The role of interpretability in Protein Language Models

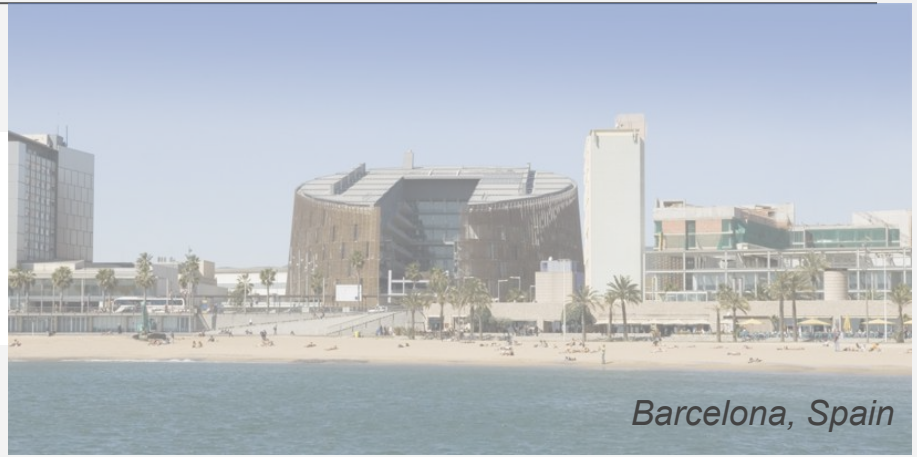
Alex Vicente
16/06/2026

About the speaker

Presenting today: Alex Vicente-Sola

ML Research Scientist, AI PhD

Centre for Genomic Regulation



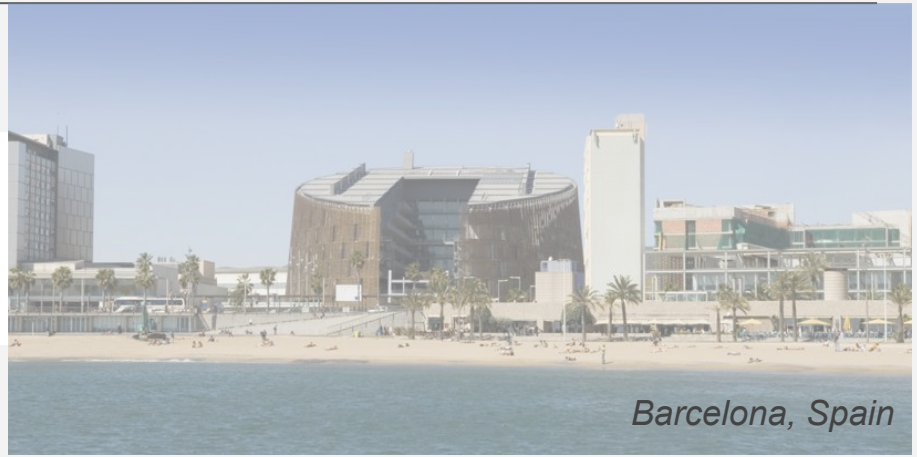
Barcelona, Spain

About the speaker

Presenting today: Alex Vicente-Sola

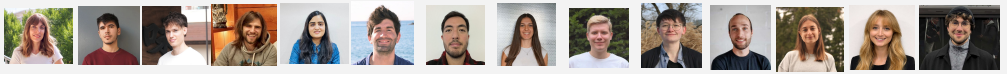
ML Research Scientist, AI PhD

Centre for Genomic Regulation



Our Lab:

AI for Protein Design - *Noelia Ferruz Lab*

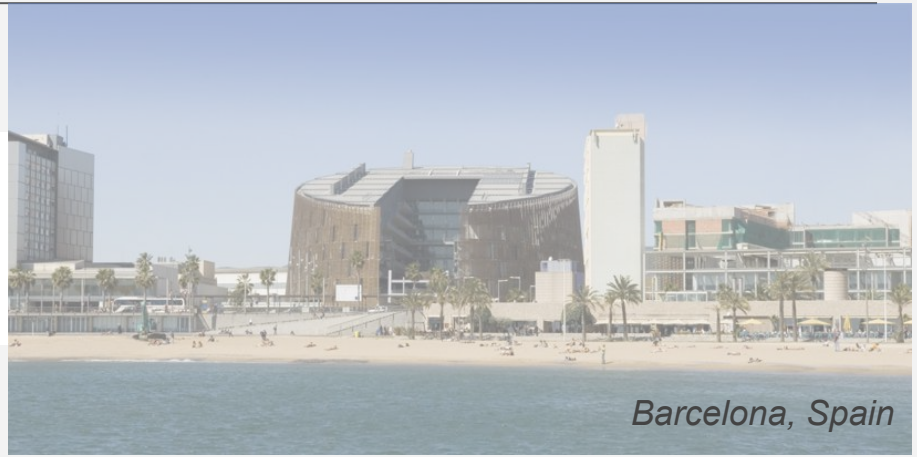


About the speaker

Presenting today: Alex Vicente-Sola

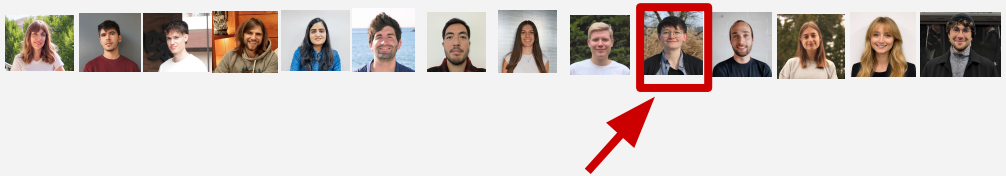
ML Research Scientist, AI PhD

Centre for Genomic Regulation



Our Lab:

AI for Protein Design - *Noelia Ferruz Lab*

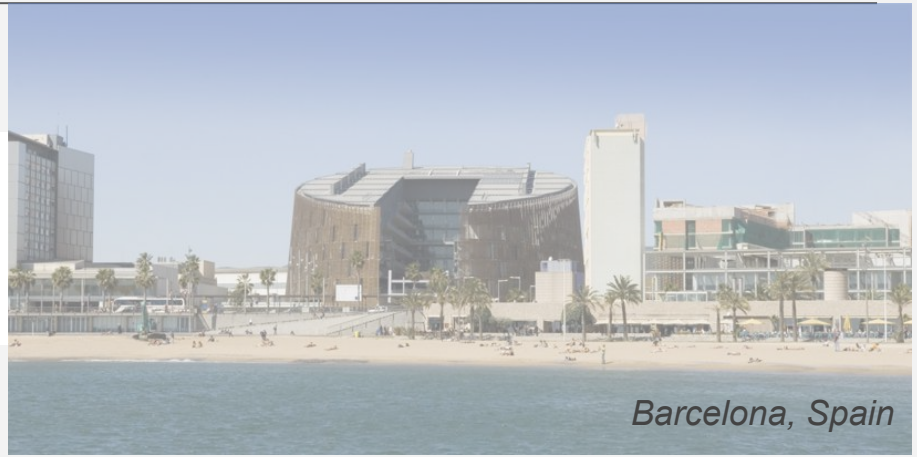


About the speaker

Presenting today: Alex Vicente-Sola

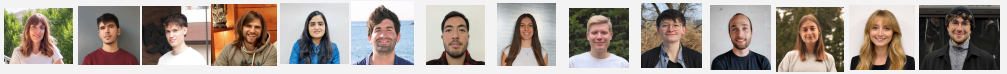
ML Research Scientist, AI PhD

Centre for Genomic Regulation



Our Lab:

AI for Protein Design - *Noelia Ferruz Lab*



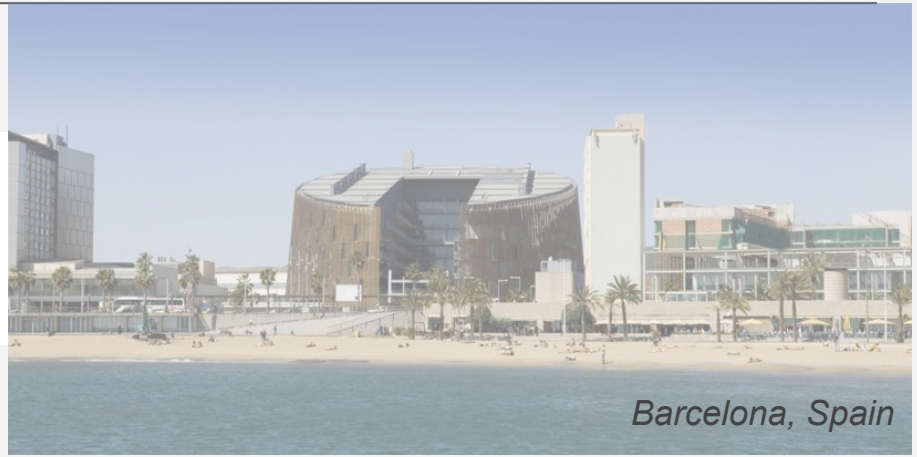
Design of custom-tailored proteins

About the speaker

Presenting today: Alex Vicente-Sola

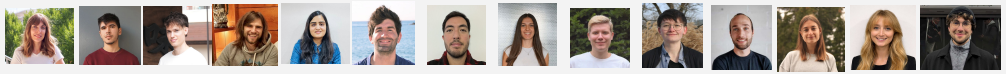
ML Research Scientist, AI PhD

Centre for Genomic Regulation

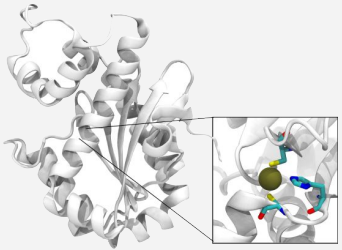


Our Lab:

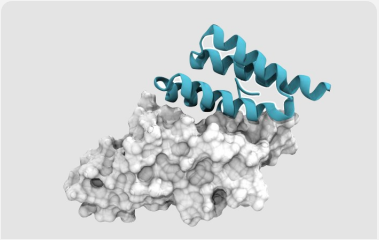
AI for Protein Design - *Noelia Ferruz Lab*



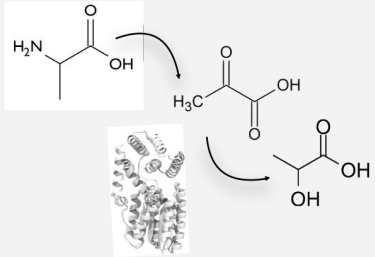
Design of custom-tailored proteins



CO2 capture



Immune checkpoints



Enzymatic activity

AI for Protein Design - *Noelia Ferruz Lab*

Mission: design of custom-tailored proteins

AI for Protein Design - *Noelia Ferruz Lab*

Mission: design of custom-tailored proteins

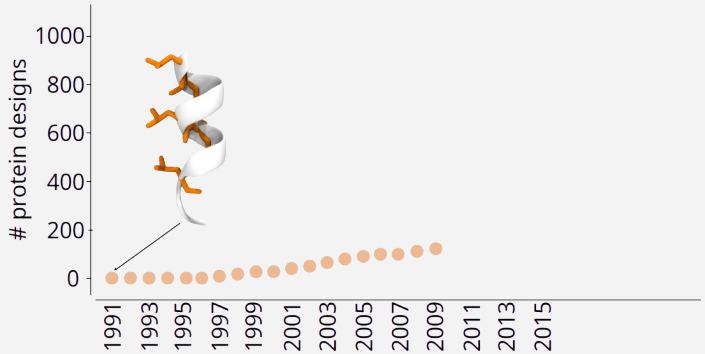
How: *AI-based design*

AI for Protein Design - *Noelia Ferruz Lab*

Mission: design of custom-tailored proteins

How: *AI-based design*

Rational Approaches

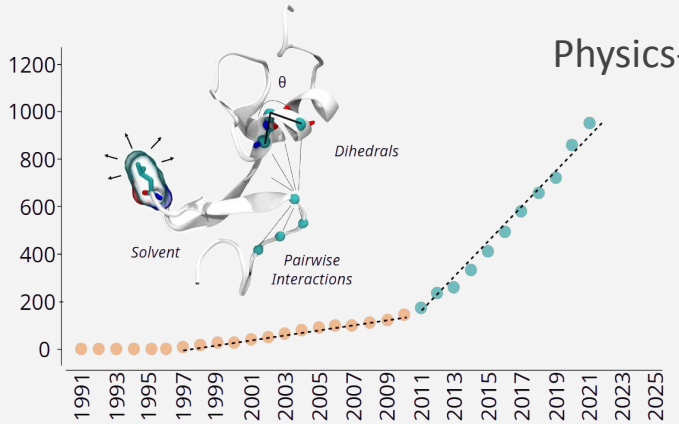


Chronowska et al. *The Protein Design Archive (PDA): insights from 40 years of protein design* (2025) Nature Biotechnology

AI for Protein Design - *Noelia Ferruz Lab*

Mission: design of custom-tailored proteins

How: *AI-based design*



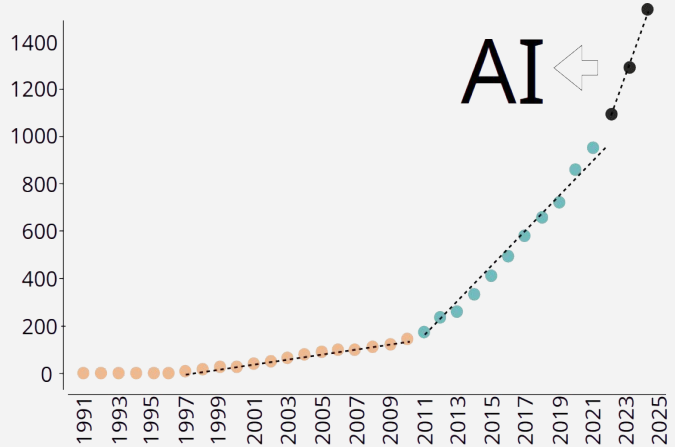
Physics-based Approaches

Chronowska et al. *The Protein Design Archive (PDA): insights from 40 years of protein design* (2025) Nature Biotechnology

AI for Protein Design - *Noelia Ferruz Lab*

Mission: design of custom-tailored proteins

How: *AI-based design*

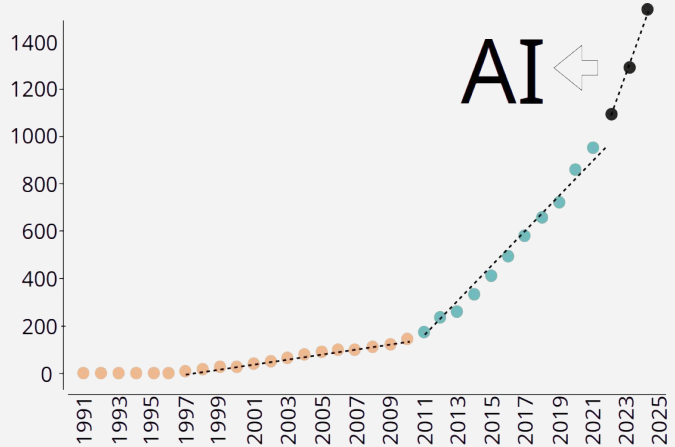


Chronowska et al. *The Protein Design Archive (PDA): insights from 40 years of protein design* (2025) Nature Biotechnology

AI for Protein Design - *Noelia Ferruz Lab*

Mission: design of custom-tailored proteins

How: *AI-based design*



Lab expertise: ***Protein Language Models***

Protein Language Models

Protein Language Models

Language Modeling with protein grammar

Protein Language Models

Language Modeling with protein grammar

Input: Amino acid sequence

*MDRAPQRQHRASR_*

Protein Language Models

Language Modeling with protein grammar

Input: Amino acid sequence

*MDRAPQRQHRASR_*

Protein Language Models

Language Modeling with protein grammar

Input: Amino acid sequence

*MDRAPQRQHRASR_*

Output: Next amino acid

MDRAPQRQHRASRE

Protein Language Models

Training: Predict the next amino acid given the previous

Input: Amino acid sequence

*MDRAPQRQHRASR_*

Output: Next amino acid

MDRAPQRQHRASRE

Protein Language Models

Deployment:

Protein Language Models

Deployment: Sequence completion

*MDRAPQ_*

MDRAPQ

Protein Language Models

Deployment: Sequence completion

*MDRAPQ_*

MDRAPQR

Protein Language Models

Deployment: Sequence completion

*MDRAPQR_*

MDRAPQR

Protein Language Models

Deployment: Sequence completion

*MDRAPQR_*

MDRAPQRQ

Protein Language Models

Deployment: Sequence completion

*MDRAPQRQ_*

MDRAPQRQ

Protein Language Models

Deployment: Sequence completion

*MDRAPQRQ_*

MDRAPQRQH

Protein Language Models

Deployment:
- Sequence completion
- Unconditional protein generation

Protein Language Models

Deployment:
- Sequence completion
- Unconditional protein generation

Protein Language Models

Deployment:
- Sequence completion
- Unconditional protein generation

M_

M

Protein Language Models

Deployment:
- Sequence completion
- Unconditional protein generation

M_

MD

Protein Language Models

Deployment:
- Sequence completion
- Unconditional protein generation

*MD_*

MD

Protein Language Models

Deployment:
- Sequence completion
- Unconditional protein generation

MD_

MDR

Protein Language Models

Deployment:
- Sequence completion
- Unconditional protein generation

*MDR_*

MDR

Protein Language Models

Deployment:
- Sequence completion
- Unconditional protein generation

*MDR_*

MDRA

Protein Language Models

Deployment:
- Sequence completion
- Unconditional protein generation
- Scoring

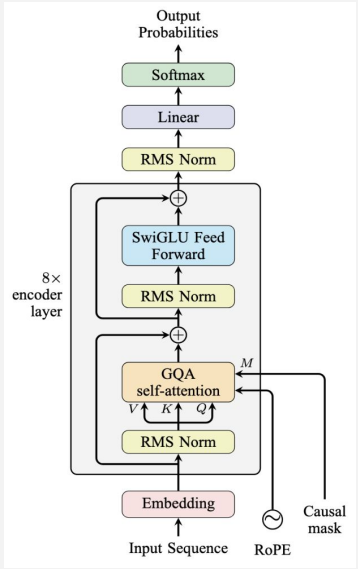
MDRAPQRQHRASRE…

	Position 1	Position 2	Position 3	Position 4	Position 5	Position 6
Token 1	0.18	0.07	0.13	0.06	0.24	**0.31**
Token 2	0.09	**0.42**	0.28	0.17	0.08	0.12
Token 3	0.61	0.11	**0.35**	0.09	0.14	0.16
Token 4	**0.47**	0.26	0.08	**0.53**	0.32	0.23
Token 5	0.15	0.14	0.16	0.15	**0.27**	0.18

Protein Language Models (pLMs)

Deployment:
- Sequence completion
- Unconditional protein generation
- Scoring

MDRAPQRQHRASRE…



	Position 1	Position 2	Position 3	Position 4	Position 5	Position 6
Token 1	0.18	0.07	0.13	0.06	0.24	**0.31**
Token 2	0.09	**0.42**	0.28	0.17	0.08	0.12
Token 3	0.61	0.11	**0.35**	0.09	0.14	0.16
Token 4	**0.47**	0.26	0.08	**0.53**	0.32	0.23
Token 5	0.15	0.14	0.16	0.15	**0.27**	0.18

Transformer

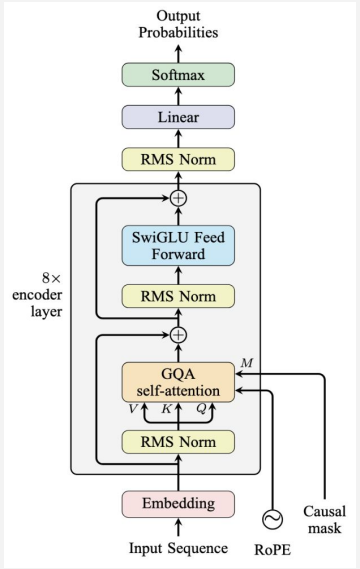
Decoder-only / Encoder-only / Encoder-Decoder

Protein Language Models (pLMs)

Deployment:
- Sequence completion
- Unconditional protein generation
- Scoring

Reusing their feature extraction → Embedding

MDRAPQRQHRASRE…

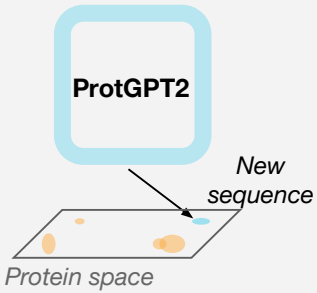


Transformer

Decoder-only / Encoder-only / Encoder-Decoder

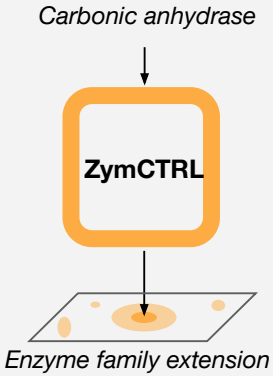
AI for Protein Design - *Noelia Ferruz Lab*

Protein Language Models developed by the lab

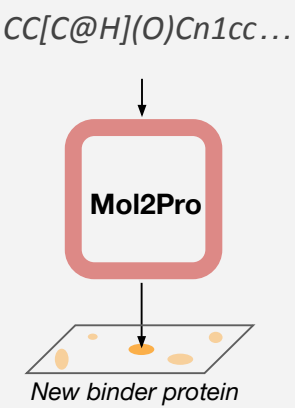


Ferruz N et al. ***ProtGPT2 Is a Deep Unsupervised Language Model for Protein Design.*** Nat. Commun., **13**, 4348 (2022)

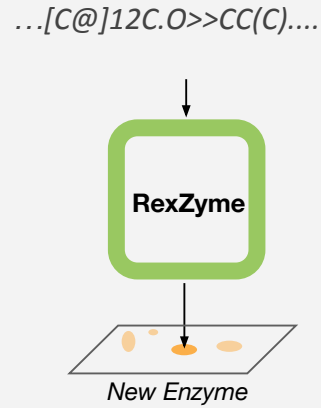
Version 3: Garibbo M et al. *ProtGPT3: an Open-source family of Promptable and Aligned Protein Language Models* BioRxiv (2026)



Munsamy G et al. ***Conditional language models enable the efficient design of proficient enzymes.*** (2024) BioRxiv



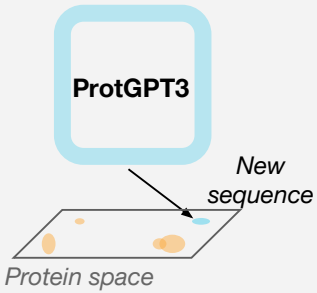
Vicente A et al. ***Generalise or Memorise? Benchmarking Ligand-Conditioned Protein Generation from Sequence-Only Data*** (2026) BioRxiv



Mimbrero-Pelegrí N et al. ***Translation Machines for the Design of Enzymes Tailored to Specific Reactions.*** (Manuscript in preparation)

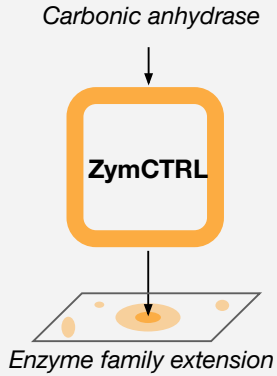
AI for Protein Design - *Noelia Ferruz Lab*

Protein Language Models developed by the lab

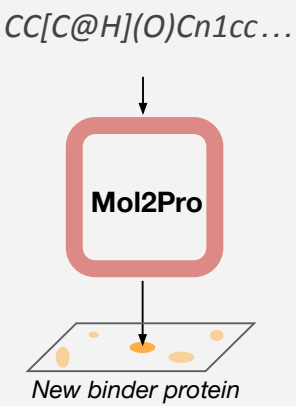


Ferruz N et al. ***ProtGPT2 Is a Deep Unsupervised Language Model for Protein Design.*** Nat. Commun., **13**, 4348 (2022)

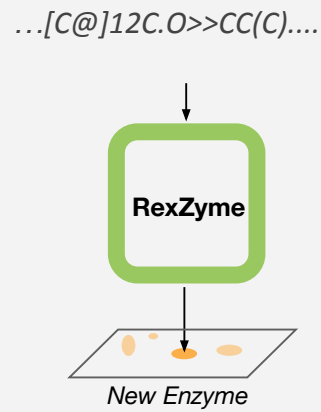
Version 3: Garibbo M et al. *ProtGPT3: an Open-source family of Promptable and Aligned Protein Language Models* BioRxiv (2026)



Munsamy G et al. ***Conditional language models enable the efficient design of proficient enzymes.*** (2024) BioRxiv https://doi.org/10.1101/2024.05.03.592223



Vicente A et al. ***Generalise or Memorise? Benchmarking Ligand-Conditioned Protein Generation from Sequence-Only Data*** (2026) BioRxiv



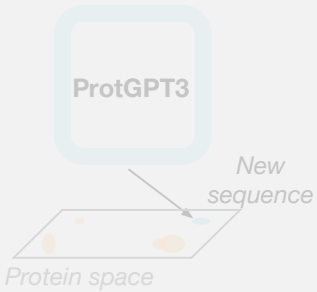
Mimbrero-Pelegrí N et al. ***Translation Machines for the Design of Enzymes Tailored to Specific Reactions.*** (Manuscript in preparation)

Also exploring other directions:

Structure-based, **Physics**-Based, and **multi-modal** design

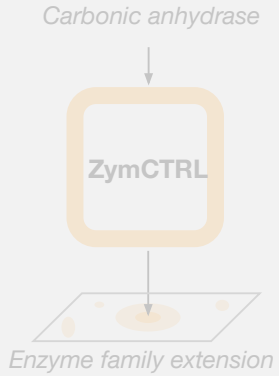
AI for Protein Design - *Noelia Ferruz Lab*

Protein Language Models developed by the lab

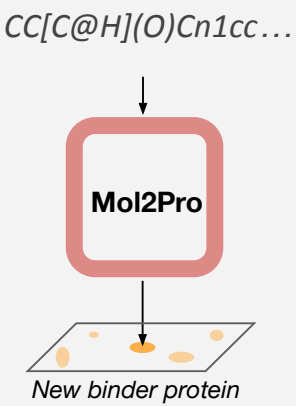


Garibbo M et al. *ProtGPT3: an Open-source family of Promptable and Aligned Protein Language Models* BioRxiv (2026)

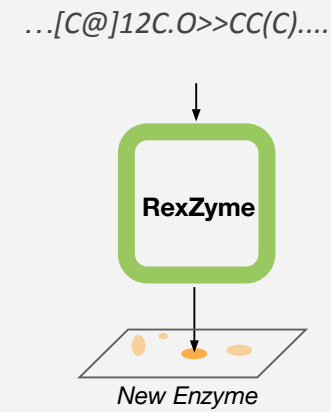
Ferruz N et al. *ProtGPT2 Is a Deep Unsupervised Language Model for Protein Design.* Nat. Commun., 13, 4348 (2022)



Munsamy G et al. *Conditional language models enable the efficient design of proficient enzymes.* (2024) BioRxiv https://doi.org/10.1101/2024.05.03.592223



Vicente A et al. *Generalise or Memorise? Benchmarking Ligand-Conditioned Protein Generation from Sequence-Only Data* (2026) BioRxiv



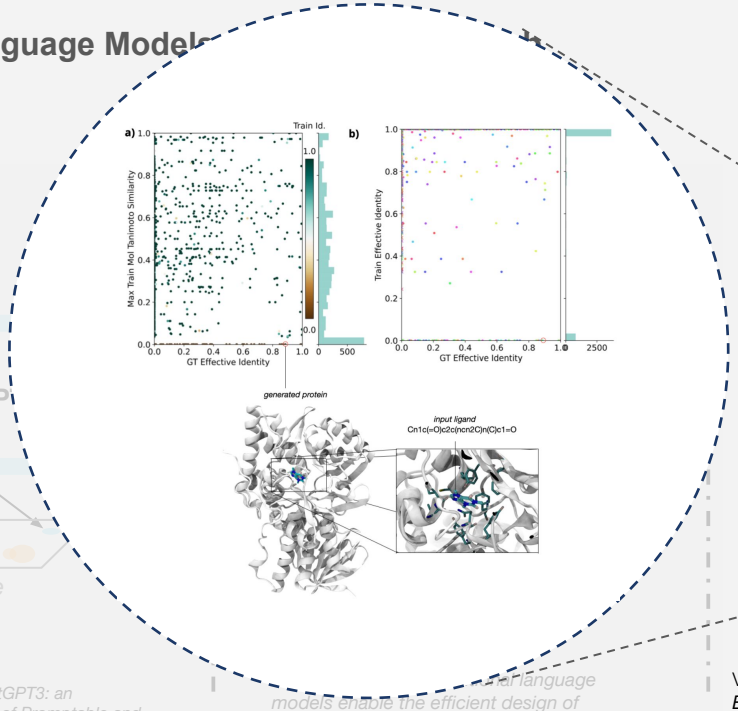
Mimbrero-Pelegrí N et al. *Translation Machines for the Design of Enzymes Tailored to Specific Reactions.* (Manuscript in preparation)

Also exploring other directions:

Structure-based, **Physics**-Based, and **multi-modal** design

AI for Protein Design - *Noelia Ferruz Lab*

Protein Language Models



ProtGP

Protein space

Garibbo M et al. *ProtGPT3: an Open-source family of Promptable and Aligned Protein Language Models* BioRxiv (2026)

Ferruz N et al. *ProtGPT2 Is a Deep Unsupervised Language Model for Protein Design.* Nat. Commun., 13, 4348 (2022)

…al language models enable the efficient design of proficient enzymes. (2024) BioRxiv https://doi.org/10.1101/2024.05.03.592223

CC[C@H](O)Cn1cc…

Mol2Pro

New binder protein

Vicente A et al. *Generalise or Memorise? Benchmarking Ligand-Conditioned Protein Generation from Sequence-Only Data* (2026) BioRxiv

…[C@]12C.O>>CC(C)….

RexZyme

New Enzyme

Mimbrero-Pelegrí N et al. *Translation Machines for the Design of Enzymes Tailored to Specific Reactions.* (Manuscript in preparation)

Also exploring other directions:

Structure-based, **Physics**-Based, and **multi-modal** design

AI for Protein Design - *Noelia Ferruz Lab*

Protein Language Models developed by the lab

ProtGPT3

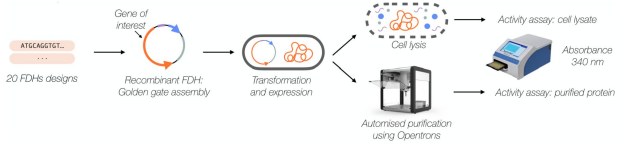
New sequence

Protein space

Carbonic

Zym

Enzyme family extens



	FDH	specific activity (mU/mg)
Wild Type	**CtFDH**	34.9 ± 4.1
	B2	72.7 ± 7.6
	C2	11 ± 0.4
	A3	32.1 ± 3.4
	B3	18.5 ± 2.9
	C3	29.4 ± 0.4

…[C@]12C.O>>CC(C)….

