

پیش بینی فعالیت بیولوژیکی برخی آفت کش ها با استفاده از روش های QSAR و مطالعات داکینگ مولکولی

فاطمه ایزدی^۱، الهام باهر^{۲*}، سجاد قرقانی^۳

۱-دانشجو کارشناسی ارشد، گروه شیمی، دانشکده علوم، دانشگاه گلستان، گرگان

۲-استادیار، گروه شیمی، دانشکده علوم، دانشگاه گلستان، گرگان

۳-دانشیار، گروه بیوانفورماتیک، آزمایشگاه کموانفورماتیک، مرکز تحقیقات بیوشیمی و بیوفیزیک (IBB)، دانشگاه تهران، تهران، ایران

*صندوق پستی: ۴۹۱۳۸۱۵۷۷۵۹، گلستان، ایران

e.baher@gu.ac.ir

پذیرش: ۱۴۰۵/۰۲/۲۴

دریافت: ۱۴۰۵/۰۲/۱۶

چکیده

در پژوهش حاضر، دوز سمیت (LD50) برخی از آفت کش ها با استفاده از مطالعات کمی ساختار- فعالیت (QSAR) مدل سازی و پیش بینی شد. به این منظور پس از رسم ساختار مولکولی، توصیف کننده های مولکولی محاسبه شد و میزان سمیت آفت کش ها با روش خطی (MLR) و غیرخطی ماشین بردار پشتیبان (SVM) با استفاده از توصیف کننده های انتخابی به عنوان ورودی، مورد پیش بینی قرار گرفت. از الگوریتم ژنتیک (GA) برای انتخاب بهترین توصیف کننده مرتبط استفاده شد. در ادامه برای بررسی برهمکنش های موثر لیگاند-پروتئین و محاسبه انرژی آزاد اتصال بین آفت کش ها و پروتئین آلبومین، داکینگ مولکولی انجام شد. در بخش اول با مقایسه نتایج آماری به دست آمده، مدل MLR-SVM با دارا بودن مقادیر R و SE به ترتیب برای مجموعه آموزشی برابر با ۰/۹۹ و ۰/۰۱۴ و برای مجموعه آزمون برابر با ۰/۷۱۸ و ۰/۰۸۲ پیش بینی بسیار بهتری نسبت به سایر مدل ها داشتند. این نتایج، نزدیکی مقادیر محاسباتی به مقادیر تجربی LD50 را نشان داد. همچنین، برای سنجش بیشتر اعتبار مدل MLR-SVM، از آزمون در هم آمیختگی- Y و آزمون ارزیابی تقاطعی استفاده شد که آماره های Q و SPRESS به ترتیب برابر ۰/۴۹۸ و ۰/۰۴۷ به دست آمدند. در بخش دوم پژوهش، آفت کش هایی با بیشترین (ترکیب ۵۷ با LD50 برابر با ۲/۶۸) و کمترین (ترکیبات شماره ۳۵ و ۶۳ با LD50 برابر با ۲/۲۱) مقادیر سمیت، جهت بررسی نحوه برهمکنش با پروتئین آلبومین، توسط مطالعات داکینگ مولکولی با نرم افزار autodock-vina مورد ارزیابی قرار گرفتند. برسی نتایج نشان داد که ترکیباتی که تمایل بیشتری به تشکیل پیوندهای آگریز دارند، فعال تر هستند و سمیت بیشتری از خود نشان می دهند. همچنین، حضور برخی عناصر، مانند فلوئور در ساختمان مولکولی آنها باعث افزایش سمیت بیشتر آنها می شود.

کلمات کلیدی: مطالعات کمی ساختار- فعالیت، داکینگ مولکولی، رگرسیون خطی چندگانه (MLR)، ژنتیک الگوریتم (GA)، ماشین بردار پشتیبان (SVM)، آفت کش ها، LD50.

۱-مقدمه

در کشاورزی، استفاده از آفت‌کش‌ها با هدف حفاظت از گیاهان و محصولات زراعی و افزایش ماندگاری دانه‌های ذخیره شده، امری رایج و ضروری است. سموم دفع آفات، مولکول‌های شیمیایی و سمی هستند که برای کنترل آفات، علف‌های هرز و عوامل بیماری‌زای گیاهی طراحی شده‌اند. با این حال، این ترکیبات ممکن است خطراتی را برای گونه‌های غیر هدف، از جمله انسان و اکوسیستم داشته باشند [۱].

آفت‌کش‌ها طیف گسترده‌ای از گونه‌های شیمیایی، مانند هیدروکربن‌های کلر، ترکیبات فسفری آلی، کربامات‌ها، ترکیبات سولفید آلی، مولکول‌های هتروسیکلیکی نیترو آروماتیک هالوژن و تعدادی از مواد معدنی هستند. در این بین، کربامات‌ها و ارگانوفسفرها به دلیل فعالیت حشره‌کشی بالا و ماندگاری نسبتاً کم در محیط در کشاورزی کاربرد گسترده‌ای دارند [۲]. سارین^۱ و سومان^۲ نیز در این دسته از ترکیبات قرار دارند. تماس طولانی مدت با این سموم می‌تواند بر سیستم عصبی، غدد درون‌ریز، سیستم ایمنی، تولیدمثل، کلیه، قلب و عروق و دستگاه تنفسی جانداران تأثیر منفی بگذارد [۳].

یکی از روش‌های رایج و پرکاربرد برای اندازه‌گیری و شناسایی این ترکیبات، روش‌های الکتروشیمیایی مبتنی حسگرهای زیستی است. این روش‌ها در کنار مزایای خاص خود، معایبی مانند ماندگاری کم حسگرها و هزینه‌های اقتصادی زیاد را به دنبال دارد [۴].

برای توسعه روش‌هایی سریع‌تر، کم هزینه‌تر و کارآمدتر برای پیش‌بینی سمیت، محققان استفاده از روش رابطه کمی ساختار-فعالیت (QSAR) را پیشنهاد می‌دهند. در این روش، با استفاده از قوانین و روابط ریاضی و آمار، ارتباط میان ساختار شیمیایی یک ترکیب و فعالیت زیستی آن به صورت کمی و کیفی بررسی می‌شود. امروزه، سازمان‌های مختلف نظارت بر غذا، دارو و محیط‌زیست در سراسر جهان از مدل‌های QSAR برای ارزیابی و کسب اطلاعات در مورد اثرات مواد شیمیایی بر سلامت انسان و محیط زیست استفاده می‌کنند [۵-۱۰].

مطالعات متعددی در زمینه توسعه و کاربرد مدل‌های QSAR انجام شده است. برای نمونه، در یکی از این پژوهش‌ها از روش 3D-SAR برای بررسی توانایی یون‌های فلزی موجود در آفت‌کش‌ها در ایجاد اختلال غدد درون‌ریز و ارائه مدل‌هایی برای آنالیز فعالیت‌های زیستی در سم‌شناسی استفاده شده است [۵].

در پژوهش دیگری، توسعه کمی مدل‌های ارتباط ساختاری سرطان و یافتن ویژگی‌های متمایزکننده برای سرطان‌زایی در جوندگان، با استفاده از دستورالعمل Organisation for Economic Co-operation and Development (OECD) گزارش شده است [۶].

همچنین، گوها و جرس در مطالعه‌ی دیگری به توسعه مدل‌های رابطه کمی ساختار-فعالیت (QSAR) برای پیش‌بینی فعالیت بیولوژیکی ۱۷۹ آنالوگ آرتمیزینین پرداختند. آنها دو مدل خطی (رگرسیون خطی چندگانه) و غیرخطی (شبکه عصبی محاسباتی) برای پیوند ساختارها به فعالیت بیولوژیکی این ترکیبات ارائه دادند [۱۰].

داکینگ مولکولی، همگام با تحولات عظیم در فیزیک، شیمی، فناوری اطلاعات، بیوشیمی و علوم رایانه، از آغاز فعالیت آن در دهه ۱۹۶۰، به یک ابزار قدرتمند و یک فناوری اساسی تبدیل شده است. برای انجام داکینگ، نخست باید ساختار سه بعدی پروتئین هدف در دسترس باشد. این ساختار پروتئینی و یک لیگاند که معمولاً از پایگاه داده معتبر استخراج می‌شود، به عنوان ورودی به برنامه داکینگ است. موفقیت برنامه داکینگ به دو مولفه‌ی، الگوریتم جستجو و تابع امتیازدهی بستگی دارد.

فضای جستجو از نظر تئوری شامل کلیه جهت‌ها و ترکیبات احتمالی پروتئین مرتبط با لیگاند است. با این حال، در عمل با ابزارهای محاسباتی فعلی، جستجوی جامع این فضا غیرممکن است. زیرا فضای جستجو شامل برشمردن تمام انحراف‌های مولکولی است (مولکول‌ها پویا هستند و در مجموعه‌ای از حالت‌های ساختاری وجود دارند) و تمام جهت‌گیری‌های چرخشی و انتقالی لیگاند نسبت به پروتئین در سطح مشخصی از دانه‌بندی است. بیشتر برنامه‌های داکینگ، کل فضای ساختاری لیگاند (لیگاند انعطاف پذیر) را در بر می‌گیرند و برخی برنامه‌های داکینگ تلاش می

² soman¹ sarin

در این روش، یک رابطه خطی بین پاسخ‌های مورد مطالعه (y_i) و مقادیر انتخاب شده توصیف‌کننده‌ها (X_{ij}) محاسبه می‌شود. در این معادله e_i خطای تصادفی است (باقیمانده مدل نیز نامیده می‌شود). b_0 عرض از مبدأ و b_j ضرایب توصیف‌کننده‌ها می‌باشند.

