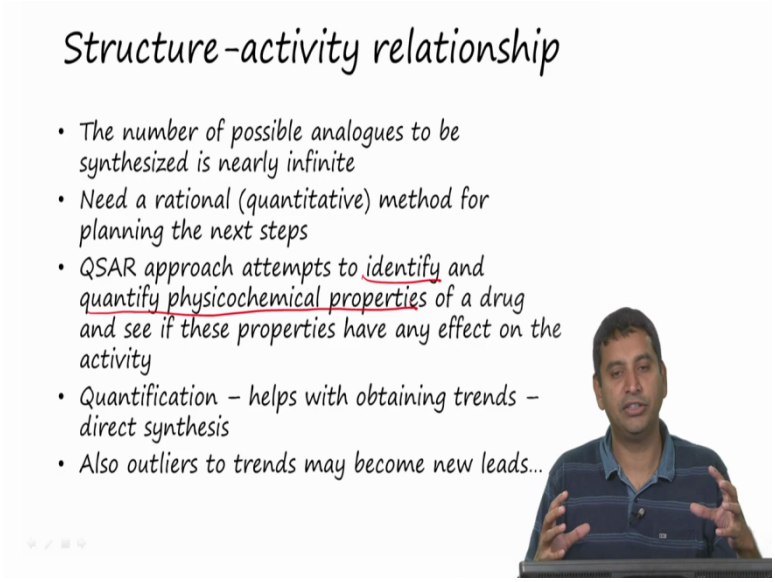
Medicinal Chemistry
Professor Harinath Chakrapani
Department of Chemistry
Indian Institute of Science and Engineering Research, Pune
Quantitative Structure-Activity Relationship (QSAR) Part-1

(Refer Slide Time: 00:16)

So welcome back, we have looked at how to identify a lead compound and now once we have a lead compound we need to further modify it so that we can make the compound more drug like okay. So we will be looking later as what are the ways in which that compound is taken forward once we identify a lead and how clinical trials are conducted and so on. But before we get there we need to have a good understanding of whether we have the best compound to take it forward because clinical trials are very expensive and therefore we need to make sure that we have the best compound okay.

So traditionally we are trying to synthesise more compounds and using traditional structure-activity relationship we have some idea about what are the next step forward. Today we will look at quantitative way of studying this and this is called quantitative structure-activity relationship, so this is QSAR. So let us start from structure-activity relationship.

(Refer Slide Time: 1:09)



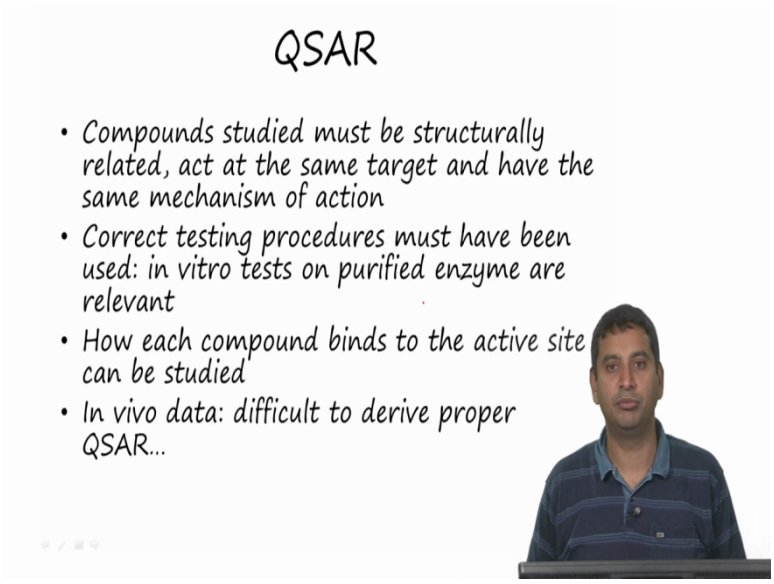
So here you have a number of possible analogues that can be synthesised, sold you do not have good idea about how to go forward we will look more or less from intuition as to maybe this part of the structure should be changed or the other part of the structure should be changed and so on. But if I add up these combinations it will be nearly infinity and clearly we do not have the resources to synthesise infinite number of compounds, so what we need is that we need a rational method and preferably quantitative methods for planning the next step. And since synthesises is actually quite expensive to carry out, we would need to take help from quantitative methods. So yet used the approach that we use, we use QSAR approach where we aim to identify and quantify physicochemical properties.

So after all a drug or small molecule is consist of various functional groups, it consists of various structures in the molecule which have some physicochemical properties. So let us say there is an enzyme which we are targeting and there is an inhibitor that we have synthesised and we know that inhibitor has to bind to the enzyme and inhibited okay. So all these binding we have already looked that is all majority of them are non-covalent interactions and those interactions are important for the binding properties okay. Now can we identify various physicochemical properties of this drug molecule or of this inhibitor and see which of these properties if I change any of these properties what effect it has on activity, so that is the aim of doing typical structure activity relationship but we will attempt to do it in a quantitative manner ok.

So once we have quantitative manner so let us say we have a straight line or curve, I now know that if I make a particular change to a particular physicochemical properties of the molecule then it is going to have this particular effect okay. So this will help it obtaining friends and now once I have this trend then I can in some predictable manner said that you are the next 10 compounds that you need to make ok, so that will help the synthetic chemist to synthesise new molecules.

Now when we are doing the structural activity relationship that is QSAR, we may also find that there are certain molecules which do not appear in this trend but they are still active, so those molecules will also become important because now there may be some other physicochemical property that may become important. So together these are the reasons why people attempt to or we look at quantitative structure-activity relationship.

(Refer Slide Time: 3:38)

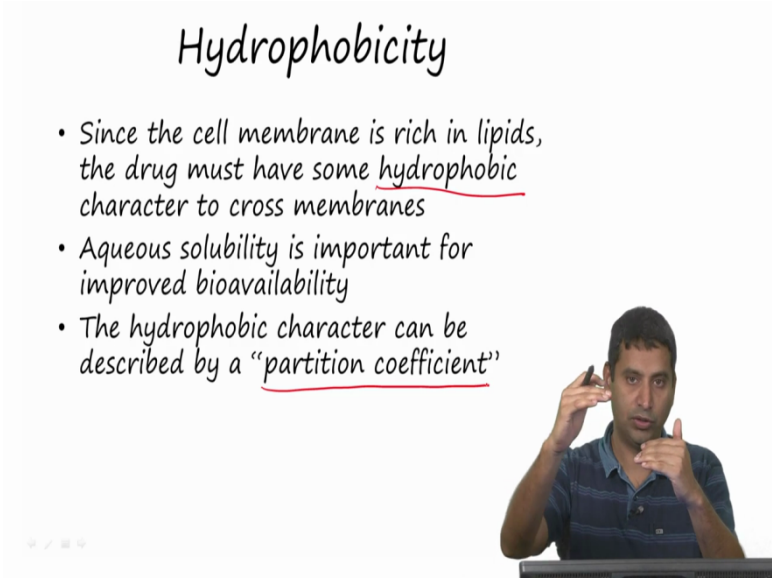


So what are the conditions? We can't do QSAR on all types of compounds, but we can do it on a small subset of compounds and the condition is that it must be structurally related ok. So if I have a bunch of inhibitors that I have studied and I have identified a lead compound and now I use SAR that is structure-activity relationship to identify or modify that lead compound and then I have let us say 20, 30, 40 compounds in that series and I studied the inhibitory effect. Now these compounds that I am studying our variants of the parent compound okay, so therefore they are structurally related okay. And since they are structurally related it is quite likely that they will act at the same target and perhaps have the same mechanism of action, so QSAR can be reliably conducted on the series of compounds.

