

AlphaFold

Structural Bioinformatics

Alberto Meseguer

Course 2022-2023

Meet my friend Altair



<http://gallegolab.org/>

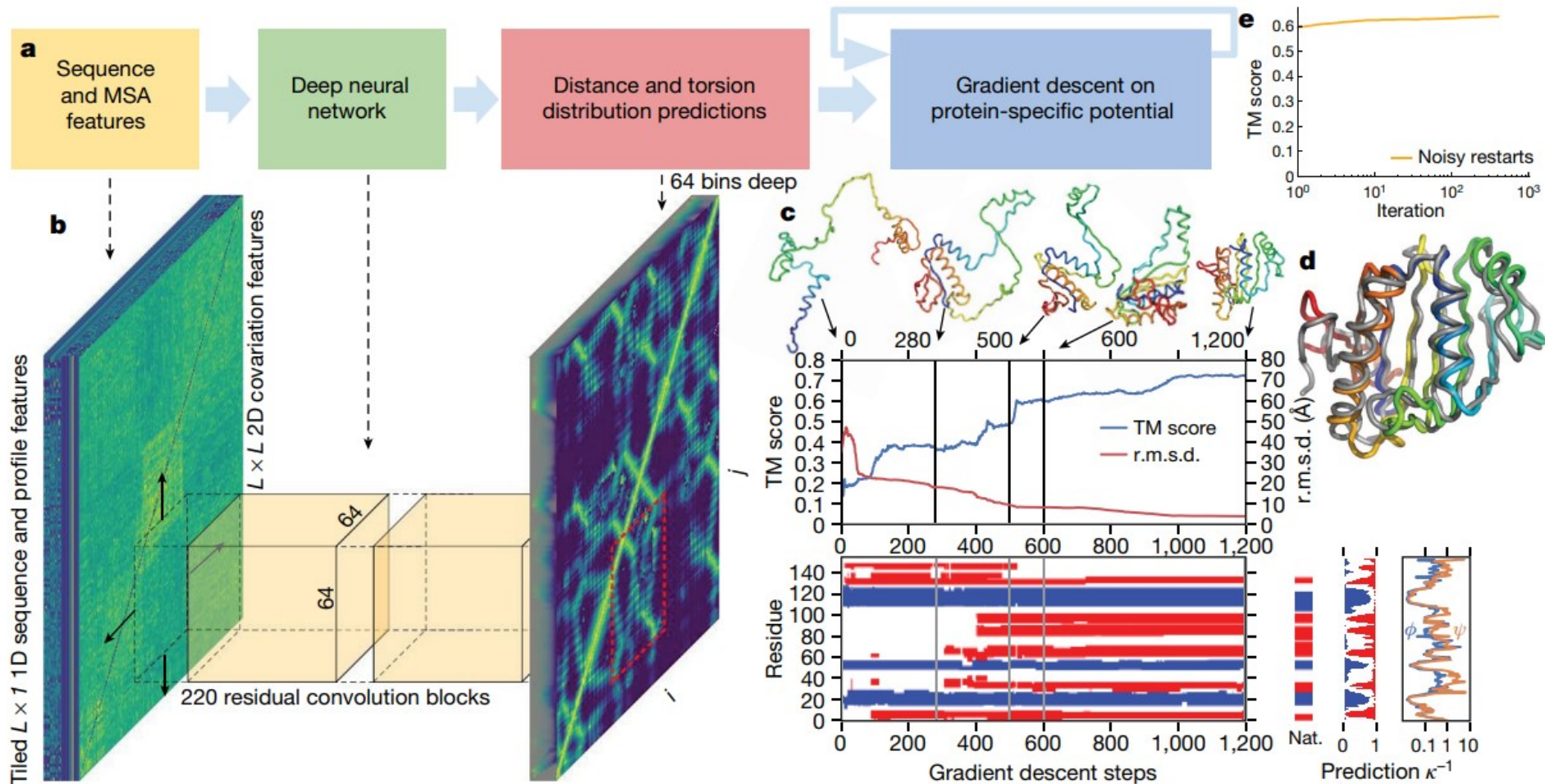
Why is alphafold important?

Alphafold2 has achieved the best performance ever at predicting protein structure *de novo*

Median Free-Modelling Accuracy

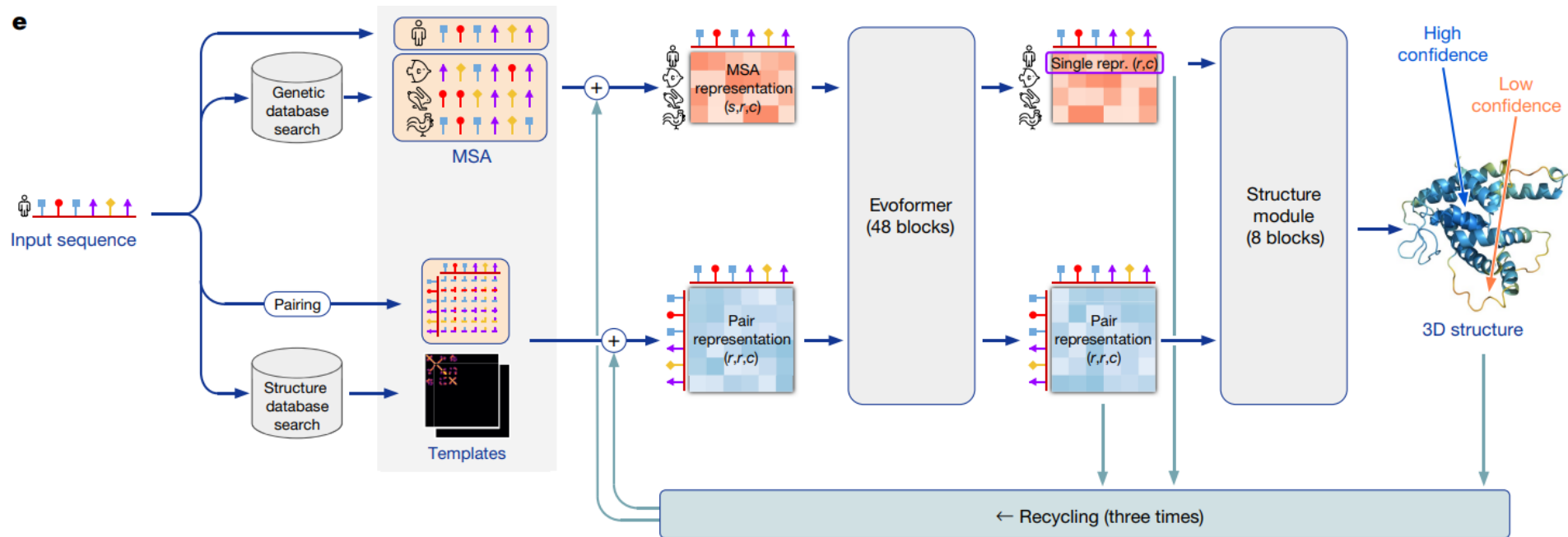


The alphafold1 workflow



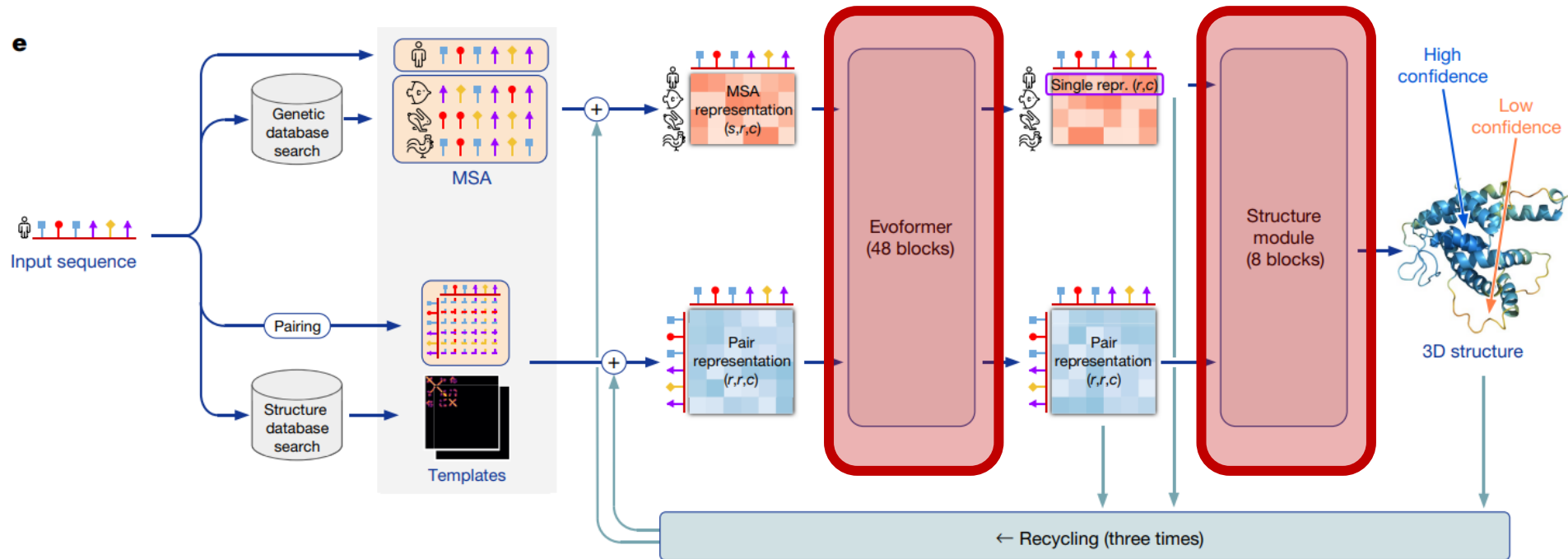
Senior AW, et al. 2020.

The alphafold2 workflow



Jumper J, et al. 2021.