ماشین بردار پشتیبان (SVM) به‌عنوان یک ابزار مدل‌سازی قدرتمند، برای شناسایی الگوها و مدل‌سازی توسعه پیدا کرده است. SVM یکی از الگوریتم‌های نظارت‌شده یادگیری ماشین است [۱۶-۱۷]. اگر داده‌ها به‌صورت خطی قابل جداسازی باشند از ماشین‌های خطی برای تولید یک سطح بهینه استفاده می‌شود که داده‌ها را بدون خطا و با حداکثر فاصله‌ی میان صفحه و نزدیکترین نقاط آموزشی (بردارهای پشتیبان) تفکیک می‌کند. در غیر این‌صورت، برای داده‌هایی با رفتار غیرخطی از توابع کرنل مختلف، مانند کرنل خطی، چند جمله‌ای، تابع پایه شعاعی و تابع سیگموئید، می‌توان استفاده کرد [۱۸]. در مقایسه با روش‌های کلاسیک رگرسیون و شبکه‌های عصبی، SVM ها از مزایایی مانند بهینه‌سازی بهتر، توانایی تعمیم مناسب، اجرای ساده و کاربردهای چندبعدی برخوردار هستند. در روش SVM، تابع رگرسیون به‌صورت معادله ۲ بر روی داده‌ها تعریف می‌شود.

معادله ۲

$$y = f(x) = \sum_{i=1}^n w_i \phi(x_i) + b$$

که در آن $\{\phi(x_i) + b\}$ مشخص‌کننده ورودی‌ها $\{w_i\}_{i=1}^l$ و b مقادیر ثابت می‌باشند. ثابت‌های b و w با حداقل کردن تابع ۳ به‌دست می‌آید.

معادله ۳

$$R_{SVM}(C) = C \frac{1}{n} \sum_{i=1}^n L_{\varepsilon}(d_i, y_i) + \frac{1}{2} \|w\|^2$$

معادله ۴

$$L_{\varepsilon}(d, y) = \begin{cases} |d - y| & |d - y| \geq \varepsilon \\ 0 & \text{otherwise} \end{cases}$$

در معادله ۳، R_{SVM} تابع خطای منظم شده و $C \frac{1}{n} \sum_{i=1}^n L_{\varepsilon}(d_i, y_i)$ خطای تجربی است که توسط تابع

کنند برای مدل‌سازی گیرنده پروتئین را انعطاف پذیر در نظر بگیرند [۱۱ و ۱۲].

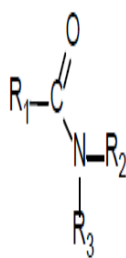
یکی از مطالعات انجام شده در زمینه داکینگ مولکولی، به مدل‌سازی QSAR مشتقات کامپوتوسین در برابر DNA توپوایزومراز (DNA Topo-I) با استفاده از رگرسیون خطی چندگانه و ارزیابی تقاطعی پرداخته است. نتایج این مطالعه نشان داد که مشتقات فعال پیش‌بینی‌شده دارای فعالیت ضدسرطانی مشخص و میل اتصال بالا به پروتئین هدف و توانایی تشکیل پیوند H-H هستند [۱۱].

در مطالعه‌ای دیگر، عملکرد مطالعات داکینگ بر روی سایت‌های فعال پروتئین ساخته‌شده توسط ساختمان مدل همولوژی با استفاده از فاکتور VIIa برای درمان و جلوگیری از خونریزی در بیماران مبتلا به هموفیلی A یا B استفاده شد [۱۳]. در مطالعه دیگری، توانایی برنامه داکینگ مولکولی که عمدتاً برای کمپلکس‌های لیگاند-پروتئین توسعه یافته‌اند، در بازتولید لیگاند DNA مورد ارزیابی قرار گرفت [۱۴].

هدف از این پژوهش، توسعه یک مدل تئوری مبتنی بر اصول QSAR برای مدل‌سازی دوز سمیت (LD50) برخی از مولکول‌های کربامات‌ها، S-آلکیل کربامات‌ها، اوره‌ها، کلرواستامیدها و مشتقات آنها می‌باشد. این مدل علاوه بر قابلیت پیش‌بینی LD50 ترکیباتی که داده‌های تجربی آنها در دسترس نمی‌باشد، می‌تواند به درک کامل و بهتر مکانیزم و فرایندهای موثر بر فعالیت این مولکول‌ها کمک کند. برای دستیابی به این هدف، برخی از روش‌های خطی و غیرخطی در انتخاب توصیف‌کننده استفاده شد و سپس توانایی تبیین آنها در پیش‌بینی DL50 توسط مدل‌ها خطی و غیرخطی مورد مقایسه قرار خواهد گرفت. در این راستا، رگرسیون خطی چندگانه (MLR) به‌عنوان روش خطی و ماشین بردار پشتیبان^۱ (SVM) به‌عنوان روش غیر خطی برای مدل‌سازی استفاده شد. در روش رگرسیون خطی چندگانه یک معادله ریاضی طرح‌ریزی می‌شود که شکل کلی آن به‌صورت معادله ۱ است [۱۵].

$$y_i = b_0 + \sum_{j=1}^n b_j X_{ij} + e_i \quad \text{معادله ۱}$$

^۱ Support vector machines



شکل ۱ ساختار شیمیایی ترکیبات مورد مطالعه

۲-۲ محاسبه توصیف کننده‌ها

ابتدا ساختار مولکولی ترکیبات با استفاده از نرم افزار hyperchem(7.0) رسم شد و پس از بهینه سازی، توصیف کننده‌های مولکولی با استفاده از نرم افزار Padel descriptor محاسبه شدند.

در گام بعد، توصیف کننده‌هایی انتخاب شدند که علاوه بر برخورداری از تفسیرپذیری فیزیکی و شیمیایی، قابلیت ایجاد یک مدل پیش‌بینی مناسب را داشته باشند. به همین منظور، ابتدا غربالگری اولیه انجام شد که شامل حذف متغیرهای ثابت و دارای همبستگی خطی بالا بود. پس از این مرحله، ۱۳۰ توصیف کننده باقی ماند. در ادامه از MLR به عنوان روش خطی و از GA به عنوان روش غیرخطی برای انتخاب توصیف کننده‌های مناسب استفاده شد.

۲-۳ انتخاب توصیف کننده‌ها با روش MLR

برای انتخاب متغیر به روش خطی، از تکنیک Stepwise MLR در نرم افزار SPSS (22.0) استفاده شد. در مجموع، ۱۱ مدل با تعداد توصیف کننده‌های مختلف به دست آمد. برای انتخاب بهترین توصیف کننده‌ها، نمودار مقادیر R و SE بر حسب تعداد توصیف کننده‌ها رسم شد. شکل ۲ نمودار مقادیر R و SE بر حسب تعداد توصیف کننده‌ها را نشان می‌دهد. با توجه به این شکل، مدل شماره ۱۰ با ۸ توصیف کننده، با ضریب همبستگی ۰/۹۳۵ و انحراف استاندارد ۰/۰۳۹ انتخاب شد. توصیف کننده‌های انتخاب شده با این روش در جدول ۲ آورده شده است.

اتلاف $L_E(d,y)$ (معادله ۴) محاسبه می‌شود و نشان دهنده این است که در مقادیر کمتر از ϵ هیچ خطایی وجود نخواهد داشت. همچنین، $\frac{1}{2}||w||^2$ نشان دهنده پیچیدگی مدل است. مقدار C ثابت تنظیم است و مصالحه بین خطای تجربی و پیچیدگی مدل را برقرار می‌کند و اندازه لوله نامیده می‌شود. این ثابت تنظیم، بیانگر فاصله ϵ و $+\epsilon$ حول خط رگرسیون است و مشخص کننده میزان صحت داده‌های سری آموزشی است. داده‌هایی که درون این لوله‌ی فرضی قرار بگیرند، خطا محسوب نشده و داده‌های خارج از این محدوده به عنوان مقادیر خطا در نظر گرفته می‌شوند [۱۷]. برای انجام این پژوهش، ابتدا ساختار مولکولی رسم و پس از بهینه سازی، توصیف کننده‌های مولکولی با استفاده از نرم افزارهای مربوطه محاسبه می‌شوند. سپس، روش‌های غربالگری خطی و غیرخطی برای انتخاب انواع متغیرها به کار گرفته می‌شوند. در ادامه بر اساس توصیف کننده‌های موثر، مدل نهایی برای پیش‌بینی، ارائه و ارزیابی می‌شود. در پایان برای بررسی کامل ارتباط ساختار مولکولی و برهمکنش آن از برنامه docking استفاده می‌شود.

