مجله شیمی کوانتومی و اسپکتروسکوپی
۱۴:۴۵ (۱۴۰۳) ۱۵-۳۰
Journal of Quantum Chemistry and Spectroscopy (JQCS)

مطالعات ارتباط کمی ساختار-فعالیت آگونیست های SST2 با استفاده از رگرسیون خطی چندگانه و ماشین بردار پشتیبان

مهدی نکوئی*[۱]، سپهر محمدپور[۲]، اسلام پوربشیر[۲]

[۱]گروه شیمی، واحد شاهرود، دانشگاه آزاد اسلامی، شاهرود، ایران

[۲]گروه شیمی، دانشکده علوم، دانشگاه محقق اردبیلی، اردبیل

تاریخ ثبت اولیه:۱۴۰۲/۱۲/۰۵، تاریخ دریافت نسخه اصلاح شده:۱۴۰۳/۰۲/۲۳ ، تاریخ پذیرش قطعی:۱۴۰۳/۰۳/۱۲

چکیده

مطالعه ارتباط کمی-ساختار فعالیت(QSAR) جهت پیش‌بینی فعالیت دارویی(EC_{50}) برخی از مشتقات 3*H*-pyrido[2,3-d]pyrimidin-4-ones به عنوان آگونیست های SST2 با استفاده از روشهای رگرسیون خطی چندگانه (MLR) و ماشین بردار پشتیبان (SVM) انجام شد. در ابتدا ساختار ترکیبات مورد نظر توسط نرم افزار هایپرکم رسم و بهینه گردید. ترکیبات رسم شده جهت محاسبه توصیف‌کننده‌های مولکولی به نرم افزار دراگون وارد و تعداد ۱۴۸۱ توصیف کننده برای هر مولکول محاسبه شد. جهت انتخاب مناسب ترین توصیف کننده‌ها از روش رگرسیون مرحله‌ایی استفاده گردید. توصیف‌کننده‌های انتخاب شده Mor24v, G2e, G3e, F08[N-O], F08[N-F] می‌باشند که تاثیر الکترونگاتیوی، فاصله هندسی جفت اتم های دوتایی و حجم واندروالس را نشان می دهند. پس از انتخاب مناسب ترین توصیف کننده‌ها، از دو روش MLR و SVM جهت مدلسازی و پیش‌بینی فعالیت دارویی ترکیبات، استفاده شد. عملکرد هر مدل توسط چندین پارامتر آماری مورد ارزیابی قرار گرفت. نتایج بدست آمده نشان از برتری روش SVM نسبت به MLR دارد.

واژه های کلیدی : ارتباط کمی ساختار- فعالیت، مشتقات 3H-pyrido[2,3-d]pyrimidin-4-ones - رگرسیون خطی چند گانه، ماشین بردار پشتیبان

۱. مقدمه

با توجه به گسترش انواع بیماری ها، یکی از موضوعات مورد توجه و با اهمیت، طراحی، سنتز و تولید دارو می باشد. روندی که در گذشته منجر به کشف و توسعه داروهای جدید می‌شد به روش آزمون و خطا صورت می‌گرفت که روشی وقت گیر و هزینه‌بر

***عهده دار مکاتبات:** مهدی نکوئی

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

تلفن:۰۲۳۳۲۳۹۴۲۸۹ **پست الکترونیک:**E-mail:m_nekoei1356@yahoo.com

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 مهم ترین 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. ظهور علم کمومتریکس توانسته راه حلی برای رفع این مشکلات باشد[۴-۱]. یکی از مهم‌ترین زمینه‌های کاربرد روش‌های کمومتریکس، مطالعه ارتباط بین خواص مولکول‌ها با ویژگی‌های ساختاری آن‌هاست. این نوع مطالعات که با عنوان ارتباط کمی ساختار- فعالیت (QSAR)[1] معروف شده‌اند، به بررسی نحوه ارتباط بین خواص مختلف مولکول‌ها با مشخصات ساختاری و ذاتی آن‌ها می‌پردازند[۱۰-۵]. بررسی ساختار شیمیایی و فعالیت ترکیبات، پیش بینی فعالیت ترکیبات جدید را بر اساس اطلاعات مرتبط به ساختار شیمیایی آن‌ها امکان‌پذیر می‌سازد. EC_{50}[2] غلظتی موثری از یک ترکیب دارویی است که منجر به ۵۰٪ اثر مهار رشد بیماری می‌گردد. روشهای مختلفی برای اندازه گیری این پارامتر وجود دارد. اما از آنجاییکه این اندازه‌گیری‌ها، وقت گیر و هزینه بر می باشند، استفاده از روش‌هایی برای تخمین این پارامتر(EC_{50}) ضروری به نظر می‌رسد. برای پیش بینی فعالیت دارویی ترکیبات شیمیایی، روش ارتباط کمی ساختار - فعالیت (QSAR) روش مطمئن و مناسبی می باشد. QSAR ارتباط ریاضی بین ساختار و فعالیت دسته‌ایی از ترکیبات دارویی را توصیف می‌کند. در سالیان اخیر، روند رو به افزایش مقالات منتشره در منابع علمی بر اساسQSAR، دلالت بر جایگاه منحصر به فرد این دیدگاه در شیمی نظری و تعمیق بینش دانشمندان در فهم و توجیه آن دارد[۱۴-۱۱]. از جمله روشهایی که جهت مدل سازی و پیش‌بینی فعالیت ترکیبات داروئی مورد استفاده قرار می‌گیرند، می‌توان به روش‌های خطی از جمله رگرسیون خطی چندگانه، کمترین مربعات جزیی و روش‌های غیرخطی مانند شبکه‌های عصبی مصنوعی، ماشین بردار پشتیبان، روش های فازی عصبی و... اشاره نمود[۱۹-۱۵].

هدف از این تحقیق ارائه‌ی مدل‌های مناسب جهت پیش‌بینی فعالیت داروئی برخی مشتقات 3H-pyrido[2,3-d]pyrimidin-4-ones به عنوان آگونیست های SST2 با استفاده از روش های SW- MLR و SW-SVM می‌باشد.

۲. روش‌های محاسباتی

۲-۱. انتخاب سری داده‌ها

سری داده‌ها شامل فعالیت دارویی۲۸ ترکیب از مشتقات 3H-pyrido[2,3-d]pyrimidin-4-ones می‌باشد[۲۰]. فعالیت دارویی این ترکیبات به صورت EC_{50} گزارش شده است. EC_{50} عبارتست از غلظتی موثری از یک ترکیب دارویی است که منجر به ۵۰٪ اثر مهار رشد بیماری می‌گردد. این مقادیر به مقیاس لگاریتمی(pEC_{50}) تبدیل و مورد استفاده قرار گرفت. در این کار این ترکیبات به صورت تصادفی به دو گروه سری آموزش و تست تقسیم شدند، سری آموزش شامل ۲۲ مولکول(۸۰٪) و سری تست شامل ۶

[1] Quantitative structure activity relationship
[2] Effective concentration 50%

مولکول(٪۲۰) می‌باشد. مقادیر pEC_{50} به عنوان متغیر وابسته و توصیف کننده‌ها 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.

