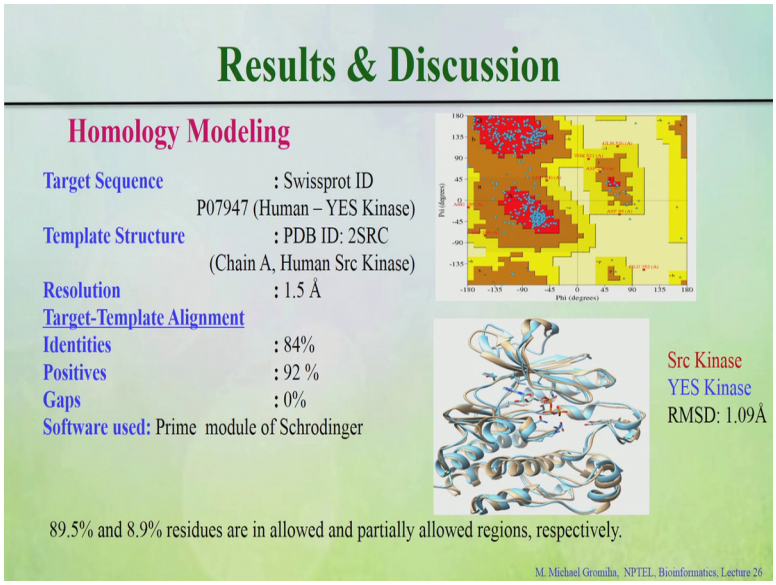
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 26b
Virtual screening II

(Refer Slide Time: 00:19)

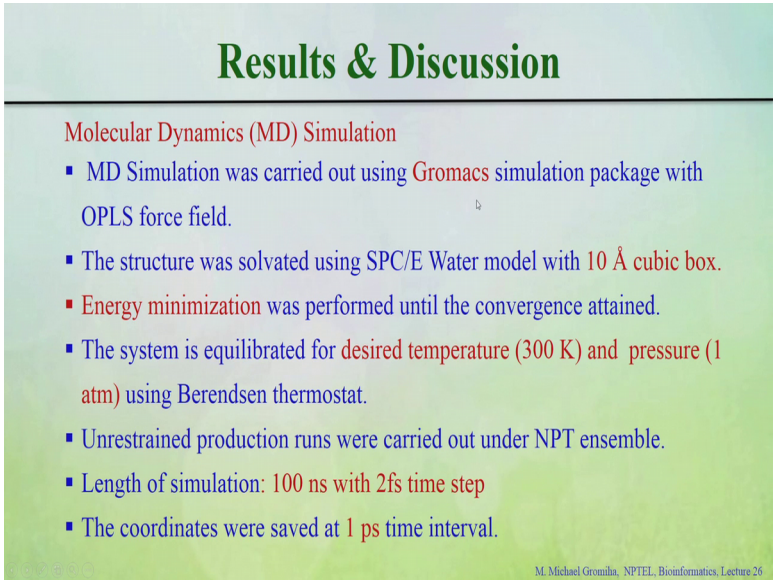


In the protein side, as we discussed earlier; so the sequence identity, very high identity is 84, positive is 92 percent. In this case you can do the homology modeling; so, you can see the homology modeling. Here this is the one is SRC Kinase and the blue one is the C-Yes kinase and RMSD is 1.09 angstrom; so, what is RMSD?

Student: Root mean square deviation

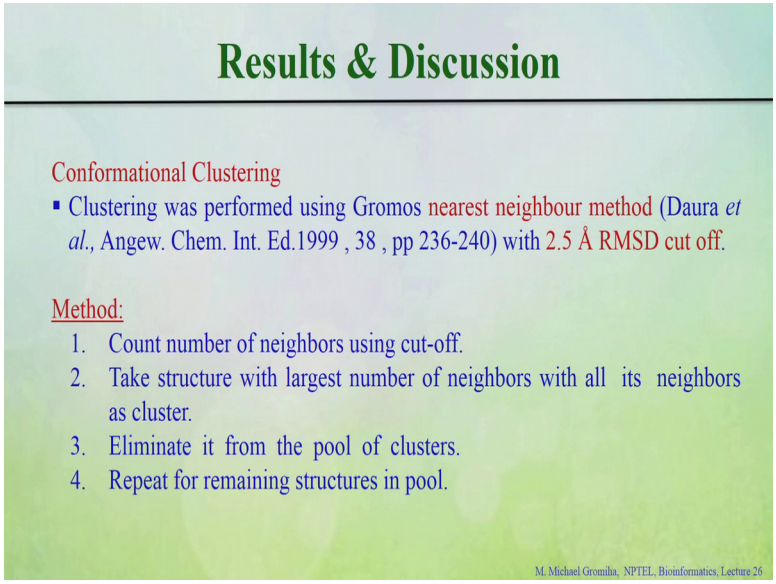
Root mean square deviation; how far these two structures we need superimpose, they are vary from each other. So, this is 1.09 angstrom; so we can see Ramachandran plot most of the cases in the allowed region. So, you can see the structures; they can validate and say the structures are in good shape. Now once this is done; you have the one static structure, now we need to do some dynamic structures; some different conformations. In this case we can do model dynamic simulations; so, what the MD works, how MD works? So, you can apply some temperature and pressure and the different time frame; you will get different types of structures.

(Refer Slide Time: 01:17)



So, we can use various packages available with the literature like Amber or the Gromacs; different force fields. You can simulate the particular protein and at their various time frame; you can get the conformation. So, you can see once we get several conformations, these are the conditions used for getting these structures; so finally, we use sampling.

(Refer Slide Time: 01:38)

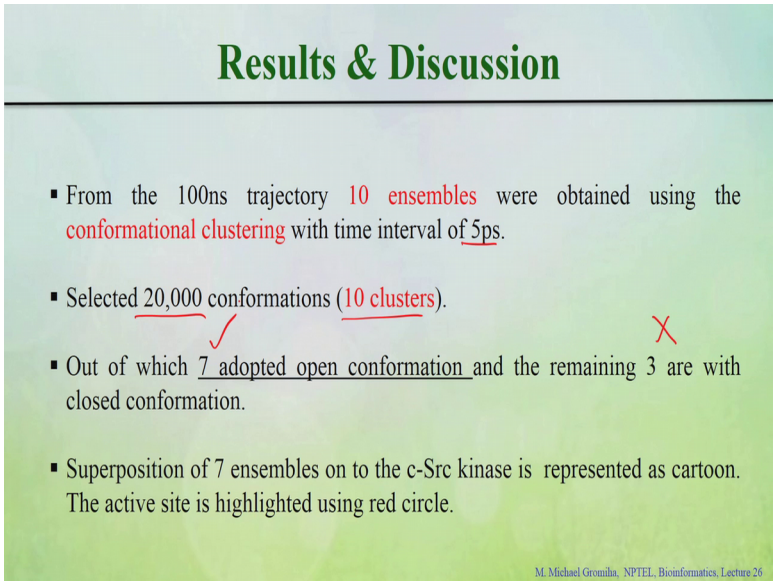


Once we do with the different conformation, but the conformation clustering; you can cluster only using nearest neighbour method. What they do? First they contact; count the number of neighbours, any specific cut off we can count the neighbours. Then they take

the structure with largest number of neighbours; with all the neighbors as a cluster, similar type of cluster; this were the how the clustering works.

