Pharmaceutical Chemistry

QSAR through the OCHEM portal

SAPIENZA
UNIVERSITÀ DI ROMA

Introduction

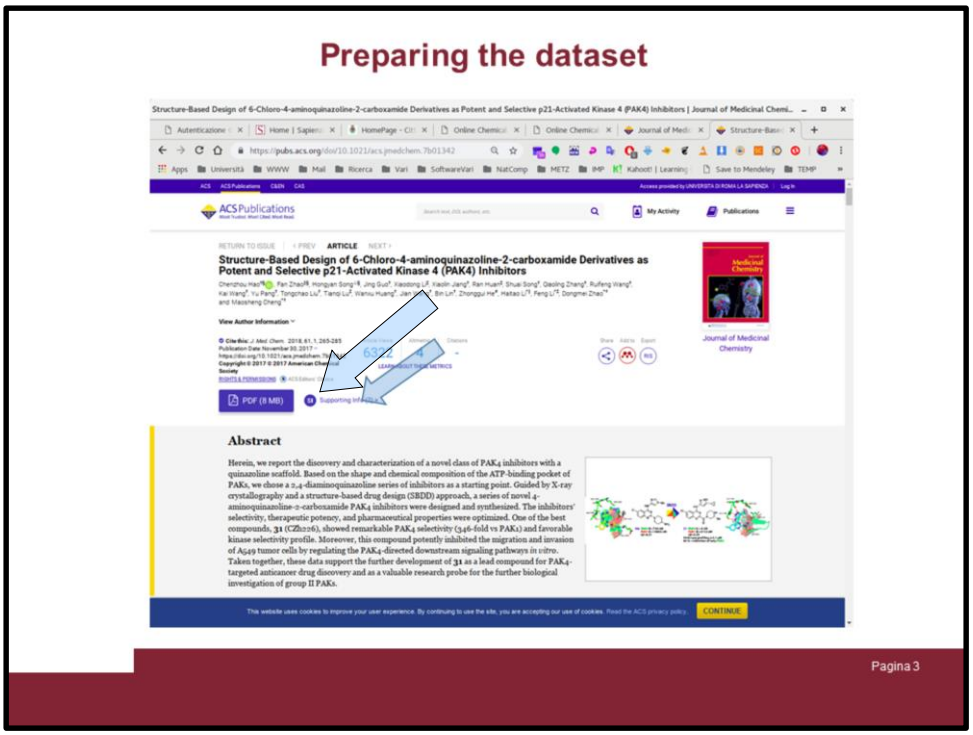
J Comput Aided Mol Des (2011) 25:533–554
DOI 10.1007/s10822-011-9440-2

Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information

The Online Chemical Modeling Environment is a web-based platform that aims to automate and simplify the typical steps required for QSAR modeling. The platform consists of two major subsystems: the database of experimental measurements and the modeling framework. A user-contributed database contains a set of tools for easy input, search and modification of thousands of records. The OCHEM database is based on the wiki principle and focuses primarily on the quality and verifiability of the data. The database is tightly integrated with the modeling framework, which supports all the steps required to create a predictive model: data search, calculation and selection of a vast variety of molecular descriptors, application of machine learning methods, validation, analysis of the model and assessment of the applicability domain. As compared to other similar systems, OCHEM is not intended to re-implement the existing tools or models but rather to invite the original authors to contribute their results, make them publicly available, share them with other users and to become members of the growing research community. Our intention is to make OCHEM a widely used platform to perform the QSPR/QSAR studies online and share it with other users on the Web. The ultimate goal of OCHEM is collecting all possible chemoinformatics tools within one simple, reliable and user-friendly resource. The OCHEM is free for web users and it is available online at http://www.ochem.eu

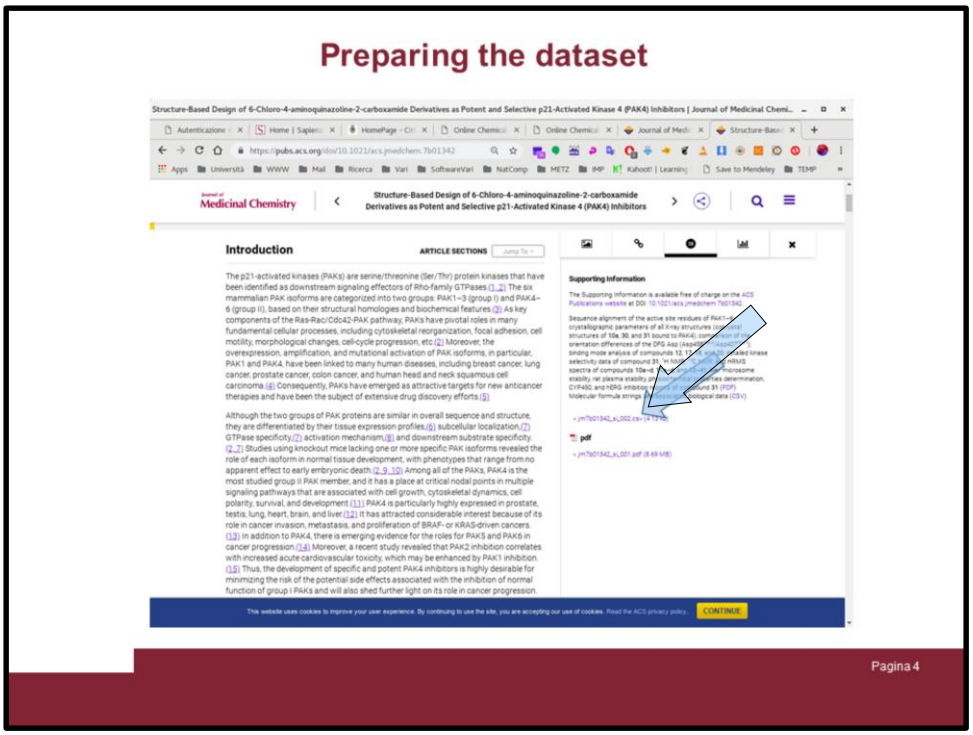
.

In this tutorial you will be guided to use OCHE.eu platform to build QSAR models focusing in the development of PLS-based models



First you need a dataset of molecules, as example the molecules available from the J Med Chem publication on PAK4 inhibitor will be retrieved from the article web page.

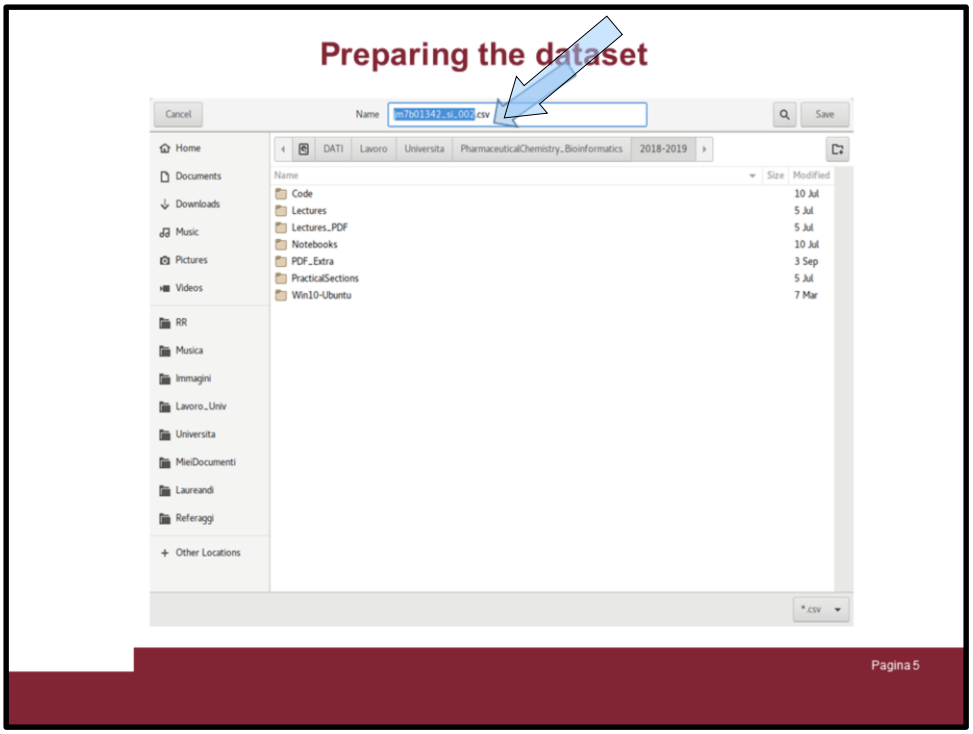
To do this go to pubs.acs.org/doi/abs/10.1021/acs.jmedchem.7b01342 and click on the supporting info “SI” icon ....



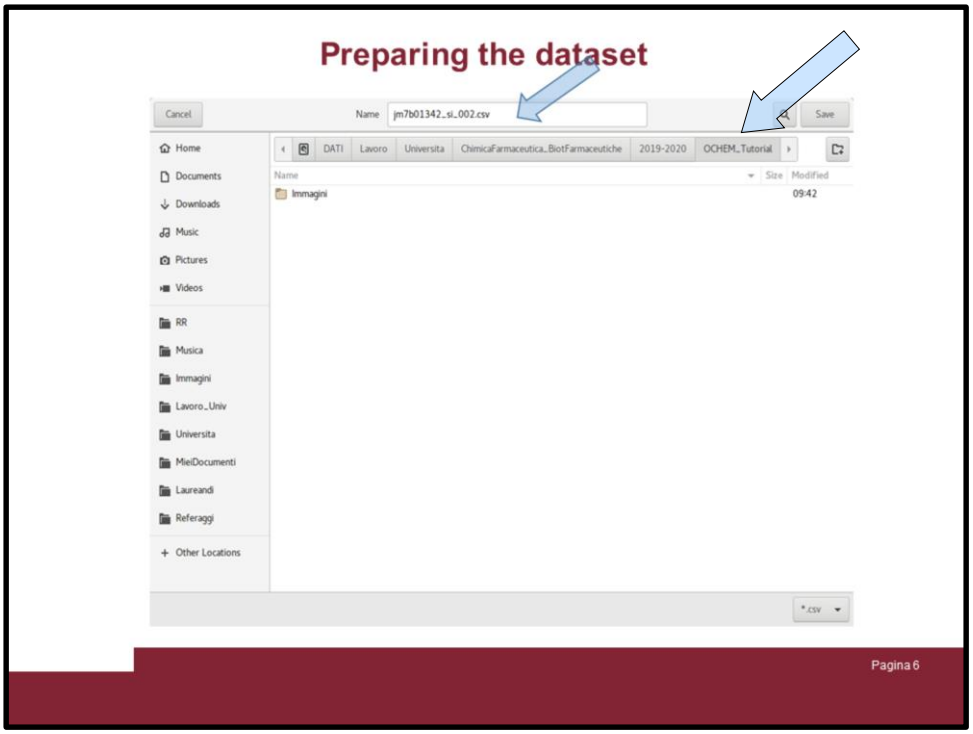
…. the browser will go to the page (in different way dependin on the browser you use) and it will present the link to download a comma separated variable (csv) file:

jm7b01342_si_002.csv

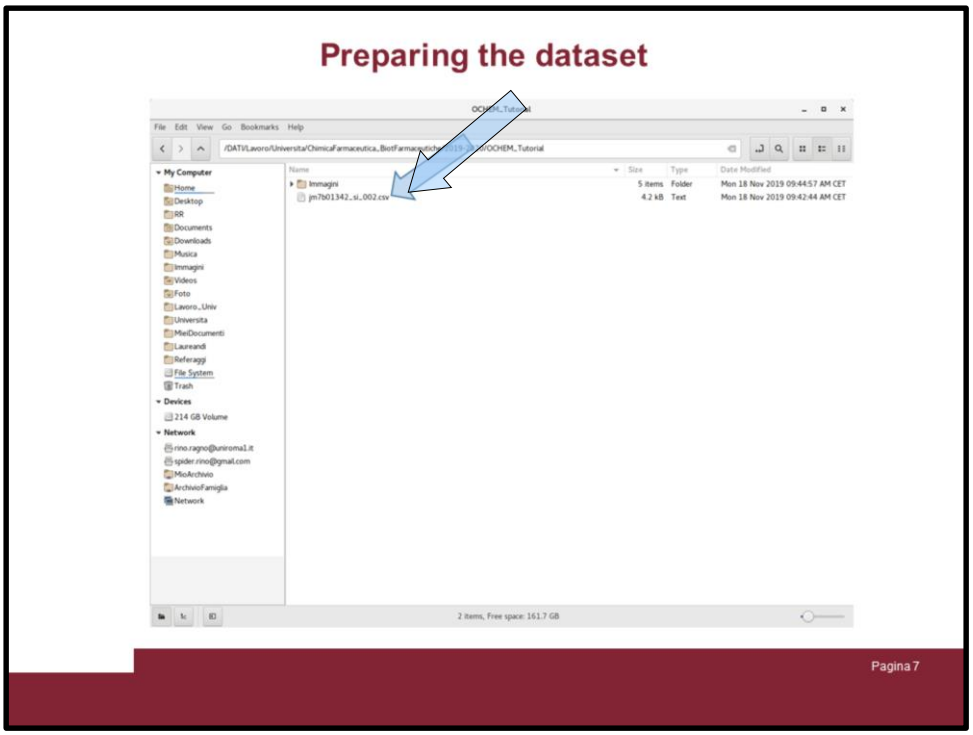
Click on it and the file will be downloaded on your PC (Take a note were it is going to be saved)



If your browser was set to ask where to save the under downloading files you will be asked for a folder to save the csv file

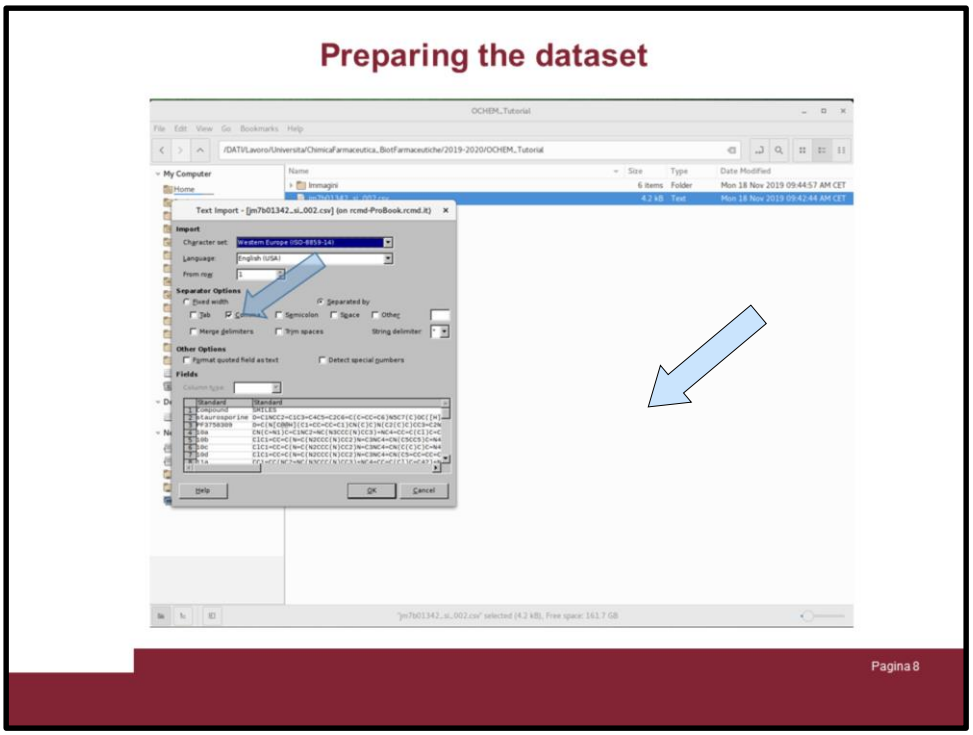


Here a new folder callse “OCHEM_Tutorial” was prepared to save the csv file



ALWAYS CHECK THE FILE IS THERE!

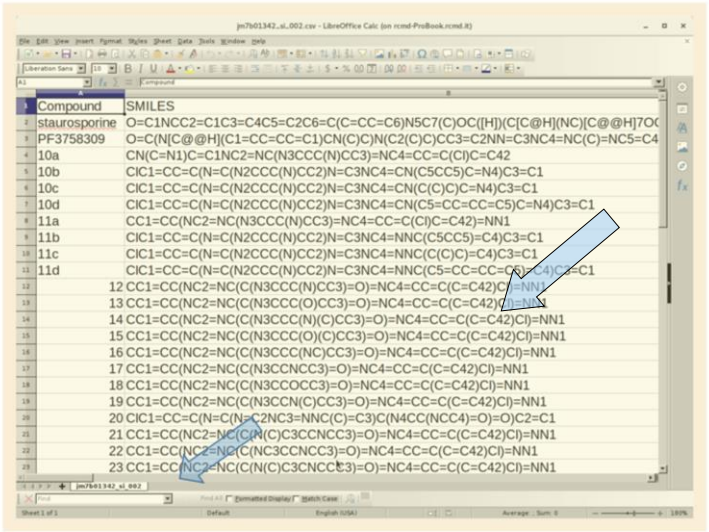
Now right click on the csv file and open it with libreoffice or import in excel. Here libreoffice was used



A dialog window will pop up asking how the file should be interpreted.

Check comma as separator and uncheck all the other options. You have an example on the windows how the data will be read. Pay attention on the way the number will be interpreted (Italian or International)

Preparing the dataset

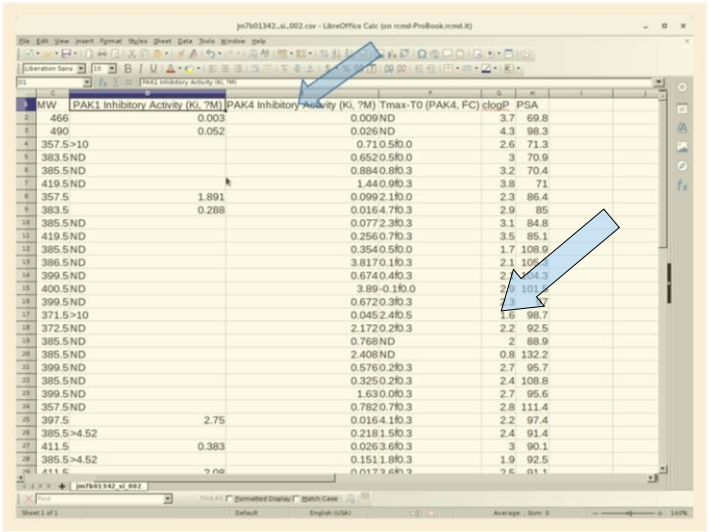


	Compound	SMILES
2	staurosporine	O=C1NCC2=C1C3=C4C5=C2C6=C(C=CC=C6)N5C7(C)OC([H])(C[C@H](NC)[C@@H]7OC
3	PF3758309	O=C(N[C@@H](C1=CC=CC=C1)CN(C)C)N(C2(C)C)CC3=C2NN=C3NC4=NC(C)=NC5=C4
4	10a	CN(C=N1)C=C1NC2=NC(N3CCC(N)CC3)=NC4=CC=C(Cl)C=C42
5	10b	ClC1=CC=C(N=C(N2CCC(N)CC2)N=C3NC4=CN(C5CC5)C=N4)C3=C1
6	10c	ClC1=CC=C(N=C(N2CCC(N)CC2)N=C3NC4=CN(C(C)C)C=N4)C3=C1
7	10d	ClC1=CC=C(N=C(N2CCC(N)CC2)N=C3NC4=CN(C5=CC=CC=C5)C=N4)C3=C1
8	11a	CC1=CC(NC2=NC(N3CCC(N)CC3)=NC4=CC=C(Cl)C=C42)=NN1
9	11b	ClC1=CC=C(N=C(N2CCC(N)CC2)N=C3NC4=NNC(C5CC5)=C4)C3=C1
10	11c	ClC1=CC=C(N=C(N2CCC(N)CC2)N=C3NC4=NNC(C(C)C)=C4)C3=C1
11	11d	ClC1=CC=C(N=C(N2CCC(N)CC2)N=C3NC4=NNC(C5=CC=CC=C5)=C4)C3=C1
12		CC1=CC(NC2=NC(C(N3CCC(N)CC3)=O)=NC4=CC=C(C=C42)Cl)=NN1
13		CC1=CC(NC2=NC(C(N3CCC(O)CC3)=O)=NC4=CC=C(C=C42)Cl)=NN1
14		CC1=CC(NC2=NC(C(N3CCC(N)(C)CC3)=O)=NC4=CC=C(C=C42)Cl)=NN1
15		CC1=CC(NC2=NC(C(N3CCC(O)(C)CC3)=O)=NC4=CC=C(C=C42)Cl)=NN1
16		CC1=CC(NC2=NC(C(N3CCC(NC)CC3)=O)=NC4=CC=C(C=C42)Cl)=NN1
17		CC1=CC(NC2=NC(C(N3CCNCC3)=O)=NC4=CC=C(C=C42)Cl)=NN1
18		CC1=CC(NC2=NC(C(N3CCOCC3)=O)=NC4=CC=C(C=C42)Cl)=NN1
19		CC1=CC(NC2=NC(C(N3CCN(C)CC3)=O)=NC4=CC=C(C=C42)Cl)=NN1
20		ClC1=CC=C(N=C(N=C2NC3=NNC(C)=C3)C(N4CC(NCC4)=O)=O)C2=C1
21		CC1=CC(NC2=NC(C(N(C)C3CCNCC3)=O)=NC4=CC=C(C=C42)Cl)=NN1
22		CC1=CC(NC2=NC(C(NC3CCNCC3)=O)=NC4=CC=C(C=C42)Cl)=NN1
23		CC1=CC(NC2=NC(C(N(C)C3CNCCC3)=O)=NC4=CC=C(C=C42)Cl)=NN1

Once the file is imported in libreoffice you can check it is correctly read.

