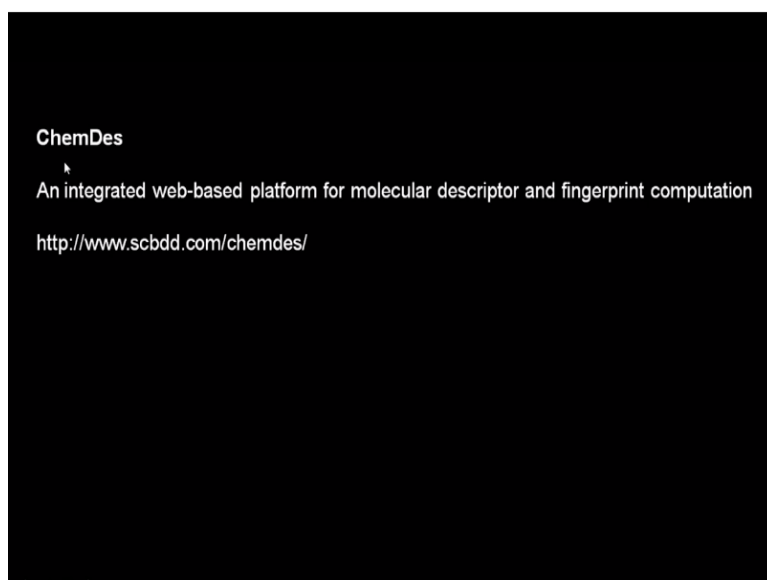


Computer Aided Drug Design
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology - Madras

Lecture - 28
Quantitative Structure Activity Relationship (QSAR)

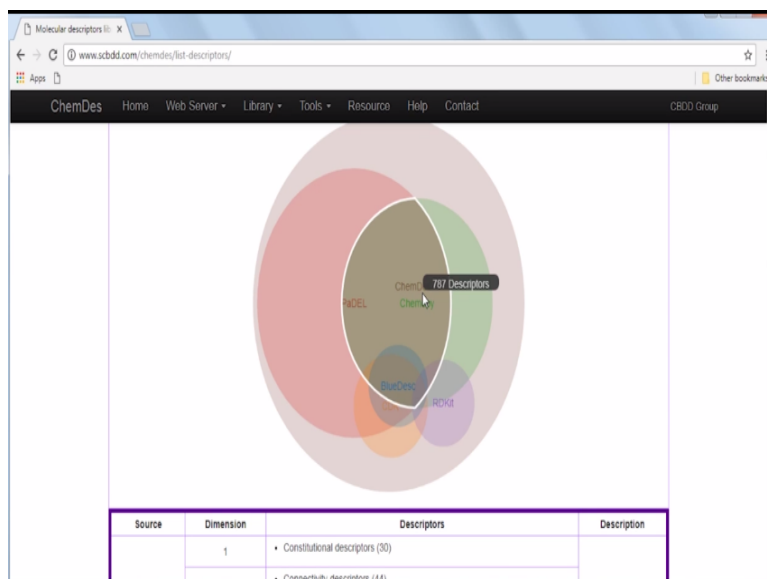
Hello everyone, welcome to the course on computer-aided drug design. We will continue on the topic of QSAR. Today, we will look at couple of softwares which can give us certain descriptors even the structures. In fact, there are many softwares in the net, which is free of charge, web servers are there from where we can get 1000s of descriptors. I am going to show you couple of them and you guys can explore further.

(Refer Slide Time: 00:44)



One of them is called ChemDes okay. This is an integrated web-based platform for molecular descriptor and fingerprint computation is called ChemDes. Let us have a look at that okay.

(Refer Slide Time: 01:09)

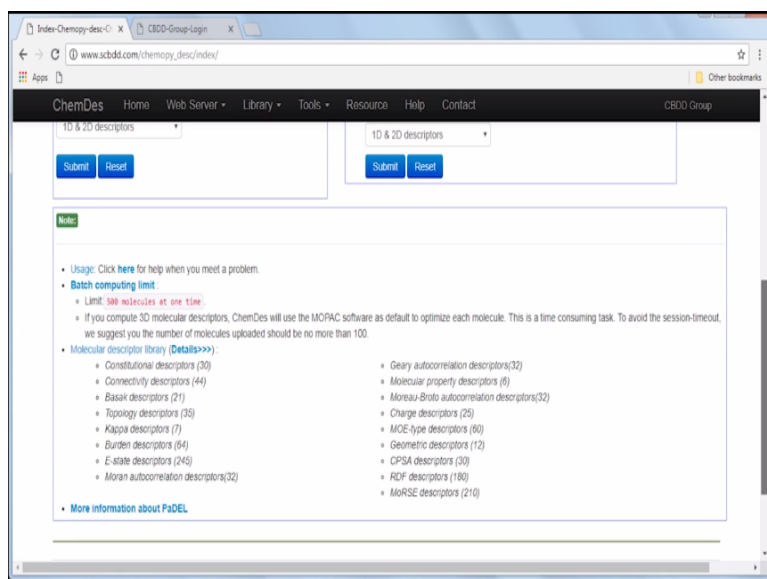


This is the ChemDes okay as you can see this is the address of this. It has got lot of descriptors, look at this descriptors library. See look at this so ChemDes it can calculate 787 descriptors okay. There is another software is there and so many different softwares are there okay, this software can calculate 103 descriptors, this can 196 descriptors, 787 so there are many softwares PaDEL and so on okay.

Let us look at ChemDes okay as I said ChemDes can calculate almost 1135 descriptors. There are zero dimensional descriptors, 1-dimensional descriptors, 2-dimensional descriptors, 3-dimensional descriptors and as you can see the next column okay these are the descriptors calculated by each of these softwares shown in this picture above okay Chemopy, ChemDes and so on.

