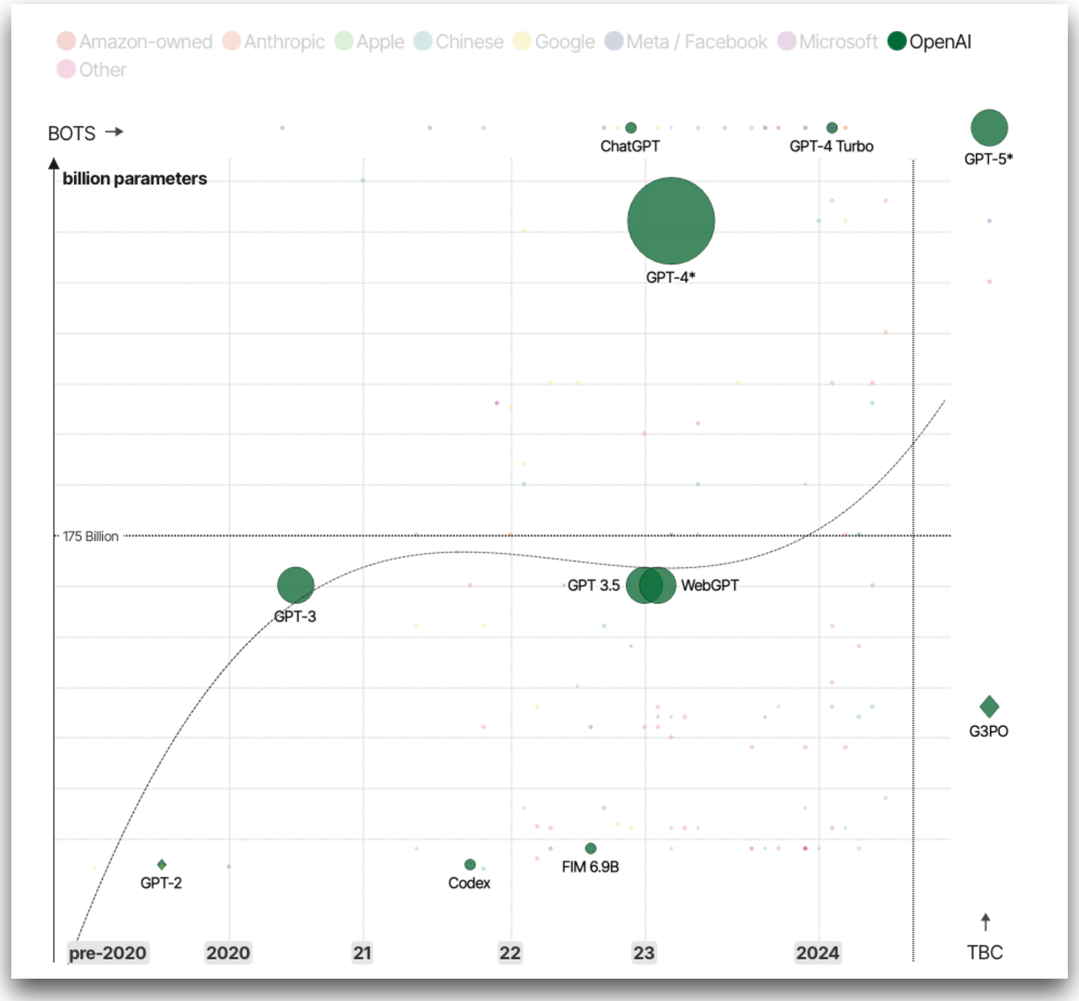
explainability and mechanistic interpretability in AI

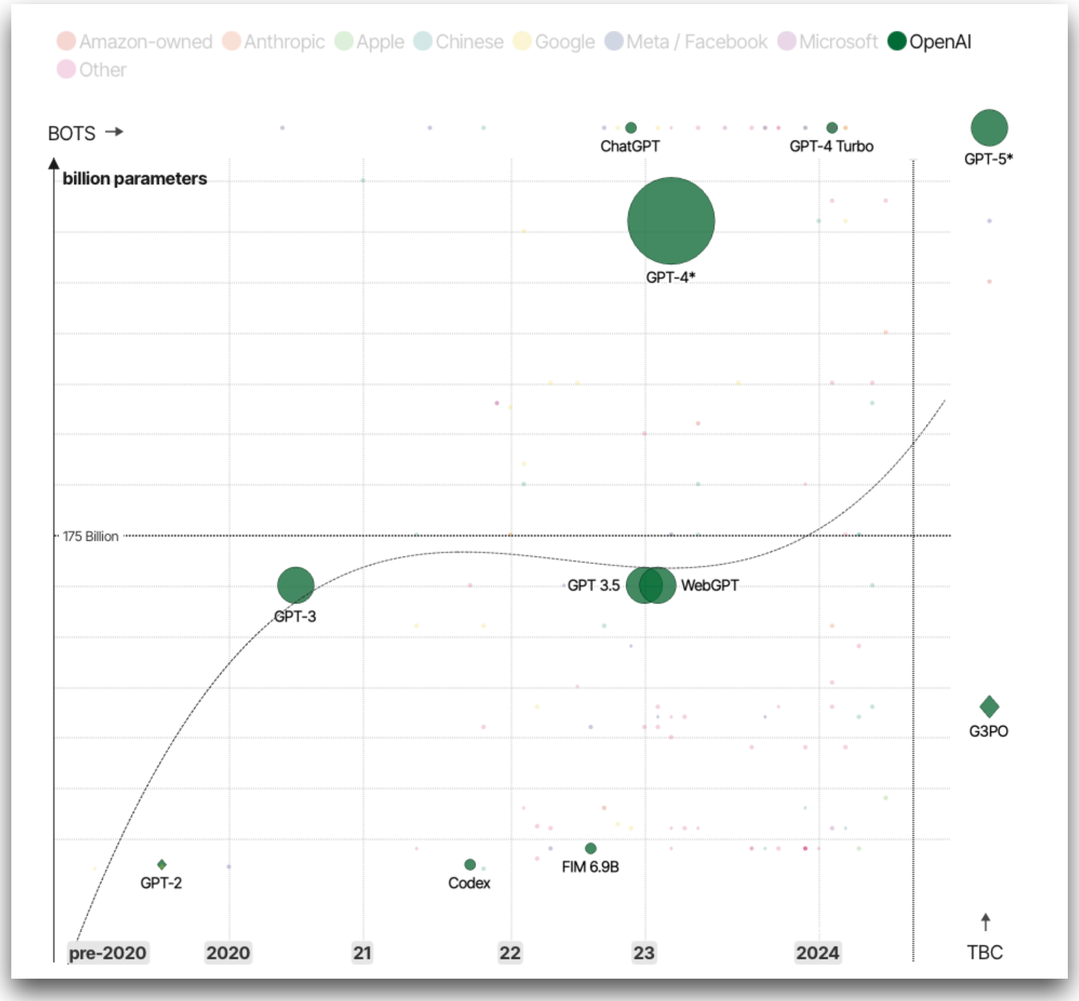
Darin Tsui

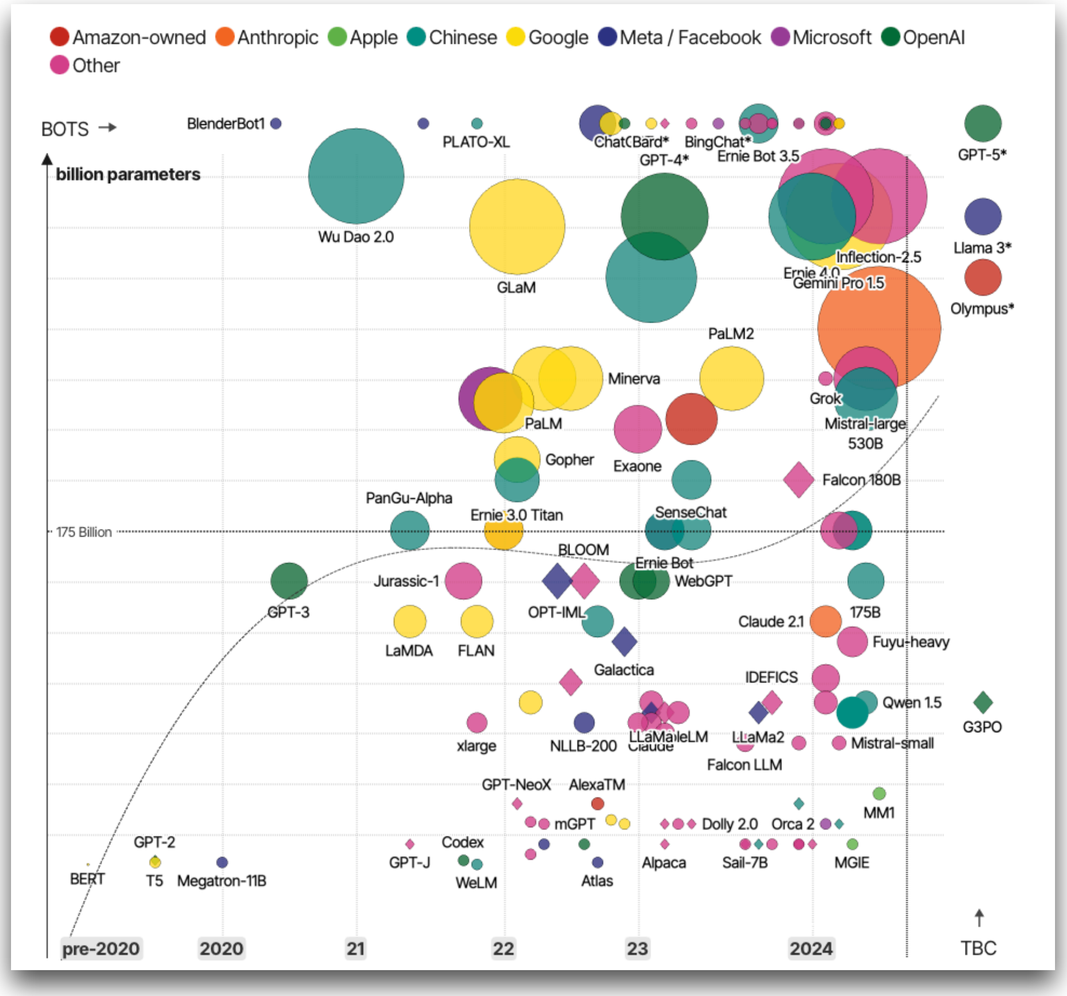
the rise and rise of AI



https://informationisbeautiful.net

the rise and rise of AI





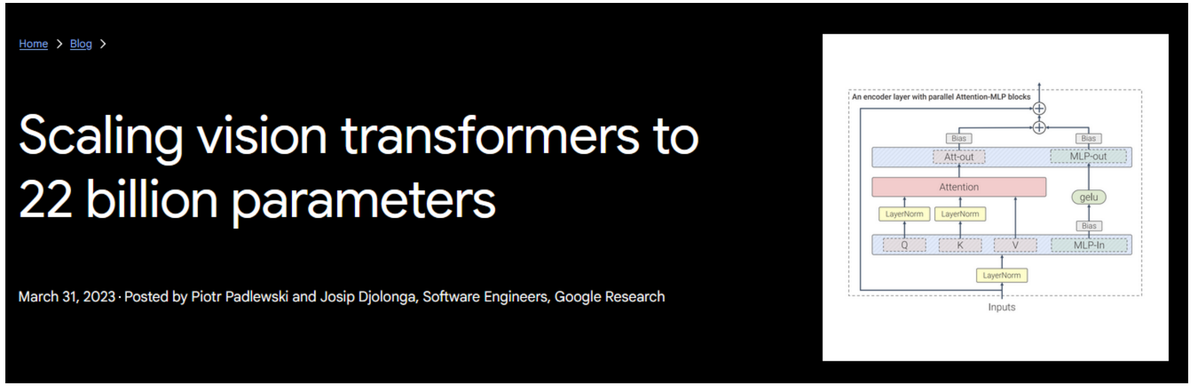
https://informationisbeautiful.net

the rise and rise of AI

120 billion parameters

the rise and rise of AI

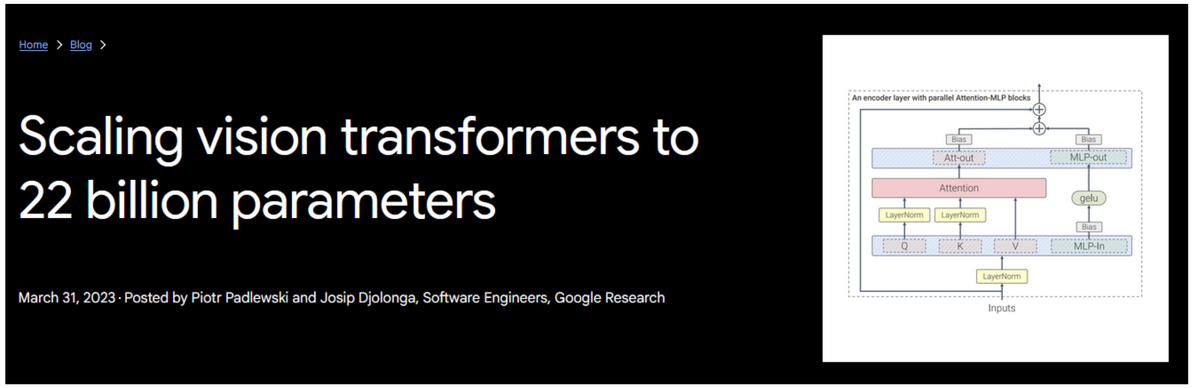
120 billion parameters



22 billion parameters

the rise and rise of AI

120 billion parameters



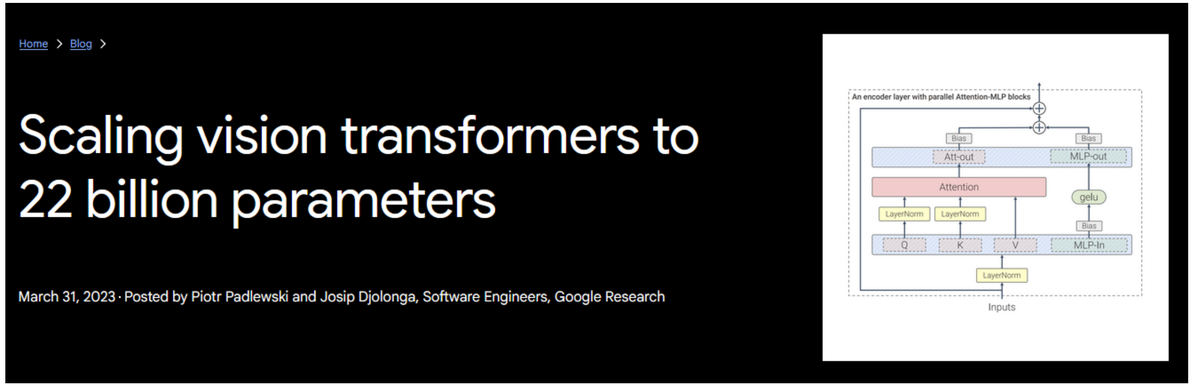
22 billion parameters

Trained on **2.78 billion** natural proteins

96 billion parameters

the rise and rise of AI

120 billion parameters



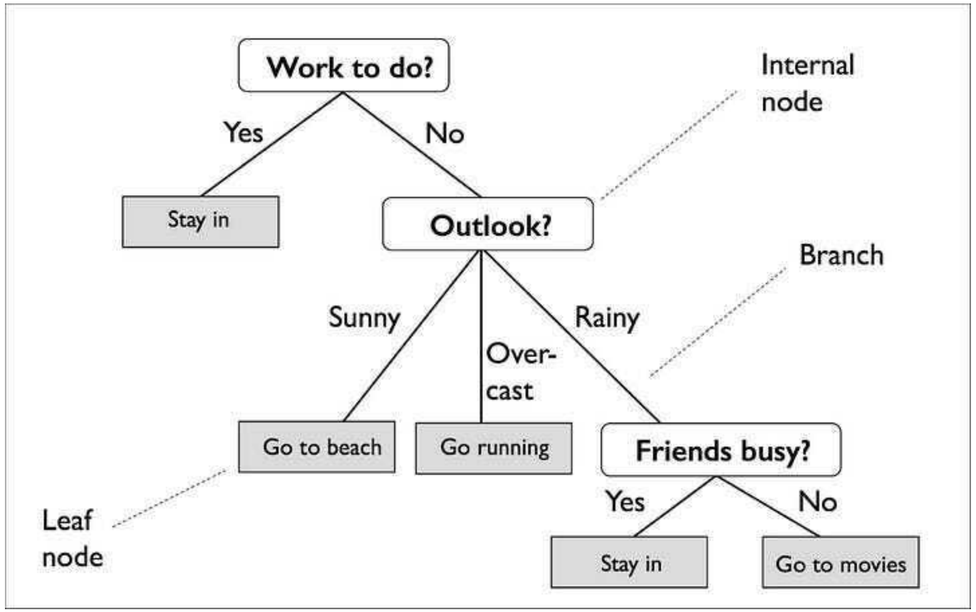
22 billion parameters

Trained on **2.78 billion** natural proteins

96 billion parameters

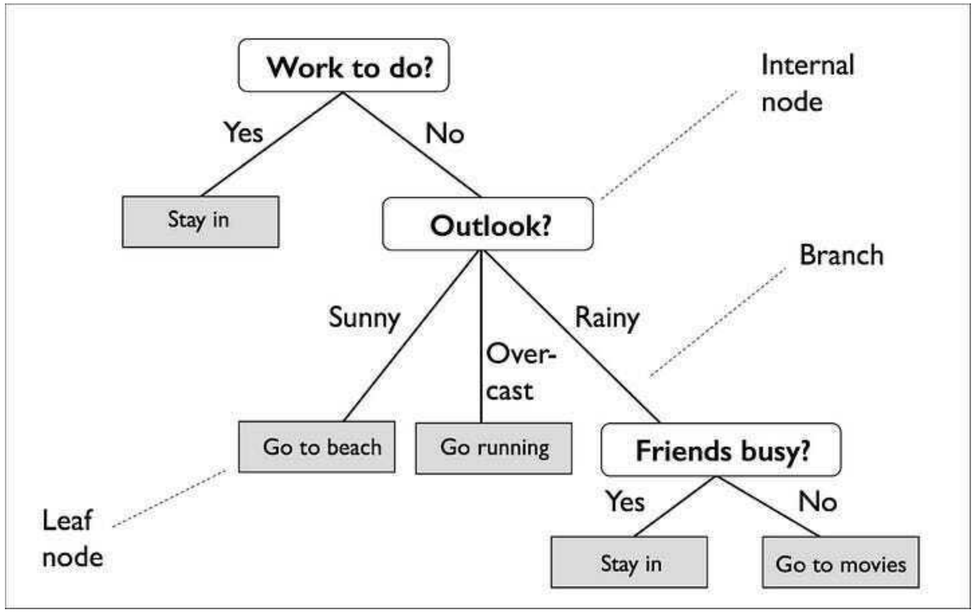
But does our power to **explain** them also **scale**?

