4. Lecture Molecular Scoring

0:01
All right, we're now starting to think about molecular scoring and how does it tie in with generative design.

0:09
So molecular scoring for me is one of the hardest, the most critical parts of generative design, as this is how we transform a chemical structure into a numerical value which will go ahead and optimise.

0:22
And the key thing here is that numerical value has to represent the chemical and biological problem at hand.

0:30
Without proper scoring, we certainly won't be able to find molecules that we will want to make and get closer to being able to deliver a drug.

0:38
So in terms of scoring, again, it very much depends on the problem at hand.

0:43
So I've taken here, for example, three different drugs and depending upon the chemical biology problem, they may score very differently.

0:51
So these are all approved medications.

0:53
However, depending a problem upon the problem that we're trying to solve, they may score differently.

0:58
So for example, if we want to try to find an anti malarial, that scoring function, that scoring strategy that we will have to use will be very different to the scoring strategy that we'll be using if we want to try to find molecules that will cure a particular type of brain cancer.

1:19
Broadly, however, scoring falls into two areas.

1:25
The first area is explicit scoring, and that explicit sort of scoring is what we will define.

1:31
So these are the computational sort of tools, the scores that we would generate through, for example, molecular docking or molecular dynamics or free energy perturbation.

1:44
On the other hand, we also have this idea of implicit scores, and implicit scores is very much this idea.

1:50
Then of we have a gut feeling, we have a idea of what is tractable and what is not.

1:57
What does the project want to make?

2:00
Implicit scoring is very much a subjective score.

2:04
Sometimes it is hard to be able to encode what implicitly we like about a molecule.

2:11
However, what we would say is that certainly implicit scores are what we expect our molecules to have.

2:19
So we would expect certain types of properties and that is implicit scoring.

2:25
Now, on the slide here, I've shown for example, a simple function explicit score plus implicit score.

2:33
However, that need not necessarily be the case.

2:36
There are multiple different ways to combine explicit and implicit scoring, and that is very much a wide variety of tools which we're not going to talk about too much today.

2:47
Instead, the real focus is about understanding the difference between explicit and implicit scoring.

2:54
So just to give you a little bit of a flavour of the differences, then explicit scoring.

2:58
Really thinking then about calculated properties.

3:01
So here I have two molecules with the same explicit score and a log P of 2.64.

3:08

So the molecule on the left hand side is very very small and the molecule on the right hand side is a prodrug and they both have a similar a log P.

3:18
So if we went ahead and said I want a molecule and I will say a molecule is good if it has the value of a log P of 2.64 explicit score, both of these will be generated and they will both score very well.

3:35
However, if we start to think then about the implicit score, which one would you prefer to make?

3:40
Would you prefer to make the left hand side in terms of synthetic tractability, in terms of ligand deficiency, in terms of what chemistry you may know?

3:50
Or would you prefer the right hand side?

3:52
Would you prefer the pro drug?

3:53
Would you prefer something that might be a little bit harder to make, but may satisfy more of your solubility issues?

4:00
Maybe it satisfies more of the DMPK properties that you're wanting from your molecule.

4:07
So again, this is where we start to think about explicit versus implicitly defined scoring objectives.

4:15
Moving on then we then can then sort of look at a more a more complete example.

4:22
On the left hand side we have a small molecule and that small molecule is a one step drug.

4:31
So this has I believe a market value of $330 million.

4:35
And on the right hand side, we have a molecule that is made after 107 steps, and this molecule is worth $1.2 billion.

4:45
So with that in mind, implicitly, which one would we prefer?

4:50
Now of course, it will very much depend upon the individual.

4:54
You might prefer the left hand side because it has an easier synthetic scheme.

5:02
However, maybe you go to speak to a natural product chemist and they will say, Oh no, I much prefer the right hand side.

5:09
And again, implicit scoring will very much depend upon the time of day that we ask this question.

5:15
So maybe if we've had a long day, it might be, yes, I prefer the left hand side compared to the right hand side.

5:22
But implicit scoring is incredibly important because it captures that sort of experience, that gut feeling, that assessment that chemist looking at a chemical structure might have.

5:36
Briefly then talking about explicit scoring.

5:39
Explicit scoring separates itself into sort of two areas.

5:44
The 1st is receptor based scoring.

5:47
So in receptor based scoring, this is where we have a protein structure, we have a ligand.

5:51
We would do, for example docking, we would form a complex, we may calculate a docking score.

