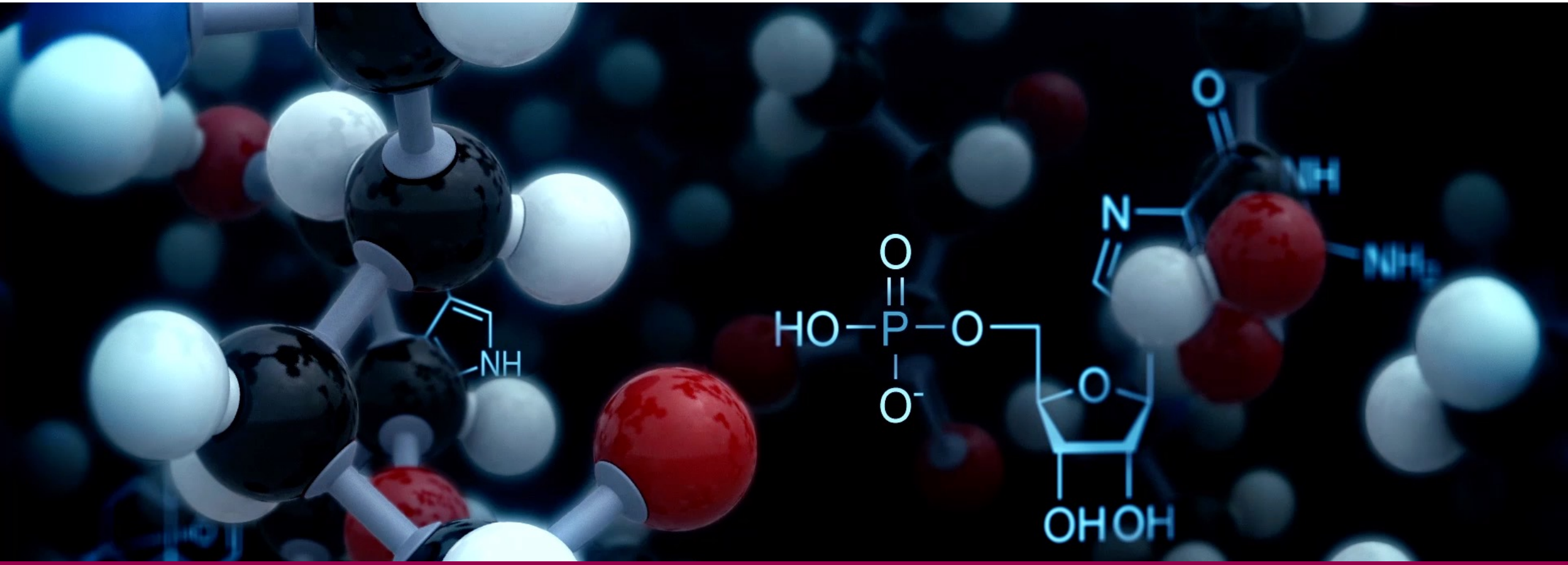
Large Language Models: from Drug Discovery to Clinical Trials

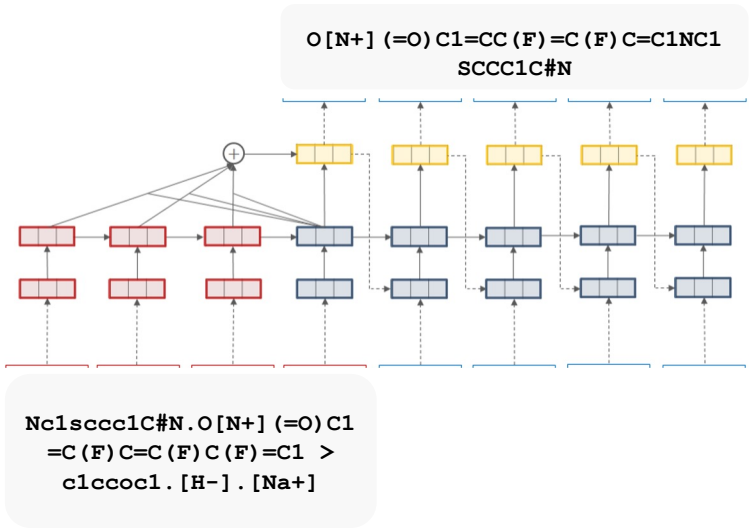


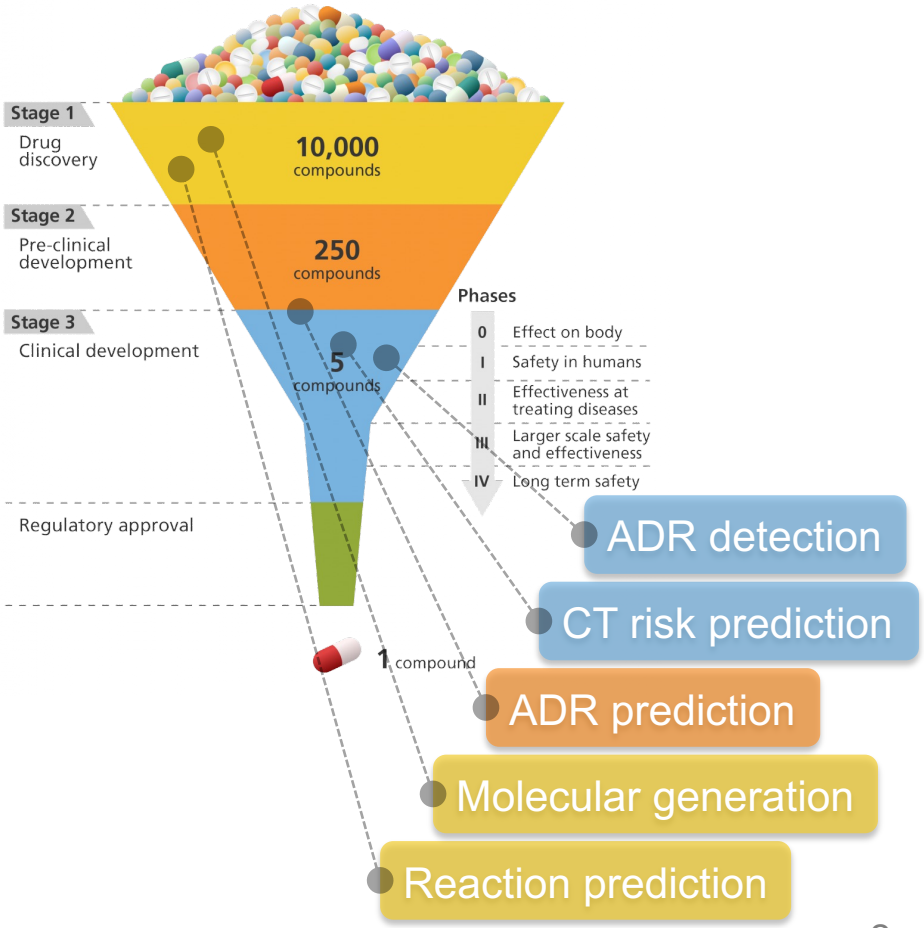
Douglas Teodoro

UNIVERSITÉ DE GENÈVE

Outline

the chef cooked the meal

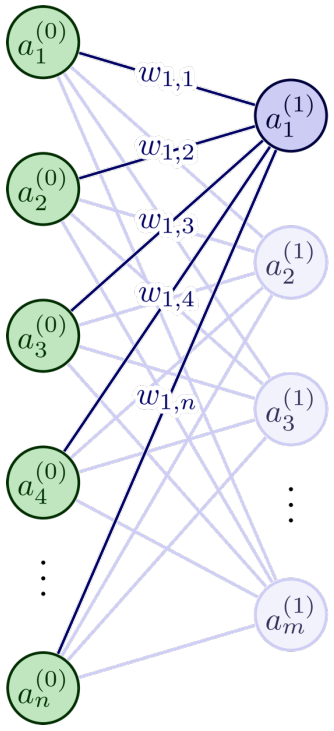




the chef cooked the meal

LANGUAGE MODELLING VIA NEURAL NETWORKS

What is an artificial neural network?

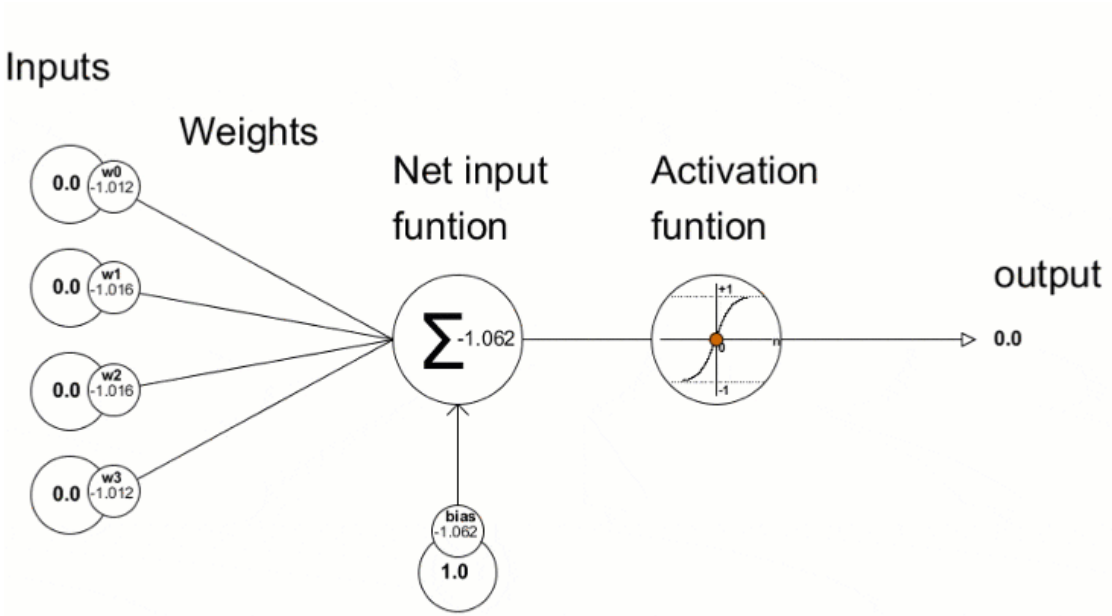


$$= \sigma\left(w_{1,0}a_0^{(0)} + w_{1,1}a_1^{(0)} + \ldots + w_{1,n}a_n^{(0)} + b_1^{(0)}\right)$$

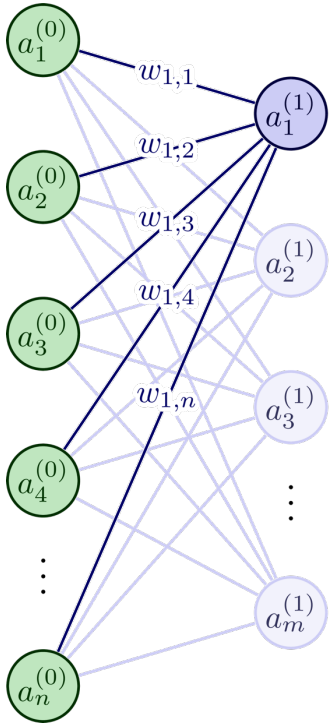
$$= \sigma\left(\sum_{i=1}^{n} w_{1,i}a_i^{(0)} + b_1^{(0)}\right)$$

$$\begin{pmatrix} a_1^{(1)} \\ a_2^{(1)} \\ \vdots \\ a_m^{(1)} \end{pmatrix} = \sigma\left[\begin{pmatrix} w_{1,0} & w_{1,1} & \cdots & w_{1,n} \\ w_{2,0} & w_{2,1} & \cdots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,0} & w_{m,1} & \cdots & w_{m,n} \end{pmatrix}\begin{pmatrix} a_1^{(0)} \\ a_2^{(0)} \\ \vdots \\ a_n^{(0)} \end{pmatrix} + \begin{pmatrix} b_1^{(0)} \\ b_2^{(0)} \\ \vdots \\ b_m^{(0)} \end{pmatrix}\right]$$

$$a^{(1)} = \sigma\left(\mathbf{W}^{(0)}a^{(0)} + \mathbf{b}^{(0)}\right)$$



What is a *deep* neural network?

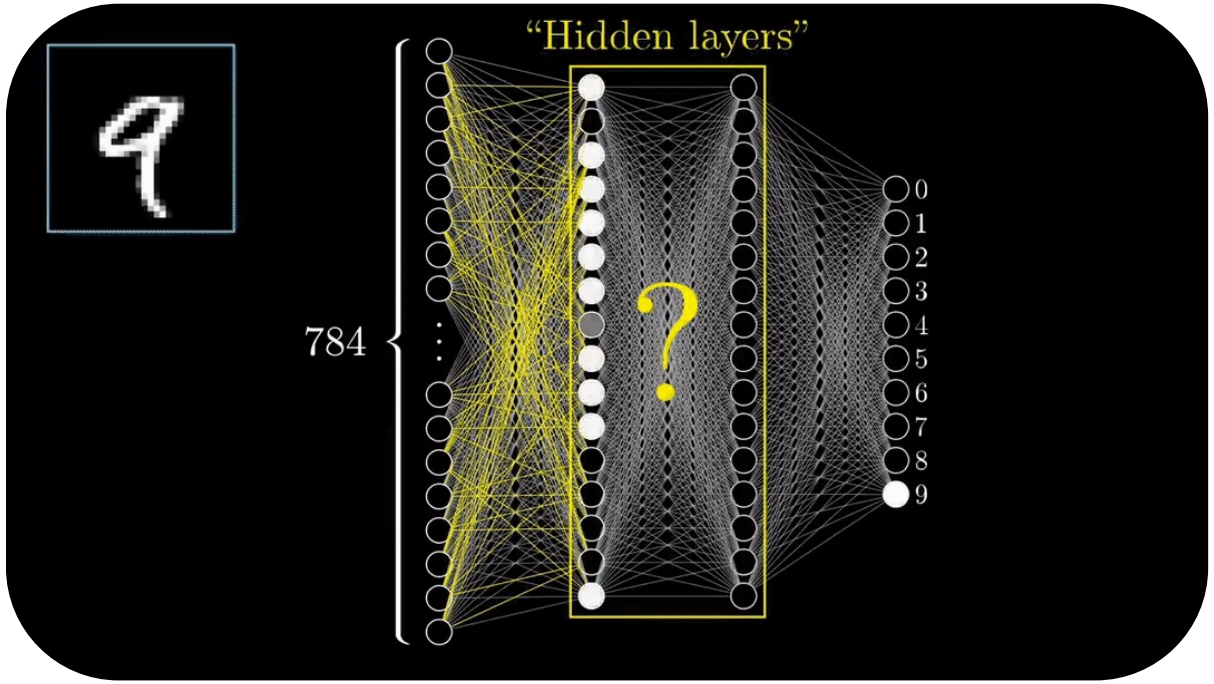


$$= \sigma\left(w_{1,0}a_0^{(0)} + w_{1,1}a_1^{(0)} + \ldots + w_{1,n}a_n^{(0)} + b_1^{(0)}\right)$$

$$= \sigma\left(\sum_{i=1}^{n} w_{1,i}a_i^{(0)} + b_1^{(0)}\right)$$

$$\begin{pmatrix} a_1^{(1)} \\ a_2^{(1)} \\ \vdots \\ a_m^{(1)} \end{pmatrix} = \sigma\left[\begin{pmatrix} w_{1,0} & w_{1,1} & \cdots & w_{1,n} \\ w_{2,0} & w_{2,1} & \cdots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,0} & w_{m,1} & \cdots & w_{m,n} \end{pmatrix}\begin{pmatrix} a_1^{(0)} \\ a_2^{(0)} \\ \vdots \\ a_n^{(0)} \end{pmatrix} + \begin{pmatrix} b_1^{(0)} \\ b_2^{(0)} \\ \vdots \\ b_m^{(0)} \end{pmatrix}\right]$$

$$a^{(1)} = \sigma\left(\mathbf{W}^{(0)}a^{(0)} + \mathbf{b}^{(0)}\right)$$



www.3blue1brown.com

Deep learning. LeCun, Bengio & Hinton. 2015.

