**Algorithms for Protein Modelling and Engineering**
**Professor Doctor Pralay Mitra**
**Department of Computer Science and Engineering**
**Indian Institute of Technology, Kharagpur**
**Lecture 50**
**Protein Modification (Contd.)**

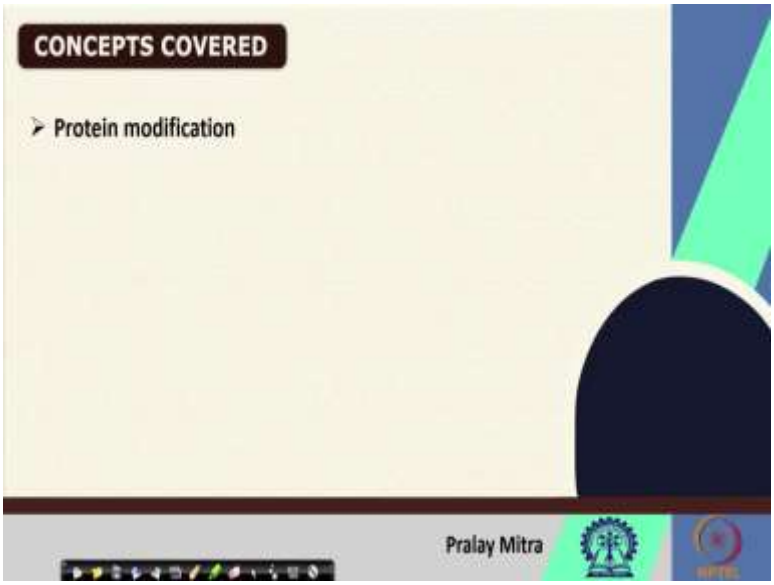
Welcome back to this class Algorithms for Protein Modelling and Engineering. So we are at the stage of protein modification. So as of now, as part of the protein engineering we started with the protein design. And then in the design we mentioned about the algorithm which will go for ab-initio protein design. We also mentioned that there are possibilities that you can customize that algorithm so that instead of full protein design you can go for selective protein design, specifically useful for protein interactive design. That we will deal again in later detail later.

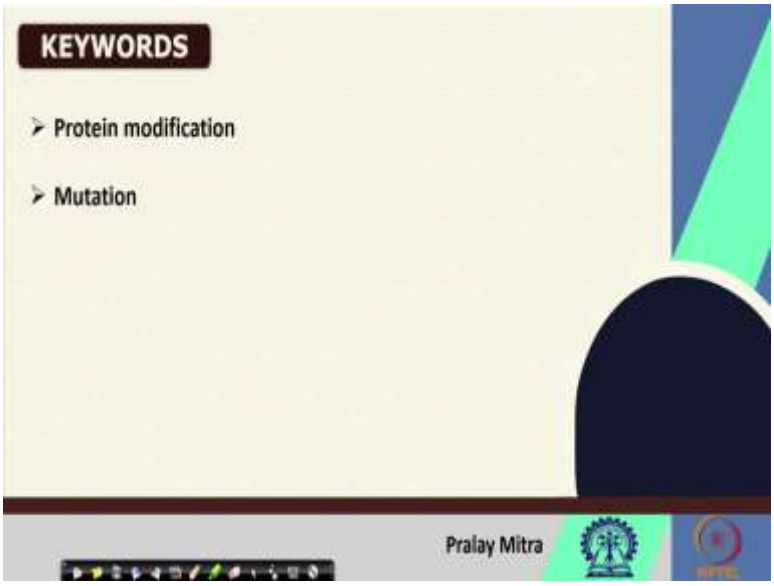
Next we discussed about the insertion and deletion operation. We mentioned that insertion and deletion is going to be same, because from where to where we are moving, based upon that whether it is insertion or deletion, that will be. So that is why instead of insertion we are taking deletion. Single point insertion and multi point deletion we discussed.

Now here in this specific protein modification or protein engineering we are going to discuss the mutation. So for the mutation I know that protein design we discussed but that protein design will design a number of amino acids from the protein. But if we are interested for single point mutation, then is it required; that to go for that protein design algorithm which will take several hours to finish? No, probably.

Similar to the way we have designed that protein insertion and deletion method, so we will also, can have another machine learning technique. That machine learning technique will take care about this protein mutation, specifically single point mutation.

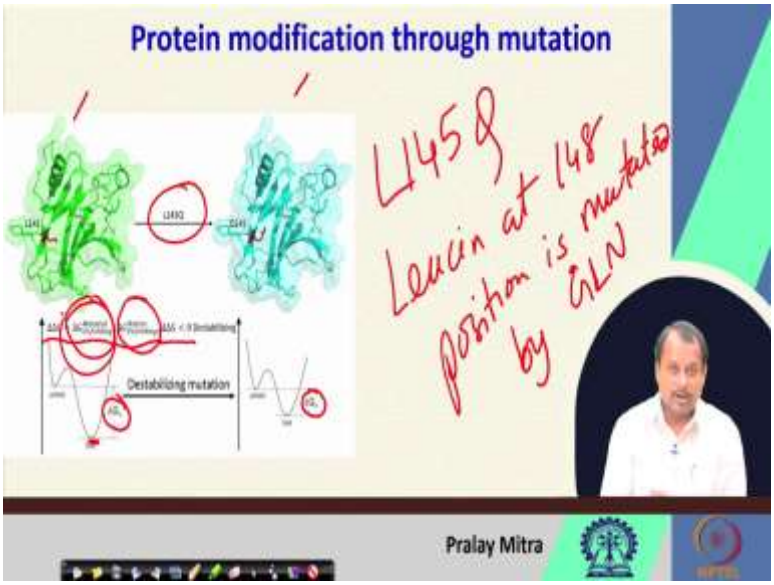
(Refer Slide Time: 02:25)




So the topic we are going to cover is protein modification and the mutation. So protein modification and the mutation, that is the keyword also I have picked up.

