

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего профессионального образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Кафедра теоретической и прикладной химии

Григорьева Е.А.

ХЕМОИНФОРМАТИКА

Лабораторный практикум

Информационная система ISIDA/QSPR

Челябинск

Издательский центр ЮУрГУ

2022

Информационная система ISIDA/QSPR.

Цель работы: познакомиться с возможностями модулей информационной системы ISIDA/QSPR. Построить регрессионные модели на основе фрагментных дескрипторов. Оформить отчет в файле Microsoft Word.

Программа ISIDA/QSPR включает:

- 1) Редактор EdiSDF, предназначенный для визуализации и редактирования файлов, содержащих химические 2D и 3D формулы в формате Structure-Data Files (SDF). Входные данные для программы ISIDA/QSPR представлены файлами SDF.
- 2) Программу FMF для предсказания физических, химических и биологических свойств с использованием готовых моделей ISIDA/QSPR.
- 3) Модуль MolFrag для анализа субструктурных молекулярных фрагментов (SMF) и их вкладов.

В программе ISIDA/QSPR используются субструктурные молекулярные фрагменты (SMF) и многомерный линейный регрессионный анализ (MLR) для QSPR и QSAR моделирования и предсказания физических, химических и биологических свойств.

В качестве исходных данных в ISIDA/QSPR служат известные экспериментальные величины моделируемого свойства для ряда соединений обучающего набора данных. Субструктурные молекулярные фрагменты как подграфы молекулярных графов соединений применяются в качестве дескрипторов (независимых переменных) в QSPR моделях. Как правило, применяются кратчайшие топологические пути. Кратность фрагмента в соединении – величина дескриптора. Дескрипторы вычисляются исключительно из данных о структурной (2D) формуле соединения.

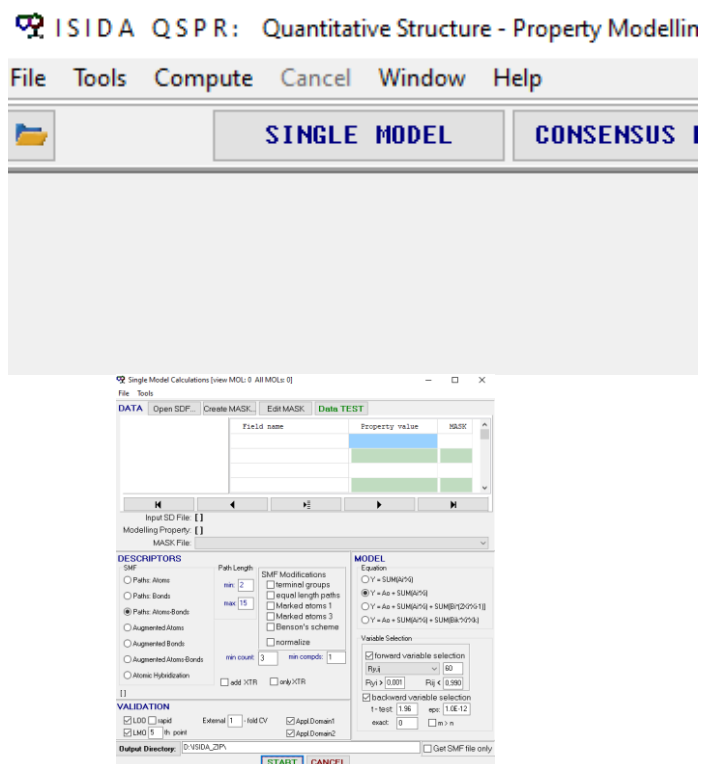
Для построения модели привлекается оригинальная комбинированная пошаговая техника увеличения и уменьшения числа SMF дескрипторов в модели для выбора переменных из их исходного множества, обеспечивающих наилучшее предсказание свойства.

ISIDA/QSPR генерирует множество устойчивых MLR моделей, каждая из которых соответствует применяемому типу SMF дескрипторов и методам ступенчатого отбора переменных. Для надежного предсказания свойств используется консенсус модель. Консенсус модель объединяет предсказания множества индивидуальных моделей. Программа рассчитывает свойство как арифметическое среднее величин, вычисленных с помощью наиболее устойчивых индивидуальных моделей, исключая выпадающие величины и применяя методы по оценке области применимости каждой индивидуальной модели.

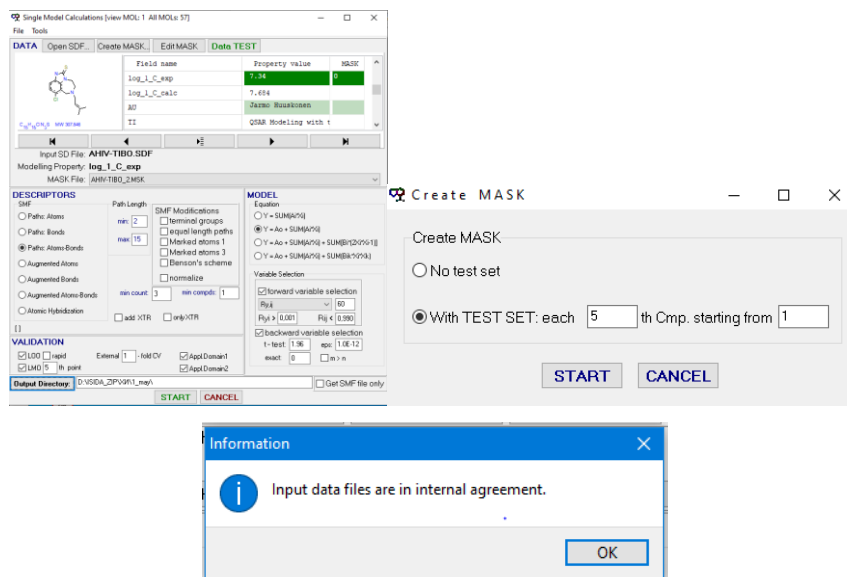
1.1 Построение индивидуальной модели (MLR) на одном наборе дескрипторов SMF, предсказание свойств на тестовом наборе

Переименовать файл ANIV-TIBO.SDF в папке ISIDA_QSPR на 20_BD_Familia.*, где Familia это ваша фамилия, латиницей. Входной файл должен находиться в программном каталоге ISIDA_QSPR.

Запустить ПП ISIDA/QSPR – ISIDA_QSPR.exe, открыть модуль «SINGLE MODEL».



Пройти во вкладку «Open SDF». Загрузить файл с данными 20_BD_Familia.* (бывший ANIV-TIBO.SDF), содержащий структуры 57 производных тетрагидроимидазобензодиазепинона (TIBO), экспериментальные значения анти-ВИЧ активности $\log(1/IC_{50})$ и другие данные. Обозначить моделируемое свойство, нажав на поле с моделируемым свойством, проверить название «log_1_C_exp» в строке «Modeling Property». Запустить формирование обучающей выборки «Create MASK».

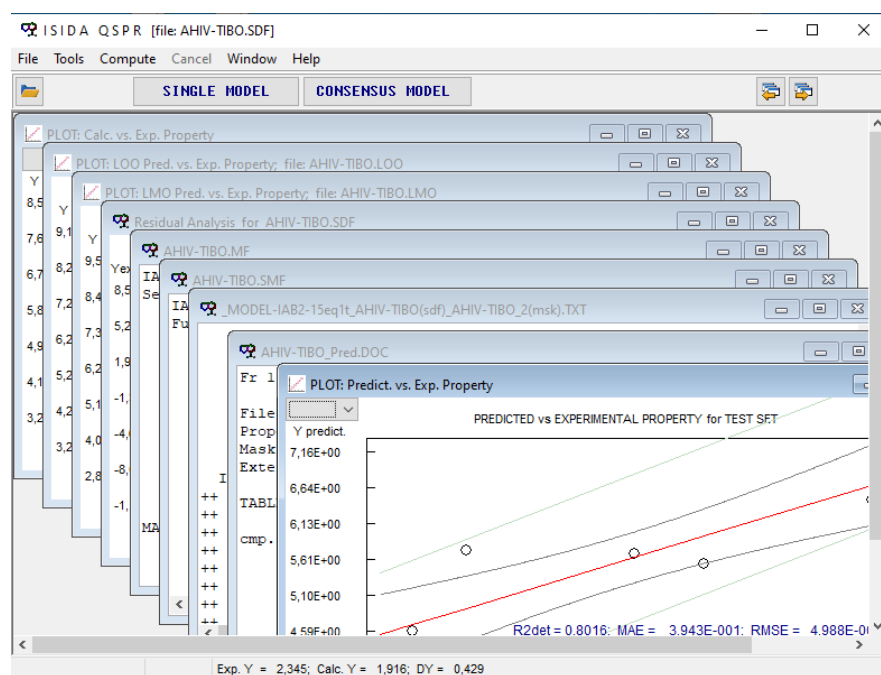


Сохранить файл маски под предложенным именем (*.msk). Просмотреть содержимое полученного файла в текстовом редакторе.