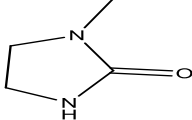
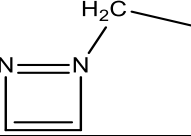
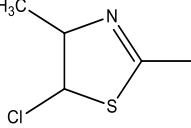
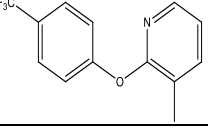
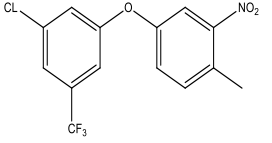
۲-۲ مواد و روش‌ها

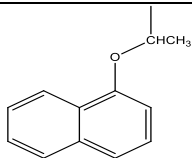
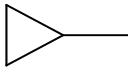
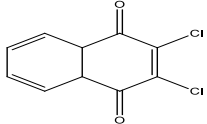
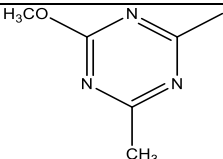
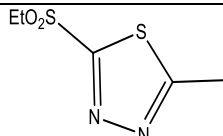
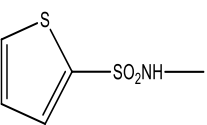
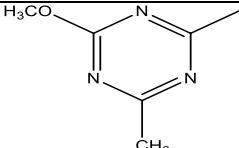
۲-۱ مواد

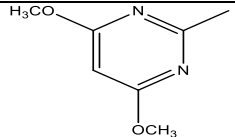
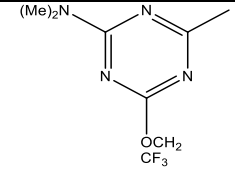
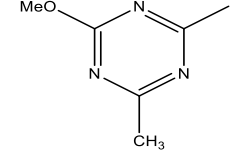
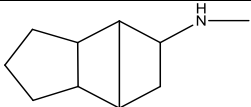
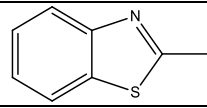
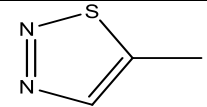
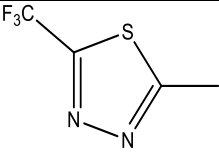
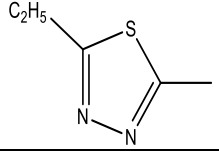
در این پژوهش، سری داده‌های مورد استفاده برای پیش‌بینی فعالیت بیولوژیکی برخی آفت کش‌ها از داده‌های موجود در مطالعه آدانا و همکاران (۲۰۱۳) انتخاب شده است [۱۹].

داده‌های این مقاله شامل ۷۶ ترکیب است که سمیت آنها با شاخص LD50 نمایش داده می‌شود. در این پژوهش LD50 به مقیاس لگاریتمی تبدیل شده و به صورت Log LD50 مورد استفاده قرار گرفته است. محدوده داده‌های سمیت از ۲,۲۱ تا ۲,۶۲ است. ساختار پایه‌ای این ترکیبات در شکل ۱ و مشخصات ساختاری آنها در جدول ۱ ارائه شده است.

جدول ۱ سمیت مشاهده شده ترکیبات مورد مطالعه

ترکیب	R1	R2	R3	Log LD50
1	2,6-diC(CH ₃) ₃ -4-CH ₃ -C ₆ H ₂ -O	H	Me	2/6
2	C ₂ H ₅ -NHCOCH(CH ₃)O	H	Ph	2/45
3	C ₆ H ₅ -NHCOOCH(CH ₃)	H	Et	2/5
4	3-NHCOOC ₂ H ₅ -C ₆ H ₄ -O	H	Ph	2/51
5	2-COOH-C ₆ H ₄	H	Naphtyl	2/45
6	4-NHCOOCH ₃ -C ₆ H ₄ -O	H	3-CH ₃ -C ₆ H ₄	2/51
7	C ₂ H ₅ S	(CH ₃) ₂ CHCH ₂	(CH ₃) ₂ CHCH ₂	2/41
8	(CH ₃) ₃ CCCHBr	H	C ₆ H ₅ -C(CH ₃) ₂	2/51
9	C ₆ H ₅	CH ₃ OCOCH(CH ₃)	3-Cl,4-F-C ₆ H ₃	2/54
10	CH ₃ O	H	3,4-diCl-C ₆ H ₃	2/38
11	4-NHCOOCH(CH ₃) ₂ -C ₆ H ₄ -O	C ₂ H ₅	Ph	2/55
12	n-C ₃ H ₇ C(CH ₃) ₂	H	4-Cl-C ₆ H ₄	2/41
13	(CH ₃) ₂ CHO	H	3-Cl-C ₆ H ₄	2/36
14	C ₂ H ₅ S	C ₂ H ₅	c-Hx	2/4
15		H	(CH ₃) ₂ CHCH ₂	2/3
16	ClCH ₂		2,6-diCH ₃ -C ₆ H ₃	2/46
17	C ₂ H ₅	H		2/33
18		H	2,4-diF-C ₆ H ₃	2/6
19	n-C ₃ H ₇ S	C ₂ H ₅	n-Bu	2/31
20	n-C ₃ H ₇ S	n-C ₃ H ₇	n-Pr	2/32
21	ClCH ₂	CH ₃ OCH ₂	2,6-diC ₂ H ₅ -C ₆ H ₃	2/44
22	C ₂ H ₅ S	n-C ₃ H ₇	n-Pr	2/29
23	C ₆ H ₅	C ₂ H ₅ OCOCH(CH ₃)	3,4-diCl-C ₆ H ₃	2/57
24		H	CH ₃ SO ₂	2/6
25	C ₂ H ₅	H	3,4-diCl-C ₆ H ₃	2/35
26	ClCH ₂ CCCH ₂ O	H	3-Cl-C ₆ H ₄	2/42
27	4-Cl-C ₆ H ₄ -CH ₂ S	C ₂ H ₅	C ₂ H ₅	2/42
28	ClCH ₂	HCCCH(CH ₃)	C ₆ H ₅	2/35

29	$(C_2H_5)_2CH$	CH_3	CH_3	2/38
30	$2-Cl-C_6H_4-CH_2S$	C_2H_5	C_2H_5	2/42
31	$CH_2C(Cl)CH_2S$	C_2H_5	C_2H_5	2/32
32	$ClCH_2$	$CH_2=CHCH_2$	$CH_2=CHCH_2$	2/25
33	$ClCH_2$	$(CH_3)_2CH$	C_6H_5	2/33
34		C_2H_5	C_2H_5	2/43
35	C_2H_5S	C_2H_5	C_2H_5	2/21
36	$ClCHC(Cl)CH_2S$	$(CH_3)_2CH$	$(CH_3)_2CH$	2/43
37		H	3,4-diCl-C ₆ H ₃	2/36
38	C_6H_5	H	$HOOCCH_2O$	2/35
39	C_6H_5	$(CH_3)_2NCO$	3,4-diCl-C ₆ H ₃	2/54
40	Cl_2CH	$CH_2=CHCH_2$	$CH_2=CHCH_2$	2/34
41	Cl_2CH	H		2/56
42	$(CH_3)_2C=N-O$	H	C_6H_5	2/3
43	$(CH_3)_2CHO$	H	C_6H_5	2/31
44	$HCCCH(CH_3)O$	H	3-Cl-C ₆ H ₄	2/37
45	3,4-diCl-C ₆ H ₃ NH-	CH_3	CH_3	2/58
46	2-Cl,4-CH ₃ -C ₆ H ₃ NH-	CH_3	CH_3	2/41
47	$C_6H_5C(CH_3)_2NH-$	C_6H_5	CH_3	2/48
48	3-CF ₃ -C ₆ H ₄ NH-	CH_3	CH_3	2/43
49	4-OCH ₃ -C ₆ H ₄ OC ₆ H ₄ NH-	CH_3	CH_3	2/49
50	C_6H_5NH-	CH_3	CH_3	2/31
51	4-Cl-C ₆ H ₄ NH-	CH_3	$HCCCH(CH_3)$	2/4
52	2-Cl-C ₆ H ₄ -SO ₂ NH-	H		2/57
53	3-Cl,4-Br-C ₆ H ₃ NH-	OCH_3	CH_3	2/5
54	CH_3NH-	CH_3		2/45
55		H		2/54

56	2-CO ₂ CH ₃ -C ₆ H ₄ -CH ₂ SO ₂ NH-	H		2/62
57	2-CO ₂ CH ₃ ,6-CH ₃ -C ₆ H ₄ -SO ₂ NH-	H		2/68
58	2-CO ₂ CH ₃ -C ₆ H ₄ -SO ₂ NH-	H		2/59
59	3,4-diCl-C ₆ H ₄ NH-	CH ₃	OCH ₃	2/42
60	4-Cl-C ₆ H ₄ -O-C ₆ H ₄ NH-	CH ₃	CH ₃	2/48
61	4-Cl-C ₆ H ₄ NH-	CH ₃	CH ₃	2/32
62	3-Cl,4-CH ₃ O-C ₆ H ₃ NH-	CH ₃	CH ₃	2/38
63	4-Br-C ₆ H ₄ NH-	CH ₃	CH ₃	2/21
64	4-Cl-C ₆ H ₄ NH-	OCH ₃	CH ₃	2/33
65		CH ₃	CH ₃	2/34
66	4-(CH ₃) ₂ CH-C ₆ H ₄ NH-	CH ₃	CH ₃	2/33
67	CH ₃ NH-	H		2/32
68	3-OCF ₂ CHF ₂ -C ₆ H ₄ NH-	CH ₃	CH ₃	2/45
69	3-Cl,4-C(F ₂ Cl)-S-C ₆ H ₃ NH-	CH ₃	CH ₃	2/49
70	3,4-diCl-C ₆ H ₃ NH-	HCCCH ₂	(CH ₃) ₂ CH	2/45
71	c-octyl-NH-	CH ₃	CH ₃	2/32
72	3-OCONHC(CH ₃) ₃ -C ₆ H ₄ NH-	CH ₃	CH ₃	2/46
73	3,4-diCl-C ₆ H ₃ NH-	n-C ₄ H ₉	CH ₃	2/5
74	C ₆ H ₅ NH-	H		2/4
75	CH ₃ NH-	CH ₃		2/41
76	CH ₃ NH-	CH ₃		2/38



