

МОДЕЛИРОВАНИЕ КОЛИЧЕСТВЕННОЙ ВЗАИМОСВЯЗИ СТРУКТУРА-СВОЙСТВО С ПОМОЩЬЮ ПРОГРАММЫ ISIDA_QSPR

В. П. Соловьев *

Лаборатория новых физико-химических проблем, Институт физической химии и электрохимии им. А. Н. Фрумкина Российской академии наук, Ленинский пр. 31/4, Москва, 119991 Россия

* E-mail: solovev-vp@mail.ru

В этом обучающем курсе на примерах нескольких задач иллюстрируется моделирование количественных взаимосвязей структура-свойство (Quantitative Structure-Property Relationships, QSPR) с помощью программы ISIDA_QSPR ^[1-5], реализующей в качестве метода машинного обучения множественную линейную регрессию (МЛР) (Multiple Linear Regression, MLR) с использованием субструктурных молекулярных фрагментов (СМФ) (Substructural Molecular Fragments, SMF) ^[6] в качестве дескрипторов (независимых переменных).

QSPR модели для предсказания физических и химических свойств, биологической активности строят с помощью методов машинного обучения, применяемого к набору данных о веществах или молекулах с известным изучаемым свойством и структурами молекул, описанными молекулярными дескрипторами. Построенные модели используются для предсказания свойств новых веществ.

СМФ дескрипторы являются подграфами молекулярных графов. Каждый уникальный подграф рассматривается как дескриптор, а его кратность (количество вхождений) в молекуле используется как значение дескриптора. Программа ISIDA_QSPR включает два принципиальных класса фрагментов: топологические пути и атомно-центрированные фрагменты – атомы со связанными с ними ближайшими соседями. Фрагменты могут быть представлены с явным указанием атомов и связей, только атомов или связей. При этом атомы водорода могут быть указаны в структуре молекулы, –

выражены явно, - и присутствовать в молекулярных фрагментах или выражены неявно и отсутствовать в СМФ.

Для повышения надежности предсказаний программа ISIDA_QSPR выполняет ансамблевое QSPR моделирование [7-14]. Оно включает генерацию большого числа индивидуальных моделей, выбор из них статистически значимых и применение выбранных моделей к тестируемым/новым данным для получения средних предсказанных значений, исключая выбросы (*консенсус-модель*) [7]. Каждая индивидуальная модель соответствует определенному типу СМФ и определенным параметрам метода машинного обучения. Программа строит MLR модели, сочетая методы прямого [10] и обратного [15] пошагового отбора переменных. В процессе моделирования применяется *n*-кратный внешний перекрестный контроль (*n-fold external cross-validation, nCV*) для тестирования выбранных наиболее надежных предсказательных моделей.

Программа ISIDA_QSPR представляет собой графический интерфейс, позволяющий достаточно легко выполнять задачи QSPR моделирования и поддерживающий графический анализ результатов, связанных с графическим представлением структур химических соединений. Она работает под операционной системой Windows различных версий.

Пошаговые инструкции упражнений обучающего курса по использованию программы ISIDA_QSPR выделены слева вертикальной зелёной полосой. Обучающий курс включает небольшой теоретический и вспомогательный материал, представленный без зеленой полосы.

Инсталляция программы ISIDA_QSPR на компьютере под управлением операционных систем WINDOWS различных версий: а) загрузите zip-архив программы ISIDA_QSPR с сайта <http://vpsolovev.ru/programs/isidaqspr/>, б) распакуйте zip-архив, содержащий директорию ISIDA_QSPR. Для Windows 7 и Windows Vista строго рекомендуется размещать директорию ISIDA_QSPR не на системном диске. После этого программа готова к использованию. Главный исполняемый файл программы ISIDA_QSPR.exe.

В этом обучающем курсе рассматриваются следующие упражнения:

1. Построение индивидуальной модели множественной линейной регрессии (МЛР) на одном наборе дескрипторов СМФ и предсказание свойства на тестовом наборе данных.
2. Анализ дескрипторов СМФ индивидуальной модели МЛР: вклады фрагментов в моделируемое свойство, матрица корреляций вкладов фрагментов и критерий сходства молекул на основе СМФ.

3. Построение индивидуальной модели с выполнением внешнего k -кратного перекрестного контроля.

4. Расчет средних предсказаний, исключая выбросы, на основе ансамбля моделей множественной линейной регрессии с использованием различных типов СМФ дескрипторов (*консенсус-модель*).

5. Прогнозирование свойств и виртуальный скрининг с использованием построенных предсказательных QSPR моделей.

ДААННЫЕ ДЛЯ МОДЕЛИРОВАНИЯ

Исходные данные для моделирования с использованием программы ISIDA_QSPR должны быть представлены в формате структура-данные (Structure Data File, SDF) [16], который может содержать структуры молекул, количественные и качественные данные. В этом курсе используются подготовленные файлы исходных данных: **GdL_logK_TRAINING.SDF** и **GdL_logK_TEST.SDF**. Файл GdL_logK_TRAINING.SDF содержит экспериментальные значения логарифмов *констант устойчивости* $\log K$ 111-ти комплексов $Gd^{3+}L$ иона гадолиния (Gd^{3+}) с органическими молекулами (L) в воде при температуре 298 K и ионной силе 0.1 моль/л для построения QSPR моделей, связывающих константу устойчивости комплекса $Gd^{3+}L$ со структурой лиганда L. Файл GdL_logK_TEST.SDF содержит величины констант $\log K$ для аналогичных комплексов $Gd^{3+}L$ 57-ми органических молекул для тестирования полученных моделей. В SDF файлах экспериментальные значения логарифмов констант устойчивости представлены полем под именем LogK.

Исходные данные для моделирования в формате структура-данные SDF должны быть расположены в главном каталоге программы ISIDA_QSPR.

О константе устойчивости комплексов

Константа устойчивости K является константой равновесия обратимой химической реакции комплексообразования катиона металла (M) с лигандом L (органической или неорганической молекулой или их анионом):



Согласно закону действующих масс константа устойчивости K равна отношению равновесной концентрации C_{ML} продукта реакции – комплекса ML к произведению равновесных концентраций реагентов C_M и C_L :

$$K = \frac{C_{ML}}{C_M C_L}$$

Обычно используют логарифм константы устойчивости $\log K$, т.к. он пропорционален стандартной энергии Гиббса реакции $\Delta G = -RT \cdot \ln K = -RT \cdot \ln 10 \cdot \log K$, которая широко применяется в изучении комплексообразования. Таким образом, указанные выше файлы подготовлены для моделирования характеристики реакции – константы устойчивости. Поскольку в случае реакции (1) ион металла Gd^{3+} и условия реакции постоянны, то константа $\log K$ зависит только от свойств лиганда. Поэтому мы будем строить модели структура – свойство, которые связывают константу $\log K$ со структурой лиганда. Это позволит выбирать или конструировать лиганды, обеспечивающие требуемую величину константы $\log K$, т.е. необходимую устойчивость комплекса. Комплексы гадолиния с органическими лигандами используются в медицине [17].

Создание SDF файлов формата структура-данные

Данные в формате SDF для QSPR моделирования можно создавать и редактировать с помощью инструментальных средств EdiSDF [18] и EdChemS [19], входящих в состав программы ISIDA_QSPR. Этот разносторонний инструментарий позволяет создавать, объединять, дробить файлы структура-данные, добавлять в них новые записи, добавлять новые поля данных или редактировать существующие, а также редактировать структурные формулы.

Указанные выше SDF файлы не нуждаются в подготовке для моделирования, но читатель может попробовать использовать инструменты EdiSDF и EdChemS для редактирования или создания SDF файлов: Выберите (щелкните) **Tools** → **SDF Editor** в главном меню ISIDA_QSPR (рис. 1), чтобы открыть менеджер файлов EdiSDF (рис. 2а). Щелкните на кнопке **New SD File** менеджера EdiSDF, введите названия полей данных в виде отдельных строк для нового SDF файла в текстовой области появившегося окна **Construction of blank SD File** (рис. 3), где затем щелкните на кнопке **Ok** и сохраните новый пустой SDF файл. Щелкните **File** → **Open** в главном меню EdiSDF, чтобы открыть сохраненный пустой SDF файл. Отметьте галочкой метку **Edit field** слева внизу в EdiSDF (рис. 2а), чтобы иметь возможность вводить и редактировать содержимое полей данных. Щелкните на кнопке **Edit Mol** (рис. 2а) для ввода или редактирования 2D молекулярной

структуры открывшимся редактором EdChemS (рис. 2б). Используйте помощь в подготовке структурных формул: щелкните по пункту меню **Help** → **EdChemS Help** главного меню EdChemS, чтобы использовать файл помощи. Если химическая формула подготовлена, выберите в главном меню **File** → **Return to EdiSDF**, чтобы вернуться в EdiSDF из EdChemS. В EdiSDF используйте таблицу под структурной формулой, включающую два столбца с названиями **FIELD NAME** и **PROPERTY VALUE** (рис. 2а), для ввода или редактирования значений полей данных: дважды щелкните в ячейке под **PROPERTY VALUE**, введите с клавиатуры цифровое значение или текст для выбранного поля данных и кликните по навигационной кнопке ►, чтобы сохранить введенные данные. Щелкните на кнопке **Add Record** для подготовки следующей записи (структуры и данных) SDF файла: новая запись будет создана и открыта как копия текущей записи на экране EdiSDF и далее готова к редактированию. Сохраните подготовленный SDF файл: используйте **File** → **Save as** в главном меню EdiSDF.

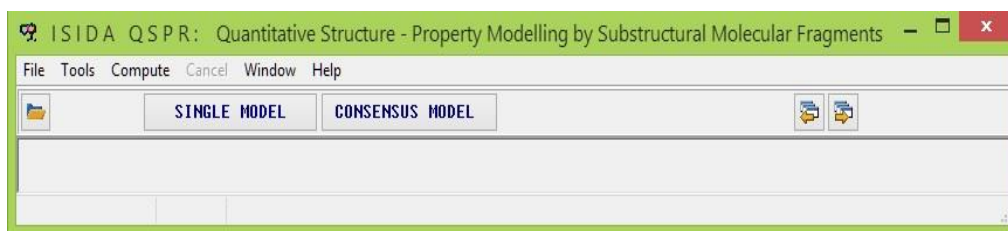
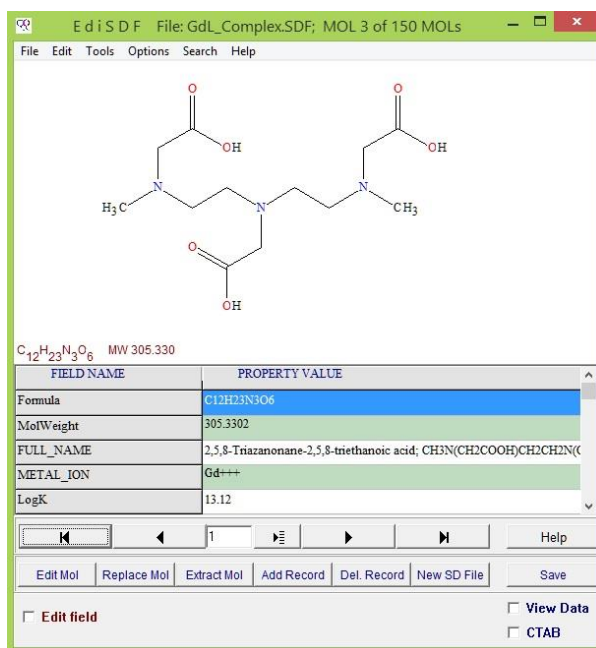
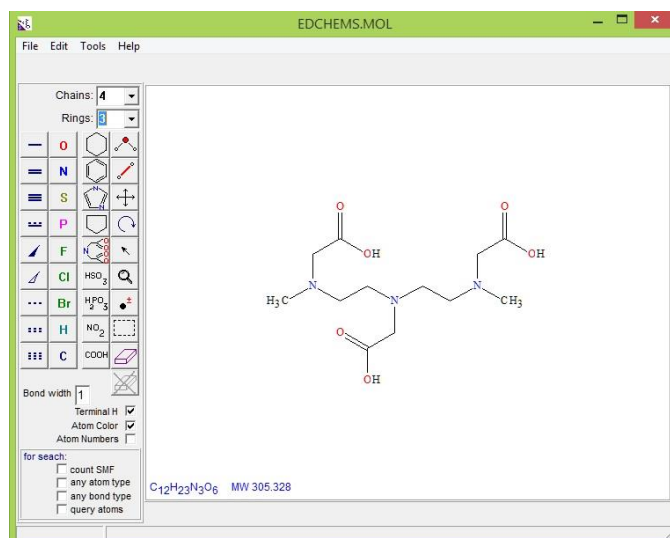


Рис. 1. Рабочий стол (главное окно) программы ISIDA_QSPR.



а



б

Рис. 2. Рабочий стол (главное окно) программ EdiSDF (а) и EdChemS (б) для работы с файлами формата структура-данные (SDF).

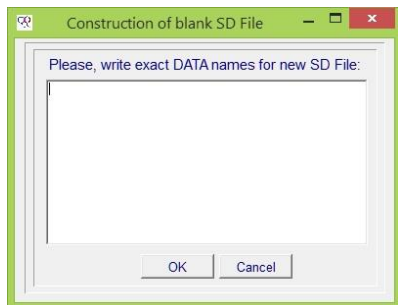


Рис. 3. Графическое окно процедуры для создания бланка SDF файла.

УПРАЖНЕНИЕ 1

1. Построение индивидуальной модели множественной линейной регрессии и предсказание свойства на тестовом наборе данных

Задача:

В этом упражнении мы строим индивидуальную МЛР модель на одном наборе СМФ дескрипторов с их отбором программой, чтобы продемонстрировать простое QSPR моделирование с помощью программы ISIDA_QSPR. Для проведения моделирования и прогнозирования (предсказания) пользователю необходим только входной файл в формате структура – данные (Structure-Data File, SDF).

Управление программой ISIDA_QSPR осуществляется манипулятором типа мышь и её левой кнопкой. Запустите главный исполняемый файл программы ISIDA_QSPR.exe. Щелкните мышью на кнопке **Single Model** рабочего стола (главного окна) программы ISIDA_QSPR (рис. 1), чтобы открыть диалоговое окно **Single Model Calculations** (рис. 4) для начала ввода входных данных и параметров для расчета индивидуальной модели. Диалоговое окно включает панель **Data** для ввода данных, панель **Descriptors** для выбора типа СМФ дескрипторов, панель **Model** для выбора типа уравнения МЛР и методов прямого и обратного пошагового отбора переменных, а также панель **Validation** для выбора методов внутренней и внешней проверки модели на качество предсказаний (рис. 4). Графические элементы программы имеют подсказки: при задержке указателя мыши на элементе всплывает строка подсказки.

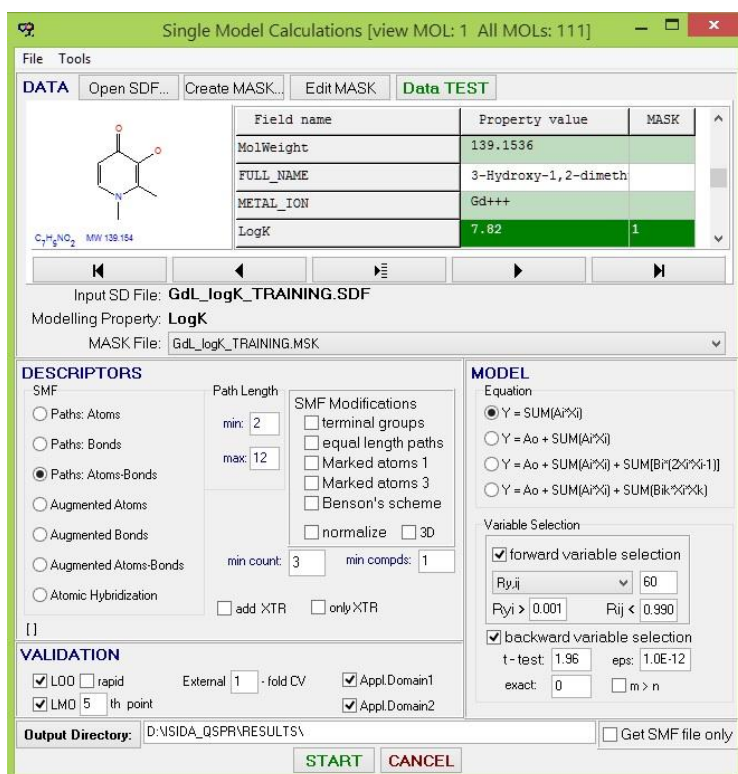


Рис. 4. Диалоговое окно расчета индивидуальной (единичной) модели.

