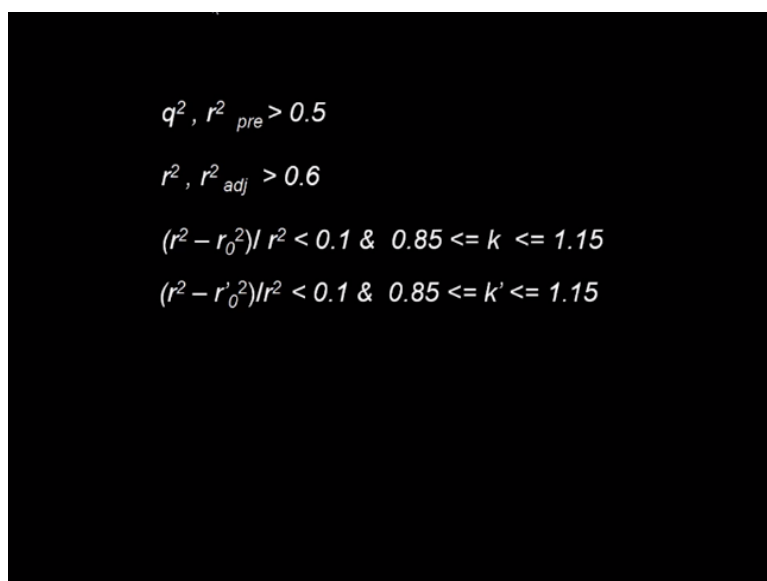


Computer Aided Drug Design
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology - Madras

Lecture - 30
Quantitative Structure Activity Relationship (QSAR)

Hello everyone, welcome to the course on computer-aided drug design. We will continue on the topic of QSAR. The previous class we talked about what are the important statistics we need to obtain in order to say the regression is good and it has got good predictability.

(Refer Slide Time: 00:35)



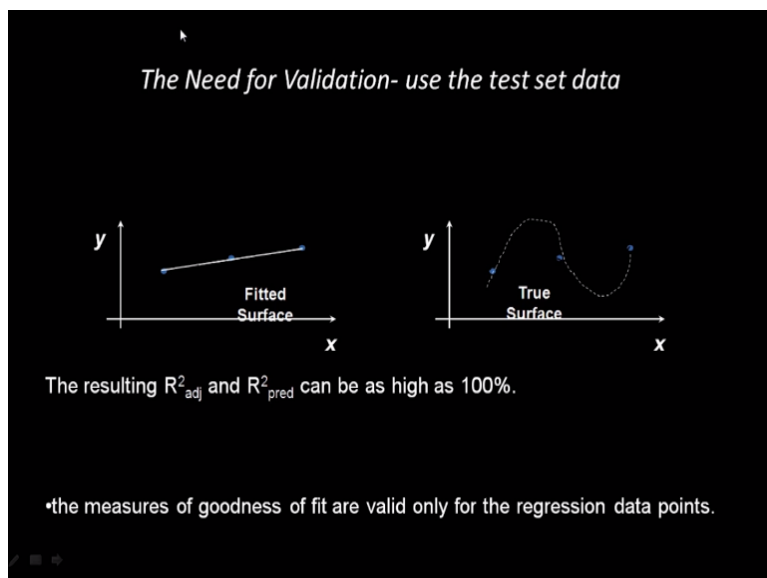
$q^2, r^2_{pre} > 0.5$
 $r^2, r^2_{adj} > 0.6$
 $(r^2 - r_0^2)/r^2 < 0.1 \ \& \ 0.85 \leq k \leq 1.15$
 $(r^2 - r'_0{}^2)/r^2 < 0.1 \ \& \ 0.85 \leq k' \leq 1.15$

R square, r square adjusted should be at least >0.6 , q square that is cross validated r square or leave one out method r square like that, r square predicted they should be >0.5 . Then r square- r_0 square that means when you force it to pass the regression line through the origin we call that r_0 square/r square should be <0.1 okay and then when you force it to pass through the origin, the slope should be almost 1 that is in this region.

And then we do the other way, instead of $y=$ we do $x=$ okay then we use again r dash 0 square make it pass through the origin and the corresponding slope is given by k dash that also should be 1. It is quite stringent. We need to have lot of conditions to be satisfied before we say that the regression model is good okay. It is very, very important that you have a validation that is the test set.

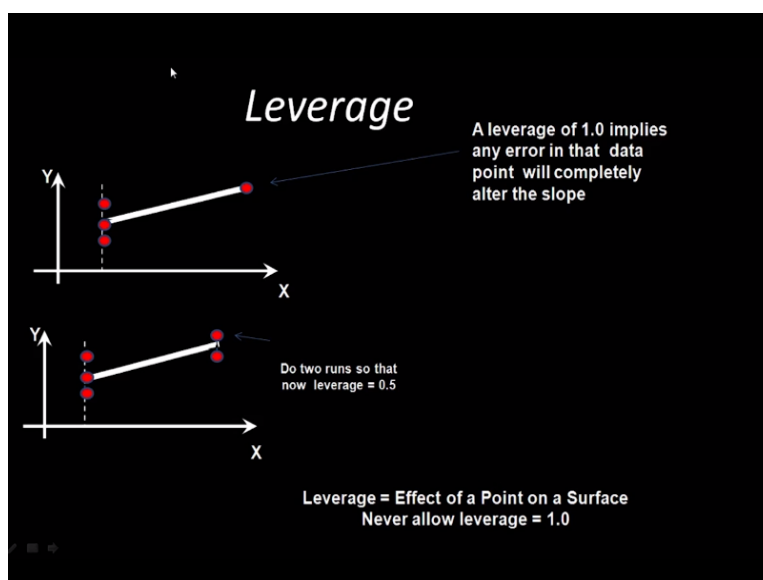
Because you may fit 3 points like this but in reality the data may be going like this okay. So there could be going up or going down.

(Refer Slide Time: 01:55)



So it is very, very important that we fit actually. Your r square adjusted, r square predicted or cross validated for the training set may be very high but when you do for the validation the whole thing could be in a big surprise. So you need to always have a validation or a test set okay. There is something called leverage okay. Imagine that this is your descriptor x and this is your activity y.

