AD-MDI

Applicability domain–Model Disturbance Index Tool

NANOBRIDGES
-A Collaborative Project

UNIWERSYTET GDAŃSKI

Marie Curie Actions
EUROPEAN UNION

THE AUTHORS ARE GRATEFUL FOR THE FINANCIAL SUPPORT FROM THE EUROPEAN COMMISSION THROUGH THE MARIE CURIE **IRSES** PROGRAM, NANOBRIDGES PROJECT (FP7-PEOPLE-2011-IRSES, GRANT AGREEMENT NUMBER 295128).

Applicability domain- Model Disturbance Index (AD-MDI)

AD is simply defined as "the response and chemical structure space in which the model makes predictions with a given reliability".

Theoretical Background

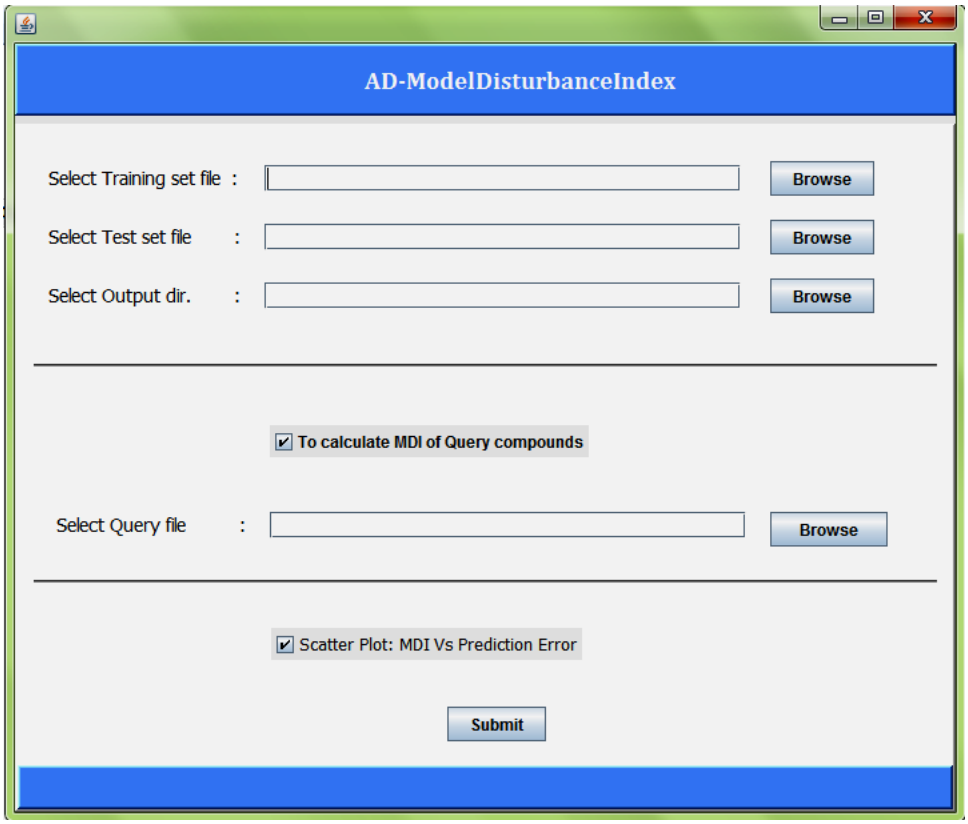
The Applicability domain- Model Disturbance Index **(AD-MDI)** method differs from other methods, which usually uses X information (*i.e.* independent variables) in the descriptor space or Y information (*i.e*. dependent variables) in the properties space individually. This method proposes a novel approach using X and Y information simultaneously. It assumes that the inclusion of a compound, which is similar to the training set compounds, will generate a lower disturbance (*i.e.* a measure to assess the difference of model predictive ability between before-and-after the compound introduction) to the model; otherwise, a higher disturbance may result. Indeed, this assumption can be easily understood from the influence on QSAR model by outliers, where a model is substantially improved if the outlier is excluded, and because the outlier has a significant influence on the data set. Likewise, the influence on the model from an external compound follows the same rules.