1.1. Ввод данных для моделирования

Выше было отмечено, что данные для моделирования программой ISIDA_QSPR должны быть подготовлены в формате файла структура-данные (Structure-Data File, SDF) [16], где моделируемое количественное свойство представлено полем данных. Это поле для обучающего набора должно содержать вещественные числа для всех соединений в файле SDF, хотя значения свойства для тестируемых соединений для этого поля могут отсутствовать. Молекулярные структурные формулы могут быть представлены в виде двухмерных или трехмерных структур. Как правило, атомы водорода химических структур не указываются, хотя молекулярные данные с явными атомами водорода в некоторых случаях позволяют получить модели структура – свойство с более высокими предсказательными характеристиками по сравнению с моделями без явного указания водородных атомов [20]. Входной файл должен находиться в главном каталоге программы ISIDA_QSPR.

На панели **Data** диалогового окна построения одной модели (рис. 4) щелкните на кнопке **Open SDF** и откройте файл GdL_logK_TRAINING.SDF в главном каталоге программы ISIDA_QSPR. На панели **Data** появится первая химическая структура файла и строка Input SD File: GdL_logK_TRAINING.SDF и таблица (рис. 4). Таблица содержит информацию о полях данных открытого SDF файла в виде двух колонок с названиями: **Field name** (имя поля) и **Property value** (значение свойства) (рис. 4). Щелкните на ячейке LogK в столбце **Field name**, чтобы выбрать моделируемое свойство. На панели **Data** внизу появится строка: Modelling Property: LogK (рис. 4). Навигационные кнопки под структурой и таблицей позволяют просматривать SDF файл.

1.2. Деление данных на обучающий и тестовый наборы

Программа ISIDA_QSPR позволяет разделить исходный набор данных на два подмножества: соответственно, обучающий и тестовый наборы для построения и проверки модели. Тестовый набор служит для внешнего контроля качества QSPR модели. Программа создает файл-маску (*.MSK), чтобы указать, какие соединения в исходном SDF файле включены в обучающий и тестовый наборы. В текстовом файле-маске первая строка содержит общее количество молекул в SDF-файле и число 1, разделенные пробелом. Вторая и остальные строки содержат целые числа 0 или 1, разделенные пробелом: 1, если соединение включено в обучающий набор, иначе 0, если оно принадлежит тестовому набору. Таким образом, порядковый номер нуля или единицы, начиная со второй строки MSK файла, соответствует номеру соединения в SDF файле. Каждая строка MSK файла может содержать любое количество чисел, но удобно использовать 10 чисел на строку.

На панели **Data** диалогового окна расчета одной модели (рис. 4) щелкните мышью на кнопке **Create MASK**. Появится диалоговое окно Create MASK (рис. 5): щелкните на радиокнопке (переключателе) **Create MASK With TEST SET**; введите число 5 в редактируемое поле **each** и 3 в редактируемое поле **starting from**; в данном случае, начиная с третьего соединения файла SDF, каждое пятое соединение будет использоваться для тестового набора; щелкните на кнопке **START**, чтобы сохранить/перезаписать файл-маску GdL_logK_TRAINING.MSK в главном каталоге ISIDA_QSPR с помощью открывшегося окна Save mask file Dialog. Имя файла GdL_logK_TRAINING.MSK появится внизу панели **Data** (рис. 4); щелкните на кнопке **Data**

TEST для проверки входных данных. Появится информационное диалоговое окно с сообщением: “Input data files are in internal agreement” (Файлы входных данных находятся во внутреннем согласии); щелкните на кнопке **OK** информационного диалогового окна, чтобы его закрыть.



Рис. 5. Диалоговое окно Create Mask для создания файла-маски.

1.3. Субструктурные молекулярные фрагменты в качестве дескрипторов

Программа ISIDA_QSPR содержит модуль для генерации дескрипторов. Генерируемые дескрипторы - субструктурные молекулярные фрагменты (СМФ) [5,6] представляют собой подграфы молекулярного графа (рис. 6). Каждый уникальный подграф рассматривается как дескриптор, а его кратность (количество вхождений) в молекуле используется как значение дескриптора (табл. 1). Дескрипторы вычисляются исключительно из 2D или 3D химических структурных формул SDF файла структурных данных. ISIDA_QSPR может генерировать два основных класса СМФ (рис. 6): топологические пути (I) и атомно центрированные фрагменты - атомы с ближайшими соседями (II) с указанием типов атомов и связей (AB), только типов атомов (A), только типов связей (B). Для атома может быть указан только символ атома и дополнительно - состояние его гибридизации, нотация Бенсона [21] и метка [5,22] (рис. 7). Химическая связь молекулярных фрагментов может иметь следующие характеристики: тип (ковалентная σ -связь, координационная для нековалентных связей и динамическая для реакций), порядок (одинарная, двойная, тройная, ароматическая) и топологию (циклическая или ациклическая) (рис. 8). Топологический путь может характеризоваться оптимальностью (кратчайший или все пути), длиной (минимальную и максимальную) и детальностью описания (указаны все атомы или указаны только концевые атомы). Для топологических путей определяются минимальное ($m_{min} = 2, 3... 15$) и максимальное ($m_{max} = 2, 3... 15$) количество составляющих атомов. Для выбранных значений m_{min} и m_{max} сгенерированные СМФ включают все промежуточные пути с t атомами: $m_{min} \leq t \leq m_{max}$.

Таким образом, для построения QSPR модели исходный набор дескрипторов генерируется программой ISIDA_QSPR в соответствии с заданной пользователем схемой фрагментации и характеристиками СМФ.

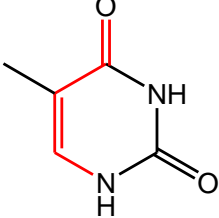
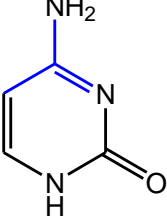
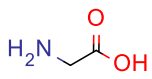
Топологические пути	Атомно центрированные фрагменты
	
последовательность атомов и связей $N - C = C - C = O$ последовательность атомов $N C C C O$ последовательность связей $- = - =$	ближайшие атомы и связи $C(- C; - N; = N)$ ближайшие атомы $C(C; N; N)$ ближайшие связи $C(-; -; =)$

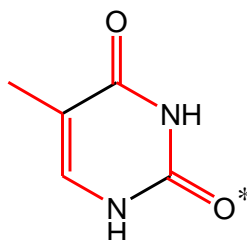
Рис. 6. Два принципиальных класса субструктурных молекулярных фрагментов: топологические пути и атомно центрированные фрагменты. Обозначения связей: ‘-’ одинарная, ‘=’ двойная.

Таблица 1. Молекула глицина и ее молекулярные фрагменты - кратчайшие топологические пути для каждой пары атомов - кратчайшие цепочки от одного атома до другого из соединенных химическими связями атомов молекулы.

молекула	атомы водорода не указаны		атомы водорода указаны явно				
	фрагмент	n^a	фрагмент	n	фрагмент	n	
	C-N	1	C-N	1	H-C-N	2	
	C-C	1	C-C	1	C-C-H	2	
	C-O	1	C-O	1	H-C-H	1	
	C=O	1	C=O	1	C-O-H	1	
	C-C-O	1	C-C-O	1	H-C-C-O	2	
	C-C-N	1	C-C-N	1	H-C-C=O	2	
	C-C=O	1	C-C=O	1	H-C-N-H	4	
	O-C=O	1	O-C=O	1	C-C-N-H	2	
	N-C-C-O	1	N-C-C-O	1	C-C-O-H	1	
	N-C-C=O	1	N-C-C=O	1	H-O-C=O	1	
				H-N	2	H-O-C-C-N	1
				C-H	2	H-N-C-C-O	2
				H-O	1	H-N-C-C=O	2
			C-N-H	2	H-C-C-O-H	2	
			H-N-H	1	H-N-C-C-O-H	2	

^a n - число фрагментов

АТОМ			
Символ химического элемента	Гибридизационное состояние	Система обозначений Бенсона	Помеченный атом
C	CD для C_{sp^2}	CN для $C=N$	C*
N	CT для C_{sp}	CO для $C=O$	N*
O	CB для $C_{sp^2 \text{ aromatic}}$	NO для $N=O$	O*
...



примеры молекулярных фрагментов			
O=C-N-C=O	OD=CD-N-CD=OD	CO-N-CO	O=C-N-C=O*
C-C=C-N	C-CD=CD-N	C-CD=CD-N	

Рис. 7. Атомные характеристики субструктурных молекулярных фрагментов.

СВЯЗЬ		
Тип	Кратность	Топология
ковалентная	одинарная	циклическая
нековалентная	двойная	ациклическая
динамическая	тройная	
	ароматическая	

реакция

конденсированный граф реакции

динамические связи:
образующиеся (I)
исчезающие (II)

нековалентная связь:
водородная (---)

циклическая связь:
явно указанная (---), чтобы
учитывать в молекулярных
фрагментах

примеры молекулярных фрагментов		
$C-ds-C-s-C-C=O$	$S-C=O \cdots H-O$	$C \cdots N \cdots C-C-O$

Рис. 8. Характеристики химических связей субструктурных молекулярных фрагментов.

На панели *Descriptors* диалогового окна Single Model Calculations (рис. 4) щелкните на радиокнопке (переключателе) *Paths: Atoms-Bonds*, введите число 2 в поле редактирования *min* и 12 в поле редактирования *max* для указания минимальной и максимальной длины СМФ дескрипторов, соответственно. Убедитесь, что в группе полей SMF Modifications панели *Descriptors* отсутствуют отметки галочкой для включения специальных характеристик для СМФ. Используйте по умолчанию число 3 в поле редактирования *min count* и 1 в поле редактирования *min compds* (рис. 4). Таким образом, для этого упражнения в качестве дескрипторов выбраны кратчайшие топологические пути с явным указанием типов атомов и связей с минимальным $m_{min} = 2$ и максимальным $m_{max} = 12$ числом атомов в путях. Обратите внимание, что программа также сгенерирует все промежуточные пути с m атомами: $m_{min} \leq m \leq m_{max}$. Убедитесь, что в правом нижнем углу диалогового окна (рис. 4) не установлена отметка галочкой *Get SMF file only*. Эта опция используется для генерации только файла СМФ дескрипторов и сохранения его в рабочем каталоге без моделирования.

1.4. Регрессионное уравнение

В качестве метода машинного обучения в ISIDA_QSPR используется множественная линейная регрессия (МЛР) для построения взаимосвязи между независимыми переменными (СМФ дескрипторами) x_i и зависимой переменной (моделируемым свойством) y . В программе заложена возможность построения четырех типов уравнений:

$$y \approx \hat{y} = \sum_i a_i x_i + \Gamma \quad (2)$$

$$y \approx \hat{y} = a_0 + \sum_i a_i x_i + \Gamma \quad (3)$$

$$y \approx \hat{y} = a_0 + \sum_i a_i x_i + \sum_i b_i (2x_i^2 - 1) + \Gamma \quad (4)$$

$$y \approx \hat{y} = a_0 + \sum_i a_i x_i + \sum_{i,k} b_{ik} x_i x_k + \Gamma \quad (5)$$

Здесь a_i и b_i (b_{ik}) – коэффициенты - вклады молекулярных фрагментов в свойство y , x_i - число вхождений фрагмента i -го типа. Свободный терм a_0 не зависит от дескрипторов. ISIDA_QSPR строит модели как включая терм a_0 , так и без него. Дополнительный вклад $\Gamma = \sum_m c_m D_m$ может использоваться для описания любой особенности химических соединений с

помощью внешних по отношению к СМФ дескрипторов D_m ; по умолчанию $\Gamma = 0$. Следует отметить, что наблюдаемое свойство y известно с некоторой погрешностью, однако значения переменных x_i - числа вхождений каждого фрагмента для каждой молекулы известны точно.

При построении уравнения МЛР задача состоит в том, чтобы для заданного набора n объектов, характеризующихся свойством y и описываемых независимыми переменными x , найти такие значения коэффициентов уравнения, чтобы оно наилучшим образом описывало моделируемое свойство y , т.е. чтобы различия между наблюдаемыми значениями свойства y и величинами \hat{y} , рассчитанными по уравнению (2) – (5), были как можно меньше. Наиболее часто в качестве критерия такого согласия используют минимум суммы квадратов отклонений наблюдаемых значений y от соответствующих значений модели \hat{y} : $\min \sum_{i=1}^n (y_i - \hat{y}_i)^2$. Метод поиска значений коэффициентов a_i и b_i , удовлетворяющих такому критерию, называется методом наименьших квадратов [23]. Высоко надежным и быстрым методом вычисления коэффициентов a_i и b_i (b_{ik}) уравнений (2) – (5) является *сингулярное разложение* (Singular Value Decomposition, SVD) [24-26], которое применяется в программе ISIDA_QSPR.

На панели **Model** (рис. 4) щелкните на радиокнопке $Y = \text{SUM}(A_i * X_i)$ для выбора линейного уравнения без свободного коэффициента.

1.5. Прямой и обратный пошаговый выбор переменных

Молекула и построенное из молекул вещество могут быть описаны многими тысячами (N) дескрипторов – количественных и качественных характеристик x_i , $i = 1, 2, \dots, N$. Сложной проблемой построения QSPR моделей является поиск ряда значимых дескрипторов x_1, x_2, \dots, x_m и оптимального их числа m , которые обеспечивают с определенной точностью функциональную взаимосвязь с моделируемым свойством y : $y \approx \hat{y} = F(x_1, x_2, \dots, x_m)$. При этом, эта взаимосвязь должна быть справедлива не только для обучающего набора данных, на основе которого эта взаимосвязь построена, но и для большого круга тестируемых данных, а в идеальном случае, для всех веществ, т.е. носить характер закона природы. Таким образом, искомая взаимосвязь в определенной мере должна описывать реально существующую закономерность в природе.

Для отбора значимых дескрипторов из исходного множества сгенерированных СМФ в программе используется: а) фильтрация дескрипторов; б) прямой пошаговый отбор переменных – итерационный процесс накопления дескрипторов, которые значительно коррелируют со моделируемым свойством [10]; в) обратное пошаговое исключение

переменных – итерационный процесс удаления дескрипторов, которые вносят большие ошибки в модель ^[15].

а) фильтрация фрагментных дескрипторов

На стадии фильтрации исключаются СМФ дескрипторы: 1) редкие, встречающиеся реже, чем в трех молекулах обучающей выборки, 2) присутствующие во всех соединениях обучающего набора с постоянным числом кратности, 3) связанные СМФ дескрипторы, всегда присутствующие в молекулах обучающего набора в одной и той же комбинации, интерпретируются как один расширенный молекулярный фрагмент; 4) имеющие коэффициент корреляции с моделируемым свойством ниже заданного порога, 5) один из каждой пары, если коэффициент корреляции между ними выше заданного порога.