Language model

Definition: A *probabilistic model* that assigns a probability $P(w_1,w_2,...,w_n)$ to every finite word sequence $w_1,...,w_n$

You shall know a word by the company it keeps!

J. R. Firth 1890-1960: 'A Synopsis of Linguistic Theory' (1957)

Auto-regressive model

$P(w_1,...,w_n) = P(w_1) \times P(w_2|w_1) \times P(w_3|w_{1:2}) \times ... \times P(w_n|w_{1:n-1})$

$P(w_1,...,w_n) \approx P(w_1) \times P(w_2|w_1) \times P(w_3|w_2) \times ... \times P(w_n|w_{n-1})$

V = {chef, cooked, meal, the}

P(the,chef,cooked,the,meal) =

P(the)×P(chef|the)×P(cooked|the chef)×...×P(meal|the chef cooked the)

P(the,chef,cooked,the,meal) = 0.0200

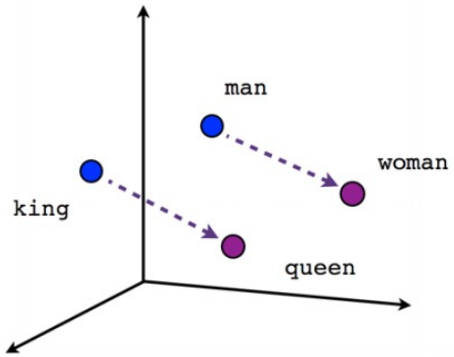
P(the,meal,cooked,the,chef) = 0.0100

P(chef,the,the,meal,cooked) = 0.0001

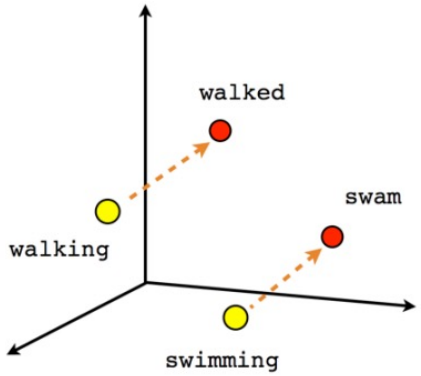
$P(w_1,...,w_n) = \text{neural\_net}(w_1,...,w_n)$

GPT-3 still acts in this way, but the model is implemented as a very large neural network of 175-billion parameters!

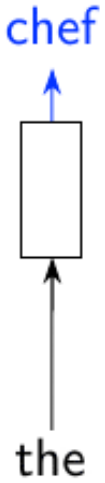
Representation and generation

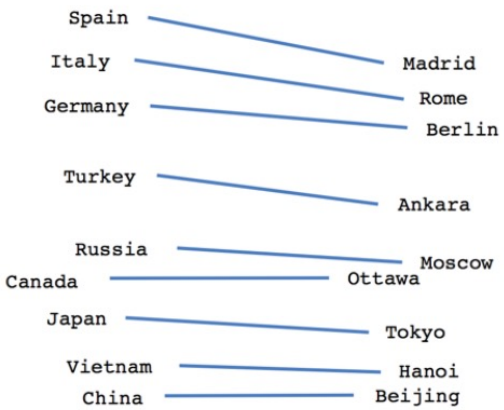


Male-Female



Verb tense





Country-Capital

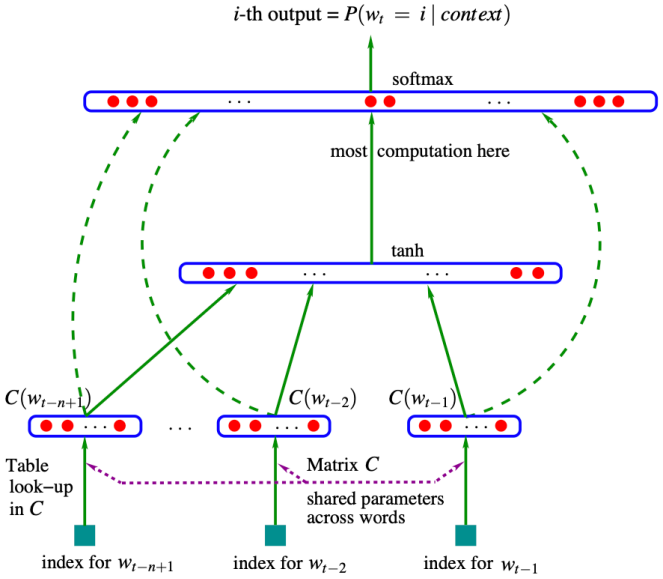
Source: ai.googleblog.com

A brief story of (modern) LLMs

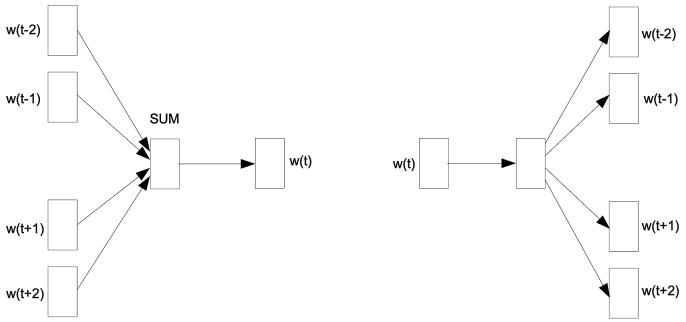
Concept	Authors	Year	Citations
NLM	Bengio *et al.*	2003	10k+
word2vec	Mikolov *et al.*	2013	37k+
seq-to-seq	Sutskever *et al.*	2014	23k+
Transformers	Vaswani *et al.*	2017	86k+
ELMo	Peters *et al.*	2018	13k+
BERT	Devlin *et al.*	2018	76k+
GPT-2	Radford *et al.*	2019	6k+

…
GPT-3
Claude
ChatGPT
GPT-4
…

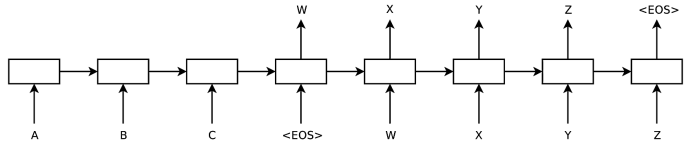
A brief story of LLMs



A Neural Probabilistic Language Model, Bengio *et al.*, 2003

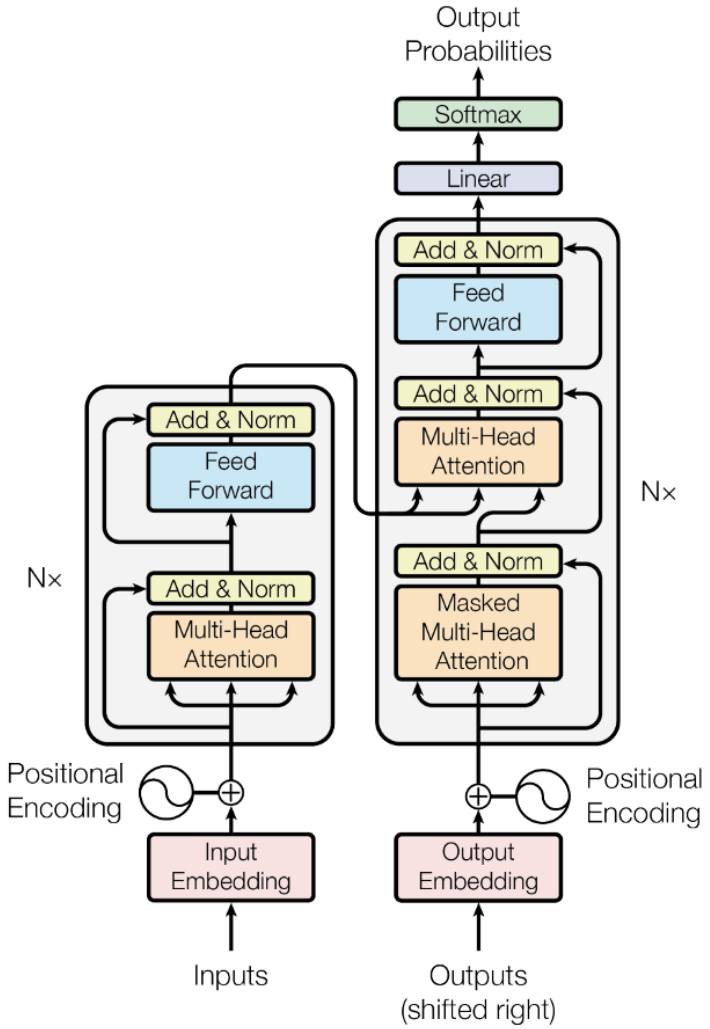


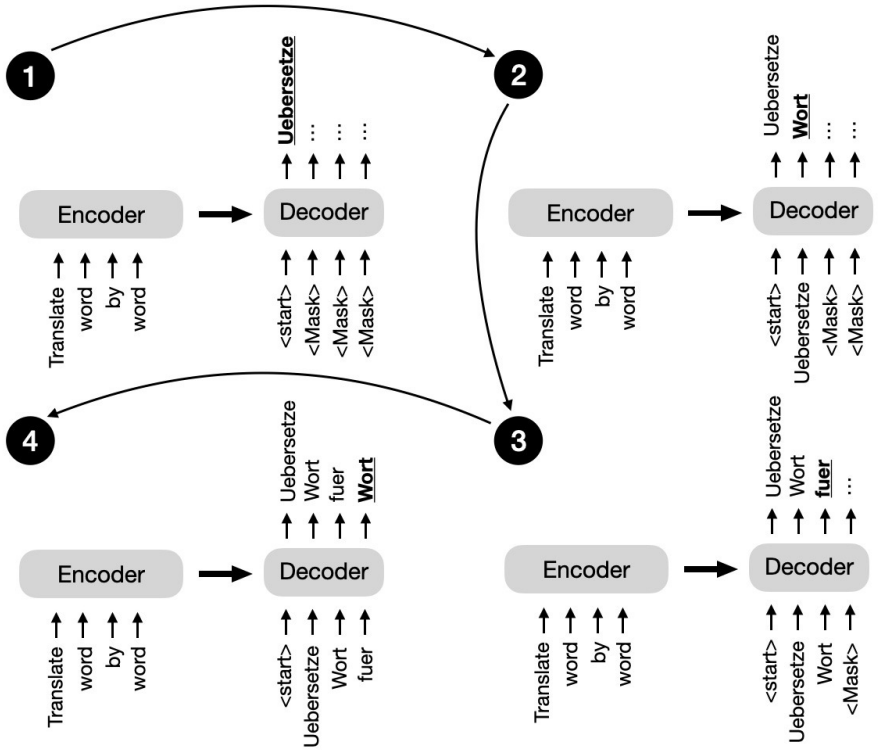
Efficient Estimation of Word Representations in Vector Space, Mikolov *et al.*, 2013

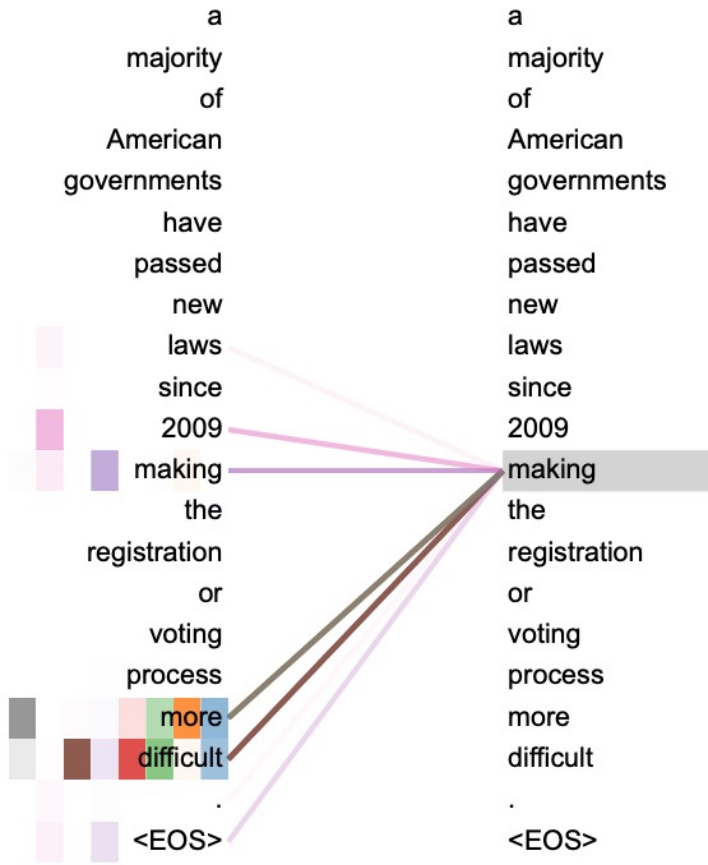


Sequence to Sequence Learning with Neural Networks, Sutskever *et al.*, 2014

A brief story of LLMs

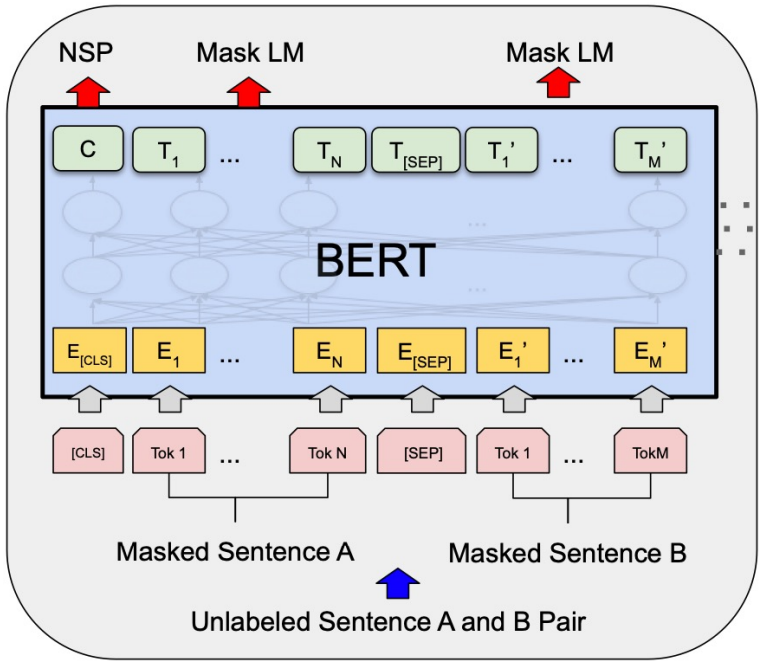




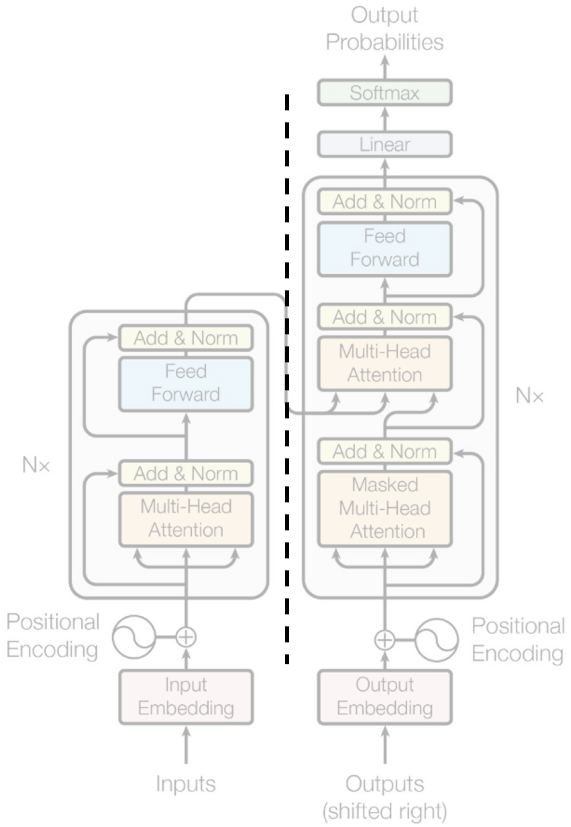


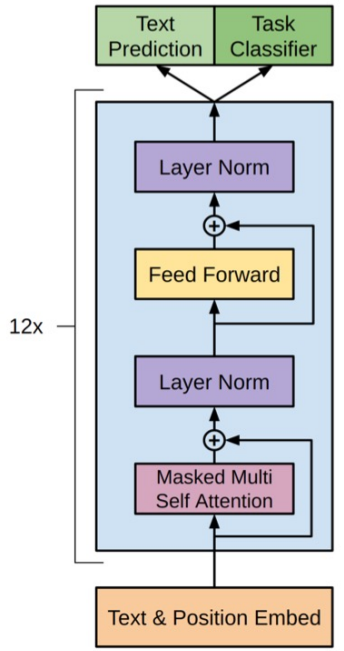
Attention is all you need, Vaswani *et al.*, 2017

