

Molecular language models

Raul Fernandez-Diaz

What are language models?

Natural language processing basics

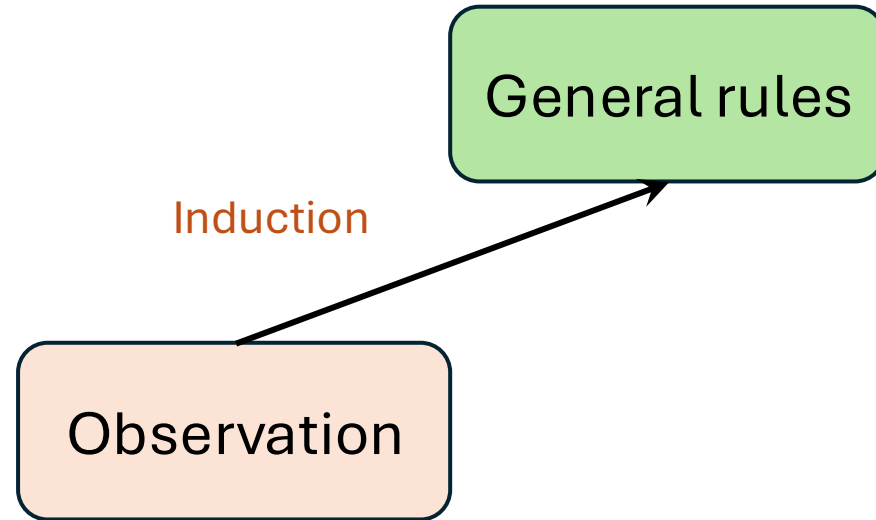
Natural language

The patient complaints of severe Severity left-sided Location chest Anatomy pain Problem. He underwent angioplasty Procedure & had 2 stents Medical Device placed a year ago. His BP Body Measurement and cardiac enzymes test Lab Data were normal. He is on aspirin & plavix Medicine. He has a history of alcohol & marijuana abuse Substance Abuse.

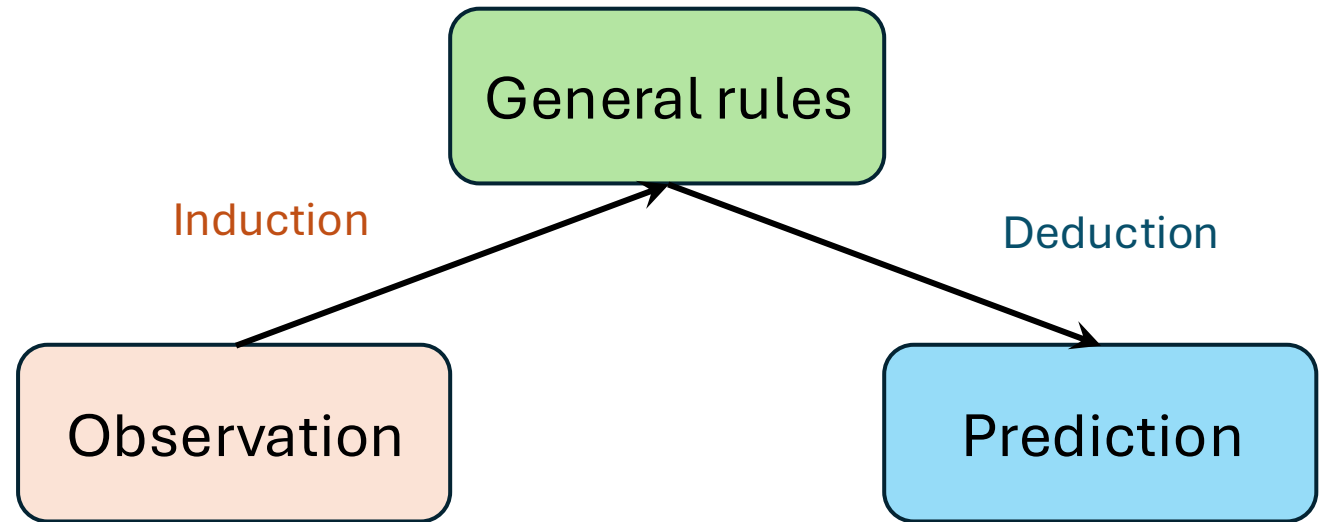
How do we learn language?



How do we learn language?



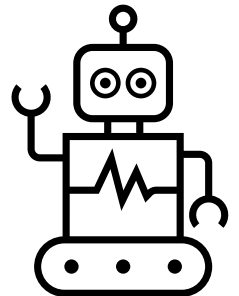
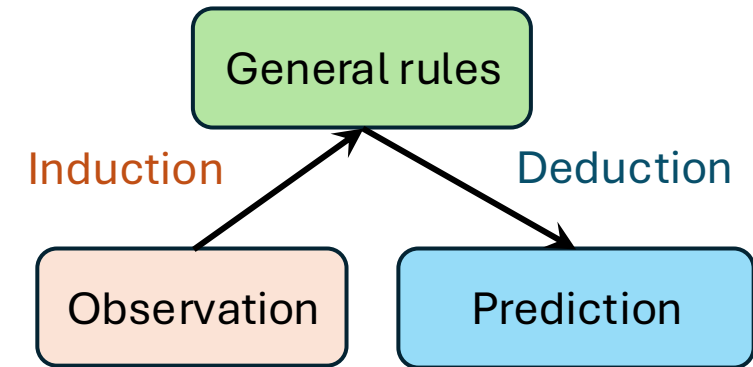
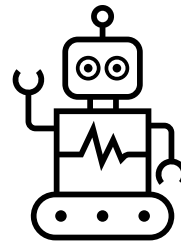
How do we learn language?