Согласно ^[27,28] квадрат коэффициента корреляции R_{yi}^2 дескриптора x_i с моделируемым свойством y рассчитывается по формуле:

$$R_{yi}^2 = \frac{s_{yi}^2}{s_{yy} s_{ii}} = \frac{\{\sum_{k=1}^n y_k x_{ki} - \frac{1}{n}(\sum_{k=1}^n y_k)(\sum_{k=1}^n x_{ki})\}^2}{(\sum_{k=1}^n y_k^2 - \frac{1}{n}(\sum_{k=1}^n y_k)^2)(\sum_{k=1}^n x_{ki}^2 - \frac{1}{n}(\sum_{k=1}^n x_{ki})^2)} \quad (6)$$

Суммирование выполняется по всем n молекулам обучающего набора данных. Если коэффициента корреляции $|R_{yi}|$ моделируемого свойства с дескриптором x_i ниже заданного порога R_{yi}^0 , то дескриптор исключается для дальнейшего рассмотрения в качестве независимой переменной уравнения МЛР.

Квадрат коэффициента корреляции R_{ij}^2 между дескрипторами x_i и x_j рассчитывается по формуле ^[27,28]:

$$R_{ij}^2 = \frac{s_{ij}^2}{s_{ii} s_{jj}} = \frac{\{\sum_{k=1}^n x_{ki} x_{kj} - \frac{1}{n}(\sum_{k=1}^n x_{ki})(\sum_{k=1}^n x_{kj})\}^2}{(\sum_{k=1}^n x_{ki}^2 - \frac{1}{n}(\sum_{k=1}^n x_{ki})^2)(\sum_{k=1}^n x_{kj}^2 - \frac{1}{n}(\sum_{k=1}^n x_{kj})^2)} \quad (7)$$

Если коэффициента корреляции $|R_{ij}|$ между дескрипторами x_i и x_j выше заданного порога R_{ij}^0 , то один из них исключается из рассмотрения. Исключается менее информативный молекулярный фрагмент, либо более короткий, либо лексикографически более младший. Например, среди фрагментов С-С-С-О и С-С-С будет исключен второй, а среди С-С-С-О и N-С-С-О – первый.

Величины R_{yi}^0 и R_{ij}^0 в программе могут быть заданы пользователем, по умолчанию $R_{yi}^0 = 10^{-3}$ и $R_{ij}^0 = 0.99$.

б) прямой пошаговый отбор переменных

На этой стадии происходит накопление наиболее значимых дескрипторов, в определенной мере коррелирующих со свойством. ISIDA_QSPR имеет несколько таких алгоритмов, обозначенных как группа Forward variable selection (FVS).

В алгоритме *max t-Test*, основанном на максимальной величине критерия Стьюдента *t-test*, строится линейное регрессионное уравнение $y \approx \hat{y} = c_0 + c_s x_s$, связывающее известное свойство y с таким дескриптором x_s , который обеспечивает максимальный критерий $t-test = c_s / \delta c_s$, где δc_s – стандартное отклонение коэффициента c_s , \hat{y} – свойство, рассчитанное согласно приведенному регрессионному уравнению. Найденный дескриптор x_s запоминается, из величины свойства y обучающего набора вычитается соответствующее значение уравнения $\hat{y} = c_0 + c_1 x^{(s)}$: $\Delta y = y - \hat{y}$ согласно [29], затем в качестве y берется остаток Δy , и вновь строится подобное линейное регрессионное уравнение с целью поиска следующего лучшего дескриптора. Этот цикл повторяется заданное пользователем число раз $m = 0.1n, 0.2n, \dots, 0.8n$, где n – размер обучающей выборки. Поскольку один и тот же дескриптор может выбран алгоритмом несколько раз, то может выбрано не более m дескрипторов. Расчет коэффициентов c_0 , c_s и стандартного отклонения δc_s выполняется с помощью алгоритма сингулярного разложения [24-26].

Три других эффективных алгоритма [10] основаны на расчете максимального коэффициента корреляции R_{yi}^2 , R_{yij}^2 или R_{yijk}^2 , соответственно, между моделируемым свойством и одной, двумя или тремя дескрипторами. Эти коэффициенты корреляции рассчитываются аналитически, не вычисляя коэффициентов большого числа МЛР уравнений, что ускоряет процесс отбора переменных. В выпадающем списке Forward variable selection программы ISIDA_QSPR эти три алгоритма обозначены как $R_{y,i}$, $R_{y,ij}$ и $R_{y,ijk}$. Псевдокод алгоритма $R_{y,i}$ прямого пошагового отбора переменных:

известно N переменных x и свойство y для n объектов
задано число m выбираемых переменных, при этом $N > m$
 $\Delta y^{(1)} = y; i = 0;$
цикл 1: пока не выбрано m переменных x

$i + +;$

цикл 2: среди всех N переменных x выбрать такую x_s , что

коэффициент корреляции $R_{y_s}^2$ между $\Delta y^{(i)}$ и x_s максимален

конец цикла 2

вычислить коэффициенты c_0 и c_s уравнения $\hat{y}^{(i)} = c_0 + c_s x_s$

$\Delta y^{(i+1)} = \Delta y^{(i)} - \hat{y}^{(i)}$

запомнить переменную x_s

если за заданное число циклов не выбрано новой переменной,

то выйти из цикла

конец цикла 1.

Алгоритмы $R_{y,ij}$ и $R_{y,ijk}$ отличаются выполнением цикла 2 и уравнением для $\hat{y}^{(i)}$:

алгоритм $R_{y,ij}$:

цикл 2: среди всех N переменных x выбрать две такие x_s и x_t , что

коэффициент корреляции $R_{y_{st}}^2$ между $\Delta y^{(i)}$ и (x_s, x_t) максимален

конец цикла 2

вычислить коэффициенты c_0 , c_s и c_t уравнения $\hat{y}^{(i)} = c_0 + c_s x_s + c_t x_t$

алгоритм $R_{y,ijk}$:

цикл 2: среди всех N переменных x выбрать три такие x_s , x_t и x_u , что

коэффициент корреляции $R_{y_{stu}}^2$ между $\Delta y^{(i)}$ и (x_s, x_t, x_u) максимален

конец цикла 2

вычислить коэффициенты c_0 , c_s , c_t и c_u уравнения $\hat{y}^{(i)} = c_0 + c_s x_s + c_t x_t + c_u x_u$

Число выбираемых дескрипторов m задается пользователем: $m = 0.1n, 0.2n, \dots, 0.8n$, где n – размер обучающей выборки. По умолчанию $m = 0.6n$, т.е. 60% от n .

Расчет коэффициента корреляции $R_{y_i}^2$ представлен формулой (6), коэффициенты корреляции $R_{y_{ij}}^2$ и $R_{y_{ijk}}^2$ рассчитываются по формулам [27,28]:

$$R_{y_{ij}}^2 = \frac{R_{y_i}^2 + R_{y_j}^2 - 2R_{y_i}R_{y_j}R_{ij}}{1 - R_{ij}^2}$$

$$R_{yijk}^2 = 1 - (1 - R_{yi}^2)(1 - R_{yj.i}^2)(1 - R_{yk.i}^2)$$

где

$$R_{yk.i}^2 = \frac{(R_{yk.i} - R_{yj.i} R_{jk.i})^2}{(1 - R_{yj.i}^2)(1 - R_{jk.i}^2)}$$

$$R_{yj.i} = \frac{R_{yj} - R_{yi} R_{ij}}{\sqrt{(1 - R_{yi}^2)(1 - R_{ij}^2)}}$$

$$R_{jk.i} = \frac{R_{jk} - R_{ij} R_{ik}}{\sqrt{(1 - R_{ij}^2)(1 - R_{ik}^2)}}$$

в) обратное пошаговое исключение переменных

На предыдущей стадии было выбрано m дескрипторов последовательно по одному, двум или трем. Теперь необходимо проверить, как они все вместе коррелируют со свойством, и удалить те дескрипторы, вклады которых в уравнение МЛР имеют наибольшую относительную погрешность. Строится линейное регрессионное уравнение $y \approx \hat{y} = c_0 + \sum_{i=1}^m c_i x_i$, включающее все m фрагментных дескрипторов, выбранных на предыдущей стадии. Определяется дескриптор c_e , для которого относительная ошибка $\delta c_e / c_e$ превышает заданный порог, т.е. расчетный критерий Стьюдента $t_{calc} = c_e / \delta c_e > t-test$ для данного коэффициента c_e превышает заданное (табличное) значение $t-test$. Этот дескриптор удаляется, регрессионное уравнение с оставшимися дескрипторами строится заново. Этот цикл повторяется до тех пор, пока для всех оставшихся k дескрипторов расчетный критерий Стьюдента t_{calc} не станет ниже заданного: $t_{calc} = c_e / \delta c_e \leq t-test$. Расчет коэффициентов c_i и их стандартных отклонений $\delta c_i, i = 0, 1, \dots, m$ выполняется с помощью алгоритма сингулярного разложения [24-26]. Величина $t-test$ выбирается пользователем, по умолчанию $t-test = 1.96$. Влияние $t-test$ на качество моделей изучено в интервале $t-test = 1.96 - 3.3$. Как правило, его оптимальное значение лежит в интервале $t-test = 1.96 - 2.7$.

На панели **Model** (рис. 4) проверьте, что установлены флажки (отмечены галочкой) **forward variable selection** и **backward variable selection** области **Variable Selection**. Выберите алгоритм $R_{y,ij}$ из выпадающего списка комбинированного окна

алгоритмов *forward variable selection*. Справа от комбинированного окна введите 60 в поле редактирования для количества предварительно отобранных переменных в процентах от размера обучающего набора данных. Введите 0.001 в поле редактирования R_{y_i} и 0.99 в поле редактирования R_{ij} для пороговых значений коэффициентов корреляции. Введите 1.96 в поле редактирования *t-test*, 1E-12 в поле редактирования *eps* и 0 в поле редактирования *exact*. Убедитесь, что флажок (галочка) $m > n$ не установлен (рис. 4).

1.6. Параметры внутреннего и внешнего контроля качества модели

Внутренний контроль качества модели выполняется с использованием обучающей выборки. Применяется разновидность перекрестного контроля (cross-validation) - метод отбрасывания по одному – скользящий контроль (Leave One Out, LOO) или по нескольку (Leave Many Out, LMO) объектов. В процедуре метода LOO (LMO) каждое соединение (k соединений) обучающей выборки однократно исключается, регрессионные коэффициенты МЛР модели рассчитываются на основе остающихся соединений, для исключенного соединения (k исключенных соединений) выполняется предсказание \hat{y} , применяя эту модель. Далее, используя величины y и \hat{y} , вычисляют квадрат коэффициента детерминации скользящего контроля Q^2 :

$$Q^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \langle y \rangle)^2} \quad (8)$$

Здесь y_i и \hat{y}_i – соответственно, известное и предсказанное значения свойства для i -го соединения обучающей выборки, n - число точек. $\langle y \rangle = \frac{1}{n} \sum_{i=1}^n y_i$ - среднее значение известного свойства. Необходимое условие для предсказательной модели $Q^2 > 0.5$ [30], наивысшее значение Q^2 равно 1.

Внешний контроль качества модели выполняется с использованием тестовой выборки, никак не участвующей в построении модели и отборе дескрипторов. Модель, разработанная на обучающем наборе данных, применяется к тестовому набору для предсказания моделируемого свойства. Для оценки качества предсказаний вычисляют такие важные статистические характеристики, как квадрат коэффициента детерминации R_{det}^2 , среднеквадратичная ошибка *RMSE* и средняя абсолютная ошибка *MAE*:

$$R_{det}^2 = 1 - \frac{\sum_{i=1}^m (y_i - \hat{y}_{pred,i})^2}{\sum_{i=1}^m (y_i - \langle y \rangle)^2} \quad (8)$$

$$RMSE = \left[\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2 \right]^{1/2} \quad (9)$$

$$MAE = \frac{1}{m} \sum_{i=1}^m |y_i - \hat{y}_i| \quad (10)$$

Здесь y_i и $\hat{y}_{pred,i}$ – соответственно, известное (экспериментальное) и предсказанное для тестовой выборки значения свойства для i -го соединения, m - число точек тестовой выборки, $\langle y \rangle = \frac{1}{m} \sum_{i=1}^m y_i$ - среднее значение экспериментального свойства тестовой выборки. Необходимое условие для предсказательной модели $R_{det}^2 > 0.5$, наивысшее значение R_{det}^2 равно 1. $RMSE$ предсказаний в определенной мере должна соответствовать экспериментальной погрешности в оценке моделируемого свойства.

На панели **Validation** (рис. 4) установите флажок (отметьте галочкой) **LOO** для вычисления коэффициента корреляции метода отбрасывания по одному объекту. Убедитесь, что флажок **rapid** не установлен. Установите флажок **LMO** для расчета коэффициента корреляции метод отбрасывания по нескольку объектов. Введите 5 в поле редактирования ***i-th point*** для вычислений LMO: каждый пятый объект исключается один и только один раз, по оставшимся данным строится модель, и только исключенные молекулы предсказываются по этой модели. Введите 1 в поле редактирования ***External n-fold CV***, что означает выполнение моделирования без внешнего k -кратного перекрестного контроля.

1.7. Область применимости модели

Область применимости модели (Applicability Domain, AD) определяет область химического пространства, в котором модель предполагается точной в пределах ее стандартного отклонения. Три определения AD могут использоваться в ISIDA_QSPR одновременно или по отдельности: 1) контроль неизвестных фрагментов, 2) область изменения фрагментных дескрипторов^[20] и 3) "кворум-контроль"^[11]. Контроль фрагментов заключается в отбрасывании предсказаний для соединений, содержащих молекулярные фрагменты, не встречающиеся в начальном наборе СМФ, сгенерированном для обучающей выборки (AD1). Контроль области изменения (Bounding Box) рассматривает AD как многомерное дескрипторное пространство, ограниченное минимальными и максимальными значениями дескрипторов, задействованных в индивидуальной модели (AD2). "Кворум-контроль" (AD3) - это порог для количества моделей, принятых AD1 и

AD2. Если это число меньше заданного пользователем порога, консенсус-предсказание игнорируется. AD3 применяется при построении ансамбля QSPR моделей для расчета средних предсказанных значений, исключая выбросы (*консенсус-модель*).

На панели **Validation** (рис. 4) установите флажки **Appl. Domain1** и **Appl. Domain2** соответственно для контроля новых фрагментов и области их изменения в тестируемых молекулах, чтобы учесть область применимости модели.

1.8. Сохранение и загрузка результатов моделирования

Файлы результатов моделирования сохраняются в выбранном пользователем каталоге с помощью кнопки **Output Directory** в нижнем левом углу (рис. 4). Появится диалоговое окно выбора каталога, в котором щелчком по кнопке **Open** выбирается открытый каталог, например, D:\ISIDA_QSPR\RESULTS.

В дальнейшем файлы результатов можно открыть снова, выбрав File → Open главного меню ISIDA_QSPR (рис. 1) и открыв файл *.OUT, имя которого начинается с даты и времени проведения расчетов и включает имя входного SDF файла.

1.9. Результаты расчетов и основные статистические характеристики индивидуальной QSPR модели

Теперь для выполнения расчетов кликните на кнопке **Start** диалогового окна **Single Model Calculations** (рис. 4). ISIDA_QSPR создаст и откроет 9 файлов: 4 текстовых файла и 5 файлов графического представления результатов.

Первый текстовый файл _MODEL-...TXT (выбор в главном меню ISIDA_QSPR: Window → _MODEL-...TXT) содержит информацию, касающуюся QSPR модели:

- a) полный начальный список СМФ дескрипторов;
- b) исходные параметры QSPR моделирования, включая имена входных файлов и моделируемое свойство;
- c) группы связанных фрагментов, всегда встречающиеся в одной и той же комбинации в ряде соединений обучающей выборки;
- d) статистические характеристики множественной линейной регрессии в качестве QSPR модели: коэффициент корреляции Пирсона R , критерий Фишера F , среднеквадратичная

ошибка *RMSE*, средняя абсолютная ошибка *MAE*, коэффициент детерминации скользящего контроля *Q* и др.;

e) СМФ дескрипторы МЛР модели, коэффициенты уравнения регрессии a_i - вклады СМФ и их случайные ошибки Δa_i для 95% доверительного интервала;

f) матрица парных корреляций для вкладов СМФ;

g) сингулярные числа s_i сингулярного разложения (SVD) матрицы дескрипторов, отношения между которыми является критерием устойчивости решения: например, если $s_{max}/s_{min} > 10^8$, то это редкий сигнал о том, что некоторые дескрипторы следует исключить из модели;

h) таблица экспериментальных (Y_{exp}) и расчетных (Y_{calc}) согласно МЛР модели (fitted) значений моделируемого свойства и остатков $Y_{exp} - Y_{calc}$ для обучающего набора данных.