5:56
This is explicit scoring using receptor based information.

6:02
On the other hand, we also have ligand based explicit scoring methods, so this can be calculating descriptors, pharmacophores.

6:10
This is very much thinking about the molecule incompleteness, not worrying about additional target information.

6:18
Certainly when we would be thinking about a larger scoring function, we may wish to mix and match multiple parts.

6:24
So we may wish to have objectives that are related to, for example, calculated physical chemical descriptors.

6:33
On the other hand, we may think about combining docking or molecular dynamics.

6:38
So very much it depends upon the project and the project needs and the information that we have.

6:46
We would also at this stage be thinking about using, for example, machine learning approaches.

6:51
So machine learning approaches where we would provide a small molecule, we would featurize it and then we would try to predict a corresponding score for that molecule.

7:03
OK.

7:03
So now when we're starting to think about combining multiple scores together, the key thing here is that how do we balance those different objectives?

7:13
How do we find a molecule which balances satisfies all of our requirements?

7:19
How do we find a balance between synthesizability and predicted potency?

7:24
How do we find that?

7:25
And one of the things is that we will have to combine multiple objectives together and to do that in a relatively clever way.

7:33
So here the idea is that really when we're thinking about effective molecular design, it requires balancing multiple objectives simultaneously.

7:42
And what I show here is a plot.

7:45
And this plot captures objective 1 and objective 2.

7:48
We could have whatever objectives we like here, molecular weight, potency, synthesizability, solubility, whatever.

7:56
But the key take away here is that when we are looking for the best solutions or the sweet spot, the place where we can't improve one objective without sacrificing another, or where is it optimal trade off between the two objectives, these lie upon what we call the Pareto frontier.

8:17
And the Pareto frontier is interesting because it contains potentially multiple good choices.

8:24
That is, it depends upon every objective and every sort of trade off.

8:32
We may wish to have a trade off between objective 1 and objective 2.

8:35
So we set the value of objective one and then we get a particular value of objective 2.

8:43
And what that means is that we get multiple solutions that sit on the Pareto frontier depending on how much we're willing to have that trade off.

8:51
These are known as non dominated solutions.

8:55
On the other hand, we also have dominated solutions and dominated solutions are really the worst choice.

9:02
So this is where we can certainly find a better solution that has a trade off between our objectives and therefore we kind of want to avoid these particular solutions.

9:14
The final sort of area which we have to worry about is this area known as infeasible solutions.

9:21
And infeasible solutions is where we cannot simply design A molecule with those properties.

9:27
It is impossible to do.

9:28
That is, we cannot find a solution in the infeasible zone any further forward from the Pareto frontier.

9:37
In essence, this is the perfect solution and it wouldn't be possible to be able to create a perfect solution.

9:45
So often we have to rely upon instead the Pareto frontier and the non dominated solutions.

9:52
Lastly, I'm going to now talk about filtering.

9:55
And filtering is another form of scoring and I think it's a little bit overlooked, but filtering is really where we try to start to combine multiple binary scoring approaches to be able to remove molecules that we don't like.

10:12
And this often occurs after generation.

10:15
So after we have generated our chemical structures and what we do is that we take our molecules and we say one, we like that molecule or zero, we do not like that molecule.

10:27
And an example of a philtre will be the rule of five.

10:30
So if we satisfy the rule of five, so we have the right molecular weight, the right number of hydrogen bond donors or acceptors, what we can have is we can then say one, it satisfies the rule of five or zero.

10:42
It does not satisfy the rule of five and that trade off can lead us to then better compounds overall.

10:51
This is because we are again capturing more information that we actually want our final solutions to have.

10:58

Filtering often happens because the sheer number of chemical structures that a generative design method would produce is far in excess of what can ever be made.

11:08
Therefore, we have to philtre down various different sets.

11:13
Finally, I'm going to point out one additional area in this idea of post generation scoring.

11:19
And post generation scoring happens after the generative design method has been completed.

11:25
On the other hand, you can also have pre generation scoring.

11:28
This is where we may score seed compounds or during search scoring.

11:33
This is where we actually score as the search is progressing and depending upon which area of the search and which area of the generative design software is executing will then depend strongly which approach we take and what our scoring function looks like.

11:51
But overall, the key take away is that molecular scoring is all about how do we transform that chemical structure into a number and then use that number to design new compounds.

12:03
But again, a score is only as good if we can have them made and tested.