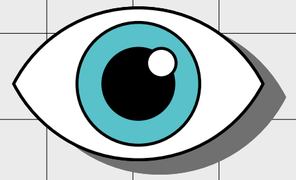
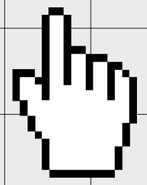
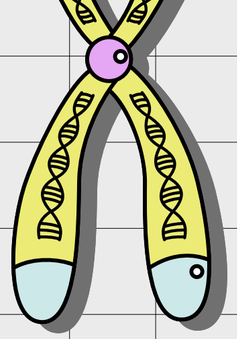
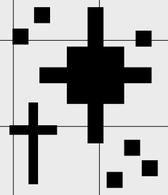
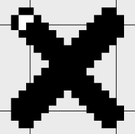
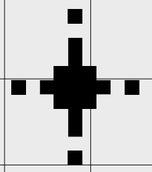
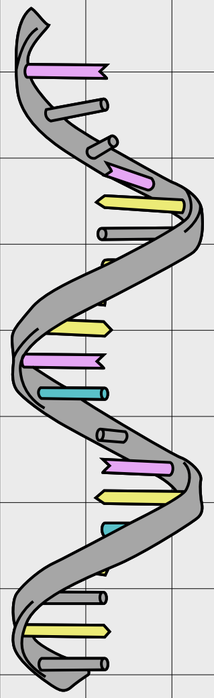
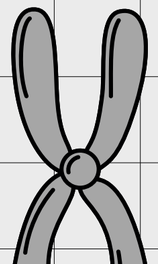
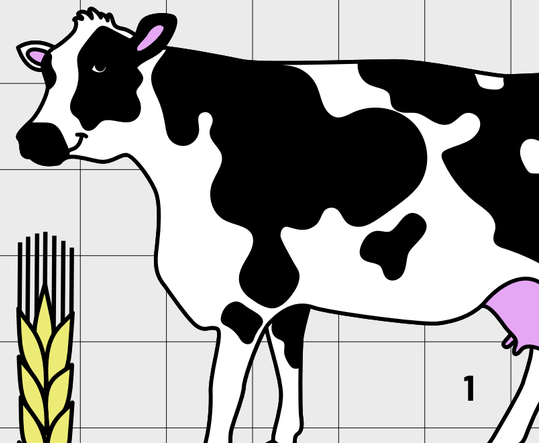








APPLYING INTERPRETABLE MACHINE LEARNING IN COMPUTATIONAL BIOLOGY

Pitfalls, Recommendations, and Opportunities for New Developments

Biological Data, December 2024

Rana İşlek

OUTLINE

1. +Introduction
2. +IML Methods: Post Hoc & By-Design
   +Evaluation Metrics
   +Common Pitfalls
   +Opportunities & Recommendations
3. +Conclusion
   +Q/A

INTRODUCTION

Recent advances in machine learning (ML) is transforming computational biology by enabling high-throughput data acquisition and complex predictions.

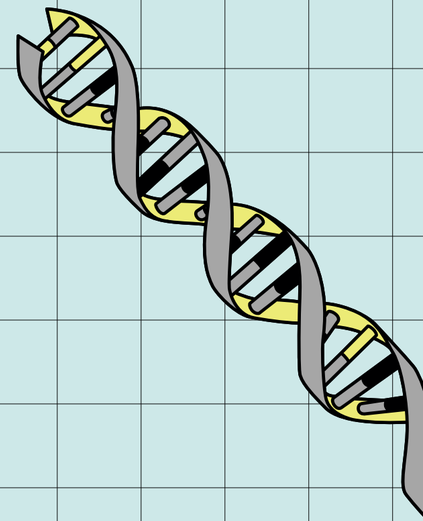
With this rapid development of new models, there is a growing need for techniques that can provide interpretation or understanding of model behavior, enabling researchers to verify that the proposed model reflects actual biological mechanisms. **-> Interpretable Machine Learning (IML) bridges the gap between black-box predictions and actionable biological insights.**

However, guidelines for using IML in computational biology are generally underdeveloped.