Второй текстовый файл *.SMF содержит полный набор сгенерированных СМФ дескрипторов и матрицу их значений для молекул обучающей выборки:

Full Set of Fragments.

1.	C-O
2.	C=C
3.	C-C
4.	C-N
5.	C=O
6.	C=C-N
7.	C=C-O

...

MATRIX: Compound (Line) x Number_of_Fragment (Column).

	1	2	3	4	5	6	7	8	9	10	11	12...
1	1	2	3	3	1	2	1	1	1	3	2	3...
2	1	1	2	0	1	0	0	0	1	2	1	0...
4	2	0	2	2	2	0	0	2	0	0	2	1...
5	2	0	6	0	1	0	0	0	7	0	1	0...
6	4	0	2	0	2	0	0	0	1	0	2	0...
7	4	0	6	2	4	0	0	4	4	0	4	1...

...

Третий текстовый файл *.MF содержит список СМФ дескрипторов (независимых переменных) модели и матрицу их значений:

Set of Fragments for the Model.

2.	C=C
11.	C-C=O
18.	C-C-N-C
32.	C-C-C=O
34.	N-C-C-O
37.	C-C-N-C-C
38.	C-C-N-C-C-O

...

MATRIX: Compound (Line) x Number_of Fragment (Column).

	2	11	18	32	34	37	38	44	58...
1	2	2	2	0	0	0	0	0	0...
2	1	1	0	1	0	0	0	0	0...
4	0	2	2	0	2	1	2	0	0...
5	0	1	0	2	0	0	0	1	0...
6	0	2	0	2	0	0	0	4	0...
7	0	4	4	4	2	4	4	0	2...

...

Четвертый текстовый файл *_Pred.DOC содержит таблицу предсказанных свойств (правая колонка с названием модели, здесь IAB2-120) для соединений тестового набора, определенного MASK файлом. Колонка Datum содержит экспериментальные данные. Если соединение определено как находящееся за пределами области применимости модели (область задана, используя AD1 и AD2), предсказанное значение для этого соединения исключается (см. соединение под номером 18):

TABLE P1. Test set: Predicted property LogK for the compounds from the GdL_logK_TRAINING.SDF file.

cmp. no.	Datum	IAB2-120
3	2.4	2.25
8	17.54	17.43
13	4.57	4.84
18	4.43	
23	19.17	20.62

...

Первый графически представленный файл Residual_Analysis_...* включает анализ остатков: график зависимости разности ($Y_{exp} - Y_{calc}$) от рассчитанного (fitted) свойства (Y_{calc}) для соединений обучающего набора с известным (экспериментальным) свойством Y_{exp} (рис. 9). Красная пунктирная линия соответствует нулевому отклонению. Чтобы увидеть остаток ($Y_{exp} - Y_{calc}$) и соответствующую молекулярную структуру, наведите указатель мыши на точку данных (кружок) и нажмите. Структура появляется на желтом фоне. Номер записи в файле SDF, значения Y_{exp} , Y_{calc} и ($Y_{exp} - Y_{calc}$) отобразятся в нижней строке окна программы (рис. 9).

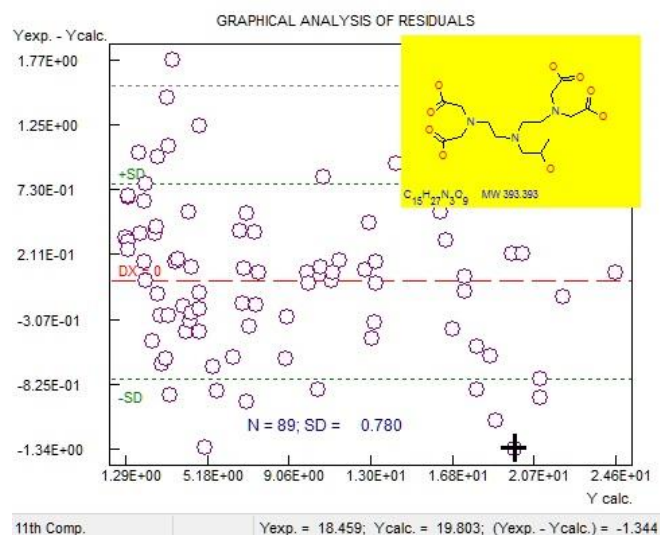


Рис. 9. Графическое окно анализа остатков.

Второй график PLOT_Calc_vs_Exp... отображает для обучающей выборки сопоставление известных Y_{exp} и рассчитанных (fitted) Y_{calc} величин свойства в форме линейной взаимосвязи $Y_{calc} = a + b \cdot Y_{exp}$ с указанием числа данных (n), квадрат коэффициента детерминации (R_{det}^2), среднеквадратичной ошибки ($RMSE$) и средней абсолютной ошибки (MAE) (рис. 10 а).

На следующих двух графиках PLOT_LOO... и PLOT_LMO... представлено сравнение экспериментальных Y_{exp} и предсказанных Y_{pred} величин свойства в форме линейной взаимосвязи $Y_{pred} = a + b \cdot Y_{exp}$ для методов отбрасывания по одному (LOO) или по пять (LMO) на обучающей выборке (рис. 10 б и в). Для предсказательной модели коэффициенты a и b этих двух уравнений и для модели $Y_{calc} = a + b \cdot Y_{exp}$ очень близки (рис. 10 а, б и в).

На пятом графике PLOT_Predict... выполнено сравнение экспериментальных Y_{exp} и предсказанных Y_{pred} величин моделируемого свойства в форме линейной взаимосвязи $Y_{pred} = a + b \cdot Y_{exp}$ для соединений внешнего тестового набора, сформированного с помощью MASK файла. Для соединений, идентифицированных как находящиеся за пределами области применимости (AD) модели, предсказанные значения исключены (рис. 10 г).

Молекулярную структуру, соответствующую точке данных на графиках, можно визуализировать, щелкнув по выбранной точке данных. Структура отображается на желтом фоне; номер структуры в файле SDF, значения Y_{exp} и Y_{calc} (Y_{pred}) появляются в строке состояния в нижней части окна программы.

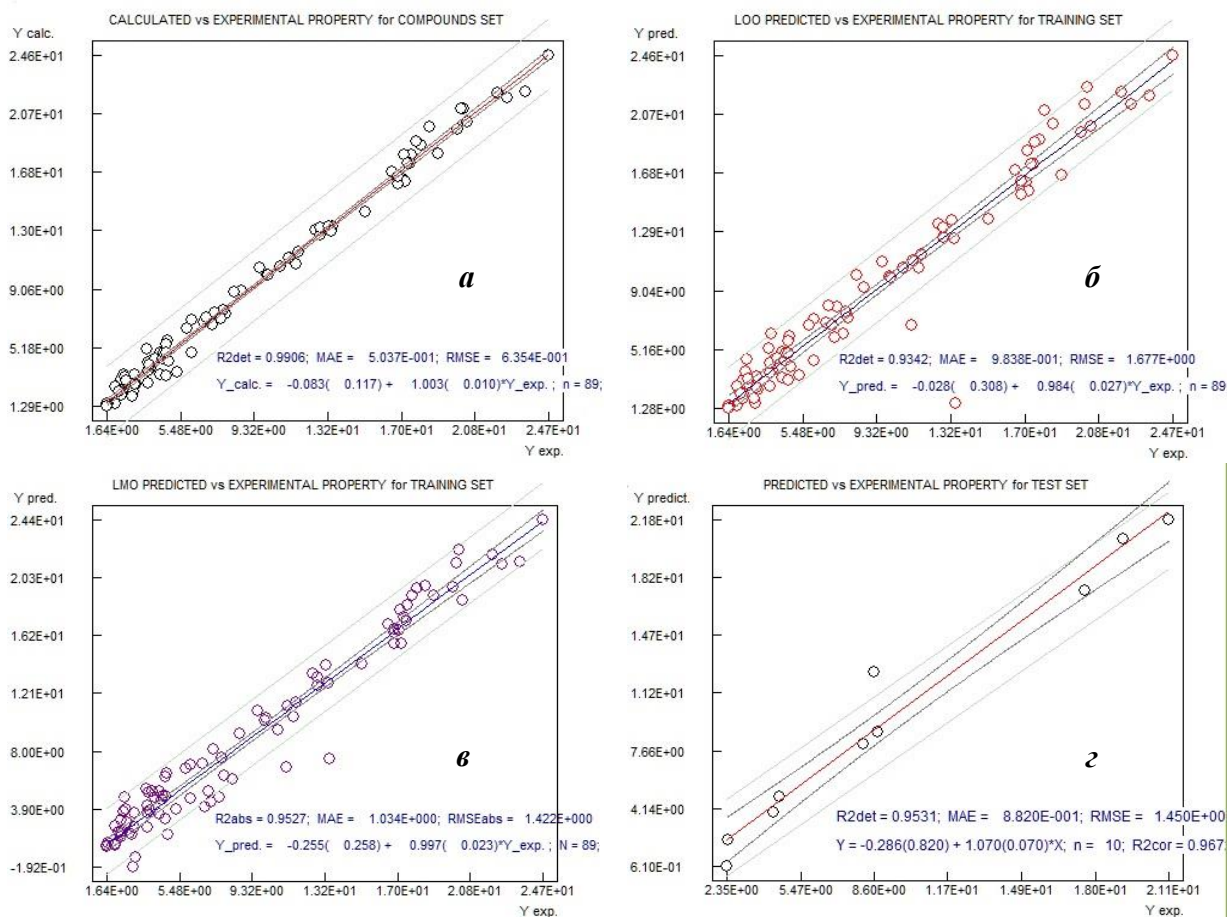


Рис. 10. Моделирование константы устойчивости ($Y = \log K$) комплексов $Gd^{3+}L$ иона гадолиния Gd^{3+} с органическими молекулами L в воде. Четыре графика отображают сопоставление экспериментальных Y_{exp} , рассчитанных (fitted) Y_{calc} и предсказанных Y_{pred} величин в форме линейной взаимосвязи Y_{calc} (Y_{pred}) = $a + b \cdot Y_{exp}$, представленной для расчетных и экспериментальных констант устойчивости для модели (а), расчетных и предсказанных констант устойчивости для методов отбрасывания по одному LOO (б) или по пять LMO (в) на обучающей выборке, расчетных и предсказанных констант устойчивости для внешней тестовой выборки (г).

1.9. Среднеквадратичная ошибка: подгонка, внутренний и внешний перекрестный контроль

Среднеквадратичная ошибка $RMSE = \left[\frac{1}{n} \sum_{i=1}^n (Y_{exp,i} - Y_i)^2 \right]^{1/2}$ характеризует способность модели количественно воспроизводить экспериментальные данные, где Y_i – рассчитанное (подогнанное, fitted) $Y_{calc,i}$ или предсказанное $Y_{pred,i}$ значение свойства для i -той точки данных. Как правило, значения $RMSE$ увеличиваются в порядке $RMSE$ (модель) < $RMSE$ (LOO) < $RMSE$ (LMO) < $RMSE$ (внешний тестовый набор или k -кратный внешний

перекрестный контроль), как показано на рисунке 10. Причина очевидна: для расчета *RMSE* (модель) программа применяет QSPR модель к соединениям, использованным для построения этой модели, тогда как в других процедурах программа на этапе построения модели лишь частично (в LOO и LMO на этапах выбора переменных) или никогда (во внешнем тестовом наборе или *k-fold CV*) не использует информацию о предсказываемых соединениях. В частных случаях возможны отклонения от приведенного неравенства: на рис. 10 *g* для внешнего набора данных *RMSE* оказалась ниже *RMSE* (LOO) и *RMSE* (LMO) вследствие достаточно строгого определения области применимости модели (см. AD1 и AD2), что привело к надежному предсказанию свойства, но лишь для 10 из 22 тестируемых соединений (см. файл *_Pred.DOC).

УПРАЖНЕНИЕ 2

2. Анализ молекулярных фрагментов индивидуальной МЛР модели

Задача:

Это упражнение является второй частью предыдущего упражнения. Вслед за выполненными расчетами индивидуальной модели (см. Упражнение 1) открывается подпрограмма MolFrag, которая выводит подробную информацию о молекулярных фрагментах модели. Присутствие определенных структурных мотивов увеличивает или уменьшает величину моделируемого свойства. Эти данные полезны при конструировании новых соединений с желаемыми свойствами. Инструмент MolFrag как окно Statistics of Substructural Molecular Fragments (рис. 11) предоставляет пользователю:

- a) список фрагментных дескрипторов и их вкладов, минимальную и максимальную кратность фрагментов в соединениях обучающей выборки: вкладка *Model Parameters*;
- b) оценку парного молекулярного сходства на основе фрагментов, участвующих в модели: вкладка *Similarity*;
- c) парная корреляция вкладов фрагментов: вкладка *Correlations*;
- d) список фрагментов и их значений, участвующих в модели, и их вклады для каждой молекулы обучающей выборки: вкладка *SMF table*.

Некоторые подробности по использованию этих опций приведены ниже.

File Edit Help

Type of SMF: IAB2-12
30 fragments

SMF Table Correlations Similarity Model Parameters

File: GdL_logK_TRAINING.LRN
Total number of SMF: 1005

id.	name	contrib.	SD	min	max	mols
1	C-O	0.0				
2	C=C	1.37011	0.21990	0	3	4
3	C-C	0.0				
4	C-N	0.0				
5	C=O	0.0				
6	C=C-N	0.0				
7	C=C-O	0.0				
8	C-C-N	0.0				
9	C-C-C	0.0				
10	C-C=C	0.0				
11	C-C=O	1.41641	0.10681	0	5	76
12	C-N-C	0.0				
13	C-C-O	0.0				
14	C-C=C-O	0.0				

Concatenated SMF: group 3 of 121

no.	id.	count	name	contrib.	main
1	59	2	C-N-C-C-C=O	-0.55818	+
2	56	2	C-N-C-C-C-O	0.00000	

Рис. 11. Графический интерфейс программы MolFrag.

На вкладке **Model Parameters** отображаются две таблицы (рис. 11). Верхняя показывает полный список СМФ дескрипторов, сгенерированных на основе обучающей выборки, в ней для каждого молекулярного фрагмента, вошедшего в МЛР модель, указано: идентификационный номер (*id*), название в форме обозначений атомов и связей, вклад - соответствующий коэффициент МЛР модели (*contrib.*) и его стандартное отклонение (*SD*), минимальное (*min*) и максимальное (*max*) значение дескриптора в соединениях обучающей выборки, а также число соединений, содержащих данный фрагмент - дескриптор (*mols*). Нижняя таблица содержит группы взаимосвязанных фрагментов, которые всегда встречаются в одинаковой комбинации в ряде соединений обучающего набора. Главный фрагмент (самый длинный или лексикографически старший по порядку) в группе обозначен знаком "+" в столбце *main*. Кнопки навигации внизу используются для просмотра групп фрагментов (рис. 11).