And the details of these descriptors, if you go here it gives you all the details. This is how we can get the details here okay. So we can calculate lot of descriptors okay.

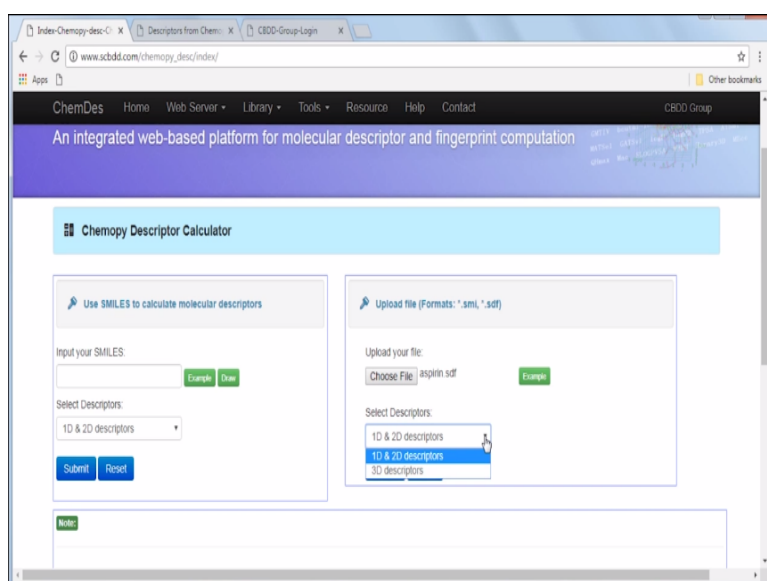
(Refer Slide Time: 02:54)



It is called Constitutional descriptors, Connectivity descriptors, Basak descriptors, Topology, Kappa descriptors so on, so on, so on and each one there are so many numbers. This is the number here and if you look here, it gives you the details of these descriptors okay. So if you want to go into more details, we can click on them, it will tell you for example Topology descriptors, very lot of Topology parameters are there, lot of Topology descriptors are there okay.

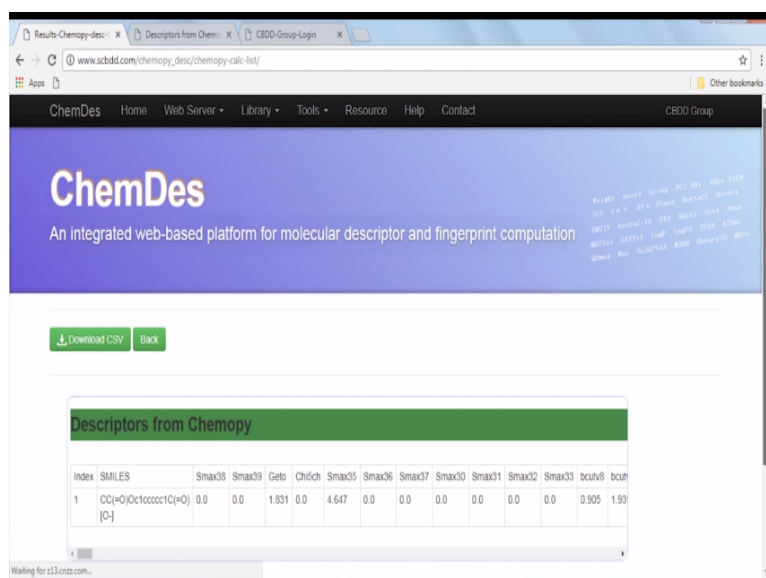
So we can upload a file for example we can upload a SDF file and so choose file. I have I think aspirin or something somewhere stored, yeah I have aspirin I can get that.

(Refer Slide Time: 04:26)



We can do either 1-D, 2-D or 3-D descriptor, 1-D, 2-D submit, so it will do some calculations and then give the details about that particular.

(Refer Slide Time: 04:34)

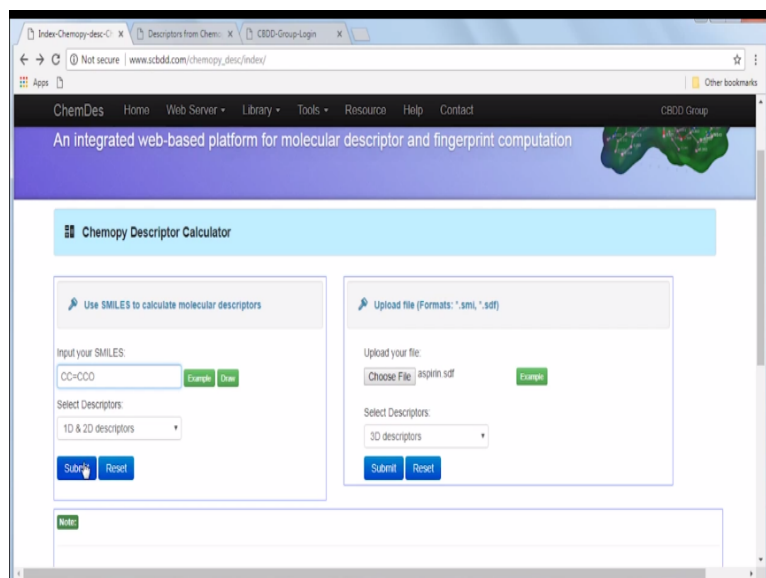


The screenshot shows the ChemDes website interface. At the top, there is a navigation bar with 'ChemDes' and 'CBDD Group'. Below the navigation bar, the ChemDes logo and tagline are displayed. A 'Download CSV' button is visible. The main content area features a table titled 'Descriptors from Chemistry' with the following data:

Index	SMILES	Smax38	Smax39	Geto	Chl6ch	Smax35	Smax36	Smax37	Smax30	Smax31	Smax32	Smax33	boutv6	boutv
1	CC(=O)Oc1ccccc1C(=O)[O-]	0.0	0.0	1.831	0.0	4.647	0.0	0.0	0.0	0.0	0.0	0.0	0.905	1.93

Yeah, as you can see that is giving large number of descriptors. Aspirin is there, so I can do a 3-D descriptor calculation also. Submit okay so it has given some 3-D information about the descriptor molecule as you can see. So you want to get more, you can download this CSV file and then save it for further calculation. If you want to know more about the details of what each descriptor means I did show you how to find out right.

