

**Практическое занятие по курсу
«Основы компьютерного молекулярного моделирования и OSAR»**

**Моделирование связи структуры и активности органических соединений и
прогнозирование их свойств с помощью онлайн-платформы OCHEM**

Задачи

1. Знакомство с онлайн-платформой моделирования OCHEM
2. Регистрация в системе
3. Загрузка данных о структуре и активности/свойствах
4. Построение моделей связи структуры и активности
5. Прогнозирование свойств

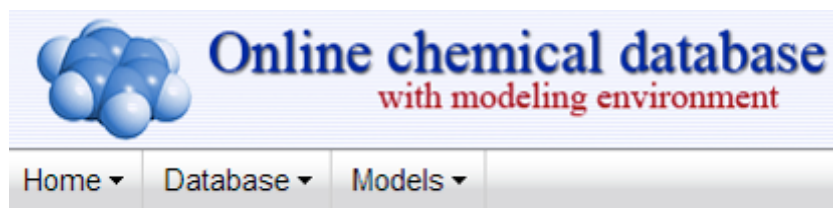
1. Знакомство с онлайн-платформой моделирования OCHEM

OCHEM – Online Chemical Modeling Environment

Разрабатывается Исследовательским центром им. Гельмгольца (Мюнхен, Германия) и компанией eADMET в партнерстве с ведущими исследовательскими группами в области хемоинформатики по всему миру

<https://www.ochem.eu/>

- Распределенная интернет-платформа
- Ввод, хранение, поиск данных по структуре и активности/свойствам
- Построение предсказательных моделей: широкий выбор дескрипторов, методов машинного обучения и процедур валидации
- Прогнозирование активности/свойств для новых структур
- Контроль области применимости моделей
- Возможность публикации данных и моделей для общего доступа
- Дополнительно: скрининг структур по токсифорным группам (ToxAlerts)
- Дополнительно: сравнение выборок соединений (SetCompare)
- (Режим интеллектуальной собственности)



2. Регистрация в системе

- В правом верхнем углу страницы щелкните [create account](#)
- Введите обязательную (отмечена звездочками) и при желании дополнительную информацию
- Прочитайте условия использования

- Нажмите кнопку [I ACCEPT. CREATE MY ACCOUNT.](#)
- Еще раз подтвердите согласие с условиями

Registration Information

Login* (min. 4 characters and max. 20 characters)

e-mail*

Password*

Confirm password*

Personal Information

Title* -- please select -- Please, select a form of address!

First name*

Last name*

Affiliation Lomonosov Moscow Stat

Form of organization* Academic Please, select a form of organization!

City

State

Country

Zip

Phone

Occupation

Company

WebSite

- Для входа в систему в правом верхнем углу страницы щелкните [log in](#)
- Введите идентификатор и пароль

Already have an account?

If you already have an account, please enter your login and password below:

Login ID

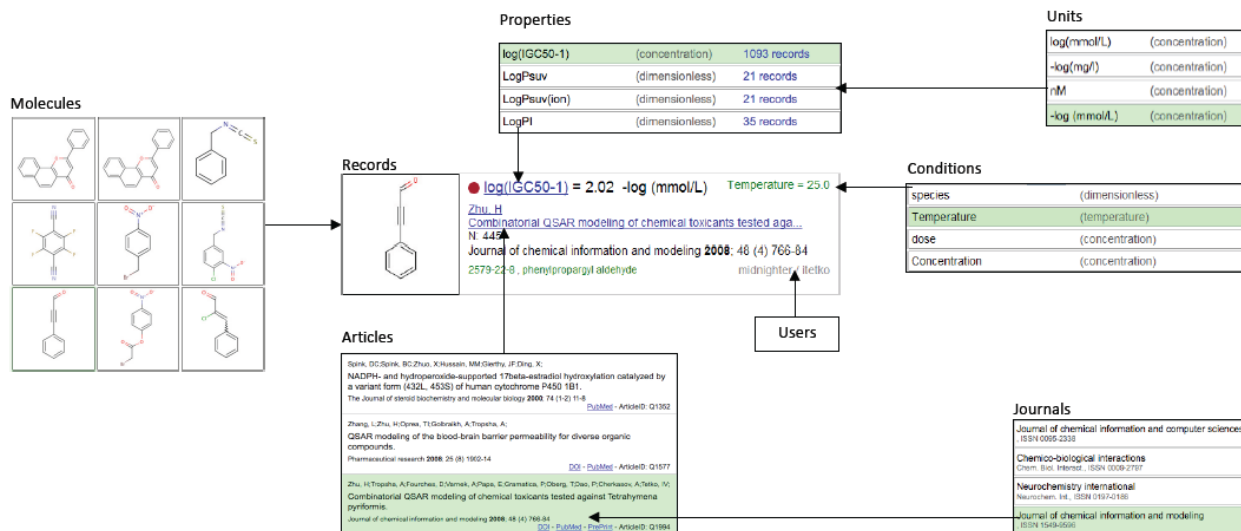
Password



- Нажмите кнопку [LOGIN](#)

3. Загрузка данных о структуре и активности/свойствах

Общая схема хранения информации о структуре и активности соединений в системе OCHEM



Мы загрузим данные по упрощенной процедуре (в частности, без ссылок на оригинальные публикации). Процесс загрузки включает несколько шагов.

- В меню Database выберите Batch data upload.
- Выберите файл со своим заданием и установите указанные на рисунке флажки (не искать данные в PubChem/PubMed, не публиковать загруженную базу).
- Нажмите кнопку Upload.

Select a file to upload

Upload file

Choose File

20GS_hCHRM3_98_Ki.sdf

Settings

☐ Allow molecule lookup by name on PubChem
☐ Allow article lookup by PMID on PubMed
☒ Make the uploaded records hidden

Upload

- Появляется экран выбора столбцов (полей) для импорта.

20GS_hCHRM3_98_Ki.sdf

<input checked="" type="checkbox"/> MOLECULE	<input type="checkbox"/> CdId	<input type="checkbox"/> ACTIVITY_ID	<input type="checkbox"/> COMPOUND_ID	<input checked="" type="checkbox"/> KI	<input type="checkbox"/> Compound Synonyms
1 2 3 25 29 0 0 0 0 999 V2000 4.02	1	2885019	433188	2760.0	
1 2 3 16 18 0 0 0 0 999 V2000 4.06	2	2885021	125692	1300.0	
1 2 3 17 20 0 0 0 0 999 V2000 -0.6	3	2880828	109405	67.0	
1 2 3 15 17 0 0 0 0 999 V2000 3.19	4	2904637	144434	1400.0	
1 2 3 29 32 0 0 0 0 999 V2000 -0.8	5	2983754	346194	14.0	
1 2 3 21 23 0 0 1 0 999 V2000 0.07	6	2972544	195	1.0	Isopto Atropine, Atropine
1 2 3 19 22 0 0 0 0 999 V2000 0.06	7	2976998	109477	1360.0	
1 2 3 15 17 0 0 0 0 999 V2000 1.83	8	2977042	423746	67000.0	
1 2 3 18 21 0 0 0 0 999 V2000 -1.4	9	2971588	322144	204.0	
1 2 3 23 27 0 0 0 0 999 V2000 -2.0	10	2971727	113397	56.0	

- Щелкните заголовок столбца с целевым свойством (в данном случае Ki) и выберите вид столбца Property.
- Появляется список зарегистрированных в системе свойств, но они нам не подходят, поэтому нажимаем [create new property](#).
- Введите (скопируйте из файла *Targets.doc*) информацию о моделируемой активности: имя, тип, скрытое свойство (только для себя), единицы измерения (нМ), описание (должно содержать не менее 50 символов). Затем нажмите [SAVE](#).