(Refer Slide Time: 02:33)




Now, in the protein modification through mutation it will be something like this pictorially if I wish to demonstrate. So here on the left hand side there is one protein given to you. On the right hand side another protein is given to you. From this left hand side protein I got this right hand side protein by performing one mutation operation L145Q which means leucine at 145 position is mutated by glutamine.

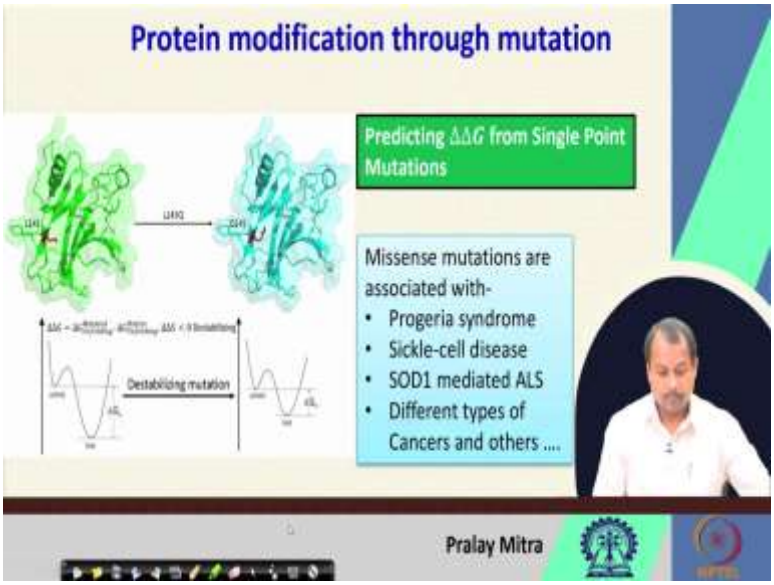
And when it is mutated then the question we are going to ask, that after this mutation whether this cyan color or mutated protein will be stable or not. That is also kind of similar to kind of it will fold or not. So that is what we are going to test. And this testing will be done again using some machine learning tool.

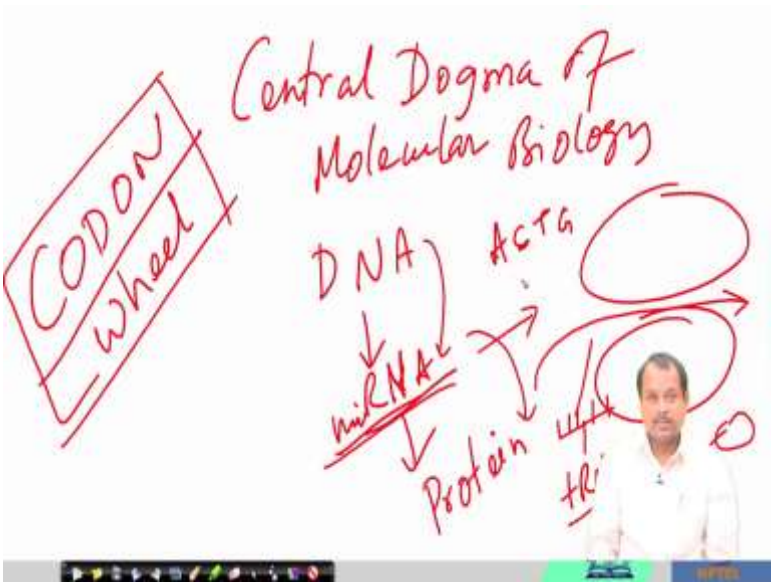
So, here delta delta G of mutated unfolding minus delta delta G of native unfolding we will compute. So delta delta G means change in Gibbs Free Energy, that we will compute for the mutated one and the native one. Then we will take the difference in order to check that whether it is going to be stable or not.

So, for that the equation we are considering is that, say this graph is basically unfolded state and this is my folded state. So this is my delta G. this is my delta G of mutated one. This is the delta G of the native one. Now the mutated minus, this is mutated minus native. If I take the difference, after taking the difference if I find that delta delta G is going to be negative, which means this

mutated unfolding delta G is less than native unfolding delta G, then it is going to be stable. So that is what we wish to do.

(Refer Slide Time: 04:53)




In this context we will address missense mutations. What is this missense mutation? So missense mutation means, if you let me go to the white board, if you look at the Central Dogma of Molecular Biology then DNA to RNA then protein, well this is miRNA, this is fine, to protein. Now when it is converting from miRNA to protein then what will happen that three consecutive miRNA, DNA and miRNA ACTG, and in this case there is uracil for RNA. Now for this mRNA, following the CODON wheel, three consecutive nucleotides will convert to some amino acid.

