توسعه مدل‌های رابطه‌های کمی ساختار–ویژگی برای پیش‌بینی حد اشتعال‌پذیری بالای ترکیب‌های آلی

الهام آل کثیر، فاطمه عباسی تبار*+

گروه شیمی، واحد مرودشت، دانشگاه آزاد اسلامی، مرودشت، ایران

چکیده: در این مطالعه، معادله کمی ساختار–ویژگی (QSPR) برای پیش‌بینی حد اشتعال‌پذیری بالای (UFL) ۵۸۸ ترکیب آلی شامل ترکیب‌های هیدروکربنی، ترکیب‌های هالوژن دار، الکل‌ها، اترها، استرها، آلدهیدها، کتون‌ها، اسیدها، آمین‌ها، آمیدها، نیتریل‌ها، و ترکیب‌های نیترو مورد مطالعه قرار گرفت. طیف گسترده‌ای از توصیف کننده‌ها برای هر مولکول محاسبه شد. الگوریتم کولونی مورچگان حافظه‌دار (M-ACO) همراه با برازش خطی چند متغیره (MLR) برای انتخاب بهترین زیر مجموعه توصیف کننده‌هایی که سهم چشمگیری در ویژگی UFL دارند به کار برده شد. تبدیل متغیرهای متفاوتی در هر دو متغیر وابسته و مستقل به منظور دستیابی به مدل‌های برازش خطی چندمتغیره با کارایی بهتر انجام شد. بهترین مدل یک مدل چهار متغیره بود که با استفاده از توصیف کننده‌های محاسبه شده به عنوان متغیر مستقل و لگاریتم مقدارهای UFL به عنوان متغیر وابسته به دست آمد. این مدل دارای گستره کاربردی بسیار وسیعی شامل مقدارهای UFL بین ۲/۷ تا ۱۰۰ درصد حجمی است. خطای آموزش مدل ۰/۱ واحد log UFL (R^2=۰/۸۰) و خطای پیش‌بینی ۰/۱۲ واحد log UFL (R^2=۰/۷۵) است. بنابراین، مدل از درستی خوبی برخوردار است و می‌تواند برای پیش‌بینی UFL گستره وسیعی از ترکیب‌های آلی به کار رود.

واژگان کلیدی: حد اشتعال‌پذیری بالا، الگوریتم کولونی مورچگان، تبدیل متغیر، روابط کمی ساختار–ویژگی، برازش خطی چند متغیره.

KEYWORDS: *Upper Flammability Limit, Ant Colony Algorithm, Variable Transformation, Quantitative Structure-Property Relationships, Multivariate Linear Regression.*

مقدمه

حدود اشتعال‌پذیری گستره‌ای از غلظت سوخت بر حسب درصد حجمی در ۲۹۸ کلوین است که در آن واکنش احتراق در حضور منبع افروزش خارجی انجام می‌شود. پایین‌تر از حد اشتعال‌پذیری پایین[۱] (LFL) سوخت کافی برای اشتعال وجود ندارد. بالاتر از حد اشتعال‌پذیری بالا[۲] (UFL)، مخلوط سوخت/هوا بسیار غنی از سوخت است و اکسیژن کافی برای سوختن وجود ندارد [۱–۳]. بنابراین، برای جلوگیری از آتش‌سوزی و انفجار یک گاز، دانستن LFL و UFL آن ضروری است.

* عهده‌دار مکاتبات

+Email: fabbasitabar@gmail.com

(۱) Lower Flammability Level

(۲) Upper Flammability Level

UFL یکی از مهم‌ترین اندیس‌های استفاده شده برای سنجیدن اشتعال‌پذیری مواد شیمیایی در صنایع شیمیایی است. دانستن مقدارهای UFL باعث بالا بردن ایمنی در طراحی فرایند و کارهای روزمره مانند ذخیره کردن یا بارگیری فرآورده‌های قابل اشتعال می‌شود. مقدارهای UFL تجربی، عمده‌ترین و دقیق‌ترین داده‌های UFL استفاده شده در تولید است اما UFL مطلق نیست و به فاکتورهای گوناگونی مانند ماهیت ترکیب‌ها، ساختار هندسی دستگاه، نوع و قدرت منبع افروزش، دما و فشار آزمایش، درجه اختلاط، غلظت اکسیژن و غلظت رقیق کننده‌ها بستگی دارد [۴,۵]. از این رو، مقدارهای UFL گزارش شده، بیش‌تر متناقض و گاهی وقت‌ها به طور کامل متفاوت هستند. افزون بر این، اندازه‌گیری تجربی دقیق UFL نیاز به دستگاه‌های استاندارد و چندین شرط دارد که در ASTM E۶۸۱ به آن‌ها اشاره شده است. همچنین، این کار وقت‌گیر و هزینه‌بر است و همیشه عملی نیست. بنابراین، نیاز به توسعه روش‌های محاسبه‌ای دقیق برای پیش‌بینی UFL ترکیب‌ها است [۶].

روش‌های گوناگونی برای پیش‌بینی UFL ترکیب‌های خالص در مقاله‌های منتشر شده موجود است. در پژوهش‌های اولیه از روش‌های مشارکت گروهی برای پیش‌بینی UFL ترکیب‌ها استفاده شد. با وجود این که کاربرد این روش‌ها ساده است و پارامترهای مورد نیاز را می‌توان تنها از ساختار شیمیایی هر ترکیب به دست آورد، ولی این مدل‌ها دارای معایبی نیز می‌باشند. به عنوان مثال، دامنه کاربرد این مدل‌ها محدود به مجموعه داده‌های مورد مطالعه است و ترکیب‌های شیمیایی با گروه‌های عاملی جدید که در ایجاد مدل استفاده نشده‌اند، توسط این مدل‌ها پیش‌بینی نمی‌شوند. افزون بر این، توانایی این مدل‌ها در تشخیص ترکیب‌های ایزومری ضعیف است [۷,۸].

به‌تازگی، برخی از مدل‌های همبستگی تجربی برای پیش‌بینی UFL ترکیب‌های آلی با همبستگی برخی ویژگی‌های فیزیکی–شیمیایی دیگر با UFL ایجاد شده‌اند [۹,۱۰]. قدرت پیش‌بینی این مدل‌ها بهتر ازمدل‌های مشارکت گروهی بود ولی این مدل‌ها نیز اشکال‌های مهمی داشتند. اول این که، کاربرد این مدل‌ها به ویژگی‌های فیزیکی–شیمیایی غیر متعارف نیاز دارد که در دسترس بودن یا نبوود آن‌ها ممکن است باعث محدودیت‌هایی در دامنه کاربرد آن‌ها شود. افزون بر این، درستی پیش‌بینی این مدل‌ها به طور کامل به درستی ویژگی‌های فیزیکی–شیمیایی مورد نیاز بستگی دارد.

در سال‌های اخیر، روش ارتباط کمی ساختار–ویژگی[۱] (QSPR) به طور گسترده‌ای برای پیش‌بینی اشتعال‌پذیری ترکیب‌های گوناگون استفاده شده است [۱۳–۱۱]. در این روش بین UFL ترکیب‌های خالص و ساختار مولکولی آن‌ها، که توسط انواع توصیف‌کننده‌های مولکولی نمایش داده می‌شوند، ارتباط برقرار می‌شود. توصیف‌کننده‌های مولکولی پارامترهای نظری هستند که می‌توانند با استفاده از الگوریتم‌های ریاضی شناخته شده، تنها از ساختارهای مولکولی محاسبه شوند. طیف گسترده‌ای از توصیف‌کننده‌ها مانند توصیف‌کننده‌های توپولوژیکی، توصیف‌کننده‌های هندسی، توصیف‌کننده‌های الکترواستاتیکی و توصیف‌کننده‌های شیمیایی کوانتومی برای آنالیز QSPR ارایه شده است [۱۴,۱۵]. از بین انبوه توصیف‌کننده‌های محاسبه شده، بهترین توصیف‌کننده‌ها برای به دست آوردن مدلی که ویژگی‌های ترکیب مورد نظر را با استفاده از ساختار مولکولی آن با بیش‌ترین درستی پیش‌بینی کند، انتخاب می‌شوند [۱۶–۲۰]. روش‌های گوناگونی مانند الگوریتم کولونی مورچگان [۲۶–۲۱]، ژنتیک الگوریتم، انتخاب مرحله به مرحله، الگوریتم طرح‌ریزی پی‌درپی برای انتخاب توصیف‌کننده استفاده می‌شوند. QSPR دارای برتری‌هایی است که می‌توان به موارد زیر اشاره کرد. تعداد توصیف‌کننده‌های استفاده شده در مقایسه با روش‌های مشارکت گروهی کم‌تر است [۲,۲۷]. توصیف‌کننده‌های استفاده شده دارای مفهوم فیزیکی هستند و می‌توان اطلاعات فیزیکی–شیمیایی که در ویژگی مورد نظر سهیم هستند را دنبال کرد. از آن جا که فقط توصیف‌کننده‌های تئوری که از ساختار مولکولی به دست می‌آیند را شامل می‌شود، مدل‌های QSPR را می‌توان به طور نظری برای هر ترکیب آلی و تعیین ویژگی‌های فیزیکی–شیمیایی گوناگون به کار برد [۳۰–۲۸].