A brief story of LLMs

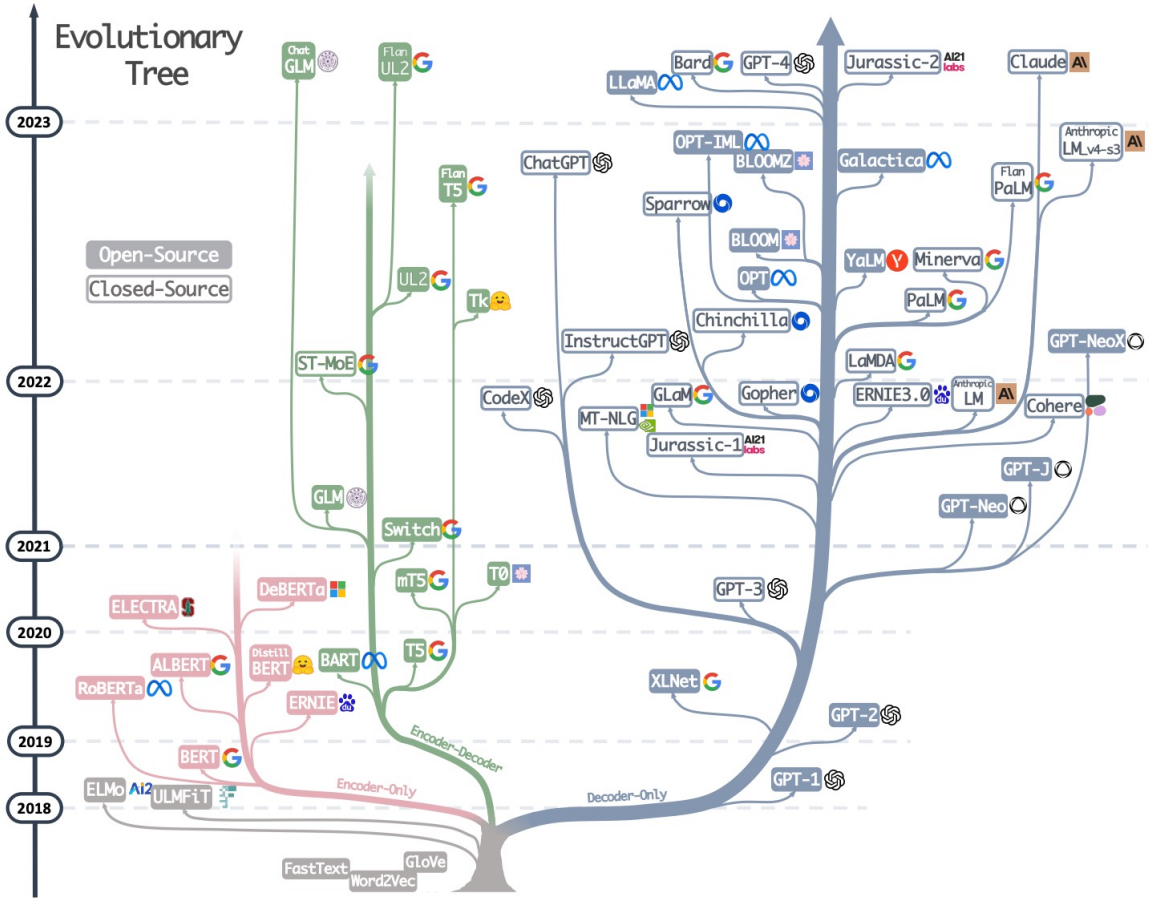


BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin *et al.*, 2018

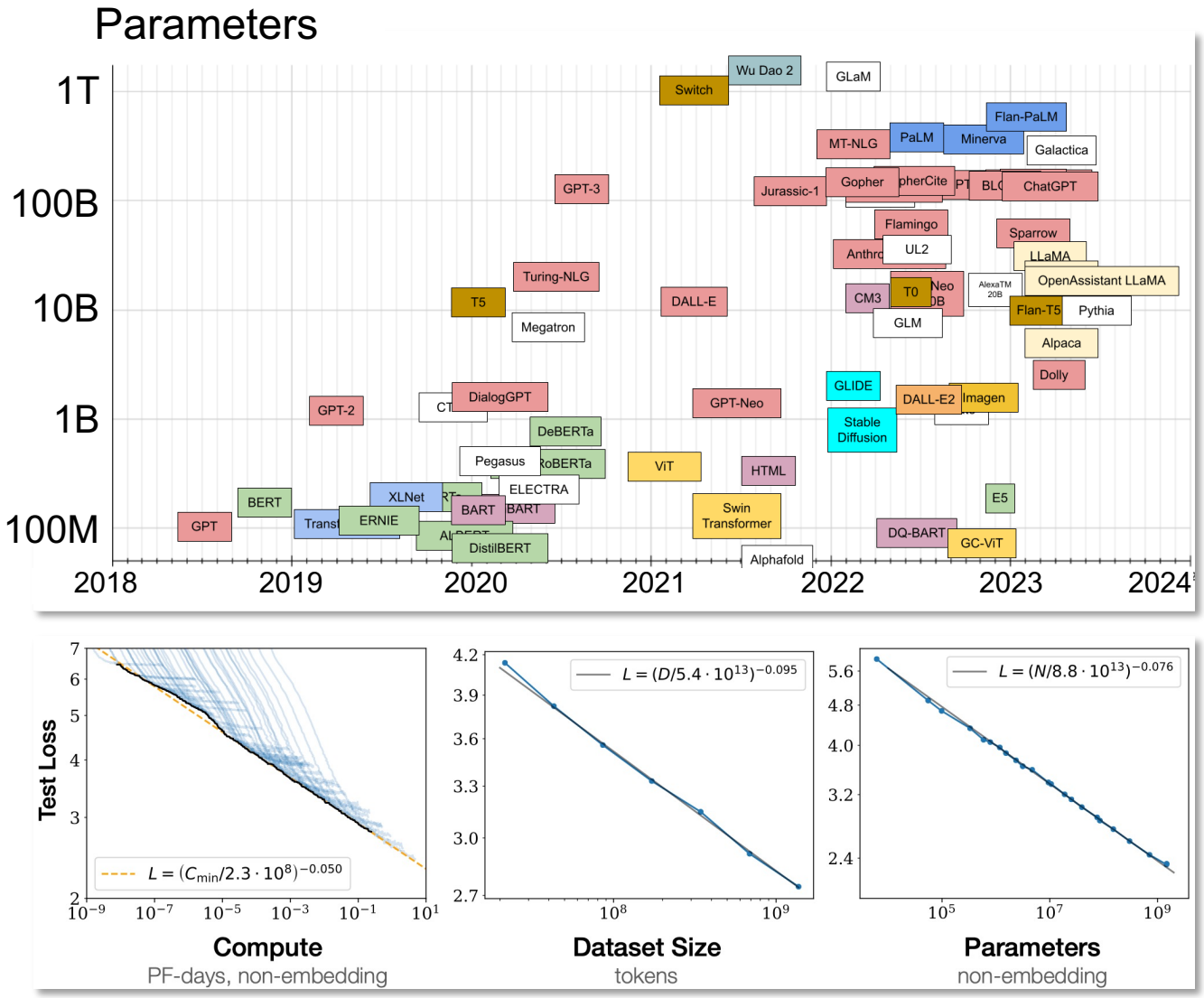




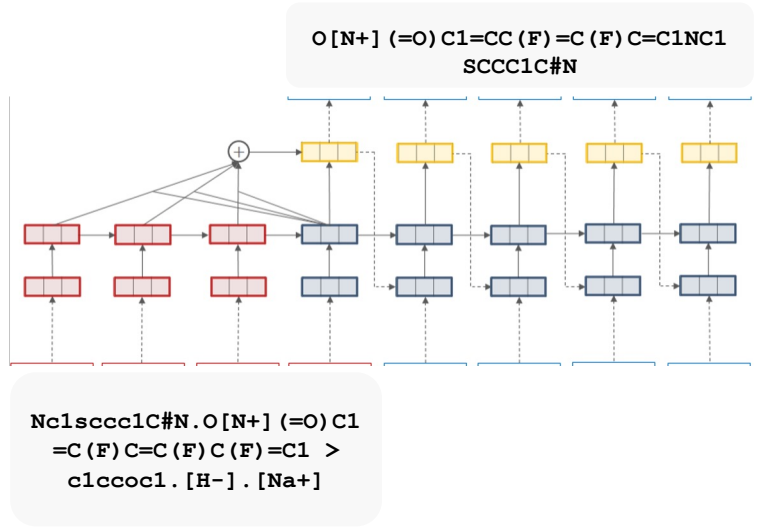
Language Models are Unsupervised Multitask Learners, Radford *et al.*, 2019



Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond
Yang *et al.*, 2023

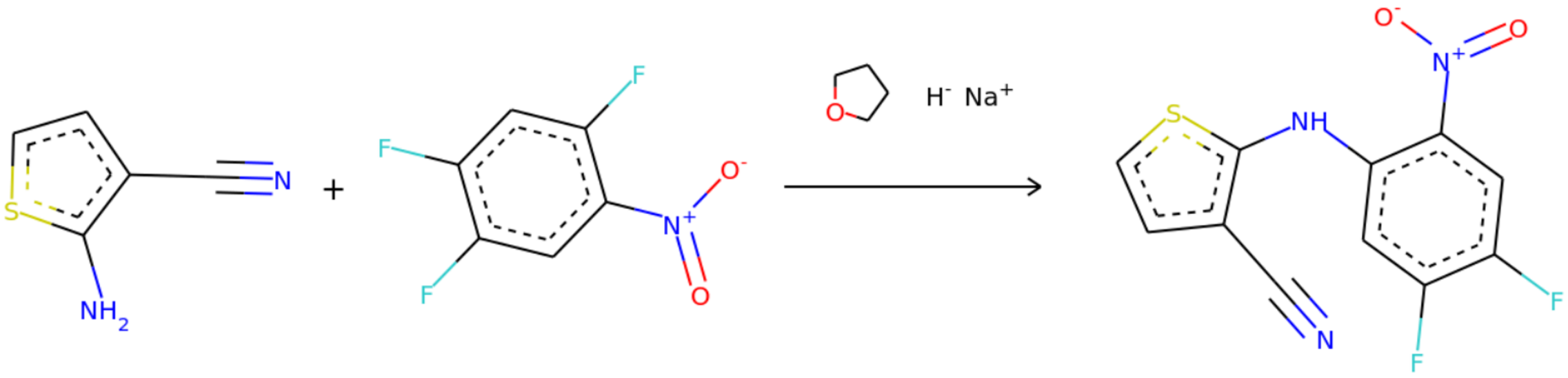


Scaling Laws for Neural Language Models,
Kaplan *et al.*, 2020



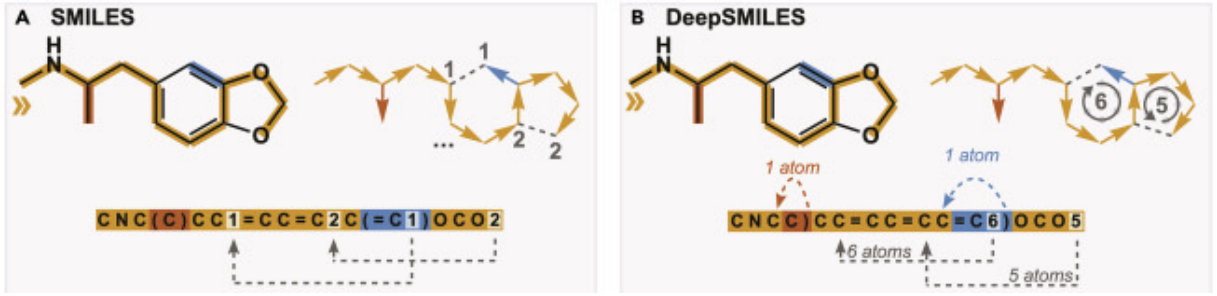
FROM NATURAL TO CHEMICAL LANGUAGE

Natural language			Chemical language	
Character	*a-zA-Z0-9*	←→	Atom	*C, H, N, …*
Word	*water*	←→	Molecule	H_2O
Sentence	*I want water*	←→	Reaction	$2H_2 + O_2 \rightarrow 2H_2O$



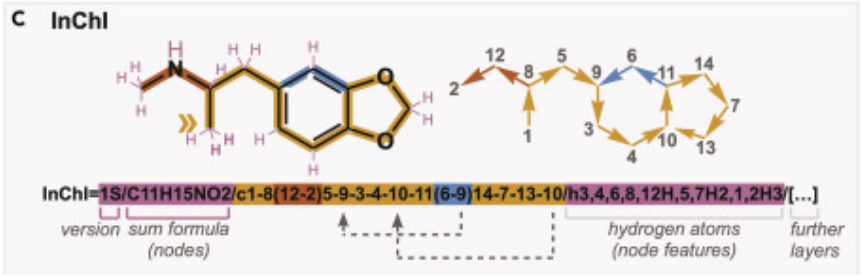

```
Nc1sccc1C#N.O[N+](=O)C1=C(F)C=C(F)C(F)=C1 >
c1ccoc1.[H-].[Na+] >
O[N+](=O)C1=CC(F)=C(F)C=C1NC1SCCC1C#N
```