(Refer Slide Time: 02:28)



You are trying to get few data points and you are plotting between the x and the y and for most of the compounds the x that means the descriptor values almost same except for one compound, which is far away okay. This type of fitting is not very good because your entire

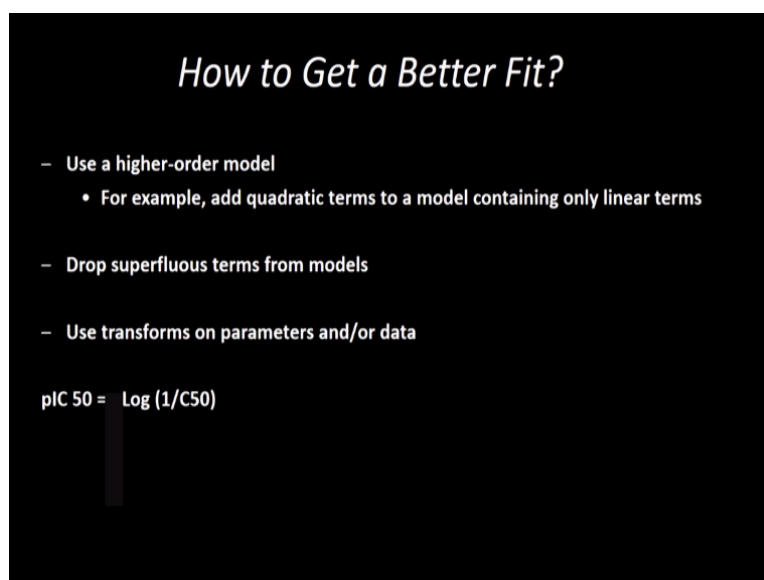
slope will depend upon this particular point okay. This is called leverage of 1 which is not very good.

That means you have all the data points, the descriptor value is almost same and only one has something different. So never have your QSAR like this. You should have at least one or 2 points here and here like this you know that means the descriptor values for the compound should be sort of distributed along rather than all of them clustered at one place and only one of them far away.

Because that could be leveraging your entire slope and errors in this particular point could completely affect your QSAR. So always select a descriptor, which is distributed in certain range okay if you are trying to develop a QSAR that is another important point actually okay. That is another important point in addition to other points, which we talked about in selecting descriptors okay.

So how to get a better fit, you can have a higher-order model, quadratic terms in the model, drop superfluous terms, use transforms for example what do we do?

(Refer Slide Time: 04:03)



How to Get a Better Fit?

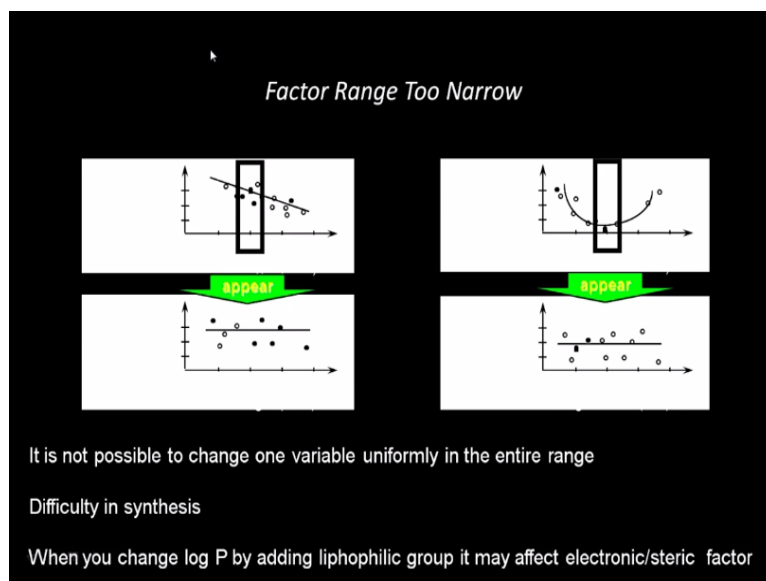
- Use a higher-order model
 - For example, add quadratic terms to a model containing only linear terms
- Drop superfluous terms from models
- Use transforms on parameters and/or data

pIC 50 = $\text{Log}(1/C_{50})$

Suppose we measure the concentration at which 50% of the cells die if you are doing an anti-cancer activity. If you are doing antibacterial concentration at which 50% or 90% of the bacteria die. We do not use that Log of 1/C50 is pIC 50 okay. This is called transformation because what happens is these concentrations could be in micromole or millimole so 10 power -4, 10 power -3, all those terms come in.

But when you take logarithm and put a – in front of that they will be all positive numbers. So the regression will be better okay. So never use your dependent variable in the order of 10 power -4, 10 power -3 and so on but try to take logarithm so that they will be in whole number okay. Then, another important point factors range too narrow like suppose you take a set of components and the descriptor values are almost same.

(Refer Slide Time: 05:18)



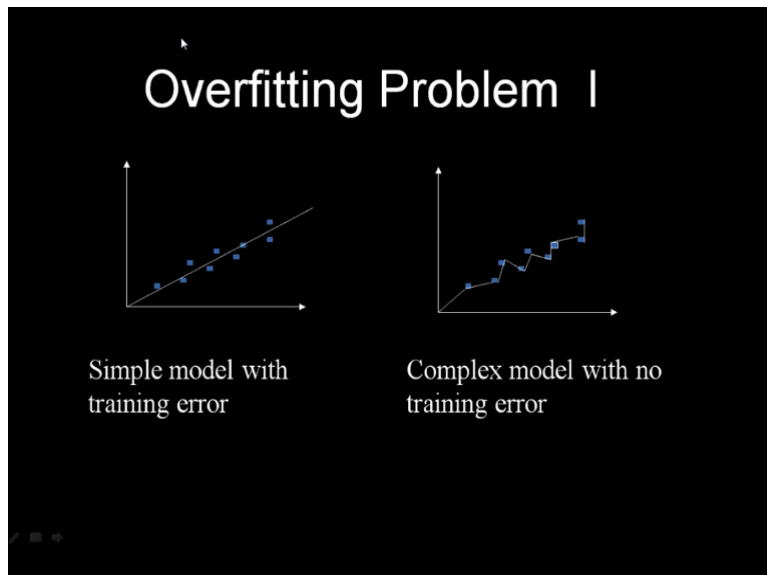
And then when you try to fit, you may feel that there is no difference in the activity, the activity does not change with the descriptor but in reality it may be changing like this but because your range is so small that they all may appear as if there is no effect on the activity okay. So it may be going like this but range so when you select the descriptors you need to see whether they are covering a reasonably good range or they are narrow all clustered together because that may give leverage that may give this type of errors actually okay.

Of course, when we are talking about synthetic chemistry it may be difficult to synthesize so that a particular descriptor changes in over a wide range. For example, I want to change Log P, I may think of putting OCH₃, CH₃, fluorine, chlorine so that Log P changes and so that is not bad but it may be difficult to synthesize those type of derivatives you know that may be a problem.