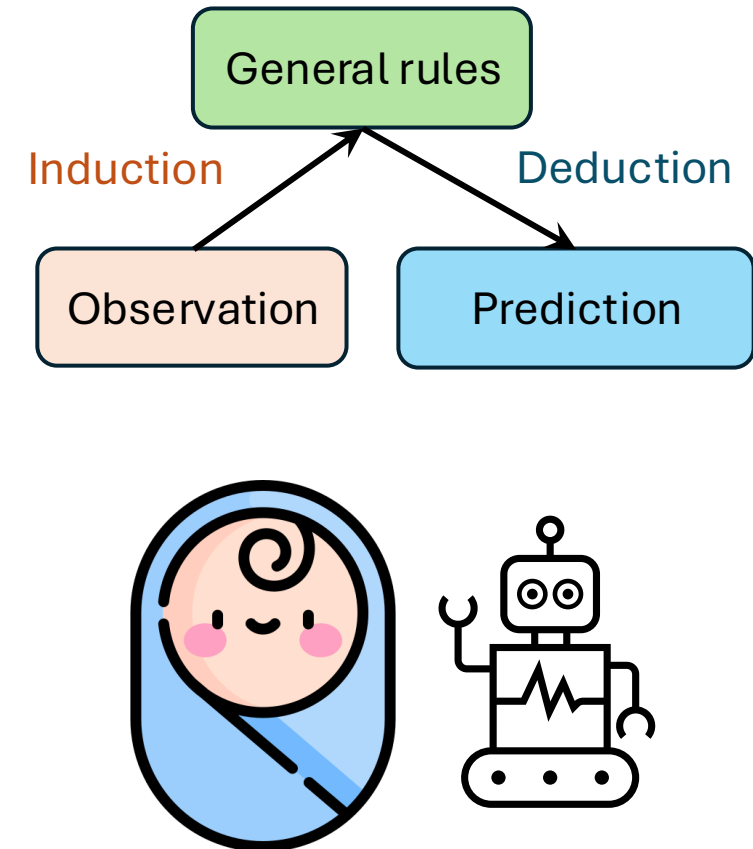
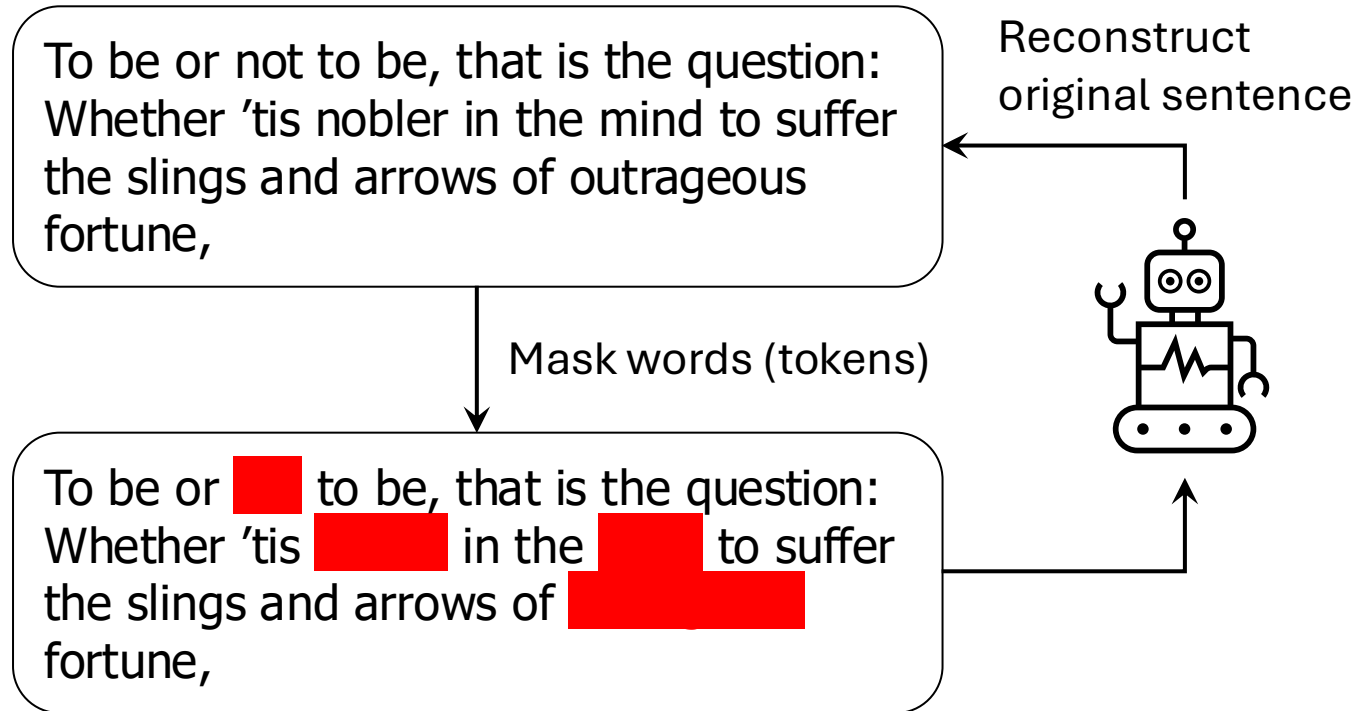
- **Induction**: Process of inferring general rules from discrete observations
- **Deduction**: Process of deriving discrete predictions from general rules

How can we teach a model languages?

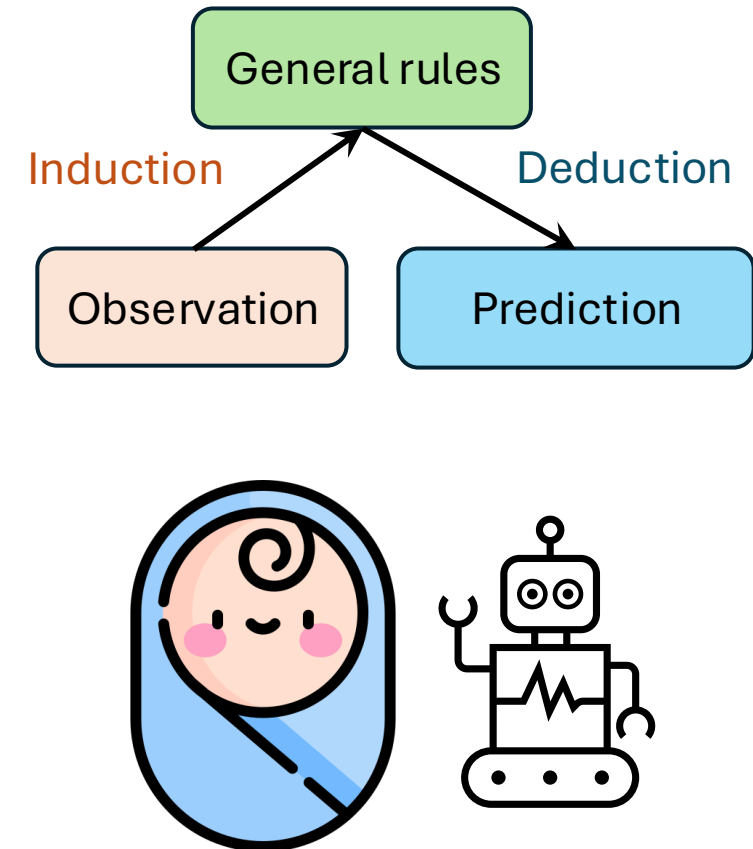
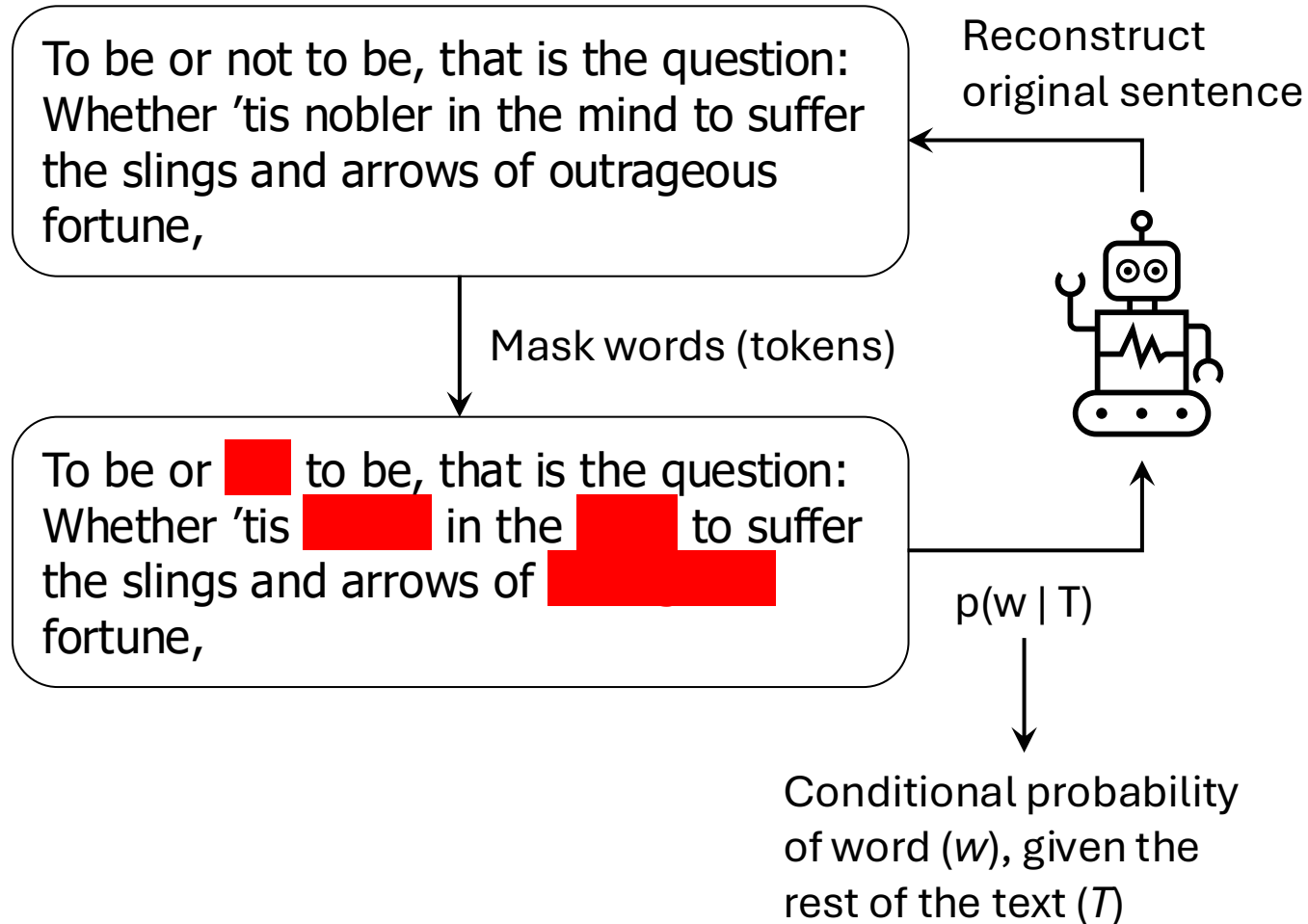
To be or not to be, that is the question:
Whether 'tis nobler in the mind to suffer
the slings and arrows of outrageous
fortune,