(Refer Slide Time: 05:49)

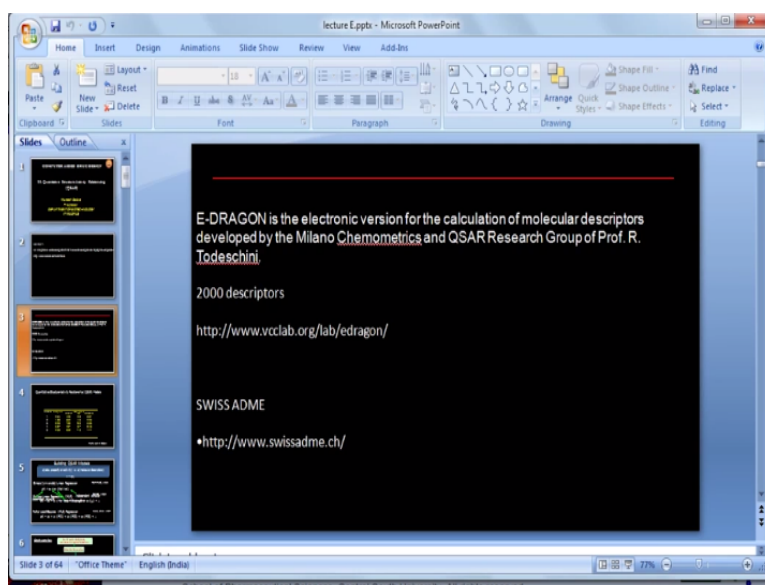


The screenshot shows the 'Chemopy Descriptor Calculator' interface. It has two main input sections: 'Use SMILES to calculate molecular descriptors' and 'Upload file (Formats: *.smi, *.sdf)'. The SMILES section has an input field with 'CC=CCO' and a 'Submit' button. The file upload section has a 'Choose File' button and a 'Submit' button. Both sections have a 'Select Descriptors' dropdown menu set to '1D & 2D descriptors' and '3D descriptors' respectively. There are also 'Reset' buttons for both sections.

We can also give the smile notation of the molecule also. For example, like that you know there is a smile notation and then give submit okay again it calculates the descriptors for that molecule. So this is CH₂-CH double bond CH-CH₂-OH okay. This is the unsaturated alcohol okay (()) (06:16) okay. So this is a very nice software okay so the descriptors library as I showed you it can show you lot of descriptors and I can calculate okay.

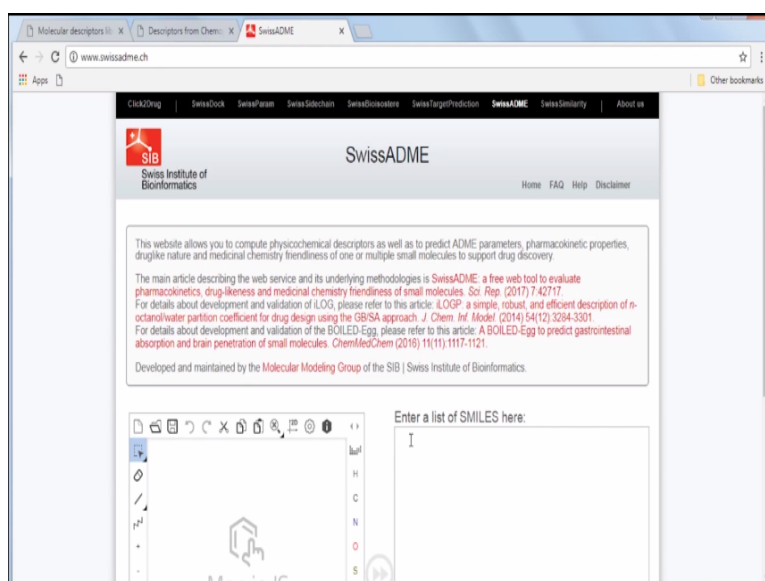
And as I said there were apart from ChemDes which can calculate 1135 we also have other softwares like BlueDesc and PaDEL and RDKit, each of them can calculate some amount of descriptors. So you have a huge number of descriptor calculation softwares, which are free for you so you do not need to pay at all. Remember you do not need to pay at all, we can get lot of descriptor calculator okay using softwares okay.

(Refer Slide Time: 07:19)



Similarly, we also have the SwissADME. If you remember SwissADME, we have seen that many times, that also can calculate a few descriptors for you okay.