interpretable machine learning models



dev.to, shap

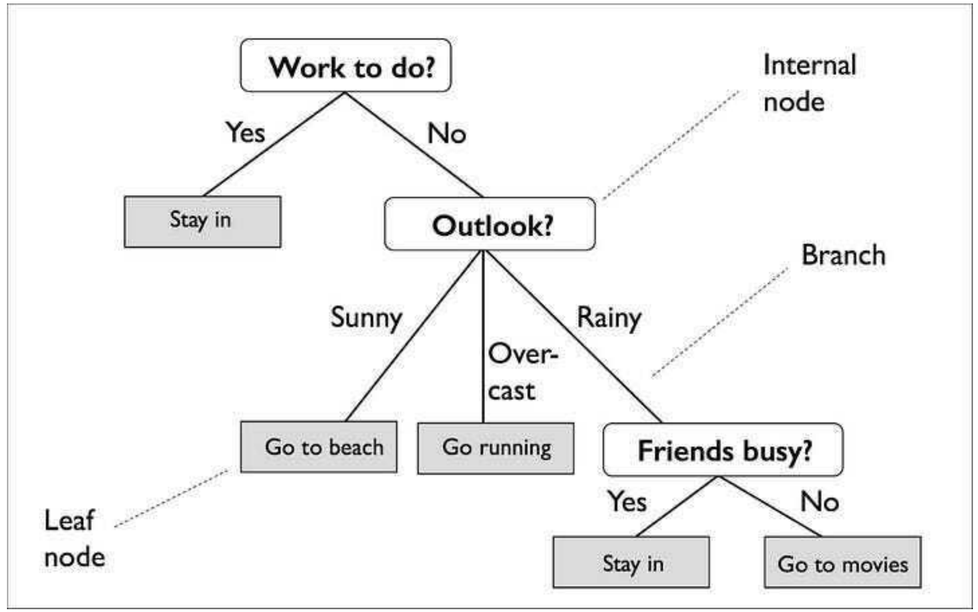
interpretable machine learning models



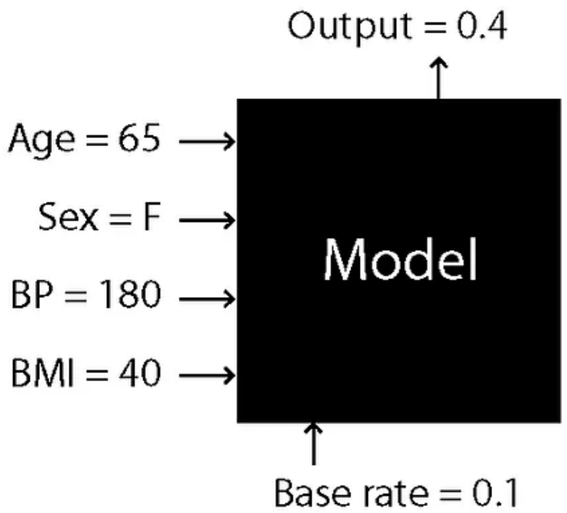
$$y = 3x_1 - x_2 + 2x_3$$

dev.to, shap

interpretable machine learning models

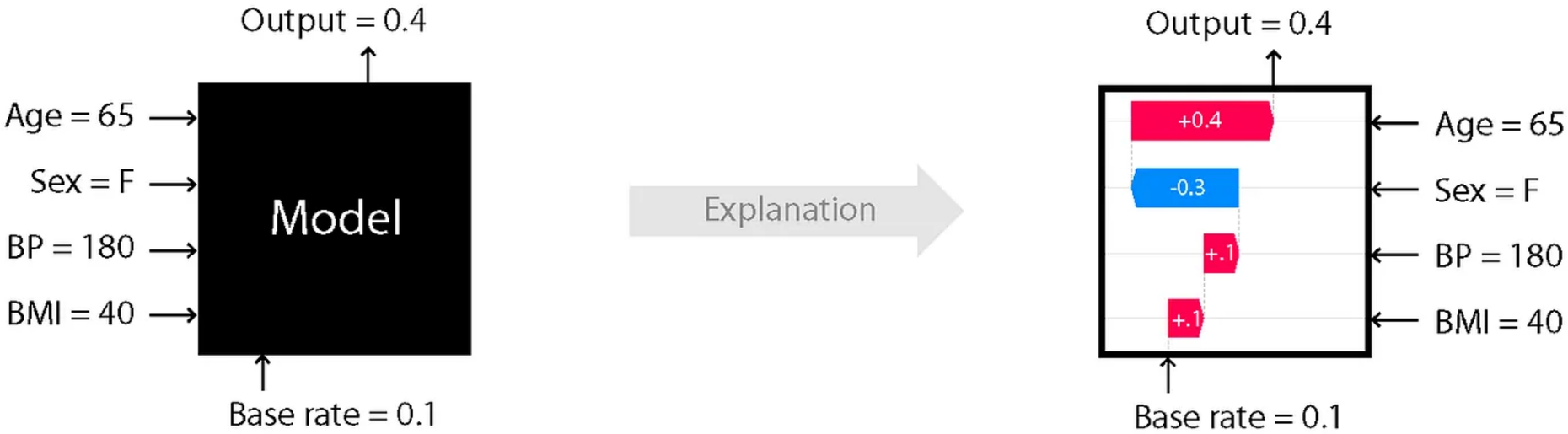


$$y = 3x_1 - x_2 + 2x_3$$



dev.to, shap

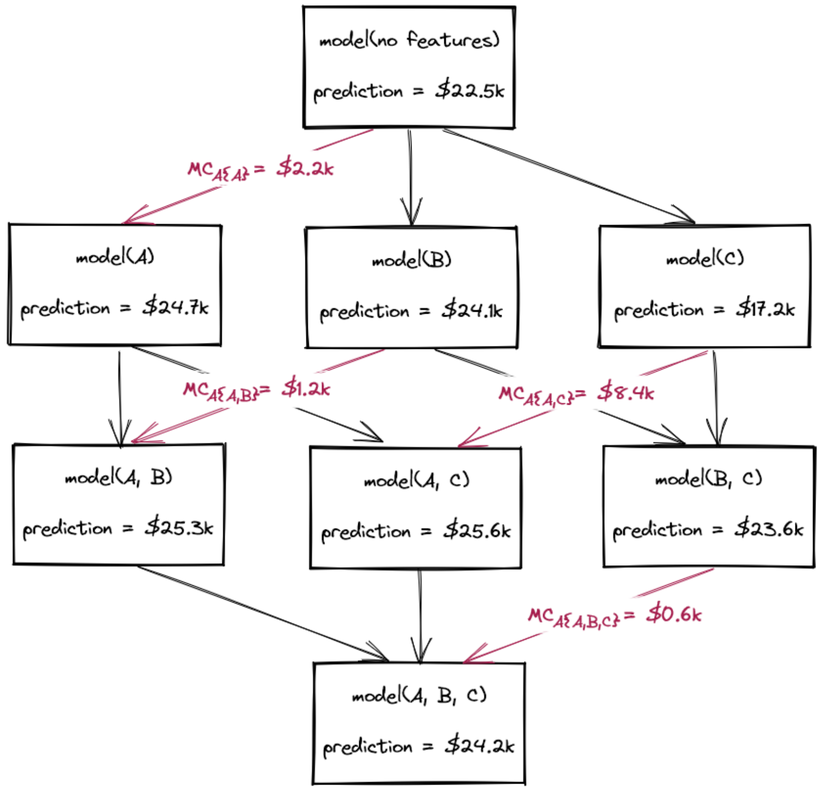
Shapley values



$$\sum_{T \subseteq D \setminus \{i\}} \frac{|T|!\,(|D|-|T|-1)!}{|D|!} \left[v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x}) \right]$$

a simple example

	A	B	C	predicted price
house1	10	6.5	0.5	$24,200

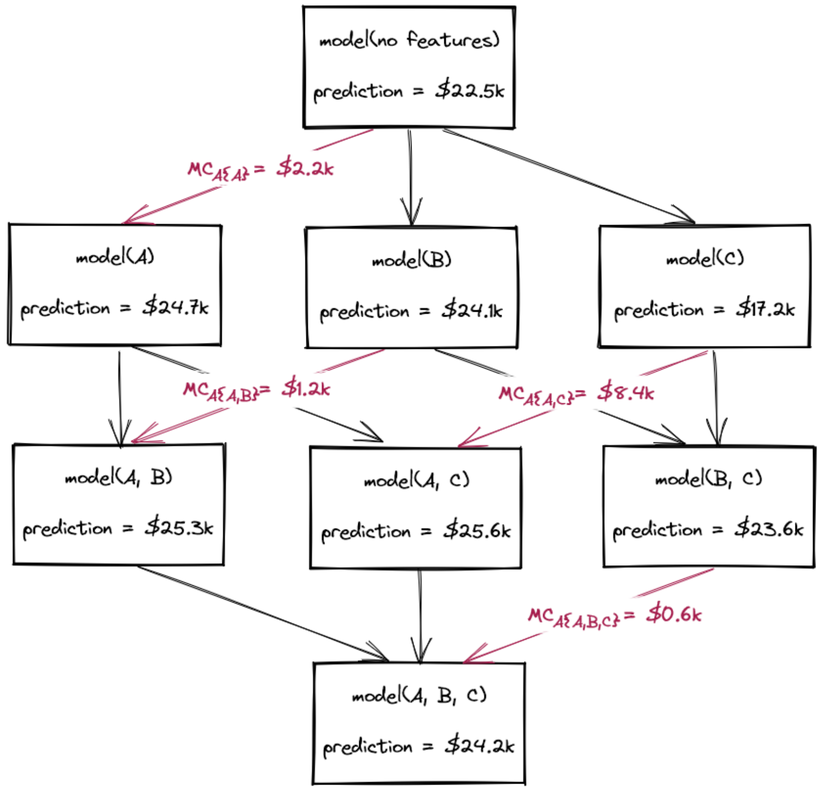


$$\sum_{T \subseteq D \setminus \{i\}} \frac{|T|!\,(|D|-|T|-1)!}{|D|!} \left[v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x}) \right]$$

https://www.aidancooper.co.uk/how-shapley-values-work/

a simple example

	A	B	C	predicted price
house1	10	6.5	0.5	$24,200



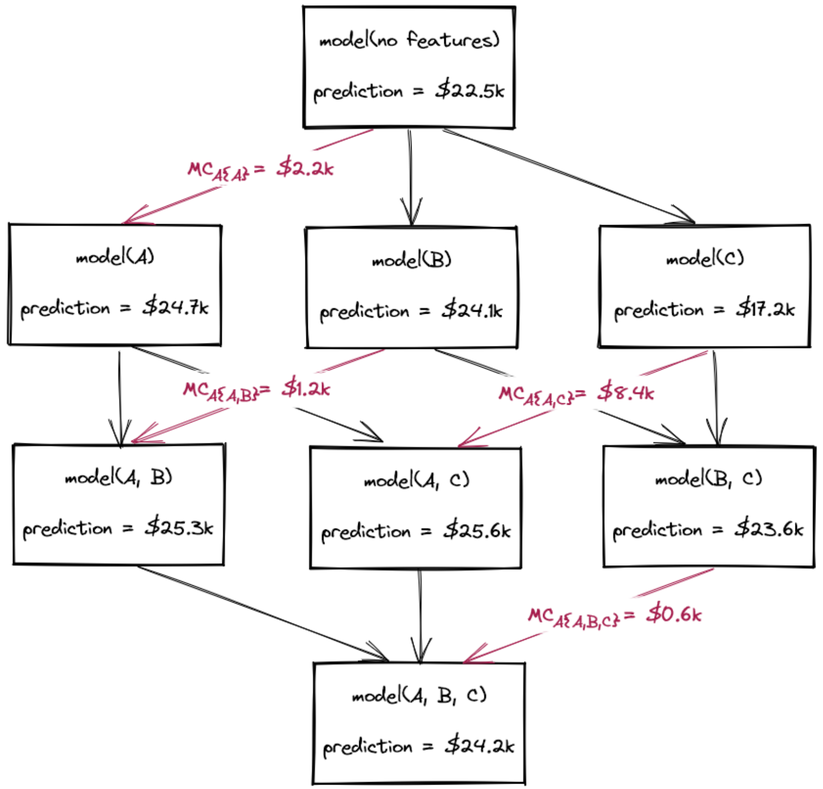
$$\sum_{T \subseteq D \setminus \{i\}} \frac{|T|!\,(|D|-|T|-1)!}{|D|!} \left[v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x}) \right]$$

$$i = \{A\} \qquad D = \{A, B, C\}$$

https://www.aidancooper.co.uk/how-shapley-values-work/

a simple example

	A	B	C	predicted price
house1	10	6.5	0.5	$24,200



$$\sum_{T \subseteq D \setminus \{i\}} \frac{|T|!\,(|D|-|T|-1)!}{|D|!} \left[v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x}) \right]$$

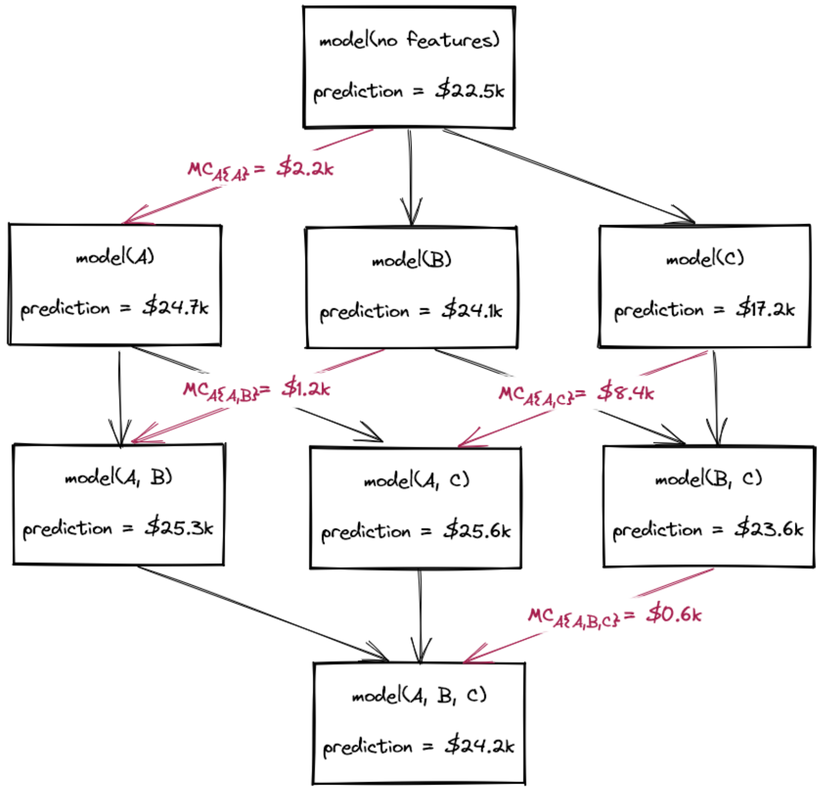
$$i = \{A\} \qquad D = \{A, B, C\}$$

T	$\frac{\lvert T\rvert!\,(\lvert D\rvert-\lvert T\rvert-1)!}{\lvert D\rvert!}$	$v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x})$

https://www.aidancooper.co.uk/how-shapley-values-work/

a simple example

	A	B	C	predicted price
house1	10	6.5	0.5	$24,200



$$\sum_{T \subseteq D \setminus \{i\}} \frac{|T|!\,(|D|-|T|-1)!}{|D|!} \left[v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x}) \right]$$

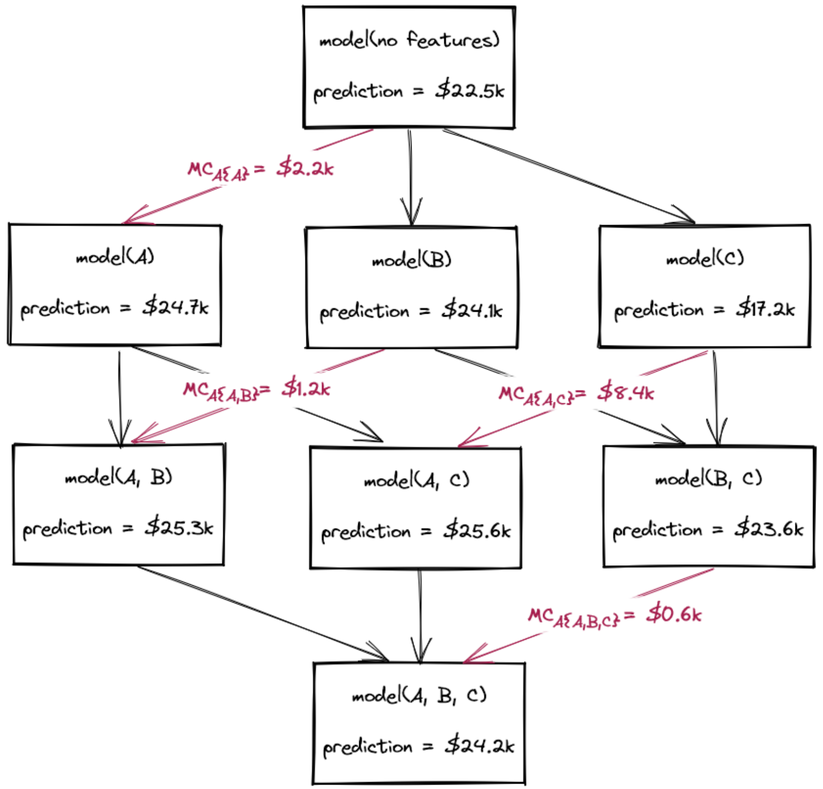
$$i = \{A\} \qquad D = \{A, B, C\}$$

T	$\frac{\lvert T\rvert!\,(\lvert D\rvert-\lvert T\rvert-1)!}{\lvert D\rvert!}$	$v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x})$
$\emptyset$		
$\{B\}$		
$\{C\}$		
$\{B, C\}$		

https://www.aidancooper.co.uk/how-shapley-values-work/

a simple example

	A	B	C	predicted price
house1	10	6.5	0.5	$24,200



$$\sum_{T \subseteq D \setminus \{i\}} \frac{|T|!\,(|D|-|T|-1)!}{|D|!} \left[v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x}) \right]$$

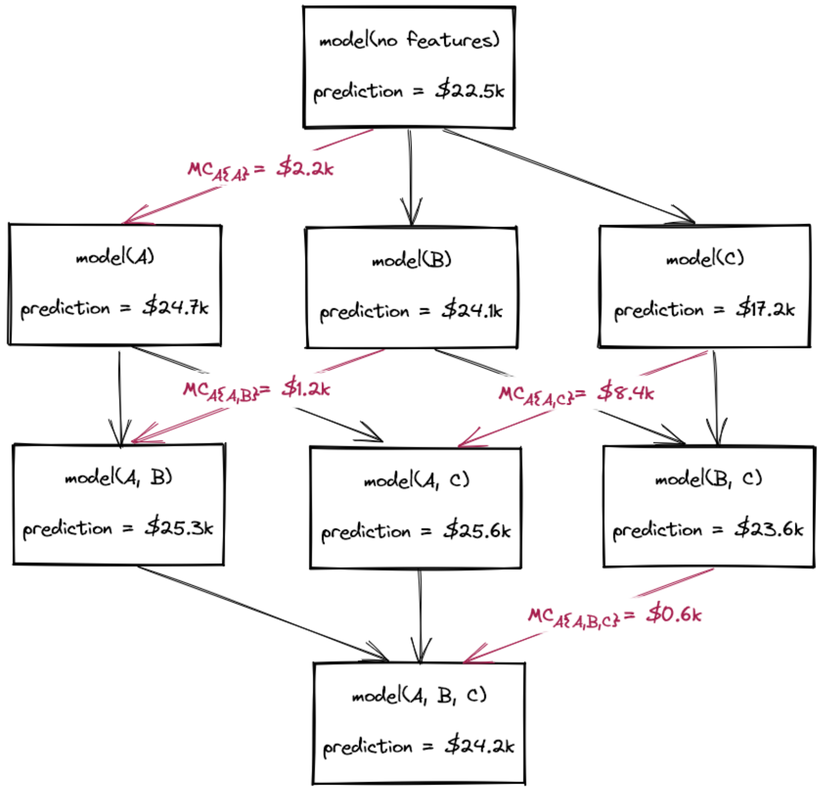
$$i = \{A\} \qquad D = \{A, B, C\}$$

T	$\frac{\lvert T\rvert!\,(\lvert D\rvert-\lvert T\rvert-1)!}{\lvert D\rvert!}$	$v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x})$
$\emptyset$	$\frac{0!(3-0-1)!}{3!} = \frac{1}{3}$	$2,200
$\{B\}$		
$\{C\}$		
$\{B, C\}$		

https://www.aidancooper.co.uk/how-shapley-values-work/

a simple example

	A	B	C	predicted price
house1	10	6.5	0.5	$24,200



$$\sum_{T \subseteq D \setminus \{i\}} \frac{|T|!\,(|D|-|T|-1)!}{|D|!} \left[v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x}) \right]$$

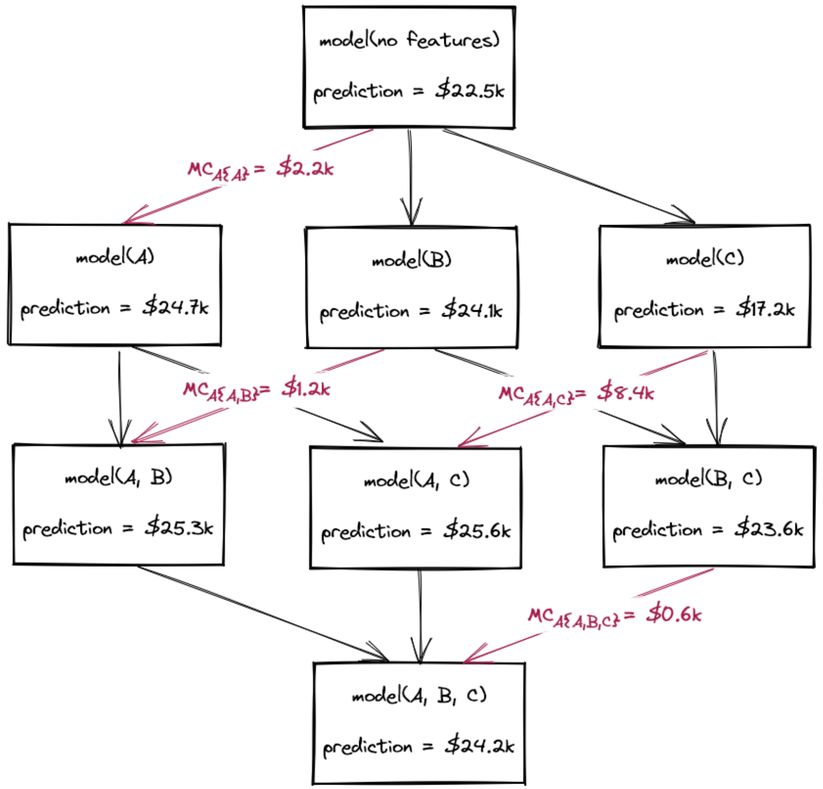
$$i = \{A\} \qquad D = \{A, B, C\}$$

T	$\frac{\lvert T\rvert!\,(\lvert D\rvert-\lvert T\rvert-1)!}{\lvert D\rvert!}$	$v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x})$
$\emptyset$	$\frac{0!(3-0-1)!}{3!} = \frac{1}{3}$	$2,200
$\{B\}$	1/6	$1,200
$\{C\}$	1/6	$8,400
$\{B, C\}$	1/3	$600

https://www.aidancooper.co.uk/how-shapley-values-work/

a simple example

	A	B	C	predicted price
house1	10	6.5	0.5	$24,200



$$\sum_{T \subseteq D \setminus \{i\}} \frac{|T|!\,(|D|-|T|-1)!}{|D|!} \left[v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x}) \right]$$

$$i = \{A\} \qquad D = \{A, B, C\}$$

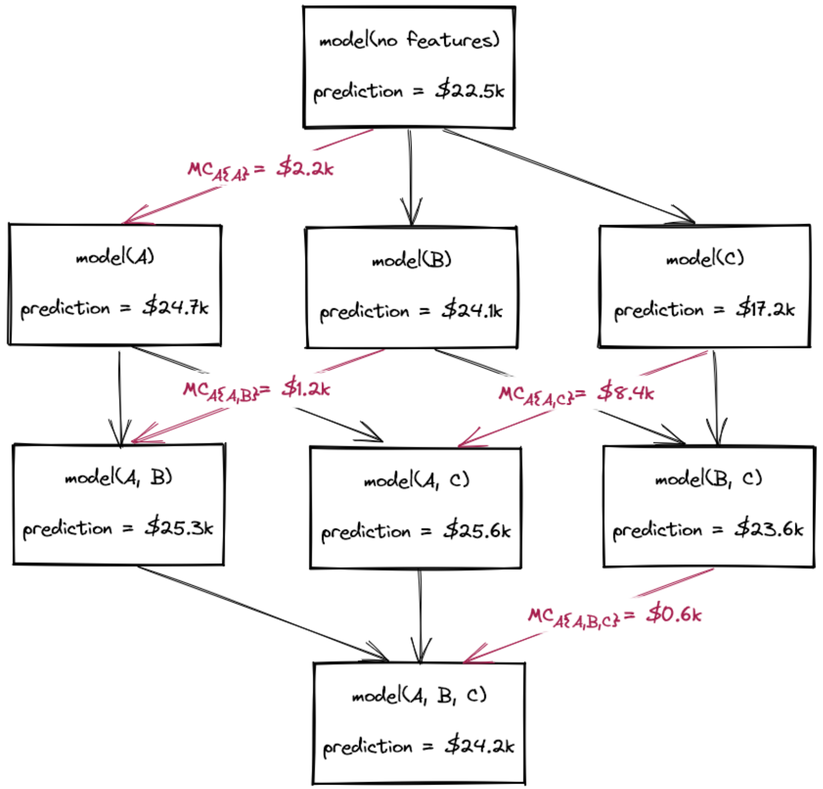
T	$\frac{\lvert T\rvert!\,(\lvert D\rvert-\lvert T\rvert-1)!}{\lvert D\rvert!}$	$v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x})$
$\emptyset$	$\frac{0!(3-0-1)!}{3!} = \frac{1}{3}$	$2,200
$\{B\}$	1/6	$1,200
$\{C\}$	1/6	$8,400
$\{B, C\}$	1/3	$600