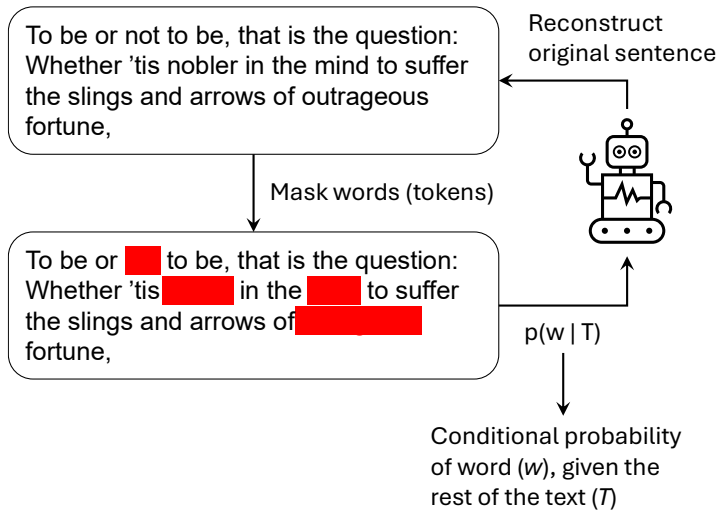
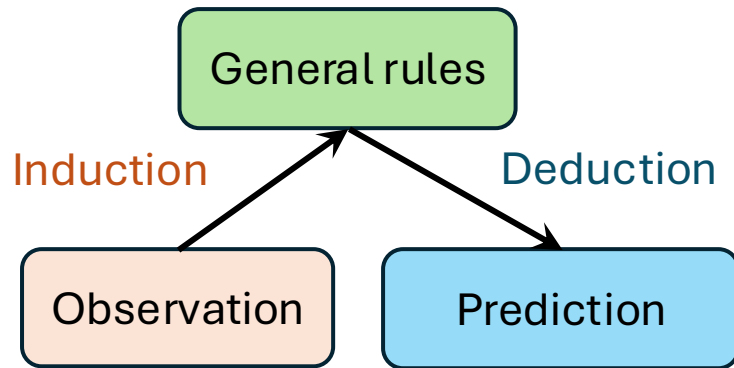
How can we teach a model languages?



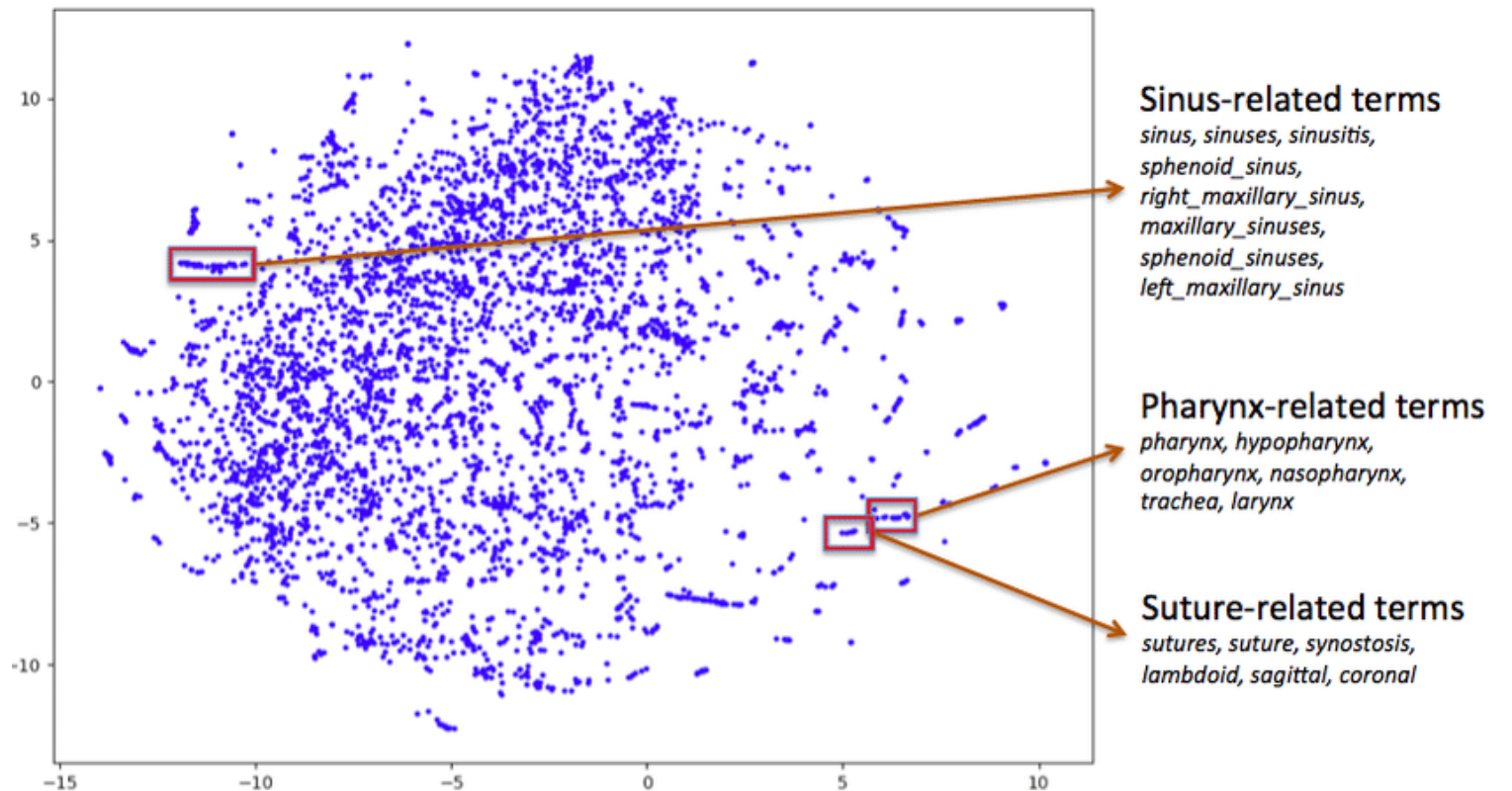
How can we teach a model languages?



Words are embedded as vectors



2D representation of multidimensional space



Similar words are close together in the embedding/representation space

Resources

Model comparison

- <https://arena.ai/>

Language modelling visual explanation

- <https://www.youtube.com/watch?v=LPZh9BOjkQs>
- <https://www.youtube.com/watch?v=eMlx5fFNoYc>

How can we model language?

Transformers and the attention mechanism

Objective: pay attention to the right words

- The girl went to her grandma's <mask>. <mask> had made cookies.
- The ship was lost in the <mask>.
- There was <mask> in the M50.

Objective: pay attention to the right words

- The girl went to her grandma's house. <mask> had made cookies.
- The ship was lost in the <mask>.
- There was <mask> in the M50.

Objective: pay attention to the right words

- The girl went to her **grandma's** house. **She** had made cookies.
- The ship was lost in the <mask>.
- There was <mask> in the M50.

Objective: pay attention to the right words

- The girl went to her grandma's house. She had made cookies.
- The **ship** was **lost** in the **sea/ocean**.
- There was <mask> in the M50.