Slide on the right …

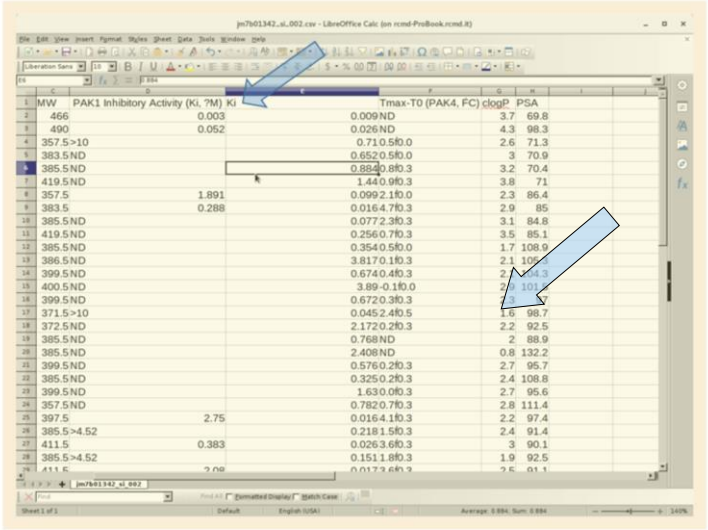
Preparing the dataset



MW	PAK1 Inhibitory Activity (Ki, ?M)	PAK4 Inhibitory Activity (Ki, ?M)	Tmax-T0 (PAK4, FC)	clogP	PSA
466	0.003	0.009	ND	3.7	69.8
490	0.052	0.026	ND	4.3	98.3
357.5	>10	0.71	0.5	f0.0	2.6	71.3
383.5	ND	0.652	0.5	f0.0	3	70.9
385.5	ND	0.884	0.8	f0.3	3.2	70.4
419.5	ND	1.44	0.9	f0.3	3.8	71
357.5	1.891	0.099	2.1	f0.0	2.3	86.4
383.5	0.288	0.016	4.7	f0.3	2.9	85
385.5	ND	0.077	2.3	f0.3	3.1	84.8
419.5	ND	0.256	0.7	f0.3	3.5	85.1
385.5	ND	0.354	0.5	f0.0	1.7	108.9
386.5	ND	3.817	0.1	f0.3	2.1	108.9
399.5	ND	0.674	0.4	f0.3	2.7	104.2
400.5	ND	3.89	-0.1	f0.0	2.6	104.2
399.5	ND	0.672	0.3	f0.3	2.7	104.2
371.5	>10	0.045	2.4	f0.5	1.6	98.7
372.5	ND	2.172	0.2	f0.3	2.2	92.5
385.5	ND	0.768	ND	2	88.9
385.5	ND	2.408	ND	0.8	132.2
399.5	ND	0.576	0.2	f0.3	2.7	95.7
385.5	ND	0.325	0.2	f0.3	2.4	108.8
399.5	ND	1.63	0.0	f0.3	2.7	95.6
357.5	ND	0.782	0.7	f0.3	2.8	111.4
397.5	2.75	0.016	4.1	f0.3	2.2	97.4
385.5	>4.52	0.218	1.5	f0.3	2.4	91.4
411.5	0.383	0.026	3.6	f0.3	3	90.1
385.5	>4.52	0.151	1.8	f0.3	1.9	92.5
411.5	2.08	0.017	3.6	f0.3	2.5	91.1

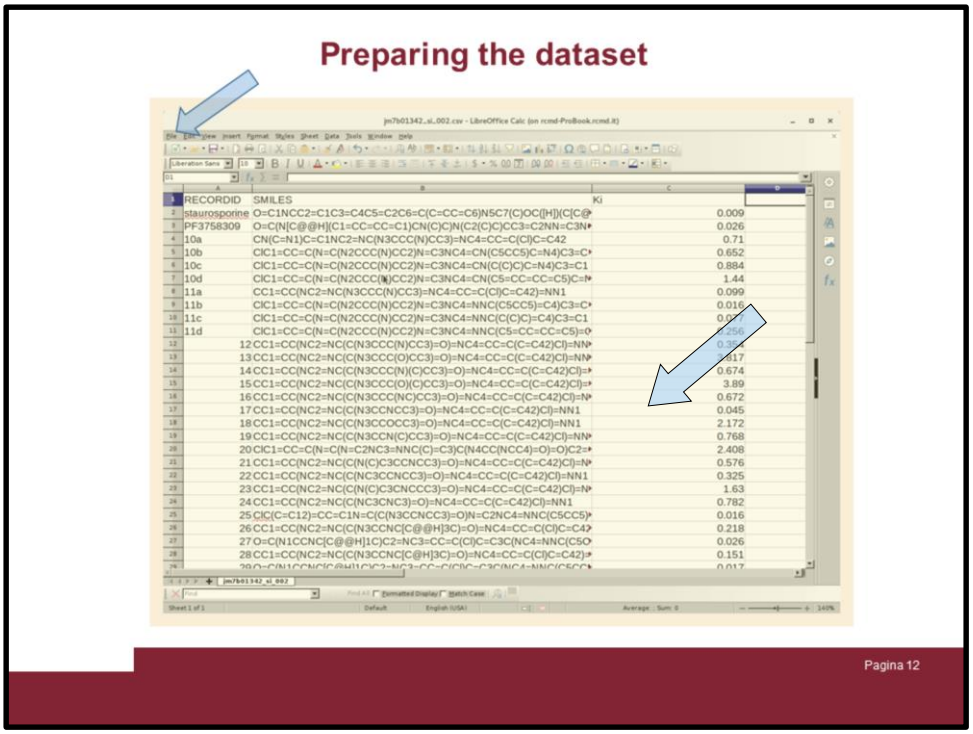
… and focus on the PAK4 Ki colum

Preparing the dataset



MW	PAK1 Inhibitory Activity (Ki, ?M)	Ki	Tmax-T0 (PAK4, FC)	clogP	PSA
466		0.003	0.009	ND	3.7 69.8
490		0.052	0.026	ND	4.3 98.3
357.5	>10		0.71	0.5	f0.0 2.6 71.3
383.5	ND		0.652	0.5	f0.0 3 70.9
385.5	ND		0.884	0.8	f0.3 3.2 70.4
419.5	ND		1.44	0.9	f0.3 3.8 71
357.5		1.891	0.099	2.1	f0.0 2.3 86.4
383.5		0.288	0.016	4.7	f0.3 2.9 85
385.5	ND		0.077	2.3	f0.3 3.1 84.8
419.5	ND		0.256	0.7	f0.3 3.5 85.1
385.5	ND		0.354	0.5	f0.0 1.7 108.9
386.5	ND		3.817	0.1	f0.3 2.1 108.4
399.5	ND		0.674	0.4	f0.3 2.1 104.3
400.5	ND		3.89	-0.1	f0.0 2.6 101.6
399.5	ND		0.672	0.3	f0.3
371.5	>10		0.045	2.4	f0.5 1.6 98.7
372.5	ND		2.172	0.2	f0.3 2.2 92.5
385.5	ND		0.768	ND	2 88.9
385.5	ND		2.408	ND	0.8 132.2
399.5	ND		0.576	0.2	f0.3 2.7 95.7
385.5	ND		0.325	0.2	f0.3 2.4 108.8
399.5	ND		1.63	0.0	f0.3 2.7 95.6
357.5	ND		0.782	0.7	f0.3 2.8 111.4
397.5		2.75	0.016	4.1	f0.3 2.2 97.4
385.5	>4.52		0.218	1.5	f0.3 2.4 91.4
411.5		0.383	0.026	3.6	f0.3 3 90.1
385.5	>4.52		0.151	1.8	f0.3 1.9 92.5

Rename the column into just Ki and ...

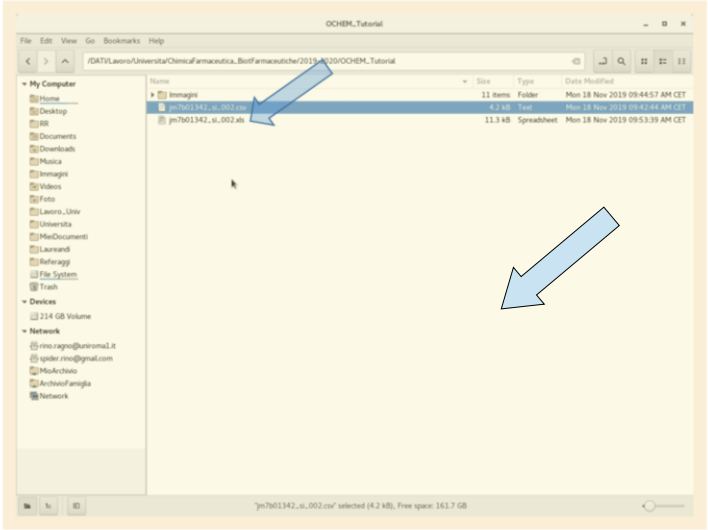


Delete the other columns. Rename the compound column header into RECORDID.

You need just three columns: RECORDID, SMILES and Ki

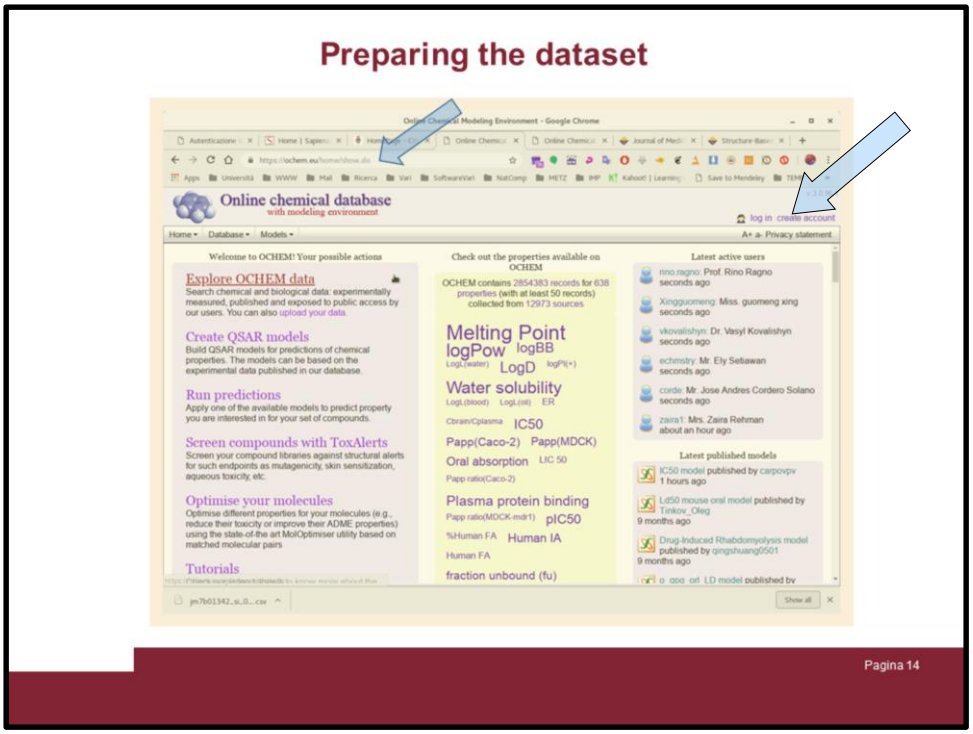
Save the file as excel type, the one with .xls extension is only valid

Preparing the dataset



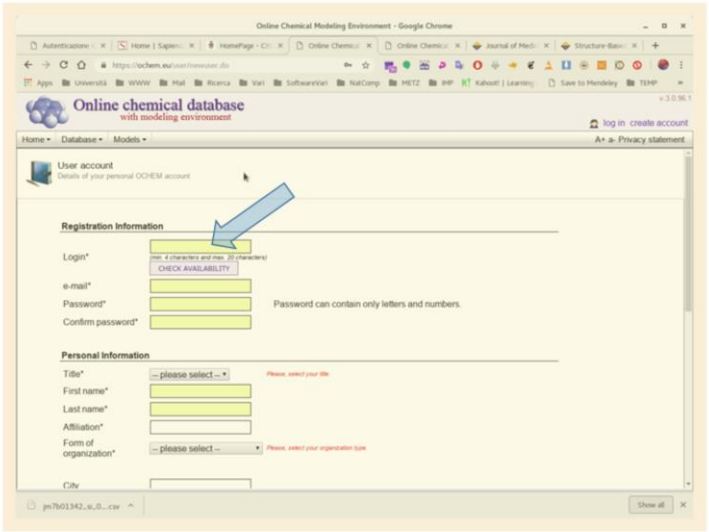
Pagina 13

CHECK THE XLS FILE IS THERE!



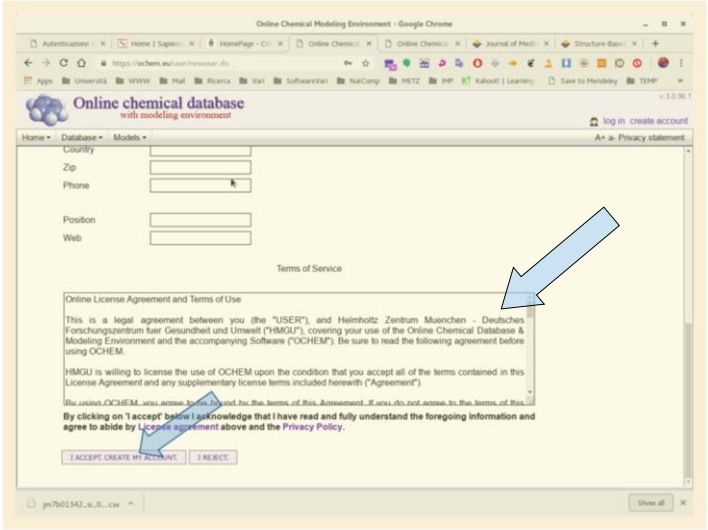
Go to OCHEM.eu and click onto “Create account” if you do not have any

Preparing the dataset



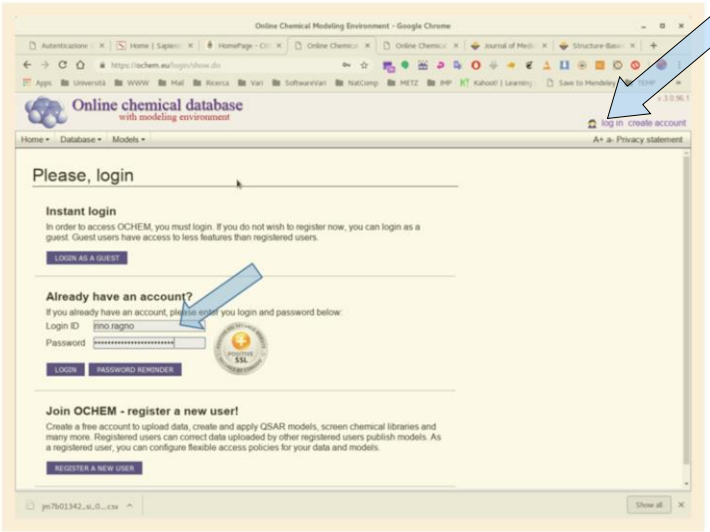
Fill in all the data ...

Preparing the dataset



…. and then click on “I ACCEPT CREATE MY ACCOUNT”

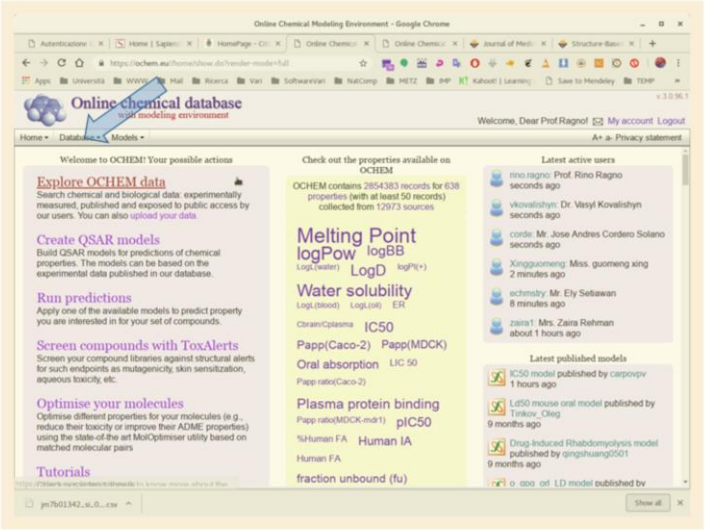
Preparing the dataset



Pagina 17

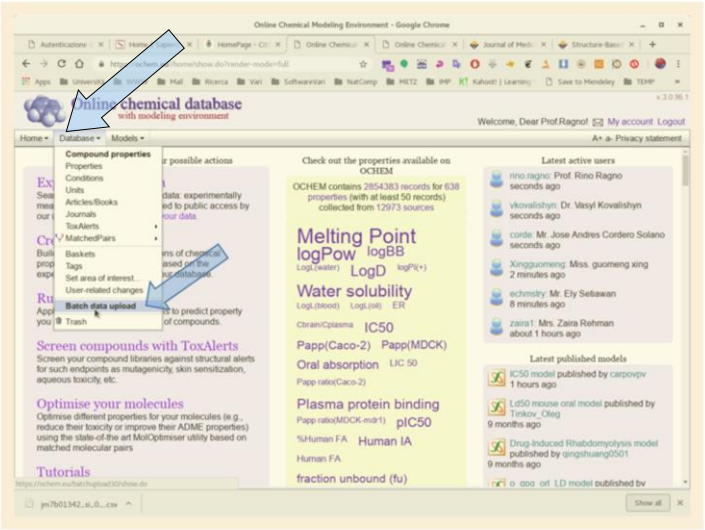
Once the OCHE account is activated log in in the portal using the username and password you indicated during the registration

Preparing the dataset



Now you have logged into OCHEM

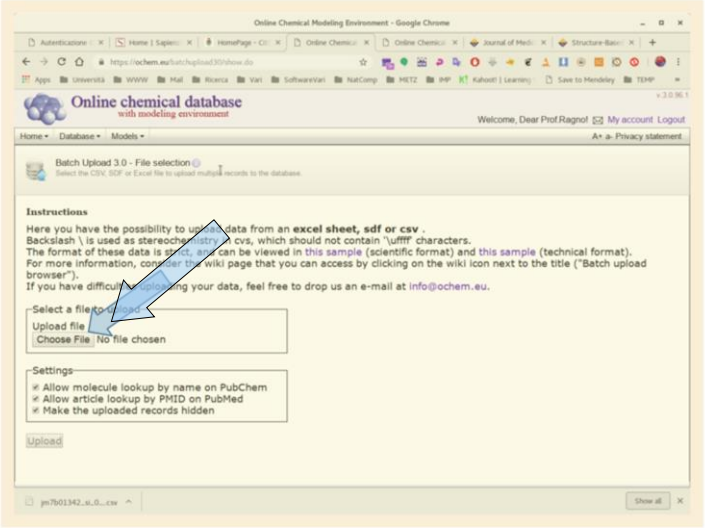
Preparing the dataset



Now let's to import the xls file.

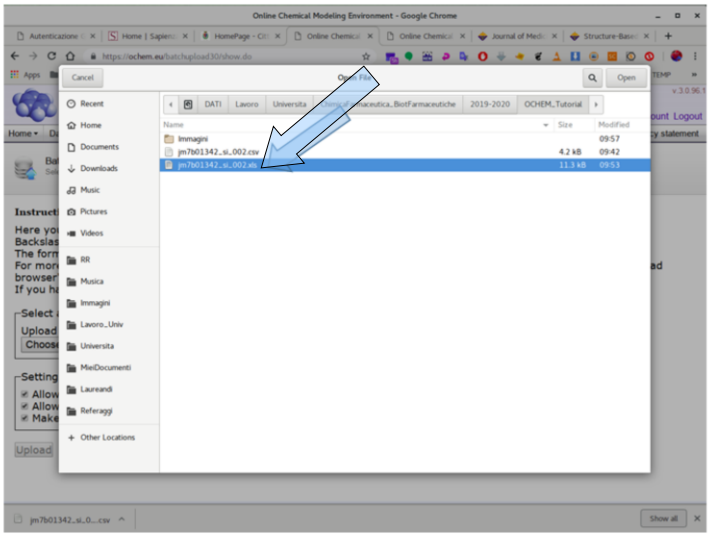
Click on the Database tab and select "Batch data upload"

Preparing the dataset



A window will give you some direction. Click on “ChooseFile”

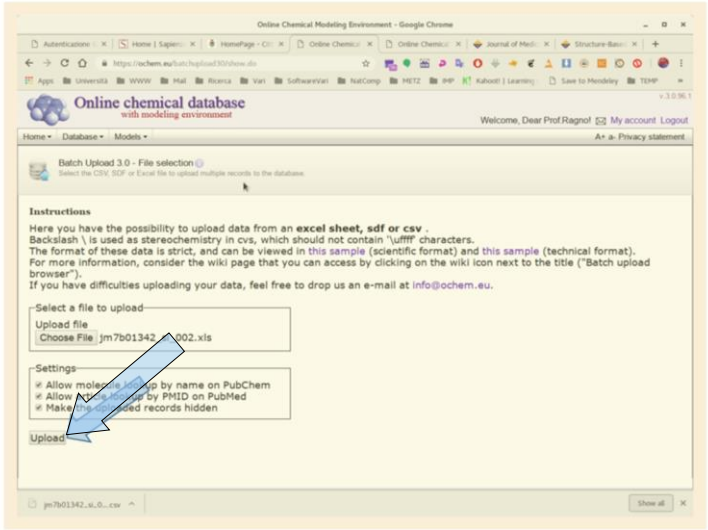
Preparing the dataset



Pagina 21

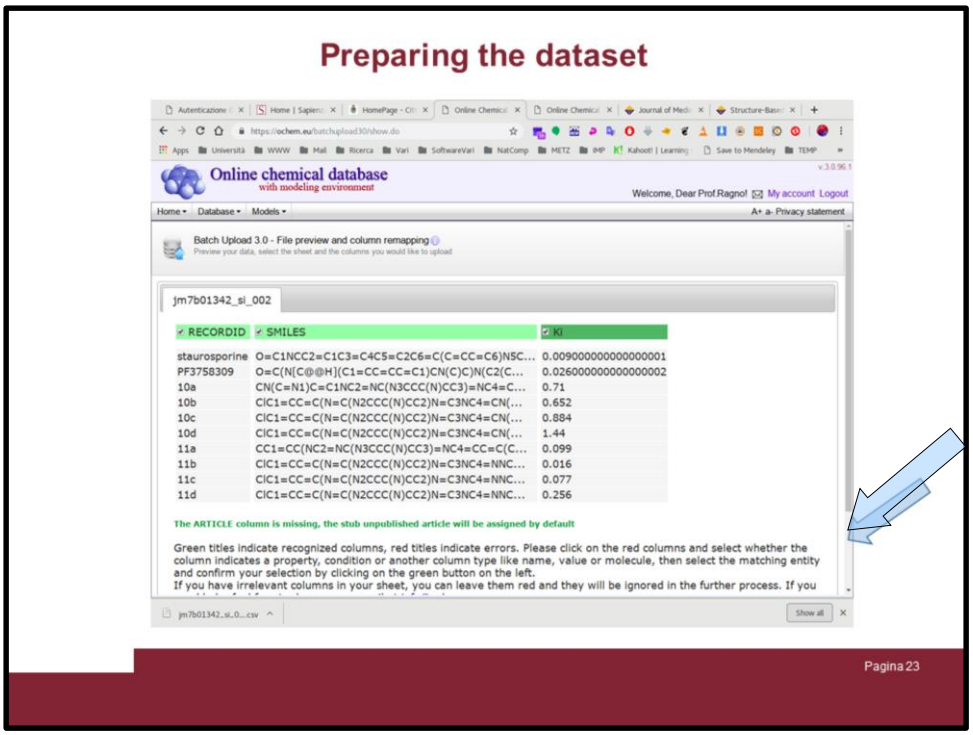
A window will popo up and ask you to select the file you prepared before. Just click on it twice.

