

AI@MPA toolbox: AI for medicines regulations

Seamus Doyle

Data Scientist - AI unit, Swedish Medical Products Agency

AI@MPA toolbox

AI-toolbox for support in medicines regulation

- **M-RECON** – medicines regulatory context engine
- **PICROSS** – product information cross search
- **PACKSIM** – package similarity search
- **SCHEMA** – SMILES chemical embeddings map
- **REGULUS** – regulatory universal support



AI@MPA

AI for medicines regulation provided by the Swedish Medical Products Agency

M-RECON

Medicines regulatory context engine

Available

PICROSS

Product information cross search

Available

PACKSIM

Package similarity search

Available

RADAR

Raw data augmented review

Coming soon

SCHEMA

SMILES chemical embeddings map

Available

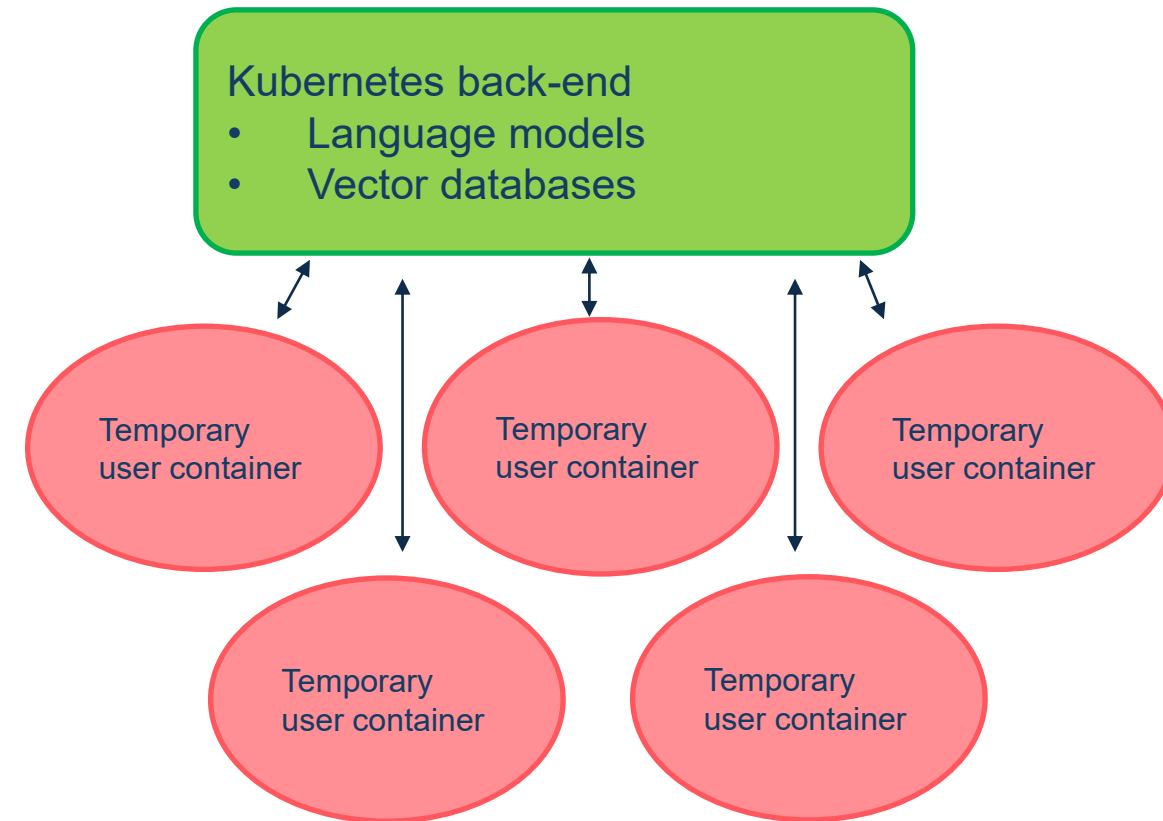
REGULUS EU

Regulatory universal support

Available

IT and information security

- Use of several overlapping safety principles
 - **encryption, access restriction and immediate deletion of uploaded information**
 - documents from the dossiers containing intellectual property from applicants, and assessment reports with non-public regulatory positions, can be safely used as index documents for the semantic mapping of regulatory context.