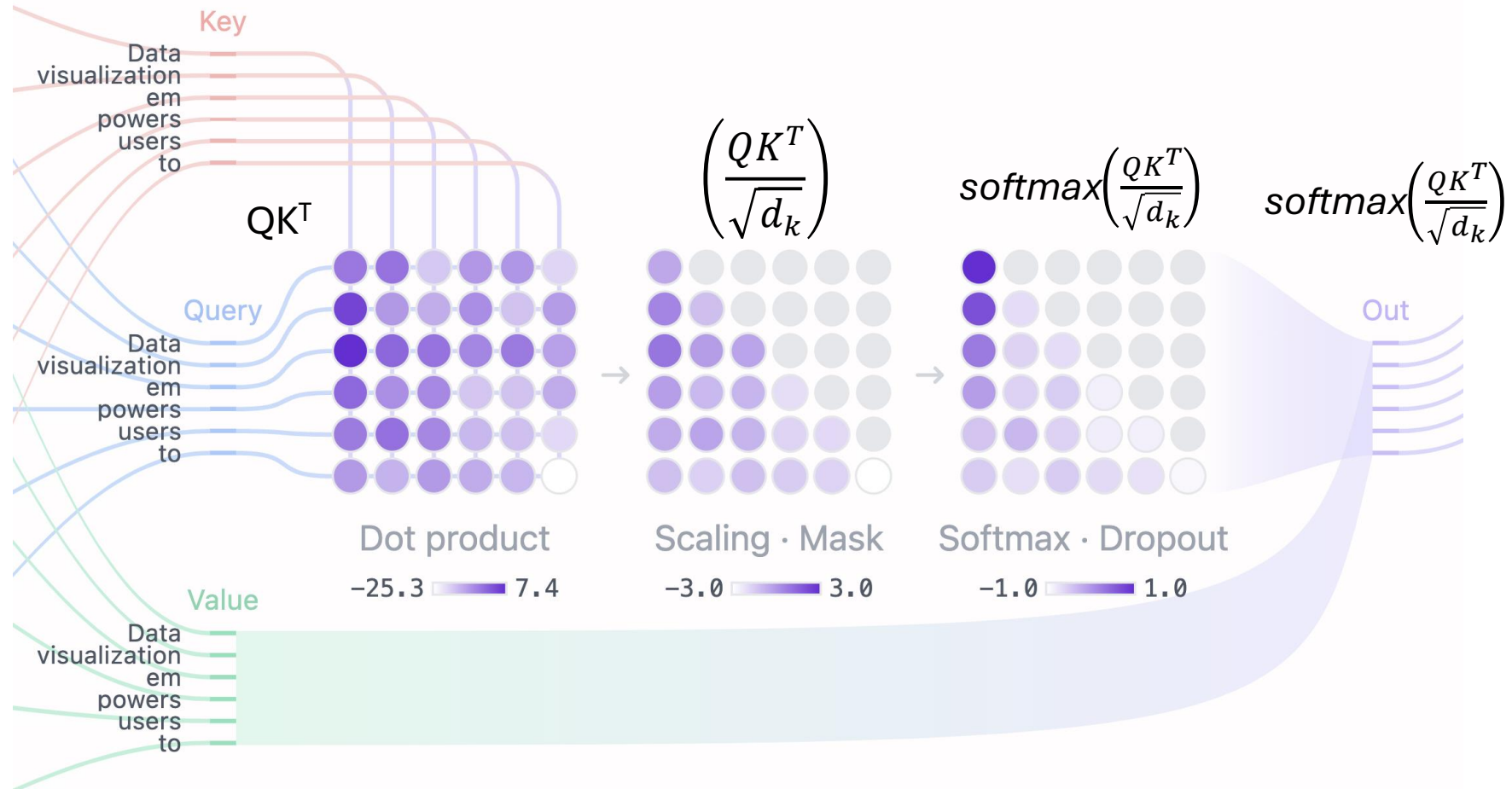
Objective: pay attention to the right words

- The girl went to her grandma's house. She had made cookies.
- The ship was lost in the sea/ocean.
- There was traffic/an accident in the M50.

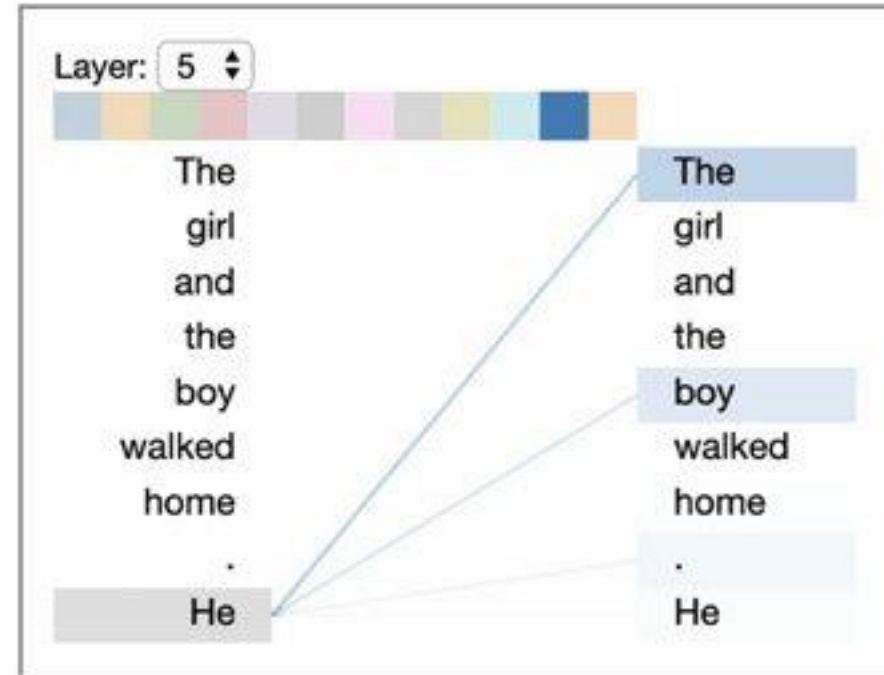
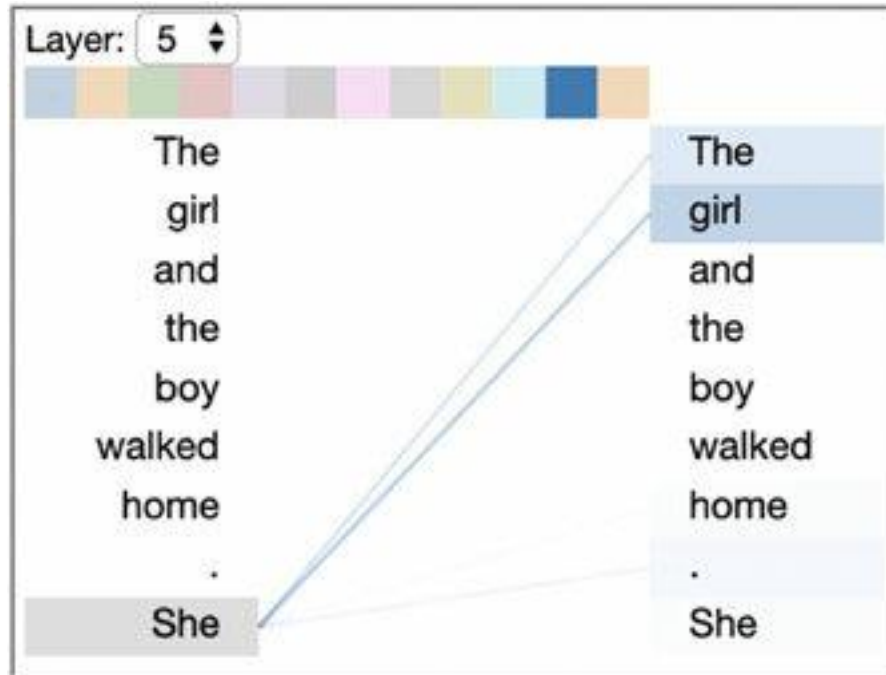
Attention Layer: Query, Key, Value

Attention: $\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$

- **Key:** embedding of word we are interested in
- **Query:** embedding of words we want to see if they are worth attending to
- **Value:** embedding of all words