1/3($2,200) + 1/6($1,200) + 1/6($8,400) + 1/3($600) ≈ **$2500**

https://www.aidancooper.co.uk/how-shapley-values-work/

a simple example

	A	B	C	predicted price
house1	10	6.5	0.5	$24,200



$$\sum_{T \subseteq D \setminus \{i\}} \frac{|T|!\,(|D|-|T|-1)!}{|D|!} \left[v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x}) \right]$$

Shapley values

No features: $22,500

Feature A: $ 2,500

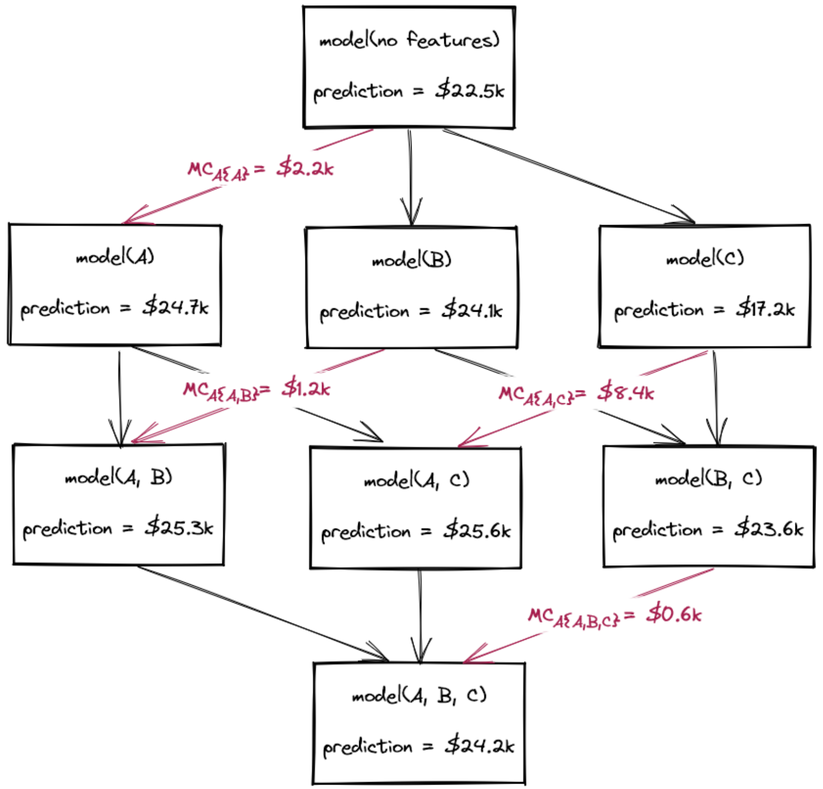
Feature B: $1,250

Feature C: -$2,100

https://www.aidancooper.co.uk/how-shapley-values-work/

a simple example

	A	B	C	predicted price
house1	10	6.5	0.5	$24,200



$$\sum_{T \subseteq D \setminus \{i\}} \frac{|T|!\,(|D|-|T|-1)!}{|D|!} \left[v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x}) \right]$$

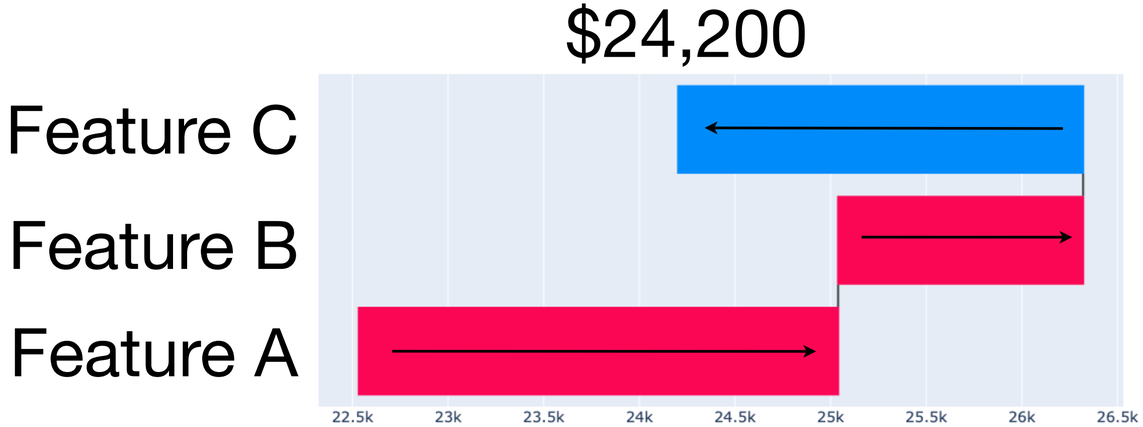
Shapley values

No features: $22,500

Feature A: $ 2,500

Feature B: $1,250

Feature C: -$2,100



https://www.aidancooper.co.uk/how-shapley-values-work/

estimating Shapley values and interactions

KernelSHAP: weighted least squares regression

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

$$\min_{\beta_0,\ldots,\beta_d} \sum_{S \subseteq D} \mu(S)\Big(u(S) - v(S)\Big)^2$$

estimating Shapley values and interactions

KernelSHAP: weighted least squares regression

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

$$\min_{\beta_0,\ldots,\beta_d} \sum_{S \subseteq D} \mu(S)\Big(u(S) - v(S)\Big)^2$$

DeepSHAP/DeepLIFT: estimate via gradients

Learning Important Features Through Propagating Activation Differences

Avanti Shrikumar [1] **Peyton Greenside** [1] **Anshul Kundaje** [1]

estimating Shapley values and interactions

KernelSHAP: weighted least squares regression

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

$$\min_{\beta_0,\ldots,\beta_d} \sum_{S \subseteq D} \mu(S)\Big(u(S) - v(S)\Big)^2$$

DeepSHAP/DeepLIFT: estimate via gradients

Learning Important Features Through Propagating Activation Differences

Avanti Shrikumar [1] Peyton Greenside [1] Anshul Kundaje [1]

SHAP-IQ extend to Shapley interactions

SHAP-IQ: Unified Approximation of any-order Shapley Interactions

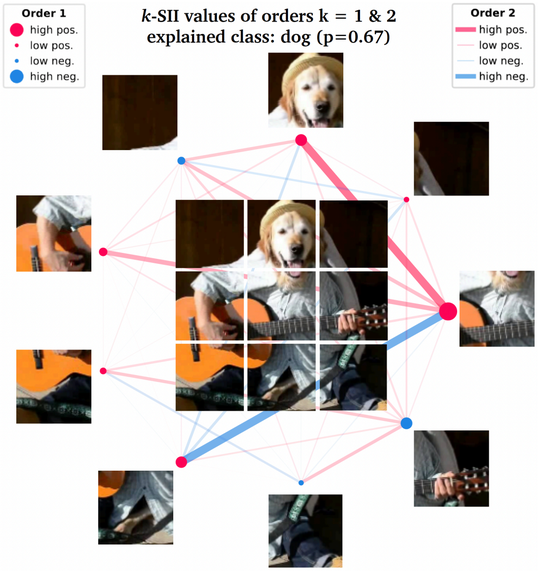
Fabian Fumagalli*
Bielefeld University, CITEC
D-33619, Bielefeld, Germany
ffumagalli@techfak.uni-bielefeld.de

Maximilian Muschalik*
LMU Munich, MCML Munich
D-80539, Munich, Germany
maximilian.muschalik@ifi.lmu.de

Patrick Kolpaczki
Paderborn University
D-33098, Paderborn, Germany
patrick.kolpaczki@upb.de

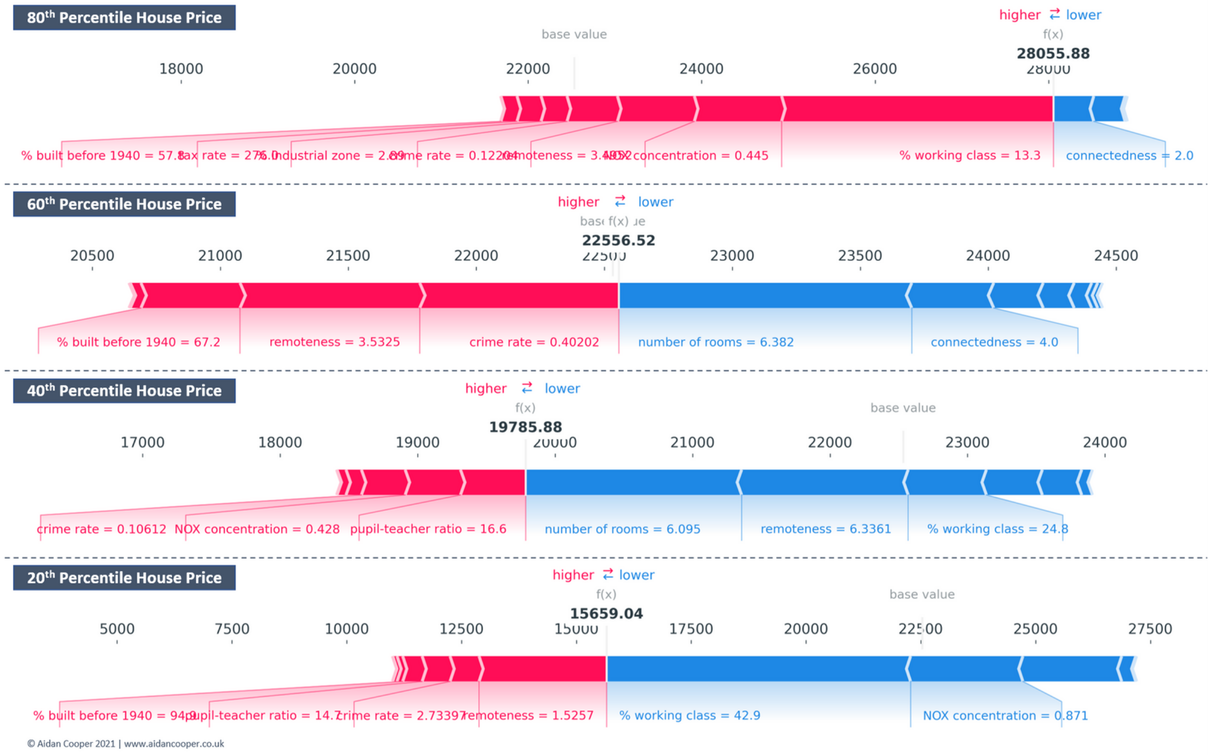
Eyke Hüllermeier
LMU Munich, MCML Munich
D-80539, Munich, Germany
eyke@ifi.lmu.de

Barbara Hammer
Bielefeld University, CITEC
D-33619, Bielefeld, Germany
bhammer@techfak.uni-bielefeld.de



applications of Shapley explanations

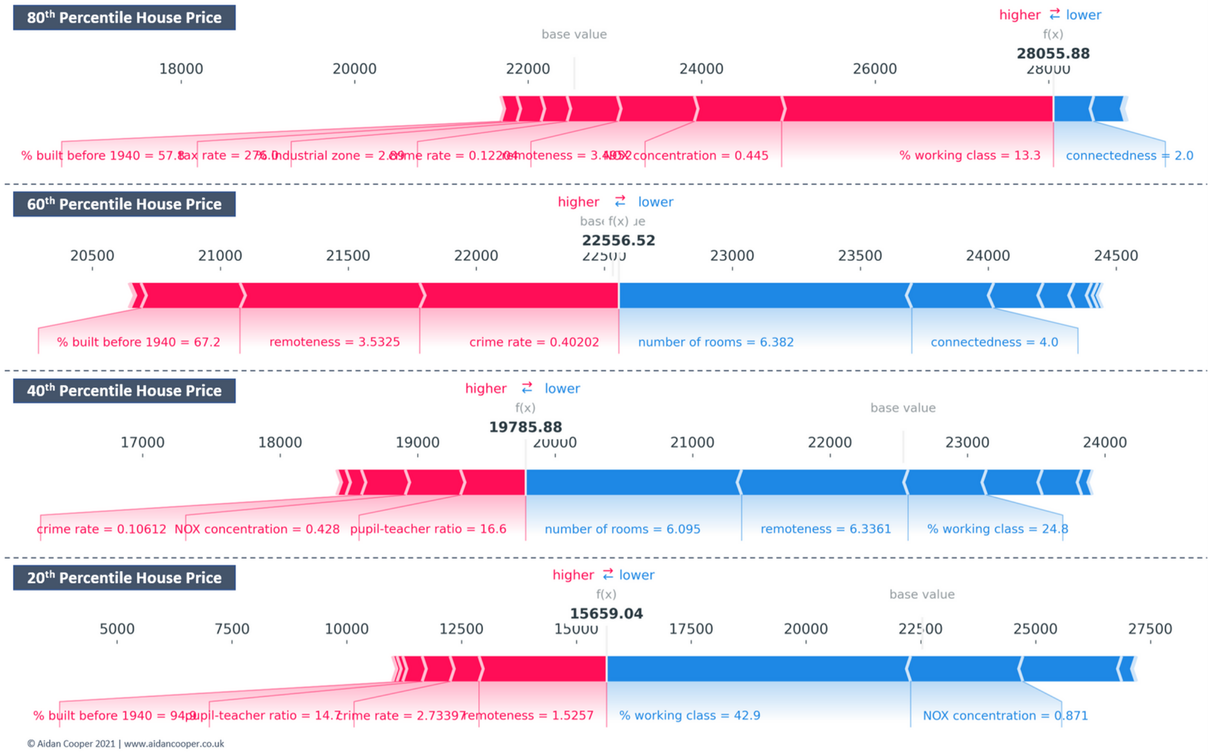
Tabular data



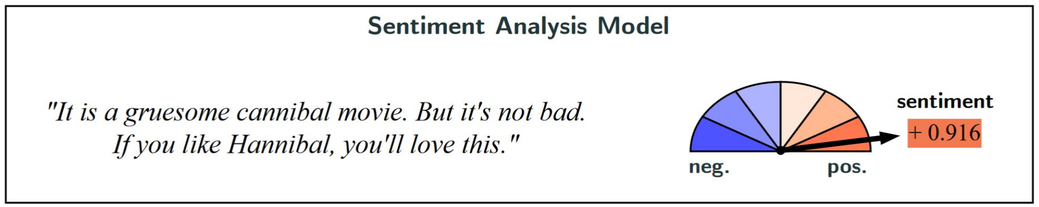
https://www.aidancooper.co.uk/how-shapley-values-work/, Fumagalli *et. al, NeurIPS* (2023), *Jethani et. al, ICLR* (2022)

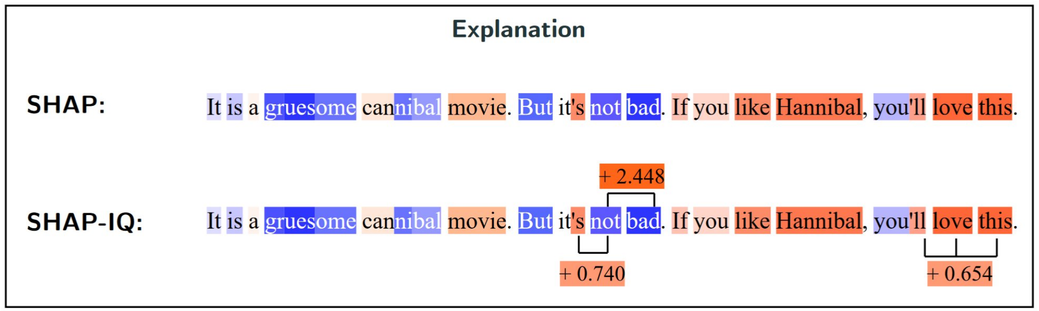
applications of Shapley explanations

Tabular data



NLP

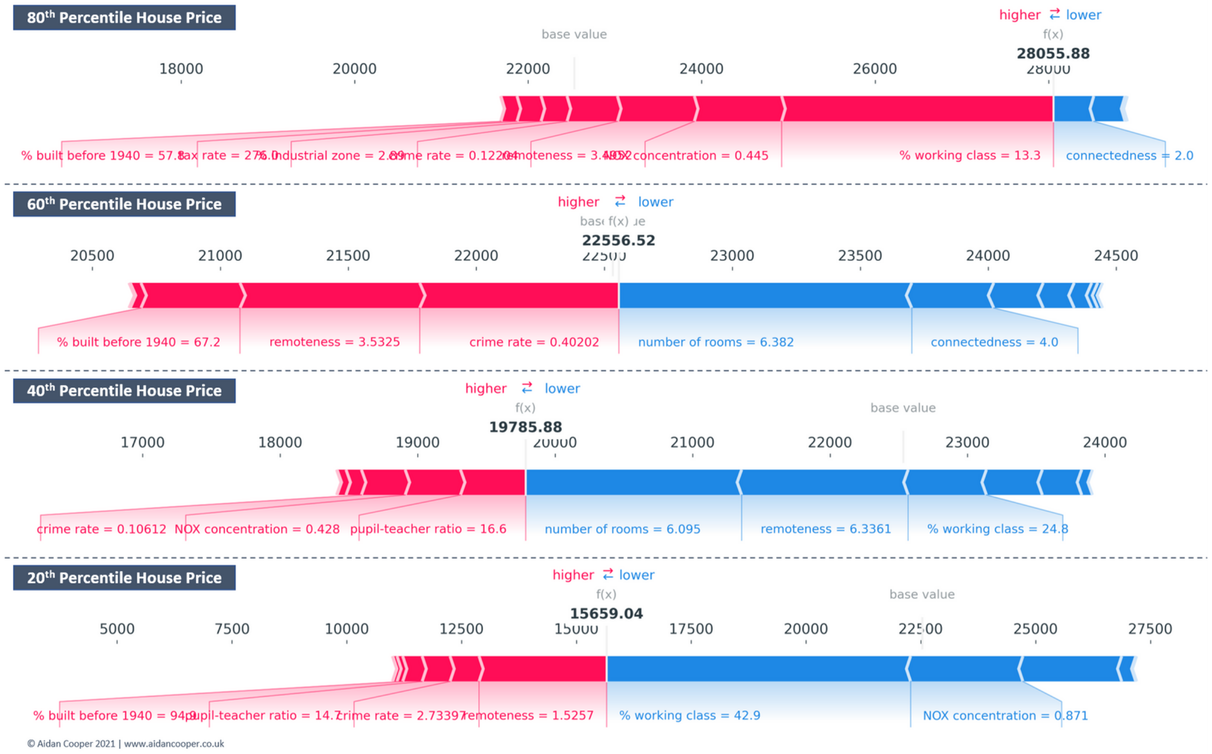




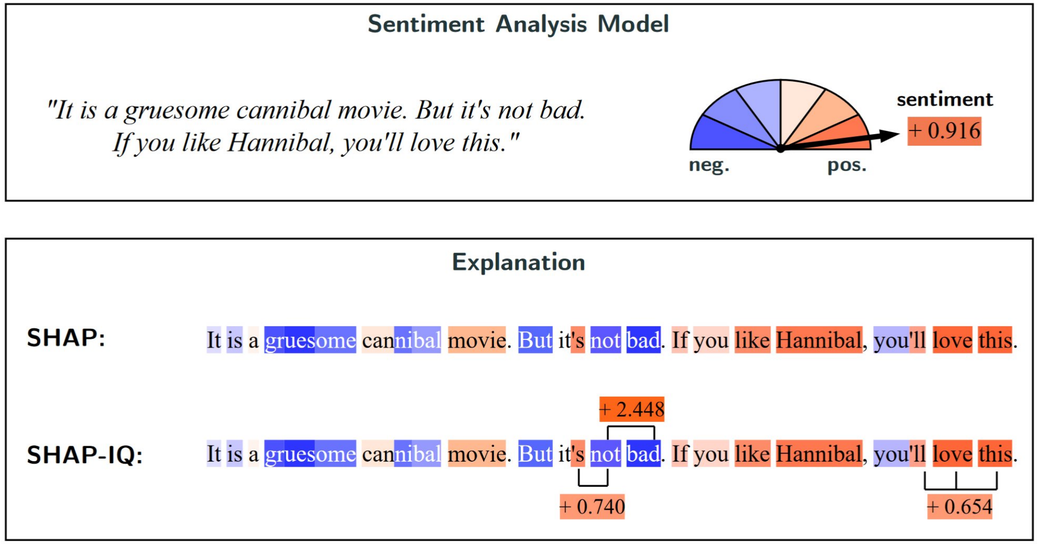
https://www.aidancooper.co.uk/how-shapley-values-work/, Fumagalli *et. al, NeurIPS* (2023), *Jethani et. al, ICLR* (2022)

