دارویی (مقادیر EC_{50}) برخی از مشتقات دی آریل پیریمیدین ها به عنوان مهارکننده های ویروس HIV، با استفاده از روش های رگرسیون خطی چند گانه (MLR) و شبکه های عصبی مصنوعی (ANN) به عنوان روش های خطی و غیر خطی انجام شد. در ابتدا، ساختار ترکیبات مورد نظر با استفاده از نرم افزار هایپر کم رسم و بهینه سازی شد. سپس، ترکیبات رسم شده به نرم افزار دراگون وارد شدند و ۱۴۸۱ توصیف کننده ی مولکولی برای هر مولکول محاسبه گردید. برای انتخاب مناسب ترین توصیف کننده ها، از روش رگرسیون گام به گام استفاده شد. پس از انتخاب توصیف کننده های بهینه، از دو روش MLR و ANN برای مدل سازی و پیش بینی فعالیت دارویی ترکیبات استفاده گردید. عملکرد هر مدل با استفاده از چندین پارامتر آماری ارزیابی شد. نتایج به دست آمده نشان دهنده ی برتری روش ANN نسبت به MLR در پیش بینی فعالیت دارویی ترکیبات بود.

واژه های کلیدی: ارتباط کمی ساختار-فعالیت، دی آریل پیریمیدین ها، رگرسیون خطی چند گانه، شبکه های عصبی مصنوعی

۱. مقدمه

طراحی دارو یکی از حوزه های حیاتی و پیشرفته علوم زیستی و پزشکی است که نقش بسیار مهمی در بهبود سلامت و کیفیت زندگی انسان ها دارد. این فرآیند شامل کشف، توسعه و ارزیابی داروهای جدید برای درمان بیماری ها است و به دانشمندان و محققان این امکان را می دهد تا با ترکیب علم و فناوری، راه حل های نوآورانه ای برای مقابله با بیماری ها ارائه دهند. اهمیت طراحی

*عهده دار مکاتبات: مهدی نکویی

نشانی: گروه شیمی، واحد شاهرود، دانشگاه آزاد اسلامی، شاهرود، ایران

پست الکترونیک: E-mail: m_nekoei@iau.ac.ir

تلفن: ۰۲۳۳۲۳۹۴۲۸۹

دارو به چندین جنبه مختلف مرتبط است. اولاً، طراحی دارو می‌تواند به تولید داروهای جدید و موثر برای درمان بیماری‌های مختلف از جمله بیماری‌های نادر و صعب‌العلاج کمک کند. این داروها می‌توانند زندگی میلیون‌ها نفر را نجات دهند و بهبود بخشند. ثانیاً، طراحی دارو می‌تواند به کاهش عوارض جانبی داروها کمک کند. با استفاده از روش‌های مدرن و فناوری‌های پیشرفته، محققان می‌توانند ترکیباتی با کارایی بالا و عوارض جانبی کمتر تولید کنند که منجر به افزایش کیفیت زندگی بیماران می‌شود. همچنین، طراحی دارو نقش مهمی در پیشگیری از بیماری‌ها دارد. واکسن‌ها و داروهای پیشگیرانه می‌توانند از گسترش بیماری‌ها جلوگیری کنند و به حفظ سلامت عمومی جامعه کمک کنند. این امر به ویژه در مقابله با بیماری‌های واگیردار و اپیدمی‌ها بسیار مهم است. از دیدگاه اقتصادی نیز، صنعت داروسازی یکی از صنایع پررونق و حیاتی است که به توسعه اقتصادی کشورها کمک می‌کند. طراحی و تولید داروهای جدید می‌تواند فرصت‌های شغلی جدیدی ایجاد کند و به رشد اقتصادی کمک کند [۱-۵].

در نتیجه، طراحی دارو نه تنها از دیدگاه بهبود سلامت و درمان بیماری‌ها بلکه از جنبه‌های اقتصادی و اجتماعی نیز اهمیت بسزایی دارد و به پیشرفت علوم زیستی و پزشکی کمک می‌کند. این فرآیند می‌تواند زندگی‌های بسیاری را تغییر دهد و آینده‌ای سالم‌تر و بهتر برای بشریت به ارمغان آورد. به همین دلایل، یکی از موضوعات مورد توجه و با اهمیت، طراحی، سنتز و تولید دارو می‌باشد. روندی که در گذشته منجر به کشف و توسعه داروهای جدید می‌شد به روش آزمون و خطا صورت می‌گرفت که روشی وقت‌گیر و هزینه‌بر بوده است. مشکل دیگری که پیش‌روی محققان است، عدم اطلاع آنها از فعالیت داروئی ترکیبات، قبل از انجام سنتز و بررسی تجربی آنها بوده و به همین دلیل یکی از مهم‌ترین اهداف شیمیدان‌ها و محققان دارویی پیش‌بینی فعالیت ترکیبات دارویی، قبل از سنتز آنها می‌باشد. چرا که انجام بسیاری از آزمایشات مستلزم صرف زمان و هزینه‌های زیادی است. از این رو نیاز به استفاده از روش‌های تئوری و محاسباتی که بدون انجام آزمایش بتوانند ویژگی و یا فعالیت ترکیبات دارویی را پیش‌بینی کنند ضروری به نظر می‌رسد [۶-۱۰]. ارتباط کمی ساختار-فعالیت (QSAR) یک روش محاسباتی در شیمی دارویی است که به بررسی ارتباط بین ساختار شیمیایی مولکول‌ها و فعالیت بیولوژیکی آنها می‌پردازد. QSAR به محققان کمک می‌کند تا با استفاده از مدل‌های ریاضی و آماری، پیش‌بینی کنند که چگونه تغییرات در ساختار مولکولی می‌تواند بر فعالیت بیولوژیکی ترکیبات تأثیر بگذارد. در QSAR، مولکول‌ها با استفاده از خصوصیات مختلف شیمیایی و فیزیکی توصیف می‌شوند، مثل ویژگی‌های الکترونی، هیدروفوبیک و اندازه مولکولی. سپس این خصوصیات با داده‌های تجربی از فعالیت بیولوژیکی ترکیبات مرتبط می‌شوند. با استفاده از روش‌های QSAR، محققان مدل‌هایی ایجاد می‌کنند که می‌توانند فعالیت بیولوژیکی ترکیبات جدید را پیش‌بینی کنند. این روش به ویژه در طراحی داروهای جدید بسیار مفید است، زیرا به محققان اجازه می‌دهد تا ترکیبات با پتانسیل بالا را شناسایی کنند و از انجام آزمایش‌های پرهزینه و زمان‌بر جلوگیری کنند. QSAR همچنین می‌تواند به بهبود کارایی و کاهش عوارض جانبی داروها نیز کمک کند. به طور کلی، QSAR یکی از ابزارهای قدرتمند در شیمی دارویی و طراحی دارو است که به بهبود فرآیندهای تحقیق و توسعه در این زمینه کمک می‌کند [۱۱-۱۵].