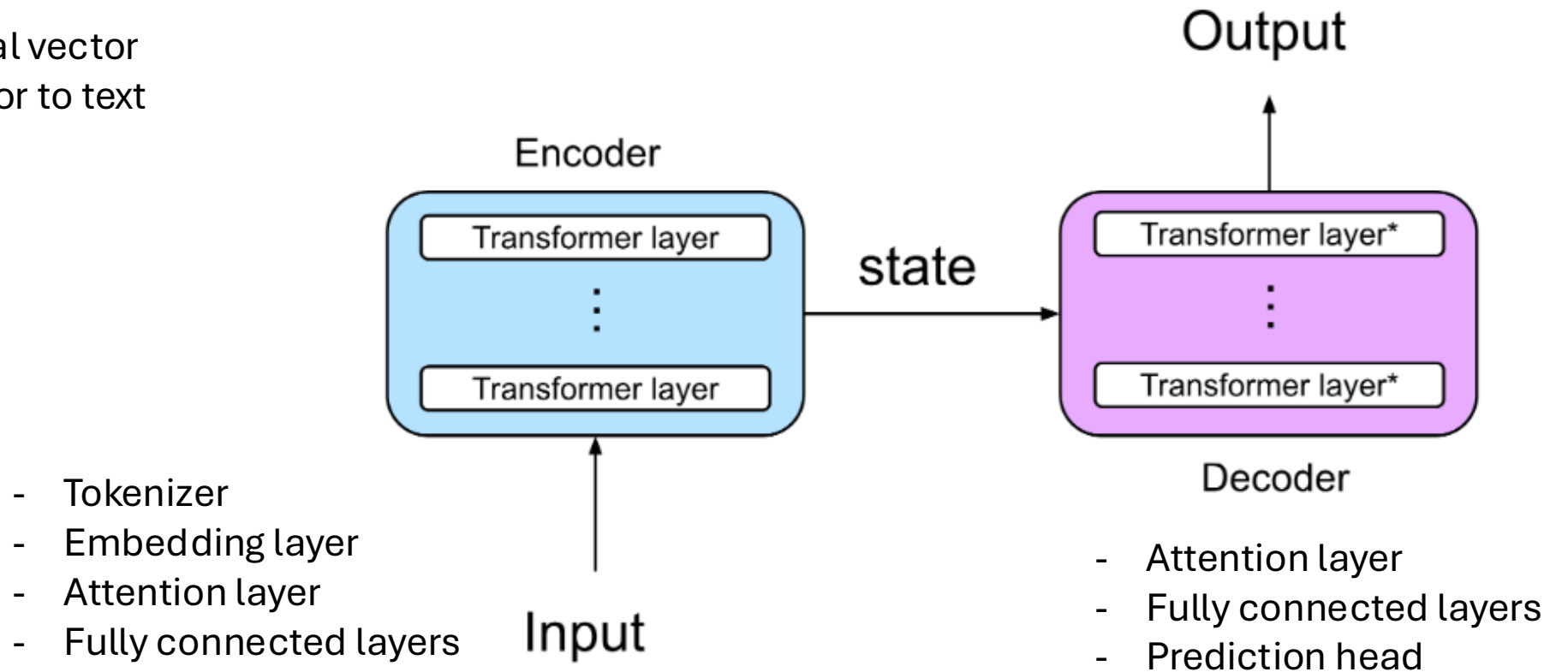


Objective: pay attention to the right words



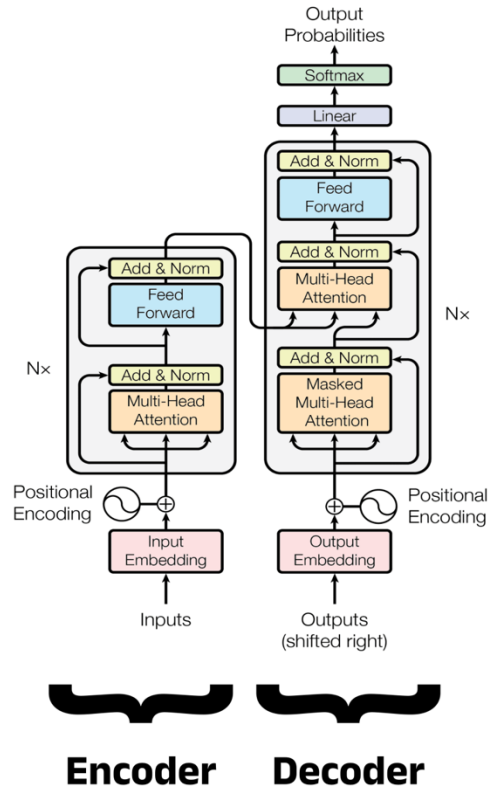
Transformer architecture

- **Encoder:** Text to numerical vector
- **Decoder:** Numerical vector to text

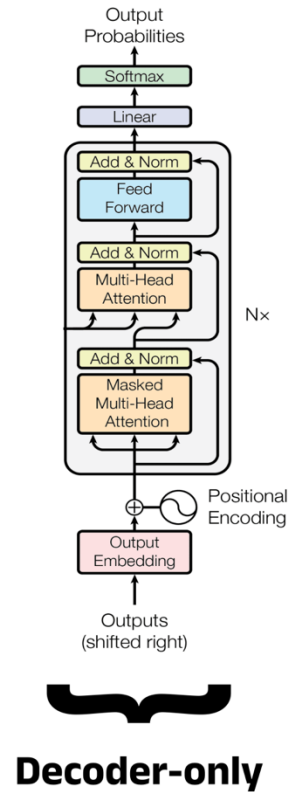


Types of architecture