Name: Type: ☐ Hidden property ☐ Make public [?]

property name is good and length is 9

System of units ⁱ

Concentration
Default unit:

Tags [+]
No tags selected

Obligatory Conditions ⁱ

[+]

Aliases:

(comma separated list of the other names, i.e. synonyms, for this property)

Description:

warning: property description length is 57 and may not be informative

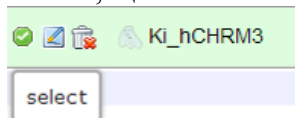
SAVE

CANCEL

- Скопируйте имя нового свойства в поле поиска и нажмите **search**

Type part of name to filter: **search**

- Чтобы выбрать нужную активность, щелкните зеленый значок ✓



- Вернувшись на экран выбора столбцов для импорта, щелкните заголовок ACTIVITY_ID и выберите вид столбца Known column. В списке выберите recordid и нажмите ОК.
- Остальные столбцы нам не потребуются.

20GS_hCHRM3_98_Ki.sdf

<input checked="" type="checkbox"/> MOLECULE	<input type="checkbox"/> CdId	<input checked="" type="checkbox"/> recordid	<input type="checkbox"/> COMPOUND_ID	<input checked="" type="checkbox"/> Ki_hCHRM3	<input type="checkbox"/> Compound Synonyms
1 2 3 25 29 0 0 0 0 999 V2000 4.02	1	2885019	433188	2760.0	
1 2 3 16 18 0 0 0 0 999 V2000 4.06	2	2885021	125692	1300.0	
1 2 3 17 20 0 0 0 0 999 V2000 -0.6	3	2880828	109405	67.0	
1 2 3 15 17 0 0 0 0 999 V2000 3.15	4	2904637	144434	1400.0	
1 2 3 29 32 0 0 0 0 999 V2000 -0.8	5	2983754	346194	14.0	
1 2 3 21 23 0 0 1 0 999 V2000 0.07	6	2972544	195	1.0	Isopto Atropine, Atropine,
1 2 3 19 22 0 0 0 0 999 V2000 0.00	7	2976998	109477	1360.0	
1 2 3 15 17 0 0 0 0 999 V2000 1.83	8	2977042	423746	67000.0	
1 2 3 18 21 0 0 0 0 999 V2000 -1.4	9	2971588	322144	204.0	
1 2 3 23 27 0 0 0 0 999 V2000 -2.0	10	2971727	113397	56.0	

- Нажмите кнопку Upload this sheet. Система выполняет предварительный анализ базы данных.
- На следующем экране (экран соответствия свойств) нажмите Submit.

Database entities remapping page

Property: **Ki_hCHRM3**

Values
Unit: **default**, min value: 0.0, max value: 86000.0

Article: **unpublished**

Molecule set: **default**

submit

- Система выполняет загрузку базы данных. Процесс может занять несколько минут.


Row 32 processed
[interrupt]

- После его завершения появляется экран предварительного просмотра загружаемой базы данных. Система указывает общее число записей, число пропущенных записей (в частности, дубликатов соединений) и число корректных записей. При необходимости можно просматривать записи и включать/выключать их вручную. Закончив проверку, нажмите кнопку Proceed with upload внизу страницы.

Batch upload preview browser

Summary:

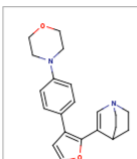
All rows in the sheet	Count: 826
Status: duplicate_internal,	Count: 20
Status: valid,	Count: 806

Filter by row number: and row type:

1 - 10 of 826 10 items on page 1 of 83 >>

Row 1

☒ Save ☐ Skip



Ki_hCHRM3 = 2760.0 (in nM)

Radchenko
20GS_hCHRM3_98_Ki.sdf...

N: AUTO_1
Unpublished 0; ()

MoleculeID: M534355

- После окончательной загрузки данных выводится информация о числе записей.