در سالیان اخیر، روند رو به افزایش مقالات منتشره در منابع علمی بر اساس QSAR، دلالت بر جایگاه منحصر به فرد این دیدگاه در شیمی نظری و تعمیق بینش دانشمندان در فهم و توجیه آن دارد [۱۱-۱۴]. از جمله روشهایی که جهت مدل سازی و پیش بینی فعالیت ترکیبات دارویی مورد استفاده قرار می گیرند، می توان به روش های خطی از جمله رگرسیون خطی چندگانه (MLR)، کمترین مربعات جزئی و روش های غیرخطی مانند شبکه های عصبی مصنوعی (ANN)، ماشین بردار پشتیبان، روش های فازی عصبی و... اشاره نمود [۱۶-۲۰].

هدف اصلی این تحقیق ارائه ی مدل های مناسب جهت پیش بینی فعالیت دارویی برخی از مشتقات دی آریل پیریمیدین ها بعنوان مهارکننده های ویروس HIV با استفاده از روش انتخاب مرحله ای^۱ به عنوان روش انتخاب متغیر (انتخاب مناسب ترین توصیف کننده ها) و روش های SW-MLR و SW-ANN به عنوان روش های مدلسازی می باشد.

۲. روش های محاسباتی

۲-۱. انتخاب سری داده ها

سری داده ها شامل فعالیت دارویی ۵۲ ترکیب از مشتقات دی آریل پیریمیدین بعنوان مهارکننده های ویروس HIV مورد بررسی قرار گرفت [۲۱]. قدرت بازدارندگی این ترکیبات به صورت EC₅₀ گزارش شده است. EC₅₀ عبارتست از غلظتی از ترکیب دارویی است که در آن ۵۰٪ حداکثر پاسخ دارو ایجاد گردد. این مقادیر به مقیاس لگاریتمی (pEC₅₀) تبدیل و مورد استفاده قرار گرفت. در این کار این ترکیبات به صورت تصادفی به دو گروه سری آموزش و سری تست تقسیم شدند، سری آموزش شامل ۴۲ مولکول (۸۰٪) و سری تست شامل ۱۰ مولکول (۲۰٪) می باشد. مقادیر pEC₅₀ به عنوان متغیر وابسته و توصیف کننده ها به عنوان متغیر مستقل انتخاب شدند. سری آموزش جهت ایجاد یک مدل مناسب و سری تست جهت ارزیابی مدل مورد استفاده قرار گرفت. لازم به ذکر است جهت مقایسه، سری پیش بینی (تست) در هر دو روش دارای ترکیبات یکسان می باشد. (در ANN از سری ارزیابی برای بررسی و جلوگیری از برازش اضافی استفاده گردید).

۲-۲. رسم و بهینه سازی ساختار مولکول ها

در ابتدا ساختار مولکولی هر ترکیب در نرم افزار هایپرکم^۲ ترسیم شد. سپس ساختار سه بعدی ترکیبات با استفاده از روش نیمه تجربی کوانتومی AM1 بهینه گردید. این بهینه سازی تا زمانی ادامه یافت که جذر میانگین مربعات گرادیان انرژی به ۰/۰۰۱ کیلوکالری بر مول رسید. هایپرکم یک نرم افزار قدرتمند برای رسم، بهینه سازی و مدل سازی مولکولی و شیمی محاسباتی است. این نرم افزار ابزارهای محاسباتی متنوعی از جمله مکانیک مولکولی، دینامیک مولکولی و مکانیک کوانتومی را ارائه می دهد. با استفاده از این نرم افزار می توان اطلاعات فراوانی نظیر زوایای پیوندی، طول پیوندها، زوایای پیچش، بار اتم ها، انرژی تشکیل مولکول و...

¹ Stepwise

² Hyperchem

را بدست آورد. برخی دیگر از قابلیت‌های این نرم افزار عبارتند از: توانایی نمایش ساختار مولکولی با قابلیت کنترل آن (از جمله انتخاب، چرخش، تبدیل و تغییر اندازه ساختار مولکولی)، دارای ابزارها و متدهای محاسباتی مختلف (از جمله تعیین تراز انرژی) و امکان تعریف نوع اتم، جرم اتمی و سایر ویژگی‌ها و

۲-۳. محاسبه توصیف کننده‌های مولکولی

توصیف کننده‌های مولکولی^۱ مقادیر عددی هستند که ویژگیهای مختلف مولکول را نشان می‌دهند. توصیف کننده‌های مولکولی نتیجه نهایی یک استدلال و روش ریاضی است که اطلاعات شیمیایی را به رمز تبدیل می‌کند و آنها را به صورت یک نماد نشان می‌دهد که ارائه دهنده یک ویژگی مولکول به صورت یک عدد مفید می‌باشد. هر یک از این توصیف کننده‌ها، اطلاعات خاصی از مولکول را در اختیار می‌گذارند. توصیف کننده‌های مولکولی مختلفی برای اهداف گوناگون به کار برده شده‌اند. اختلاف این توصیف کننده‌ها در پیچیدگی اطلاعات رمزگزاری شده و زمان مورد نیاز برای محاسبه می‌باشد.

این توصیف کننده‌ها در مدل‌سازی و پیش‌بینی فعالیت‌های بیولوژیکی و فیزیکی-شیمیایی ترکیبات بکار می‌روند. برخی از انواع اصلی توصیف کننده‌های مولکولی بصورت زیر می‌باشند.

توصیف کننده‌های توپولوژیکال (این توصیف کننده‌ها شامل ویژگی‌های مربوط به ساختار دو بعدی مولکول‌ها هستند، مانند شاخص‌های اتصال، شاخص‌های مسیر و شاخص‌های وابسته به اتم‌ها). توصیف کننده‌های هندسی (این توصیف کننده‌ها به اطلاعات سه‌بعدی مولکول‌ها اشاره دارند، مانند زوایای پیوند، طول پیوندها و زاویه‌های دی‌هیدرال). توصیف کننده‌های الکترونی (این توصیف کننده‌ها ویژگی‌های الکترونی مولکول‌ها را مشخص می‌کنند، مانند بارهای جزئی اتمی، انرژی اوربیتال‌های مولکولی و لحظات دو قطبی). توصیف کننده‌های فیزیکوشیمیایی (این توصیف کننده‌ها ویژگی‌های فیزیکی و شیمیایی مولکول‌ها را شامل می‌شوند، مانند نقطه ذوب، فشار بخار، ضریب تقسیم و حلالیت). توصیف کننده‌های طیف‌سنجی (این توصیف کننده‌ها شامل اطلاعات طیفی مولکول‌ها هستند، مانند طیف‌های IR، NMR، UV-Vis و... [۲۱]).