۲-۲. رسم و بهینه سازی ساختار مولکول‌ها

در ابتدا ساختار مولکولی ترکیبات در نرم افزار HyperChem ترسیم شد. سپس با احتساب اتم های هیدروژن، ساختار سه بعدی ترکیبات با استفاده از روش نیمه تجربی کوانتمیAM1 بهینه گردید. این بهینه سازی تا زمانی ادامه یافت که جذر میانگین مربعات گرادیان انرژی به ۰/۰۰۱ کیلوکالری بر مول رسید. با استفاده از این نرم افزار می توان اطلاعات فراوانی نظیر زوایای پیوندی، طول پیوندها، زوایای پیچش، بار اتم ها، انرژی تشکیل مولکول و... 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 ویژگی ها و ....

۲-۳. محاسبه توصیف کننده‌های مولکولی

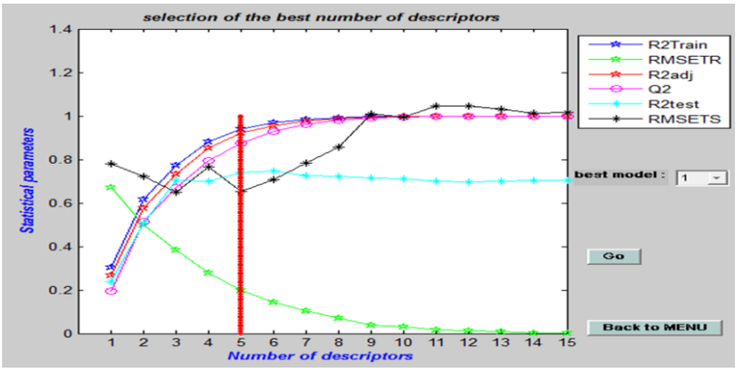
توصیف کننده‌ها مقادیر عددی هستند که ویژگیهای مختلف مولکول را نشان می‌دهند. توصیف‌کننده‌های مولکولی نتیجه نهایی یک استدلال و روش ریاضی است که اطلاعات شیمیایی را به رمز تبدیل می‌کند و آنها را به صورت یک نماد نشان می‌دهد که ارائه دهنده یک ویژگی مولکول به صورت یک عدد مفید می‌باشد. هر یک از این توصیف کننده‌ها، اطلاعات خاصی از مولکول را در اختیار می‌گذارند. توصیف‌کننده‌های مولکولی مختلفی برای اهداف گوناگون به کار برده شده‌اند. اختلاف این توصیف‌کننده‌ها در پیچیدگی اطلاعات رمزگزاری شده و زمان مورد نیاز برای محاسبه می‌باشد. اولین نوع توصیف‌کننده‌ها، توصیف‌کننده‌های توپولوژی می‌باشند. این توصیف‌کننده‌ها از روی گراف‌های مولکولی بدست می‌آیند و جزء ساده‌ترین نوع توصیف‌کننده‌ها می‌باشند و به ساختار فضایی مولکول ارتباطی نداشته و تنها به نوع اتم، نوع پیوندها و نحوه ارتباط اتم‌ها به یکدیگر وابسته است. از جمله این توصیف کننده‌ها می‌توان به تعداد اتم‌ها، شاخص‌های ارتباطی مولکولی و وزن مولکولی و ... اشاره کرد. دومین نوع توصیف کننده‌ها، توصیف کننده‌های هندسی است. این توصیف‌کننده‌ها با ساختار سه بعدی مولکول‌ها در ارتباط می‌باشند. برای محاسبه این توصیف‌کننده‌ها ابتدا می‌بایست ساختار فضایی مولکول‌ها بهینه شود. برخی از این توصیف کننده‌ها عبارتند از: حجم مولکولی، مساحت سطح و مساحت سطح در دسترس حلال. توصیف‌کننده‌های شیمی کوانتومی از جمله توصیف‌کننده‌های دیگری هستند که با استفاده از بهینه‌سازی نیمه تجربی ساختار مولکول‌ها در نرم‌افزارهای مختلف بدست می‌آیند. از جمله این توصیف کننده‌ها می‌توان به انرژی بالاترین تراز اشغال شده، انرژی پایین ترین تراز اشغال نشده، بار و الکترونگاتیویته اتم‌ها و ... اشاره کرد. توصیف‌کننده‌های دیگر، توصیف کننده های فیزیکوشیمیایی هستند که این توصیف‌کننده‌ها بیانگر بعضی از خواص فیزیکوشیمیایی مولکول‌ها می‌باشند که به ساختار مولکول وابستگی شدیدی نشان می‌دهند. از قبیل: ضریب تقسیم آب-اکتانول، ویسکوزیته، میزان

حلالیت ترکیبات در آب، شکست مولکولی، نقطه ذوب و نقطه جوش. توصیف‌کننده‌های ارتباطی مولکولی نیز از جمله توصیف کننده‌های مهم دیگری هستند که اطلاعاتی از جمله اندازه و ساختار مولکول، مرتبه شاخه‌دار شدن و نحوه ارتباط اتم‌ها در مولکول را بیان می‌کنند[۲۱].

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

۲-٤. انتخاب مناسب‌ترین توصیف‌کننده‌ها

جهت انتخاب مناسب ترین توصیف کننده ها از روش رگرسیون مرحله ایی[1] استفاده شد. در روش رگرسیون مرحله ای، 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 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 مرحله‌ایی 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[2](خطای جذر میانگین مربعات)کاهش می‌یابد. اما بدلیل پیچیدگی مدل، نمی‌توانیم تعداد زیادی توصیف کننده را جهت مدل‌سازی انتخاب کنیم. بدین منظور و جهت انتخاب تعداد توصیف کننده‌های مناسب، نمودار پارامترهای مختلف آماری برحسب تعداد توصیف کننده‌ها رسم گردید که در شکل۱ نشان داده شده است. بر طبق این شکل، تعداد ۵ توصیف کننده به عنوان توصیف کننده‌هایی که بیشترین ارتباط را با فعالیت دارویی(pEC_{50}) دارند، انتخاب شدند. این ۵ توصیف کننده به همراه مفهوم و نوع آنها در جدول۱ ارائه شده است.