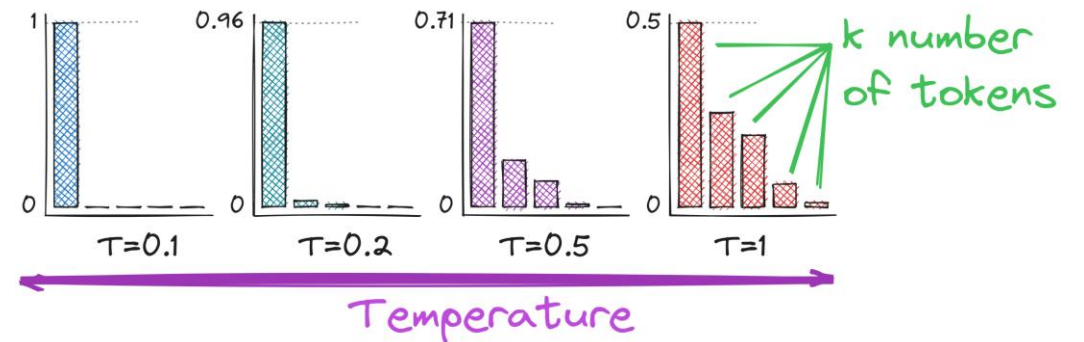
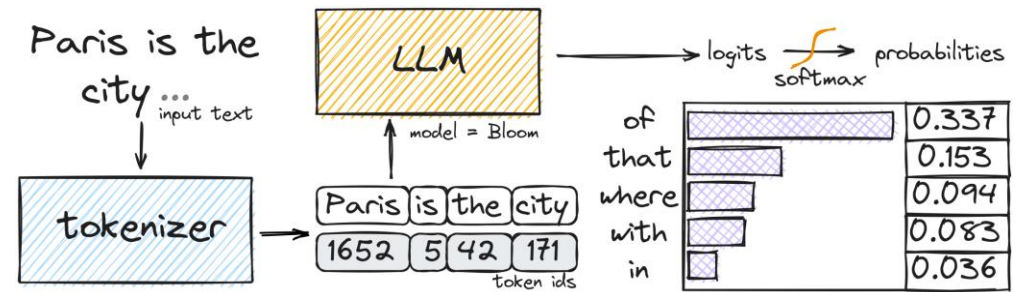
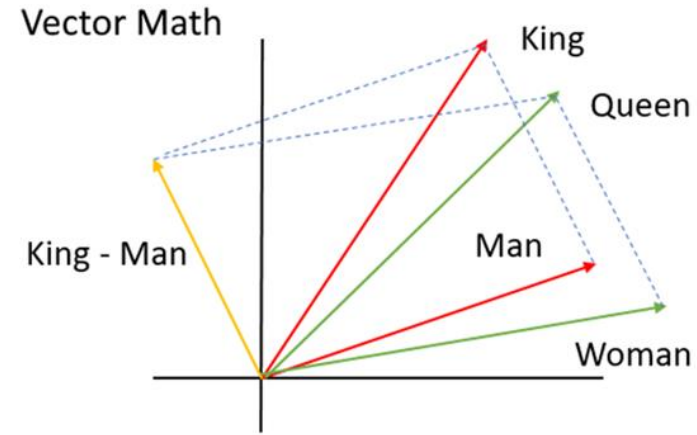
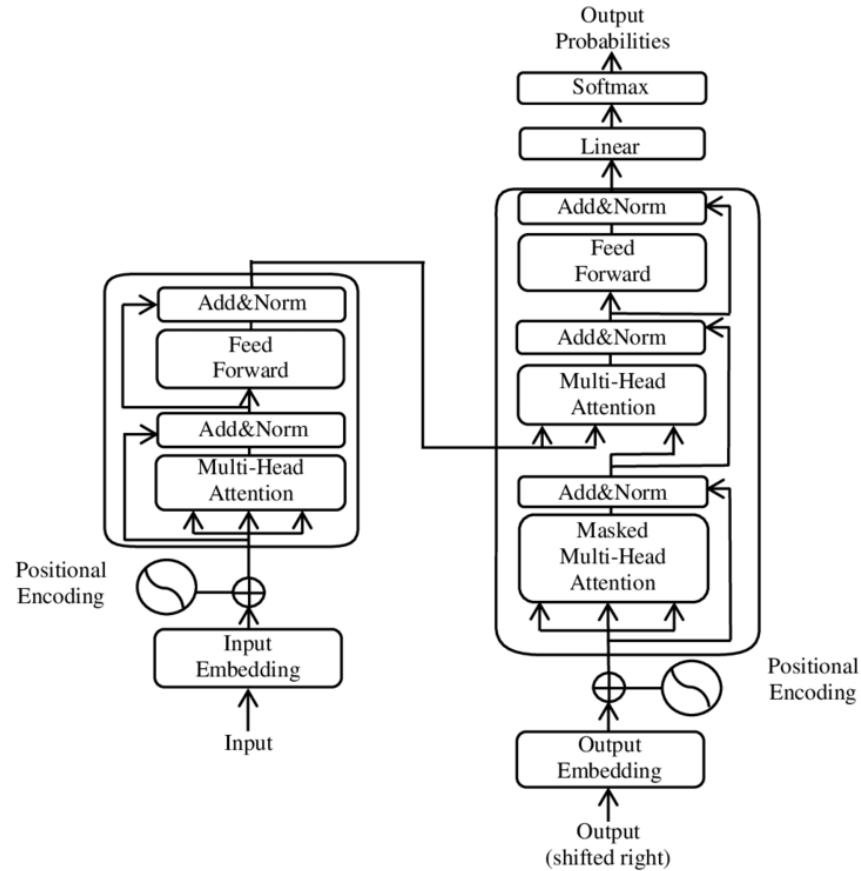


AI@MPA toolbox – Our users

- Toolbox is used by 24 agencies in the European network
- To get access
 - All EU/EES NCAs invited
 - Email: gabriel.westman@lakemedelsverket.se

Swedish MPA
FIMEA
FAGG/AFMPS
PEI
MEB
SAM
NOMA
AGES
ANSM
HALMED
EOF
AEMPS
BfArM
SÚKL
EMA
DKMA
Medicines Authority
IMA
JAZMP
INFARMED
AIFA
Ministry of Health
HPRA
BDA

Key LLM concepts



LLM capabilities and limitations

- LLMs are primarily trained for **plausible text generation** – not to be truthful
- Work best as conversational agents and text generators
- Reasoning skills vary
- Hallucinations are common

You

Give me list of references on proteomic pathway analysis in herpes simplex encephalitis

ChatGPT

Here's a list of references on proteomic pathway analysis in herpes simplex encephalitis:

1. Lubick, K. J., & Robertson, S. J. (2015). The proteomic profile of HSV-1-infected human T cells reveals protein changes occurring in immunological pathways and cell cycle regulation. *Journal of Proteomics*, 123, 44-58.
2. Noriega, V. M., & Hayashi, K. (2017). Immunoproteomic analysis of antibody response to herpes simplex virus type 1 glycoprotein D epitope in patients with Herpes Simplex Encephalitis. *Proteomics Clinical Applications*, 11(5-6), 1600091.
3. Zhou, G., Du, T., & Roizman, B. (2013). The role of the amino terminus of herpes simplex virus 1 protease in the formation of the virus-induced capsid structure. *Journal of Biological Chemistry*, 288(4), 2978-2982.
4. Sullivan, M. A., & Kim, C. (2019). Pathways of apoptosis and importance in development. *Journal of Cellular Biology*, 28(3), 234-245.
5. Verjans, G. M., & Hintzen, R. Q. (2012). Host cell protein interactions with herpes simplex virus: proteomic analysis of cell proteins bound to the viral protein ICPO. *Proteomics*, 12(10), 1608-1613.
6. Schoen, A., & Struyf, F. (2016). Early dynamics of protein complexes during herpes simplex virus 1 infection. *Journal of Virology*, 90(9), 4151-4163.
7. Hansen, H. J., & Caswell, R. J. (2014). Proteomic analysis of herpes simplex virus type 1 primary enveloped virions. *Journal of Virology*, 88(21), 13176-13191.
8. Sandri-Goldin, R. M., & Iizasa, H. (2011). Interaction between herpes simplex virus regulatory

REGULUS - Regulatory universal support

- In-house deployed and secure service for generative AI
- Based on open-weights LLMs
- Three assistant services: *Open assistant, Document assistant and Regulatory assistant*
- Reads text documents (Word, PDF etc.), images (scans and photos) and audio for transcription

⚙️ Advanced settings

↺ Reset chat

Assistant context: 2/2 messages

Explain the difference between Pairwise Alignment and Multiple Sequence Alignment for aligning nucleotide sequences.