برای محاسبه توصیف کننده‌ها، بعد از رسم ساختارهای مولکولی به کمک نرم افزار Hyper Chem و بهینه سازی ساختار آنها، این ساختارها به نرم افزار Dragon وارد شده و توصیف کننده‌های مولکولی به تعداد ۱۴۸۱ مورد به وسیله این نرم افزار محاسبه شدند.

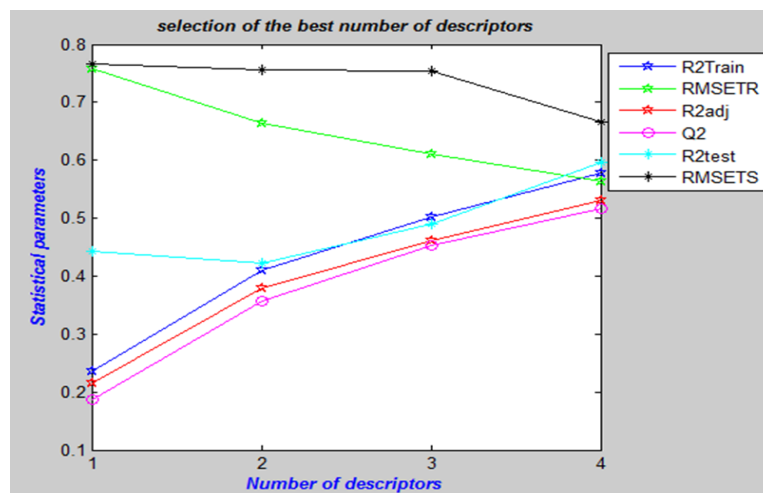
۲-۴. انتخاب مناسب‌ترین توصیف کننده‌ها

جهت انتخاب مناسب‌ترین توصیف کننده‌ها از روش رگرسیون مرحله‌ای^۲ استفاده شد. در روش رگرسیون مرحله‌ای، متغیرها یکی پس از دیگری وارد مدل شدند در این حالت، ابتدا متغیری وارد مدل می‌شود که بالاترین میزان همبستگی را با متغیر وابسته دارد. با ورود هر متغیر جدید، کلیه متغیرهای موجود در معادله بررسی شده و اگر هر کدام از آنها سطح معناداری خود را از دست

¹ Molecular Descriptors

² Stepwise

بدهد، قبل از ورود متغیر جدید از مدل خارج می‌شود. به این ترتیب داده‌های pEC_{50} به عنوان متغیر وابسته و توصیف‌کننده‌ها به عنوان متغیر مستقل در نظر گرفته شده و تکنیک رگرسیون مرحله‌ای انجام شد. همانطور که می‌دانیم روش رگرسیون مرحله‌ای تعداد زیادی مدل ارائه می‌کند. که مدل اول شامل یک توصیف‌کننده، مدل دوم شامل دو توصیف‌کننده و ... می‌باشد. با افزایش تعداد توصیف‌کننده‌ها بالطبع مقدار R^2 افزایش و $RMSE^1$ (خطای جذر میانگین مربعات) کاهش می‌یابد. اما بدلیل پیچیدگی مدل، نمی‌توانیم تعداد زیادی توصیف‌کننده را جهت مدل‌سازی انتخاب کنیم. بدین منظور و جهت انتخاب تعداد توصیف‌کننده‌های مناسب، نمودار پارامترهای مختلف آماری از جمله R^2 ، ریشه میانگین مربعات خطا (RMSE) و ... برحسب تعداد توصیف‌کننده‌ها رسم گردید که در شکل ۱ نشان داده شده است. بر طبق این شکل، تعداد ۴ توصیف‌کننده به عنوان توصیف‌کننده‌هایی که بیشترین ارتباط را با فعالیت دارویی (pEC_{50}) دارند، انتخاب شدند. توصیف‌کننده‌های انتخاب شده به همراه مفهوم و نوع آنها در جدول ۱ ارائه شده است.



شکل ۱. نمودار پارامترهای آماری (R^2_{train} , R^2_{test} , $RMSE_{train}$, $RMSE_{test}$) برحسب تعداد توصیف‌کننده‌ها

جدول ۱. توصیف‌کننده‌های انتخاب شده توسط رگرسیون خطی چندگانه مرحله به مرحله

نوع توصیف‌کننده	مفهوم توصیف‌کننده	نشانه توصیف‌کننده
2D autocorrelations.	Geary autocorrelation - lag 8 / weighted by atomic van der Waals volumes.	GATS4v
Charge descriptors	Maximum positive charge.	qpos
RDF descriptors.	Radial Distribution Function – 8.0/ unweighted.	RDF080u
GETAWAY	Leverage-weighted autocorrelation of lag 4/ unweighted.	HATS4u

¹Root-mean-square-error

۲-۵. ارزیابی توصیف کننده ها

به منظور ارزیابی توصیف کننده های انتخاب شده مبنی بر مستقل بودن از همدیگر در جدول ۲ ماتریس همبستگی توصیف کننده - های انتخاب شده آورده شده است. همانطور که در این جدول مشاهده می شود ضریب همبستگی بین توصیف کننده های انتخاب شده همگی کمتر از ۰/۷۹۰ می باشد. لذا نتایج جدول نشان می دهد که بین توصیف کننده های انتخاب شده همبستگی چندانی وجود نداشته و توصیف کننده های تقریباً مستقل از هم هستند.

جدول ۲. ماتریس ضرایب همبستگی توصیفگرهای انتخاب شده

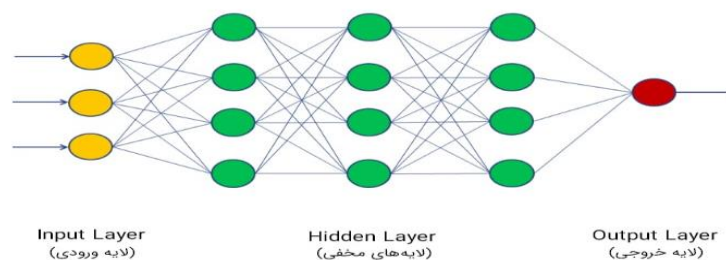
	GATS4v	qpos	RDF080u	HATS4u
GATS4v	1	0	0	0
qpos	-0.0064	1	0	0
RDF080u	-0.1763	-0.1325	1	0
HATS4u	0.0932	0.0780	-0.2586	1