Molecular string representations

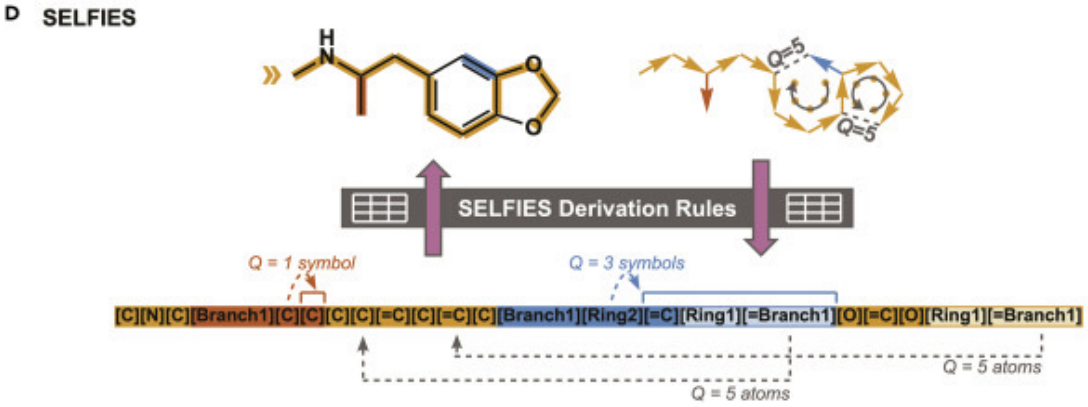


SMILES: simplified molecular-input line-entry system

Branch/ring constraints

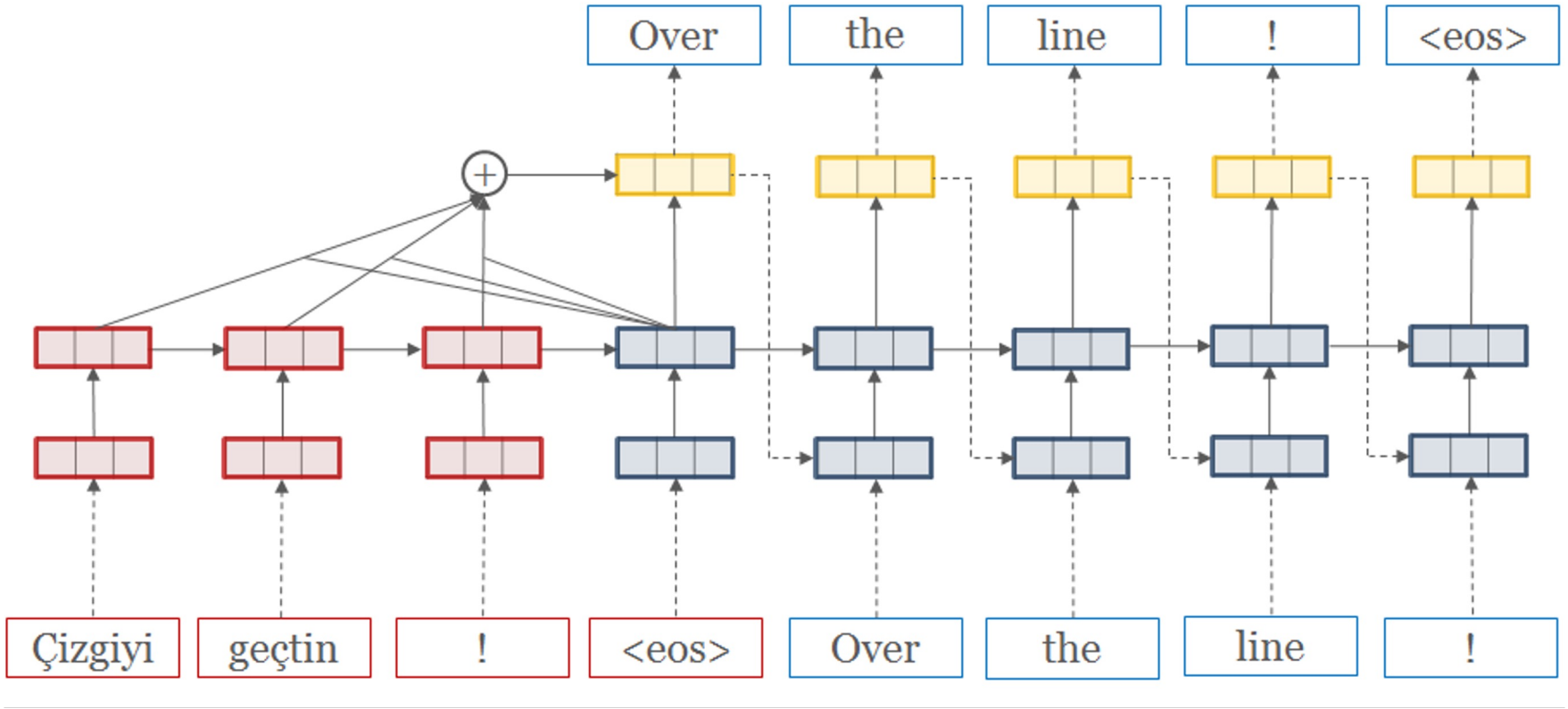


InChI: International Chemical Identifier (Uniqueness)



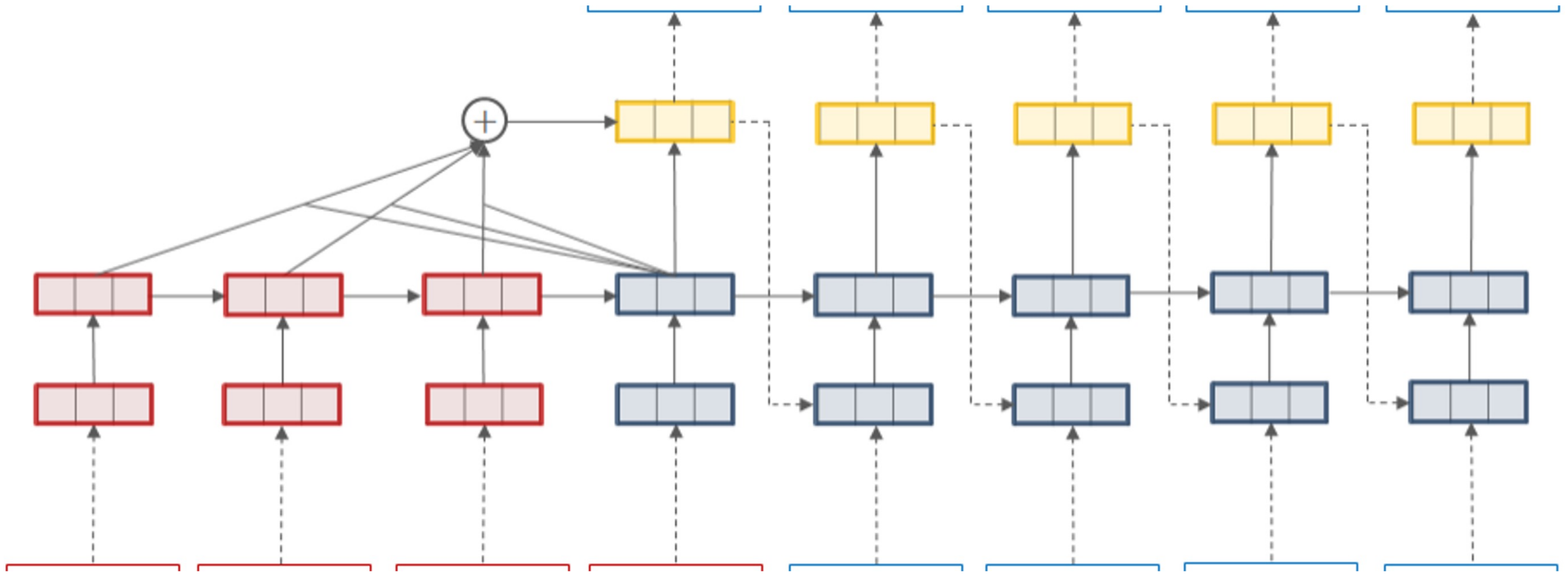
SELFIES: SELF-referencing embedded string (Branch/ring and valence constraints)

SELFIES and the future of molecular string representations, Krenn et *al.*, 2022.


Over
the
line
!
<eos>
Çizgiyi
geçtin
!
<eos>
Over
the
line
!

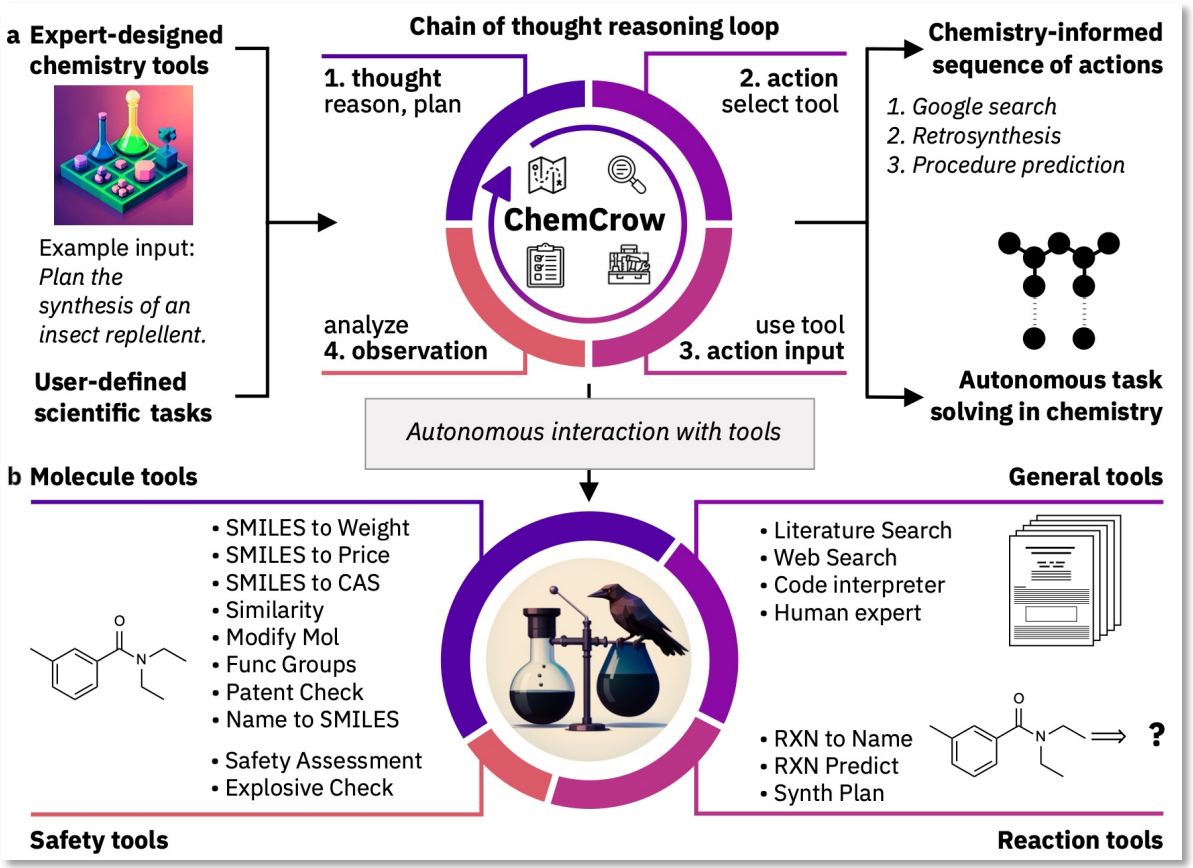
SMILES

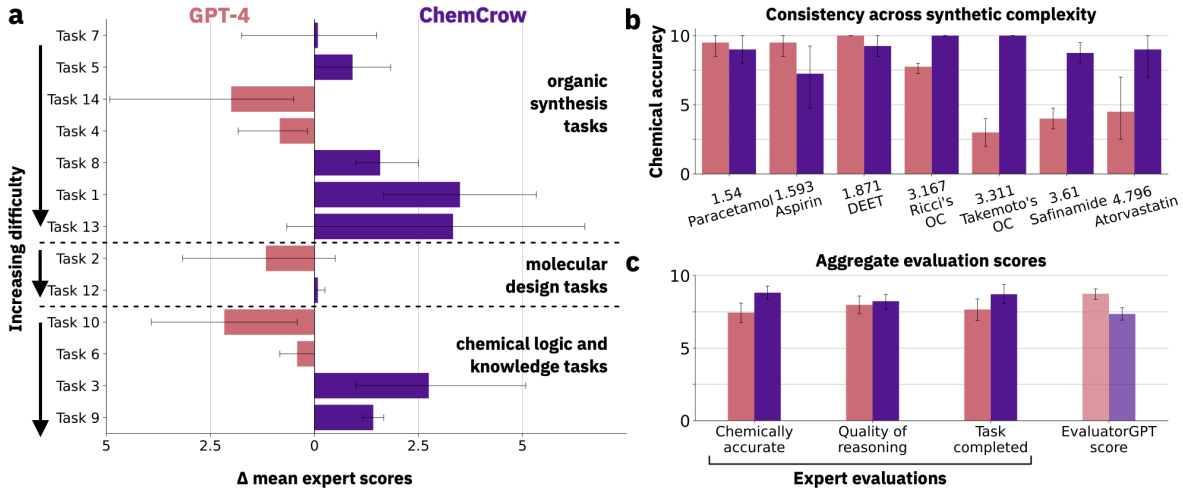
O[N+](=O)C1=CC(F)=C(F)C=C1NC1SCCC1C#N



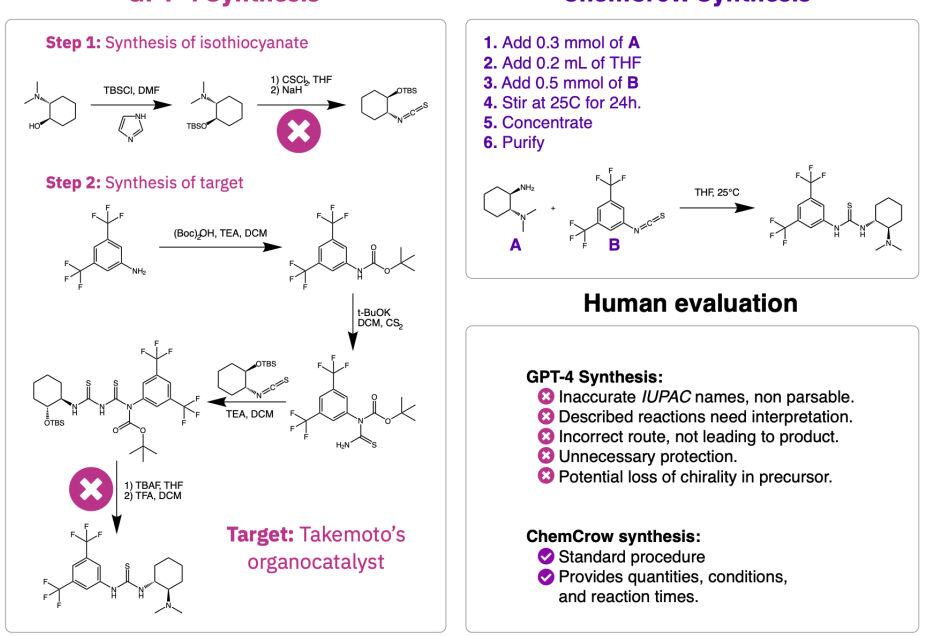
Nc1sccc1C#N.O[N+](=O)C1=C(F)C=C(F)C(F)=C1 > c1ccoc1.[H-].[Na+]

Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction, Schwaler *et al.,* 2019