Preparing the dataset

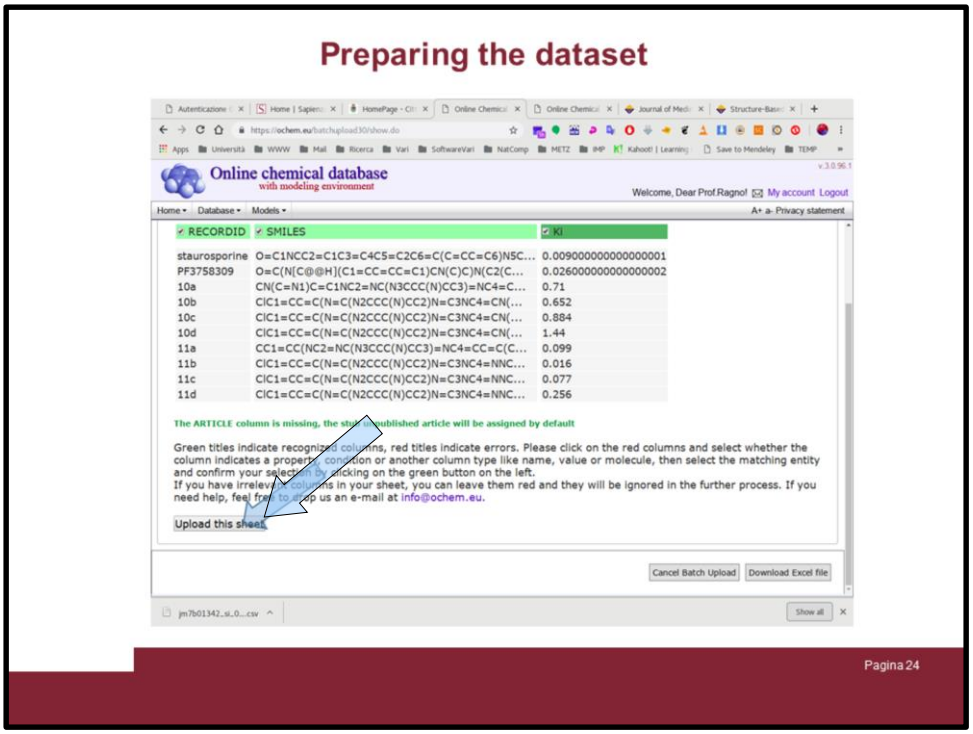


Pagina 22

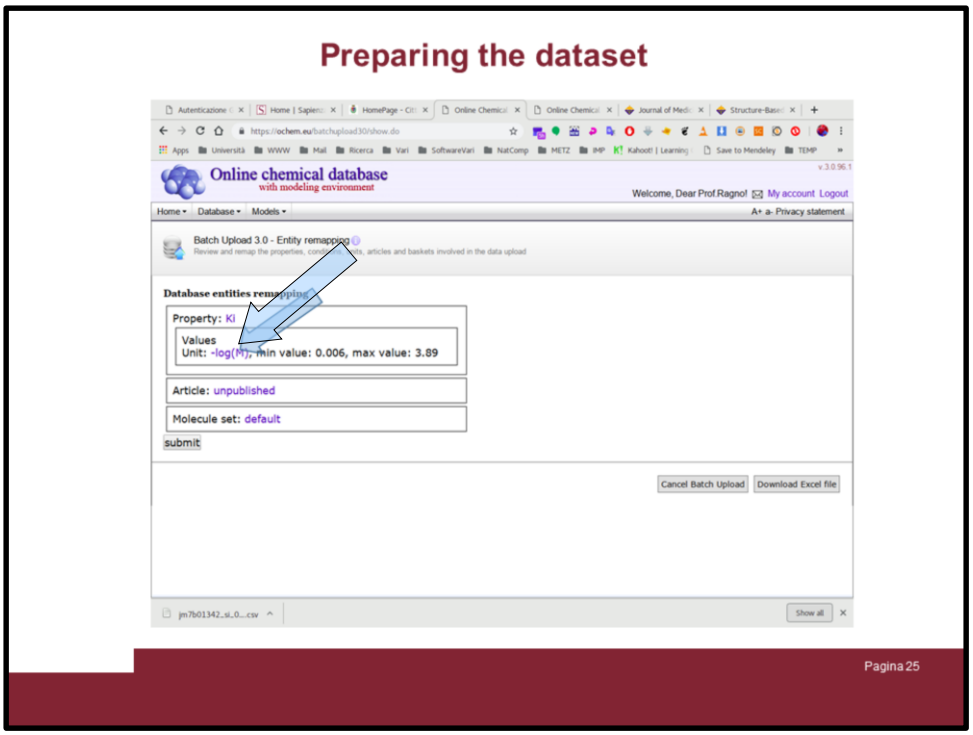
Now click on “Upload” button



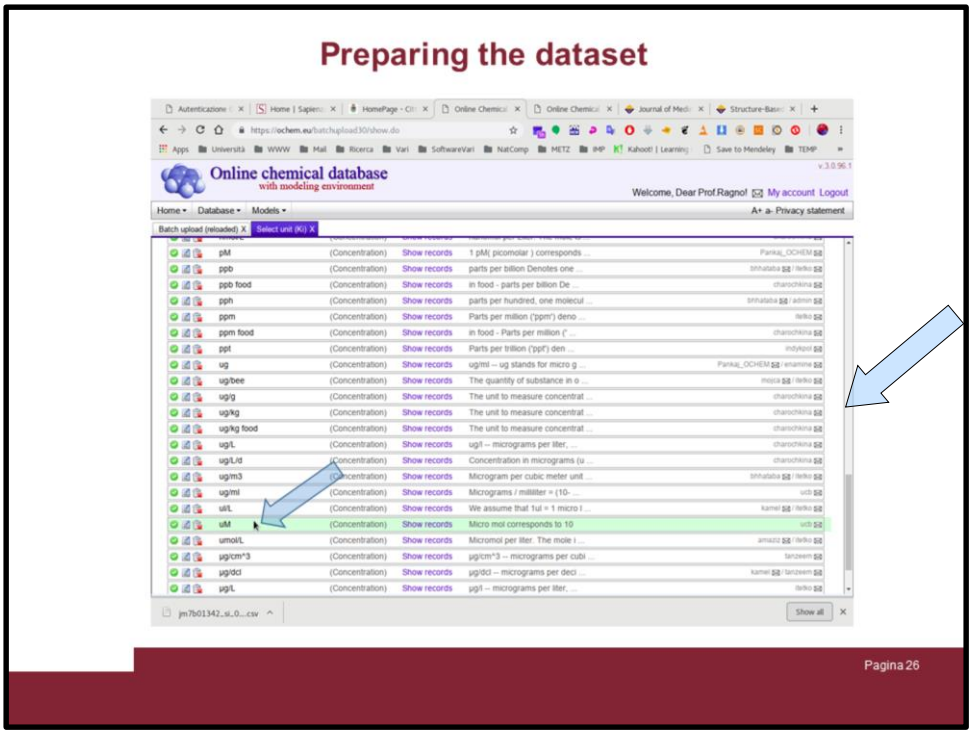
You will have a preview on the data under uploading, slide down ...



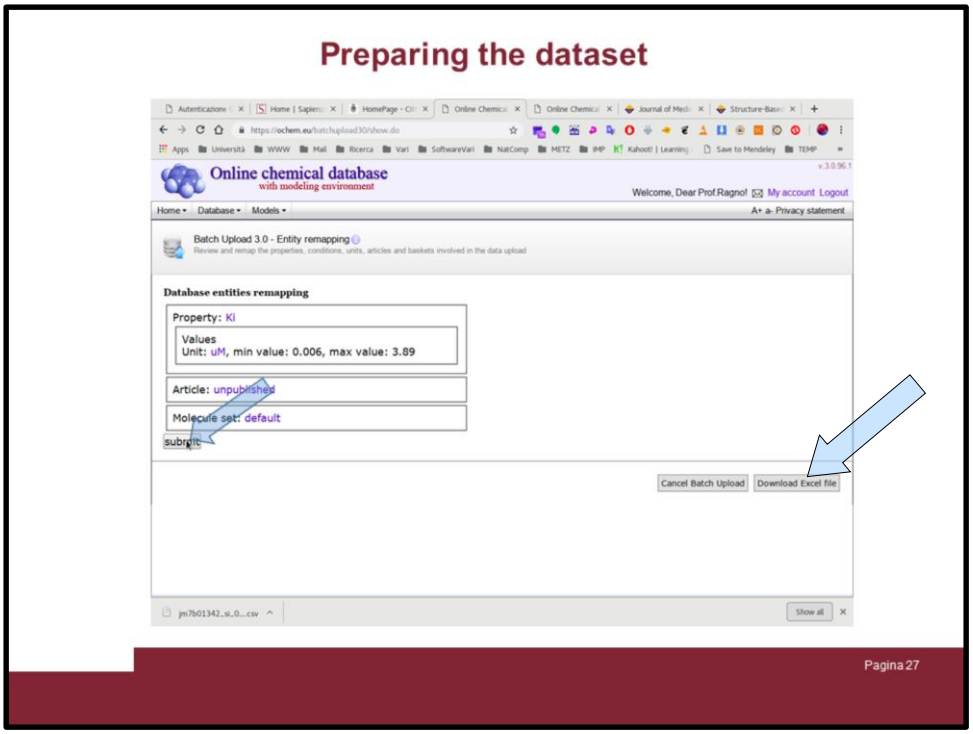
And click on “Upload this sheet”



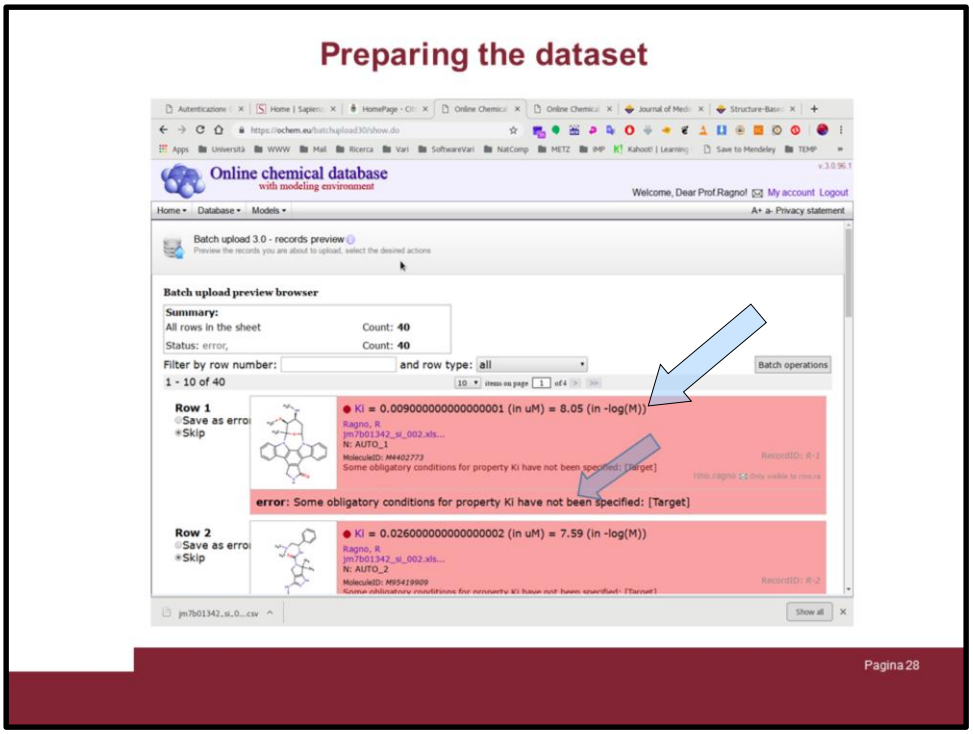
Next the portal will show some info on the data, pay attention on how the activity potency (IC50 or Ki) will be interpreted, here the program suggest Kis as log(M). You know it is not true so it have to be fixed, therefore click on the “log(M)” link ...



A list of measures will be shown. Slide down till you find uM or the unit you know is your activity ppotency.

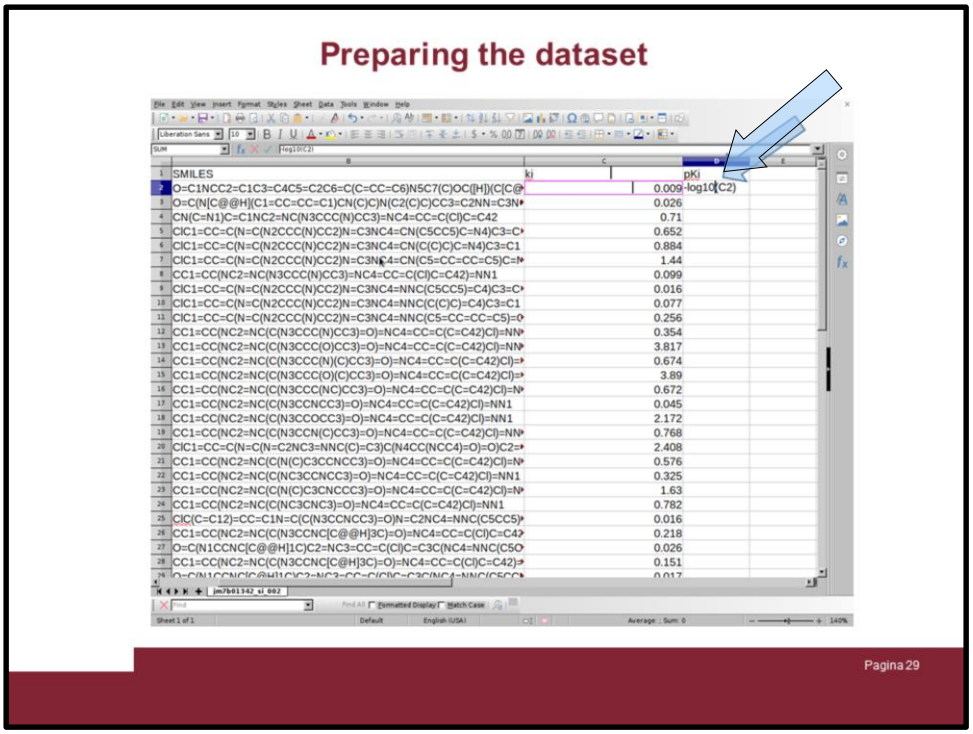


At this poit almos everything is set. Click on “Download excel file”

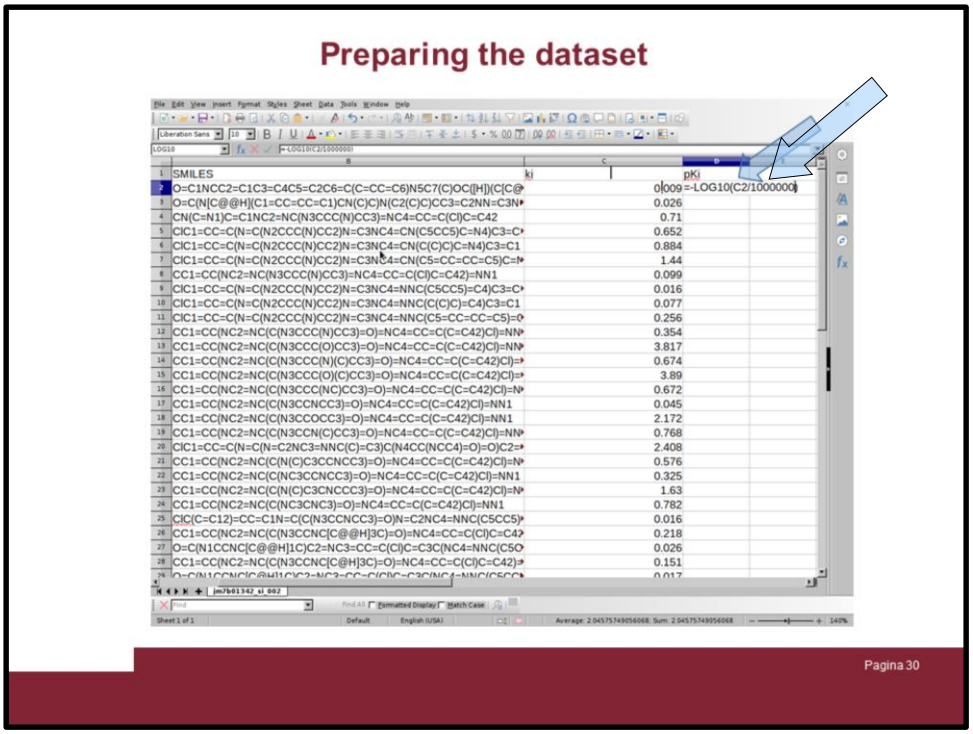


And it can happen the program complain about something, although everything was correctly made.

In this case the complain is about the Ki.



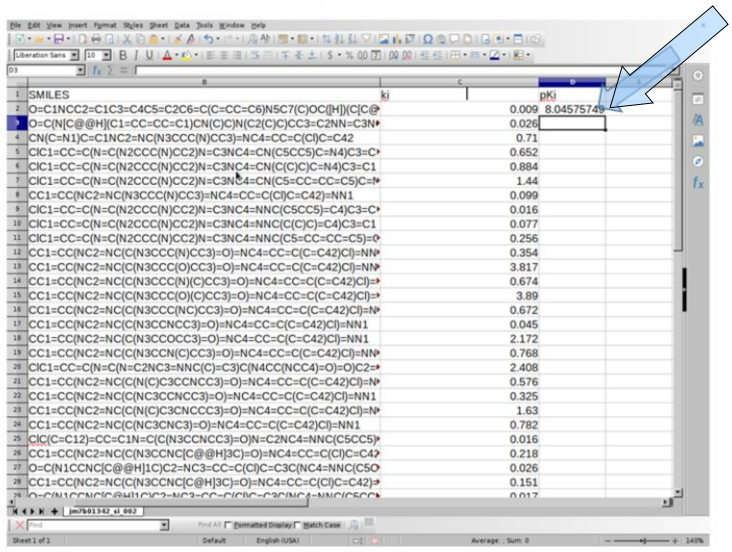
So go back in your excel file and convert the Ki into pki as follows



Click on the cell aside the first Ki value end insert"=log10(C2/1000000).

This means you are going to transform the ki from micromolar to molar and make the log10.

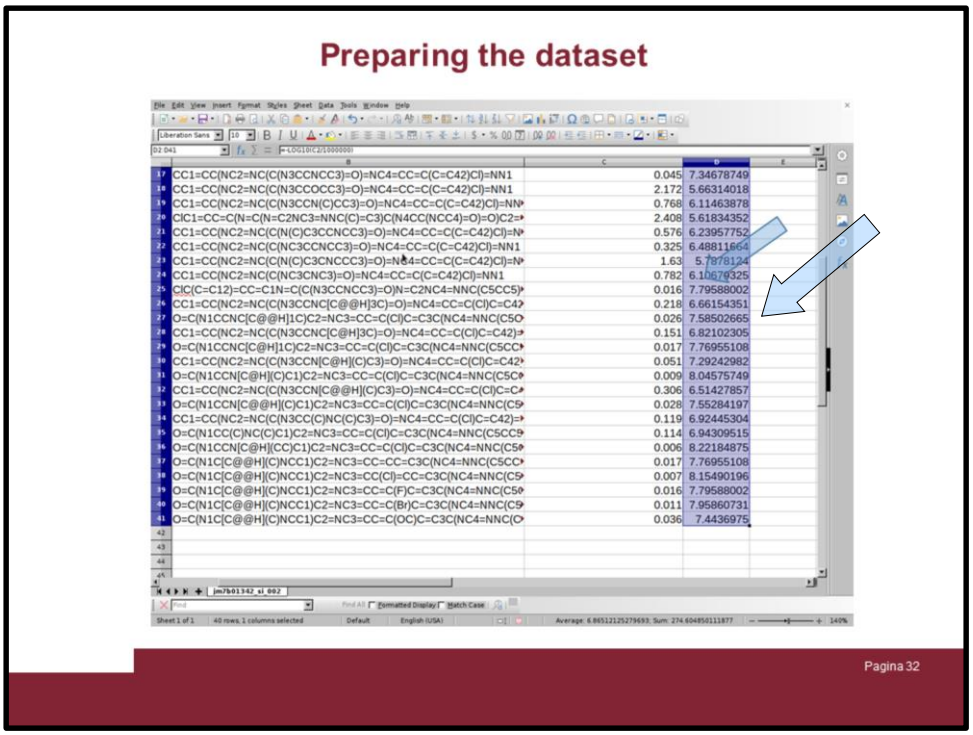
Preparing the dataset



	SMILES	ki	pKi
2	O=C1NCC2=C1C3=C4C5=C2C6=C(C=CC=C6)N5C7(C)OC([H])(C[C@	0.009	8.04575749
3	O=C(N[C@@H](C1=CC=CC=C1)CN(C)C)N(C2(C)C)CC3=C2NN=C3N	0.026	
4	CN(C=N1)C=C1NC2=NC(N3CCC(N)CC3)=NC4=CC=C(Cl)C=C42	0.71	
5	ClC1=CC=C(N=C(N2CCC(N)CC2)N=C3NC4=CN(C5CC5)C=N4)C3=C	0.652	
6	ClC1=CC=C(N=C(N2CCC(N)CC2)N=C3NC4=CN(C(C)C)C=N4)C3=C1	0.884	
7	ClC1=CC=C(N=C(N2CCC(N)CC2)N=C3NC4=CN(C5=CC=CC=C5)C=N	1.44	
8	CC1=CC(NC2=NC(N3CCC(N)CC3)=NC4=CC=C(Cl)C=C42)=NN1	0.099	
9	ClC1=CC=C(N=C(N2CCC(N)CC2)N=C3NC4=NNC(C5CC5)=C4)C3=C	0.016	
10	ClC1=CC=C(N=C(N2CCC(N)CC2)N=C3NC4=NNC(C(C)C)=C4)C3=C1	0.077	
11	ClC1=CC=C(N=C(N2CCC(N)CC2)N=C3NC4=NNC(C5=CC=CC=C5)=C	0.256	
12	CC1=CC(NC2=NC(C(N3CCC(N)CC3)=O)=NC4=CC=C(C=C42)Cl)=NN	0.354	
13	CC1=CC(NC2=NC(C(N3CCC(O)CC3)=O)=NC4=CC=C(C=C42)Cl)=NN	3.817	
14	CC1=CC(NC2=NC(C(N3CCC(N)(C)CC3)=O)=NC4=CC=C(C=C42)Cl)=	0.674	
15	CC1=CC(NC2=NC(C(N3CCC(O)(C)CC3)=O)=NC4=CC=C(C=C42)Cl)=	3.89	
16	CC1=CC(NC2=NC(C(N3CCC(NC)CC3)=O)=NC4=CC=C(C=C42)Cl)=N	0.672	
17	CC1=CC(NC2=NC(C(N3CCNCC3)=O)=NC4=CC=C(C=C42)Cl)=NN1	0.045	
18	CC1=CC(NC2=NC(C(N3CCOCC3)=O)=NC4=CC=C(C=C42)Cl)=NN1	2.172	
19	CC1=CC(NC2=NC(C(N3CCN(C)CC3)=O)=NC4=CC=C(C=C42)Cl)=NN	0.768	
20	ClC1=CC=C(N=C(N=C2NC3=NNC(C)=C3)C(N4CC(NCC4)=O)=O)C2=	2.408	
21	CC1=CC(NC2=NC(C(N(C)C3CCNCC3)=O)=NC4=CC=C(C=C42)Cl)=N	0.576	
22	CC1=CC(NC2=NC(C(NC3CCNCC3)=O)=NC4=CC=C(C=C42)Cl)=NN1	0.325	
23	CC1=CC(NC2=NC(C(N(C)C3CNCCC3)=O)=NC4=CC=C(C=C42)Cl)=N	1.63	
24	CC1=CC(NC2=NC(C(NC3CNC3)=O)=NC4=CC=C(C=C42)Cl)=NN1	0.782	
25	ClC(C=C12)=CC=C1N=C(C(N3CCNCC3)=O)N=C2NC4=NNC(C5CC5)	0.016	
26	CC1=CC(NC2=NC(C(N3CCNC[C@@H]3C)=O)=NC4=CC=C(Cl)C=C42	0.218	
27	O=C(N1CCNC[C@@H]1C)C2=NC3=CC=C(Cl)C=C3C(NC4=NNC(C5C	0.026	
28	CC1=CC(NC2=NC(C(N3CCNC[C@H]3C)=O)=NC4=CC=C(Cl)C=C42)	0.151	