شکل۱. نمودار پارامتر آماری مختلف برحسب تعداد توصیف کننده‌ها

[1] Stepwise
[2]Root-mean-square-error

جدول۱. توصیف‌کننده‌های انتخاب شده توسط رگرسیون خطی چندگانه مرحله به مرحله

نشانه توصیف کننده	مفهوم توصیف کننده	نوع توصیف کننده
Mor24v	3D MoRSE-signal 24/ weighted by atomic van der Waals volumes.	3D MoRSE
G2e	2st component symmetry directional WHIM index/ weighted by atomic Sanderson electronegativities.	WHIM descriptors
G3e	3st component symmetry directional WHIM index/ weighted by atomic Sanderson electronegativities.	WHIM descriptors
F08[N-O]	Frequency of N – O at topological distance 8	2D Atom Pairs
F08[N-F]	Frequency of N – F at topological distance 8	2D Atom Pairs

۲-۵. ارزیابی توصیف کننده ها

به منظور ارزیابی توصیف‌کننده‌های انتخاب شده مبنی بر مستقل بودن از همدیگر ماتریس همبستگی توصیف‌کننده‌های انتخاب شده در جدول۲ آورده شده است. همانطور که در این جدول مشاهده می‌شود ضریب همبستگی بین توصیف‌کننده‌های انتخاب شده همگی کمتر از ۰/۳۱۰ می‌باشد. لذا نتایج جدول نشان می‌دهد که بین توصیف‌کننده‌های انتخاب شده همبستگی چندانی وجود نداشته و توصیف‌کننده‌های تقریبا مستقل از هم هستند.

جدول۲. ماتریس ضرایب همبستگی توصیفگرهای انتخاب شده

	Mor24v	G2e	G3e	F08[N-O]	F08[N-F]
Mor24v	1				
G2e	0.171	1			
G3e	-0.234	-0.183	1		
F08[N-O]	0.288	0.114	-0.310	1	
F08[N-F]	0.031	0.281	0.189	0.204	1

۲-۶. ماشین بردار پشتیبان

ماشین بردار پشتیبان یکی از روش‌های یادگیری تحت نظارت است که هم برای دسته‌بندی و هم رگرسیون قابل استفاده است. این روش توسط وپنیک[1] بر پایه تئوری یادگیری آماری بنا نهاده شده است. ماشین بردار پشتیبان روشی برای طبقه‌بندی دوتائی در

[1] Vapnic

فضای ویژگی‌های دلخواه است و از این رو روشی مناسب برای مسائل پیش بینی به شمار می رود[۲۲]. ماشین بردار پشتیبان در اصل یک دسته بندی کننده دو کلاسه است که کلاس‌ها را توسط یک مرز خطی از هم جدا می‌کند. در این روش نزدیکترین نمونه‌ها به مرز تصمیم گیری را بردارهای پشتیبان می‌نامند. این بردارها معادله مرز تصمیم گیری را مشخص می‌کنند[۲۳].

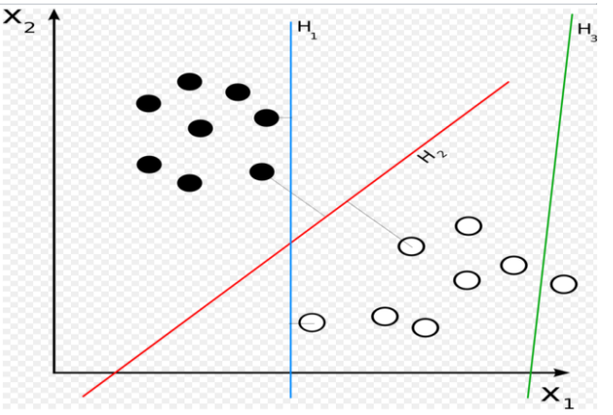
در سال ۱۹۹۶، واپنیک و همکارانش، نسخه‌ای از SVM را پیشنهاد دادند که به جای طبقه‌بندی، عمل رگرسیون را انجام می‌دهد. این مورد به Support Vector Regression یا SVR معروف است. همانند SVM در این مدل نیز از تابع کرنل و ابرپارامتر C استفاده می‌شود[۲۴].

بردارهای پشتیبان به زبان ساده، مجموعه‌ای از نقاط در فضای n بعدی داده‌ها هستند که مرز دسته‌ها را مشخص می‌کنند و مرزبندی و دسته بندی داده‌ها براساس آنها انجام می‌شود و با جابجایی یکی از آنها، خروجی دسته بندی ممکن است تغییر کند. در فضای دوبعدی، بردارهای پشتیبان، یک خط، در فضای سه بعدی یک صفحه و در فضای n بعدی یک ابر صفحه را شکل خواهند داد. ماشین بردار پشتیبان، یک دسته بند یا مرزی است (شکل۲) که با معیار قرار دادن بردارهای پشتیبان، بهترین دسته بندی و تفکیک بین داده ها را برای ما مشخص می کند[۲۴].

برای مجموعه داده‌های آزمایشی شامل n عضو(نقطه) رابطه زیر برقرار می‌باشد:

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p,\ y_i \in \{-1, 1\}\}_{i=1}^{n} \qquad (۱)$$

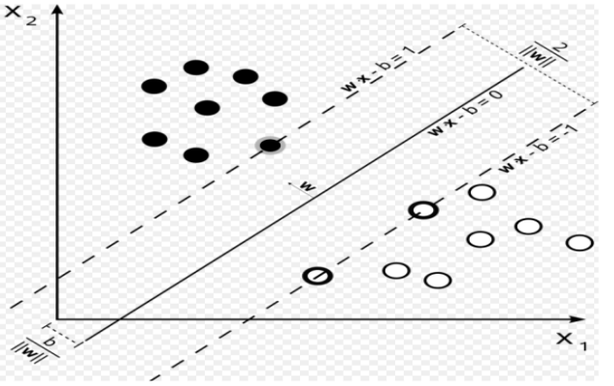
به ازای داده‌های موجود، تعداد زیادی مرزبندی می‌توانیم داشته باشیم که سه تا از این مرزبندی‌ها در زیر نمایش داده شده است.