۲-۶. شبکه های عصبی مصنوعی

شبکه های عصبی مصنوعی یکی از شاخه های اصلی در حوزه یادگیری ماشین و هوش مصنوعی هستند که از ساختار و عملکرد مغز انسان الهام گرفته اند. این شبکه ها با تقلید از نحوه ارتباط نورون های مغز، توانایی یادگیری از داده ها و حل مسائل پیچیده را دارند. شبکه های عصبی مصنوعی از لایه های متعددی تشکیل شده اند که شامل لایه ورودی، لایه های پنهان و لایه خروجی هستند. هر لایه شامل تعدادی نورون (نود) است که با دریافت ورودی ها، آن ها را پردازش کرده و خروجی تولید می کنند. این خروجی ها به لایه های بعدی منتقل می شوند تا در نهایت نتیجه نهایی در لایه خروجی ارائه شود. شبکه های عصبی با استفاده از توابع فعال سازی^۱ و به روزرسانی وزن ها، توانایی یادگیری الگوهای پیچیده از داده ها را پیدا می کنند. فرآیند آموزش این شبکه ها معمولاً از طریق الگوریتم هایی مانند گرادیان کاهشی^۲ و انتشار معکوس^۳ انجام می شود. در این فرآیند، شبکه با استفاده از داده های آموزشی، وزن های خود را به روز می کند تا خطای بین پیش بینی های شبکه و مقادیر واقعی به حداقل برسد. شکل ۲ نمایی از یک شبکه عصبی مصنوعی را نشان می دهد. شبکه های عصبی مصنوعی در انواع مختلفی مانند شبکه های عصبی پیش خور، شبکه های عصبی بازگشتی و شبکه های عصبی کانولوشن وجود دارند که هر یک برای کاربردهای خاصی مانند تشخیص تصویر، پردازش زبان طبیعی، پیش بینی سری های زمانی و غیره استفاده می شوند. با وجود قدرت و انعطاف پذیری بالای این شبکه ها، چالش هایی مانند نیاز به حجم

¹ Activation Functions² Gradient Descent³ Backpropagation

زیاد داده‌های آموزشی، زمان محاسباتی طولانی و خطر بیش‌برازش^۱ نیز وجود دارد. با این حال، شبکه‌های عصبی مصنوعی به عنوان یکی از قدرتمندترین ابزارهای هوش مصنوعی، نقش مهمی در پیشرفت فناوری‌های مدرن ایفا می‌کنند.



شکل ۲. نمایی از یک شبکه عصبی مصنوعی

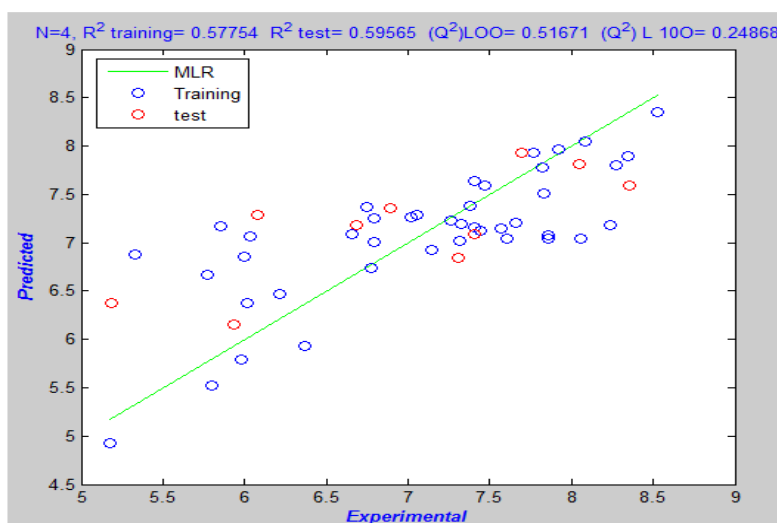
۳. نتایج و بحث

۳-۱. مدل سازی به روش رگرسیون خطی چندگانه (MLR)

پس از انتخاب مناسب ترین توصیف کننده‌ها توسط روش مرحله‌ای، مرحله بعدی، ایجاد مدل میان توصیف کننده‌های انتخاب شده و pEC_{50} می‌باشد. بین توصیف کننده‌ها و فعالیت دارویی ترکیبات برای سری آموزش با استفاده از روش MLR رابطه زیر به عنوان مدل خطی بدست آمد:

$$pEC_{50} = -2.8443 + 3.4377(GATS4v) - 8.2561(qpos) + 0.1418(RDF080u) + 7.6867(HATS4u)$$

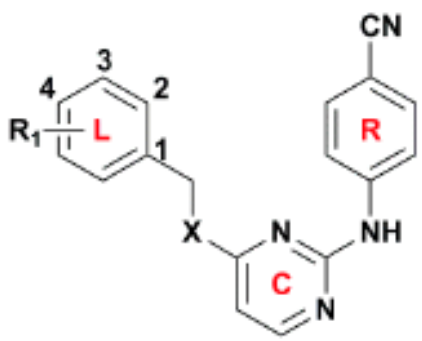
سپس از معادله بدست آمده برای پیش‌بینی فعالیت دارویی ترکیبات سری تست استفاده گردید. مقادیر تجربی و پیش‌بینی شده pEC_{50} برای کلیه ترکیبات مجموعه آموزش و تست در جدول (۳) آورده شده است. شکل (۳) نمودار مقادیر پیش‌بینی شده بر حسب مقادیر تجربی را نشان می‌دهد. در این شکل نزدیکی نتایج به خط راست قدرت پیش‌بینی مدل را نشان می‌دهد.



شکل ۳. نمودار مقادیر پیش‌بینی شده pEC_{50} بر حسب مقادیر تجربی برای سری آموزش و تست به روش SW-MLR

¹ Overfitting

جدول ۳. مقادیر تجربی و محاسبه شده pEC_{50} ترکیبات مختلف برای مجموعه‌های آموزشی و پیش‌بینی در مدل‌های SW-MLR، SW-ANN

 I : DAPYs (X = O, NH)				 II : CAPYs (X = O, S, NH)				
No.	Type	X	R ₁	R ₂	C _y	pEC_{50}	SW-MLR	SW-ANN
1	I	O	2-Cl	-	-	7.469	7.589	7.889
2*		O	3-Cl	-	-	7.310	6.842	6.730
3*		O	4-Cl	-	-	6.076	7.286	7.235
4		O	2-Br	-	-	7.444	7.128	7.180
5		O	3-Br	-	-	7.319	7.014	6.614
6		O	4-Br	-	-	5.328	6.873	6.548
7		O	2-F	-	-	8.237	7.184	7.236
8		O	3-F	-	-	7.854	7.072	7.087
9		O	4-F	-	-	6.036	7.065	7.069
10		O	2-CH ₃	-	-	6.796	7.007	6.961
11		O	3-CH ₃	-	-	8.056	7.038	7.052
12		O	4-CH ₃	-	-	6.796	7.247	6.961
13		O	2-OCH ₃	-	-	7.569	7.145	7.245
14		O	3-OCH ₃	-	-	7.602	7.039	7.035