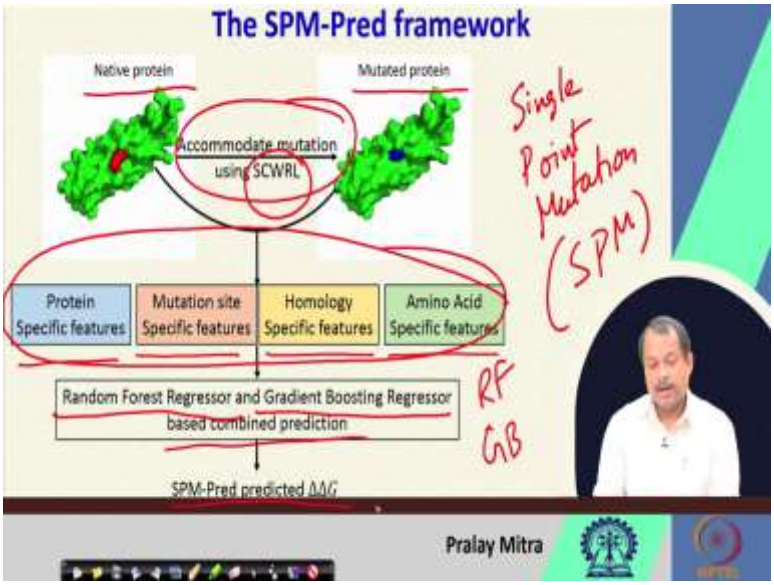
If you look at the simulation video of any protein folding method then you will see that after stripping out of DNA to miRNA then miRNA is going through the ribosome. There is a larger subunit and smaller subunit in the ribosomes. And this miRNA is passing through that one. While it is passing through that one, inside that one, what is happening?

So, at the environment there are some tRNA. So it goes and binds with this miRNA. So in one side of the tRNA there is a, there are that ACTU, and in this case ACGU, and in the another side there is amino acid. So while the complemented of tRNA is binding with miRNA then this amino acid is binding with another amino acid and that way protein synthesis will take place.

Now, this CODON wheel will say that if there is a small change in that nucleotide and because of that one if the amino acid will change then that is called as the missense mutation. So that is called as the missense mutation. So that deals with a number of diseases like Progeria syndrome, Sickle-cell disease, SOD1 mediated ALS, different types of cancers and others. So it deals with that one.

So, specifically in this specific protein engineering topic we will be interested to know that whether any sort of such missense mutation or single nucleotide polymorphism will lead to some modification at the protein level which can be devastating, in the sense that because of that small mutation or changes it will unfold the protein or it will misfold the protein so that it will not able to perform its own job. So that is our intention and we will see that one. So finally our aim is predicting delta delta G from single point mutations. Again this is going to be our computational prediction system.

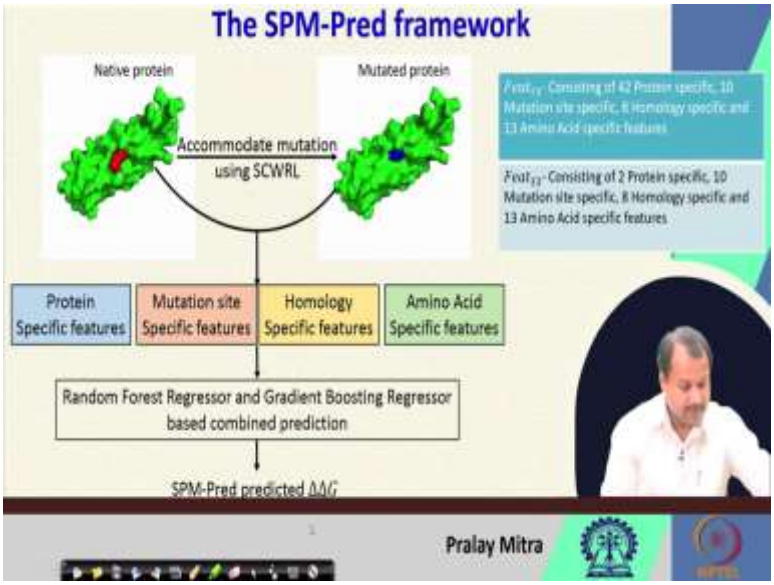
(Refer Slide Time: 08:47)




The single point mutation, in short SPM, the framework is something like this. Native protein is given to you. This is your mutated protein. Whether it will accommodate the mutation or not, that we need to check.

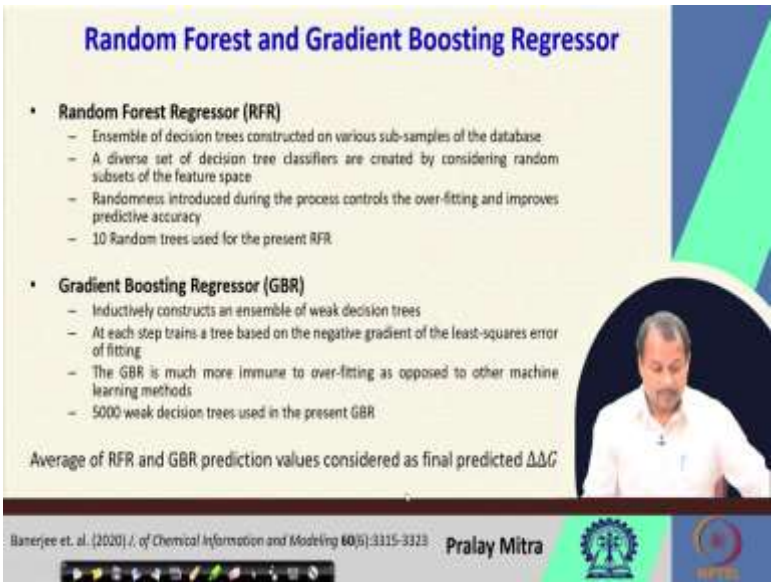
Here, SCWRL, s c w r l, is a software which performs side chain fitting. Now the features we are going to consider; protein specific features, mutation site specific features, homology specific features, amino acid specific features. Then when we will combine all those features using Random Forest Regressor and Gradient Boosting Regressor, in short, RF Regressor and GB Regressor, based combined prediction, then they will have SPM-Pred predicted delta delta G. And that is our goal.

(Refer Slide Time: 10:01)




So, let us go to that step by step. So Feat 73 consisting of 42 protein specific, 10 mutation site specific, 8 homology specific and 13 amino acid specific features. So Feat is the short form of feature feet. Feat 33 consisting of 2 protein specific, 10 mutation site specific, 8 homology specific and 13 amino acid specific features.