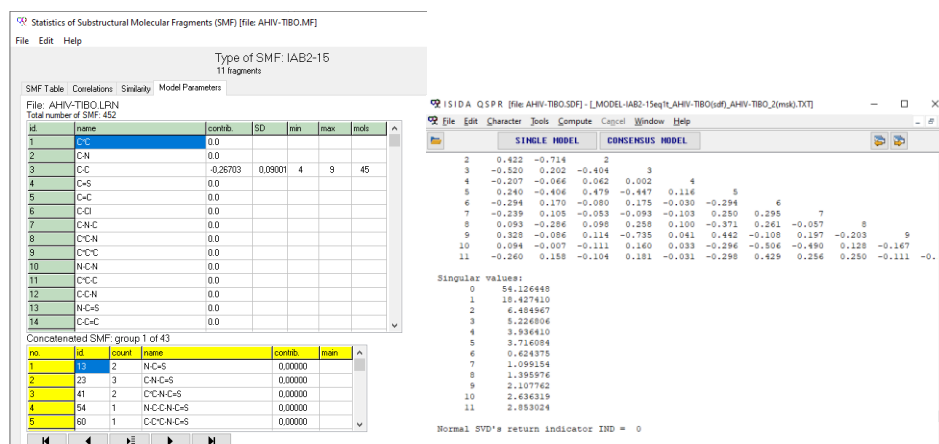
Запустить анализ данных «Data Test», чтобы проверить корректность введенных данных, получить подтверждение от системы о внутреннем согласии данных входного файла. Нажать кнопку ОК, чтобы закрыть диалоговое окно «Информация».

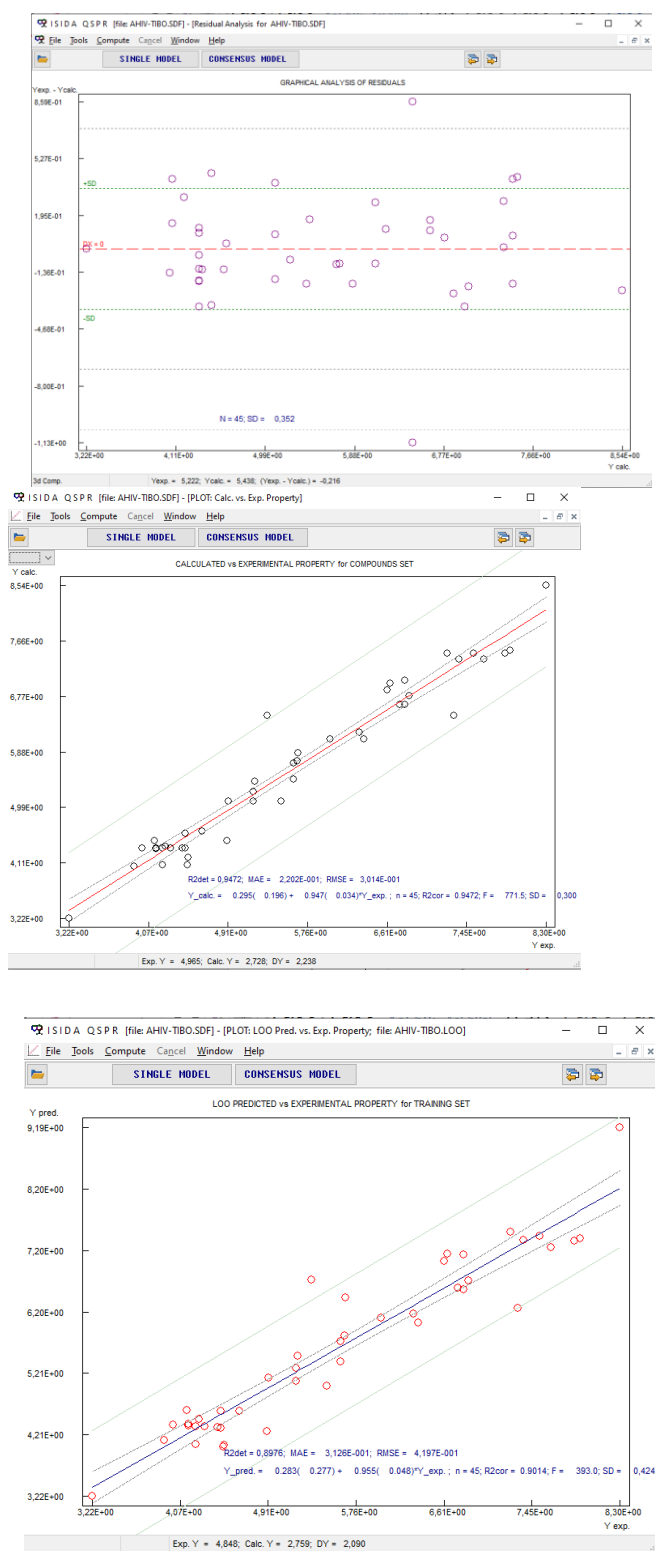
Настройка в диалоговом окне «Single Model Calculation» полей: «Descriptors», «Model»... для построения модели. Установить флажки и значения параметров в полях, как предложено на рисунке (!?или по умолчанию).

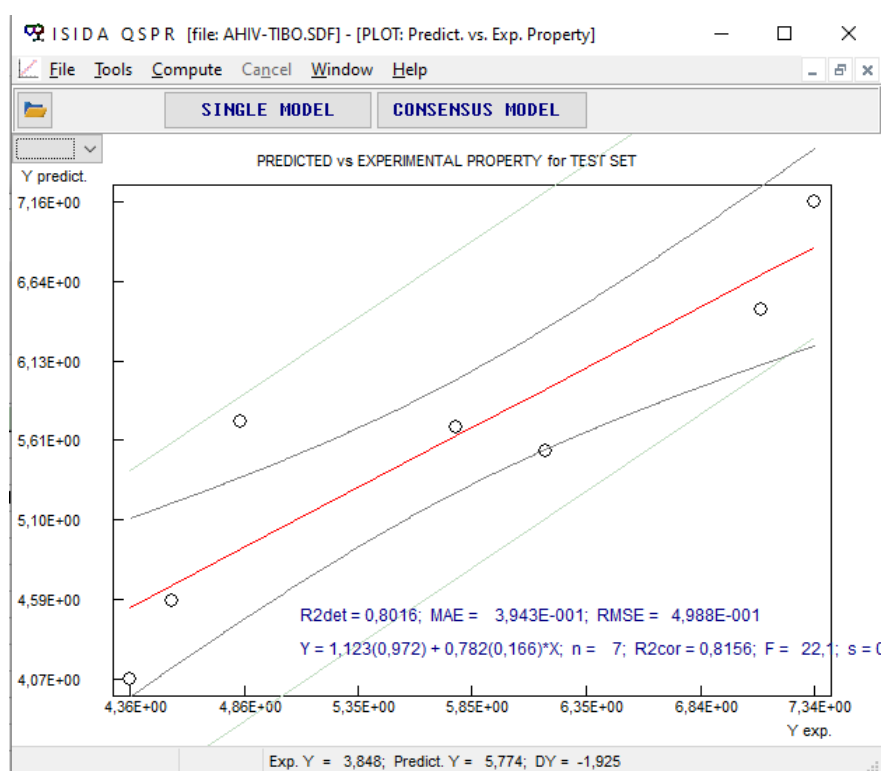
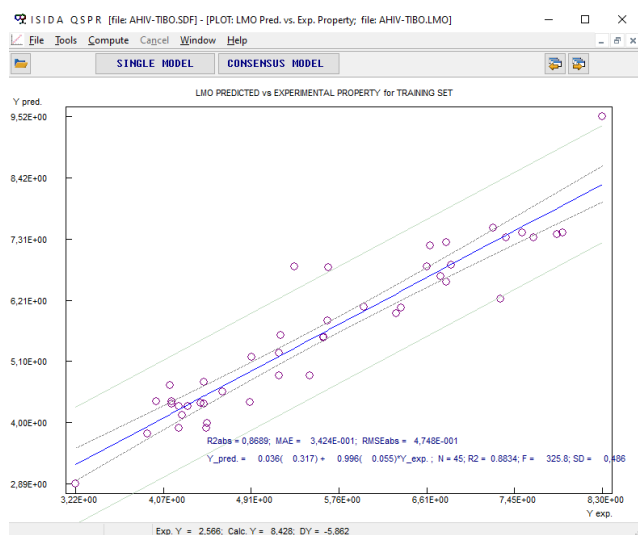
Выбрать место на диске для хранения выходных файлов. Нажать кнопку «Start» в диалоговом окне «Single Model Calculation». Программа создает 9 выходных файлов с представлением результатов: 4 текстовых и 5 графических.



