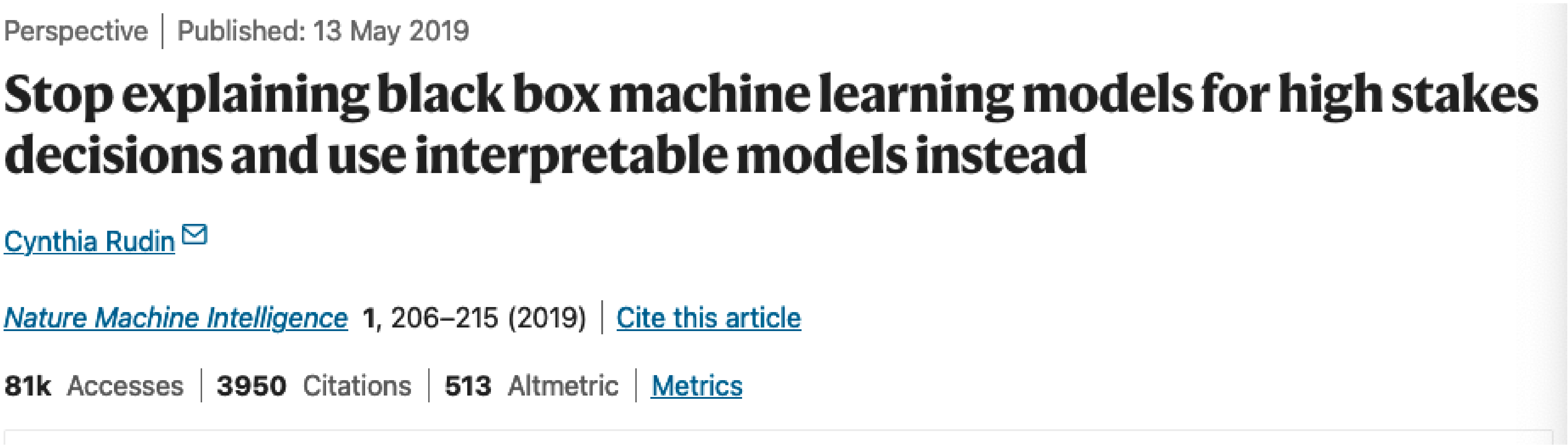
Interpretability in Deep Learning for Genomics

1/7

Deep Learning Reading Group

Gus Fridell



Perspective | Published: 13 May 2019

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

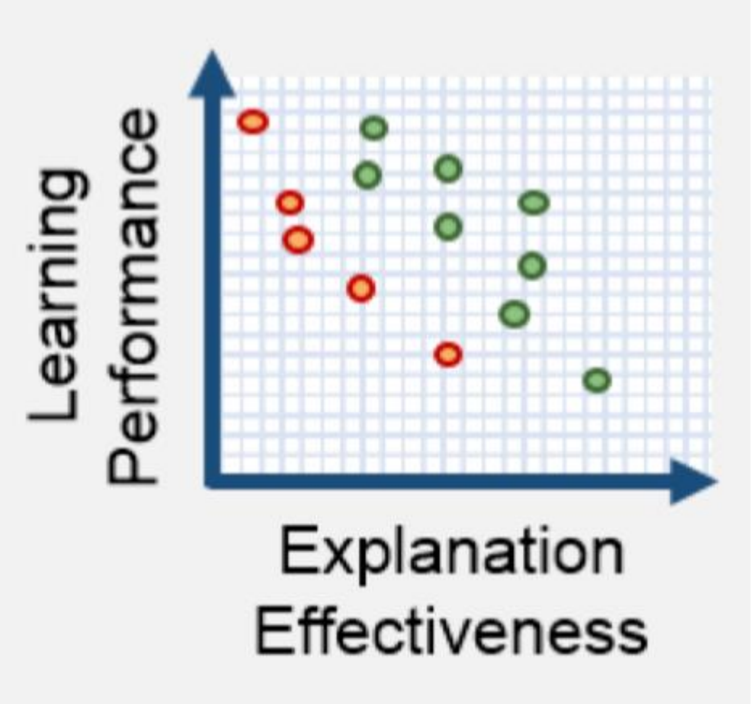
Nature Machine Intelligence **1**, 206–215 (2019) | Cite this article

81k Accesses | **3950** Citations | **513** Altmetric | Metrics

https://www.youtube.com/watch?v=n_mwYWfI_sI&ab_channel=TheTWIMLAIPodcastwithSamCharrington

The interpretability-accuracy trade off is a myth

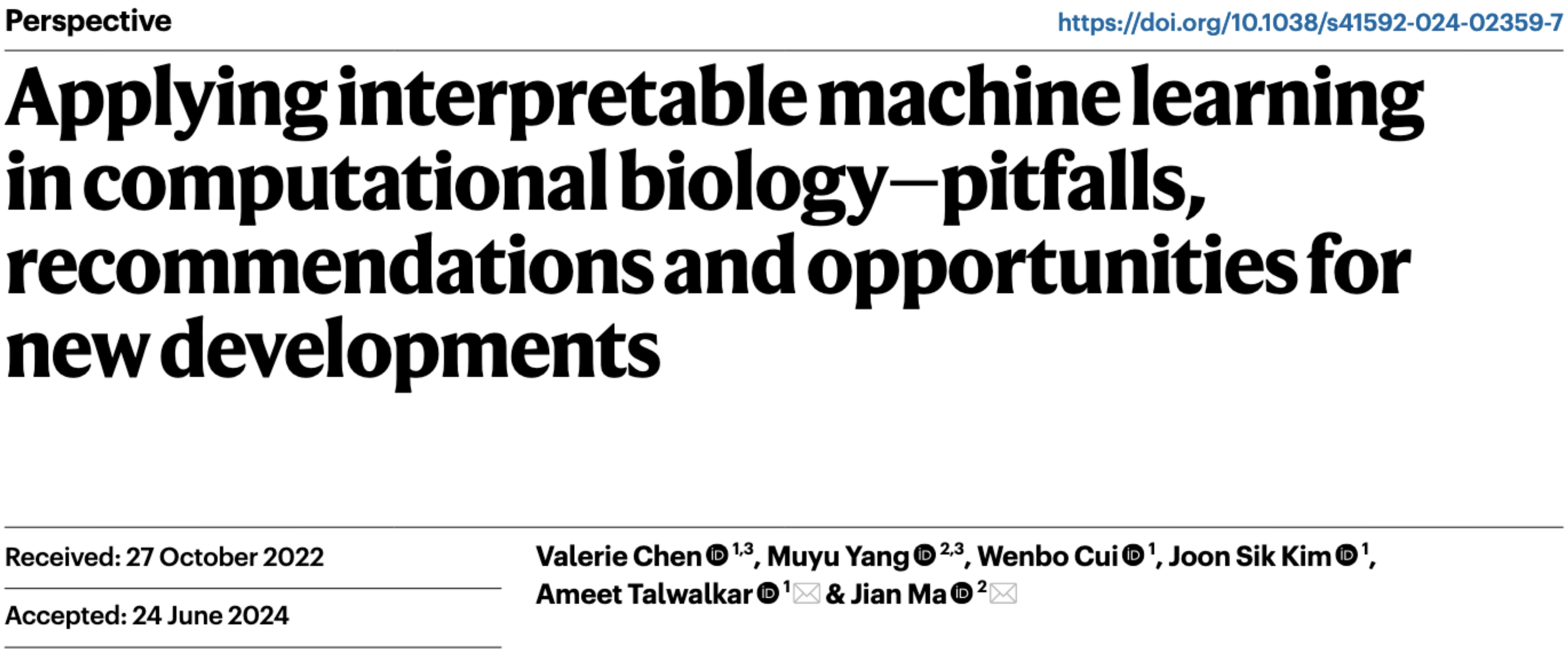
- Publication Bias
  - Papers are often rejected based on performance gains
- BBM is much easier to train and apply “out of the box”
- There is no “secret ability” in BBMs to find “hidden patterns”
  - If a pattern in the data was important enough that a BBM could leverage it, then an interpretable model should also be able to locate it
- Because no single model *dominates* in terms of accuracy, we’re still operating within a Rashomon set of models, and an interpretable model is likely to be included in that set



