Critiquing Protein Family Classification Models Using Sufficient Input Subsets

Brandon Carter, Maxwell Bileschi, Jamie Smith, Theo Sanderson, Drew Bryant, David Belanger, Lucy Colwell

MIT Computer Science & Artificial Intelligence Laboratory
Google Research

bcarter@csail.mit.edu

Gifford Lab

Motivation

- Deep learning models achieve high accuracy in many scientific applications, but are often **difficult to interpret**, hindering their adoption

Motivation

- Deep learning models achieve high accuracy in many scientific applications, but are often **difficult to interpret**, hindering their adoption

- In order to trust them, practitioners must validate they behave according to **underlying scientific principles**

Motivation

- Deep learning models achieve high accuracy in many scientific applications, but are often **difficult to interpret**, hindering their adoption
- In order to trust them, practitioners must validate they behave according to **underlying scientific principles**
- We study the problem of **predicting protein function from amino acid sequence**

Model Interpretability

- **Sufficient input subset** (Carter et al. 2019): minimal feature subset whose values alone suffice for the model to reach the same decision

Model Interpretability

- **Sufficient input subset** (Carter et al. 2019): minimal feature subset whose values alone suffice for the model to reach the same decision
- Simple to compute and can interpret any black-box model

Model Interpretability

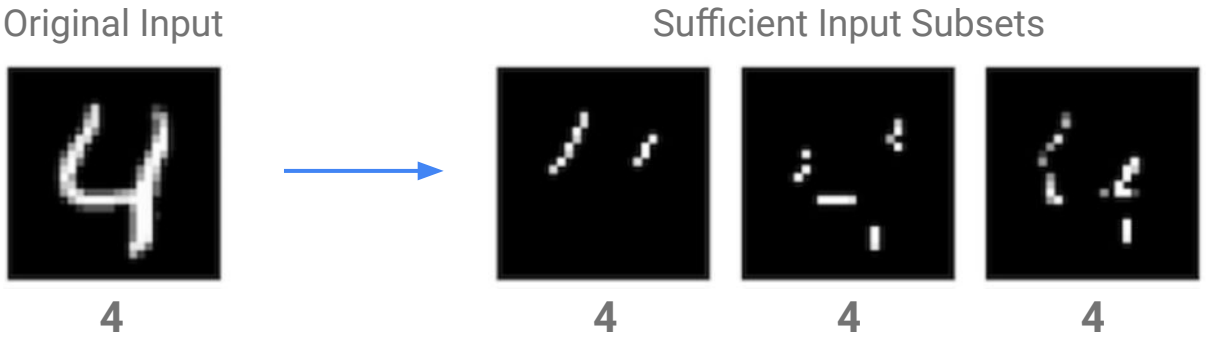
- **Sufficient input subset** (Carter et al. 2019): minimal feature subset whose values alone suffice for the model to reach the same decision
- Simple to compute and can interpret any black-box model

Original Input

4

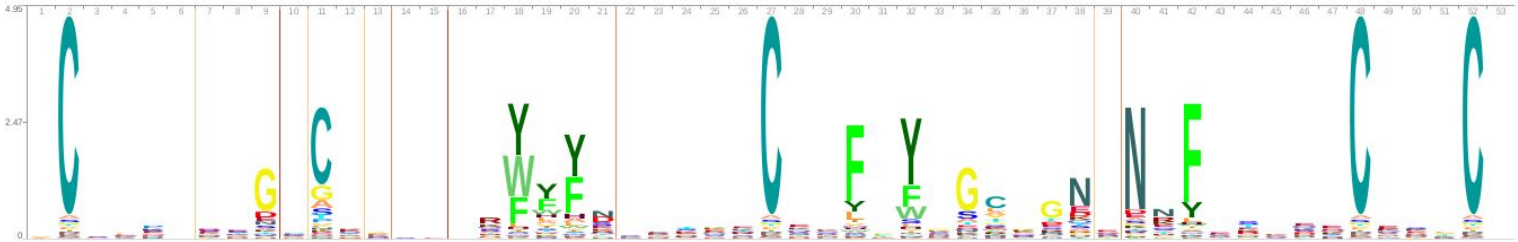
Model Interpretability

- **Sufficient input subset** (Carter et al. 2019): minimal feature subset whose values alone suffice for the model to reach the same decision
- Simple to compute and can interpret any black-box model



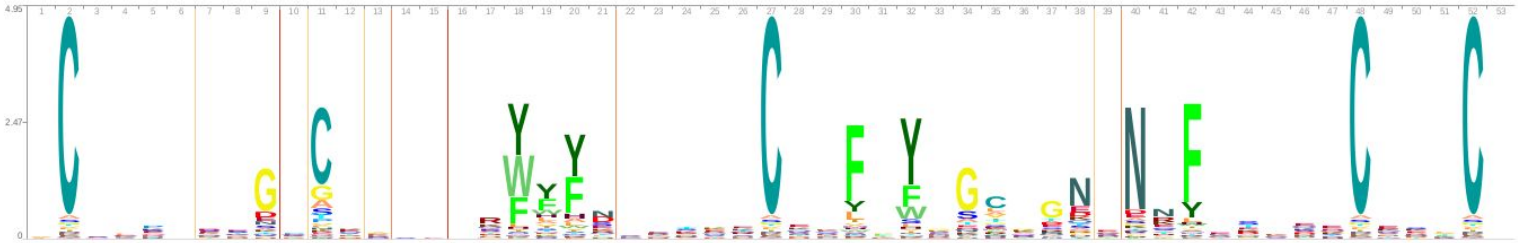
Contribution #1: SIS Logo Visualization

HMM Logo (Enzyme inhibitors from Pfam)