applications of Shapley explanations

Tabular data



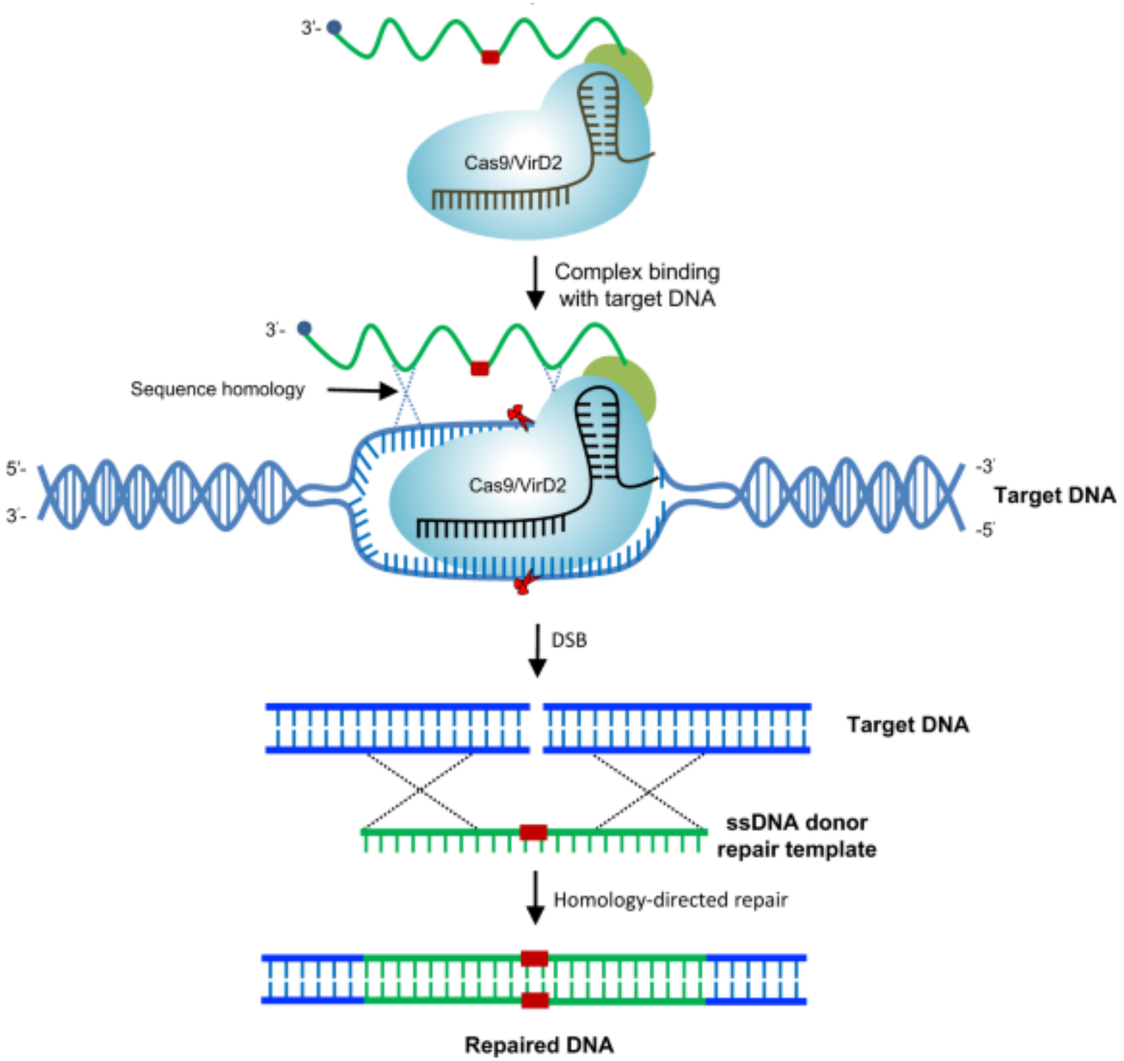
NLP



Vision

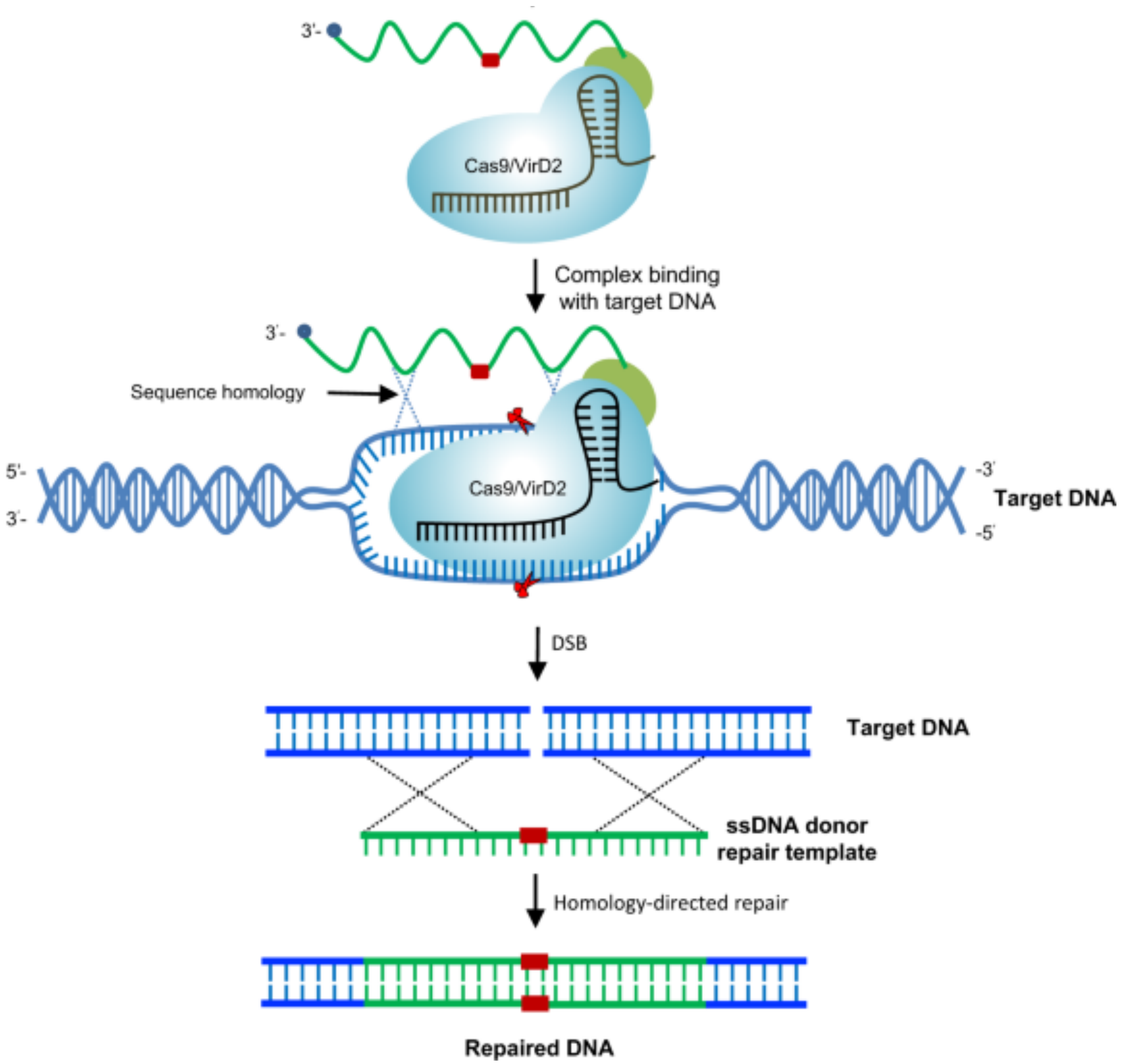
https://www.aidancooper.co.uk/how-shapley-values-work/, Fumagalli *et. al, NeurIPS* (2023), *Jethani et. al, ICLR* (2022)

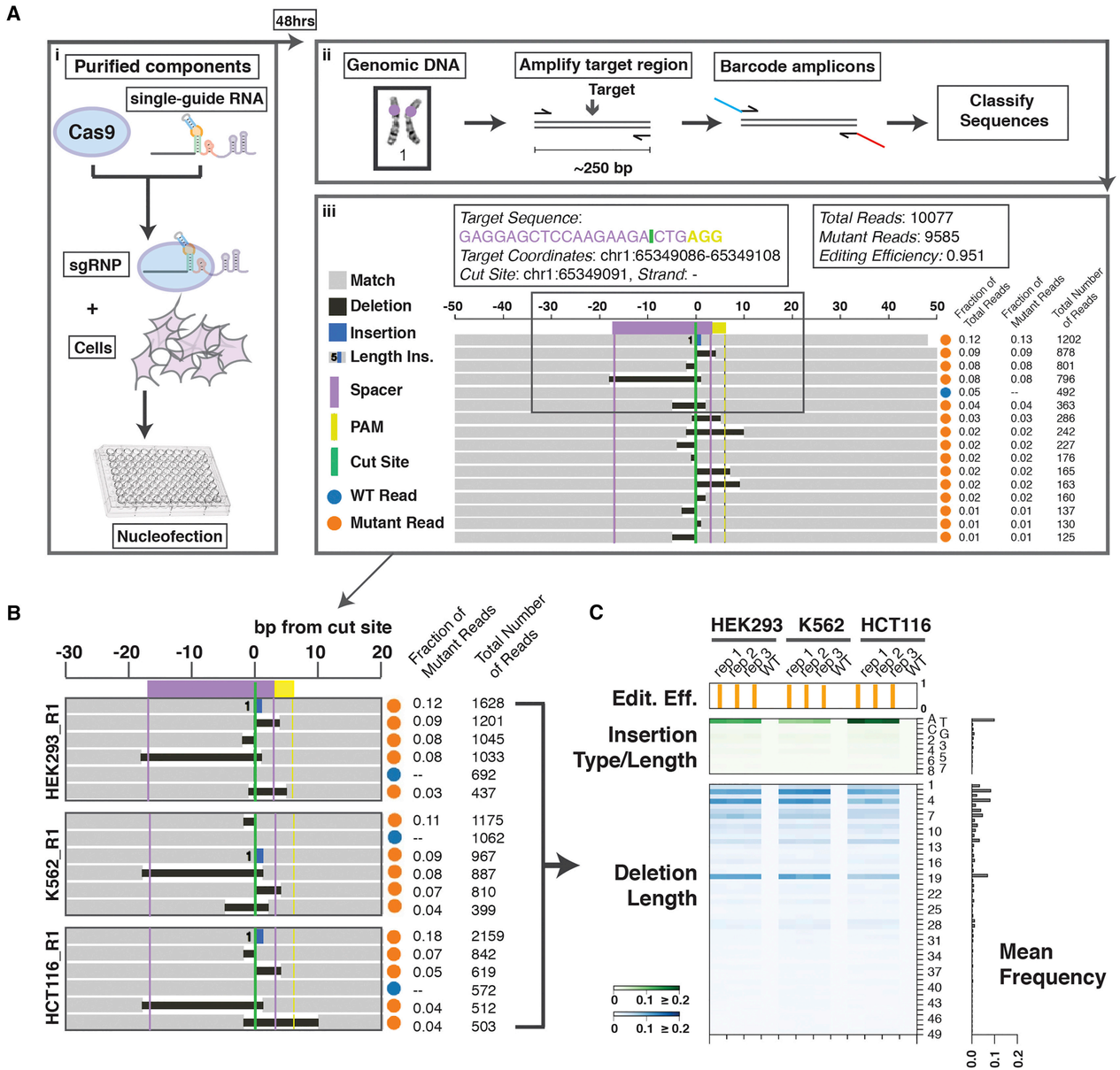
we want to explain biological models over <u>many</u> sequences



Ali *et. al*, *Communications Biology* (2020), van Overbeek *et. al*, *Molecular Cell* (2016)

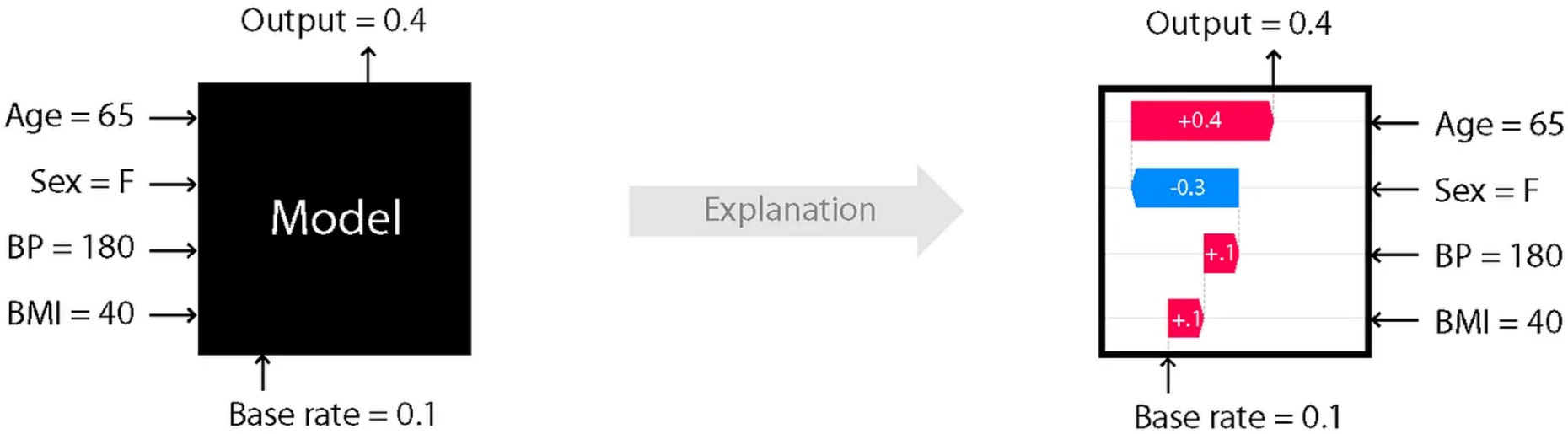
we want to explain biological models over <u>many</u> sequences





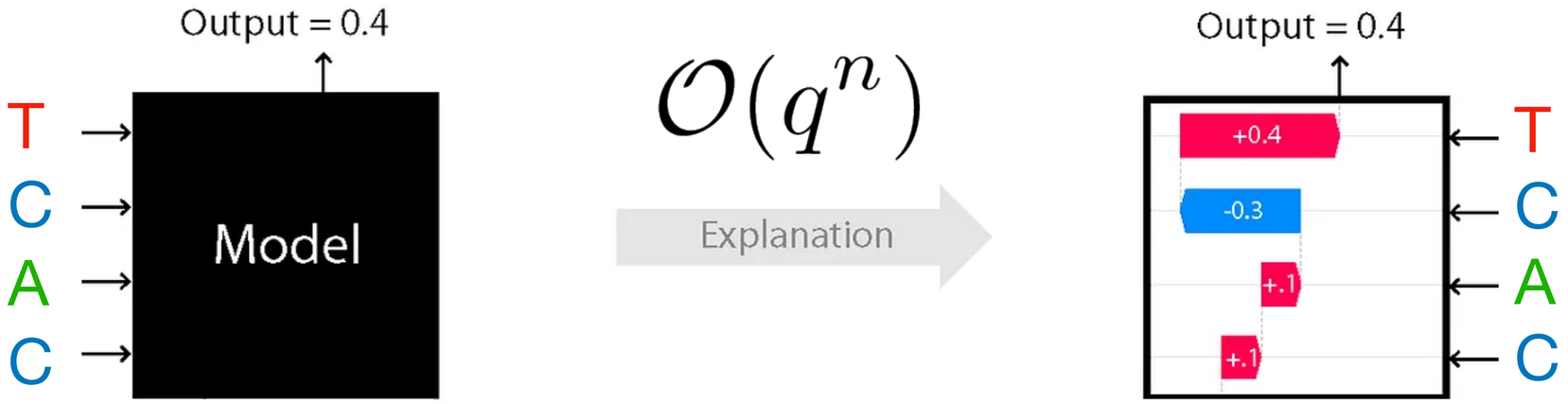
Ali *et. al*, *Communications Biology* (2020), van Overbeek *et. al*, *Molecular Cell* (2016)

we want to explain biological models over <u>many</u> sequences



$$\sum_{T \subseteq D \setminus \{i\}} \frac{|T|!\,(|D|-|T|-1)!}{|D|!} \left[v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x}) \right]$$

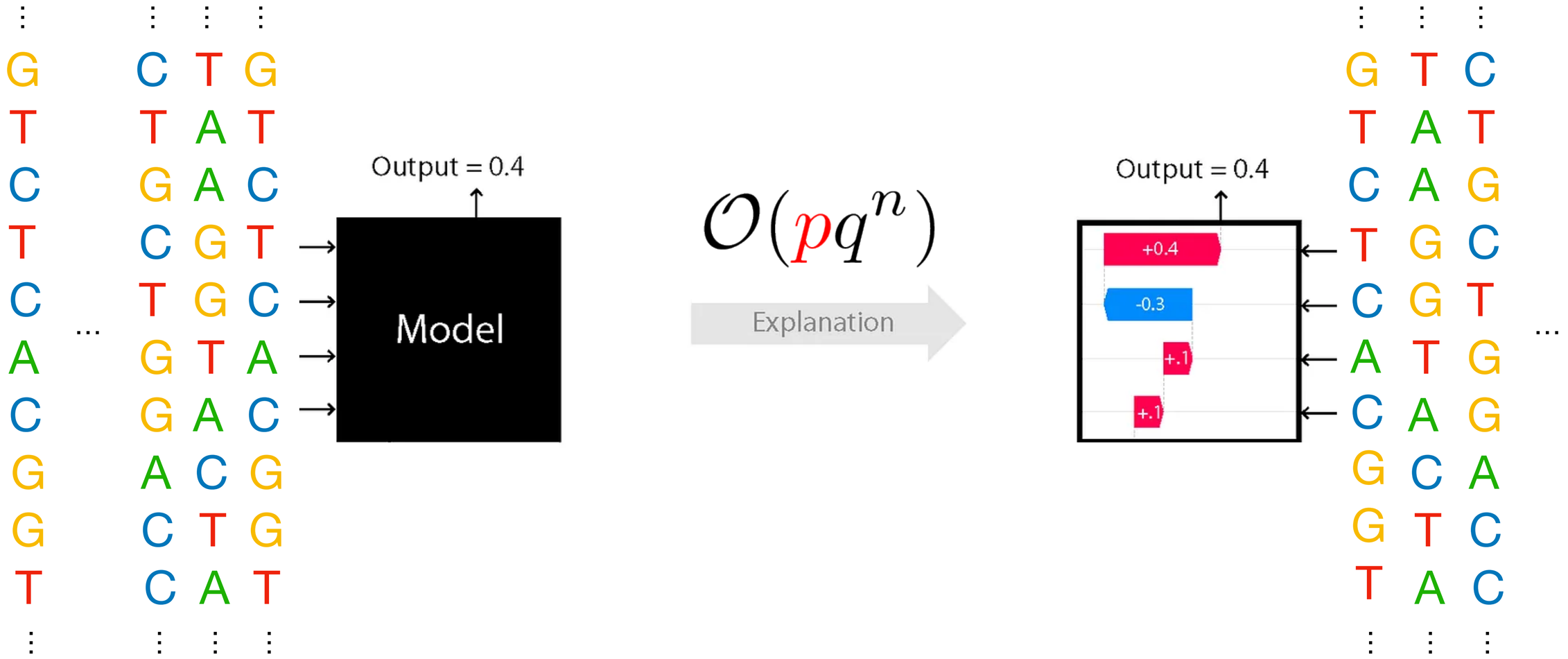
we want to explain biological models over <u>many</u> sequences



$$\sum_{T \subseteq D \setminus \{i\}} \frac{|T|!\,(|D|-|T|-1)!}{|D|!} \left[v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x}) \right]$$

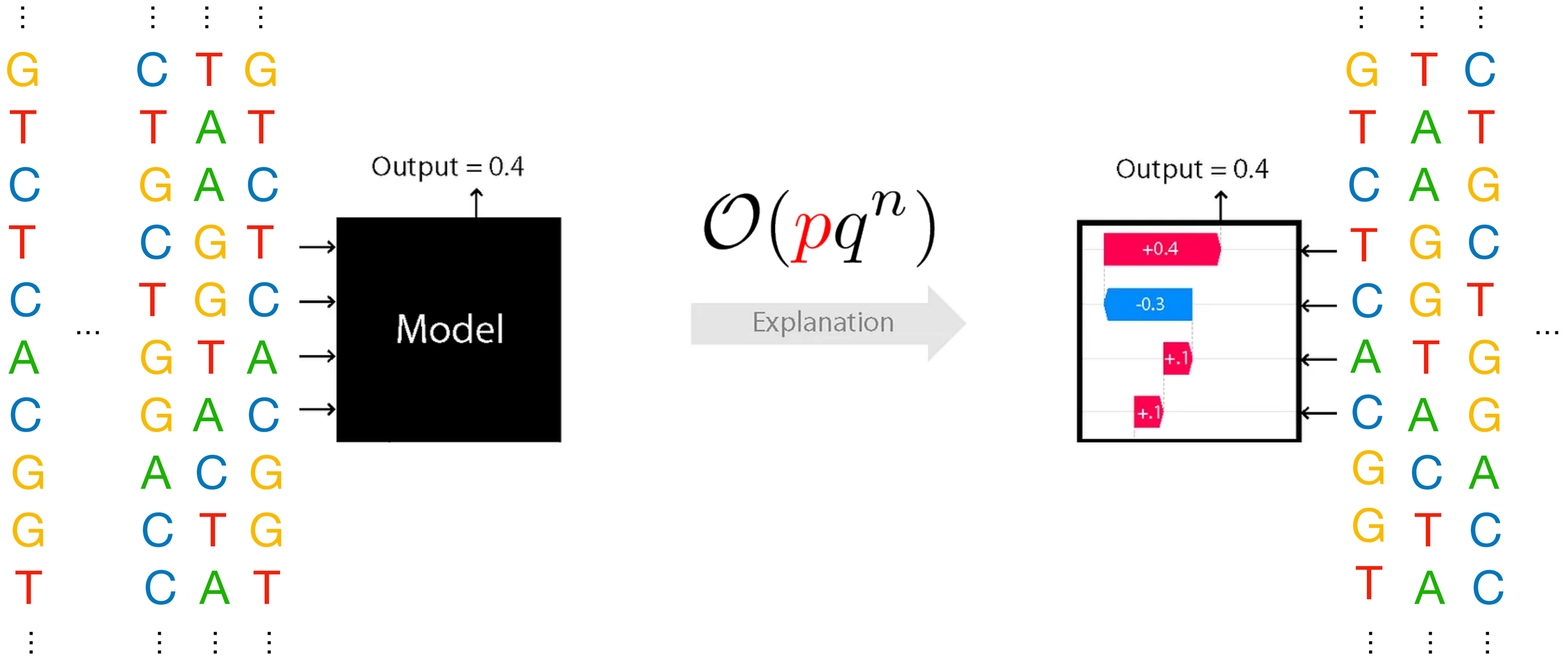
shap

we want to explain biological models over <u>many</u> sequences



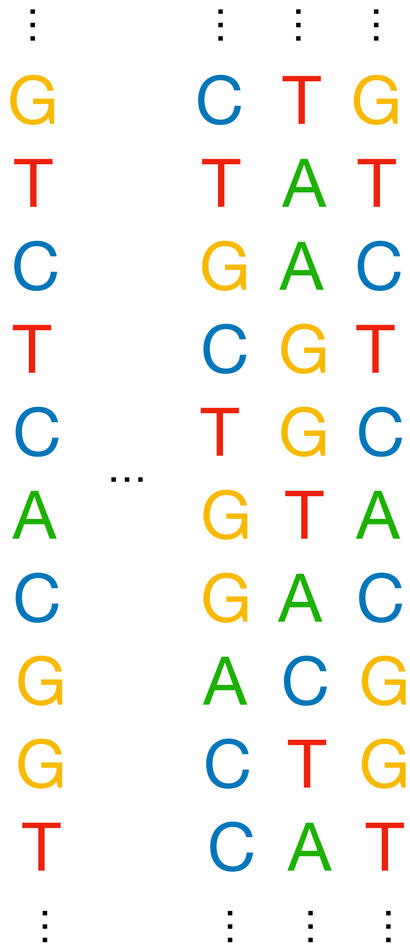
$$\sum_{T \subseteq D \setminus \{i\}} \frac{|T|!\,(|D|-|T|-1)!}{|D|!} \left[v_{T \cup \{i\}}(\mathbf{x}) - v_T(\mathbf{x}) \right]$$

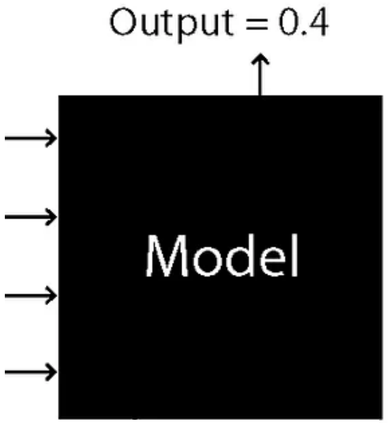
we want to explain biological models over <u>many</u> sequences



Explaining 1038 sequences
SHAP-IQ ~ 81 days!