روش QSPR به طور گسترده برای تعیین UFL ترکیب‌ها استفاده شده است [۲۶–۲۱]. *قراقزی* به کمک ژنتیک الگوریتم یک مدل MLR بر پایه ۵ توصیف‌کننده برای پیش‌بینی UFL ۸۶۵ ترکیب ایجاد نمود [۲]. مدل ایجاد شده دارای قدرت پیش‌بینی بسیار خوبی با $R^2_{\text{آموزش}}$ و $R^2_{\text{تست}}$ به ترتیب برابر با ۰/۹۲ و ۰/۹۳ بود. با این حال مدل یادشده تنها برای پیش‌بینی UFL در رنج ٪ ۲۵–۲/۵ کاربری داشت. *ونگ* و *همکاران* از QSPR برای پیش‌بینی UFL ۷۸ مخلوط دوتایی استفاده کردند [۱۷]. مخلوط‌های دوتایی از ۹ ترکیب خالص تشکیل شده بودند. از ماشین بردار پشتیبان برای مدل‌سازی بر پایه ۴ توصیف‌کننده استفاده شد. برای مدل به دست آمده $R^2_{\text{آموزش}}$ و $R^2_{\text{تست}}$ به ترتیب برابر بودند با ۰/۹۲ و ۰/۹۷. مدل یادشده تنها برای پیش‌بینی

(۱) Quantitative-Structure-Property Relationship

UFL در بازه ٪ ۴۳ -۱۰ کاربرد داشت. *یوان* و *همکاران* برای پیش‌بینی UFL ۷۹ ترکیب از فناوری‌های گوناگونی از جمله MLR، KNN، SVM و RF استفاده کردند [۳۱]. هرچند در این کار از ترکیب‌هایی با بازه گستره‌ای از UFL استفاده شده است (٪۸۰-۵) ولی پارامترهای آماری مدل‌های به دست آمده خوب نبودند. $R^2_{\text{آموزش}}$ و $R^2_{\text{تست}}$ برای مدل‌های MLR، KNN، SVM و RF به ترتیب برابر بودند با (۰/۶۶ و ۰/۵۳)، (۰/۶۸ و ۰/۰۶)، (۰/۷۸ و ۰/۶۶) و (۰/۹۲ و ۰/۵۶). هدف از این مطالعه، ایجاد یک مدل قابل اعتماد با دامنه کاربرد گسترده برای پیش‌بینی UFL یک مجموعه بزرگ ترکیب‌های آلی متنوع با استفاده از ساختار شیمیایی آن‌ها است. ترکیب سامانه کولونی مورچگان حافظه‌دار-برازش خطی چند متغیره برای انتخاب زیر مجموعه بهینه از توصیف‌کننده‌ها که سهم چشمگیری در ویژگی حد اشتعال‌پذیری بالا دارند استفاده شد. اعتبار و قدرت پیش‌بینی مدل نیز توسط فناوری‌های گوناگونی ارزیابی شد.

بخش تجربی

سری داده‌ها

مجموعه داده‌ها و اطلاعات برای این پژوهش از مجموعه‌ای متنوع از ۵۸۸ ترکیب آلی تشکیل شده است که همگی آن‌ها از منابع معتبر گرفته شده‌اند [۱۶]. این مجموعه شامل ترکیب‌های هیدروکربنی، ترکیب‌های هالوژن‌دار، الکل‌ها، اترها، استرها، آلدهیدها، کتون‌ها، اسیدها، آمین‌ها، آمیدها، نیتریل‌ها، و ترکیب‌های نیترو می‌باشد. مقدارهای UFL این ترکیب‌ها در بازه ۲/۷ تا ۱۰۰ درصد حجمی هستند. لیست کامل این ترکیب‌ها به همراه ساختار شیمیایی و مقدارهای UFL تجربی مرتبط در جدول S1 فایل پیوست ارایه شده است.

محاسبه توصیف‌کننده‌های مولکولی

برای به دست آوردن یک مدل QSPR، ترکیب‌ها باید توسط توصیف‌کننده‌های مولکولی نمایش داده شوند. مجموعه گسترده‌ای از توصیف‌کننده‌ها شامل توصیف‌کننده‌های توپولوژیکی، توصیف‌کننده‌های هندسی، توصیف‌کننده‌های الکترواستاتیکی و توصیف‌کننده‌های شیمیایی کوانتومی محاسبه شد. در این کار، برای محاسبه توصیف‌کننده‌ها، نخست ساختار ترکیب‌های مورد مطالعه در ChemDraw, 2D ترسیم و سپس با استفاده از روش تبدیل پیش فرض که در CS Chem3D Ultra قرار دارد به ساختارهای سه بعدی تبدیل شدند. در ادامه کار، با توجه به این که برای محاسبه برخی از توصیف‌کننده‌ها نیاز به پایدارترین ساختار است، ساختارهای سه بعدی تولید شده توسط نرم افزار HyperChem به روش نیمه‌تجربی AM1 بهینه شدند [۳۵-۳۲]. خروجی نرم افزار HyperChem برای محاسبه توصیف‌کننده‌ها به عنوان ورودی نرم افزارهای Dragon و MOE مورد استفاده قرار گرفت. در مجموع، ۳۴۷۵ توصیف‌کننده برای هر ترکیب محاسبه شد.

پس از محاسبه توصیف‌کننده‌های مولکولی، نخست توصیف‌کننده‌هایی که مقدارهای ثابت و یا نزدیک به ثابت برای همه مولکول‌ها دارند حذف شدند، زیرا این توصیف‌کننده‌ها تفاوت‌های ساختاری میان ترکیب‌هایی که دارای مقدارهای UFL متفاوت می‌باشند را رمزگذاری نمی‌کنند. سپس، توصیف‌کننده‌هایی که همبستگی بالایی با یکدیگر دارند شناسایی ($R > ۰/۹۰$) و از بین آن‌ها توصیف‌کننده‌هایی که بیشترین همبستگی با ویژگی مورد مطالعه دارد نگه داشته و بقیه حذف شدند. این کاهش‌ها منجر به مجموعه‌ای کاهش یافته از ۴۱۹ توصیف‌کننده برای مطالعه‌های بعدی شد. پیش از انجام مدل‌سازی توسط برازش خطی چند متغیره، مقیاس‌گذاری[۱] و مرکزسازی[۲] [۳۶] روی توصیف‌کننده‌ها نیز انجام شد.

مدل‌سازی

در این مطالعه، مدل‌سازی با روش برازش خطی چند متغیره[۳] انجام شد. MLR به دلیل سادگی و تفسیر آسان به عنوان یک روش متداول، جذاب و مفید برای ایجاد مدل در روش‌های QSPR استفاده می‌شود [۳۷]. کاربرد این روش مستلزم انتخاب توصیف‌کننده‌های مربوطه از میان مجموعه بزرگی از توصیف‌کننده‌ها است و این اشکال اصلی این روش است.

مطالعه‌ها نشان داده است که معادله بین ویژگی و ویژگی‌های ساختاری مولکول‌ها در بسیاری از موارد خطی نیست [۳۸]. آنالیز برازش خطی چند متغیره که در آن از مرتبه‌های بالاتری از توصیف‌کننده‌ها استفاده می‌شود یک روش مناسب برای مدل‌سازی در این گونه موردها است [۳۶]. از این رو، در این مطالعه، برای به دست آوردن مدل با کیفیت بهتر و مقایسه دقیق‌تر، مدل‌سازی به سه روش انجام شد. در روش اول، توصیف‌کننده‌ها و مقدارهای UFL ترکیب‌ها مورد استفاده قرار گرفتند. مدل‌سازی دوم با تبدیل

(۱) Scaling
(۲) Mean Centering
(۳) Multiple Linear Regression

متغیر روی توصیف‌کننده‌های محاسبه شده انجام شد. برای انجام این کار، ماتریسی شامل درجه‌های اول تا سوم و لگاریتم مقدارهای مطلق توصیف‌کننده‌ها ساخته و پس از کاهش توصیف‌کننده‌ها، الگوریتم ACO حافظه‌دار برای انتخاب بهترین توصیف‌کننده‌ها و ساخت مدل MLR اجرا شد. در روش سوم، تبدیل متغیر تنها روی متغیر وابسته انجام و لگاریتم مقدارهای UFL ترکیب‌ها به عنوان متغیر وابسته برای مدل‌سازی مورد استفاده قرار گرفت.