Next we should use the correct testing procedures that have been used, so if I have a purified enzyme we have already looked that even in the past as to how to identify enzyme inhibitors, so we typically verify the enzyme and then screen for inhibition of the enzyme activity. So this correct testing procedure must be used on the Prairie fight enzyme with the cities of compounds okay, and keep in mind that it is preferable to get data which is comparable with the enzyme under the same conditions okay so we can then study how each compound binds to the active site.

And now QSAR cannot be very easily used to derive good conclusion from NVO data so you may have a lot of NVO data with a particular compound because NVO data is going to have more difficulties as because you have to think about pharmacokinetics, you have to think about bio distribution and all these properties are not very easy to predict and therefore QSAR will largely fail when it comes to trying to explain NVO data. So the condition under which QSAR works is that we should have a series of compounds which are structurally related which have been in the main by appropriate testing procedures and they should be on a purified enzyme and not in a NVO model ok.

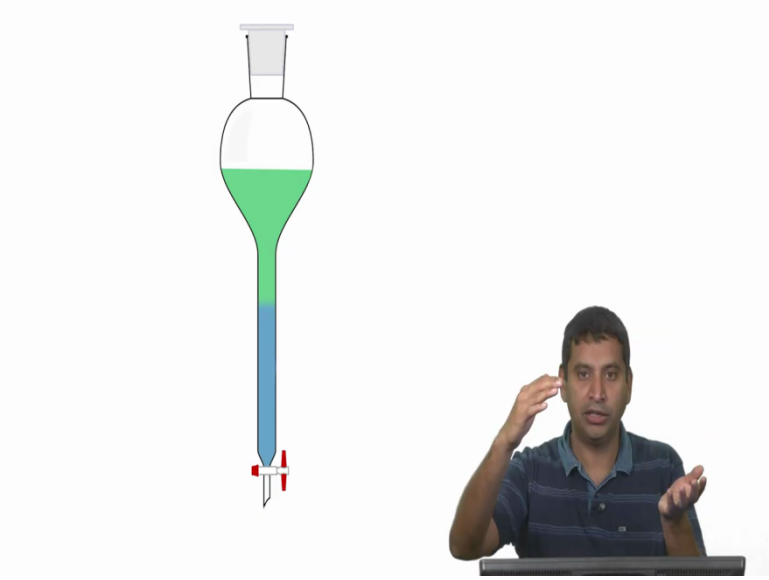
(Refer Slide Time: 5:45)



So one of the 1st things that people look at when it comes to understanding how a series of compounds work is hydrophobicity ok. As you know, since the cell membrane is rich in lipid ok. As you know, since the cell membrane is rich in lipids, the drug must have some hydrophobic character for it to cross membrane and act in the target ok. At the same time we need aqueous solubility for improved bioavailability. So the hydrophobic character of a molecule can be determined by what is known as a "partition coefficient" ok. Partition

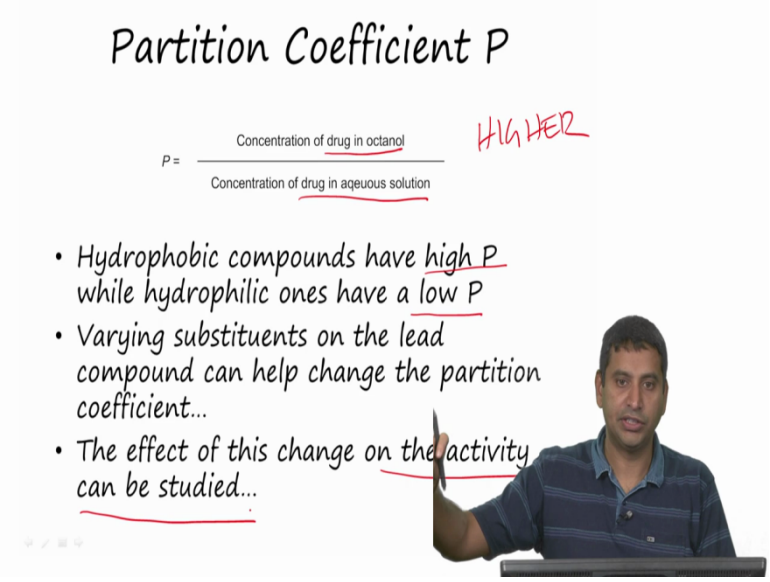
coefficient is nothing but the level to which a particular drug molecule okay partitions in 2 different solvents ok.

(Refer Slide Time: 6:14)



So during sentences for example, aqueous soluble impurities can be washed off using a separate funnel which many of view may have used. Similarly, what we are going to do is we are going to determine what is the partition between octanol and aqueous solution ok.

(Refer Slide Time: 6:29)

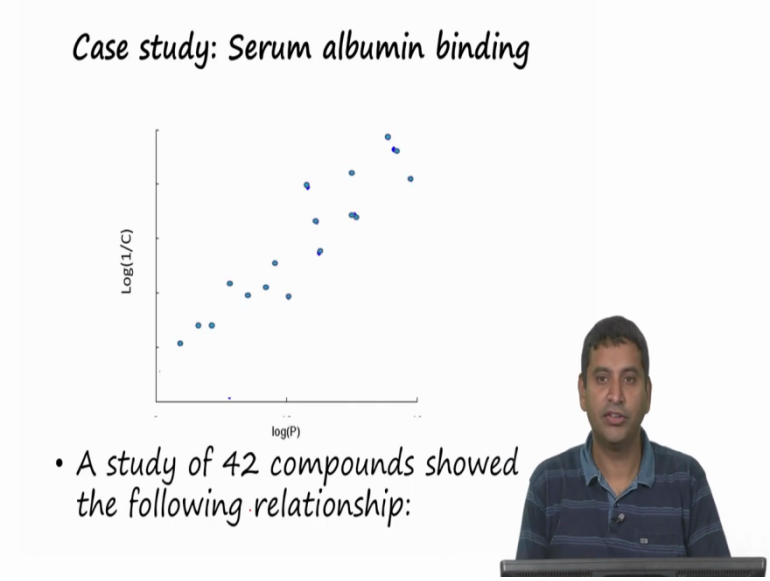


So P is nothing but concentration of the drug in octanol divided by concentration of the drug in aqueous solution ok. So in a separate module we will be looking at how to determine partition coefficient and therefore I will spare the experimental details over here. Now if I

have a mixture of octanol and water or buffer, hydrophobic compounds will preferentially partition into the organic layer ok. Whereas, the hydrophilic compounds or the aqueous soluble compounds will preferentially partition into the aqueous layer, so therefore hydrophobic compounds will have higher concentration of the drug in octanol and therefore the P value becomes high, whereas in the case of hydrophilic molecule it will have low P value as a converse.