DARPA XAI BAA

The real trade-off in IML

- Challenges:
  - Domain knowledge must be used to inform architecture/model choice to be interpretable, while being flexible enough to fit the data accurately
    - Applying a model is simply not enough, you must have some expertise in what your model tries to do to make an interpretable model
  - Solving a constrained problem is harder than solving an unconstrained problem
- Advantages and Accuracy
  - Small differences in accuracy are overwhelmed by the ability to interpret results and improve data processing
    - Interpretable models can be *more* accurate if they allow us to correct false assumptions, reveal errors in data processing, and better iterate on feature engineering
  - Better extendibility to unseen datasets (the test set is not the same as an unseen set!)
  - For high stakes, interpretable decisions are more valuable



Perspective

https://doi.org/10.1038/s41592-024-02359-7

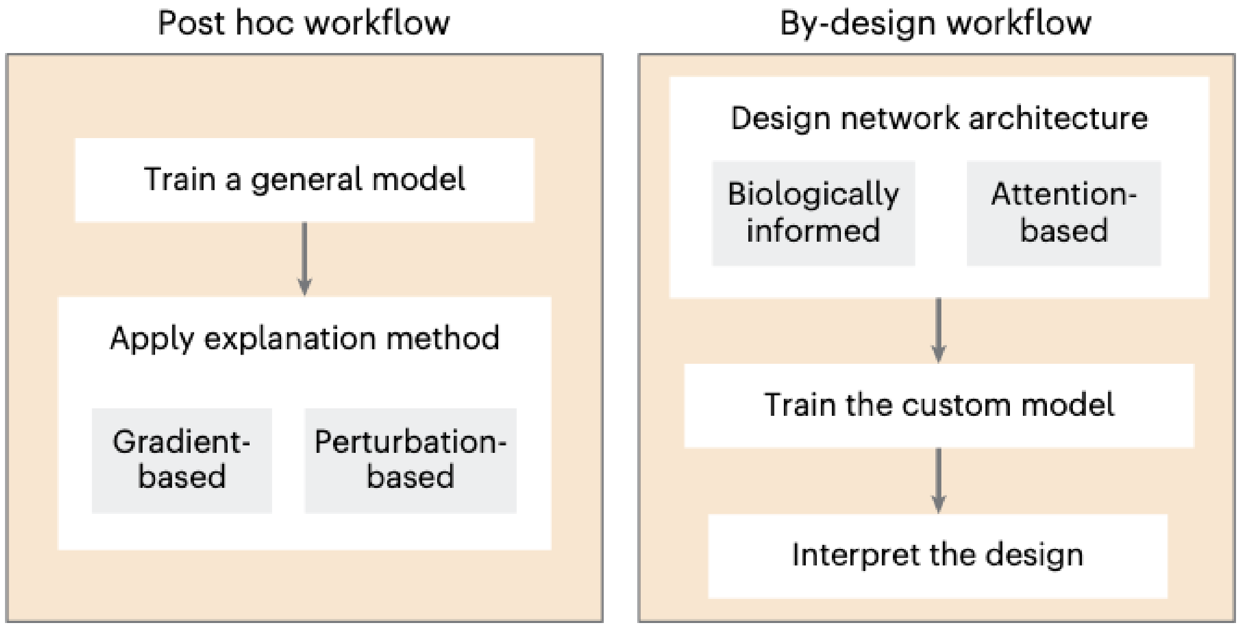
Applying interpretable machine learning in computational biology—pitfalls, recommendations and opportunities for new developments

Received: 27 October 2022

Accepted: 24 June 2024

Valerie Chen [1,3], Muyu Yang [2,3], Wenbo Cui [1], Joon Sik Kim [1], Ameet Talwalkar [1] ✉ & Jian Ma [2] ✉

Two main approaches



Post hoc workflow

Train a general model

Apply explanation method

Gradient-based

Perturbation-based

By-design workflow

Design network architecture

Biologically informed

Attention-based

Train the custom model

Interpret the design

‘Feature importance’ methods

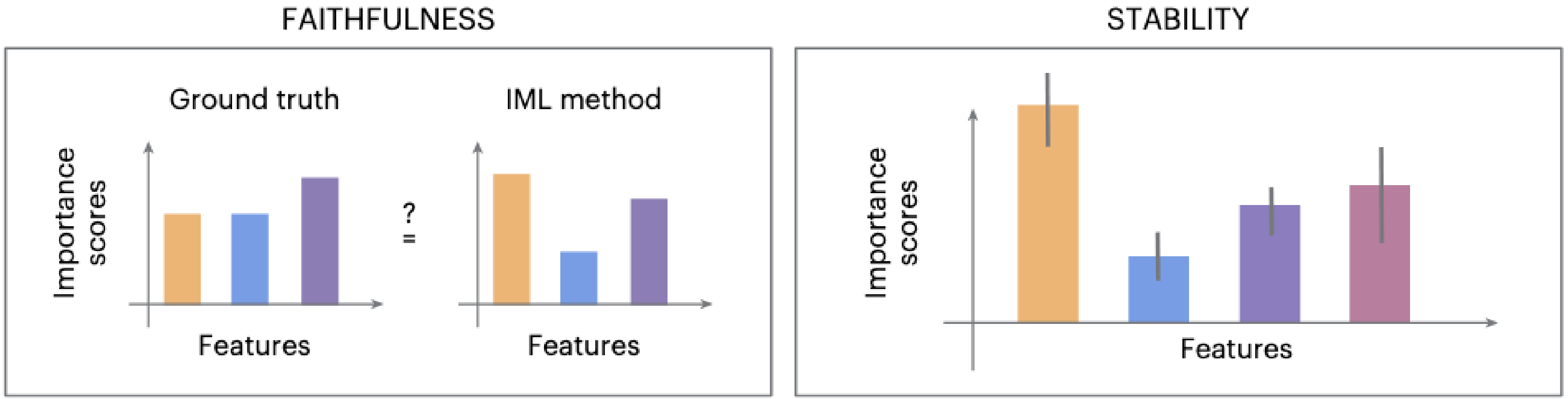
- Gradient-based or Perturbation-based
- Shows prediction-level behavior of what the model used, but doesn’t capture how or why
- Limited debugging capability
- May not align with human-understandable concepts (up to us to decide what actual features or patterns are being detected)

‘Naturally interpretable’ methods

- Biologically informed
  - Hidden nodes in the NN correspond to biological entities (genes, or pathways)
- Attention: learned weights are often considered as an explanation*

How do we evaluate IML? (Post-hoc workflows)