شکل ۲ نمودار مقادیر R و SE بر حسب تعداد توصیف کننده ها

جدول ۲ توصیف کننده های انتخاب شده به روش MLR

شماره متغیر	توصیف کننده ها	تعریف
V ₁	SaasC	Sum of atom-type E-State: :C:-
V ₂	mindO	Minimum atom-type E-State: =O
V ₃	nHBa	Count of E-States for (strong) Hydrogen Bond acceptors
V ₄	GATS3i	Geary autocorrelation - lag 3 / weighted by first ionization potential
V ₅	AATS8i	Average Broto-Moreau autocorrelation - lag 8 / weighted by first ionization potential
V ₆	GATS5m	Geary autocorrelation - lag 5 / weighted by mass
V ₇	nsssN	Count of atom-type E-State: >NH-+
V ₈	MDEC-33	Molecular distance edge between all tertiary carbons

مانند، اطلاعات حیاتی ذخیره شده در این ساختارهای داده ای را حفظ می کند [۲۰-۲۱]. مسئله ای که توسط الگوریتم ژنتیک حل می شود به پیش نیاز خاصی نیاز ندارد. این ویژگی، الگوریتم ژنتیک را به یک ابزار بهینه سازی قدرتمند تبدیل کرده است. برای انتخاب توصیف کننده ها با روش GA از GA-PLS toolbox تحت MATLAB استفاده شد. توصیف کننده های انتخاب شده با این روش در جدول ۳ آمده است.

۲-۴ انتخاب توصیف کننده ها با روش GA

الگوریتم ژنتیک، خانواده ای از مدل های محاسباتی^۱ است که از مفهوم تکامل^۲ الهام گرفته شده اند. این دسته از الگوریتم ها، جواب های محتمل^۳ یا جواب های کاندید^۴ برای یک مسأله خاص را در یک ساختار داده ای کروموزوم مانند^۵ کدبندی می کنند. الگوریتم ژنتیک از طریق اعمال عملگرهای باز ترکیب^۶ روی ساختارهای داده ای کروموزوم

⁴ Candidate

⁵ Chromosome-like

⁶ Recombination Operators

¹ Computational Models

² Evolution

³ Potential Solutions

جدول ۳ توصیف‌کننده‌های انتخاب‌شده به روش GA

شماره متغیر	توصیف‌کننده‌ها	تعریف
V ₁	ATSC3i	Centered Broto-Moreau autocorrelation - lag 3 / weighted by first ionization potential
V ₂	SPMax8_Bhm	Largest absolute eigenvalue of Burden modified matrix - n 8 / weighted by relative mass
V ₃	SPMax1_Bhp	Largest absolute eigenvalue of Burden modified matrix - n 1 / weighted by relative mass
V ₄	nHBa	Count of E-States for (strong) Hydrogen Bond acceptors
V ₅	minssS	Minimum atom-type E-State: >Be<-2
V ₆	IC0	Information content index (neighborhood symmetry of 0-order)

۲-۵ مدل‌سازی با روش MLR

$$-\log LD50 = 1.213 + 0.010(v_1) + 0.097(v_2) + 0.040(v_3) + 0.110(v_4) - 0.001(v_5) - 0.076(v_6) - 0.036(v_7) + 0.008(v_8)$$

مدل به‌دست‌آمده با توصیف‌کننده‌های به‌دست آمده به روش GA در معادله ۶ آمده است.

$$-\log LD50 = 1.085 - 0.001(v_1) + 0.099(v_2) + 0.161(v_3) + 0.891(v_4)$$

پیش از ساخت مدل با روش MLR، مجموعه داده‌ها به دو بخش شامل، سری آموزشی (۶۱ مولکول) برای توسعه مدل و سری آزمون (۱۵ مولکول) برای ارزیابی مدل تقسیم شدند. سپس، با استفاده از توصیف‌کننده‌های انتخاب شده، مدل MLR توسعه داده شد. این مدل در معادله ۵ ارائه شده است.

معادله ۵

جدول ۴ پارامترهای آماری مدل MLR

	Train set				Test set			
	R	Radj	SE	F	R	Radj	SE	F
MLR-MLR	0.932	0.867	0.037	391.385	0.347	0.053	0.1	1.781
MLR-GA	0.856	0.728	0.054	161.511	0.641	0.366	0.079	9.068

جدول ۵ پارامترهای آماری مدل SVM

	Train set		Test set		Model	
	R	SE	R	SE	Q ²	SPRESS
SVM-MLR	0/99	0/014	0/718	0/082	0/498	0/047
SVM-GA	0/99	0/017	0/462	0/136	-0/384	0/06

حالت الکترونیکی ذاتی، تحت تأثیر نیروی الکترونیکی تمامی اتم‌های اطراف در مولکول قرار دارد و به میزان الکترونگاتیوی اتم‌ها وابسته است

توصیف‌کننده‌های GATS (GATS3i و GATS5m) از دسته توصیف‌کننده‌های توپولوژی از نوع Geary می‌باشد. ضریب Geary یک تابع از نوع فاصله است که مقادیر آن از صفر تا بی‌نهایت متغیر است. AATS از دسته توصیف‌کننده‌های توپولوژی خود همبسته Moreau-Broto است. این همبستگی، یک همبستگی مکانی است که از مجموع مربعات خواص اتمی (جرم اتمی، قابلیت قطبیت، بار و الکترونگاتیویته) در گراف مولکولی به‌دست می‌آید [۲۳].

توصیف‌کننده nssN از دسته ساده‌ترین توصیف‌کننده‌های مولکولی است که بر پایه شمارش اتم‌های تشکیل‌دهنده ترکیب است و اطلاعات توپولوژیکی و الکترونیکی مربوط به آن را نشان می‌دهد [۲۴].

توصیف‌کننده MDEC-33 از دسته توصیف‌کننده‌های Molecular distance و از نوع توپولوژیکی است. این توصیف‌کننده به ترازبندی فضایی ساختار مولکولی مربوط است [۲۵].

توصیف‌کننده mindO و SaasC نیز از دسته توصیف‌کننده‌های الکترونتوپولوژیکی هستند که با توجه به گروه‌های عاملی در همسایگی هر اتم در گراف مولکولی به‌دست می‌آیند [۲۲]. همبستگی خطی این توصیف‌کننده‌ها در جدول ۶ آورده شده است. همانطور که مشاهده می‌شود همبستگی معناداری بین توصیف‌کننده‌ها وجود ندارد.

جدول ۶ همبستگی توصیف‌کننده‌های مدل MLR

	V1	V2	V3	V4	V5	V6	V7	V8
V1	1	.098	-.224	.220	-.004	-.420	.202	-.055
V2		1	.199	.055	.063	-.142	.341	.360
V3			1	-.061	.208	-.208	-.172	.594
V4				1	-.134	-.227	.148	.076
V5					1	-.108	.034	.037
V6						1	-.188	-.277
V7							1	-.156
V8								1

جدول ۴ پارامترهای آماری مدل MLR با استفاده از توصیف‌کننده‌های MLR و GA را نشان می‌دهد.

۲-۶ مدل‌سازی با روش SVM

از آنجایی که تابع شعاعی کرنل رایج‌ترین تابع مورد استفاده در SVM است، در این پژوهش برای مدل‌سازی $-\log$ LD50 از این تابع استفاده شد و تمامی محاسبات با استفاده از برنامه نوشته‌شده تحت نرم‌افزار MATLAB انجام شد. در این مدل‌سازی، مقادیر $-\log$ LD50 به‌عنوان متغیر وابسته و توصیف‌کننده‌های انتخاب‌شده با روش‌های MLR و GA به‌عنوان ورودی (متغیرهای مستقل) SVM به‌طور جداگانه برای انجام مدل‌سازی وارد شدند. پارامترهای آماری به‌دست‌آمده از مدل‌سازی SVM با استفاده از توصیف‌کننده‌های انتخابی توسط روش‌های انتخاب متغیر MLR و GA در جدول ۵ ارائه شده است.