Then eliminate it from the pool of clusters and the remaining the repeat for the remaining clusters in this pool. So, likewise you can create several clusters and from this clusters; you can identify the sample structures; you can use it for the docking.

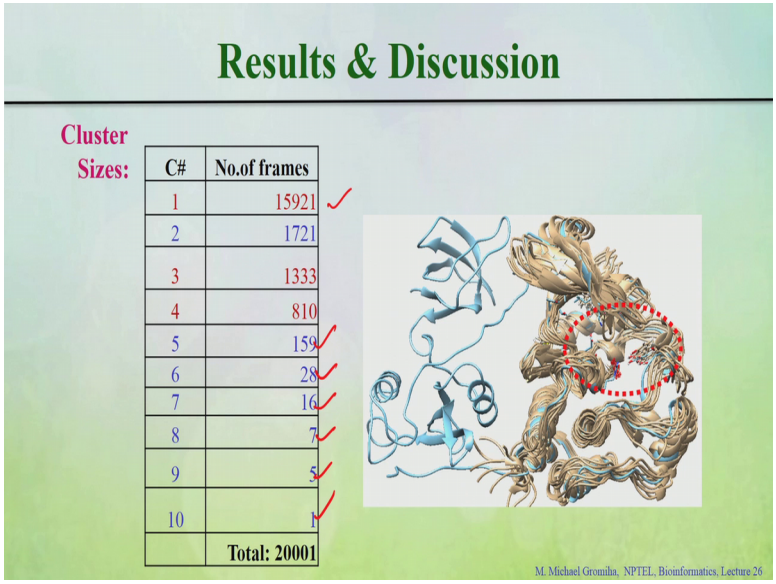
(Refer Slide Time: 02:23)



So, if you use the 10 nanosecond trajectory within ensembles; you can, so get the 20000 conformations with the interval of 5 picoseconds because even if we 5 picoseconds go with the 100 nanoseconds; you will get about 20000 conformations; 20000 conformations gives you the clustering techniques in 10 clusters. Take the protein, do the MD simulations with different time frame right for example, 5 picoseconds get conformations and like arrange the conformations in different clusters.

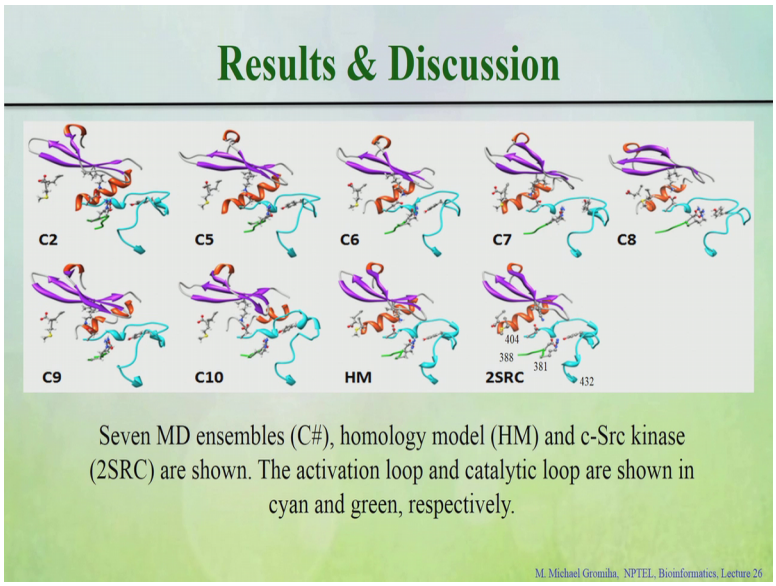
We have 10 clusters and we look at the 10 clusters; 7 are open conformation and the 3 are in the close conformation. So, if it is a open conformation then the ligand go and bind; if it is a closed conformation; we cannot bind, so this is not good for the docking.

(Refer Slide Time: 03:08)



So, that is fine; so if you see this is the cluster number and this is number of a frames, some of them; these three are the closed conformation. So, we take all the open conformation data.

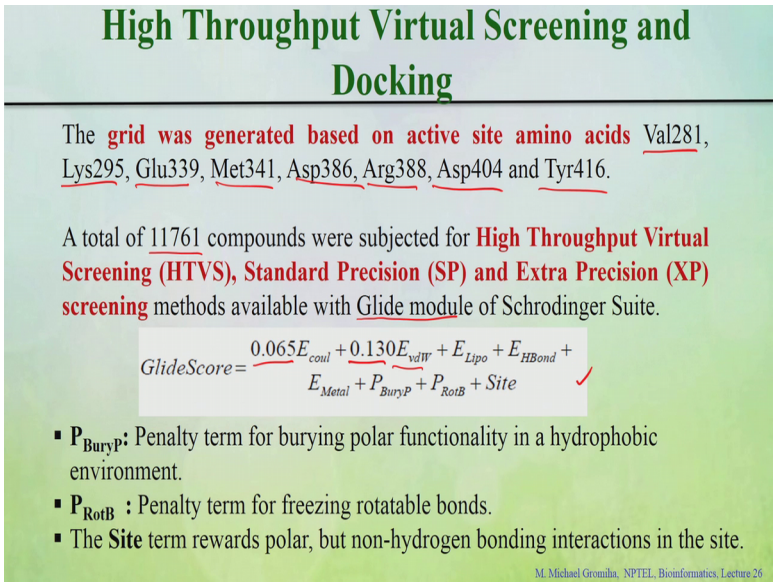
(Refer Slide Time: 03:18)



So, we have a 7; 1, 2, 3, 4, 5, 6, 7 you can see C 2; C 5, C 6 because this is fine; here you can see the superimposition of all the seven structures, some of them are; this is the ligand binding side; you can see the superimposition of the structures. So, 7 conformation you can see the seven structures 1, 2, 3, 4, 5, 6, 7.

This is a homology modeling one and here this is the actual one. Now we have 9 structures; so, protein side we have now we had sets 9 structures and the ligand side how many ligands we have? Around 11000 ligands because 5000 from this library and some decoys and the known actives.