RexZyme

New Enzyme

Garibbo M et al. *ProtGPT3: an Open-source family of Promptable and Aligned Protein Language Models* BioRxiv (2026)

Ferruz N et al. *ProtGPT2 Is a Deep Unsupervised Language Model for Protein Design.* Nat. Commun., 13, 4348 (2022)

Munsamy G et al. *Conditional language models enable the efficient design of proficient enzymes.* (2024) BioRxiv https://doi.org/10.1101/2024.05.03.592223

Vidal A et al. *Generalise or Memorise? Benchmarking Ligand-Conditioned Protein Generation from Sequence-Only Data* (2026) BioRxiv

Mimbrero-Pelegrí N et al. *Translation Machines for the Design of Enzymes Tailored to Specific Reactions.* (Manuscript in preparation)

Also exploring other directions:

Structure-based, **Physics**-Based, and **multi-modal** design

AI Interpretability and Explainability

AI Interpretability and Explainability

Interpretability

Information on how the model works internally

Target: Developers

AI Interpretability and Explainability

Interpretability

Information on how the model works internally

Target: Developers

Explainability

Human-understandable explanation for a particular output

Target: End user

AI Interpretability and Explainability

Interpretability

Information on how the model works internally

Target: Developers

Explainability

Human-understandable explanation for a particular output

Target: End user

Uses for interpretability

AI Interpretability and Explainability

Interpretability

Information on how the model works internally

Target: Developers

Explainability

Human-understandable explanation for a particular output

Target: End user

Uses for interpretability

🔎 *Through the lens of:* ***"Toward the Explainability of Protein Language Models"***

Hunklinger A & Ferruz, Nature Machine Intelligence 2026

AI Interpretability and Explainability

Interpretability

Information on how the model works internally

Target: Developers

Explainability

Human-understandable explanation for a particular output

Target: End user

Uses for interpretability

🔎 *Through the lens of:* ***"Toward the Explainability of Protein Language Models"***

Hunklinger A & Ferruz, Nature Machine Intelligence 2026

Roles of interpretability

AI Interpretability and Explainability

Interpretability

Information on how the model works internally

Target: Developers

Explainability

Human-understandable explanation for a particular output

Target: End user

Uses for interpretability

🔎 *Through the lens of:* ***"Toward the Explainability of Protein Language Models"***

Hunklinger A & Ferruz, Nature Machine Intelligence 2026

Roles of interpretability

1. **Evaluator:** Validate patterns used by AI

AI Interpretability and Explainability

Interpretability

Information on how the model works internally

Target: Developers

Explainability

Human-understandable explanation for a particular output

Target: End user

Uses for interpretability

🔎 *Through the lens of:* ***"Toward the Explainability of Protein Language Models"***

Hunklinger A & Ferruz, Nature Machine Intelligence 2026

Roles of interpretability

1. **Evaluator:** Validate patterns used by AI (matching prev knowledge)

AI Interpretability and Explainability

Interpretability

Information on how the model works internally

Target: Developers

Explainability

Human-understandable explanation for a particular output

Target: End user

Uses for interpretability

🔎 *Through the lens of:* ***"Toward the Explainability of Protein Language Models"***

Hunklinger A & Ferruz, Nature Machine Intelligence 2026

Roles of interpretability

1. **Evaluator:** Validate patterns used by AI (matching prev knowledge)
2. **Engineer:** Optimise architecture / algorithm

AI Interpretability and Explainability

Interpretability

Information on how the model works internally

Target: Developers

Explainability

Human-understandable explanation for a particular output

Target: End user

Uses for interpretability

🔎 *Through the lens of:* ***"Toward the Explainability of Protein Language Models"***

Hunklinger A & Ferruz, Nature Machine Intelligence 2026

Roles of interpretability

1. **Evaluator:** Validate patterns used by AI (matching prev knowledge)
2. **Engineer:** Optimise architecture / algorithm
3. **Coach:** Modify inference

AI Interpretability and Explainability

Interpretability

Information on how the model works internally

Target: Developers

Explainability

Human-understandable explanation for a particular output

Target: End user

Uses for interpretability

🔎 *Through the lens of:* ***"Toward the Explainability of Protein Language Models"***

Hunklinger A & Ferruz, Nature Machine Intelligence 2026

Roles of interpretability

1. **Evaluator:** Validate patterns used by AI (matching prev knowledge)
2. **Engineer:** Optimise architecture / algorithm
3. **Coach:** Modify inference
4. **Multi-tasker:** Perform other tasks not trained for

AI Interpretability and Explainability

Interpretability

Information on how the model works internally

Target: Developers

Explainability

Human-understandable explanation for a particular output

Target: End user

Uses for interpretability

🔎 *Through the lens of:* ***"Toward the Explainability of Protein Language Models"***

Hunklinger A & Ferruz, Nature Machine Intelligence 2026

Roles of interpretability

1. **Evaluator:** Validate patterns used by AI (matching prev knowledge)
2. **Engineer:** Optimise architecture / algorithm
3. **Coach:** Modify inference
4. **Multi-tasker:** Perform other tasks not trained for
5. **Teacher:** Discover new knowledge

AI Interpretability and Explainability

Interpretability

Information on how the model works internally

Target: Developers

Explainability

Human-understandable explanation for a particular output

Target: End user

Uses for interpretability

🔎 *Through the lens of:* ***"Toward the Explainability of Protein Language Models"***

Hunklinger A & Ferruz, Nature Machine Intelligence 2026

Roles of interpretability

1. **Evaluator:** Validate patterns used by AI (matching prev knowledge)
2. **Engineer:** Optimise architecture / algorithm
3. **Coach:** Modify inference
4. **Multi-tasker:** Perform other tasks not trained for
5. **Teacher:** Discover new knowledge

These apply to all applications

We will focus on the current state for Protein Language Models

AI Interpretability and Explainability

Interpretability

Information on how the model works internally

Target: Developers

Explainability

Human-understandable explanation for a particular output

Target: End user

Uses for interpretability

🔎 *Through the lens of:* ***"Toward the Explainability of Protein Language Models"***

Hunklinger A & Ferruz, Nature Machine Intelligence 2026

Roles of interpretability

1. **Evaluator:** Validate patterns used by AI (matching prev knowledge)
2. **Engineer:** Optimise architecture / algorithm
3. **Coach:** Modify inference
4. **Multi-tasker:** Perform other tasks not trained for
5. **Teacher:** Discover new knowledge

Where are they applied

1. **Input** (parts)
2. **Input** (full) **- output pairs**
3. **Model**
4. **Dataset**

These apply to all applications

We will focus on the current state for Protein Language Models

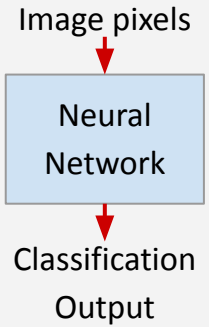
Input Interpretability

*How much the output depends on **each part** of the input*

Input Interpretability

*How much the output depends on **each part** of the input*

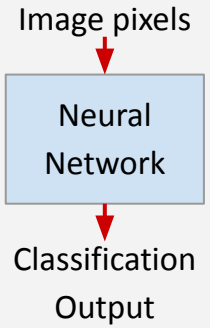
Example:

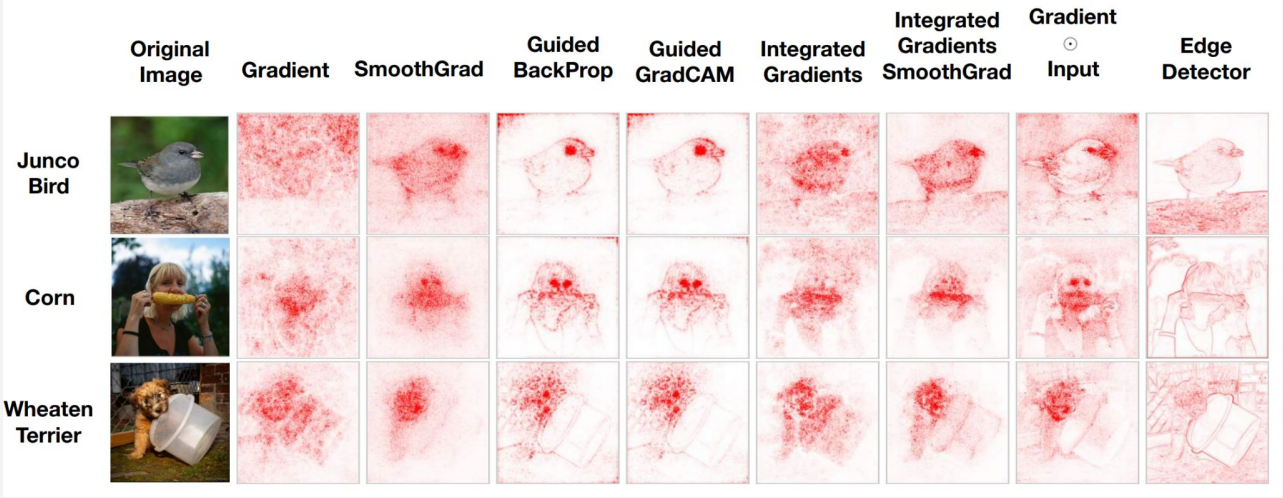


Input Interpretability

*How much the output depends on **each part** of the input*

Example:



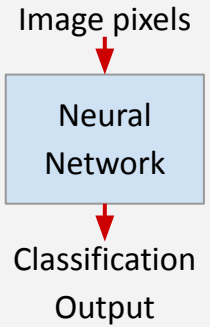


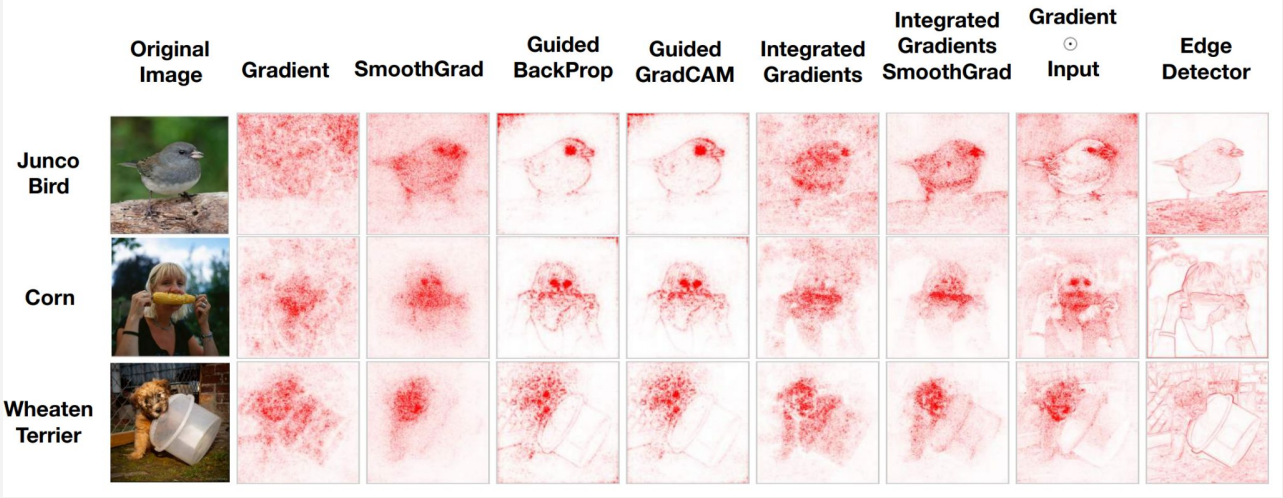
Adebayo, J. et al., *Sanity checks for saliency maps*. NeurIPS, 31. 2018

Input Interpretability

*How much the output depends on **each part** of the input*

Example:





Adebayo, J. et al., *Sanity checks for saliency maps*. NeurIPS, 31. 2018

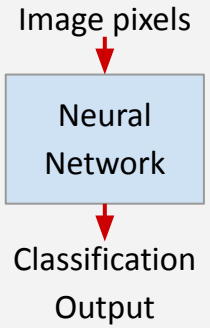
Integrated Gradients

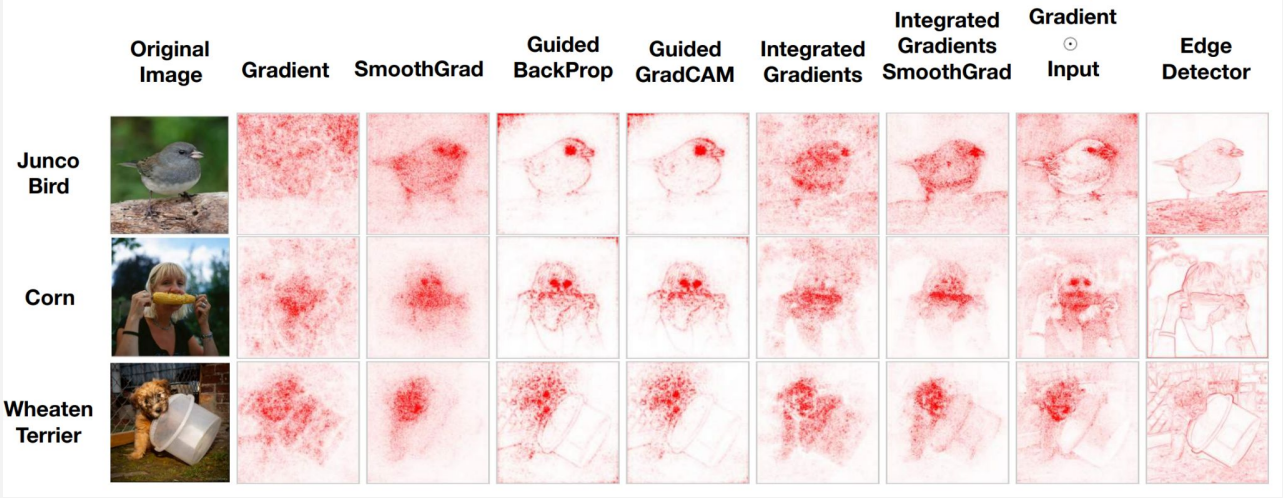
Axiomatic Attribution for Deep Networks,Sundararajan et al. ICML 2017

Input Interpretability

*How much the output depends on **each part** of the input*

Example:





Adebayo, J. et al., *Sanity checks for saliency maps*. NeurIPS, 31. 2018

<u>Integrated Gradients</u>

Given:

Input: ***x***

Baseline: ***x'*** (absence of signal)

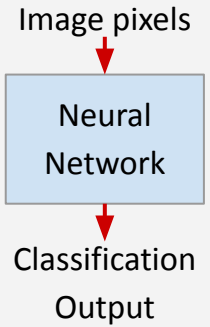
Model output: ***F(x)***

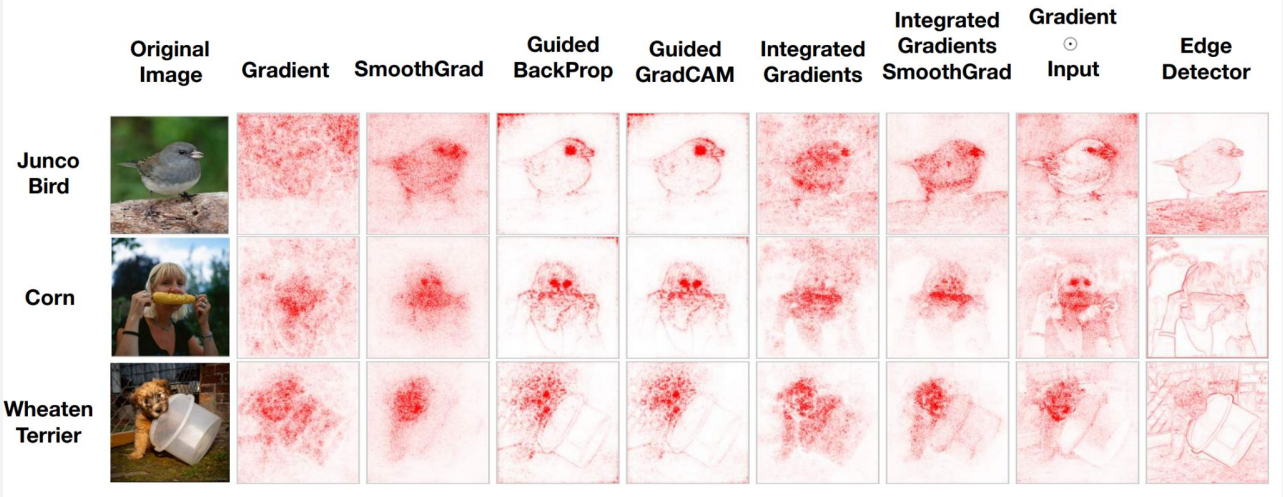
Axiomatic Attribution for Deep Networks,Sundararajan et al. ICML 2017