В отчете провести описание всех сгенерированных файлов. Переключение между окнами с разными файлами можно осуществлять при помощи стрелок, расположенных в верхней правой части экрана. Текстовые файлы: *.SMF; *.TXT; *.MF; *.DOC. Графические файлы: *.RAC; *.PLT – 2 шт.; *.LOO; *.LMO.





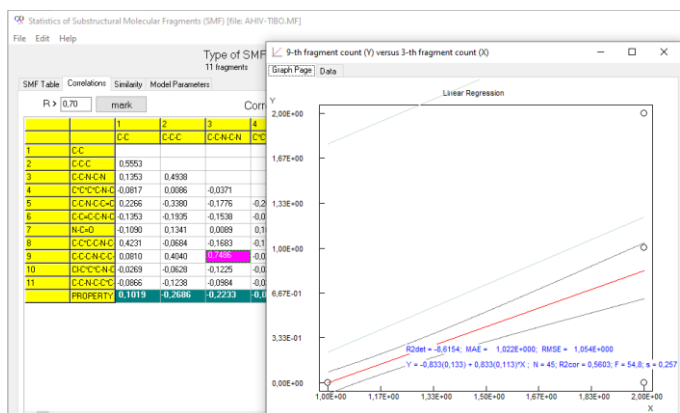
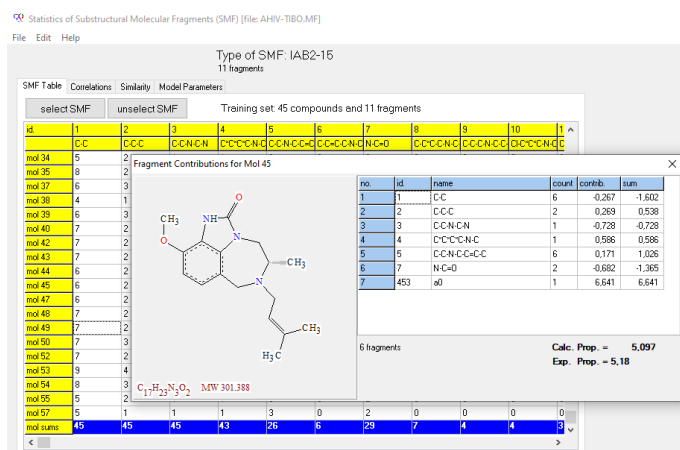


Для построенного регрессионного уравнения в отчете привести значения параметров. Из какой таблицы можно получить результирующие значения параметров? Какие фрагменты исследуемых объектов вносят существенный вклад в уравнение регрессии? Показать расчет прогнозируемого свойства для 1 из объектов обучающей выборки.

1.2 Модуль MolFrag.

Анализ вкладов субструктурных молекулярных фрагментов (SMF) в построенную модель

Запустить модуль MolFrag из главного меню ISIDA/QSPR – Tools → MolFrag (или Ctrl + F). При каждом новом запуске программы/модуля MolFrag, необходимо загружать выходной файл *.SMF. Модуль MolFrag открывает инструмент, который позволяет работать с табличными данными: “SMF Table”, “Correlations”, “Similarity”, “Model Parameters”. Что представляют собой эти таблицы, основное назначение?



Statistics of Substructural Molecular Fragments (SMF) [file: AHIV-TIBO.MF]

File Edit Help

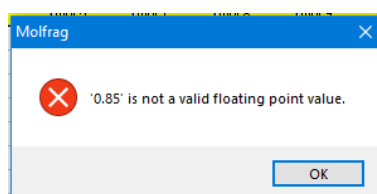
Type of SMF: IAB2-15
11 fragments

SMF Table Correlations Similarity Model Parameters

TC > 0.85 Mark

Similarity: Tanimoto coefficient (TC)

	mol 2	mol 3	mol 4	mol 5	mol 7	mol 8	mol 9	mol 10	mol 12	mol 13
mol 24	0.5600	0.6786	0.5000	0.4074	0.4286	0.4828	0.4828	0.5600	0.7200	0.7308
mol 25	0.3636	0.5200	0.7647	0.6111	0.6316	0.7000	0.7000	0.8750	0.5455	0.5000
mol 27	0.6000	0.6667	0.4545	0.4091	0.4348	0.3846	0.3846	0.4545	0.8000	0.8095
mol 28	0.6667	0.7200	0.5217	0.4783	0.3846	0.4444	0.4444	0.5217	0.8571	0.8636
mol 29	0.7778	0.6667	0.4545	0.4091	0.4348	0.3846	0.3846	0.4545	0.8000	0.7273
mol 30	0.4706	0.5000	0.7857	0.8462	0.5294	0.6111	0.6111	0.7857	0.6111	0.5500
mol 32	0.4706	0.5000	0.7857	0.8462	0.5294	0.6111	0.6111	0.7857	0.6111	0.5500
mol 33	0.6000	0.8182	0.6000	0.5900	0.4348	0.5000	0.5000	0.6000	1.0000	0.9000
mol 34	0.4706	0.5000	0.7857	0.8462	0.5294	0.6111	0.6111	0.7857	0.6111	0.5500
mol 35	0.6500	0.6875	0.6250	0.6700	0.6700	0.6000	0.6000	0.7778	0.7500	0.6000



Statistics of Substructural Molecular Fragments (SMF) [file: AHIV-TIBO.MF]

File Edit Help

Type of SMF: IAB2-15
11 fragments

SMF Table Correlations Similarity Model Parameters

File: AHIV-TIBO.LRN
Total number of SMF: 452

id.	name	contrib.	SD	min	max	mols
1	C-C	0.0				
2	C-N	0.0				
3	C-C	-0.26703	0.09001	4	9	45
4	C=S	0.0				
5	C=C	0.0				
6	C-Cl	0.0				
7	C-N-C	0.0				
8	C-C-N	0.0				
9	C-C-C	0.0				
10	N-C-N	0.0				
11	C-C-C	0.0				
12	C-C-N	0.0				
13	N-C-S	0.0				
14	C-C=C	0.0				

Concatenated SMF: group 1 of 43

no.	id.	count	name	contrib.	main
1	13	2	N-C=S	0.00000	
2	23	3	C-N-C=S	0.00000	
3	41	2	C-C-N-C=S	0.00000	
4	54	1	N-C-C-N-C=S	0.00000	
5	60	1	C-C-C-N-C=S	0.00000	