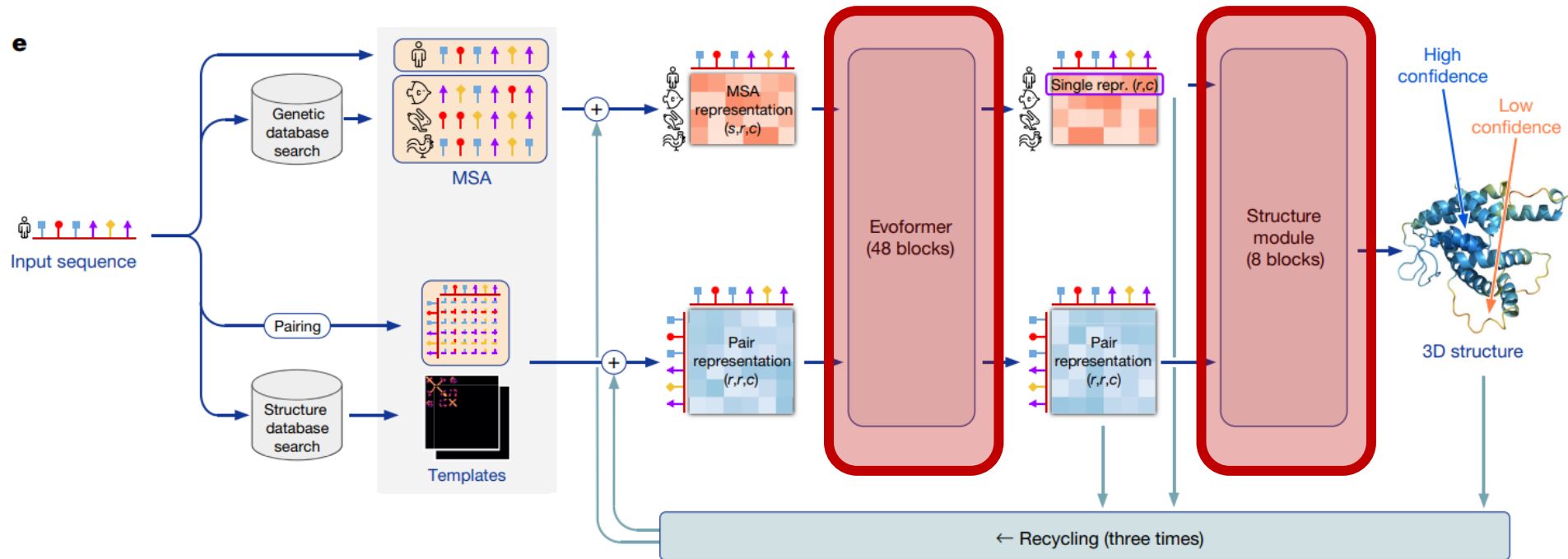
The alphafold2 workflow



Jumper J, et al. 2021.

What is happening here??

The alphafold2 workflow



What is happening here??
Deep neural networks!!

Jumper J, et al. 2021.

Understanding machine learning

Machine learning consists on developing algorithms that are able to learn by themselves in order to improve performance at some task

In the case of alphafold, the task is to predict protein structure from sequence

Deep neural networks are one of the more advanced algorithms for machine learning

Understanding machine learning

Machine learning consists on developing algorithms that are able to learn by themselves in order to improve performance at some task

In the case of alphafold, the task is to predict protein structure from sequence

Deep neural networks are one of the more advanced algorithms for machine learning

To better understand machine learning, we are going to understand one of the first machine learning algorithms ever:

The perceptron



Understanding machine learning: The perceptron

The perceptron is an algorithm that takes an input and with that is able to make a binary prediction

For example, we could use a perceptron to predict if tomorrow will rain using as input the atmospheric pressure and the humidity percentage of the day before

INPUT

AP

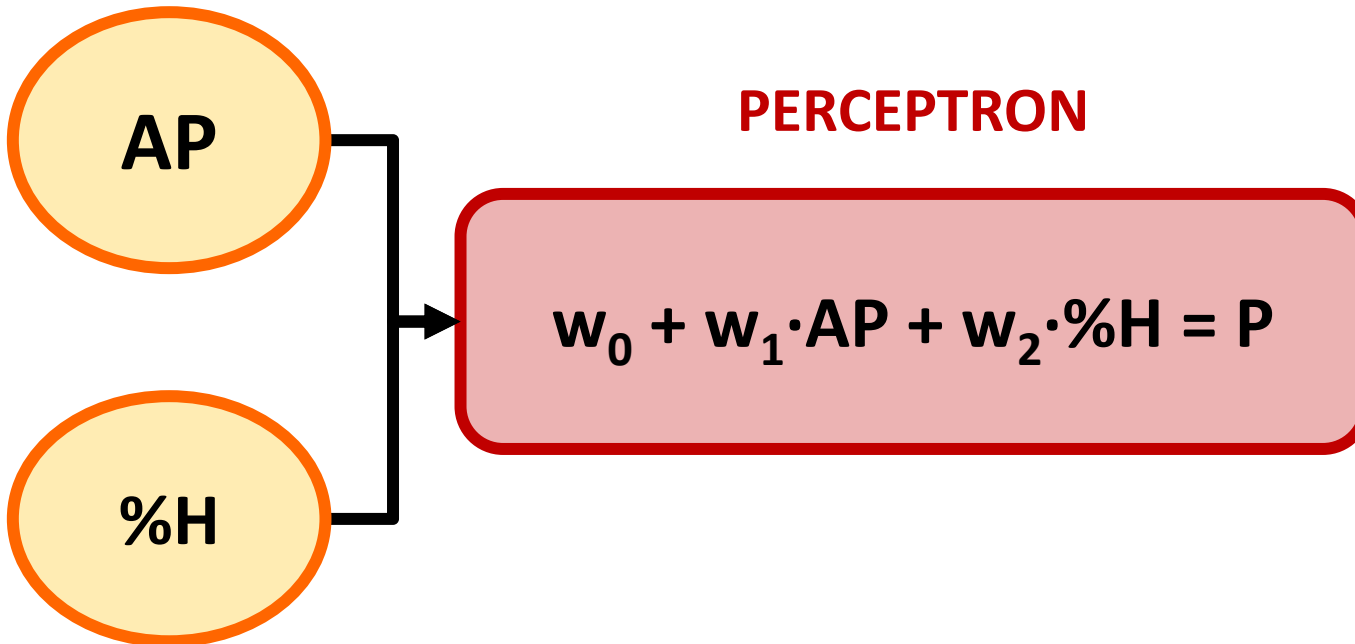
%H

Understanding machine learning: The perceptron

The perceptron is an algorithm that takes an input and with that is able to make a binary prediction

For example, we could use a perceptron to predict if tomorrow will rain using as input the atmospheric pressure and the humidity percentage of the day before

INPUT

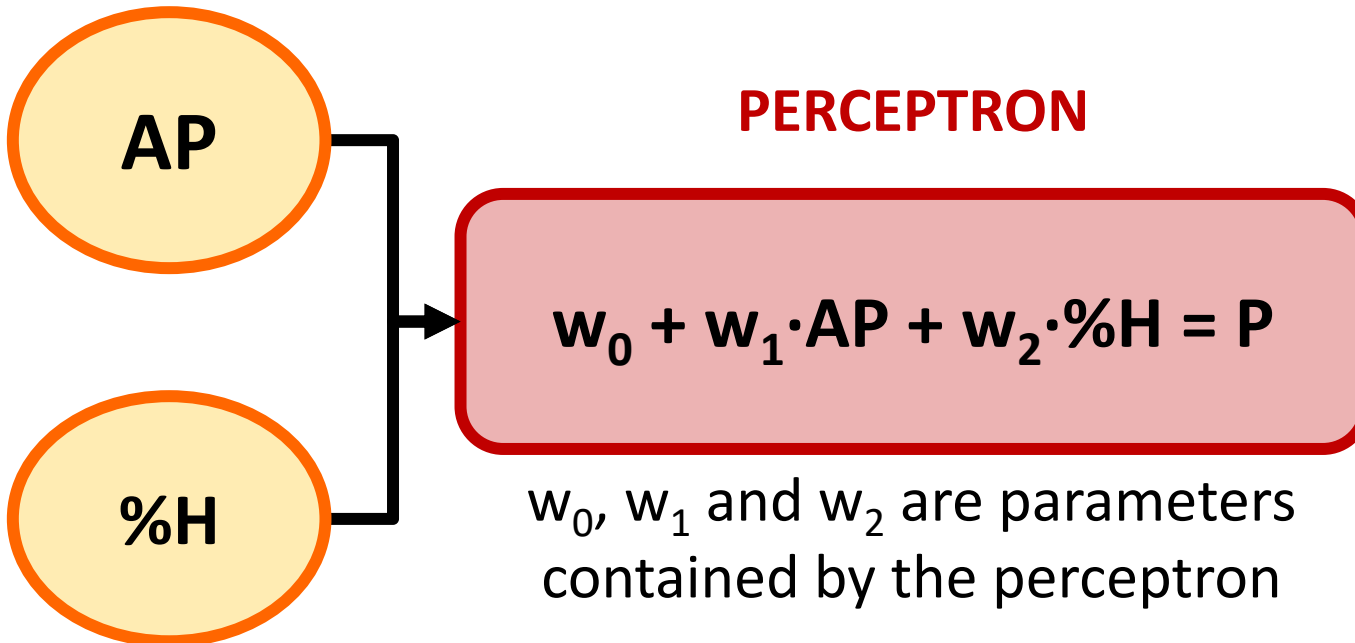


Understanding machine learning: The perceptron

The perceptron is an algorithm that takes an input and with that is able to make a binary prediction

For example, we could use a perceptron to predict if tomorrow will rain using as input the atmospheric pressure and the humidity percentage of the day before