مدل‌سازی به کمک مجموعه آموزش انجام و توانایی پیش‌بینی مدل به دست آمده با استفاده از مجموعه آزمون شامل ۲۰٪ مولکول‌های اولیه ارزیابی شد. برای ایجاد مجموعه آموزش و آزمون، مجموعه داده‌ها به طور راندوم به دو زیر مجموعه آموزش (۴۷۰ مولکول) و آزمون (۱۱۸ مولکول) تقسیم شد. ضریب‌های برازش مدل با استفاده از مجموعه آموزش محاسبه و سپس برای تعیین مقدارهای UFL مولکول‌ها در مجموعه آزمون استفاده شد. به منظور بررسی احتمال همبستگی شانسی در مدل نهایی، آزمون به هم ریختگی Y[1] براساس تغییر تصادفی متغیر وابسته به کار برده شد [۳۹]. روش ارزیابی متقاطع به طور گسترده برای بررسی توانایی پیش‌بینی مدل استفاده می‌شود [۴۰]. در اینجا توانایی پیش‌بینی مدل توسط روش‌های ارزیابی متقاطع مونته کارلو[2] (MCCV)، ارزیابی متقاطع پنج برابری[3]، ارزیابی متقاطع دو حلقه‌ای[4] (D-CV)، و ارزیابی متقاطع یکتایی[5] (LOOCV) مورد بررسی قرار گرفت. روش ارزیابی متقاطع یکتایی در یک زمان، یک داده از داده‌ها را از سری آموزش حذف می‌کند، مدل را با داده‌های باقی‌مانده می‌سازد و این مدل برای پیش‌بینی ویژگی مربوط به ترکیب بیرون گذاشته به کار می‌رود. این شیوه می‌بایست تکرار شود تا این که ویژگی مربوط به همه ترکیب‌ها توسط این روش پیش‌بینی شود. در روش ارزیابی متقاطع مونته کارلو، ترکیب‌های مجموعه آموزش به طور راندوم و با نسبت معین از پیش تعیین شده به دو دسته آموزش و تست تقسیم می‌شوند. مدل‌سازی برای دسته آموزش جدید انجام و مقدارهای UFL دسته آزمون پیش‌بینی می‌شود و مقدارهای مجذور ضریب همبستگی مقدارهای پیش‌بینی شده و مقدارهای تجربی محاسبه و ثبت می‌شود [۲۰]. در ارزیابی متقاطع پنج برابری، مجموعه داده به ۵ زیر مجموعه تقسیم شده، هر بار یکی از ۵ زیر مجموعه به عنوان مجموعه آزمون در نظر گرفته شده و چهار زیر مجموعه دیگر برای تشکیل یک مجموعه آموزشی جمع شده و متوسط خطا در پنج ارزیابی محاسبه می‌شود. روش ارزیابی متقاطع پنج برابری و مونته کارلو ۱۰۰ مرتبه تکرار می‌شوند و از مقدارهای محاسبه شده میانگین گرفته و گزارش می‌شوند. ارزیابی متقاطع دو حلقه روشی مناسب هم برای انتخاب مدل و هم ارزیابی مدل است [۴۱]. این روش شامل دو حلقه درونی و بیرونی است. نخست، کل داده‌های آموزش به طور تصادفی به دو زیر مجموعه آموزش و آزمون تقسیم می‌شوند. مجموعه آزمون تنها برای ارزیابی مدل استفاده می‌شود. مجموعه آموزش در حلقه درونی استفاده شده و به مجموعه داده‌های آموزش و ارزیابی به طور مکرر تقسیم می‌شود (برای مثال ۵۰ مرتبه). مدل توسط مجموعه داده آموزش ساخته شده ولی مجموعه داده تست برای تخمین خطای مدل استفاده می‌شود. کل الگوریتم D-CV، 100 مرتبه تکرار شد.

همه محاسبه‌ها بر روی یک رایانه شخصی با Core™ i5 2.67GHz Intel CPU و رم ۴ گیگابایتی اجرا شد. سیستم عامل مایکروسافت ویندوز XP بود. همه برنامه‌های لازم در متلب (MathWorks) نوشته شده است.

انتخاب توصیف‌کننده بر پایه الگوریتم کولونی مورچگان

در این کار، الگوریتم کولونی مورچگان حافظه‌دار [۲۴] برای انتخاب بهترین توصیف‌کننده‌هایی که به خوبی معادله بین ساختار و UFL ترکیب‌ها را نشان دهند استفاده شد.

همه الگوریتم‌های ACO مرحله‌های مشابهی در هر تکرار دارند: ایجاد کولونی مورچه‌ها؛ ارزیابی هر مورچه؛ به روز کردن مقدارهای فرومون توسط بهترین مورچه (مورچه‌ها)؛ و تبخیر فرومون. شایان ذکر است که هر مورچه یک راه حل مشکل مورد بررسی است. در این حالت، هر مورچه نمایش رشته‌های بیتی[6] از همه توصیف‌کننده‌ها است که در آن عناصر برای توصیف‌کننده‌های غیرمنتخب صفر تنظیم می‌شوند در حالی که توصیف‌کننده‌های انتخاب شده روی یک تنظیم می‌شوند. انتخاب توصیف‌کننده‌ها بر اساس بردار احتمال انجام و بر اساس مقدارهای فرومون محاسبه می‌شود. نخست همه توصیف‌کننده‌ها مقدارهای فرومون همانندی دارند (به معمول روی یک تنظیم می‌شوند). مدل برازش MLR با استفاده از توصیف‌کننده‌های انتخاب شده ایجاد و سپس هر مورچه توسط تابع برازش از پیش تعریف شده ارزیابی می‌شود. در این کار، تابع برازش، مجذور ضریب

(۱) Y-Randomization Test
(۲) Monte Carlo Cross Validation
(۳) 5-Fold Cross Validation
(۴) Double Cross Validation
(۵) Leave-one-out Cross Validation
(۶) bitstring representation

جدول ۱ - پارامترهای آماری مدل‌های QSPR با استفاده از توصیف‌کننده‌های محاسبه شده به عنوان متغیر مستقل و مقدارهای UFL ترکیب‌ها به عنوان متغیر وابسته

تعداد توصیف‌کننده‌ها	$R^2_{آموزش}$	RMSE-آموزش	q^2	RMSE-CV	5-CV[b]		MCCV[a]		D-CV[c]			$R^2_{تست}$	RMSE-تست	F
					RMSE	Q^2	$R^2_{تست}$	$R^2_{آموزش}$	$R^2_{ارزیابی}$	$R^2_{تست}$	$R^2_{آموزش}$			
۲	۰/۴۷	۷/۰۸	۰/۴۵	۷/۲۳	۷/۲۴	۰/۴۵	۰/۵۴	۰/۴۸	۰/۵۶	۰/۵۴	۰/۴۹	۰/۴۳	۹/۱۲	۲۰۷/۳۴
۳	۰/۶۱	۶/۰۶	۰/۵۶	۶/۴۹	۶/۵۳	۰/۵۵	۰/۵۹	۰/۶۱	۰/۵۸	۰/۶۱	۰/۶۱	۰/۳۸	۹/۵۲	۲۴۵/۵۰
۴	۰/۶۳	۵/۹۲	۰/۵۸	۶/۳۳	۶/۳۸	۰/۵۷	۰/۶۱	۰/۶۳	۰/۶۰	۰/۶۳	۰/۶۳	۰/۳۹	۹/۵۲	۱۹۷/۹۲
۵	۰/۶۲	۵/۹۸	۰/۵۷	۶/۴۱	۶/۴۷	۰/۵۶	۰/۶۰	۰/۶۲	۰/۵۸	۰/۶۲	۰/۶۲	۰/۳۸	۹/۵۱	۱۹۱/۴۸
۶[d]	**۰/۶۸**	**۵/۵۲**	**۰/۶۳**	**۵/۹۵**	**۶/۰۲**	**۰/۶۲**	**۰/۶۶**	**۰/۶۸**	**۰/۶۵**	**۰/۶۸**	**۰/۶۸**	**۰/۳۹**	**۹/۵۴**	**۱۶۳/۲۰**
۷	۰/۶۵	۵/۷۴	۰/۵۸	۶/۳۴	۶/۴۲	۰/۵۷	۰/۶۲	۰/۶۶	۰/۶۱	۰/۶۵	۰/۶۵	۰/۳۹	۹/۴۶	۱۲۳/۷۹
۸	۰/۶۹	۵/۴۰	۰/۶۱	۶/۰۷	۶/۱۶	۰/۶۰	۰/۶۴	۰/۶۹	۰/۶۱	۰/۶۶	۰/۷۰	۰/۴۱	۹/۳۲	۱۴۸/۷۸
۹	۰/۶۸	۵/۴۸	۰/۶۳	۵/۹۶	۶/۰۵	۰/۶۱	۰/۶۵	۰/۶۸	۰/۶۳	۰/۶۷	۰/۶۸	۰/۴۰	۹/۴۳	۱۱۰/۰۱
۱۰	۰/۶۷	۵/۵۷	۰/۶۱	۶/۰۵	۶/۱۴	۰/۶۰	۰/۶۵	۰/۶۸	۰/۶۳	۰/۶۷	۰/۶۷	۰/۳۹	۹/۵۲	۱۳۵/۸۴

a مقدارهای گزارش شده میانگین ۱۰۰ تکرار MCCV است. b مقدارهای گزارش شده میانگین ۱۰۰ تکرار 5-fold CV است. c مقدارهای گزارش شده میانگین ۱۰۰ تکرار D-CV است. d بهترین مدل به منظور تاکید با حروف برجسته نشان داده شده است.

همبستگی به دست آمده از ارزیابی متقاطع (q^2) است. به روز رسانی فرومون توسط منتخب‌ترین مورچه که بیشینه مقدار q^2 را دارد انجام می‌شود. تبخیر فرومون (یعنی جایی که درآن شدت فرومون در طول زمان کاهش می‌یابد) نیز برای جلوگیری از تجمع نامحدود فرومون و همگرایی زودرس احتمالی به دست آمده از آن به راه حل ناخواسته، انجام می‌شود. بنابراین، این سازوکار کل فضای جستجو را بررسی می‌کند. در این مطالعه، این کار با ضرب کردن همه مقدارهای فرومون در ضریب کم‌تر از یک انجام می‌شود.

الگوریتم ACO حافظه‌دار از حافظه خارجی مبتنی بر دانش تکرارهای پیشین ACO استفاده می‌کند. نخست حافظه خالی است، ولی با اجرای چندین بار الگوریتم ACO پر می‌شود. پس از هر ACO، بهترین مورچه منتخب در حافظه ذخیره می‌شود و فرایند تا پر شدن کل حافظه ادامه می‌یابد. به روز رسانی فرومون توسط کل مورچه‌های منتخب جمع‌آوری شده در حافظه انجام می‌شود. سپس حافظه خالی شده و با انجام چندین الگوریتم ACO با استفاده از مسیرهای فرومون به روز شده دوباره پر می‌شود. این روند چندین مرتبه تکرار می‌شود. در پایان، حافظه شامل چندین راه حل برتر مشکل است. تعداد دفعه‌های ظاهر شدن هر توصیف‌کننده در حافظه خارجی ملاک اهمیت آن است. سرانجام، پیش‌بینی توسط منتخب‌ترین مورچه و تفسیر با درنظر گرفتن اهمیت هر توصیف‌کننده انجام می‌شود.