(Refer Slide Time: 10:37)



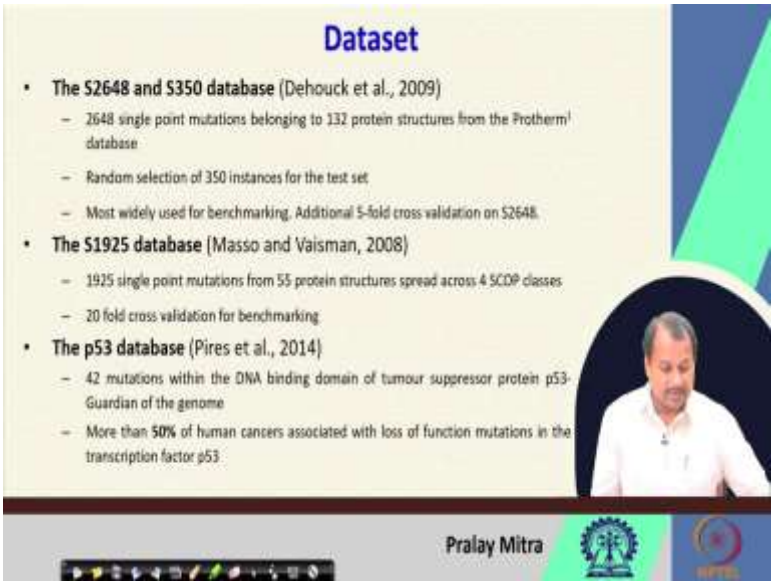

Now, these features are fed to your Random Forest and Gradient Boosting Regressor. So, in short, Random Forest Regressor or RFR is ensemble of decision trees constructed on various sub-samples of the database. A diverse set of decision tree classifiers are created by considering

random subsets of the feature space. Randomness introduced during the process controls the over-fitting and improves predictive accuracy. 10 random trees used for the present Random Forest Regressor.

Now, in case of Gradient Boosting Regressor or GBR inductively constructs an ensemble of weak decision trees. At each step trains a tree based on the negative gradient of the least square error of fitting. The GBR is much more immune to over-fitting as opposed to other machine learning methods. That is a good thing. And 5000 weak decision trees are used in the present GBR. So average of RFR and GBR prediction values considered as final predicted delta delta G. So two different regressors are used.

So, please note it down. Our aim is not to classify whether after the mutation it is going to be stable or not. When we are interested to predict that what will be the delta delta G, based upon that one some decision we will take. But I need delta delta G, change of Gibbs Free Energy after the mutation then I have to go for Regressor. That is why the Regressor Model. And instead on relying only one Regressor, which every Regressor model has some pros and cons, so it might be good idea, and here also we are demonstrating one algorithm where GBR and RBR are combined and their average predictions are taken for the purpose. So those features are fed into this GBR and RFR and the average prediction is considered.

(Refer Slide Time: 12:46)



Incidentally lot of work has done for this missense mutation or single point mutation or say, single nucleotide polymorphism. There exist several datasets. So there is one dataset. Here S followed by the number indicates how many number of single point mutation datasets are there. So that is why S2648 will indicate 2648 single point mutations belonging to 132 protein structures from the Protherm database.

So Protherm is a very widely used mutation database. So you can look at that one. And another S350 database is the random selection of 350 instances for the test set from that S2648. It is done by Dehouch and published in 2009, most widely used for benchmarking and additional five-fold cross validation on S2648.

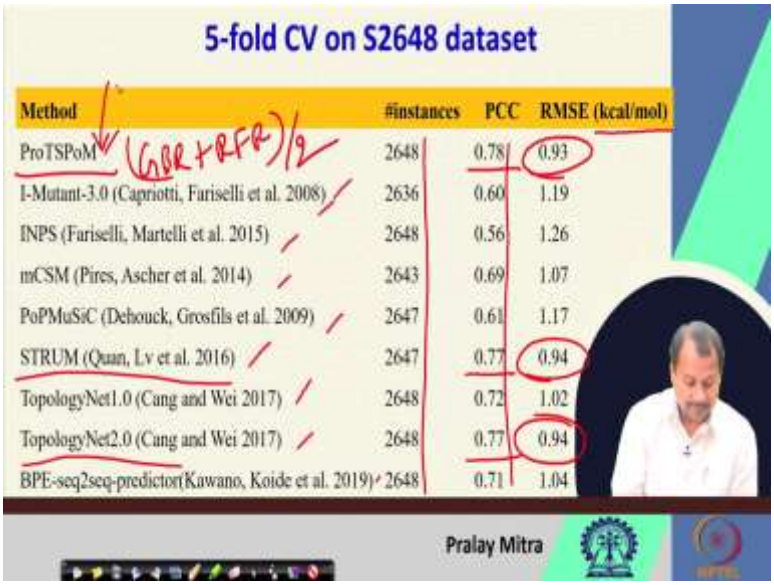
Another dataset S1925 is given by Masso and Vaisman which contains 1925 single point mutations from 55 protein structures spread across four SCOP classes. Now you know what are the SCOP classes. And also you know that what will be the effect if the SCOP classes are different. SCOP classes are different means at the fold level, at the family, Superfamily and the domain level they are different.

Specifically at the fold level when they will be different they are belonging to different classes, then you may expect that some mutation will take place in helix, some in sheets, some in coil, some in position where after helix there is a sheet. So possibilities are there. So that way it may give you a varied or diverse nature of the dataset. 20-fold cross validation for benchmarking was used for the dataset S1925.