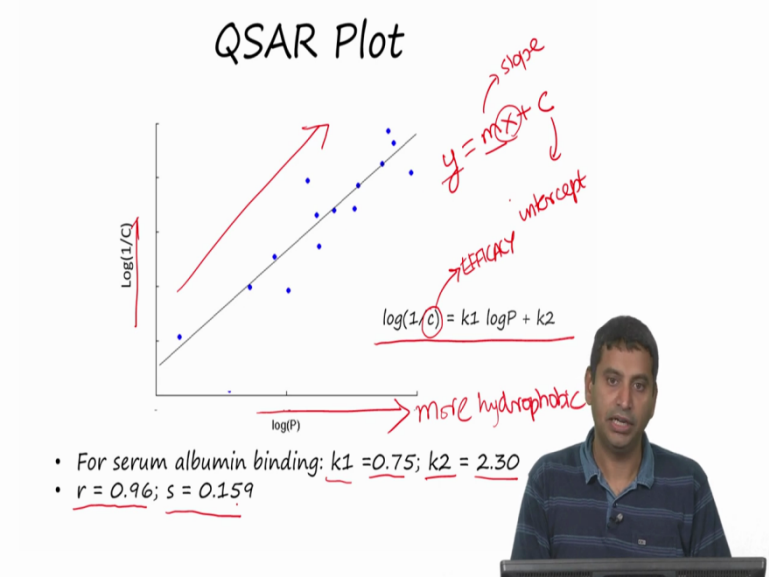
Now, once I have a lead compound I can vary the substrate wends and then study the effect of the change on the partition coefficient P. This effect is what we would want to study and understand whether it has any effect on the activities of the molecules. So keep in mind we are designing let us say an inhibitor for enzyme, and so we want to study how hydrophobicity affects the activity ok.

(Refer Slide Time: 7:42)



So here is the case study, so here they have taken about 42 compounds and they have studied the ability of this group of molecules to bind to serum albumin ok. So they find that if I take the ability to bind serum albumin as C, and then take a log of it and do log of 1 by C versus log of P because log of P is the partition coefficient okay, now I can study what is the trend. So in this particular study they found that they see all these dots are each individual compounds, each individual compounds represented by these dots and as we discussed previously by the experiment that I have discussed earlier, these are the values that are determined okay.

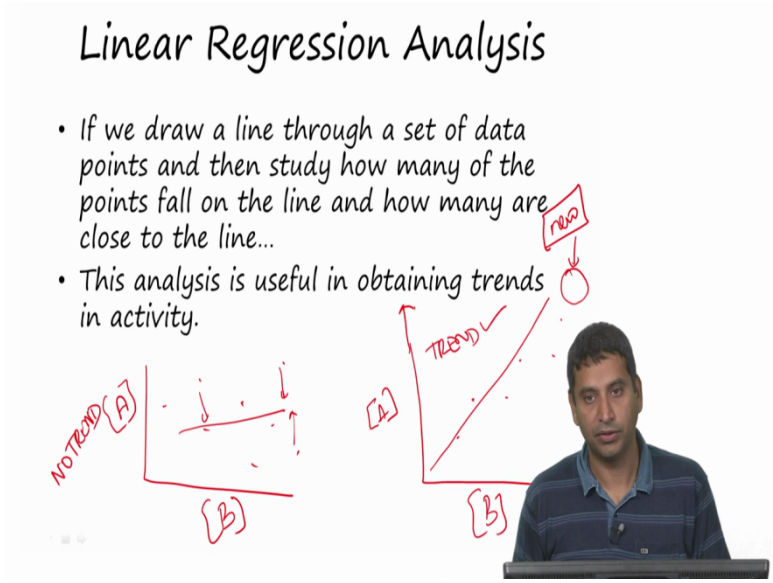
(Refer Slide Time: 8:24)



Now if I fit a straight line to this then I find that I get an equation called as log of 1 by C equals K 1 into log P + K 2 ok. So this is of the form Y equals M X + C, Now M is the slope and C is the intercept ok. Now using this equation I can study 2 parameters; Y and X variable Y and X ok. Now if I translate this to the equation written below, C is the efficacy of the molecule and log of 1 by C on the Y axis it gives me the increasing value here, and similarly log P goes up as you can see here from left to right, and higher the log P value the more hydrophobic the molecule is right. So therefore when I go from left to right, I see that the activity goes up okay so if I said this to occur I get K 1 which is the slope as 0.75 and K2 which is intercept is 2.30 ok.

And there are other parameters that we obtained while we are fitting this curve, one is called the linear regression coefficient which is R and the 2nd one is called the standard deviation ok or standard estimate of error right.

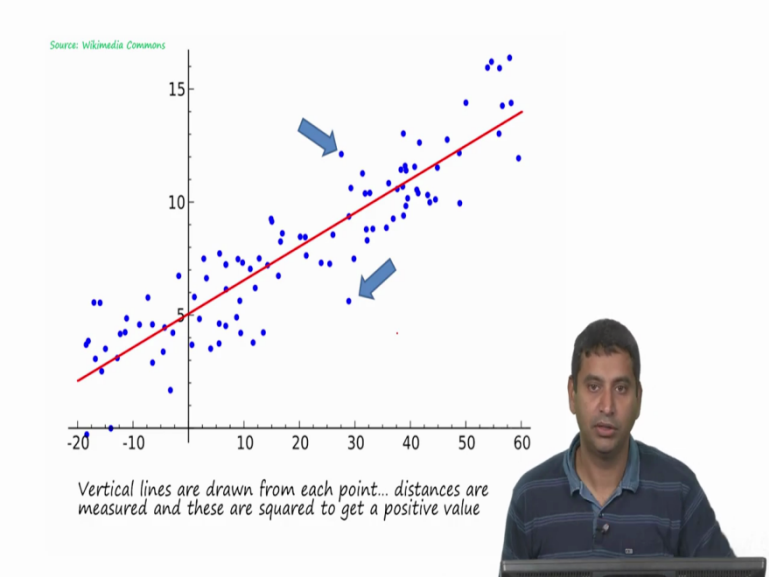
(Refer Slide Time: 9:48)



Now let us look at us linear regression analysis, now if we have a set of data points, what happens is that you can actually draw a set of data points. Now I can fit this to a straight line okay, and the best fit here is obtained through software or using numerical analysis, but you can take any 2 parameters A and B and fit a straight line to it ok. Now you can also have a situation where you have absolutely no trend, if I set up straight line, actually you do not get a straight line, I get something like this each of this point is very far from the straight line ok. So therefore you can have 2 contrasting situations where there is no trend okay, keep in mind this again parameter A and parameter B and here there is a trend okay.

Now the linearity here is obtained I will explain later but linearity can be obtained by looking at the regression coefficient, but if I have a linear relationship here I can then tell you that if I make a new compound in this direction then A can be improved okay, so therefore this is useful in obtaining trends for activity ok.