Another important aspect is how to evaluate the performance of AD. The universally accepted rule is that the prediction error (PE) of the compound inside the AD should be smaller than compound that are outside the AD. So, the ideal indicator to define AD should depend linearly on prediction error. However, because of the limitation of the coverage of training set, the ambiguity of the relationship between molecular structure and property, and the huge diverse structure space of query compounds, it is not easy to establish an ideal linear-relationship between AD (*MDI*) and prediction error. Thus, the samples of test set are placed into four quadrants (*i.e. true positive, true negative, false positive and false negative prediction*) divided by PE and AD (*MDI*) threshold. Those test set compounds that appears in true positive quadrant are inside applicability domain and those compounds that are in true negative quadrant are outside the applicability domain (***outliers***). Those test set compounds present in false positive (*the unreliable sample wrongly judged as reliable*) or false negative (*the reliable sample is wrongly judged as unreliable*) quadrants (which should be *ideally empty*) are sometimes inevitable, which may be due to the variables used in model cannot totally represent the structure characteristics for all the compounds. For query compounds, MDI can be

calculated and using the already developed MDI-PE (*using test set*) relationship, PE for query compounds can be predicted. Then one can find the quadrant in which the query compound will be present.

The AD-MDI method is proposed by Yan *et. al.* and the algorithm is well explained in literature **[1]**. AD-MDI Tool is standalone software to perform this method and is available at NanoBRIDGES project (http://nanobridges.eu/) official website.

Snapshot 1: AD-MDI Tool



AD-MDI Program Folder

The program folder consists of three folders "**Data**", "**Lib**" and "**Output**". For convenience, user may keep input file in "**Data**" folder and may save output files in "**Output**" folder, since *by default,* clicking on the browse button will open these folder. "**Lib**" folder consists of library files required for running the program. Hence try not to move or delete or rename these library files.