For p53 database, so p53 is a widely, is a very well-known protein in the research of the cancer. So this p53 has one dataset. It is published by Pires. So p53 database contains 42 mutations within the DNA binding domain of tumor suppressor protein p53 Guardian of the genome. So p53 is called as the Guardian of the genome.

And it is also established that in case of cancer, so most of the cases it is the p53 which goes wrong. Or in p53 there is a mutation occurred in most of the cancer cases. So more than 50 percent of human cancers are associated with loss of function mutations in the transcription factor p 53. And that way p53 is a very wide studied protein.

(Refer Slide Time: 16:01)




Now, if the testing is done on the five-fold cross validation for S2648 dataset several algorithms exist. We are looking at their individual power, predictive power. Along this the ProTSPoM actually is the one for which GBR and RFR, average of this, on those 73, and 33 and 72 feature set has been considered.

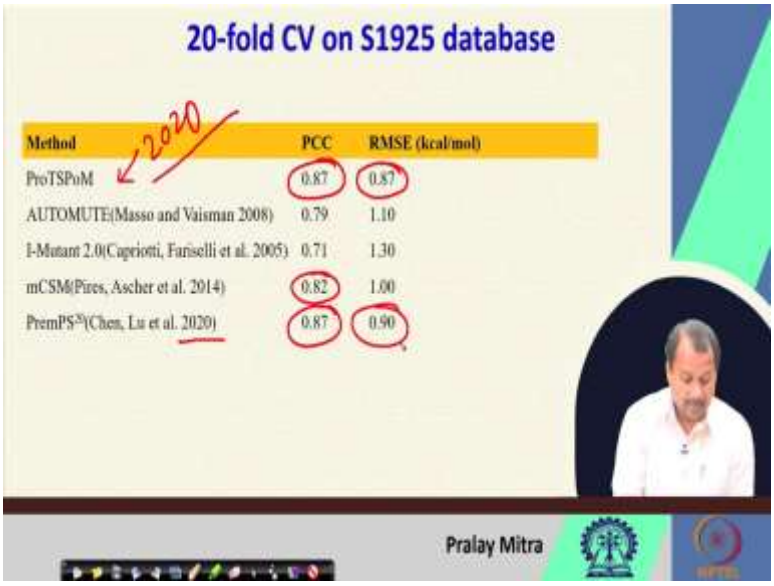
So the number of instances which are considered is mentioned here. Almost everybody has used all the instances except few. That is negligible. You can ignore that one. For Pearson correlation coefficient you can see that, so high is actually 0.78 here. 0.77, 0.77 is also for this STRUM. And this TopologyNet2, so they are also using different machine learning tool. So they got that.

Now, RMSE, so root mean square error, is calculated here. That is kcal per mol. So how is it calculated? It is a regressor as I mentioned. So what it will predict? It will predict the change in Gibbs Free Energy. Now for this dataset, so experimental data for the change in Gibbs Free Energy is also there. So compare these two. And what is the root mean square error? So between these two, I mean the experimental data and the computationally predicted data, by each of the methods. So several methods are there that you can note.

So people are working in this area, say since 2008, earliest paper I-Mutant-3.0. Then that is the RMSE. So lower the RMSE better is the method, and higher the PCC the better is the method. For this ProTSPoM it is 0.93. For say, TopologyNet2.0 it is 0.94. For STRUM it is 0.94. So

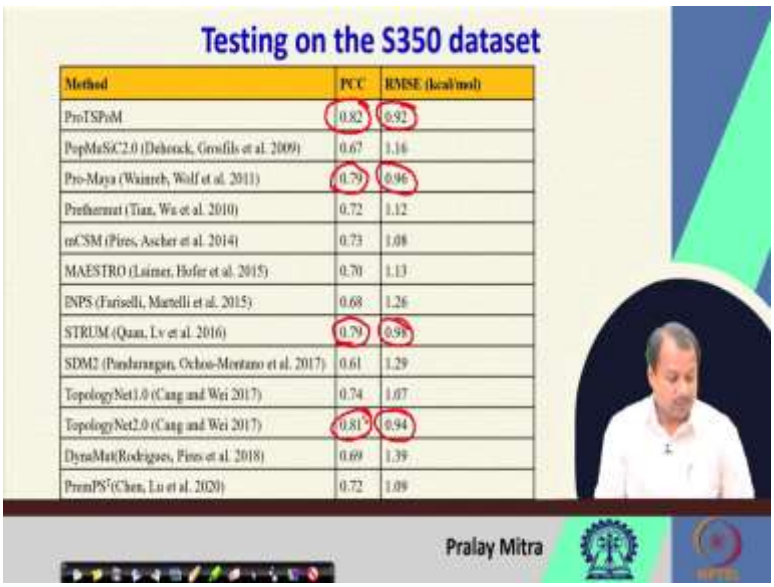
these are the values. And rest are greater than 1. This GBR, this ProTSPoM is actually published recently.

(Refer Slide Time: 18:30)