شکل۲. انواع مرز در تفکیک و دسته بندی داده ها

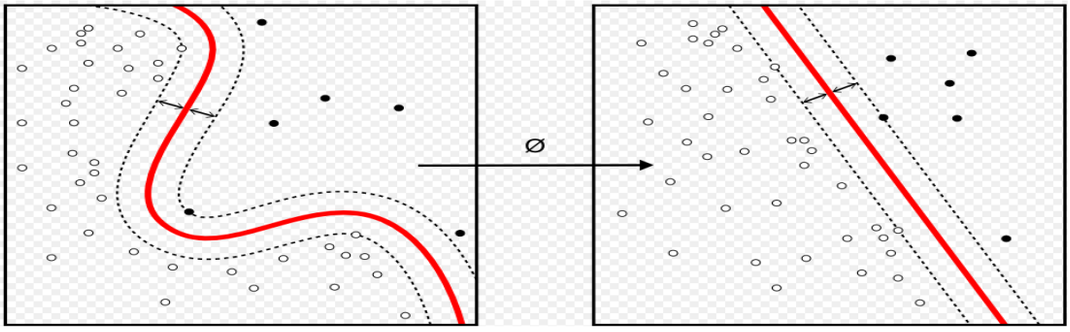
یک راه ساده برای بدست آوردن بهترین مرزبندی و ساخت یک دسته بند بهینه، محاسبه فاصله‌ی مرزهای به دست آمده با بردارهای پشتیبان هر دسته (مرزی ترین نقاط هر دسته یا کلاس) و در نهایت انتخاب مرزیست که از دسته‌های موجود، مجموعاً بیشترین فاصله را داشته باشد(شکل۳). این عمل تعیین مرز و انتخاب خط بهینه (در حالت کلی، ابر صفحه مرزی) به راحتی با انجام محاسبات ریاضی نه چندان پیچیده قابل پیاده سازی است. به عبارت دیگر، اگر داده‌های آموزشی جدایی پذیر خطی باشند، ما می‌توانیم دو ابر

صفحه در حاشیه نقاط به‌طوری‌که هیچ نقطه مشترکی نداشته باشند، در نظر بگیریم و سپس سعی کنیم، فاصله آن‌ها را ماکسیمم کنیم[۲۲].



شکل۳. نمایش دو ابرصفحه در حاشیه نقاط دو مجموعه از داده ها و بدست آوردن بهترین بردار پشتیبان

اگر داده‌ها به صورت خطی قابل تفکیک باشند، الگوریتم فوق می‌تواند بهترین بردار یا ابرصفحه را برای تفکیک داده‌ها ایجاد کند اما اگر داده‌ها به صورت خطی توزیع نشده باشند نیاز داریم داده‌ها را به کمک یک تابع ریاضی (Kernel functions) به یک فضای دیگر ببریم (نگاشت کنیم) که در آن فضا، داده‌ها تفکیک پذیر باشند و بتوان SVM آنها را به راحتی تعیین کرد(شکل۴). ماشین های بردار پشتیبان برای حل مسائل غیرخطی، ابعاد مسئله را از طریق توابع کرنل تغییر می‌دهند. انتخاب کرنل برای SVM به حجم داده‌های آموزشی و ابعاد بردار ویژگی بستگی دارد. به عبارت دیگر، بایستی با توجه به این پارامترها تابع کرنلی را انتخاب نمود که توانایی آموزش برای ورودی‌های مساله را داشته باشد. در عمل چهار نوع کرنل خطی، کرنل چند جمله ای، کرنل تانژانت هیپربولینک و کرنل RBF به کار گرفته می‌شوند. در جدول ۳ معادلات برخی از کرنل‌های رایج ارائه شده اند[۲۲].



شکل٤. استفاده از تابع کرنل برای تغییر فضای داده ها

جدول ۳. توابع کرنل رایج در ماشین های بردار پشتیبان

نوع تابع	تابع کرنل
خطی	$K(x_i, x_j) = x_i^T, x_j$
چند جمله ای	$K(x_i, x_j) = (\gamma x_i^T . x_j + C)d$

تانژانت هیپربولیک	$K(x_i, x_j) = tanh(\gamma x_i^T . x_j + C)$
RBF	$K(x_i, x_j) = exp(-\gamma \lvert x_i - x_j \rvert 2)$

۳. نتایج و بحث

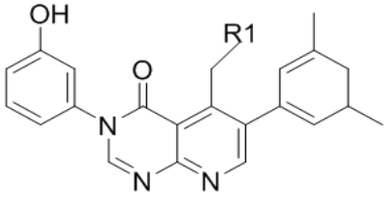
۳-۱. مدل‌سازی به روش رگرسیون خطی چندگانه(MLR)

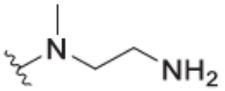
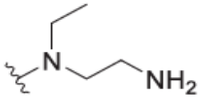
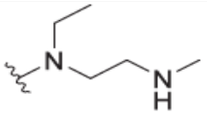
پس از انتخاب مناسب‌ترین توصیف‌کننده‌ها توسط روش مرحله‌ای، مرحله بعدی، ایجاد مدل میان توصیف‌کننده‌های انتخاب شده و pEC_{50} می‌باشد. بین توصیف کننده‌ها و فعالیت دارویی ترکیبات برای سری آموزش با استفاده از روش MLR رابطه زیر به عنوان مدل خطی بدست آمد:

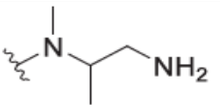
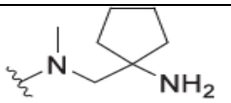
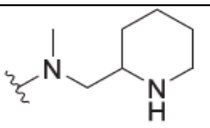
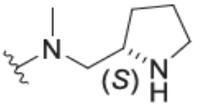
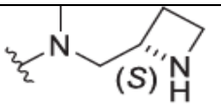
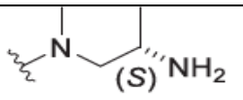
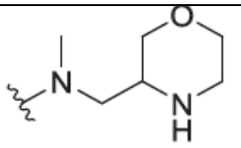
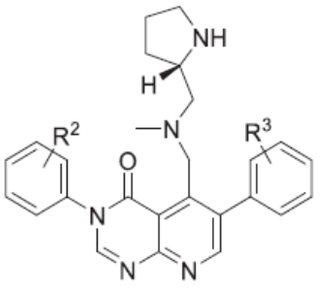
$$\mathbf{pEC_{50}} = 24.587 - 2.213(\mathbf{Mor24v}) - 56.438(\mathbf{G2e}) - 47.505(\mathbf{G3e}) + 0.467\ (\mathbf{F08[N\text{-}O]}) + 0.209\ (\mathbf{F08[N\text{-}F]})$$

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 (۴) آورده شده است. شکل(۵) نمودار مقادیر پیش بینی شده بر حسب مقادیر تجربی را نشان می دهد. در این شکل نزدیکی نتایج به خط راست قدرت پیش بینی مدل را نشان می دهد.