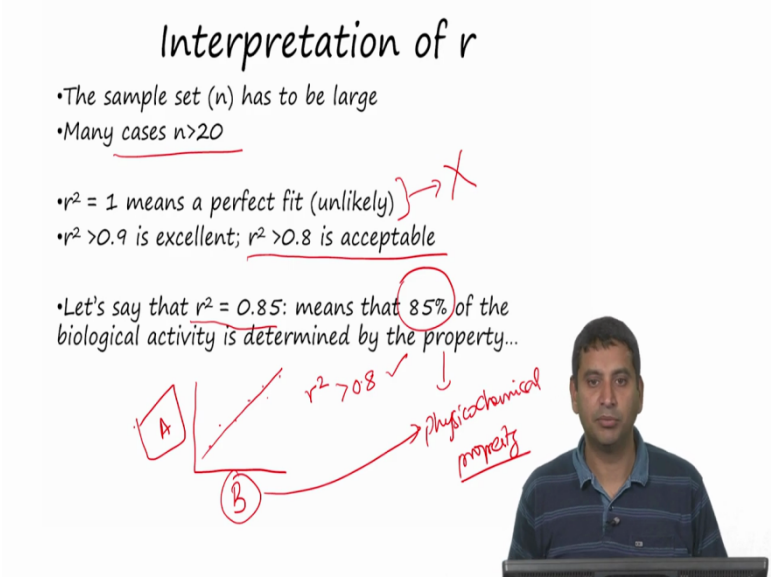
(Refer Slide Time: 11:07)



Now here is the data point for example that I fit to a straight line and I have got let us say hundreds of data points and so from this analysis you can say these 2 arrows here, these are what are known as out layers which are not really fitting on the line very well okay. But you can see that the majority of the points are actually fitting very nicely, there are a few out layers here as well and majority of points are fitting very close to the straight line okay. Now when I get this, I can determine the slope ok and from this I can also find out what is the value that I get for regression coefficient okay.

Now if we do this analysis, I can do our best fit and then I can do a regression or correlation coefficient and closer the value is to 1, so R should be ideally 1 but rarely you get this situation when R equals 1 so you would probably end up getting R square which is the square of the regression coefficient, it would be typically greater than 0.9 would be considered an excellent correlation or even greater than 0.8 is acceptable ok, so this is how we go about and obtain what is known as the linear regression coefficient ok.

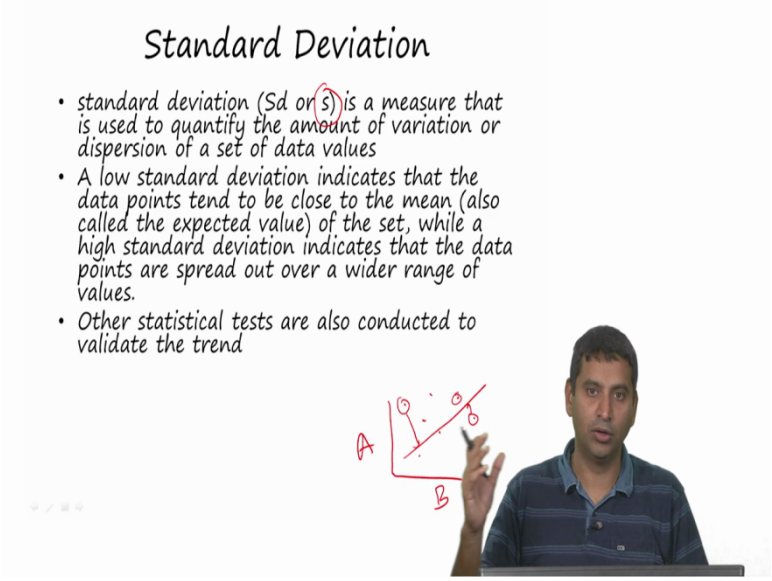
(Refer Slide Time: 12:29)



And what it means is that if I have a sample set of n and typically the sample set has to be greater than 20, R square effective 1 this is highly unlikely, it in fact almost never occurs that you have a straight line that happens between a biological property and physicochemical property of a drug. For example, inhibitor activity versus hydrophobicity you almost never get 1, but there are cases where you get R square of greater than 0.8 ok. So let us say R square is greater than 0.85 in a particular example that means that 85 percent of the biological activity is determined by the physicochemical property of the molecule ok, so this helps us understand how the activity of a particular series of compounds can be tuned ok.

So just to recap, so if I have an activity A or in this case 1 by A, and if I have a physicochemical property B, if we see that the trend is like this and if we fit this to a straight line ok, we can then determine the linear regression coefficient R square and is R square is greater than 0.8 that would be all right for us and is what it means is that if particular physicochemical property B has a major effect on the activity ok.

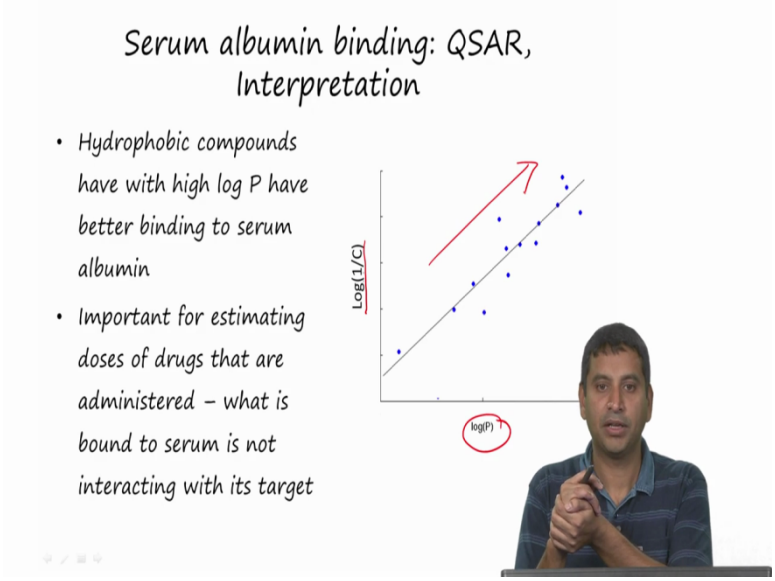
(Refer Slide Time: 14:02)



Now there is another term that comes in which are described over here which is standard deviation. So standard deviation is a concept we have studied many years back even in 12th grade it is being taught but I will very briefly explain to you what standard deviation is without going into too much detail I would recommend this as a self-study, but standard deviation is basically a measure that is used to quantify the amount of variation ok. So when you have data points which are let say something like this and you try to fit a straight line to it for example right, I would you can see there is a small variation here, where as there is a huge variation in this case okay. So therefore even looking at it I can tell you that the fit is not very good okay, but if I want to put a number to it and I want to determine what is known as the standard deviation.