Input Interpretability

*How much the output depends on **each part** of the input*

Example:





Adebayo, J. et al., *Sanity checks for saliency maps*. NeurIPS, 31. 2018

Integrated Gradients

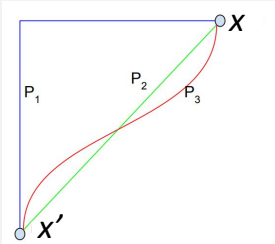
Given:

Input: ***x***

Baseline: ***x'*** (absence of signal)

Model output: ***F(x)***

Transforming X' into X in small steps

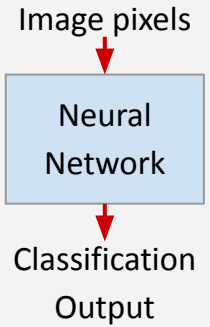


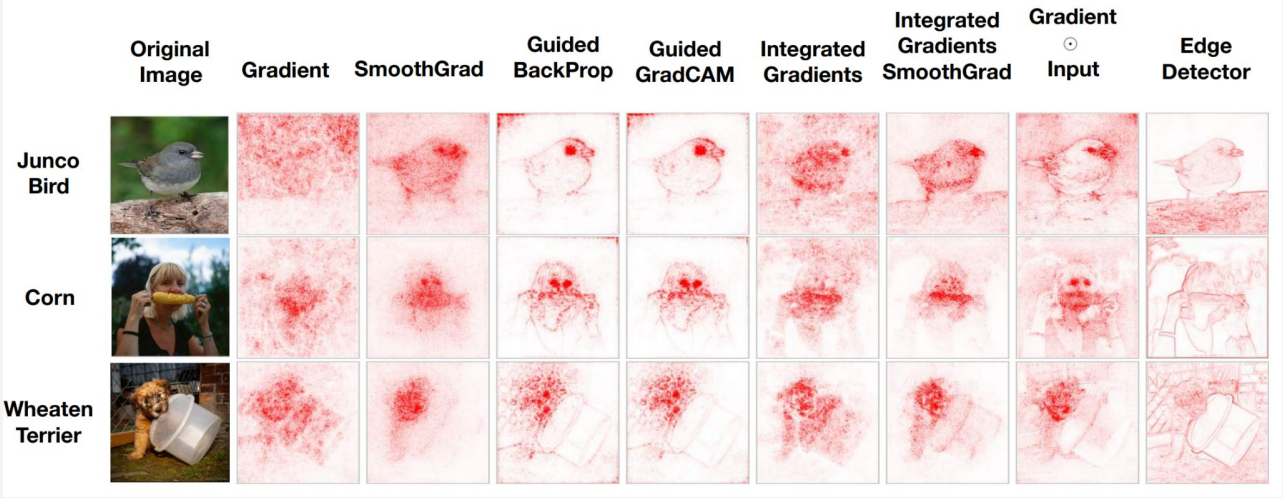
Axiomatic Attribution for Deep Networks,Sundararajan et al. ICML 2017

Input Interpretability

*How much the output depends on **each part** of the input*

Example:





Adebayo, J. et al., *Sanity checks for saliency maps*. NeurIPS, 31. 2018

Integrated Gradients

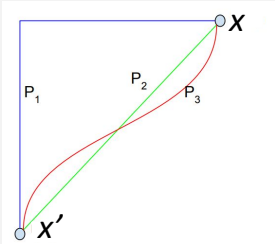
Given:

Input: ***x***

Baseline: ***x'*** (absence of signal)

Model output: ***F(x)***

Transforming X' into X in small steps



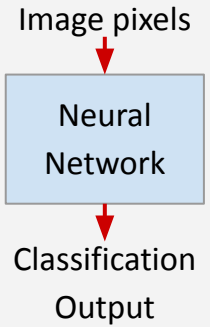
$$(x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

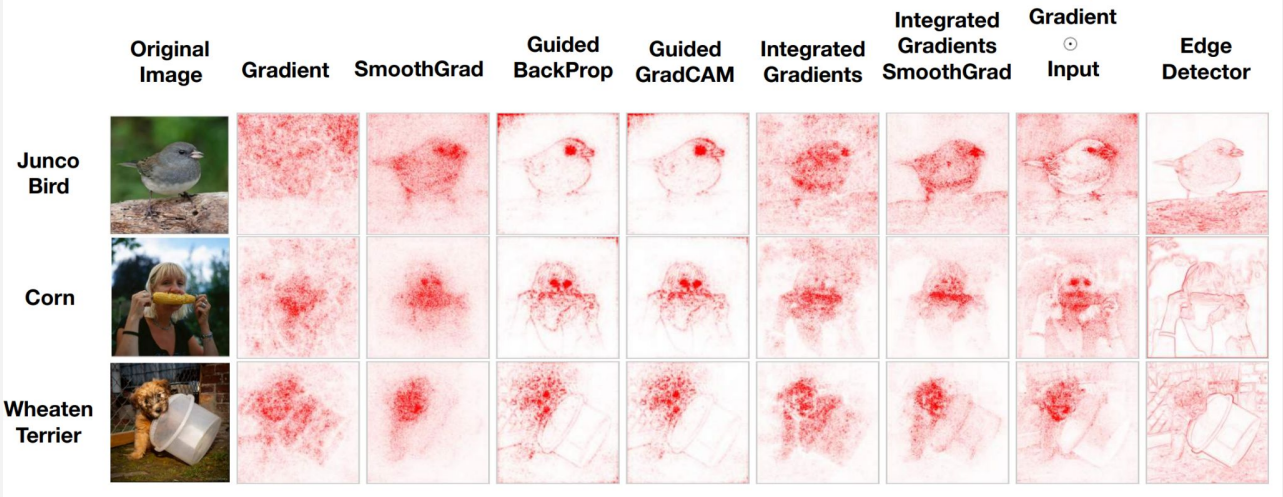
Axiomatic Attribution for Deep Networks,Sundararajan et al. ICML 2017

Input Interpretability

*How much the output depends on **each part** of the input*

Example:





Adebayo, J. et al., *Sanity checks for saliency maps*. NeurIPS, 31. 2018

Integrated Gradients

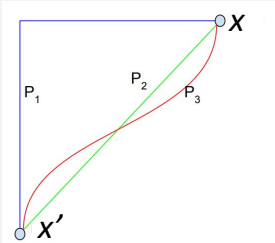
Given:

Input: ***x***

Baseline: ***x'*** (absence of signal)

Model output: ***F(x)***

Transforming X' into X in small steps



$$(x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

In practice: ***n*** Interpolation steps

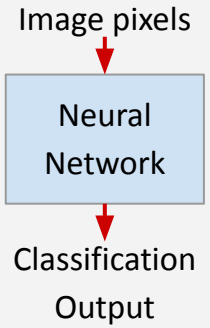
$$\Sigma_{i=1}^{n} \text{IntegratedGrads}_i(x) = F(x) - F(x')$$

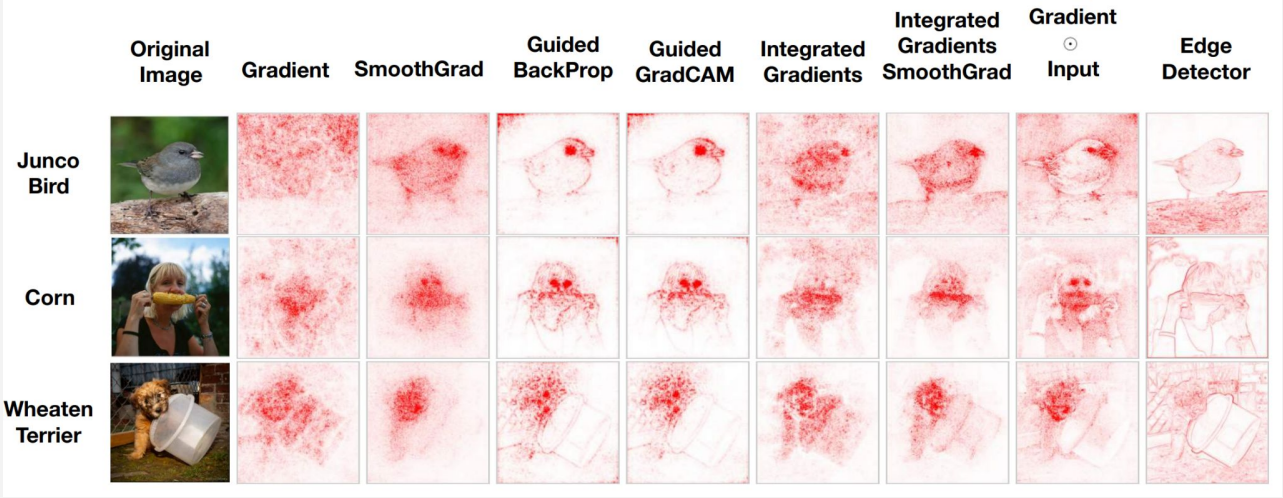
Axiomatic Attribution for Deep Networks,Sundararajan et al. ICML 2017

Input Interpretability

*How much the output depends on **each part** of the input*

Example:





Adebayo, J. et al., *Sanity checks for saliency maps*. NeurIPS, 31. 2018

Integrated Gradients

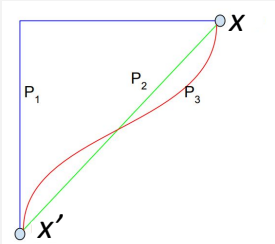
Given:

Input: ***x***

Baseline: ***x'*** (absence of signal)

Model output: ***F(x)***

Transforming X' into X in small steps



$$(x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

In practice: ***n*** Interpolation steps

$$\Sigma_{i=1}^{n} \text{IntegratedGrads}_i(x) = F(x) - F(x')$$

For each feature…

Positive Value:

Contributes towards output

Negative Value:

Contributes against output

Axiomatic Attribution for Deep Networks,Sundararajan et al. ICML 2017

Examples of Input Interpretability

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)
2. **Integrated Gradients** → Find most statistically significant motifs (2 or more AA) with high attribution

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)
2. **Integrated Gradients** → Find most statistically significant motifs (2 or more AA) with high attribution

Results: Find highest importance regions + list of consistently important motifs

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)
2. **Integrated Gradients** → Find most statistically significant motifs (2 or more AA) with high attribution

Evaluator

Results: Find highest importance regions + list of consistently important motifs

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)
2. **Integrated Gradients** → Find most statistically significant motifs (2 or more AA) with high attribution

Results: Find highest importance regions + list of consistently important motifs

Evaluator

Validation: Motif is consistent with prior work

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)
2. **Integrated Gradients** → Find most statistically significant motifs (2 or more AA) with high attribution

Results: Find highest importance regions + list of consistently important motifs

Evaluator

BioAutoMATED: an end-to-end automated machine learning tool for explanation and design of biological sequences (2023)

Convolutional Neural Network **predicting binding-enrichment** of peptides binding ranibizumab

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)
2. **Integrated Gradients** → Find most statistically significant motifs (2 or more AA) with high attribution

Results: Find highest importance regions + list of consistently important motifs

Evaluator

BioAutoMATED: an end-to-end automated machine learning tool for explanation and design of biological sequences (2023)

Convolutional Neural Network **predicting binding-enrichment** of peptides binding ranibizumab

Results: High importance FDY motif + lack of importance of methionine, cysteine, and asparagine residues

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)
2. **Integrated Gradients** → Find most statistically significant motifs (2 or more AA) with high attribution

Results: Find highest importance regions + list of consistently important motifs

Evaluator

BioAutoMATED: an end-to-end automated machine learning tool for explanation and design of biological sequences (2023)

Convolutional Neural Network **predicting binding-enrichment** of peptides binding ranibizumab

Results: High importance FDY motif + lack of importance of methionine, cysteine, and asparagine residues

Evaluator

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)
2. **Integrated Gradients** → Find most statistically significant motifs (2 or more AA) with high attribution

Results: Find highest importance regions + list of consistently important motifs

Evaluator

BioAutoMATED: an end-to-end automated machine learning tool for explanation and design of biological sequences (2023)

Convolutional Neural Network **predicting binding-enrichment** of peptides binding ranibizumab

Results: High importance FDY motif + lack of importance of methionine, cysteine, and asparagine residues

Evaluator

ESM-DBP: "Improving prediction performance of general protein language model by domain-adaptive pretraining on DNA-binding protein" (2024)

Results: Gradient attribution correlates with annotation of DNA binding residues

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)
2. **Integrated Gradients** → Find most statistically significant motifs (2 or more AA) with high attribution

Results: Find highest importance regions + list of consistently important motifs

Evaluator

BioAutoMATED: an end-to-end automated machine learning tool for explanation and design of biological sequences (2023)

Convolutional Neural Network **predicting binding-enrichment** of peptides binding ranibizumab

Results: High importance FDY motif + lack of importance of methionine, cysteine, and asparagine residues

Evaluator

ESM-DBP: "Improving prediction performance of general protein language model by domain-adaptive pretraining on DNA-binding protein" (2024)

Results: Gradient attribution correlates with annotation of DNA binding residues

Evaluator

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)
2. **Integrated Gradients** → Find most statistically significant motifs (2 or more AA) with high attribution

Results: Find highest importance regions + list of consistently important motifs

Evaluator

BioAutoMATED: an end-to-end automated machine learning tool for explanation and design of biological sequences (2023)

Convolutional Neural Network **predicting binding-enrichment** of peptides binding ranibizumab

Results: High importance FDY motif + lack of importance of methionine, cysteine, and asparagine residues

Evaluator

ESM-DBP: "Improving prediction performance of general protein language model by domain-adaptive pretraining on DNA-binding protein" (2024)

Results: Gradient attribution correlates with annotation of DNA binding residues

Evaluator
Multitasker

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)
2. **Integrated Gradients** → Find most statistically significant motifs (2 or more AA) with high attribution

Results: Find highest importance regions + list of consistently important motifs

Evaluator

BioAutoMATED: an end-to-end automated machine learning tool for explanation and design of biological sequences (2023)

Convolutional Neural Network **predicting binding-enrichment** of peptides binding ranibizumab

Results: High importance FDY motif + lack of importance of methionine, cysteine, and asparagine residues

Evaluator

ESM-DBP: "Improving prediction performance of general protein language model by domain-adaptive pretraining on DNA-binding protein" (2024)

Results: Gradient attribution correlates with annotation of DNA binding residues

Evaluator
Multitasker

EC-LMGraph "Accurate proteome-wide prediction of enzymes and catalytic sites using graph deep learning and protein language model" (2026)

Results: Predict catalytic-site from gradient attribution

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)
2. **Integrated Gradients** → Find most statistically significant motifs (2 or more AA) with high attribution

Results: Find highest importance regions + list of consistently important motifs

Evaluator

BioAutoMATED: an end-to-end automated machine learning tool for explanation and design of biological sequences (2023)

Convolutional Neural Network **predicting binding-enrichment** of peptides binding ranibizumab

Results: High importance FDY motif + lack of importance of methionine, cysteine, and asparagine residues

Evaluator

ESM-DBP: "Improving prediction performance of general protein language model by domain-adaptive pretraining on DNA-binding protein" (2024)

Results: Gradient attribution correlates with annotation of DNA binding residues

Evaluator
Multitasker

EC-LMGraph "Accurate proteome-wide prediction of enzymes and catalytic sites using graph deep learning and protein language model" (2026)

Results: Predict catalytic-site from gradient attribution

Evaluator

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)
2. **Integrated Gradients** → Find most statistically significant motifs (2 or more AA) with high attribution

Results: Find highest importance regions + list of consistently important motifs

Evaluator

BioAutoMATED: an end-to-end automated machine learning tool for explanation and design of biological sequences (2023)

Convolutional Neural Network **predicting binding-enrichment** of peptides binding ranibizumab

Results: High importance FDY motif + lack of importance of methionine, cysteine, and asparagine residues

Evaluator

ESM-DBP: "Improving prediction performance of general protein language model by domain-adaptive pretraining on DNA-binding protein" (2024)