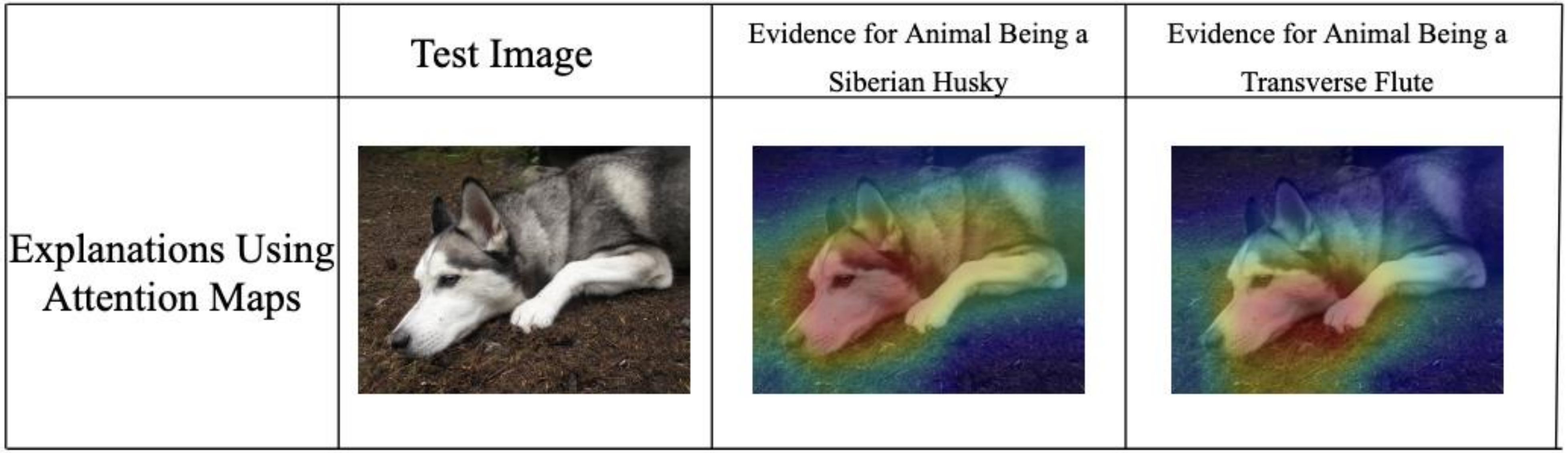
- Two common techniques: Fidelity and Stability



- Fidelity (faithfulness) is the most common metric which captures the degree to which an explanation reflects the ground truth (is what your IML method identifies as important actually being used by the model?)
- Stability measures the consistency of explanations across similar inputs

Fidelity (Post-hoc workflows)

- Take a trained sequence to expression model → feed it to a feature attribution method like DeepLIFT to find out how much each input feature (like a nucleotide) contributed to the model's prediction
  - DeepLIFT compares the activation of each neuron to a "reference" activation and backpropagates the differences to determine input importance
- Challenges:
  - Ground truth (GT) is hard to know and difficult to simulate (esp. for genomics) so the authors suggest using real data where GT mechanisms are known
  - "There is no method which generally outperforms other methods across the board, pointing to general unreliability of existing methods"
- Problem (!!)
  - If our underlying model is a black box model, then the only way to achieve high fidelity in an explainable method is for it to be as complex as the model itself
  - An explanation method has an upper limit on fidelity based on the input model complexity

