

# Predicting enzymatic function of protein sequences with attention

Journées du Réseau Inrae PEPI IBIS 2023

Nicolas Buton, François Coste and Yann Le Cunff

DYLISS Team - Univ Rennes, Inria, CNRS, IRISA



This work by F. Coste is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

# The Dyliss bioinformatics team

<http://www.irisa.fr/dyliss>



*Inria*

UMR IRISA



# Unlocking Nature's Code: The Linguistic Approach to Biological Sequences

- **Decoding Molecular Language**
  - DNA, RNA, proteins: Nature's intricate messages.
  - Viewing sequences through the lens of linguistics.
- **Revealing Hidden Patterns**
  - Unveiling motifs, structures, and meaningful repetitions.
  - Extracting concealed insights from sequences.
- **Predicting Protein Functions**
  - The challenge of discerning protein roles.
  - Harnessing linguistic principles for predictive power.
- **Bridging Disciplines**
  - Merging biology and linguistics to enrich understanding.
  - Impacting genomics, medicine, and diverse fields.

Quite a nice introduction

# Quite a nice introduction, quickly written

Default (GPT-3.5)


FR

Please write an introductory beamer slide on the linguistic approach applied to biological sequences, the last item will focus on protein function prediction (output only the frame)



Of course, here's the requested Beamer slide content:

latex

 Copy code

```
\begin{frame}
\frametitle{Unlocking Nature's Code: The Linguistic Approach to Biological S

\begin{itemize}
  \item \textbf{Decoding Molecular Language}
  \begin{itemize}
    \item DNA, RNA, proteins: Nature's intricate messages.
```

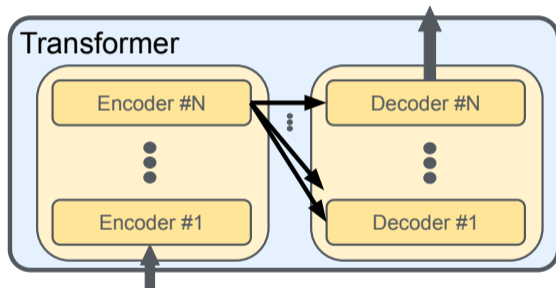
## Agenda

- An introduction to Transformers for Natural Languages Processing (NLP)
- Transformers for Biological Language Processing (BLP):

Predicting enzymatic function of sequences

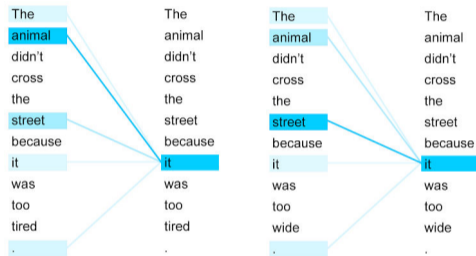
# Transformers: Attention Is All You Need<sup>1</sup>

L'animal n'a pas traversé la rue parce qu'il était trop fatigué.



The animal didn't cross the street because it was too tired.

## Self-attention



*The animal didn't cross the street because **it** was too tired.  
L'animal n'a pas traversé la rue parce qu'**il** était trop fatigué.*

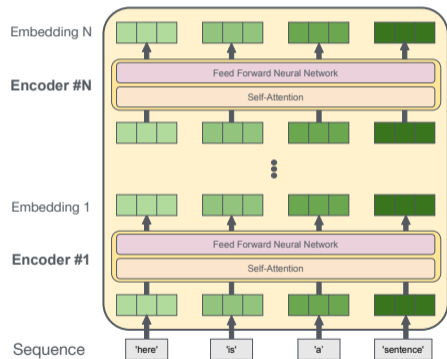
*The animal didn't cross the street because **it** was too wide.  
L'animal n'a pas traversé la rue parce qu'**elle** était trop large.*

Source: ai.googleblog.com

<sup>1</sup>A. Vaswani et al. [NIPS](#). 2017.

# BERT: Bidirectional Encoder Representations from Transformers<sup>2</sup>

Transformer encoder layers project the words into contextual embedding representations

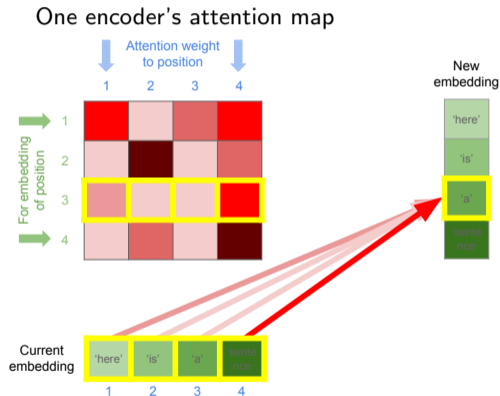
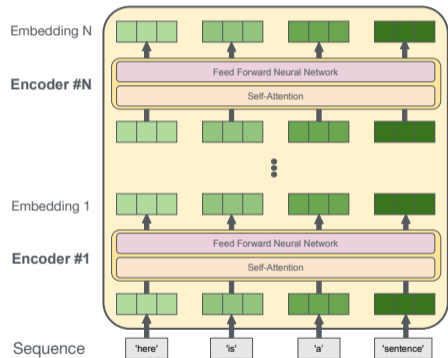


<sup>2</sup>J. Devlin et al. 2018.



# BERT: Bidirectional Encoder Representations from Transformers<sup>2</sup>

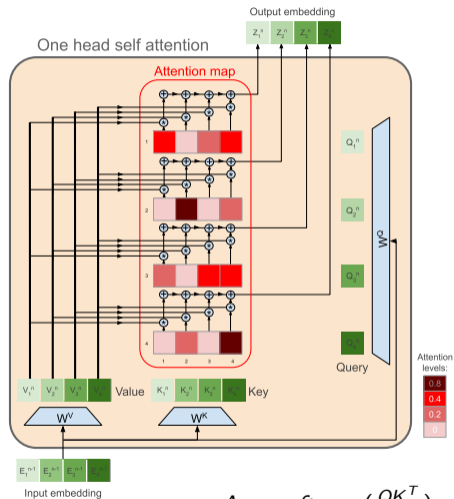
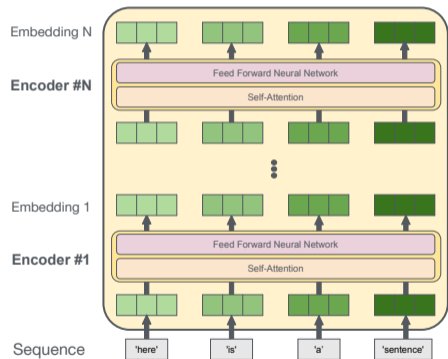
Transformer encoder layers project the words into contextual embedding representations



<sup>2</sup>J. Devlin et al. 2018.