На вкладке **Similarity** отображаются коэффициенты Танимото (*TC*) сходства каждой пары молекул обучающего набора, рассчитанные с помощью фрагментов, задействованных в модели. Пользователь может ввести пороговое значение *TC* в поле редактирования *TC*,

затем нажать кнопку **Mark**, чтобы выделить зеленым фоном коэффициенты Танимото, превышающие этот порог. Мало различающиеся структурные формулы имеют коэффициент TC близкий к единице. Например, значение $TC = 0.9127$ ячейки матрицы выделено зеленым фоном, указывая на то, что соответствующие соединения 21 (строка mol 21) и 11 (столбец mol 11) достаточно сходны и имеют близкие значения моделируемого свойства: 16.48 и 18.48. В программе заложены две формулы расчета коэффициенты Танимото для оценки сходства молекул A и B :

а) применение к фрагментным дескрипторам формулы для непрерывных переменных (флажок **Binary TC** снят)

$$TC = \frac{\sum_{i=1}^n A_i \cdot B_i}{\sum_{i=1}^n A_i \cdot A_i + \sum_{i=1}^n B_i \cdot B_i - \sum_{i=1}^n A_i \cdot B_i} \quad (11)$$

б) применение к фрагментным дескрипторам формулы для бинарных переменных (флажок **Binary TC** включен)

$$TC = \frac{\sum_{i=1}^n \min(A_i, B_i)}{\sum_{i=1}^n A_i + \sum_{i=1}^n B_i - \sum_{i=1}^n \min(A_i, B_i)} \quad (12)$$

Здесь A_i и B_i – число вхождений i -того фрагмента, соответственно, в молекулах A и B , n – число независимых переменных (различных фрагментов) модели. Читатель может сравнить структуры и соответствующие коэффициенты Танимото, вычисленные с использованием приведенных двух формул, чтобы сделать заключение, какая из формул более корректна.

На вкладке **Correlations** отображается матрица парных корреляций дескрипторов, включенных в модель. Пользователь может ввести пороговое значение коэффициента корреляции $|R|$ в поле редактирования **R >**, затем щелкнуть на кнопке **Mark**, чтобы увидеть в открывшемся окне **Correlated Fragments...** все пары фрагментных дескрипторов, для которых коэффициент корреляции превышает этот порог. Отметим, что топологические пути являются слабо коррелирующими дескрипторами, способствуя низким погрешностям их вкладов - коэффициентов МЛР модели. Затем закройте окно **Correlated Fragments**. Щелчок на любой ячейке матрицы с указанным коэффициентом корреляции для пары дескрипторов с их номерами и названиями, данными на желтом фоне, открывает окно **...-th fragment count versus ...-th fragment count**, в котором отображается линейное уравнение корреляции между ними, указаны статистические параметры уравнения: квадрат коэффициента корреляции (R^2_{cor}), критерий Фишера (F) и стандартное отклонение (s).

Вкладка **SMF Table** отображает для обучающей выборки матрицу молекулярных фрагментов, являющихся дескрипторами модели. При щелчке на ячейке левого столбца, соответствующей определенной молекуле (например, на ячейке *mol 1*), открывается окно **Fragment Contributions** с изображением молекулярной структуры с таблицей фрагментов и их вкладов в моделируемое свойство согласно модели (рис. 12).

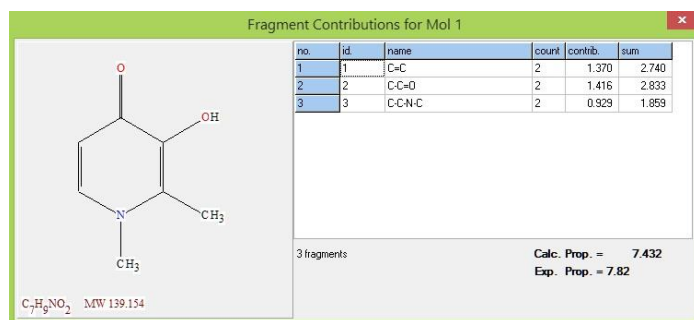


Рис. 12. Соединение обучающей выборки, его фрагменты, входящие в модель, и их вклады в моделируемое свойство.

УПРАЖНЕНИЕ 3

3. Внешний k -кратный перекрестный контроль: построение и проверка модели

Задача:

Целью данного упражнения является реализация внешнего пятикратного перекрестного контроля (5-fold external cross-validation, 5-fold CV) для демонстрации стандартного протокола для оценки предсказательной эффективности QSPR модели.

Процедура k -кратного внешнего перекрестного контроля (external k -fold cross-validation, k -fold CV) часто используется [7,14,31] в качестве стандартного протокола для оценки предсказательной эффективности модели. Согласно этой процедуре, весь набор, включающий n точек данных, разбивается на k непересекающихся пар обучающих и тестовых наборов. В каждой паре обучающий набор охватывает $(k - 1)/k$ часть данных, а соответствующее тестовый набор охватывает оставшуюся $1/k$ часть данных. Тестовый набор никак не участвует в построении модели и отборе дескрипторов. Построив модели k раз, предсказания для всех k тестовых наборов объединяются, и таким образом предсказываются все точки данных во всем наборе данных, для которых вычисляют квадрат

коэффициента детерминации R_{det}^2 , среднеквадратичную ошибку $RMSE$ и среднюю абсолютную ошибку MAE выполненного внешнего перекрестного контроля:

$$R_{det}^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \langle y \rangle)^2} \quad (13)$$

$$RMSE = \left[\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right]^{1/2} \quad (14)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (15)$$

Здесь y_i и \hat{y}_i – соответственно, экспериментальное и предсказанное на внешних выборках значение свойства для i -го соединения, $\langle y \rangle = \frac{1}{n} \sum_{i=1}^n y_i$ – средняя величина экспериментального свойства. Необходимое условие для предсказательной модели $R_{det}^2 > 0.5$, наивысшее значение R_{det}^2 равно 1. Чем больше k , тем больше соответствующий обучающий набор, что означает, что шанс встретить на этапе обучения соединения, похожие на тестовые молекулы, увеличивается. Таким образом, чем больше k , тем более "оптимистичными" становятся результаты перекрестного контроля. Перекрестный контроль при $k = 2$ требует от модели, обученной на половине исходного набора, предсказать другую половину, и поэтому может быть слишком пессимистичный, если только не используются очень большие наборы данных. С другой стороны, перекрестный контроль при $k = n$ (исключение по одному, LOO) определенно слишком оптимистичен, он генерирует QSPR уравнения, наиболее близкие к тем, которые можно было бы получить на основе всего набора. Обычно используют $k = 3, 5, 10$ или 20 . Значения $k = 10$ или 20 часто применяют для небольших выборок с $n \sim 50 \dots 100$.

3.1. Настройка параметров

Задайте параметры для моделирования, как указано выше в Упражнении 1 в разделах 1.1 - 1.8. Затем щелкните на кнопке **Create MASK** диалогового окна расчета индивидуальной (единичной) модели (рис. 4). Появится диалоговое окно Create Mask (рис. 5), где щелкните на радиокнопке **Create MASK No test set**, затем щелкните на кнопке **START**, чтобы сохранить или перезаписать файл-маску **GdL_logK_TRAINING.MSK** обязательно в каталоге ISIDA_QSPR с помощью диалогового окна Save mask file Dialog. Щелкните на кнопке **Data TEST** для проверки входных данных (рис. 4). Если данные корректны, то в диалоговом окне **Information** появится

сообщение: "Файлы входных данных находятся во внутреннем согласии". Закройте диалоговое окно *Information*.

На панели *Validation* (рис. 4) введите **5** в поле редактирования *External n-fold CV*, чтобы выполнить моделирование с внешним 5-кратным перекрестным контролем (5-fold CV).

3.2. Результаты расчетов и статистические параметры внешнего пятикратного перекрестного контроля

Щелкните на кнопке *Start* диалогового окна *Single Model Calculations* (рис. 4) для выполнения расчетов. Программа создаст и откроет 6 выходных файлов: 4 текстовых файла и 2 файла графического представления результатов.

Первый текстовый файл *.ТОМ содержит таблицу статистических параметров пяти индивидуальных МЛР моделей для каждого блока внешнего 5-кратного перекрестного контроля (5-fold CV):

5-Fold External Cross-Validation Procedure.

Th 19/5/2022 14:06:32.313

File of Mol Structures: GdL_logK_TRAINING.SDF; 111 compounds in training set.

Modeling Property Name: LogK

Mask File: GdL_logK_TRAINING.MSK

Exter.Descriptors File: -

...

no	fragment	fitting	n	k	R2	F	testR2det	testRMSE	HRF	Q2
	type	equation								
1	IAB2-12	0	88	29	0.987216	162.72	0.923	1.80E+00	6.520	0.961740
2	IAB2-12	0	89	27	0.985314	159.99	0.960	1.53E+00	6.839	0.959954
3	IAB2-12	0	89	30	0.990562	213.54	0.953	1.45E+00	5.627	0.934244
4	IAB2-12	0	89	26	0.984162	156.59	0.810	2.31E+00	7.133	0.956273
5	IAB2-12	0	89	25	0.984889	173.81	0.950	1.38E+00	7.107	0.968771

В таблице приведены следующие статистические параметры:

- число точек n в обучающей выборке,
- число коэффициентов (фрагментов) k уравнения МЛР,
- квадрат коэффициента корреляции Пирсона R^2 ,
- критерий Фишера F ,
- квадрат коэффициента детерминации $testR_{det}^2$ для тестовой выборки,
- среднеквадратичная ошибка $testRMSE$ для тестовой выборки,
- R -фактора Гамильтона ^[32] в процентах, HRF ,
- квадрат коэффициента детерминации скользящего контроля Q^2 для обучающей выборки.

Второй текстовый файл *.DOC содержит таблицу экспериментальных и рассчитанных величин свойства индивидуальных МЛР моделей для каждого блока внешнего 5-кратного перекрестного контроля:

```
Date...
File of Mol Structures: GdL_logK_TRAINING.SDF
Property Name:          LogK
Mask File:              GdL_logK_TRAINING.MSK
Exter.Descriptors File: -
...
5-Fold External Cross-Validation procedure.
```

TABLE L1. Training set: Calculated property LogK for the compounds from the GdL_logK_TRAINING.SDF file.

cmp. no.	Exp.	1 IAB2-120	2 IAB2-120	3 IAB2-120	4 IAB2-120	5 IAB2-120
1	7.82		7.86	7.43	7.45	6.72
2	2.39	3.58		3.05	1.76	2.34
3	2.40	2.85	2.02		3.92	2.64
4	6.68	7.11	7.04	6.87		7.30
5	2.84	1.57	1.38	2.21	1.76	
6	4.07		4.02	4.39	3.52	2.99
7	11.20	11.20		11.14	9.82	11.02

...

Третий текстовый файл *.AVE содержит таблицу средних величин (Average) свойства и его стандартного отклонения (STDEV) согласно моделям процедуры 5-fold CV по данным файла *.DOC, указанного выше:

```
File of Mol Structures: GdL_logK_TRAINING.SDF
Property Name:          LogK
Mask File:              GdL_logK_TRAINING.MSK
Exter.Descriptors File: -
...
TABLE LA. Training set: Average calculated property LogK for the compounds from the
GdL_logK_TRAINING.SDF file using models selected by ISIDA_QSPR.
```

cmp. no.	Exp.	Average	STDEV	Nm	Exp.- Ave.
1	7.82000E+000	7.36500E+000	4.735E-001	4	4.550E-001
2	2.39000E+000	2.68250E+000	7.977E-001	4	-2.925E-001
3	2.40000E+000	2.85750E+000	7.911E-001	4	-4.575E-001
4	6.68000E+000	7.08000E+000	1.780E-001	4	-4.000E-001
5	2.84000E+000	1.73000E+000	3.556E-001	4	1.110E+000
6	4.07000E+000	3.73000E+000	6.087E-001	4	3.400E-001
7	1.12000E+001	1.07950E+001	6.543E-001	4	4.050E-001

...

Четвертый текстовый файл *.ECV содержит таблицу предсказанных величин свойства (колонка IAB2-120) для соединений пяти тестовых наборов процедуры 5-fold CV. Колонка Datum содержит экспериментальные данные. Если соединение идентифицировано как находящееся за пределами области применимости модели, предсказанное значение для этого соединения исключено (см. соединения 1 и 6):

```
Date...
5-Fold External Cross-Validation procedure.
File of Mol Structures: GdL_logK_TRAINING.SDF
```

Property Name: LogK
Mask File: GdL_logK_TRAINING.MSK
Exter.Descriptors File: -

TABLE P1. Test set: Predicted property LogK for the compounds from the GdL_logK_TRAINING.SDF file.

cmp. no.	Datum	IAB2-120
1	7.82	
6	4.07	
11	18.48	23.15
...		
2	2.39	
7	11.2	
12	2.37	2.63
...		
3	2.4	2.25
8	17.54	17.43
13	4.57	4.84
...		
4	6.68	6.05
9	4.44	4.92
14	2.73	3.52
...		
5	2.84	2.02
10	2.4	1.90
15	2.08	1.49
...		

Первый и второй графические файлы PLOT_Predict... и PLOT_Calc... отображают сравнение экспериментальных Y_{exp} и предсказанных Y_{pred} , Y_{exp} и рассчитанных Y_{calc} величин свойства в форме линейных уравнений $Y_{pred} = a + b \cdot Y_{exp}$ и $Y_{calc} = a + b \cdot Y_{exp}$ для всех пяти тестовых и обучающих наборов процедуры 5-fold CV. На графиках приведены статистические параметры: квадрат коэффициента детерминации R_{det}^2 , среднеквадратичная ошибка $RMSE$ и средняя абсолютная ошибка MAE для объединенных тестовых, обучающих наборов данных. Молекулярную структуру, соответствующую точке данных на графиках, можно визуализировать, щелкнув по выбранной точке. Структура отображается на желтом фоне; номер молекулы во входном файле SDF, значения Y_{exp} и Y_{pred} (Y_{calc}) появятся в строке состояния в нижней части окна программы.

УПРАЖНЕНИЕ 4

4. Консенсус-модель: построение и проверка

Задача:

Целью данного упражнения является построение консенсус-модели для более надежных и точных предсказаний по сравнению с индивидуальной моделью: генерация ансамбля индивидуальных моделей множественной линейной регрессии на основе различных типов СМФ дескрипторов, выбор статистически значимых моделей и

применение их к тестируемым молекулам для получения средних предсказанных величин свойства, исключая выбросы.

Программа ISIDA_QSPR строит большое число уравнений МЛР, комбинируя методы прямого ^[10] и обратного ^[15] пошагового отбора переменных, т.е. создается ансамбль предсказательных моделей, полученных сочетанием различных типов дескрипторов и различных алгоритмов их отбора. При этом используется процедура k -кратного внешнего перекрестного контроля (см. упражнение 3) в качестве стандартного протокола для оценки предсказательной эффективности моделей. Индивидуальная модель входит в состав консенсус-модели в соответствии с двумя критериями: квадрат коэффициента детерминации Q^2 скользящего контроля (LOO) должен быть больше порога Q_{lim}^2 , определенного пользователем, а остаток $(R^2 - Q^2)$ между квадратом коэффициента корреляции R^2 и Q^2 также должен быть больше заданного порога $(R^2 - Q^2)_{lim}$ (рис. 13).