Pagina 31

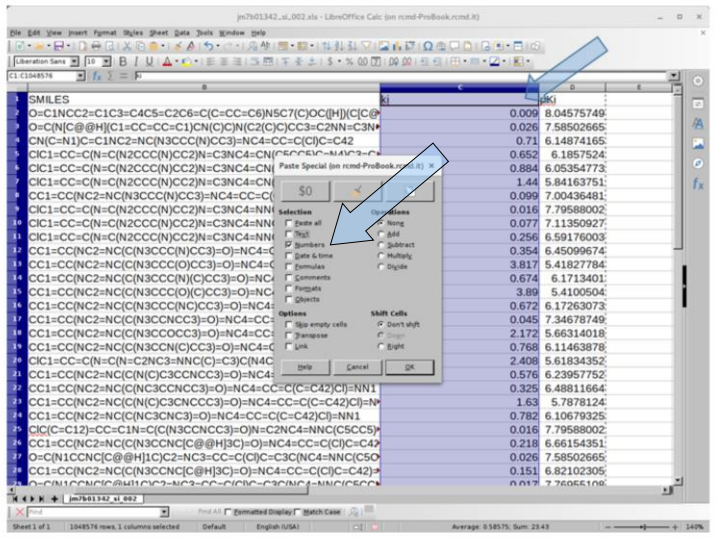
As soon as you click on enter the pKi value is calculated



Now select the cell containing the firs pKi value and by clicking on the little square on the right-botton and slide down to apply the calculation of the pKi to all the other cells.

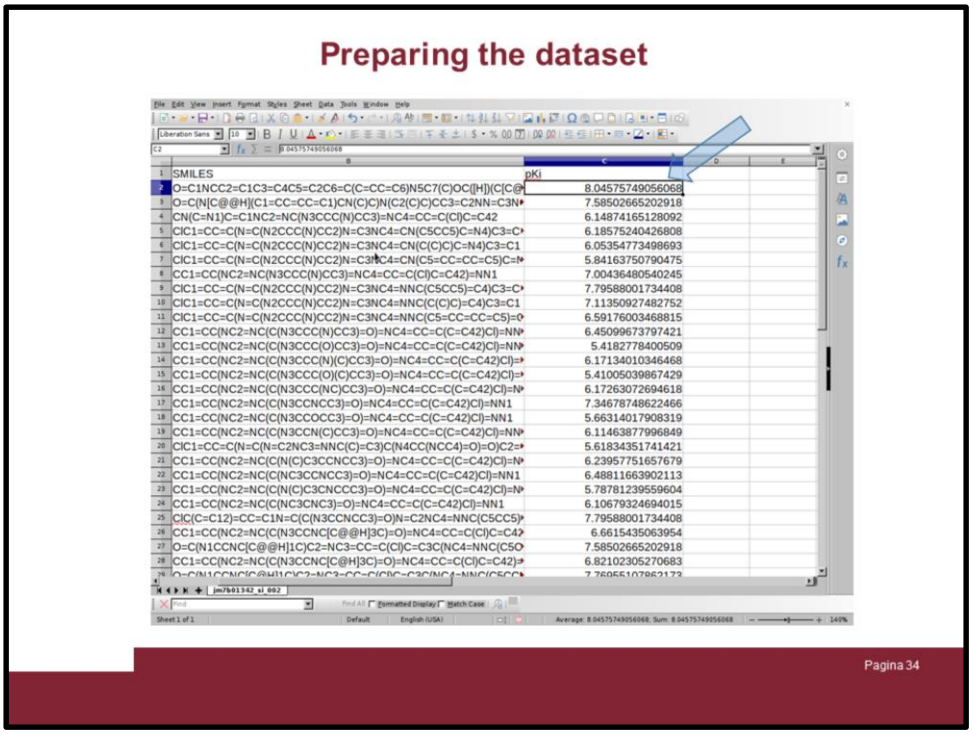
Select the pKi column and copy into the memory ...

Preparing the dataset



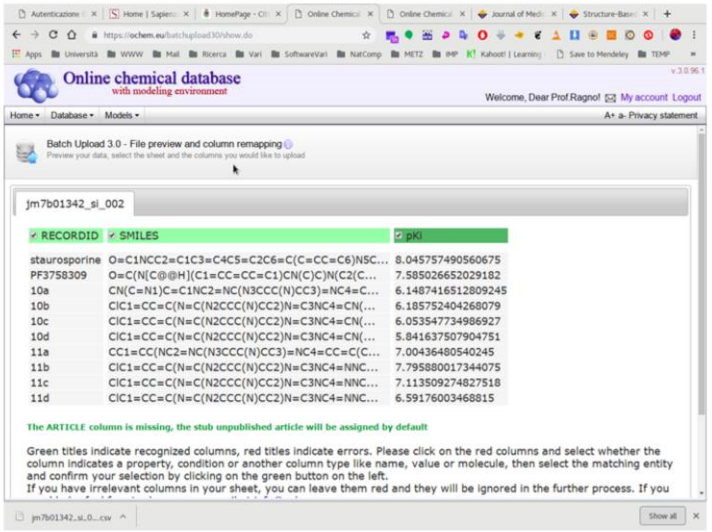
Pagina 33

… then paste it on the Ki column by choose “paste special” and indicating just the numbers.

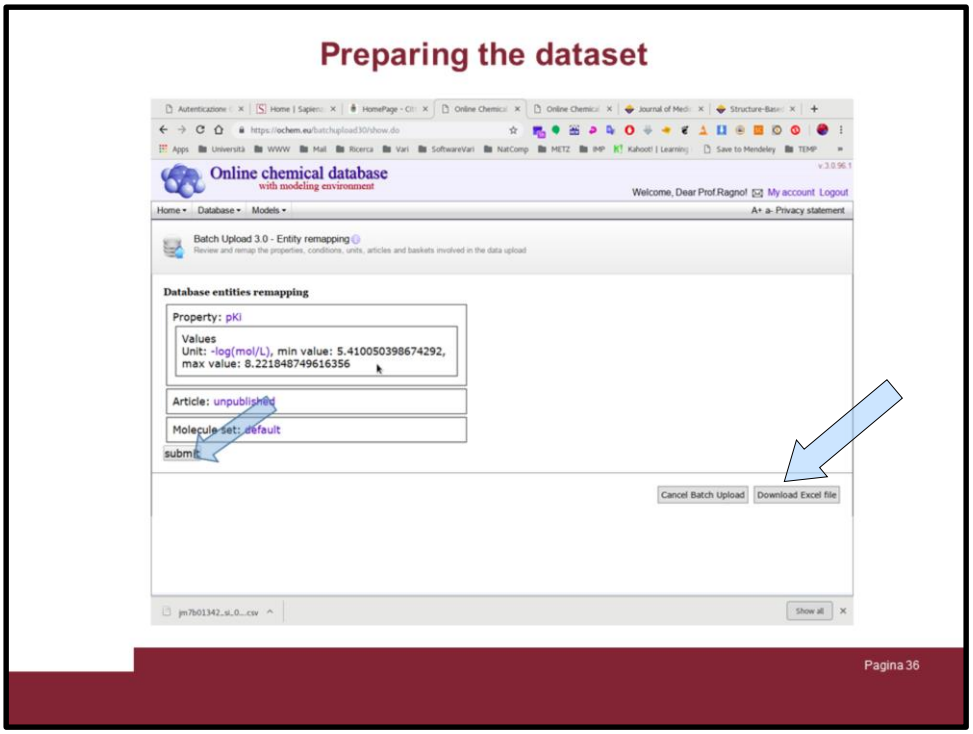


Now you have the excel file containing RECORDID, SMILES and pKi columns

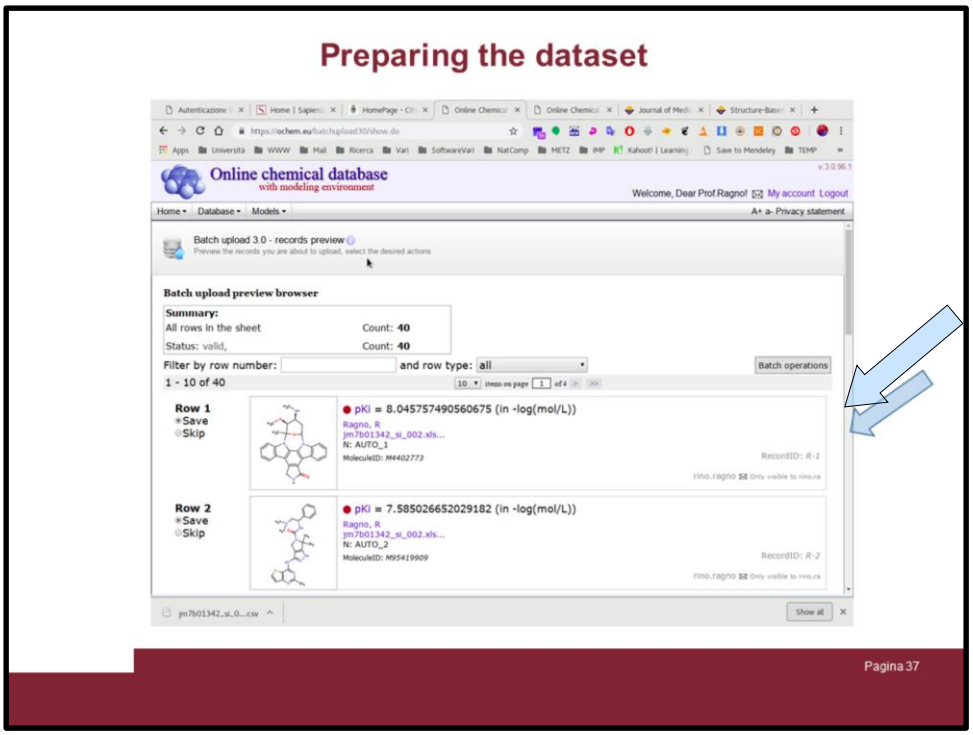
Preparing the dataset



Go to OCHEM.eu and repeat the uploading of the excel file as above described



Now the unity of the activity potency is correctly reported, click on “Download Excel file”

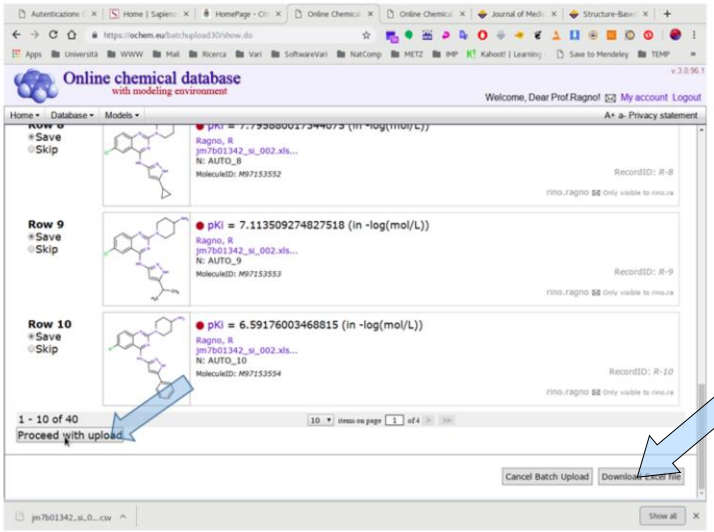


And everything is OK, slide down to check all the rows ...

In this tutorial the OCHEM.eu portal will complain about data duplication with the user rino.ragno@uniroma1.it.

Ignore it and go on.

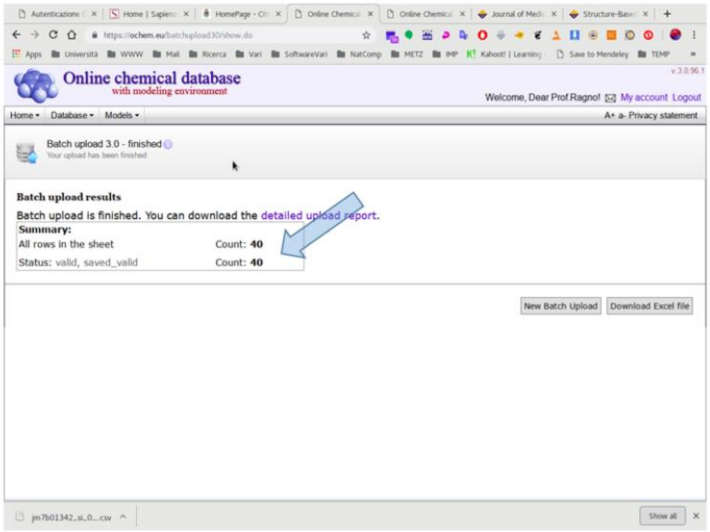
Preparing the dataset



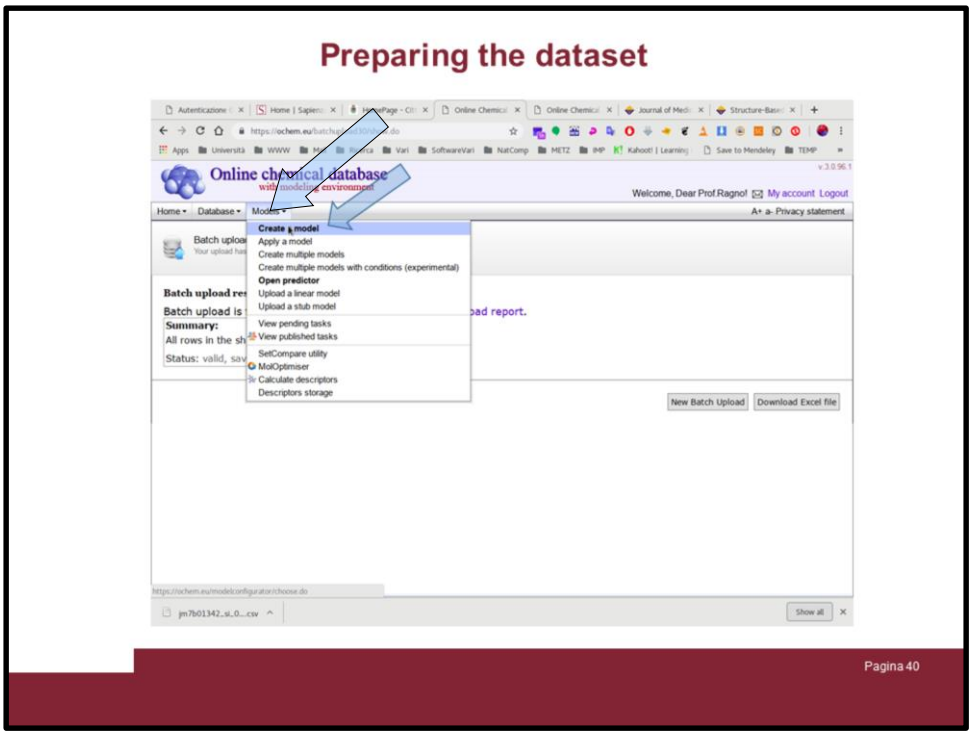
Pagina 38

… and if everything is OK click on “Proceed with upload” again

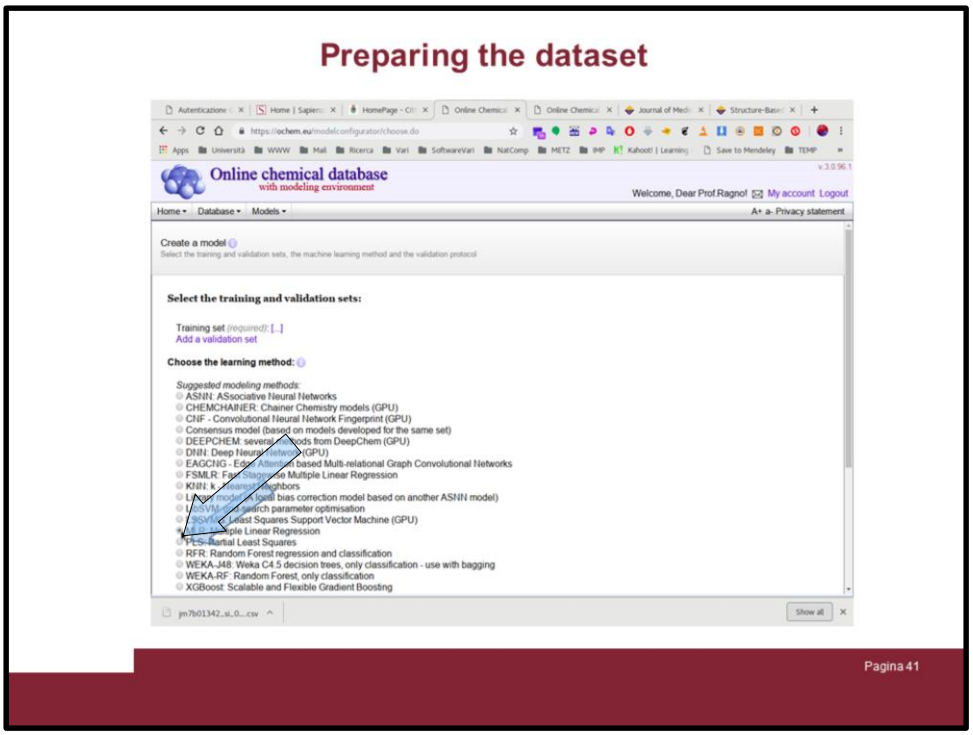
Preparing the dataset



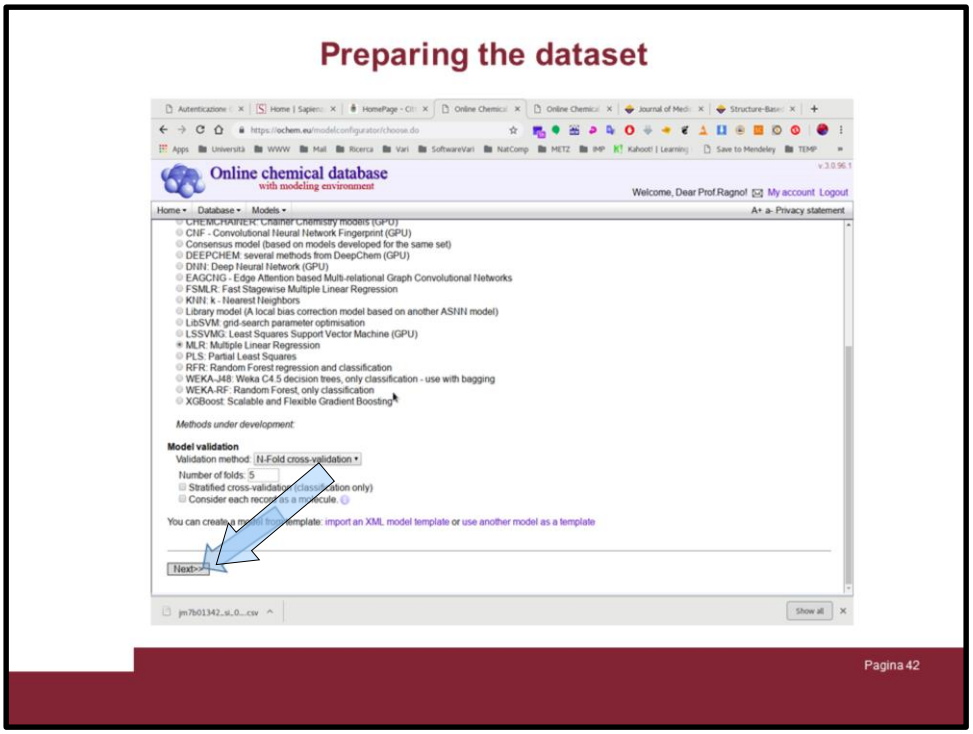
The portal will report on the data you have uploaded.