۳- نتایج

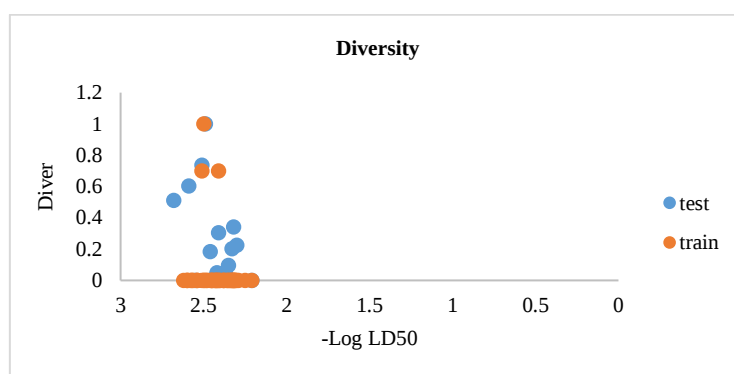
۳-۱ تفسیر توصیف‌کننده‌ها

با توجه به نتایج به‌دست آمده از مدل‌سازی ارائه شده در جدول‌های ۴ و ۵، مدل SVM-MLR به‌دلیل برخورداری از شاخص‌های آماری مطلوب‌تر، به‌عنوان بهترین مدل انتخاب شد. بنابراین، توصیف‌کننده‌های انتخاب‌شده با روش MLR مورد بررسی قرار گرفتند و آزمون پراکندگی چگونگی توزیع مولکول‌ها برای این مدل بررسی شد. توصیف‌کننده nHBa از دسته توصیف‌کننده‌های E-State است که توسط کی‌یر و هال برای خصوصیات الکترونیکی و توپولوژیکی اتم، معرفی شد [۲۲]. در این توصیف‌کننده،

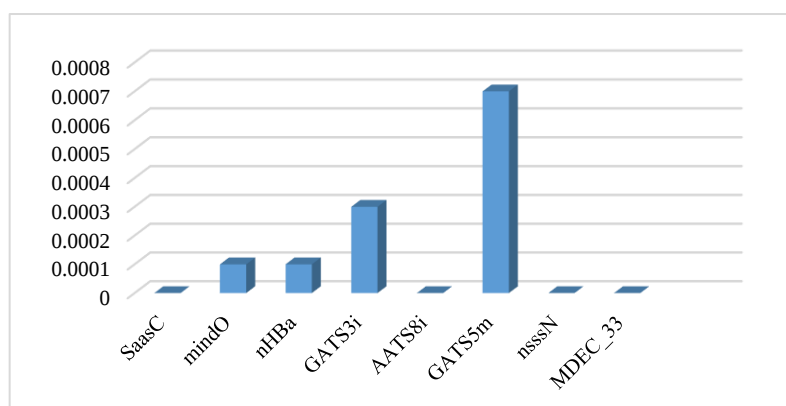
۳-۲ آنالیز حساسیت

روش آنالیز حساسیت بر پایه صفر کردن ترتیبی متغیرها برای بهترین مدل (مدل MLR-SVM) است. برای متغیرهایی که صفر شده‌اند، $RMSE_i$ به عنوان خطای پیش‌بینی محاسبه می‌شود. مقدار $RMSE_i$ معمولاً در حالت صفر شدن هر توصیف‌کننده افزایش پیدا می‌کند. سپس، تفاوت‌های بین $RMSE_i$ و $RMSE$ اولیه مدل ایجاد شده محاسبه شده و به صورت $RM_{(diff)}$ برای هر توصیف‌کننده به دست می‌آید [۲۶]. شکل ۴ مقادیر $RM_{(diff)}$ را برای هر یک از توصیف‌کننده‌های انتخاب شده با روش MLR نشان می‌دهد.

شکل ۳، آزمون پراکندگی چگونگی توزیع مولکول‌ها بر اساس توصیف‌کننده‌های انتخاب شده را نشان می‌دهد. این تحلیل یا بررسی بر این واقعیت استوار است که ساختار سری آموزشی و آزمون، نماینده کل ساختارها در سری داده‌ها می‌باشند [۲۶]. در این آزمون ابتدا اسکور فاصله (d_{ij}) برای دو ترکیب X_i و X_j با محاسبه فاصله اقلیدسی بر اساس توصیف‌کننده‌های انتخابی به دست آمد و سپس میانگین فواصل نمونه‌ها (Diver) در سری داده در مقابل $-\log LD50$ رسم شد. برای انجام این آزمون از برنامه نوشته شده تحت نرم‌افزار MATLAB استفاده شد. همانطور که در شکل‌ها مشاهده می‌شود توزیع مناسبی از سری آزمون در میان کل داده‌های آموزش وجود دارد. توزیع مناسب سری آزمون در میان داده‌های آموزشی قدرت پیش‌بینی مدل را افزایش می‌دهد.



شکل ۳ نمودار آزمون پراکندگی برای سری آموزش و آزمون روش MLR



شکل ۴ نتیجه آنالیز حساسیت توصیف‌کننده‌های انتخاب شده با روش MLR

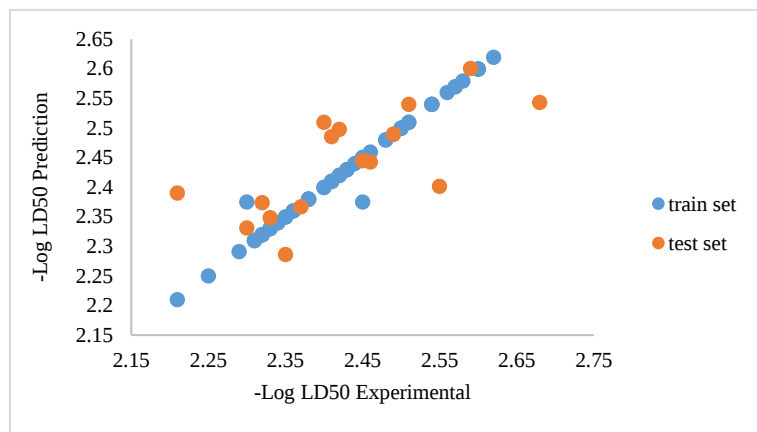
۳-۳ ارزیابی اعتبار مدل

شکل ۵ مقادیر پیش‌بینی $-\log LD50$ به دست آمده توسط مدل SVM-MLR بر حسب مقادیر تجربی را نشان می‌دهد. همان‌طور که مشاهده می‌شود، همبستگی خوبی میان نتایج پیش‌بینی و تجربی وجود دارد. همچنین، در شکل ۶ اختلاف میان مقادیر تجربی و پیش‌بینی (باقیمانده‌ها) بر حسب مقادیر تجربی رسم شده است. توزیع یکنواخت و تصادفی باقی‌مانده‌ها حول محور صفر نشان دهنده عدم وجود خطای سیستماتیک در مدل SVM-MLR است.

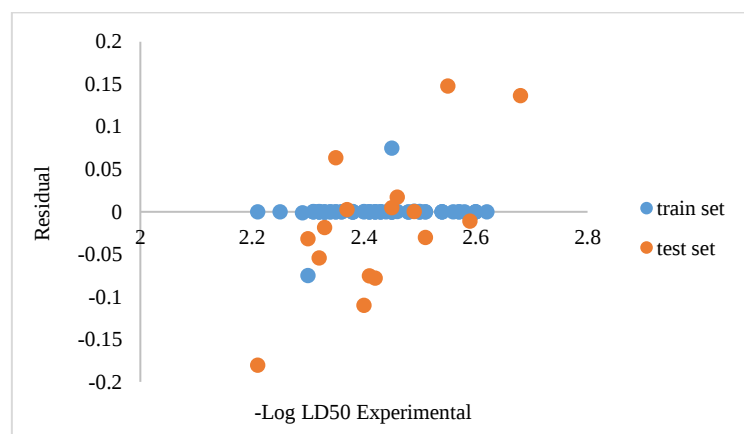
اهمیت توصیف‌کننده‌های انتخاب شده به روش خطی به ترتیب به صورت زیر می‌باشد:

$GATS5m > GATS3i > mindO > nHBa > SaasC = AATS8i = nsssN$

همان‌طور که ملاحظه می‌شود توصیف‌کننده‌های $GATS$ که از خانواده توصیف‌کننده‌های خود همبسته Geary و الکترونی می‌باشند نسبت به سایر توصیف‌کننده‌ها از اهمیت بالاتری برخوردار هستند. همچنین، توصیف‌کننده‌های $nHBa$ و $mindO$ که از دسته توصیف‌کننده‌های الکترونی هستند، اهمیت قابل توجهی در پیش‌بینی میزان سمیت دارند.



شکل ۵ نمودار مقادیر پیش‌بینی شده بر حسب مقادیر تجربی در مدل SVM-MLR



شکل ۶ نمودار مقادیر باقی مانده بر حسب مقادیر تجربی در مدل SVM-MLR

می‌توان اطمینان حاصل کرد که مدل، حاصل همبستگی تصادفی بین مقادیر متغیر وابسته و متغیرهای مستقل نیست [۲۷].

۳-۴ دامنه کاربرد مدل

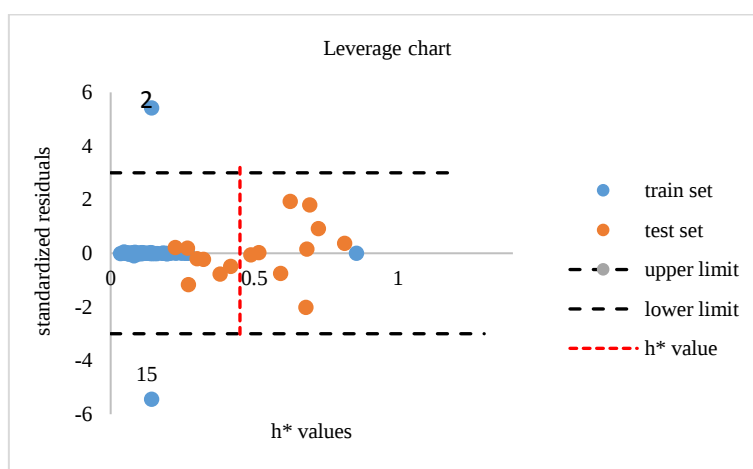
یکی از مهم‌ترین مراحل در توسعه مدل QSAR، مشخص کردن دامنه کاربرد (AD) مدل است که برای ارزیابی آن از نمودار ویلیام استفاده می‌شود.