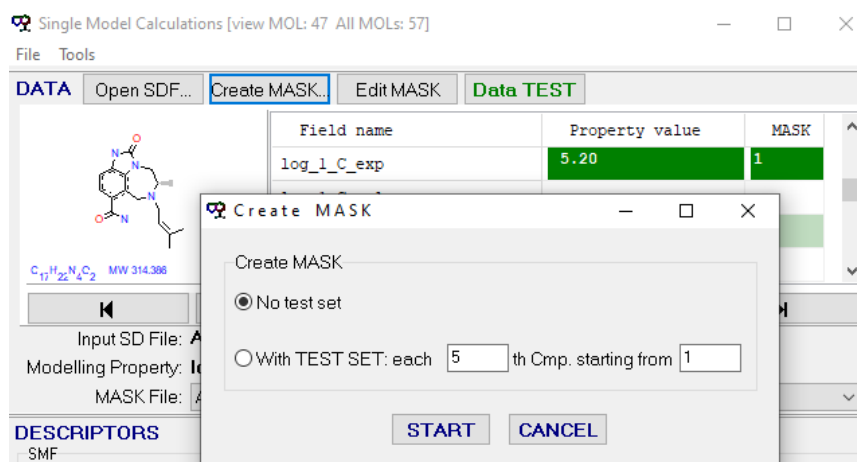
Navigation buttons: ⏮ ⏪ ⏩ ⏭

1.3 Внешняя n-кратная перекрестная проверка

Прогностическую способность модели оценивают на внешней выборке при помощи процедуры перекрестного контроля. В главном меню ISIDA/QSPR запустить формирование контрольной выборки «Create MASK». В диалоговом окне «Create MASK», установить флажок «No test set», нажать кнопку «START», проверить название файла маски (*.msk).

Запустить анализ данных «Data Test», чтобы проверить корректность введенных данных, получить подтверждение от системы о внутреннем согласии данных входного файла. Нажать кнопку ОК, чтобы закрыть диалоговое окно «Информация».

В диалоговом окне «Single Model Calculation», в поле «Validation» установить значение «5», что соответствует 5-ти кратному перекрестному контролю. Нажать кнопку «START».



5-Fold External Cross-Validation Procedure.

Fr 1/5/2020 15:36:11.971

File of Mol Structures: AHIV-TIBO.SDF; 57 compounds in training set.

Modeling Property Name: log₁C_{exp}

Mask File: AHIV-TIBO_3.MSK

Exter.Descriptors File: -

Using t-Test; Calc. Q2; Rare frag. occurs = 1; Min frag. count = 3;

no fragment	fitting	n	k	R2	F	FIT	s
type	equation						
1 IAB2-15	1	45	12	0,947204	53,82	3,567	3,52E-01
2 IAB2-15	1	45	15	0,964007	57,39	3,334	2,83E-01
3 IAB2-15	1	46	10	0,922052	47,32	3,353	4,09E-01
4 IAB2-15	1	46	18	0,978742	75,83	3,848	2,30E-01
5 IAB2-15	1	46	11	0,929812	46,37	3,176	3,96E-01

AHIV-TIBO.DOC

Fr 1/5/2020 15:35:54.766

File of Mol Structures: AHIV-TIBO.SDF

Property Name: log₁C_{exp}

Mask File: AHIV-TIBO_3.MSK

Exter.Descriptors File: -

Using t-Test; Tab. value 1.96; Calc. Q2 for each 1 point; Calc. LMO: each 5

5-Fold External Cross-Validation procedure.

TABLE L1. Training set: Calculated property log₁C_{exp} for the compounds from th

cmp. no.	Exp.	1	2	3	4	5
1	7,34	IAB2-151	IAB2-151	IAB2-151	IAB2-151	IAB2-151
2	6,80	7,02	7,39	7,63	7,49	7,58
3	5,20	4,61	5,22	6,97	7,00	7,20
4	4,64	4,61	5,00	4,28	5,32	5,07
5	4,49	4,19	4,51	4,28	4,40	4,28
6	6,17	6,35	6,03	6,15	5,67	5,67
7	5,66	5,87	6,03	5,65	5,67	5,67
8	4,13	4,46	4,11	4,27	4,58	4,58
9	4,90	4,46	4,51	4,28	4,58	4,58
10	4,48	4,07	4,39	4,58	4,31	4,31
11	4,32	4,51	4,28	4,27	4,05	4,05
12	4,92	5,10	5,22	5,25	5,04	5,04
13	6,84	6,77	6,93	6,88	7,04	7,04
14	6,80	6,63	6,93	6,72	6,66	6,66
15	5,61	5,70	5,89	5,52	5,67	5,67
16	7,11	7,50	6,54	6,46	7,11	6,77
17	7,92	7,50	7,63	8,01	8,14	8,14
18	7,64	7,36	7,65	7,52	7,76	7,76
19	4,25	4,37	4,43	4,09	4,27	4,27

AHIV-TIBO.AVE

File of Mol Structures: AHIV-TIBO.SDF

Property Name: log₁C_{exp}

Mask File: AHIV-TIBO_3.MSK

Exter.Descriptors File: -

Using t-Test; Tab. value 1.96; Calc. Q2 for each 1 point; Calc. LMO: each 5 point;

TABLE LA. Training set: Average calculated property log₁C_{exp} for the compounds from th

cmp. no.	Exp.	Average	STDEV	Nm	Exp.- Ave.
1	7,34000E+000	7,52250E+000	1,086E-001	4	-1,825E-001
2	6,80000E+000	7,04750E+000	1,037E-001	4	-2,475E-001
3	5,20000E+000	5,25500E+000	1,457E-001	4	-5,500E-002
4	4,64000E+000	4,54250E+000	3,424E-001	4	9,750E-002
5	4,49000E+000	4,34500E+000	1,386E-001	4	1,450E-001
6	6,17000E+000	6,05000E+000	2,857E-001	4	1,200E-001
7	5,66000E+000	5,80500E+000	1,799E-001	4	-1,450E-001
8	4,13000E+000	4,35500E+000	2,073E-001	4	-2,250E-001
9	4,90000E+000	4,45750E+000	1,282E-001	4	4,425E-001
10	4,48000E+000	4,33750E+000	2,112E-001	4	1,425E-001
11	4,32000E+000	4,27750E+000	1,879E-001	4	4,250E-002
12	4,92000E+000	5,15250E+000	9,912E-002	4	-2,325E-001
13	6,84000E+000	6,90500E+000	1,121E-001	4	-6,500E-002
14	6,80000E+000	6,73500E+000	1,353E-001	4	6,500E-002
15	5,61000E+000	5,69500E+000	1,820E-001	4	-8,500E-002
16	7,11000E+000	6,72000E+000	2,913E-001	4	3,900E-001
17	7,92000E+000	7,82000E+000	3,039E-001	4	1,000E-001
18	7,64000E+000	7,57250E+000	1,723E-001	4	6,750E-002
19	4,25000E+000	4,39000E+000	1,488E-001	4	-4,000E-002
20	5,65000E+000	5,39250E+000	2,460E-001	4	2,575E-001
21	4,84000E+000	5,31750E+000	1,162E-001	4	-4,775E-001
22	5,18000E+000	5,48500E+000	1,330E-001	4	-6,500E-002
23	4,22000E+000	4,26250E+000	1,365E-001	4	-4,250E-002
24	6,64000E+000	6,76250E+000	1,471E-001	4	-1,225E-001
25	5,18000E+000	5,24250E+000	8,261E-002	4	-6,250E-002
26	3,80000E+000	3,72500E+000	1,330E-001	4	7,500E-002
27	8,30000E+000	8,20500E+000	2,674E-001	4	9,500E-002

AHIV-TIBO_Pred.ECV

Fr 1/5/2020 15:35:55.24

5-Fold External Cross-Validation procedure.