Another point is when you are changing Log P, it may affect electronic or steric factors, we saw in the previous, previous classes. So it may be not completely possible to just change one particular property without affecting other property. So these are some challenges when you

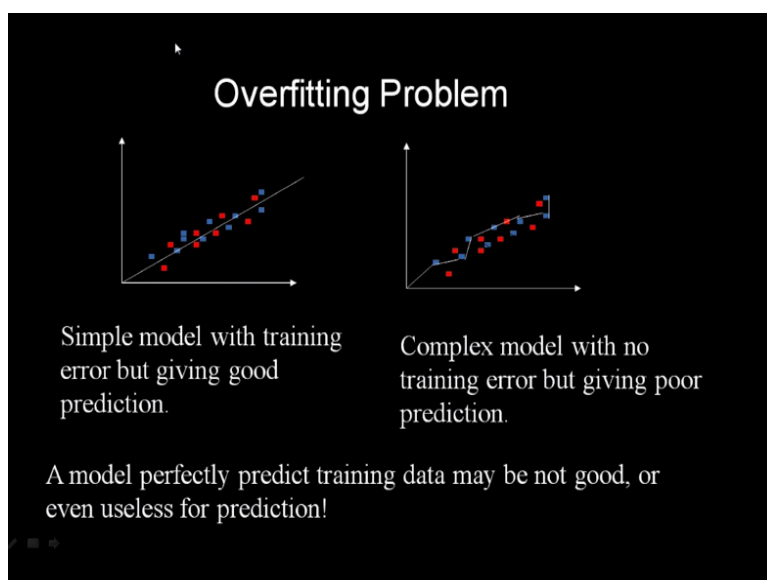
do QSAR. Overfitting okay, so the data points is like this, it is a simple model for fitting but then you can have very complex model try to fit each and every point that is called an overfitting, you do not need to have that okay.

(Refer Slide Time: 07:00)



It is good enough to have like this. This is called an overfitting problem. So when you try to fit simple model with training error but giving good prediction, complex model very complex, is going up and down, up and down, trying to fit each and every data point but it may give a very poor prediction.

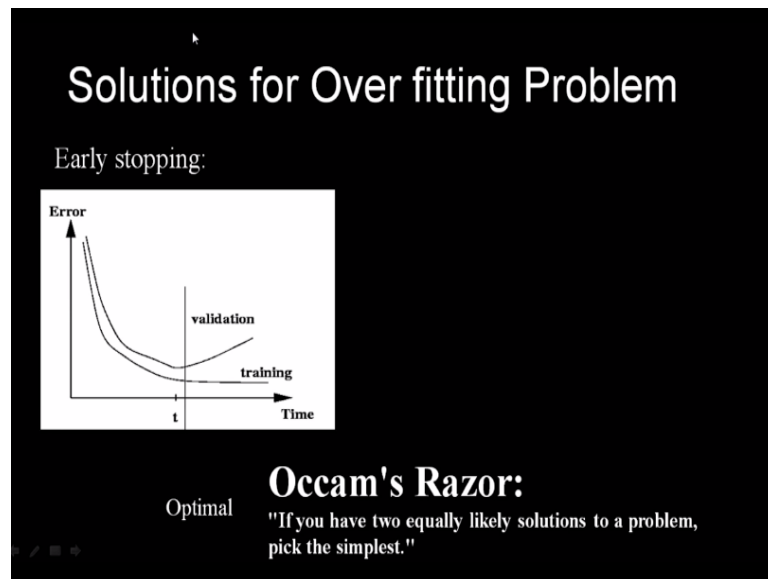
(Refer Slide Time: 07:25)



A model perfectly predict training data may be not good or even useless for prediction okay. So you have to be very careful about overfitting problem. There are many problems you

know there is something called Occam's razor. If you have 2 equally likely solutions to a problem, pick the simplest okay so always pick the simplest, do not go for very complicated.

(Refer Slide Time: 07:54)



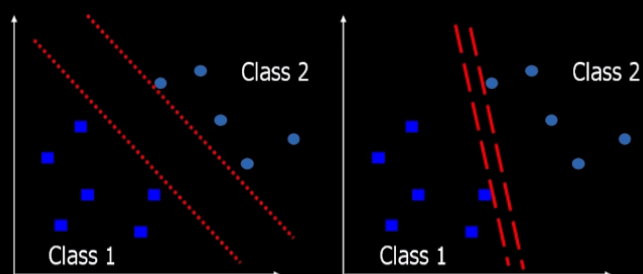
If you get almost same answers with 2 different say linear model and a quadratic model you get r^2 and r^2 adjusted q^2 almost same, little difference, always go for the linear model. Do not have to go for a quadratic model. If you have 2 descriptor model and a single descriptor model, there is only small change in the q^2 or the validation r^2 , go for the one descriptor model that is what this Occam's razor tells you.

If you have 2 equally likely solutions to problem, pick the simplest okay likely they are very small difference but if you see big difference yes you may go for a complex model but if it is a small difference do not do that.

(Refer Slide Time: 08:41)

- For a very simple two-class binary classification problem, one can have infinite many decision boundaries to separate these two classes

- Which one should we choose?



Look at this, for a very simple 2-class binary classification problem okay you have some data, data, data you want to separate these data into 2 different groups okay. You can draw the line like this so this becomes one group or you can draw the line like this then these groups change. So this could be low active compound, this could be high active compound. So where do you draw the line?

How do you differentiate okay between the low and the high active? You could have one QSAR for low active compound and one QSAR for high active compound. So depending upon how do you classify, some of the data points may go here or some of the data points may be going into the other group, so that is the problem.

(Refer Slide Time: 09:27)

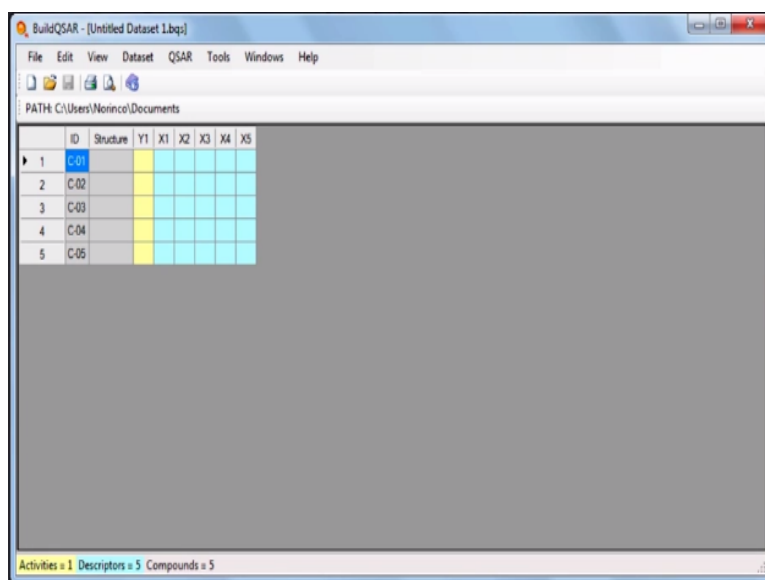
BuildQSAR

<http://www.profanderson.net/files/buildqsar.php>

spreadsheet, in which the user can enter with the data set composed by the structure definition of the compounds, one or more types of biological activity values and many physicochemical properties.