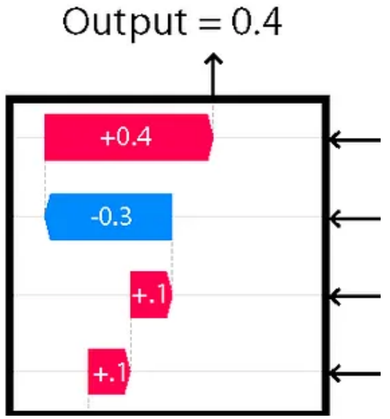
we want to explain biological models over <u>many</u> sequences

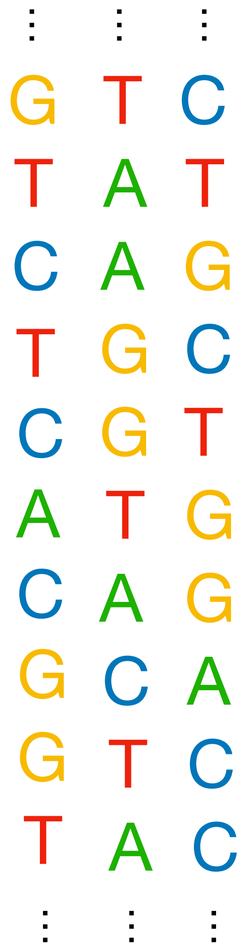


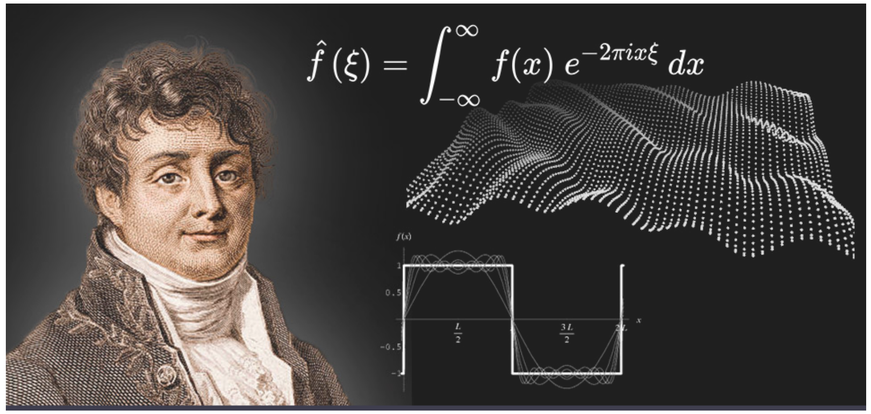


$$\mathcal{O}(pq^n)$$

Explanation

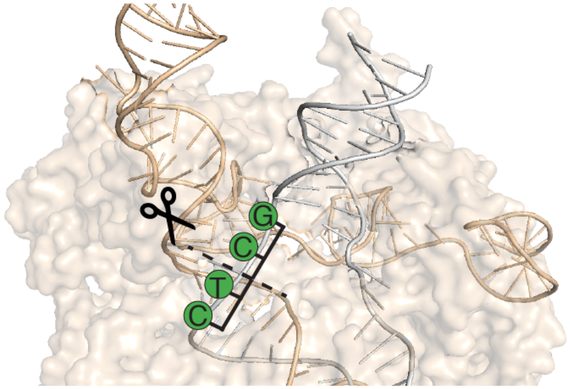




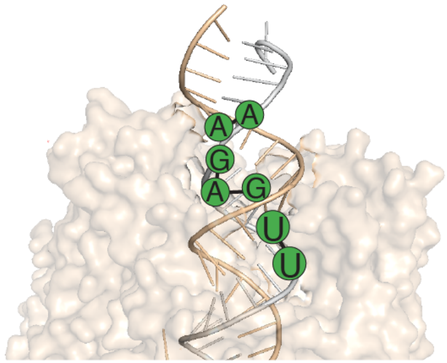


biological sequence models are fundamentally sparse

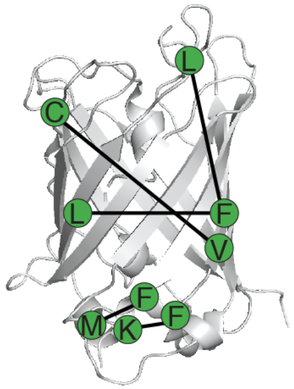
a Interactions in DNA repair



b Interactions in guide RNA binding

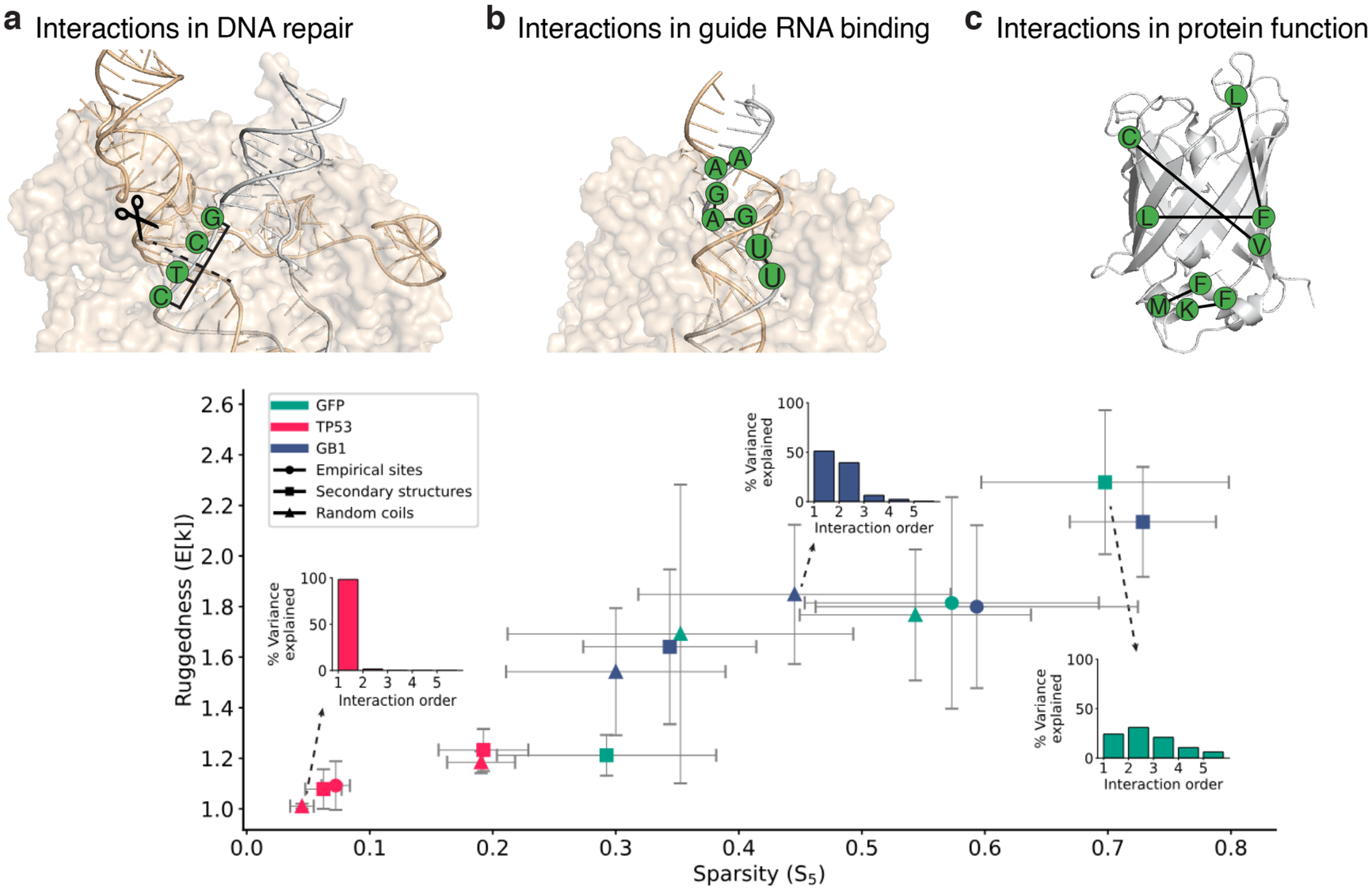


c Interactions in protein function



Kang, **Tsui**, Erginbas, Butler, Aghazadeh, Ramchandran, *under review* (2025), **Tsui**, Aghazadeh, ICLR GEM Bio (2024)

biological sequence models are fundamentally sparse



Kang, **Tsui**, Erginbas, Butler, Aghazadeh, Ramchandran, *under review* (2025), **Tsui**, Aghazadeh, ICLR GEM Bio (2024)

SHAP zero amortizes Shapley explanations

Algorithm. SHAP zero in three steps:

- **Step 1**: Pay a **one-time cost** to sketch the model globally via the Fourier transform with **sample complexity** $\mathcal{O}(sn^2)$ and **computational complexity** $\mathcal{O}(sn^3)$

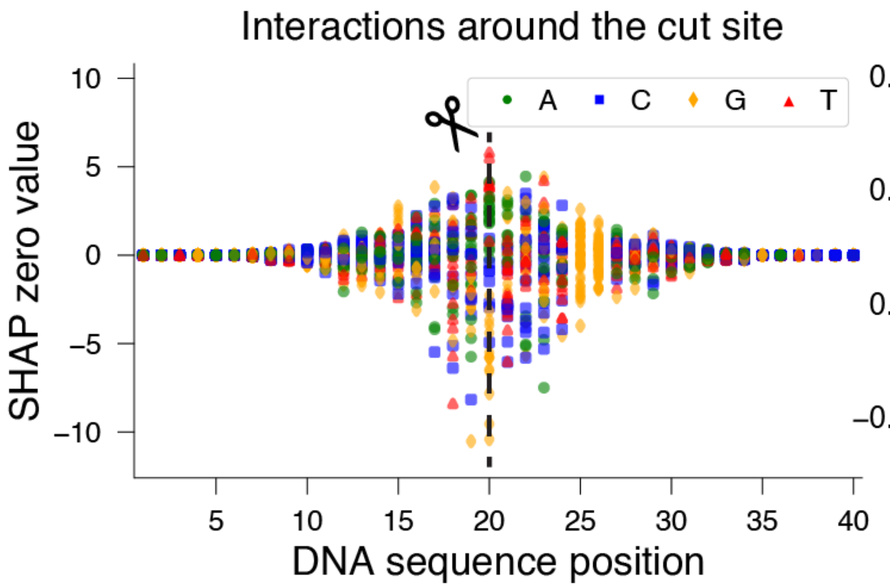
For each future sequence:

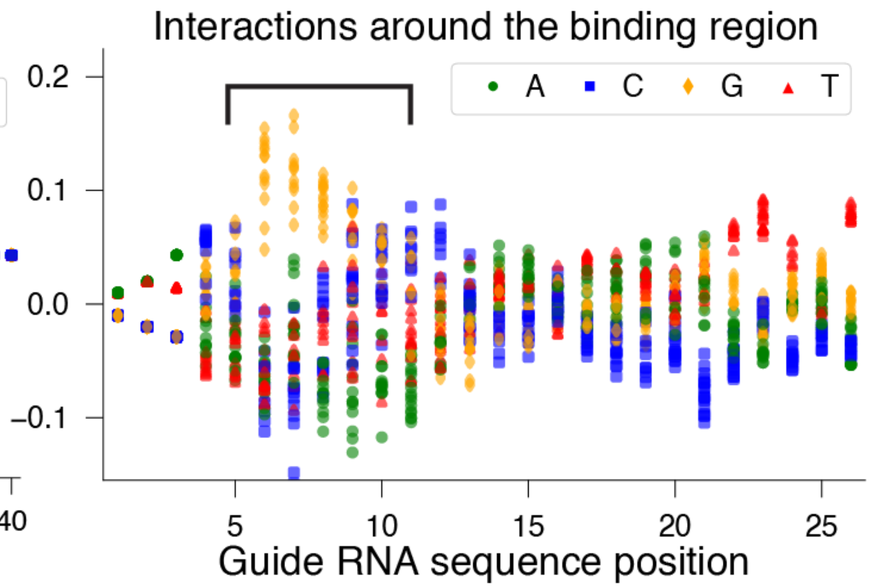
- **Step 2**: Map global sketch to the local Möbius transform to isolate interacting features with complexity $\mathcal{O}(s^2(2q)^\ell)$
- **Step 3**: Map Möbius transform to Shapley values and interactions with complexity $\mathcal{O}(s^2(2q)^\ell)$

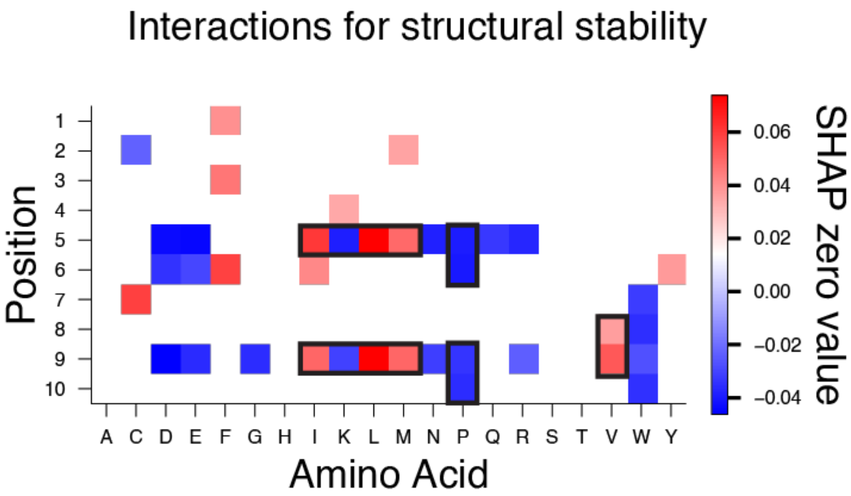
- After Step 1, SHAP zero explains for **essentially free**!

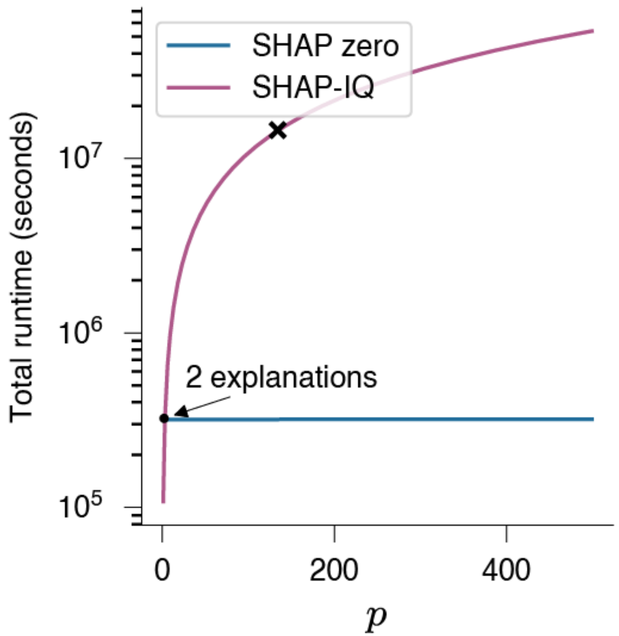
Tsui, Musharaf, Erginbas, Kang, Aghazadeh, *NeurIPS* (2025)

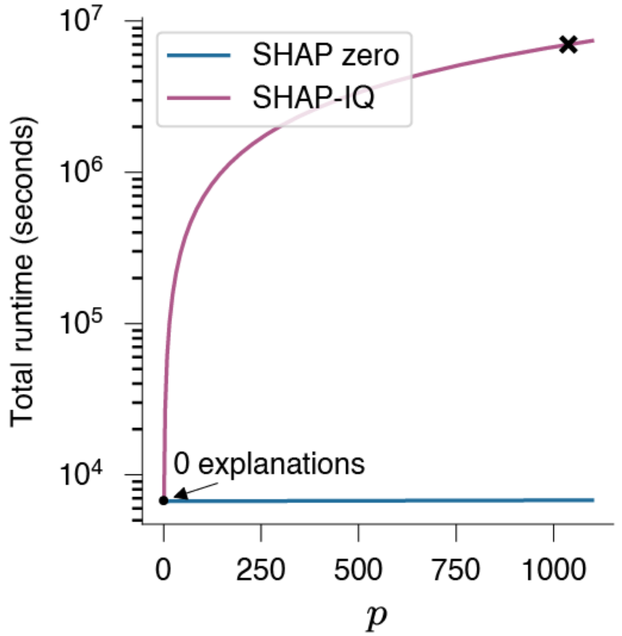
SHAP zero amortizes Shapley explanations

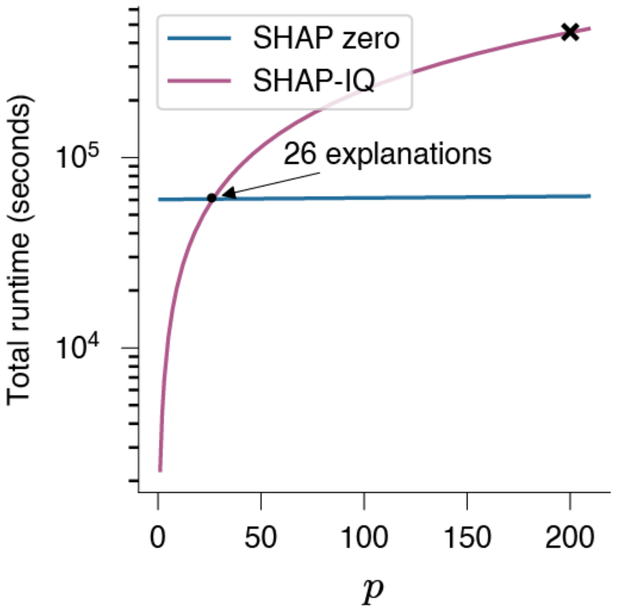












Tsui, Musharaf, Erginbas, Kang, Aghazadeh, *NeurIPS* (2025)

limitations of Shapley explanations

- **Model-agnostic:** Explains which features matter, but not *how* the model processed that input internally

limitations of Shapley explanations

- **Model-agnostic:** Explains which features matter, but not *how* the model processed that input internally

- **Untouched black-box**: Does not explain the internal representations or the specific computational steps

limitations of Shapley explanations

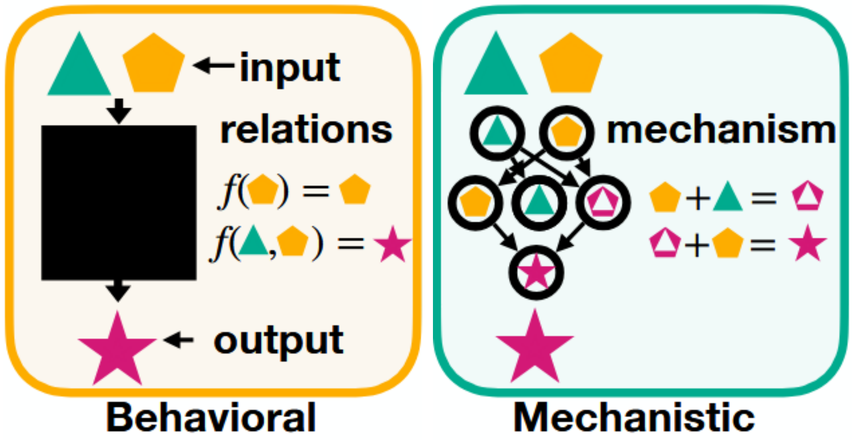
- **Model-agnostic:** Explains which features matter, but not *how* the model processed that input internally
- **Untouched black-box**: Does not explain the internal representations or the specific computational steps
- **Computational drawbacks**: Shapley explanations are slow as length of input grows and require approximations

limitations of Shapley explanations

- **Model-agnostic:** Explains which features matter, but not *how* the model processed that input internally
- **Untouched black-box**: Does not explain the internal representations or the specific computational steps
- **Computational drawbacks**: Shapley explanations are slow as length of input grows and require approximations

mechanistic interpretability aims to understand
the internal computation of the model

limitations of Shapley explanations



mechanistic interpretability aims to understand the internal computation of the model

Bereska *et. al*, *TMLR* (2024)

sparse autoencoders (SAEs)

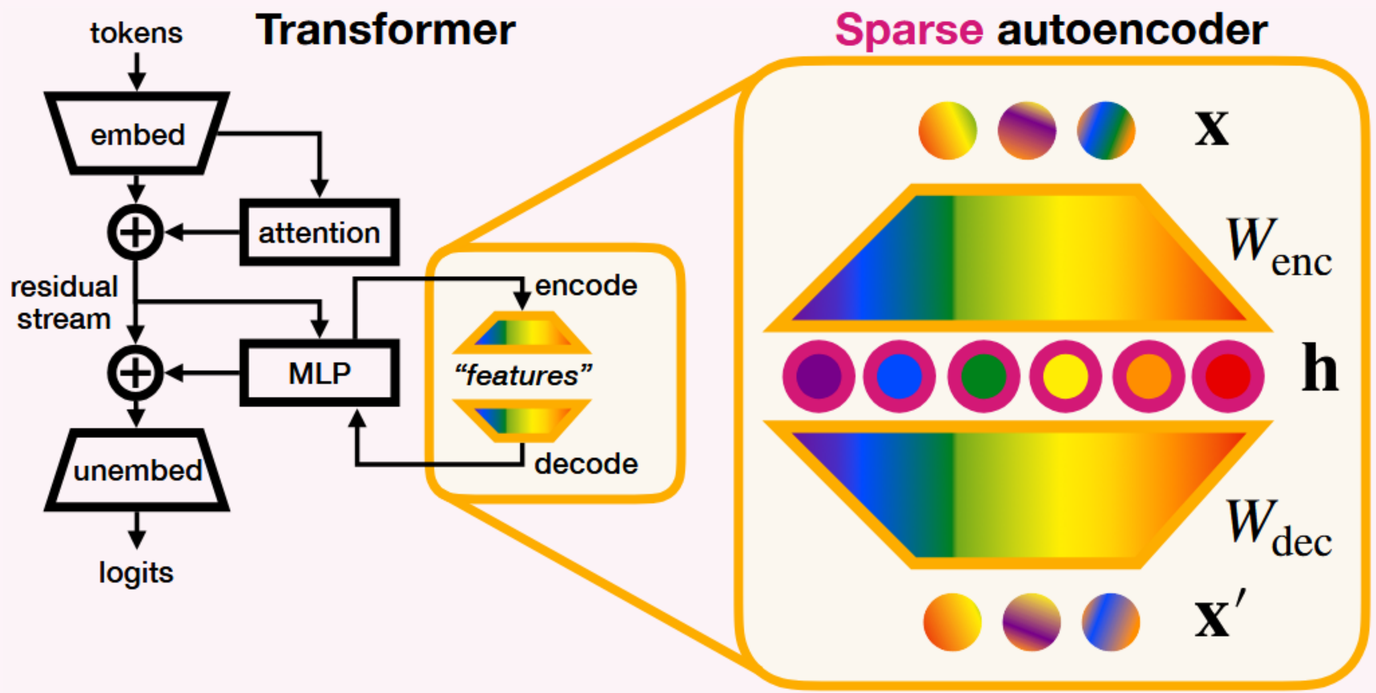


Figure 8: Illustration of a sparse autoencoder applied to the MLP layer activations, consisting of an encoder that increases dimensionality while emphasizing sparse representations and a decoder that reconstructs the original activations using the learned feature dictionary.