File of Mol Structures: AHIV-TIBO.SDF

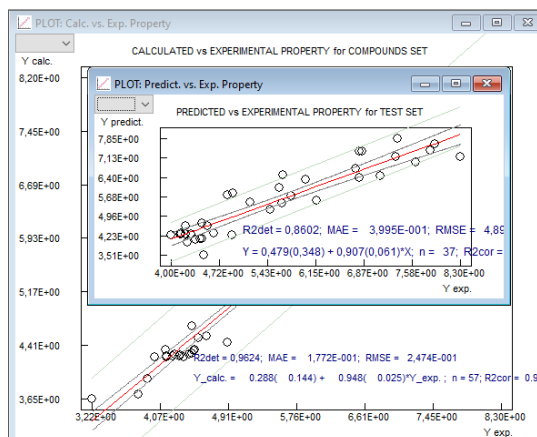
Property Name: log₁C_{exp}

Mask File: AHIV-TIBO_3.MSK

Exter.Descriptors File: -

TABLE P1. Test set: Predicted property log₁C_{exp}

cmp. no.	Datum	IAB2-151
1	7,34	7,16
6	6,17	5,55
11	4,32	
16	7,11	6,46
21	4,84	5,74
26	3,80	
31	5,78	5,70
36	4,54	4,58
41	5,94	
46	6,73	
51	5,23	
56	4,36	4,07
2	6,80	7,39
7	5,66	6,51
12	4,92	5,81
17	7,92	7,65
22	5,18	
27	8,30	7,16
32	4,46	4,11
37	4,43	4,11
42	7,25	
47	7,32	
52	6,31	
57	4,46	4,71
3	5,20	
8	4,13	4,28
13	6,84	7,38
18	7,64	6,97
23	4,22	4,58
28	6,74	6,72
33	5,48	5,22
38	4,15	4,28
43	6,00	6,32



1.4 Построение консенсусной модели

Программа ISIDA/QSPR, используя определенный пользователем метод отбора дескрипторов генерирует множество различных линейных моделей на основе набора дескрипторов SMF. При формировании консенсусной модели среди построенных моделей, используется определенный отбор, соответствующий ряду критериев.

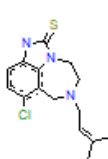
Запустить ПП ISIDA/QSPR. В главном меню ISIDA/QSPR «SINGLE MODEL» загрузить входной файл, сформировать контрольную выборку «Create MASK». В диалоговом окне «Create MASK», установить флажок «No test set», нажать кнопку «START», проверить название файла маски (*.msk).

Запустить анализ данных «Data Test», чтобы проверить корректность введенных данных, получить подтверждение от системы о внутреннем согласии данных входного файла. Нажать кнопку ОК, чтобы закрыть диалоговое окно «Информация».

Single Model Calculations [view MOL: 1 All MOLs: 57]

File Tools

DATA Open SDF... Create MASK... Edit MASK Data TEST



C₁₅H₁₆ClN₂S MW 307.846

Field name	Property value	MASK
log ₁ C _{exp}	7.34	1
log ₁ C _{calc}	7.684	
AU	Jarmo Huuskonen	
TI	QSAR Modeling with t	

Input SD File: AHIV-TIBO.SDF
Modelling Property: log₁C_{exp}
MASK File: AHIV-TIBO_4.MSK

DESCRIPTORS

SMF

☐ Paths: Atoms

☐ Paths: Bonds

☒ Paths: Atoms-Bonds

☐ Augmented Atoms

☐ Augmented Bonds

☐ Augmented Atoms-Bonds

☐ Atomic Hybridization

Path Length

min: 3

max: 12

min count: 3

min comps: 1

SMF Modifications

☐ terminal groups

☐ equal length paths

☐ Marked atoms 1

☐ Marked atoms 3

☐ Benson's scheme

☐ normalize

☐ add XTR

☐ only XTR

[]

MODEL

Equation

☒ Y = SUM(Ai*Xi)

☐ Y = Ao + SUM(Ai*Xi)

☐ Y = Ao + SUM(Ai*Xi) + SUM[Bi*(Zi*Xi-1)]

☐ Y = Ao + SUM(Ai*Xi) + SUM[Bik*Xi*Xk]

Variable Selection

☒ forward variable selection

Ry,ij

Ry > 0,001

Rij < 0,990

☒ backward variable selection

t-test: 1.96

eps: 1.0E-12

exact: 0

☐ m > n

VALIDATION

☒ LOO ☐ rapid

External 1 - fold CV

☒ Appl.Domain1

☒ LMD 5 th point

☒ Appl.Domain2

Output Directory: D:\SIDA_ZIP\X\1_may\

☐ Get SMF file only

START CANCEL

В диалоговом окне «Single Model Calculation» нажать кнопку «CANCEL». Перейти в диалоговое окно «Consensus Modelling Calculations», нажав кнопку «Consensus model» на стартовом экране ISIDA/QSPR. Заполнить все активные поля в соответствии с рисунком. Нажать кнопку «START».

Consensus Modelling Calculations

File

DATA

Structure Data File
AHIV-TIBO.SDF

Modeling Property
log_1_C_exp

Mask File
AHIV-TIBO_4.MSK

DESCRIPTORS

SMF

☐ Paths: Atoms

☐ Paths: Bonds

☒ Paths: Atoms-Bonds

☐ Augmented Atoms

☐ Augmented Bonds

☐ Augmented Atoms-Bonds

☐ Atomic Hybridization

☐ add XTR

☐ []

Path Length

min: 2-3

max: 7-12

SMF Modifications

☐ terminal groups

☒ equal length paths

☐ Benson's scheme

☐ Marked atoms 1

☐ Marked atoms 2

☐ Marked atoms 3

☐ normalize

min count: 3

min compds: 1

MODEL

Equation

☒ $Y = \sum(A_i \cdot X_i)$

☐ $Y = A_0 + \sum(A_i \cdot X_i)$

Variable Selection

☒ FVS methods:

3 Ry.i

Ryi > 0.001

Rij < 0.990

☒ backward variable selection

t-test: 1.96

eps: 1.0E-12

exact: 0

☐ m > n

CONSENSUS MODEL

Q2 lim > 0.70

R2-Q2 < 0.30

VALIDATION

☒ LOO ☒ rapid

External 5-fold CV

☒ Appl.Domain1

☐ Appl.Domain2

☐ Appl.Domain3: > 15 % models

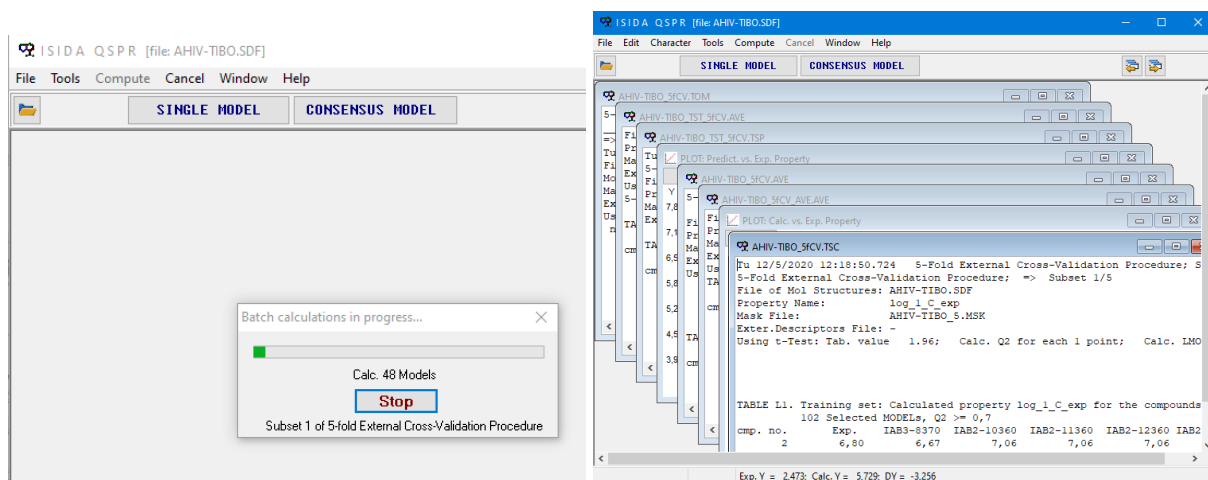
☐ Save Models: **Select Directory**

Select Directory for output files:

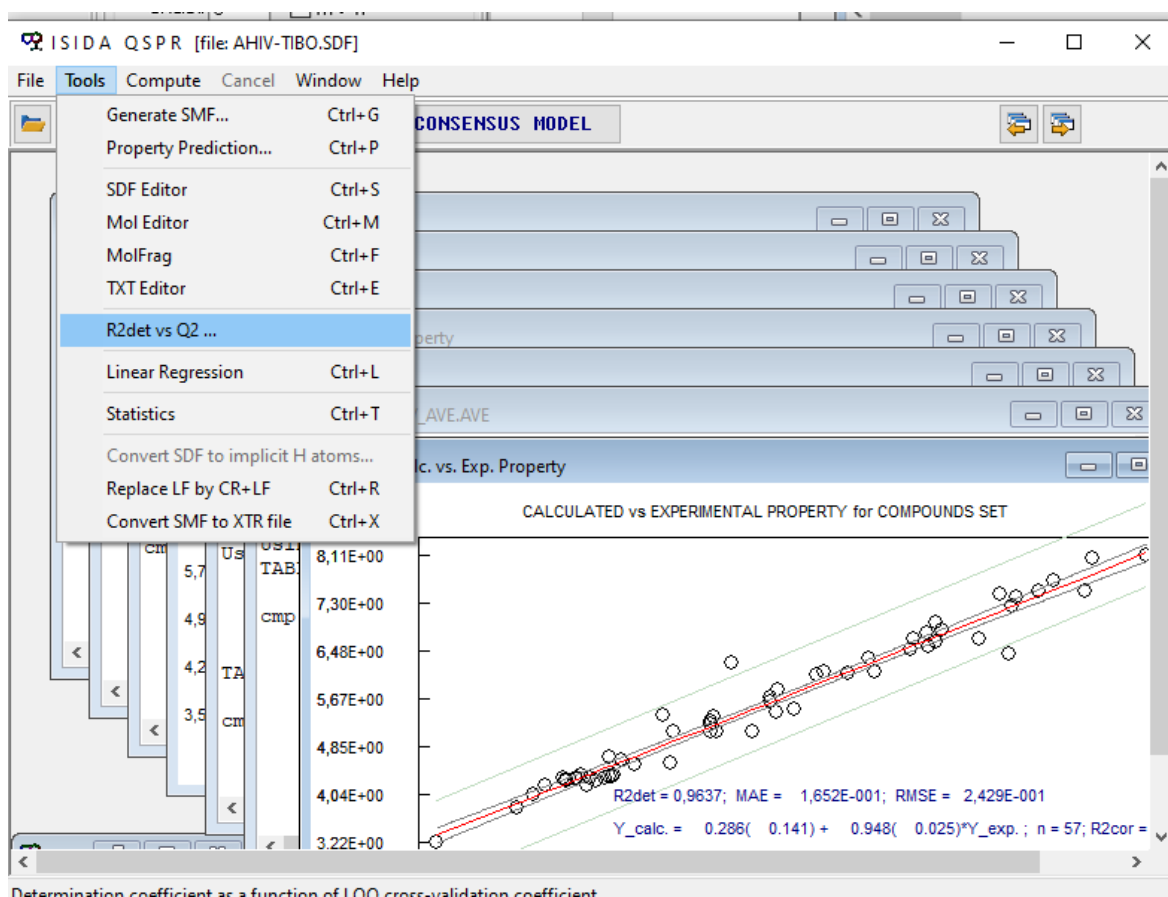
D:\ISIDA_ZIP\X\1_may\

START **CANCEL**

Программа начинает расчет моделей, который в зависимости от выбранных условий может составить от 10 до 30 минут. Программа создает 8 выходных файлов с представлением результатов: 6 текстовых и 2 графических.

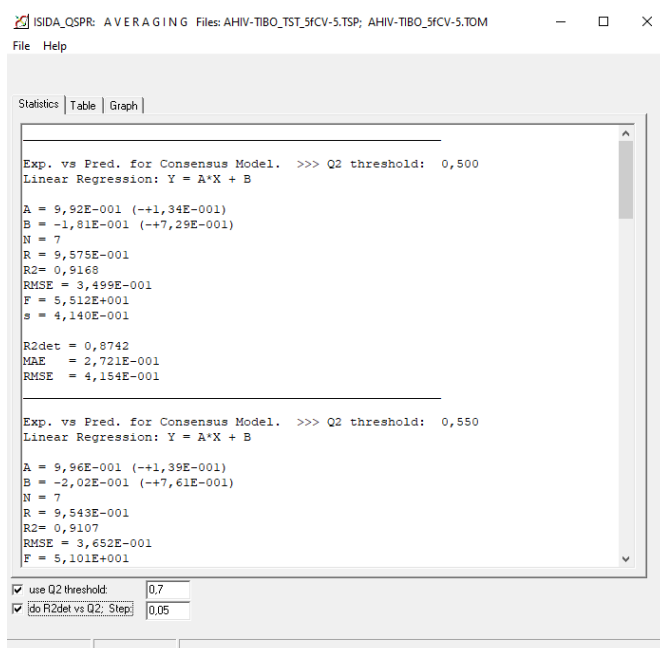
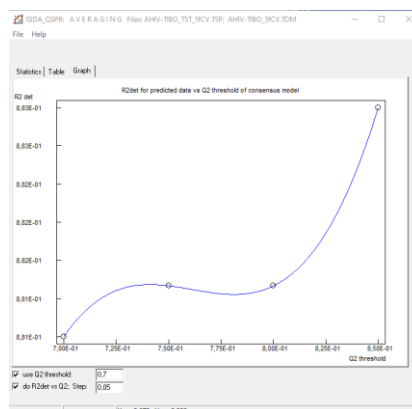
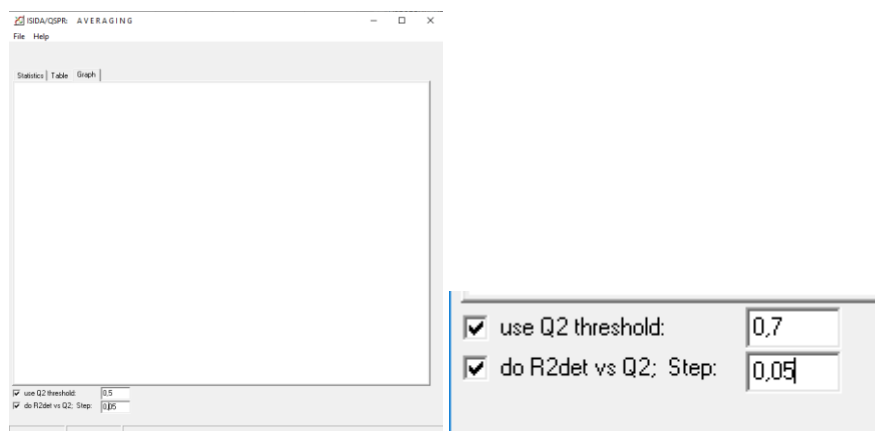


Определение оптимального значения коэффициента детерминации, Q^2_{lim} . Построить график зависимости коэффициентов R^2_{det} от Q^2_{lim} . Для этого перейти из главного меню пройти по ссылке «Tools → R2det vs Q2».



На экране отображается пустое поле. Поставить флажки перед полями: «use Q2 threshold» и «do R2det vs Q2», ввести значения 0,7 и 0,05 соответственно. Для построения графика R^2_{det} от Q^2_{lim} , загрузить по запросу (File → Open), последовательно

файлы с данными *_5fCV.TSP и *_5fCV.TOM. Инструмент Averaging проводит вычисления после чего в информационном окне «Averaging» становятся доступными данные: 2 таблицы «Statistic», «Table» и график «Graph».