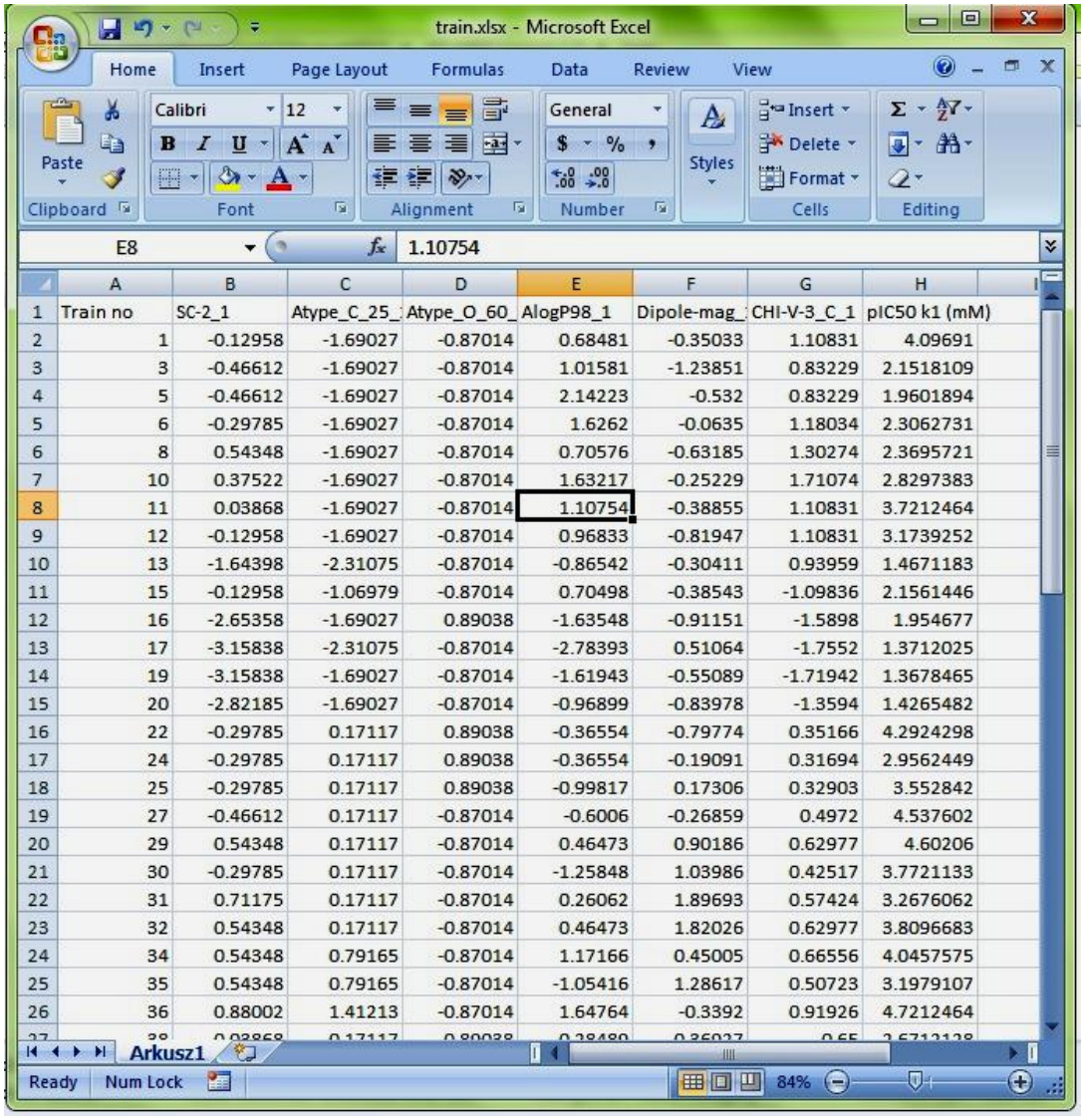
Input file format

Three different file types are allowed *i.e.* ***xlsx, xls*** and ***csv*** as input file. The input file

(*snapshot 2 (A) and (B)*) should consist of compound number (*first column*), descriptor values (*subsequent columns*) and the endpoint values (*last column*) for each object/compound. The format in which this information should be placed in the file is as follows (*see snapshot 2*):

Snapshot 2

A) Training and Test set File Format (.xlsx, .xls or .csv)



	A	B	C	D	E	F	G	H
1	Train no	SC-2_1	Atype_C_25_	Atype_O_60_	AlogP98_1	Dipole-mag_	CHI-V-3_C_1	pIC50 k1 (mM)
2	1	-0.12958	-1.69027	-0.87014	0.68481	-0.35033	1.10831	4.09691
3	3	-0.46612	-1.69027	-0.87014	1.01581	-1.23851	0.83229	2.1518109
4	5	-0.46612	-1.69027	-0.87014	2.14223	-0.532	0.83229	1.9601894
5	6	-0.29785	-1.69027	-0.87014	1.6262	-0.0635	1.18034	2.3062731
6	8	0.54348	-1.69027	-0.87014	0.70576	-0.63185	1.30274	2.3695721
7	10	0.37522	-1.69027	-0.87014	1.63217	-0.25229	1.71074	2.8297383
8	11	0.03868	-1.69027	-0.87014	1.10754	-0.38855	1.10831	3.7212464
9	12	-0.12958	-1.69027	-0.87014	0.96833	-0.81947	1.10831	3.1739252
10	13	-1.64398	-2.31075	-0.87014	-0.86542	-0.30411	0.93959	1.4671183
11	15	-0.12958	-1.06979	-0.87014	0.70498	-0.38543	-1.09836	2.1561446
12	16	-2.65358	-1.69027	0.89038	-1.63548	-0.91151	-1.5898	1.954677
13	17	-3.15838	-2.31075	-0.87014	-2.78393	0.51064	-1.7552	1.3712025
14	19	-3.15838	-1.69027	-0.87014	-1.61943	-0.55089	-1.71942	1.3678465
15	20	-2.82185	-1.69027	-0.87014	-0.96899	-0.83978	-1.3594	1.4265482
16	22	-0.29785	0.17117	0.89038	-0.36554	-0.79774	0.35166	4.2924298
17	24	-0.29785	0.17117	0.89038	-0.36554	-0.19091	0.31694	2.9562449
18	25	-0.29785	0.17117	0.89038	-0.99817	0.17306	0.32903	3.552842
19	27	-0.46612	0.17117	-0.87014	-0.6006	-0.26859	0.4972	4.537602
20	29	0.54348	0.17117	-0.87014	0.46473	0.90186	0.62977	4.60206
21	30	-0.29785	0.17117	-0.87014	-1.25848	1.03986	0.42517	3.7721133
22	31	0.71175	0.17117	-0.87014	0.26062	1.89693	0.57424	3.2676062
23	32	0.54348	0.17117	-0.87014	0.46473	1.82026	0.62977	3.8096683
24	34	0.54348	0.79165	-0.87014	1.17166	0.45005	0.66556	4.0457575
25	35	0.54348	0.79165	-0.87014	-1.05416	1.28617	0.50723	3.1979107
26	36	0.88002	1.41213	-0.87014	1.64764	-0.3392	0.91926	4.7212464