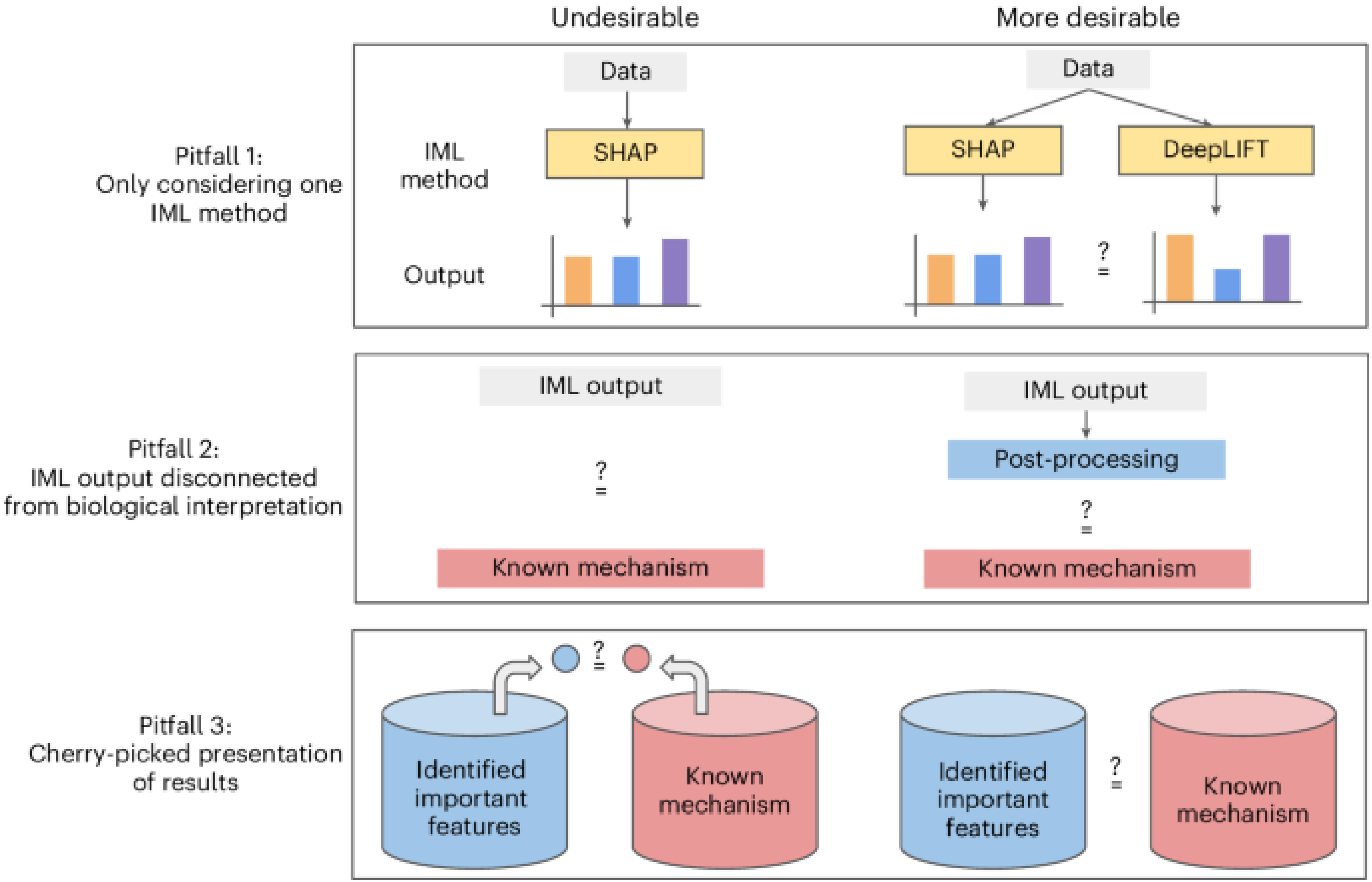


	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps			

Rudin, 2019

Stability

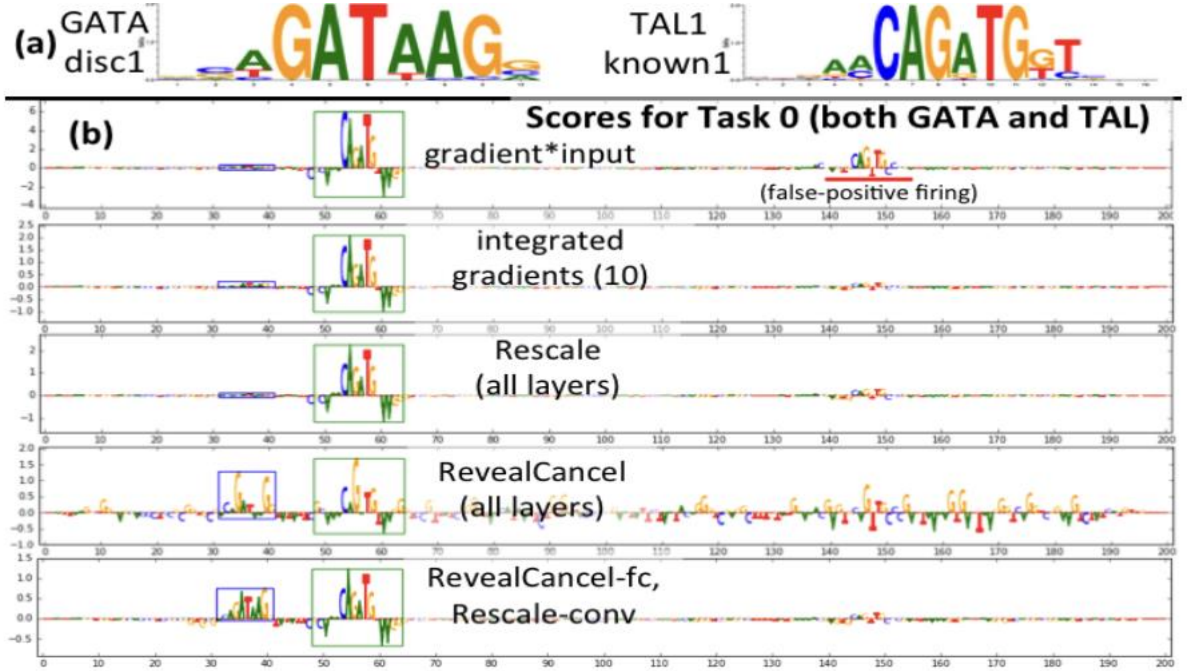
- Feature importance often varies substantially when small perturbations are applied to an input
- Popular methods like SHAP and LIME have been shown to cause unstableness
    - SHAP is a game theoretic model that ensures each feature's contribution is calculated fairly. Contributions of all features add up to the model's output
    - LIME perturbs inputs and observes how prediction's change. It fits an interpretable model (linear regression, decision tree) to the model's response to perturbations
- No single method that is most stable across multiple real-world datasets
- The author's suggest that stability should be evaluated alongside other metrics and domain knowledge (how permissive should the model be to perturbations? E.g. SNPs can either deleterious or harmless!)


Undesirable
More desirable
Pitfall 1:
Only considering one IML method
Data
IML method
SHAP
Output
Data
SHAP
DeepLIFT
?
=
Pitfall 2:
IML output disconnected from biological interpretation
IML output
?
=
Known mechanism
IML output
Post-processing
?
=
Known mechanism
Pitfall 3:
Cherry-picked presentation of results
?
=
Identified important features
Known mechanism
Identified important features
?
=
Known mechanism

Pitfall 1: Only considering one IML method

- Relying on a single run of one IML method may result in biased feature importance
  - e.g. Not all IML methods identify key motifs for TF binding
  - Different hyperparameters or multiple runs of the same IML method lead to different outputs

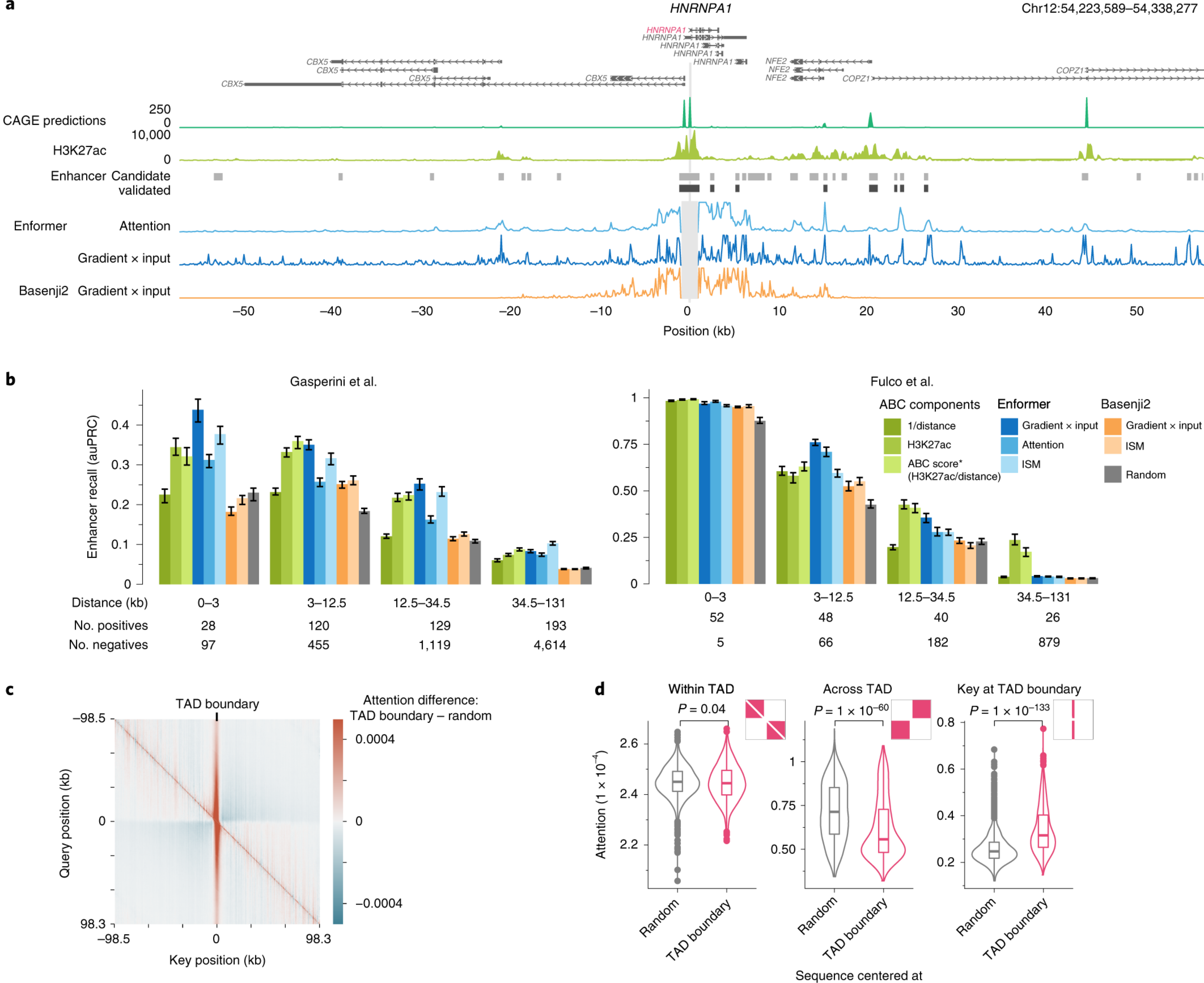
We should use a diversity of methods and hyperparameters



Shrikumar, 2017

Enformer:

- Post-hoc: Gradient x input, ISM
- By-design: Attention maps



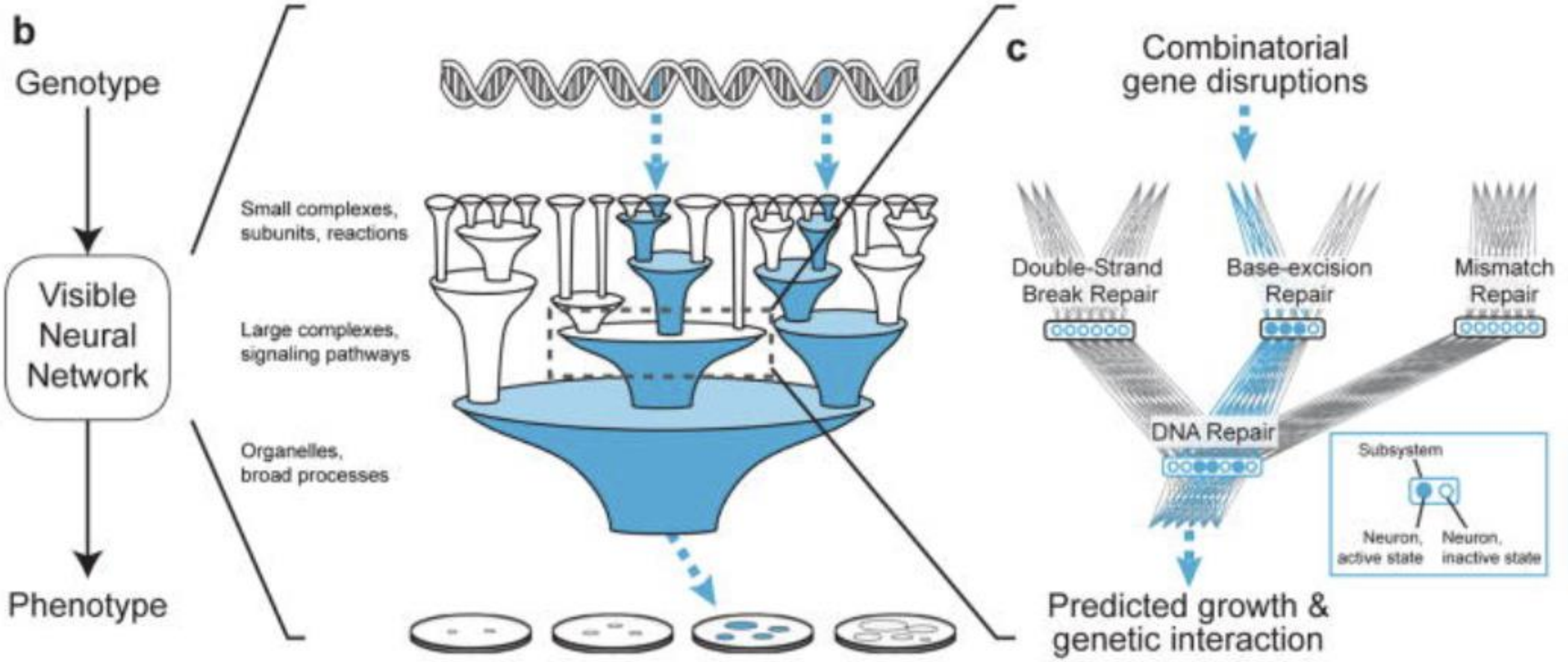
Avsec, Ž. 2021

If methods disagree:

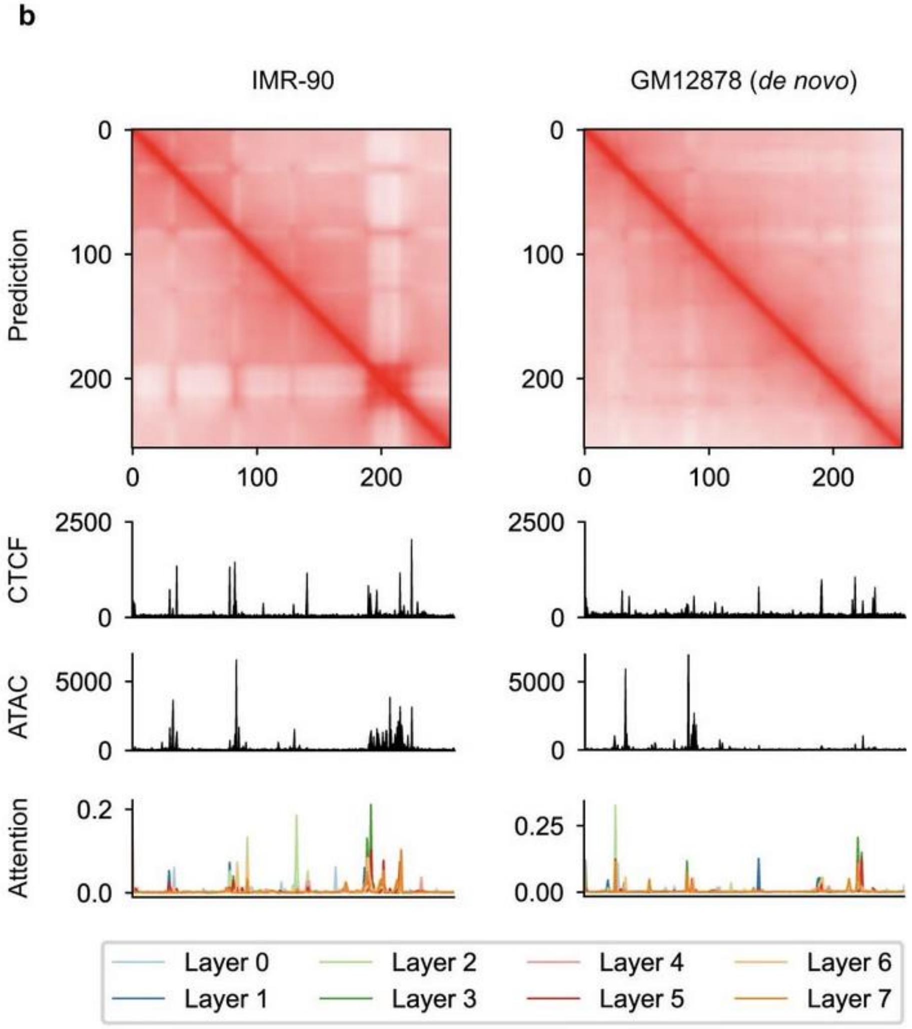
- Need some evaluation mechanism to test faithfulness
    - It could be that the features being detected by IML methods are spurious, or the model is not actually using this information how you would expect it to
- Can be applied to problems where there is prior knowledge or GT mechanisms
- In the absence of GT, experimental validation should be done to verify predictions
    - ‘human-in-the-loop’/‘lab-in-the-loop’