ISIDA_QSPR: AVERAGING Files: AHIV-TIBO_TST_5FCV.TSP; AHIV-TIBO_5FCV.TOM

File Help

Statistics Table Graph

TABLE C. R2det for predicted data vs Q2 threshold of consensus model.

Cmpd_sum	R2det	MAE	RMSE	Q2thre	MDL_sum
45	0,8811	3,438E-001	4,385E-001	0,700	27 109
45	0,8815	3,428E-001	4,378E-001	0,750	27 109
45	0,8815	3,437E-001	4,379E-001	0,800	27 109
45	0,8829	3,434E-001	4,352E-001	0,850	13 102

☒ use Q2 threshold: 0.7
☒ do R2det vs Q2: Step: 0.05

X = 0,679; Y = 0,882

Построение консенсусной модели для всего набора данных. Выполнить заново вход в окно «Consensus model», в поле «External n-fold CV» ввести значение «1». Проверить место размещения выходных файлов (результатов анализа). Поставить флажок в поле «Save Models». Нажать кнопку «START». Программа создает 5 выходных файлов с представлением результатов: 4 текстовых и 1 графический.

Consensus Modelling Calculations

File

DATA

Structure Data File: AHIV-TIBO.SDF

Modeling Property: log_1_C_exp

Mask File: AHIV-TIBO_5.MSK

DESCRIPTORS

SMF

☐ Paths: Atoms

☐ Paths: Bonds

☒ Paths: Atoms-Bonds

☐ Augmented Atoms

☐ Augmented Bonds

☐ Augmented Atoms-Bonds

☐ Atomic Hybridization

☐ add XTR

Path Length

min: 2-3

max: 7-12

SMF Modifications

☐ terminal groups

☒ equal length paths

☐ Benson's scheme

☐ Marked atoms 1

☐ Marked atoms 2

☐ Marked atoms 3

☐ normalize

min count: 3

min comps: 1

MODEL

Equation

☒ $Y = \text{SUM}(A_i^{\alpha} X_i)$

☐ $Y = A_0 + \text{SUM}(A_i^{\alpha} X_i)$

Variable Selection

☒ FVS methods:

3 Ry,ij

Ryi > 0,001

Rij < 0,990

☒ backward variable selection

t-test: 1,96

eps: 1,0E-12

exact: 0

☐ m > n

CONSENSUS MODEL

Q2 lim > 0,70

R2-Q2 < 0,10

VALIDATION

☒ LOO

☒ rapid

External: 1 -fold CV

☒ Appl.Domain1

☐ Appl.Domain2

☐ Appl.Domain3: 15 % models

☒ Save Models

Select Directory for output files:

F:\ISIDA_ZIP\X\12may\

START CANCEL

[illegible]

FILE OF HAIL STRUCTURES...

File of Hail Structures: ARTV-TIMO.REP
 Registry Name: log_1200
 Base File: ARTV-TIMO_5_HMR
 Power-Designation File:
 Using T-Test: No
 Using T-Test: No

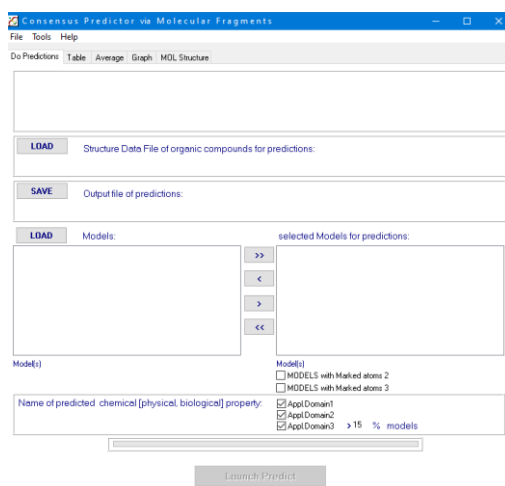
[illegible]

1.5 Прогнозирование свойств, виртуальный скрининг с использованием моделей консенсуса

Использование ранее полученных консенсусных моделей для прогноза свойств на внешнем наборе данных с помощью инструмента «Consensus Predictor».

Запустить ПП ISIDA/QSPR, из главного меню пройти «Tools → Property Prediction». В диалоговом окне «Consensus Predictor via Molecular Fragments» нажать на кнопку «LOAD» для загрузки структурных данных новых объектов, TEST_ANIV-TIBO.SDF. Автоматически формируется название выходного файла *.TSP, в той же директории, либо нажав на клавишу «SAVE» можно изменить название, путь для сохранения выходного файла *.TSP.

Выполнить загрузку списка ранее полученных консенсусных моделей, нажать на кнопку «LOAD», из предложенного списка файлов *.SPE выбрать предпочтительную модель или загрузить все предложенные. Список из 117 моделей появляется в левой части экрана, кнопкой «>>» перенести все файлы в правую часть экрана. Поставить флажки в 3-х полях с областью применимости, и ввести число «50» для AD № 3. Нажать кнопку «Launch Predict». Расчет приводит к формированию 4-х вкладок в диалоговом окне «Consensus Predictor via Molecular Fragments».



Consensus Predictor via Molecular Fragments

File Tools Help

Do Predictions Table Average Graph MOL Structure

LOAD Structure Data File of organic compounds for predictions:
F:\SIDA_ZIP\TEST_AHIV-TIBO.SDF

SAVE Output file of predictions:
F:\SIDA_ZIP\TEST_AHIV-TIBO_FMF.TSP

LOAD Models:

selected Models for predictions:

118 model(s)

118 model(s)

☐ MODELS with Marked atoms 2

☐ MODELS with Marked atoms 3

Name of predicted chemical [physical, biological] property:
log₁C_{exp}

☒ Appl Domain1

☒ Appl Domain2

☒ Appl Domain3 > 15 % models

Folder F:\SIDA_ZIP\

Launch Predict

Consensus Predictor via Molecular Fragments

File Tools Help

Do Predictions Table Average Graph MOL Structure

LOAD Structure Data File of organic compounds for predictions:
F:\SIDA_ZIP\TEST_AHIV-TIBO.SDF

SAVE Output file of predictions:
F:\SIDA_ZIP\TEST_AHIV-TIBO_FMF.TSP

LOAD Models:

selected Models for predictions:

118 model(s)

118 model(s)

☐ MODELS with Marked atoms 2

☐ MODELS with Marked atoms 3

Name of predicted chemical [physical, biological] property:
log₁C_{exp}

☒ Appl Domain1

☒ Appl Domain2

☒ Appl Domain3 > 50 % models

Folder F:\SIDA_ZIP\

Launch Predict

Consensus Predictor via Molecular Fragments

File Tools Help

Do Predictions Table Average Graph MOL Structure

File: F:\SIDA_ZIP\XH\TEST_AHIV-TIBO_FMF.TSP

File of Mol Structures: F:\SIDA_ZIP\TEST_AHIV-TIBO.SDF

Property Name: log₁C_{exp}

TABLE PP. Predicted property log₁C_{exp} for the complexes

cmp. no.	Datum	IAB2-10250	IAB2-10260	IAB2-10270
1	-	7,9449E+00	7,9449E+00	7,9449E+00
2	-	7,7441E+00	7,7441E+00	7,7441E+00
3	-	8,2914E+00	8,2914E+00	8,2914E+00
4	-	7,4990E+00	7,4990E+00	7,4990E+00
5	-	7,4990E+00	7,4990E+00	7,4990E+00
6	-	5,9007E+00	5,9007E+00	5,9007E+00
7	-	7,1955E+00	7,1955E+00	7,1955E+00
8	-	7,5702E+00	7,5702E+00	7,5702E+00
9	-	8,9698E+00	8,9698E+00	8,9698E+00
10	-	8,9698E+00	8,9698E+00	8,9698E+00
11	-	8,9698E+00	8,9698E+00	8,9698E+00
12	-	8,9698E+00	8,9698E+00	8,9698E+00