Introduction to Sequence Alignment

Sequence alignment is a fundamental technique in bioinformatics used to compare nucleotide or amino acid sequences. This comparison helps identify similarities and differences between sequences, which can provide insights into their evolutionary relationships, functional similarities, and structural properties.

Pairwise Alignment

Pairwise sequence alignment involves comparing two sequences at a time. This method is used to identify regions of similarity between two sequences, which can indicate common ancestry or functional similarities. Pairwise alignment algorithms, such as the Needleman-Wunsch algorithm or the Smith-Waterman algorithm, use dynamic programming to find the optimal alignment between two sequences. The output of pairwise alignment is typically a score that reflects the degree of similarity between the sequences and the aligned sequences themselves, showing matching and mismatching regions.

Multiple Sequence Alignment

Multiple Sequence Alignment (MSA) extends the concept of pairwise alignment by comparing more than two sequences simultaneously. MSA is crucial for analyzing the relationships among a group of sequences, identifying conserved regions that may be functionally important, and reconstructing phylogenetic trees.

Multiple sequence alignment algorithms, such as CLUSTALW or MUSCLE, aim to find the best alignment among

Enter a message...



Open assistant

- Open, multi-stage chatbot that can answer **general-knowledge questions**, process **user-provided information** and provide simple **drafts**
- Due to model data-lock, it is not useful to ask for recent information or current events
 - Proposed use case #1 – ask for a help in understanding a scientific concept that you are not fully familiar with. Ask follow-up questions to adjust the level of detail in the answer to match your learning needs.
 - Proposed use case #2 – paste a long incoming e-mail and ask for action points.
 - Proposed use case #3 – ask for a linguistic review of a short text you have written, possibly including translation

REGULUS

Regulatory universal support

Open assistant

Document assistant

Regulatory assistant

⚙️ Advanced Settings

Upload Text Files

Browse...

RAC25_Programr

Upload complete

Uploaded Documents:

RAC25_Programme (3).pdf

Context Utilization: 3%

🗑️ Remove Uploads

Summarise the agenda below and describe the demography of the speakers



Agenda Summary

The agenda consists of 7 sessions and 3 webinars, covering various topics related to the pharmaceutical industry, including:

1. Regulatory landscape and future direction of the pharmaceutical sector
2. Union list of critical medicines
3. Regulatory affairs in the digital age
4. Leveraging new technologies in regulatory science
5. Revised Variations Regulation
6. Impact of revised pharmaceutical legislation on product information
7. Predictability of submissions

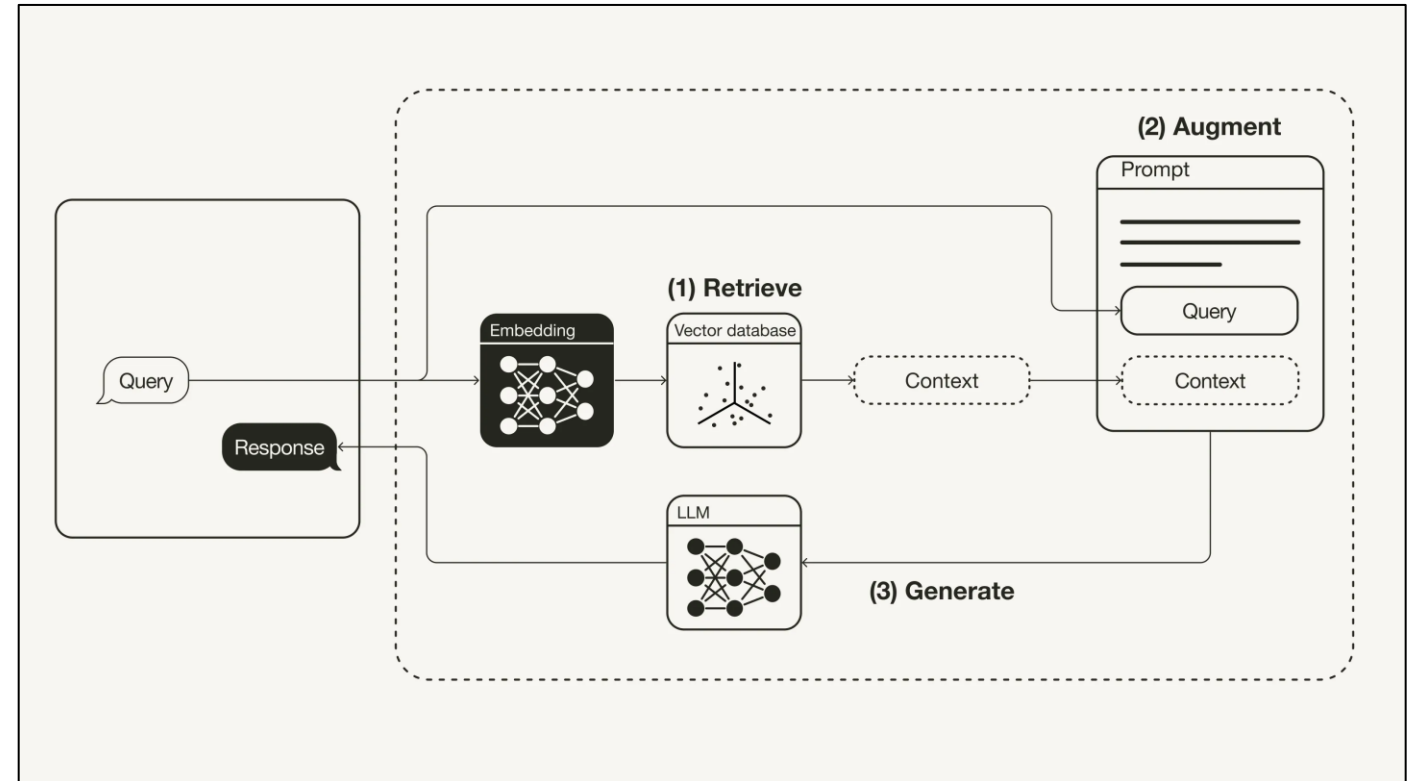
The webinars cover topics such as shortages reporting, pharmacovigilance, and environmental risk assessment.