Let us look at this particular software BuildQSAR. It is a spreadsheet type. The user can enter the data set composed by the structure definition of the compounds. So it is quite good and we will look at it that is called BuildQSAR. Yeah this is the software BuildQSAR okay. So as you can see this is your activity y and here it is giving 5 descriptors, we can keep increasing the descriptors also no problem.

(Refer Slide Time: 10:08)



The screenshot shows the BuildQSAR software interface. The window title is "BuildQSAR - [Untitled Dataset 1.lbp]". The menu bar includes File, Edit, View, Dataset, QSAR, Tools, Windows, and Help. The file path is "PATH: C:\Users\Noninco\Documents". The main area is a spreadsheet with the following data:

	ID	Structure	Y1	X1	X2	X3	X4	X5
▶	1	C-01						
	2	C-02						
	3	C-03						
	4	C-04						
	5	C-05						

At the bottom of the window, it says "Activities = 1 Descriptors = 5 Compounds = 5".

See you can add okay compounds or you can add variables, add items okay so you can remove items okay so you can remove items, compounds, all those things we can do and so we can give some title for this for compounds then we can have many descriptor data results here and then try to develop regression relationship. So as I said if you have only 5 data points you can have only one descriptor model.

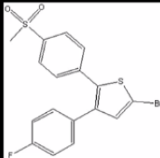
If you have 10 data points may be you can think of 2 descriptor model. So if you have some data for example let us look at some data. Let us look at an example. These are called anti-inflammatory compounds. These are anti-inflammatory drugs. They are also called as selective cyclooxygenase drugs.

(Refer Slide Time: 11:05)

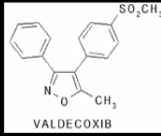
Anti inflammatory drugs

* pIC₅₀ COX-2 inhibitory values in μM

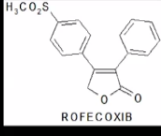
S.No	Compound Name	pIC ₅₀ *
1	Celecoxib	1.1549
2	Rofecoxib	0.301
3	Valdecoxib	0.7375
4	Nimesulide	-0.1139
5	DUP697	1.2218



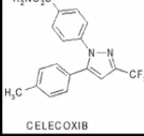
DuP-697



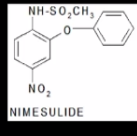
Valdecoxib
Bextra
(Pfizer)



Rofecoxib
Vioxx
(Merck)



Celecoxib
Celebrex
(Pfizer)

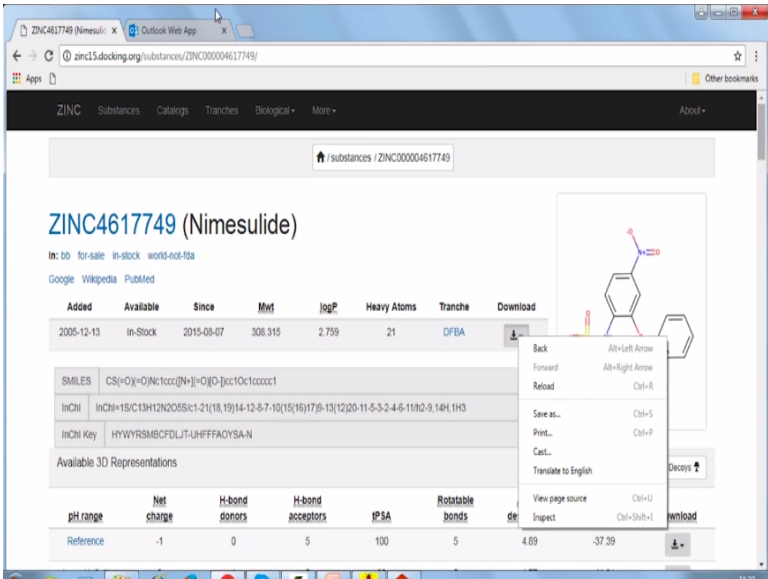


Nimesulide

Cyclooxygenase comes in the arachidonic pathway of the inflammation and I think I talked about it long time back okay and these are marketed by companies like Pfizer, Merck and so on. Bextra, Vioxx, Celebrex, DuP, these are all very selective whereas Nimesulide is not a selective drug and it goes and binds to cyclooxygenase-2, cyclooxygenase-1 and many other enzymes in the pathway and the activity I know I took it up from the literature okay.

So now I have these compounds, I want to develop a QSAR so I can get the descriptors for each one of these compounds okay and then I can get descriptors for each one of these compounds using E-DRAGON okay and then I can look at which descriptors have good correlation, highest correlation with this and then select those descriptors and put it in the BuildQSAR and then try to develop a QSAR model okay.

(Refer Slide Time: 12:45)



ZINC4617749 (Nimesulide)

In: [db](#) [for-sale](#) [in-stock](#) [world-not-fda](#)

Google [Wikipedia](#) [PubMed](#)

Added	Available	Since	Mwt	JogP	Heavy Atoms	Tranche	Download
2005-12-13	In-Stock	2015-08-07	308.315	2.759	21	DFBA	

SMILES: CS(=O)C(=O)Nc1ccc(NC(=O)O)cc1

InChI: InChI=1S/C13H12N2O5S1c1-2(1,18,19)14-12-8-7-10(15,16)17/9-13(12,20-11-5-3-2-4-6-11)02-9,14H,1H3

InChI Key: HYWYRSMBCFDLIT-UHFFFAOYSA-N

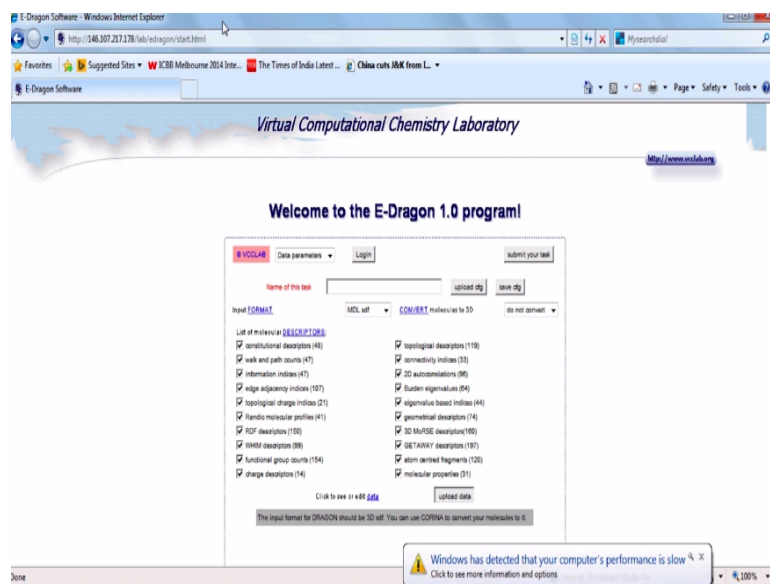
Available 3D Representations

pH range	Net charge	H-bond donors	H-bond acceptors	TPSA	Rotatable bonds	TPSA	TPSA
Reference	-1	0	5	100	5	4.69	-37.99

So first we can use if you remember Zinc database Nimesulide okay, this is Nimesulide, then I can download the SDF file here okay, which already have it okay so Nimesulide SDF file is there okay. Then, if you want to convert into SDF also this Open Babel is there. Please remember if hydrogen's are not there, you must add hydrogen, make explicit because Zinc might not put hydrogen as you can see here explicitly.