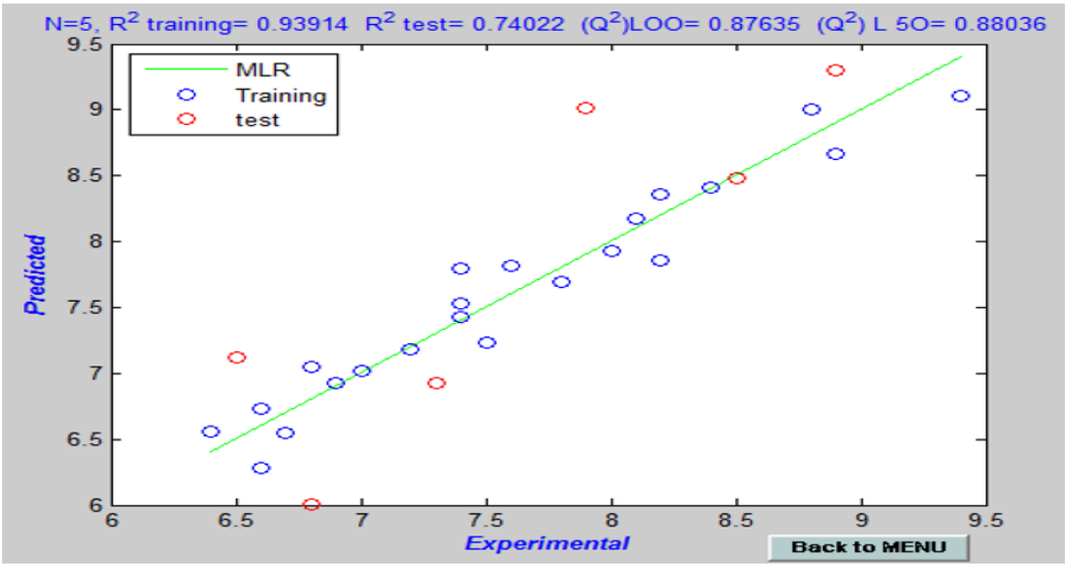
جدول٤. مقادیر تجربی و محاسبه شده pIC_{50} ترکیبات مختلف برای مجموعه‌های آموزشی و پیش‌بینی در مدل‌های SW-MLR, SW-SVM



Name	R1	pIC_{50}	SW-MLR	SW-SVM
M2		7.4	7.4	7.6
M4		7.8	7.7	7.6
M5		7.4	7.8	7.6

*M6			6.8	6.0	6.7
M7			6.9	6.9	7.1
M11			8.4	8.4	8.2
M12			8.2	7.9	8.0
M13			8.2	8.4	8.3
*M14			7.3	6.9	7.2
M15			6.4	6.6	6.6
					
	R^2	R^3			
M16	H	3-5-di-Me	6.6	6.7	6.8
M17	3-F	3-5-di-Me	6.8	7.1	7.0
M18	3-OMe	3-5-di-Me	7.2	7.2	7.2
M19	3-CONH2	3-5di-Me	7.4	7.5	7.5
M20	3-CONHMe	3-5-di-Me	6.7	6.5	6.9
M21	3-CH2OH	3-5-di-Me	6.6	6.3	6.8
M22	3-CN	3-5-di-Me	7.5	7.2	7.3

*M23	4-OH	3-5-di-Me	6.5	7.1	7.2
M24	3-OH-5-Me	3-5-di-Me	8	7.9	7.8
M25	3-F-5-OH	3-5-di-Me	8.9	8.7	8.7
M26	3-OH-5-CF3	3-5-di-Me	9.4	9.1	9.0
M29	3-OH	3-5-di-F	7	7.0	7.2
M30	3-OH	3-Cl-5-F	7.6	7.8	7.7
M31	3-OH	3-F-5-Me	8.1	8.2	8.0
*M32	3-OH	3-Cl-5-Me	8.9	9.3	8.5
*M33	3-OH	3-5-di-Cl	8.5	7.1	8.3
M34	3-OH	3-Cl-5-CH2OH	8.8	9.0	8.6
*M35	3-OH	2-F-3-5-di-Me	7.9	9.0	8.5



شکل۵. نمودار مقادیر پیش‌بینی شده pIC_{50} بر حسب مقادیر تجربی برای سری آموزش و تست به روش SW-MLR

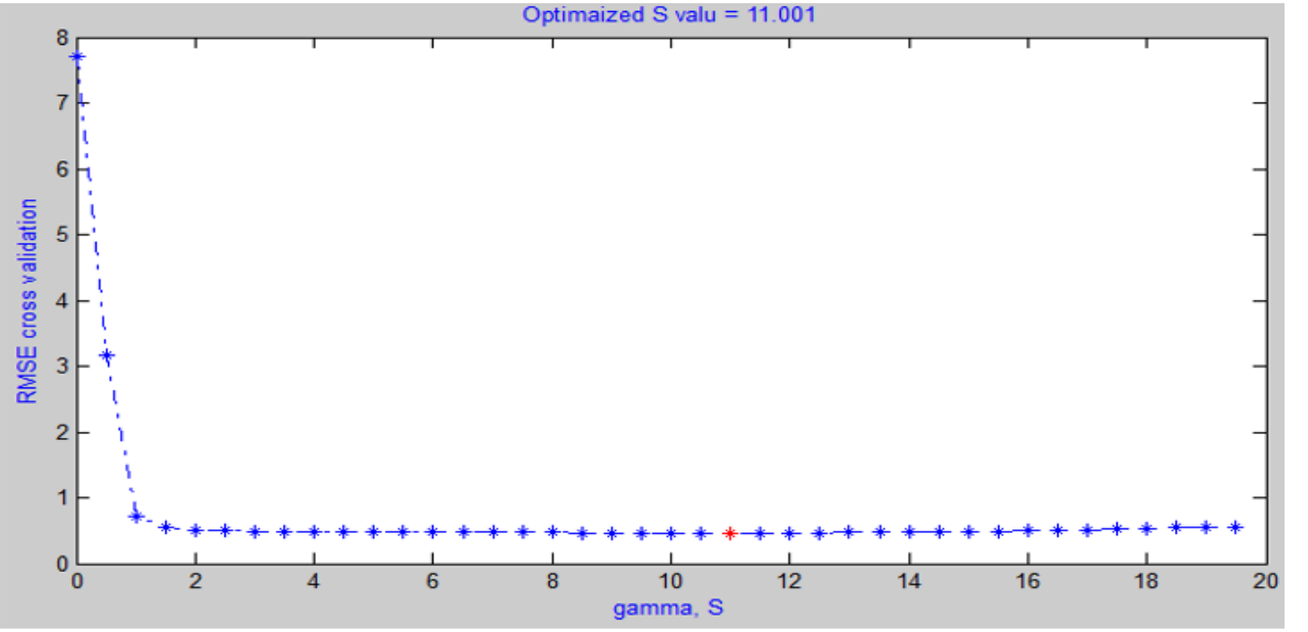
۳-۲. مدل‌سازی با استفاده از ماشین بردار پشتیبان