Now therefore what we do is, we typically said that a low standard deviation means that the data point tends to be close to the mean or the expected value while a higher standard deviation indicates that the data is spread out over a wider range of values okay. So in a QSAR analyses, one would like the R square which is the linear regression coefficient to be let us say 0.8 or close to 0.8 or greater than 0.8, and in addition to that we also do not want a large dispersion because that makes it difficult to interpret the data. Ideally the standard deviation should be low in value, so you would typically find S values that are reported along the QSAR and we can also use other statistical tests which I will not go into detail but you would want R square value greater than 0.8 and then S value to be as low as possible if you want to make good interpretation from QSAR ok.

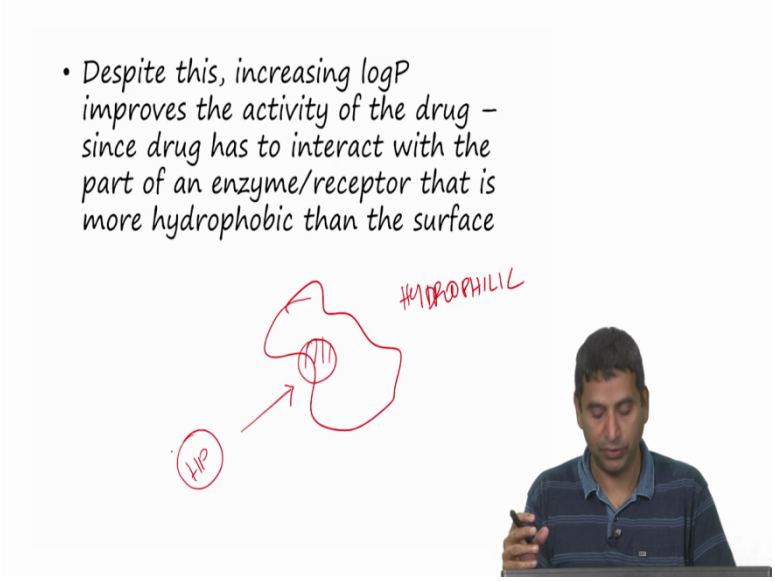
(Refer Slide Time: 15:45)



Now in the example that we have looked at, where we looked at the serum albumin binding, the interpretation of QSAR is that hydrophobic compounds which have a high log P have better binding to serum albumin okay. And it is important because the more hydrophobic the molecule is the better it binds to serum, is molecule binds better to serum then what happens is that it is not interacting with the target okay. So you want the molecule to be interacting with the target and not interacting with the serum and therefore you would prefer a molecule which has appropriate log P ok. So despite this including the log P actually improves the activity of the drug, in many cases we observed, so in that case study that we looked at which is serum albumin binding, we find that as the hydrophobicity of the molecule goes up, the activity of the compound goes up okay.

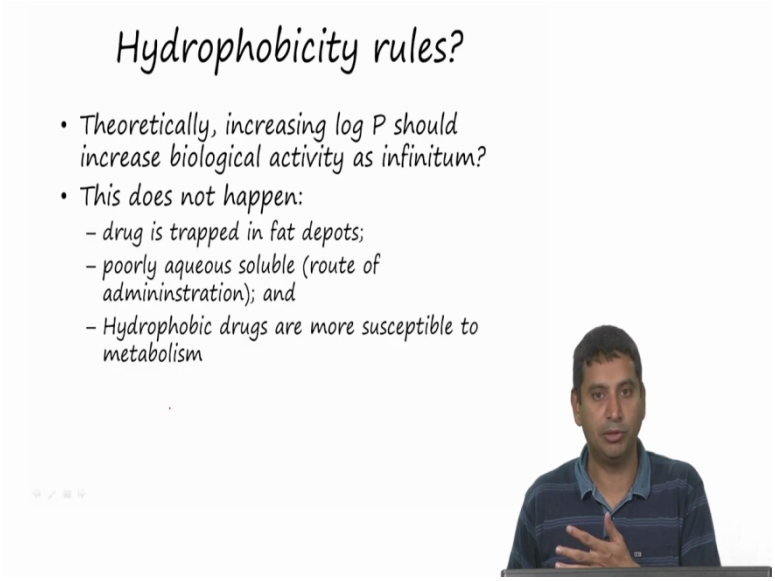
So this is understandable because you have serum binding to hydrophobic molecule but the problem is that what happens is that serum is going to bind to your molecule that is less molecule interacting with its target okay. So this is something that one needs to keep in mind while we are developing and designing new molecules right.

(Refer Slide Time: 17:02)



Second thing is that if you improve the log of P of the molecule, clearly it improves the activity of the drug which there are other examples of this, but the drug has to interact with the part of the enzyme let us say there is an enzyme, and the enzyme typically or the receptor typically we have studied before has a hydrophilic exterior and a more hydrophobic interior. So let us say this is the surface at which it has to react, now the drug has to come and interact over here and therefore this would be hydrophobic, therefore one would expect that the more hydrophobic the surface, the better it is interacting with the enzyme or receptor okay. So does it mean that hydrophobicity is the only thing that is important?

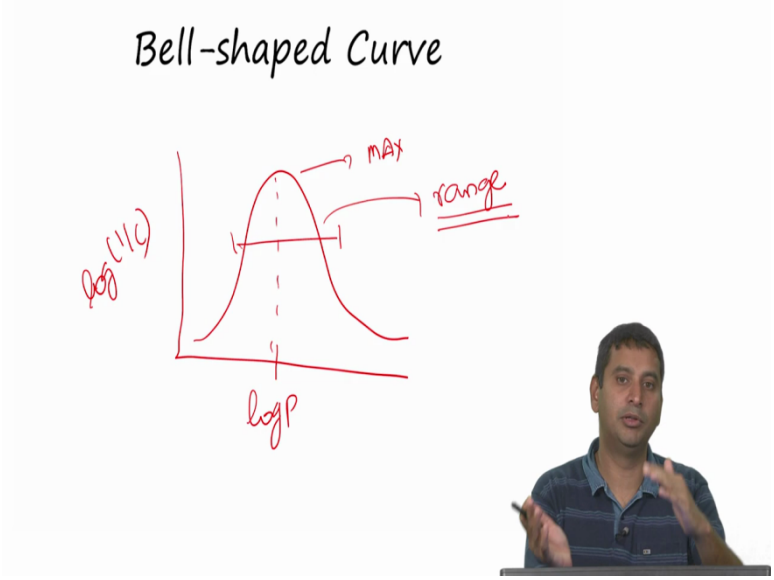
(Refer Slide Time: 17:45)



Well, theoretically according to our plot when you increase the log P value, biological activity should go on increasing ok. But in reality this does not happen because what happens is that in an NVO model the drug is actually trapped in fat depots, so there are generous fat depots in people where which is used during starvation and so on and the drug ends up beings trapped in these fat depots ok. Second thing is that depending upon the route of administration, so let us say you want to give an injection to a person, the injections are typically given in aqueous media, and if the molecule is extremely hydrophobic it is almost impossible to get it into aqueous media and therefore increasing log P value beyond a particular point is not desirable ok.