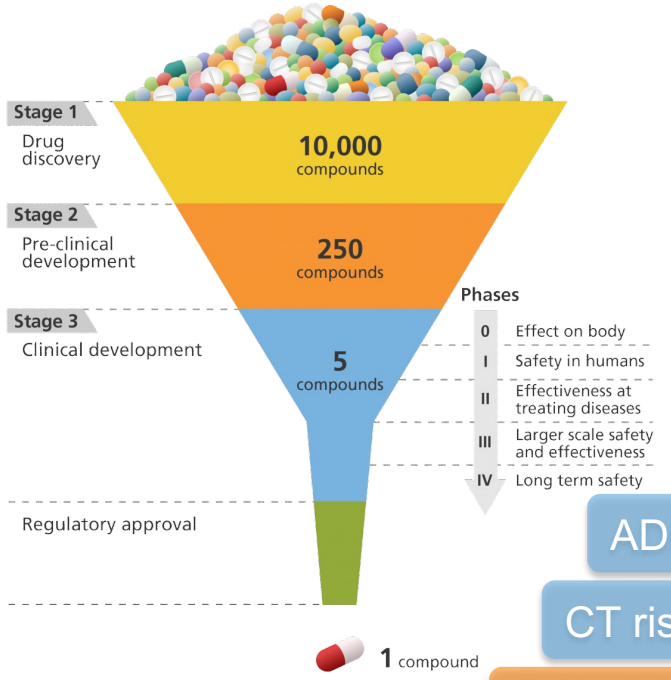




Task: synthesize Takemoto's organocatalyst



Augmenting large language models with chemistry tools
Bran *et al.*, 2023



ADR detection

CT risk prediction

ADR prediction

Molecular generation

Reaction prediction

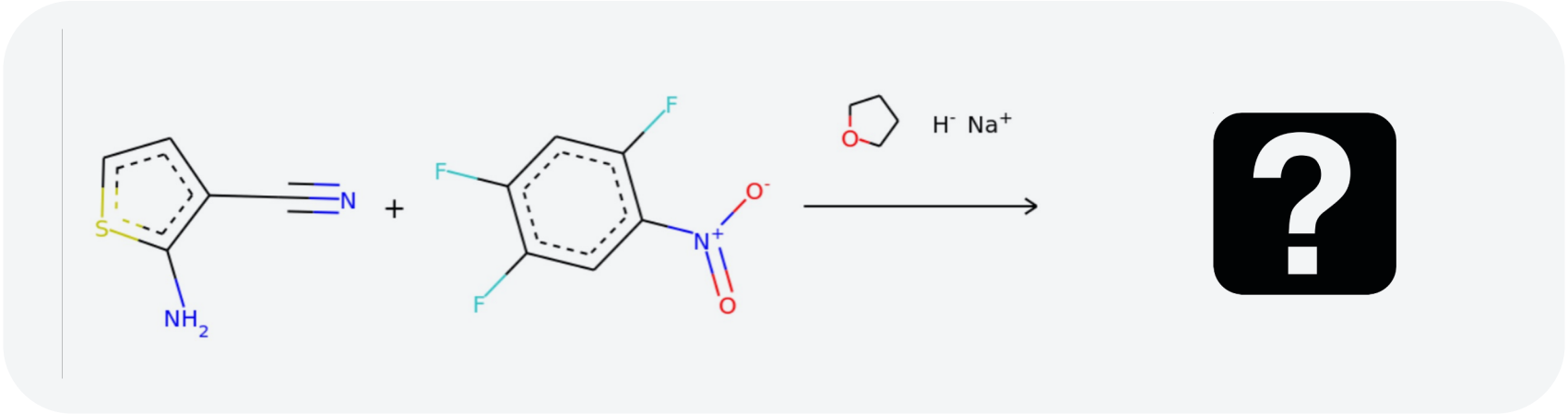
Natural and chemical language models

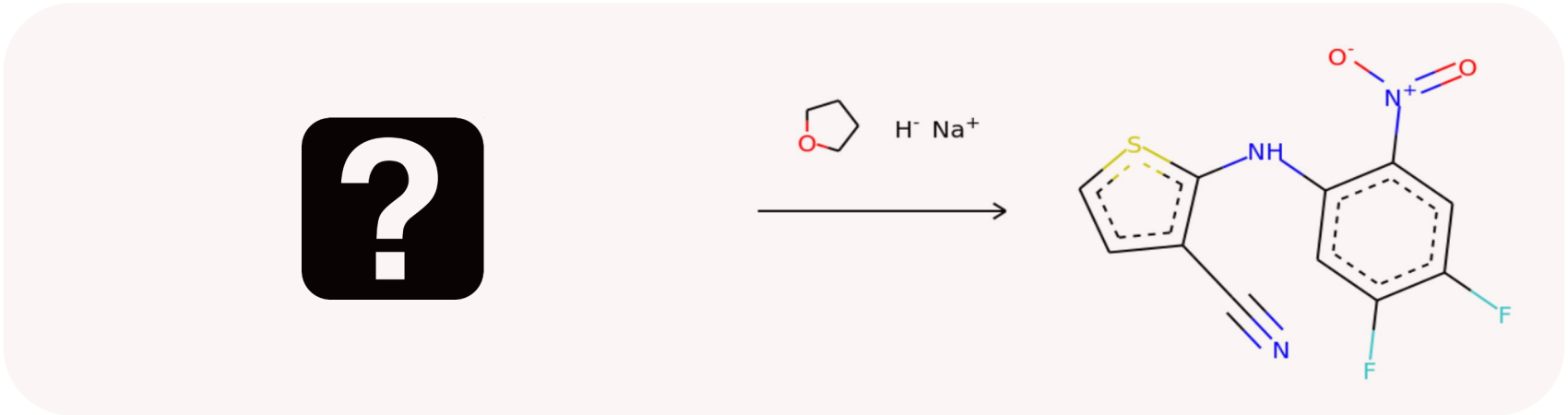
APPLICATIONS OF LLMS IN DRUG DISCOVERY AND DEVELOPMENT

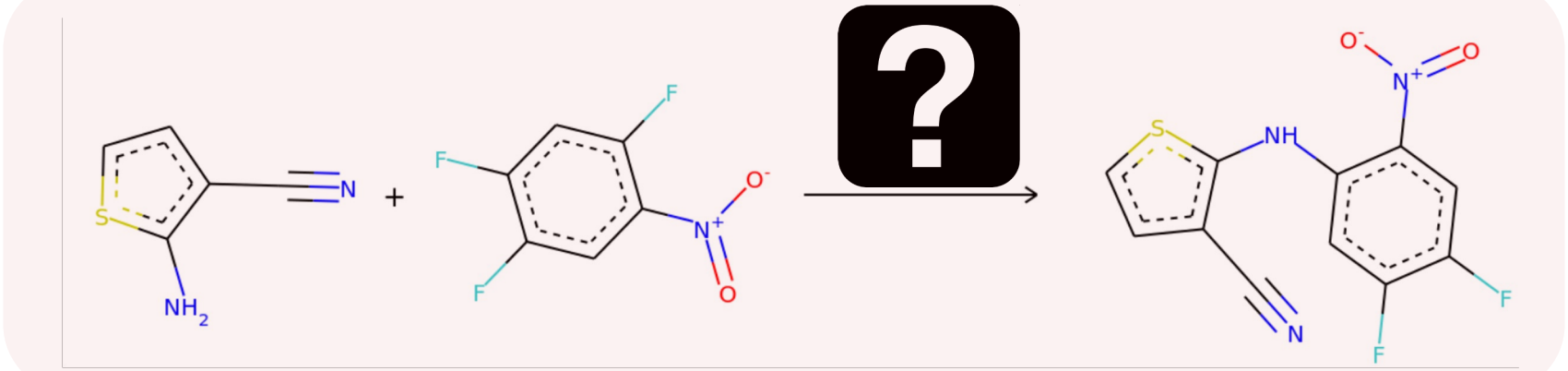
Transformer Performance for Chemical Reactions: Analysis of Different Predictive and Evaluation Scenarios

Jaume-Santero, Bornet *et al.*, JCIM, 2023

SYNTHESIS AND ONE-STEP RETROSYNTHESIS PREDICTION


F
F
O⁻
N⁺
O
F
S
N
NH₂
O
H⁻ Na⁺
?

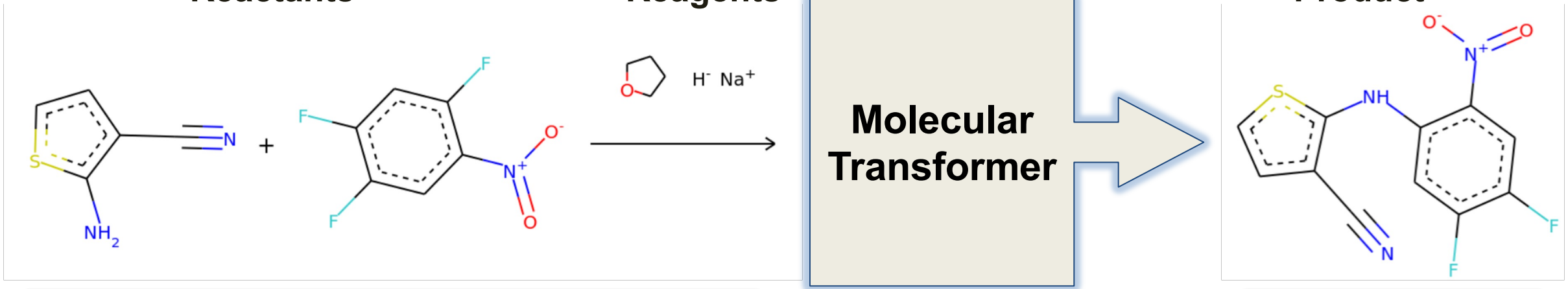

?
O
H⁻ Na⁺
O⁻
N⁺
O
S
NH
F
N
F


F
F
O⁻
N⁺
O
F
S
N
NH₂
?
O⁻
N⁺
O
S
NH
F
N
F

Reactants

Reagents

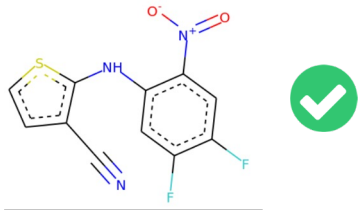
Product

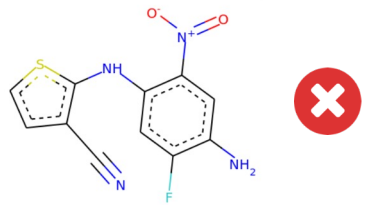


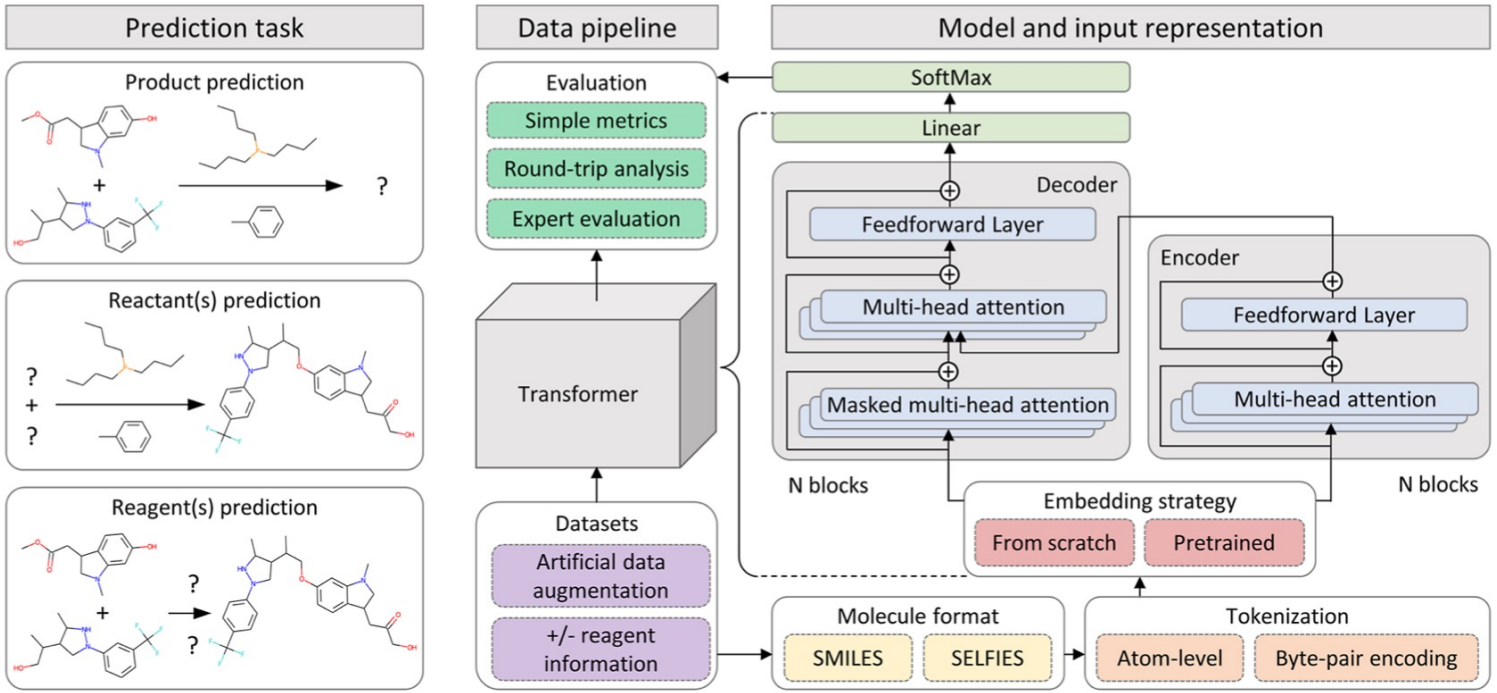

```
Nc1sccc1C#N.O[N+](=O)C1=C(F)C=C(F)C(F)=C1 >
c1ccoc1.[H-].[Na+]
```

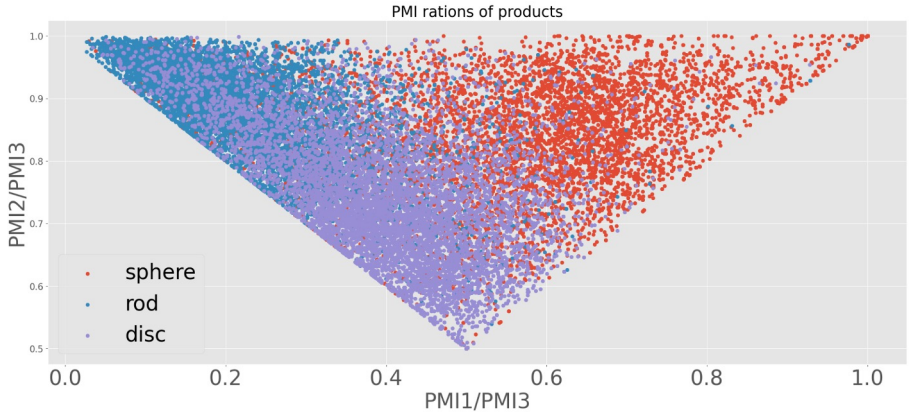
```
O[N+](=O)C1=CC(F)=C(F)C=C1NC1SCCC1C#N
```