WHAT IS INTERPRETABLE MACHINE LEARNING (IML)?

Interpretable Machine Learning (IML) methods are techniques and approaches used **to make machine learning models** more understandable and ***explainable to humans***.

These methods aim **to clarify how a machine learning model makes its predictions** or **decisions**, making them ***transparent*** and **trustworthy**, especially *in high-stakes domains* like healthcare, **biology**, and finance.



WHAT IS INTERPRETABLE MACHINE LEARNING (IML)?

Interpretable Machine Learning transforms black-box models into transparent systems, revealing the 'why' behind predictions for trust and actionable insights.

"Why does the model predict what it predicts?"

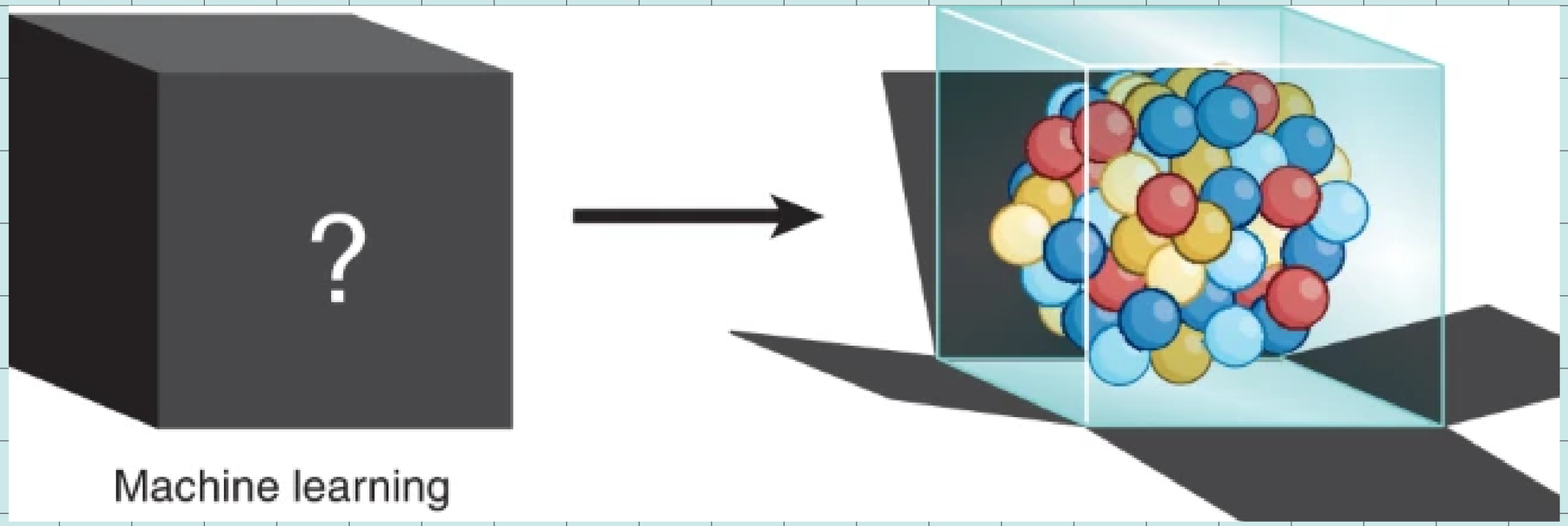


Image retrieved from: Nature Communications. (2022). Explaining the Predictions of Machine Learning Models in Chemistry.

WHAT IS IML?
AN EXAMPLE FROM BIOLOGY

We should not trust machine learning blindly; understanding the 'why' behind predictions through explanations is crucial for ensuring reliability, transparency, and actionable insights.