Batch upload results

Batch upload is finished. You can download the [detailed upload report](#).

Summary:

All rows in the sheet	Count: 826
Status: duplicate_internal, skipped	Count: 20
Status: valid, saved_valid	Count: 806

- Загрузка данных закончена, и информация о наборе данных («корзине») появляется в списке корзин (меню Database | Baskets).

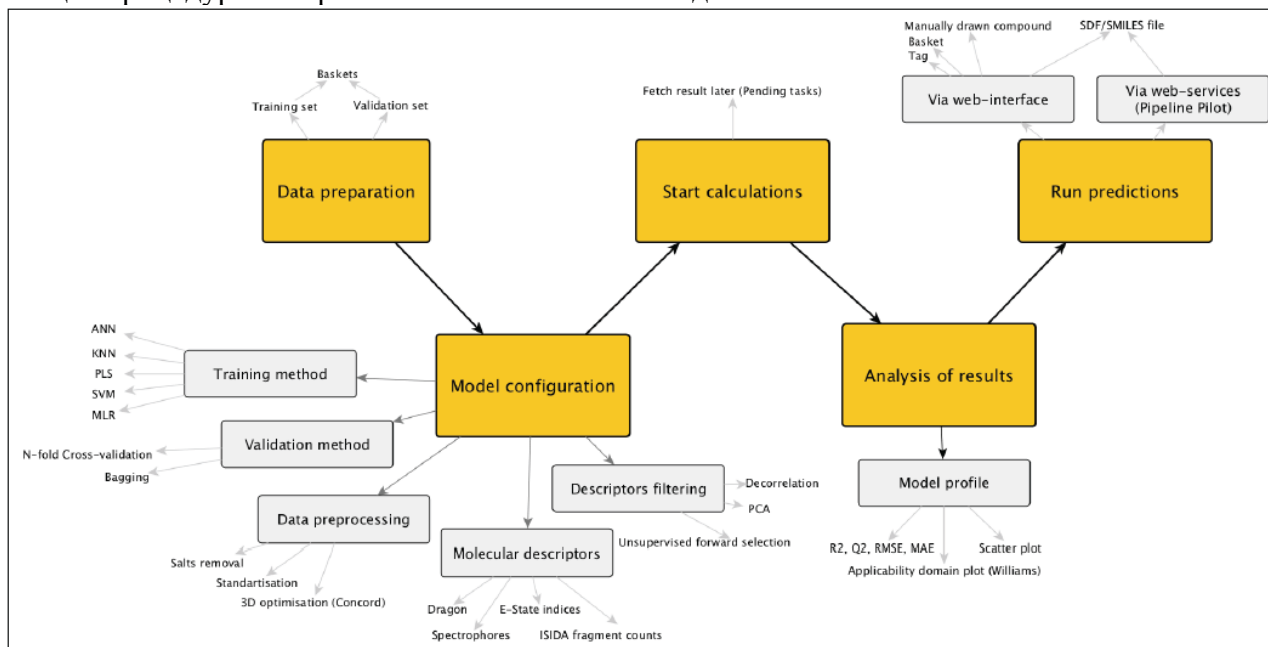




20GS_hCHRM3_98_Ki.sdf

806 records

4. Построение моделей связи структуры и активности

Общая процедура построения и использования моделей в OCHEM



Эта процедура может включать следующие этапы:

- Подготовка данных
- Задание параметров построения модели (тип модели / метод обучения, метод валидации, предварительная обработка структур, набор дескрипторов, метод отбора дескрипторов)

- Запуск расчетов и контроль их выполнения
- Анализ результатов
- Прогнозирование свойств для новых структур

Мы построим серию моделей на основе различных дескрипторов и методов обучения. Для задания необходимых параметров используется «мастер», включающий несколько экранов. Сейчас нам не нужно изменять большинство установок по умолчанию (это может потребоваться при более детальном анализе для построения наиболее точных моделей, в том числе с помощью функции Create model).