So you can use Open Babel and convert that SDF file. For example, yeah Nimesulide open, yeah hydrogen is already there but if you do not have hydrogen then we can add hydrogen and then create another file. Okay once you do that we have the SDF file and as I mentioned E-Dragon requires SDF file okay. So we can load Nimesulide okay.

(Refer Slide Time: 14:56)



So here we upload data then we upload the Nimesulide SDF file okay, browse then we upload, yeah CADD QSAR and upload Nimesulide. Yeah Nimesulide gets uploaded and then we run E-Dragon, get all the results like I showed you in the previous class okay, successfully loaded, now I say submit your task here okay and then you get all the results okay, lot of descriptors.

And you can do the same thing for DuP-697, Valdecoxib, Rofecoxib, Celecoxib okay and then create an excel file with all the descriptors okay. So with all the descriptors as you can see here I created a file DuP-697, Valdecoxib, Rofecoxib, Celecoxib, Nimesulide and these are the activities. Nimesulide has the lowest activity here okay, so which I have put in here okay.

(Refer Slide Time: 16:18)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
12 No.		DUP697	Valde	Cele	Rofe	Nim														
13 pIC50		1.22	0.7375	1.15	0.301	-0.11														
14 MW		411.33	314.39	381.41	314.38	307.33	0.861605													411.33
15 AMW		11.75	8.73	9.54	8.73	9.6	0.507307													314.39
16 Sv		25.58	24.2	26.61	24.32	21.32	0.917243													381.41
17 Se		35.73	36.56	41.76	36.57	33.39	0.652283													314.38
18 Sp		27.82	25.59	27.72	25.8	22.36	0.938064													307.33
19 Ss		59.17	55.42	75.17	56.92	62.08	0.370935													
20 Mv		0.73	0.67	0.67	0.68	0.67	0.508661													
21 Me		1.02	1.02	1.04	1.02	1.04	-0.22531													
22 Mp		0.79	0.71	0.69	0.72	0.7	0.486165													
23 Ms		2.57	2.52	3.04	2.59	2.96	-0.14628													
24 nAT		35	36	40	36	32	0.690963													
25 nSK		23	22	25	22	21	0.755729													

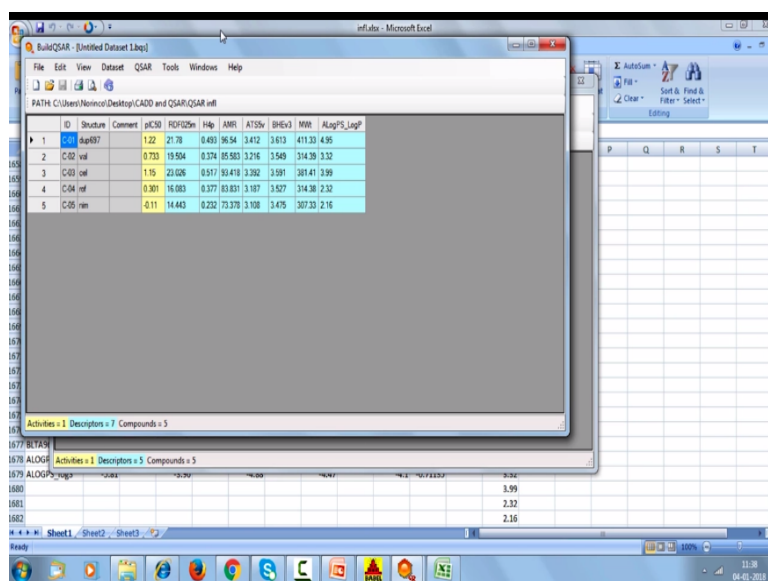
And pIC50 I have put and these are the descriptors, so many descriptors and then I tried to find out the correlation between activity as you can see correlation between activity and the descriptor. So you can see molecular weight point, it has got a correlation coefficient of 0.86 and Sp another descriptor. So as I said E-DRAGON can calculate almost 2000 descriptors okay.

So molecular weight I have taken here. These are the 5 molecular weights for these 5 compounds okay. Then, I looked at some other descriptors here, yeah this is another descriptor called PIPC05, we want to know more details about it, we need to look into E-DRAGON to know what these descriptors are okay. These are the values for these 5 compounds like that you know.

Then, I have another one, so I picked up some of them which has got very good correlation 0.988 BeHe3 that is another descriptor like that you know go down and down and down and down. Then ALOGPS logP 4.95 so I took these and then I went to here and then I selected 5 descriptors and then I put the activities of these 5 compounds okay. Let me see whether the file is there which I had kept.

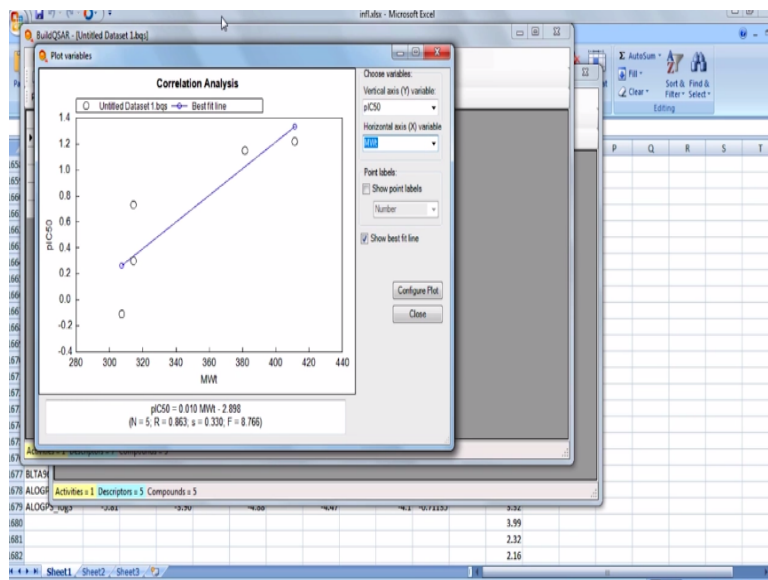
So IC50 I have it here, so I took 3, 4, 5, 6, 7 descriptors right. So this is molecular weight of this DuP-697, Nimesulide, pIC 50. Then these are descriptors. I selected the descriptors based on the correlation coefficient, which had reasonably good correlation. Of course, you can select more if you have more. Of course, here I will be able to create only one descriptor model because I have only 5 data points remember that.