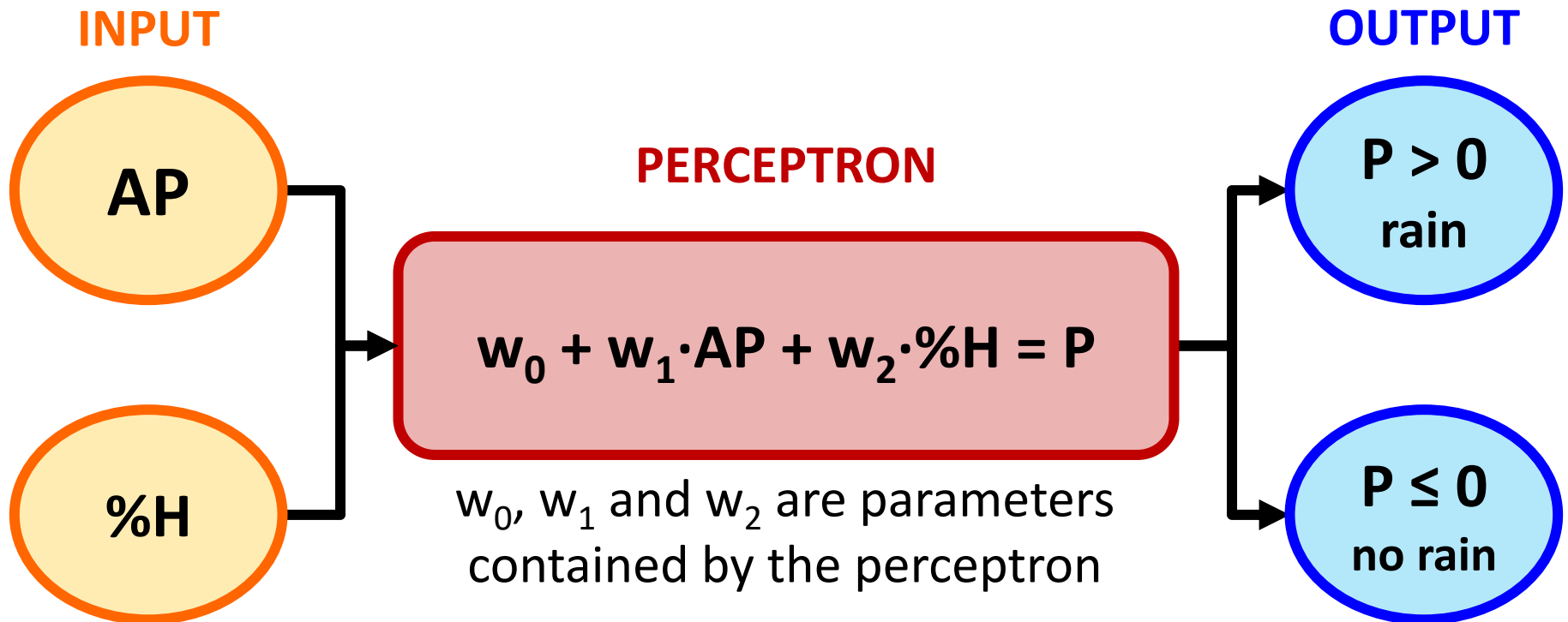
INPUT



Understanding machine learning: The perceptron

The perceptron is an algorithm that takes an input and with that is able to make a binary prediction

For example, we could use a perceptron to predict if tomorrow will rain using as input the atmospheric pressure and the humidity percentage of the day before



Understanding machine learning: The perceptron

The perceptron is able to learn (optimize the parameters w_0 , w_1 and w_2) by itself by training on a reference dataset

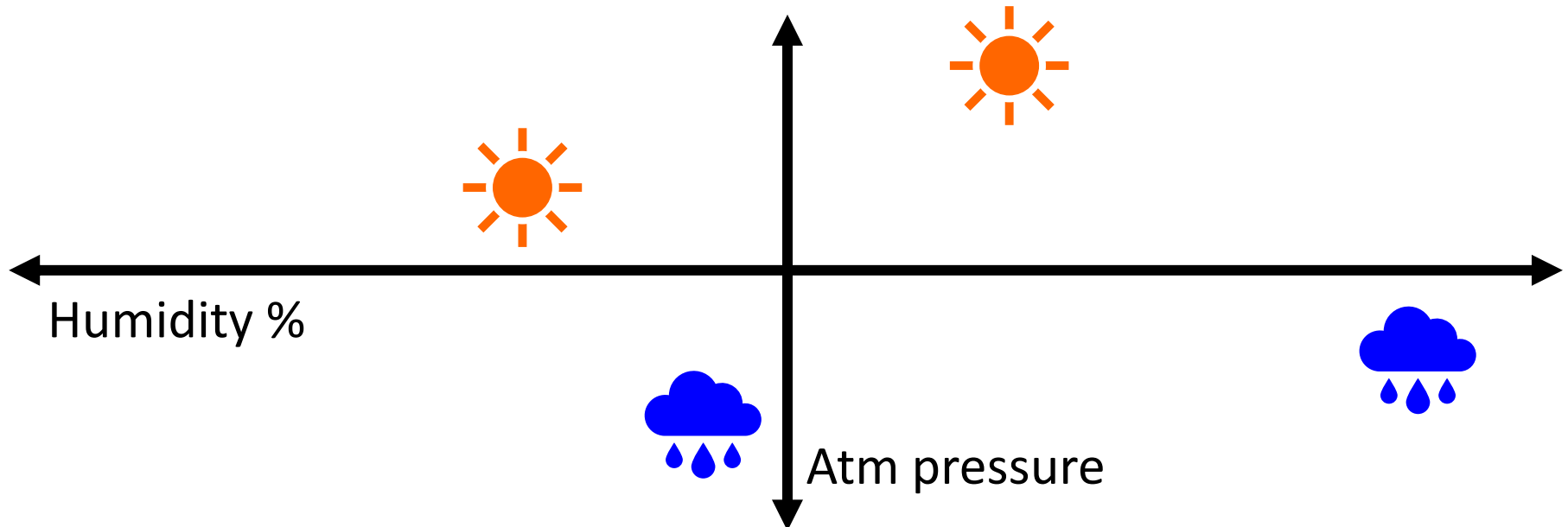
A training dataset consists on a collection of data with the variables and the outcomes we are trying to predict. In our example it would contain atmospheric pressure, humidity percentage and the weather of the day after.

Day	Atmospheric pressure	Humidity percentage	Weather the day after
Day 1	0.96 atm	70%	Rain
Day 2	0.93 atm	40%	Rain
Day 3	1.1 atm	30%	No rain
Day 4	1.23 atm	50%	No rain

Understanding machine learning: The perceptron

The perceptron is able to learn (optimize the parameters w_0 , w_1 and w_2) by itself by training on a reference dataset

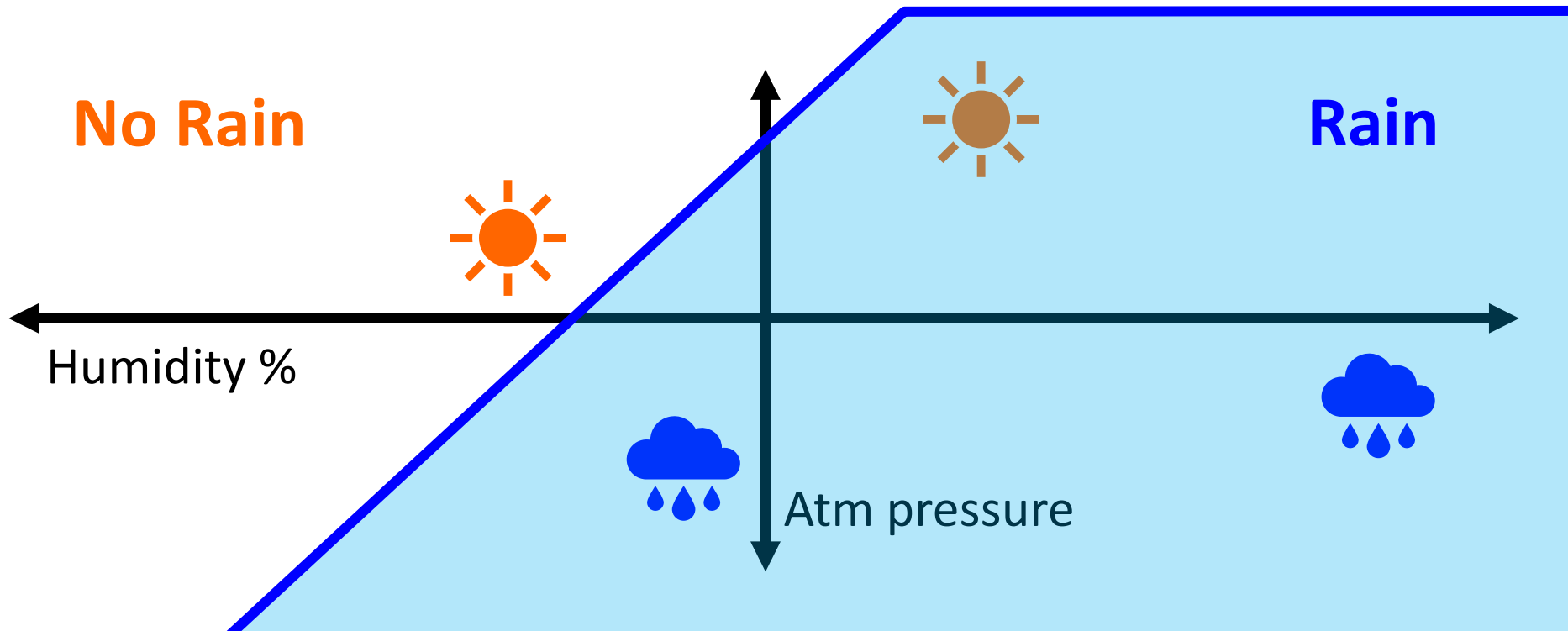
A training dataset consists on a collection of data with the variables and the outcomes we are trying to predict. In our example it would contain atmospheric pressure, humidity percentage and the weather of the day after.