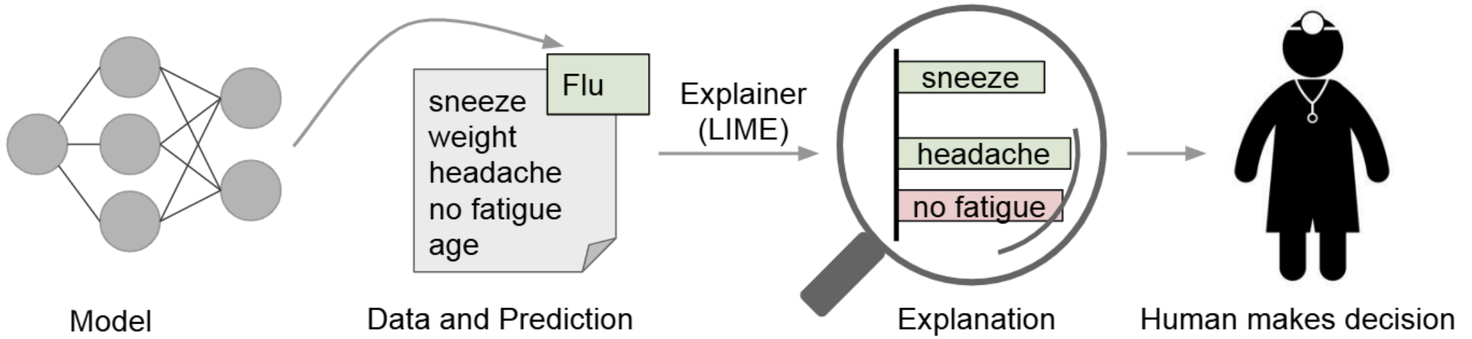


Figure 1: Explaining individual predictions. A model predicts that a patient has the flu, and LIME highlights the symptoms in the patient's history that led to the prediction. Sneeze and headache are portrayed as contributing to the "flu" prediction, while "no fatigue" is evidence against it. With these, a doctor can make an informed decision about whether to trust the model's prediction.

IML allows a doctor to critically assess the model's reasoning and make a more informed decision rather than relying solely on the model's output.

Figure retrieved from: Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). **"Why Should I Trust You?"** Explaining the Predictions of Any Classifier.

IML METHODS

IML - **I**nterpretable **M**achine **L**earning

2 DIFFERENT METHODS & EXPLANATIONS

These are applied ***after* a model has been trained** and **predictions have been made**. They focus on explaining the behavior of a black-box model **without changing its structure**.

These methods **assign each input feature**, *such as a pixel in a cellular image or a nucleotide in a DNA sequence*, **an importance value** based on its contribution to the model prediction.