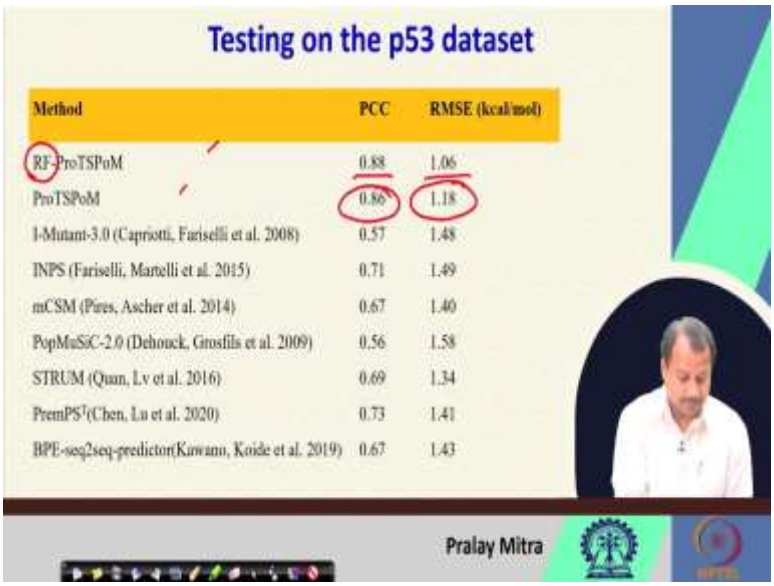

Now 20-fold cross validation on second dataset is 1925. So the ProTSPoM again, the PCC value is 0.87, then RMSE is 0.87, then PremPS 0.87, then mCSM 0.82. So they are with the high PCC values. And here the PremPS published in 2020. This is published I guess in 2021 or in 2020 itself, yes 2020, it is also in 2020 this ProTSPoM is published. So this is also 0.09.

(Refer Slide Time: 19:29)



Now, for another dataset, S350 dataset. So again if I look at the data then ProTSPoM is 0.82. Then RMSE is, PCC is 0.82, RMSE is 0.92. Then above 8, who is above 8? Here there is above 8, one guy. So TopologyNet2 then nobody else is above 8. It is close to 8, 0.79. This is also closed to 8, 0.79. So these methods are notable. And less than 1 RMSE; 0.92, 0.96 and then 0.98, then 0.94, so these methods are really good on this prediction power.

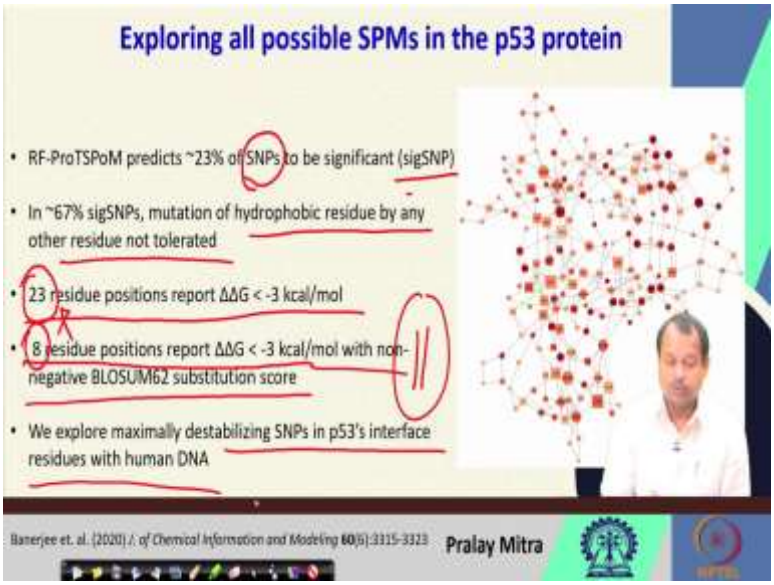
(Refer Slide Time: 20:18)




Then testing on the p53 dataset. Again, so if I look at, then ProTSPoM is 0.86, RMSE is high 1.18.In case of other methods actually PCC is not that much good, and RMSE is also not that much good. Here one interesting fact is that, only RFR if you use instead of combining with GBR then the PCC will increase little bit and RMSE will decrease a little bit. But what is a reason behind that?

Why RFR will give better result, and if you combined with the GBR then it will not get that much good result, it is not apparently clear that why it is so. So two results are given here. So in this case you can see that most of the work is published recently 2019, 20, and this is also as I mentioned published in 2020.

(Refer Slide Time: 21:27)




Next in order to apply this method, so application. So we design some technique. Very good. So when you design some technique then where we can apply that one? It is not only predicting that, predicting on the dataset and checking that what is the PCC value or what is the RMSE value, but if you go for applying that method, here is one instance, exploring all possible SPMs in p53 protein.

So what you can do? RF-ProTSPoM predicts about 23 percent of SNPs to be significant. So this is SNP, single nucleotide polymorphism to be significant or SigSNP. In 67 percent SigSNPs mutation of hydrophobic residues by any other residue not tolerated which means if you go for all possible mutations of p53 protein and look at the capability; so last time we did protein design and their application domain was different. In this case we are going for missense mutation or single point mutation. And for the single point mutation we should go for that protein design algorithm that we discussed on the last week.

So here we are designing one machine learning based technique, basically Regressor model we design. So we are discussing one Regressor model. So Random Forest Regressor and Gradient Boosting Regressor, specifically for this p53 protein, it is noted that Random Forest Regressor is doing good. So that is why, if you go for all possible mutations then the number of such mutations will be huge. Now 20 different amino acids, leaving the original one apart, then 19 possibilities are there for the mutation at each position.