Consensus Predictor via Molecular Fragments

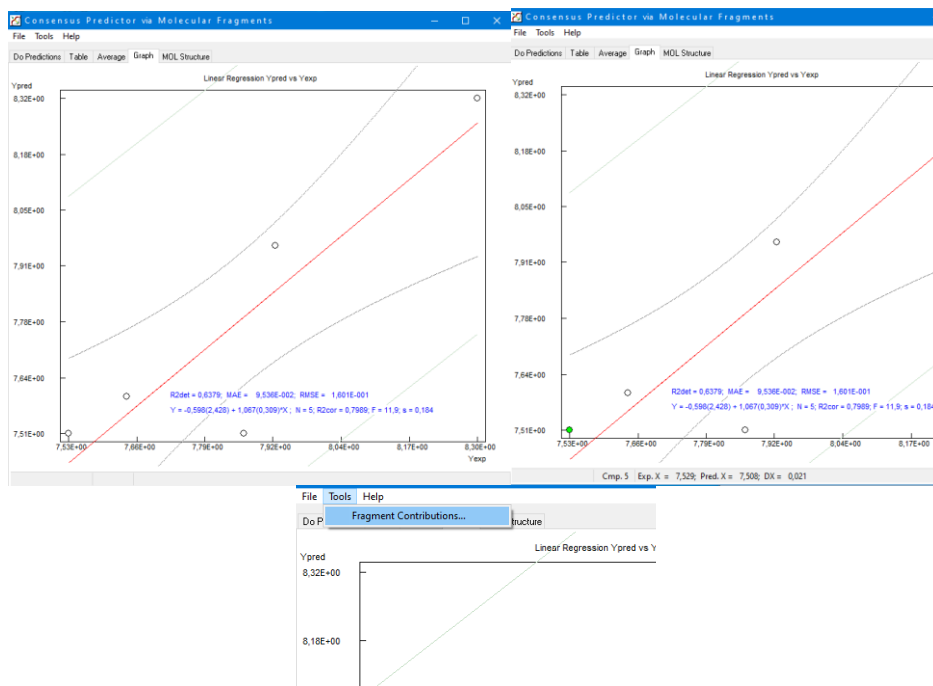
File Tools Help

Do Predictions Table Average Graph MOL Structure

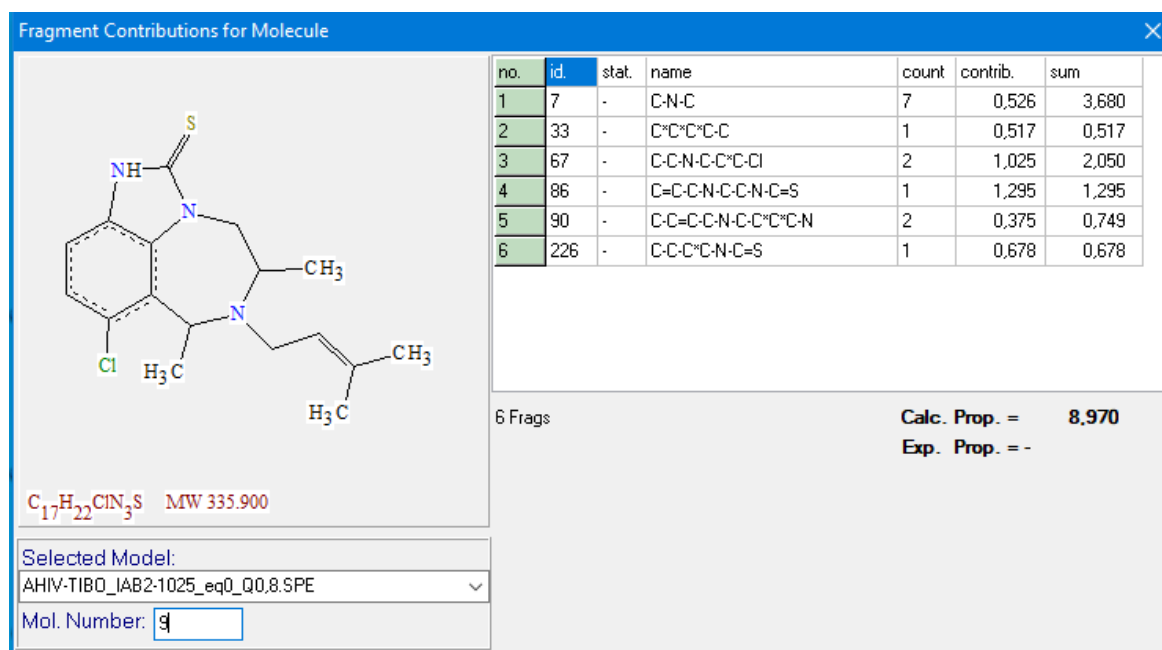
File: F:\SIDA_ZIP\TEST_AHIV-TIBO_FMF.AVE

TABLE PA. Average predicted property log₁C_{exp} for the compounds

cmp. no.	Datum	Average	STDEV	Nm
1	7,92	7,96224E+000	5,925E-002	74
2	7,64	7,59814E+000	1,160E-001	99
3	8,30	8,31903E+000	6,017E-002	74
4	7,86	7,50816E+000	3,707E-002	81
5	7,53	7,50816E+000	3,707E-002	81
6	-	6,67168E+000	4,470E-001	117
7	-	7,07149E+000	1,500E-001	95
8	-	7,53993E+000	1,079E-001	87
9	-	8,9858E+000	1,957E-001	91
10	-	7,48592E+000	4,201E-001	111
11	-	7,48592E+000	4,201E-001	0
12	-	8,47588E+000	2,735E-001	98



Нажатие на выбранную точку данных на графике визуализирует соответствующую молекулярную структуру и значения Y_{exp} и Y_{pred} (Y_{calc}). Для каждого объекта на графике можно просмотреть информацию о вкладе каждого молекулярного фрагмента в моделируемое свойство «Tools → Fragment Contributions».



Сделать вывод по прогнозу свойств на внешнем наборе данных с помощью инструмента «Consensus Predictor». На сколько точными оказались предсказанные значения свойств?

1. Количественные показатели качества регрессионных моделей.
2. Показатели описательной способности моделей.
3. Показатели предсказательной способности моделей.
4. Показатели качества статистических моделей.
5. Количественные показатели качества классификационных моделей.
6. Понятие о внутреннем и внешнем, перекрестном и скользящем контроле.
7. Сравнение качества моделей.
8. Моделирование на основе векторов дескрипторов фиксированной длины.
9. Построение моделей «структура-свойство» без использования векторов дескрипторов фиксированной длины.
10. Общие принципы валидации моделей.
11. Понятие об оверфиттинге и принцип оптимальной сложности моделей.
12. Дескрипторы на основе цепочек атомов.
13. Дескрипторы на основе атом-центрированных фрагментов.
14. Дескрипторы на основе связь-центрированных фрагментов.
15. Дескрипторы на основе случайных фрагментов.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Введение в хемоинформатику: Моделирование 'структура-свойство' / Т. И. Маджидов, И. И. Баскин, И. С. Антипин, А. А. Варнек.- Казань: [Казанский университет], 2015.- 302 с.
2. Информационная система ISIDA/QSPR для построения количественных моделей «структура-свойство» – официальный веб-сайт: <http://vpsolovev.ru/programs/isidaqspr/>
3. Введение в хемоинформатику: Компьютерное представление химических структур / Т. И. Маджидов, И. И. Баскин, И. С. Антипин, А. А. Варнек.- Казань: [Казанский университет], 2013.- 173 с.

ОГЛАВЛЕНИЕ

Обзор информационной системы ISIDA/QSPR	2
1.1 Построение индивидуальной модели (MLR) на одном наборе дескрипторов SMF, предсказание свойств на тестовом наборе	3
1.2 Модуль MolFrag. Анализ вкладов субструктурных молекулярных фрагментов (SMF) в построенную модель	8
1.3 Внешняя n-кратная перекрестная проверка	10
1.4 Построение консенсусной модели	13
1.5 Прогнозирование свойств, виртуальный скрининг с использованием моделей консенсуса	19
Контрольные вопросы по работе в ИС «ISIDA/QSPR»	22
Библиографический список	22