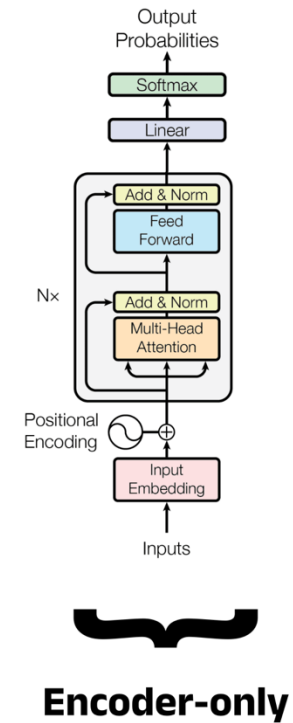
Transformer



GPT*



BERT*

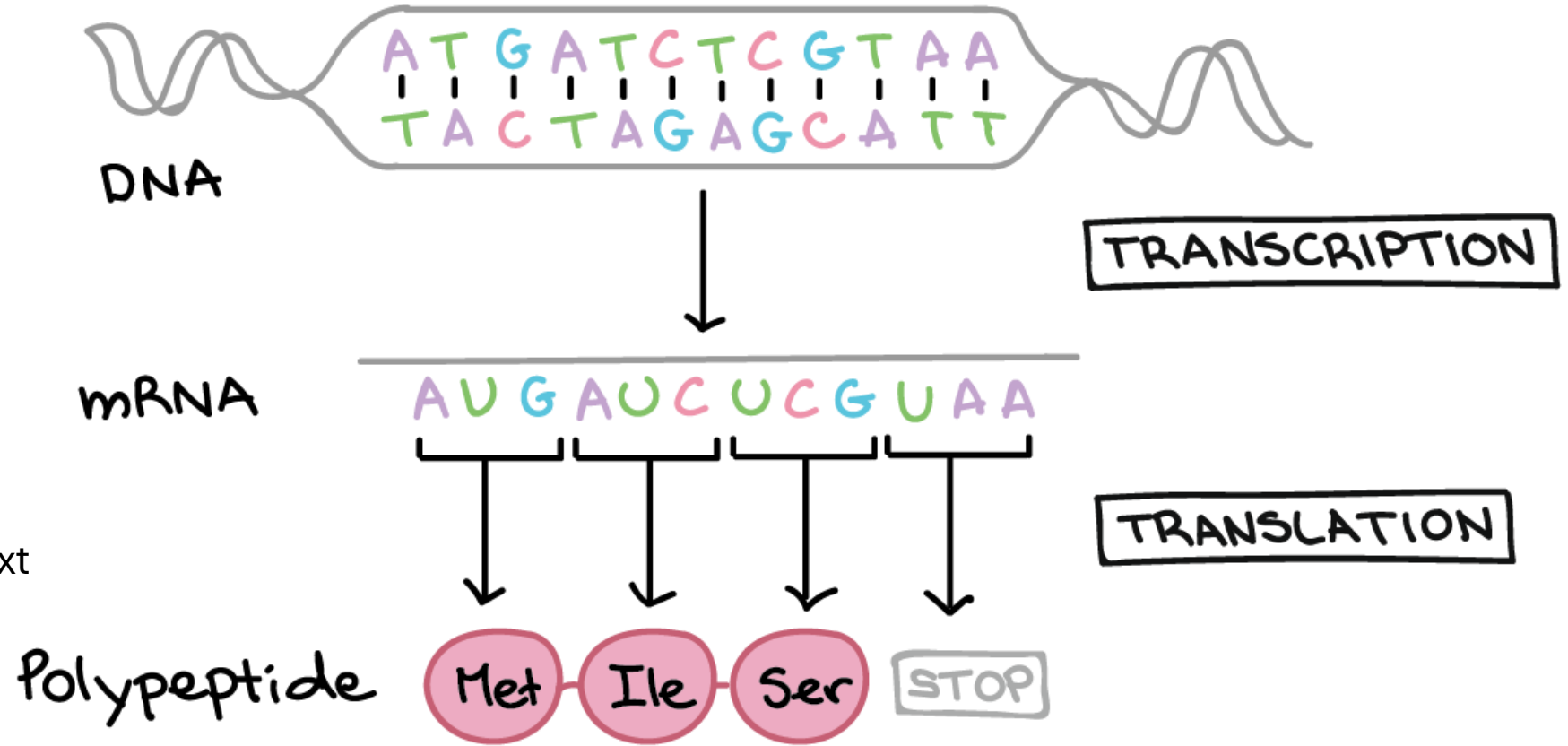


What are the molecular languages?

Transforming biomolecules into strings

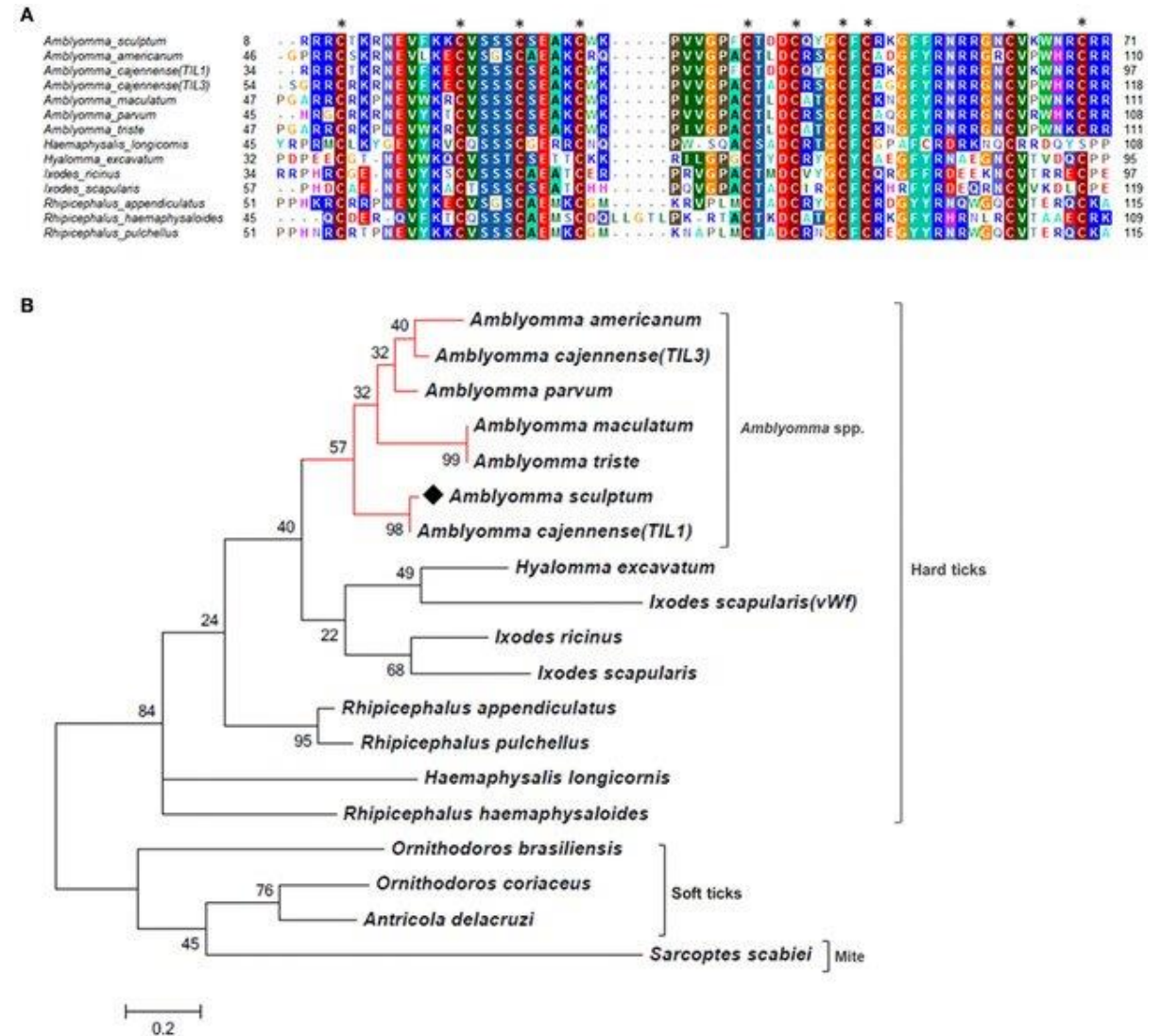
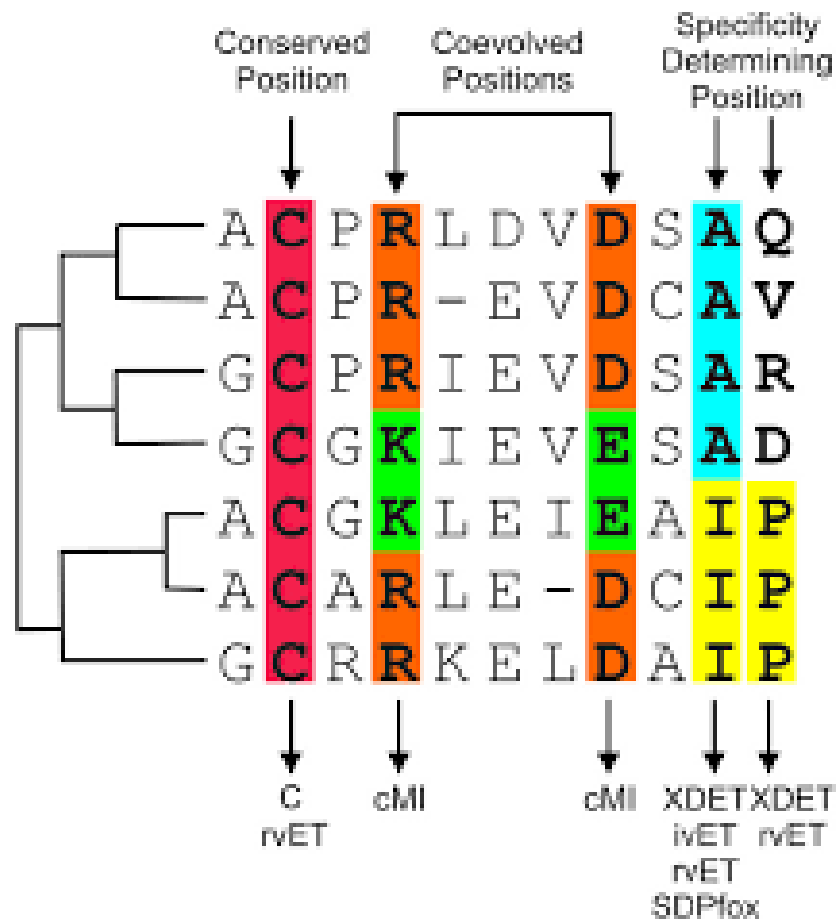
Biosequences