# BERT: Bidirectional Encoder Representations from Transformers<sup>2</sup>

Transformer encoder layers project the words into contextual embedding representations

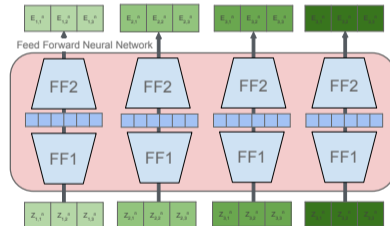
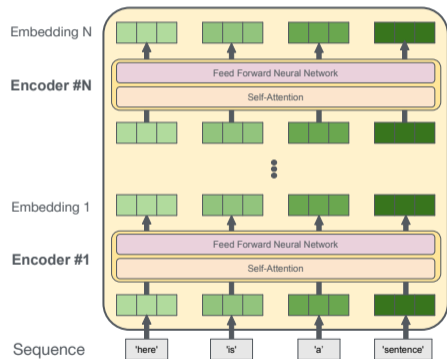


$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

<sup>2</sup>J. Devlin et al. 2018.

# BERT: Bidirectional Encoder Representations from Transformers<sup>2</sup>

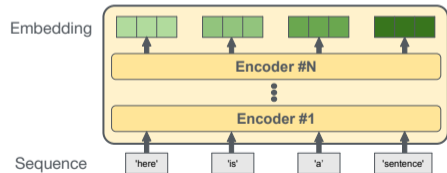
Transformer encoder layers project the words into contextual embedding representations



<sup>2</sup>J. Devlin et al. 2018.

# BERT: Bidirectional Encoder Representations from Transformers<sup>2</sup>

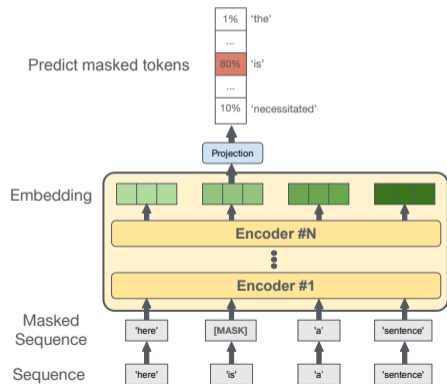
Transformer encoder layers project the words into contextual embedding representations



<sup>2</sup>J. Devlin et al. 2018.

# BERT: Bidirectional Encoder Representations from Transformers<sup>2</sup>

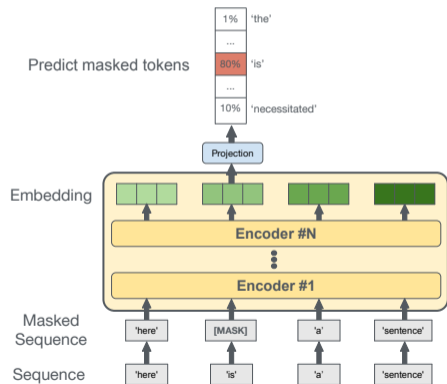
1. **Pre-training:** self-supervised learning of encoder from all the sequences of the domain



<sup>2</sup>J. Devlin et al. 2018.

# BERT: Bidirectional Encoder Representations from Transformers<sup>2</sup>

1. **Pre-training:** self-supervised learning of encoder from all the sequences of the domain



↪ different embeddings of 'bank' in:

## Sentences

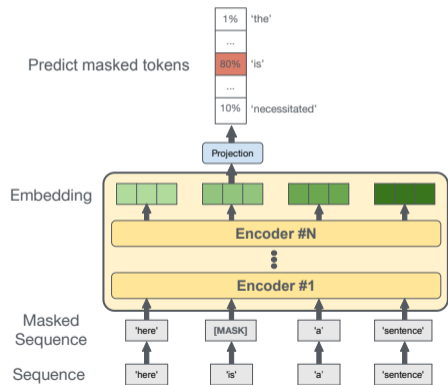
- "bank"
- "he eventually sold the shares back to the bank at a premium."
- "the bank strongly resisted cutting interest rates."
- "the bank will supply and buy back foreign currency."
- "the bank is pressing us for repayment of the loan."
- "the bank left its lending rates unchanged."
- "the river flowed over the bank."
- "tall, luxuriant plants grew along the river bank."
- "his soldiers were arrayed along the river bank."
- "wild flowers adorned the river bank."
- "two fox cubs romped playfully on the river bank."
- "the jewels were kept in a bank vault."
- "you can stow your jewellery away in the bank."
- "most of the money was in storage in bank vaults."
- "the diamonds are shut away in a bank vault somewhere."
- "thieves broke into the bank vault."
- "can I bank on your support?"
- "you can bank on him to hand you a reasonable bill for your services."
- "don't bank on your friends to help you out of trouble."
- "you can bank on me when you need money."
- "i bank on your help."

Source: r3d\_robot

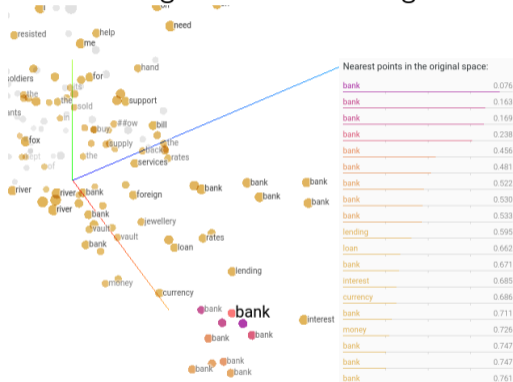
<sup>2</sup>J. Devlin et al. 2018.