(Refer Slide Time: 04:02)



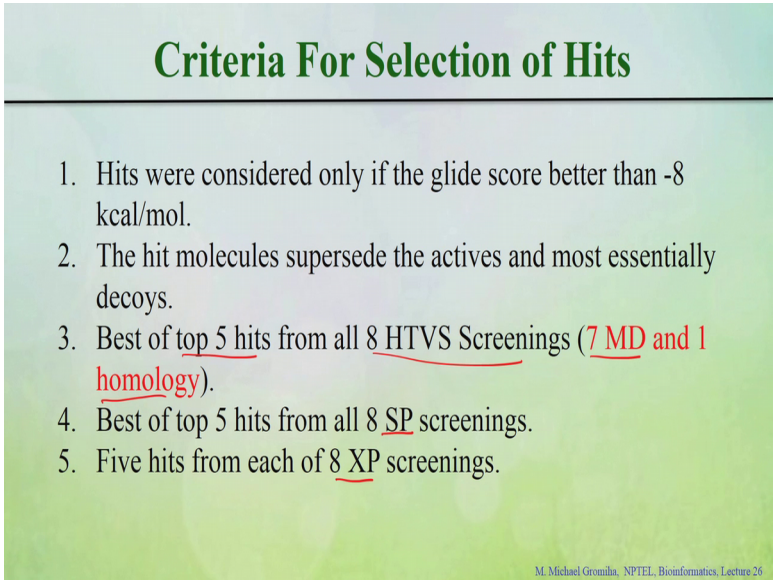
So, now we can do the docking; so, if you know the protein. So, we know the active site now there are several residues in the active site. For example, these are the main residues valine 281, lysine 295, glutamine 339; 41, 86; arg 88, aspartic acid 404, tyrosine 416. So, now, we can generate a grid around this active site amino acid residues because the ligands probably they potentially they bind only that region. You can generate a grid and we can use all these 11000 compounds to dock; kind of virtual screening.

So, in this case we can do the high throughput virtual screening or you can do a standard precision and the extra precision available. They are available in the glide module; in this Schrodinger then finally, they get the score using this equation. Glide score is given as they have different components these are columbic interaction and we have van der Waals interactions; here this is a lipophilic interaction, this is hydrogen bonds, but this is the penalty time given for the burying the polar functionally in the hydrophobic environment, because, if you have the polar residues then mainly their surface interact with the residues of the surface.

If it is goes with the buried ones then we give some penalty. Here you can see the rotatable bonds and you can see the other interactions. So, we get these glide score to get the interaction between protein and the ligand. So most of the cases, we get the good score, but here they already optimized with some numbers. So, in this case we also have some chances; so that it may fail to identify the best hit also because it is optimized with some set of data.

So, now how to get the hits? So, we can use the glide score and we use with different types of screening. You can use the high throughput screening or with the standard precision and the extra precision.

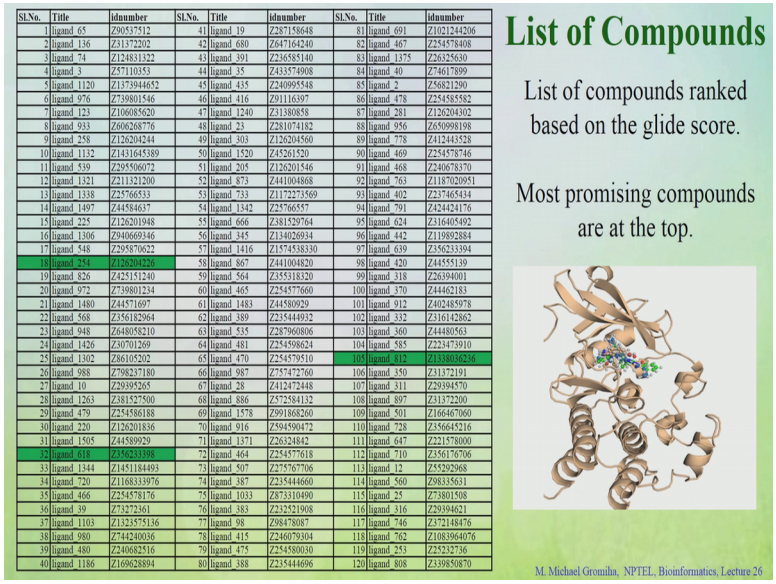
(Refer Slide Time: 06:10)



So, each case you can get some data and you can rank all the data. So, you can see the best of top 5 hits from all the screen methods that is a 7 MD and 1 homology model structures. So, 8 methods right and 8 screenings you can get the 5 hits.

So, we will get the 40 hits for the virtual screening; likewise 40 from this standard precision and 40 from the extra precision; so, totally we will get 120 ligands.

(Refer Slide Time: 06:36)