در مرحله دوم، جهت حصول نتایج بهتر از ماشین بردار پشتیبان برای ایجاد مدل و پیش‌بینی فعالیت دارویی ترکیبات موردنظر استفاده شد. در ابتدا پارامترهای مربوط به SVM بهینه گردید:

۱- نوع تابع کرنل[۱] ۲- پارامتر ظرفیت[۲] ۳- فاکتور حساسیت (سیگما[۳]) و ۴- اپسیلون یا گاما[۴]

[1] kernel function type
[2] Capacity parameter (C)
[3] Sensitive factor (ε)
[4] Gamma (eps)

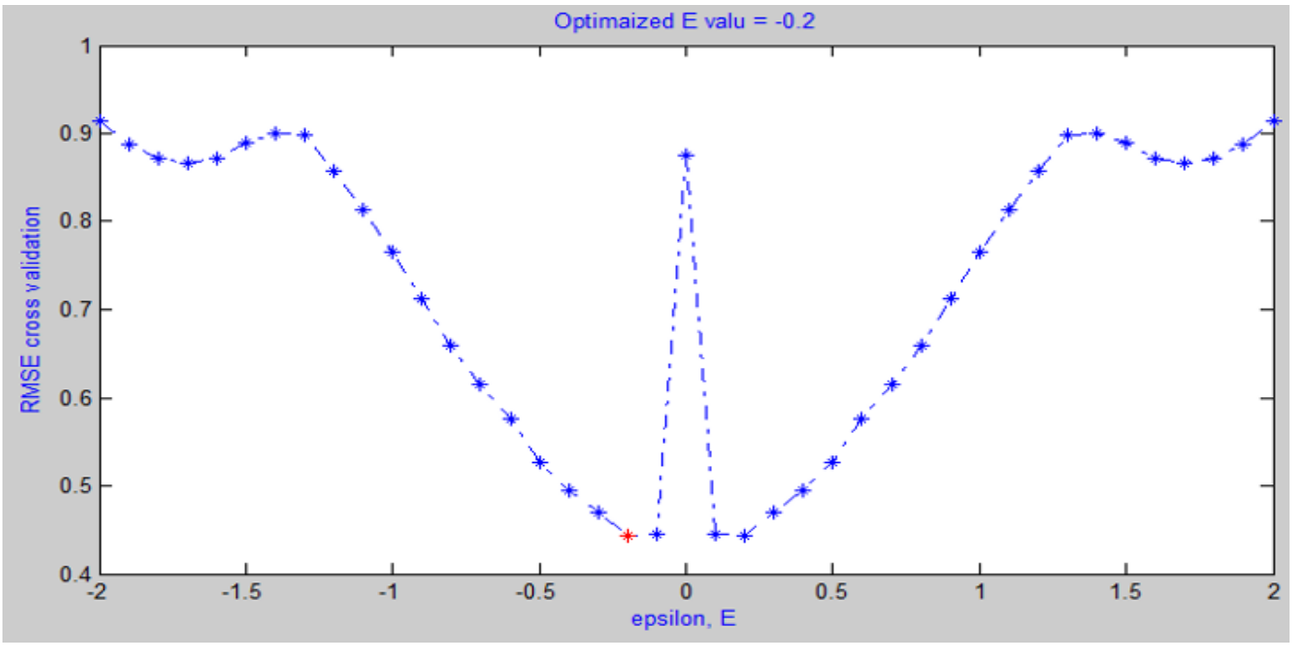
نوع تابع کرنل، نحوه توزیع نمونه‌ها را در فضا مشخص می‌کند. RBF یکی از توابعی است که به طور معمول استفاده می‌شود و نتایج خوبی نیز می‌دهد در این پروژه از RBF به عنوان تابع برای SVM استفاده گردید. بعلاوه پارامتر متناظر با نوع تابع یعنی Gamma که روی تعداد بردارهای پشتیبان تاثیر می‌گذارد نیز باید بهینه شود. تعداد بردارهای پشتیبان بر زمان آموزش مدل تاثیر می‌گذارد طوری که افزایش مقدار Gamma و در نتیجه تعداد بردار پشتیبان می‌تواند به افزایش زمان آموزش و همچنین Overfitting منجر شود. مقدار Gamma توانایی و قدرت SVM را در پیشگویی کنترل می‌کند. در شکل ۶ نمودار مقادیر متفاوت Gamma بر حسب RMSE نمایش داده شده است.



شکل ٦. نمودار تغییرات مقدار Gamma برحسب مقدار RMSE برای سری آموزش

همانطوری‌که از این شکل ملاحظه می‌شود مقدار Gamma از ۰ تا ۲۰ تغییر می یابد و از مقدار ۱۱ به بعد با افزایش Gamma مقدار RMSE کمی افزایش می یابد. بنابرین مقدار ۱۱ به عنوان نقطه بهینه برای Gamma انتخاب شد.

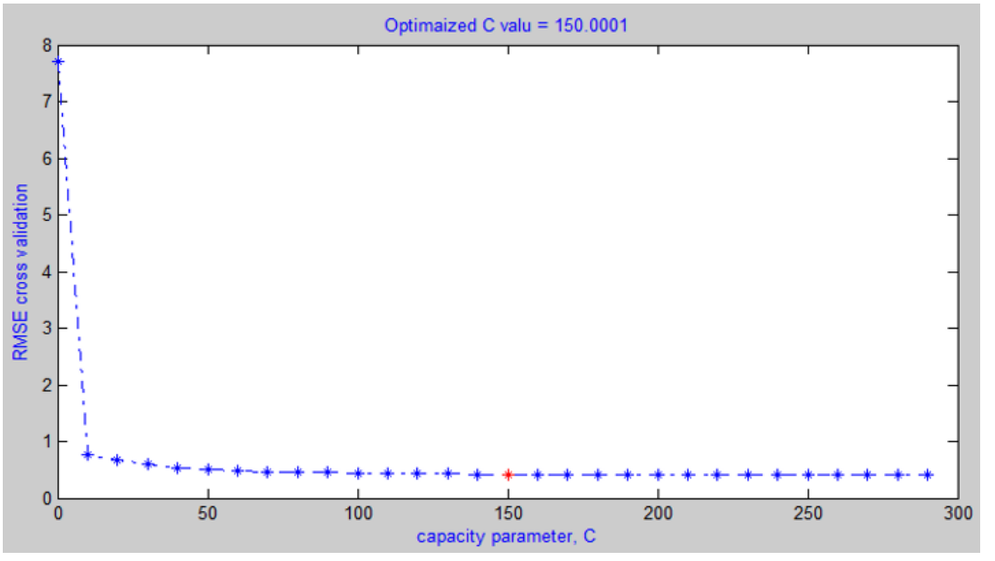
فاکتور حساسیت یکی دیگر از پارامترهایی است که باید بهینه شود. فاکتور حساسیت به نویزهای موجود در داده‌ها مربوط می‌شود که معمولاً ناشناس هستند. در شکل ۷ نمودار تغییرات RMSE برحسب epsilon (ε) نمایش داده شده است.