نتیجه‌ها و بحث

مجموعه داده‌های در نظر گرفته شده، یک مجموعه داده بزرگ دارای ۵۸۸ ترکیب آلی متنوع است. حدود ۳۴۷۵ توصیف‌کننده برای هر مولکول محاسبه شد. همان‌گونه که پیش تر نیز اشاره شد، معادله بین ساختار مواد شیمیایی و ویژگی‌های آن‌ها در بسیاری از موارد غیرخطی است. برای غلبه بر این رفتار غیرخطی، از تجزیه و تحلیل برازش خطی چند متغیره بر روی ماتریس توصیف‌کننده‌های تبدیل متغیر یافته و از لگاریتم مقدارهای مطلق UFL به جای مقدارهای تجربی UFL استفاده شد. الگوریتم حافظه‌دار ACO برای انتخاب مناسب‌ترین توصیف‌کننده‌ها به کار برده شد. تمام پارامترهای الگوریتم توسط روش سعی و خطا بهینه شد [۴۲]. این پارامترها شامل تعداد مورچه‌ها، تعداد دوره‌ها، اندازه حافظه خارجی و مقدار ρ و τ استفاده شده در به ترتیب مرحله‌های به روز رسانی و تبخیر فرومون بودند. پارامترهای فوق به ترتیب در ۴۰، ۵۰، ۲۵، ۰/۱، و ۰/۲ تنظیم شدند. مقدار q^2 که ضریب همبستگی ارزیابی متقابل است به عنوان مقدار برازش در این الگوریتم استفاده شد.

در بررسی اول، یک مدل با استفاده از توصیف‌کننده‌های محاسبه شده به دست آمد. مقدارهای UFL ترکیب‌ها به عنوان متغیر وابسته در نظر گرفته شد. در جدول ۱ پارامترهای آماری مدل‌های MLR به دست آمده با اندازه‌های گوناگون به کمک

جدول ۲ - پارامترهای آماری مدل‌های QSPR با استفاده از توصیف‌کننده‌های تبدیل یافته به عنوان متغیر مستقل و مقدارهای UFL ترکیب‌ها به عنوان متغیر وابسته

تعداد توصیف‌کننده‌ها	$R^2_{آموزش}$	RMSE-آموزش	q^2	RMSE-CV	5-CV[b]		MCCV[a]		D-CV[c]			$R^2_{تست}$	RMSE-تست	F
					RMSE	Q^2	$R^2_{تست}$	$R^2_{آموزش}$	$R^2_{ارزیابی}$	$R^2_{تست}$	$R^2_{آموزش}$			
۲	۰/۵۹	۶/۲۰	۰/۵۵	۶/۵۱	۶/۵۲	۰/۵۵	۰/۶۱	۰/۶۰	۰/۶۲	۰/۶۱	۰/۶۱	۰/۴۹	۸/۶۹	۳۴۲/۸۴
۳	۰/۶۷	۵/۵۶	۰/۶۲	۵/۹۷	۶/۰۷	۰/۶۱	۰/۶۳	۰/۶۷	۰/۶۱	۰/۶۰	۰/۶۹	۰/۴۵	۹/۲۱	۳۲۱/۰۴
۴	۰/۶۱	۶/۰۵	۰/۵۶	۶/۴۲	۶/۴۵	۰/۵۶	۰/۶۵	۰/۶۱	۰/۶۳	۰/۶۱	۰/۶۳	۰/۴۹	۸/۷۷	۲۴۶/۹۱
۵[d]	**۰/۶۵**	**۵/۷۵**	**۰/۵۷**	**۶/۳۵**	**۶/۴۰**	**۰/۵۷**	**۰/۶۱**	**۰/۶۵**	**۰/۶۱**	**۰/۶۱**	**۰/۶۷**	**۰/۴۸**	**۸/۸۰**	**۱۷۳/۵۹**
۶	۰/۶۲	۵/۹۹	۰/۵۸	۶/۲۹	۷/۳۴	۰/۵۴	۰/۶۴	۰/۶۲	۰/۶۱	۰/۵۹	۰/۶۴	۰/۴۸	۸/۹۸	۱۲۶/۹۳
۷	۰/۷۰	۵/۳۲	۰/۶۵	۵/۷۷	۵/۸۸	۰/۶۴	۰/۶۹	۰/۷۰	۰/۶۷	۰/۶۸	۰/۷۰	۰/۴۰	۹/۴۹	۲۱۷/۷۰
۸	۰/۷۴	۴/۹۹	۰/۶۳	۵/۹۰	۶/۱۷	۰/۶۱	۰/۶۶	۰/۷۴	۰/۶۳	۰/۶۳	۰/۷۷	۰/۵۱	۸/۴۳	۲۵۹/۷۹
۹	۰/۷۱	۵/۲۸	۰/۶۴	۵/۸۰	۵/۹۴	۰/۶۳	۰/۶۸	۰/۷۰	۰/۶۴	۰/۶۲	۰/۷۳	۰/۴۹	۸/۹۶	۱۵۷/۹۱
۱۰	۰/۶۹	۵/۳۹	۰/۶۴	۵/۸۳	۵/۹۳	۰/۶۴	۰/۶۷	۰/۷۰	۰/۶۷	۰/۶۸	۰/۶۹	۰/۴۱	۹/۴۵	۳۵۱/۹۱

a مقدارهای گزارش شده میانگین ۱۰۰ تکرار MCCV است. b مقدارهای گزارش شده میانگین ۱۰۰ تکرار 5-fold CV است. C مقدارهای گزارش شده میانگین ۱۰۰ تکرار D-CV است. d بهترین مدل به منظور تاکید با حروف برجسته نشان داده شده است.

الگوریتم ACO حافظه‌دار ارایه شده است. از این جدول پیدا است که مقدار q^2 با افزایش تعداد توصیف‌کنندهای وارد شده افزایش می‌یابد و در مدل با ۶ توصیف‌کننده بیش‌ترین مقدار را دارد. با توجه به این که پارامترهای آماری دیگر این مدل نیز بیشینه مقدار را دارد، بنابراین، مدل با ۶ توصیف‌کننده به عنوان بهترین مدل انتخاب شد. برای این مدل، مقدارهای R^2 آموزش و R^2 تست به ترتیب ۰/۶۸ و ۰/۳۹ بود. این مقدارهای کم ما را به سمت کار بیش‌تر در این مجموعه داده سوق می‌دهد. اهمیت آماری مدل ممکن است با در نظرگرفتن معادله غیر خطی بین توصیف کننده‌های محاسبه شده و متغیر وابسته بهبود یابد.

از نتیجه‌های فوق نتیجه‌گیری می‌شود که یک معادله غیر خطی بین مقدارهای UFL ترکیب‌ها و توصیف‌کننده‌های محاسبه شده وجود دارد. برای مدل‌سازی این معادله غیر خطی، تبدیل متغیر روی توصیف‌کننده‌های محاسبه شده اعمال شد. برای انجام این کار، ماتریسی شامل درجه‌های اول تا سوم و لگاریتم مقدارهای مطلق توصیف‌کننده‌های محاسبه شده به عنوان متغیر مستقل و مقدارهای UFL ترکیب‌ها به عنوان متغیر وابسته ساخته شد. الگوریتم ACO حافظه‌دار برای انتخاب بهترین زیرمجموعه از توصیف‌کننده‌ها اجرا و مدل‌های MLR با اندازه‌های گوناگون ساخته شد. نتیجه‌ها در جدول ۲ داده شده است.

مقایسه نتیجه‌های جداول ۱ و ۲ نشان می‌دهد که با استفاده از توصیف‌کننده‌های تبدیل یافته مدل‌هایی با توانایی پیشگویی بهتری به دست می‌آید. $R^2_{آموزش}$ و $R^2_{تست}$ بهترین مدل به ترتیب ۰/۶۵ و ۰/۴۸ بود. در بررسی بعدی، یک مدل بر اساس لگاریتم متغیر وابسته (لگاریتم مقدار UFL) بدون تبدیل متغیر مستقل (توصیف‌کننده‌ها) به دست آمد. به طور کلی، تبدیل روی متغیر وابسته زمانی انجام می‌شود که داده‌ها غیر نرمال، غیر خطی، یا ناهمواریانس هستند و برای تبدیل آن‌ها به نرمال، خطی، و یا هم واریانسی انجام می‌شود [۴۳]. شایان ذکر است که تبدیل متغیر پاسخ نسبت به متغیرهای مستقل متداول‌تر است و اغلب تبدیل لگاریتمی برای تبدیل توزیع غیر نرمال به توزیع تقریبا نرمال پاسخ استفاده می‌شود. مدل‌های MLR گوناگون که متغیر وابسته جدید ($\log \mathrm{UFL}$) را به توصیف‌کننده‌های مولکولی محاسبه شده نسبت می‌دهد نیز به دست آمد. جدول ۳ پارامترهای آماری مدل‌های به دست آمده را نشان می‌دهد. معادله ریاضی مدل در معادله (۱) نمایش داده شده است.