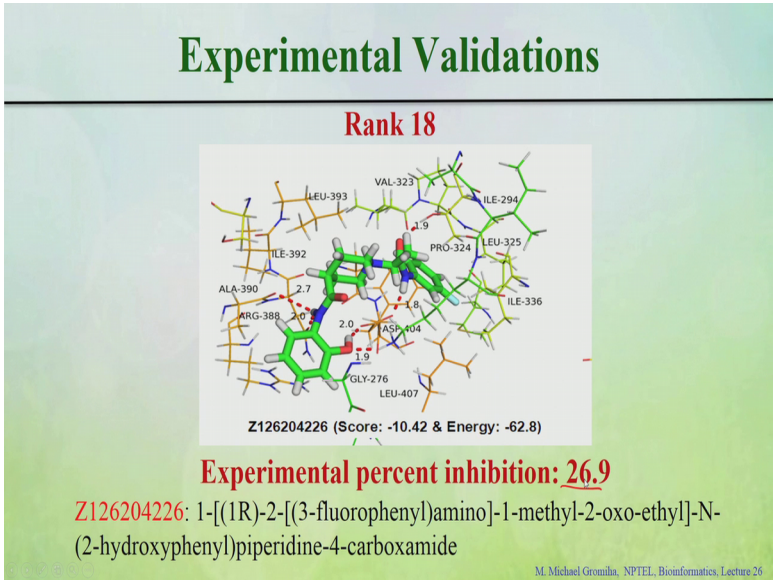
Sl.No.	Title	idnumber	Sl.No.	Title	idnumber	Sl.No.	Title	idnumber
1	ligand_65	Z90537512	41	ligand_19	Z58718648	81	ligand_691	Z1021244206
2	ligand_136	Z31372202	42	ligand_680	Z64716240	82	ligand_467	Z25578408
3	ligand_74	Z124831322	43	ligand_391	Z236585140	83	ligand_1375	Z26325630
4	ligand_3	Z57110353	44	ligand_35	Z433574908	84	ligand_40	Z74617899
5	ligand_1120	Z1373944652	45	ligand_435	Z240995548	85	ligand_2	Z56821290
6	ligand_976	Z739801546	46	ligand_416	Z91116397	86	ligand_478	Z254585582
7	ligand_123	Z106085620	47	ligand_1240	Z31380858	87	ligand_281	Z126204302
8	ligand_933	Z606268776	48	ligand_23	Z281074182	88	ligand_956	Z650998198
9	ligand_258	Z126204244	49	ligand_303	Z126204560	89	ligand_778	Z412443528
10	ligand_1132	Z1431645389	50	ligand_1520	Z45261520	90	ligand_469	Z254578746
11	ligand_539	Z295506072	51	ligand_205	Z126201546	91	ligand_468	Z240678370
12	ligand_1321	Z211321200	52	ligand_873	Z441004868	92	ligand_763	Z1187020951
13	ligand_1338	Z25766533	53	ligand_733	Z1172273569	93	ligand_402	Z237465434
14	ligand_1497	Z44584637	54	ligand_1342	Z25766557	94	ligand_791	Z424424176
15	ligand_225	Z126201948	55	ligand_666	Z381529764	95	ligand_624	Z316405492
16	ligand_1306	Z940669346	56	ligand_345	Z134026934	96	ligand_442	Z119892884
17	ligand_548	Z295870622	57	ligand_1416	Z1574538330	97	ligand_639	Z356233394
18	ligand_254	Z126204226	58	ligand_867	Z441004820	98	ligand_420	Z44555139
19	ligand_826	Z425151240	59	ligand_564	Z355318320	99	ligand_318	Z26394001
20	ligand_972	Z739801234	60	ligand_465	Z254577660	100	ligand_370	Z44462183
21	ligand_1480	Z44571697	61	ligand_1483	Z44580929	101	ligand_912	Z402485978
22	ligand_568	Z356182964	62	ligand_389	Z235444932	102	ligand_332	Z316142862
23	ligand_948	Z648058210	63	ligand_535	Z287960806	103	ligand_360	Z44480563
24	ligand_1426	Z30701269	64	ligand_481	Z254598624	104	ligand_585	Z223473910
25	ligand_1302	Z86105202	65	ligand_470	Z254579510	105	ligand_812	Z1338036236
26	ligand_988	Z798237180	66	ligand_987	Z75747760	106	ligand_350	Z31372191
27	ligand_10	Z29395265	67	ligand_28	Z412472448	107	ligand_311	Z29394570
28	ligand_1263	Z381527500	68	ligand_886	Z572584132	108	ligand_897	Z31372200
29	ligand_479	Z254586188	69	ligand_1578	Z991868260	109	ligand_501	Z166467060
30	ligand_220	Z126201836	70	ligand_916	Z594590472	110	ligand_728	Z356645216
31	ligand_1505	Z44589929	71	ligand_1371	Z26324842	111	ligand_647	Z221578000
32	ligand_618	Z356233398	72	ligand_464	Z254577618	112	ligand_710	Z356176706
33	ligand_1344	Z1451184493	73	ligand_507	Z75767706	113	ligand_12	Z55292968
34	ligand_720	Z1168333976	74	ligand_387	Z235444660	114	ligand_560	Z98335631
35	ligand_466	Z254578176	75	ligand_1033	Z873310490	115	ligand_25	Z73801508
36	ligand_39	Z73272361	76	ligand_383	Z232521908	116	ligand_316	Z29394621
37	ligand_1103	Z1323575136	77	ligand_98	Z98478087	117	ligand_746	Z372148476
38	ligand_980	Z744240036	78	ligand_415	Z246079304	118	ligand_762	Z1083964076
39	ligand_480	Z240682516	79	ligand_475	Z254580030	119	ligand_253	Z25232736
40	ligand_1186	Z169628894	80	ligand_388	Z235444696	120	ligand_808	Z339850870

We can see the 120 ligands; so, we get the MD derived structures 7; homology model structure 8; other side we have the ligands, we have three different types of screening methods. Each method we take the top 5 so; that means, 8 proteins into 5 equal to 40 each different screening method 40, 40, 40 that give 120.

So, this is a 120; so if the Schrodinger could get the best one. So, the top one should be at the top most one, but if you look into this experimental data; these are the ligands, we showed some potential inhibition. Like a rank number 18, 32 and 105; so, you can see this here these are the compounds which are identified to show some set of inhibition.

They are at these different ranks 18, 32 and 105 at least we can see all these compounds in the top 120 hits; among these 2.2 million. Because we rank from 2.2 million, we get 120; among this we can see the three compounds which are in the list of a top 120.

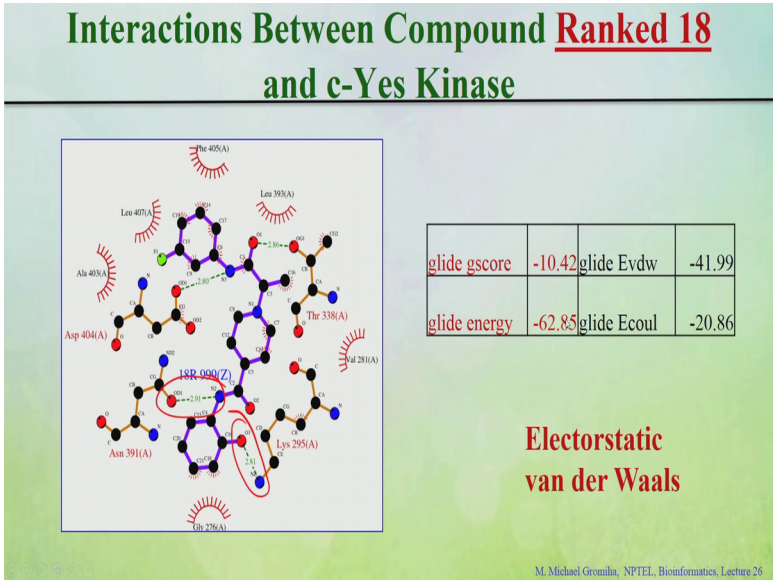
(Refer Slide Time: 07:57)



Now, we will see how they interact and how they could be a potential lead compound. So, take the one which rank 18; this is the one rank 18 this showed the inhibition rate of 26.9. So, you have the glide score of 10.42 and the energy of minus 62.8, so what is the difference between score and energy? So, the score means they have this empirical equation to get the value; by combining different factors to the put the empirical equation and finally, get a number that it is called the glide score.

Energy they calculate the van der Waals energy and the electrostatic energy like the ampere force field, then they give the total energy of this thing; this the energy using the glide. So, here the inhibition is 26 percentage and then see this is the compound and you can see the surrounded with the protein amino acids in the protein.

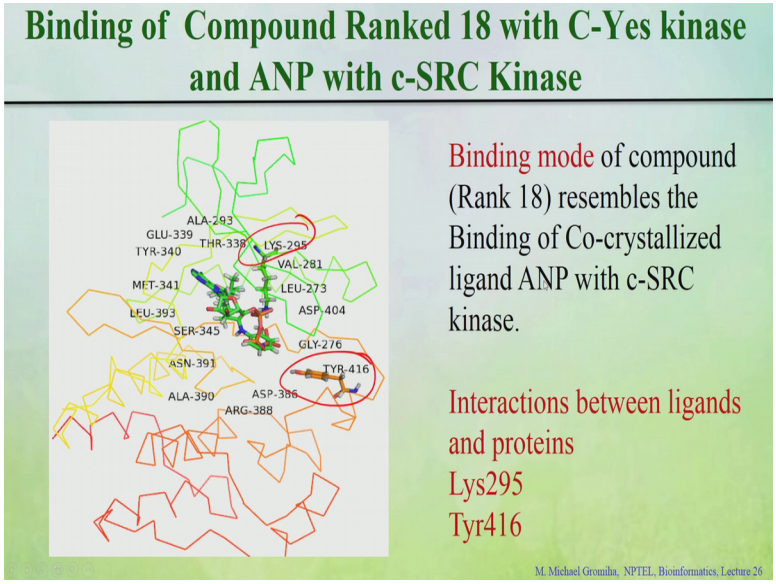
(Refer Slide Time: 08:42)



Now, some of the interactions I will show here; you can see these hydrophobic interactions. And some of them are having the hydrogen bonds, you can see this is the OD 1 and this N 2. They are having this electrostatic interaction and here also you can see the interaction and the here are the hydrophobic interactions the CH 2 groups.