- **DNA and RNA:** 4 letter alphabet, huge context



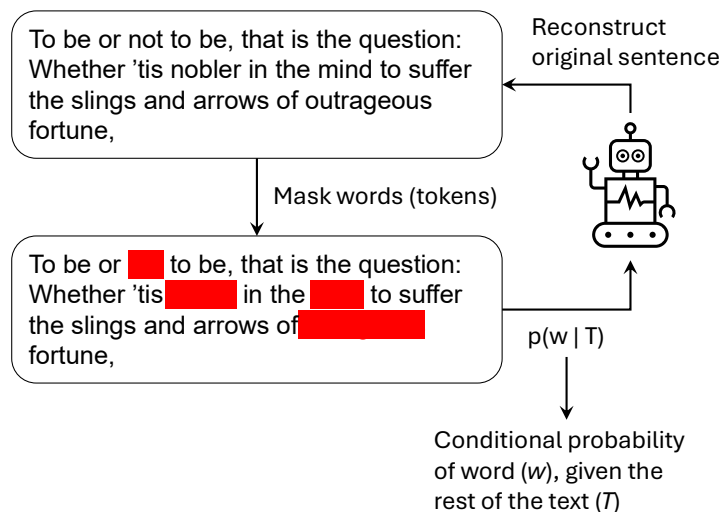
- **Proteins:** ~20 letter alphabet, smaller context

Biosequence evolution



Biosequences

Language models learn evolutionary conservation



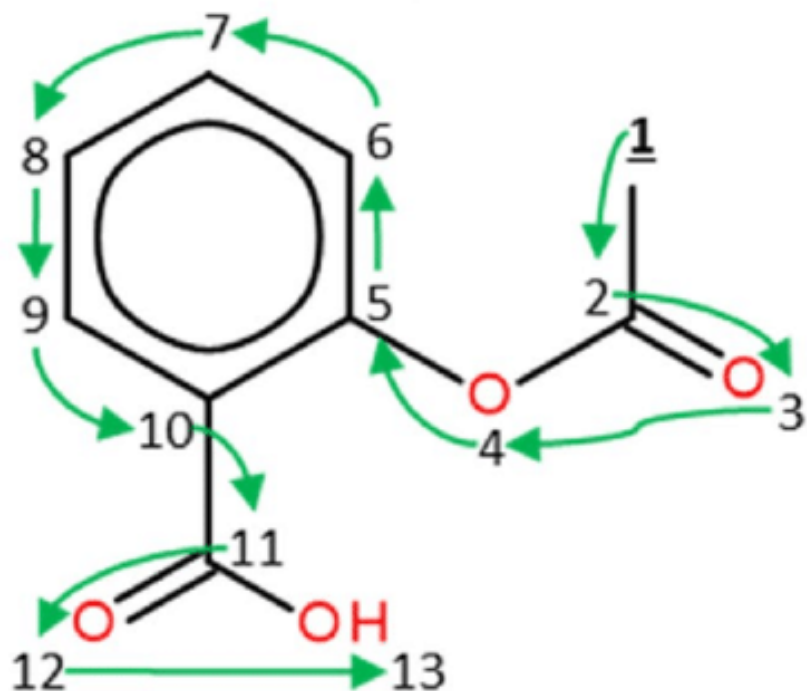
Q5E940_BOVIN	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_HUMAN	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_MOUSE	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_RAT	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_CHICK	-----MPREDRATWKSNYFMKIIQLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_RANSY	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVYLMGKNTMMRKAIRGHLENN--SALE	76
Q7ZUG3_BRARE	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMOTIRLSLRGK-AVYLMGKNTMMRKAIRGHLENN--PALE	76
RLA0 ICTPU	-----MPREDRATWKSNYFLKIIQLNDYPKCFIVGADNVGSKOMOTIRLSLRGK-AIYLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_DROME	-----MVRENAKAAQAQYFIKVVELFDEFPPKCFIVGADNVGSKOMONRTSLRGK-AVYLMGKNTMMRKAIRGHLENN--POLE	76
RLA0_DICDI	-----MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSLOLKIIRKSIRGI-GAVLMGKKTMIRKIVIRDLDASK--PELD	75
Q54LP0_DICDI	-----MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSLOLKIIRKSIRGI-GAVLMGKKTMIRKIVIRDLDASK--PELD	75
RLA0_PLAFB	-----MAKLSKQQKKQMYYIELSSLIIQYSKTILVHVDNVSQNMASVRKSLRGK-ATILMGKNTIRIALTKKNLTAAV-PDIE	76
RLA0_SULAC	----MIGLAVTITKKIAKKWDEVAELTEKLKTHSTIIIANIEGFAPADKLHRIRKKLRGK-ADIKVTKNNFNLIANKNAG----DYIK	79
RLA0_SULTO	---MRIMAVITQERKIAKWKEIEVKLEOKLREYHTIIIANIEGFAPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG----LDVS	80
RLA0_SULSO	---MKRLALALKQRKVASWGLEEVKELTELKNSNTLILNEGFPADKLHRIRKKLRGK-ATTIKVTNTLFFKIAAKNAG----IDIE	80
RLA0_AERPE	MSVVSVIGQMYYKREKIPDEWTKLMLRELLEELFSKHRYVLFDITGTPTFPVVRVRKKLKWK-YPMVMAKKRILLRAMKAAGLE-LDDN	86
RLA0_PYRAE	MMLAIKGRRYVRTROYTPARKVKIVSEATELLQKYPYVFLFDLHLGLSRILHEYYRRLRYGVIKIKIPLFKIAFTKVIYGG--IPAE	85
RLA0_METAC	-----MAEERHHTHEHPQWKDEIENIKELIQSHKVFPGMVGLEGILATKMCKIRDLKDVAVLKVSRTTLTERALNQLG--ETIP	78
RLA0_METMA	-----MAEERHHTHEHPQWKDEIENIKELIQSHKVFPGMVRLEGILATKIOKIRDLKDVAVLKVSRTTLTERALNQLG--ESIP	78
RLA0_ARCFU	-----MAAVRGS-----PPEYKVRVAVEEIKRMISSKPVVAIVSPFNPVAGOMQKIRREFRGK-AEKIVVKNTLLERALDALG--GDYL	75
RLA0_METKA	MAVKAKGQPSPSGYEPKVAEWKRRREVKEKELMDVEYNVLGVLDLEGIPAPLOEIRAKLERDTTIIRMSRNTLMRIALEEKLDER--PELE	88
RLA0_METHH	-----MAHVAEWKKEVQELHDLIKGYEVVGIANLADIPARLOKMQRTLRDSALTIRMSKFFLISLALEKAGREL--ENV D	74
RLA0_METTL	-----MITAESEHKIAPWKIEEVNKLLKELKNGQIVALVDMMVEVPAROLOEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNFEFA	82
RLA0_METVA	-----MIDAKSEHKIAPWKIEEVNALKEKLLSANVIALIDMMVEVPALOEOEIRDKIR-DQMTLKMSRNTLIERAKEVAEETGNFEFA	82
RLA0_METJA	-----METKVAHVAPWKIEEVKTLKGLISKSPVAIVDMMDVPALOEOEIRDKIR-DVKVIRMSRNTLIIIRALKEAAEELNNPKLA	81
RLA0_PYRAB	-----MAHVAEWKKEVEVELANLKSYPVALVDYSSMPAYPLSQMRRLIRENGGLRVSRNTLIELATKKAAGELGKPELE	77
RLA0_PYRHQ	-----MAHVAEWKKEVEVELAKLTSYVVALVDYSSMDPAYPLSQMRRLIRENGGLRVSRNTLIELATKKAAGELGKPELE	77