شکل ۷. نمودار تغییرات مقدار(ε) epsilon برحسب مقدار RMSE برای سری آموزش

با تغییر مقدار epsilon از ۲- تا ۲ مقدار بهینه ۰/۲- برای epsilon انتخاب شد.

و در نهایت پارامتر ظرفیت(C) باید بهینه شود اگر مقدار C پایین باشد یک پراکندگی در پیش‌بینی دیده خواهد شد و در بعضی اوقات با افزایش بیش از حد C، Overfitting رخ می‌دهد هرچند که مقدار زیادی C تاثیر چندانی روی پیش‌بینی ندارد ولی با این حال مقدار این پارامتر نیز باید بهینه شود. در شکل ۸ نمودار تغییرات(Capacity parameter (C بر حسب RMSE نمایش داده شده است.

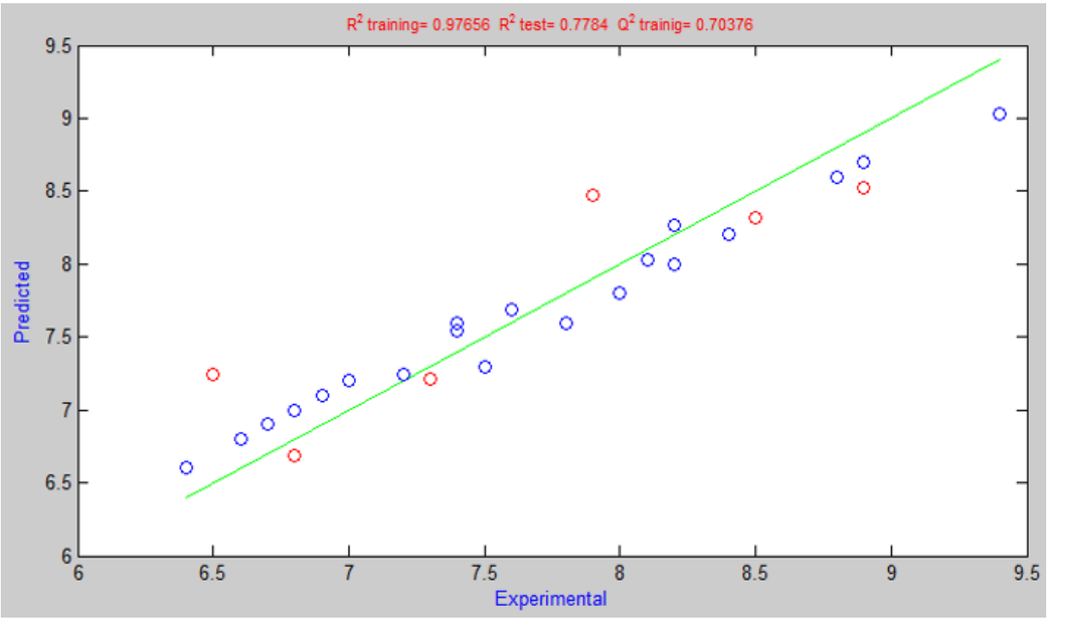


شکل ۸. نمودار تغییرات مقدار(C) Capacity parameter برحسب مقدار RMSE برای سری آموزش

با تغییر مقدار Capacity parameter از ۰ تا ۳۰۰ مقدار بهینه ۱۵۰ برای Capacity parameter انتخاب شد.

در مرحله آخر با استفاده از تمامی پارامترهای بهینه شده، مدل SVM ساخته شده و مقادیر فعالیت دارویی ترکیبات پیش بینی شد. با استفاده از مدل SVM بهینه شده مقادیر فعالیت دارویی ترکیبات مورد نظر در مجموعه آموزشی و پیش‌بینی مورد محاسبه قرار گرفت و در جدول۴ نشان داده شده است. همانطور که در این جدول مشاهده می‌شود SVM توانسته است پیش‌بینی‌های بسیار خوبی را برای مقادیر فعالیت دارویی ترکیبات مورد نظر نشان دهد.

مقادیر فعالیتهای دارویی محاسبه‌شده و تجربی برای ترکیبات براساس مدل SVM در دو مجموعه آموزشی و تست در شکل ۹ آورده شده است در این شکل میزان نزدیکی داده‌ها به خط راست قدرت پیش‌بینی مدل را نشان می‌دهد.



شکل ۹. نمودار مقادیر فعالیت دارویی محاسبه شده برای ترکیبات براساس مدل SVM در دو مجموعه آموزشی و تست برحسب مقادیر تجربی

۳-۳. نتایج پارامترهای آماری جهت مقایسه مدل‌های انتخاب شده

مطابق جدول ۵، سه پارامتر آماری، جهت ارزیابی توانایی پیش‌بینی مدل‌های ساخته شده به روش‌های SW-MLR و SW-SVM به کار گرفته شد. نتایج جدول نشان دهنده برتری روش SVM به روش MLR می‌باشد.

جدول ۵. پارامترهای آماری برای مدل های انتخاب شده

	RMSE		R^2		F	
	آموزش	تست	آموزش	تست	آموزش	تست
SW-MLR	۰/۱۹۸	۰/۶۵۲	۰/۹۳۹	۰/۷۴۰	۴۹/۳۸۰	۰/۰
SW-SVM	۰/۱۹۲	۰/۴۲۱	۰/۹۷۶	۰/۷۷۸	۳۶/۴۰۳	۰/۰