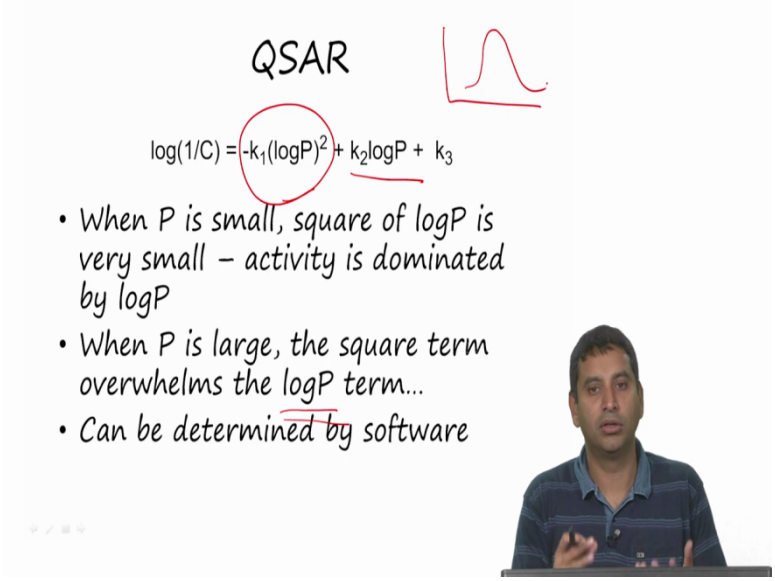
Now the other part is that hydrophobic molecules are also more susceptible to metabolism we have looked at previously ADME, which tells us how to understand metabolism but the more hydrophobic the molecule, the more comfortable it is to metabolism.

(Refer Slide Time: 18:39)



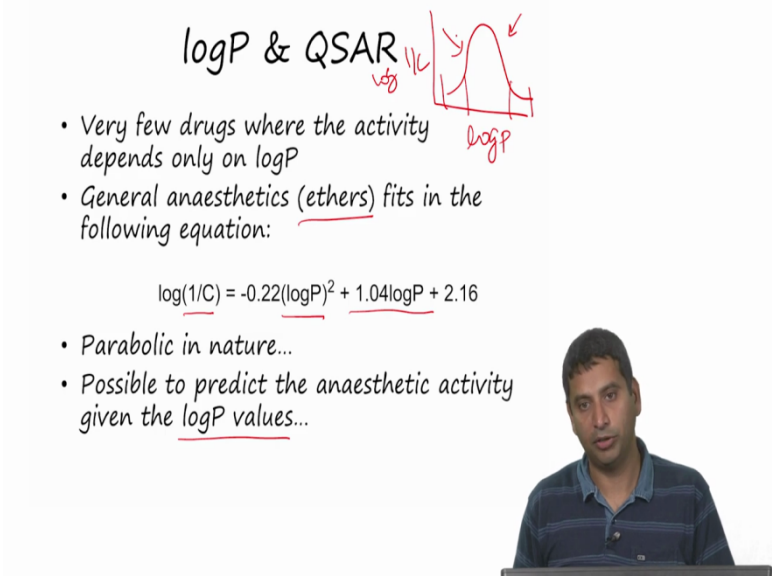
So in reality what we look at is we actually look at a bell shaped curve right. So if you look at log P versus log of 1 by C, so there is a particular point for molecules where the activity is the maximum. And these bell shaped curves are what normally people see, so you want to look for this range of log P values where you have good range of log P so that you can also developed a molecule further for NVO studies okay.

(Refer Slide Time: 19:08)



Now when we do the QSAR for this series of compounds, you will find that there is a square term in the molecule which is not unusual because you have a parabolic relationship okay. So but what it means is that when P value is small, then the square of log P is very small and the activity is dominated by the log P value ok, whereas when P value is large, the square of P becomes larger and it starts over helming the log P term and therefore you get a curve like this ok. Now this helps us understand what happens when you change log P over a series of compounds right. Now typically most of the field determined by software there are sophisticated softwares that are available which can help you with biblically understand and fitted to a particular model ok.

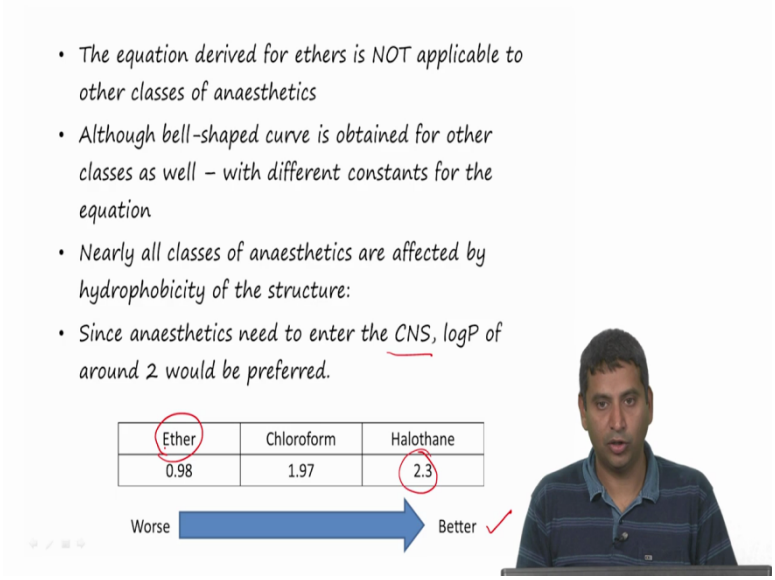
(Refer Slide Time: 20:03)



Now when we look at log P and QSAR, there are very few drugs whose activity depends only on log P although log P is an important parameter, and for example general anaesthetics fits well in the following equation. So you have typically ether which fits well in this equation, log of 1 by C is - 0.22 into log P the whole square + 1.04 into log P + 2.16. So if I plot this I would get a parabola like what we have seen previously okay. Now what happens is that because I am able to write out an equation which describes the activity, now because of this I can now find out when I get maximum activity.

So this helps us predict the anaesthetic activity given the log P value. This is my x-axis, if I have a very low log P then it is quite likely that the activity would be low ok, whereas conversely if I have a very high log P then it is quite likely that I would get a low activity and therefore the new set of compounds that I synthesised would have to have this intermediary log P values ok.

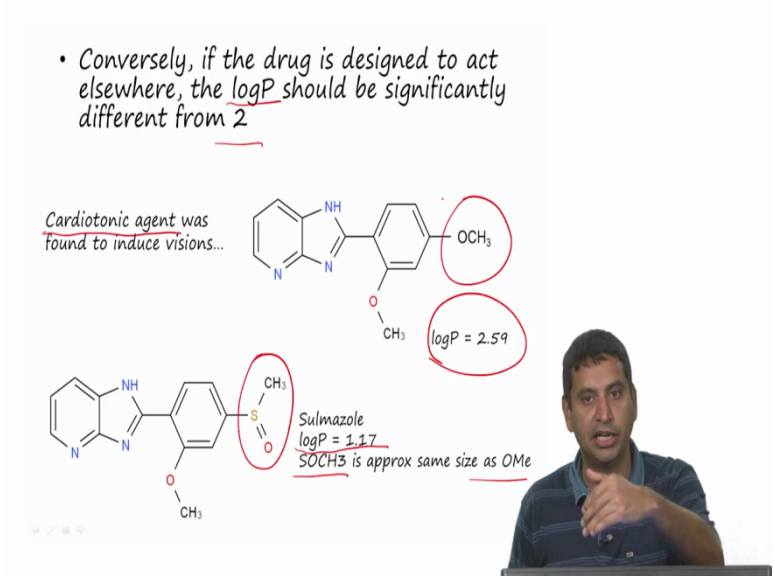
(Refer Slide Time: 21:12)



Now this equation that we derived for this particular set of ethers is not applicable to other classes of antibiotics, but there are typically bell shaped curves that are obtained for other classes as well but they have different constants in their equations ok. So nearly all classes of anaesthetics are affected by hydrophobicity of the structure and therefore it is possible to understand for each class of an aesthetic what is better it is getting into the central nervous system define here as CNS. And what we typically find is a log P of around 2 would be highly preferred okay. And this kind of a study has been applied to 3 examples I have seen here, where you see Halothane, chloroform and ether, the log P value goes from 2.3 to 1.97 to 0.98 ok.