Post Hoc Explanations

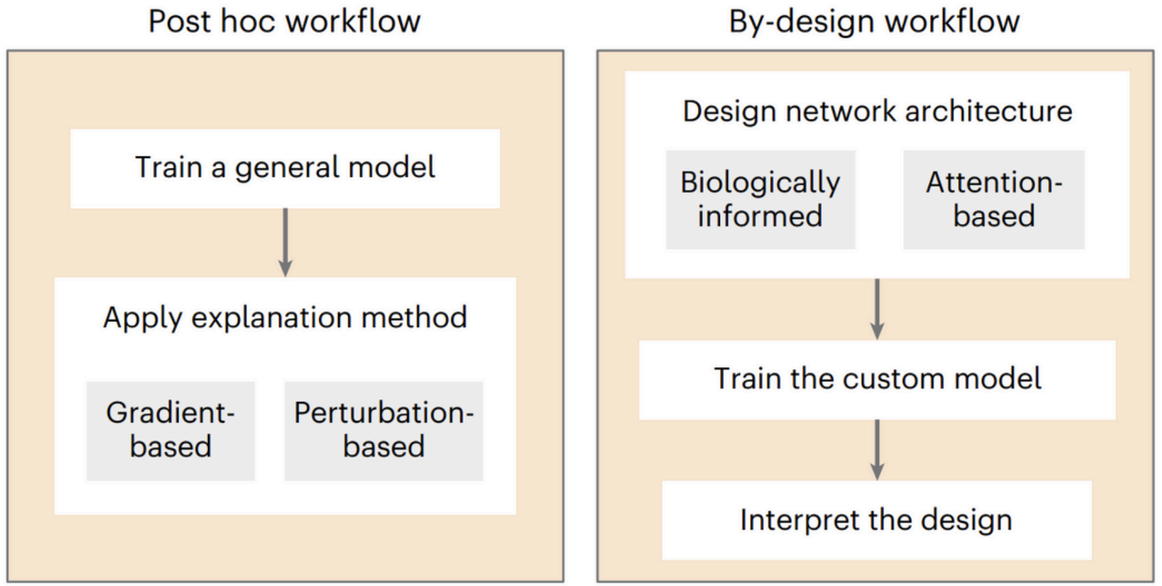


Fig. 1 | The two main IML approaches used to explain prediction models are post hoc explanations and by-design explanations. Each approach has its canonical workflows and popular types of IML methods: post hoc explanations are model agnostic and are applied after a model is trained, while by-design explanations are typically built into or inherent to the model architecture.

These are models that are ***naturally interpretable***. These methods build interpretability directly into the model architecture, so **the model's predictions are inherently understandable**.

For instance, **linear models** *(coefficient weights to ascertain the importance of each feature to the prediction outcome)* and **decision trees** *(interpretable as one can examine the splits in the tree)*.

By-Design Explanations

CLASSICAL BY-DESIGN MODELS

Classic examples include linear models and decision trees, where coefficients or splits directly explain feature importance.