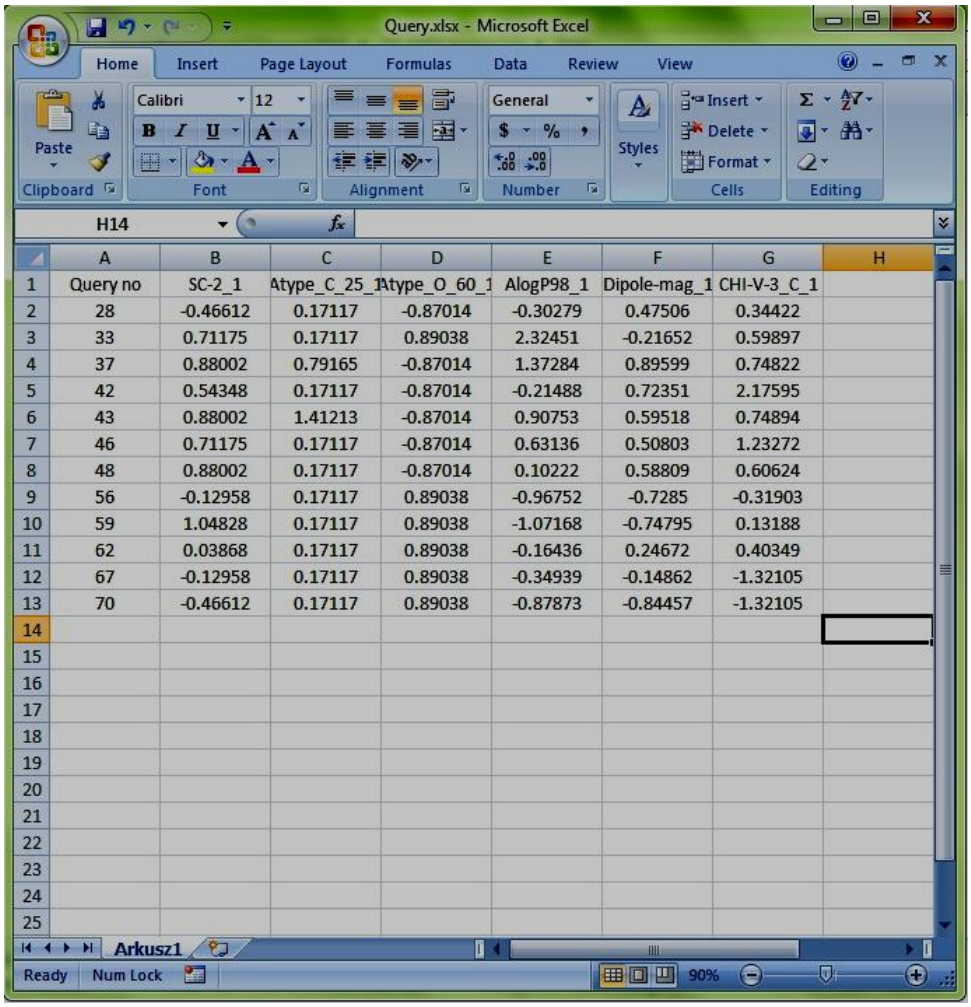
First Row: Header *i.e.* name for each column, for instances, descriptor names, and endpoint name. It can be numerical, alphabet or alphanumerical in nature.

First column: Serial number/Compound number (*only numerical values*)

Subsequent columns: Property/Independent variables/Descriptor values present in the developed QSAR (***MLR***) model (*only numerical values*).