Now let try to make some models. Click on the “Models” tab and select “Create a model”

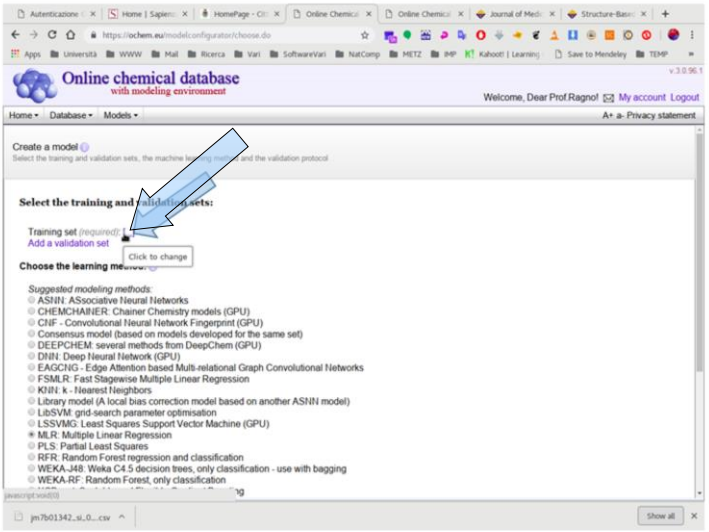


A window will give you the opportunity to select the statistical engine, for this purpose select PLS (although in the image is show selected MLR)



… slide down and click on “Next>>”

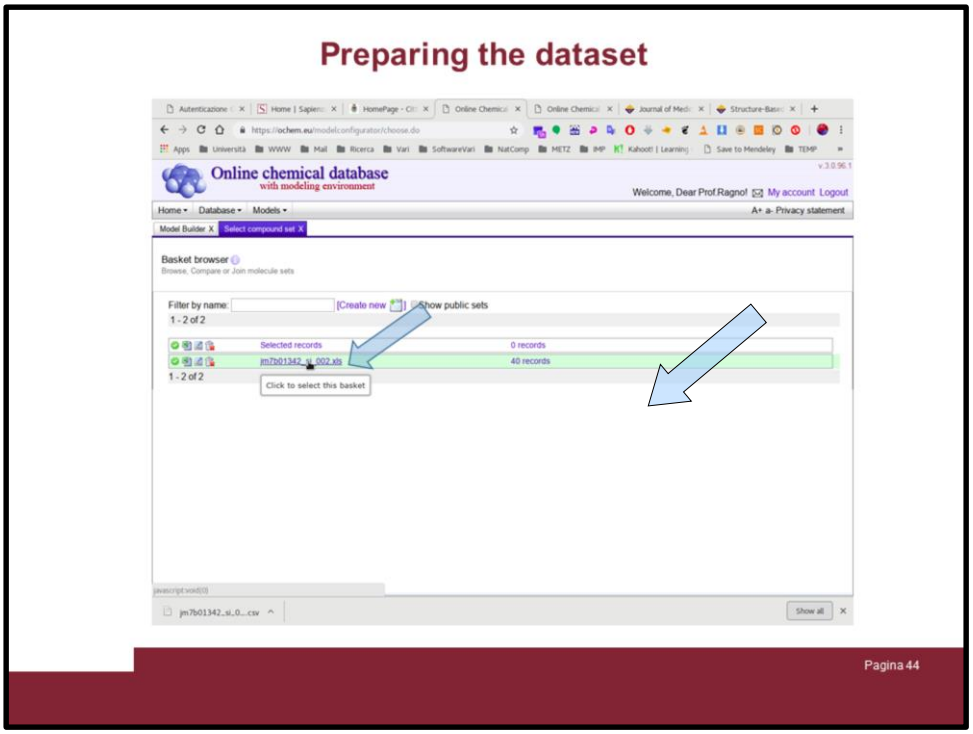
Preparing the dataset



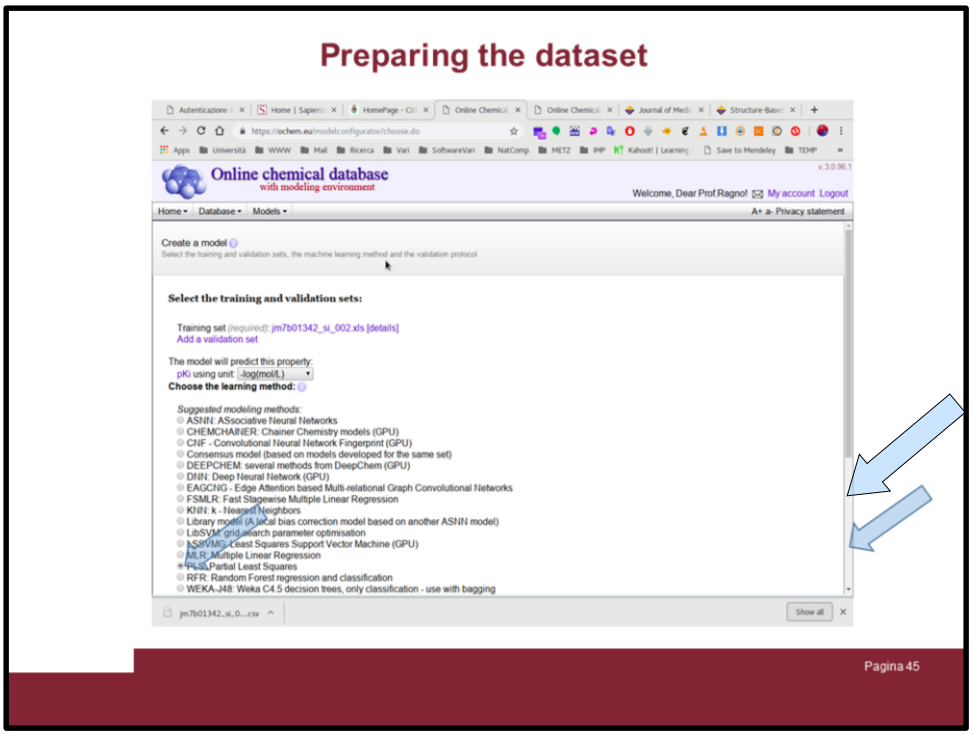
Pagina 43

… but it will ask you to select a training set.

Then click in the “[...]” link ...



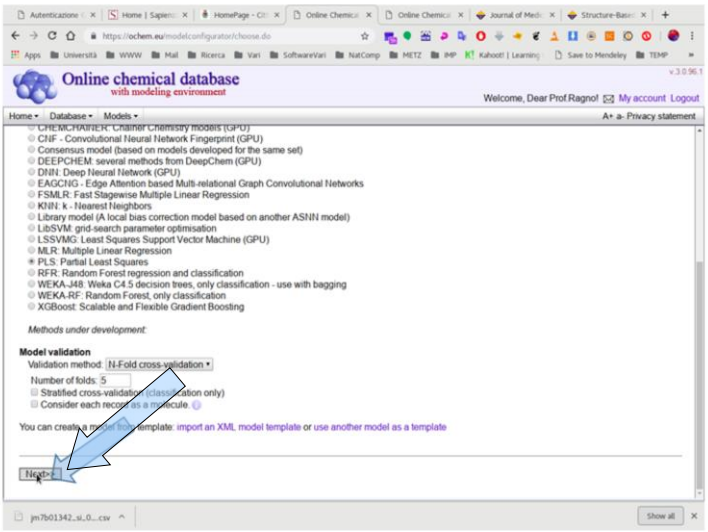
… and select the excel file containing the dataset as prepared above



Now the correct statistical method is selected (PLS) and there is also the training set.

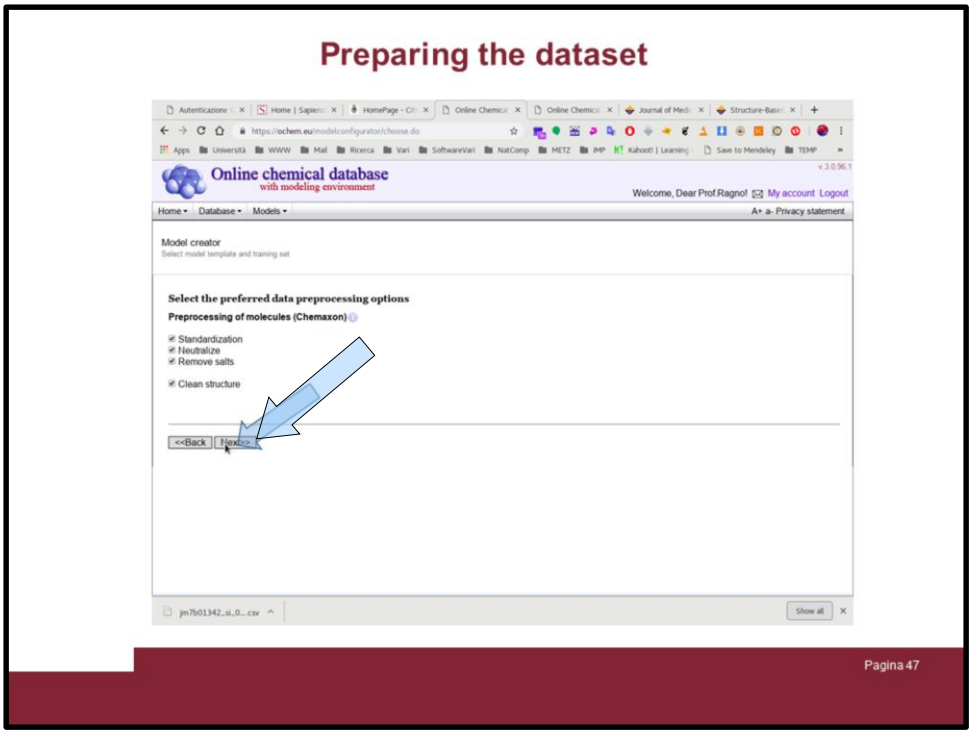
Slide down ...

Preparing the dataset

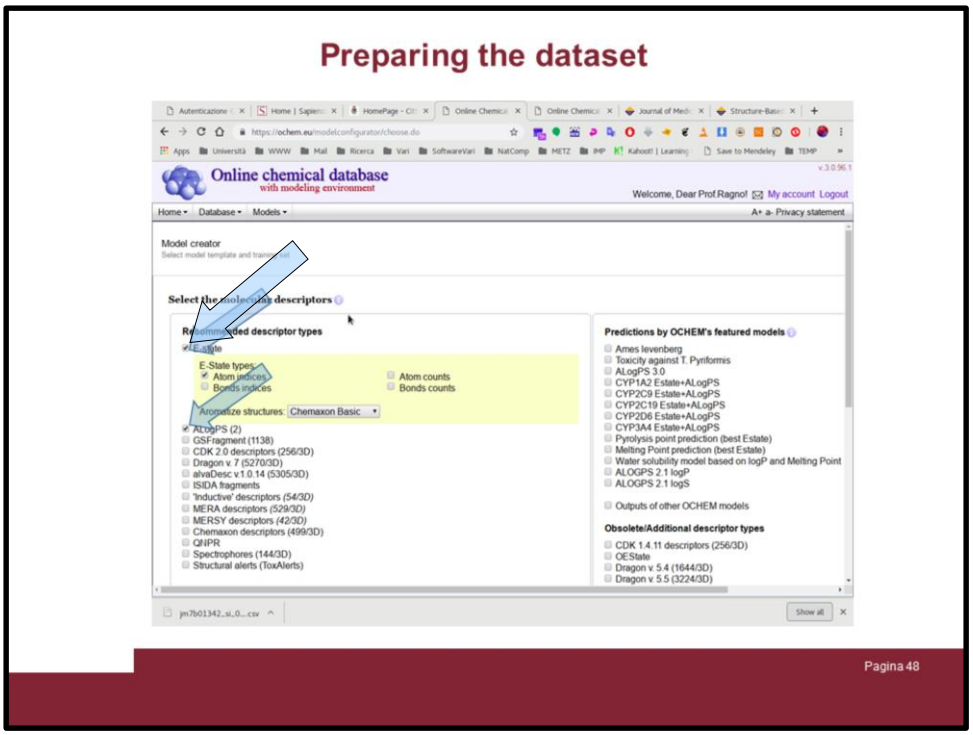


Pagina 46

... and click on “Next>>”

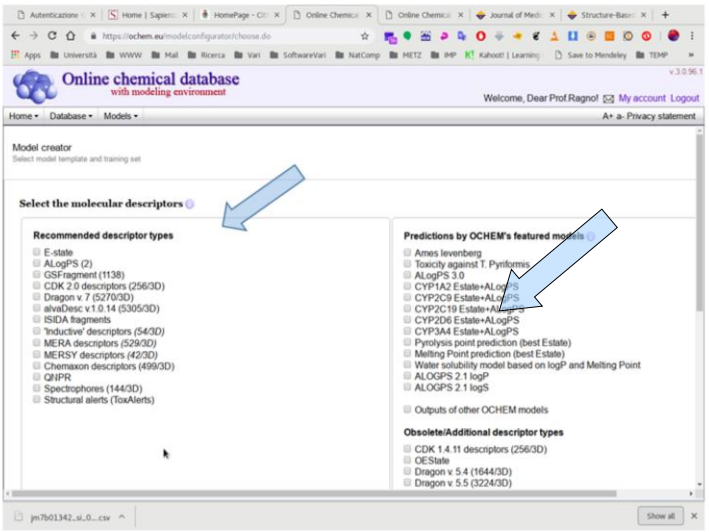


A window will ask you how to treat the molecules, leave everything as suggested and click on “Next>>”.



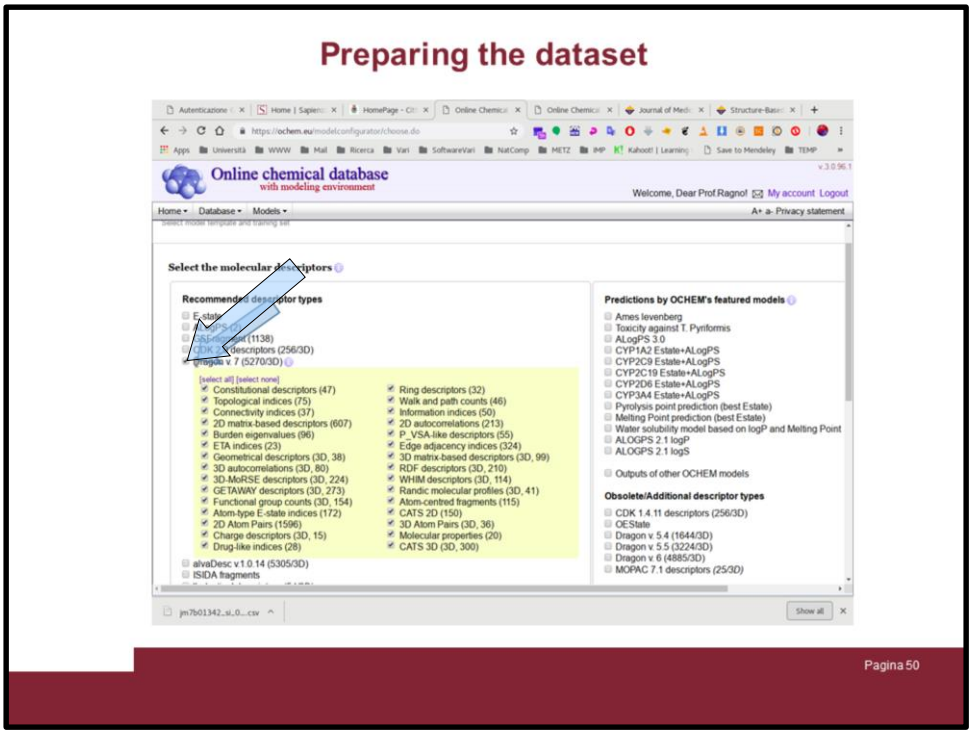
In this windows the portal let you select which descriptors you want to use. Uncheck everything ...

Preparing the dataset

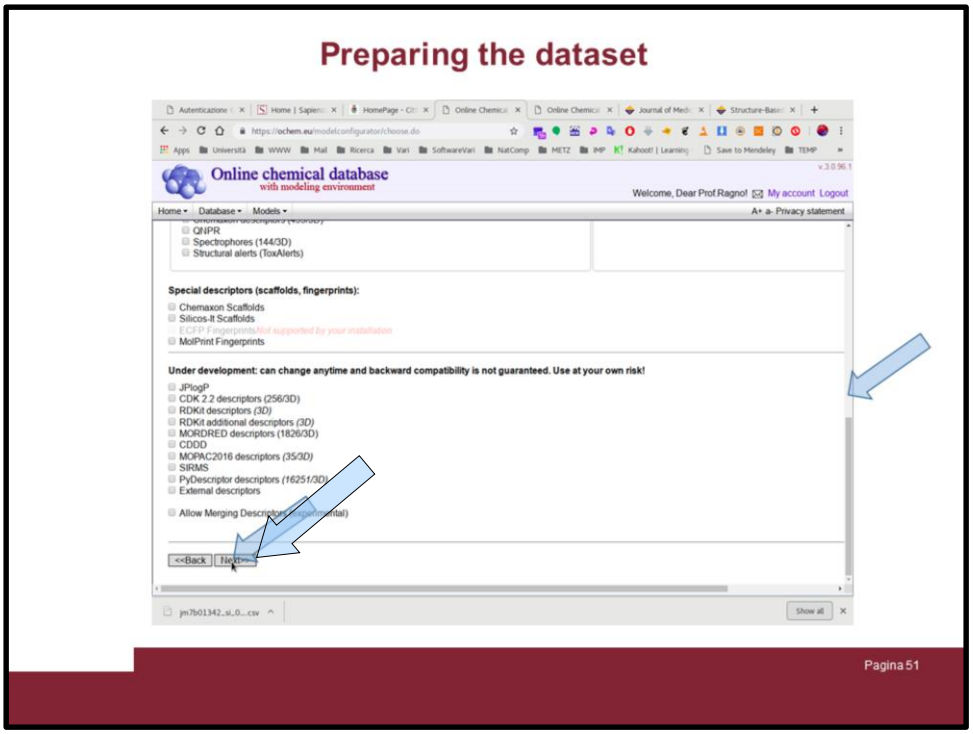


Pagina 49

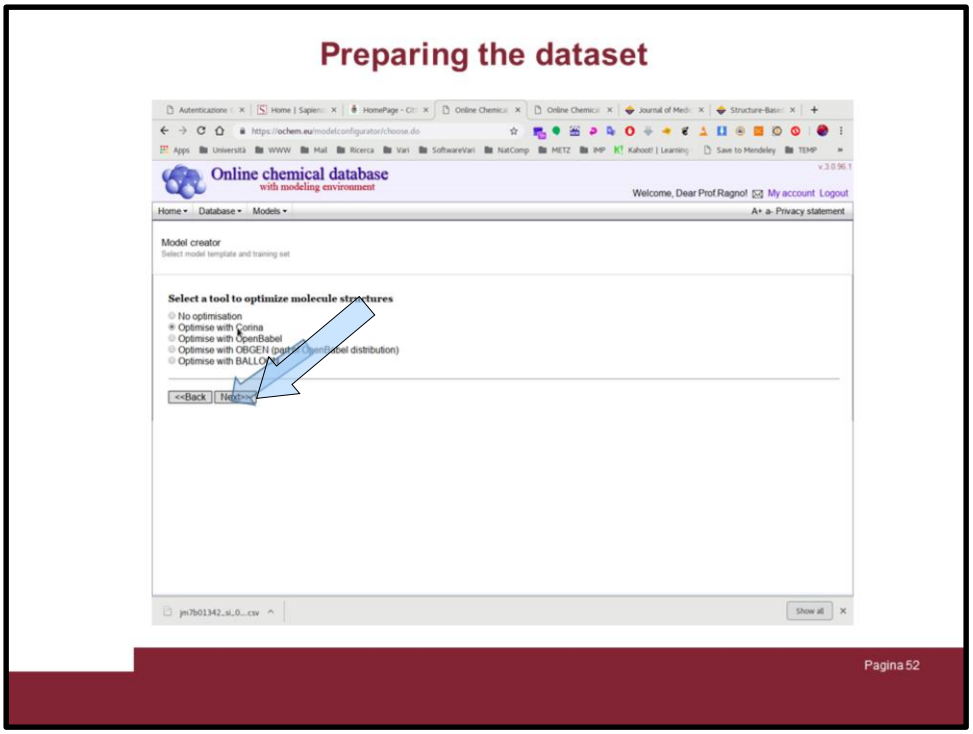
Once all is unchecked ...



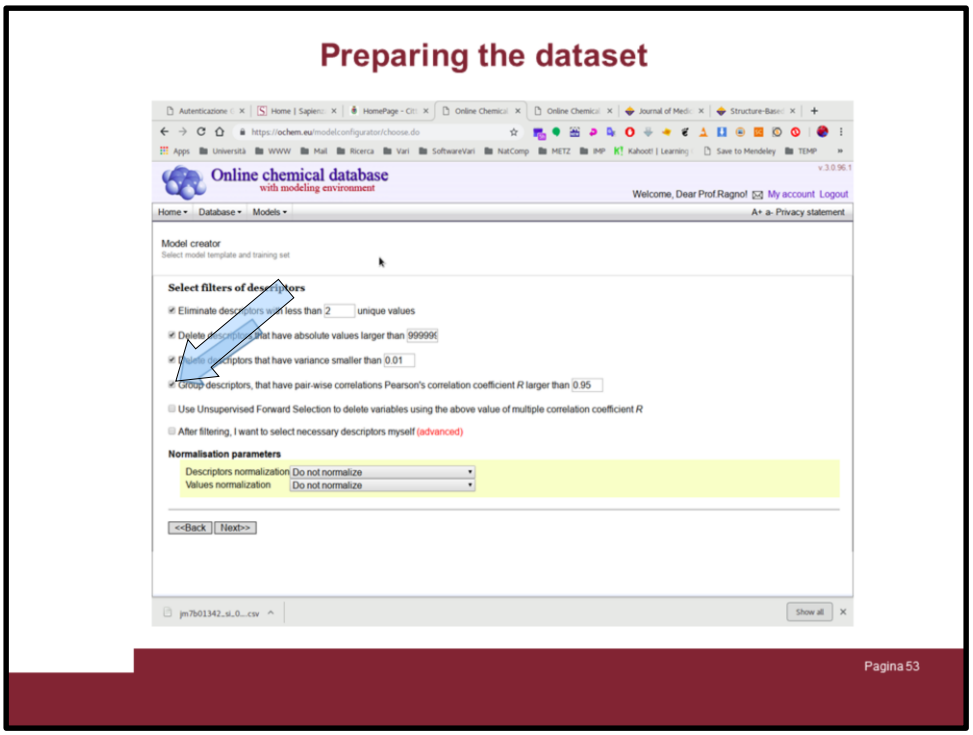
Select the Dragoin v. 7 descriptors. Make sure all are selected.