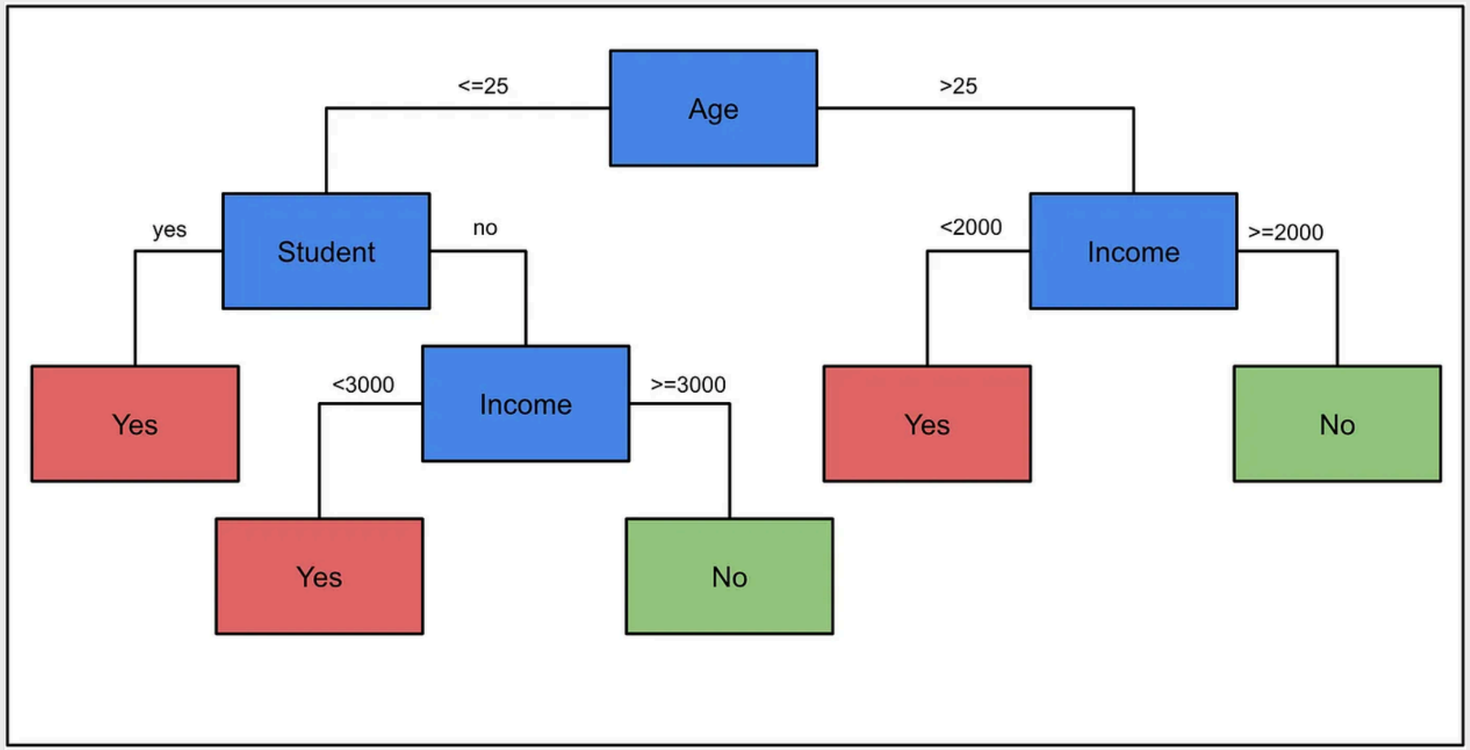


Figure retrieved from: Towards Data Science. (n.d.). Interpretable vs. Explainable Machine Learning.

BY-DESIGN EXPLANATIONS

BIOLOGICALLY INFORMED NEURAL NETWORKS

BINN are specific models for computational biology, embedding domain knowledge into the architecture; enabling insights by interpreting the hidden layers, **such as using node weights to infer biological relevance.**

-**DCell:** Models cell subsystems, linking hidden nodes to biological entities like genes
-**P-NET:** Incorporates biological pathways into the network design.
-**KPNN:** Integrates gene regulatory and protein signaling networks for interpretable predictions.

ATTENTION MECHANISMS

Attention handles sequence-based inputs. It assigns **importance weights to input features**, which **do not incorporate domain knowledge,** are **automatically learned** as part of the training process and have been shown empirically to assist the network in **focusing on the correct parts of the input sequence.**

-**Transformers:** Use self-attention (e.g., BERT) to identify relationships in sequence-based data.
-**Enformer:** Leverages attention to identify enhancers regulating gene expression.

POST HOC EXPLANATIONS

GRADIENT-BASED METHODS

It computes the gradient of the model's output with respect to each input feature to determine how changes in the input affect predictions.

-**DeepLIFT**: In NN, compares the activation of each neuron to its reference and assigns contribution scores according to the difference.
-**Grad-CAM:** Uses gradients to generate visual explanations for predictions in convolutional neural networks.

PERTURBATION-BASED METHODS

Assess feature importance by systematically modifying or perturbing input features and observing the impact on predictions.

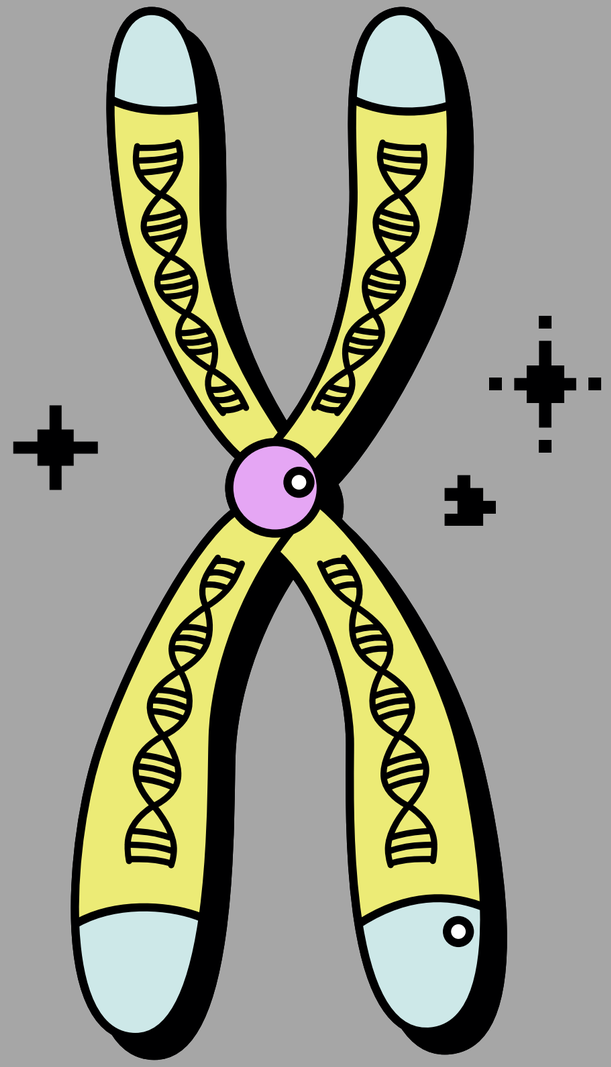
-**SHAP (Shapley Additive Explanations):** It uses a game theoretic approach that measures each player's contribution to the final outcome.
-**LIME (Local Interpretable Model-Agnostic Explanations):** Modify input features locally to approximate the behavior of the model.