15		O	4-OCH ₃	-	-	6.000	6.857	6.460
16		O	2-CF ₃	-	-	5.770	6.666	5.978
17		O	3-CF ₃	-	-	6.215	6.471	6.110
18*		O	3,5-diMeO	-	-	8.051	7.812	8.374
19*		O	5-F-2-Br	-	-	7.409	7.084	7.133
20		O	2-F-4-Br	-	-	6.658	7.084	7.085
21		O	2-CN	-	-	7.854	7.039	7.087
22		O	3-CN	-	-	6.745	7.370	6.776
23		O	4-CN	-	-	5.854	7.170	7.186
24		O	H	-	-	7.658	7.210	7.267
25		O	1-Naph	-	-	7.824	7.780	8.033
26		O	2-Naph	-	-	6.770	6.743	6.901
27		NH	H	-	-	8.276	7.807	8.153
28		NH	2-Cl	-	-	7.827	7.509	7.886
29		NH	3-Cl	-	-	7.924	7.965	7.907
30		NH	4-Cl	-	-	7.322	7.193	7.092
31		NH	2-Br	-	-	7.376	7.375	7.780
32*		NH	3-Br	-	-	7.693	7.931	8.060
33		NH	4-Br	-	-	7.405	7.633	7.682
34*		NH	3-F	-	-	8.357	7.586	5.188
35		NH	4-F	-	-	7.143	6.920	7.089
36		NH	4-CH ₃	-	-	8.086	8.052	7.151
37		NH	2-OCH ₃	-	-	7.058	7.288	7.263
38		NH	3-OCH ₃	-	-	8.347	7.890	8.253
39		NH	4-OCH ₃	-	-	7.406	7.164	7.281
40		NH	2-CF ₃	-	-	5.976	5.796	5.929
41*		NH	3-CF ₃	-	-	6.687	7.182	6.928

42*		NH	4-CF ₃	-	-	6.895	7.360	7.479
43		NH	3,4-diMe	-	-	8.523	8.357	8.580
44		NH	2,4-diF	-	-	7.772	7.933	8.114
45*	II	O	H	H	Cyclohexyl	5.936	6.158	5.349
46		O	H	H	Cyclopentyl	5.173	4.932	5.142
47*		O	H	CH ₃	Cyclohexyl	5.188	6.377	5.425
48		NH	H	H	Cyclohexyl	6.018	6.372	5.461
49		NH	H	H	Cyclopentyl	5.801	5.521	5.188
50		S	H	H	Cyclohexyl	7.260	7.223	7.422
51		S	H	CH ₃	Cyclohexyl	6.367	5.927	6.413
52		S	CH ₃	H	Cyclohexyl	7.018	7.260	7.190

*Used as test set

۲-۳. مدل سازی و پیش بینی توسط شبکه های عصبی مصنوعی

در این روش توصیف کننده های انتخاب شده وارد شبکه عصبی مصنوعی می شوند. پردازش داده ها در محیط ویندوز ۱۰ و با استفاده از نرم افزار MATLAB انجام شد. یک شبکه سه لایه با تابع انتقال سیگموئیدی برای نرون ها طراحی شده است. مقادیر اولیه وزن ها بطور تصادفی از بازه [0, 1] بوده و قبل از عمل آموزش مقادیر ورودی و خروجی در فاصله [0.1, 0.9] نرمال شده است. بهینه سازی و بهنگام کردن وزن ها و بایاس ها بوسیله الگوریتم BP^۱ انجام شده است. مجموعه داده ها به سه گروه تقسیم شده است: مجموعه آموزش، مجموعه ارزیابی و مجموعه تست. مجموعه آموزش (۵۰٪ داده ها) جهت آموزش دادن شبکه عصبی مصنوعی، مجموعه ارزیابی (۳۰٪ داده ها) برای ارزیابی مدل در طی آموزش دادن شبکه و ایجاد مدل مناسب و مجموعه پیش بینی (۲۰٪ داده ها) برای تست مدل ایجاد شده به کار رفت.

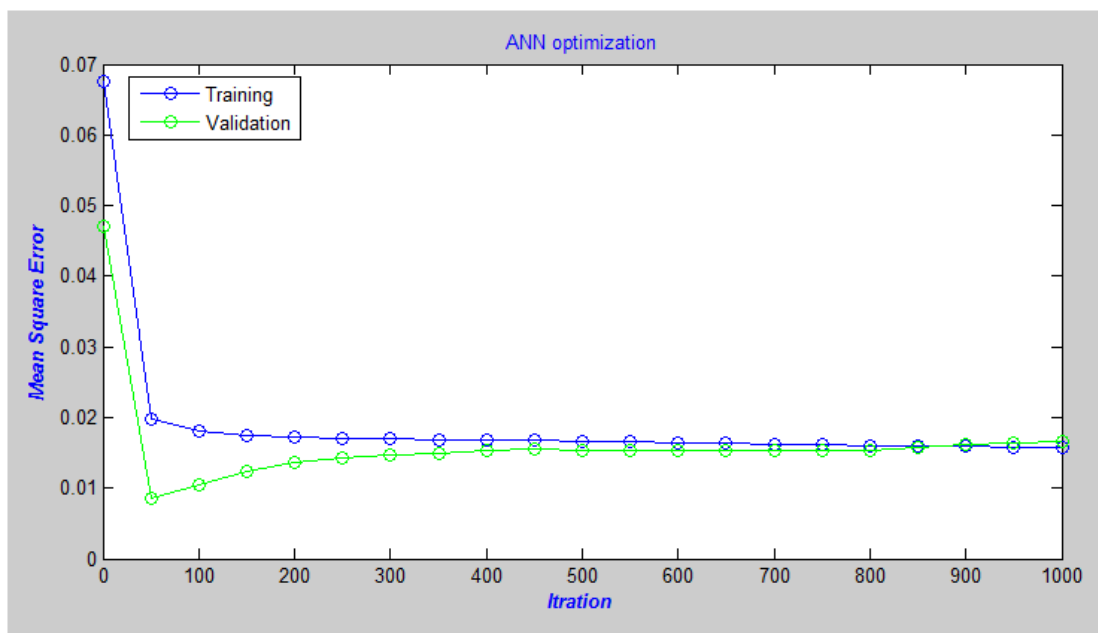
تعداد نرون ها در لایه ورودی با تعداد توصیف کننده های وارد شده به شبکه های عصبی مصنوعی برابر است. به ازای هر تعداد توصیف کننده وارد شده به شبکه عصبی، تعداد نرون ها در لایه مخفی بهینه می شود. بدین ترتیب که به ازای هر مدل ANN، تعداد نرون ها در لایه مخفی از ۱ تا ۱۰ تغییر داده شده و مقادیر RMSE برای مجموعه های آموزشی و پیش بینی محاسبه گردید. از رسم مقادیر RMSE بر حسب تعداد نرون ها در لایه مخفی، تعداد نرون های لایه مخفی بهینه شد. سپس بقیه پارامترها از جمله وزن ها و بایاس ها، سرعت یادگیری و مومتوم نیز بهینه گردید. جدول ۴ مشخصات شبکه عصبی مصنوعی بهینه شده را نشان می دهد.