Understanding machine learning: The perceptron

For each day in our dataset, the perceptron calculates X and predicts if that day will be rainy or not

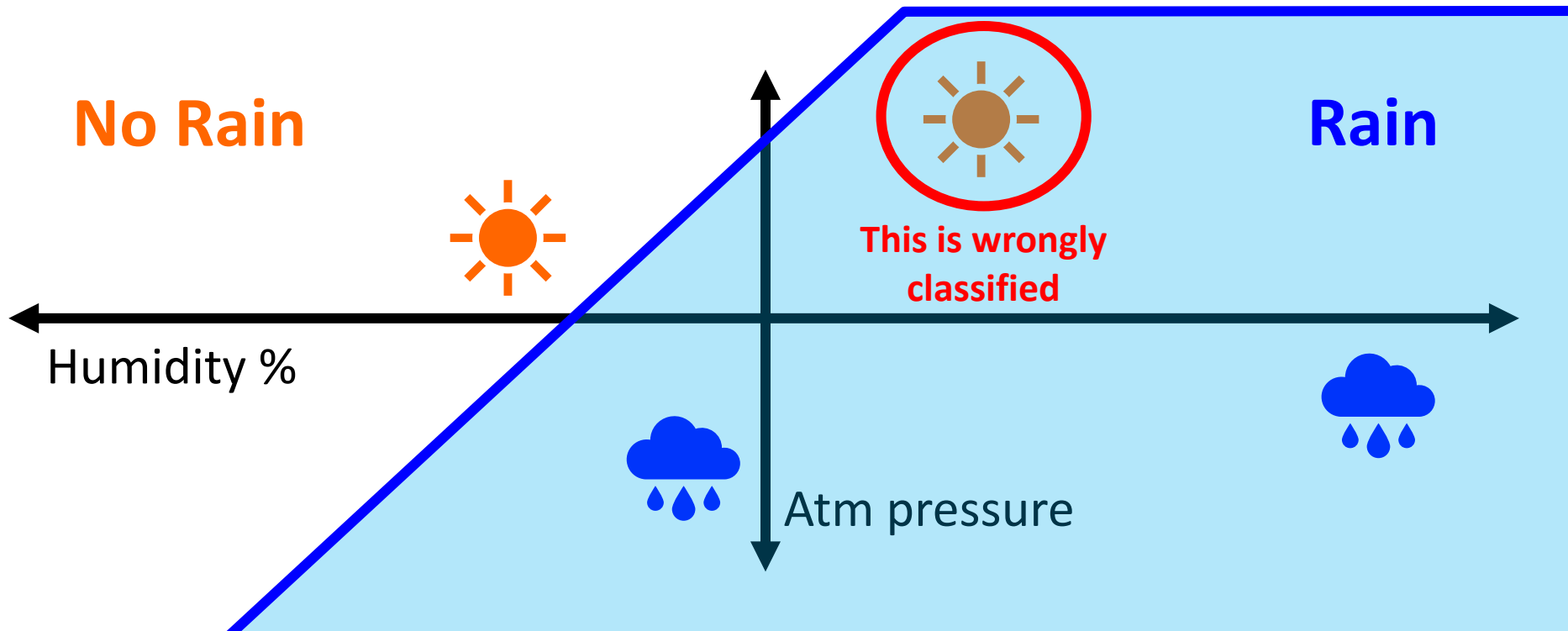
Imagine that we get the following results:



Understanding machine learning: The perceptron

For each day in our dataset, the perceptron calculates X and predicts if that day will be rainy or not

Imagine that we get the following results:



Understanding machine learning: The perceptron

For each day that is wrongly classified, the perceptron updates its parameters (w_0 , w_1 and w_2) by applying the next mathematical formula:

In this example we are going to update w_1 , but the same procedure applies to all the parameters in the perceptron:

$$w_{1,\text{new}} = w_{1,\text{old}} + LR \cdot D \cdot X_1$$

Understanding machine learning: The perceptron

For each day that is wrongly classified, the perceptron updates its parameters (w_0 , w_1 and w_2) by applying the next mathematical formula:

In this example we are going to update w_1 , but the same procedure applies to all the parameters in the perceptron:

$$W_{1,\text{new}} = W_{1,\text{old}} + LR \cdot D \cdot X_1$$

This is the value of the parameter on the next iteration

This is the value of the parameter in the previous iteration

Understanding machine learning: The perceptron

For each day that is wrongly classified, the perceptron updates its parameters (w_0 , w_1 and w_2) by applying the next mathematical formula:

In this example we are going to update w_1 , but the same procedure applies to all the parameters in the perceptron:

$$w_{1,\text{new}} = w_{1,\text{old}} + \text{LR} \cdot D \cdot X_1$$

This is the learning rate and it determines how much the parameters change from iteration to iteration.

Understanding machine learning: The perceptron

For each day that is wrongly classified, the perceptron updates its parameters (w_0 , w_1 and w_2) by applying the next mathematical formula:

In this example we are going to update w_1 , but the same procedure applies to all the parameters in the perceptron:

$$w_{1,\text{new}} = w_{1,\text{old}} + LR \cdot D \cdot x_1$$

Remember that the perceptron calculates a value P to make the prediction

$$w_0 + w_1 \cdot AP + w_2 \cdot \%H = P$$

$P > 0$
rain

$P \leq 0$
no rain

Understanding machine learning: The perceptron

For each day that is wrongly classified, the perceptron updates its parameters (w_0 , w_1 and w_2) by applying the next mathematical formula:

In this example we are going to update w_1 , but the same procedure applies to all the parameters in the perceptron:

$$w_{1,\text{new}} = w_{1,\text{old}} + LR \cdot D \cdot x_1$$

Remember that the perceptron calculates a value P to make the prediction

If P should be higher, $D = 1$

If P should be lower, $D = -1$

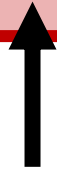
Understanding machine learning: The perceptron

For each day that is wrongly classified, the perceptron updates its parameters (w_0 , w_1 and w_2) by applying the next mathematical formula:

In this example we are going to update w_1 , but the same procedure applies to all the parameters in the perceptron:

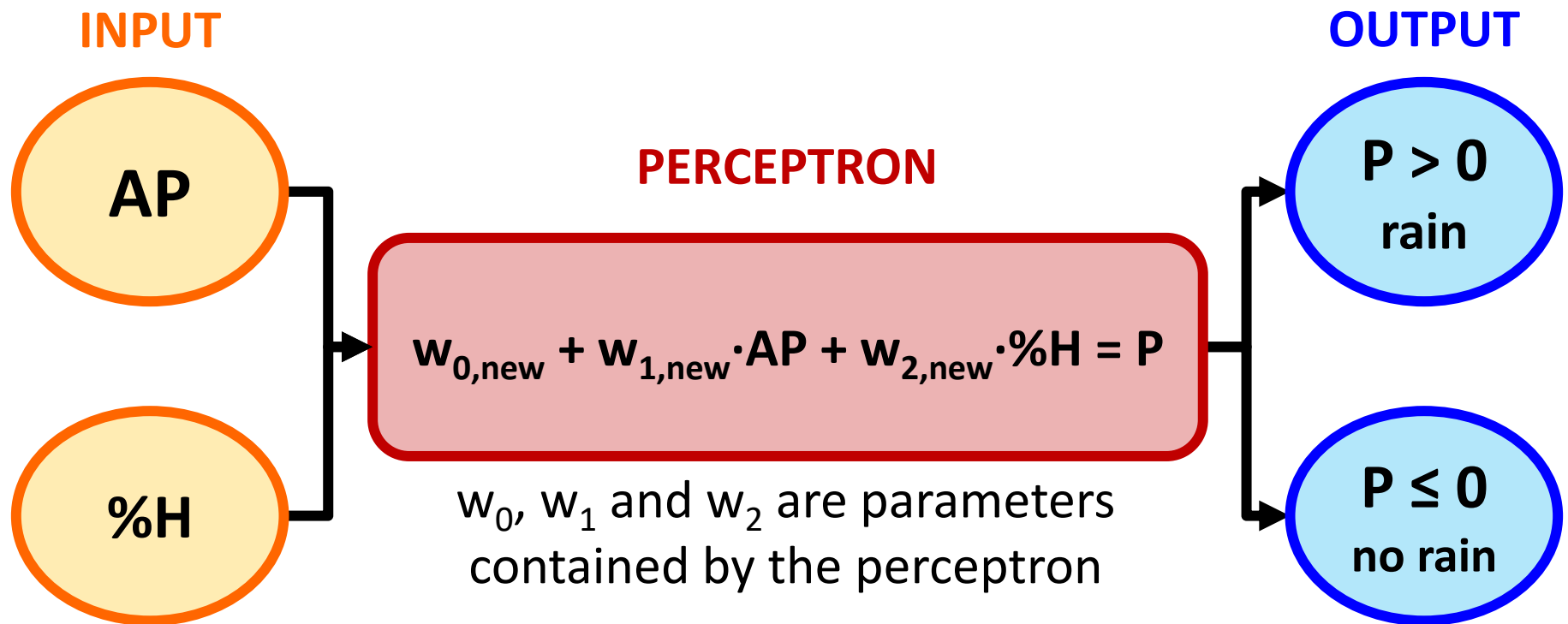
$$w_{1,\text{new}} = w_{1,\text{old}} + \text{LR} \cdot D \cdot X_1$$