همچنین، آزمون درهم آمیختگی-Y برای اطمینان از عدم وجود همبستگی تصادفی در مدل ایجاد شده و ارزیابی تقاطعی و محاسبه Q^2 انجام شد. برای آزمون درهم آمیختگی-Y، ترتیب مقادیر متغیر وابسته (Y) در سری آموزش به صورت تصادفی و بدون ایجاد تغییری در ماتریس توصیف‌کننده‌ها عوض می‌شود. سپس، مدل جدید ایجاد شده و نتایج آماری آن ارزیابی می‌شود. این کار ۲۰ بار انجام شد و نتایج آن در جدول ۷ آمده است.

از آنجایی که مقدار میانگین Q^2 مدل‌های ایجاد شده بسیار کمتر از مقدار Q^2 مدل اصلی است، با استفاده از این آزمون

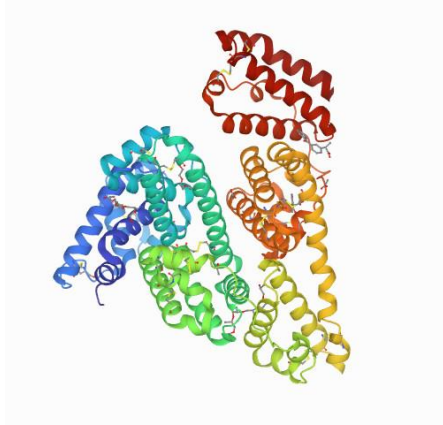
جدول ۷ آزمون در هم آمیختگی-Y برای مدل SVM-MLR

Q^2	شماره تکرار	Q^2	شماره تکرار
0/321	11	-1/857	1
-0/719	12	-1/196	2
-1/272	13	-2/444	3
0/305	14	-0/791	4
-3/476	15	-3/197	5
0/836	16	-1/516	6
-0/76	17	-0/242	7
-1/16	18	-1/45	8
-0/00062	19	-1/111	9
-3/056	20	-0/209	10



شکل ۷ نمودار دامنه کاربرد مدل SVM-MLR، خطوط تیره افقی نمایانگر $\pm 3\sigma$ هستند و خط تیره عمودی نمایانگر مقدار بحرانی حد هشدار ($h^*=0/45$) است.

شکل ۸ ساختار بلوری آلبومین سرم لیورین در کمپلکس با کتوپروپن با کد 6OCK است که از بانک اطلاعاتی پروتئین با فرمت PDB گرفته شده است. این پروتئین با قدرت تفکیک^۱ برابر با ۱/۹۰ آنگستروم و مقدار R (R-Value Observed) برابر با ۰/۱۹۲ به عنوان مدل پروتئینی مناسب انتخاب شد. مقدار R میزان توافق بین مدل کریستالوگرافی داده تجربی حاصل از پراش اشعه ایکس را نشان می‌دهد.



شکل ۸ ساختار بلوری آلبومین سرم لیورین در کمپلکس با کتوپروپن
کیفیت ساختار بلوری پروتئین نقش تعیین کننده‌ای در صحت نتایج داکینگ مولکولی دارد. از آنجایی که، در بهترین ساختارهای بلوری به دست آمده از اشعه ایکس، اتم‌های هیدروژن حذف می‌شوند، بنابراین باید قبل از انجام داکینگ، هیدروژن‌هایی که در ساختار کریستالوگرافی دیده نمی‌شوند به ساختار اضافه شوند. همچنین، مولکول‌های آب و یون‌های همراه پروتئین نیز باید حذف شوند. فایل پروتئین و فایل لیگاند با فرمت PDB برای داکینگ آماده می‌شوند.

در این نمودار باقی مانده‌های استاندارد شده بر حسب مقادیر حد هشدار (h_i) رسم می‌شوند. نتایجی با باقیمانده استاندارد بزرگتر از ۳ واحد انحراف استاندارد یا مولکول‌های موثر از لحاظ ساختاری در مدل که $h_i > h^*$ دارند، داده‌های پرت مولکول هستند.

دامنه کاربرد مدل MLR-SVM در شکل ۷ نشان داده شده است. همانطور که در شکل ۷ مشاهده می‌شود، مولکول شماره ۱۵ خارج از حد -3σ و مولکول شماره ۲ خارج از حد $+3\sigma$ می‌باشد و به عنوان داده پرت شناسایی می‌شوند. اما این مسئله الزاماً به دلیل نادرست بودن ساختار مولکولی نیست، بلکه ممکن است ناشی از خطاهای موجود در داده‌های تجربی باشد. پراکندگی مولکول‌های سری آزمون در راستای محور X در دسته نقاط Leverage خوب قرار می‌گیرند که ناشی از ساختار و ماهیت ترکیبات می‌باشند و باعث خطا در محاسبات آماری نمی‌شوند. مدل پیشنهادی در پژوهش حاضر با مطالعه آدادا و همکارانش مقایسه شد و نتایج آن در جدول ۸ آمده است.

با مقایسه پارامترهای آماری به دست آمده در جدول ۸ و عدم استفاده از توصیف‌کننده‌های تجربی و گستردگی نوع توصیف‌کننده‌های مولکولی به کار رفته در مدل MLR-SVM، به نظر می‌رسد مدل MLR-SVM در پیش‌بینی و تبیین عوامل ساختاری موثر در میزان سمیت آفت‌کش‌های مورد بررسی موفق‌تر عمل می‌کند.

۳-۵ مطالعات داکینگ مولکولی

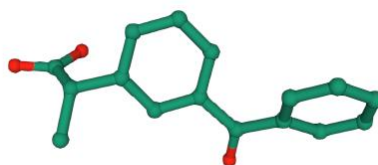
محدوده داده‌های سمیت بر اساس مقادیر $-\log LD50$ از ۲,۲۱ تا ۲,۶۲ است که از بین آنها مولکول‌های ۳۵ و ۶۳ با ضعیف‌ترین و مولکول ۵۷ با قوی‌ترین اثر سمیت برای مطالعه داکینگ انتخاب شد.

جدول ۸ مقایسه پارامترهای آماری مدل MLR-SVM با مدل‌های مشابه

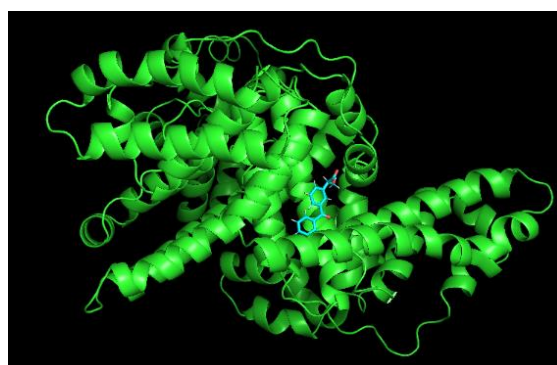
مدل‌ها	تعداد توصیف‌کننده‌های تجربی	تعداد توصیف‌کننده‌ها	R	SE	Q2
MLR-SVM	-	۸	0/99	0/014	0/498
PLS[19]	۱	۸	0/84	0/26	0/54
ANN[19]	۱	۸	0/81	0/11	0/66

^۱ Resolution

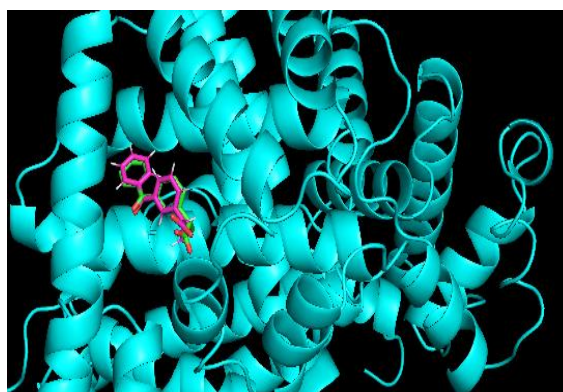
ایکس آلومین لپورین حذف شد و برای تایید فرایند داکینگ انجام شده مجدداً داکینگ با این لیگاند انجام شد. این مرحله با هدف ارزیابی اعتبار فرایند داکینگ انجام شد تا مشخص شود آیا لیگاند 9KL پس از داکینگ، مجدداً در همان محل اتصال اولیه قرار می‌گیرد یا خیر (شکل ۹ب). پس از فرایند ریداک و جستجوی تمام مکان‌های ممکن در پروتئین مطابق شکل ۱۰ و ۱۱، باز هم لیگاند به مکان قبلی خود بازگشت که این امر نشان دهنده انجام یک داکینگ موفق است.