- В меню Models выберите Create multiple models.
- На первом экране щелкните многоточие в поле Training set и выберите нужный набор (Basket), щелкнув зеленый значок ✓
- Выберите логарифмическую шкалу для моделируемой активности.
- Выберите методы обучения ASNN (ассоциативные нейронные сети), FSMLR (пошаговая линейная регрессия) и PLS (регрессия частичных наименьших квадратов).
- Выберите дескрипторные блоки “OEstate and ALogPS” и “ISIDA Fragments (Length 2 - 4)” (электротопологические состояния, прогнозируемые значения липофильности и растворимости, фрагментные дескрипторы размером от 2 до 4 атомов).
- По умолчанию выбран алгоритм отбора дескрипторов Unsupervised forward selection и 5-кратный перекрестный контроль.

Select the training and validation sets:

Training set (required): 20GS_hCHRM3_98_Ki.sdf [details]
 Add a validation set

The model will predict this property:

Ki_hCHRM3 using unit: -log(M)

Select the methods you want to use for the modeling:

Method	Descriptors	Descriptor selection	Model validation
<input type="checkbox"/> [all] <input type="checkbox"/> [none] <input type="checkbox"/> ANN <input checked="" type="checkbox"/> ASNN (bias correction) <input type="checkbox"/> KNN <input type="checkbox"/> LibSVM <input checked="" type="checkbox"/> FSMLR <input type="checkbox"/> MLRA <input checked="" type="checkbox"/> PLS +add a custom template	<input type="checkbox"/> [all] <input type="checkbox"/> [none] <input type="checkbox"/> CDK <input type="checkbox"/> Dragon v.6 (all blocks) <input checked="" type="checkbox"/> OEstate and ALogPS <input checked="" type="checkbox"/> ISIDA Fragments (Length 2 - 4) <input type="checkbox"/> GSFrag <input type="checkbox"/> Mera and Mersy <input type="checkbox"/> Chemaxon descriptors <input type="checkbox"/> Inductive Descriptors <input type="checkbox"/> Adriana <input type="checkbox"/> Spectrophores <input type="checkbox"/> QNPR (SMILES - length 1 - 3 threshold 5) <input type="checkbox"/> Two simple descriptors (MW+Number of carbons) +add a custom template	<input type="checkbox"/> [all] <input type="checkbox"/> [none] <input checked="" type="checkbox"/> Unsupervised forward selection <input type="checkbox"/> Simple pairwise decorrelation (r < 0.95) +add a custom template	<input type="checkbox"/> [all] <input type="checkbox"/> [none] <input checked="" type="checkbox"/> 5-fold cross-validation <input type="checkbox"/> 5-fold cross-validation (stratified) <input type="checkbox"/> Bagging with 64 models +add a custom template

Show advanced options>>

Considering the selection above, **6 models** will be created.

Create the models

- В текущей конфигурации будет построено 6 вариантов моделей. Выбрав параметры, нажмите кнопку Create the models.

- Система создает задания для построения моделей и помещает их в очередь на выполнение. Процесс расчета может занять некоторое время. Для контроля за его ходом щелкните ссылку [basket summary page](#). Затем для обновления информации нажимайте кнопку Refresh.

Predicted property: Ki_hCHRM3
Training set: 20GS_hCHRM3_98_Ki.sdf

Metrics for

	ASNN	FSMLR	PLS
ALogPS, OEstate	running	ready	running
Fragmentor (Length 2 - 4)	running	ready	running

[Refresh](#) [Fetch statistics for 2 ready task\(s\)](#)

- Когда будут построены все 6 моделей (состояние ready), нажмите ссылку Fetch statistics for 6 ready task(s). Появляется сводная таблица статистических параметров моделей. Для выбора нужного параметра используйте раскрывающийся список Metrics.