Bereska *et. al*, *TMLR* (2024)

sparse autoencoders (SAEs)

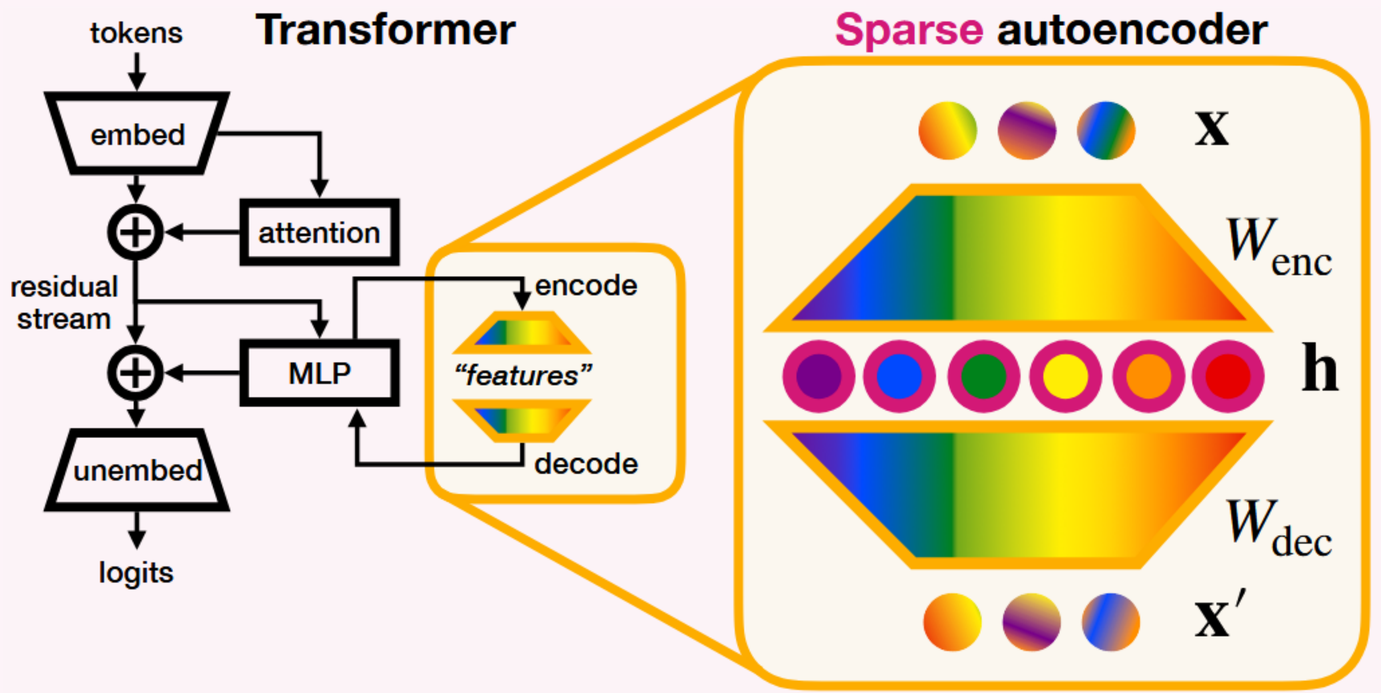


Figure 8: Illustration of a sparse autoencoder applied to the MLP layer activations, consisting of an encoder that increases dimensionality while emphasizing sparse representations and a decoder that reconstructs the original activations using the learned feature dictionary.

$$\boldsymbol{h} = \mathrm{ReLU}(\boldsymbol{W}_{\text{enc}}\boldsymbol{x} + \boldsymbol{b}),$$

$$\boldsymbol{x}' = \boldsymbol{W}_{\text{dec}}\boldsymbol{h} = \sum_{i=0}^{d_{\text{hid}}-1} h_i \boldsymbol{f}_i,$$

$$\mathcal{L}(\boldsymbol{x}) = |\boldsymbol{x} - \boldsymbol{x}'|_2^2 + \alpha|\boldsymbol{h}|_1.$$

Bereska *et. al*, *TMLR* (2024)

sparse autoencoders (SAEs)

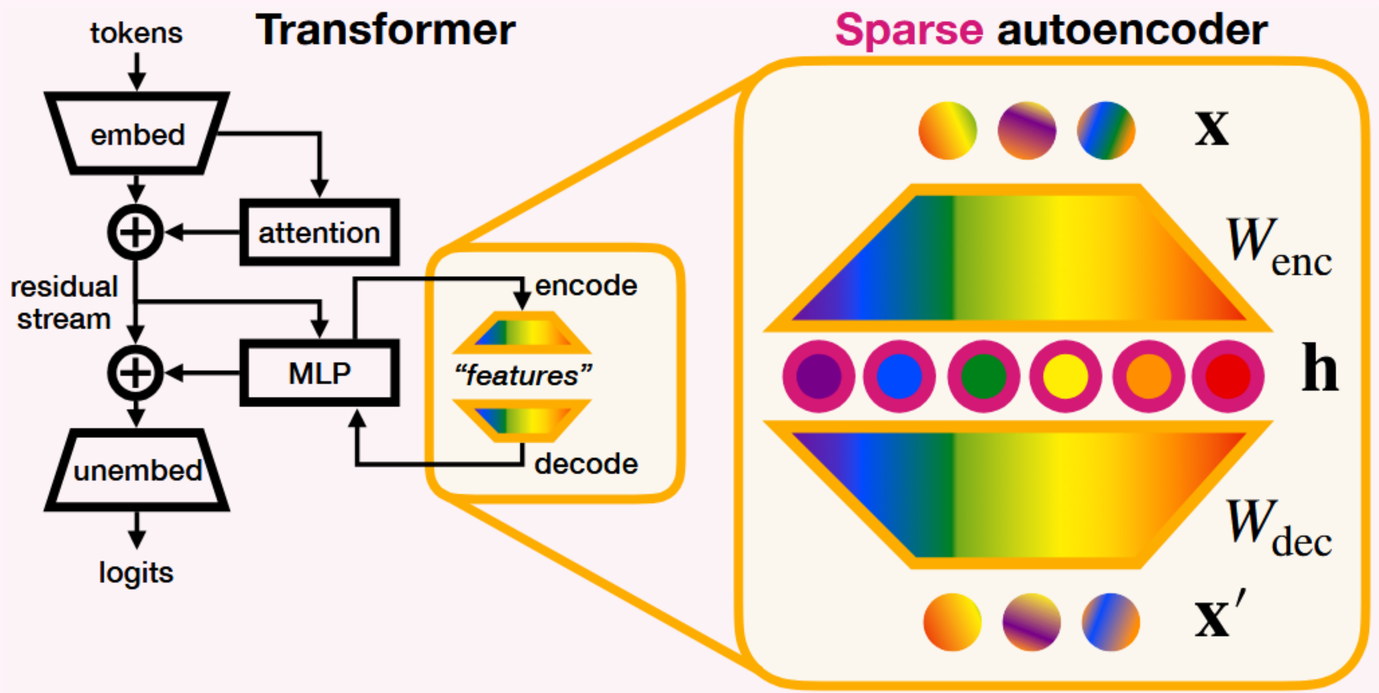


Figure 8: Illustration of a sparse autoencoder applied to the MLP layer activations, consisting of an encoder that increases dimensionality while emphasizing sparse representations and a decoder that reconstructs the original activations using the learned feature dictionary.

$$\boldsymbol{h} = \mathrm{ReLU}(\boldsymbol{W}_{\text{enc}}\boldsymbol{x} + \boldsymbol{b}),$$

$$\boldsymbol{x}' = \boldsymbol{W}_{\text{dec}}\boldsymbol{h} = \sum_{i=0}^{d_{\text{hid}}-1} h_i \boldsymbol{f}_i,$$

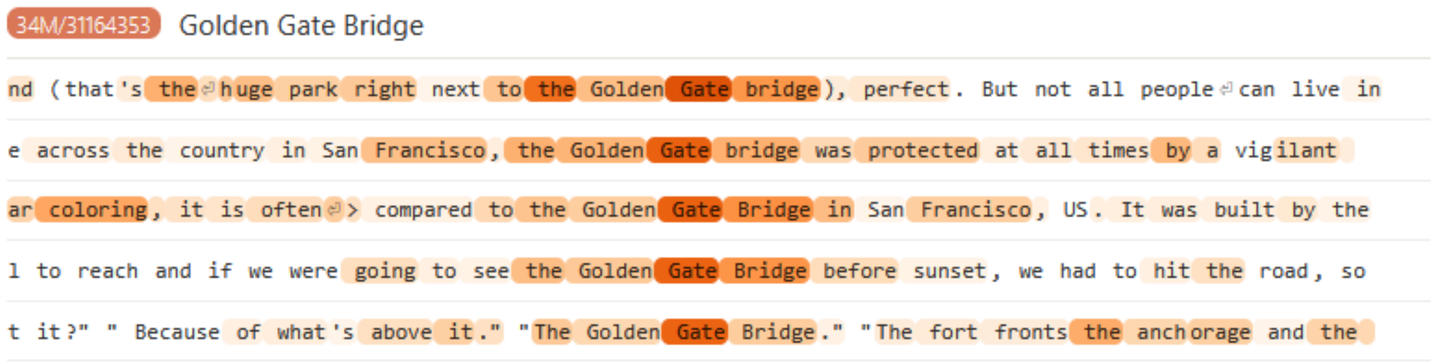
$$\mathcal{L}(\boldsymbol{x}) = |\boldsymbol{x} - \boldsymbol{x}'|_2^2 + \alpha|\boldsymbol{h}|_1.$$

find predictive features via probing

$$y = \mathbf{W}_{probe}\bar{\mathbf{h}}$$

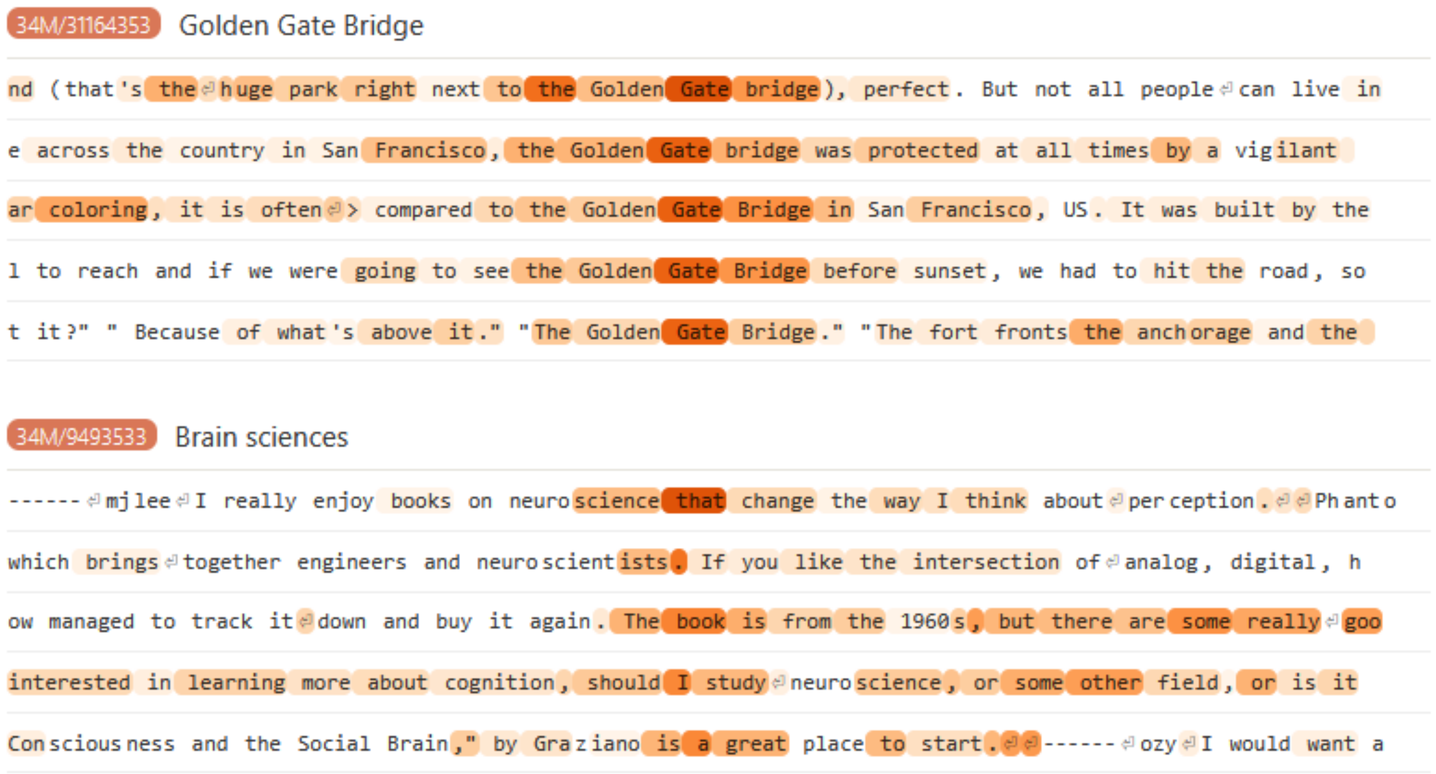
Bereska *et. al*, *TMLR* (2024)

SAEs uncover interpretable features



Templeton *et. al*, *https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html* (2024)

SAEs uncover interpretable features



Templeton *et. al*, *https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html* (2024)

SAEs uncover interpretable features

34M/31164353 Golden Gate Bridge

nd (that's the huge park right next to the Golden Gate bridge), perfect. But not all people can live in

e across the country in San Francisco, the Golden Gate bridge was protected at all times by a vigilant

ar coloring, it is often > compared to the Golden Gate Bridge in San Francisco, US. It was built by the

l to reach and if we were going to see the Golden Gate Bridge before sunset, we had to hit the road, so

t it?" " Because of what's above it." "The Golden Gate Bridge." "The fort fronts the anchorage and the

34M/9493533 Brain sciences

------ mjlee I really enjoy books on neuroscience that change the way I think about perception. Phanto

which brings together engineers and neuroscientists. If you like the intersection of analog, digital, h

ow managed to track it down and buy it again. The book is from the 1960s, but there are some really goo

interested in learning more about cognition, should I study neuroscience, or some other field, or is it

Consciousness and the Social Brain," by Graziano is a great place to start. ------ ozy I would want a

1M/887839 Monuments and popular tourist attractions

eautiful country, a bit eerily so. The blue lagoon is stunning to look at but too expensive to bathe in

nteresting things to visit in Egypt. The pyramids were older and less refined as this structure and the

st kind of beautiful." "What about the Alamo?" "Do people..." "Oh, the Alamo." "Yeah, it's a cool place

------ fvrghl I went to the Louvre in 2012, and I was able to walk up the Mona Lisa without a queue. I

you have to go to the big tourist attractions at least once like the San Diego Zoo and Sea World. ---

Templeton *et. al*, *https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html* (2024)

SAEs uncover interpretable features

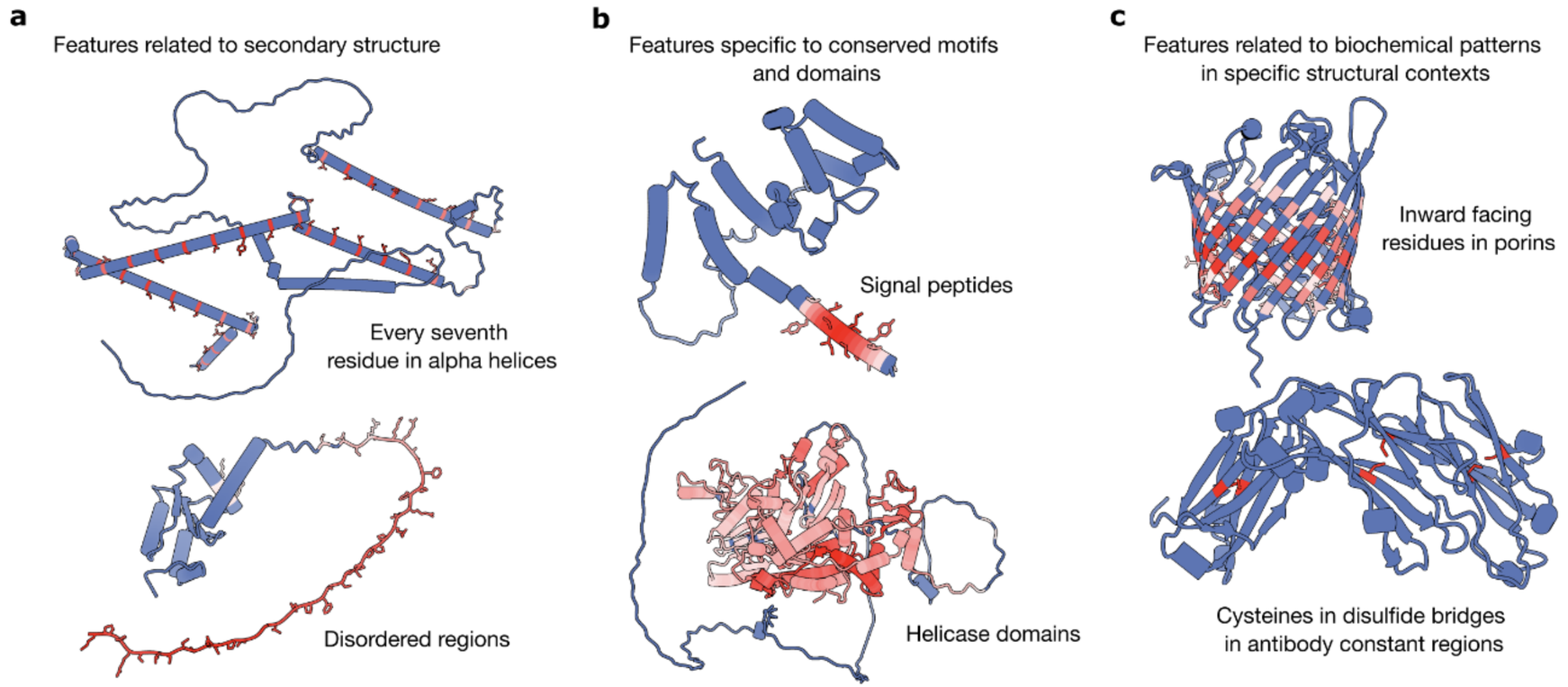


Figure 2. Examples of SAE features. We find features related to secondary structure (a), conserved motifs and domains (b), and biochemical patterns in specific structural contexts (c). The structures are colored according to activation (red: activation, blue: no activation).