Slide down and click on “Next>>”

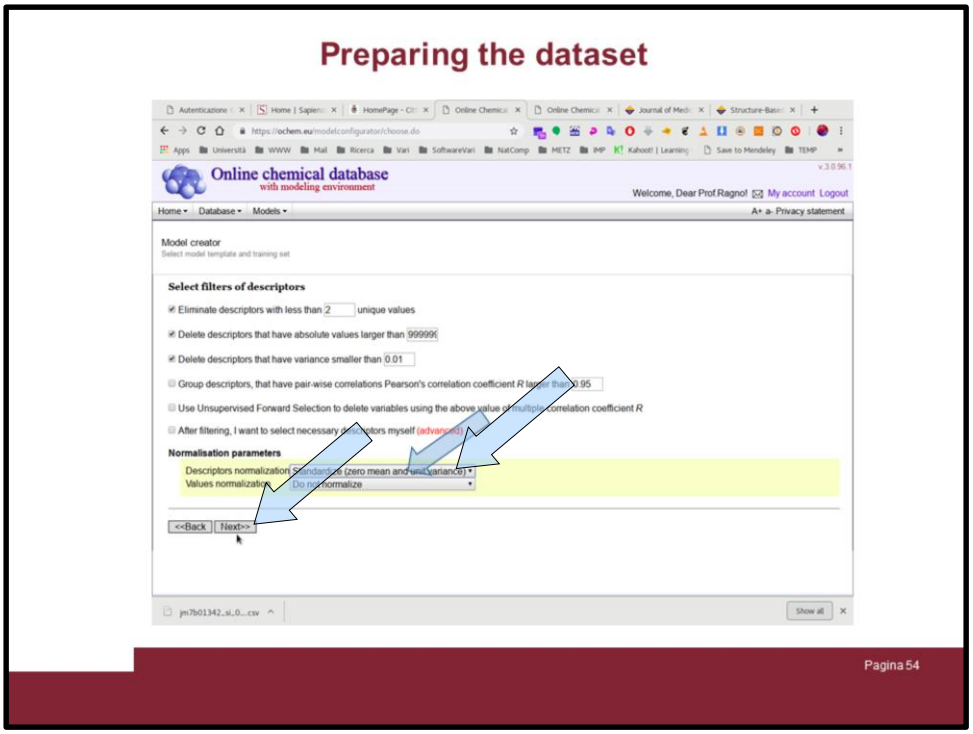


Here you can choose hoe to convert into 3D the smiles, accept “optimise with Corina” and click on “Next>>”



In this next window you can transform and filter the descriptor data ….

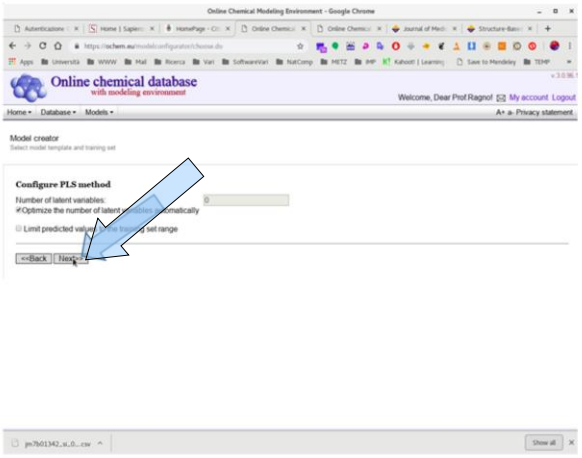
Deselect “Group descriptors” and ...



… set the normalization as shown in the image.

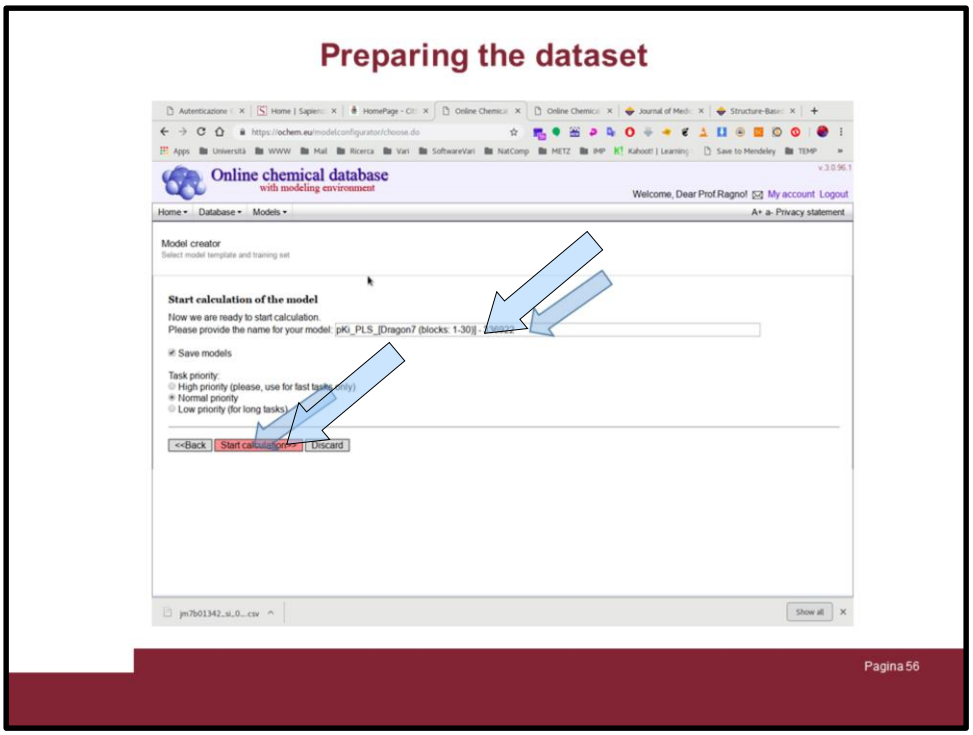
Then click on “Next>>”

Preparing the dataset

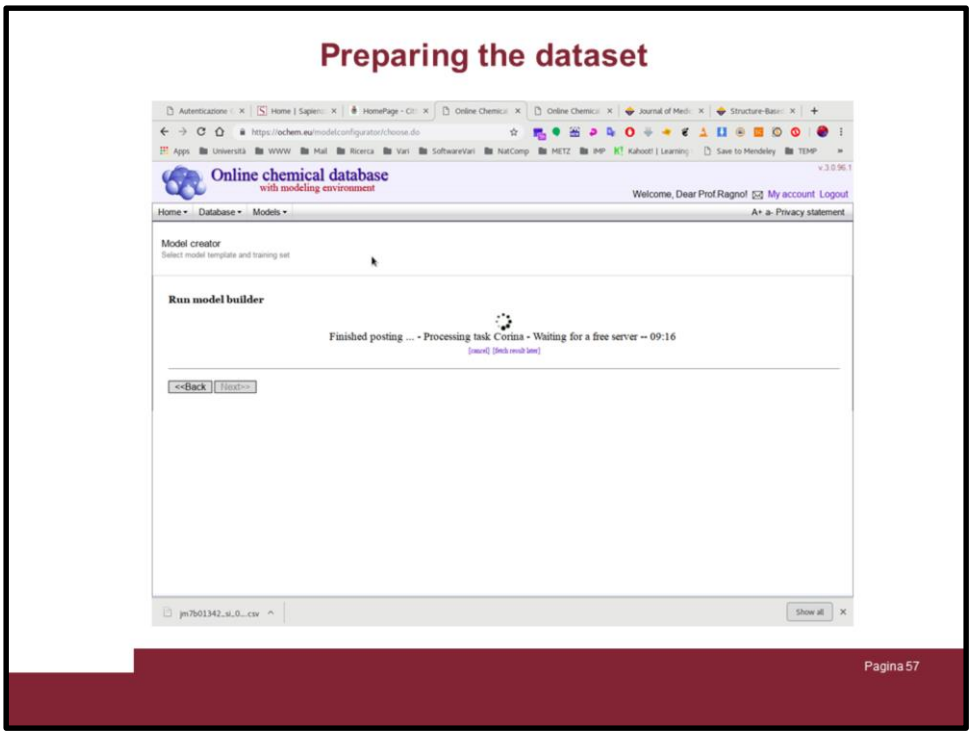


Pagina 55

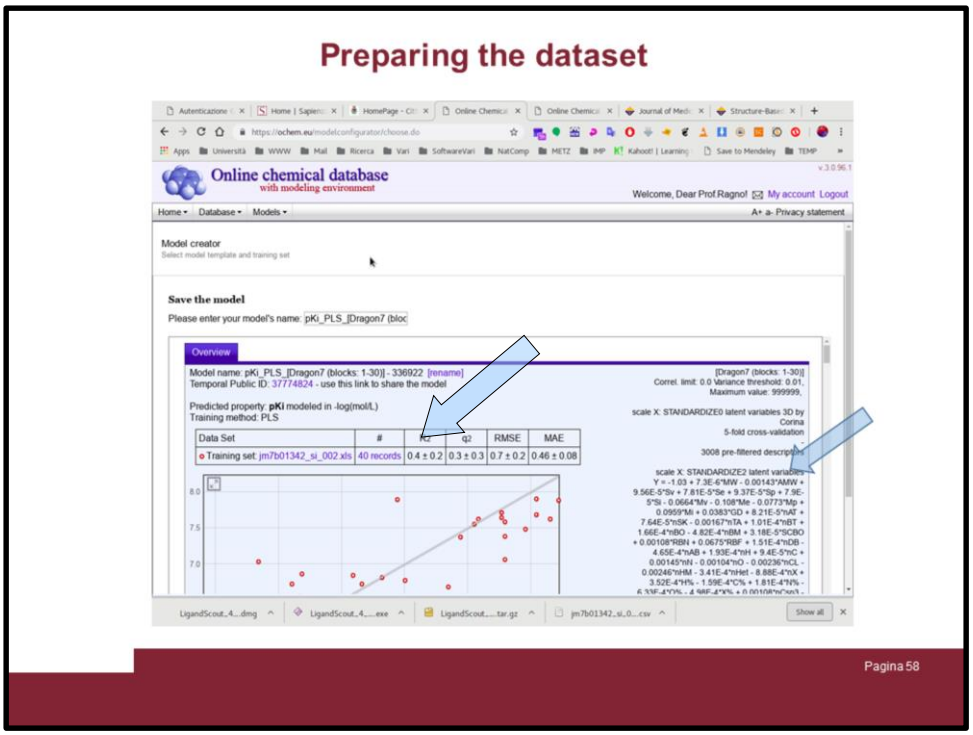
… everything is ready, just click on “Next>>” to launch the calculation.



Befor launching the portal will ask to set the name of the under creating model. Modify it as you like and click on “Start calculation”



A windows wiil open and showing the process ...



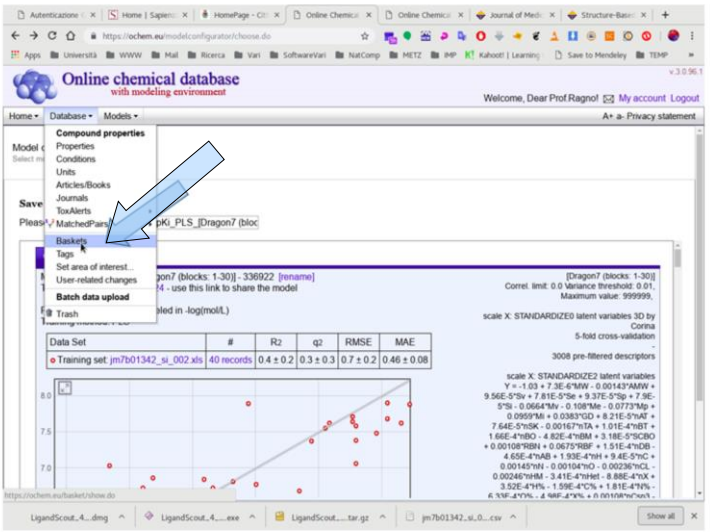
After some minutes the results will be shown.

Here you can see the R2 is not so high (2 Principal components).

But if you have payed attention in the training set were also included the staurosporine and PF3758309 molecules that were used as reference drug in the article.

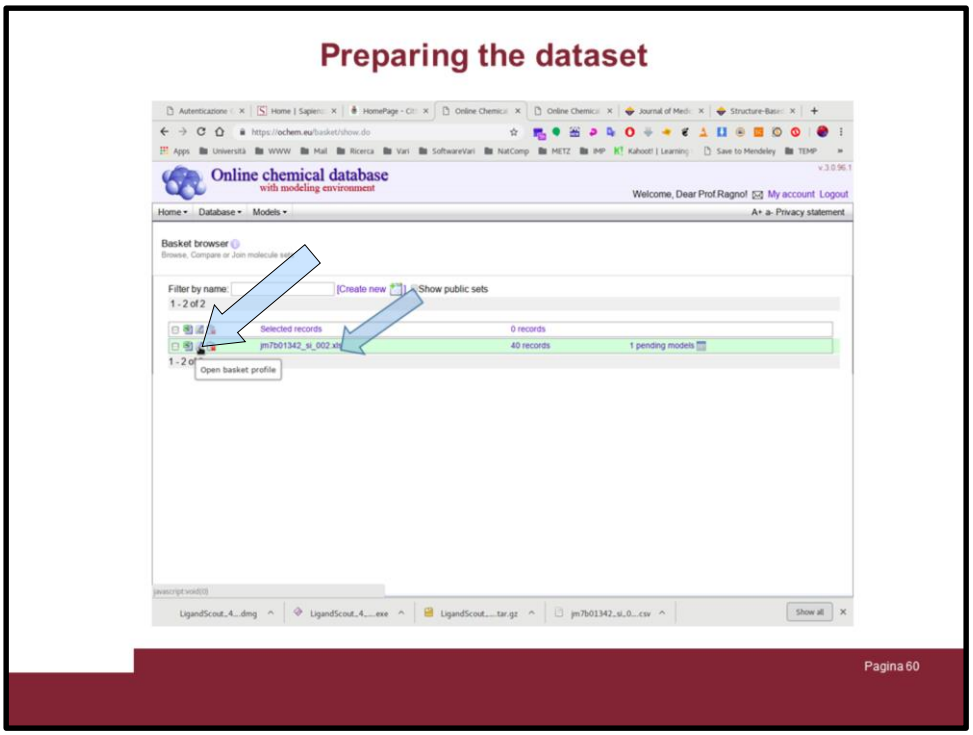
Therefore it could be assumed that those two molecules being quite different from the new molecules they should be removed from the training set.

Preparing the dataset

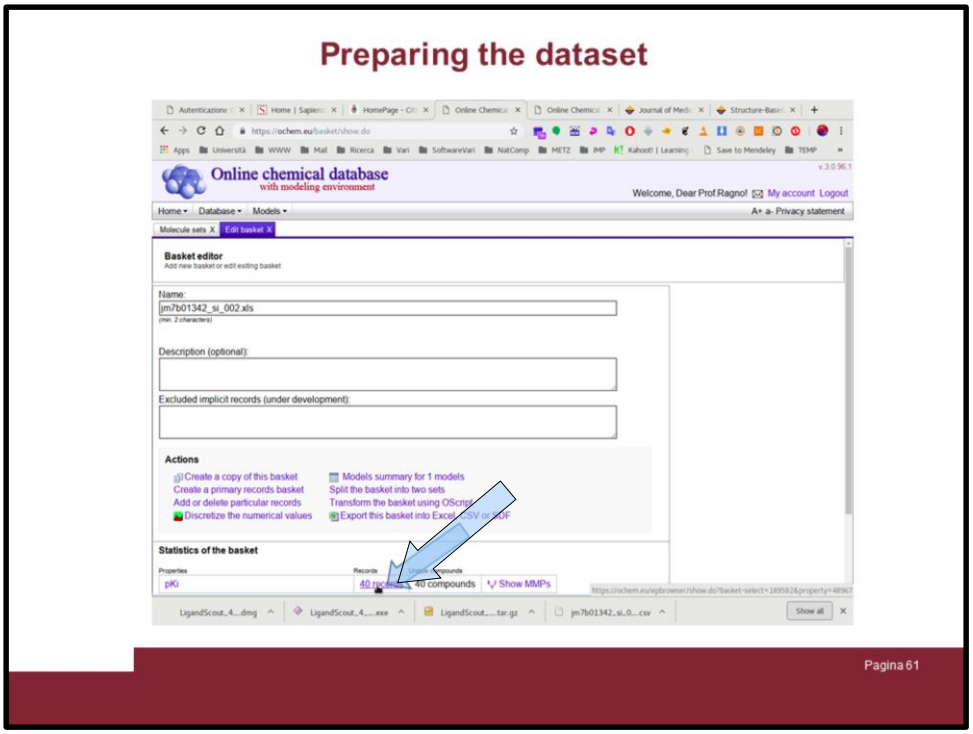


To remove molecules from the training set it is possible to create a selection of the original dataset and work on it.

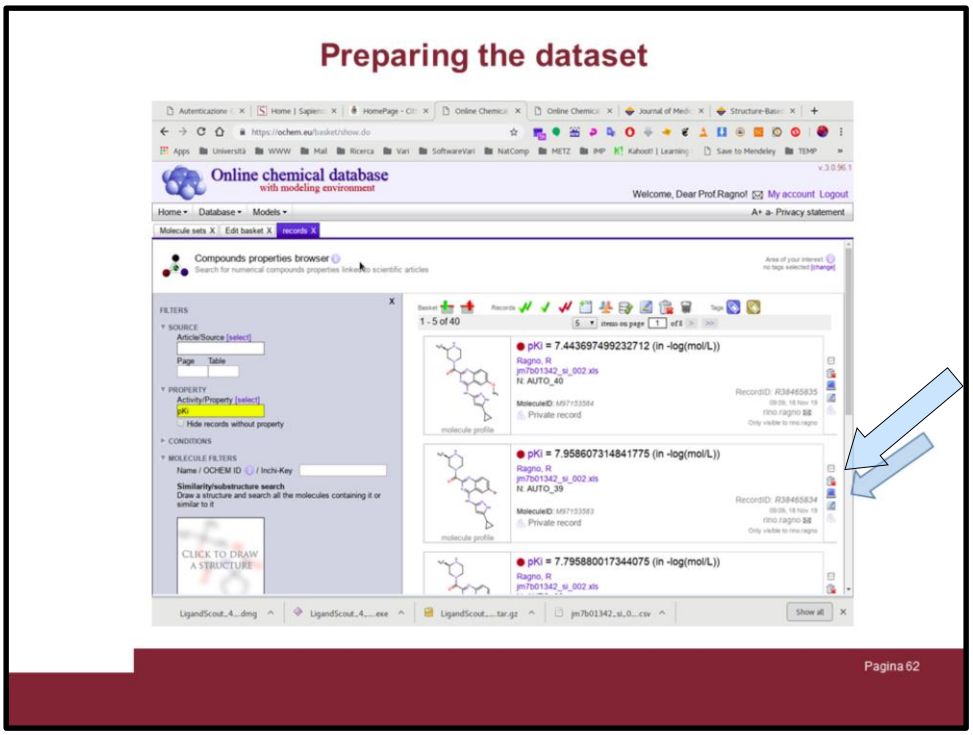
To do that click on “Database” tab and select baskets



The Baskets (Datasets) wil be listed. Let's click on the one we are working on ...

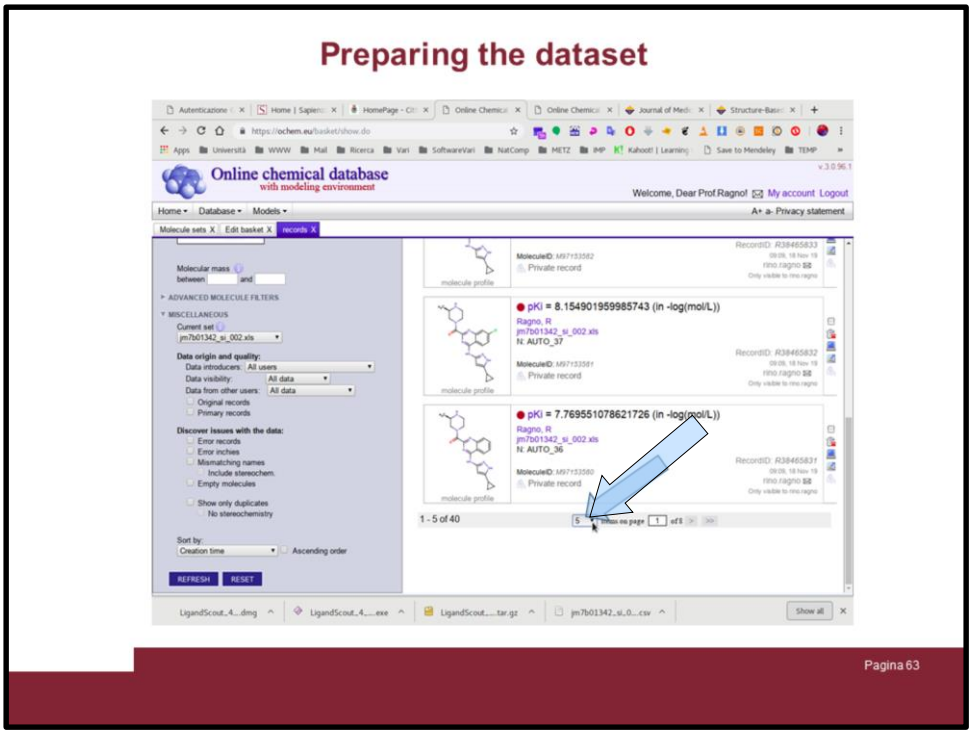


… and after the windows have changed to the one similat to the image click on the “records” link

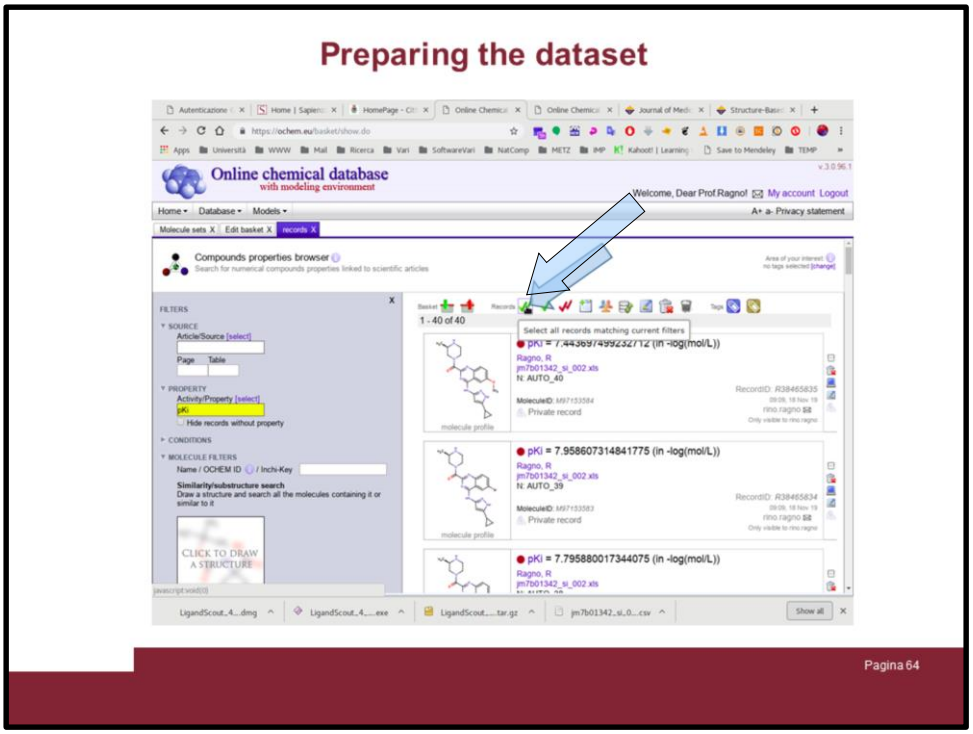


… the window will now show all the records (molecules) in groups of 5.