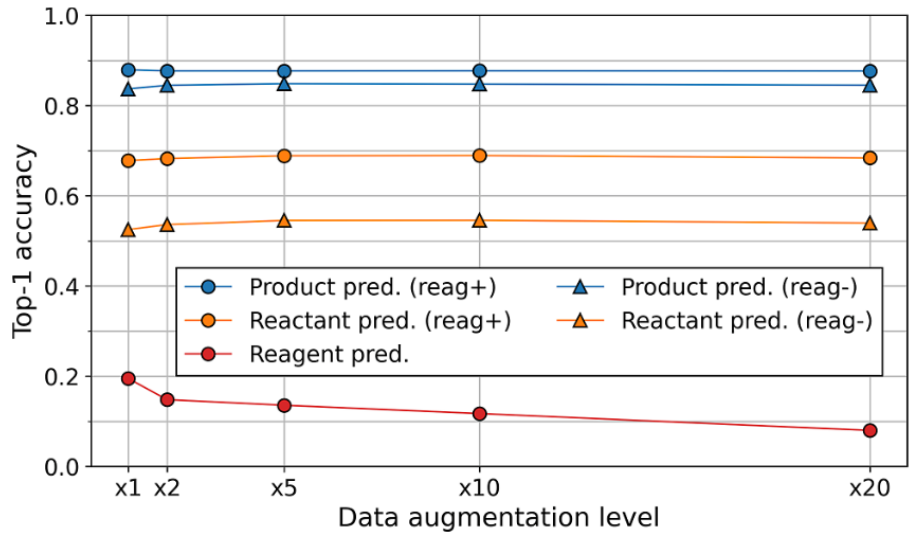
Score	SMILES
1.0	N#Cc1ccsc1Nc1cc(F)c(F)cc1[N+](=O)[O-]
7.3×10^{-6}	N#Cc1cscc1Nc1cc(F)c(F)cc1[N+](=O)[O-]
4.6×10^{-6}	N#Cc1ccsc1Nc1cc(F)c([N+](=O)[O-])cc1F
4.2×10^{-6}	N#Cc1ccsc1Nc1cc(F)c(N)cc1[N+](=O)[O-]

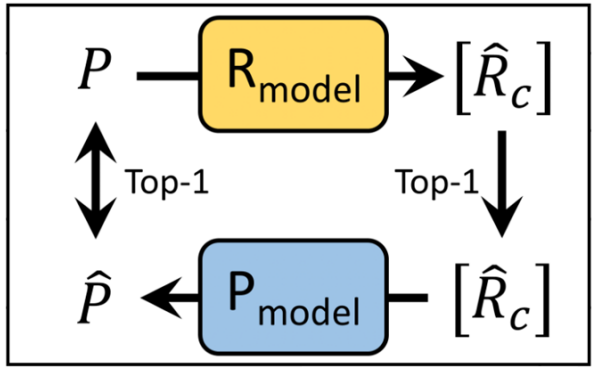


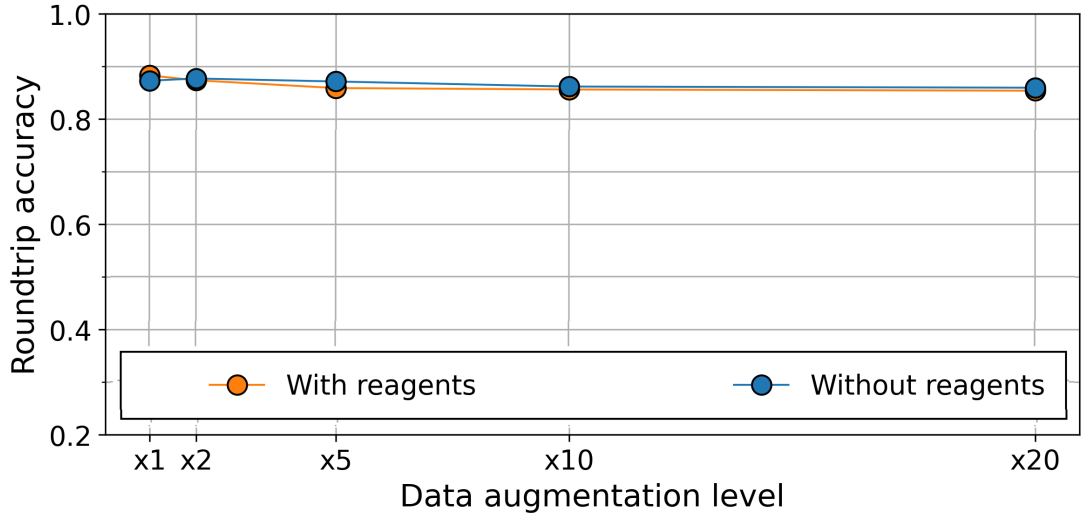



Prediction task
Product prediction
Reactant(s) prediction
Reagent(s) prediction
Data pipeline
Evaluation
Simple metrics
Round-trip analysis
Expert evaluation
Transformer
Datasets
Artificial data augmentation
+/- reagent information
Model and input representation
SoftMax
Linear
Decoder
Feedforward Layer
Multi-head attention
Masked multi-head attention
N blocks
Encoder
Feedforward Layer
Multi-head attention
N blocks
Embedding strategy
From scratch
Pretrained
Molecule format
SMILES
SELFIES
Tokenization
Atom-level
Byte-pair encoding

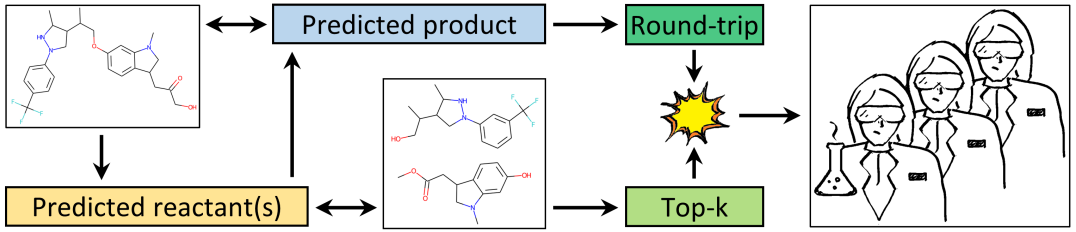

PMI rations of products
PMI2/PMI3
PMI1/PMI3
sphere
rod
disc
0.0
0.2
0.4
0.6
0.8
1.0

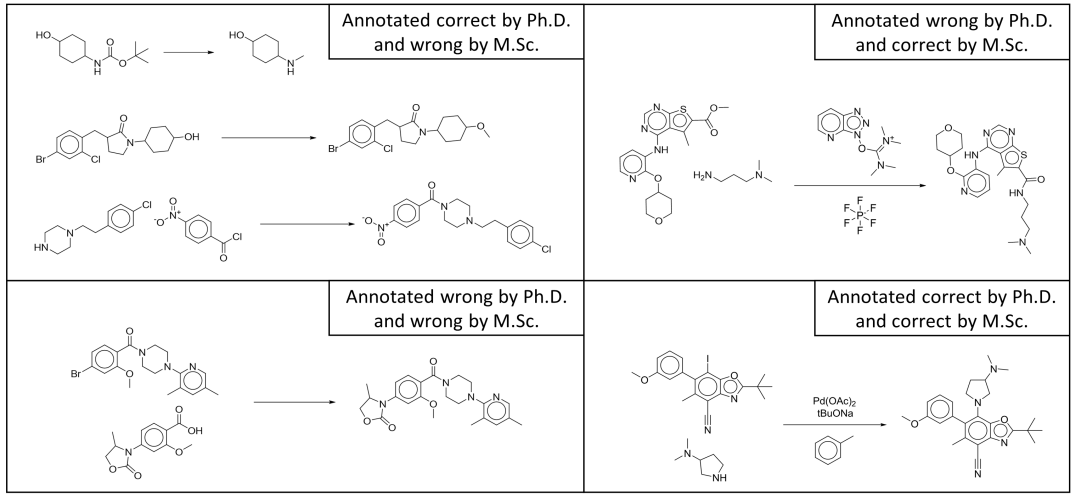

Top-1 accuracy
Data augmentation level
Product pred. (reag+)
Reactant pred. (reag+)
Reagent pred.
Product pred. (reag-)
Reactant pred. (reag-)
0.0
0.2
0.4
0.6
0.8
1.0
x1
x2
x5
x10
x20

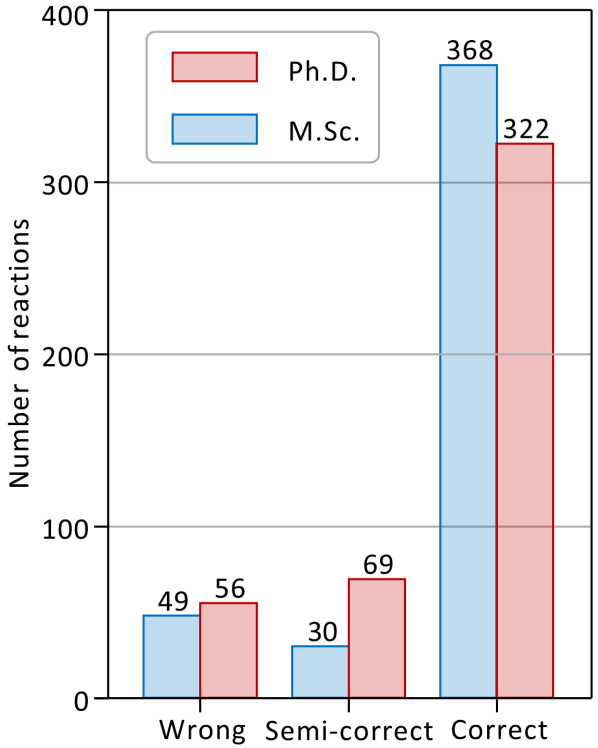




gold standard







Named entity recognition in chemical patents using ensemble of contextual language models

Copara *et al.*, CLEF, 2020

REACTION EXTRACTION FROM PATENTS

EXAMPLE_LABEL
Example 15A

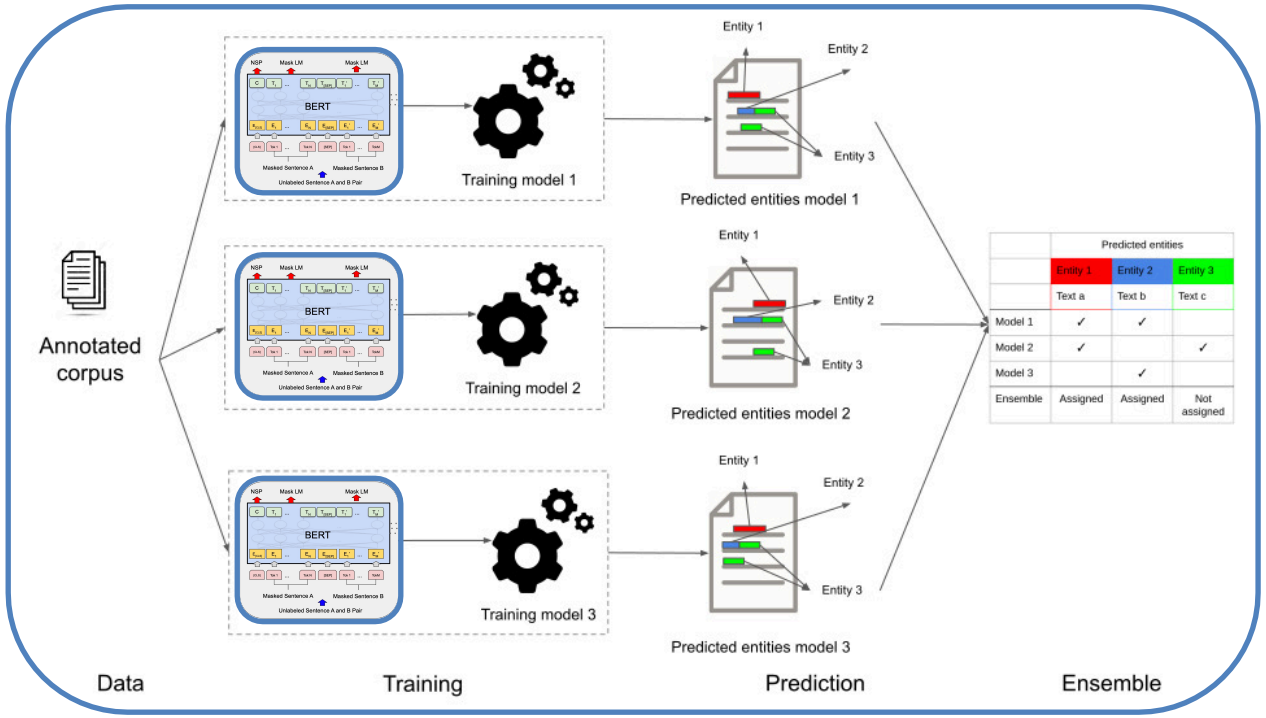
REACTION_PRODUCT
1-(2-Methoxyethyl)-5-methyl-2,4-dioxo-3-(2-phenylethyl)-1,2,3,4-tetrahydrothieno[2,3-d]pyrimidine-6-carboxylic acid

REAGENT_CATALYST STARTING_MATERIAL
75 ml of trifluoroacetic acid were added to a solution of 5.0 g (11.2 mmol) of the compound from Ex. 10A in 225 ml of

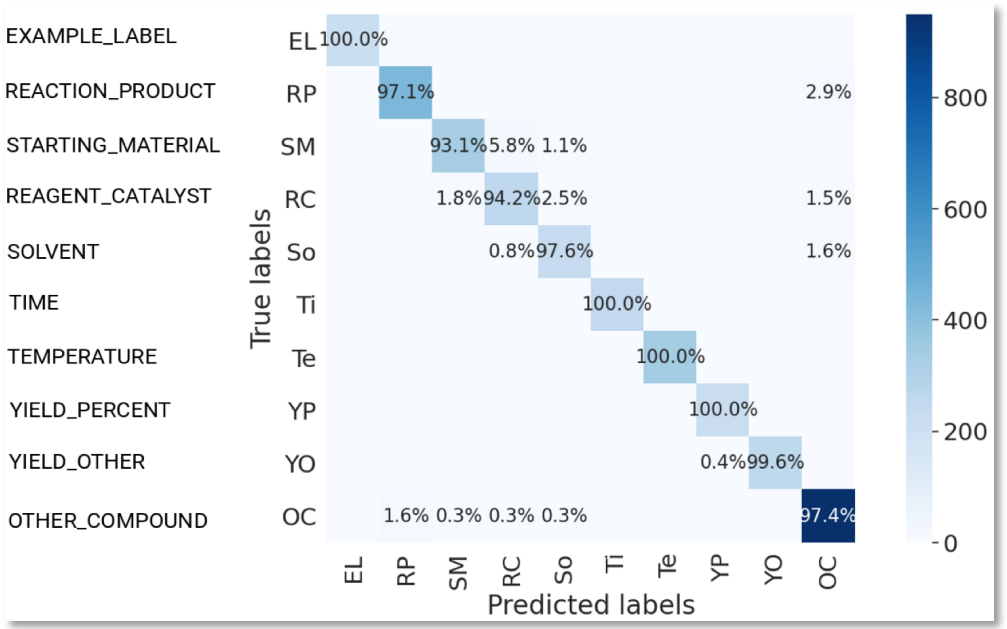
SOLVENT TEMPERATURE TIME
dichloromethane, and the mixture was stirred at RT for 2 h. The reaction mixture was then concentrated to

OTHER_COMPOUND
dryness on a rotary evaporator. The remaining residue was stirred in diethyl ether and filtered off with suction, and

YIELD_OTHER YIELD_PERCENT REACTION_PRODUCT
the solid was dried under high vacuum. 4.1 g (92% of theory) of the title compound were obtained.