X is the value of the missclassified day that is multiplying for the parameter that we are updating. In this case, the parameter that multiplies w_1 is atmospheric pressure

$$w_0 + w_1 \cdot \text{AP} + w_2 \cdot \%H = P$$


Understanding machine learning: The perceptron

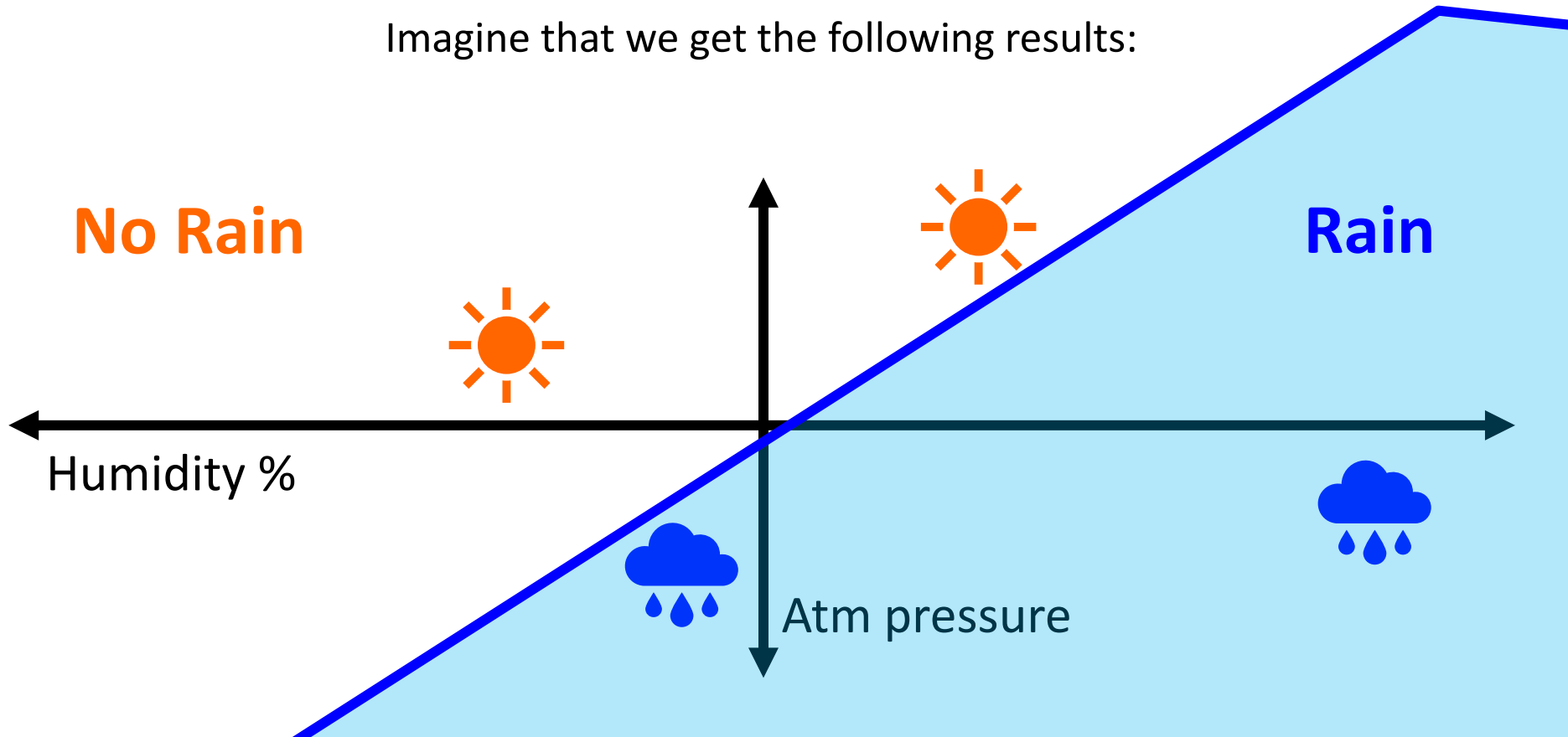
Now we execute the perceptron again with the new parameters



Understanding machine learning: The perceptron

After updating the parameters (w_0 , w_1 and w_2), the perceptron repeats the classification

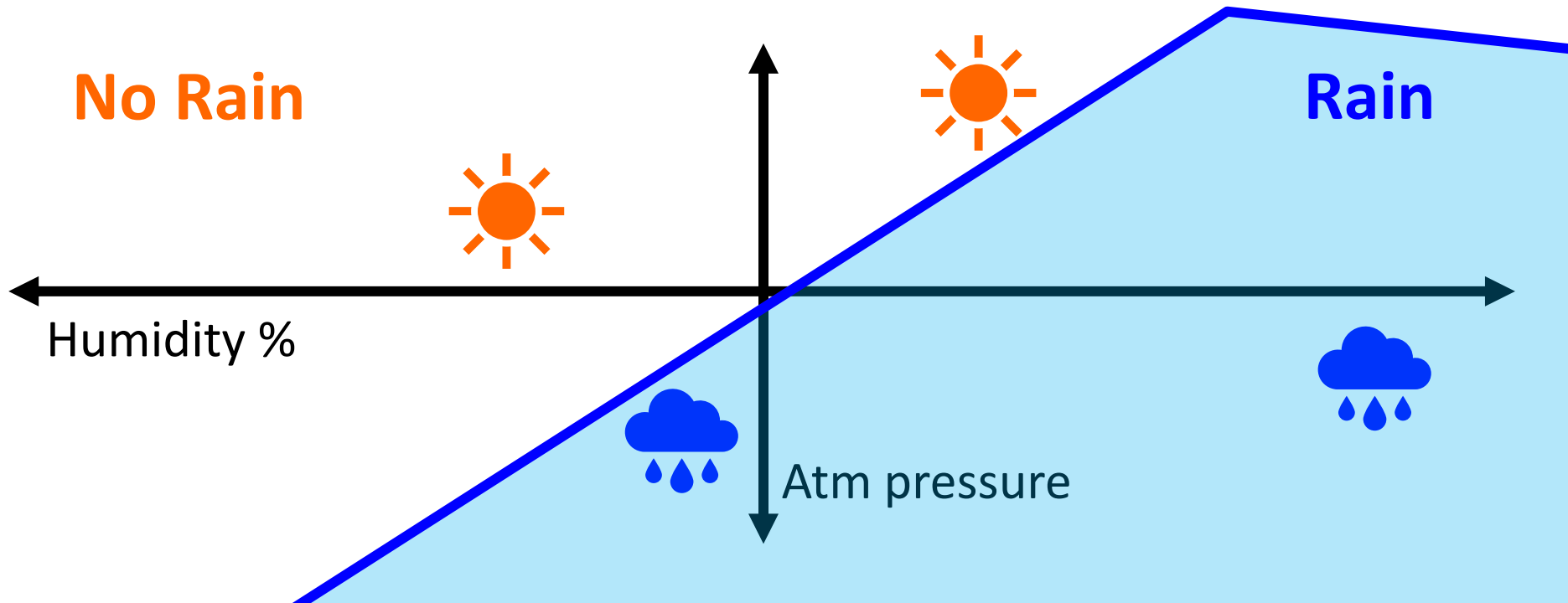
Imagine that we get the following results:



Understanding machine learning: The perceptron

The training of the perceptron finishes after all the elements in the dataset are correctly classified (or we reach a maximum number of iterations)

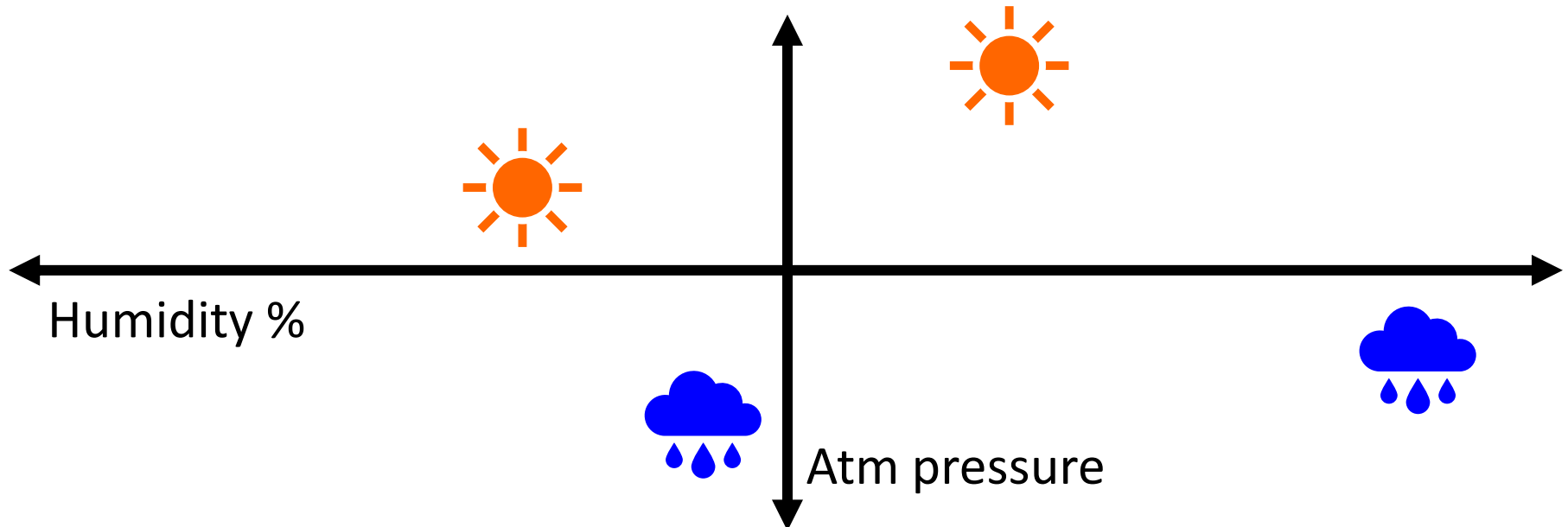
After the training, we can use the perceptron to make predictions on cases for which we don't know the outcome



Understanding machine learning: The perceptron

The quality and size of the training dataset will determine the predictive power of the resulting perceptron