Last column: Endpoint values/Dependent variables (*only numerical values*)

B) Query compound File Format (.xlsx, .xls or .csv)



	A	B	C	D	E	F	G	H
1	Query no	SC-2_1	Atype_C_25_1	Atype_O_60_1	AlogP98_1	Dipole-mag_1	CHI-V-3_C_1	
2	28	-0.46612	0.17117	-0.87014	-0.30279	0.47506	0.34422	
3	33	0.71175	0.17117	0.89038	2.32451	-0.21652	0.59897	
4	37	0.88002	0.79165	-0.87014	1.37284	0.89599	0.74822	
5	42	0.54348	0.17117	-0.87014	-0.21488	0.72351	2.17595	
6	43	0.88002	1.41213	-0.87014	0.90753	0.59518	0.74894	
7	46	0.71175	0.17117	-0.87014	0.63136	0.50803	1.23272	
8	48	0.88002	0.17117	-0.87014	0.10222	0.58809	0.60624	
9	56	-0.12958	0.17117	0.89038	-0.96752	-0.7285	-0.31903	
10	59	1.04828	0.17117	0.89038	-1.07168	-0.74795	0.13188	
11	62	0.03868	0.17117	0.89038	-0.16436	0.24672	0.40349	
12	67	-0.12958	0.17117	0.89038	-0.34939	-0.14862	-1.32105	
13	70	-0.46612	0.17117	0.89038	-0.87873	-0.84457	-1.32105	
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								
24								
25								

First Row: Header *i.e.* name for each column, for instances, descriptor names, and endpoint name. It can be numerical, alphabet or alphanumerical in nature.

First column: Serial number/Compound number (*only numerical values*)

Subsequent columns: Property/Independent variables/Descriptor values present in the developed QSAR (***MLR***) model (*only numerical values*).

Note: No Property/Activity/ dependent variable column in Query file.

How to run the program

It's simple! Just click/double click on the jar file (ApplicabilityDomainMDI.*jar*) present in the AD-MDI folder. A window will open as shown in *Snapshot 1*, with few queries, which a user has to fill before clicking on '***Submit***' button to run the program.

"***Select Training set File***": Click on 'browse' button to select the training set file. By default, it will open the "Data" folder present in AD-MDI folder. So for convenience, user can keep the input files in the "Data" folder.

"***Select Test set File***": Click on 'browse' button to select the test set file file. By default, it will also open the "Data" folder present in AD-MDI folder.

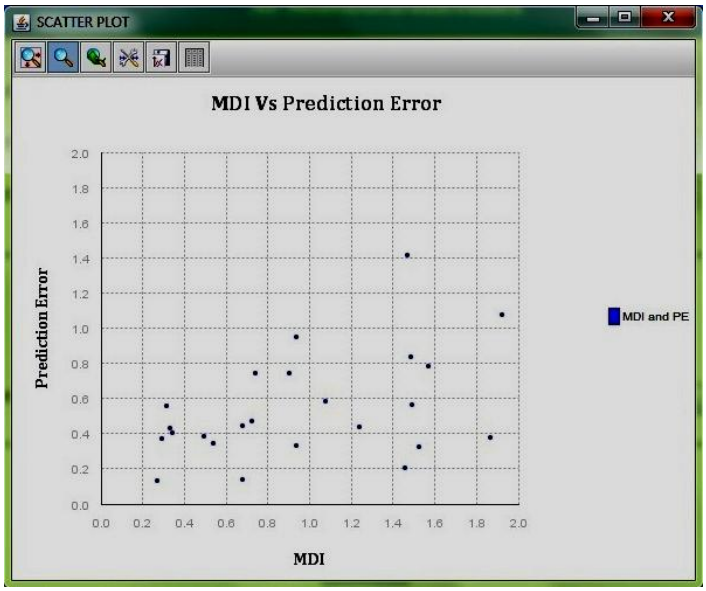
"***Select Output Directory***": Click on 'browse' button to select the destination/output file directory and define output file name (*without any extension*). By default, it will open the "Output" folder present in AD-MDI folder. So for convenience, user can save the output files in the "Output" folder.