شکل ۹ الف ساختار لیگاند 9KL کمپلکس شده با آلومین سرم لپورین

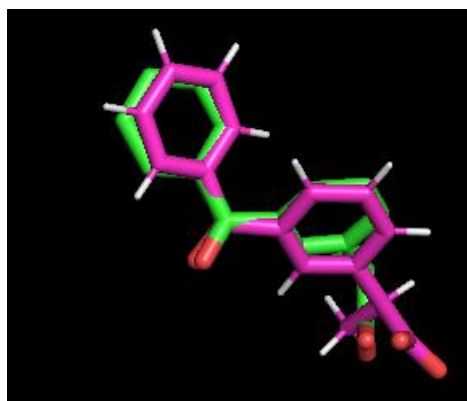


شکل ۹ ب ساختار آلومین سرم لپورین کمپلکس شده با لیگاند 9KL



شکل ۱۰ تصویر انطباقی لیگاند قبل و بعد از ریداک

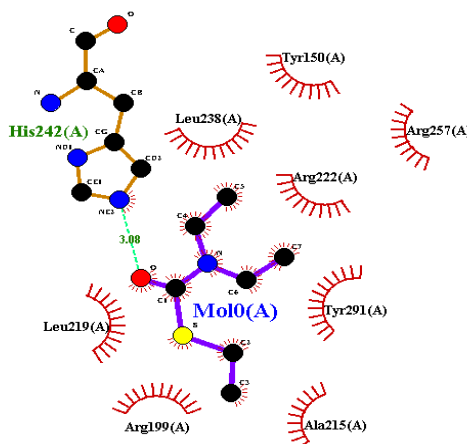
در پژوهش حاضر برای محاسبات داکینگ مولکولی از نرم افزار Autodock vina استفاده شد. لیگاند به صورت انعطاف‌پذیر (Flexible ligand) و پروتئین به صورت صلب در نظر گرفته شد. برای پیدا کردن جایگاه فعال آلومین سرم لپورین از لیگاند 9KL استفاده شد. این لیگاند از بانک اطلاعاتی پروتئین با فرمت PDB گرفته شده است و در کمپلکس با این پروتئین است. شکل ۹ الف، ساختار لیگاند 9KL را نشان می‌دهد. برای اطمینان بیشتر از جایگاه فعال پروتئین، لیگاند 9 KL از ساختار کریستالوگرافی اشعه



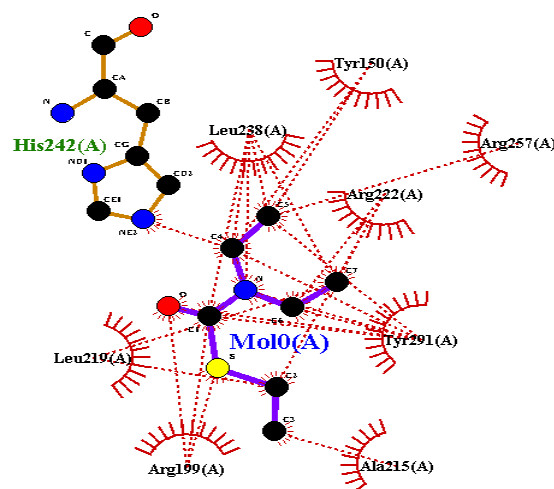
شکل ۱۱ تصویر انطباقی لیگاند قبل و بعد از ریداک

به ترتیب پیوند هیدروژنی و برهمکنش‌های آبگریز ترکیبات ۳۵ و ۶۳ (با کمترین فعالیت) و ۵۷ (بیشترین فعالیت) را نشان می‌دهند. بررسی این شکل‌ها نشان می‌دهد که ترکیب شماره ۳۵ در مجموع ۹ برهمکنش با سایت فعال آلومین برقرار می‌کند. اسیدهای آمینه محل اتصال شامل، Tyr150، Arg257، Arg222، Leu238، His242، Leu219، Tyr291، Arg199، Ala215 هستند که از این میان اسیدهای آمینه Tyr150، Arg257، Arg222، Leu238، Leu219، Tyr291، Arg199، Ala215 با ترکیب ۳۵ برهمکنش‌های آبگریز تشکیل می‌دهند. در حالی که، اسید آمینه His242 از طریق پیوند هیدروژنی با این ترکیب برهمکنش دارد.

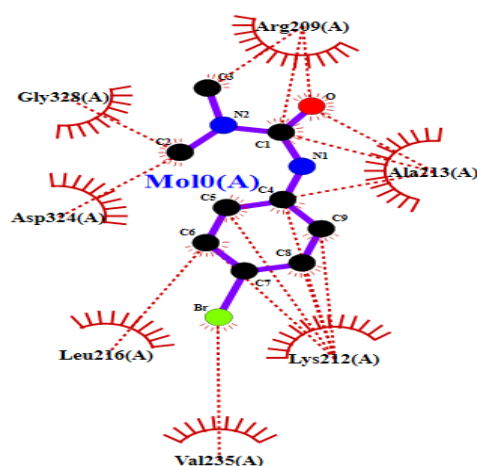
با توجه به محل اتصال آلومین سرم لپورین، شبکه‌ای با ابعاد (Grid) $66 \times 66 \times 56$ و فاصله شبکه‌ای (Grid spacing) برابر با 0.375 آنگستروم در اطراف محل اتصال این آلومین برای ایجاد فایل‌های شبکه گیرنده (Receptor grid) تعریف شد. در این فضای شبکه‌ای، لیگاند می‌تواند آرایش و شکل‌های مختلفی را به خود بگیرد. در نهایت، از میان تمامی صورت‌بندی‌های حاصل از کد نویسی در محیط commat prompt، تنها یک صورت بندی با منفی‌ترین انرژی پیوند (Binding energy) به عنوان بهترین صورت‌بندی انتخاب می‌شود. برای نشان دادن پیوند هیدروژنی و برهمکنش‌های آبگریز، از نرم‌افزار Ligplot⁺ استفاده شد. شکل‌های ۱۲ تا ۱۶



شکل ۱۲ طرح دو بعدی پیوندهای هیدروژنی بین ترکیب ۳۵ و اسیدهای آمینه آلومین سرم لپورین



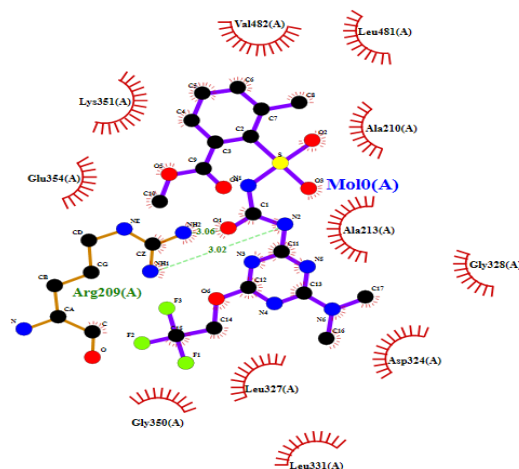
شکل ۱۳ طرح دو بعدی نرم‌افزار Ligplot+ از برهمکنش‌های پیوندهای آبگریز بین ترکیب ۳۵ و اسیدهای آمینه آلبومین سرم لبورین



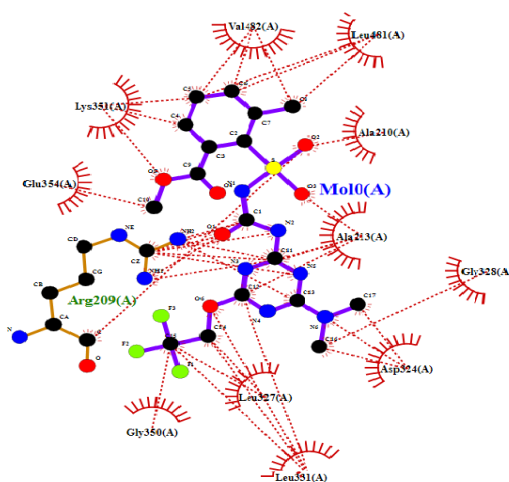
شکل ۱۴ طرح دو بعدی نرم‌افزار Ligplot+ از برهمکنش‌های پیوندهای آبگریز بین ترکیب ۶۳ و اسیدهای آمینه آلبومین سرم لبورین

می‌باشند که از این میان، اسیدهای آمینه Leu481، Leu331، Gly350، Glu354، Lys351، Val482، Leu327، Asp324، Gly328، Ala213، Ala210 برهمکنش‌های آبگریز با ترکیب ۵۷ تشکیل می‌دهند. در حالی که، اسید آمینه Arg209 از طریق پیوند هیدروژنی با این ترکیب برهمکنش دارد. از مقایسه تصاویر به‌دست آمده از نرم‌افزار Ligplot+ می‌توان نتیجه گرفت که افزایش برهمکنش‌های آبگریز رابطه مستقیمی با میزان دوز سمیت آفتکش‌ها دارد و ترکیب ۵۷ با بیشترین تعداد پیوندهای آبگریز دارای بیشترین اثر سمیت است.