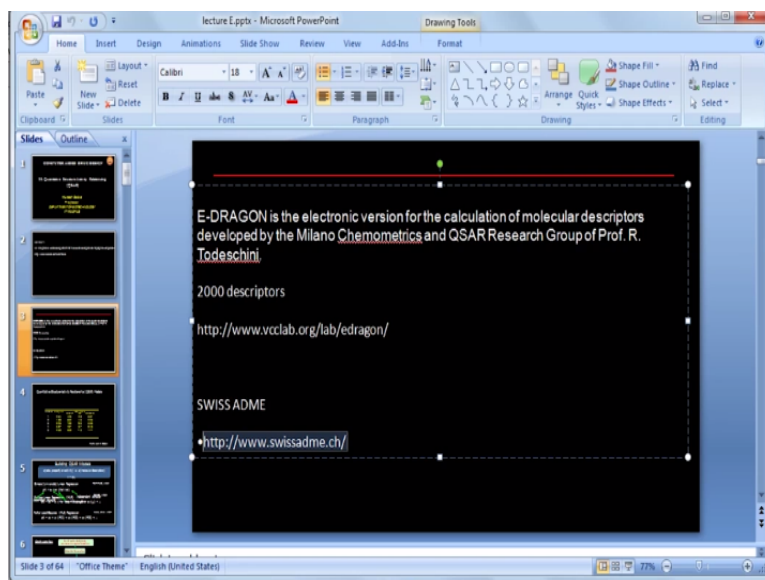
(Refer Slide Time: 07:34)



SwissADME can calculate lot of descriptors so you remember the software I think many times I have shown you this. So we can draw this, we can do this then we can say run but the

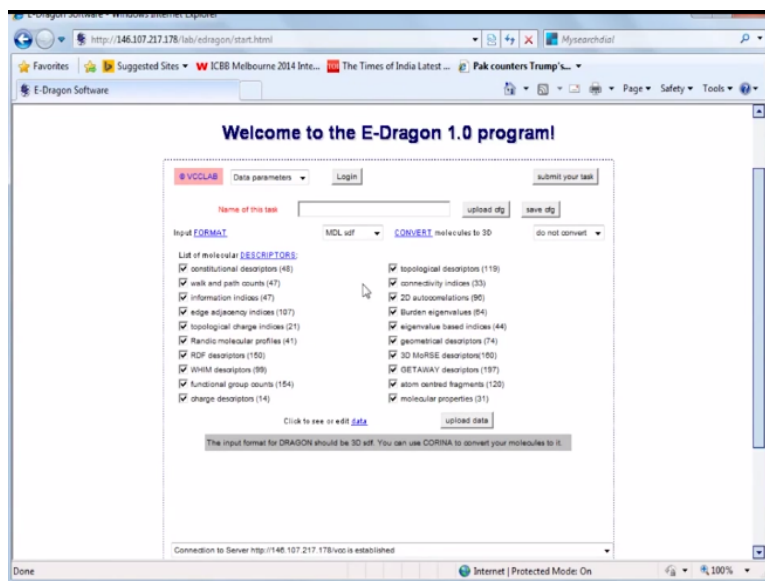
number of descriptors it calculates are very limited. See Log S, solubility and that sort of thing actually, not many, rotatable bonds and things like that but if you want to do a real QSAR, ChemDes is better option or we can go to the E-DRAGON okay.

(Refer Slide Time: 08:11)



E-DRAGON calculates 2000+ descriptors okay for a QSAR. SwissADME use only important parameters like solubility, Log P and bioavailability and so on but if you are interested in QSAR you want large number of descriptors then ideally you should go to this E-DRAGON, is developed by Milano Chemometrics by Professor Todeschini, it is a huge number of descriptors.

(Refer Slide Time: 08:48)

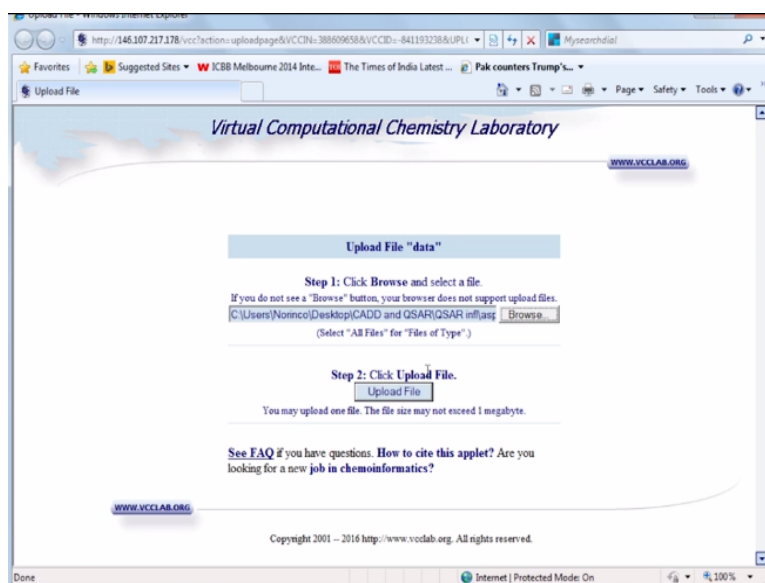


So you can do it online. Look at this, so when you come to E-DRAGON, this is the thing, it has got so many descriptors okay, thousands and thousands and thousands, look at this huge

number of descriptors numbers, list of molecular descriptors it says okay. Now how do we run molecules? It is quite simple, we can upload data, we need to upload SDF file, SDF file we can get it downloaded from if you remember your Zinc database.