So, these interactions are mainly dominant electrostatic and van der Waals interactions. So, glide score is minus 10 and the glide energy is 62; we check the contribution van der Waals is 41 and the columbic energy is minus 21; if we add up these two, we will get the 62.

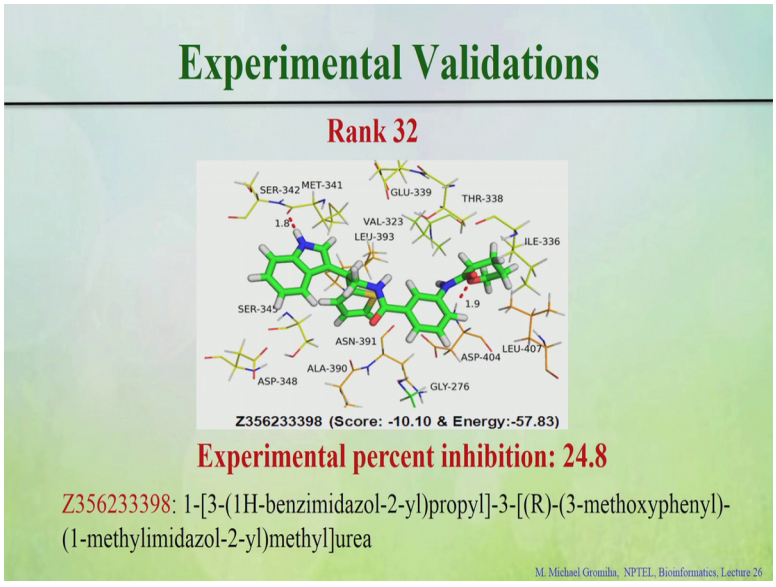
(Refer Slide Time: 09:18)



Then we can compare; this is the compound we obtained from C-Yes Kinase with our molecule; this is the zinc id for the compound. If you see the crystal structure these are similar type of compound available with the C-Yes Kinase and ANP.

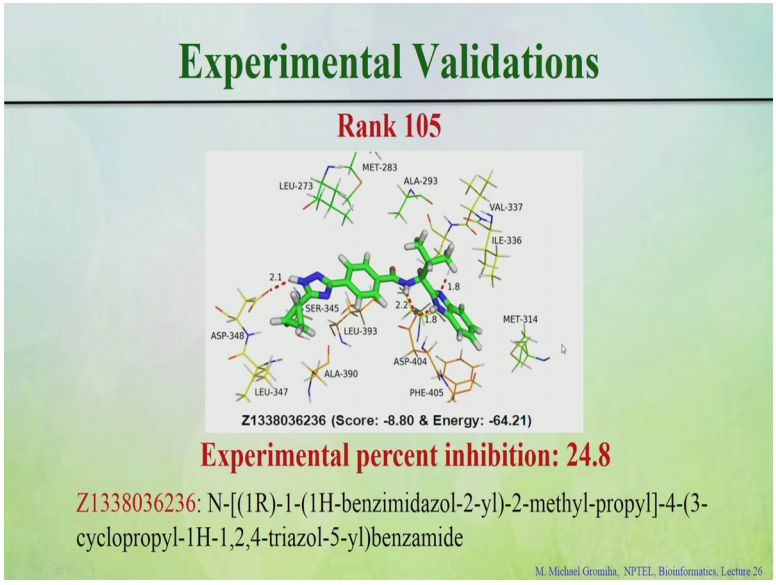
So, in this case you can see the interactions mainly the lysine 295 this is a interactions and tyrosine 416; those three interactions are main interactions when this co crystallize with the ligand ANP. So, you can see similar type of interactions with the co crystallize structure as well as the one, we identified using virtual screening.

(Refer Slide Time: 09:54)



Then this is a another one this rank 32 and here the inhibition is 24.8 percent; you can see the type of interactions; this is the ligand and you can see how they are interacting with the amino acids in a particular protein.

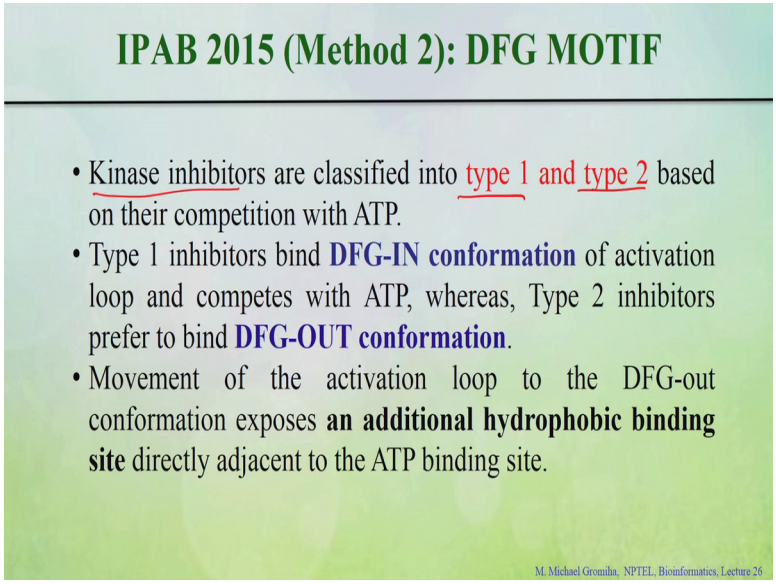
(Refer Slide Time: 10:11)



So, this is a another one rank 105; this is also inhibition 24.8 percent and here you can see the interaction electrostatic interactions, as well as the hydrogen bonds and the other hydrophobic interactions for this particular ligand. So, if the score is minus 8.8 and the energy is almost similar that is minus 64.21 kilocal per mole. So, energy is one step; so here what we did? You take the protein site; take the ligand, get the properties and reduce the number do the screening and then we identified the proper ligands.