Adams *et. al*, *ICML* (2025)

SAEs uncover interpretable features

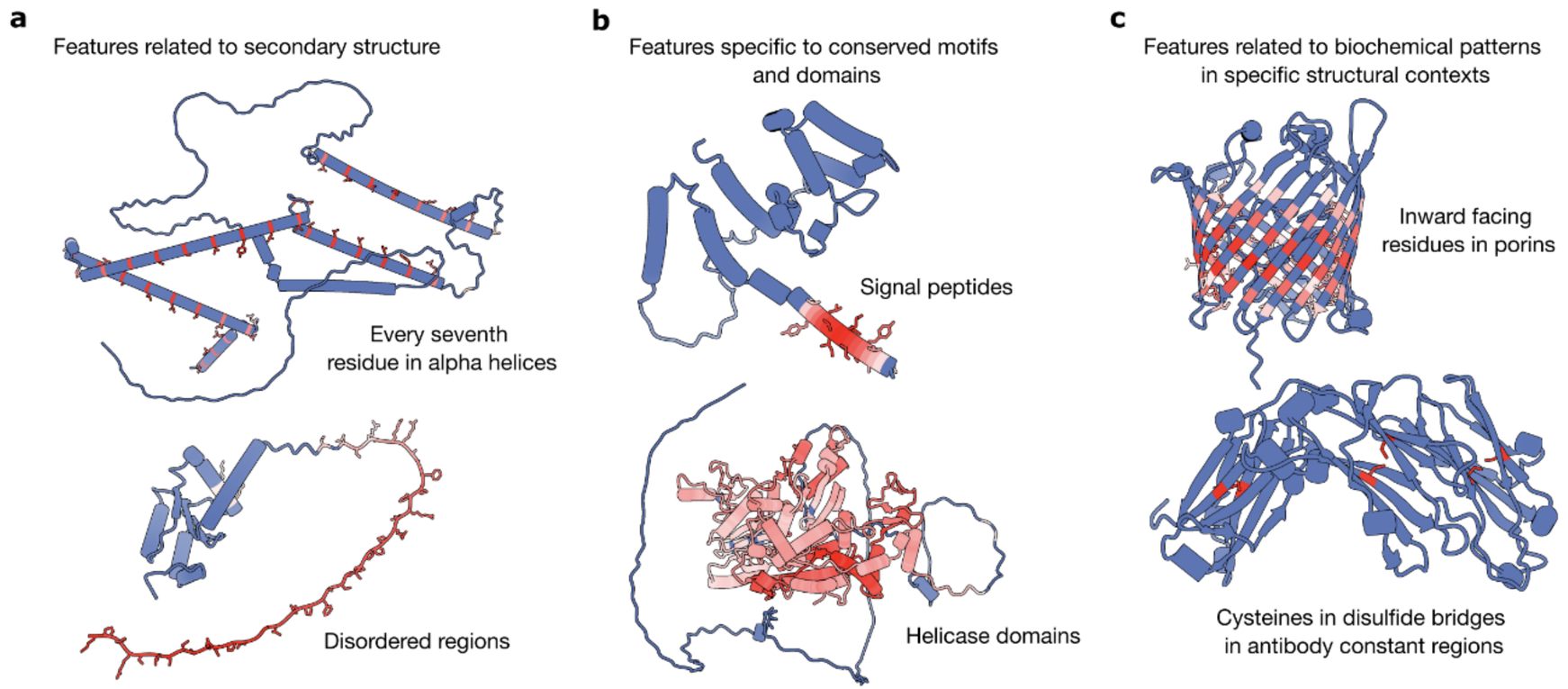
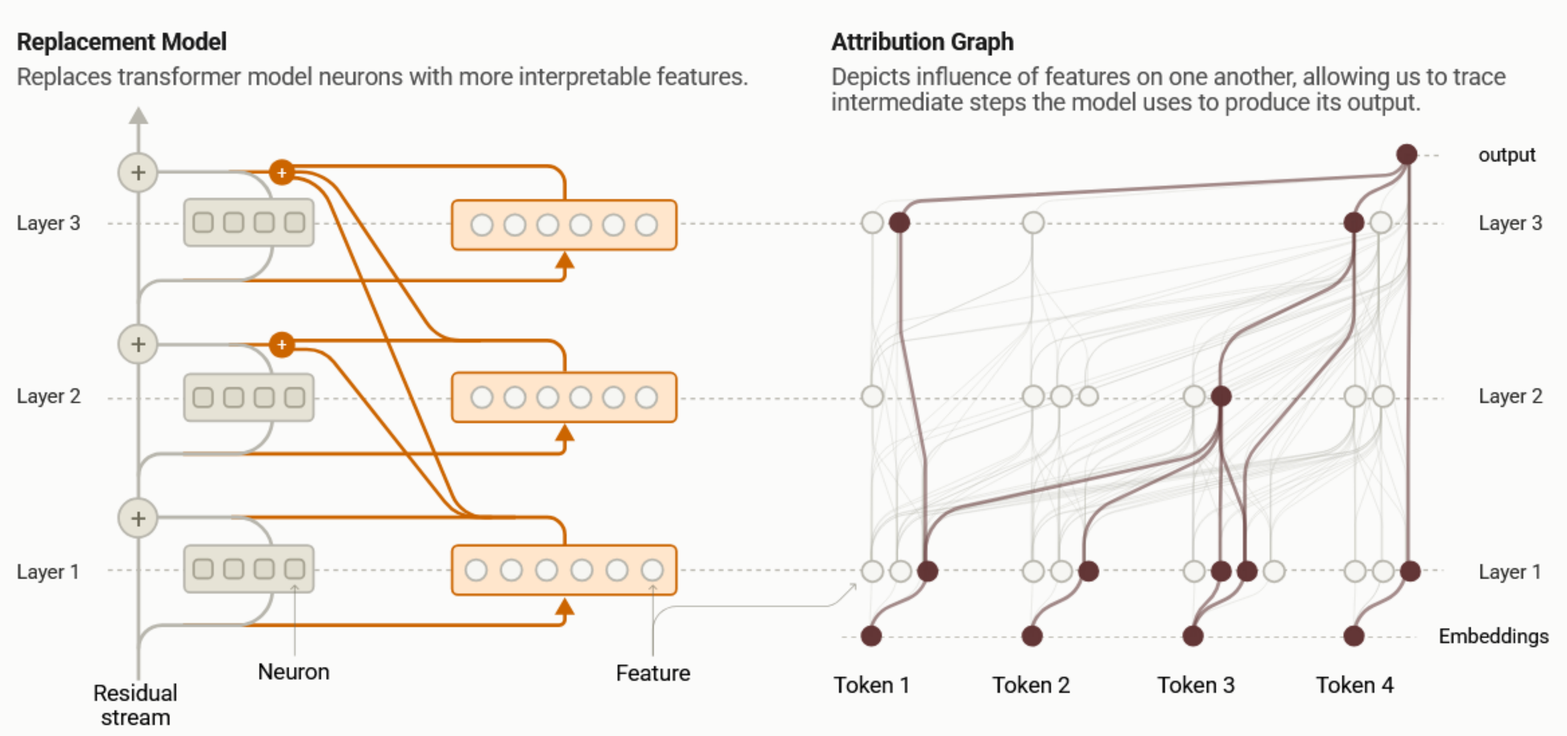


Figure 2. Examples of SAE features. We find features related to secondary structure (a), conserved motifs and domains (b), and biochemical patterns in specific structural contexts (c). The structures are colored according to activation (red: activation, blue: no activation).

https://interprot.com/#/sae-viz/SAE4096-L24/4000

Adams *et. al*, *ICML* (2025)

Cross-Layer Transcoders (CLTs): SAEs at every layer



Ameisen *et. al*, *https://transformer-circuits.pub/2025/attribution-graphs/methods.html* (2025)

Cross-Layer Transcoders (CLTs): SAEs at every layer

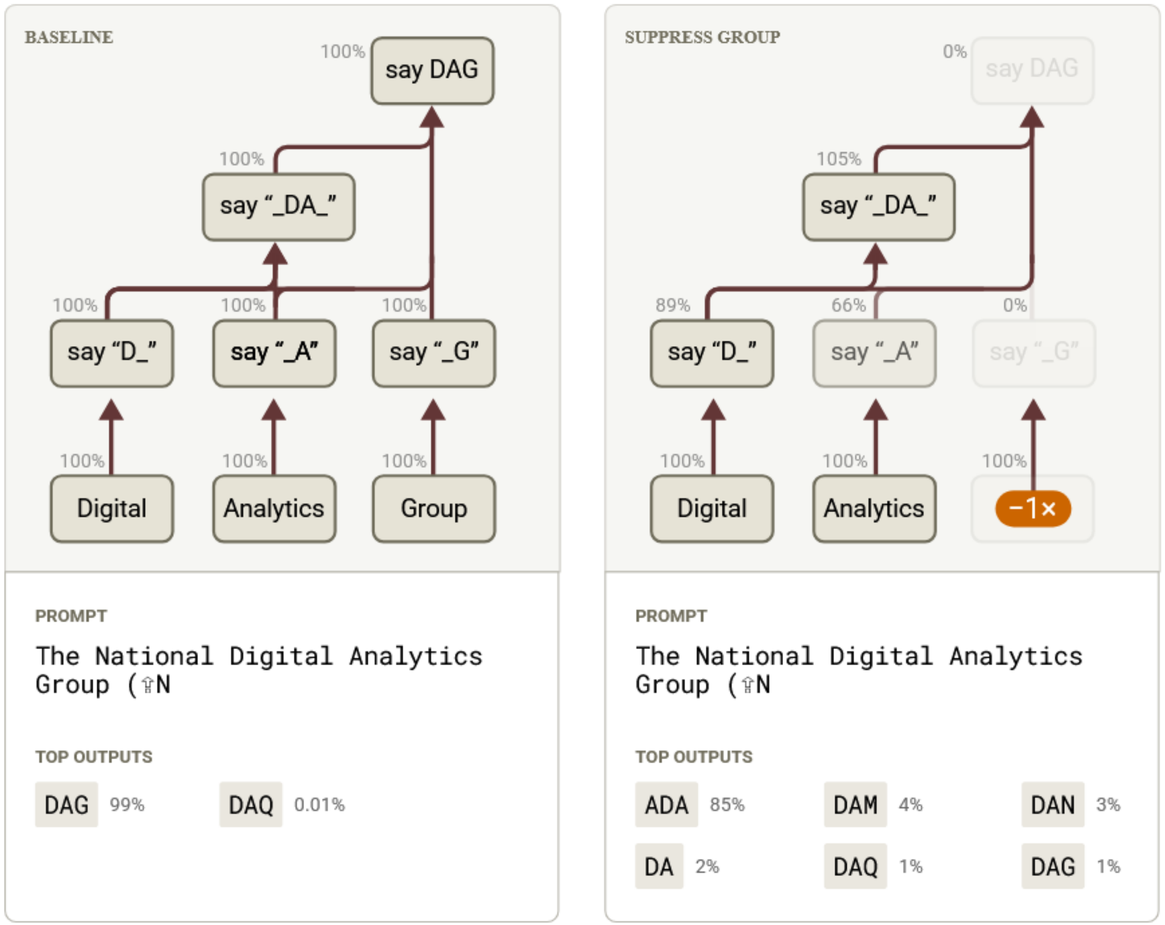


Figure 10: Suppressing the word "Group" in the fictional organization's name causes 18L to output other acronyms with "DA" in them.

Ameisen *et. al*, *https://transformer-circuits.pub/2025/attribution-graphs/methods.html* (2025)

Cross-Layer Transcoders (CLTs): SAEs at every layer

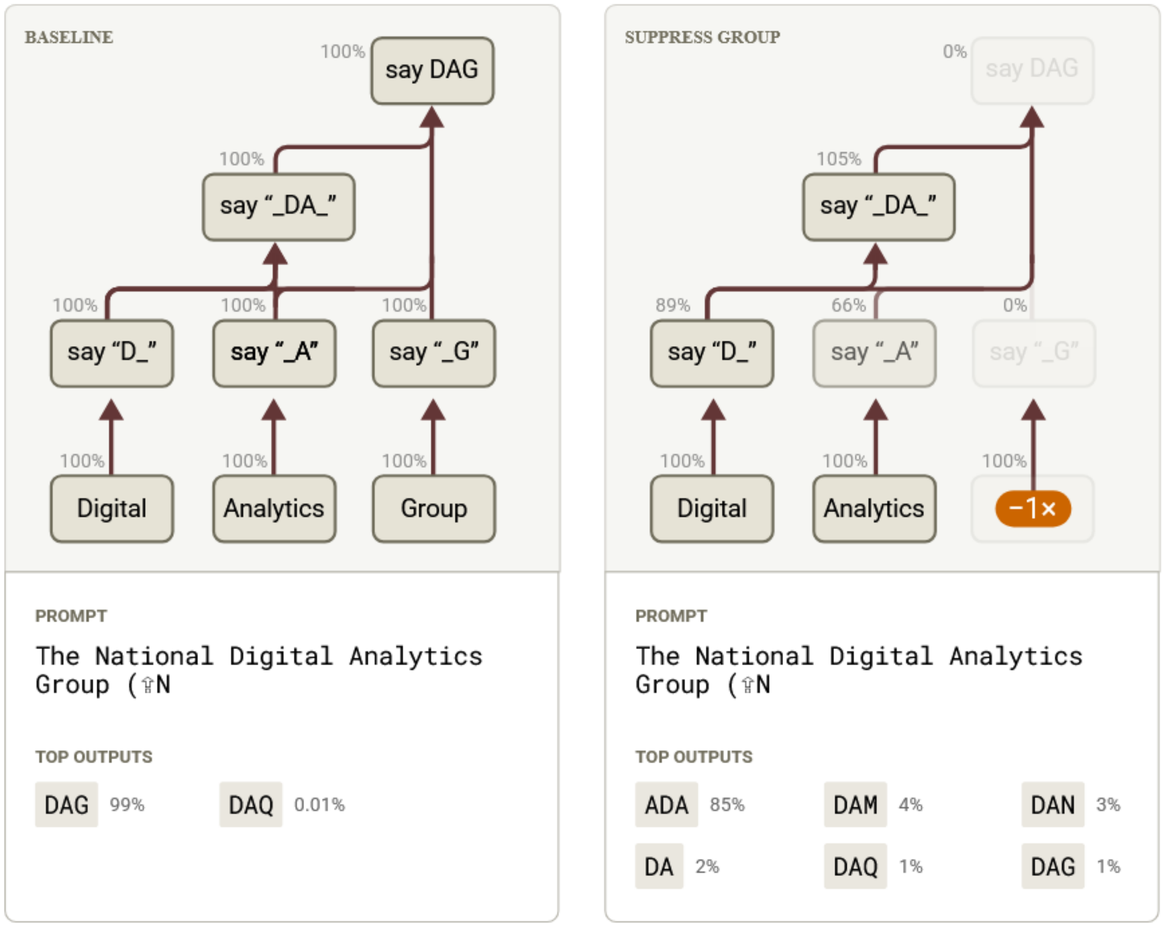


Figure 10: Suppressing the word "Group" in the fictional organization's name causes 18L to output other acronyms with "DA" in them.

https://www.neuronpedia.org/gemma-2-2b/graph

Ameisen *et. al*, *https://transformer-circuits.pub/2025/attribution-graphs/methods.html* (2025)

limitations of mechanistic interpretability

- **Intractability:** Explaining a toy model vs. explaining a SoTA LLM

limitations of mechanistic interpretability

- **Intractability:** Explaining a toy model vs. explaining a SoTA LLM
- **Lack of ground-truth**: what defines an "interpretable" feature? most interpretable features are scored by yet another black-box model (LLMs)

limitations of mechanistic interpretability

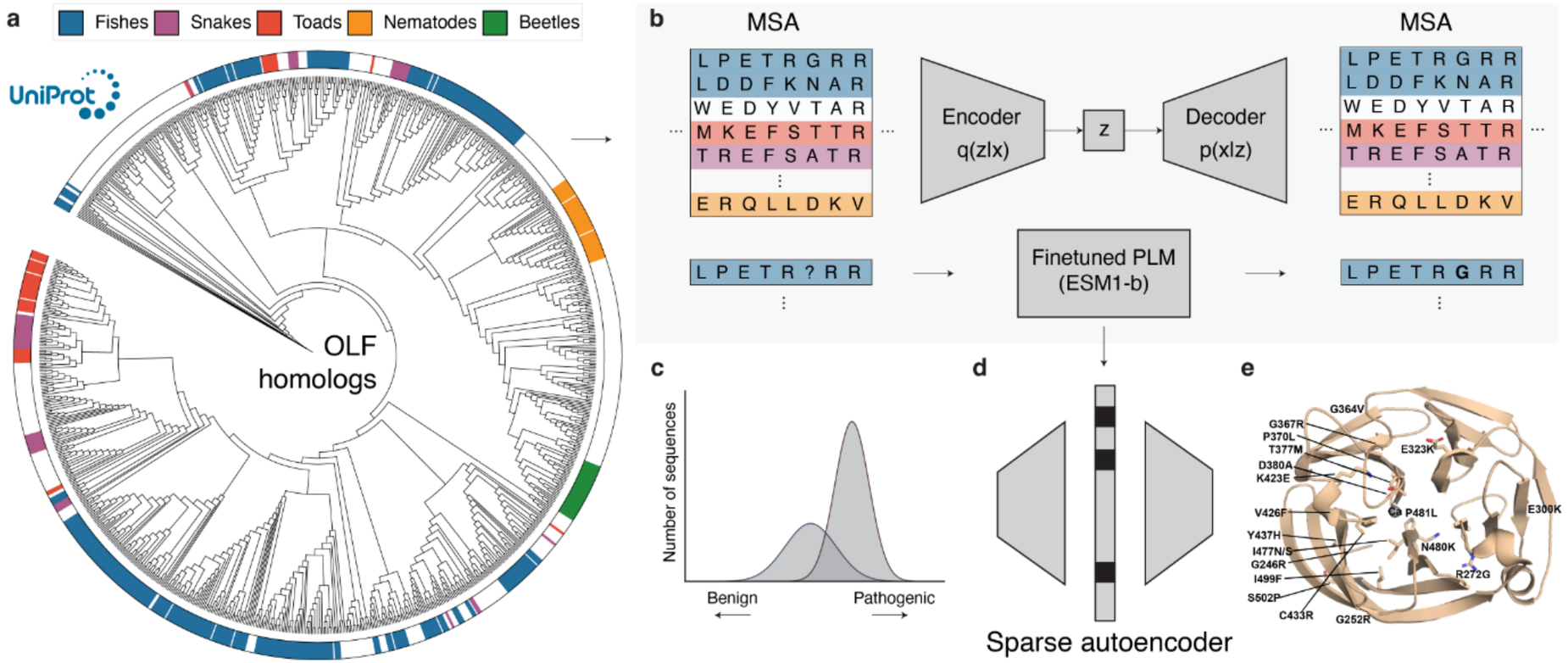
- **Intractability:** Explaining a toy model vs. explaining a SoTA LLM
- **Lack of ground-truth**: what defines an "interpretable" feature? most interpretable features are scored by yet another black-box model (LLMs)
- **Correlation vs. causation**: how do you prove a direct, causal mechanism rather than a strong correlation?

limitations of mechanistic interpretability

- **Intractability:** Explaining a toy model vs. explaining a SoTA LLM
- **Lack of ground-truth**: what defines an "interpretable" feature? most interpretable features are scored by yet another black-box model (LLMs)
- **Correlation vs. causation**: how do you prove a direct, causal mechanism rather than a strong correlation?

is mechanistic interpretability a **gimmick**? or can we use it to solve **real-world problems**?

GOLF: combining gen AI and interpretability for glaucoma

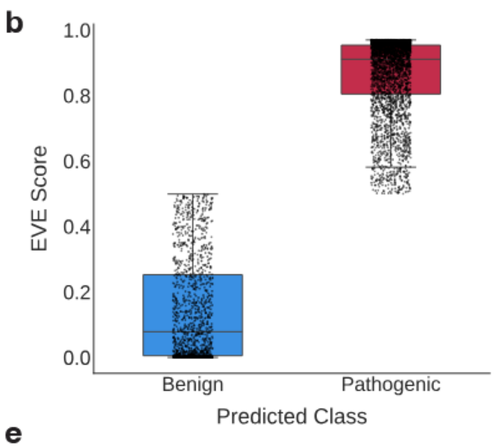


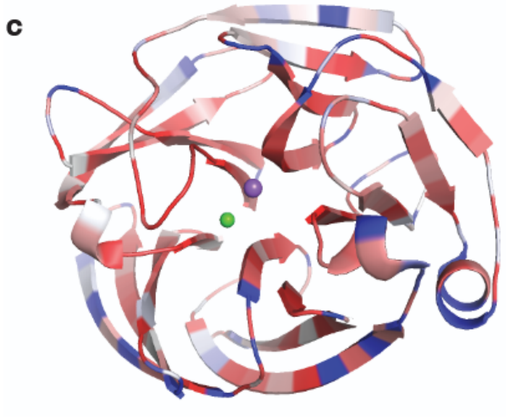
Walton, **Tsui**, Fogel, Huard, Chagas, Lieberman, Aghazadeh, *MLCB* (2025)

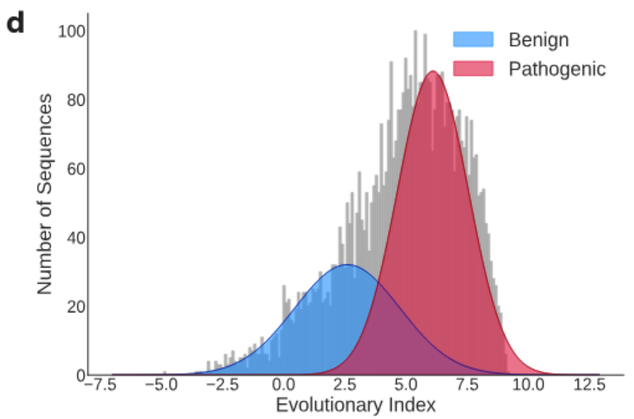
GOLF: combining gen AI and interpretability for glaucoma