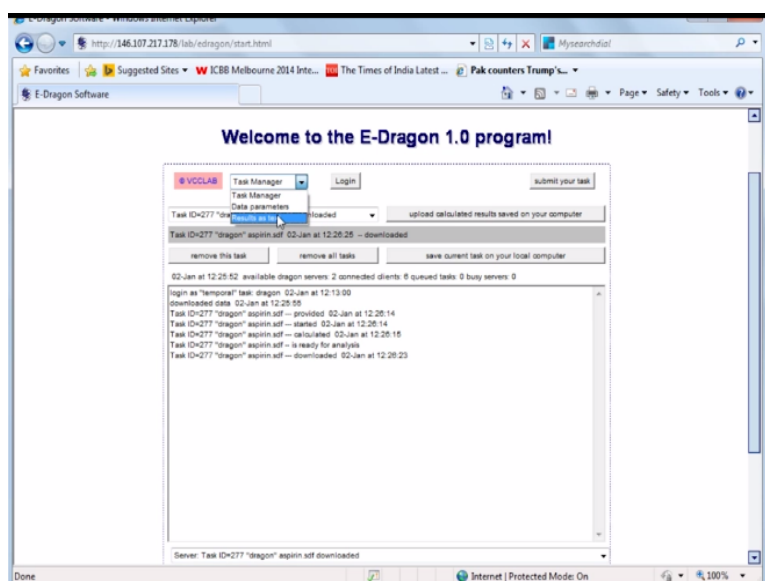
Zinc database gives you SDF file right. Let us look into I think we can run, I can do aspirin, open, being loaded, so we say upload file.

(Refer Slide Time: 10:03)



So the file has been uploaded. We close this, so aspirin has been loaded then we say submit your task okay. So it will calculate all these descriptors, doing calculations so doing calculations goes to the server and does it okay. So download is done okay. So what do we do?

(Refer Slide Time: 10:43)



We go here, we say results as text okay so we can see the results here okay. This is your input file as you can see for aspirin okay. This is the input file for aspirin okay.

(Refer Slide Time: 11:15)

dragonK: Descriptors

No.	MOL_ID	MW	AMW	Sv	Se	Sp	Ss	Mv	Me	Mp	Ms	nAT	nSK	nBT	nB
1	ZINC0000053	179.16	8.96	13.14	20.9	13.48	41.17	0.66	1.05	0.67	3.17	20	13	20	13
		0.022	0.008	0.003	0.071	1.816	1.774	1.631	1.102	0.98	1.033	0.62	0.066	18.044	0.185

So results have come, so control X, control A, control C, I have copied the entire result then I can go to excel for example and then I can say control V okay so all your results have come so this is your molecule okay. So you can see this molecule, we can convert this text data to numbers in columns (()) (12:07) simply we converted that into text to columns okay. So it is easy for us to view as you can see it is very nice to view now okay.

(Refer Slide Time: 12:51)

No.	MOL_ID	MW	AMW	Sv	Se	Sp	Ss	Mv	Me	Mp	Ms	nAT	nSK	nBT	nBO	nBM	SCBO	AP
1	NC000000	179.16	8.96	13.14	20.9	13.48	41.17	0.66	1.05	0.67	3.17	20	13	20	13	8	18	0.4

So molecular weight is 179, number of carbon, number of nitrogen's and so on okay, aspirin this is, it also gives you some Log PS values okay and log solubility also. So lot of descriptors you can calculate okay. So use of this, imagine I want to do one more and so on.

Now I want to upload another file, say okay we looked at aspirin, we can look at another molecule for example, it is called Zinc53.

Let us see what is the result. I do not remember that let us look at Zinc6694 okay. Yeah this is a coxib,

(Refer Slide Time: 15:25)

The screenshot shows the ZINC database entry for ZINC6694 (Cox). The page includes a chemical structure, a table of properties, and a table of available 3D representations.

Added	Available	Since	Mwt	logP	Heavy Atoms	Tranche	Download
2005-09-29	In-Stock	2015-08-07	314.366	2.954	22	DFAA	Download

SMILES: Cc1onc(-c2ccccc2)c1-c3ccc(S(=O)(=O)O)cc3

InChI: InChI=1S/C16H14N2O3S1c1-11-15(12-7-3-14)(10-8-12)(22)(17,19)(20)(16)(10-21-11)13-5-3-2-4-6-13R(2-10H,1H3,(H2,17-

InChI Key: LNPDTQAFDNKSHKUHFFACYSAN

pH range	Net charge	H-bond donors	H-bond acceptors	IPSA	Rotatable bonds	Apolar desolvation	Polar desolvation	Download
Reference	0	1	4	86	3	4.93	-13.7	Download

This is a selective cyclooxygenase inhibitor okay. This is coxib as you can see it has got sulphur, two oxygen's here. There is selective cyclooxygenase-2 inhibitor actually. So we will load that, upload data, browse, upload file, yeah well it has been uploaded, now submit your task, results as text okay, open results in a browser. Yeah so control X, control A, control C.

So we can convert data into text to columns okay, fix with delimited width, next space, so 314 is result, so this is the coxib and this is the selective cyclooxygenase-2 inhibitor we can remove this column also okay.

(Refer Slide Time: 17:50)

No.	MOL_ID	MW	AMW	Sv	Se	Sp	Ss	Mv	Me	Mp	Ms	nAT	nSK	nBT	nBO	nBM	SCBO	AR
1	NC000002	179.16	8.96	13.14	20.9	13.48	41.17	0.86	1.05	0.67	3.17	20	13	20	13	8	18	0.4
2	NC000066	314.39	8.73	24.2	36.56	25.59	55.42	0.87	1.02	0.71	2.52	36	22	38	24	19	34.5	0.7

