Interpretability for computational biology

An-phi Nguyen, Maria Rodriguez-Martinez

IBM **Research**

ETH zürich

Motivation

Why do we need interpretability to unveil the decision process of a machine learning model?

Trust

for high-risk scenarios, e.g. healthcare, the user needs to trust the decision taken.

Debugging

the model may be badly trained or there might be an unfair bias in either the dataset or the model itself.

Hypothesis generation

surprising results might be consequences of new mechanisms or patterns unknown even to field experts.

What is interpretability?

From Merriam-webster (2019):

"tell the meaning of:present in understandableterms",
"to make known, plain or understandable"

But...

NO AGREED UPON **FORMAL** DEFINITION

Evaluating interpretability

What should we consider for evaluation? [1]

Expressive power: are explanations presented in an understandable format?

Usability: should the interpretability method be agnostic with respect to the model-to-be-explained?

Runtime: how fast can the interpretability method produce explanations?

Fidelity and consistency: are the explanations produced the actual decision process of the model-to-be-explained?

Representativeness: how well do the explanations "generalize"?

Uniqueness and stability: is the explanation always unique? If so, should "similar" samples have "similar" explanations?

How should we evaluate? [2]

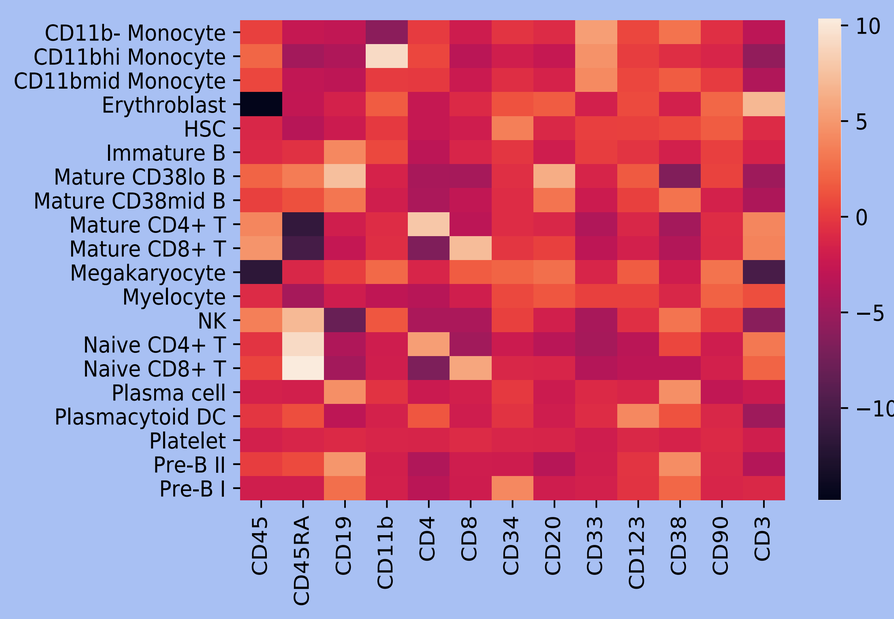
Human evaluation: should field experts or laymen evaluate the explanations?

Programmatic evaluation: is it possible to define proxy metrics that somehow correlates with what we consider interpretable?

Interpret me!

Cell Line classification

Which markers (columns) are important to classify a cell line (rows)?
- Visualizing the weights of a shallow neural network.

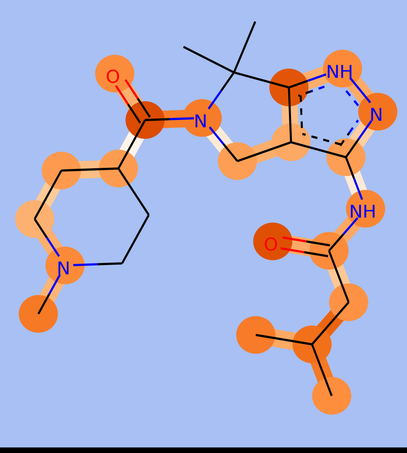


Transcription factor binding prediction

Which nucleotides are important for binding CTCF?
- Visualizing importance of nucleotides as computed by LIME [3] and Anchors [4]. At the bottom the motif as provided by JASPAR [5].

IC50 score prediction

Which chemical components of the drug PHA-793887 are important for predicting its IC50 score?
- Visualizing the importance score of the components as computed by an attention mechanism [6].



Follow our GitHub Repo!

For further discussion, contact me at
uye@zurich.ibm.com

[1] Molnar C., https://christophm.github.io/interpretable-ml-book/, 2019.
[2] Doshi-Velez F. et al., arXiv:1702.08608, 2017.
[3] Ribeiro M. T. et al., KDD ,2016.
[4] Ribeiro M. T. et al., AAAI, 2018.
[5] Khan A. et al., NAR, 2018.
[6] Manica M. et al., WCB ICML, 2019.

Individualized PaediatricCure

The project leading to this publication has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826121