So what we find in an anaesthetic values is this is actually the better compound compared to ether therefore now the next set of compounds if I want to prepare would be around this range weather values are between 1.5 and 2.5 okay. Keep in mind that if I go very high that also creates a problem.

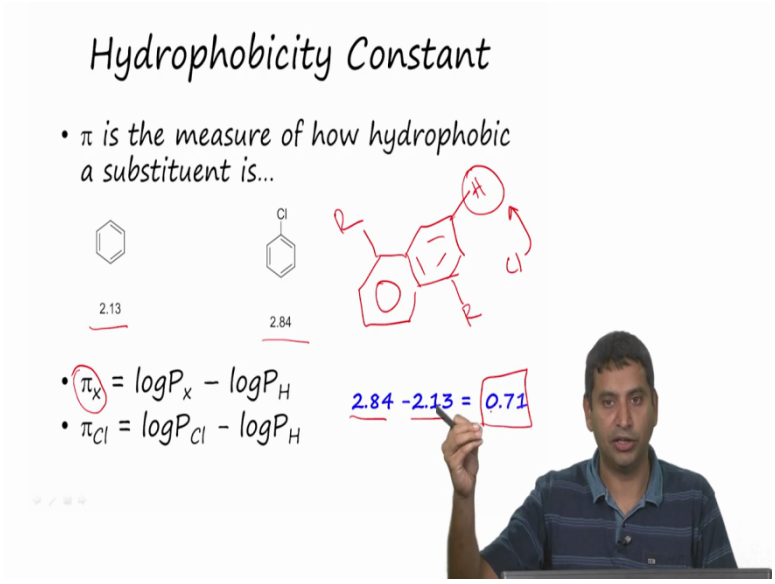
(Refer Slide Time: 22:17)



But it also helps us do is that if there is a compound with log P of around 2, one could predict that this molecule will enter the central nervous system and perhaps start acting the brain or in the spinal cord. Here is an example of something whose problem has been addressed by using quantitative structure-activity relationship. There is an example of or molecule which was considered as a cardio tonic agent, but it was found in many patients to induce visions ok. And so these visions are symptomatic of your drug interacting with the central nervous system. So now based on the previous study we know that log P of 2.59 can create trouble and it may help us get into the CNS therefore, next set of molecules that they synthesised one with the log P of 1.17.

And we look later about size but SOCH 3 is of similar size to OMe and therefore this molecule which has basically difference in the side chain here was found to have excellent activity cardio tonic activity but does not induce vision. So this is an example of how understanding or good structure-activity relationship can help us design better compounds.

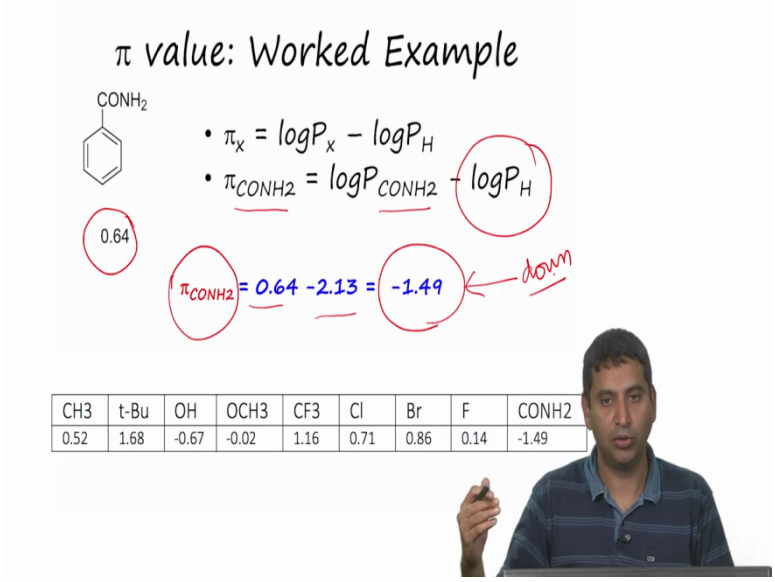
(Refer Slide Time: 23:26)



Now what we will do is, we will look at how to determine what is known as a hydrophobicity constant okay. So let us start with benzene, so if I want to study what happens when I put a substituent, the example I am taking is chloro benzene. So here benzene's log P value is 2.13, whereas chloro benzene is 2.84, now I define a parameter called as Pi which is basically nothing but the difference between the log P of the substituted benzene and benzene ok. So here in this particular case the Pi of Cl would be log P of CL minus log P of H right, which is basically 2.84 - 2.13 and that gives me a value of 0.71. Now the interpretation from the 0.71 is that let us say I have a complicated structure with many substituents, let us say I have R groups here and so on.

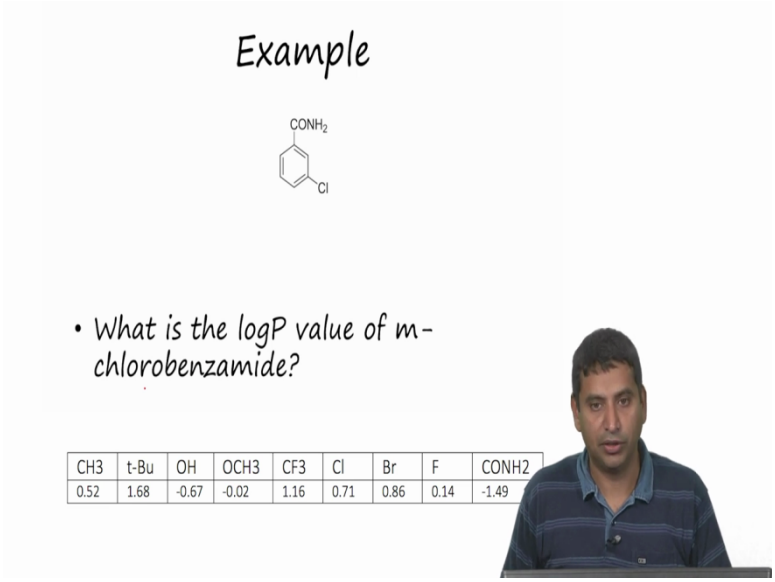
Now if I substitute one of these positions which has a hydrogen with Cl okay, I can suggest that log P value may change by around 0.71 ok. So therefore this hydrophobicity constant is very useful to understand how the subsequent affects the log P value okay.