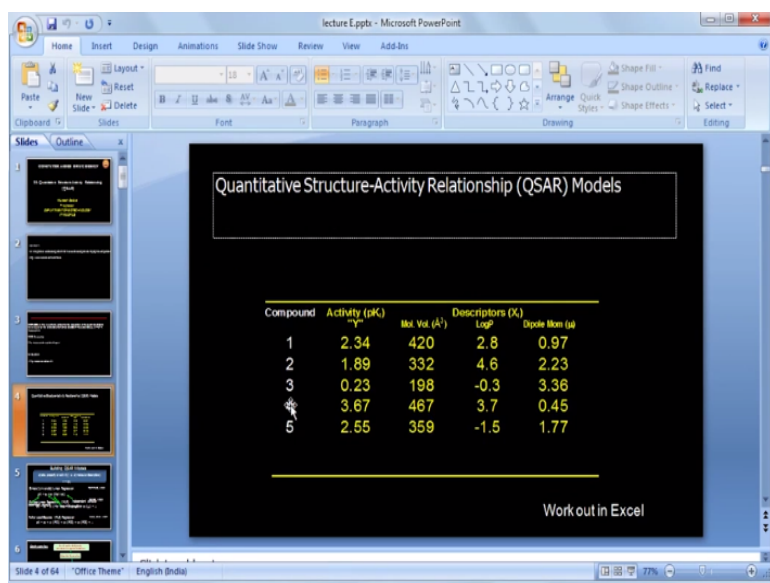
So this is another molecule okay, there is a different molecular weight and so on so it goes descriptors. So if I know the activity that is my y for my QSAR and these are the x's okay which many molecules like these activities are known I can develop a regression relationship with the one descriptor model or 2 descriptor model depending upon the number of data you understood.

So DRAGON is also very powerful software for getting large number of descriptors okay. The details of the descriptors we can get it as I said you can get the details of the descriptors from this okay. Details of descriptors we can get it from the main page of the software okay so like I said if I have the activity details and that will be my y my regression relation.

I can select the best descriptor and then develop a regression relation $y = mx + c$, x could be my one of the descriptors. So depending upon the number of data points I can have one descriptor model or two descriptor model or many descriptor models okay. So we talked about DRAGON. Then, I showed you the other software ChemDes that also calculate descriptors okay.

Huge number of softwares are there okay, so you do not need any commercial software as you can see so many different softwares are there, which can calculate descriptors for you okay.

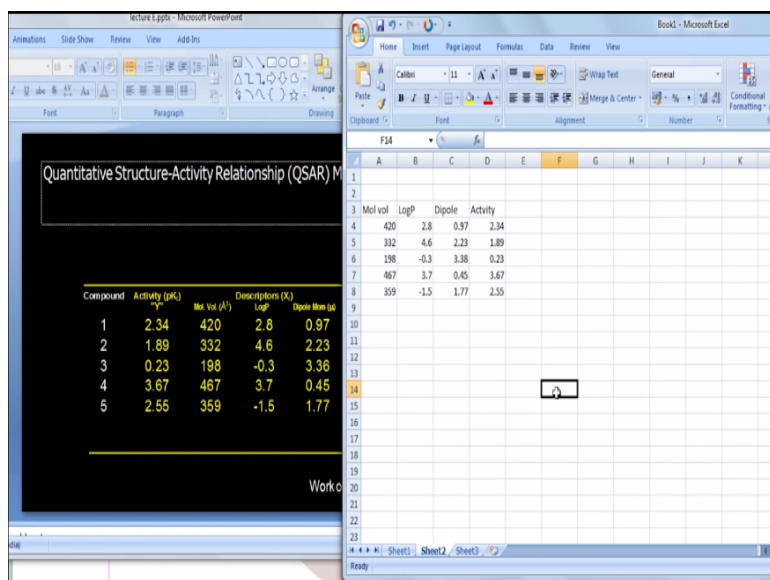
(Refer Slide Time: 19:49)



So we have imagined some activity, we have some descriptors okay. So how do you go about doing this business of calculating the regression relation? Simple, if it is a simple model then it is not very difficult for us to do. We can use excel for example okay. So I can say molecular volume as you can see 420, 332, 198, 467, 359 okay. Then activity could be 2.34, 1.89, 0.23, 3.67, 2.55 okay.

Log P 2.8, 4.6, -0.3, 3.7, -1.5, it is quite hydrophilic and then your dipole 0.97, 2.23, 3.36, 0.45, 1.77 okay.

(Refer Slide Time: 21:24)



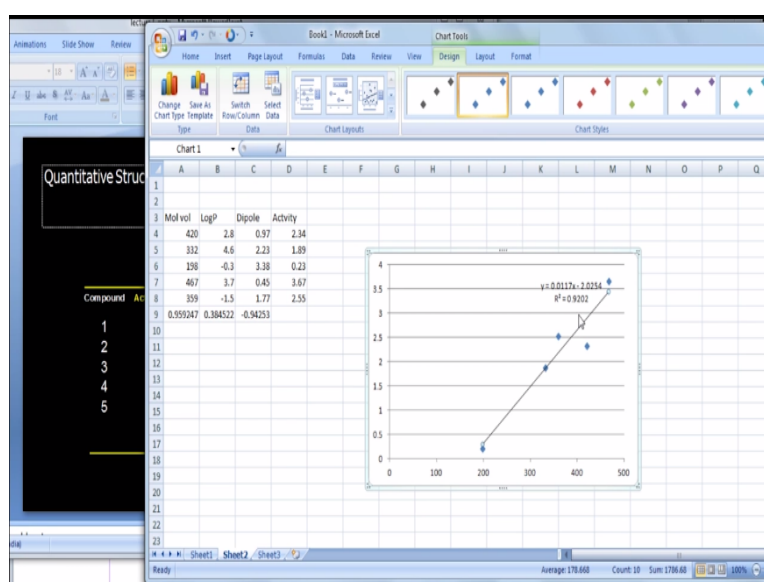
Now we have only 5 data points, so at the most we can develop only one independent variable regression relationship either with molecular volume or Log P or dipole only. So how do I find out? I can use correlation to see whether there is a correlation between the X

and activity let me see what is there, it is quite good molecular volume, it shows quite good correlation. Let us see whether Log P also has a correlation, not so good, see dipole, how could the dipole has the correlation okay.