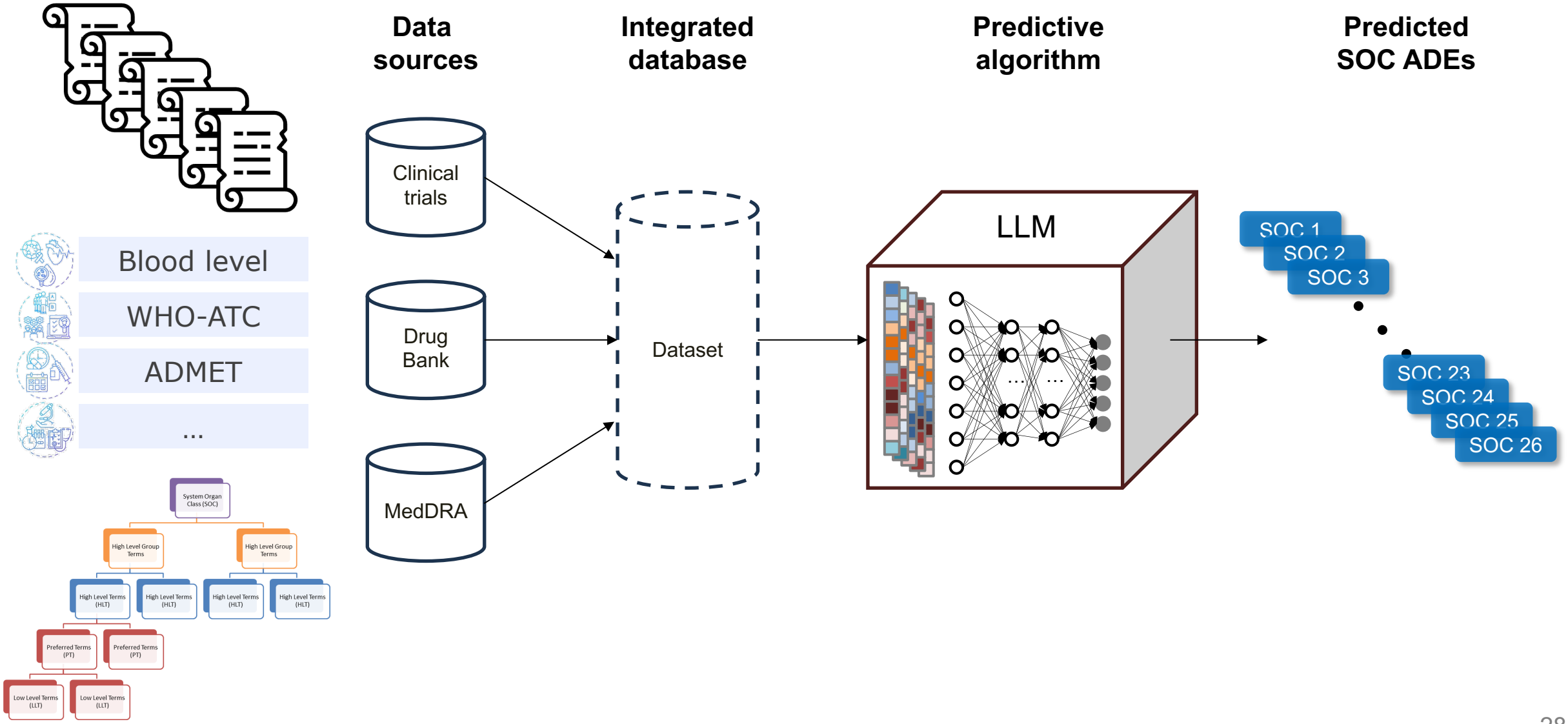
Team	Precision		Recall		F_1-score	
	exact	relaxed	exact	relaxed	exact	relaxed
Melaxtech	0.9571	0.9690	0.9570	0.9687	0.9570	0.9688
DS4DH (run 3)	0.9378	0.9692	0.9087	0.9558	0.9230	0.9624
DS4DH (run 2)	0.9083	0.9510	0.9114	0.9684	0.9098	0.9596
Baseline (BANNER)	0.9071	0.9219	0.8723	0.8893	0.8893	0.9053

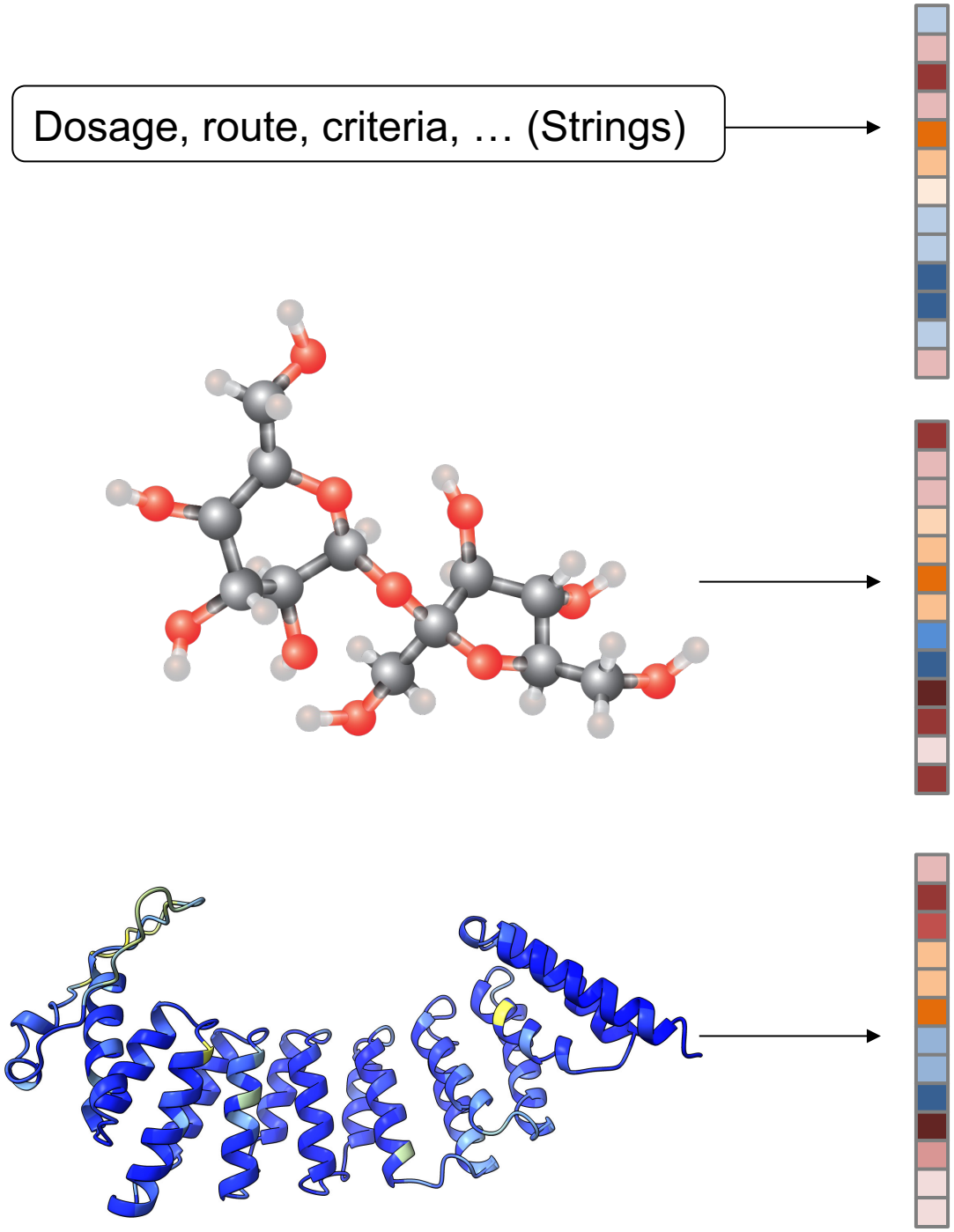


Prédiction des effets indésirables potentiels des médicaments pour les nouvelles molécules en cours d'essais cliniques

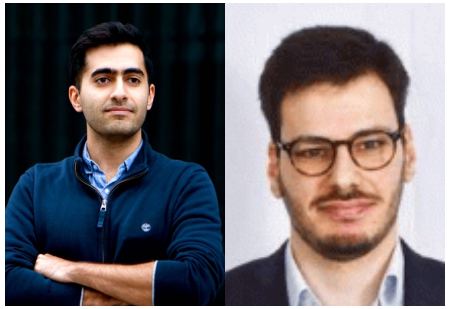
Vicente Alvarez & Yazdani, 2022

PREDICTING POTENTIAL ADRS FOR NEW MOLECULES


Data sources
Integrated database
Predictive algorithm
Predicted SOC ADEs
Blood level
WHO-ATC
ADMET
...
Clinical trials
Drug Bank
MedDRA
Dataset
LLM
SOC 1
SOC 2
SOC 3
SOC 23
SOC 24
SOC 25
SOC 26
System Organ Class (SOC)
High Level Group Terms
High Level Terms (HLT)
Preferred Terms (PT)
Low Level Terms (LLT)



Model	F1-score
Naïve	38%
SMILES (ChemBERTa)	58%
Baseline (CT+DB)	64%
Baseline + MACCS	64%
Baseline + GROVER	63%
Baseline + ESMFold-2	64%
Dosage/route + SMILES (ChemBERTa)	**69%**



Deep learning-based risk prediction for interventional clinical trials based on protocol design: A retrospective study

Ferdowsi *et al.*, Patterns, 2023

RISK PREDICTION OF CLINICAL TRIALS

Protocol

NCT04372602

COVID-19

Intervention

0

Duvelisib

2

Placebo

1

Peripheral blood draw

Eligibility

Exclusion

1. Known allergy
2. Pregnant and/or breastfeeding
…

Inclusion

1. Age ≥ 18 years or older
2. A diagnosis of advanced COVID-19
3. Adequate hematologic function
…

Design

Phase

II

Enrolment

28

Outcome

Primary

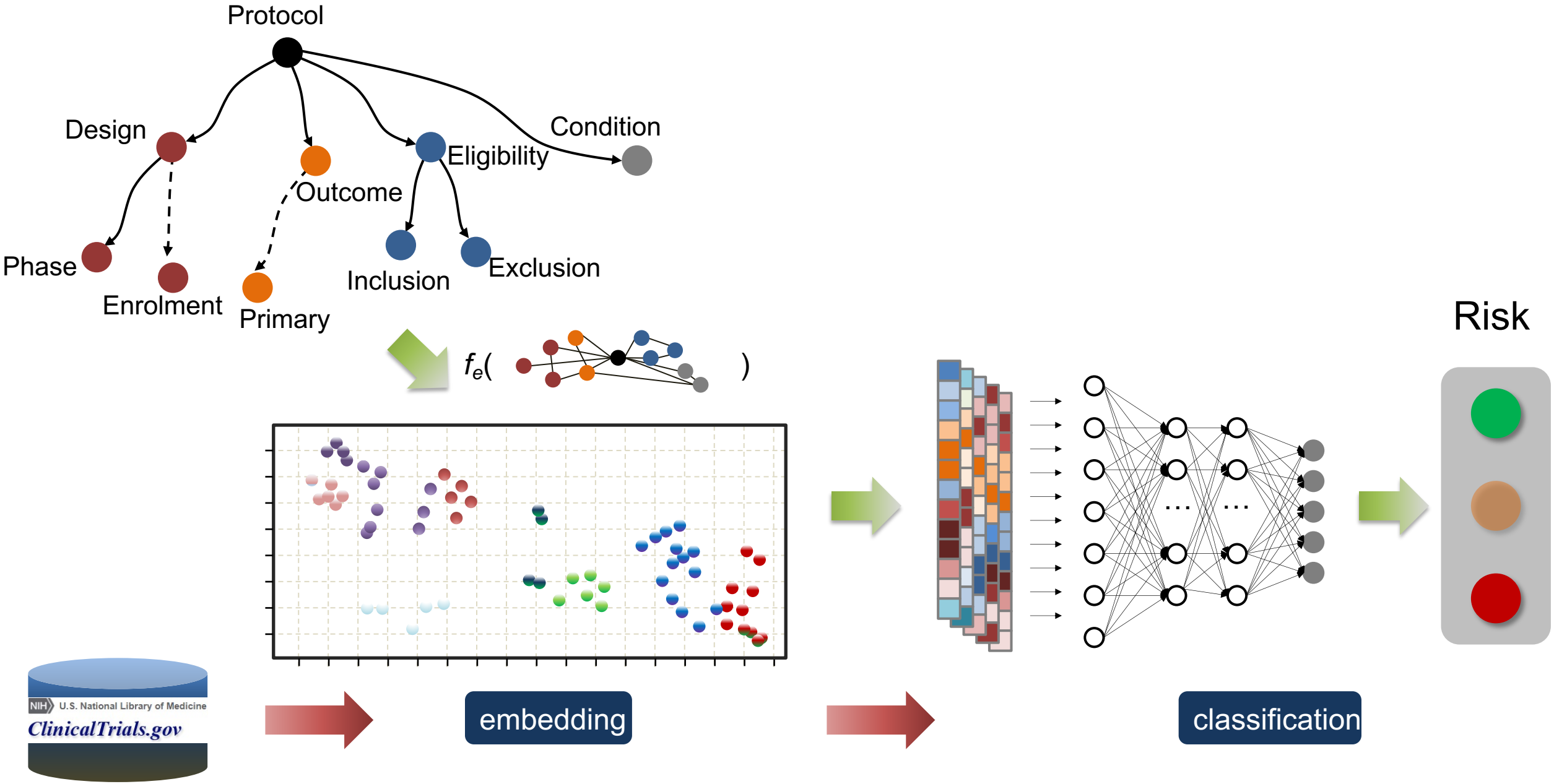
Overall Survival Through 28 Days

Secondary

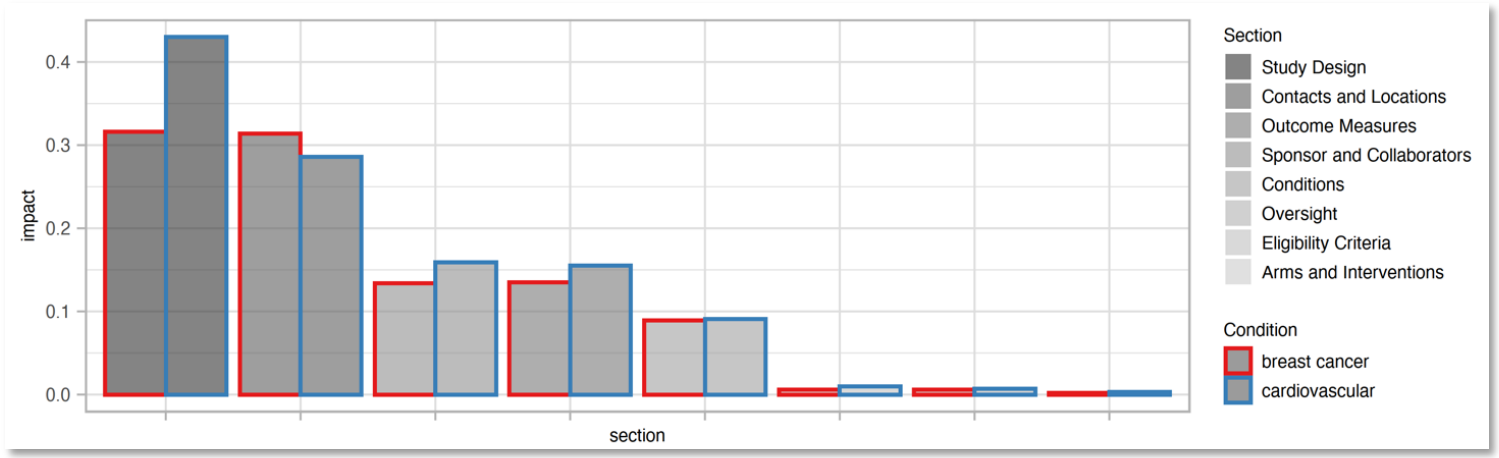
Length of ICU stay
…

Status

Completed


Protocol
Design
Condition
Eligibility
Outcome
Phase
Enrolment
Primary
Inclusion
Exclusion
$f_e($)
Risk
NIH U.S. National Library of Medicine
ClinicalTrials.gov
embedding
classification