ترکیب ۶۳ نیز ۶ برهمکنش آبگریز با سایت اتصال آلبومین دارد. اسیدهای آمینه محل اتصال شامل Arg209، Gly328، Lys212، Leu216، Asp324، Ala213، Val235 می‌باشند. همه این اسیدهای آمینه با ترکیب ۶۳ برهمکنش‌های آبگریز دارند و هیچگونه پیوند هیدروژنی بین این اسیدهای آمینه و ترکیب ۶۳ مشاهده نشده است. ترکیب ۵۷ نیز ۱۲ برهمکنش با سایت اتصال آلبومین دارد. اسید آمینه‌های محل اتصال شامل Leu481، Val482، Leu331، Gly350، Arg209، Glu354، Lys351، Leu327، Asp324، Gly328، Ala213، Ala210



شکل ۱۵ طرح دو بعدی نرم افزار Ligplot⁺ از بر همکنش های پیوندهای هیدروژنی بین ترکیب ۵۷ و اسیدهای آمینه آلبومین سرم لیورین



شکل ۱۶ طرح دو بعدی نرم افزار Ligplot⁺ از بر همکنش های پیوندهای آبریز بین ترکیب ۵۷ و اسیدهای آمینه آلبومین سرم لیورین

۴- نتیجه گیری

منتخب، که شامل توصیف کننده های توپولوژیکی و الکترونی هستند، قابلیت مناسبی در تبیین رابطه ی میان دوز سمیت و ساختار مولکولی این ترکیبات برخوردار هستند. با توجه به ساختار ترکیبات مورد مطالعه، به نظر می رسد حضور عناصری مانند فلوئور در ساختار مولکولی، با تاثیر بر توصیف کننده های الکترونی، موجب افزایش سمیت می شود. همچنین، نتایج مطالعات داکینگ مولکولی نشان داد که مولکول هایی توانایی بیشتری در برقراری پیوندهای آبریز، فعالیت بیشتری از خود نشان داده و در نتیجه سمیت بالاتری دارند.

در این پژوهش، از روش های stepwise- MLR و GA برای انتخاب متغیر و از روش های MLR و SVM به عنوان روش های مدل سازی خطی و غیرخطی استفاده شد. نتایج آماری به دست آمده از ارزیابی مدل ها نشان می دهد که مدل SVM-MLR در مقایسه با سایر روش ها، دقت و توانایی بیشتری برای پیش بینی میزان سمیت آفت کش های مورد بررسی دارد. همچنین، گستردگی توصیف کننده های

۵- منابع

- [1] Aktar, M. W.; Sengupta, D.; Chowdhury, A. Impact of pesticides use in agriculture: their benefits and hazards. *Interdiscip. Toxicol.* 2009, 2, 1-12.
- [2] Del carlo .M, Mascini .M, Pepe .A, Diletti .G, Compagnone .D, Food Chemistry, 2004, 651-656.
- [3] Mostafalou, S.; Abdollahi, M. Pesticides and human chronic diseases: Evidences, mechanisms, and perspectives. *Toxicol. Appl. Pharmacol.* 2013, 268, 157-177.
- [4] Fernández.H, Arévalo.F.J, Granero.A.M, Robledo.S.N, Díaz.C.H, Riberi.W.I, Zon.M.A, Electrochemical Biosensors for the Determination of Toxic Substances Related to Food Safety Developed in South America: Mycotoxins and Herbicides, 2017
- [5] McKinney, J.D.; Richard, A.; Waller, C.; Newman, M.C.; Gerberick, F. *Toxicol. Sci.* 2000, 56, 8-17.
- [6] Supratik K., Kunal R., *Chemosphere in press.* (2012).
- [7] Dharmendra K. Yadav Feroz Khan. QSAR, docking and ADMET studies of camptothecin derivatives as inhibitors of DNA topoisomerase-I, Volume 27, Issue 1-2 January-February 2013 Pages 21-33.
- [8] P.R. Duchowicz, D.O. Bennardi, E.V. Ortiz & N.C. Comelli, SAR AND QSAR IN ENVIRONMENTAL RESEARCH 2021, VOL. 32, NO. 5, 395-410, QSAR models for insecticidal properties of plant essential oils on the housefly (*Musca domestica* L.)
- [9] S. Raimondo & M.G. Barron (2019): Application of Interspecies Correlation Estimation (ICE) models and QSAR in estimating species sensitivity to pesticides, SAR and QSAR in Environmental Research, DOI: 10.1080/1062936X.2019.1686716
- [10] *J. Chem. Inf. Comput. Sci.* 2004, 44, 4, Development of QSAR Models To Predict and Interpret the Biological Activity of Artemisinin Analogues, Rajarshi Guha and Peter C. Jurs
- [11] Oshiro.C, Bradley.EK, Eksterowicz.J, Evensen.E, Lamb.M.L, Lancot.JK, Putta.S, Stanton.R and Grootenhuis.DJ, Performance of 3D-Database Molecular Docking Studies into Homology Models (2003).
- [12] Gilad.Y, Senderowitz.H, Docking Studies on DNA Intercalators (2014), *Journal of Chemical Information and Modeling*, 2014, 54, 96-107.
- [13] Kearsley SK, Underwood DJ, Sheridan RP, Miller MD, Flexibase: a way to enhance the use of molecular docking methods (1994). *Journal of Computer-Aided Molecular Design.* 8 (5): 565-82.
- [14] Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS, Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy (2004), *Journal of Medicinal Chemistry.* 47 (7): 1739-49.
- [15] Gramatica Paola, Chirico Nicola, Papa Ester, Cassani Stefano, and Kovarich Simona. QSARINS: A New Software for the Development, Analysis, and Validation of QSAR MLR Models (2013).
- [16] Furey, T. S.; Cristianini, N.; Duffy, N.; Bednarski, D. W.; Schummer, M. and Haussler, D. Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. *Bioinformatics* 2000, 16, 906-914.
- [17] Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. Support Vector Machines for Predicting HIV Protease Cleavage Sites in Protein. *J. Comput. Chem.* 2002, 23, 267-274.
- [18] Hua, S. J.; Sun, Z. R. Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach. *J. Mol. Biol.* 2001, 308, 397-407.
- [19] Adada .A, Larifb .M, Hmammouchia .R, Taghikia .A, Bouachrinec M, Lakhlifia .T, DFT-QSAR models to predict biological activities of carbamates, s-alkylcarbamathioates, ureas and chloroacetamides derivatives, 2013.
- [20] S. N. Sivandam, and S. N. Deepa, *Introduction to Genetic Algorithm*, Springer-Verlag, 2008.
- [21] Andries P. Engelbrecht, *Computational Intelligence An Introduction*, John Wiley & Sons Ltd, 2007, Second Edition
- [22] Kier, L. B.; Hall, L. H. An Atom-Centered Index for Drug QSAR Models. In *Advances in Drug Design*; Testa, B., Ed.; Academic Press: 1992; Vol. 22.

- [23] Broto, p, moreau. G. and vandycke, c. (1984). molecular structures: perception, autocorrelation descriptor and SAR studies. Autocorrelation descriptor. Eur. j. med. chem., 19, 66-70.
- [24] Roberto Todeschini, Viviana Consonni, Handbook of descriptprs-Weinheim- Newyork, Wiley-VCH, 2000.
- [25] Zefirov. N.S. Palyulin. V.A., Skvortsova, M.I. and Baskin, I.I. (1995). Inverse Problems in OSAR. In QSAR and Molecular Modelling: Cocepts, Computational Tools and Biological Applications (Sanz, F., Giraldo, J. and Manaut, F., eds.), Prous Science, Barcelona (Spain), pp.
- [26] Elham Baher, Naser Darzi, Iranian Journal of Pharmaceutical Research (2017), 16 (1): 146-157
- [27] Rücker, Christoph, Gerta Rücker, and Markus Meringer. "y-Randomization and its variants in QSPR/QSAR." Journal of chemical information and modeling 47.6 (2007): 2345-2357

Machine Learning-Based QSAR and Molecular Docking Analysis for Toxicity Prediction of Pesticides: A Tool for Environmental Risk Management

Fateme Ezadi¹, Elham Baher^{2*}, sajjad gharaghani³

1. Department of chemistry, faculty of science, Golestan University, Gorgan, Iran

2. Department of chemistry, faculty of science, Golestan University, Gorgan, Iran

3. Department of Bioinformatics, Laboratory of Chemoinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran

e.baher@gu.ac.ir

Receipt: 2026/05/06

Accepted: 2026/05/14

Abstract

In the present study, the lethal dose (LD50) of some pesticides was predicted using quantitative structure-activity (QSAR) studies. For this purpose, after drawing the molecular structure, molecular descriptors were calculated and LD50 by multiple linear regression (MLR) and support vector machine (SVM) was predicted. Genetic algorithm (GA) and MLR-stepwise were used to select the related descriptors. By comparing the obtained statistical results, the MLR-SVM model ($Q^2 = 0.498$ and $SPRESS = 0.047$) was best model with $R = 0.99$ and $SE = 0.014$ and $R = 0.718$ and $SE = 0.082$ for the training and test set, respectively. Then, molecular docking was performed to investigate the ligand-protein interactions. For this part, the pesticides with the highest ($LD50 = 2.68$) and the lowest ($LD50 = 2.21$) toxicity were selected. Based on the obtained results, compounds that tend to form more hydrophobic bonds are causing more toxicity, and also the presence of some elements such as fluoride in their molecular structure increases their toxicity.

Keywords: QSAR studies, Molecular docking, Genetic Algorithm, Support vector machine