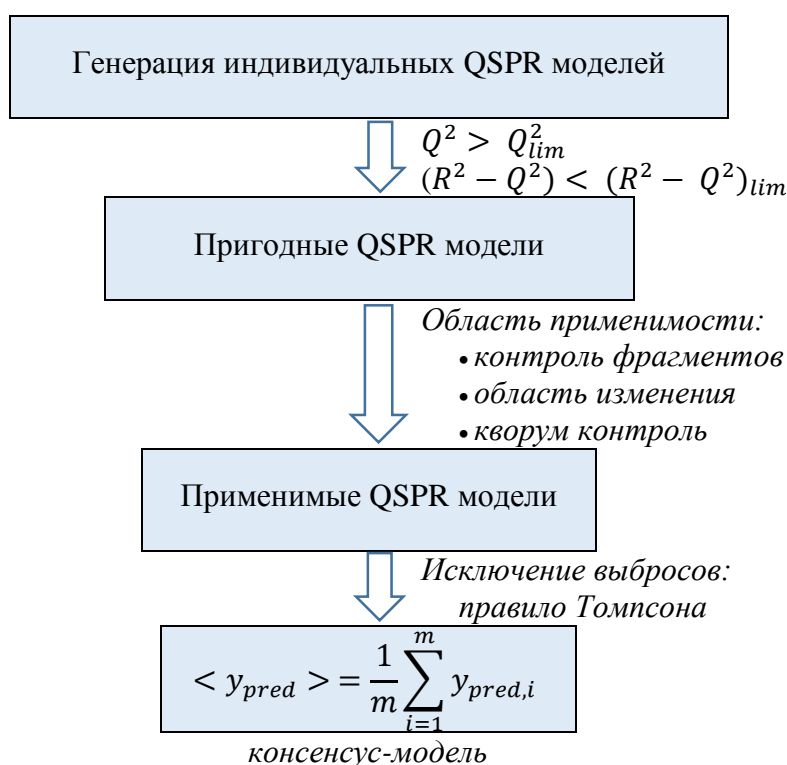


Рис. 13. Построение консенсус-модели на основе ансамбля m выбранных индивидуальных моделей.

Затем программа применяет выбранные индивидуальные модели к каждому тестируемому соединению и предсказывает целевое свойство как среднее арифметическое значений, оцененных выбранными моделями, за исключением тех моделей, которые дают выбросы

согласно правилу Томпсона^[33] или не могут быть применены к данному соединению из-за проблемы области применимости модели (AD) (рис. 13). Три критерия AD могут быть использованы одновременно или по отдельности: контроль неизвестных фрагментов, область изменения фрагментных дескрипторов модели^[20] и "кворум-контроль"^[11] (см. раздел 1.7).

Предсказанные величины целевого свойства получаются значительно более надежными и точными при включении в консенсус-модель нескольких сотен индивидуальных моделей. В этом упражнении мы используем лишь несколько типов дескрипторов и один тип уравнения МЛР для оперативной демонстрации построения ансамбля 144 индивидуальных моделей, применяемых для предсказаний консенсус-моделью.

4.1. Загрузка и проверка входного файла структура-данные

В ISIDA_QSPR откройте диалоговое окно построения одной модели кнопкой ***Single Model*** (рис. 4). Загрузите входной SDF файл GdL_logK_TRAINING.SDF и выберите свойство LogK для моделирования, как показано в разделе "1.1. Ввод данных для моделирования" упражнения 1. Щелкните на кнопке ***Create MASK*** (рис. 4), появится диалоговое окно Create MASK (рис. 5), в котором щелкните на радиокнопке (переключателе) ***Create MASK No test set***; щелкните на кнопке ***START***, чтобы сохранить/перезаписать файл-маску GdL_logK_TRAINING.MSK в главном каталоге ISIDA_QSPR с помощью диалогового окна Save mask file Dialog. Имя файла GdL_logK_TRAINING.MSK появится внизу панели ***Data*** (рис. 4). Щелкните на кнопке ***Data TEST*** для проверки входных данных (рис. 4). Если данные корректны, то в диалоговом окне ***Information*** появится сообщение: "Файлы входных данных находятся во внутреннем согласии". Закройте диалоговое окно ***Information***, затем закройте диалоговое окно Single Model Calculations, щелкнув на кнопке ***CANCEL*** (рис. 4).

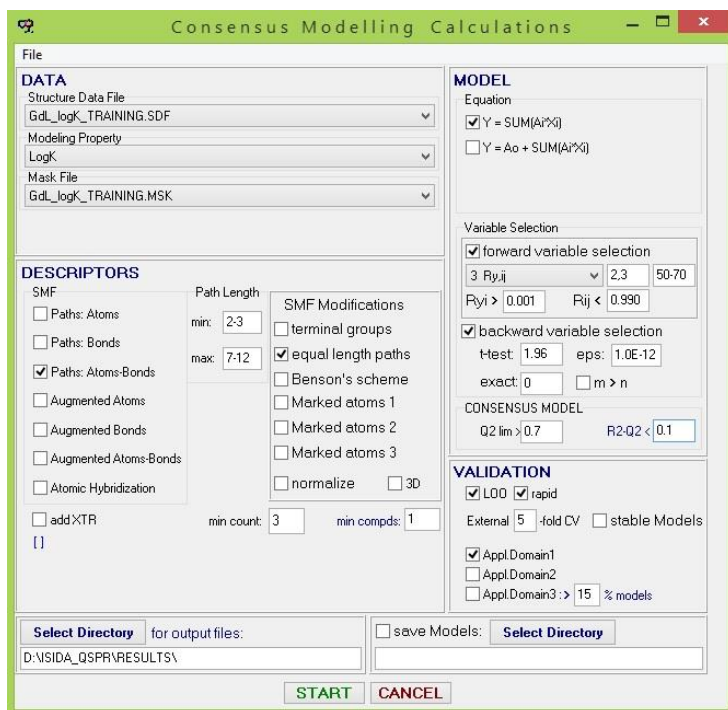


Рис. 14. Диалоговое окно расчета консенсус-модели.

4.2. Выбор дескрипторов и уравнений МЛР

Щелкните на кнопке *Consensus Model* программы ISIDA_QSPR (рис. 1), чтобы открыть диалоговое окно *Consensus Modelling Calculations* (рис. 14), используемое для ввода параметров консенсус-модели. Диалоговое окно включает панель *Data* для ввода данных, панель *Descriptors* для выбора типов СМФ дескрипторов, панель *Model* для выбора типов уравнений МЛР и методов прямого и обратного пошагового выбора переменных, а также панель *Validation* для выбора параметров внутренней и внешней проверки индивидуальных моделей (см. подробности о проверках в упражнениях 1 и 3). На панели *Descriptors* (рис. 14) установите флажок (галочку) *Paths: Atoms-Bonds*, затем установите флажок *equal length paths*, введите 2-3 в поле редактирования *min* и 7-12 в поле редактирования *max* для задания минимальной и максимальной длины пути. Убедитесь в отсутствии дополнительных флажков на панели *Descriptors*. По умолчанию используйте 3 в поле редактирования *min count* и 1 в поле редактирования *min compds* (рис. 14). Такая настройка приводит к генерации двух классов дескрипторов: а) кратчайших топологических путей с явным указанием атомов и связей и б) аналогичные кратчайшие пути, включая пути равной длины. Для каждого класса путей мы задали минимальное ($2 \leq m_{min} \leq 3$) и максимальное ($7 \leq m_{max} \leq 12$) число составляющих атомов (m). Топологические пути включают все промежуточные

кратчайшие пути с m атомами: $m_{min} \leq m \leq m_{max}$. Наш выбор дескрипторов приводит к генерации 24 типов СМФ. СМФ каждого типа являются начальной совокупностью дескрипторов для построения нескольких МЛР моделей с использованием различных методов отбора переменных.

На панели *Model* (рис. 14) установите флажок $Y = SUM(Ai*Xi)$ для выбора только одного типа уравнения линейной регрессии - без свободного коэффициента.

4.3. Методы отбора переменных

На панели *Model* (рис. 14) в её разделе Variable Selection установите флажки *forward variable selection* и *backward variable selection*. Справа от комбинированного окна (выпадающего списка) методов forward variable selection введите 2,3 в первое поле редактирования для выбора алгоритмов $R_{y,i}$ и $R_{y,ij}$ отбора переменных ^[10], во второе поле редактирования введите 50-70 для масштабируемого количества отобранных переменных, представленного в процентах от размера обучающей выборки (n). В данном случае $0.5n$, $0.6n$ и $0.7n$ переменных будут отобраны алгоритмами $R_{y,i}$ и $R_{y,ij}$ и последовательно применены к построению моделей. Введите 0.001 в поле редактирования $R_{y,i}$ и 0.99 в поле редактирования R_{ij} для пороговых значений коэффициентов корреляции (см. раздел 1.5a о корреляциях). Введите 1.96 в поле редактирования *t-test*, $1E-12$ в поле редактирования *eps* и 0 в поле редактирования *exact*. Убедитесь, что снят флажок $m > n$ (рис. 14).

4.4. Консенсус-модель

На панели *Model* в разделе Consensus Model (рис. 14) введите 0.7 в поле редактирования *Q2 lim* для минимального значения коэффициента корреляции Q^2 скользящего контроля (см. раздел 1.6). Введите 0.1 в поле редактирования $R^2 - Q^2$ для максимального значения разности между квадратом коэффициента корреляции R^2 модели и Q^2 скользящего контроля. Согласно этим критериям индивидуальные модели выбираются для консенсус-модели.

4.5. Область применимости модели

На панели **Validation** (рис. 14) пользователь может использовать одновременно или по отдельности три подхода, определяющих область применимости модели: контроль неизвестных фрагментов (AD1), область изменения фрагментных дескрипторов модели (AD2) и "кворум-контроль" (AD3) (см. раздел 1.7). Здесь установите флажок **Appl. Domain 1** и снимите флажки **Appl. Domain 2** и **Appl. Domain 3**.

4.6. Параметры внутреннего и внешнего контроля качества модели

На панели **Validation** (рис. 14) установите флажки **LOO** и **rapid** для быстрого расчета коэффициента корреляции скользящего контроля Q , введите 5 в поле редактирования **External n-fold CV** для выполнения внешнего пятикратного перекрестного контроля.

4.7. Каталог для сохранения результатов консенсус-моделирования

Программа сохраняет файлы результатов расчета в определенном пользователем каталоге: щелкните на кнопке **Select Directory** внизу в левом углу (рис. 14), появится диалоговое окно выбора каталога, войдите в каталог ...ISIDA_QSPR\RESULTS и щелкните на кнопке **Open** диалогового окна, чтобы его выбрать. Убедитесь, что флажок **Save Models** снят в правом углу диалогового окна (рис. 14).

Файлы результатов расчета можно в дальнейшем вновь открыть, выбрав в главном меню ISIDA_QSPR (рис. 1): File → Open → каталог с файлами результатов, и открыв файл *.OUT, имя которого начинается с даты и времени проведения расчетов и включает имя входного SDF файла.

4.8. Результаты моделирования и статистические параметры консенсус-модели

Для выполнения расчетов щелкните на кнопке **START** диалогового окна **Consensus Modelling Calculations** (рис. 14). Пятикратно в соответствии с пятикратным перекрестным контролем (5-fold CV) будет построено по 144 индивидуальных модели согласно выбранным параметрам моделирования: генерация 24 типов дескрипторов, два метода их предварительного отбора и три масштабируемого количества

отобранных дескрипторов. Программа создаст восемь выходных файлов: шесть текстовых файлов и два файла графического представления результатов, что описано ниже.

Файл ***.ТОМ** содержит статистические параметры индивидуальных моделей МЛР для каждого блока пятикратного перекрестного контроля (5-fold CV):

5-Fold External Cross-Validation Procedure. Table of models: stat. parameters.

=> **Subset 1/5**

File of Mol Structures: GdL_logK_TRAINING.SDF; 88 compounds in training set.
Modeling Property Name: LogK
Mask File: GdL_logK_TRAINING.MSK
Exter.Descriptors File: -

no	fragment	fitting	n	k	R2	F	testR2det	testRMSE	HRF	Q2
	type	equation								
1	IAB3-937	0	88	34	0.994027	272.30	0.940	1.54E+00	4.457	0.983188
2	IAB2-936	0	88	26	0.991633	293.92	0.958	1.29E+00	5.275	0.981164
3	IAB2-11a37	0	88	32	0.993365	270.46	0.907	1.88E+00	4.697	0.979663
4	IAB2-11a35	0	88	25	0.989293	242.53	0.933	1.60E+00	5.967	0.974314
5	IAB2-8a36	0	88	29	0.990500	219.69	0.840	2.33E+00	5.621	0.973377

...

=> **Subset 2/5**

File of Mol Structures: GdL_logK_TRAINING.SDF; 89 compounds in training set.
Modeling Property Name: LogK
Mask File: GdL_logK_TRAINING.MSK
Exter.Descriptors File: -

no	fragment	fitting	n	k	R2	F	testR2det	testRMSE	HRF	Q2
	type	equation								
1	IAB2-836	0	89	31	0.992046	241.14	0.962	1.48E+00	5.033	0.968968
2	IAB2-837	0	89	32	0.992992	260.52	0.963	1.46E+00	4.725	0.967241
3	IAB2-835	0	89	27	0.989108	216.54	0.961	1.51E+00	5.890	0.966892
4	IAB2-12a35	0	89	27	0.982853	136.69	0.952	1.78E+00	7.390	0.961315
5	IAB2-1136	0	89	33	0.989789	169.63	0.896	2.56E+00	5.703	0.960585

...

Здесь для очередного блока 5-fold CV для каждой индивидуальной модели представлены следующие параметры: число точек n в обучающей выборке, число коэффициентов (фрагментов) k уравнения МЛР, квадрат коэффициента корреляции Пирсона R^2 , критерий Фишера F , квадрат коэффициента детерминации $testR_{det}^2$ для тестовой выборки, среднеквадратичная ошибка $testRMSE$ для тестовой выборки, R -фактора Гамильтона ^[32] в процентах HRF , квадрат коэффициента детерминации скользящего контроля Q^2 для обучающей выборки. Модели отсортированы в порядке убывания Q^2 . Индивидуальные модели, не вошедшие в консенсус-модель, помечены величинами $Q^2 \leq 0.499999$.

Файл ***_TST_5fCV.AVE** для всех пяти внешних тестовых наборов процедуры 5-fold CV включает средние предсказанные значения свойства (колонка *Average*) и их стандартные отклонения (*STDEV*), оцененные консенсус-моделями, и количество индивидуальных

моделей (Nm), использованных для расчета среднего значения. Если соединение определено как находящееся за пределами области применимости всех индивидуальных моделей, входящих в состав консенсус-модели, предсказанное значение для этого соединения исключается (например, см. соединение 1; область применимости была задана как AD1):

TABLE PA. Test set: Average predicted property LogK for the compounds from the GdL_logK_TRAINING.SDF file using models selected by ISIDA_QSPR.

cmp. no.	Datum	Average	STDEV	Nm	Dat.- Ave.
1	7.82			0	
6	4.07			0	
11	18.48	1.98739E+001	5.963E-001	122	-1.394E+000
16	3.18	3.54161E+000	1.081E+000	124	-3.616E-001
21	16.48			0	
26	12.54	1.25386E+001	7.704E-001	66	1.364E-003
31	2.5	2.55175E+000	3.486E-001	80	-5.175E-002
36	3.09	1.71098E+000	1.247E-001	82	1.379E+000
...					

Файл ***_TST_5fCV.TSP** содержит значения свойства, предсказанные каждой индивидуальной МЛР моделью для каждого блока пятикратного перекрестного контроля:

5-Fold External Cross-Validation Procedure; => **Subset 1/5**
 File of Mol Structures: GdL_logK_TRAINING.SDF
 Property Name: LogK
 Mask File: GdL_logK_TRAINING.MSK
 Exter.Descriptors File: -

TABLE P1. Test set: Predicted property LogK for the compounds from the GdL_logK_TRAINING.SDF file
 142 Selected MODELS, Q2 >= 0.7

cmp. no.	Datum	IAB3-9370	IAB2-9360	IAB2-11a370	IAB2-11a350	IAB2-8a360	IAB2-8a370...
1	7.82						...
6	4.07						...
11	18.48	19.09	19.69	19.04	19.93	19.15	19.02...
16	3.18	3.87	1.87	3.30	1.98	3.77	2.41...
...							

5-Fold External Cross-Validation Procedure; => **Subset 2/5**
 File of Mol Structures: GdL_logK_TRAINING.SDF
 Property Name: LogK
 Mask File: GdL_logK_TRAINING.MSK
 Exter.Descriptors File: -

TABLE P1. Test set: Predicted property LogK for the compounds from the GdL_logK_TRAINING.SDF file
 144 Selected MODELS, Q2 >= 0.7

cmp. no.	Datum	IAB2-8360	IAB2-8370	IAB2-8350	IAB2-12a350	IAB2-11360	IAB3-11360...
2	2.39						...
7	11.2	9.98	9.88	10.01			...
12	2.37	2.55	2.57	2.58	1.96	2.71	2.71...
...							