¹Back-propagation

جدول ۴. ساختار و مشخصات ANN تولید شده

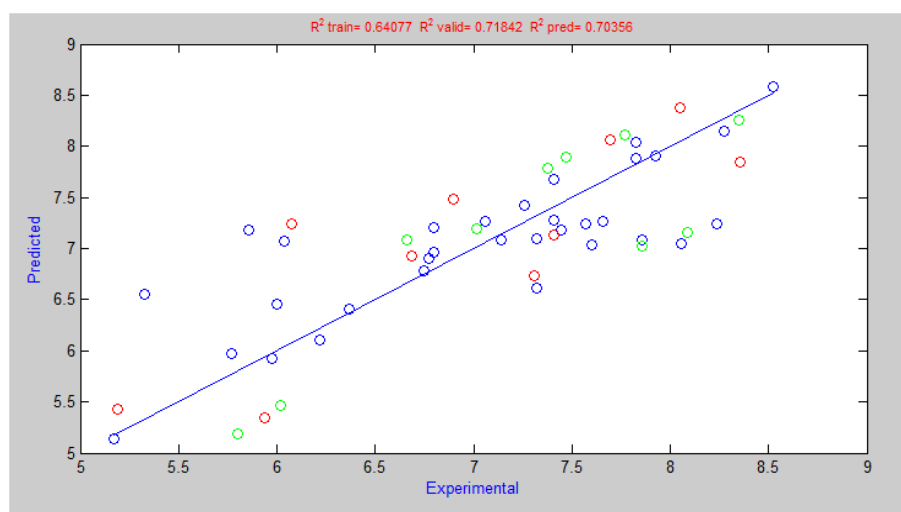
تعداد گره ها (نرون ها) در لایه ورودی	۴
تعداد گره ها (نرون ها) در لایه مخفی	۵
تعداد گره ها (نرون ها) در لایه خروجی	۱
سرعت آموزش	۰/۳
مومنتوم	۰/۳
تعداد تکرارها یا چرخه های آموزشی	۱۰۰۰
تابع انتقال	تابع سیگموئیدی

برای جلوگیری از Overfitting در طول آموزش مقادیر RMSE بعد از هر ۵۰ بار چرخه آموزشی، محاسبه و ثبت گردید. شکل ۴ نمودار میزان خطا بر حسب تعداد دورها برای این داده ها را نشان می دهد. همانطور که ملاحظه می شود میزان خطا برای سری آموزش همواره در حال کاهش است اما برای سری ارزیابی در ۵۰ چرخه آموزش، کمترین خطا مشاهده میشود و بعد از آن افزایش می یابد. بنابراین این مقدار به عنوان مقدار بهینه تعداد چرخه های آموزش انتخاب شد.



شکل ۴. مقادیر RMSE برای مجموعه های آموزشی و ارزیابی بر حسب تعداد چرخه های آموزش

با استفاده از مدل ANN بهینه شده مقادیر فعالیت های بازدارندگی (PEC_{50}) ترکیبات مورد نظر در مجموعه آموزشی، ارزیابی و پیش بینی (تست) مورد محاسبه قرار گرفت و در جدول ۳ نشان داده شده است. در شکل ۵ نیز مقادیر محاسبه شده PEC_{50} ترکیبات مورد نظر در مجموعه های مختلف بر حسب مقادیر تجربی رسم شده اند. در این شکل، میزان نزدیکی داده ها به خط راست قدرت پیش بینی مدل را نشان می دهد.



شکل ۵. نمودار مقادیر pEC_{50} محاسبه شده براساس مدل ANN در مجموعه های آموزش (رنگ آبی)، ارزیابی (رنگ سبز) و تست (رنگ قرمز) برحسب مقادیر تجربی

۳-۳. ارزیابی مدل با استفاده از پارامترهای آماری

اعتبار و اهمیت مدل های ساخته شده وقتی مشخص می گردد که فعالیت مولکول هایی که در سری تست هستند را به خوبی و بطور رضایت بخش و قابل قبول پیش بینی کند. در این کار چندین روش به منظور ارزیابی و مقایسه توانایی مدل های ارائه شده در پیش بینی مقادیر pEC_{50} ذکر شده است. مطابق جدول (۵) پنج پارامتر آماری، جهت ارزیابی توانایی پیش بینی مدل های ساخته شده به روش های ANN و MLR به کار گرفته شد. این نتایج نشان دهنده آن است که شبکه عصبی مصنوعی نسبت به روش MLR توانمندی بیشتری جهت پیش بینی فعالیت دارویی ترکیبات را دارد.

جدول ۵. پارامترهای آماری برای مدل های انتخاب شده

		SW-MLR	SW-ANN
R^2	سری آموزش	۰/۵۷۷	۰/۶۴۰
	سری تست	۰/۵۹۵	۰/۷۰۳
RMSE	سری آموزش	۰/۵۶۲	۰/۵۲۹
	سری تست	۰/۶۶۵	۰/۵۵۲
REP	سری آموزش	۷/۹۳۱	۷/۴۵۹
	سری تست	۹/۵۵۴	۷/۹۳۶
SEP	سری آموزش	۰/۵۶۲	۰/۵۲۹
	سری تست	۰/۶۶۵	۰/۵۵۲
AARD	سری آموزش	۶/۳۶۵	۵/۷۳۰
	سری تست	۸/۶۸۰	۷/۲۱۱

۱-۳-۳. ارزیابی مدل‌ها توسط روش رد مرحله‌ای تک تک و گروهی

به منظور بررسی بیشتر قدرت پیش‌بینی مدل‌های خطی و غیر خطی، تکنیک رد مرحله‌ای تک تک و گروهی مورد استفاده قرار گرفت. در روش رد مرحله‌ای تک تک، هر بار یکی از ترکیبات به طور تصادفی از سری داده‌ها حذف شدند و در روش رد مرحله‌ای گروهی، هر بار یک گروه از ترکیبات (۵ ترکیب) به طور تصادفی از سری داده‌ها حذف شدند. سپس با استفاده از مدل ساخته شده توسط بقیه ترکیبات، فعالیت دارویی ترکیب یا ترکیبات حذف شده، پیش‌بینی شدند. این فرایند برای تمام اعضای سری داده‌ها تکرار شد. نتایج حاصل از رد مرحله‌ای و گروهی در جدول (۶) ارائه شده است.