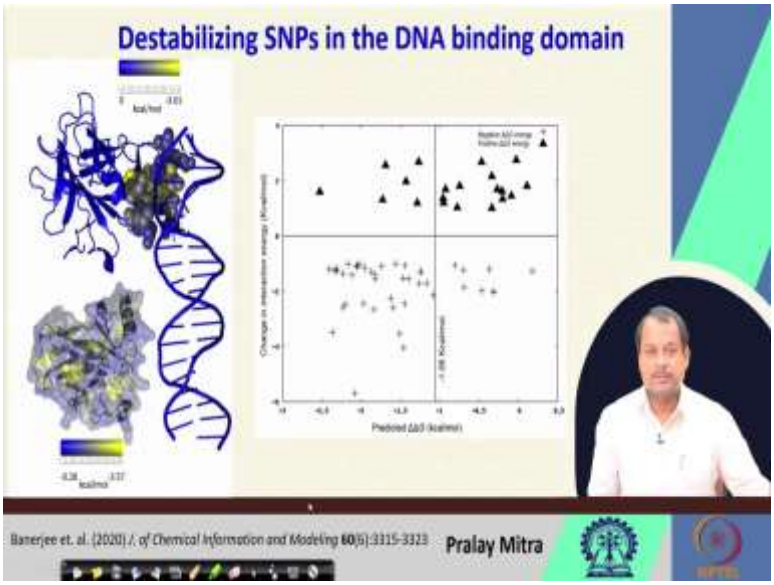
Even if you consider that 30 positions are there then accordingly 19 divided by 30 will be the total number of mutations which are possible. If you run the your protein design algorithm which takes several hours then it is not feasible, whereas if you run this Regressor model which takes fraction of second or about a second then easily you can calculate that what is the predicted value?

Once you will have that one then you can be ambitious to check whether those predictions are going to be the stable one or not. That way it is predicted that 23 residue positions report delta delta G greater than, less than minus 3 kcal per mole. 8 residue positions report delta delta G minus 3 less than kcal per mole with non-negative Blossom62 substitution score, which means evolutionary also it is very encouraging.

And here this 23 and this 8, so this 31 cases where delta delta G value is changing a lot because of the mutation is actually a very good candidate for the experiment. And to check that what is the reason behind this, whether this can give us any new insight or not, so that way, we can model and we can engineer the protein using our computational tool and technique using our algorithm or machine learning or deep learning technique. That technique can be utilized to filter out and output something like 23 plus 8 on which the experimentalists can do the experiment and mention that whether it is going to be a very good prediction, or I mean that, it has some biological significance or not.

And for that you need not have to go for 19 to the power the length of the positions for which you need the mutation. I am giving you, say 31, 23 plus 8. We explore maximally destabilizing SNPs in 53 interface residues with human DNA. So that is the area where we can really contribute by doing computational technique.

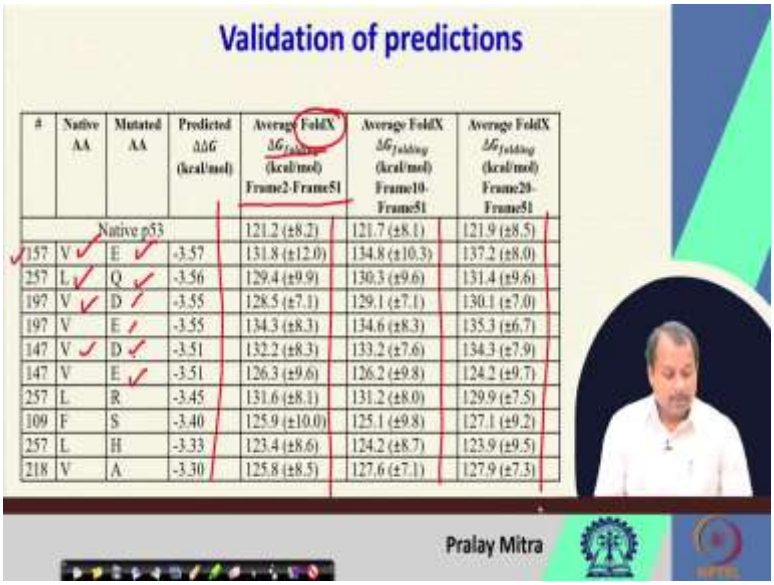
(Refer Slide Time: 25:45)




So here destabilizing SNP in the DNA binding domain. So basically it is the pictorial view. So only the DNA binding domain has been considered. And when you consider that one, in this graph also we can see that few cases are there which are with very less delta delta G kcal per mol. That means they need special attention by the biologists. So that is what we can do for the biologists.

So we can design the algorithm. We can run that algorithm. We can design that algorithm. We can implement that algorithm. We can run the algorithm and get the result for the biologists who can perform the experiment on that in order to get the final output. Of course we can also do our analysis by ourselves or collaborate in order to get the result.

(Refer Slide Time: 26:35)