Оставшиеся три текстовых файла ***_5fCV_AVE.AVE**, ***_5fCV.AVE** и ***_5fCV.TSC** аналогично описанным выше файлам для тестовых наборов ***_TST_5fCV.AVE** и ***_TST_5fCV.TSP** содержат информацию о средних рассчитанных значениях свойства и рассчитанных каждой индивидуальной моделью для соединений обучающих наборов каждого блока пятикратного перекрестного контроля.

Первый и второй графические файлы PLOT_Calc... и PLOT_Predict... (рис. 15) отображают сравнение экспериментальных Y_{exp} и рассчитанных Y_{calc} , Y_{exp} и предсказанных Y_{pred} средних величин свойства согласно консенсус-модели в форме линейных уравнений $Y_{calc} = a + b \cdot Y_{exp}$ и $Y_{pred} = a + b \cdot Y_{exp}$ для всех пяти тестовых и обучающих наборов процедуры 5-fold CV. На графиках приведены статистические параметры: квадрат коэффициента детерминации R_{det}^2 , среднеквадратичная ошибка $RMSE$ и средняя абсолютная ошибка MAE для объединенных тестовых и обучающих наборов данных. Молекулярную структуру, соответствующую точке данных на графиках, можно визуализировать, щелкнув по выбранной точке. Структура отображается на желтом фоне. Номер структуры во входном файле, значения Y_{exp} , Y_{pred} и Y_{calc} появятся в строке состояния в нижней части окна программы.

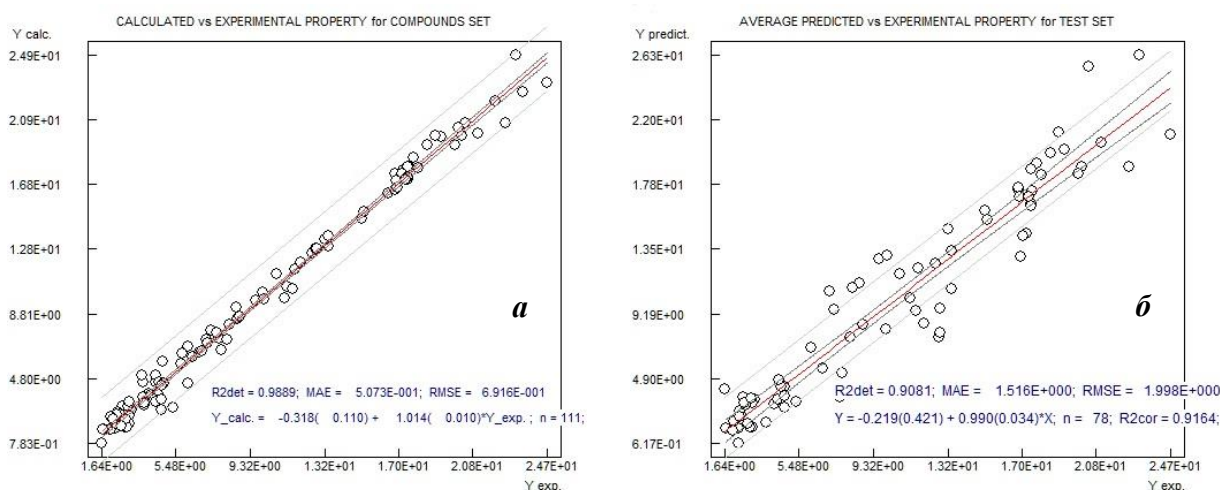


Рис. 15. Сравнение экспериментальных Y_{exp} , рассчитанных Y_{calc} и предсказанных Y_{pred} средних величин моделируемого свойства $Y = \log K$ согласно консенсус-модели для объединенных обучающих (а) и тестовых (б) наборов пятикратного перекрестного контроля: сопоставление экспериментальных и рассчитанных (fitted) (а), экспериментальных и предсказанных (б) величин в форме линейной взаимосвязей $Y_{calc} = a + b \cdot Y_{exp}$ и $Y_{pred} = a + b \cdot Y_{exp}$.

4.9. Эффективность консенсус-модели как функция порога принятия индивидуальных моделей

Предсказательная эффективность консенсус-модели зависит от числа и качества индивидуальных моделей, выбор которых, в свою очередь, зависит от определяемого пользователем порогового значения коэффициента детерминации LOO Q_{lim}^2 (см. введение

к упражнению 4). Анализируя результаты консенсус-моделирования, пользователь может построить график зависимости коэффициента детерминации R_{det}^2 предсказаний процедуры k -fold CV как функцию Q_{lim}^2 , что может помочь определить оптимальное значение Q_{lim}^2 , обеспечивающее приемлемый R_{det}^2 .

Щелкните **Tools** → **R2det vs Q2** в главном меню ISIDA_QSPR (рис. 1), чтобы открыть окно инструмента Averaging (рис. 16), который позволяет проанализировать коэффициент детерминации R_{det}^2 k -кратного внешнего перекрестного контроля как функцию Q_{lim}^2 . В этом окне установите флажки (галочки) **use Q2 threshold** и **do R2det vs Q2**, в соответствии с Q_{lim}^2 , как было для консенсус-модели, введите 0.7 в поле редактирования **use Q2 threshold** и введите значение инкремента 0.05 в поле редактирования **Step**. Щелкните **File** → **Open** в главном меню инструмента Averaging (рис. 16), в открывшемся диалоговом окне выберите файл анализируемой консенсус-модели <дата_время>**GdL_logK_TRAINING_TST_5fCV.TSP** из списка файлов *.TSP в каталоге, где были сохранены результаты консенсус-моделирования (например, в каталоге C:\ISIDA_QSPR\RESULTS), а затем нажмите кнопку **Open**. Будет предложено открыть еще один файл с той же датой и временем <дата_время>**GdL_logK_TRAINING_5fCV.TOM** из списка .TOM файлов в той же директории: выберите и нажмите кнопку **Open**. После того как инструмент Averaging выполнит расчеты, перейдите на его вкладку **Graph** (рис. 16), где на графике отображается зависимость между R_{det}^2 и Q_{lim}^2 . Дополнительная информация, связанная с этим графиком, доступна на вкладке **Table** (рис. 15), где видно, что R_{det}^2 практически не меняется с ростом Q_{lim}^2 от 0.7 до 0.9: $R_{det}^2 \approx 0.908$, затем R_{det}^2 незначительно возрастает до 0.913 при $Q_{lim}^2 = 0.95$. Однако в последнем случае резко уменьшается число индивидуальных моделей, входящих в консенсус-модель, с 135 до 34 (см. вкладку **Table**). Таким образом, примененное значение $Q_{lim}^2 = 0.7$ достаточно оптимально для изучаемой выборки.

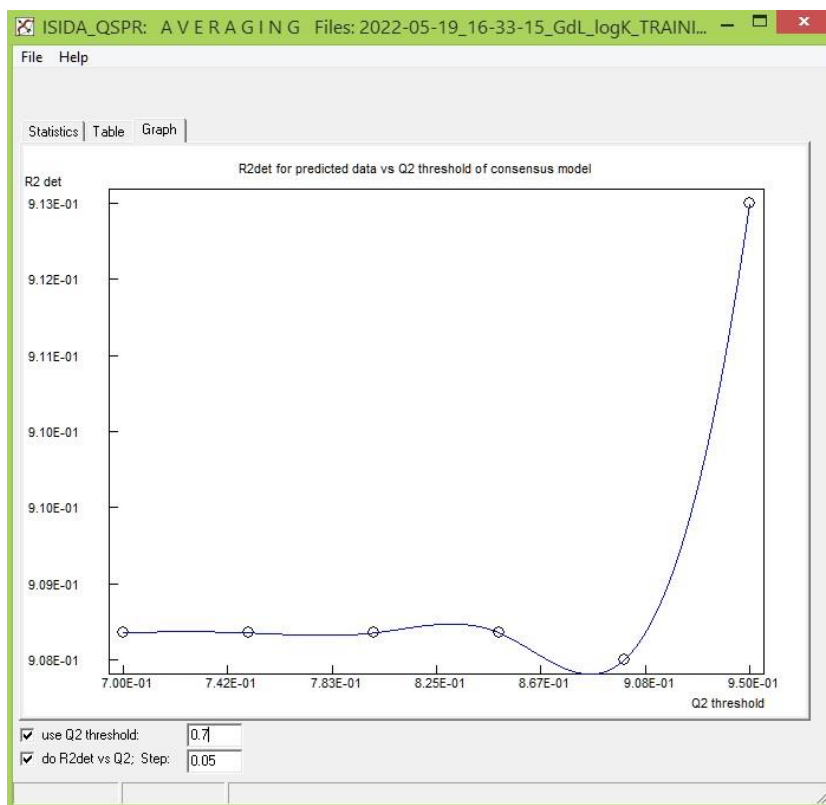


Figure 16. Графический интерфейс программного инструмента Averaging.

4.10. Построение консенсус-модели на всем наборе данных и ее сохранение

В этом разделе объясняется, как построить и сохранить консенсус-модель на всем наборе данных. Щелкните на кнопке **Consensus Model** программы ISIDA_QSPR (рис. 1), чтобы открыть диалоговое окно **Consensus Modelling Calculations** (рис. 14); используйте все ранее примененные настройки (рис. 14) этого окна также и в этом упражнении 4, кроме числа блоков *k*-fold CV и имени каталога для выходных файлов. Введите 1 в поле редактирования **External n-fold CV**, чтобы деактивировать *k*-fold CV. Внизу справа (рис. 14) установите флажок **save Models** и щелкните на кнопке **Select Directory**. В появившемся диалоговом окне щелчком по кнопке **Open** выберите открытый каталог, например, C:\ISIDA_QSPR\GdL_logK_MODELS. Щелкните на кнопке **START** (рис. 14), чтобы построить и сохранить набор индивидуальных МЛР моделей, входящих в состав консенсус-модели.

Программа создаст и откроет пять выходных файлов: четыре текстовых файла и один файл графического представления результатов, которые аналогичны файлам для обучающих наборов данных в *k*-fold CV (см. раздел "4.8. Результаты моделирования и статистические параметры консенсус-модели"). В выбранном каталоге, как указано выше

C:\ISIDA_QSPR\GdL_logK_MODELS, будут сохранены файлы *.SPE индивидуальных МЛП моделей, входящих в состав консенсус-модели, и файл <дата_время>GdL_logK_TRAINING_...TOM со статистическими параметрами этих моделей.

УПРАЖНЕНИЕ 5

5. Предсказание свойств и виртуальный скрининг с использованием консенсус-модели

Задача:

Это упражнение демонстрирует программный инструмент Consensus Predictor ^[34], который применяет ранее полученные консенсус-модели к набору реальных или виртуальных молекул для предсказания их свойств.

В качестве входных данных программный инструмент Consensus Predictor, как и программа ISIDA_QSPR (см. раздел “Данные для моделирования”), использует файл в формате SDF ^[16], содержащий структуры молекул и данные. Этот файл может также включать экспериментальные или расчетные значения свойства в поле данных, названном так же, как моделируемое свойство, для которого и была построена QSPR консенсус-модель. В этом случае входное значение свойства сравнивается с предсказываемым, после чего оценивается эффективность прогнозирования. В этом упражнении мы используем следующий подготовленный файл структура-данные: **GdL_logK_TEST.SDF**.

5.1. Загрузка входных данных

Щелкните в главном меню ISIDA_QSPR (рис. 1): **Tools** → **Property Prediction**, чтобы открыть графический интерфейс предиктора Consensus Predictor (рис. 17). Щелкните на его верхней кнопке **LOAD**, затем выберите в открывшемся диалоговом окне кнопкой **Open** файл **GdL_logK_TEST.SDF** в главном каталоге программы ISIDA_QSPR. Под верхней кнопкой **LOAD** появится строка <...>\ISIDA_QSPR\GdL_logK_TEST.SDF, указывающая на то, что выбранный файл загружен. В то же время под кнопкой **SAVE** появится по умолчанию имя выходного файла <...>\ISIDA_QSPR\GdL_logK_TEST.TSP.

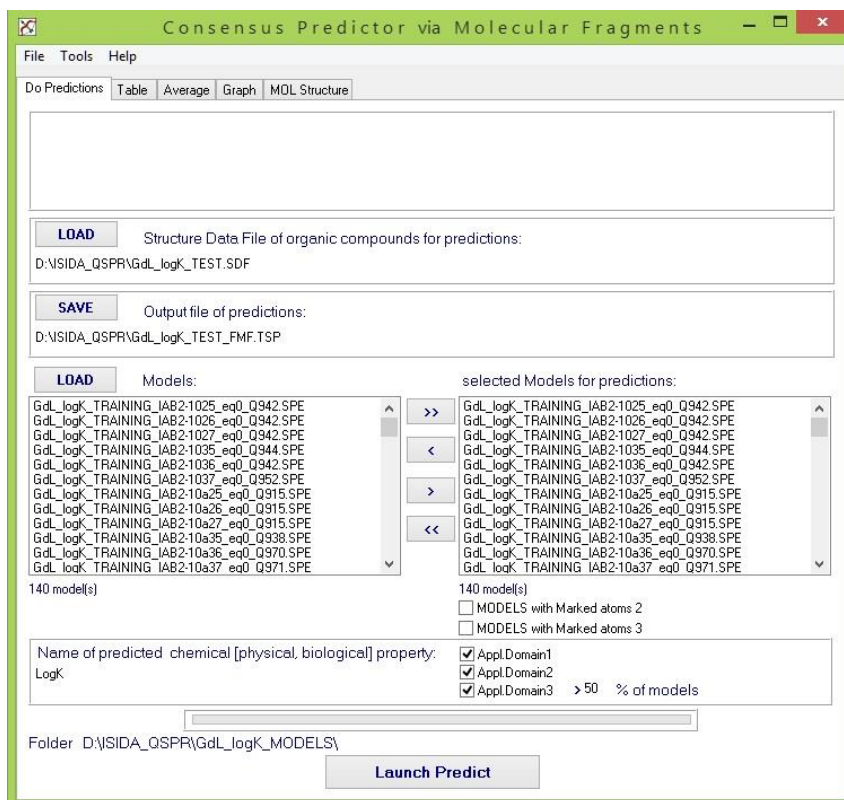


Рис. 17. Графический интерфейс программного инструмента Consensus Predictor.

5.2. Загрузка выбранных моделей и выбор области их применимости

Щелкните на нижней кнопке **LOAD** предиктора (рис. 17). В открывшемся диалоговом окне откройте каталог C:\SIDA_QSPR\GdL_logK_MODELS, содержащий сохраненные QSPR модели (см. раздел 4.10 "Построение консенсус-модели на всем наборе данных"), выберите и откройте любой файл *.SPE из списка. Имена всех файлов *.SPE консенсус-модели появятся в левом окне списка моделей Model(s). Щелкните на кнопке >>, чтобы выбрать все файлы *.SPE. Список из 140 моделей появится в правом окне Model(s). Убедитесь, что флажки (галочки) **MODELS with Marked atoms 2** и **MODELS with Marked atoms 3** сняты.

Для обеспечения надежного прогноза используйте одновременно три подхода, определяющих область применимости модели: контроль неизвестных фрагментов (AD1), область изменения фрагментных дескрипторов модели (AD2) и "кворум-контроль" (AD3) (см. раздел 1.7). Установите флажки **Appl. Domain 1**, **Appl. Domain 2** и **Appl. Domain 3**. Введите 50 в поле редактирования для минимального числа в

процентах применимых индивидуальных моделей, прошедших контроль AD1 и AD2. Если фактическое число меньше заданного значения 50%, предсказание консенсус-моделью игнорируется согласно AD3.