(Refer Slide Time: 18:50)



So what do we do? We can say QSAR, we can say variable selection, systematic search, here there is no 2 variable problems so we did plot variables okay. So we can see pIC 50 versus RDF025. This is $R=0.98$ then I can say okay this is with respect to H4P another descriptor again it gives you $R=0.95$. Its molecular weight is not very good 0.863, then this is ALOGPS logP 0.86 okay. So we can have AMR another descriptor is quite good 0.975 okay good.

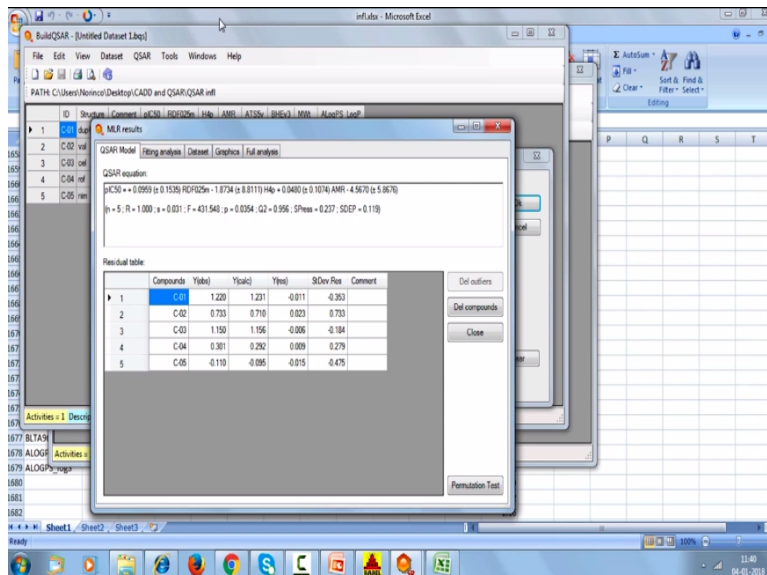
(Refer Slide Time: 19:45)



We can look at them, then we can do tools, options okay then we start doing a QSAR, biological activity so we can do RDF. So you see this is the mathematical regression relation, pIC 50=this RDF this gives you the + or - 95% confidence - so $y=mx+c$ type okay. In fact, it is trying to create a model with 3 parameters okay. So you have 4 parameters, 5 data points, so degrees of freedom are only 1.

It is not very good model okay so as you can see this is a QSAR, it gives you $R=1$ so obviously degrees of freedom are very poor okay.

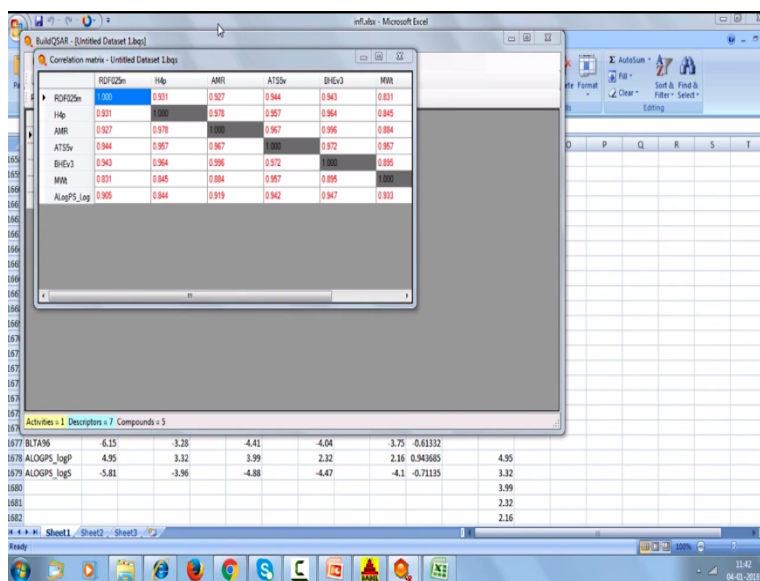
(Refer Slide Time: 21:26)



So we can take a 2 descriptors at a time may be so obviously now it is better because we have 1, 2, 3 so degrees of freedom are 2. We can see q^2 is 0.5 here, F value these are statistics which you need to understand if you attend a course of mine called biostatistics and design of experiments. You will come to know many of these things. You can look at correlation matrix.

We can look at is there a correlation between these descriptors okay. As I said the descriptor should not be cross correlated, which is bad then okay. See some of the descriptors are quite highly correlated okay.

(Refer Slide Time: 22:29)



So if you are developing a 2 parameter model, you should not select those descriptors which are correlated amongst themselves that is cross correlated, which is not good okay remember that okay. So we can have again if you take each descriptor one at a time, it gives you what is the q square, what is the r square so we can select that one which looks good. This AMR looks like a good model okay.

So you get a good q square here and you get a good r so this is one parameter model with AMR is good. So you need to do little bit of changes, modifications okay before you arrive at the best model. So we can add columns that means we can add some more descriptors, we can add some more data points. So many things we can do, we can even scale data for example if you think you need to take logarithm or you need to take square or square root.

Then, we can do that sort of things also so this software is extremely user-friendly. It works like excel, so it is very good. So what is the step-by-step procedure? So you can download your structure from Zinc, check whether hydrogen's are also there, otherwise you can use Open Babel and add hydrogen and then you can use the E-Dragon, SDF file and then get all the descriptors.

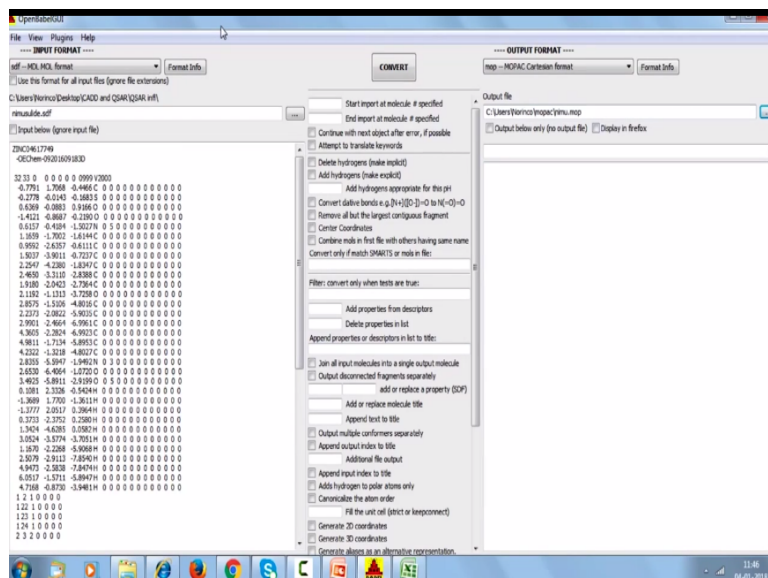
You can go to excel and then if you know the activity you can see which descriptors have good high correlation with activity. You can select those which have more than 0.9 or 0.95, select only them and then copy them into the BuildQSAR model like this. The activity is coming here and the descriptor value is coming here. Here I have shown only 5, you may have more.