Pitfall 2: IML output disconnected from biological interpretation

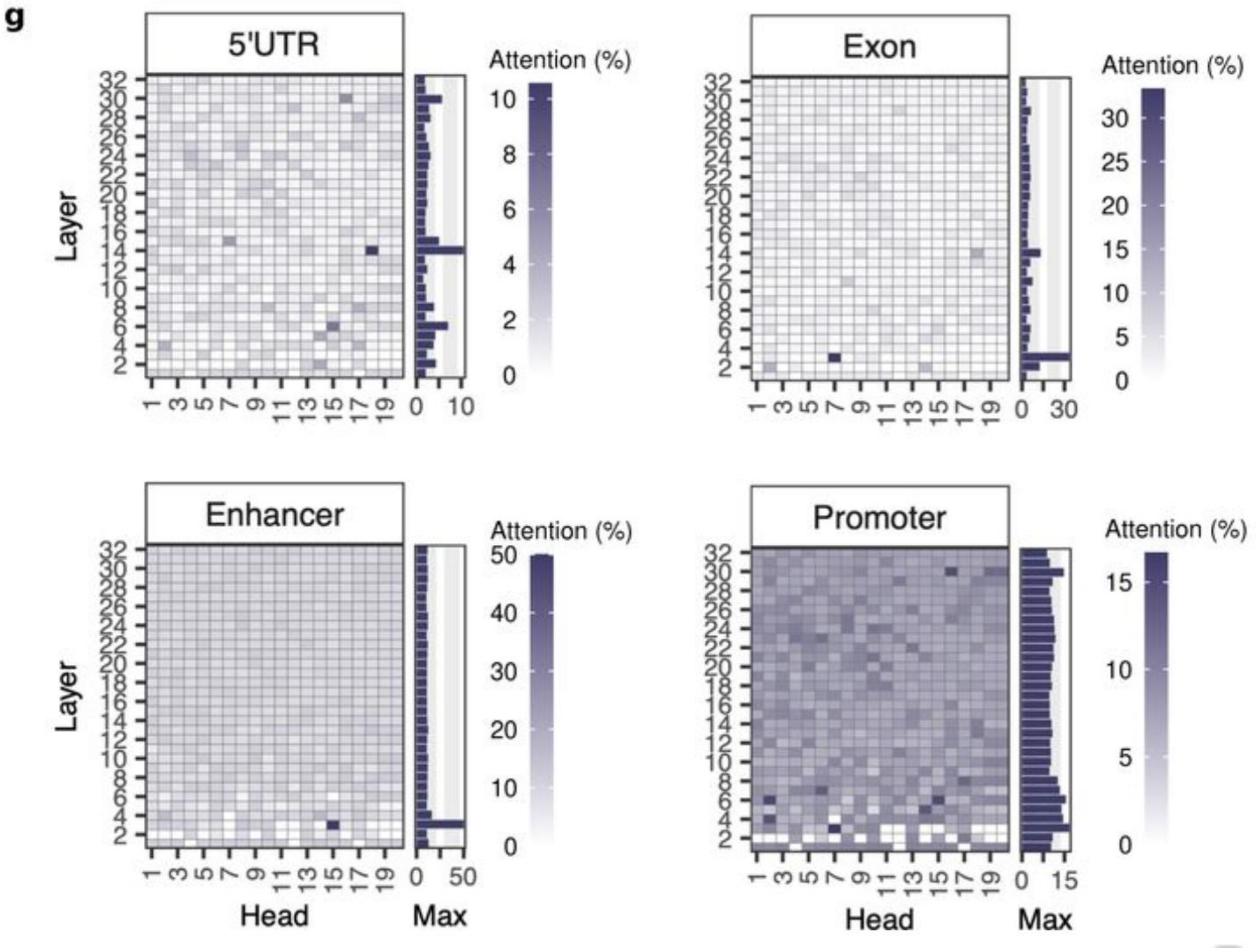
- IML methods can identify predictive features, but don’t automatically provide biological meaning
    - DNA sequences: gradient-based methods give nucleotide-level importance, but need post-processing to reveal meaningful patterns (e.g. TF-MoDISco for motif discovery)
- Post-hoc methods require domain-specific post-processing techniques
- Interpretable-by-design networks are easier to interpret, but some decisions must be made
    - Attention flow and attention rollout can also be used to process attention matrices


b
Genotype
Visible Neural Network
Phenotype
Small complexes, subunits, reactions
Large complexes, signaling pathways
Organelles, broad processes
c
Combinatorial gene disruptions
Double-Strand Break Repair
Base-excision Repair
Mismatch Repair
DNA Repair
Subsystem
Neuron, active state
Neuron, inactive state
Predicted growth & genetic interaction

Dcell (Ma, J. 2018) embeds GO hierarchical structures into network architecture and further identifies genotype-phenotype relationships by identifying Boolean logic gates of cellular subsystems



Tan, J. 2023



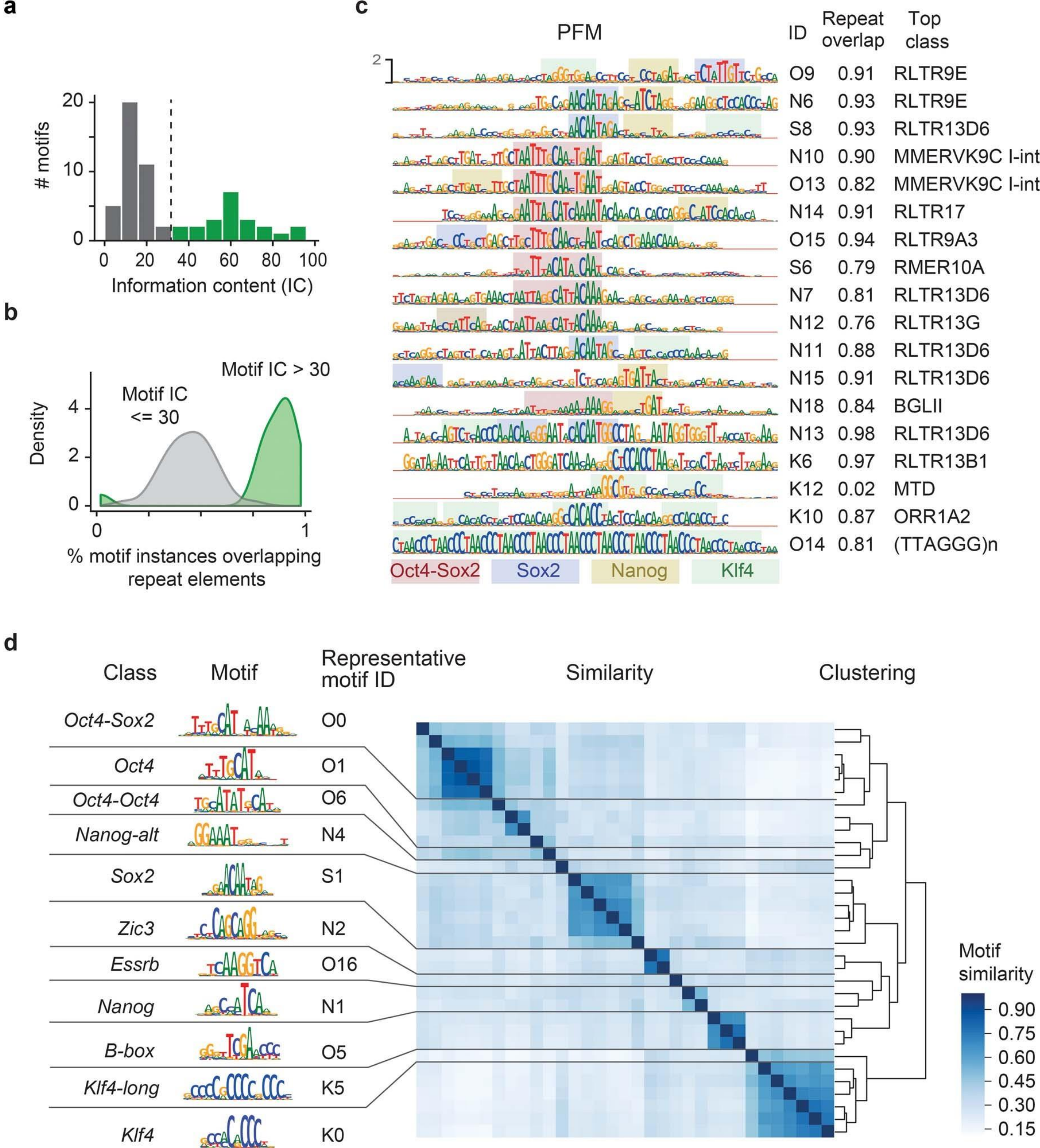
Dalla-Torre, H. 2023

Attention interpretations can be averaged across layers (Enformer), within-layer (C.Origami), or individually (Nucleotide Transformer using “BERTology” method)