Metrics for Metrics for

	ASNN	FSMLR	PLS
ALogPS, OEstate	0.99	1.1	1.1
Fragmentor (Length 2 - 4)	1	1.1	1.1

	ASNN	FSMLR	PLS
ALogPS, OEstate	0.46	0.29	0.31
Fragmentor (Length 2 - 4)	0.41	0.36	0.38

- Выберите оптимальную модель, исходя из требований минимальности средней ошибки прогноза $RMSE$ и максимальности параметра перекрестного контроля Q^2 . В данном случае оптимальной можно считать нейросетевую модель на основе дескрипторов OEstate и ALogPS. Для более детального ее анализа щелкните на значении в соответствующей ячейке таблицы.
- На экран выводится основная информация об условиях построения и параметрах модели, а также график соответствия экспериментальных величин активности и значений, прогнозируемых в ходе перекрестного контроля. Опираясь на числовые параметры и графические данные, оцените качество модели.

Save the model

Please enter your model's name:

Overview
Applicability domain

Model name: Ki_hCHRM3_ASNN_[ALogPS, OEstate] [\[rename\]](#)
Private ID is 21830025

Predicted property: Ki_hCHRM3 modeled in -log(M)
Training method: ASNN

Data Set	#	R2	q2	RMSE	MAE
Training set: 20GS_hCHRM3_98_Ki.sdf	805 records	0.51 ± 0.05	0.46 ± 0.07	0.99 ± 0.06	0.74 ± 0.05

Number of compounds ignored because of errors in original model = 1

[Download model statistics](#) [Create a copy of this model](#) [View configuration XML](#) [Export configuration XML](#)

[ALogPS, OEstate]
Correl. limit: 0.95 Variance threshold: 0.01,
Maximum value: 999999, using UFS
Supersab, 1000 iterations, 3 neurons
ensemble=64 additional param
PARTITION=3, SELECTION=2
5-fold cross-validation
-
89 pre-filtered descriptors
Supersab, 1000 iterations, 3 neurons
ensemble=64 k=15 additional param
PARTITION=3, SELECTION=2

Calculated in 297 seconds
Size: 102 Kb

- Для загрузки детальной информации о модели щелкните ссылку Download model statistics. Не изменяя установки по умолчанию, нажмите кнопку Get Excel file. Сохраните файл на рабочий стол и откройте его. Создайте дополнительный лист таблицы и скопируйте на него информацию из ОСНЕМ с вкладки Model profile (вкладку загрузки можно закрыть). Сохраните и закройте файл.

	A	B	C	D	E	F
1	Model name: Ki_hCHRM3_ASNN [ALogPS, OEstate] [rename]					
2	Private ID is 21830025					
3						
4	Predicted property: Ki_hCHRM3 modeled in -log(M)					
5	Training method: ASNN					
6						
7	Data Set					
8		#	R ²	q ²	RMSE	MAE
9	Training set: 20GS_hCHRM3_98_Ki.sdf	805 records	0.51 ± 0.05	0.46 ± 0.07	0.99 ± 0.06	0.74 ± 0.05
10						

- Нажмите кнопку Save внизу страницы, чтобы сохранить модель для последующего прогнозирования активности.

5. Прогнозирование свойств

Для прогнозирования активности и других свойств соединений можно использовать как самостоятельно построенные и сохраненные модели, так и другие модели, опубликованные в системе. Мы выполним прогноз целевого параметра Ki, а также липофильности и растворимости для загруженной выборки соединений. В дальнейшем такое прогнозирование можно проводить и для новых структур.

- В меню Models выберите Open predictor. В верхней части страницы щелкните Browse the full list of public models. Открывается список доступных моделей. Просматривая список, выберите с помощью флажков только что построенную и сохраненную нами модель, а также модель ALogPS. Затем нажмите кнопку Next внизу страницы.