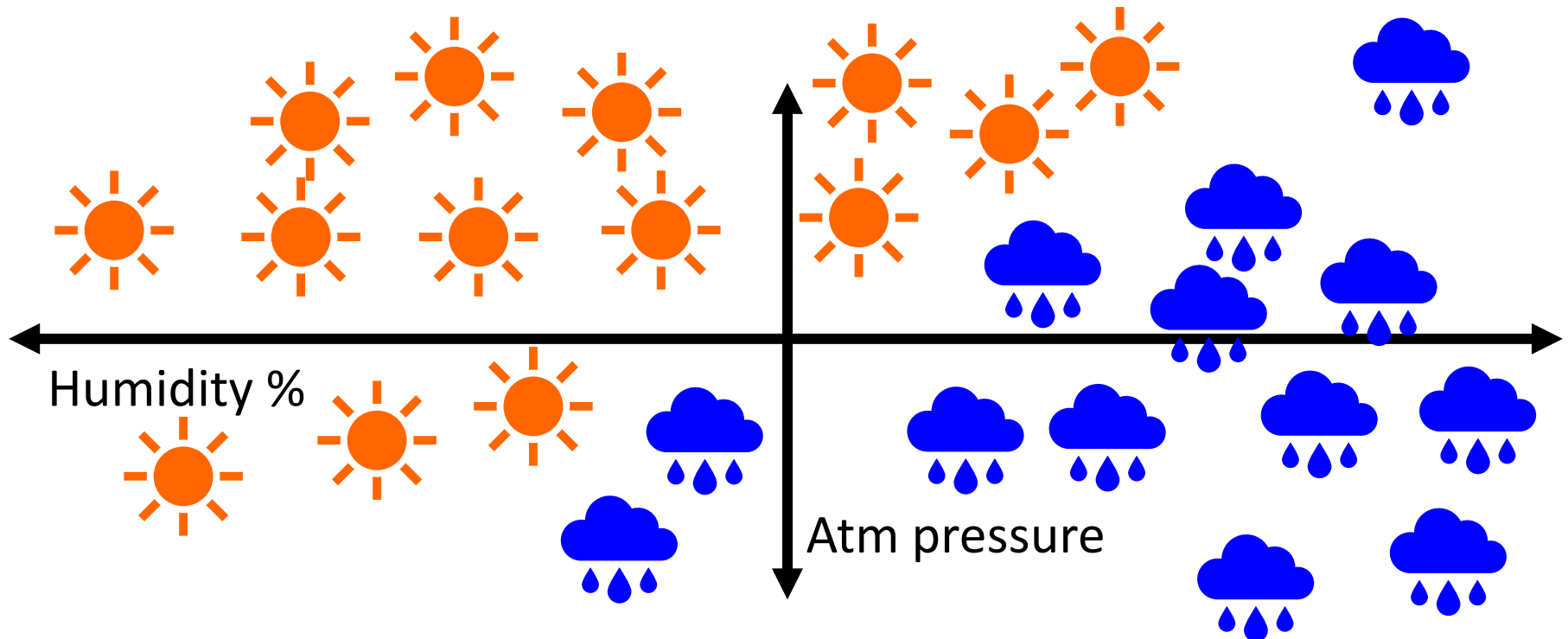
It is not the same obtaining a perceptron with this dataset:



Understanding machine learning: The perceptron

The quality and size of the training dataset will determine the predictive power of the resulting perceptron

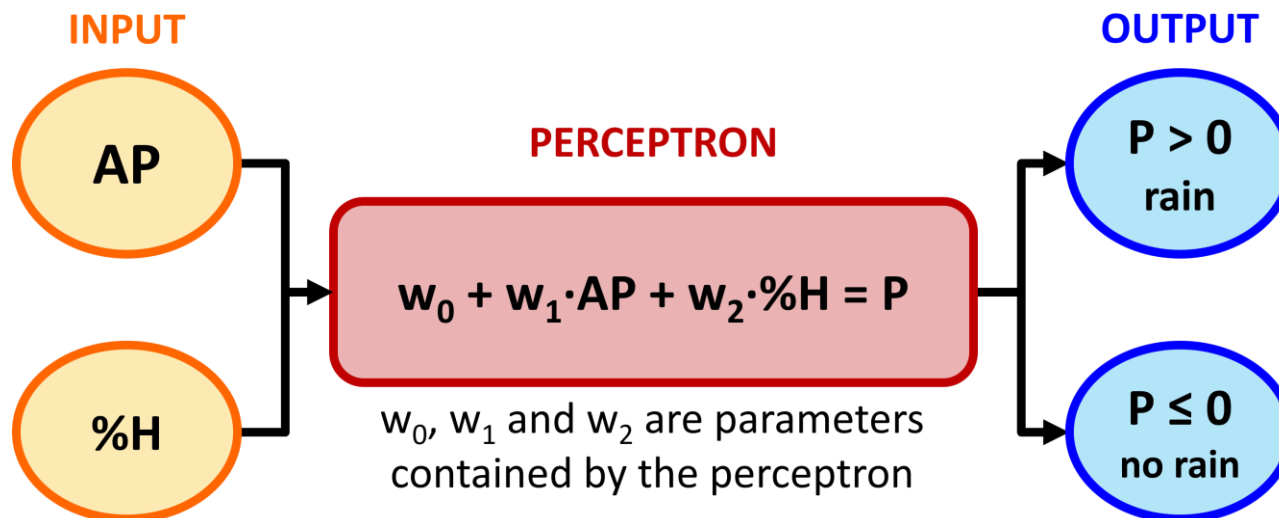
Than with this dataset:



Understanding machine learning: The perceptron

Take home messages from the perceptron:

- Takes an input, performs mathematical operations on it and returns an output
- It needs to train on a dataset of instances with known outcome
- It can self-modify its parameters (w_0, w_1, w_2) to optimize its performance
- The quality and size of the dataset will determine how good the perceptron is optimized



If you want to know more about the perceptron, check out this video:

<https://www.youtube.com/watch?v=4Gac5I64LM4&t=479s>

From the perceptron to deep neural networks

One perceptron

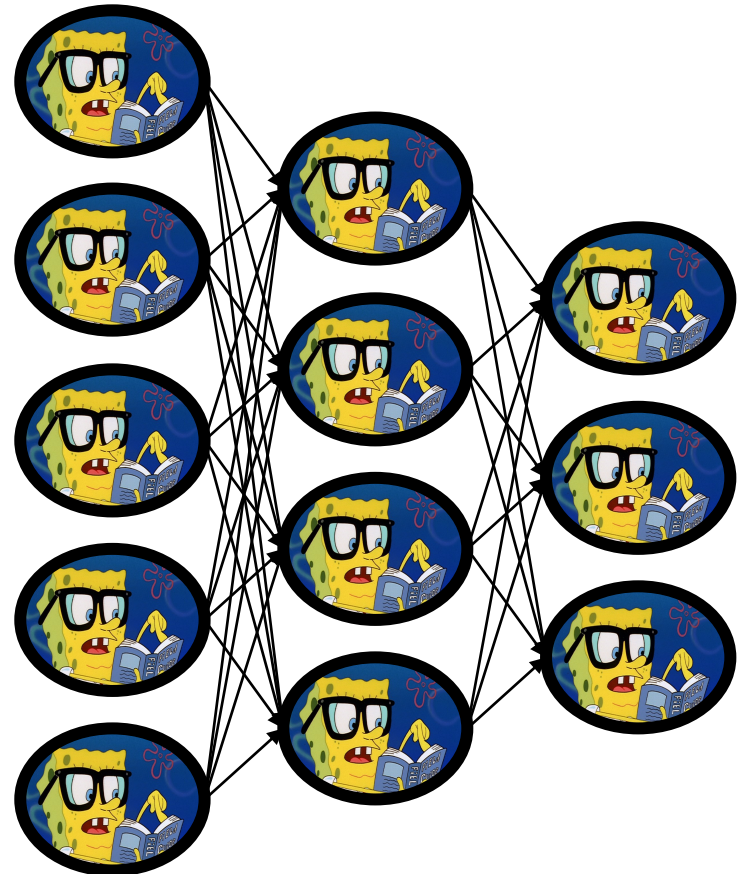
One more complex neuron



From the perceptron to deep neural networks

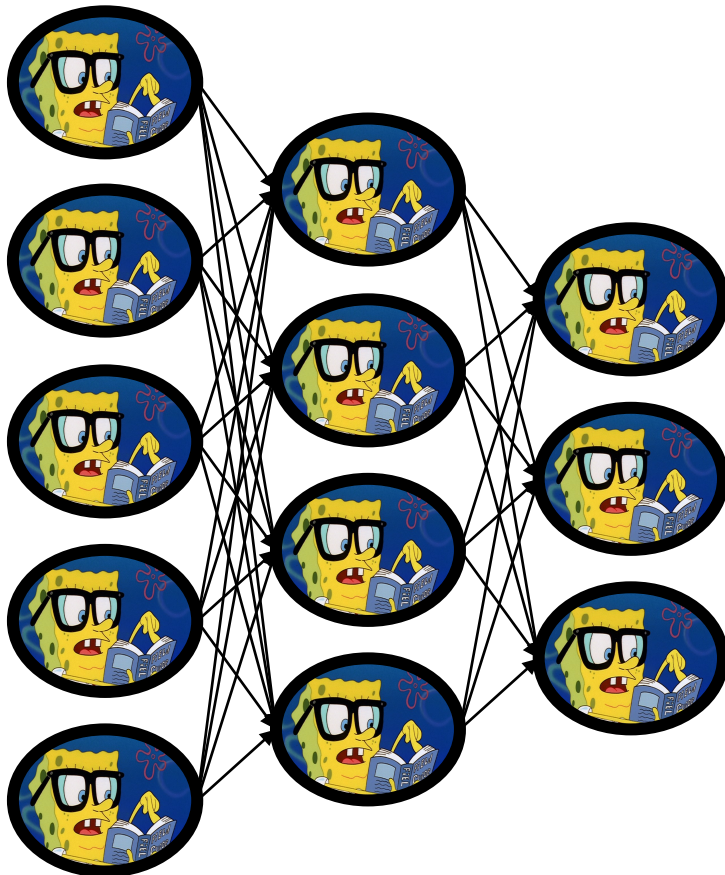
One more complex neuron

Many neurons organized in
a network

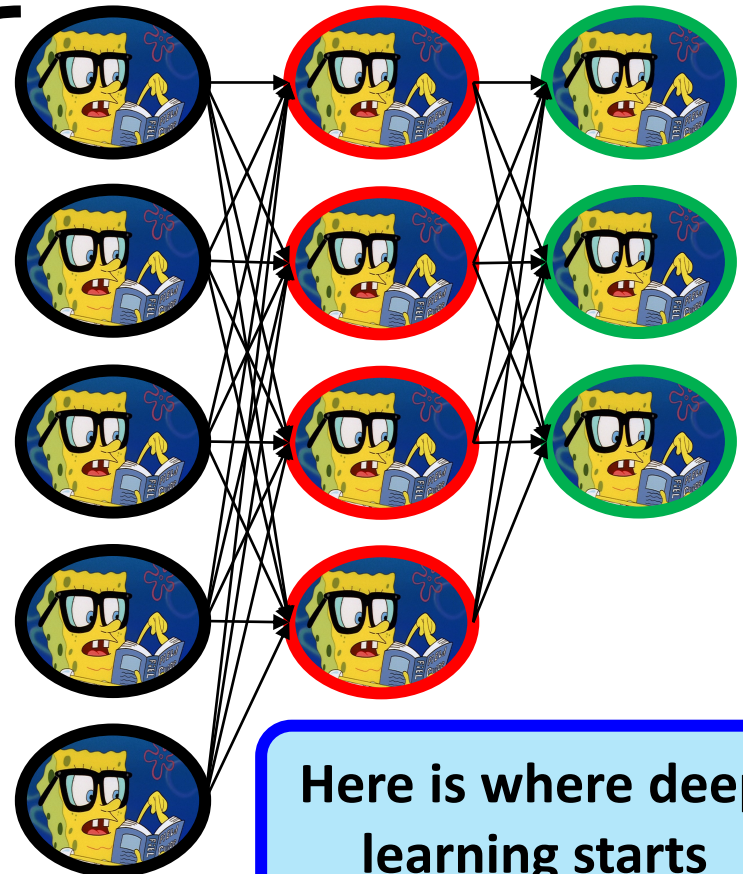


From the perceptron to deep neural networks

Many neurons organized in a network



Networks of complex architecture



Here is where deep learning starts

Where can I use alphafold2?

In this class we will use AlphaFold2 via a google collab session called **CollabFold**:

<https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb>

ColabFold v1.5.2: AlphaFold2 using MMseqs2

Easy to use protein structure and complex prediction using [AlphaFold2](#) and [AlphaFold2-multimer](#). Sequence alignments/templates are generated through [MMseqs2](#) and [HHsearch](#). For more details, see [bottom](#) of the notebook, checkout the [ColabFold GitHub](#) and read our manuscript. Old versions: [v1.4](#), [v1.5.1](#)

[Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: Making protein folding accessible to all. Nature Methods, 2022](#)



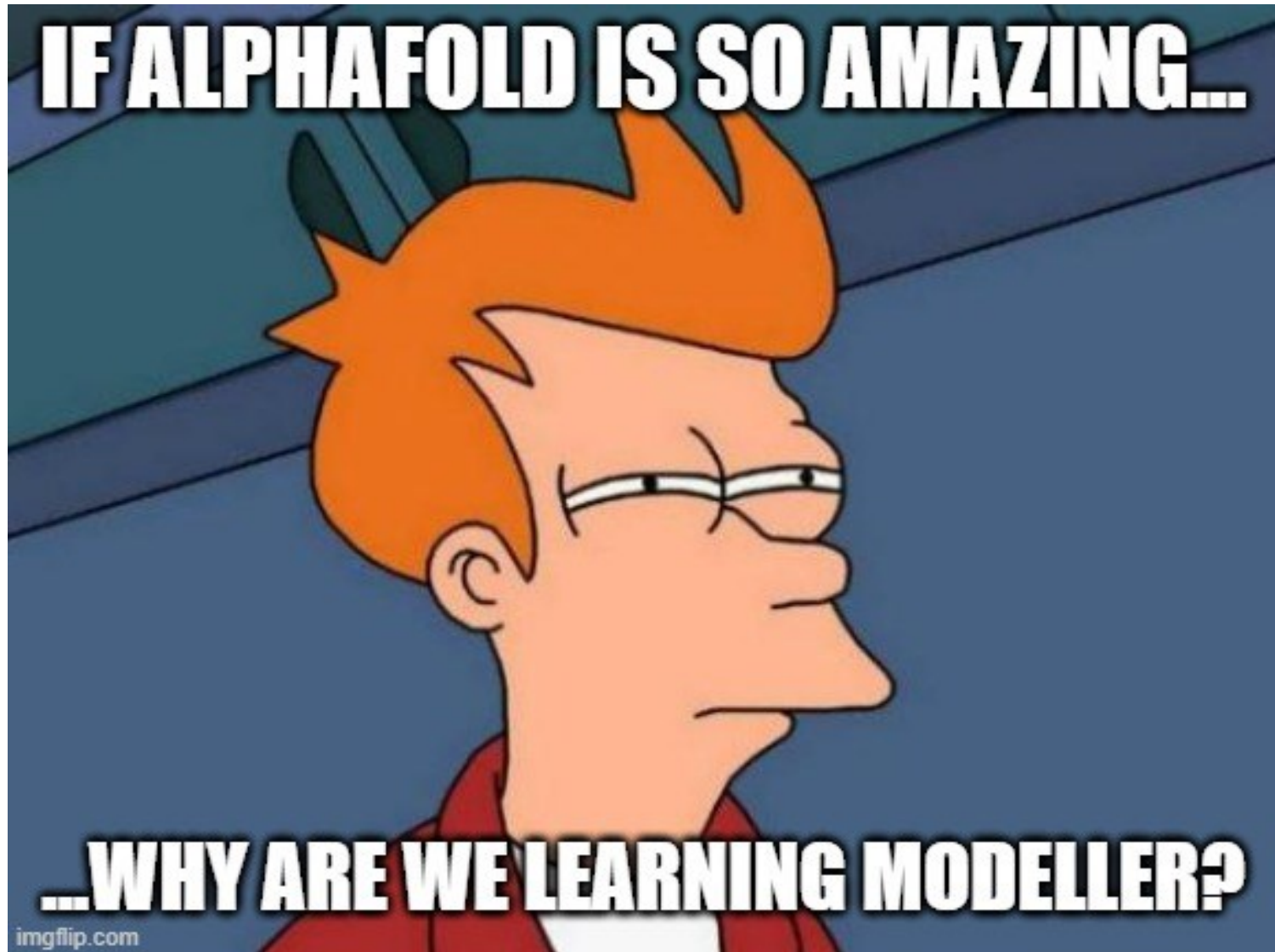
Showing that alphafold is amazing

Step 1: Modeling a globular protein

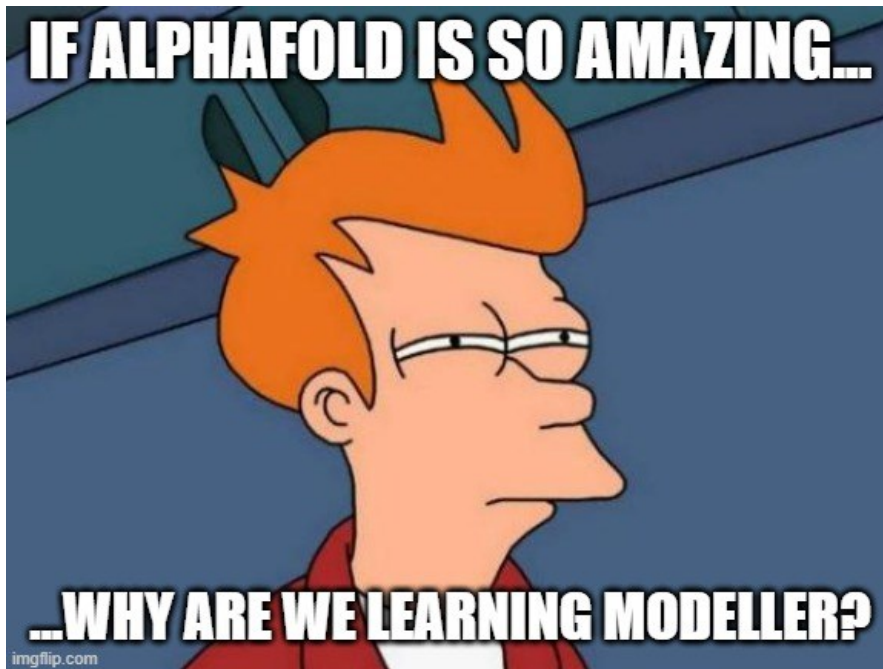
Step 2: Modeling a protein with two independent domains

Step 3: Modeling a chimeric protein

Comparing alphafold to other methods



Comparing alphafold to other methods



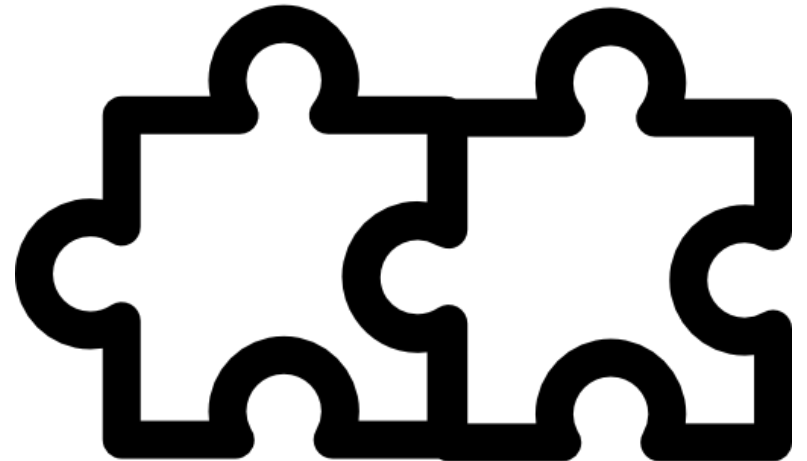
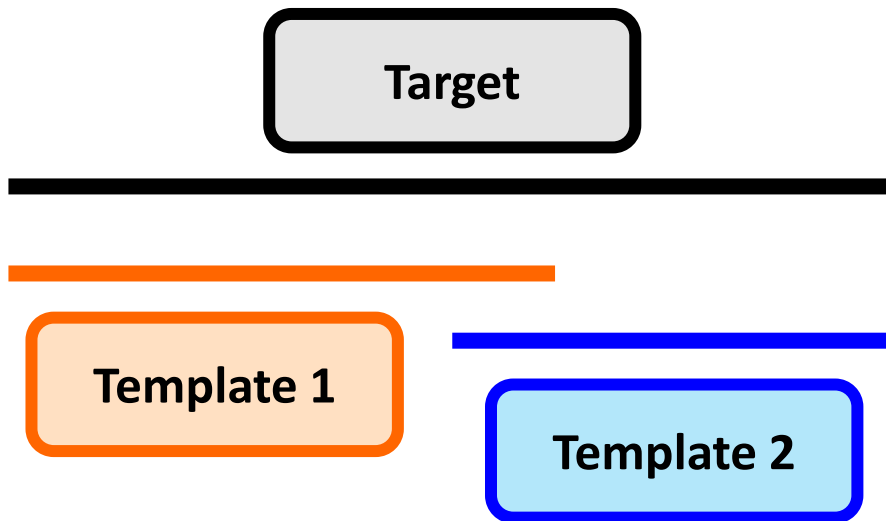
Alphafold uses some principles of homology modeling, but it is not the best way to learn what homology modeling is