Document assistant

- Intended for long-context document processing
- Maximum context length 100 000 tokens (approx. 75 000 words)
 - Proposed use case #1 – upload a document and ask for a summary
 - Proposed use case #2 – ... ask for a translation into a language of your choice
 - Proposed use case #3 – upload two documents and ask for a comparison from an aspect of your choice (e.g., congruence, conformance, pros- and cons)

Retrieval-augmented generation

- Technique for grounding LLM output in a pre-defined information source
- The input question is used to perform a semantic search in a vector database
- Output is generated from the semantic matches only



REGULUS EU

Regulatory universal support

Open assistant

Document assistant

Regulatory assistant

Context of use

⚙️ Advanced settings

Include:

- ☒ EMA guidelines
- ☐ EMA qualification
- ☐ EPARs
- ☐ EU product information
- ☐ FDA guidance
- ☐ US product information
- ☐ Eudralex
- ☐ MDCG

[Database status](#)

When is prospective validation required for AI endpoints in pre-clinical studies?



[The following output was created using generative AI. Keep this statement until human review of factual statements, completeness and reasoning has been completed.]

Prospective Validation for AI Endpoints

Prospective validation for AI endpoints in pre-clinical studies is required when the AI model is intended to be used as part of an integrated risk assessment for regulatory purposes [3]. According to the EMA guidelines, prior to model deployment in high regulatory impact settings, such as in relation to the primary endpoint, performance should be tested with prospectively generated data (future calendar time) that is acquired in a setting or population representative of the intended context of use [3]. This implies that prospective validation is necessary to ensure the reliability and accuracy of AI endpoints in pre-clinical studies, especially when they have a significant impact on regulatory decision-making [6].

References

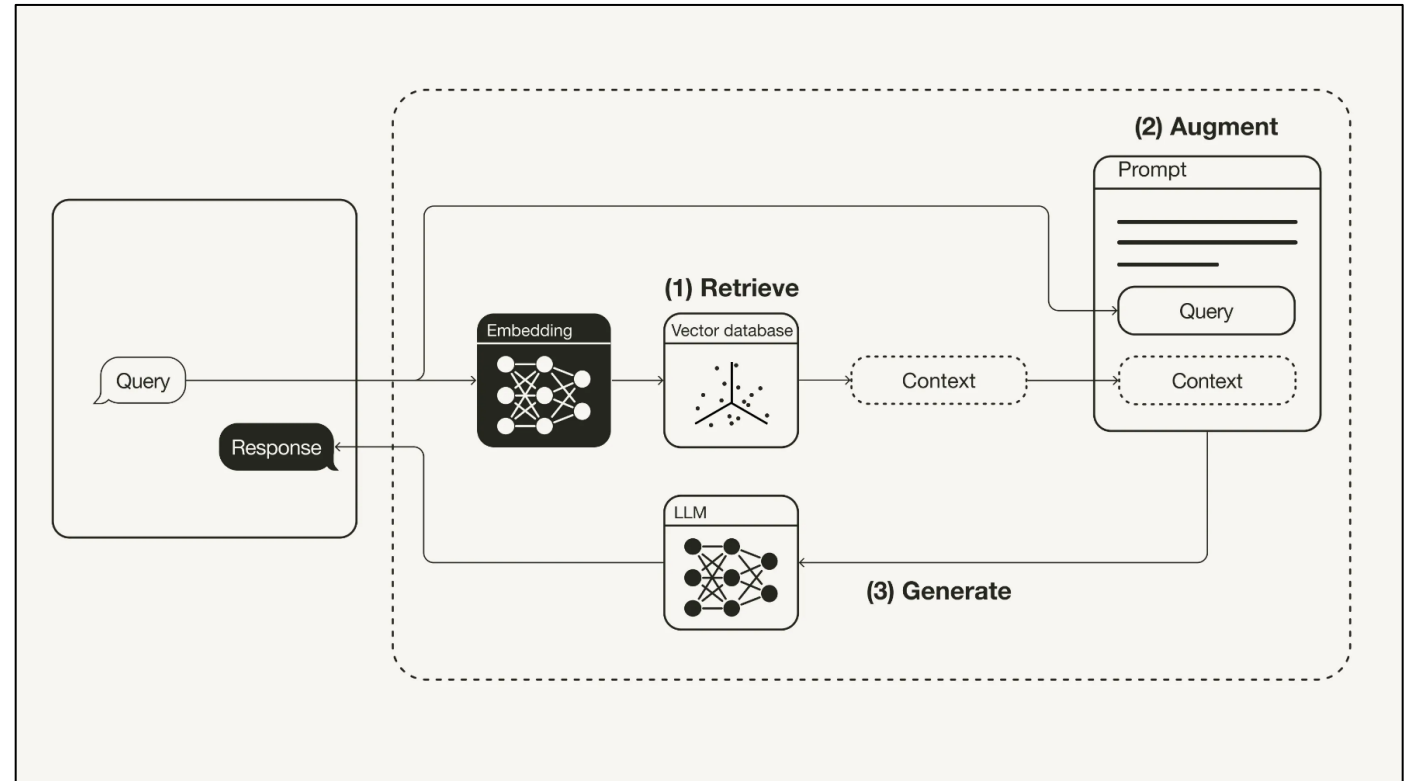
- [Overview of comments received on 'Reflection paper on regulatory requirements for the development of medicinal products for chronic non-infectious liver diseases \(PBC, PSC, NASH\)' \(EMA/CHMP/299976/2018\) page 24 \(EMA guidelines\)](#)
- [Overview of comments received on 'Guideline on quality, non-clinical and clinical requirements for investigational advanced therapy medicinal products in clinical trials' \(EMA/CAT/123573/2024\) - second public consultation page 214 \(EMA guidelines\)](#)

Enter a message...