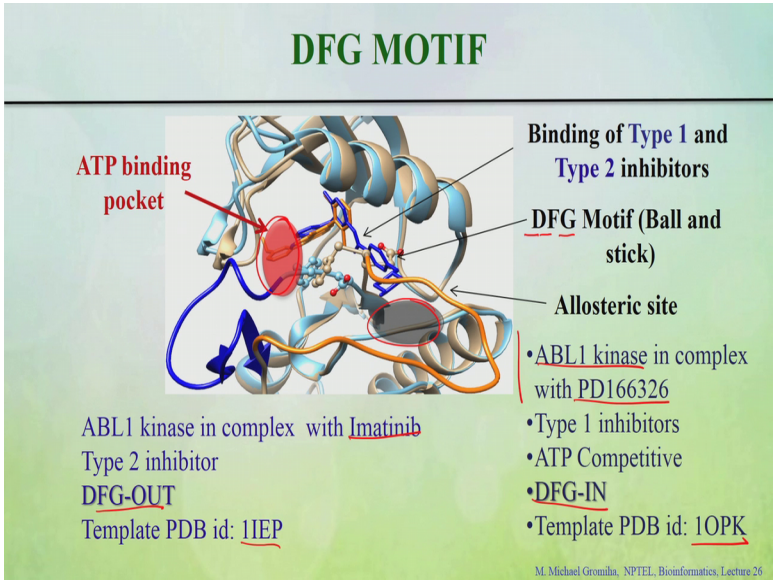
It is also possible to look into this binding site information for example, if they have the Kinase inhibitors.

(Refer Slide Time: 10:55)



Because we also in the first competition, we also learned the proper potential compounds. So, they classified in two different types one is type 1 and type 2 based on the competition with ATP. So, in this protein if you see there are two different conformations one is the DFG-IN.

(Refer Slide Time: 11:11)



You can see here this is a DFG motif; D is aspartic acid, F is a phenylalanine, G is the glycine you can see here.

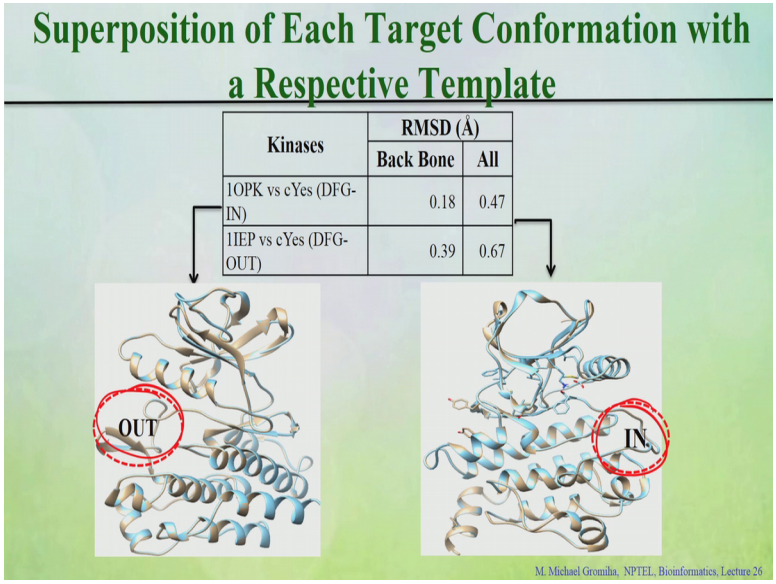
So, this DFG motifs either it is in the ATP binding pocket or they block ATP binding pocket; now that is the allosteric site here that is the competitive inhibitors. So, this protein they have two different types of motifs; one is the DFG-IN conformation. In this case type 1 inhibitors bind with in conformation and type 2 inhibitors; they bind with DFG-OUT conformation.

This information you can obtain from the literature and the movement of the activation loop to this out conformation also exposes an additional hydrophobic binding, directly adjacent to the ATP binding site; that is another hydrophobic binding site adjacent to the ATP binding sites.

So, in this case if you know the conformation DFG in and DFG out; even if the homology is less, structurewise you can get a similar structures and you can model the structures and mainly focus on the conformation. So, here this is the two conformation; one is a DFG in, one is DFG out conformation. This is example; one example is ABL 1 link kinase, this is the complex with the ligand. This is a kind of type 1 inhibitor that is ATP competitive, this in DFG in conformation.

So, this is the PDB id for that particular protein; likewise this is another complex ABL 1 Kinase with Imatinib, this is a type 2 inhibitor; this is in DFG out conformation, this is the template. If you see these two templates, the sequence identity is less but structurally these binding pockets are having similar structures.

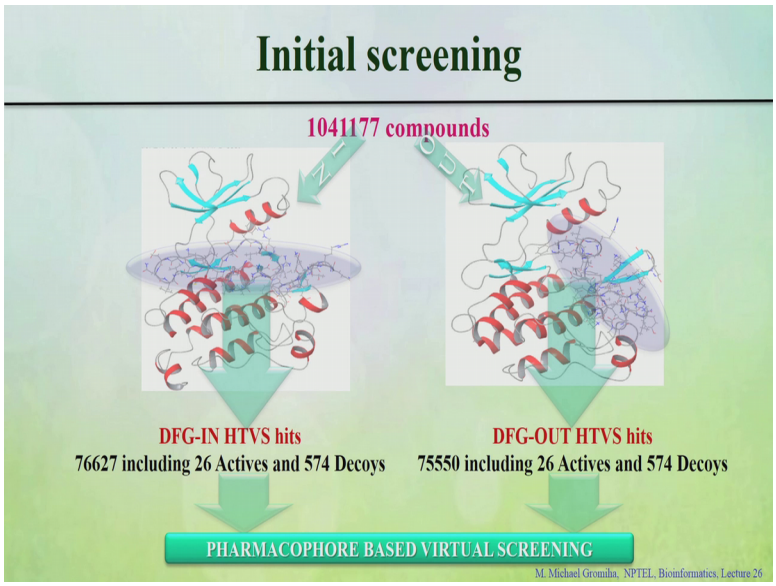
(Refer Slide Time: 12:55)



So, then we can easily model the structures with the c yes, you can see the RMSD is very less at this particular region; this is the in conformation and this out conformation. So, now, here we have only 2 structures; earlier we used 8 structures; 7 from MD and then 1 from the homology modeling.

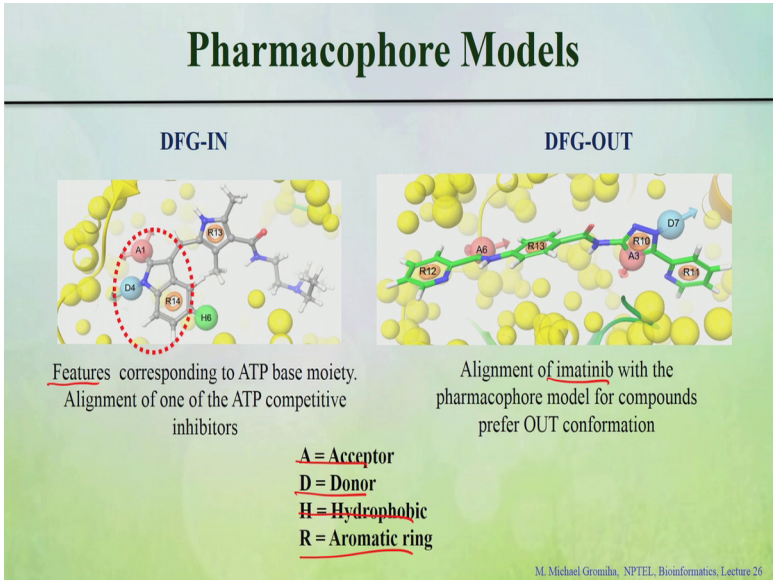
Then we use a screening, so here now we have the two conformation the in conformation out conformation. And as I discussed earlier we need to prepare the same; you can prepare the ligands and get the properties and we can filter with the average property values and then we use a tanimoto coefficient. So, we will do the same then do with the screening.