Pitfall 3: Cherry picking

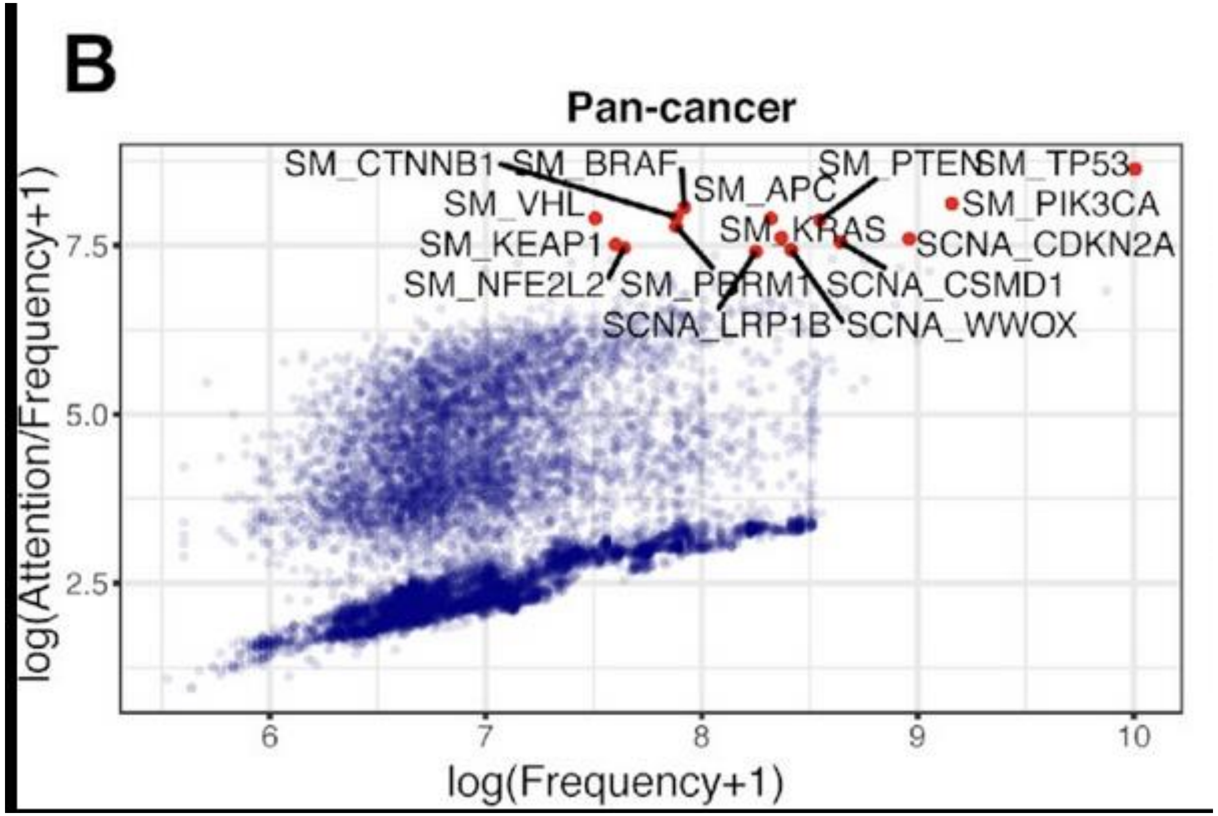
- Often a result of publication bias (highlighting your ‘best’ result)
- E.g. presenting in DNA/RNA/protein sequence tasks results that only showcase local regions where importance scores are consistent with existing annotations
- Recommend conducting quantitative analysis of faithfulness of the importance score with prior knowledge across the ***entire*** dataset
    - “It is crucial not to overlook the nontrivial feature importance attributions that may appear inconsistent with prior knowledge”
- Recommend including a measure of stability
    - Vary hyperparameters, perturb inputs, compare methods
    - Verify whether most important features identified in one dataset are consistent across independent datasets

BPNet found 51 sequences from gradient-based feature importance, then justified decisions to include them into representative set, and why they were returned as important features even if they were not validated

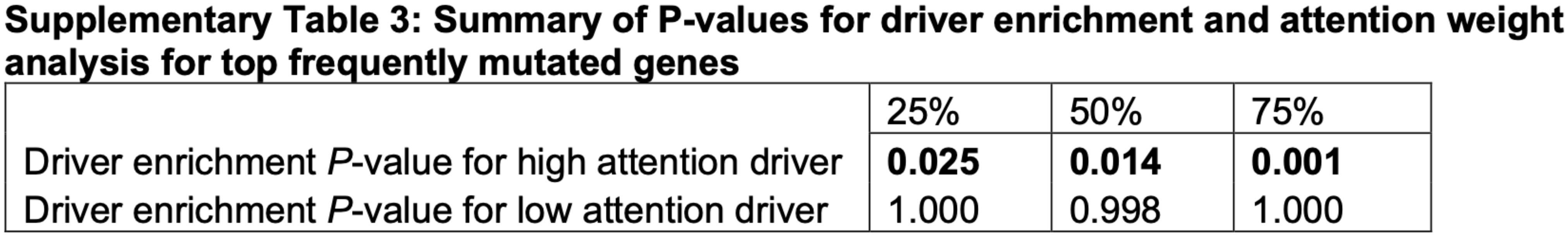


Avsec, Ž. 2019

CITRUS, which models the impact of somatic alterations on transcription, showed a biologically meaningful split between high and low attention weights for given genes



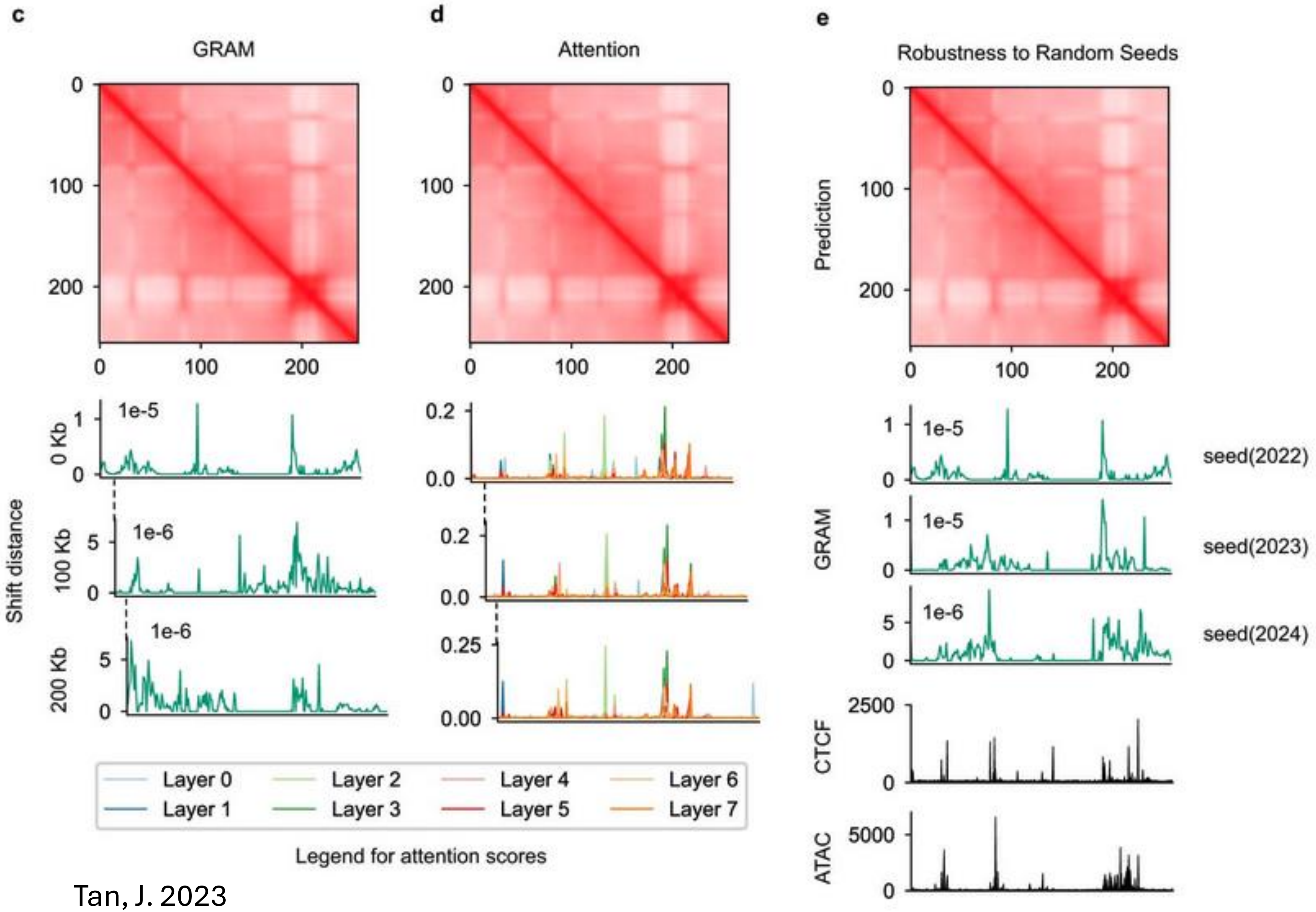
Tao, Y. 2022



Supplementary Table 3: Summary of P-values for driver enrichment and attention weight analysis for top frequently mutated genes