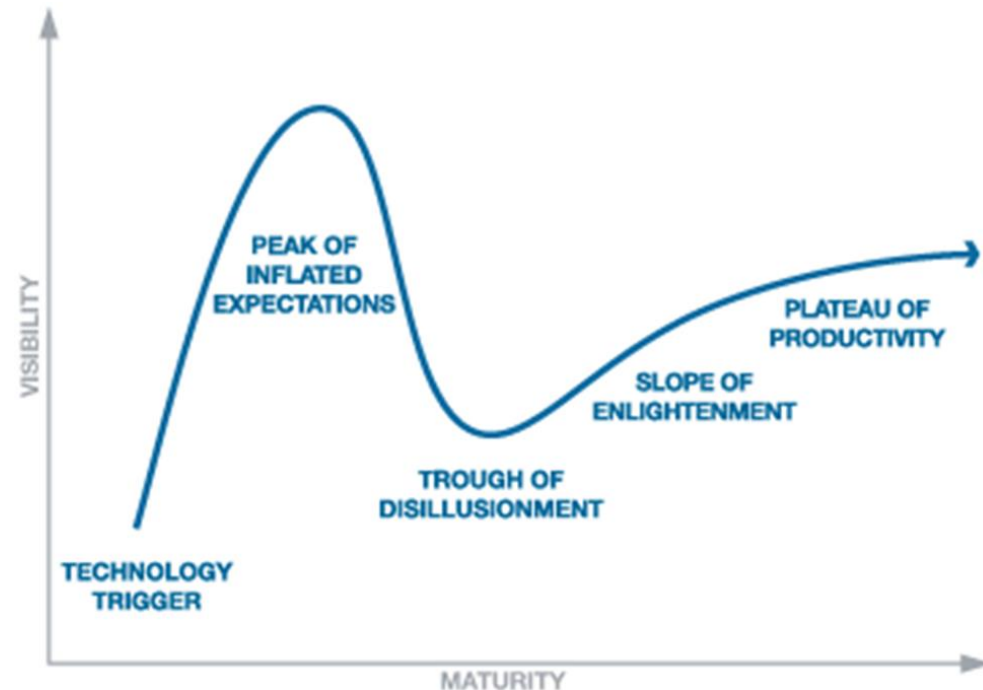
Regulatory assistant

- Vector database with EU and US regulatory landscape for medicines and medical device regulation
- Great for searching for information contained in specific documents
- Not good for horizontal comparisons



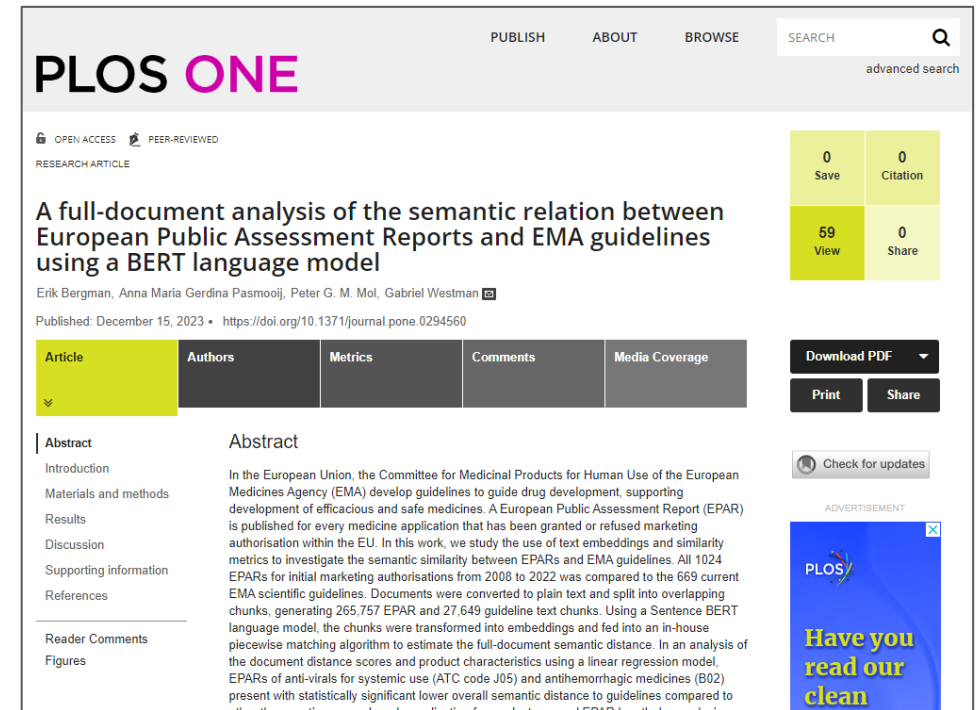
Notes on generative AI

- LLM output must be **reviewed** to mitigate risks of incorrect or incomplete output
 - See it as a junior colleague
- Aim for use cases where "90% right" is a good start (rather than a 10% risk of catastrophic failure)
- Beware of difference between valuable output and "imitation of work"



M-RECON - semantic document-level matching for medicines regulation

- Aim - to identify relevant guidelines (*and approved products of relevance for a specific procedure*)
- Ambition to include **EMA product information (v)**
- Search is best initiated from documents in the application dossier
- Search results most valuable for non-experienced assessors and those working horizontally across therapeutic areas.
- Academic paper published (Bergman et al., PLOS ONE, 2023)



M-RECON

Medicines regulatory context engine

File based search

Text based search

Database info

Context of use

Choose file

Browse...

Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle_240903.pdf

Upload complete

Search

Search domains (human - h, veterinary - v)

- ☐ EMA guidance (h)
- ☐ EU products (h)
- ☐ EMA qualifications (h)
- ☐ FDA guidance (h)
- ☐ US products (h)
- ☒ EMA guidance (v)

Results



EMA guidance (h)			EU products (h)			EMA qualifications (h)			FDA guidance (h)			US products (h)			EMA guidance (v)		
Score	Document												Source page				
75%	Draft concept paper on guidance on statistical principles for clinical trials for immunological veterinary medicinal products												https://www.ema.europa.eu/en/guidance-statistical-principles-clinical-trials-immunological-veterinary-medicinal-products-scientific-guideline				
75%	Guideline on data requirements for applications for immunological veterinary medicinal products intended for limited markets submitted under Article 23 of Regulation (EU) 2019/6 - Revision 1												https://www.ema.europa.eu/en/data-requirements-applications-immunological-veterinary-medicinal-products-intended-limited-markets-submitted-under-article-23-regulation-eu-2019-6-scientific-guideline				
74%	Guideline on quality data requirements for applications for veterinary medicinal products other than biologicals intended for limited markets												https://www.ema.europa.eu/en/quality-data-requirements-applications-veterinary-medicinal-products-other-biologicals-intended-limited-markets-scientific-guideline				
74%	Concept paper for the revision of residues guidelines to align with the definitions for withdrawal periods provided in Regulation (EU) 2019/6												https://www.ema.europa.eu/en/concept-paper-revision-residues-guidelines-align-definitions-withdrawal-periods-provided-regulation-eu-2019-6-scientific-guideline				
73%	Guideline on quality data requirements for applications for biological veterinary medicinal products intended for limited markets												https://www.ema.europa.eu/en/quality-data-requirements-applications-biological-veterinary-medicinal-products-intended-limited-markets-scientific-guideline				
73%	Concept paper on a guideline on data requirements for post-authorisation studies for antimicrobial veterinary medicinal products under Article 36(2) of Regulation (EU) 2019/6												https://www.ema.europa.eu/en/antimicrobial-resistance-surveillance-post-marketing-authorisation-commitment-scientific-guideline				
73%	Guideline on efficacy and target animal safety data requirements for applications for non-immunological veterinary medicinal products intended for limited markets submitted under Article 23 of Regulation (EU) 2019/6 - Revision 1												https://www.ema.europa.eu/en/efficacy-target-animal-safety-data-requirements-applications-non-immunological-veterinary-medicinal-products-intended-limited-markets-submitted-under-article-23-regulation-eu-2019-6-scientific				
73%	Question and answer document on requirements for pre-clinical studies submitted in support of a marketing authorisation application for a veterinary medicinal product												https://www.ema.europa.eu/en/questions-and-answers-requirements-pre-clinical-studies-submitted-support-marketing-authorisation-application-veterinary-medicinal-product-scientific-guideline				
73%	Note on the removal of the target animal batch safety test (TABST) from European Pharmacopoeia monographs												https://www.ema.europa.eu/en/note-removal-target-animal-batch-safetytest-tabst-european-pharmacopoeia-monographs-scientific-guideline				
72%	Concept paper for the development of a reflection paper on the availability and characteristics of diagnostic tests to improve the responsible use of antibiotics in animals												https://www.ema.europa.eu/en/development-reflection-paper-availability-and-characteristics-diagnostic-tests-improve-responsible-use-antibiotics-animals				

M-RECON database content

- EMA guidelines (h)
- EPARs (h)
- EMA qualification opinions/letters (h)
- FDA regulatory guidance (h)
- US product information (h)
- **EMA guidelines (v)**

Example: EMA guidelines matched to the clinical meropenem/vaborbactam dossier

Guideline

Guideline on the evaluation of medicinal products indicated for treatment of bacterial infections - Revision 3

Addendum to the guideline on the evaluation of medicinal products indicated for treatment of bacterial infections

Guideline on the use of pharmacokinetics and pharmacodynamics in the development of antimicrobial medicinal products

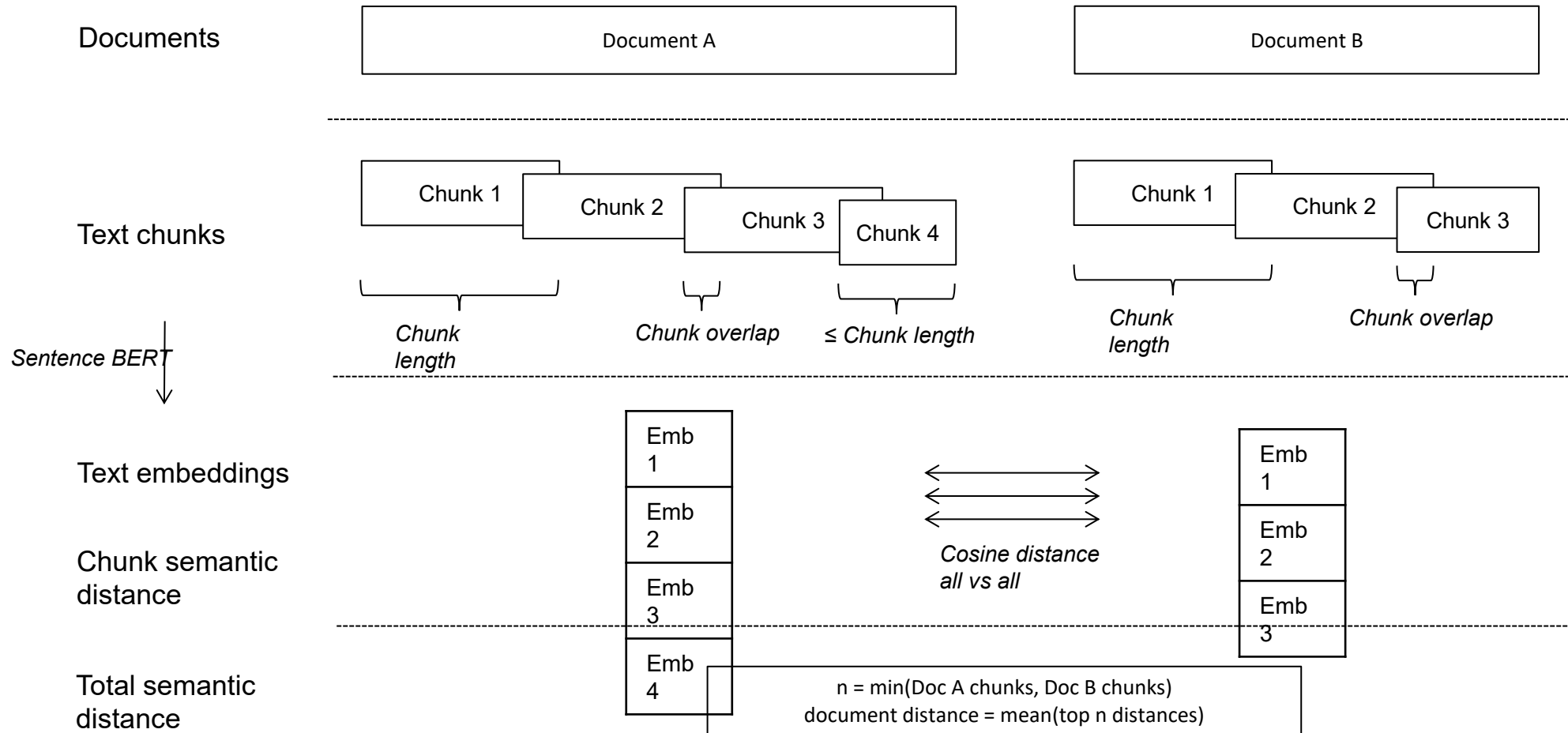
Points to consider on pharmacokinetics and pharmacodynamics in the development of antibacterial medicinal products