جدول ۶. پارامترهای آماری برای مدل‌های انتخاب شده

		SW-MLR	SW-ANN
Q^2_{LOO}	کل داده‌ها	۰/۵۱۶	۰/۵۴۷
Q^2_{LGO}	کل داده‌ها	۰/۴۷۰	۰/۵۲۹

۲-۳-۳. ارزیابی مدل‌های ارائه شده با استفاده از آزمون Y-تصادفی

این تکنیک ارزیابی مدل، با هدف بررسی هر گونه ارتباط تصادفی بین داده‌ها انجام شد. در این آزمون، متغیر وابسته بطور تصادفی بهم ریخته شد. مدل QSAR جدید با استفاده از ماتریکس متغیرهای مستقل اصلی و مقادیر تصادفی از متغیر وابسته توسعه یافت. اگر در مدل اصلی هیچ گونه ارتباط تصادفی وجود نداشته باشد، تفاوت قابل توجهی بین مقدار ضریب تعیین مدل اصلی و مدل QSAR که با پاسخ تصادفی توسعه یافته، وجود خواهد داشت. نتایج حاصل از چندین بار اجرای آزمون Y-تصادفی در جدول (۷) نشان داده شده است. مقادیر کوچک ضریب تعیین (R^2) بیانگر عدم ارتباط شانس در مدل توسعه یافته توسط رگرسیون خطی چندگانه می‌باشد.

جدول ۷. نتایج حاصل از چندین بار اجرای آزمون Y-تصادفی

تکرار	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
R^2_{test}	۰/۰۲۴	۰/۰۵۴	۰/۰۴۸	۰/۰۹۶	۰/۲۰۱	۰/۰۳۴	۰/۱۴۹	۰/۲۴۳	۰/۰۸۵	۰/۰۱۲۴
	۰/۰۸۵	۰/۱۴۲	۰/۰۳۴	۰/۰۸۵	۰/۰۵۲	۰/۱۲۴	۰/۰۵۶	۰/۰۶۳	۰/۱۰۸	۰/۰۸۷
	ANN									
	MLR									

۳-۳-۳. ارزیابی توصیف کننده های انتخاب شده

اولین توصیف کننده انتخاب شده در مدل ها، GATS4v بوده که از نوع توصیف کننده های خودهمبستگی دوبعدی^۱ می باشد. این توصیف کننده ها به بررسی چگونگی توزیع ویژگی های مولکولی (مانند بار اتمی، قطبش پذیری، یا خواص دیگر) در فواصل مشخصی درون مولکول می پردازد. این توصیف کننده ها با محاسبه همبستگی بین ویژگی های اتم ها یا گروه های اتمی در فواصل مختلف، یک بردار عددی تولید می کند که نمایانگر ساختار مولکول است. این بردار می تواند به عنوان ورودی برای مدل های یادگیری ماشین استفاده شود تا خواص مولکولی پیش بینی شود. توصیف کننده های خودهمبستگی دوبعدی ابزاری قدرتمند در شیمی محاسباتی هستند که با ارائه یک نمای کمی از ساختار مولکولی، به پژوهشگران در درک و پیش بینی خواص مولکول ها کمک می کنند. تاثیر این توصیف کننده در مدل مثبت بوده یعنی با افزایش مقدار این توصیف کننده، pEC₅₀ نیز افزایش می یابد.

دومین توصیف کننده وارد شده در مدل، qpos می باشد که از نوع توصیف کننده های مبتنی بر بار^۲ می باشد. این توصیف کننده ها، یکی از انواع توصیف کننده های شیمیایی هستند که برای توصیف توزیع بار الکتریکی درون یک مولکول استفاده می شوند که اطلاعاتی درباره نحوه توزیع بارهای مثبت و منفی در اتم ها یا بخش های مختلف مولکول ارائه می دهند و در پیش بینی خواص شیمیایی و فیزیکی مولکول ها بسیار مفید هستند. بارهای اتمی و توزیع بار در مولکول ها نقش مهمی در تعیین رفتار مولکول ها در واکنش های شیمیایی، برهمکنش های بین مولکولی و خواص فیزیکی مانند قطبیت و حلالیت دارند. تاثیر این توصیف کننده در مدل منفی بوده یعنی با افزایش مقدار این توصیف کننده، pEC₅₀ کاهش می یابد.

RDF080u سومین توصیف کننده وارد شده در مدل است که از نوع توصیف کننده های تابع توزیع شعاعی^۳ می باشد. این توصیف کننده ها یکی از ابزارهای مهم در شیمی محاسباتی و طراحی دارو هستند که برای تحلیل ساختار اتمی و مولکولی سیستم های مختلف استفاده می شوند و اطلاعاتی درباره چگونگی توزیع اتم ها یا مولکول ها در فاصله های مشخص از یک اتم مرجع ارائه می دهند. توصیف کننده های RDF ابزاری قدرتمند برای تحلیل ساختار و برهمکنش های اتمی و مولکولی هستند و در بسیاری از زمینه های علمی و فناوری کاربرد دارند. اثر این توصیف کننده در مدل مثبت بوده که نشان می دهد با افزایش مقدار این توصیف کننده، pEC₅₀ نیز افزایش می یابد.

چهارمین توصیف کننده وارد شده در مدل، HATS4u می باشد که از نوع توصیف کننده های GETAWAY^۴ می باشد. این توصیف کننده ها بر پایه ی اطلاعات هندسی، توپولوژیکی و وزنی اتم ها در یک مولکول استوار هستند و برای پیش بینی فعالیت بیولوژیکی، خواص فیزیکوشیمیایی و سایر ویژگی های مولکولی به کار می روند. اثر این توصیف کننده در مدل مثبت بوده که نشان می دهد با افزایش مقدار این توصیف کننده، pEC₅₀ نیز افزایش می یابد.

¹ 2D autocorrelations

² Charge Descriptors

³ Radial Distribution Function: RDF

⁴ Geometry, Topology, and Atoms-Weighted Assembly

۴. نتیجه گیری

در این تحقیق از دو روش مختلف برای پیش‌بینی فعالیت ترکیبات دارویی (EC_{50}) مشتقات دی آریل پیریمیدین‌ها استفاده شد. از آنجایی که اندازه‌گیری فعالیت دارویی بیشتر ترکیبات به صورت تجربی با صرف هزینه، زمان و پیچیدگی زیاد همراه است، دستیابی به مقادیر EC_{50} ، با روش‌های تجربی مقرون به صرفه نیست. بنابراین، پیش‌بینی آن با استفاده از روش‌های محاسباتی از اهمیت بالایی برخوردار است. برای انتخاب توصیفگرهای مناسب از روش رگرسیون مرحله‌ای استفاده شد. سپس این توصیفگرها برای مدل‌سازی خطی و غیرخطی مورد استفاده قرار گرفتند. نتایج نشان داد که از بین روش‌های استفاده شده، روش ANN روش مناسب‌تری برای پیش‌بینی فعالیت این ترکیبات دارویی است. نتایج تحقیق جهت طراحی مشتقات دارویی جدید و موثر غیر نوکلئوزیدی خانواده دارویی دی آریل پیریمیدین‌ها در مهار آنزیم کلیدی Reverse transcriptase ویروس HIV پیشنهاد می‌شود.