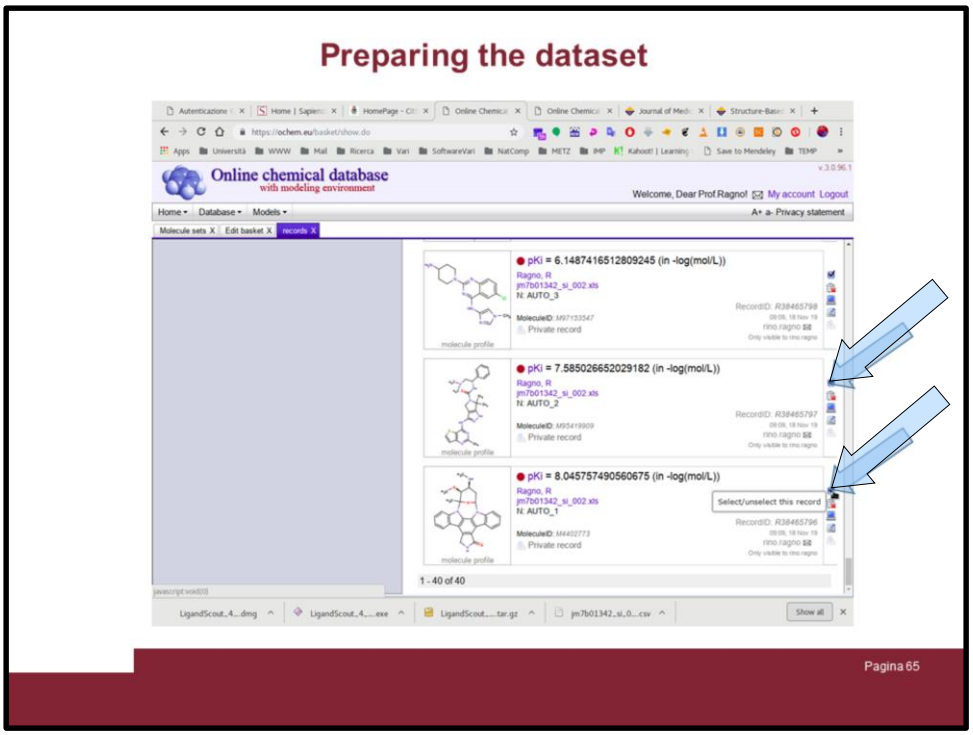
Slide down ...



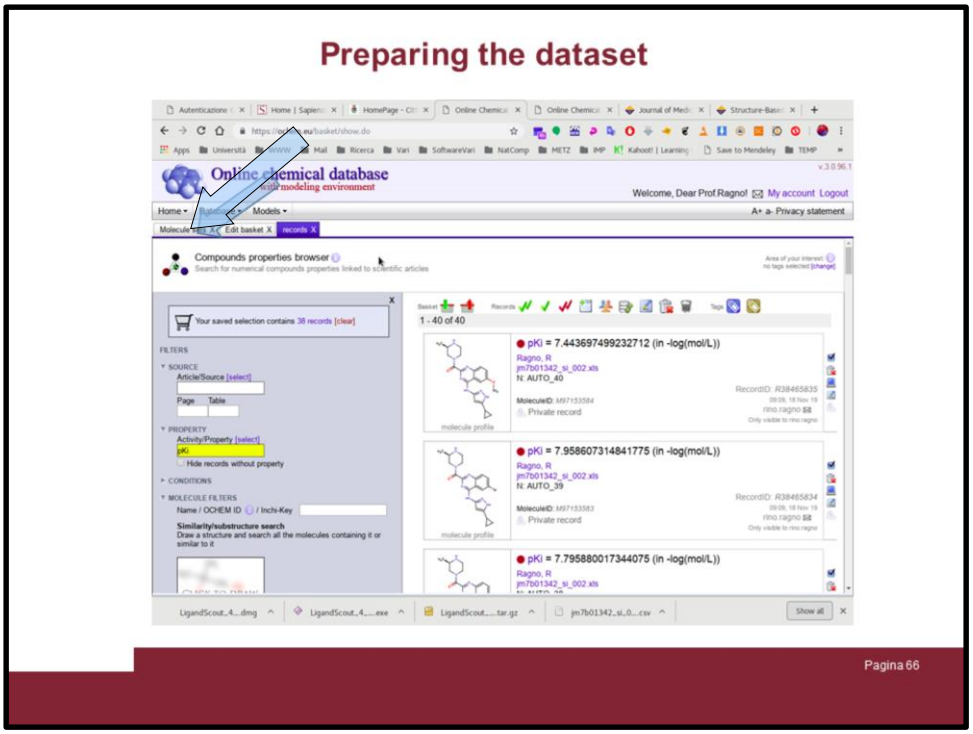
… and change it from 5 to 50 (a number greater of your molecule number)



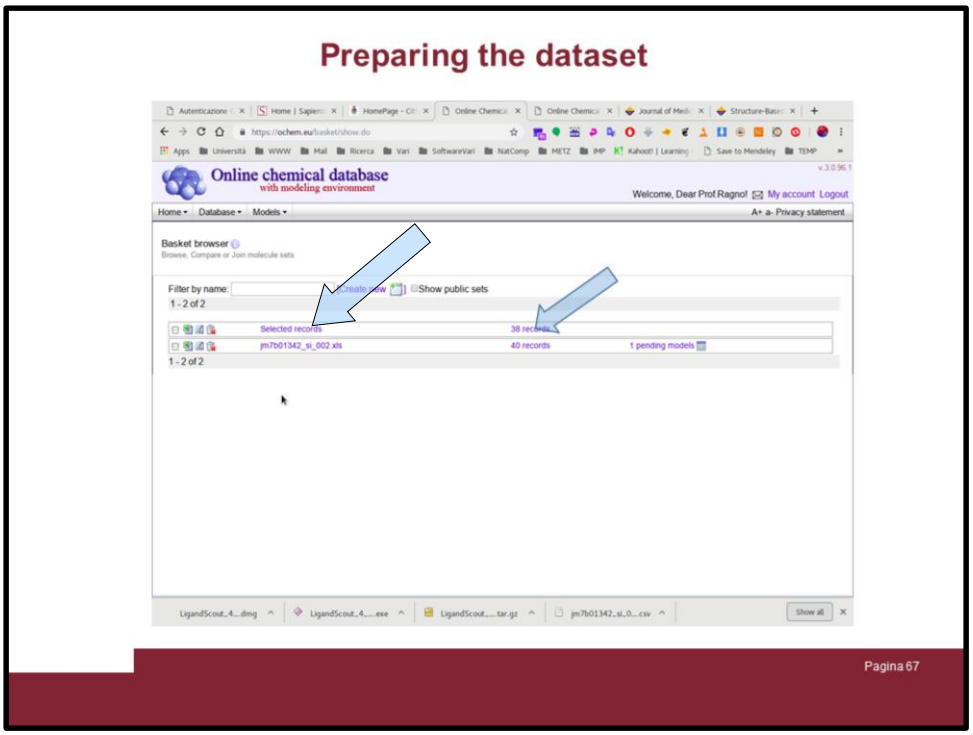
Now you have all the 40 molecules displayed (sorted in reverse mode from the last to the first) , look for the staurosporine (AUTO_1) and PF3758309 (AUTO_2) molecules (theys should be in the bottom. But first select all the record by clickn on the double check green icon



Then slide down and deselect the two moelcules by clicking in the little boxes as shown

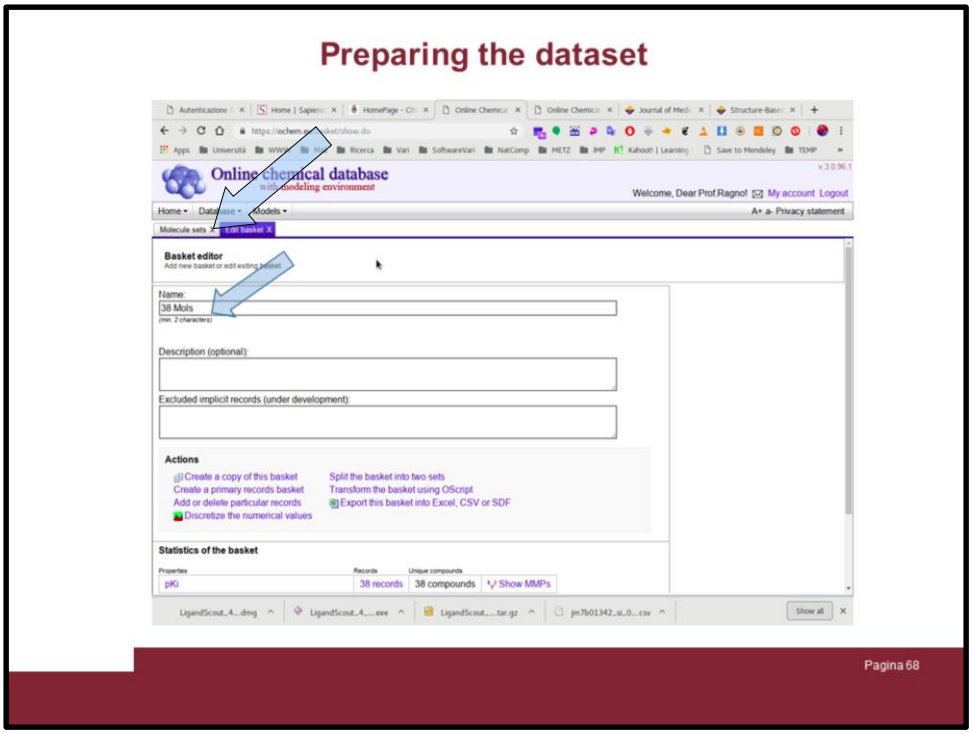


Now click on the “Molecule sets” tab and ...

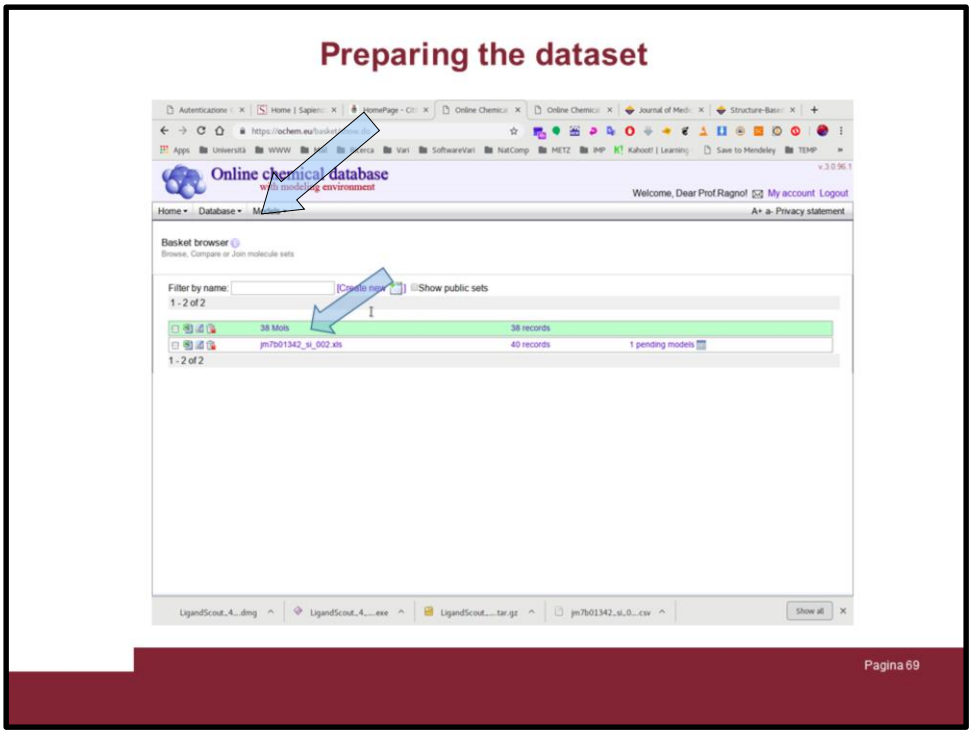


… the “Selected records” will indicate 38 records.

Click on it ...

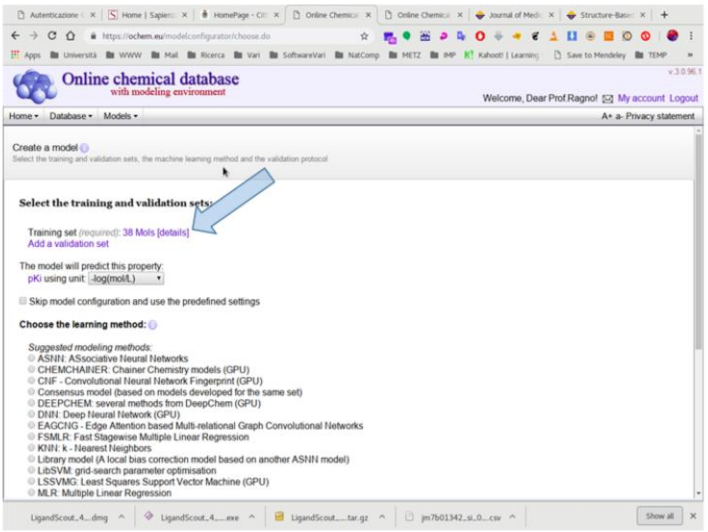


And rename it as you like and click again on the “Molecule sets” tab



Now create a model using the “38 Mols” set

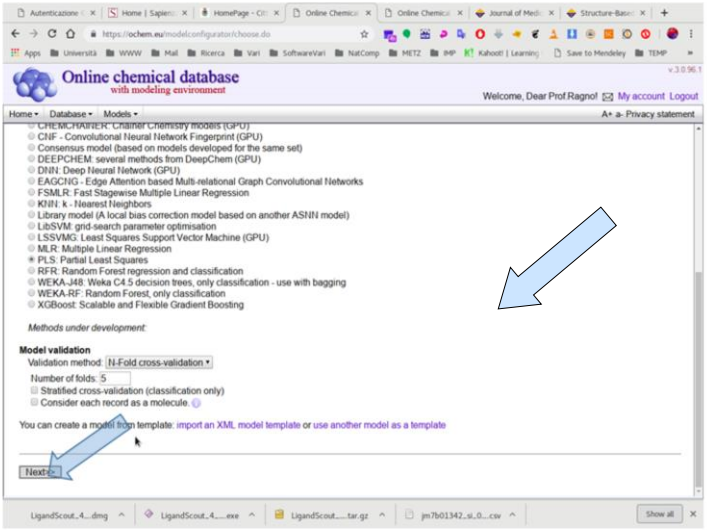
Preparing the dataset



Pagina 70

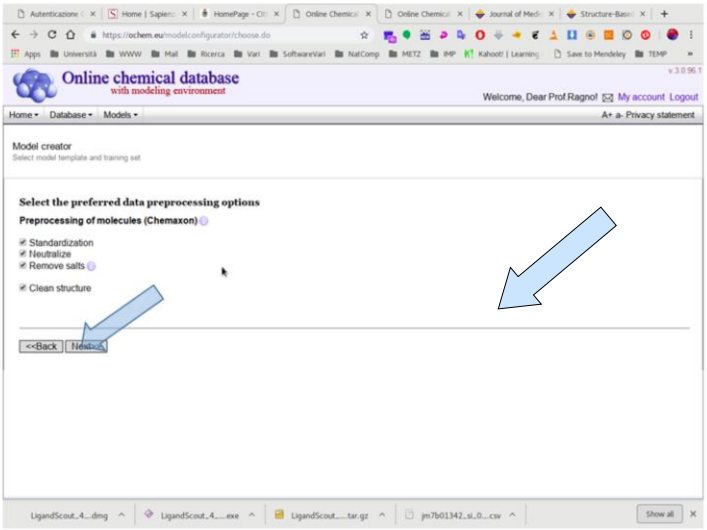
Everything should be already set, but CHECK IT!

Preparing the dataset



Click on “Next>>”

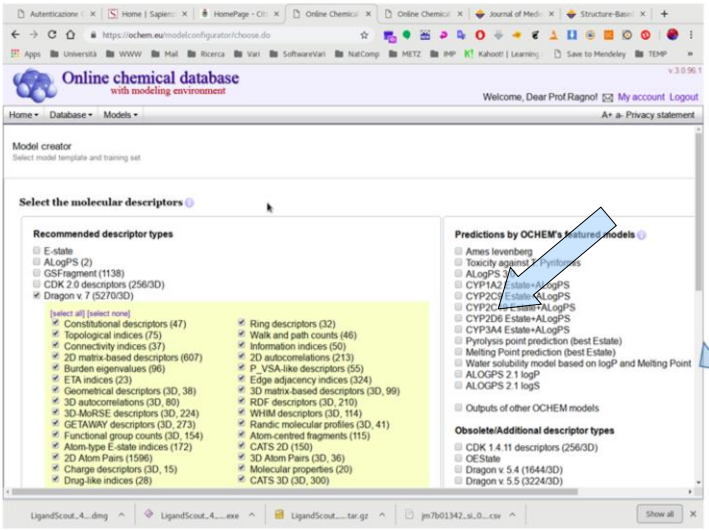
Preparing the dataset



Pagina 72

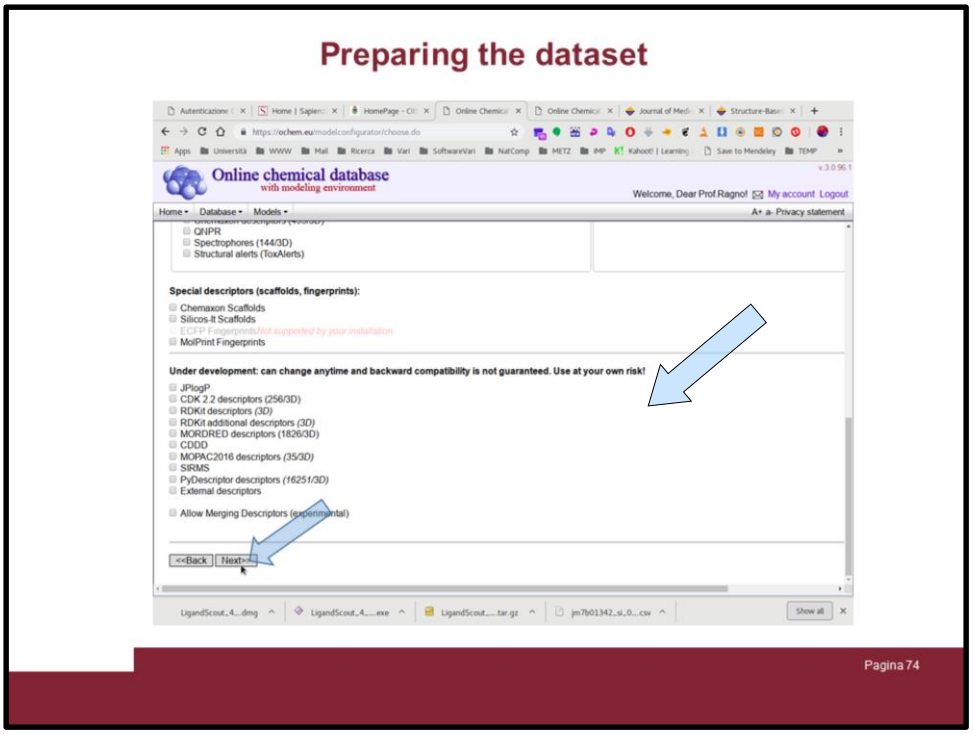
Click on “Next>>”

Preparing the dataset



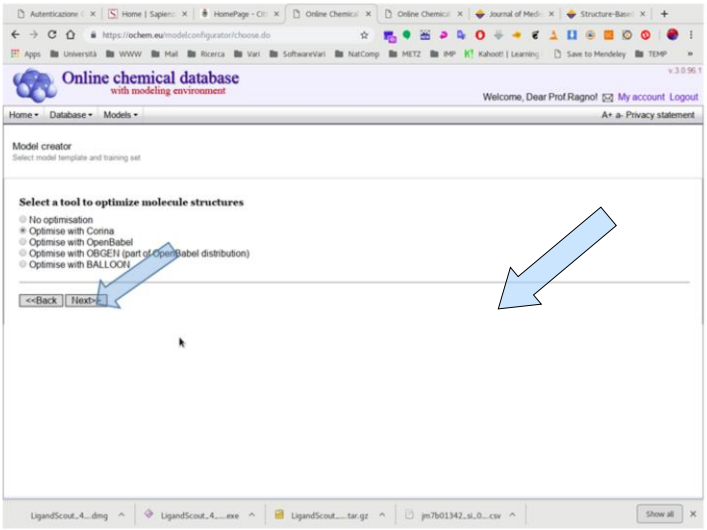
Pagina 73

Click on “Next>>”



Click on “Next>>”

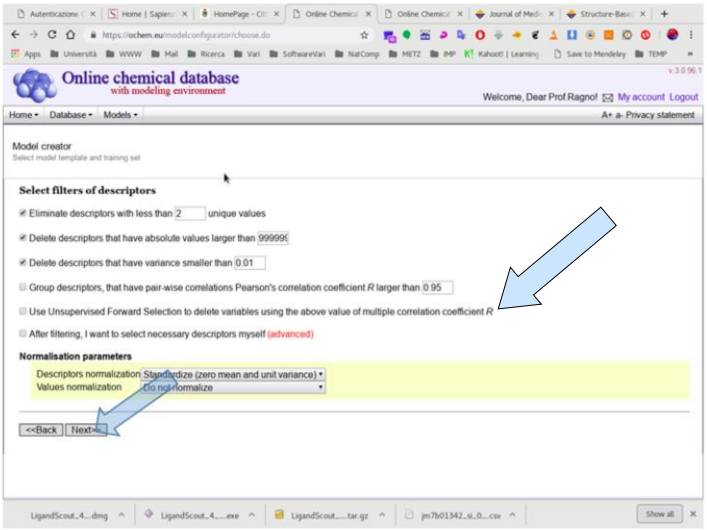
Preparing the dataset



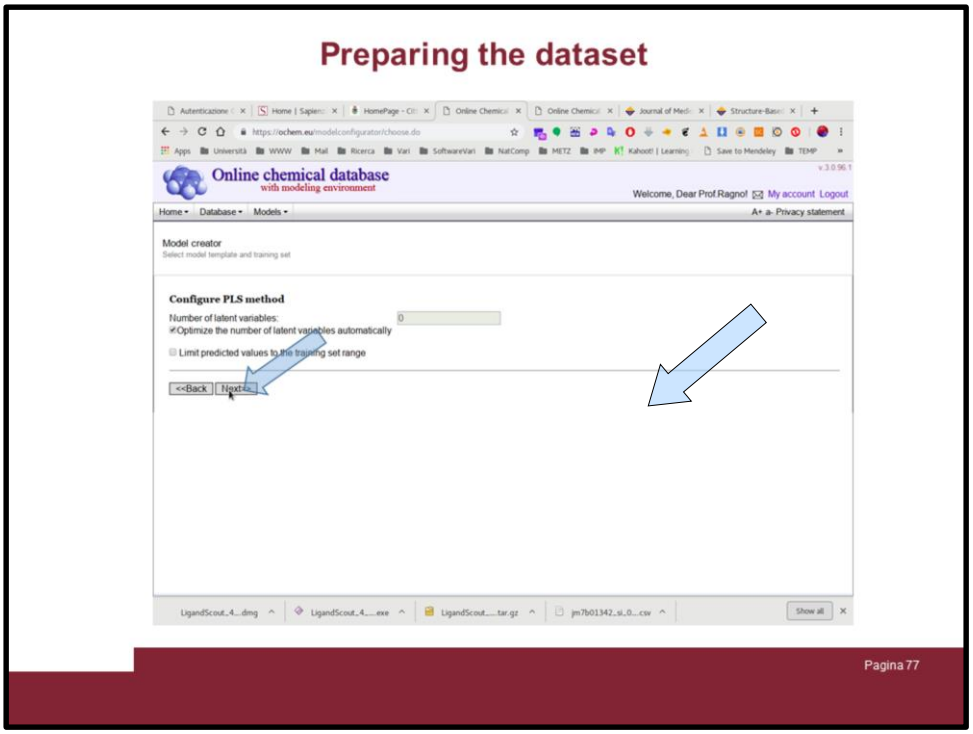
Pagina 75

Click on “Next>>”