# BERT: Bidirectional Encoder Representations from Transformers<sup>2</sup>

1. **Pre-training:** self-supervised learning of encoder from all the sequences of the domain



Embeddings of 'bank' and neighbors:

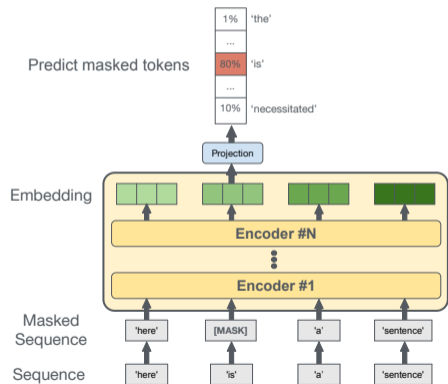


Financial bank  
Source: r3d\_robot

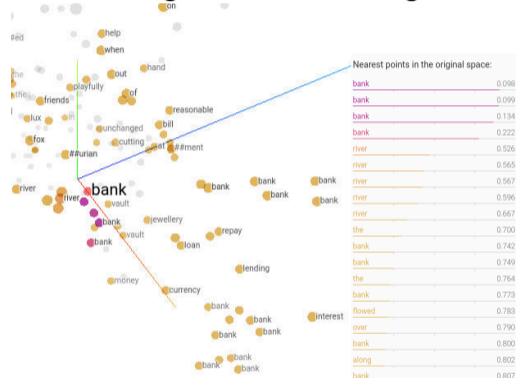
<sup>2</sup>J. Devlin et al. 2018.

# BERT: Bidirectional Encoder Representations from Transformers<sup>2</sup>

1. **Pre-training:** self-supervised learning of encoder from all the sequences of the domain



Embeddings of 'bank' and neighbors:



River bank

Source: r3d\_robot

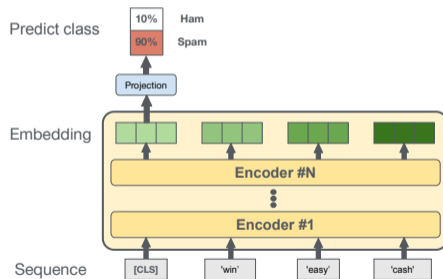
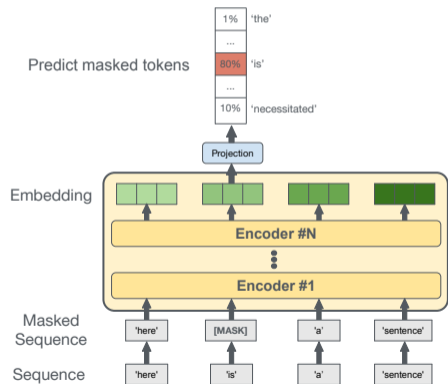
<sup>2</sup>J. Devlin et al. 2018.



# BERT: Bidirectional Encoder Representations from Transformers<sup>2</sup>

1. **Pre-training**: self-supervised learning of encoder from all the sequences of the domain

2. **Fine-tuning**: supervised learning of a specific task from input-output examples



<sup>2</sup>J. Devlin et al. 2018.

# From NLP to protein function prediction

- Encoders: new **state-of-the-art in NLP** (BERT<sup>3</sup>, GPT3<sup>4</sup>, T5<sup>5</sup>,...)
- Attention for proteins?
  - Tasks Assessing Protein Embeddings (TAPE)<sup>6</sup>  
Secondary structure, contact, homology, fluorescence, stability prediction
  - Protein language models ESM-1b<sup>7</sup>, ProtTrans<sup>8</sup>  
Remote homology, secondary structure, long-range residue-residue contacts, mutational effect, sub-cellular location, membrane vs water-soluble
  - AlphaFold2<sup>9</sup>  
Structure prediction
  - **Function prediction???** in 2022!

---

<sup>3</sup>J. Devlin et al. 2018.

<sup>4</sup>T. B. Brown et al. 2020.

<sup>5</sup>C. Raffel et al. arXiv 2020.

<sup>6</sup>R. Rao et al. NIPS. 2019.

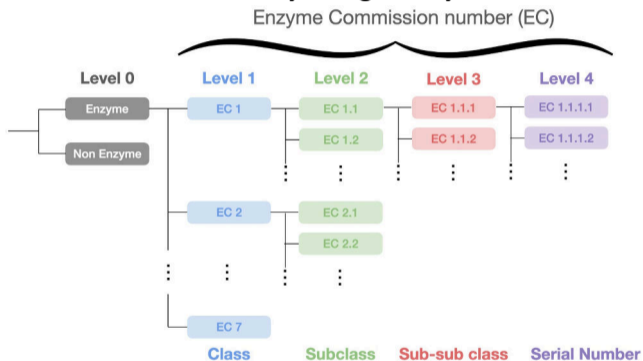
<sup>7</sup>A. Rives et al. Proceedings of the National Academy of Sciences 2021.

<sup>8</sup>A. Elnaggar et al. IEEE Trans. Pattern Anal. Mach. Intell. 2022.

<sup>9</sup>J. Jumper et al. Nature 2021.