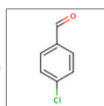
<input checked="" type="checkbox"/>	Ki_hCHRM3_ASNN [ALogPS, OEstate]	predicts Ki_hCHRM3 using 20GS_hCHRM3_98_Ki.sdf (806)	ASNN	2013-12-09
<input checked="" type="checkbox"/>	ALogPS 3.0 ★, published by itetko	predicts logPow, Aqueous Solubility using ALogPS 3.01 (26014)	ASNN	2013-10-23

- Теперь необходимо указать соединения для прогноза. Вы можете загрузить структуры из файла, ввести идентификатор для конкретного соединения, нарисовать структуру в графическом редакторе или выбрать загруженные ранее соединения. Щелкните многоточие в поле Choose a previously prepared set и выберите загруженную ранее базу данных. Затем нажмите кнопку Next.

Provide the compound(s) to predict

Please provide compounds for which you want to predict the target property
Several options are available:

- ☐ Upload compounds from a file
(SDF/MOL2/SMILES/Excel sheet) No file chosen
- ☐ Provide a Name/CAS-RN/SMILES
- ☐ Draw Molecule
(click on depiction to the right to draw)
- ☒ Choose a previously prepared set: 20GS_hCHRM3_98_Ki.sdf
- ☐ Select molecules by a tag: [...]

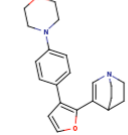
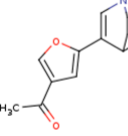


Additional options

Prediction scenario:

☐ Disable prediction cache

- Система создает и запускает задания на прогноз. Расчет может занять некоторое время. После его завершения результаты выводятся на экран. Для каждого соединения указываются экспериментальные величины параметров (если имеются), а также прогнозируемые значения для каждой из выбранных моделей (с оценкой ошибки). Дополнительно помечаются соединения, которые не попадают в область применимости той или иной модели (OUT OF AD). Результаты прогнозирования также можно экспортировать в файл.

1 - 15 of 806 15 items on page 1 of 54 > >>
<div style="display: flex; align-items: center;">  <div style="margin-left: 10px;"> <p> $K_i_hCHRM3 (K_i_hCHRM3_ASNN_ [AlogPS, OEstate]) = 7.16 \cdot \log(M) \pm 2.78$ (ASNN-STDEV = 0.37, estimated RMSE = 1.42) $K_i_hCHRM3(measured) = 2760.0 \text{ nM} = 5.56 \cdot \log(M)$ $\log Pow (AlogPS 3.0) = 3.32 \text{ Log unit} \pm 1.50$ (ASNN-STDEV = 0.68, estimated RMSE = 0.76) $Aqueous Solubility (AlogPS 3.0) = 4.69 \cdot \log(mol/L) \pm 1.41$ (ASNN-STDEV = 1.00, estimated RMSE = 0.72) </p> <p style="color: red; font-weight: bold;">OUT OF AD</p> </div> </div>
<div style="display: flex; align-items: center;">  <div style="margin-left: 10px;"> <p> $K_i_hCHRM3 (K_i_hCHRM3_ASNN_ [AlogPS, OEstate]) = 5.6 \cdot \log(M) \pm 2.78$ (ASNN-STDEV = 0.41, estimated RMSE = 1.42) $K_i_hCHRM3(measured) = 1300.0 \text{ nM} = 5.89 \cdot \log(M)$ $\log Pow (AlogPS 3.0) = 1.48 \text{ Log unit} \pm 0.77$ (ASNN-STDEV = 0.45, estimated RMSE = 0.39) $Aqueous Solubility (AlogPS 3.0) = 2.74 \cdot \log(mol/L) \pm 1.41$ (ASNN-STDEV = 0.74, estimated RMSE = 0.72) </p> </div> </div>

- После окончания работы выйдите из системы, щелкнув ссылку Logout.

Для более детального знакомства с возможностями системы ОСНЕМ рекомендуется изучить имеющуюся на сайте документацию, а также самостоятельно экспериментировать с различными доступными функциями.