Preparing the dataset

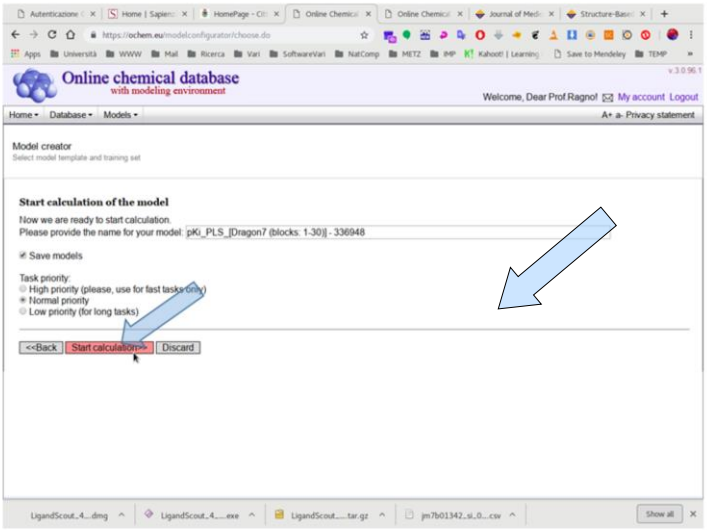


Click on “Next>>”

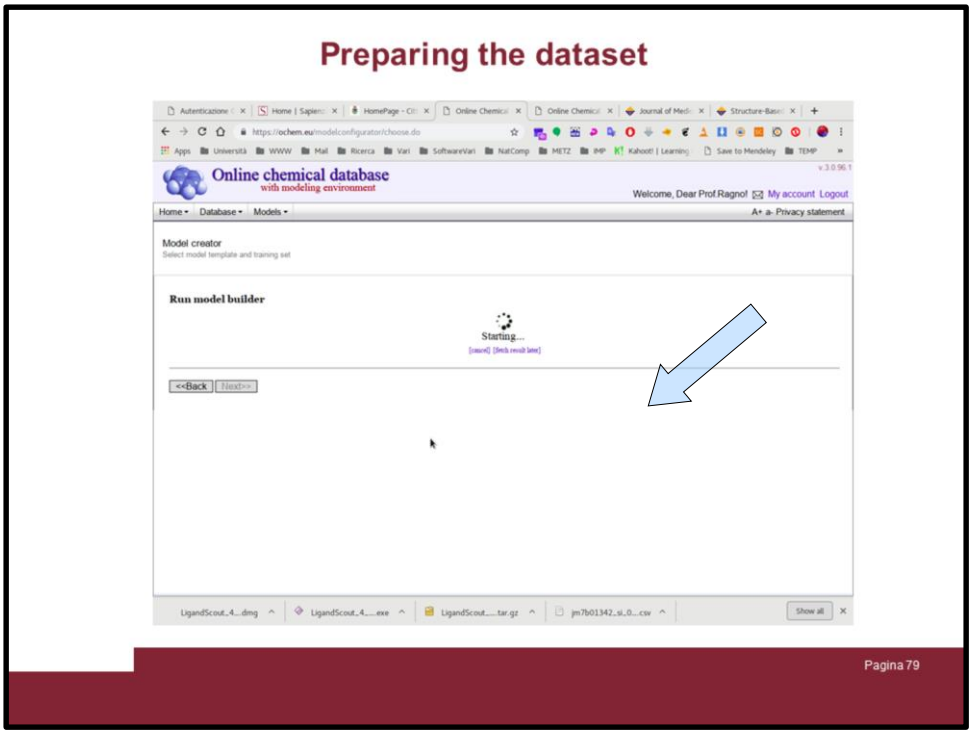


Click on “Next>>”

Preparing the dataset

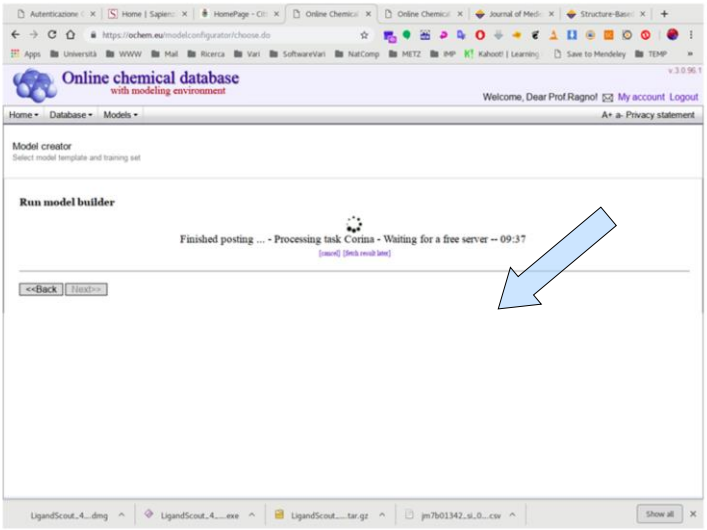


Click on “Next>>”



Wait for the model

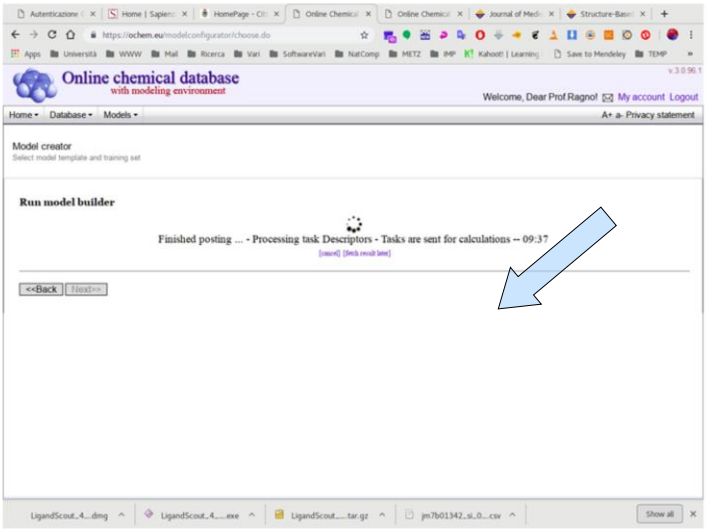
Preparing the dataset



Pagina 80

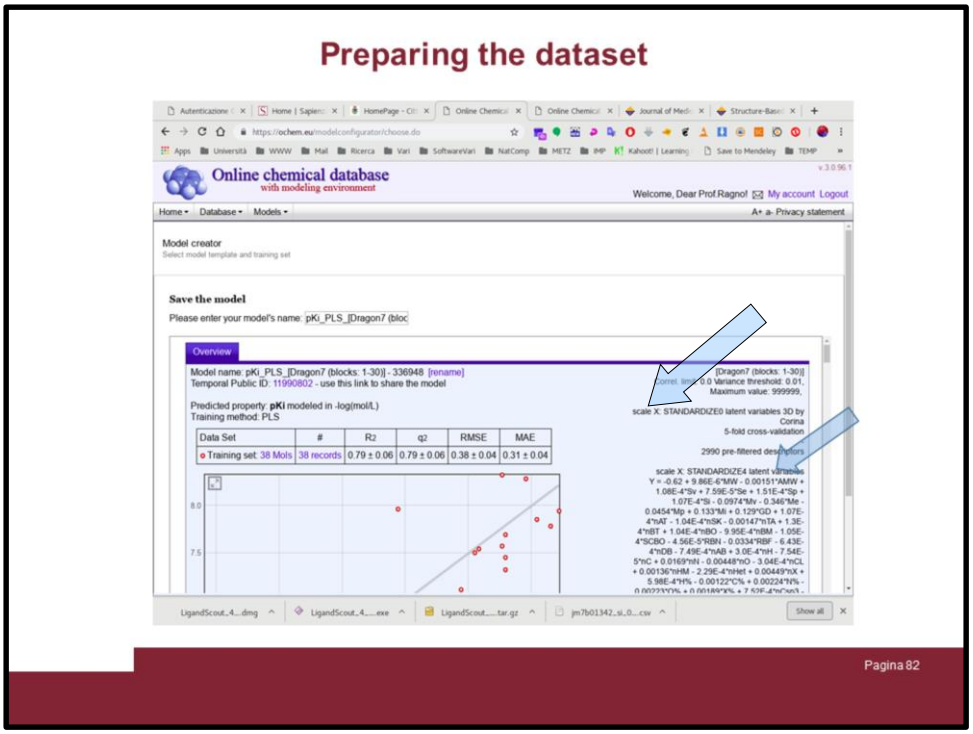
Wait for the model

Preparing the dataset



Pagina 81

Wait for the model

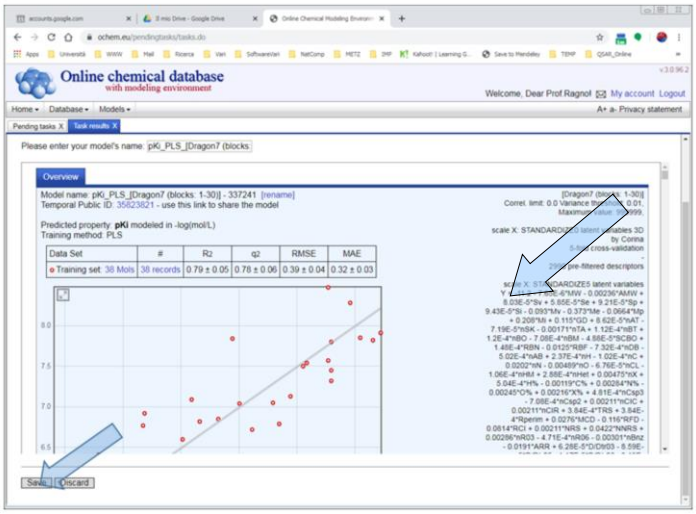


Once the model is ready you can have the results.

In the right you have the QSAR equation and on the bottom (not shown) you can download everything in an excel file and inspect it as youlike.

Slide down and save the model!

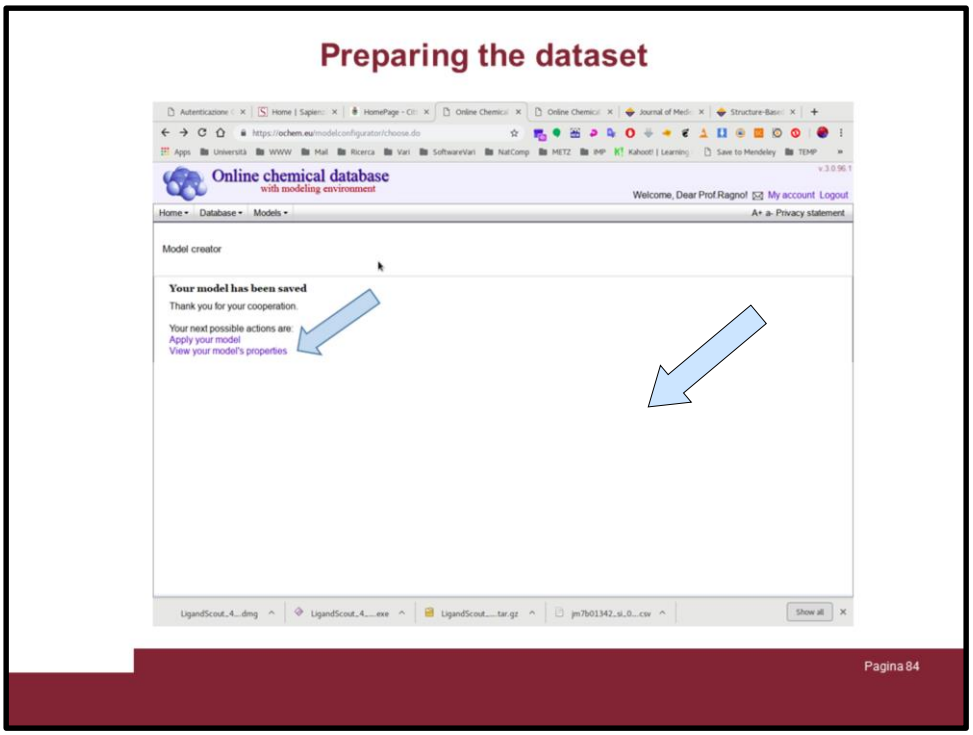
Preparing the dataset



Once the model is ready you can have the results.

In the right you have the QSAR equation and on the bottom (not shown) you can download everything in an excel file and inspect it as youlike.

Slide down and save the model!

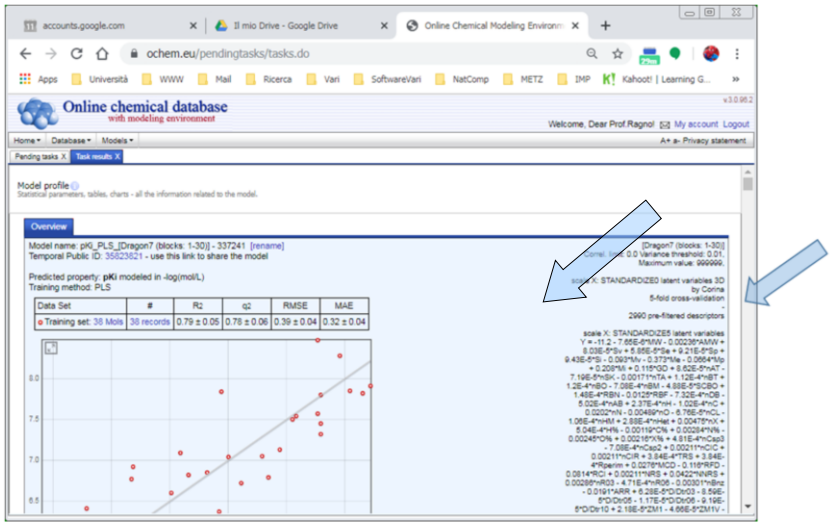


You have created a QSAR model by means of PLS.

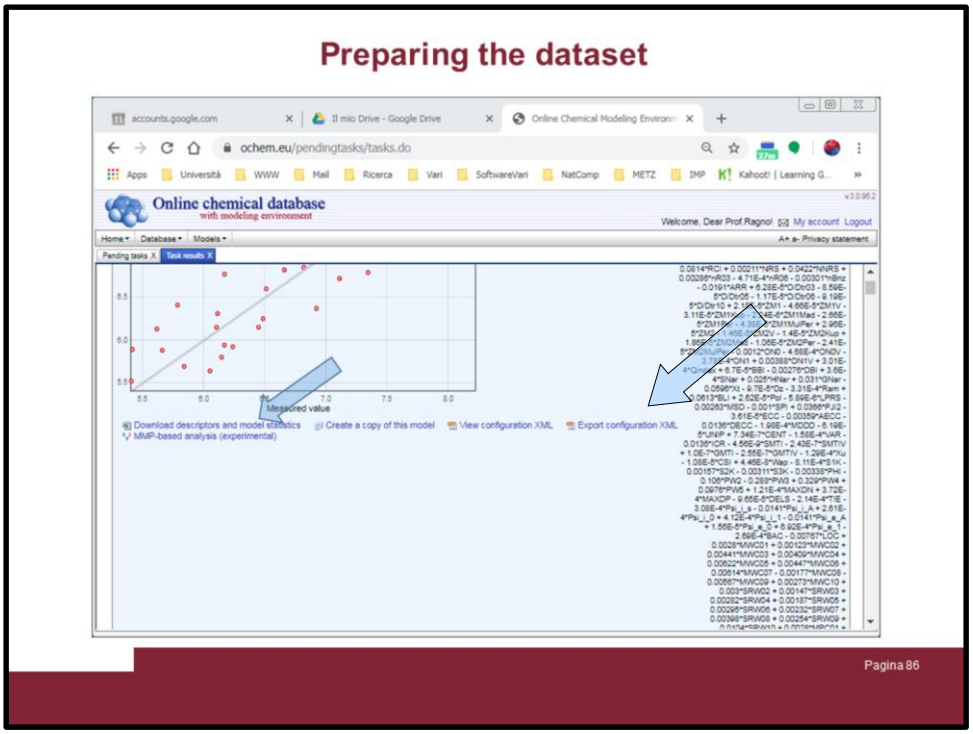
You can go back to your model later on to check the details.

Click on “View your model’s properties”

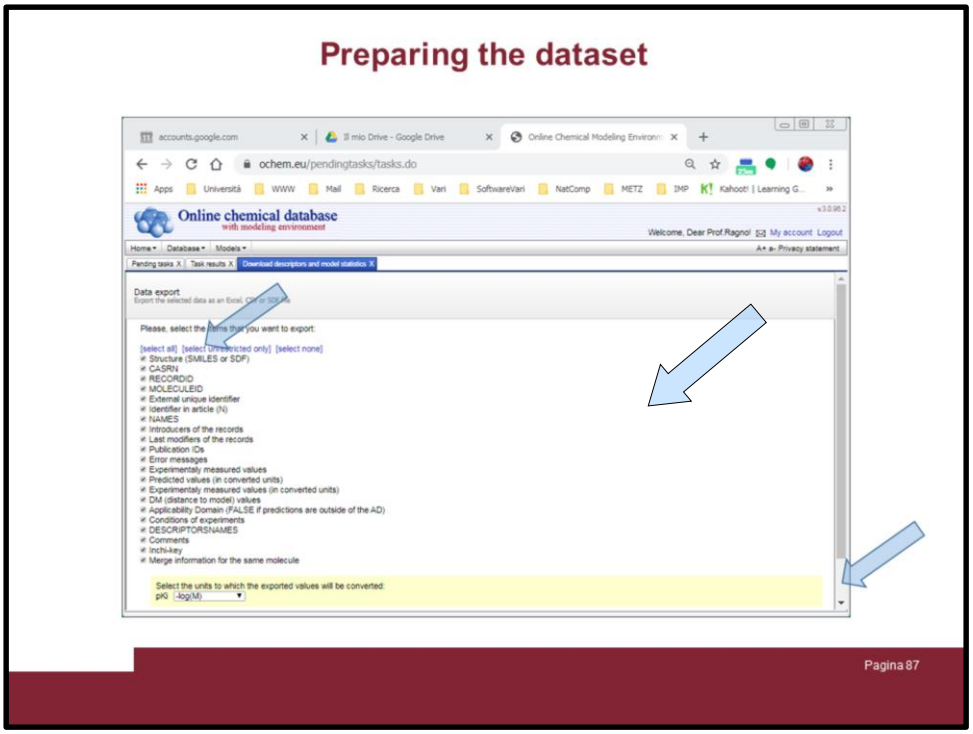
Preparing the dataset



The model again will be shown and slide down to look for a excel file to be downloaded



The model again will be shown and slide down to look for a excel file to be downloaded

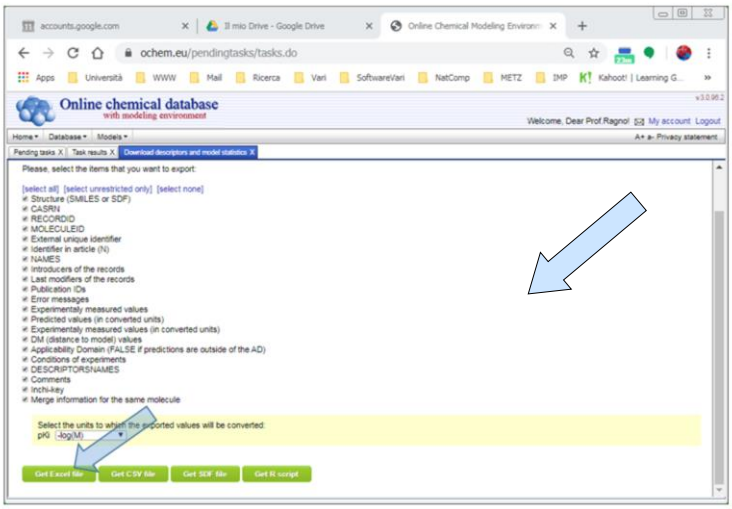


The model again will be shown.

The portal will ask what you want to save, just select everything

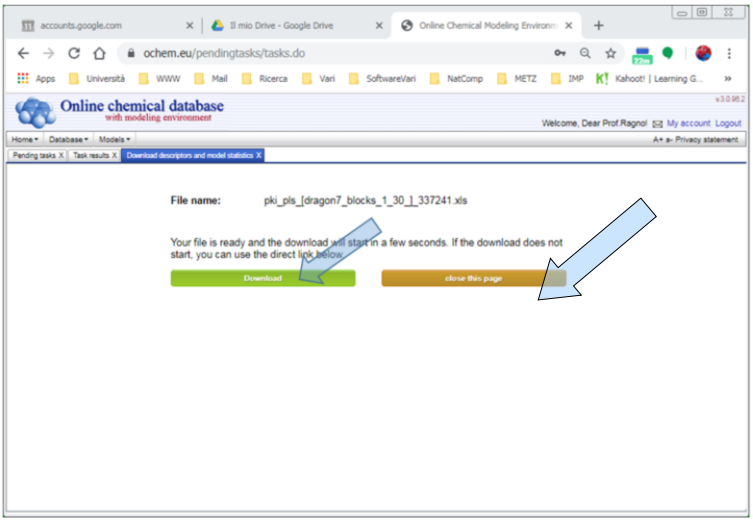
Slide down to download the file

Preparing the dataset



Click on “Get Excel File”

Preparing the dataset



Then click on “Download” and you can open it with excel or libreoffice to inspect the results