5.4. Результаты предсказаний консенсус-моделью

Щелкните на кнопке **Launch Predict** окна Consensus Predictor (рис. 16) и дайте согласие перезаписать выходной файл *.TSP, если он существует. Результаты расчетов будут представлены на четырех вкладках окна Consensus Predictor. На вкладке **Average** для каждой молекулы представлено среднее предсказанное значение свойства (*Average*), его стандартное отклонение (*STDEV*), оцененное консенсус-моделью, и количество индивидуальных моделей (*Nm*), использованных для расчета среднего значения. Если для соединений известны экспериментальные или каким-то образом оцененные значения свойства, они отображаются в колонке *Datum*. Если соединение определено как находящееся за пределами AD, прогнозируемое значение для этого соединения не приводится (например, для соединений 6 и 8):

TABLE PA. Average predicted property LogK for the compounds from the file D:\ISIDA_QSPR\GdL_logK_TEST.SDF

cmp. no.	Datum	Average	STDEV	Nm	Dat.- Ave.
1	22.3	2.17525E+001	7.306E-001	135	5.475E-001
2	13.06	1.40039E+001	9.597E-001	111	-9.439E-001
3	24.5	2.18384E+001	8.147E-001	137	2.662E+000
4	13.12	1.46170E+001	4.169E-001	83	-1.497E+000
5	19.74	1.96360E+001	8.111E-001	94	1.040E-001
6	17.15			0	
7	17.69	1.79361E+001	9.343E-001	106	-2.461E-001
8	22.58			0	
9	19.4	1.87939E+001	2.272E-001	94	6.061E-001
...					

Перейдите на вкладку **Graph**. Там для 39 соединений на графике отображается линейная корреляция между экспериментальным Y_{exp} и предсказанным Y_{pred} свойством, а также соответствующие статистические параметры. Щелкните на выбранной точке данных, затем перейдите на вкладку **MOL Structure**, чтобы посмотреть соответствующую молекулярную структуру лиганда. Для просмотра всех молекулярных структур на этой вкладке используются кнопки навигации.

5.5. Анализ вкладов фрагментов в предсказанное свойство

Щелкните **Tools** → **Fragment Contributions** в главном меню Consensus Predictor (рис. 16), чтобы открыть окно Вклад фрагментов для молекулы (Fragment Contributions for Molecule, рис. 18). Это окно включает выбранную молекулярную структуру, составляющие ее фрагменты, их число в молекуле и вклады в соответствии с выбранной индивидуальной QSPR моделью. Введите 5 в поле редактирования **Mol Number**, чтобы проанализировать этот лиганд, предсказанный как активный (высокая константа устойчивости комплекса GdL), и сопоставьте высокие вклады фрагментов в свойство для различных индивидуальных моделей с помощью выпадающего списка **Selected Model**.

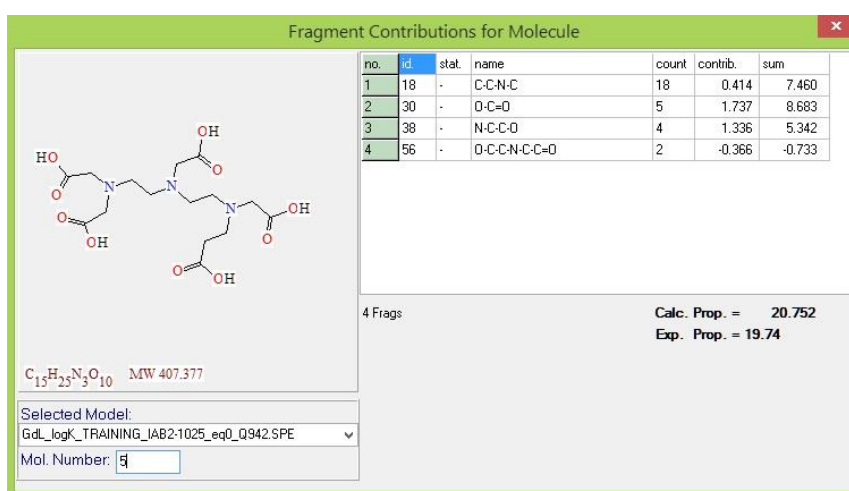


Figure 18. Графический интерфейс программного инструмента Fragment Contributions for Molecule.

Заключение

QSPR модели для предсказания физических и химических свойств, биологической активности строятся программой ISIDA_QSPR с помощью множественной линейной регрессии, применяемой к набору данных о веществах или молекулах с известным изучаемым свойством и структурами молекул, описанными субструктурными молекулярными фрагментами. Построенные модели используются для предсказания свойств новых веществ.

Программа ISIDA_QSPR, использующая в качестве метода машинного обучения множественную линейную регрессию и генерирующая субструктурные молекулярные фрагменты в качестве дескрипторов (независимых переменных), выполняет ансамблевое QSPR моделирование, повышающее надежность предсказаний. Оно включает генерацию

большого числа индивидуальных QSPR моделей, выбор из них статистически значимых и применение выбранных моделей к тестируемым/новым данным для получения средних предсказанных значений, исключая выбросы (*консенсус-модель*). Каждая индивидуальная модель соответствует определенному типу дескрипторов и определенным параметрам метода машинного обучения. Программа строит модели, сочетая методы прямого и обратного пошагового отбора переменных. В процессе моделирования применяется внутренний и внешний перекрестный контроль для выбора наиболее надежных предсказательных моделей.

Программа ISIDA_QSPR представляет собой графический интерфейс, позволяющий достаточно легко выполнять задачи QSPR моделирования и поддерживающий графический анализ результатов, связанных с графическим представлением структур химических соединений. Она работает под операционной системой Windows различных версий.

Сокращения

AD	Область применимости модели (Applicability Domain)
BVS	Пошаговое исключение переменных (Backward Variable Selection)
EdChemS	Редактор *.MOL файлов в составе программы ISIDA_QSPR
EdiSDF	Менеджер файлов *.SDF файлов в составе программы ISIDA_QSPR
<i>F</i>	Критерий Фишера
<i>FIT</i>	Критерий Кубиньи
FVS	Пошаговый накопительный отбор переменных (Forward Variable Selection)
<i>HRF</i>	R-фактора Гамильтона
ISIDA	Вычислительный дизайн и анализ данных (In Silico design and Data Analysis)
LMO	Метод отбрасывания по нескольку объектов (Leave Many Out)
LOO	Метод отбрасывания по одному – скользящий контроль (Leave One Out)
<i>MAE</i>	Средняя абсолютная ошибка
MLR	Множественная линейная регрессия
MJIP	Множественная линейная регрессия
<i>n</i>	Количество точек данных
<i>k</i> -fold CV	Внешний <i>k</i> -кратный перекрестный контроль (<i>k</i> -fold external cross-validation)
<i>Q</i>	Коэффициент детерминации скользящего контроля LOO
QSPR	Количественная взаимосвязь структура-свойство
<i>R</i>	Коэффициент корреляции Пирсона
R_{det}^2	Квадрат коэффициента детерминации

<i>RMSE</i>	Среднеквадратичная ошибка
<i>s</i>	Стандартное отклонение
<i>SDF</i>	Файл формата структура-данные
<i>SMF</i>	Субструктурные молекулярные фрагменты (Substructural Molecular Fragments)
СМФ	Субструктурные молекулярные фрагменты
<i>SVD</i>	Сингулярное разложение (Singular Value Decomposition)
<i>TC</i>	Коэффициенты сходства Танимото (The Tanimoto Similarity Coefficient)
<i>Y_{calc}</i>	Расчетное (fitted - подогнанное под МЛР модель) свойство
<i>Y_{exp}</i>	Экспериментальное (измеренное) свойство
<i>Y_{pred}</i>	Предсказанное свойство (тестируемого, нового соединения)

Список литературы

- [1] Solov'ev, V.; Varnek, A., QSPR Models on Fragment Descriptors. In *Tutorials in Chemoinformatics*, Varnek, A., Ed. John Wiley & Sons Ltd: Strasbourg, 2017; pp 135 - 162.
- [2] Solov'ev, V. P.; Varnek, A. A. ISIDA QSPR (In Silico Design and Data Analysis for Quantitative Structure-Property Relationships). <http://vpsolovev.ru/programs/> (accessed 20 May 2022).
- [3] Varnek, A.; Solov'ev, V. P. "In Silico" Design of Potential Anti-HIV Actives Using Fragment Descriptors. *Combinat. Chem. High Throughput Screening* 2005, 8 (5), 403-416.
- [4] Katritzky, A. R.; Kuanar, M.; Fara, D. C.; Karelson, M.; Acree, W. E. J.; Solov'ev, V. P.; Varnek, A. QSAR Modeling of Blood:Air and Tissue:Air Partition Coefficients Using Theoretical Descriptors. *Bioorganic and Medicinal Chemistry* 2005, 13 (23), 6450-6463.
- [5] Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural Fragments: An Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *Journal of Computer-Aided Molecular Design* 2005, 19 (9-10), 693-703.
- [6] Solov'ev, V. P.; Varnek, A. A.; Wipff, G. Modeling of Ion Complexation and Extraction Using Substructural Molecular Fragments. *Journal of Chemical Information and Computer Sciences* 2000, 40 (3), 847-858.
- [7] Solov'ev, V.; Baulin, D.; Tsivadze, A. Design of phosphoryl containing podands with Li⁺/Na⁺ selectivity using machine learning. *SAR and QSAR in Environmental Research* 2021, 32 (7), 521-539. DOI: 10.1080/1062936X.2021.1929462.
- [8] Solov'ev, V.; Kireeva, N.; Ovchinnikova, S.; Tsivadze, A. The complexation of metal ions with various organic ligands in water: prediction of stability constants by QSPR ensemble modelling. *J. Incl. Phenom. Macrocycl. Chem.* 2015, 83, 89-101. DOI: 10.1007/s10847-015-0543-6.
- [9] Solov'ev, V.; Varnek, A.; Tsivadze, A. QSPR ensemble modelling of the 1:1 and 1:2 complexation of Co²⁺, Ni²⁺, and Cu²⁺ with organic ligands. Relationships between stability constants. *J. Comput. Aided Mol. Des.* 2014, 28 (5), 549-564.
- [10] Solov'ev, V. P.; Kireeva, N.; Tsivadze, A. Y.; Varnek, A. QSPR ensemble modelling of alkaline-earth metal complexation. *J. Incl. Phenom. Macrocycl. Chem.* 2013, 76 (1-2), 159-171.
- [11] Solov'ev, V. P.; Tsivadze, A. Y.; Varnek, A. A. New approach for accurate QSPR modeling of metal complexation: Application to stability constants of complexes of lanthanide Ions Ln³⁺, Ag⁺, Zn²⁺, Cd²⁺ and Hg²⁺ with organic ligands in water. *Macroheterocycles* 2012, 5 (4-5), 404-410.
- [12] Solov'ev, V.; Marcou, G.; Tsivadze, A.; Varnek, A. Complexation of Mn²⁺, Fe²⁺, Y³⁺, La³⁺, Pb²⁺, and UO₂²⁺ with organic ligands: QSPR ensemble modeling of stability constants. *Ind. Eng. Chem. Res.* 2012, 51 (41), 13482-13489.
- [13] Solov'ev, V.; Sukhno, I.; Buzko, V.; Polushin, A.; Marcou, G.; Tsivadze, A.; Varnek, A. Stability Constants of Complexes of Zn²⁺, Cd²⁺, and Hg²⁺ with Organic Ligands: QSPR Consensus Modeling and Design of New Metal Binders. *J. Incl. Phenom. Macrocycl. Chem.* 2012, 72 (3-4), 309-321.
- [14] Tetko, I. V.; Solov'ev, V. P.; Antonov, A. V.; Yao, X.; Doucet, J. P.; Fan, B.; Hoonakker, F.; Fourches, D.; Jost, P.; Lachiche, N.; Varnek, A. Benchmarking of linear and nonlinear approaches for quantitative structure-property relationship studies of metal complexation with ionophores. *Journal of Chemical Information and Modeling* 2006, 46 (2), 808-819.
- [15] Solov'ev, V. P.; Varnek, A. A. Structure-property modeling of metal binders using molecular fragments. *Rus. Chem. Bull.* 2004, 53 (7), 1434-1445.

- [16] Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* 1992, 32 (3), 244-255.
- [17] Ronconi, L.; Sadler, P. J. Using coordination chemistry to design new medicines. *Coordination Chemistry Reviews* 2007, 251, 1633-1648.
- [18] Solov'ev, V. P.; Varnek, A. A. EdiSDF (Editor of Structure - Data Files). <http://vpsolovev.ru/programs/> (accessed 01 March 2022).
- [19] Solov'ev, V. P.; Varnek, A. A. EdChemS (Editor of Chemical Structures). <http://vpsolovev.ru/programs/> (accessed 01 March 2022).
- [20] Solov'ev, V.; Oprisiu, I.; Marcou, G.; Varnek, A. Quantitative Structure_Property Relationship (QSPR) Modeling of Normal Boiling Point Temperature and Composition of Binary Azeotropes. *Ind. Eng. Chem. Res.* 2011, 50 (24), 14162-14167. DOI: 10.1021/ie2018614.
- [21] Reid, R. C.; Prausnitz, J. M.; Sherwood, T. K. *The Properties of Gases and Liquids*; McGraw-Hill Book Co: New York, 1977.
- [22] Ruggiu, F.; Solov'ev, V.; Marcou, G.; Horvath, D.; Graton, J.; Le Questel, J.-Y.; Varnek, A. Individual Hydrogen-Bond Strength QSPR Modelling with ISIDA Local Descriptors: a Step Towards Polyfunctional Molecules. *Molecular Informatics* 2014, 33 (6-7), 477-487.
- [23] Себер, Д. *Линейный регрессионный анализ*; Мир: М., 1980.
- [24] Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C++*. *The Art of Scientific Computing*; 2 ed ed.; Cambridge University Press: New York, 2002.
- [25] Лоусон, Ч.; Хенсон, Р. *Численное решение задач метода наименьших квадратов*; Наука: М., 1986.
- [26] Форсайт, Д.; Малькольм, М.; Моулдер, К. *Машинные методы математических вычислений*; Мир: М., 1980.
- [27] Sachs, L. *Applied Statistics. A Handbook of Techniques. Second Edition*; Springer: Berlin, 1984.
- [28] Kubinyi, H. Variable Selection in QSAR Studies. II. A Highly Efficient Combination of Systematic Search and Evolution. *Quant. Struct.-Act. Relat.* 1994, 13 (4), 393-401.
- [29] Жохова, Н. И.; Баскин, И. И.; Палюлин, В. А.; Зефиоров, А. Н.; Зефиоров, Н. С. Фрагментные дескрипторы с "выделенными" атомами и их применение в исследованиях количественных соотношений "структура-активность"/"структура-свойство". *Доклады Академии наук* 2007, 417 (5), 639-641.
- [30] Golbraikh, A.; Tropsha, A. Beware of q²! *J. Mol. Graphics and Modell.* 2002, 20, 269-276.
- [31] Varnek, A.; Kireeva, N.; Tetko, I. V.; Baskin, I. I.; Solov'ev, V. P. Exhaustive QSPR Studies of a Large Diverse Set of Ionic Liquids: How Accurately Can We Predict Melting Points? *Journal of Chemical Information and Modeling* 2007, 47 (3), 1111-1122.
- [32] Hartley, F. R.; Burgess, C.; Alcock, R. M. *Solution Equilibria*; John Wiley: Chichester, 1980.
- [33] Muller, P. H.; Neumann, P.; Storm, R. *Tafeln der mathematischen Statistik*; VEB Fachbuchverlag: Leipzig, 1979.
- [34] Solov'ev, V. P. FMF (Forecast by Molecular Fragments). Predictions of physical and chemical properties and biological activity using QSPR models of the ISIDA_QSPR program. <http://vpsolovev.ru/programs/> (accessed 19 May 2022).