(Refer Slide Time: 13:39)



You can the in conformation we will get about a 76627 hits; likewise if we take the DFG out conformation; so, we can get about 75000 hits. Then we add up together, now we will get about 1 million compounds; then what we did? So, those we are subjected to pharmacophore based model. What is a pharmacophore based model? If you see a protein; we can see the active site and see the moiety.

(Refer Slide Time: 14:06)

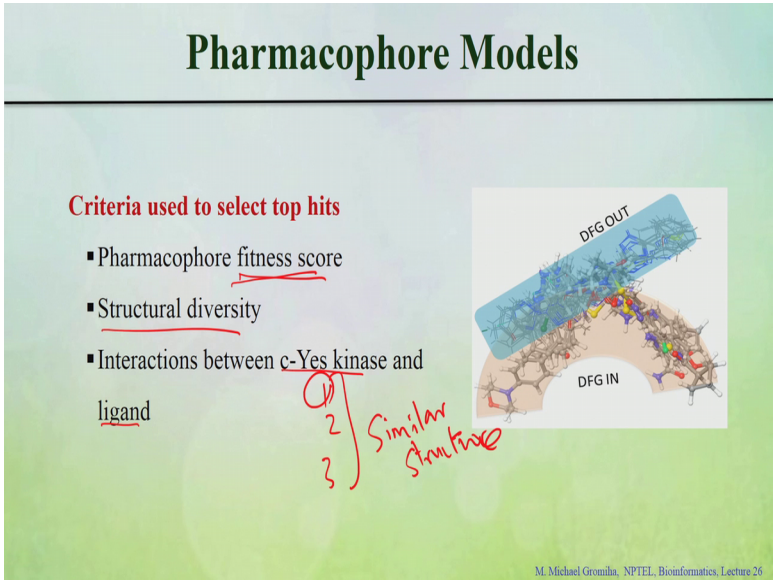


Around this particular binding site; so, what are the features corresponding to these ATP based moiety? So, in this case; if you see what are the possibilities of different types of compounds? For example, here A is the acceptor, D is the donor and H is a hydrophobic and R is the aromatic ring.

So, these are another example for the DFG out, this is the alignment of imatinib and you can see the a pharmacophore model for the compounds, which we prefer the out conformation. So, now you have the ligands; you screen these ligands here and see which ligands they fit with this pharmacophore around the active site; based on the acceptors, donors, hydrophobic as well as aromatic rings. So, here also we know that ATP compound inhibitors here also we know out conformation.

So, how we fit the ligands around the pharmacophore of this particular site? So, when you do this; then we will get different numbers and we can rank that numbers using the fitness score, you can see here this is the DFG in and you can see DFG out.

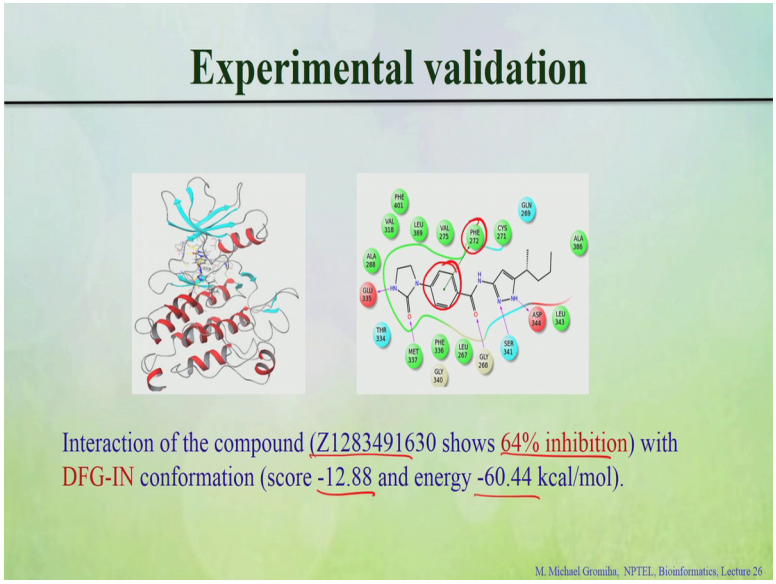
(Refer Slide Time: 15:19)



So, these are the ligands which fits the pharmacophore. So, first we can get the fitness score, we can screen; you can rank based on fitness score. Then go with the structure diversity for example, if you have the 5 ones; they are similar structures, high scores; then take only 1; for example, 1, 2, 3; so there are similar structure.

Then you take the first one; then take everything. So, then we can take the interaction between the C-Yes kinase and the ligand actual interactions. So, you have to use various criteria to identify the top fits; one is the fitness score based on the pharmacophore modeling and check the diversity of these compounds and then see the actual interaction between the ligand as well as the C-Yes kinase.

(Refer Slide Time: 16:19)

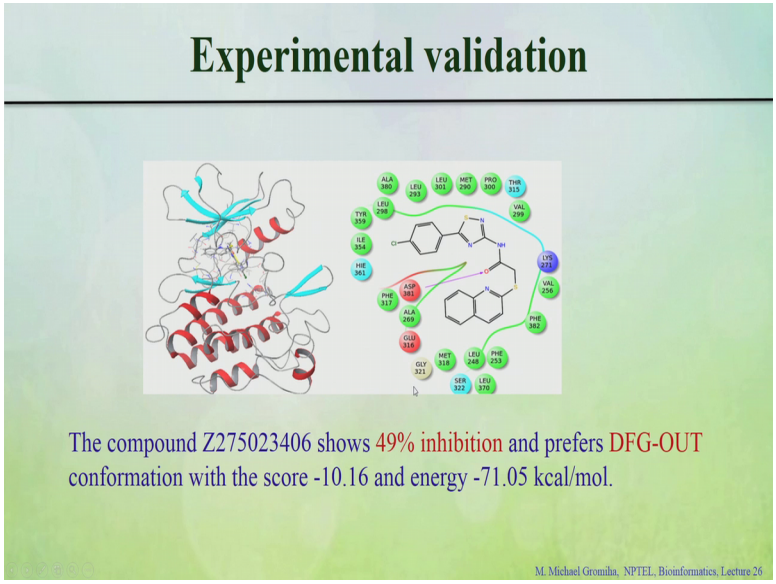


Once we do this; then we can have the compounds; in this case we could find two different compounds, the inhibitors it is very high.