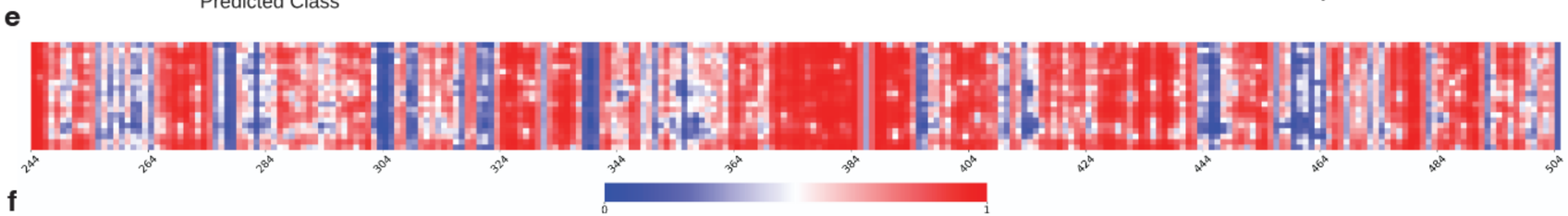
a

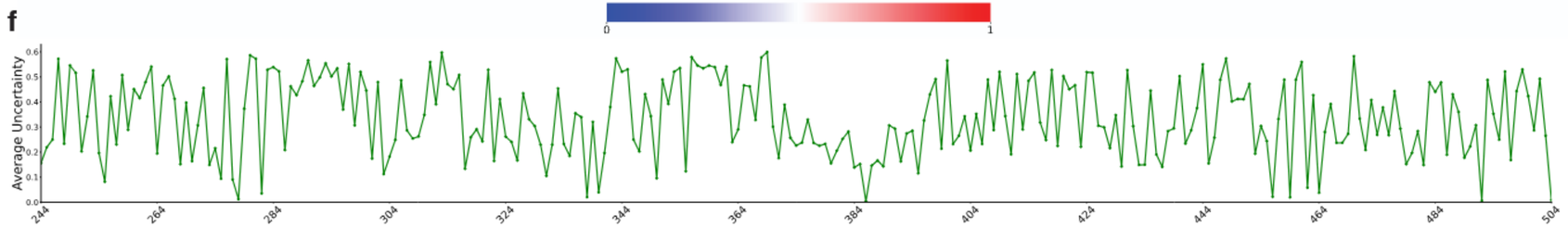
Mutation / Model	Thr 293 Lys	Val 329 Met	Glu 352 Lys	Thr 353 Ile	Lys 398 Arg	Ala 445 Val	Lys 500 Arg	Cys 245 Tyr	Gly 246 Arg	Gly 252 Arg	Arg 272 Gly	Glu 300 Lys	Glu 323 Lys	Gly 364 Val	Gly 367 Arg	Pro 370 Leu	Thr 377 Met	Thr 377 Lys	Asp 380 Ala	Lys 423 Glu	Val 426 Phe	Cys 433 Arg	Tyr 437 His	Ile 477 Asn	Ile 477 Ser	Tyr 479 His	Asn 480 Lys	Pro 481 Leu	Pro 481 Thr	Ile 499 Phe	Ile 499 Ser	Ser 502 Pro	#
EVE	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	31/32
ESM-1b	✓	✓	✓	✓	✓	✓	✓	×	×	×	×	×	×	✓	×	✓	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	22/32
Alpha Mis.	✓	✓	✓	✓	✓	✓	✓	✓	×	×	×	×	✓	×	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×	×	✓	×	22/32





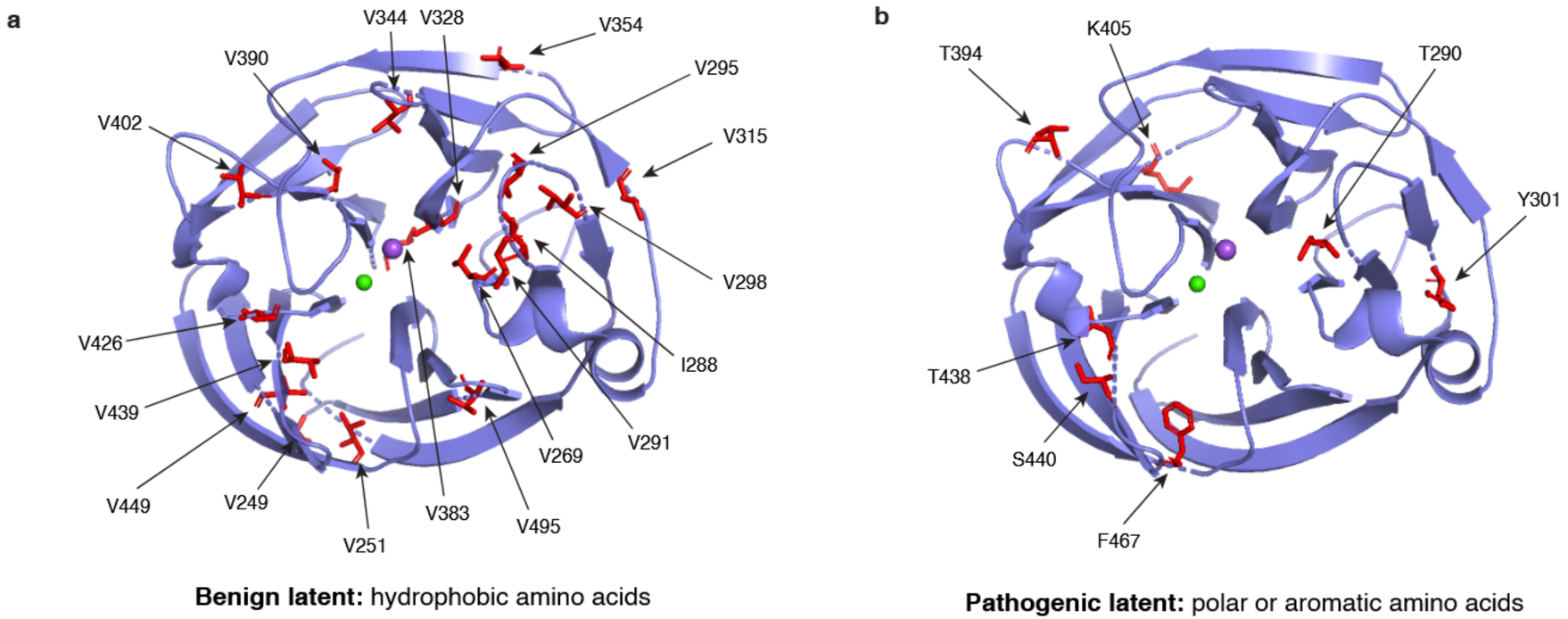






Walton, **Tsui**, Fogel, Huard, Chagas, Lieberman, Aghazadeh, *MLCB* (2025)

GOLF: combining gen AI and interpretability for glaucoma



Benign latent: hydrophobic amino acids

Pathogenic latent: polar or aromatic amino acids

Figure 3: **Mechanistic interpretability of the disease prediction models.** Visualization of the latent variables associated with **a**, benign and **b**, pathogenic predictions. Benign latent variables favored hydrophobic residues, while pathogenic latent variables favored polar or aromatic residues.

Walton, **Tsui**, Fogel, Huard, Chagas, Lieberman, Aghazadeh, *MLCB* (2025)

SAEs for low-resource protein function prediction and design

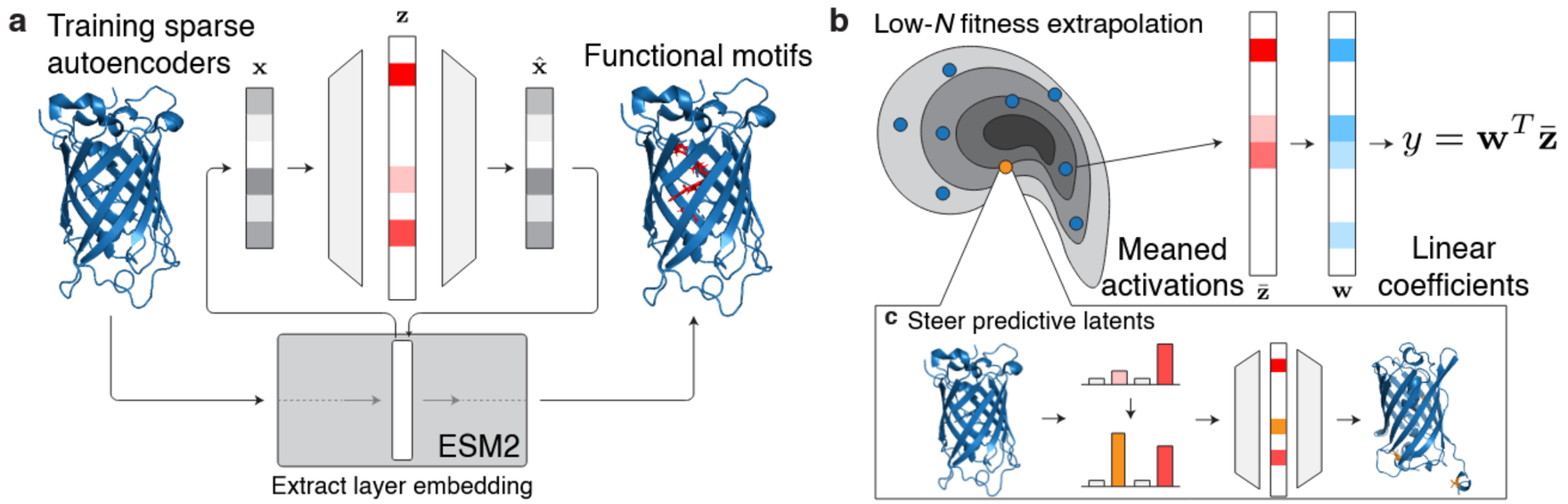
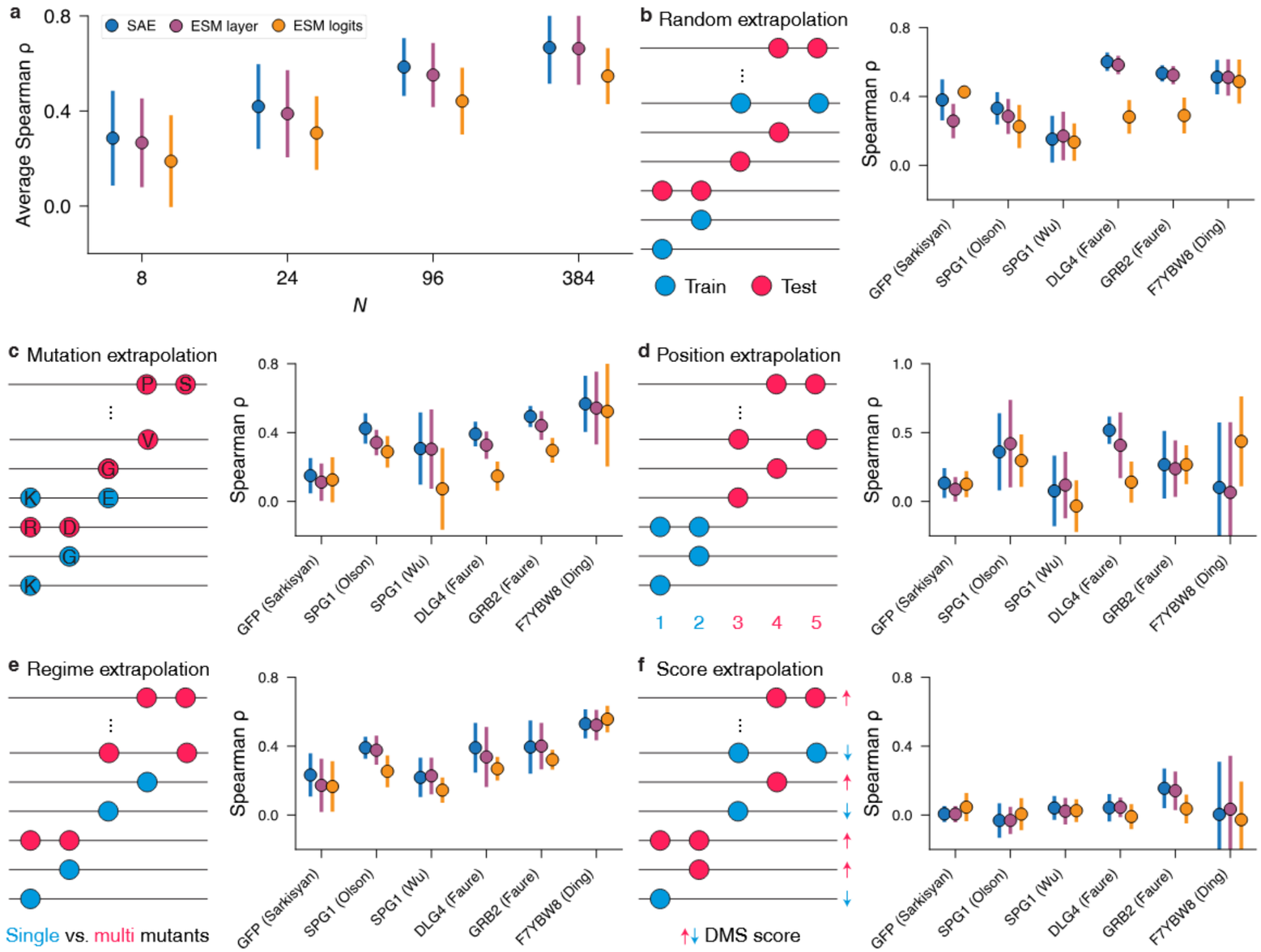


Figure 1: **Overview of downstream low-N tasks for SAEs.** **a**, We train SAEs on the layer embeddings of ESM2. By projecting the model embedding $\mathbf{x}$ to the latent representation $\mathbf{z}$, and reconstructing the model embedding as $\hat{\mathbf{x}}$, the activations in $\mathbf{z}$ correspond to specific biological motifs. **b**, In low-N fitness extrapolation, a linear probe is trained on top of the SAE's latent space to predict protein fitness from N many training sequences. **c**, Using the learned linear probe weights, we steer predictive latents to design highly-functional variants.

Tsui, Talreja, Aghazadeh, *NeurIPS AI4Science* (2025)

SAEs for low-resource protein function prediction and design



Tsui, Talreja, Aghazadeh, *NeurIPS AI4Science* (2025)

SAEs for low-resource protein function prediction and design

Table 3: Protein engineering results using $N = 24$ training sequences. All variants were constrained to a maximum of five mutations away from the wild type.

Method	DMS	Mean fitness ↑	Max fitness ↑	Top 10% fitness ↑	Top 20% fitness ↑
SAE	GFP_AEQVI_Sarkisyan	**3.49** ± 0.44	**3.87**	**3.75** ± 0.08	3.71 ± 0.07
	SPG1_STRSG_Olson	**2.75** ± 1.29	**4.53**	**4.47** ± 0.04	**4.29** ± 0.24
	SPG1_STRSG_Wu	**0.67** ± 0.94	**3.89**	**2.70** ± 0.79	**2.18** ± 0.76
	DLG4_HUMAN_Faure	**0.39** ± 0.22	**0.68**	**0.66** ± 0.02	**0.62** ± 0.05
	GRB2_HUMAN_Faure	**-0.10** ± 0.48	**0.67**	**0.59** ± 0.07	**0.49** ± 0.12
	F7YBW8_MESOW_Ding	0.81 ± 0.33	1.16	**1.15** ± 0.01	**1.13** ± 0.03
ESM layer	GFP_AEQVI_Sarkisyan	3.29 ± 0.66	3.72	3.71 ± 0.01	3.70 ± 0.01
	SPG1_STRSG_Olson	0.29 ± 1.95	3.19	2.74 ± 0.35	2.44 ± 0.39
	SPG1_STRSG_Wu	0.08 ± 0.30	1.69	0.81 ± 0.63	0.41 ± 0.60
	DLG4_HUMAN_Faure	-0.10 ± 0.41	0.63	0.45 ± 0.14	0.36 ± 0.13
	GRB2_HUMAN_Faure	-0.40 ± 0.39	0.30	0.24 ± 0.05	0.17 ± 0.10
	F7YBW8_MESOW_Ding	**1.06** ± 0.10	**1.16**	1.15 ± 0.02	1.12 ± 0.03
ESM logits	GFP_AEQVI_Sarkisyan	3.13 ± 0.86	3.76	3.73 ± 0.02	**3.72** ± 0.02
	SPG1_STRSG_Olson	-1.11 ± 2.21	2.27	2.05 ± 0.42	1.56 ± 0.60
	SPG1_STRSG_Wu	0.15 ± 0.37	1.69	1.13 ± 0.33	0.76 ± 0.50
	DLG4_HUMAN_Faure	-0.15 ± 0.38	0.53	0.36 ± 0.14	0.27 ± 0.13
	GRB2_HUMAN_Faure	-0.26 ± 0.44	0.58	0.44 ± 0.09	0.32 ± 0.14
	F7YBW8_MESOW_Ding	1.05 ± 0.06	1.12	1.11 ± 0.01	1.11 ± 0.01
Random	GFP_AEQVI_Sarkisyan	3.36 ± 0.70	3.75	3.72 ± 0.02	3.70 ± 0.02
	SPG1_STRSG_Olson	-1.25 ± 2.63	3.05	2.40 ± 0.49	1.99 ± 0.55
	SPG1_STRSG_Wu	0.33 ± 0.76	3.61	2.27 ± 0.82	1.36 ± 1.11
	DLG4_HUMAN_Faure	-0.34 ± 0.43	0.35	0.32 ± 0.04	0.24 ± 0.10
	GRB2_HUMAN_Faure	-0.96 ± 0.38	0.16	-0.13 ± 0.18	-0.30 ± 0.22
	F7YBW8_MESOW_Ding	0.66 ± 0.37	1.16	1.11 ± 0.03	1.06 ± 0.06

Tsui, Talreja, Aghazadeh, *NeurIPS AI4Science* (2025)

SAEs for low-resource protein function prediction and design

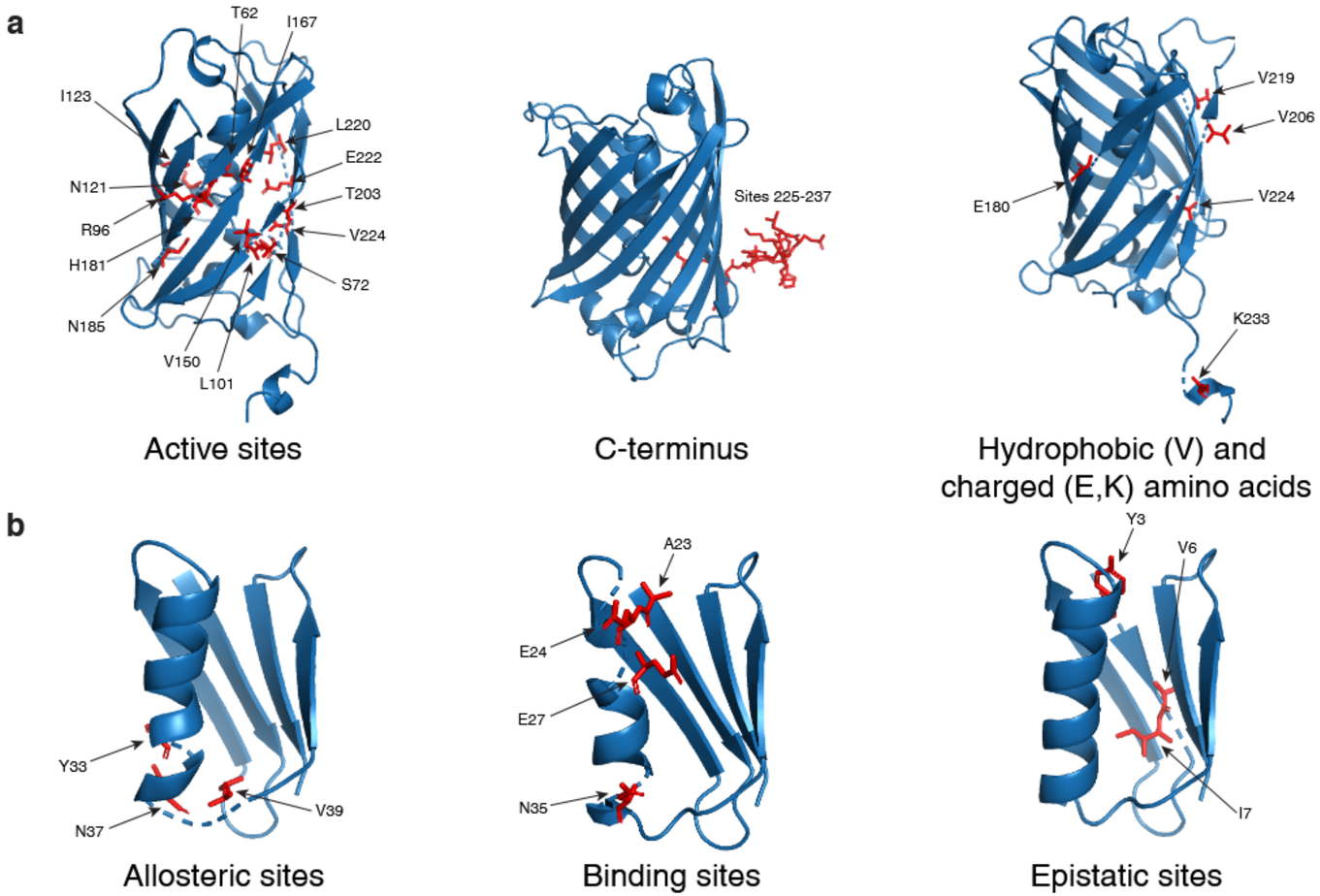


Figure 3: **Analysis of the top-performing steered variants. a**, Our analysis of the top-performing GFP variants revealed that steering activated latent features corresponding to key biological motifs, including active site amino acids, the C-terminus, and hydrophobic and charged amino acids. **b**, GB1 variants activated latent features associated with allosteric, binding, and epistatic sites.

Tsui, Talreja, Aghazadeh, *NeurIPS AI4Science* (2025)

summary

Shapley explanations

- Attributes features and interactions of features to **model predictions**
- Mathematically grounded
- Computationally expensive

Mechanistic interpretability

- Explores the **underlying mechanisms** of neural networks
- Uncovers features at scale in LLMs
- Correlation vs. causation, lack of validation

possible future research directions

Shapley explanations

- Developing novel algorithms to scale Shapley explanations

possible future research directions

Shapley explanations

- Developing novel algorithms to scale Shapley explanations
- Applying Shapley to under-explored domains (vision, biology)

possible future research directions

Shapley explanations

- Developing novel algorithms to scale Shapley explanations
- Applying Shapley to under-explored domains (vision, biology)
- Leveraging Shapley to create better data pipelines/ML architectures/validation

possible future research directions

Shapley explanations

- Developing novel algorithms to scale Shapley explanations
- Applying Shapley to under-explored domains (vision, biology)
- Leveraging Shapley to create better data pipelines/ML architectures/validation

Mechanistic interpretability

- More principled feature validation methods

possible future research directions

Shapley explanations

- Developing novel algorithms to scale Shapley explanations
- Applying Shapley to under-explored domains (vision, biology)
- Leveraging Shapley to create better data pipelines/ML architectures/validation

Mechanistic interpretability

- More principled feature validation methods
- Convincing skeptical communities, such as in the sciences, to adopt techniques

possible future research directions

Shapley explanations

- Developing novel algorithms to scale Shapley explanations
- Applying Shapley to under-explored domains (vision, biology)
- Leveraging Shapley to create better data pipelines/ML architectures/validation

Mechanistic interpretability

- More principled feature validation methods
- Convincing skeptical communities, such as in the sciences, to adopt techniques
- Leverage mechanistic interpretability to spot AI vulnerabilities

thank you!

Connect with me!