And then start doing the QSAR with the different descriptors and check out your q square value and then see whether they are cross correlated, then look at one descriptor model. If you have more data points you can look at 2 descriptor model and so on. So you can do lot of things. This is a good software. I like it because it is very user-friendly because it looks like excel.

We can copy, paste from excel whatever you have okay from E-DRAGON okay. Now what about electronic descriptor? So if you want electronic descriptors what do we do? Of course, we need to do manually as you know MOPAC can do electronic descriptors right, calculations so okay. Let us take Nimesulide SDF, so we need to get the MOPAC file. here is there Nimesulide SDF is called MOPAC file.

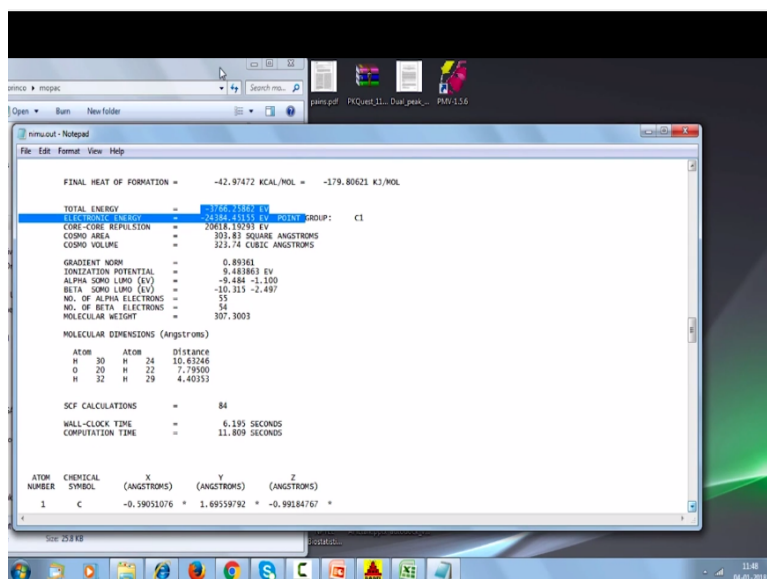
So going MOPAC, yeah you have MOPAC Cartesian coordinate okay output file, so we need to say yeah so I am going to put it here, I will say Nimesulide, I am going to save it here.

(Refer Slide Time: 27:14)



So this will be the MOPAC file for Nimesulide here okay. I can see from here okay. Now we will copy that and say convert, yeah MOPAC is there okay. Now we need to put in the file or something, save, so we need this one, we do not need this one okay. So we need to put it into this Nimesulide MOPAC yeah. Okay done Nimesulide out is here yeah this is Nimesulide, PM7 calculations have been done.

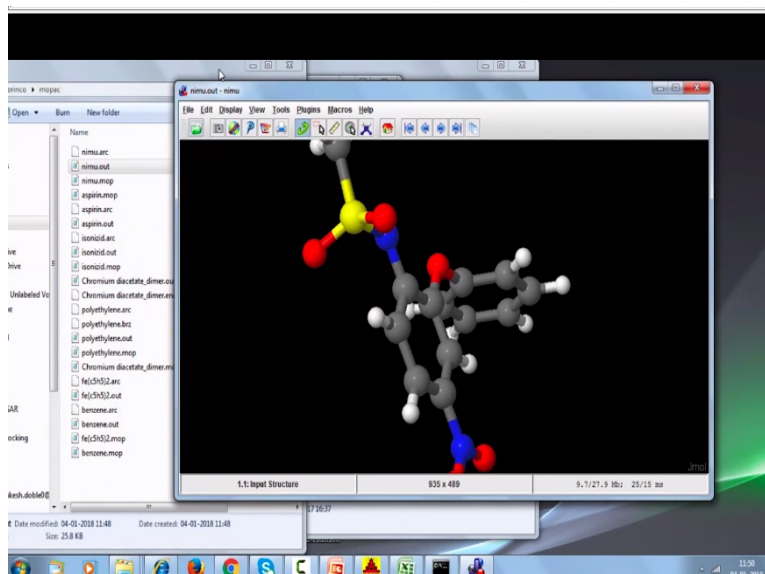
(Refer Slide Time: 29:00)



Final heat of formation kilo calories or it is given in kilo joules, total energy is available, electronic energy is there, core-core repulsion is there okay, ionization potential is there so these are electronic descriptors, which you will not get from E-DRAGON. So you can use MOPAC to collect for each of these 5 molecules anti-inflammatory drugs. We can collect it running MOPAC and then again take them to the QSAR software, which I showed you.

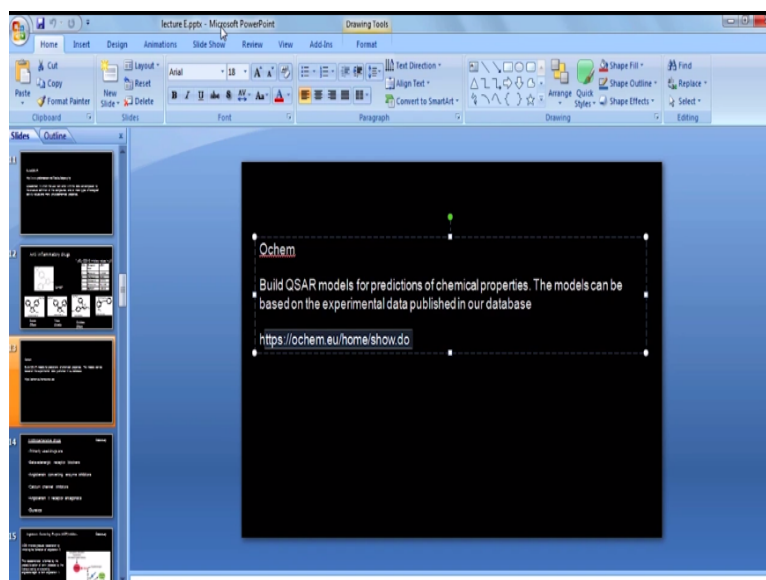
And then see whether any electronic descriptors have good correlate activity okay and I also showed you another software. If you want to look at the MOPAC output that is called the Jmol. So Nimesulide out we can put it inside and we will get the details about the structure of Nimesulide. So this is Nimesulide. So Jmol is a good software for viewing structures from MOPAC.