In the first case, we prepared the protein, we prepared the ligand and we did the docking, but different screening, but we did not consider the conformation; whether it this is a binds with the ATP site or if it is the competitive site; so, we did not consider that. So, there we got the maximum inhibition of a 31 percent, but here now we considered the conformation and the type of inhibitors.

Now you can see these compound shows about 64 percent inhibition; we can get the score 12.88; the glides score and the energy of minus 6.44 kilocal per mol. So, this is from the DFG in conformation; likewise you can get the another data.

(Refer Slide Time: 17:13)



So, here this is showing 49 percent inhibition and prefers DFG out conformation. So, here the score is minus 10.16 and the energy of minus 71.05 kilocal per mol. So, if we see the structure here; so in this case you can see this is the pharmacophore, it contains a different a residues. And you can see the different types of interactions; you can see the hydrogen bonds, you can see the hydrophobic interactions see; phenylalanine 272, this is a ring; this can form these pi-pi interactions.

And you can see this is N H; with this one you can see the hydrogen bonds with the serine, you can see different types of interactions. These compound also you can see the different types of interactions from this negative charge residues or the polar residues as well as the hydrophobic residues. So, do the virtual screening; so what are the various steps we have to follow to identify the potential compounds from the pool of compounds?

Student: we need to have a library of ligands

So, first we need to have a library; so, ligand library we have. So, now if you identify any specific target we have to know the structure. So, two types you can do if you know the structure we can directly use the structure; if a structure is not known then what you have to do? We have to do homology modeling; in this case we consider two different aspects.

One aspect is we just get the sequence identity with the same family; similar same family; SRC kinase or C-Yes kinase; fin kinase same family we did the homology modeling for the sequence identity.

So, then we picked up the templates the same family; then we did a homology modeling based on the template. Then they can see the; validate the structure and then if it is fine you can use it. And the second aspect; how did we get the structure? Here we mainly consider the conformation. So; DFG in and DFG out conformation for the two types of inhibitors inhibitor type 1 and inhibitor type 2.

So, because they are experimentally shown that these type of inhibitors; they mainly bind with the particular site. In this case even if you to homology modeling; all these process, because they do not bind anywhere they bind only that particular place.

In this case, we can focus on that region with the high sequence identity. We get similar structures and focus on these two conformations; the two sites, mainly you can see I showed this here; DFG motif. So, is ATP binding packet and this allows to the site here we focus on these two sites. So, now protein side and homology modeling we can do; plus MD simulations, why do you need MD simulation?

Student: To get more conformation.

More conformation because see a only one conformation. So, possibly many ligands may fail because this case you can focus only one conformation. So, we may different conformations; so did MD simulations and then we get 20000 conformation; classified in a different clusters and picked up the samples from each cluster. Because we emitted the discarded the close conformation because close conformation ligand cannot bind.

So, we take only the open conformation; so now, the protein site we did two different aspects is done. Now our ligand site what we have to do? First we have to get the known actives from exisiting databases. So, we get the known actives from the exisiting database; then we get the properties for that actives. Then the whole library; first we have to prepare the ligands because lot of errors and other charges and ions and all. And then we calculate these properties and we compare with the known actives with some standard deviation.

If this known actives and the known ligands vary very much then we can discard, see if any particular deviation if you take any of the all the different properties they fit; then we can filter the make the first filter and get the compounds. And then what we do? Then we add some decays and some known actives as a control.

We need to know whether the final result you should identify all the actives, but should not get any decays. Then add up these compounds with the filtered compounds and finally, use the screening and we will get the compounds. Then we rank the compounds and we can get that; that is from the homology modeling.

In the second type, if you focus only on the these sites; the type 2 inhibitors and the type 1 inhibitors, at this different sites, then we focus on that particular region. Particular region you can develop the pharmacophore models and based on the type of a residues with respect to a hydrogen bonds and acceptors, donors, hydrophobic and aromatic rings; we can fit all the ligands and see which can fit at this model and you can select based on the fitness score.

So, if we get the fitness score we can rank the compounds; when its look at the compounds many of them are having similar structures. So, in this case if 1 fails; 2 will also probably fail, 3 will also fail; first 10 compounds, the high ranked compounds and 10 compounds are structurally similar. If the first one fails then the high possibility that all the 10 will fail. So, in this reason to avoid all this confusion; we take the similar structures and check the interactions; so take the best and discard all the 9.

Likewise, you can see the diversity to select the compounds; then once you select the different second refinement is over; when we select the compound then it is exactly see the interactions between C-Yes kinase and the ligand; how they interact then if you are comfortable, then we can rank based on this order. Finally, you can select these compounds; so, when we compare the experiments you can see the direct interaction between the ligand as well as with the amino acids. And you can see the good type of interactions for example, here this is 12.88 score and the energy is 60 and you can see 64 percent inhibition.

We can likewise for any targets; so we need to prepare the protein, prepare the ligand and first you collect the experimental information as much as possible. Either the activities or the binding mode or the binding pose or if any specific residues which are involved in

the active sites collect all the experimental information as possible. So, that you can finally, validate the final model; if this is not properly interact with that particular residues then you can discard. Then second we can validate with the MD simulations; do the different MD simulation structures. In this case also you can see whether this particular ligand can fit at the different conformations using MD.

But eventually for everything you need the experimental validation; this is very important to have the experimental validation. Once you do the validation, then you can see the percentage inhibition as well as the inhibitor constant like the IC50 values and so on. In this case, you can go with the next level; the inhibitor studies and the animal studies and then finally, studies and so on. So, what the advantages having virtual screening?

Student: less number of compound need to be is screened

We can you can narrow down the compounds to a less number of compounds, so it can be virtually screen. Otherwise instead of doing 2.2 million compounds; so in this competition; several groups they participated in this competition and different groups they showed the compounds which are totally different; not many compounds are overlapping in different groups. So, then this case they selected the compounds 120 compounds from different groups; if 5 groups means that about 600 compounds; it currently screened about 10000 compounds instead of 2.2 million compounds.

Well in that case they could find about 25 to 30 compounds, which are showing a high inhibition rate. So, instead doing 2.2 million compounds using the computational strategy, they used various methodologies. They use machine learning techniques, they use the pharmacophore based modeling, they used ligand based modeling and they used a structure based design. And based on different approaches, they identified the compounds finally, they could get some of the potential novel compounds as a lead compounds and these could be as a inhibitors.

So, we discussed about the docking, we discussed about the screening. So, we can also identify the potential inhibitors using the QSAR studies; what is a QSAR?

Student: Quantitative Structure Activity Relationship

Quantitative Structure Activity Relationship; that we will discuss in next class. If you have a large number of ligands with known affinity; for any particular target then we can use the QSAR studies to fit the model and using that we can identify new compounds that also you can experimentally verify, whether these compounds could lead as an potential lead compounds as any inhibitors; that we will discuss in the next classes.

Thanks for your kind attention.