# Our study: attention for enzymatic function prediction?

- Enzymes are essential catalysts of chemical reactions in biological systems
- Well studied protein function, classified by 4 digit Enzyme Commission (EC) number:

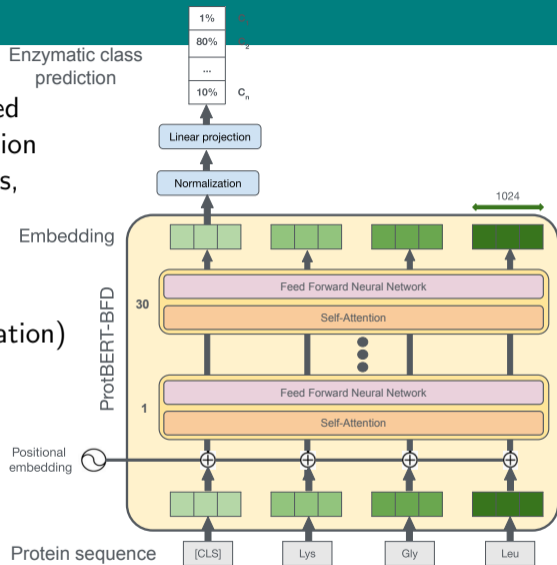


- State-of-the-art predictors from sequences only (in 2022):
  - ECPred<sup>10</sup> (classical machine learning ensemble, levels 0–4)
  - UDSMProt<sup>11</sup> (Bi-LSTM, levels 1–2)

Picture by N. Buton

- Use a **pre-trained** encoder that was trained through self-supervised learning on 2.5 billion protein sequences from reference databases, metagenomes and metatranscriptomes: **ProtBERT-BFD**<sup>12</sup>
- Fine-tune** it on protein sequences labeled by their EC numbers (no hierarchy information) using cross-entropy loss function

2-stages learning enabling the prediction of EC classes with very few proteins



<sup>12</sup>A. Elnaggar et al. *IEEE Trans. Pattern Anal. Mach. Intell.* 2022.

<sup>13</sup>N. Buton, F. Coste, and Y. Le Cunff. *submitted* 2022.

# Evaluation of EnzBert for enzymatic function prediction

- Benchmarks

- For comparison with UDSMProt: **EC40**<sup>14</sup>
- For comparison with ECPred<sup>15</sup>: **ECPred40**, similar to ECPred time-based evaluation, but ensuring < 40% sequence identity between test and training/validation sets

- Predictions on test set

Prediction	Level	#class	Accuracy
UDSMProt	1	6	0.87
<b>EnzBert<sub>EC40</sub></b>	1	6	<b>0.97</b>
UDSMProt	2	65	0.84
<b>EnzBert<sub>EC40</sub></b>	2	65	<b>0.95</b>

Prediction	Level	#class	Macro-f1
ECPred	0	2	0.77
<b>EnzBert<sub>ECPred40</sub></b>	0	2	<b>0.84</b>
ECPred	4	634	0.41
<b>EnzBert<sub>ECPred40</sub></b>	4	634	<b>0.55</b>

Significant improvement of EnzBert upon UDSMProt and ECPred

<sup>14</sup>N. Strodthoff et al. [Bioinformatics](#) 2020.

<sup>15</sup>A. Dalkiran et al. [BMC Bioinformatics](#) 2018.

# Evaluation of EnzBert for enzymatic function prediction

- Benchmarks

- For comparison with UDSMProt: **EC40**<sup>14</sup>
- For comparison with ECPred<sup>15</sup>: **ECPred40**, similar to ECPred time-based evaluation, but ensuring < 40% sequence identity between test and training/validation sets

- Predictions on test set

Prediction	Level	#class	Accuracy
UDSMProt	1	6	0.87
<b>EnzBert<sub>EC40</sub></b>	1	6	<b>0.97</b>
UDSMProt	2	65	0.84
<b>EnzBert<sub>EC40</sub></b>	2	65	<b>0.95</b>

Prediction	Level	#class	Macro-f1
ECPred	0	2	0.77
<b>EnzBert<sub>ECPred40</sub></b>	0	2	<b>0.84</b>
ECPred	4	634	0.41
<b>EnzBert<sub>ECPred40</sub></b>	4	634	<b>0.55</b>

Significant improvement of EnzBert upon UDSMProt and ECPred

Does attention also help understanding predictions?

<sup>14</sup>N. Strodthoff et al. [Bioinformatics](#) 2020.

<sup>15</sup>A. Dalkiran et al. [BMC Bioinformatics](#) 2018.

# Attention map: a built-in mechanism to see important residues

Specific type of interpretability: **Residues importance scores**

- associate one real value per residues
- higher value correspond to more important residues

0 MSMQEKIMRE LHVKPSIDPK QEIEDRVNFL KQYVKKTGAK GFVLGI **SGGQ** **D**STLAGRLAQ LAVESIREEG GDAQFIAVRL PHG  
1 DKSWKFDIKS TVSAFSDQYQ QETGDQLTDF NKGNVKARTR MIAQYAIGGQ EGLLVL **GTDH** **AAE**AVTGFFT KYGDGGADLL PLT  
2 ERLYLKEPTA DLLDEKPQQS DETELGIS **YD** EIDDYLEGKE VSAKVSEALE KRYSMTEHKR QVPASMFDDW WK

Can be derived from attentions maps, but...

# Attention map: a built-in mechanism to see important residues

Specific type of interpretability: **Residues importance scores**

- associate one real value per residues
- higher value correspond to more important residues

0 MSMQEKIMRE LHVKPSIDPK QEIEDRVNFL KQYVKKTGAK GFVLGI SGGQ DSTLAGRLAQ LAVESIREEG GDAQFIAVRL PHG  
1 DKSWKFDIKS TVSAFSDQYQ QETGDQLTDF NKGNVKARTR MIAQYAIGGQ EGLLVL GTDH AAEAVTGFFT KYGDGGADLL PLT  
2 ERLYLKEPTA DLLDEKPQQS DETELGIS YD EIDDYLEGKE VSAKVSEALE KRYSMTEHKR QVPASMFDDW WK

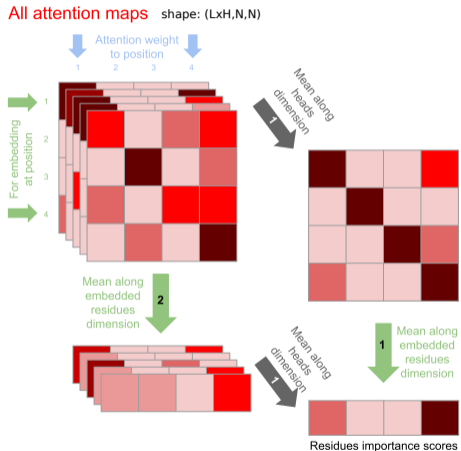
Can be derived from attentions maps, but...

...there are many attention maps!