EVALUATION METRICS

Evaluation metrics are essential in machine learning because they provide a **standardized way** to assess the *performance*, *reliability*, and **interpretability** of models.

Two metrics that frequently appear in the IML literature and increasingly adopted by computational biology publications:

- **Faithfulness:**
  - How accurately does the IML explanation reflect the actual mechanisms of the model?
- **Stability:**
  - Do similar inputs produce consistent explanations?
- Both are critical for ensuring biological reliability and model interpretability.



To algorithmically assess the quality of explanations generated by IML methods, several concepts have been proposed, which are generally agnostic to the type of IML method that is applied.

COMMON PITFALLS

As the computational biology community increasingly adopts IML methods to understand machine learning model behaviors, three pitfalls have been identified which should be avoided when using IML.

Pitfall 1: Different methods may lead to conflicting interpretations. Using multiple methods provides a more comprehensive understanding. -> **faithfulness**

Pitfall 2: Importance scores/output labels often require post-processing to connect with biological insights. Without it, interpretations may lack biological relevance.

Pitfall 3: Choosing only favorable findings while ignoring contradictory ones can lead to misleading conclusions. Robust evaluations are necessary to avoid it. -> **stability**

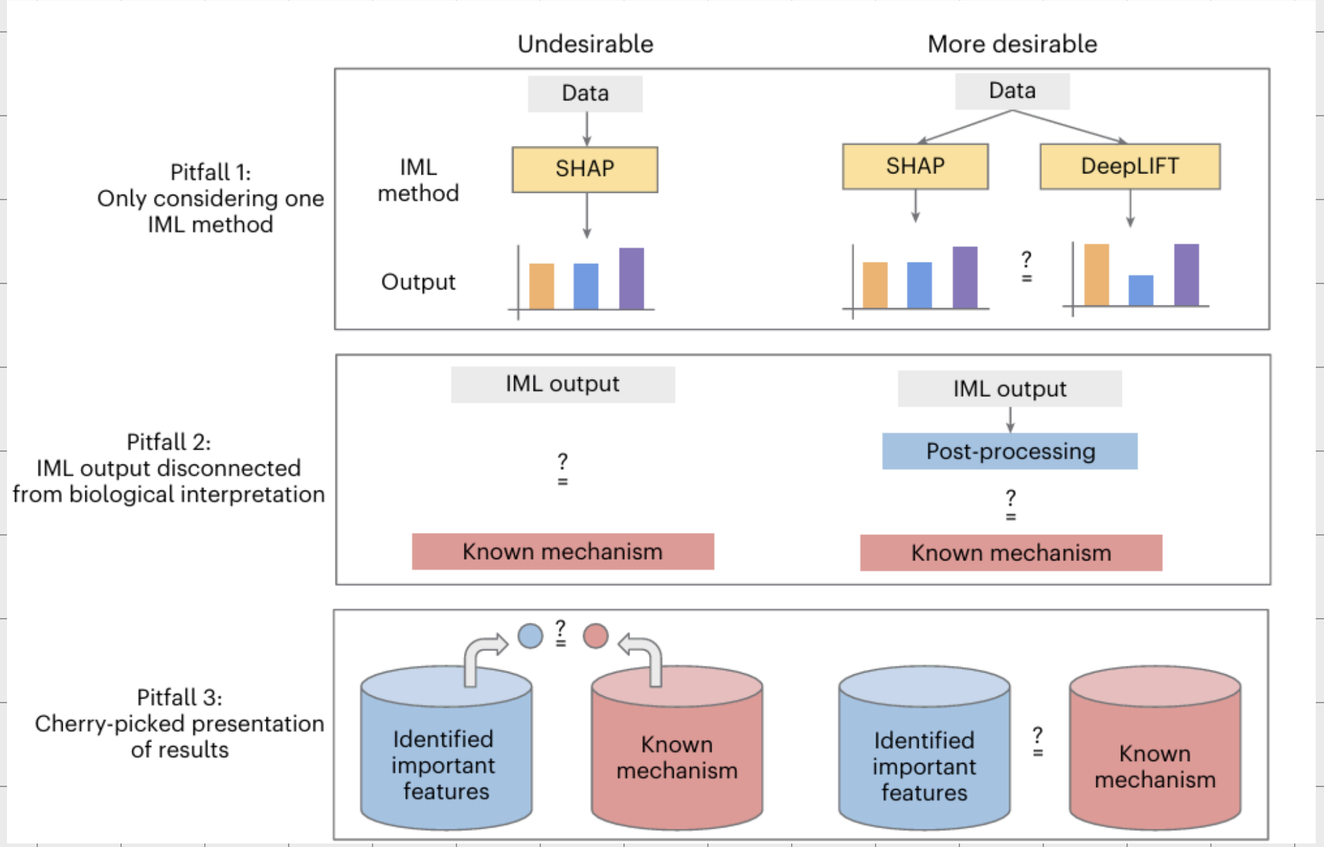


Figure retrieved from: Applying interpretable machine learning in computational biology—pitfalls, recommendations and opportunities for new developments..

RECOMMENDATIONS

Recent advancements in Large Language Models (LLMs), such as Enformer and Geneformer, have transformed predictive modeling in biology. However, IML techniques for LLMs are still underdeveloped.

Besides establishing better practices to avoid the pitfalls of IML usage, there are multiple opportunities to develop novel IML techniques for new model architectures and biological applications.

Key opportunities:

1. **Tokenization:** Optimizing how biological data, like DNA sequences, are broken down into smaller units for processing.
2. **Adaptation:** Applying LLM-specific techniques like mechanistic interpretability to biological contexts.