Results: Gradient attribution correlates with annotation of DNA binding residues

Evaluator
Multitasker

EC-LMGraph "Accurate proteome-wide prediction of enzymes and catalytic sites using graph deep learning and protein language model" (2026)

Results: Predict catalytic-site from gradient attribution

Evaluator
Multitasker

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)
2. **Integrated Gradients** → Find most statistically significant motifs (2 or more AA) with high attribution

Results: Find highest importance regions + list of consistently important motifs

Evaluator

BioAutoMATED: an end-to-end automated machine learning tool for explanation and design of biological sequences (2023)

Convolutional Neural Network **predicting binding-enrichment** of peptides binding ranibizumab

Results: High importance FDY motif + lack of importance of methionine, cysteine, and asparagine residues

Evaluator

ESM-DBP: "Improving prediction performance of general protein language model by domain-adaptive pretraining on DNA-binding protein" (2024)

Results: Gradient attribution correlates with annotation of DNA binding residues

Evaluator
Multitasker

EC-LMGraph "Accurate proteome-wide prediction of enzymes and catalytic sites using graph deep learning and protein language model" (2026)

Results: Predict catalytic-site from gradient attribution

Evaluator
Multitasker

DeepPlantAllergy "DeepPlantAllergy: deep learning for explainable prediction of allergenicity in plant proteins" (2025)

1. Predict allergy (different models: SeqVec, ProtBert, and ESM-1B embeddings)

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)
2. **Integrated Gradients** → Find most statistically significant motifs (2 or more AA) with high attribution

Results: Find highest importance regions + list of consistently important motifs

Evaluator

BioAutoMATED: an end-to-end automated machine learning tool for explanation and design of biological sequences (2023)

Convolutional Neural Network **predicting binding-enrichment** of peptides binding ranibizumab

Results: High importance FDY motif + lack of importance of methionine, cysteine, and asparagine residues

Evaluator

ESM-DBP: "Improving prediction performance of general protein language model by domain-adaptive pretraining on DNA-binding protein" (2024)

Results: Gradient attribution correlates with annotation of DNA binding residues

Evaluator
Multitasker

EC-LMGraph "Accurate proteome-wide prediction of enzymes and catalytic sites using graph deep learning and protein language model" (2026)

Results: Predict catalytic-site from gradient attribution

Evaluator
Multitasker

DeepPlantAllergy "DeepPlantAllergy: deep learning for explainable prediction of allergenicity in plant proteins" (2025)

1. Predict allergy (different models: SeqVec, ProtBert, and ESM-1B embeddings)
2. Integrated Gradients → Find motifs in peptides (the allergen)

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)
2. **Integrated Gradients** → Find most statistically significant motifs (2 or more AA) with high attribution

Results: Find highest importance regions + list of consistently important motifs

Evaluator

BioAutoMATED: an end-to-end automated machine learning tool for explanation and design of biological sequences (2023)

Convolutional Neural Network **predicting binding-enrichment** of peptides binding ranibizumab

Results: High importance FDY motif + lack of importance of methionine, cysteine, and asparagine residues

Evaluator

ESM-DBP: "Improving prediction performance of general protein language model by domain-adaptive pretraining on DNA-binding protein" (2024)

Results: Gradient attribution correlates with annotation of DNA binding residues

Evaluator
Multitasker

EC-LMGraph "Accurate proteome-wide prediction of enzymes and catalytic sites using graph deep learning and protein language model" (2026)

Results: Predict catalytic-site from gradient attribution

Evaluator
Multitasker

DeepPlantAllergy "DeepPlantAllergy: deep learning for explainable prediction of allergenicity in plant proteins" (2025)

1. Predict allergy (different models: SeqVec, ProtBert, and ESM-1B embeddings)
2. Integrated Gradients → Find motifs in peptides (the allergen)

Results

Different models highlighted different regions

Some regions overlapping with experimentally validated epitopes

Docking tests to confirm binding predictions

Examples of Input Interpretability

Accurate and efficient interpretation of quantitative amino-acid attribution for disordered proteins undergoing LLPS (2023)

Liquid-liquid phase separation (LLPS): Process in cells regulated by intrinsically disordered proteins

1. Predict if sequence will undergo phase separation (ProtBert-BFD)
2. **Integrated Gradients** → Find most statistically significant motifs (2 or more AA) with high attribution

Results: Find highest importance regions + list of consistently important motifs

Evaluator

BioAutoMATED: an end-to-end automated machine learning tool for explanation and design of biological sequences (2023)

Convolutional Neural Network **predicting binding-enrichment** of peptides binding ranibizumab

Results: High importance FDY motif + lack of importance of methionine, cysteine, and asparagine residues

Evaluator

ESM-DBP: "Improving prediction performance of general protein language model by domain-adaptive pretraining on DNA-binding protein" (2024)

Results: Gradient attribution correlates with annotation of DNA binding residues

Evaluator
Multitasker

EC-LMGraph "Accurate proteome-wide prediction of enzymes and catalytic sites using graph deep learning and protein language model" (2026)

Results: Predict catalytic-site from gradient attribution

Evaluator
Multitasker

DeepPlantAllergy "DeepPlantAllergy: deep learning for explainable prediction of allergenicity in plant proteins" (2025)

1. Predict allergy (different models: SeqVec, ProtBert, and ESM-1B embeddings)
2. Integrated Gradients → Find motifs in peptides (the allergen)

Evaluator

Results

Different models highlighted different regions

Some regions overlapping with experimentally validated epitopes

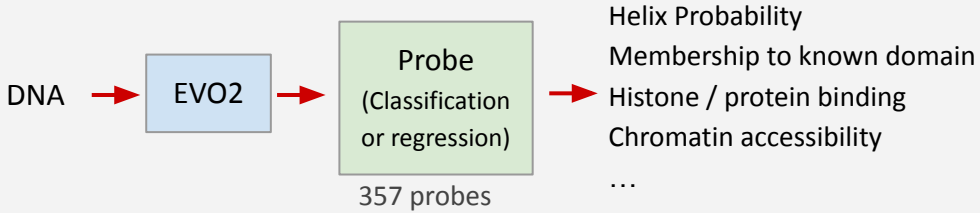
Docking tests to confirm binding predictions

Examples of Input Interpretability

EVEE: Explaining 4.2 million genetic variants with state-of-the-art, interpretable predictions (2026)

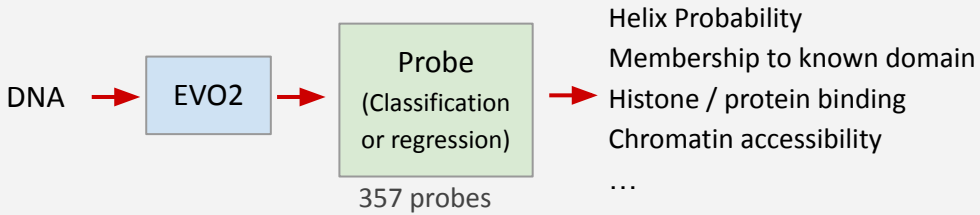
Examples of Input Interpretability

EVEE: Explaining 4.2 million genetic variants with state-of-the-art, interpretable predictions (2026)



Examples of Input Interpretability

EVEE: Explaining 4.2 million genetic variants with state-of-the-art, interpretable predictions (2026)

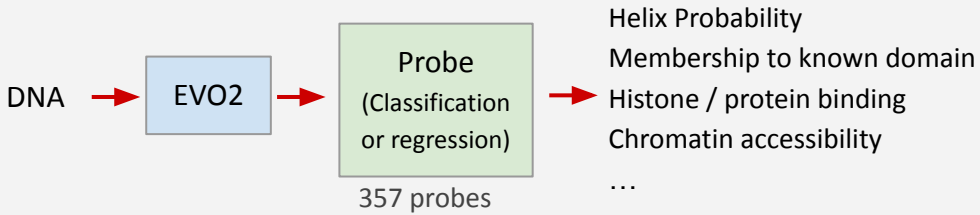


Explainability - Disruption profile

Objective: Explain disruption causes using a pretrained LLM (Claude Sonnet 4.6)

Examples of Input Interpretability

EVEE: Explaining 4.2 million genetic variants with state-of-the-art, interpretable predictions (2026)



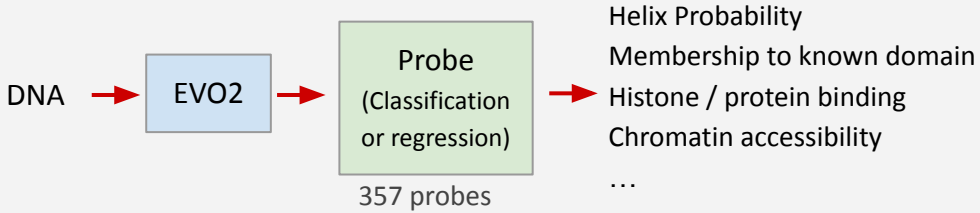
Explainability - Disruption profile

Objective: Explain disruption causes using a pretrained LLM (Claude Sonnet 4.6)

Disruption scoring: Difference in probe score for a reference vs variant

Examples of Input Interpretability

EVEE: Explaining 4.2 million genetic variants with state-of-the-art, interpretable predictions (2026)



Explainability - Disruption profile

Objective: Explain disruption causes using a pretrained LLM (Claude Sonnet 4.6)

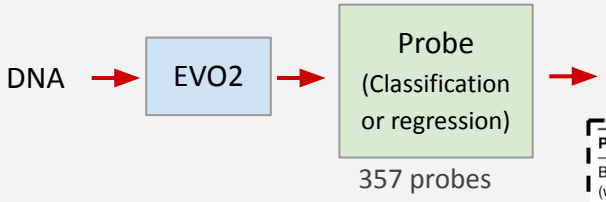
Disruption scoring: Difference in probe score for a reference vs variant

Input:

- Top-k disruptions
- Variant metadata (Gene name + variant coordinate, consequence type, genomic position …)
- Instructions prompt

Examples of Input Interpretability

EVEE: Explaining 4.2 million genetic variants with state-of-the-art, interpretable predictions (2026)



Explainability - Disruption profile

Objective: Explain disruption causes using a pretrained LLM (Claud

Disruption scoring: Difference in probe score for a reference vs vari

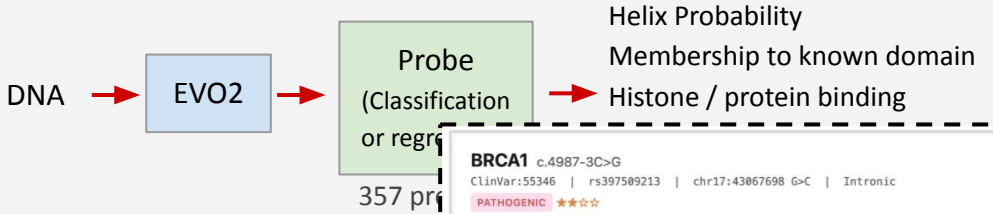
Input:

- Top-k disruptions
- Variant metadata (Gene name + variant coordinate, consequen
- Instructions prompt

Prompt	Content
Baseline (without Evo2)	You are a clinical genomics expert. Based on the variant information provided, interpret the likely biological impact and pathogenicity of this variant. Be specific about biological mechanisms when the information supports it. 3–5 sentences for summary, 1 sentence for mechanism. Writing style: Write like a senior geneticist. Be direct and concise. Never use em dashes. Never use "notably", "critically", "importantly", "striking". Don't use superlatives or hedging language. State facts plainly. Don't repeat the variant ID or coordinates in the summary. Specific numbers are good. Vague qualifiers ("profound", "massive", "broad") are not. Return JSON: {summary, mechanism, confidence, key_evidence}
Evo2 (with Evo2)	You are a clinical genomics expert interpreting variant pathogenicity predictions from Evo2, a 7-billion-parameter DNA foundation model. The system uses a probe trained on reference genome activations, then run on both reference and variant activations. For each biological feature (helix, conservation, domain membership, etc.), you see: **ref** (the probe's prediction on the reference genome), **var** (the prediction on the variant genome), and Δ (var − ref, the disruption signal). Large negative deltas mean the variant disrupts that feature. Near-zero deltas mean the feature is preserved. This is the key signal: pathogenic variants show large disruptions, benign variants show near-zero deltas. Provide a clinical interpretation: lead with the disruption story (which features were disrupted and which preserved); use the gene function context to explain why disruption of those features matters for this gene; explain why the model predicts what it does based on the disruption profile. Do not reference clinical prediction tools (CADD, AlphaMissense, SIFT, PolyPhen, etc.). 3–5 sentences for summary, 1 sentence for mechanism. Writing style: Write like a senior geneticist. Be direct and concise. Never use em dashes. Never use "notably", "critically", "importantly", "striking". Don't use superlatives or hedging language. State facts plainly. Don't repeat the variant ID or coordinates in the summary. Specific numbers are good. Vague qualifiers ("profound", "massive", "broad") are not. Return JSON: {summary, mechanism, confidence, key_evidence}

Examples of Input Interpretability

EVEE: Explaining 4.2 million genetic variants with state-of-the-art, interpretable predictions (2026)



Explainability - Disruption profile

Objective: Explain disruption causes using a pretrained

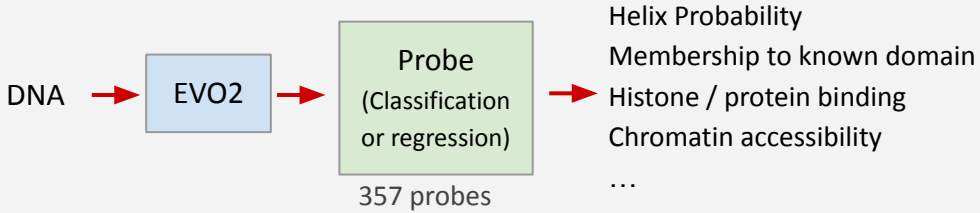
Disruption scoring: Difference in probe score for a refe

Input:

- Top-k disruptions
- Variant metadata (Gene name + variant coordinat
- Instructions prompt

Examples of Input Interpretability

EVEE: Explaining 4.2 million genetic variants with state-of-the-art, interpretable predictions (2026)



Explainability - Disruption profile

Objective: Explain disruption causes using a pretrained LLM (Claude Sonnet 4.6)

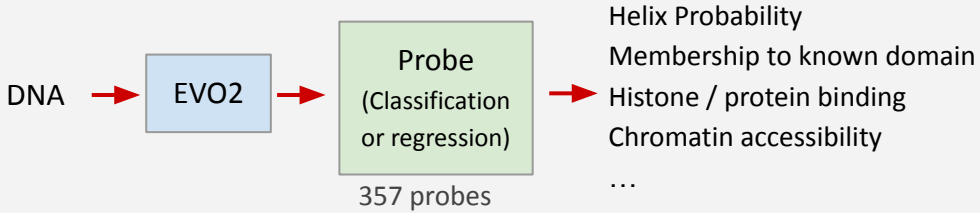
Disruption scoring: Difference in probe score for a **reference vs variant**

Input:

- Top-k disruptions
- Variant metadata (Gene name + variant coordinate, consequence type, genomic position …)
- Instructions prompt

Examples of Input Interpretability

EVEE: Explaining 4.2 million genetic variants with state-of-the-art, interpretable predictions (2026)



Explainability - Disruption profile

Objective: Explain disruption causes using a pretrained LLM (Claude Sonnet 4.6)

Disruption scoring: Difference in probe score for a **reference vs variant**

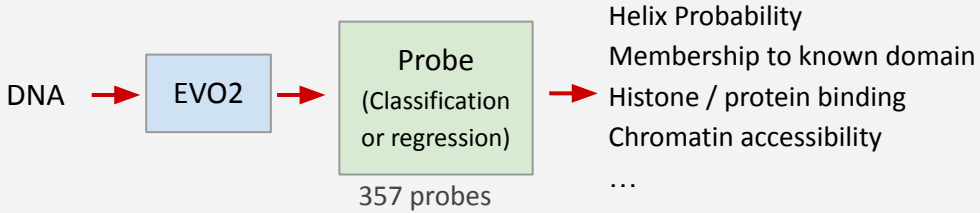
Input:

- Top-k disruptions
- Variant metadata (Gene name + variant coordinate, consequence type, genomic position …)
- Instructions prompt

Results:
- Human-readable mechanism hypotheses
- Explanations line up with established biology e.g.: BRCA1 splice disruption and LDLR/RET structural consequences

Examples of Input Interpretability

EVEE: Explaining 4.2 million genetic variants with state-of-the-art, interpretable predictions (2026)



Evaluator

We will trust prediction more because of justification

Explainability - Disruption profile

Objective: Explain disruption causes using a pretrained LLM (Claude Sonnet 4.6)

Disruption scoring: Difference in probe score for a **reference vs variant**

Input:

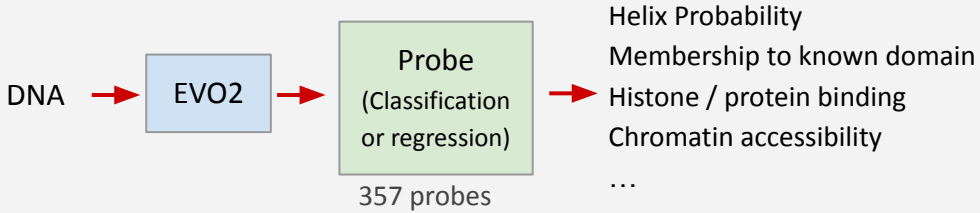
- Top-k disruptions
- Variant metadata (Gene name + variant coordinate, consequence type, genomic position …)
- Instructions prompt

Results:

- Human-readable mechanism hypotheses
- Explanations line up with established biology e.g.: BRCA1 splice disruption and LDLR/RET structural consequences

Examples of Input Interpretability

EVEE: Explaining 4.2 million genetic variants with state-of-the-art, interpretable predictions (2026)



Evaluator

We will trust prediction more because of justification

Teacher - weakly

"Probes are trained to predict known annotations; Novel molecular mechanisms of pathogenicity would not be captured by this supervised approach"

Explainability - Disruption profile

Objective: Explain disruption causes using a pretrained LLM (Claude Sonnet 4.6)

Disruption scoring: Difference in probe score for a **reference vs variant**

Input:

- Top-k disruptions
- Variant metadata (Gene name + variant coordinate, consequence type, genomic position …)
- Instructions prompt

Results:

- Human-readable mechanism hypotheses
- Explanations line up with established biology e.g.: BRCA1 splice disruption and LDLR/RET structural consequences

Model Interpretability

Understand the contribution of different parts of the model towards the output

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
    - **Sparse autoencoders:** Project to sparse concepts

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
    - **Sparse autoencoders:** Project to sparse concepts

Ante-hoc

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
    - **Sparse autoencoders:** Project to sparse concepts

Ante-hoc

- **Interpretable architectures**

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
  - **Sparse autoencoders:** Project to sparse concepts

Ante-hoc

- **Interpretable architectures**

Evaluator

Insights Into the Inner Workings of Transformer Models for Protein Function Prediction (2023)

Results: Transformer heads attribution maps correlate with ground truth sequence annotations (e.g. transmembrane regions, active sites)

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
  - **Sparse autoencoders:** Project to sparse concepts

Ante-hoc

- **Interpretable architectures**

Evaluator

Insights Into the Inner Workings of Transformer Models for Protein Function Prediction (2023)

Results: Transformer heads attribution maps correlate with ground truth sequence annotations (e.g. transmembrane regions, active sites)

BERTology Meets Biology (2020)

Results: Attention focusing on folding structure, target binding sites

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
  - **Sparse autoencoders:** Project to sparse concepts

Ante-hoc

- **Interpretable architectures**

Evaluator

Insights Into the Inner Workings of Transformer Models for Protein Function Prediction (2023)

Results: Transformer heads attribution maps correlate with ground truth sequence annotations (e.g. transmembrane regions, active sites)

BERTology Meets Biology (2020) ***Evaluator***

Results: Attention focusing on folding structure, target binding sites

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
  - **Sparse autoencoders:** Project to sparse concepts

Evaluator

Insights Into the Inner Workings of Transformer Models for Protein Function Prediction (2023)

Results: Transformer heads attribution maps correlate with ground truth sequence annotations (e.g. transmembrane regions, active sites)

BERTology Meets Biology (2020) ***Evaluator***

Results: Attention focusing on folding structure, target binding sites

Ante-hoc

- **Interpretable architectures**

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

Results: Steering Protein Language Model towards high activity enzymes ***Coach***

Model Interpretability

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

Results: Steering Protein Language Model towards high activity enzymes ***Coach***

Model Interpretability

Coach

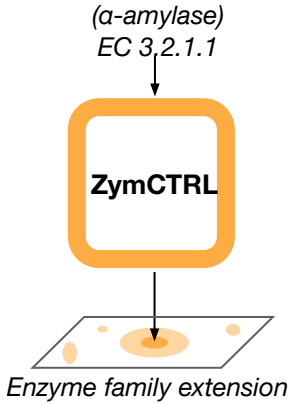
Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

Model Interpretability

Coach

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

- **ZymCTRL** generates enzymes

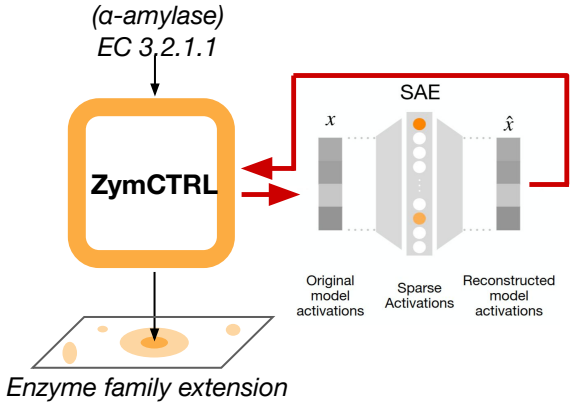


Model Interpretability

Coach

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

- **ZymCTRL** generates enzymes
- **SAE** trained to project features to **high dimensionality** & **sparse** space

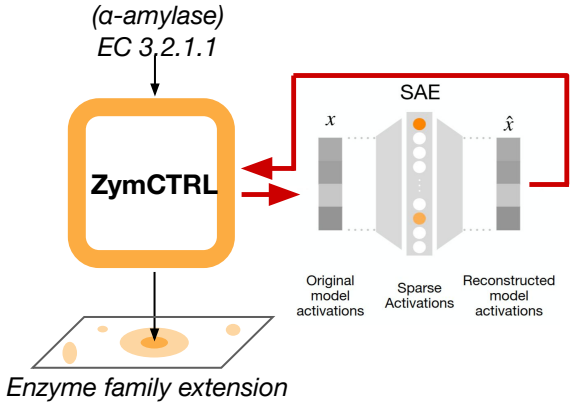


Model Interpretability

Coach

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

- **ZymCTRL** generates enzymes
- **SAE** trained to project features to **high dimensionality** & **sparse** space
  α-amylase dataset

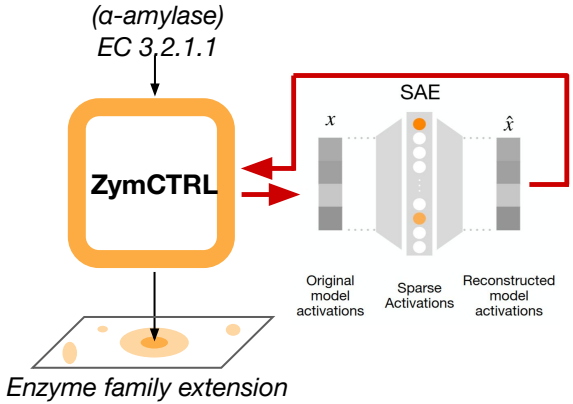


Model Interpretability

Coach

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

- **ZymCTRL** generates enzymes
- **SAE** trained to project features to **high dimensionality** & **sparse** space
  α-amylase dataset



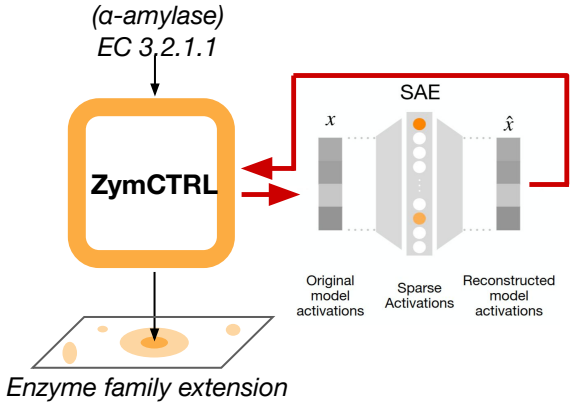
Steering

Model Interpretability

Coach

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

- **ZymCTRL** generates enzymes
- **SAE** trained to project features to **high dimensionality** & **sparse** space
  α-amylase dataset



Steering

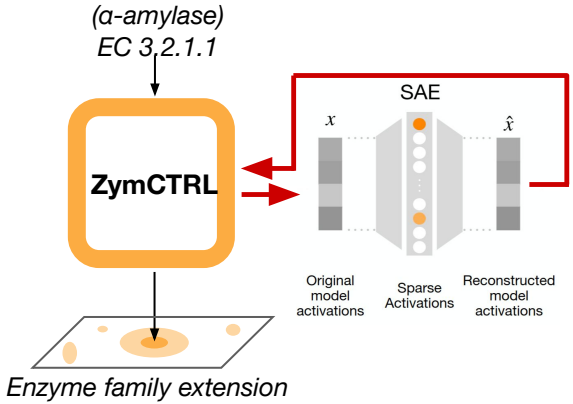
Clamping steering:

Model Interpretability

Coach

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

- **ZymCTRL** generates enzymes
- **SAE** trained to project features to **high dimensionality** & **sparse** space
  α-amylase dataset



Steering

Clamping steering:

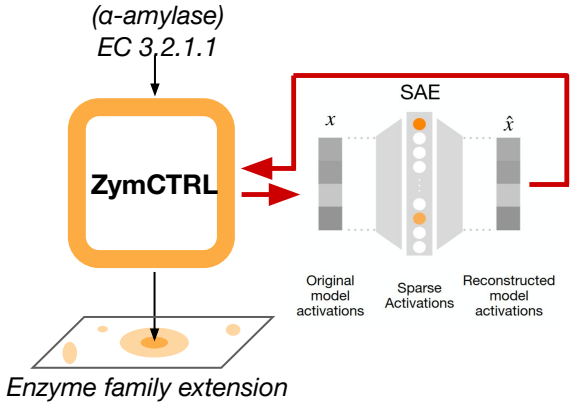
1. Logistic regression: feature to activity correlation ***β***

Model Interpretability

Coach

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

- **ZymCTRL** generates enzymes
- **SAE** trained to project features to **high dimensionality** & **sparse** space
  α-amylase dataset



Steering

Clamping steering:

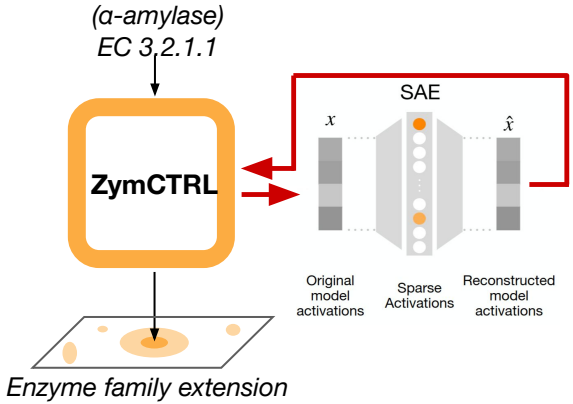
1. Logistic regression: feature to activity correlation ***β***
2. Set high ***β*** features to max
   If it naturally activated

Model Interpretability

Coach

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

- **ZymCTRL** generates enzymes
- **SAE** trained to project features to **high dimensionality** & **sparse** space
  α-amylase dataset



Steering

Clamping steering:

1. Logistic regression: feature to activity correlation ***β***
2. Set high ***β*** features to max
   If it naturally activated

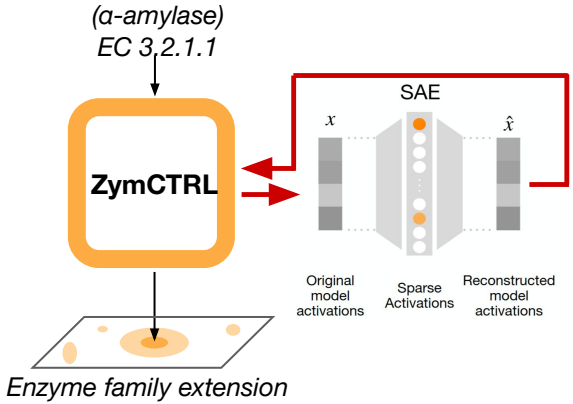
MSA steering:

Model Interpretability

Coach

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

- **ZymCTRL** generates enzymes
- **SAE** trained to project features to **high dimensionality** & **sparse** space
  α-amylase dataset



Steering

Clamping steering:

1. Logistic regression: feature to activity correlation ***β***
2. Set high ***β*** features to max
   If it naturally activated

MSA steering:

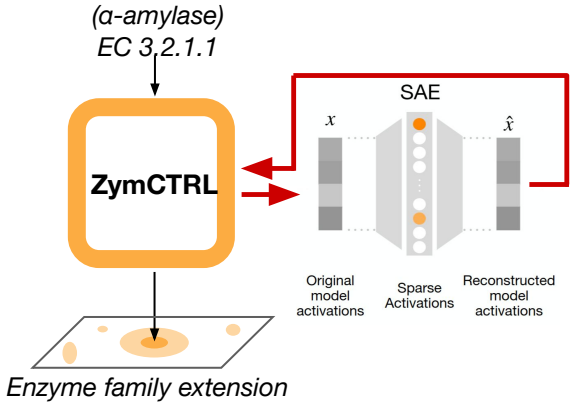
1. Align training set (MSA)

Model Interpretability

Coach

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

- **ZymCTRL** generates enzymes
- **SAE** trained to project features to **high dimensionality** & **sparse** space
  α-amylase dataset



Steering

Clamping steering:

1. Logistic regression: feature to activity correlation ***β***
2. Set high ***β*** features to max
   If it naturally activated

MSA steering:

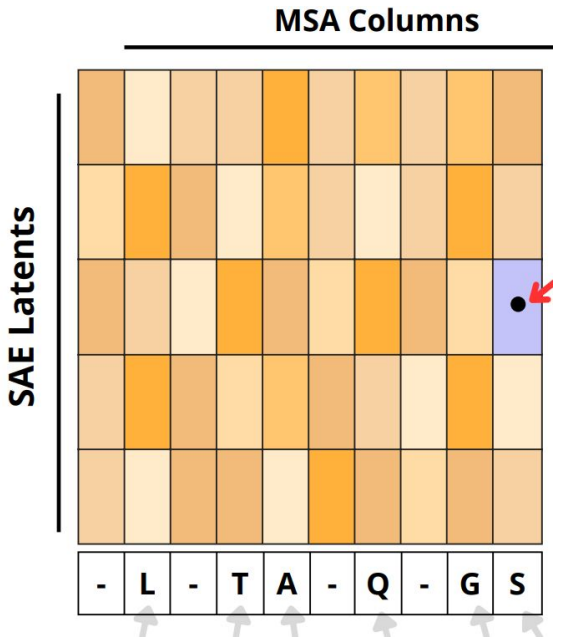
1. Align training set (MSA)
2. Logistic regression per MSA column

Model Interpretability

Coach

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

- **ZymCTRL** generates enzymes
- **SAE** trained to project features to **high dimensionality** & **sparse** space
  α-amylase dataset



Steering

Clamping steering:

1. Logistic regression: feature to activity correlation β
2. Set high β features to max
   If it naturally activated

MSA steering:

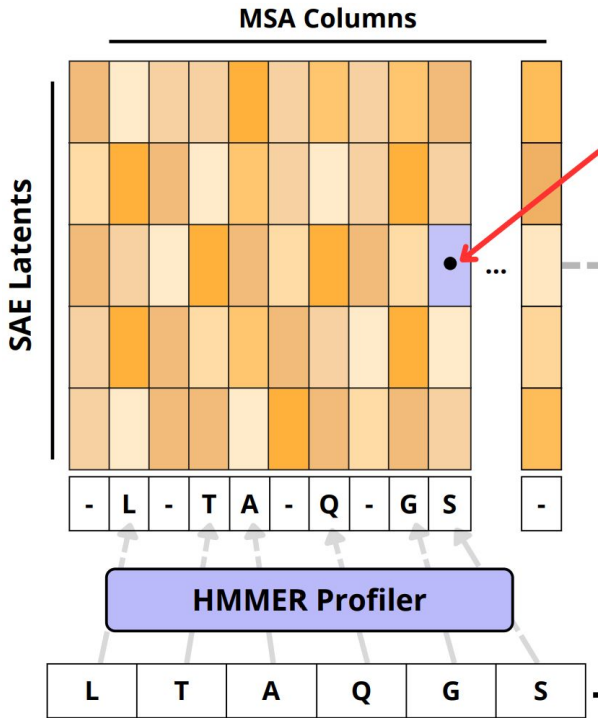
1. Align training set (MSA)
2. Logistic regression per MSA column

Model Interpretability

Coach

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

- **ZymCTRL** generates enzymes
- **SAE** trained to project features to **high dimensionality** & **sparse** space
  α-amylase dataset



Steering

Clamping steering:

1. Logistic regression: feature to activity correlation ***β***
2. Set high ***β*** features to max
   If it naturally activated

MSA steering:

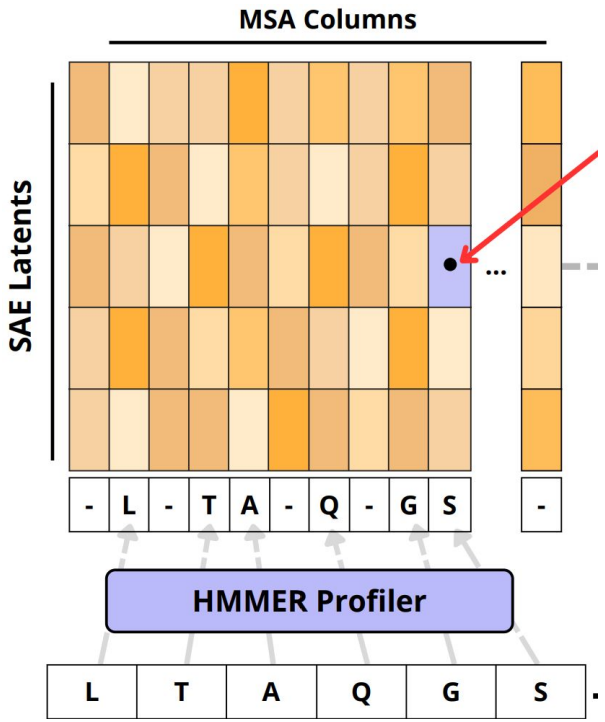
1. Align training set (MSA)
2. Logistic regression per MSA column
   Generation:
   a. Align partial sequence (HMM profiler)

Model Interpretability

Coach

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

- **ZymCTRL** generates enzymes
- **SAE** trained to project features to **high dimensionality** & **sparse** space
  α-amylase dataset



Steering

Clamping steering:

1. Logistic regression: feature to activity correlation β
2. Set high β features to max
   If it naturally activated

MSA steering:

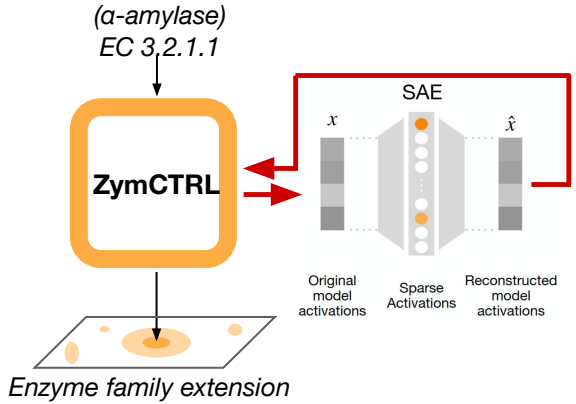
1. Align training set (MSA)
2. Logistic regression per MSA column
   Generation:
   a. Align partial sequence (HMM profiler)
   b. Set high (column-wise) β_c features to max

Model Interpretability

Coach

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

- **ZymCTRL** generates enzymes
- **SAE** trained to project features to **high dimensionality** & **sparse** space
  α-amylase dataset



Steering

Clamping steering:

1. Logistic regression: feature to activity correlation β
2. Set high β features to max
   If it naturally activated

MSA steering:

1. Align training set (MSA)
2. Logistic regression per MSA column
   Generation:
   a. Align partial sequence (HMM profiler)
   b. Set high (column-wise) β_c features to max

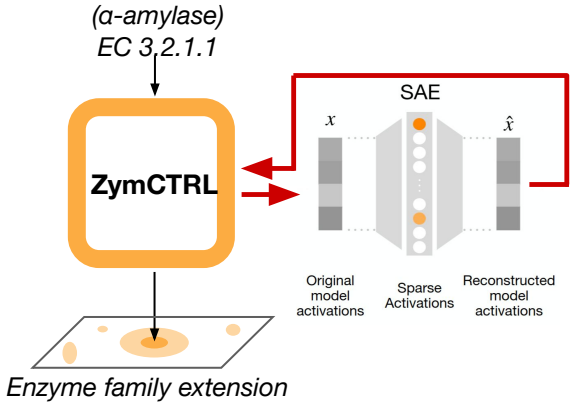
Results:

Model Interpretability

Coach

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

- **ZymCTRL** generates enzymes
- **SAE** trained to project features to **high dimensionality** & **sparse** space
  α-amylase dataset



Steering

Clamping steering:

1. Logistic regression: feature to activity correlation β
2. Set high β features to max
   If it naturally activated

MSA steering:

1. Align training set (MSA)
2. Logistic regression per MSA column
   Generation:
   a. Align partial sequence (HMM profiler)
   b. Set high (column-wise) β_c features to max

Results:

Steering **increases predicted activity**
Better than fine tuning

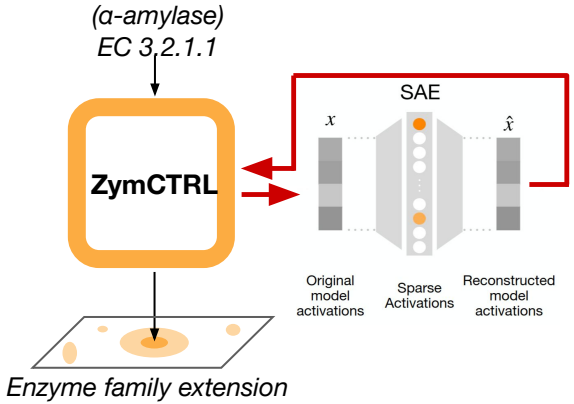
Intervention	Median Predicted Activity
Base (no steering)	1.045
Ablation	1.051
Clamping	1.139
CAA	1.058
MSA Steering	**1.995**

Model Interpretability

Coach

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

- **ZymCTRL** generates enzymes
- **SAE** trained to project features to **high dimensionality** & **sparse** space
  α-amylase dataset



Steering

Clamping steering:

1. Logistic regression: feature to activity correlation β
2. Set high β features to max
   If it naturally activated

MSA steering:

1. Align training set (MSA)
2. Logistic regression per MSA column
   Generation:
   a. Align partial sequence (HMM profiler)
   b. Set high (column-wise) β_c features to max

Results:

Steering **increases predicted activity**
- Better than fine tuning
- RL still better (work in progress)

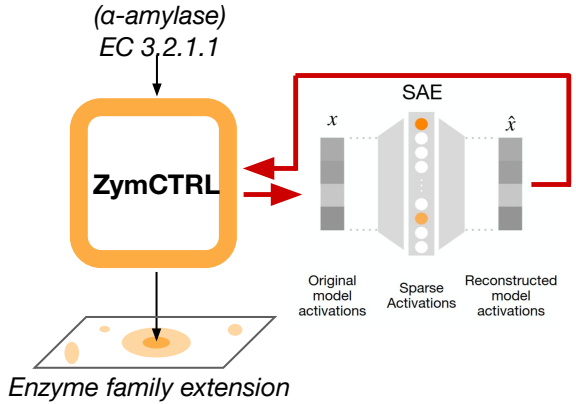
Intervention	Median Predicted Activity
Base (no steering)	1.045
Ablation	1.051
Clamping	1.139
CAA	1.058
MSA Steering	**1.995**

Model Interpretability

Coach

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

- **ZymCTRL** generates enzymes
- **SAE** trained to project features to **high dimensionality** & **sparse** space
  α-amylase dataset



Steering

Clamping steering:

1. Logistic regression: feature to activity correlation β
2. Set high β features to max
   If it naturally activated

MSA steering:

1. Align training set (MSA)
2. Logistic regression per MSA column
   Generation:
   a. Align partial sequence (HMM profiler)
   b. Set high (column-wise) β_c features to max

Results:

Steering **increases predicted activity**
- Better than fine tuning
- RL still better (work in progress)

Limitation: ***Activity is predicted***

Intervention	Median Predicted Activity
Base (no steering)	1.045
Ablation	1.051
Clamping	1.139
CAA	1.058
MSA Steering	**1.995**

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
  - **Sparse autoencoders:** Project to sparse concepts

Evaluator

Insights Into the Inner Workings of Transformer Models for Protein Function Prediction (2023)

Results: Transformer heads attribution maps correlate with ground truth sequence annotations (e.g. transmembrane regions, active sites)

BERTology Meets Biology (2020) ***Evaluator***

Results: Attention focusing on folding structure, target binding sites

Ante-hoc

- **Interpretable architectures**

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

Results: Steering Protein Language Model towards high activity enzymes ***Coach***

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
  - **Sparse autoencoders:** Project to sparse concepts

Evaluator

Insights Into the Inner Workings of Transformer Models for Protein Function Prediction (2023)

Results: Transformer heads attribution maps correlate with ground truth sequence annotations (e.g. transmembrane regions, active sites)

BERTology Meets Biology (2020) ***Evaluator***

Results: Attention focusing on folding structure, target binding sites

Ante-hoc

- **Interpretable architectures**

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

Results: Steering Protein Language Model towards high activity enzymes ***Coach***

Concept Bottleneck Language Models For protein design (2024)

Setup: Sequences annotated with predefined human-understandable concepts

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
  - **Sparse autoencoders:** Project to sparse concepts

Evaluator

Insights Into the Inner Workings of Transformer Models for Protein Function Prediction (2023)

Results: Transformer heads attribution maps correlate with ground truth sequence annotations (e.g. transmembrane regions, active sites)

BERTology Meets Biology (2020) ***Evaluator***

Results: Attention focusing on folding structure, target binding sites

Ante-hoc

- **Interpretable architectures**

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

Results: Steering Protein Language Model towards high activity enzymes ***Coach***

Concept Bottleneck Language Models For protein design (2024) ***Engineer***

Setup: Sequences annotated with predefined human-understandable concepts

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
  - **Sparse autoencoders:** Project to sparse concepts

Evaluator

Insights Into the Inner Workings of Transformer Models for Protein Function Prediction (2023)

Results: Transformer heads attribution maps correlate with ground truth sequence annotations (e.g. transmembrane regions, active sites)

BERTology Meets Biology (2020) ***Evaluator***

Results: Attention focusing on folding structure, target binding sites

Ante-hoc

- **Interpretable architectures**

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

Results: Steering Protein Language Model towards high activity enzymes ***Coach***

Concept Bottleneck Language Models For protein design (2024) ***Engineer*** ***Coach***

Setup: Sequences annotated with predefined human-understandable concepts

Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
  - **Sparse autoencoders:** Project to sparse concepts

Evaluator

Insights Into the Inner Workings of Transformer Models for Protein Function Prediction (2023)

Results: Transformer heads attribution maps correlate with ground truth sequence annotations (e.g. transmembrane regions, active sites)

BERTology Meets Biology (2020) ***Evaluator***

Results: Attention focusing on folding structure, target binding sites

Ante-hoc

- **Interpretable architectures**

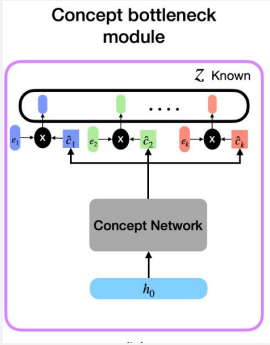
Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

Results: Steering Protein Language Model towards high activity enzymes ***Coach***

Concept Bottleneck Language Models For protein design (2024) ***Engineer*** ***Coach***

Setup: Sequences annotated with predefined human-understandable concepts

1. Sequence embedding h_0 is projected to a vector c_i per concept



Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
  - **Sparse autoencoders:** Project to sparse concepts

Evaluator

Insights Into the Inner Workings of Transformer Models for Protein Function Prediction (2023)

Results: Transformer heads attribution maps correlate with ground truth sequence annotations (e.g. transmembrane regions, active sites)

BERTology Meets Biology (2020) ***Evaluator***

Results: Attention focusing on folding structure, target binding sites

Ante-hoc

- **Interpretable architectures**

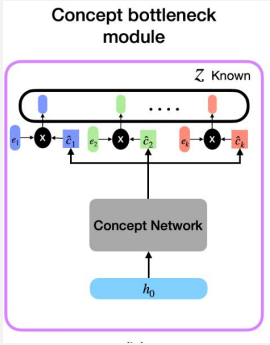
Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

Results: Steering Protein Language Model towards high activity enzymes ***Coach***

Concept Bottleneck Language Models For protein design (2024) ***Engineer*** ***Coach***

Setup: Sequences annotated with predefined human-understandable concepts

1. Sequence embedding h_0 is projected to a vector c_i per concept



Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
  - **Sparse autoencoders:** Project to sparse concepts

Evaluator

Insights Into the Inner Workings of Transformer Models for Protein Function Prediction (2023)

Results: Transformer heads attribution maps correlate with ground truth sequence annotations (e.g. transmembrane regions, active sites)

BERTology Meets Biology (2020) ***Evaluator***

Results: Attention focusing on folding structure, target binding sites

Ante-hoc

- **Interpretable architectures**

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

Results: Steering Protein Language Model towards high activity enzymes ***Coach***

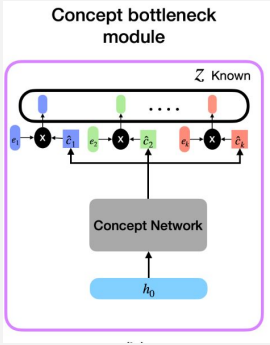
Concept Bottleneck Language Models For protein design (2024) ***Engineer*** ***Coach***

Setup: Sequences annotated with predefined human-understandable concepts

1. Sequence embedding h_0 is projected to a vector c_i per concept

2. Cross product with a learnt embedding e_i per concept = how much of this concept (z_i)

$$z_i = \hat{c}_i \times e_i.$$



Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
  - **Sparse autoencoders:** Project to sparse concepts

Evaluator

Insights Into the Inner Workings of Transformer Models for Protein Function Prediction (2023)

Results: Transformer heads attribution maps correlate with ground truth sequence annotations (e.g. transmembrane regions, active sites)

BERTology Meets Biology (2020) ***Evaluator***

Results: Attention focusing on folding structure, target binding sites

Ante-hoc

- **Interpretable architectures**

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

Results: Steering Protein Language Model towards high activity enzymes ***Coach***

Concept Bottleneck Language Models For protein design (2024) ***Engineer*** ***Coach***

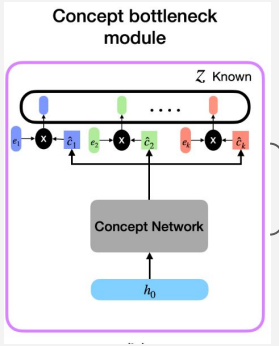
Setup: Sequences annotated with predefined human-understandable concepts