(Refer Slide Time: 24:43)



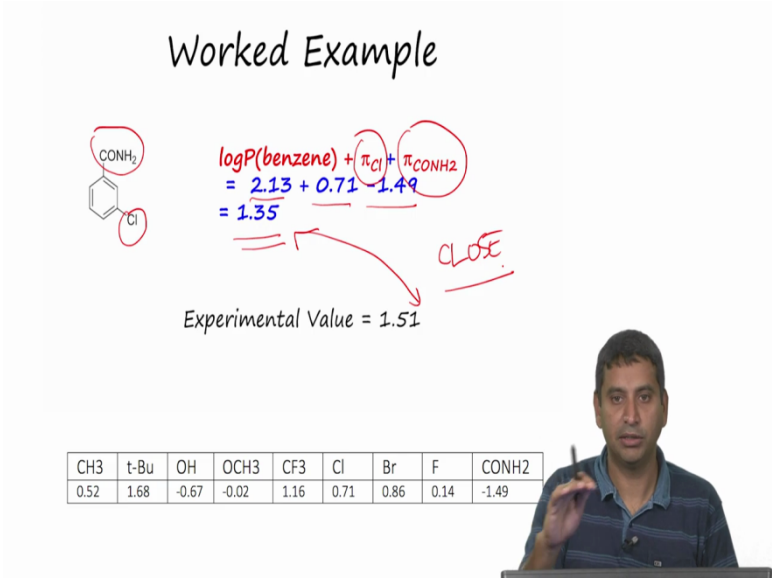
So here we look at another world example, the Pi of an Amide, the amide here is CONH2 can be determined by subtracting out the log P of benzamide from log P of benzene. So here value would be 0.64-2.13 and I get a value of - 1.49 ok. So if I introduce a subsequent CONH2 on benzene, it is likely that the hydrophobicity would go down that means the molecule will become more soluble in water ok.

(Refer Slide Time: 25:22)



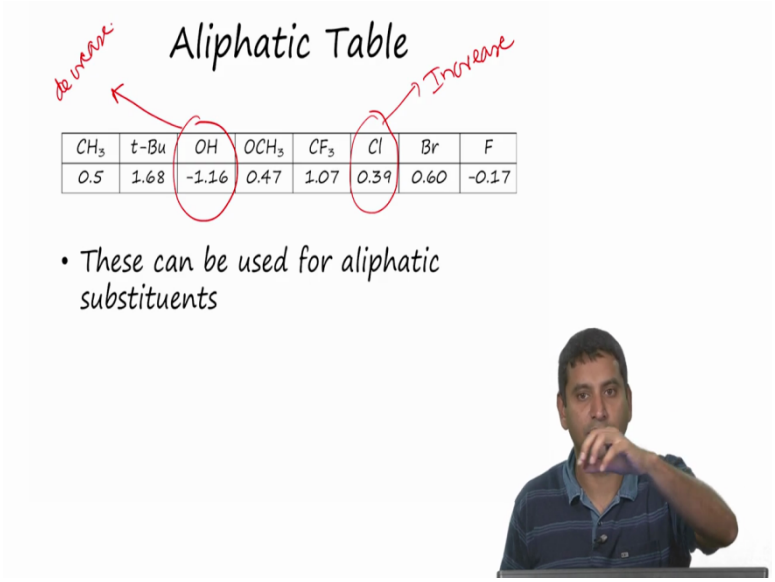
So let us look at an example now, here is a table that I have given below, what is the log P value of metachlorobenzamide?

(Refer Slide Time: 25:30)



Now if you want to look at it, the experimental value is 1.51, now let us look at when we take the values from what we have determined okay. So log P of benzene so benzamide consists of benzene, a chloro group and an amide group. So what I would do is I will start with Pi of Cl, Pi of CONH2 and then add it to log P of benzene, so this is nothing but 2.13 + 0.71 - 1.49 and the value comes to 1.35 ok so if I compare these 2 values, these are close to one another and therefore I have some predictive value by using Pi.

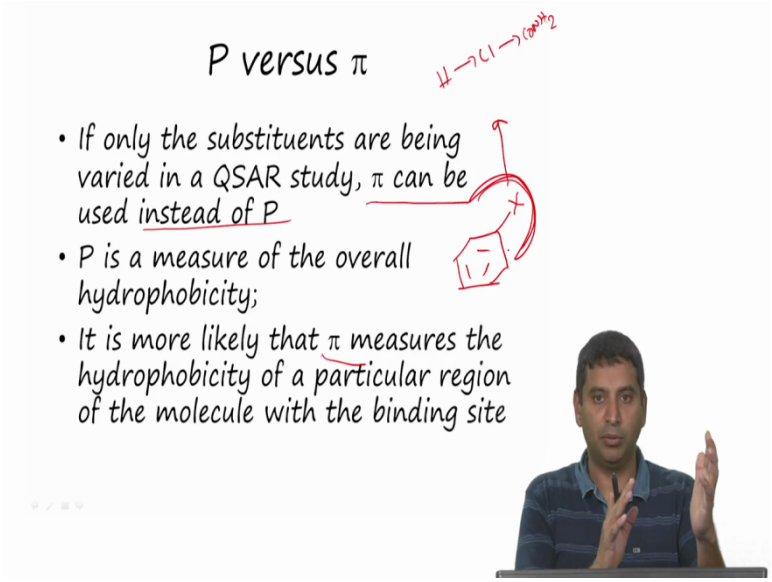
(Refer Slide Time: 26:12)



Now a similar table can be constructed for aliphatic compounds, and if I change the subsequent on the aliphatic compounds let us say I put a Cl on let us say a long chain alkane, if I put a Cl then it is going to increase the hydrophobicity by a value of 0.39. Conversely if I

put on hydroxyl group it is going to decrease the log P value by 1.16 right. So therefore by using these 2 parameters one can understand what is the influence or one can predict the influence of these functional groups on the hydrophobicity of the molecule.

(Refer Slide Time: 26:54)



Now we have looked at log P values, we have also looked at Pi, they are related obviously but they are very interesting in their use right. If only substituent are being varied in a QSAR study, Pi can be used instead of P that is no problem. So let us say I take benzene, I vary from chlorobenzene to bromobenzene to Phlorobenzene to benzamide to benzoic acid and so on, then I can use Pi instead of P ok. P is actually a measure of the overall hydrophobicity of the molecule okay, so we already looked at how to determine log P or P, it is a partition coefficient and therefore it is a measure of the entire hydrophobicity of the molecule whereas, Pi gives us locally what is the hydrophobicity.

So if we are looking at a particular site over here, let us say this X over here, now what happens if I change X from hydrogen to chlorine to let us say CONH2 and so on, this will help us understand locally what happens to the binding effect or to the interactions that are possible in this area ok. So both of them are very useful but they have to be used in the appropriate manner.