	25%	50%	75%
Driver enrichment *P*-value for high attention driver	**0.025**	**0.014**	**0.001**
Driver enrichment *P*-value for low attention driver	1.000	0.998	1.000

Stability: C.Origami tests multiple different methods, under different parameters (seeds) and under different perturbations



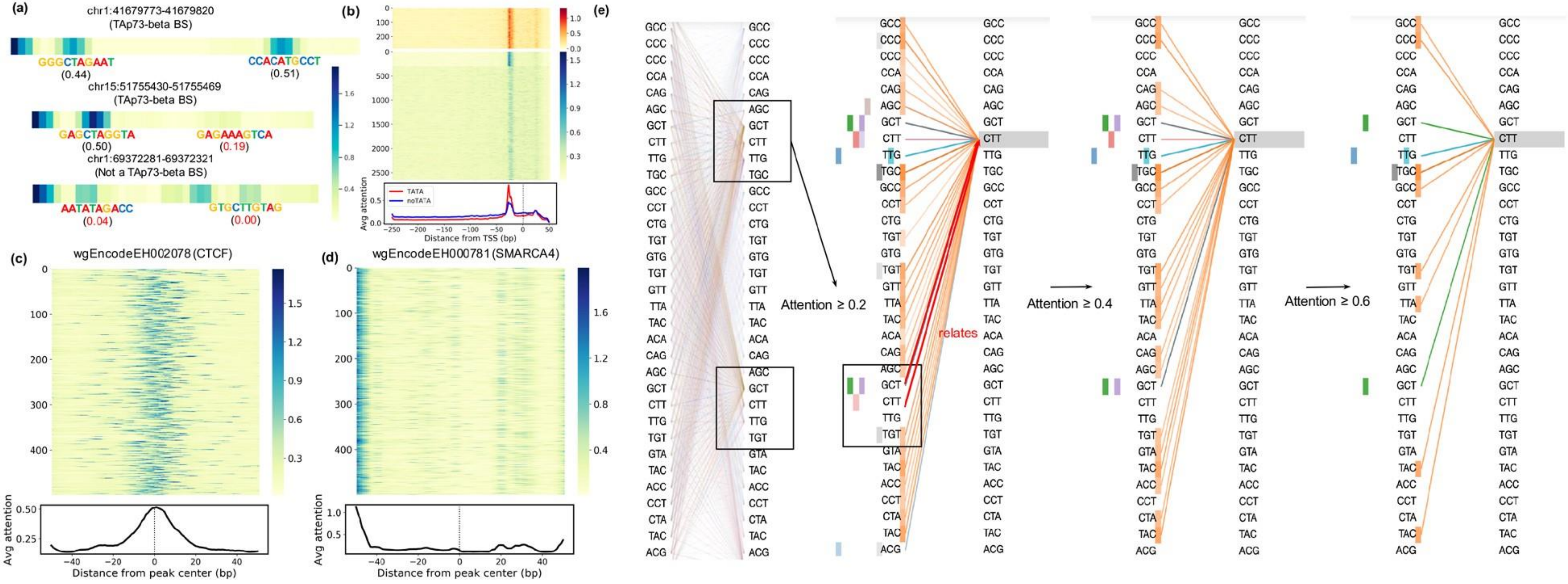
Tan, J. 2023

Opportunities for IML dev. in the age of LLMs

- Besides establishing better practices, IML development has lagged behind rapid advancement of LLMs in predictive biological modeling
- SOA models still utilize classic IML methods
- These methods are often applied in biased and incomplete ways, and their validity/reliability is a topic of open debate

- **What is the right choice of tokenization for biological applications?**
  - Commonly used: single-nucleotide tokenization, fixed k-mer tokenization, byte-pair-encoding tokenization
  - “There remain opportunities to develop…[schemes to] better represent underlying biology…by incorporating prior knowledge.”
- **How can LLM-specific IML methods be adapted to biological context?**
  - Mechanistic interpretability techniques aim to translate complex transformer models into human-understandable algorithms (circuits, human-readable programs)
  - Prompting LLMs: future LLMs that are pretrained on natural language and biological corpora to explain LLMs and generate new hypotheses

- **How do we develop IML techniques to handle multimodal applications?**
  - The paper highlights several multimodal, multi-omic models, but their IML methods are limited to unimodal explanations
  - Multimodal, biological data are very correlated in function (sequence and epigenomic features) which is challenging when assigning accurate importance scores and drawing meaningful conclusions
  - We need to define evaluation techniques to check whether explanations properly attribute importance scores to each modality
- **What types of novel visualization tools can best facilitate interpretation?**
  - Methods should be tailored to data types and applications
  - DNABERT-Viz allows users to explore important genomic regions and sequence motifs, but a suite of tools across different data types is necessary to enable standardized IML workflows


(a)
chr1:41679773-41679820
(TAp73-beta BS)
GGGCTAGAAT
(0.44)
CCACATGCCT
(0.51)
chr15:51755430-51755469
(TAp73-beta BS)
GAGCTAGGTA
(0.50)
GAGAAAGTCA
(0.19)
chr1:69372281-69372321
(Not a TAp73-beta BS)
AATATAGACC
(0.04)
GTGCTTGTAG
(0.00)
1.6
1.2
0.8
0.4
(b)
1.0
0.5
0.0
1.5
1.2
0.9
0.6
0.3
TATA
noTATA
Avg attention
Distance from TSS (bp)
(c)
wgEncodeEH002078 (CTCF)
Avg attention
Distance from peak center (bp)
(d)
wgEncodeEH000781 (SMARCA4)
Avg attention
Distance from peak center (bp)
(e)
GCC CCC CCC CCA CAG AGC GCT CTT TTG TGC GCC CCT CTG TGT GTG TGT GTT TTA TAC ACA CAG AGC GCT CTT TTG TGT GTA TAC ACC CCT CTA TAC ACG
Attention ≥ 0.2
relates
Attention ≥ 0.4
Attention ≥ 0.6

Ji, Y. 2021

Summary and Conclusion

- 2 Main classes of IML: by-design and post-hoc
  - By-design methods aim to be interpretable by design of model architecture/algorithm
  - Post-hoc methods must be tested for fidelity (model's agreement with ground truth) and stability (consistency of predictions) when applied
- Authors prefer by-design methods, and stress that you mult use a diverse set of methods and parameters when applying post-hoc methods
- Human-in-the-loop/lab-in-the-loop approaches are crucial for validation when methods disagree or ground truth is inaccessible
- There is much opportunity for developing novel, domain-specific IML methods that integrate prior biological knowledge into architectures and novel hypothesis generation