$$\log \mathrm{UFL} = 0.98 + 0.09\mathrm{X2v} - 0.14\mathrm{F01[c-c]} - 0.54\mathrm{GCUT_{SLOGP_3}} - 0.06\mathrm{SMR_{VSA7}} \qquad (۱)$$

که در آن $\log \mathrm{UFL}$ لگاریتم مقدار حد اشتعال‌پذیری بالا، $X2v$ توصیف‌کننده شاخص اتصال ظرفیت، $F01[c-c]$ توصیف‌کننده اثر انگشتی دو بعدی است که موقعیت و فراوانی پیوندهای کربن-کربن

جدول ۳ - پارامترهای آماری مدل‌های QSPR با استفاده از توصیف‌کننده‌های محاسبه شده به عنوان متغیر مستقل و مقدارهای log UFL ترکیب‌ها به عنوان متغیر وابسته

تعداد توصیف‌کننده‌ها	$R^2_{آموزش}$	RMSE-آموزش	q^2	RMSE-CV	5-CV[b]		MCCV[a]		D-CV[c]			$R^2_{تست}$	RMSE-تست	F
					RMSE	Q^2	$R^2_{تست}$	$R^2_{آموزش}$	$R^2_{ارزیابی}$	$R^2_{تست}$	$R^2_{آموزش}$			
۲	۰/۷۴	۰/۱۲	۰/۷۳	۰/۱۲	۰/۱۲	۰/۷۳	۰/۷۴	۰/۷۴	۰/۷۴	۰/۷۴	۰/۷۴	۰/۷۲	۰/۱۳	۶۵۴/۰۲
۳	۰/۷۶	۰/۱۱	۰/۷۶	۰/۱۱	۰/۱۱	۰/۷۵	۰/۷۶	۰/۷۶	۰/۷۶	۰/۷۶	۰/۷۶	/۰/۷۶	۰/۱۲	۴۹۵/۸۹
۴[d]	**۰/۸۰**	**۰/۱**	**۰/۷۹**	**۰/۱**	**۰/۱۰**	**۰/۷۹**	**۰/۸۰**	**۰/۸۰**	**۰/۸۰**	**۰/۸۰**	**۰/۸۰**	**۰/۷۵**	**۰/۱۲**	**۴۷۲/۱۶**
۵	۰/۷۸	۰/۱۱	۰/۷۷	۰/۱۱	۰/۱۱	۰/۷۷	۰/۷۷	۰/۷۸	۰/۷۷	۰/۷۷	۰/۷۸	۰/۷۵	۰/۱۲	۴۰۴/۹۵
۶	۰/۷۹	۰/۱	۰/۷۸	۰/۱۱	۰/۱۱	۰/۷۸	۰/۷۹	۰/۷۹	۰/۷۹	۰/۷۹	۰/۸۰	۰/۷۸	۰/۱۲	۲۹۶/۷۳
۷	۰/۸۱	۰/۱	۰/۸۰	۰/۱	۰/۱	۰/۷۹	۰/۸۰	۰/۸۱	۰/۸۰	۰/۸۰	۰/۸۱	۰/۷۶	۰/۱۲	۲۷۶/۴۳
۸	۰/۸۱	۰/۱	۰/۸۰	۰/۱	۰/۱	۰/۸۰	۰/۸۰	۰/۸۱	۰/۸۰	۰/۸۰	۰/۸۱	۰/۷۷	۰/۱۲	۲۸۱/۵۰
۹	۰/۸۲	۰/۱	۰/۸۰	۰/۱	۰/۱	۰/۸۰	۰/۸۱	۰/۸۲	۰/۸۱	۰/۸۱	۰/۸۲	۰/۷۸	۰/۱۲	۲۲۹/۲۷
۱۰	۰/۸۲	۰/۱	۰/۸۰	۰/۱	۰/۱	۰/۸۰	۰/۸۱	۰/۸۲	۰/۸۰	۰/۸۰	۰/۸۲	۰/۷۶	۰/۱۲	۲۰۳/۱۵

a مقدارهای گزارش شده میانگین ۱۰۰ تکرار MCCV است. b مقدارهای گزارش شده میانگین ۱۰۰ تکرار 5-fold CV است. c مقدارهای گزارش شده میانگین ۱۰۰ تکرار D-CV است. d بهترین مدل به منظور تأکید با حروف برجسته نشان داده شده است.

را نشان می‌دهد GCUT_SLOGP_3 جزء توصیف‌کننده‌های ماتریس مجاور و فاصله است و SMR_VSA7 جزء توصیف‌کننده‌های مساحت سطح است.

جدول ۴ پارامترهای آماری این مدل را می‌دهد. $R^2_{آموزش}$ و $R^2_{تست}$ برای این مدل به ترتیب ۰/۸ و ۰/۷۵ به دست آمد. همه مقدارهای t با مقدارهای کم p معنی دار و تأیید کننده اهمیت هر توصیف‌کننده هستند. برای یک مدل قابل قبول مقدارهای p هر کدام از ضریب‌ها که با استفاده از مقدارهای t محاسبه می‌شوند باید کوچک‌تر از ۰/۰۵ باشد. برای محاسبه مقدار t از مقدار خطای استاندارد (SE) که در جدول ۴ داده شده است استفاده می‌شود. برای ارزیابی سریع ضریب‌ها، می‌توان مقدارهای SE را با مقدار قدر مطلق هر ضریب برازش مقایسه نمود و در صورتی که مقدار ضریب از مقدار SE متناظرش خیلی بزرگتر باشد، قابل قبول خواهد بود. شایان ذکر است که هر چه ضریب برازش بزرگ‌تر و SE مربوط به آن کوچک‌تر باشد، t محاسبه شده بزرگ‌تر و p مرتبط با آن کوچک‌تر خواهد بود. مقدار آماری F مدل با مقدار p کم‌تر از $۱/۰۰\times۱۰^{-۲۰}$، ۴۷۲/۱۶ بود (در مقایسه با مقدار بحرانی ۳/۰۶ در سطح احتمال ۰/۰۵). مقدارهای RMSE[1] مجموعه آموزش و تست نیز به ترتیب ۰/۱ و ۰/۱۲ بود. نام‌ها و بیش‌ترین و کم‌ترین مقدارهای توصیف‌کننده‌های انتخاب شده در جدول ۵ آمده است. شکل ۱ نمودار مقدارهای log UFL دیده شده را در مقابل مقدارهای پیش‌بینی شده توسط مدل نشان می‌دهد. همان‌گونه که دیده می‌شود همبستگی خوبی بین این مقدارها وجود دارد. به منظور ارزیابی بیشتر مدل، پارامترهای Q^2_{F1}، Q^2_{F2}، Q^2_{F3}، r^2_m و ضریب همبستگی سازگاری[2] (CCC) برای مدل یادشده محاسبه شدند. برای یک مدل قابل قبول، مقدارهای Q^2_{F1}، Q^2_{F2}، Q^2_{F3} باید بزرگ‌تر از ۰/۶، مقدار r^2_m بزرگ‌تر از ۰/۵ و مقدار CCC بیش‌تر از ۰/۸۰ باشد [۴۴]. مقدارهای Q^2_{F1}، Q^2_{F2}، Q^2_{F3}، r^2_m و CCC برای مدل ارایه شده به ترتیب برابر ۰/۷۴، ۰/۷۴، ۰/۷۱، ۰/۶۳ و ۰/۸۱ به دست آمد که نشان دهنده قابل قبول بودن مدل نهایی است.

برای بررسی همبستگی شانسی، آزمون به هم ریختگی Y استفاده شد. در این روش مقدارهای متغیر وابسته بدون تغییر دادن متغیرهای مستقل به طور تصادفی تغییر می‌نماید. در صورت شانسی بودن مدل به دست آمده، پارامتر R^2 برای مدل‌های ساخته شده در این آزمون نیز مقدارهای بالایی را از خود نشان می‌دهند. ۱۰۰ بار ترتیب مقدارهای تجربی بصورت تصادفی به‌هم ریخته می‌شود و هر بار مدل برازش خطی ایجاد شده، الگوریتم ACO حافظه‌دار برای انتخاب بهترین توصیف‌کننده‌ها اجرا و ضریب همبستگی محاسبه می‌شود. سرانجام همان‌گونه که در شکل ۲ نشان داده شده است،

(۱) Root- Mean-Square Error

(۲) Concordance Correlation Coefficient

جدول ۴ - پارامترهای آماری برای بهترین مدل به دست آمده با استفاده از توصیف‌کننده‌های محاسبه شده به عنوان متغیر مستقل و log UFL به عنوان متغیر وابسته

توصیف‌کننده	Beta	SE	t	p	VIF
ثابت	۰/۹۷۶	۰/۰۰۵	۲۰۵/۵۸	۰/۰	-
X2v	۰/۰۹۳	۰/۰۰۸	۱۱/۳۳	$۱/۹\times۱۰^{-۲۶}$	۲/۹۷
F01[C-C]	۰/۱۳۶	۰/۰۰۹	−۱۴/۴۴	$۲/۶\times۱۰^{-۳۹}$	۳/۹۴
GCUT_SLOGP_3	۰/۱۵۴	۰/۰۰۹	−۱۷/۰۸	$۳/۹\times۱۰^{-۵۱}$	۳/۵۷
SMR_VSA7	۰/۰۶۰	۰/۰۰۶	−۹/۹۰	$۴/۴\times۱۰^{-۲۱}$	۱/۶۱