So this also has a very good correlation negative that means as the dipole increases activity decreases, molecular volume is positive so as the molecular volume increases activity also increases okay. Now let us draw graph also. We can insert scatter so we get a scatter plot as you can see. As the molecular volume increases, your activity also increases. This is molecular volume on the x axis activity.

So we can draw a trend line I think you all guys must be good at excel okay, display equation, display R-squared value close okay.

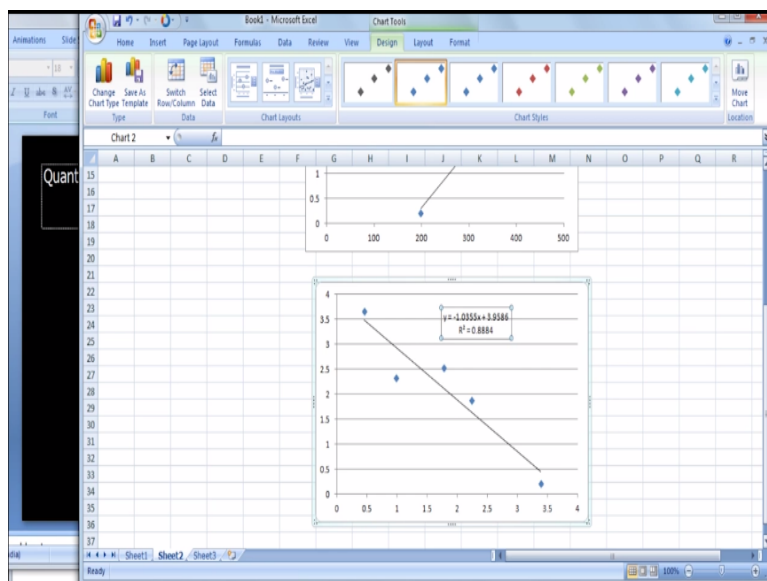
(Refer Slide Time: 23:08)



So this is your QSAR, $\text{activity} = 0.0117 \times \text{molecular volume} - 2.025$. R-squared is pretty good = 0.92 which is very very nice. Let us do with dipole also because dipole also is good. So we may have a QSAR with the dipole also. So let us look at dipole, $x = \text{dipole}$ okay, so as the dipole increases activity decreases like I showed you it is a negative correlation here and if you look at the R-squared here, R-squared this is also good.

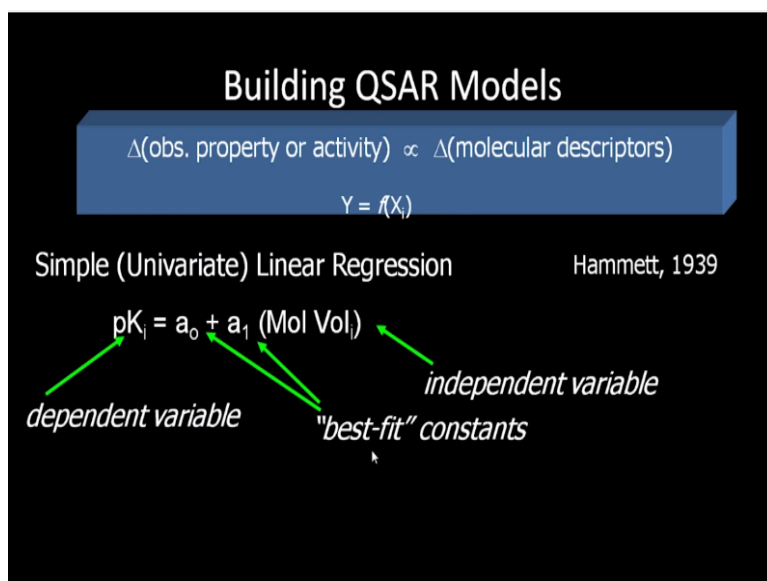
So this is the QSAR with respective dipole, $x = -1.03$ because it is going down into dipole + 3.9 and here molecular volume regression is $\text{activity} = 0.0117 \times \text{molecular volume} - 2.254$, so we can have two either one of the QSAR's. R-squared here is 0.92, R-squared here is little bit less that does not matter, it is not so greatly bad.

(Refer Slide Time: 25:04)



So we can have depending upon what we like whereas if you plot the Log P, the correlation is very poor, 0.38 only. So we will not go for this type. So excel is a good software. If you are interested in looking at regression equation with one independent variable model okay. With one independent so we can work out with excel that is not a big deal.

(Refer Slide Time: 25:32)



So we can have molecular volume or like I showed you dipole or we can have multiple linear regression.

(Refer Slide Time: 25:38)

Building QSAR Models

$$\Delta(\text{obs. property or activity}) \propto \Delta(\text{molecular descriptors})$$

$$Y = f(X_i)$$

Simple (Univariate) Linear Regression	Hammett, 1939
$pK_i = a_0 + a_1 (\text{Mol Vol}_i)$	
Multiple Linear Regression (MLR)	Hansch, 1969
$pK_i = a_0 + a_1 (\text{Mol Vol}_i) + a_2 (\log P) + a_3 (\mu_i) + \dots$	

But then multiple linear regression but then multiple linear regression is not possible because the number of data points is only 5, general rule of thumb is if you have 5 data points go with only one descriptor model, do not go for more. If you have 10 maybe so this generally is not a good idea okay.