Addendum to the guideline on the evaluation of medicinal products indicated for treatment of bacterial infections to address paediatric-specific clinical data requirements

ICH: E 3: Structure and content of clinical study reports - Step 5

Guideline on clinical investigation of medicinal products for the treatment of Sepsis

Piecewise semantic document matching



SCHEMA - SMILES chemical embedding map

- Chemical substance semantic similarity search (based on transformer-model embeddings)
- Reads molecule structure images, SDF and SMILES strings
- Intended use – complementary tool that may provide early input into safety profile of a new chemical entity

Select substance

Similarity search

Database info

Context of use

Load substance

Substance database

Substance image:

Browse...

caffeine.png

Upload complete

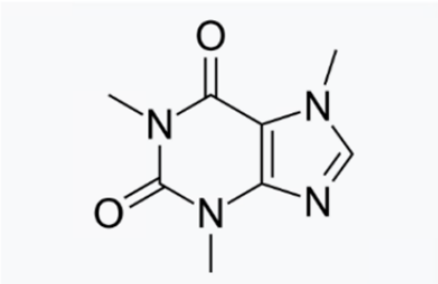
Process image

SMILES:

Cn1c(=O)c2c(ncn2C)n(C)c1=O

Process SMILES

Uploaded image



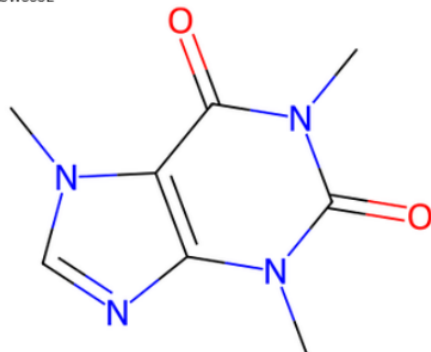
File name: caffeine.png | File size: 9696 bytes

Selected substance

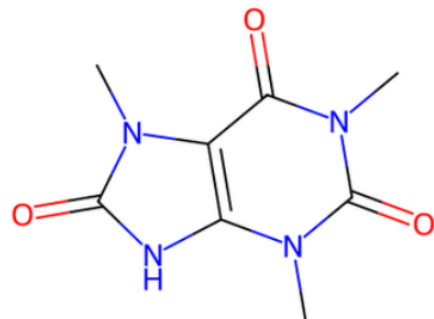


Cn1c(=O)c2c(ncn2C)n(C)c1=O

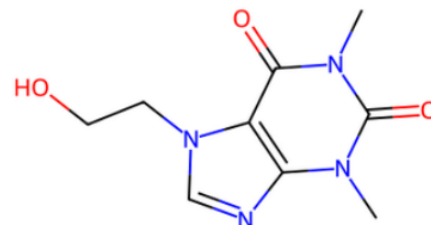
Score: 0.0
Caffeine
2519
Cn1c(=O)c2c(ncn2C)n(C)c1=O
3G8A5W338E



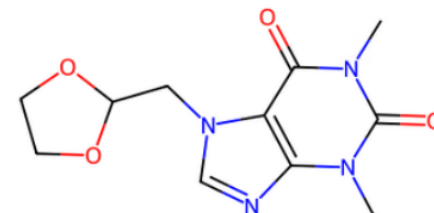
Score: 0.5887675881385803
1,3,7-Trimethyluric acid
79437
Cn1c(=O)c2c([nH]c(=O)n2C)n(C)c1=O
Z61UN6MHB7



Score: 0.6635653972625732
Etofylline
1892
Cn1c(=O)c2c(ncn2CCO)n(C)c1=O
L164909TBI



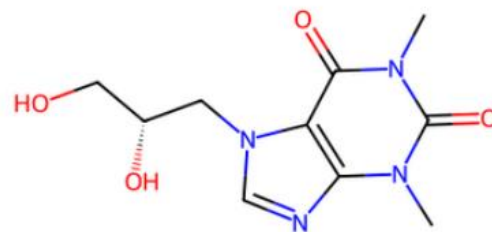
Score: 1.0210137367248535
Doxofylline
50942
Cn1c(=O)c2c(ncn2CC2OCCO2)n(C)c1=O
MPM23GMO7Z
R03DA11



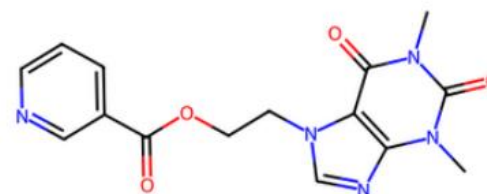
Score: 1.0806951522827148
Dyphylline
3182
Cn1c(=O)c2c(ncn2CC(O)CO)n(C)c1=O
263T0E9RF9
R03DA01