In Enzbert: 30 layers  $\times$  16 heads per layer = 480 attention maps



# Interpretability from multi-heads attention?



AttnAgg2A1A: dim2 Average, dim1 Average  
attention aggregation method

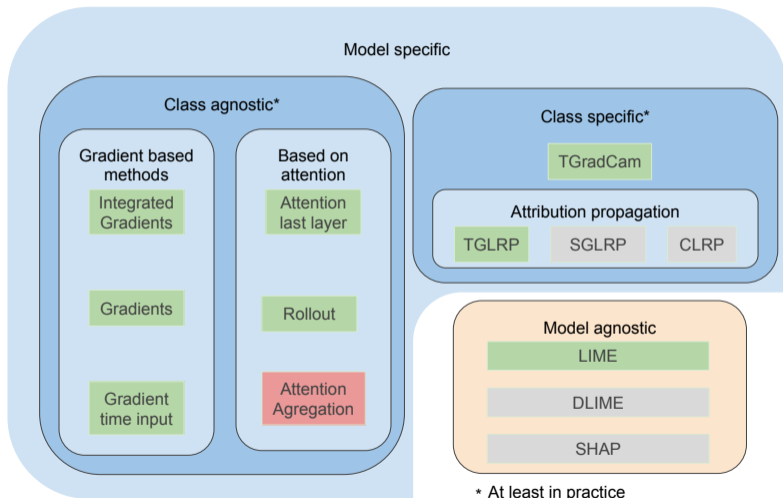
Combine the attention maps  
to derive a score per residue

But multiple way to aggregate:

- Pooling strategy: Maximum, Average
- Order of the operators: start with dimension 1, 2 or 3

13 different possibilities to evaluate  
(16 before removing duplicates)

# Comparison with other interpretability methods



Green (and red):  
methods evaluated in this  
study

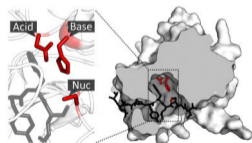
# How to evaluate interpretability methods?

```
0 MSMQEKIMRE LHVKPSIDPK QEIEDRVNFL KQYVKKTGAK GFVLGISGGQ DSTLAGRLAQ LAVESIREEG GDAQFIAVRL PHGTQQDEDD AQLALKFIKP
1 DKSWKFDIKS TVSAFSDQYQ QETGDQLTDF NKGNVKARTR MIAQYAIGGQ EGLLVLGTDHA AEAVTGGFFT KYGDGGADLL PLTGLTKRQG RTLLKELGAP
2 ERLYLKEPTA DLLDEKPPQS DETELGISVD EIDDYLEGKE VSAKVSEALE KRYSMTEHKR QVPASMFDDW WK
```

## Evaluation proxy

Sort residues by importance: evaluate retrieval of catalytic sites (~1% of all residues in enzyme)

Catalytic residues: amino acids directly involved in chemical catalysis



**Figure:** Catalytic triad of TEV protease  
source: T. Shafee (2014)

**Our reference:** residues annotated as catalytic sites from 963 enzymes by Mechanism and Catalytic Site Atlas (M-CSA) database

# Interpretability results

Method type	PRG-AUC(x100)	max F-Gain(%)	Time(s)
Random	42.54 ± 4.37	69.85 ± 1.04	–
Grad	75.01	81.27	4.64
Grad X input	63.62	78.66	7.74
Integrated grad	76.41	81.70	$2.48 \times 10^2$
Attn last layer	87.80	85.62	<b>2.87</b>
<b>AttnAgg2A1A</b>	<b>98.02</b>	<b>96.05</b>	3.72
Rollout	66.08	76.77	2.95
TGLRP	90.92	88.56	$4.05 \times 10^1$
TGradCam	81.00	76.77	$4.35 \times 10^1$
LIME	93.46	91.44	$1.73 \times 10^4$

## PRG-AUC

- Area under the precision recall gain curve
- Precision and recall **gain** are rescaled precision recall
- Higher is better

See P. A. Flach and M. Kull. [NIPS](#). 2015

# Interpretability results

Method type	PRG-AUC(x100)	max F-Gain(%)	Time(s)
Random	42.54 ± 4.37	69.85 ± 1.04	–
Grad	75.01	81.27	4.64
Grad X input	63.62	78.66	7.74
Integrated grad	76.41	81.70	$2.48 \times 10^2$
Attn last layer	87.80	85.62	<b>2.87</b>
<b>AttnAgg2A1A</b>	<b>98.02</b>	<b>96.05</b>	3.72
Rollout	66.08	76.77	2.95
TGLRP	90.92	88.56	$4.05 \times 10^1$
TGradCam	81.00	76.77	$4.35 \times 10^1$
LIME	93.46	91.44	$1.73 \times 10^4$

## max F-Gain

- F-Gain is a rescaled F1 score (average precision and recall)
- Maximum for all possible threshold on the scores
- Higher is better

See P. A. Flach and M. Kull. [NIPS](#). 2015

# Interpretability results

Method type	PRG-AUC(x100)	max F-Gain(%)	Time(s)
Random	42.54 ± 4.37	69.85 ± 1.04	–
Grad	75.01	81.27	4.64
Grad X input	63.62	78.66	7.74
Integrated grad	76.41	81.70	$2.48 \times 10^2$
Attn last layer	87.80	85.62	<b>2.87</b>
<b>AttnAgg2A1A</b>	<b>98.02</b>	<b>96.05</b>	3.72
Rollout	66.08	76.77	2.95
TGLRP	90.92	88.56	$4.05 \times 10^1$
TGradCam	81.00	76.77	$4.35 \times 10^1$
LIME	93.46	91.44	$1.73 \times 10^4$