Protein Language Models

Name	Num params	Laboratory	Year
ESM-1b	650M	Meta	2019
ProtBERT	400M	RostLab	2020
Prot-T5-XL	560M	RostLab	2020
ProtGPT2	730M	Noelia Ferruz Lab (Barcelona)	2022
ESM-2	8M, 35M, 150M, 650M, 3B y 15B	Meta	2022
ESM-Fold	3B (+ 690M)	Meta	2022
ProstT5	560M (From Prot-T5-XL)	RostLab	2023

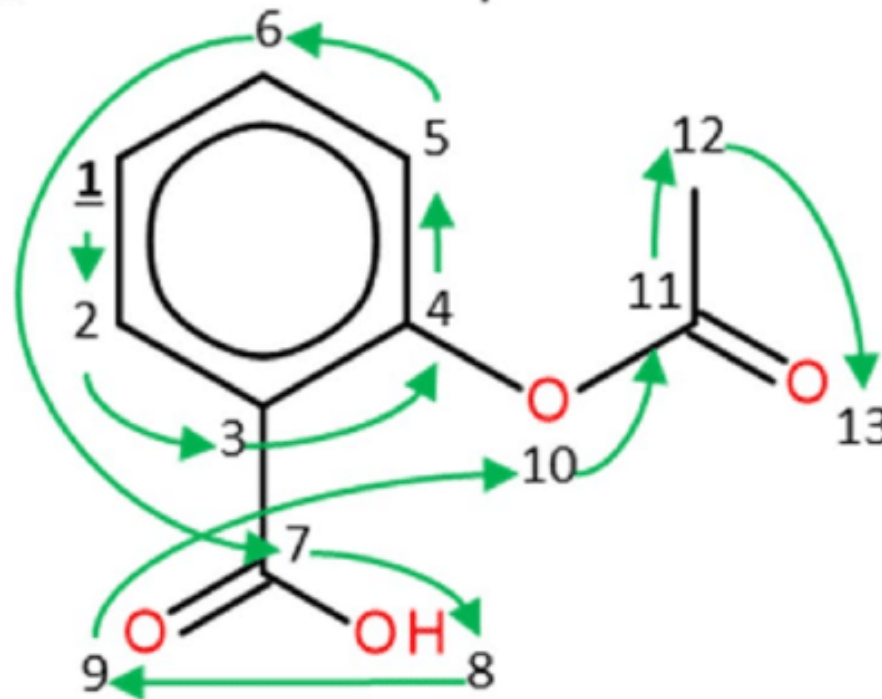
Small molecules: SMILES

a Canonical representation



CC(=O)Oc1ccccc1C(=O)O

b Randomized representation



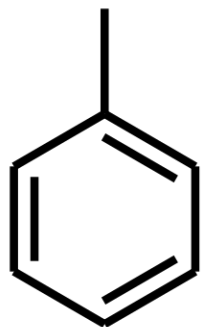
c1cc(c(cc1)C(O)=O)OC(C)=O

Simplified
Molecular
Input
Line
Entry
System

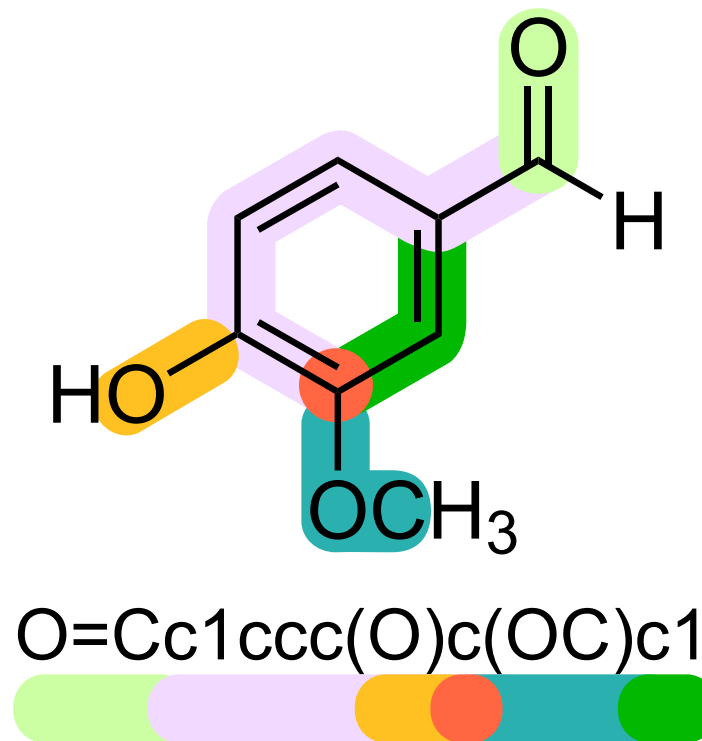
SMILES

Toluene

SMILES
Enumeration



Cc1ccccc1
c1ccccc1C
c1(C)ccccc1
c1c(C)cccc1
c1cc(C)ccc1
c1ccc(C)cc1
c1cccc(C)c1



Chemical fingerprints

1. Define circular substructures
2. Hash the substructures
3. Fold the hashes into bit-vector

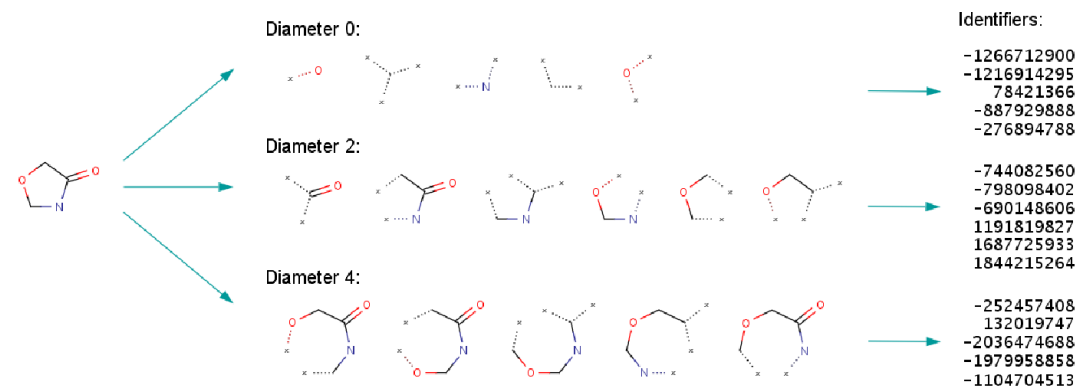


Fig. 2. ECFP generation process

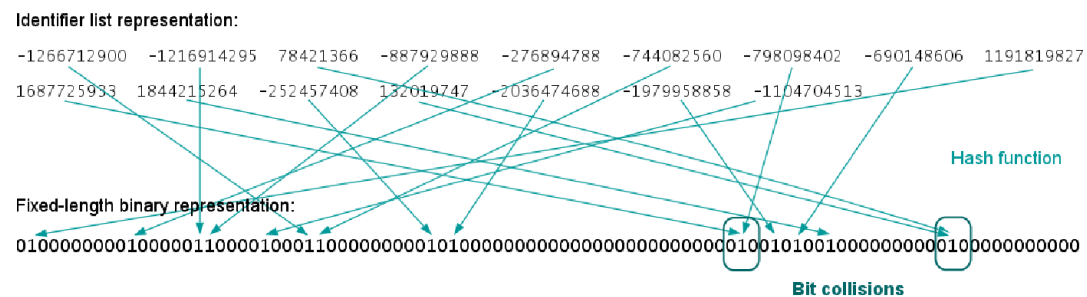


Fig. 3. Generation of the fixed-length bit string ("folding")

Chemical Language Models

Name	Num params	Laboratory	Year
ChemBERTa	3.5M	Ramsudar Lab	2020
Molformer-XL	45M	IBM	2022
ChemBERTa-2	3.5M	Ramsudar Lab	2022
ChemBERTa-3	3.5M	Ramsudar Lab	2025

Tutorial 1:

Representation learning

Transforming biomolecules into vectors

Objective

1. Get models from huggingface
2. Generate representations
 1. What's the best strategy?
 2. Can you improve efficiency?
3. Train machine learning model
 1. What's the best model?
 2. How to properly evaluate?

Tutorial materials

- <https://github.com/RaulFD-creator/ucd-teaching>
- raul.fernandezdiaz@ucdconnect.ie