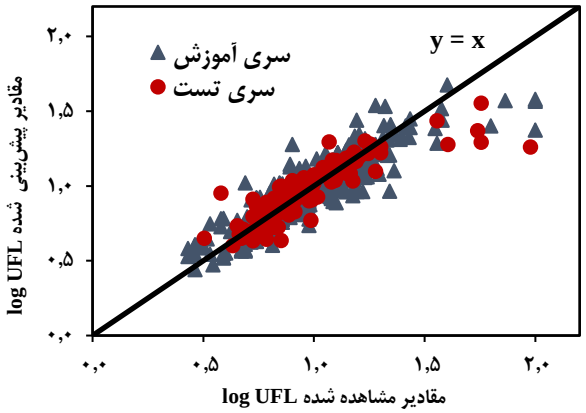
جدول ۵. بیش‌ترین و کم‌ترین مقدارهای توصیف‌کننده‌های استفاده شده در بهترین مدل به دست آمده

	X2v	F01[C-C]	GCUT_SLOGP_3	SMR_VSA7
بیش‌ترین	۰/۰	۰/۰	۰/۸	۰/۰
حداقل	۱۳/۶۰	۲۴/۰	۲/۹	۲۳۴/۹

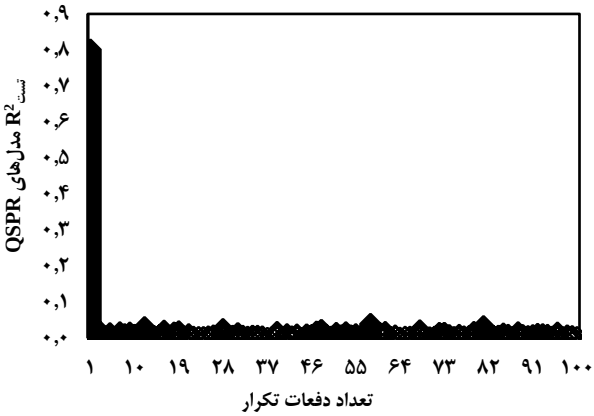
بیش‌ترین میزان همبستگی مربوط به داده‌های واقعی بوده است. همبستگی پایین مربوط به ۱۰۰ بار به‌هم ریختگی نشان‌دهنده این است که ارتباط شانسی بین توصیف‌کننده‌های محاسبه‌ای و مقدارهای log UFL وجود ندارد و اثبات می‌کند که مدل اصلی به صورت شانسی توسعه نیافته و اهمیت مدل را در پیش‌بینی مقدارهای log UFL ترکیب‌ها تأیید می‌کند.

در مقایسه با کارهای گزارش شده پیشین، مدل به دست آمده در این مطالعه پارامترهای آماری خوبی دارد. این مدل دسته گسترده‌ای از ترکیب‌های آلی متنوع را می‌پوشاند. بنابراین، همان‌گونه که در جدول ۶ دیده می‌شود، گستره بسیارگسترده‌تری از مقدارهای UFL ترکیب‌ها را می‌تواند پیش‌بینی کند.

برتری یک مدل خطی چند متغیره آن است که بسیار ساده است و مدل به دست آمده به راحتی قابل تفسیر است. بزرگی ضریب‌های مربوط به هر توصیف‌کننده در این مدل نشانگر اهمیت نسبی آن توصیف‌کننده بر ویژگی تحت مطالعه (در اینجا log UFL) می‌باشد و همچنین علامت آن ضریب‌ها می‌تواند نشان دهد که تأثیر آن توصیف‌کننده بر ویژگی مورد نظر مثبت یا منفی است. در صورتی می‌توان تعبیر درستی از مدل ساخته شده داشت که توصیف‌کننده‌های به کار گرفته شده در مدل از نظر ریاضی مستقل از یک‌دیگر (عمود بر هم) باشند. مطالعه‌های پیشین انجام شده نشان داده است که در صورتی که توصیف‌کننده‌ها با یک‌دیگر وابسته خطی باشند ممکن است ضریب‌های بزرگ‌تر از حد انتظار و یا نشانه‌های نادرست طی مدل‌سازی به دست آید [۴۵].



شکل ۱ - نمودار مقدارهای دیده شده در مقابل مقدارهای پیش‌بینی شده log UFL بهترین مدل به دست آمده



شکل ۲ - آزمون به‌هم ریختگی Y. نوار اول مقدار $R^2_{تست}$ مدل را بر اساس داده‌های واقعی نشان می‌دهد. ۱۰۰ نوار دیگر $R^2_{تست}$ مدل‌ها بر اساس داده‌های به هم ریخته را نشان می‌دهد.

به منظور بررسی وابستگی متقابل توصیف‌کننده‌های انتخاب شده به یک‌دیگر، مقدارهای ضریب همبستگی پیرسون برای آن‌ها محاسبه شد (جدول ۷). براساس نتیجه‌های این جدول، از چهار توصیف‌کننده انتخابی، سه توصیف‌کننده به یک‌دیگر وابستگی معنی‌دار دارند. البته شایان ذکر است که در تمام مطالعه‌ها QSAR/QSPR نخست توصیف‌کننده‌هایی که با یک‌دیگر بیش از ۰/۹۰ همبستگی نشان دهند (ضریب همبستگی پیرسون) حذف می‌شوند. این بدان معنی است که توصیف‌کننده‌هایی که بعضا همبستگی بالای ۰/۸۵ دارند ممکن است در مدل ظاهر و سرانجام مدل قابل پذیرشی ایجاد کنند. برای بررسی بهتر میزان وابستگی توصیف‌کننده‌ها به یکدیگر، به جای بررسی مستقیم ضریب‌های همبستگی پیرسون، مقدارهای VIF توصیف‌کننده‌ها مورد مطالعه

جدول ۶ - مقایسه مدل به دست آمده و تعدادی از مدل‌های گزارش شده پیشین برای پیش‌بینی مقدارهای UFL ترکیب‌ها

تعداد ترکیب‌ها	ابزار ساخت مدل	تعداد توصیف کننده‌ها	$R^2_{\text{آموزش}}$[a]	$R^2_{\text{تست}}$	q^2	RMSE-آموزش	RMSE-تست	گستره کاربرد (%)	مرجع
۸۶۷	GC-FFNN	۱۱۳ گروه عاملی	۰/۹۴۶۹	۰/۹۴۳۳		۰/۸۸۱	۰/۸۸۲	۲/۵ -۲۴	[۲۷]
۸۶۵	GA-MLR	۵	۰/۹۲۰۲		۰/۹۱۸۴	-	-	۲/۵ – ۲۴	[۲]
۷۹	MLR KNN SVM RF	۶	۰/۸۱۲ ۰/۸۲۱ ۰/۸۸۵ ۰/۹۶۱	۰/۶۵۹ ۰/۶۷۵ ۰/۷۸۳ ۰/۹۲۴	۰/۵۳۴ ۰/۶۱ ۰/۶۶۲ ۰/۵۶۱	۰/۱۴۸ ۰/۱۴۴ ۰/۱۱۸ ۰/۰۷	۰/۱۴۰ ۰/۱۹۹ ۰/۱۱۹ ۰/۱۳۶	۵ -۵۰	[۳۱]
۵۷۹	GA-MLR	۴	۰/۷۵۸	۰/۷۵۱	۰/۵۱۱	۲/۷	۲/۷۷	۲/۷ -۴۰/۳	[۱۶]
۵۸۸	Memorized-ACS MLR	۴	۰/۸	۰/۷۵	۰/۷۹	۰/۱	۰/۱۲	۲/۷-۱۰۰	کار ارایه شده

جدول ۷ - مقدارهای ضریب‌های همبستگی پیرسون برای چهار توصیف‌کننده انتخاب شده

توصیف‌کننده	X2v	F01[c − c]	GCUT_SLOGP_3	SMR_VSA7
X2v	۱/۰۰	۰/۶۸	۰/۷۶	۰/۲۳
F01[c − c]		۱/۰۰	۰/۸۰	-۰/۲۴
GCUT_SLOGP_3			۱/۰۰	۰/۰۲
SMR_VSA7				۱/۰۰

قرار می‌گیرد. VIF فاکتور تورم واریانس[۱] می‌باشد که برای تک تک توصیف‌کننده‌های وارد شده در مدل با توجه به معادله (۲) محاسبه شد:

$$VIF = \frac{1}{1-R^2} \qquad (۲)$$

که R^2 ضریب همبستگی یک توصیف‌کننده است که بر سایر توصیف‌کننده‌ها برازش شده است. مقدار بحرانی پیشنهاد شده برای تورم واریانس طبق متون ۵٪ در نظر گرفته می‌شود [۴۶]. اگر مقدار تورم واریانس برای یک توصیف‌کننده بزرگ‌تر از این مقدار بحرانی باشد اطلاعات توصیف‌کننده مد نظر می‌تواند در اثر همبستگی توصیف‌کننده‌ها مخفی شود. مقدار فاکتور تورم واریانس مربوط به چهار توصیف‌کننده انتخابی توسط الگوریتم ACO حافظه‌دار در جدول ۴ نشان داده شده است. همان‌گونه که در جدول دیده می‌شود تمام توصیف‌کننده‌ها VIF کم‌تر از ۵ دارند که نشان‌دهنده آن است که مدل به دست آمده معنی‌دار است.