## Time(s)

- Time to execute on a CPU

# Interpretability results

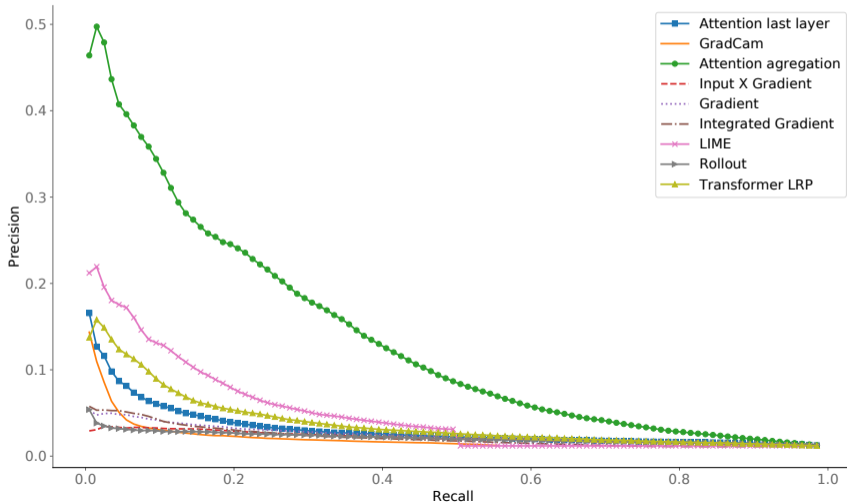
Method type	PRG-AUC(x100)	max F-Gain(%)	Time(s)
Random	42.54 ± 4.37	69.85 ± 1.04	–
Grad	75.01	81.27	4.64
Grad X input	63.62	78.66	7.74
Integrated grad	76.41	81.70	$2.48 \times 10^2$
Attn last layer	87.80	85.62	<b>2.87</b>
<b>AttnAgg2A1A</b>	<b>98.02</b>	<b>96.05</b>	3.72
Rollout	66.08	76.77	2.95
TGLRP	90.92	88.56	$4.05 \times 10^1$
TGradCam	81.00	76.77	$4.35 \times 10^1$
LIME	93.46	91.44	$1.73 \times 10^4$

## Time(s)

- Time to execute on a CPU

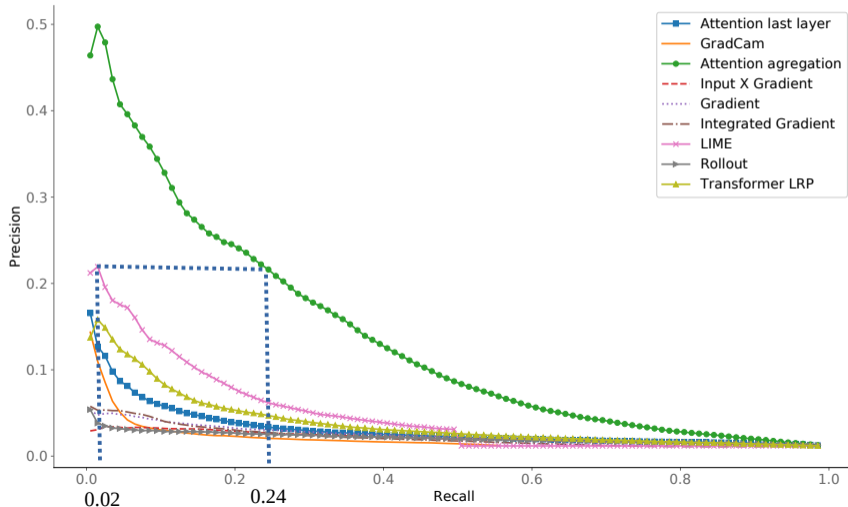
*Attention-based methods close to prediction time.*