3. **Multimodal Applications:** Integrating and explaining diverse datasets, such as genomics and imaging.
4. **Visualization Tools:** Developing platforms like DNABERT-Viz to better analyze attention scores in biological sequences.

Recommendations:

- Use domain-specific knowledge to enhance model reliability.
- Standardize evaluation metrics for LLMs and multimodal IML.
- Promote collaboration between biology and machine learning experts.

CONCLUSION

In conclusion, Interpretable Machine Learning (IML) is becoming a vital tool in computational biology, but its effective application requires standardized guidelines and best practices.

Throughout this presentation, we:

1. Explored **common IML methods** and **evaluation metrics**.
2. Identified **three major pitfalls** in current evaluation practice, applied in a the computational biology context.
3. Highlighted **the importance of improving the reliability** and **interpretability** of **IML** predictions through robust validation approaches.

The future of IML depends on creating these standards and fostering collaboration between computational biology and machine learning experts. By achieving this, we can uncover new biological insights and significantly advance biomedical research.

REFERENCES

1. **Chen, V., Yang, M., Cui, W., Kim, J. S., Talwalkar, A., & Ma, J. (2024). Applying interpretable machine learning in computational biology—pitfalls, recommendations and opportunities for new developments.**
2. Datacamp (n.d.). An introduction to SHAP values for machine learning interpretability. https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability
3. GitHub Repository for DeepLIFT. https://edwinwenink.github.io/ai-ethics-tool-landscape/tools/deeplift/
4. Molnar, C. (n.d.). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.
5. Nature Methods, 21, 1454–1461. https://doi.org/10.1038/s41592-024-02359-7
6. Online Book: Importance of Interpretability. https://christophm.github.io/interpretable-ml-book/interpretability-importance.html
7. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). Association for Computing Machinery. DOI: 10.1145/2939672.2939778
8. Shrikumar, A., Greenside, P., & Kundaje, A. DeepLIFT: Learning important features through propagating activation differences.

Q/A