چهار توصیف کننده توسط الگوریتم ACO حافظه‌دار انتخاب شدند. توصیف‌کننده X2v، جزء توصیف‌کننده‌های شاخص اتصال ظرفیت[۲] است [۴۷]. در محاسبه توصیف‌کننده‌های شاخص اتصال، مقدارهایی به کلیه اتم‌ها به جز هیدروژن اختصاص داده می‌شود. این مقدارها نشان دهنده تعداد اتم‌های همسایه پیوند شده به اتم موردنظر به جز هیدروژن است و با δ نشان داده می‌شود. برای هر پیوند طبق فرمول (۳) جمله C_k تعریف می‌شود:

$$C_k = (\delta_i \delta_j)^{-\frac{1}{2}} \qquad (۳)$$

جمع همه این ترم‌ها روی کل مولکول مقدار عددی شاخص اتصال را می‌دهد. برای محاسبه شاخص اتصال ظرفیت نیز از فرمول گفته شده در بالا استفاده می‌شود؛ مقدار δ^v به جای δ استفاده می‌شود. δ^v نشان دهنده تعداد الکترون‌های ظرفیت هر اتم منهای الکترون‌های باند شده به هیدروژن است. توصیف‌کننده شاخص اتصال ظرفیت تغییرات در شکل و سایز ترکیب‌ها را توصیف می‌کند [۴۷]. با توجه به علامت مثبت آن در مدل به دست آمده می‌توان نتیجه گرفت با افزایش حجم مولکول، برهم کنش بین مولکول‌ها افزایش یافته و log UFL ترکیب‌ها نیز افزایش می‌یابد و گستره اشتعال‌پذیری نیز افزایش می‌یابد.

(۱) Variance Inflation Factor

(۲) Valence Connectivity Index

توصیف‌کننده F01[C-C]، از جمله توصیف‌کننده‌های اثر انگشت فراوانی دو بعدی است که موقعیت و فراوانی جفت اتمی C-C را در فاصله ۱ نشان می‌دهد [۴۸]. علامت منفی این توصیف‌کننده در مولکول نشان می‌دهد که با افزایش تعداد جفت اتم‌های C-C با فاصله ۱ نسبت به یکدیگر، ترکیب‌های شیمیایی کاهش و گستره اشتعال‌پذیری کاهش می‌یابد. توصیف‌کننده GCUT_SLOGP_3، جزء توصیف‌کننده‌های ماتریس مجاور و فاصله[1] است [۴۹,۵۰]. ماتریس مجاور یک ساختار شیمیایی شامل عناصر $[M_{ij}]$ است که اگر اتم‌های i و j با هم پیوند داشته باشند $M_{ij}=1$ است و در غیر این صورت صفر است. ماتریس فاصله با عناصر $[D_{ij}]$ تعریف می‌شود که D_{ij} طول کوتاه‌ترین مسیر از اتم i تا j است و اگر i و j قسمتی از اجزاء وصل شده یکسان باشند صفر استفاده می‌شود. هر عنصر ij ماتریس مجاور فاصله برابر $1/\sqrt{D_{ij}}$ است که D_{ij} فاصله اصلاح شده بین اتم‌های i و j روی گراف مولکول است. این توصیف‌کننده‌ها از مقدارهای ویژه ماتریس مجاور–فاصله محاسبه می‌شوند. عناصر قطری مقدارهای log P هستند. بنابراین، این توصیف‌کننده سهم هر اتم در محاسبه log P (ضریب تقسیم اکتانول/آب) و میزان آب گریزی ترکیب را نشان می‌دهد. علامت توصیف‌کننده درون مدل منفی می‌باشد که نشان می‌دهد با افزایش ویژگی آبگریزی مولکول، ترکیب مورد نظر مولکول آلی بزرگ‌تری است و نیروی واندروالس قویتر بوده و فشار بخار نیز کم‌تر می‌باشد. پس، مقدار log UFL کاهش یافته و گستره اشتعال‌پذیری کاهش می‌یابد. توصیف‌کننده SMR_VSA7، از جمله توصیف‌کننده‌های مساحت سطح و بر اساس محاسبه مساحت سطح واندروالس تقریبی هر اتم (v_i) به همراه تعدادی ویژگی اتمی دیگر (p_i) است. v_i توسط تقریب جدول اتصال محاسبه می‌شود. SMR_VSA7، مجموع v_i کل اتم‌ها تعریف می‌شود طوری که p_i (ضریب شکست مولی) در یک بازه مشخص (a, b) و $R_i>0/56$ است. R_i بیانگر سهم اتم i در ضریب شکست مولی است همان‌گونه که در توصیف‌کننده SMR محاسبه شده است [۵۰]. این توصیف‌کننده اثر مساحت سطح و قطبش‌پذیری را بر اشتعال‌پذیری ترکیب‌ها نشان می‌دهد. علامت منفی ضریب برازش آن در مدل مورد نظر نشان می‌دهد که با افزایش مساحت سطح ترکیب‌ها، قطبش پذیری ترکیب‌ها بیشتر شده، نیروی جاذبه لاندن بین مولکول‌ها افزایش و log UFL کاهش می‌یابد. بنابراین، گستره اشتعال‌پذیری ترکیب‌ها کاهش می‌یابد.

نتیجه‌گیری

در این کار، ۵۸۸ ترکیب متنوع آلی برای تولید و ارزیابی مدل‌های QSPR برای پیش‌بینی مقدارهای UFL آن‌ها استفاده شد. مدل‌ها به ترتیب با استفاده از توصیف‌کننده‌های محاسبه شده، درجه‌های اول تا سوم‌های توصیف‌کننده‌های محاسبه شده و لگاریتم مقدارهای مطلق توصیف‌کننده‌ها ساخته شدند. افزون بر این، مدل با استفاده از تبدیل متغیر روی متغیر وابسته (UFL) نیز به دست آمد. برای انتخاب بهترین توصیف‌کننده‌هایی که اثر ویژگی‌های ساختاری مولکول بر متغیر وابسته (مقدارهای UFL) ترکیب‌ها را به خوبی نشان می‌دهند، الگوریتم قدرتمند کولونی ACO حافظه‌دار به کار برده شد. در صورت استفاده از درجه‌های بالاتر توصیف‌کننده‌ها به جای خود توصیف‌کننده‌ها، کیفیت مدل‌های به دست آمده اندکی بهبود یافت. اما با اعمال تبدیل متغیر روی متغیر وابسته (UFL) و استفاده از مقدارهای log UFL به جای UFL به عنوان متغیر وابسته، کیفیت مدل به طور چشمگیری بهبود یافته و مدلی با پارامترهای آماری خوب شامل $R^2_{\text{آموزش}}$=۰/۸۰، $R^2_{\text{تست}}$=۰/۷۵، q^2=۰/۷۹، F=۴۷۲/۱۶، RMSE=۰/۱۲ به دست آمد. این نتیجه‌ها نشان می‌دهند که در صورت استفاده از تبدیل متغیر، به ویژه در مورد متغیر وابسته، کیفیت مدل‌سازی بهبود خواهد یافت. همچنین مقایسه مدل ارایه شده با مدل‌های گزارش شده پیشین نشان می‌دهد که مدل حاضر گستره بسیار گسترده‌تری از UFL ترکیب‌های آلی را به خوبی پیش‌بینی می‌کند و می‌تواند برای پیش‌بینی حد اشتعال‌پذیری بالای ترکیب‌های آلی دیگر و حتی سنتز نشده استفاده شود.

تاریخ دریافت: ۱۳۹۸/۱۰/۱۰ ؛ تاریخ پذیرش: ۱۳۹۹/۰۲/۰۸

مراجع

[1] Catoire L., Naudet V., Estimation of Temperature-Dependent Lower Flammability Limit of Pure Organic Compounds in Air at Atmospheric Pressure, *Process Saf. Prog.*, **24:** 130-137 (2005).

(۱) Adjacency and Distance Matrix Descriptors

[2] Gharagheizi F., Prediction of Upper Flammability Limit Percent of Pure Compounds from Their Molecular Structures, *J. Hazard. Mater.*, **167:** 507-510 (2009).

[3] Winterbone D., Turan A., "Advanced Thermodynamics for Engineers". 2nd ed. Butterworth-Heinemann, Arnold, London, (2015).

[4] Vidal M., Rogers W., Holste J., Mannan M., A Review of Estimation Methods for Flash Points and Flammability Limits, *Process Saf. Prog.*, **23:** 47-55 (2004).

[5] Albahri T.A., Flammability Characteristics of Pure Hydrocarbons, *Chem. Eng. Sci.*, **58:** 3629-3641 (2003).

[6] Pan Y., Jiang J., Ding X., Wang R., Jiang J., Prediction of Flammability Characteristics of Pure Hydrocarbons from Molecular Structures, *AlChE J.*, **56:** 690-701 (2010).

[7] High M.S., Danner R.P., Prediction of Upper Flammability Limit by a Group Contribution Method, *Ind. Eng. Chem. Res.*, **26:** 1395-1399 (1987).

[8] Seaton W.H., Group Contribution Method for Predicting the Lower and the Upper Flammable Limits of Vapors in Air, *J. Hazard. Mater.*, **27:** 169-185 (1991).

[9] Suzuki T., Koide K., Correlation between Upper Flammability Limits and Thermochemical Properties of Organic Compounds, *Fire Mater.*, **18:** 393-397 (1994).

[10] Suzuki T., Ishida M., Neural Network Techniques Applied to Predict Flammability Limits of Organic Compounds, *Fire Mater.*, **19:** 179-189 (1995).