# Interpretability results

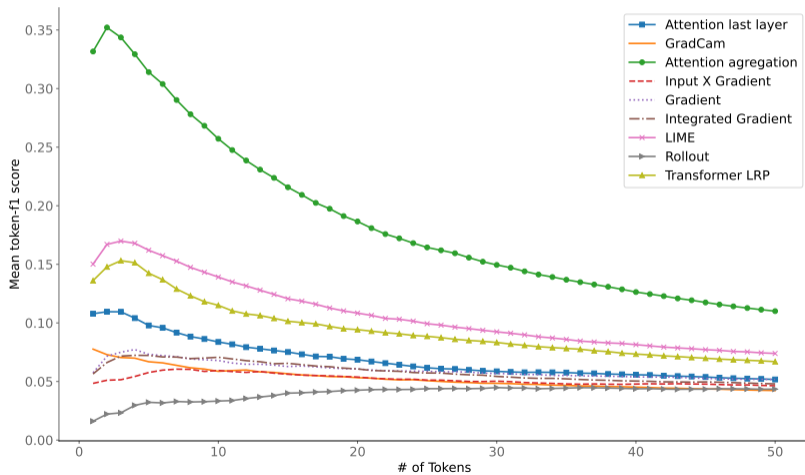




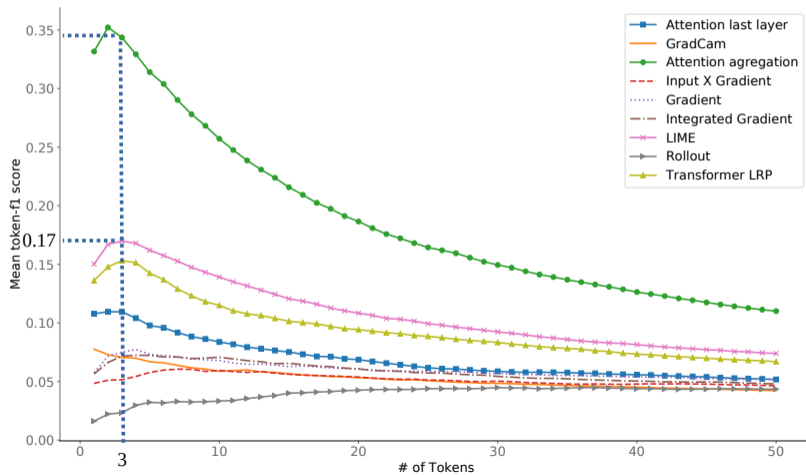
# Interpretability results



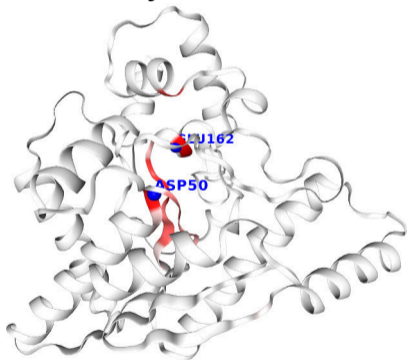
# Interpretability results



# Interpretability results

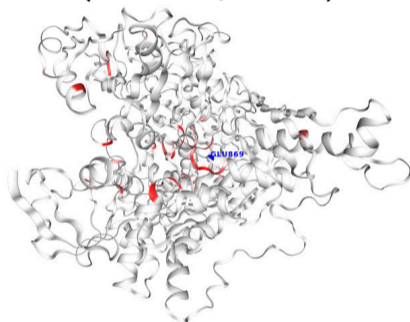


(a) Nh(3)-dependent nad(+) synthetase



0 MSBDEKIRH LHWPSIDPK GEIEDRWFL KQVWKTGK GPVLG **1** **2** LAGRLAG LAVESIREEG IDAQIAVRL PHTQDDED AQLALKFKP  
 1 DKSWPBDKS TVSAFSDQY QETADQLTDF INGVNKAIR IKAQYAIQQG EGLLY **3** **4** HTGFTT KYGGGADLL PLTGLTKRQ RTLLKELGAP  
 2 ERLYLKEPTA DLLLDEKPOOS DETELG **5** EIDDYLEGRE VSAKVSEALE KRYSPTEHRK QVPSMFDQW NK

(b) Aldehyde dehydrogenase (FAD-independent)



0 PDKQVTVNG IEDNLFVDAE ALLSDPLRQ LGLTVQVYV **1** **2** **3** **4** **5** ILDGKVVH VTKRNRVADG AQITTIIEGV QFNALPFLK ARVLRHGAS  
 1 GI **6** **7** **8** **9** **10** **11** **12** **13** **14** **15** **16** **17** **18** **19** **20** **21** **22** **23** **24** **25** **26** **27** **28** **29** **30** **31** **32** **33** **34** **35** **36** **37** **38** **39** **40** **41** **42** **43** **44** **45** **46** **47** **48** **49** **50** **51** **52** **53** **54** **55** **56** **57** **58** **59** **60** **61** **62** **63** **64** **65** **66** **67** **68** **69** **70** **71** **72** **73** **74** **75** **76** **77** **78** **79** **80** **81** **82** **83** **84** **85** **86** **87** **88** **89** **90** **91** **92** **93** **94** **95** **96** **97** **98** **99** **100** **101** **102** **103** **104** **105** **106** **107** **108** **109** **110** **111** **112** **113** **114** **115** **116** **117** **118** **119** **120** **121** **122** **123** **124** **125** **126** **127** **128** **129** **130** **131** **132** **133** **134** **135** **136** **137** **138** **139** **140** **141** **142** **143** **144** **145** **146** **147** **148** **149** **150** **151** **152** **153** **154** **155** **156** **157** **158** **159** **160** **161** **162** **163** **164** **165** **166** **167** **168** **169** **170** **171** **172** **173** **174** **175** **176** **177** **178** **179** **180** **181** **182** **183** **184** **185** **186** **187** **188** **189** **190** **191** **192** **193** **194** **195** **196** **197** **198** **199** **200** **201** **202** **203** **204** **205** **206** **207** **208** **209** **210** **211** **212** **213** **214** **215** **216** **217** **218** **219** **220** **221** **222** **223** **224** **225** **226** **227** **228** **229** **230** **231** **232** **233** **234** **235** **236** **237** **238** **239** **240** **241** **242** **243** **244** **245** **246** **247** **248** **249** **250** **251** **252** **253** **254** **255** **256** **257** **258** **259** **260** **261** **262** **263** **264** **265** **266** **267** **268** **269** **270** **271** **272** **273** **274** **275** **276** **277** **278** **279** **280** **281** **282** **283** **284** **285** **286** **287** **288** **289** **290** **291** **292** **293** **294** **295** **296** **297** **298** **299** **300** **301** **302** **303** **304** **305** **306** **307** **308** **309** **310** **311** **312** **313** **314** **315** **316** **317** **318** **319** **320** **321** **322** **323** **324** **325** **326** **327** **328** **329** **330** **331** **332** **333** **334** **335** **336** **337** **338** **339** **340** **341** **342** **343** **344** **345** **346** **347** **348** **349** **350** **351** **352** **353** **354** **355** **356** **357** **358** **359** **360** **361** **362** **363** **364** **365** **366** **367** **368** **369** **370** **371** **372** **373** **374** **375** **376** **377** **378** **379** **380** **381** **382** **383** **384** **385** **386** **387** **388** **389** **390** **391** **392** **393** **394** **395** **396** **397** **398** **399** **400** **401** **402** **403** **404** **405** **406** **407** **408** **409** **410** **411** **412** **413** **414** **415** **416** **417** **418** **419** **420** **421** **422** **423** **424** **425** **426** **427** **428** **429** **430** **431** **432** **433** **434** **435** **436** **437** **438** **439** **440** **441** **442** **443** **444** **445** **446** **447** **448** **449** **450** **451** **452** **453** **454** **455** **456** **457** **458** **459** **460** **461** **462** **463** **464** **465** **466** **467** **468** **469** **470** **471** **472** **473** **474** **475** **476** **477** **478** **479** **480** **481** **482** **483** **484** **485** **486** **487** **488** **489** **490** **491** **492** **493** **494** **495** **496** **497** **498** **499** **500** **501** **502** **503** **504** **505** **506** **507** **508** **509** **510** **511** **512** **513** **514** **515** **516** **517** **518** **519** **520** **521** **522** **523** **524** **525** **526** **527** **528** **529** **530** **531** **532** **533** **534** **535** **536** **537** **538** **539** **540** **541** **542** **543** **544** **545** **546** **547** **548** **549** **550** **551** **552** **553** **554** **555** **556** **557** **558** **559** **560** **561** **562** **563** **564** **565** **566** **567** **568** **569** **570** **571** **572** **573** **574** **575** **576** **577** **578** **579** **580** **581** **582** **583** **584** **585** **586** **587** **588** **589** **590** **591** **592** **593** **594** **595** **596** **597** **598** **599** **600** **601** **602** **603** **604** **605** **606** **607** **608** **609** **610** **611** **612** **613** **614** **615** **616** **617** **618** **619** **620** **621** **622** **623** **624** **625** **626** **627** **628** **629** **630** **631** **632** **633** **634** **635** **636** **637** **638** **639** **640** **641** **642** **643** **644** **645** **646** **647** **648** **649** **650** **651** **652** **653** **654** **655** **656** **657** **658** **659** **660** **661** **662** **663** **664** **665** **666** **667** **668** **669** **670** **671** **672** **673** **674** **675** **676** **677** **678** **679** **680** **681** **682** **683** **684** **685** **686** **687** **688** **689** **690** **691** **692** **693** **694** **695** **696** **697** **698** **699** **700** **701** **702** **703** **704** **705** **706** **707** **708** **709** **710** **711** **712** **713** **714** **715** **716** **717** **718** **719** **720** **721** **722** **723** **724** **725** **726** **727** **728** **729** **730** **731** **732** **733** **734** **735** **736** **737** **738** **739** **740** **741** **742** **743** **744** **745** **746** **747** **748** **749** **750** **751** **752** **753** **754** **755** **756** **757** **758** **759** **760** **761** **762** **763** **764** **765** **766** **767** **768** **769** **770** **771** **772** **773** **774** **775** **776** **777** **778** **779** **780** **781** **782** **783** **784** **785** **786** **787** **788** **789** **790** **791** **792** **793** **794** **795** **796** **797** **798** **799** **800** **801** **802** **803** **804** **805** **806** **807** **808** **809** **810** **811** **812** **813** **814** **815** **816** **817** **818** **819** **820** **821** **822** **823** **824** **825** **826** **827** **828** **829** **830** **831** **832** **833** **834** **835** **836** **837** **838** **839** **840** **841** **842** **843** **844** **845** **846** **847** **848** **849** **850** **851** **852** **853** **854** **855** **856** **857** **858** **859** **860** **861** **862** **863** **864** **865** **866** **867** **868** **869** **870** **871** **872** **873** **874** **875** **876** **877** **878** **879** **880** **881** **882** **883** **884** **885** **886** **887** **888** **889** **890** **891** **892** **893** **894** **895** **896** **897** **898** **899** **900** **901** **902** **903** **904** **905** **906** **907** **908** **909** **910** **911** **912** **913** **914** **915** **916** **917** **918** **919** **920** **921** **922** **923** **924** **925** **926** **927** **928** **929** **930** **931** **932** **933** **934** **935** **936** **937** **938** **939** **940** **941** **942** **943** **944** **945** **946** **947** **948** **949** **950** **951** **952** **953** **954** **955** **956** **957** **958** **959** **960** **961** **962** **963** **964** **965** **966** **967** **968** **969** **970** **971** **972** **973** **974** **975** **976** **977** **978** **979** **980** **981** **982** **983** **984** **985** **986** **987** **988** **989** **990** **991** **992** **993** **994** **995** **996** **997** **998** **999** **1000**