Alphafold has some limitations that can be overcome with modeller and/or other methods

Comparing alphafold to other methods

Modeller performs homology modeling by taking structural information of entire proteins and applying it to targets

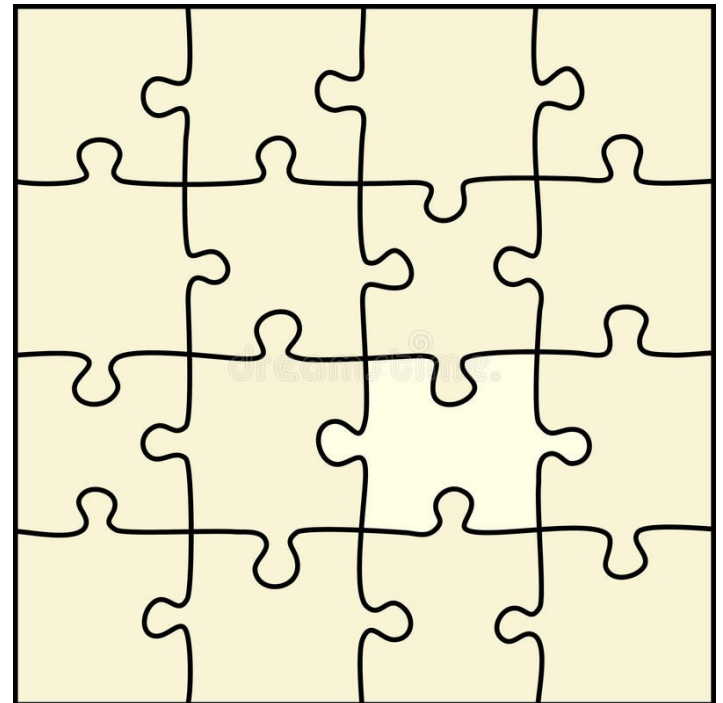
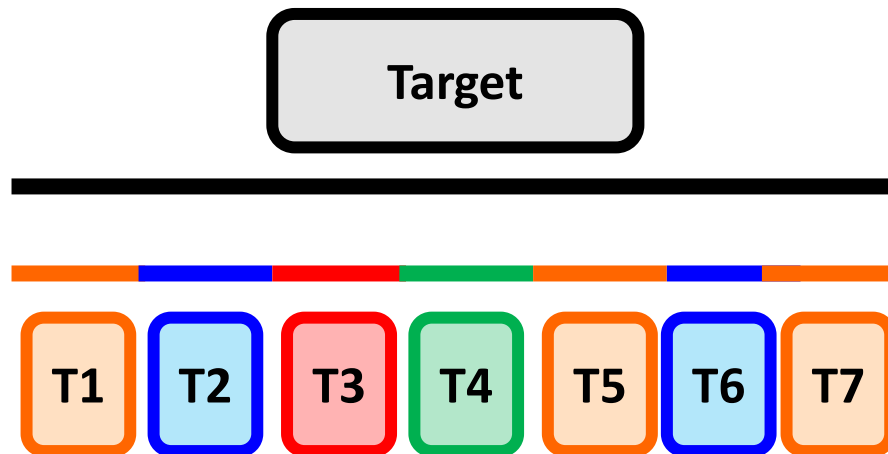
Imagine that we want to model a protein with two domains. The most straightforward approach is to use one template for each domain.



Comparing alphafold to other methods

Rosetta performs homology modeling by taking structural information of fragments of 9 amino acids and applying it to targets

In comparison with modeller, now we are using a larger number of templates and each template corresponds with a smaller region of the target



Comparing alphafold to other methods

Alphafold performs homology modeling by taking structural information and using neural networks to deconstruct this information at amino acid level

We are using structural information that is independent for each amino acid, and yet it is wholly integrated to generate a model that makes sense

Target

Structural information

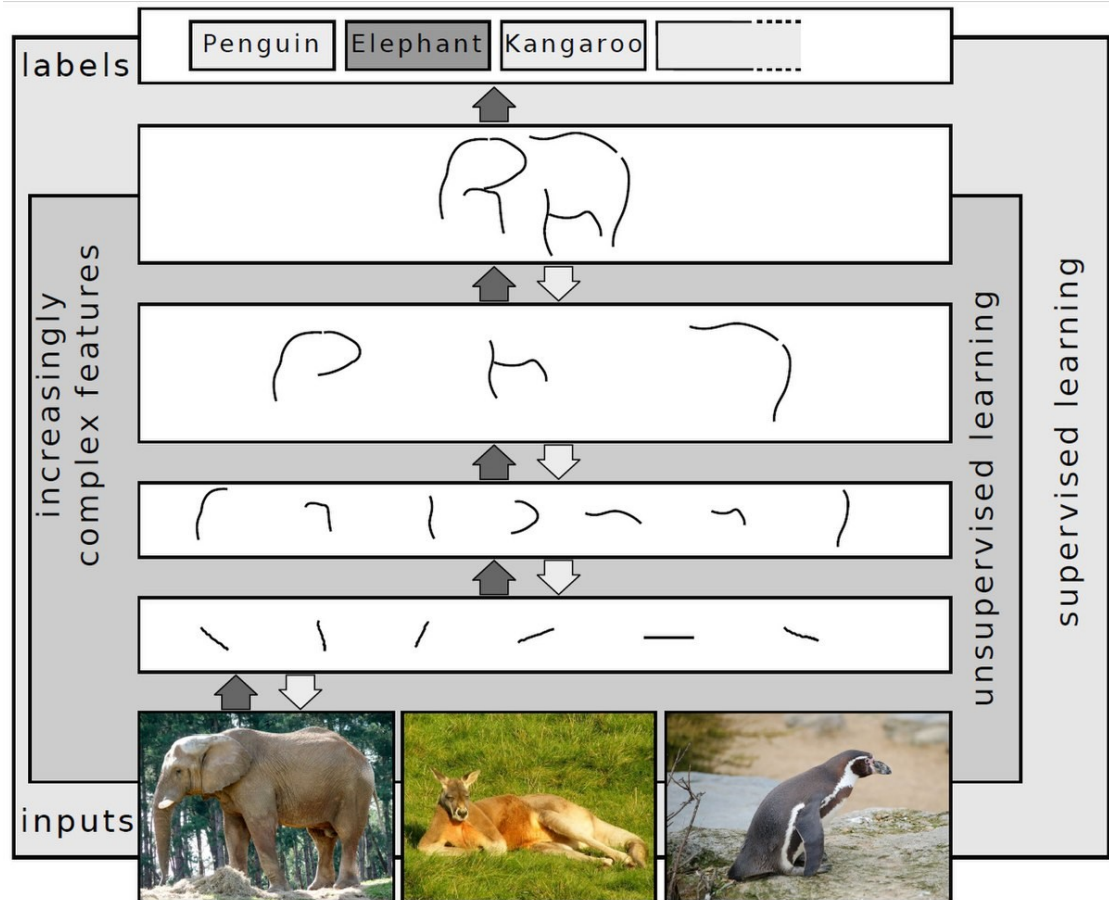


Comparing alphafold to other methods

At this point, alphafold resembles convolutional neural networks for the recognition of images

See how deep learning algorithms are able to extract features for independent elements of the image and then combine them into something meaningful.

Something similar happens with the information per amino acid in alphafold.



Comparing alphafold to other methods

Alphafold gives small freedom to the user. This makes it easier to use, but if you don't get the results you want there are few options for improvement.

Modeller gives the user control on what templates to use and gives access to the alignment

Alphafold gives the user small control on what structural information is used or how it is used

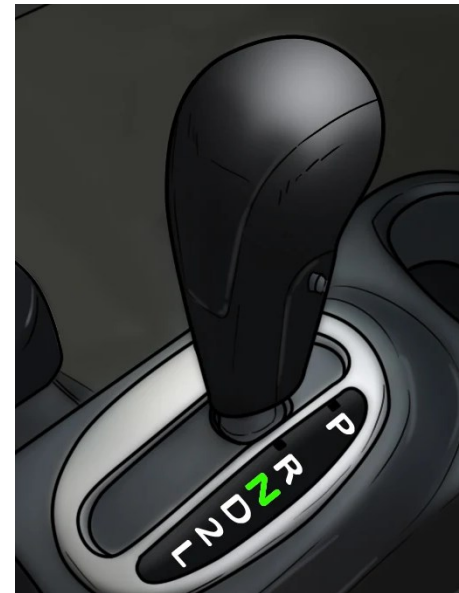
Comparing alphafold to other methods

Alphafold gives small freedom to the user. This makes it easier to use, but if you don't get the results you want there are few options for improvement.

Modeller gives the user control on what templates to use and gives access to the alignment



Alphafold gives the user small control on what structural information is used or how it is used



AlphaFold still has limitations

Step 4: AlphaFold struggles with proteins for which there is no available experimental structures

Step 5: AlphaFold doesn't take into account the biological characteristics of proteins. Handles membranes poorly.

Step 6: AlphaFold is not good making models of disordered proteins (although no program is)