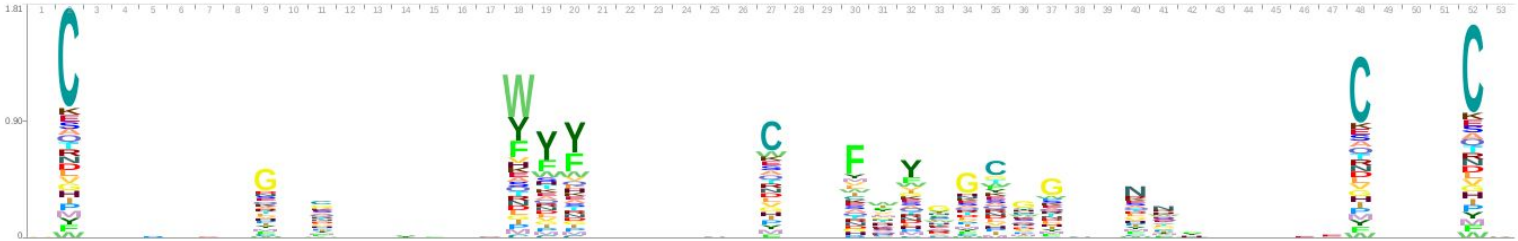


Contribution #1: SIS Logo Visualization

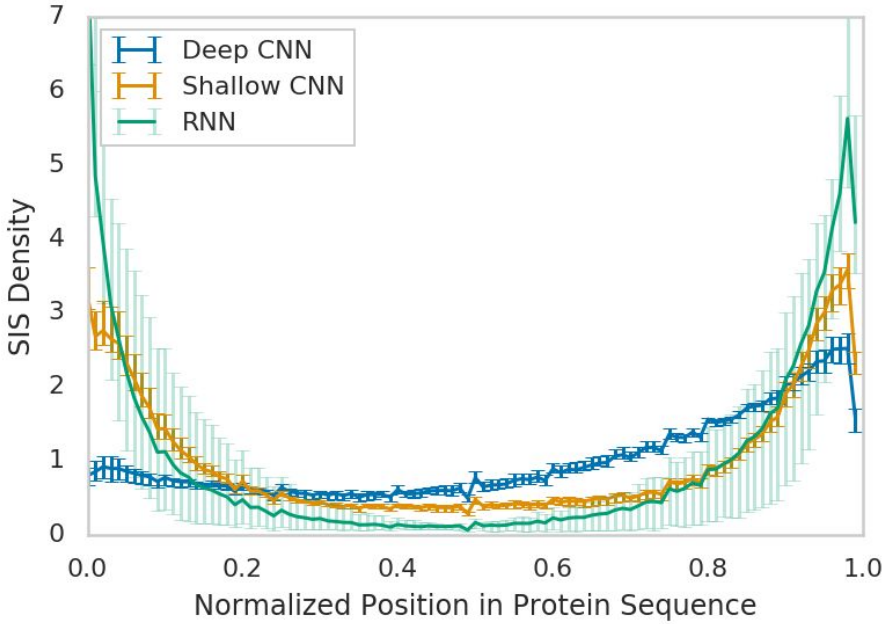
HMM Logo (Enzyme inhibitors from Pfam)



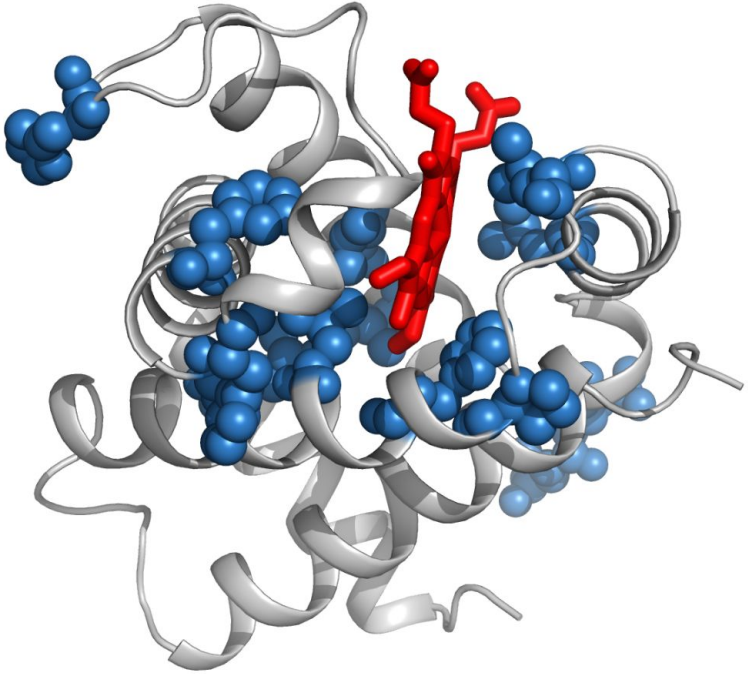
Logo of SIS Rationales from Deep CNN



Contribution #2: SIS Location



Contribution #3: 3-D Rendering of SIS



References

[1] Carter, B., Mueller, J., Jain, S., and Gifford, D. What made you do this? Understanding black-box decisions with sufficient input subsets. In *Artificial Intelligence and Statistics*, 2019.

[2] Bileschi, M. L., Belanger, D., Bryant, D. H., Sanderson, T., Carter, B., Sculley, D., DePristo, M. L., and Colwell, L. J. Using deep learning to annotate the protein universe. *bioRxiv*, pp. 626507, 2019.

SIS Code + Tutorial:
github.com/google-research/google-research/tree/master/sufficient_input_subsets