(Refer Slide Time: 26:00)

Building QSAR Models

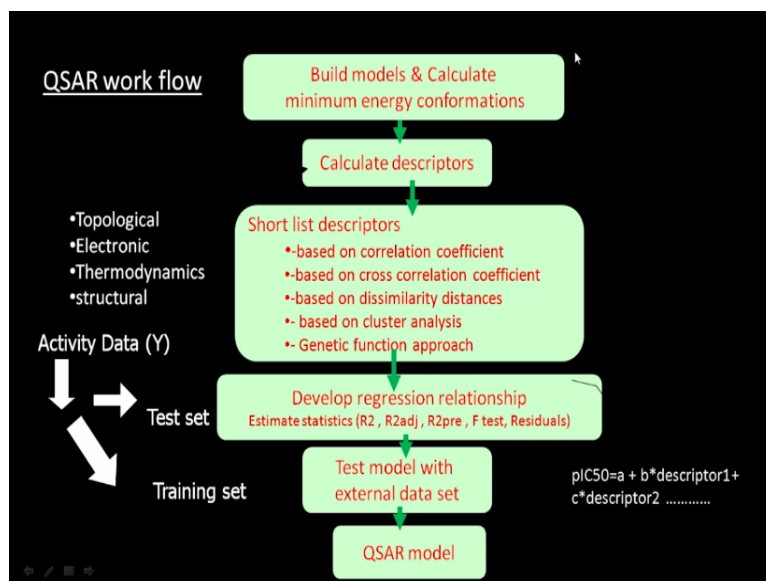
$$\Delta(\text{obs. property or activity}) \propto \Delta(\text{molecular descriptors})$$

$$Y = f(X_i)$$

Simple (Univariate) Linear Regression	Hammett, 1939
$pK_i = a_0 + a_1 (\text{Mol Vol}_i)$	
Multiple Linear Regression (MLR)	Hansch, 1969
$pK_i = a_0 + a_1 (\text{Mol Vol}_i) + a_2 (\log P) + a_3 (\mu_i) + \dots$	
Partial Least-Squares (PLS) Regression	Wold, et al. 1984
$pK_i = a_0 + a_1 (\text{PC1}) + a_2 (\text{PC2}) + a_3 (\text{PC3}) + \dots$	

Or we can have partial least square lot of statistical terms coming actually. So this is how we go about doing simple structure activity mathematical relation okay.

(Refer Slide Time: 26:13)



So what do we do? We need to build models, calculate minimum energy conformation if you are looking at minimum structural features. Then, calculate descriptors, I showed you some softwares, you may be able to find more softwares. Then, you need to shortlist descriptors, you cannot take like in DRAGON there are almost 2000 descriptors. We cannot take all of them.

So we need to based on the correlation coefficient like I showed you here right and the excel the correlation coefficient for molecular volume is very good, dipole is next good, so I may go with molecular volume. So this is what it means, shortlist descriptors, there is some correlation coefficient. These descriptors should not be cross correlated that is another important thing okay.

So what does that mean? There should not be a correlation between this, this and this and this, just look at whether there is a correlation between this and this may be then we cannot use the same descriptors. This is molecular volume okay molecular volume and dipole has very high correlation okay. So obviously if you are thinking of two descriptor model, we cannot use both, they are cross correlated you understand, two descriptor model.

You have to first check that but there is no correlation between x's, X's means independent variables understand. You can see -0.99 molecular volume and dipole whereas if you look at molecular volume and Log P I do not think they will be correlated but of course we are not going to take Log P here because the correlation of Log P with activity is quite low but definitely we will not take this and this because they are both correlated okay.

So that is the very important point. They should not be correlated, they should be dissimilar as dissimilar as possible so we can do cluster analysis, we can do so many things to shortlist descriptors. Shortlisting descriptors is a big challenge so because you have 2000 descriptors and I may have only 10 data points. So I need to select two descriptors out of this so I may make a mistake, I may not select the right descriptor, I may miss out descriptors.

So that is the biggest challenge in QSAR that is the most biggest challenge in QSAR. How to select correct descriptor, how to not select the wrong descriptors, how to not select a descriptors which are correlated with each other so all these need to be considered and there are many approaches. Once we do that we develop a regression relation, it could be a one parameter regression if you have one descriptor $y=mx+c$.

If it is two parameters $y=m_1x_1+m_2x_2+c$ and then you do a lot of statistics. There are something called R-squared, adjusted R-squared, predicted R-square, F test, residual so many things are there. Then, when you have a large number of data you do not use all of them for calculating regression, you divide the data in 2 types, training set, test set, so what do you do?

You use the training to develop the descriptor and the regression and then without regression you try to predict the values in the test set okay so if I have say 6 data points what will I do? I will take only 5 data points and develop my regression and then try to predict for the 6th one and see how good the prediction is. That is called a test set okay. So if you have say 7 or 8, I may take 5 or 6 in the training set, the remaining 2 will be in the test set.

Generally this say 20%, 20% should be in the test set, 80% of the data in the training set, so if you have a large number of data you just do not develop regression with all the data points and you follow the 80, 20 20% for test set, 80% for training set, so you train your model and calculate the various terminologies, various terms like m and c with 80% of the data and then you predict the remaining 20% of the data using that model to see how good the prediction is okay that is very, very important.

Then, you can also test with external data, you make experiments, get some results activity and then see, so you need to do all these to be sure that your model is good and it has got good predictive capability that is very, very important. So the descriptors like Topology,

electronic, thermodynamics, structural all these are descriptors which I showed you using different web servers which are free for you to use okay.

So we need to have at least 5 data points for regression equation with one descriptor model and when you have data set, do not use all of them for training and developing the model, use only 80% of them for training and remaining 20% use as testing out the test set okay so it is very, very important that you have a training and test set 80-20 and your model should be able to predict the results of the test set whereas you use the 80% of the data for developing a regression coefficient.

Without test set your regression QSAR is of no use, it is very, very important that you also have a test set okay. We will continue more on this QSAR in the next class as well. Thank you very much for your time.