Optional checkboxes

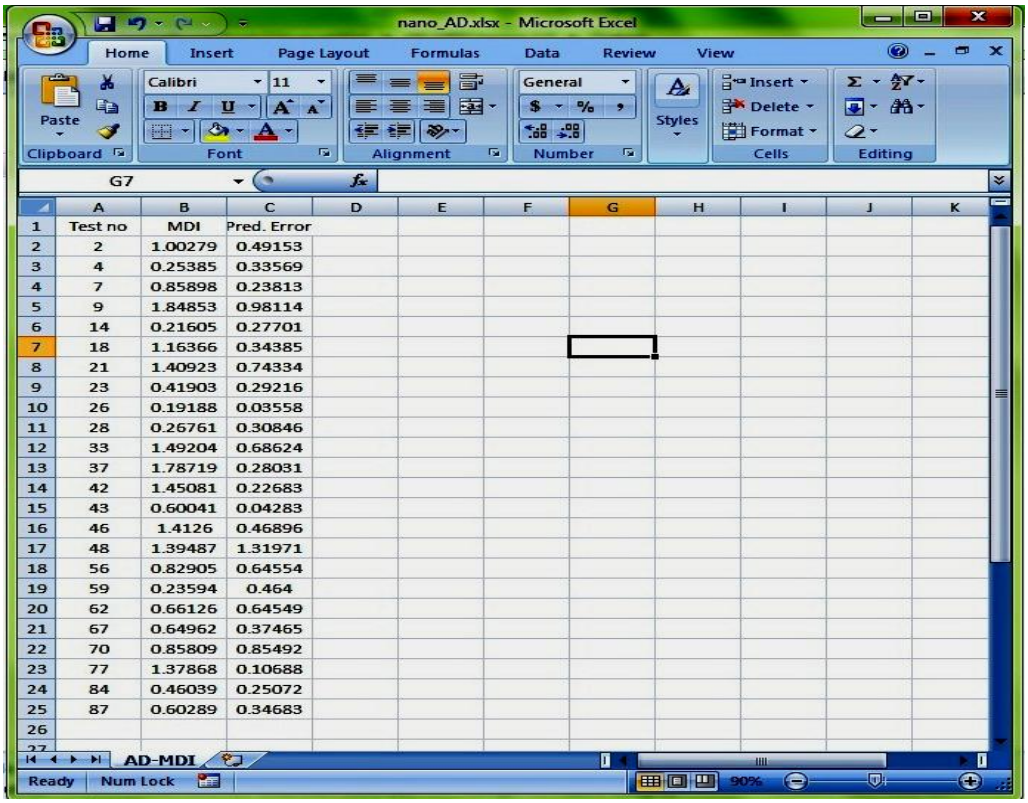
1. If user wants to calculate MDI to determine the AD of query compounds, select the checkbox labeled "*To calculate MDI of query compounds*". Once selected, user can now click on '*Browse*' button to select the query file.
2. If user likes to have scatter plot of MDI vs. PE (*test set*) that defines the AD, select the checkbox labeled "*Scatter plot: MDI vs. Prediction Error*". The same plot can be used to find whether query compounds are inside or outside the defined AD. If selected, a scatter plot (*snapshot 3*) appears in a new window after successful execution of program. The new window has few buttons to perform some functions like to change the scale (*X and Y axes),* save the scatter plot in *.png* format *etc.*