## High potential of attention and two-stage training for protein function prediction

- New state-of-the-art, *EnzBert*, for the prediction of enzyme's precise function from sequences only (! new tools since this study: CLEAN<sup>16</sup>, ...)
- Simple yet successful interpretability methods can be derived directly from attention maps

## Perspectives

- Examine other residues identified as important, besides the catalytic sites, with a 3D point of view
- Take into account and exploit the EC hierarchy
- Generalize from EC to Gene Ontology (GO) Molecular Function (MF)
- Can we learn cheaper models with explicit dependencies?

---

<sup>16</sup>T. Yu et al. [Science \(New York, N.Y.\)](#) 2023.

## High potential of attention and two-stage training for protein function prediction

- New state-of-the-art, *EnzBert*, for the prediction of enzyme's precise function from sequences only (! new tools since this study: CLEAN<sup>16</sup>, ...)
- Simple yet successful interpretability methods can be derived directly from attention maps

## Perspectives

- Examine other residues identified as important, besides the catalytic sites, with a 3D point of view
- Take into account and exploit the EC hierarchy
- Generalize from EC to Gene Ontology (GO) Molecular Function (MF)
- Can we learn cheaper models with explicit dependencies?

PhD Defense of Nicolas: Oct 18 2023!

---

<sup>16</sup>T. Yu et al. *Science* (New York, N.Y.) 2023.

Thanks for YOUR Attention

Any questions?

- [Vas+17] A. Vaswani et al. “Attention is All you Need”. [NIPS](#). 2017.
- [Dev+18] J. Devlin et al.  
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding 2018.
- [Bro+20] T. B. Brown et al. Language Models are Few-Shot Learners. 2020.
- [Raf+20] C. Raffel et al. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. [arXiv](#) 2020.
- [Rao+19] R. Rao et al. “Evaluating Protein Transfer Learning with TAPE”. [NIPS](#). 2019.
- [Riv+21] A. Rives et al. “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences”.  
Proceedings of the National Academy of Sciences 2021.



## References II

- [Eln+22] A. Elnaggar et al. “ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning”. [IEEE Trans. Pattern Anal. Mach. Intell.](#) 2022.
- [Jum+21] J. Jumper et al. “Highly accurate protein structure prediction with AlphaFold”. [Nature](#) 2021.
- [Dal+18] A. Dalkiran et al. “ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature”. [BMC Bioinformatics](#) 2018.
- [Str+20] N. Strodthoff et al. “UDSMProt: universal deep sequence models for protein classification”. [Bioinformatics](#) 2020.
- [BCL22] N. Buton, F. Coste, and Y. Le Cunff. “Predicting enzymatic function of protein sequences with attention”. [submitted](#) 2022.
- [FK15] P. A. Flach and M. Kull. “Precision-Recall-Gain curves: PR analysis done right”. [NIPS](#). 2015.

- [Yu+23] T. Yu et al. “Enzyme function prediction using contrastive learning”. [Science \(New York, N.Y.\)](#) 2023.