(Refer Slide Time: 30:10)



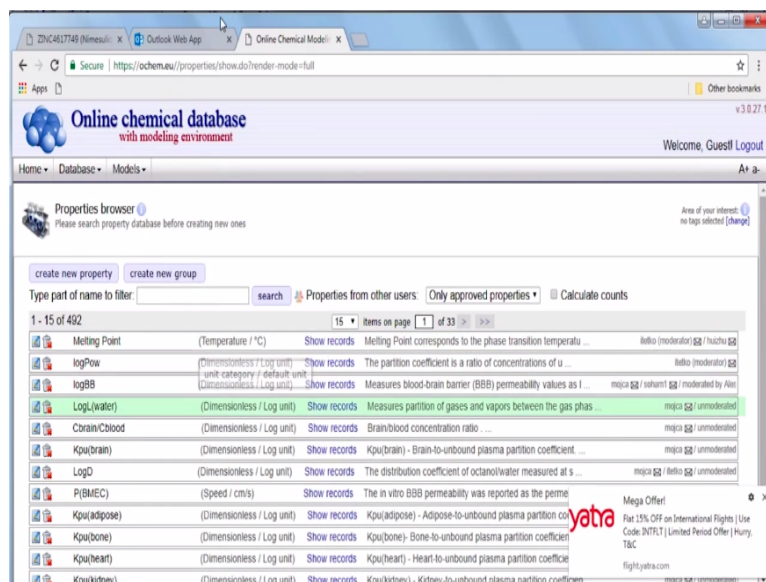
So we can get lot of details about the structural features and we can develop surfaces okay Van der Waals surface, solvent accessible surface using Jmol okay. Let us go back so this is how we can do. There is another software which is called the Ochem. This also is quite good. I would like to briefly show it to you. This also can look at descriptor calculations.

(Refer Slide Time: 30:44)



Yeah so this software also can give you, yeah it has got databases, properties. So we can get properties, log in as guest, so lot of properties are there okay.

(Refer Slide Time: 31:21)

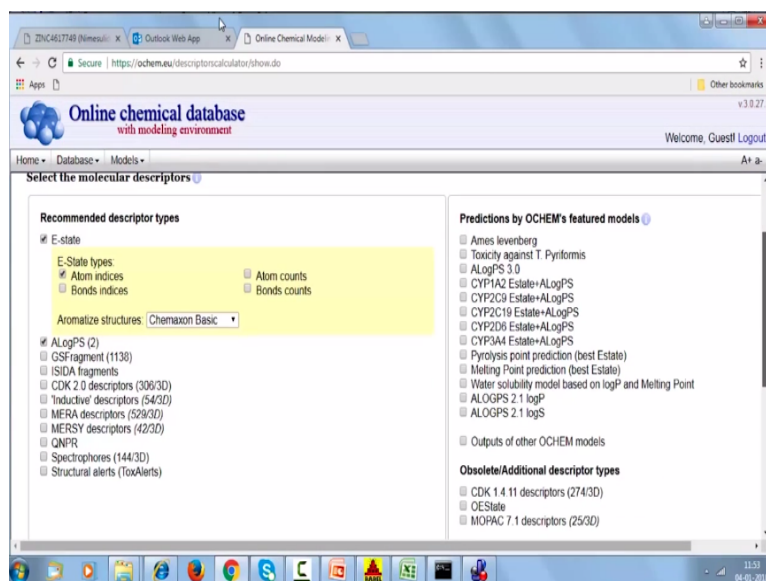


Properties you can see melting point, Log P for lot of compounds has been stored here okay. Suppose you want to know see it has got Lop P values for large number of compounds. Log P, solubility so many things okay. Now if I want to calculate for a particular model, calculate

descriptors. Yeah so I can draw the model or I can upload compound from SDF, choose file, so I can say aspirin SDF, open so it takes aspirin SDF okay.

Then we can do calculate aspirin SDF okay so lot of descriptors it can do as you can see Log P. So many descriptors are there 100s and 100s of descriptors okay. It can even give Tox alerts and so on actually.

(Refer Slide Time: 33:08)



So many descriptors it can do. So it is calculating these descriptors so I can click on many items. So it is calculating descriptor so this is also a good software for doing this type of job actually okay. Yeah so total molecules correctly calculated, you can see descriptors, see to the download page, so descriptors value. So QSAR as you can see has lot of advantages. Similarly, one can work on QSAR development for angiotensin, ACE inhibitors or beta blockers.

So all you we need is we need to have some data on the activity from literature for some of these commercial drugs and then we go about and we start developing okay. So it has given lot of results okay so many descriptors are given, so many descriptors. As you can see this talk about alerts whether there are any alerts on toxicity and so on actually okay. So that is also a very good software, the Ochem.

It can do descriptor calculations, in fact it also has a model building option, it has got the model building option because of one tough time, we will not go but we can upload a linear model. So I suggest you people have a look at these things also okay. So QSAR has

advantages and disadvantages. What are the advantages? We can understand the effect of structure on activity. When I put in a particular electron withdrawing or electron donating what is the activity?

(Refer Slide Time: 36:10)

Advantages of QSAR:

- Understanding of the effect of structure on activity
- Synthesise novel analogues.
- The results can be used to help understand interactions between functional groups in the molecules of greatest activity

We can think of synthesizing novel analogues and we can try to understand interaction between functional groups. Suppose I have one OH in the para position, I want to put another OH what will be the interaction or I want to put OH CH₃ what will be the interaction that sort of thing. Disadvantages, of course false correlation, sometimes may arise because of heavy reliance placed on biological data.

(Refer Slide Time: 36:37)

Disadvantages of QSAR:

- False correlations may arise because of heavy reliance placed on biological data, which is subject to experimental error.
- Too many descriptors and selection of the correct ones from this vast pool is a major issue
- Lack experimental design. Therefore the data collected may not reflect the complete property space.
- Various physicochemical parameters are known to be cross-correlated.

Too many descriptors and selection of the correct ones. This is very, very important point which I did talk about right. Too many descriptors, how do I select the ones which you think

is the correct one. I may be missing out the correct one, lack of experimental design. So using computational I will say I want to synthesize a molecule with the chloro in this position that position, may be it is not possible for the synthetic chemistry to design or synthesize. They will be able to synthesize only one set of molecules okay so that is the problem.

Various physicochemical parameters are known to be cross-correlated like I showed you molecular weight, size, electronic features, all of them may be cross-correlated so not all descriptors will be independent of each other okay. So we will continue more on this QSAR in the next class as well we will talk about 3-D QSAR in the next class. Thank you very much for your time.