۵. مراجع

- [1] Verma, J., Khedkar, V. M., & Coutinho, E. C. (2010). 3D-QSAR in drug design-a review. *Current topics in medicinal chemistry*, 10(1), 95-115.
- [2] Mervin, L., Voronov, A., Kabeshov, M., & Engkvist, O. (2024). QSARtuna: an automated qsar modeling platform for molecular property prediction in drug design. *Journal of Chemical Information and Modeling*, 64(14), 5365-5374.
- [3] Bassani, D., & Moro, S. (2023). Past, present, and future perspectives on computer-aided drug design methodologies. *Molecules*, 28(9), 3906.
- [4] Daoui, O., Nour, H., Abchir, O., Elkhatabi, S., Bakhouch, M., & Chtita, S. (2023). A computer-aided drug design approach to explore novel type II inhibitors of c-Met receptor tyrosine kinase for cancer therapy: QSAR, molecular docking, ADMET and molecular dynamics simulations. *Journal of Biomolecular Structure and Dynamics*, 41(16), 7768-7785.
- [5] Yu, W., & MacKerell Jr, A. D. (2016). Computer-aided drug design methods. In *Antibiotics: methods and protocols* (pp. 85-106). New York, NY: Springer New York
- [6] Haghshenas, H., Kaviani, B., Firouzeh, M., & Tavakol, H. (2021). Developing a variation of 3D-QSAR/MD method in drug design. *Journal of Computational Chemistry*, 42(13), 917-929.
- [7] Kubinyi, H. (1997). QSAR and 3D QSAR in drug design Part 1: methodology. *Drug discovery today*, 2(11), 457-467.
- [8] Tandon, H., Chakraborty, T., & Suhag, V. (2019). A concise review on the significance of QSAR in drug design. *Chemical and Biomolecular Engineering*, 4(4), 45-51.
- [9] Daoui, O., Elkhatabi, S., & Chtita, S. (2023). Design and prediction of ADME/Tox properties of novel magnolol derivatives as anticancer agents for NSCLC using 3D-QSAR, molecular docking, MOLCAD and MM-GBSA studies. *Letters in Drug Design & Discovery*, 20(5), 545-569.
- [10] Achary, P. G. (2020). Applications of quantitative structure-activity relationships (QSAR) based virtual screening in drug design: A review. *Mini Reviews in Medicinal Chemistry*, 20(14), 1375-1388.
- [11] Tandon, H., Chakraborty, T., & Suhag, V. (2019). A concise review on the significance of QSAR in drug design. *Chemical and Biomolecular Engineering*, 4(4), 45-51.
- [12] Zivkovic, M., Zlatanovic, M., Zlatanovic, N., Golubović, M., & Veselinović, A. M. (2020). The application of the combination of Monte Carlo optimization method based QSAR modeling and molecular docking in drug design and development. *Mini Reviews in Medicinal Chemistry*, 20(14), 1389-1402.

- [13] Jian-Bo, T., ZHANG, X., Shuai, B., & Ding, L. U. O. (2022). Drug design, molecular docking, and ADMET PREDICTION of CCR5 inhibitors based on QSAR Study^①. *Chinese Journal of Structural Chemistry*, 41(2), 2202001-2202013.
- [14] Bouakkadia, A., Kertiou, N., Amiri, R., Driouche, Y., & Messadi, D. (2021). Use of GA-ANN and GA-SVM for a QSPR study on the aqueous solubility of pesticides. *Journal of the Serbian Chemical Society*, 86(7-8), 673-684.
- [15] Brown, N., Ertl, P., Lewis, R., Luksch, T., Reker, D., & Schneider, N. (2020). Artificial intelligence in chemistry and drug design. *Journal of Computer-Aided Molecular Design*, 34, 709-715.
- [16] Shi, Y. (2021). Support vector regression-based QSAR models for prediction of antioxidant activity of phenolic compounds. *Scientific reports*, 11(1), 8806.
- [17] 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*, 35(5), e3338.
- [18] Vahedi, Nafiseh, Majid Mohammadhosseini, and Mehdi Nekoei. "QSAR Study of PARP Inhibitors by GA-MLR, GA-SVM and GA-ANN Approaches." *Current Analytical Chemistry* 16, no. 8 (2020): 1088-1105.
- [19] Ghanei-Nasab, S., Hadizadeh, F., Foroumadi, A., & Marjani, A. (2021). A QSAR study for the prediction of inhibitory activity of coumarin derivatives for the treatment of Alzheimer's disease. *Arabian Journal for Science and Engineering*, 46(6), 5523-5531.
- [20] Ai, H., Wu, X., Zhang, L., Qi, M., Zhao, Y., Zhao, Q., ... & Liu, H. (2019). QSAR modelling study of the bioconcentration factor and toxicity of organic compounds to aquatic organisms using machine learning and ensemble methods. *Ecotoxicology and Environmental Safety*, 179, 71-78.
- [21] Fereidoon nezahad, M., Tabaei, S. M. H., Sakhteman, A., Seradj, H., Faghih, Z., Faghih, Z., ... & Rezaei, Z. (2020). Design, synthesis, molecular docking, biological evaluations and QSAR studies of novel dichloroacetate analogues as anticancer agent. *Journal of Molecular Structure*, 1221, 128689.

Using linear and nonlinear quantitative structure-activity relationship (QSAR) models to predict the pharmacological activity of diarylpyrimidine derivatives

Nadia Nekoei, Mehdi Nekoei^{1*}, Seyed Mohammadreza Motallebi Tabaei², Nadia Nekouei³

¹Department of Chemistry, Sha. C., Islamic Azad University, Shahrood, Iran

²Department of Biochemistry, Sha. C., Islamic Azad University, Shahrood, Iran

³Department of Laboratory Sciences, Shahrood University of Medical Sciences, Shahrood, Iran

Submitted: 15 December 2024, Revised: 01 March 2025, Accepted: 10 March 2025

Abstract

A QSAR (Quantitative Structure-Activity Relationship) study was conducted to predict the pharmacological activity (EC₅₀ values) of some diaryl pyrimidine derivatives as HIV inhibitors using multiple linear regression (MLR) and artificial neural networks (ANN) as linear and nonlinear methods. Initially, the structures of the compounds were drawn and optimized using Hypercom software. Then, the drawn compounds were imported into Dragon software and 1481 molecular descriptors were calculated for each molecule. Stepwise regression was used to select the most appropriate descriptors. After selecting the optimal descriptors, two methods, MLR and ANN, were used to model and predict the pharmacological activity of the compounds. The performance of each model was evaluated using several statistical parameters. The results obtained indicated the superiority of the ANN method over the MLR in predicting the pharmacological activity of the compounds.

Keywords: Quantitative structure-activity relationship, diarylpyrimidines, multiple linear regression, artificial neural networks

*Corresponding author : Mehdi Nekoei

Address: Department of Chemistry, Shahrood Branch, Islamic Azad University, Shahrood, Iran

Tel: 02332394289

E-mail: m_nekoei@iau.ac.ir