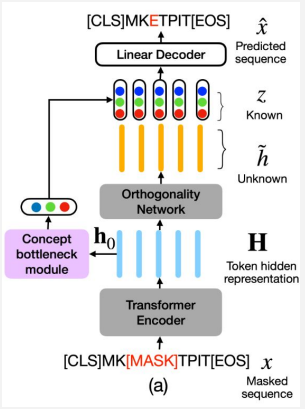
1. Sequence embedding h_0 is projected to a vector c_i per concept

2. Cross product with a learnt embedding e_i per concept = how much of this concept (z_i)

$$z_i = \hat{c}_i \times e_i.$$

3. The z_i coefficients are concatenated with "non-concept" info (***h*** unknown)





Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
  - **Sparse autoencoders:** Project to sparse concepts

Evaluator

Insights Into the Inner Workings of Transformer Models for Protein Function Prediction (2023)

Results: Transformer heads attribution maps correlate with ground truth sequence annotations (e.g. transmembrane regions, active sites)

BERTology Meets Biology (2020) ***Evaluator***

Results: Attention focusing on folding structure, target binding sites

Ante-hoc

- **Interpretable architectures**

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

Results: Steering Protein Language Model towards high activity enzymes ***Coach***

Concept Bottleneck Language Models For protein design (2024) ***Engineer*** ***Coach***

Setup: Sequences annotated with predefined human-understandable concepts

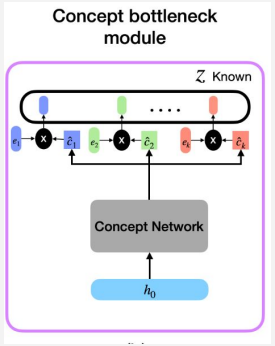
1. Sequence embedding $\boldsymbol{h_0}$ is projected to a vector $\boldsymbol{c_i}$ per concept

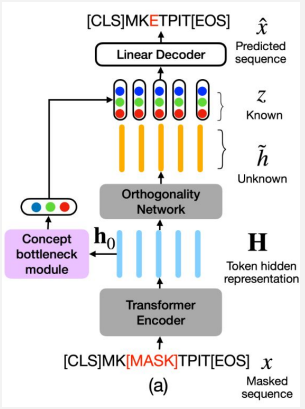
2. Cross product with a learnt embedding $\boldsymbol{e_i}$ per concept = how much of this concept ($\boldsymbol{z_i}$)

$$z_i = \hat{c}_i \times e_i.$$

3. The $\boldsymbol{z_i}$ coefficients are concatenated with "non-concept" info (***h*** unknown)

Maximise cosine distance to concept vector *(no disentanglement guarantee)*





Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
  - **Sparse autoencoders:** Project to sparse concepts

Evaluator

Insights Into the Inner Workings of Transformer Models for Protein Function Prediction (2023)

Results: Transformer heads attribution maps correlate with ground truth sequence annotations (e.g. transmembrane regions, active sites)

BERTology Meets Biology (2020) ***Evaluator***

Results: Attention focusing on folding structure, target binding sites

Ante-hoc

- **Interpretable architectures**

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

Results: Steering Protein Language Model towards high activity enzymes ***Coach***

Concept Bottleneck Language Models For protein design (2024) ***Engineer*** ***Coach***

Setup: Sequences annotated with predefined human-understandable concepts

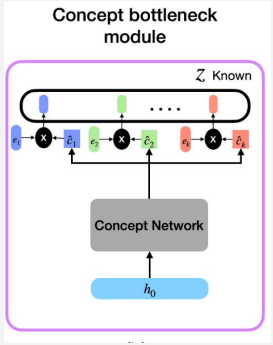
1. Sequence embedding $\boldsymbol{h_0}$ is projected to a vector $\boldsymbol{c_i}$ per concept

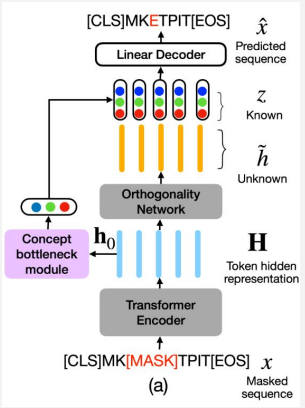
2. Cross product with a learnt embedding $\boldsymbol{e_i}$ per concept = how much of this concept ($\boldsymbol{z_i}$)

$$z_i = \hat{c}_i \times e_i.$$

3. The $\boldsymbol{z_i}$ coefficients are concatenated with "non-concept" info (***h*** unknown)
Maximise cosine distance to concept vector *(no disentanglement guarantee)*

Results: Modifying ***z*** steers the model (and validates ***z*** influences result)





Model Interpretability

Understand the contribution of different parts of the model towards the output

Post-hoc

- **Importance attribution:** To Neurons / Layers / Latents
- **Feature analysis:** What do neurons / layers look for
- **Attention analysis:** What does an attention head look for
- **Probing:** Does a representation contain ***x*** information?
- **Mechanistic / causal intervention:**
  - **Sparse autoencoders:** Project to sparse concepts

Evaluator

Insights Into the Inner Workings of Transformer Models for Protein Function Prediction (2023)

Results: Transformer heads attribution maps correlate with ground truth sequence annotations (e.g. transmembrane regions, active sites)

BERTology Meets Biology (2020) ***Evaluator***

Results: Attention focusing on folding structure, target binding sites

Ante-hoc

- **Interpretable architectures**

Sparse Autoencoders in Protein Engineering Campaigns: Steering and Model Diffing (2025)

Results: Steering Protein Language Model towards high activity enzymes ***Coach***

Concept Bottleneck Language Models For protein design (2024) ***Engineer*** ***Coach***

Setup: Sequences annotated with predefined human-understandable concepts

1. Sequence embedding h_0 is projected to a vector c_i per concept
2. Cross product with a learnt embedding e_i per concept = how much of this concept (z_i)

$$z_i = \hat{c}_i \times e_i.$$

3. The z_i coefficients are concatenated with "non-concept" info (h unknown)
   Maximise cosine distance to concept vector *(no disentanglement guarantee)*

Results: Modifying ***z*** steers the model (and validates ***z*** influences result)

Engineer **role in general AI**

Most commonly used for:

Pruning
- Projections in attention layers
- Neurons

Conclusions

Conclusions

- **Evaluator** role is the most demonstrated

Conclusions

- **Evaluator** role is the most demonstrated
  - Currently: Validates against **pre-existing field knowledge** = *Bottleneck*

Conclusions

- **Evaluator** role is the most demonstrated
  - Currently: Validates against **pre-existing field knowledge** = *Bottleneck*
  - Can **automated metrics** correlate with spurious correlations**?**

Conclusions

- **Evaluator** role is the most demonstrated
  - Currently: Validates against **pre-existing field knowledge** = *Bottleneck*
  - Can **automated metrics** correlate with spurious correlations**?**

- **Multitasker** roles mainly applied to active site identification

Conclusions

- **Evaluator** role is the most demonstrated
  - Currently: Validates against **pre-existing field knowledge** = *Bottleneck*
  - Can **automated metrics** correlate with spurious correlations**?**
- **Multitasker** roles mainly applied to active site identification
- Few examples of **Coach** / **Engineer**

Conclusions

- **Evaluator** role is the most demonstrated
  - Currently: Validates against **pre-existing field knowledge** = *Bottleneck*
  - Can **automated metrics** correlate with spurious correlations**?**
- **Multitasker** roles mainly applied to active site identification
- Few examples of **Coach** / **Engineer**
  - An approach towards better models? Or a suboptimal alternative?

Conclusions

- **Evaluator** role is the most demonstrated
  - Currently: Validates against **pre-existing field knowledge** = *Bottleneck*
  - Can **automated metrics** correlate with spurious correlations**?**
- **Multitasker** roles mainly applied to active site identification
- Few examples of **Coach** / **Engineer**
  - An approach towards better models? Or a suboptimal alternative?
    Coach: Steering ***vs*** RL or fine tuning - *What is most effective?*

Conclusions

- **Evaluator** role is the most demonstrated
  - Currently: Validates against **pre-existing field knowledge** = *Bottleneck*
  - Can **automated metrics** correlate with spurious correlations**?**

- **Multitasker** roles mainly applied to active site identification

- Few examples of **Coach** / **Engineer**
  - An approach towards better models? Or a suboptimal alternative?
    Coach: Steering ***vs*** RL or fine tuning - *What is most effective?*
    Engineer: Pruning ***vs*** distillation

Conclusions

- **Evaluator** role is the most demonstrated
  - Currently: Validates against **pre-existing field knowledge** = *Bottleneck*
  - Can **automated metrics** correlate with spurious correlations**?**
- **Multitasker** roles mainly applied to active site identification
- Few examples of **Coach** / **Engineer**
  - An approach towards better models? Or a suboptimal alternative?
    Coach: Steering ***vs*** RL or fine tuning - *What is most effective?*
    Engineer: Pruning ***vs*** distillation
- **SAEs** can improve feature attribution: **Feature decoupling**

Conclusions

- **Evaluator** role is the most demonstrated
  - Currently: Validates against **pre-existing field knowledge** = *Bottleneck*
  - Can **automated metrics** correlate with spurious correlations**?**
- **Multitasker** roles mainly applied to active site identification
- Few examples of **Coach** / **Engineer**
  - An approach towards better models? Or a suboptimal alternative?
    Coach: Steering ***vs*** RL or fine tuning - *What is most effective?*
    Engineer: Pruning ***vs*** distillation
- **SAEs** can improve feature attribution: **Feature decoupling**
- Debatable **Teacher** roles

Conclusions

- **Evaluator** role is the most demonstrated
  - Currently: Validates against **pre-existing field knowledge** = *Bottleneck*
  - Can **automated metrics** correlate with spurious correlations**?**
- **Multitasker** roles mainly applied to active site identification
- Few examples of **Coach** / **Engineer**
  - An approach towards better models? Or a suboptimal alternative?
    Coach: Steering ***vs*** RL or fine tuning - *What is most effective?*
    Engineer: Pruning ***vs*** distillation
- **SAEs** can improve feature attribution: **Feature decoupling**
- Debatable **Teacher** roles
  - Currently: Identifying known chemistry in new examples

Conclusions

- **Evaluator** role is the most demonstrated
  - Currently: Validates against **pre-existing field knowledge** = *Bottleneck*
  - Can **automated metrics** correlate with spurious correlations**?**
- **Multitasker** roles mainly applied to active site identification
- Few examples of **Coach** / **Engineer**
  - An approach towards better models? Or a suboptimal alternative?
    Coach: Steering ***vs*** RL or fine tuning - *What is most effective?*
    Engineer: Pruning ***vs*** distillation
- **SAEs** can improve feature attribution: **Feature decoupling**
- Debatable **Teacher** roles
  - Currently: Identifying known chemistry in new examples
    How can AI produce new knowledge?

Conclusions

- **Evaluator** role is the most demonstrated
  - Currently: Validates against **pre-existing field knowledge** = *Bottleneck*
  - Can **automated metrics** correlate with spurious correlations**?**
- **Multitasker** roles mainly applied to active site identification
- Few examples of **Coach** / **Engineer**
  - An approach towards better models? Or a suboptimal alternative?
    Coach: Steering ***vs*** RL or fine tuning - *What is most effective?*
    Engineer: Pruning ***vs*** distillation
- **SAEs** can improve feature attribution: **Feature decoupling**
- Debatable **Teacher** roles
  - Currently: Identifying known chemistry in new examples
    How can AI produce new knowledge?

    The new knowledge is: { The output: New protein / new chess strategy **- Does not require interpretability**

Conclusions

- **Evaluator** role is the most demonstrated
  - Currently: Validates against **pre-existing field knowledge** = *Bottleneck*
  - Can **automated metrics** correlate with spurious correlations**?**
- **Multitasker** roles mainly applied to active site identification
- Few examples of **Coach** / **Engineer**
  - An approach towards better models? Or a suboptimal alternative?
    Coach: Steering ***vs*** RL or fine tuning - *What is most effective?*
    Engineer: Pruning ***vs*** distillation
- **SAEs** can improve feature attribution: **Feature decoupling**
- Debatable **Teacher** roles
  - Currently: Identifying known chemistry in new examples
    How can AI produce new knowledge?

    The new knowledge is:
    - The output: New protein / new chess strategy **- Does not require interpretability**
    - The reasoning **- Requires interpretability**

Conclusions

- **Evaluator** role is the most demonstrated
  - Currently: Validates against **pre-existing field knowledge** = *Bottleneck*
  - Can **automated metrics** correlate with spurious correlations**?**

- **Multitasker** roles mainly applied to active site identification

- Few examples of **Coach** / **Engineer**
  - An approach towards better models? Or a suboptimal alternative?
    Coach: Steering ***vs*** RL or fine tuning - *What is most effective?*
    Engineer: Pruning ***vs*** distillation

- **SAEs** can improve feature attribution: **Feature decoupling**

- Debatable **Teacher** roles
  - Currently: Identifying known chemistry in new examples
    How can AI produce new knowledge?

    The new knowledge is: { The output: New protein / new chess strategy **- Does not require interpretability**
    The reasoning **- Requires interpretability**

    Validating new reasoning

Conclusions

- **Evaluator** role is the most demonstrated
  - Currently: Validates against **pre-existing field knowledge** = *Bottleneck*
  - Can **automated metrics** correlate with spurious correlations**?**

- **Multitasker** roles mainly applied to active site identification

- Few examples of **Coach** / **Engineer**
  - An approach towards better models? Or a suboptimal alternative?
    Coach: Steering ***vs*** RL or fine tuning - *What is most effective?*
    Engineer: Pruning ***vs*** distillation

- **SAEs** can improve feature attribution: **Feature decoupling**

- Debatable **Teacher** roles
  - Currently: Identifying known chemistry in new examples
    How can AI produce new knowledge?

    The new knowledge is: { The output: New protein / new chess strategy **- Does not require interpretability**
    The reasoning **- Requires interpretability**

    Validating new reasoning
    - **Human supervision** of new hypothesis (human knowledge bottleneck)

Conclusions

- **Evaluator** role is the most demonstrated
  - Currently: Validates against **pre-existing field knowledge** = *Bottleneck*
  - Can **automated metrics** correlate with spurious correlations**?**

- **Multitasker** roles mainly applied to active site identification

- Few examples of **Coach** / **Engineer**
  - An approach towards better models? Or a suboptimal alternative?
    Coach: Steering ***vs*** RL or fine tuning - *What is most effective?*
    Engineer: Pruning ***vs*** distillation

- **SAEs** can improve feature attribution: **Feature decoupling**

- Debatable **Teacher** roles
  - Currently: Identifying known chemistry in new examples
    How can AI produce new knowledge?

    The new knowledge is:
    - The output: New protein / new chess strategy **- Does not require interpretability**
    - The reasoning **- Requires interpretability**

    Validating new reasoning
    - **Human supervision** of new hypothesis (human knowledge bottleneck)
    - Hypothesis **testing** (cost associated to testing)

Conclusions

- **Evaluator** role is the most demonstrated
  - Currently: Validates against **pre-existing field knowledge** = *Bottleneck*
  - Can **automated metrics** correlate with spurious correlations**?**

- **Multitasker** roles mainly applied to active site identification

- Few examples of **Coach** / **Engineer**
  - An approach towards better models? Or a suboptimal alternative?
    Coach: Steering ***vs*** RL or fine tuning - *What is most effective?*
    Engineer: Pruning ***vs*** distillation

- **SAEs** can improve feature attribution: **Feature decoupling**

- Debatable **Teacher** roles
  - Currently: Identifying known chemistry in new examples
    How can AI produce new knowledge?

    The new knowledge is:
    - The output: New protein / new chess strategy **- Does not require interpretability**
    - The reasoning **- Requires interpretability**

    Validating new reasoning
    - **Human supervision** of new hypothesis (human knowledge bottleneck)
    - Hypothesis **testing** (cost associated to testing)

Thank You!
Questions?