[۱۱] پوربشیر ا.، مهاجری اول ژ.، نکوئی م.، حمیدوند س.، مطالعه ارتباط کمّی ساختار- فعالیت برای پیش بینی فعالیت مهارکنندگی PIM مشتق‌های تری آزولوپیریدین با استفاده از الگوریتم ژنتیک ـ برازش خطی چندگانه، *نشریه شیمی و مهندسی شیمی ایران*، **۳۷:** ۱۳۷ تا ۱۴۸ (۱۳۹۷).

[۱۲] قدیمی س.، رشنو طائی م.ا.، ابراهیمی ولموزویی ع.ا.، سامانی ک.ا.، جوانی ز.، نصرت زادگان ک.، پورایوبی م.، معادله ساختار و فعالیت در فسفرآمیدها، *نشریه شیمی و مهندسی شیمی ایران*، **(۳)۳۰:** ۹۱ تا ۱۰۵ (۱۳۹۰).

[۱۳] رحمان ستایش ش.، طریک ع.، زبیدی ر.، پیش‌بینی دمای ذوب مایع‌های یونی بر پایه آنیون بیس (تری‌فلورومتیل‌سولفونیل) ایمید با رویکرد QSPR، *نشریه شیمی و مهندسی شیمی ایران*، **(۱)۳۹:** ۱۴۹ تا ۱۵۸ (۱۳۹۹).

[14] Todeschini R., Consonni V., Handbook of Molecular Descriptors. John Wiley & Sons, (2008).

[15] Karelson M., Lobano, V.S., Katritzky A.R., Quantum-Chemical Descriptors in QSAR/QSPR Studies, *Chem. Rev.*, **96:** 1027-1044 (1996).

[16] Pan Y., Jiang J., Wang R., Cao H., Cui Y., Prediction of the Upper Flammability Limits of Organic Compounds from Molecular Structures, *Ind. Eng. Chem. Res.*, **48:** 5064-5069 (2009).

[17] Wang B., Xu K., Wang Q., Prediction of Upper Flammability Limits for Fuel Mixtures Using Quantitative Structure–Property Relationship Models, *Chem. Eng. Commun.*, **206:** 247-253 (2019).

[18] Katritzky A.R., Fara D.C., How Chemical Structure Determines Physical, Chemical, and Technological Properties: An Overview Illustrating the Potential of Quantitative Structure−Property Relationships for Fuels Science, *Energy Fuels*, **19:** 922-935 (2005).

[19] Taskinen J., Yliruusi J., Prediction of Physicochemical Properties Based on Neural Network Modelling, *Adv. Drug Del. Rev.*, **55:** 1163-1183 (2003).

[20] Zare-Shahabadi V., Lotfizadeh M., Gandomani A.R.A., Papari M.M., Determination of Boiling Points of Azeotropic Mixtures Using Quantitative Structure–Property Relationship (QSPR) Strategy, *J. Mol. Liq.*, **188:** 222-229 (2013).

[21] Mirjalili S., Evolutionary Algorithms and Neural Networks, Springer, 33-42 (2019).

[22] Dorigo M., Stützle T., "Handbook of Metaheuristics", Springer, 311-351 (2019).

[23] Atabati M., Zarei K., Borhani A., Ant Colony Optimization as a Descriptor Selection in QSPR Modeling: Estimation of the λmax of Anthraquinones-Based Dyes, *J. Saudi Chem. Soc.*, **20:** S547-S551 (2016).

[24] Shamsipur M., Zare-Shahabadi V., Hemmateenejad B., Akhond M., An Efficient Variable Selection Method Based on the Use of External Memory in Ant Colony Optimization. Application to QSAR/QSPR Studies, *Anal. Chim. Acta*, **646:** 39-46 (2009).

[25] Hemmateenejad B., Shamsipur M., Zare-Shahabadi V., Akhond M., Building Optimal Regression Tree by Ant Colony System–Genetic Algorithm: Application to Modeling of Melting Points, *Anal. Chim. Acta*, **704:** 57-62 (2011).

[26] Dorigo M., Stützle T., "Ant Colony Optimization", The MIT Press. The MIT Press, Cambridge, Massachusetts, (2004).

[27] Gharagheizi F., Chemical Structure-Based Model for Estimation of the Upper Flammability Limit of Pure Compounds, *Energy Fuels*, **24:** 3867-3871 (2010).

[28] Abbasitabar F., Zare-Shahabadi V., Development Predictive QSAR Models for Artemisinin Analogues by Various Feature Selection Methods: A Comparative Study, *SAR QSAR Environ. Res.*, **23:** 1-15 (2012).

[29] Chandrashekar G., Sahin F., A Survey on Feature Selection Methods, *Comput. Electr. Eng.*, **40:** 16-28 (2014).

[30] Filgueiras P.R., Portela N.A., Silva S.R.C., Castro E.V.R., Oliveira L.M.S.L., Dias J.C.M., Neto A.C., Romão W., Poppi R.J., Determination of Saturates, Aromatics, and Polars in Crude Oil by ^{13}C NMR and Support Vector Regression with Variable Selection by Genetic Algorithm, *Energy Fuels*, **30:** 1972-1978 (2016).

[31] Yuan S., Jiao Z., Quddus N., Kwon J.S., II, Mashuga C.V., Developing Quantitative Structure–Property Relationship Models to Predict the Upper Flammability Limit Using Machine Learning, *Ind. Eng. Chem. Res.*, **58:** 3531-3537 (2019).

[32] Cvetnic M., Perisic D.J., Kovacic M., Ukic S., Bolanca T., Rasulev B., Kusic H., Bozic A.L., Toxicity of Aromatic Pollutants and Photooxidative Intermediates in Water: A QSAR Study, *Ecotoxicology Environmental Safety*, **169:** 918-927 (2019).

[33] Garcia M.L., de Oliveira A.A., Bueno R.V., Nogueira V.H., de Souza G.E., Guido R.V., QSAR Studies on Benzothiophene Derivatives as Plasmodium Falciparum N-Myristoyltransferase Inhibitors: Molecular Insights into Affinity and Selectivity, *Drug Dev. Res.*, **83(2):** 264-284 (2020).

[34] Babu S., Nagarajan S.K., Madhavan T., Investigation of Empirical and Semi-Empirical Charges to Study the Effects of Partial Charges on Quality and Prediction Accuracy in 3D-QSAR, *Chemistry Select*, **4:** 3990-4002 (2019).

[35] Jiao Z., Yuan S., Zhang Z., Wang Q., Machine Learning Prediction of Hydrocarbon Mixture Lower Flammability Limits Using Quantitative Structure-Property Relationship Models, *Process Saf. Prog.*, **39(2):** e12103 (2019).

[36] Zare-Shahabadi V., Quantitative Structure–Activity Relationships of Dihydrofolatereductase Inhibitors, *Med. Chem. Res.*, **25:** 2787-2797 (2016).

[37] Faramarzi Z., Abbasitabar F., Zare-Shahabadi V., Jahromi H.J., Novel Mixture Descriptors for the Development of Quantitative Structure-Property Relationship Models for the Boiling Points of Binary Azeotropic Mixtures, *J. Mol. Liq.*, **296:** 111854 (2019).

[38] Martin Y.C., "Quantitative Drug Design: A Critical Introduction", CRC Press, (2010).

[39] Abbasitabar F., Zare-Shahabadi V., QSAR Study of Artemisinin Analogues as Antimalarial Drugs by Neural Network and Replacement Method, *Drug Res.*, **67:** 476-484 (2017).

[40] Zhu J., Lu W., Liu L., Gu T., Niu B., Classification of Src Kinase Inhibitors Based on Support Vector Machine, *QSAR & Combinatorial Science*, **28:** 719-727 (2009).

[41] Baumann D., Baumann K., Reliable Estimation of Prediction Errors for QSAR Models under Model Uncertainty Using Double Cross-Validation, *J. Cheminform.*, **6:** 47 (2014).

[42] Zare-Shahabadi V., Abbasitabar F., Application of Ant Colony Optimization in Development of Models for Prediction of Anti-HIV-1 Activity of HEPT Derivatives, *J. Comput. Chem.*, **31:** 2354-2362 (2010).

[43] Draper N.R., Smith H., "Applied Regression Analysis", John Wiley & Sons Inc , (1998).

[44] Saaidpour S., Quantitative Modeling for Prediction of Critical Temperature of Refrigerant Compounds, *Phys. Chem. Res.* **4:** 61-71 (2016).

[45] Hansch C., Quantitative Approach to Biochemical Structure-Activity Relationships, *Acc. Chem. Res.*, **2:** 232-239 (1969).

[46] Jo D.H., Lee S.G., Kim B.T., No G.T., Quantitative Structure-Activity Relationship (QSAR) Study of New Fluorovinyloxycetamides, *Bull. Korean Chem. Soc.*, **22:** 388-394 (2001).

[47] Kier L.B., Hall L.H., Derivation and Significance of Valence Molecular Connectivity, *J. Pharm. Sci.*, **70:** 583-589 (1981).

[48] Filipic S., Antic A., Vujovic M., Nikolic K., Agbaba D., A Comparative Study of Chromatographic Behavior and Lipophilicity of Selected Imidazoline Derivatives, *J. Chromatogr. Sci.*, **54:** 1137-1145 (2016).

[49] Tamiji Z., Salahinejad M., Niazi A., Molecular Modeling of Potential Pet Imaging Agents for Adenosine Receptor in Parkinson's Disease, *Struct. Chem.*, **29:** 467-479 (2018).

[50] Wildman S.A., Crippen G.M., Prediction of Physicochemical Parameters by Atomic Contributions, *J. Chem. Inf. Comput. Sci.*, **39:** 868-873 (1999).