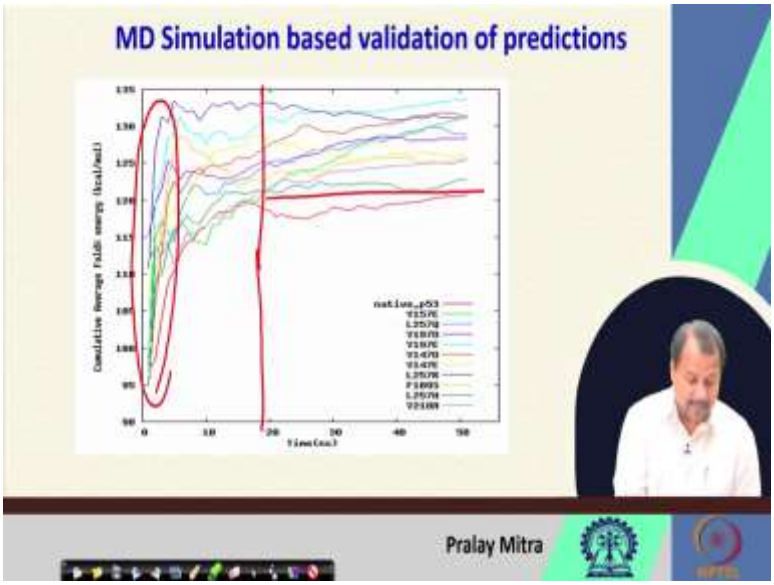

| # | Native AA | Mutated AA | Predicted ΔΔG (kcal/mol) | Average FoldX $\Delta G_{folding}$ (kcal/mol) Frame2-Frame51 | Average FoldX $\Delta G_{folding}$ (kcal/mol) Frame10-Frame51 | Average FoldX $\Delta G_{folding}$ (kcal/mol) Frame20-Frame51 |
|---|---|---|---|---|---|---|
| | | Native p53 | | 121.2 (±8.2) | 121.7 (±8.1) | 121.9 (±8.5) |
| 157 | V | E | -3.57 | 131.8 (±12.0) | 134.8 (±10.3) | 137.2 (±8.0) |
| 257 | L | Q | -3.56 | 129.4 (±9.9) | 130.3 (±9.6) | 131.4 (±9.6) |
| 197 | V | D | -3.55 | 128.5 (±7.1) | 129.1 (±7.1) | 130.1 (±7.0) |
| 197 | V | E | -3.55 | 134.3 (±8.3) | 134.6 (±8.3) | 135.3 (±6.7) |
| 147 | V | D | -3.51 | 132.2 (±8.3) | 133.2 (±7.6) | 134.3 (±7.9) |
| 147 | V | E | -3.51 | 126.3 (±9.6) | 126.2 (±9.8) | 124.2 (±9.7) |
| 257 | L | R | -3.45 | 131.6 (±8.1) | 131.2 (±8.0) | 129.9 (±7.5) |
| 109 | F | S | -3.40 | 125.9 (±10.0) | 125.1 (±9.8) | 127.1 (±9.2) |
| 257 | L | H | -3.33 | 123.4 (±8.6) | 124.2 (±8.7) | 123.9 (±9.5) |
| 218 | V | A | -3.30 | 125.8 (±8.5) | 127.6 (±7.1) | 127.9 (±7.3) |

So, validation of the predictions. So some of the predictions you can also validate by yourself. So separately when you are predicting then you are not consulting anything. You designed one Regressor model. You designed one prediction system. And you identified the number of possibilities, then you predicting one. Then you analyzed and identified, say 23 plus 8 which means 31 that identified, and this can be a very good candidate to be explored further.

Then either you can go for the experiment or you may find that some literature has already mentioned that, these I am getting with this negative, say delta delta G is very less but I do not know apparently what is the reason behind this. So computational analysis you have. You can go, and for these cases, at position 157 when valine is replaced by E, at 257 leucine is replaced by Q. At 197 valine is replaced by D E, D or E. At 147 valine is replaced by D or E. Then what are the predicted values or delta delta G, kcal per mol, that is listed here. And you see that in all cases it is very less, less than minus 3. That is something interesting.

Then average FoldX. So this FoldX actually composed the physics-based force field. So Average FoldX of delta G, it also predicts that one, from frame 2 to frame 51 which means that if you run, after that mutation if you run the molecular dynamics simulation for 50 nanoseconds then for frame 2 to 51, after the MD, so what is the variation in the energy? That you can compute here. That you can compute here. That you can compute here. And these are the results.

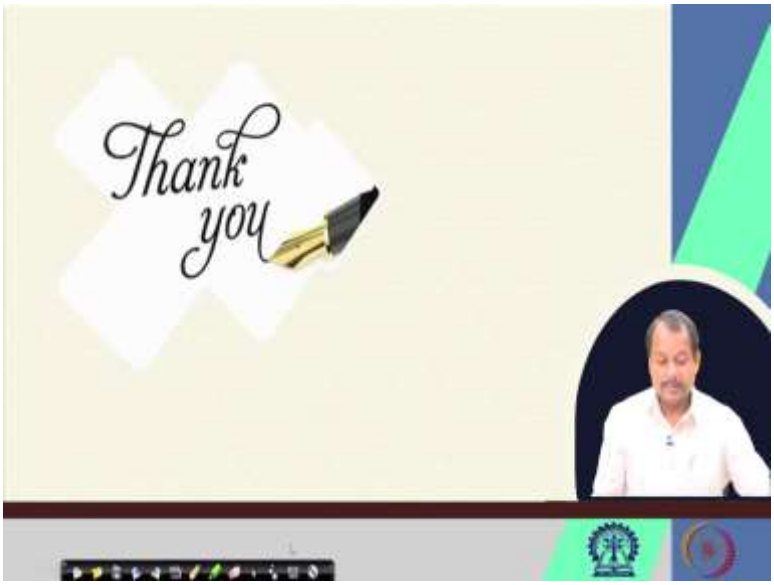
(Refer Slide Time: 28:39)




So, MD simulation based validation of prediction. So here you can see the time, so plotted along the x axis. And along the y axis cumulative average FoldX energy is plotted here. So the energy is increasing and finally it is getting stabilized or saturated basically, and that when you compute the average, here it is, the average given on, say column number 1, 2, 3, 4, 5, 6, 7; so 5, 6, 7; these three columns indicates the average of the FoldX energy.

Now if you take the average from 2 to 51, since there is a sharp increase, so since it will be also be included, so the average may not be good. But you see that after 10 or preferably after 20 it is saturating. So it is better that from 20 to 50 you take that calculation. And if you take that one, so here, so from 10 to 51, sorry, from 10 to 51, from 20 to 51 you see that the average has been computed. And that you can also validate computationally specifically when you do not have any support for experiment etc. you can also do it by yourself.

(Refer Slide Time: 30:01)




So, that is it for today's week, today's lecture and in this week. So in this week, so we started with the protein design, then protein engineering. In the protein engineering or protein modification we covered single point relation, multi point relation and single point mutation. For all the three cases we come up with some machine learning based solutions which are fast enough compared to the protein design, and based upon our requirement we find that it is perfectly suitable for our purpose. We need not have to go for computing-intensive protein design problem. But computing-intensive protein design problem has a separate scope, separate application that we demonstrated on the last week. Thank you very much.