Output

Snapshot 3: *Scatter plot in New Window*

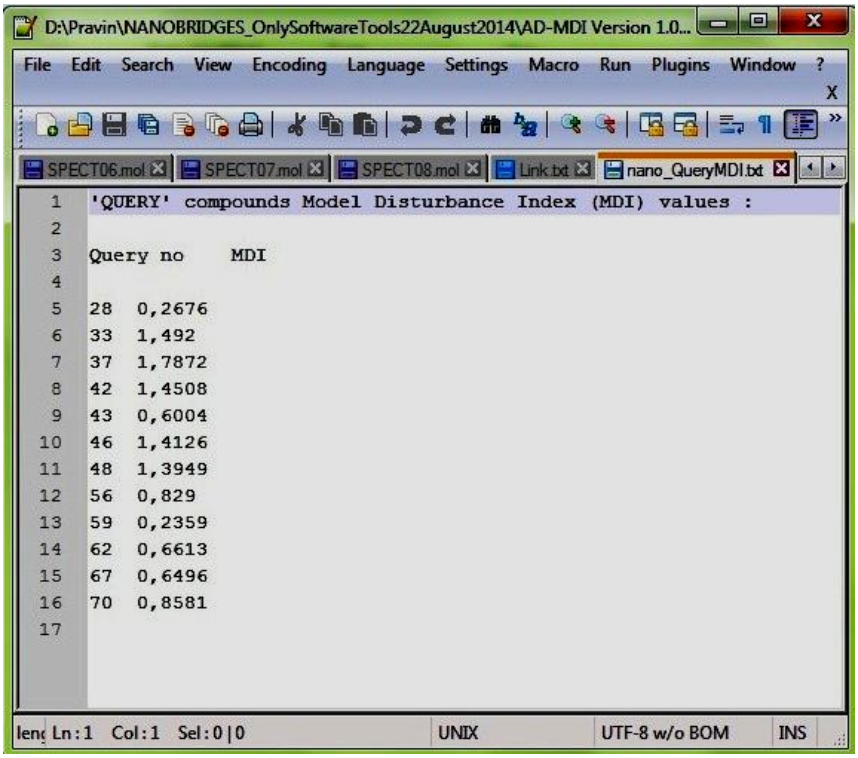


Snapshot 4: *Excel Sheet with Test set MDI and PE values*



	A	B	C	D	E	F	G	H	I	J	K
1	Test no	MDI	Pred. Error								
2	2	1.00279	0.49153								
3	4	0.25385	0.33569								
4	7	0.85898	0.23813								
5	9	1.84853	0.98114								
6	14	0.21605	0.27701								
7	18	1.16366	0.34385								
8	21	1.40923	0.74334								
9	23	0.41903	0.29216								
10	26	0.19188	0.03558								
11	28	0.26761	0.30846								
12	33	1.49204	0.68624								
13	37	1.78719	0.28031								
14	42	1.45081	0.22683								
15	43	0.60041	0.04283								
16	46	1.4126	0.46896								
17	48	1.39487	1.31971								
18	56	0.82905	0.64554								
19	59	0.23594	0.464								
20	62	0.66126	0.64549								
21	67	0.64962	0.37465								
22	70	0.85809	0.85492								
23	77	1.37868	0.10688								
24	84	0.46039	0.25072								
25	87	0.60289	0.34683								
26											

Snapshot 5: *Text file with MDI values of query compounds*



1. **Output file (.xlsx/.xls/.csv)** (*snapshot 4*): This excel sheet will consist of three columns comprising of test set compound number (extracted from input test set file), MDI and PE values. One can also plot *scatter plot* of MDI vs. PE using Microsoft excel functionalities, if user do not like to use the *in-built* scatter plot option.
2. **Text file (.txt file)** (*snapshot 4*): The text file will consist of MDI values for query compounds, if provided. This file is only generated if user has selected the checkbox labeled "*To calculate MDI of query compounds*".

Reference

1. Yan, Jun, et al. "A Combinational Strategy of Model Disturbance and Outlier Comparison to Define Applicability Domain in Quantitative Structural Activity Relationship." *Molecular Informatics*, 2014, 33 (8), 503-513.

Java External Library Used

JMathPlot library – for interactive 2D and 3D plots

- Available at https://code.google.com/p/jmathplot/

Apache POI – the Java API for Microsoft Documents

- Available at http://poi.apache.org/

XMLBeans

- Available at http://xmlbeans.apache.org/

Disclaimer

For academic purpose only.

The program AD-MDI has been developed in Java language and is platform independent. The software is validated on known data sets. Please report for discrepancy of result for any other dataset. Contact us at any of the following addresses:

Dr. Tomasz Puzyn,
NanaBRIDGES Project Coordinator,
Faculty of Chemistry,
University of Gdansk,
Gdansk,
Poland 80-952
Email Id: puzi@qsar.eu.org

Dr. Kunal Roy,
Drug Theoretics and Cheminformatics Lab.,
Dept. of Pharmaceutical Technology,
Jadavpur University,
Kolkata, West Bengal,
INDIA-700032
Email Id: kunalroy_in@yahoo.com

Software Developer details:

Pravin Ambure,
Research Scholar,
Drug Theoretics and Cheminformatics Lab.,
Dept. of Pharmaceutical Technology,
Jadavpur University,
Kolkata, West Bengal,
INDIA-700032
E-mail Id: ambure.pharmait@gmail.com (*for any queries regarding the tool)