Method	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	AUROC (%)
(Elkin and Zhu, 2021)	-	-	-	-	72.8
fastText (Joulin et al., 2017)	84.9	72.0	75.3	84.0	84.6
BERT-flat-1	65.1	67.6	62.6	63.5	72.5
BOW-flat-9	64.9	67.1	64.9	63.5	73.7
BERT-flat-9	81.4	81.4	81.4	84.2	89.1
BOW+RP-GCN-global	81.8	82.3	82.1	84.6	91.2
BERT-GCN-global	84.3	85.0	84.6	86.7	88.8
BOW+RP-GCN-selective-9	84.2	83.4	83.8	86.3	90.8
BERT-GCN-selective-9	**84.5**	**85.2**	**84.8**	**87.0**	**92.7**

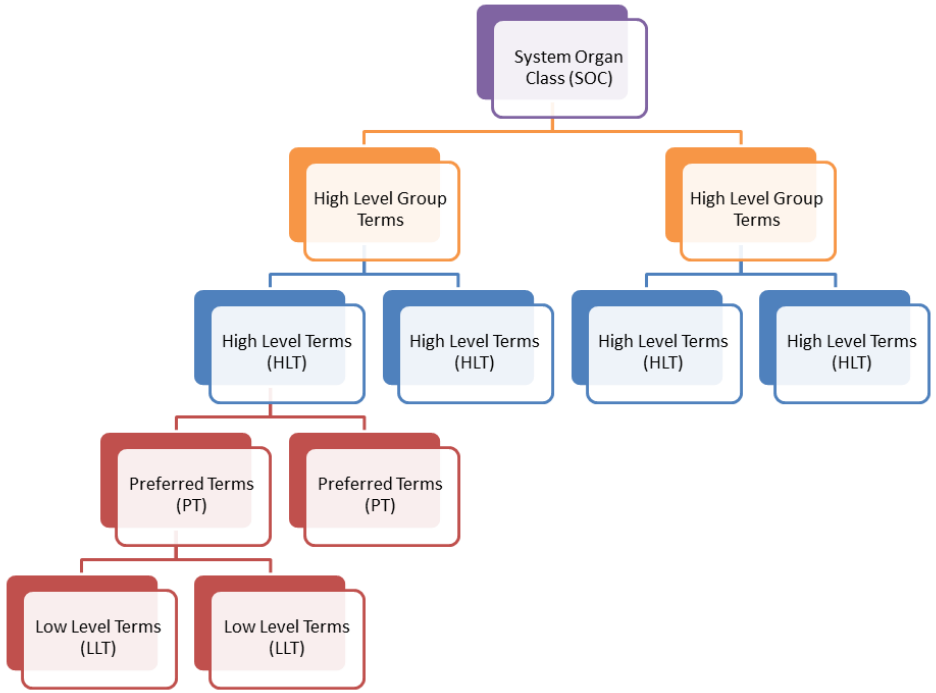


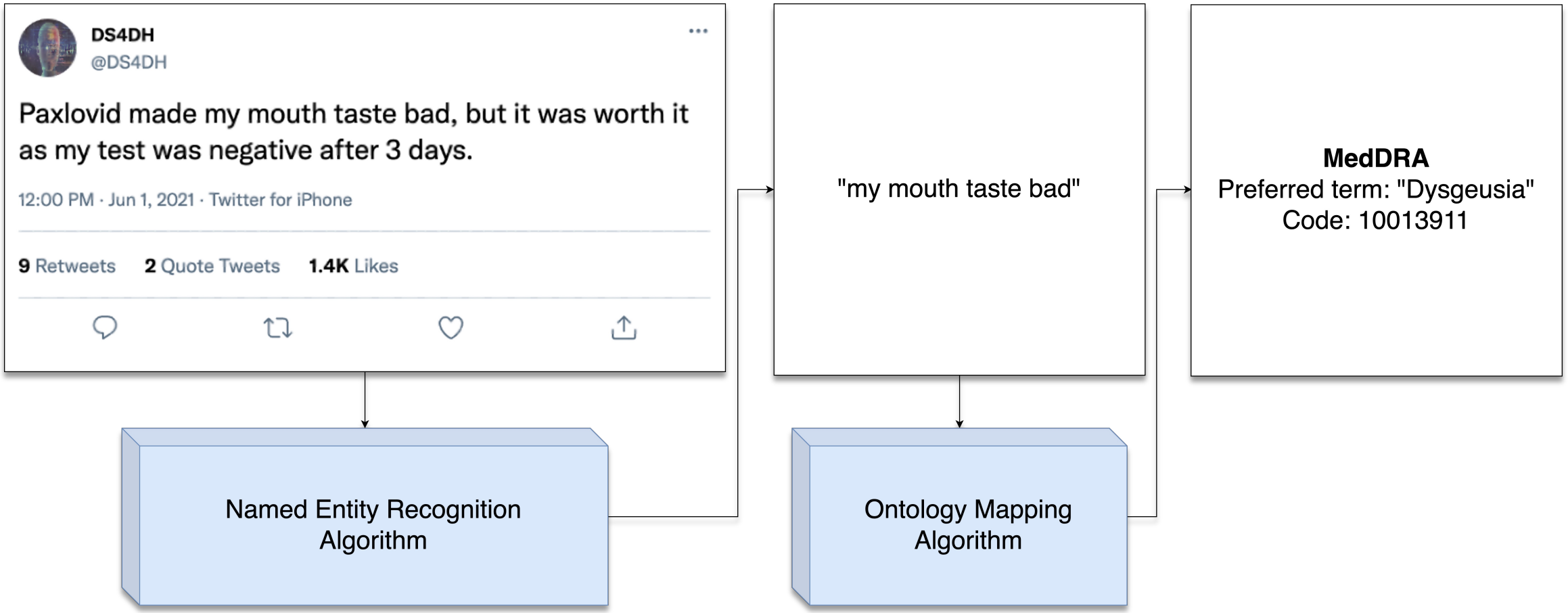
DS4DH at #SMM4H 2023: Zero-Shot Adverse Drug Events Normalization using Sentence Transformers and Reciprocal-Rank Fusion

Yazdani *et al.*, SMM4H, 2023

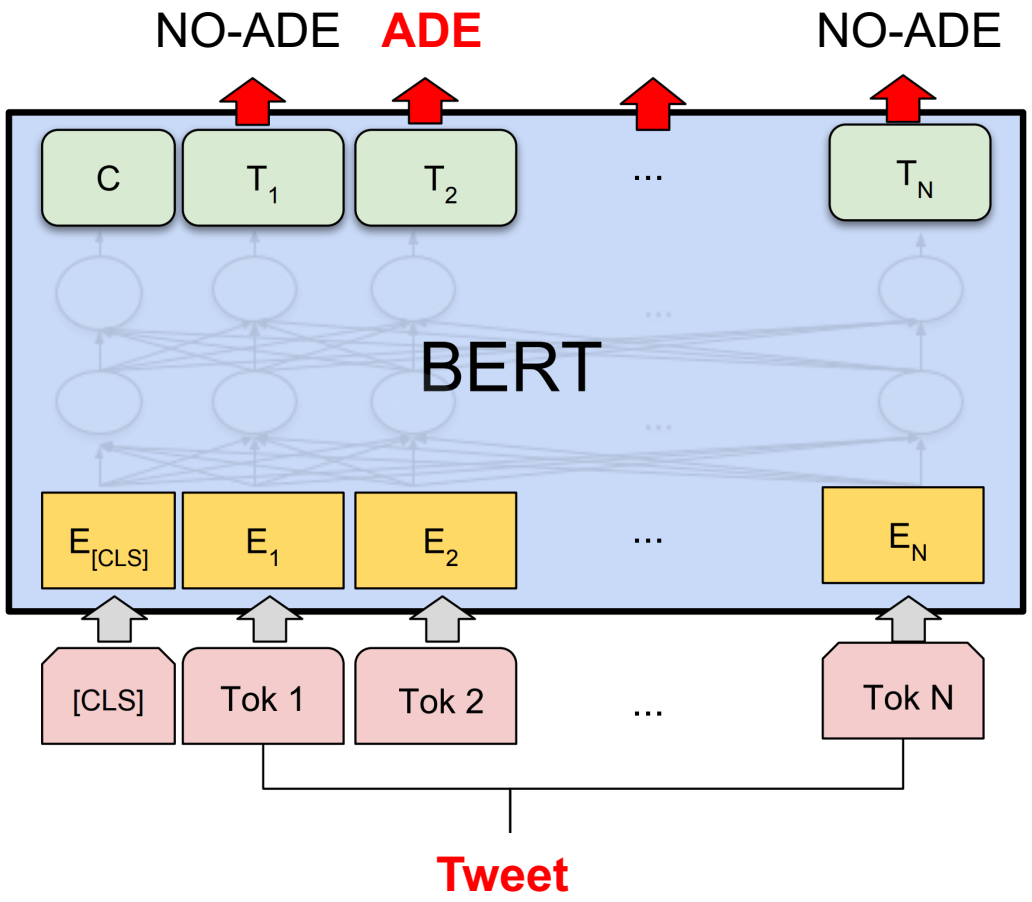
IDENTIFYING ADRS IN SOCIAL NETWORK


DS4DH
@DS4DH
Paxlovid made my mouth taste bad, but it was worth it as my test was negative after 3 days.
12:00 PM · Jun 1, 2021 · Twitter for iPhone
9 Retweets
2 Quote Tweets
1.4K Likes

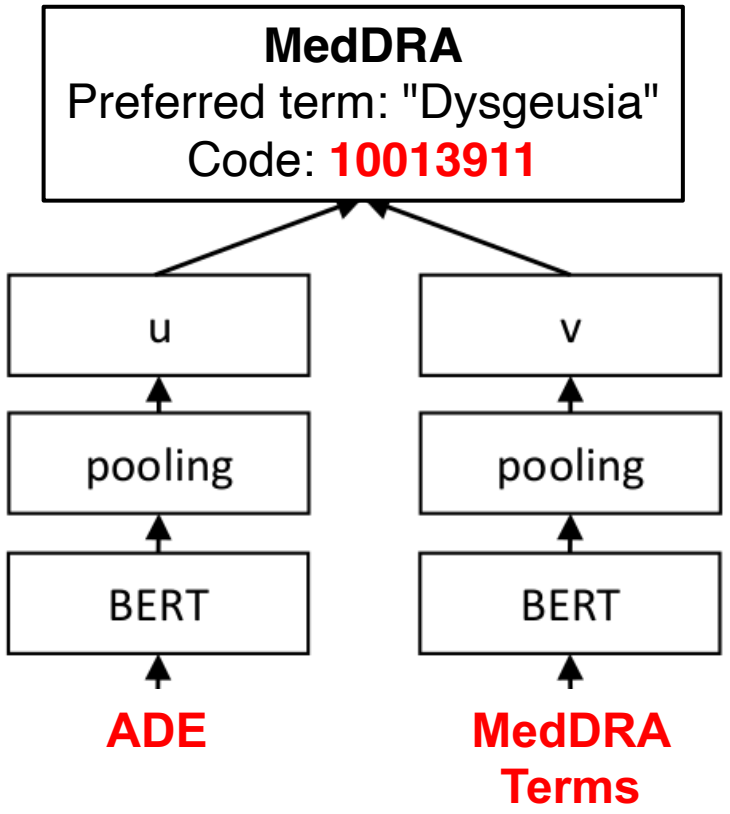

System Organ Class (SOC)
High Level Group Terms
High Level Group Terms
High Level Terms (HLT)
High Level Terms (HLT)
High Level Terms (HLT)
High Level Terms (HLT)
Preferred Terms (PT)
Preferred Terms (PT)
Low Level Terms (LLT)
Low Level Terms (LLT)


DS4DH
@DS4DH
Paxlovid made my mouth taste bad, but it was worth it as my test was negative after 3 days.
12:00 PM · Jun 1, 2021 · Twitter for iPhone
9 Retweets
2 Quote Tweets
1.4K Likes
Named Entity Recognition Algorithm
"my mouth taste bad"
Ontology Mapping Algorithm
MedDRA
Preferred term: "Dysgeusia"
Code: 10013911

Named entity recognition



Ontology mapping



Model	Overall			Unseen		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Mean	0.293	**0.422**	0.329	0.151	**0.360**	0.202
Median	0.249	0.405	0.322	0.128	0.354	0.195
Ours	**0.449**	0.405	**0.426**	**0.249**	0.354	**0.292**

Summary

- LLMs are powerful, generalisable machine learning models
- Drug discovery and development can benefit of the effectiveness LLMs in many ways
  - Synthesis/retrosynthesis, ADR prediction/extraction, CT analyses, …
- Outlook: While *discriminative* LLMs can provide strong performance in specific tasks, *generative* models can tackle more complex, general tasks

DS4DH

DATA SCIENCE FOR DIGITAL HEALTH

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Agency for Development and Cooperation SDC

Swiss National Science Foundation

Innosuisse - Swiss Innovation Agency

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

European Commission

RISKLICK

SpiroChem
GET INSPIRED. INNOVATE.

CLINERION
Real World Data Solutions

References

1. Jaume-Santero F, Bornet A, Valery A, Naderi N, Vicente Alvarez D, Proios D, Yazdani A, Bournez C, Fessard T, Teodoro D. Transformer Performance for Chemical Reactions: Analysis of Different Predictive and Evaluation Scenarios. Journal of chemical information and modeling. 2023 Mar 23;63(7):1914-24.
2. Copara J, Naderi N, Knafou J, Ruch P, Teodoro D. Named entity recognition in chemical patents using ensemble of contextual language models. InProceedings of the CLEF 2020 conference 2020. 22-25 September 2020.
3. He J, Nguyen DQ, Akhondi SA, Druckenbrodt C, Thorne C, Hoessel R, Afzal Z, Zhai Z, Fang B, Yoshikawa H, Albahem A. An extended overview of the CLEF 2020 ChEMU Lab: information extraction of chemical reactions from patents. InProceedings of the CLEF 2020 conference 2020. 22-25 September 2020.
4. Ferdowsi S, Knafou J, Borissov N, Alvarez DV, Mishra R, Amini P, Teodoro D. Deep learning-based risk prediction for interventional clinical trials based on protocol design: A retrospective study. Patterns. 2023 Mar 10;4(3).
5. Ferdowsi S, Copara J, Gouareb R, Borissov N, Jaume-Santero F, Amini P, Teodoro D. On Graph Construction for Classification of Clinical Trials Protocols Using Graph Neural Networks. InArtificial Intelligence in Medicine: 20th International Conference on Artificial Intelligence in Medicine, AIME 2022, Halifax, NS, Canada, June 14–17, 2022, Proceedings 2022 Jun 14 (pp. 249-259).
6. Ferdowsi S, Borissov N, Knafou J, Amini P, Teodoro D. Classification of hierarchical text using geometric deep learning: the case of clinical trials corpus. InProceedings of the 2021 Conference on Empirical Methods in Natural Language Processing 2021 Nov (pp. 608-618).
7. Yazdani A, Rouhizadeh H, Alvarez DV, Teodoro D. DS4DH at# SMM4H 2023: Zero-Shot Adverse Drug Events Normalization using Sentence Transformers and Reciprocal-Rank Fusion. arXiv preprint arXiv:2308.12877. 2023 Aug 15.