٤. نتیجه گیری

در این تحقیق از دو روش رگرسیون خطی چندگانه و ماشین بردار پشتیبان برای پیش‌بینی فعالیت ترکیبات دارویی(pEC_{50}) مشتقات 3*H*-pyrido[2,3-d]pyrimidin-4-ones به عنوان آگونیست های SST2 استفاده شد. از آنجایی که اندازه‌گیری فعالیت دارویی بیشتر ترکیبات به صورت تجربی با صرف هزینه، زمان و پیچیدگی زیاد همراه است، دست یابی به مقادیر pEC_{50} با روش-های تجربی مقرون به صرفه نیست. بنابراین، پیش‌بینی آن با استفاده از روش های محاسباتی از اهمیت بالایی برخوردار است. برای انتخاب توصیف‌کننده‌های مناسب از روش رگرسیون مرحله ایی استفاده شد. سپس این توصیف‌کننده‌ها برای مدل سازی خطی و غیرخطی مورد استفاده قرار گرفتند. نتایج نشان داد که از بین دو روش استفاده شده، روش SVM روش مناسب‌تری برای پیش‌بینی فعالیت این ترکیبات دارویی است. همچنین بررسی ارتباط توصیف‌کننده‌های وارد شده در مدل با اثر بازدارندگی یا pEC_{50} مورد بررسی قرار گرفت. با توجه به نتایج به دست آمده در مدل، توصیف‌کننده‌های انتخاب شده توسط روش رگرسیون مرحله‌ای شامل Mor24v, G2e, G3e, F08[N-O], F08[N-F] می‌باشند. این توصیف کننده ها تاثیر الکترونگاتیوی، فاصله هندسی جفت اتم های دوتایی و حجم واندروالس را نشان می دهند. نتایج مدل سازی نشان می دهد که مقدار pEC_{50} با الکترونگاتیوی و حجم واندروالس رابطه عکس دارد یعنی با کاهش الکترونگاتیوی و حجم واندروالس ترکیبات می‌توان pEC_{50} را افزایش داد. از طرف دیگر مدل ساخته شده نشان می دهد که فاصله هندسی جفت اتم های دوتایی نیتروژن- اکسیژن و نیتروژن- فلوئور با pEC_{50} رابطه مستقیم دارد یعنی با افزایش آن‌ها، pEC_{50} نیز افزایش می یابد. از این تحقیق استنباط می شود که با تغییر الکترونگاتیوی گروه های عاملی و تغییر حجم مولکولی ترکیبات، می توان ترکیباتی طراحی و سنتز کرد که فعالیت دارویی موثرتری داشته باشند.

٥. مراجع

[1] Daoui, O., Nour, H., Abchir, O., Elkhattabi, 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.

[2] Kubinyi, H. (1997). QSAR and 3D QSAR in drug design Part 1: methodology. *Drug discovery today*, *2*(11), 457-467.

[3] 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.

[4] Khan, A. U. (2016). Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug discovery today*, *21*(8), 1291-1302.

[5] 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.

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

[7] 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.

[8] 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.
[9] Rosell-Hidalgo, A., Young, L., Moore, A. L., & Ghafourian, T. (2021). QSAR and molecular docking for the search of AOX inhibitors: a rational drug discovery approach. *Journal of Computer-Aided Molecular Design*, *35*, 245-260.
[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] Ewees, A. A., Abualigah, L., Yousri, D., Algamal, Z. Y., Al-Qaness, M. A., Ibrahim, R. A., & Abd Elaziz, M. (2022). Improved Slime Mould Algorithm based on Firefly Algorithm for feature selection: A case study on QSAR model. *Engineering with Computers*, *38*(3), 2407-2421.
[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] 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.
[14] Triolascarya, K., Septiawan, R. R., & Kurniawan, I. (2022). QSAR Study of Larvicidal Phytocompounds as Anti-Aedes Aegypti by using GA-SVM Method. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *6*(4), 632-638.
[15] 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.
[16] Veerasamy, R. (2022). QSAR—an important in-silico tool in drug design and discovery. In *Advances in computational modeling and simulation* (pp. 191-208). Singapore: Springer Nature Singapore.
[17] Quadri, T. W., Olasunkanmi, L. O., Akpan, E. D., Fayemi, O. E., Lee, H. S., Lgaz, H., ... & Ebenso, E. E. (2022). Development of QSAR-based (MLR/ANN) predictive models for effective design of pyridazine corrosion inhibitors. *Materials Today Communications*, *30*, 103163.
[18] Abdolmaleki, A., & Ghasemi, J. B. (2019). Inhibition activity prediction for a dataset of candidates' drug by combining fuzzy logic with MLR/ANN QSAR models. *Chemical Biology & Drug Design*, *93*(6), 1139-1157.
[19] Iwaloye, O., Elekofehinti, O. O., Olawale, F., Chukwuemeka, P. O., Kikiowo, B., & Folorunso, I. M. (2023). Fragment-based drug design, 2D-QSAR and DFT calculation: Scaffolds of 1, 2, 4, triazolo [1, 5-a] pyrimidin-7-amines as potential inhibitors of Plasmodium falciparum dihydroorotate dehydrogenase. *Letters in Drug Design & Discovery*, *20*(3), 317-334.
[20] Fereidoonnezhad, 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.
[21] Consonni, V., & Todeschini, R. (2010). Molecular descriptors. *Recent advances in QSAR studies: methods and applications*, 29-102.
[22] Rodríguez-Pérez, R., & Bajorath, J. (2022). Evolution of support vector machine and regression modeling in chemoinformatics and drug discovery. *Journal of Computer-Aided Molecular Design*, *36*(5), 355-362.
[23] Janairo, J. I. B. (2023). Support vector machine in drug design. In *Cheminformatics, QSAR and Machine Learning Applications for Novel Drug Development* (pp. 161-179). Academic Press.
[24] Pramana, I. K. A. P. P., Septiawan, R. R., & Kurniawan, I. (2022). QSAR Study on Diacylgycerol Acyltransferase-1 (DGAT-1) Inhibitor as Anti-diabetic using PSO-SVM Methods. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *6*(5), 735-741.

Quantitative structure-activity relationship studies of SST2 agonists using multiple linear regression and support vector machine

Mehdi Nekoei[1]*, Sepehr Mohammadpour[2], Eslam pourbasheer[2]

[1]*Department of Chemistry, Shahrood Branch, Islamic Azad University, Shahrood, Iran*

[2]*Department of Chemistry, Faculty of Science, Mohaghegh Ardabili University, Ardabil*

Submited: 24 February 2024, Revised: 12 May 2024, Accepted: 01 Jun 2024

Abstract

A quantitative structure-activity relationship (QSAR) study to predict the pharmacological activity (EC_{50}) of some 3H-pyrido[2,3-d]pyrimidin-4-ones as SST2 agonists using multiple linear regression (MLR) methods and Support Vector Machine (SVM) was performed. At first, the structure of desired compounds was drawn and optimized by Hypercam software. The drawn compounds were entered into Dragon software to calculate molecular descriptors and the number of 1481 descriptors was calculated for each molecule. Stepwise regression method was used to select the most suitable descriptors. The selected descriptors are Mor24v, G2e, G3e, F08[N-O], F08[N-F], which show the effect of electronegativity, the geometric distance of pairs of binary atoms, and the van der Waals volume. After selecting the most appropriate descriptors, two methods, MLR and SVM, were used to model and predict the medicinal activity of the compounds. The performance of each model was evaluated by several statistical parameters. The obtained results show the superiority of SVM method over MLR.

Keywords: *Quantitative structure-activity relationship, 3H-pyrido[2,3-d]pyrimidin-4-ones, multiple linear regression, support vector machine*.

***Corresponding author :** Mehdi Nekoei

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

Tel: 02332394289 **E-mail:** m_nekoei1356@yahoo.com