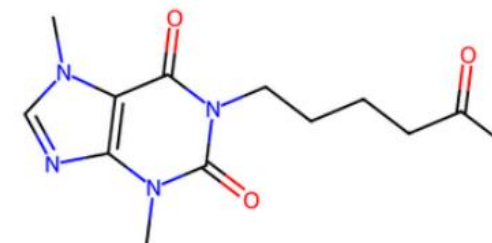
Score: 1.0806951522827148
7-[(2S)-2,3-dihydroxypropyl]-1,3-dimethyl-3,7-dihydro-1H-purine-2,6-dione
688352
Cn1c(=O)c2c(ncn2C[C@H](O)CO)n(C)c1=O
S3D1493BSS



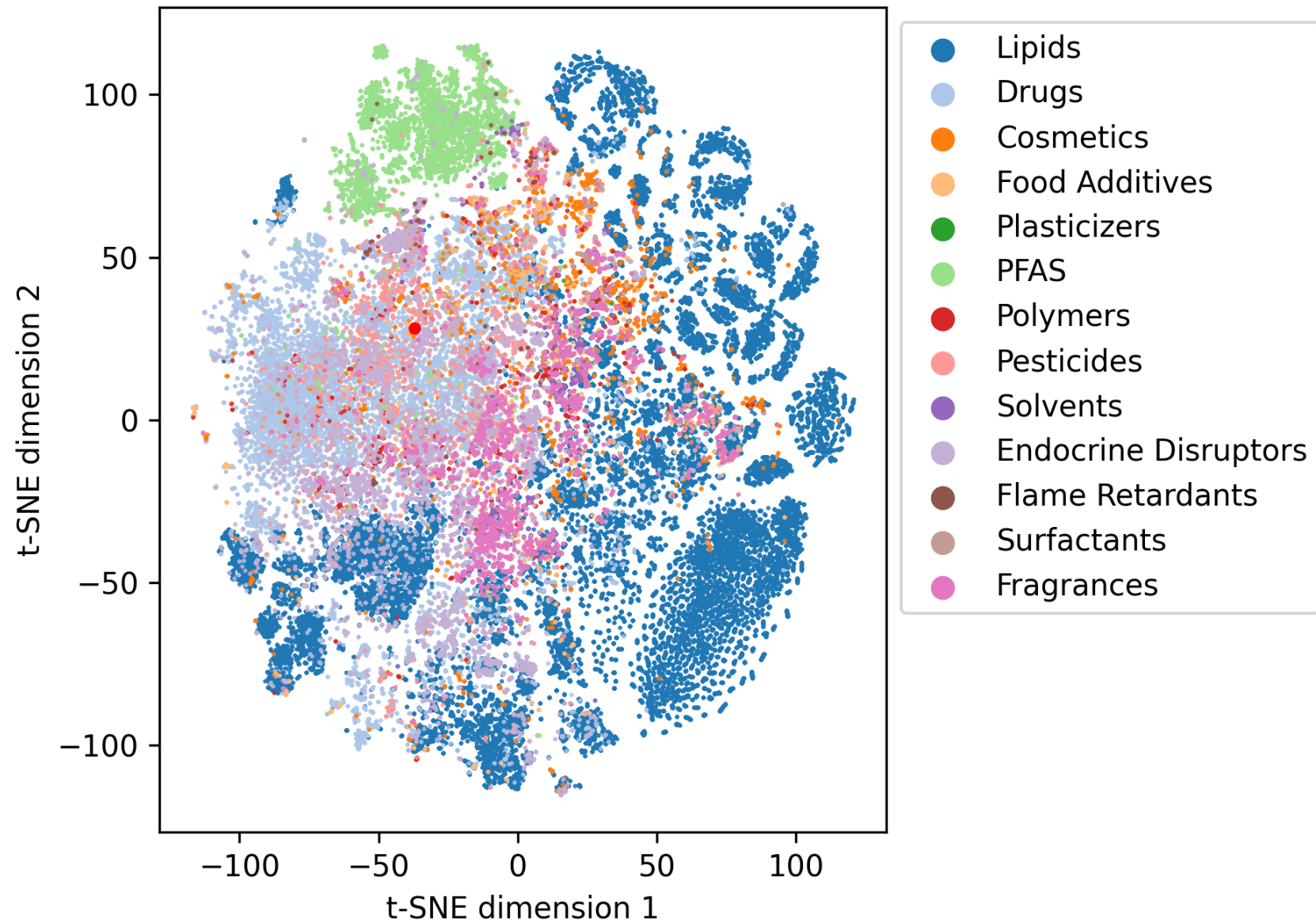
Score: 1.1497620344161987
Etofylline nicotinate
83435
Cn1c(=O)c2c(ncn2CCOC(=O)c3cccnc3)n(C)c1=O
SN6D1V68DL
C04AD04



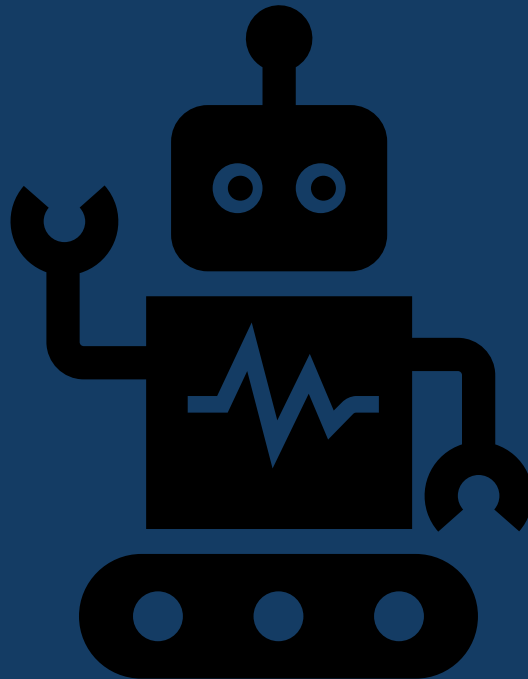
Score: 1.2073583602905273
Pentoxifylline
4740
CC(=O)CCCCn1c(=O)c2c(ncn2C)n(C)c1=O
SD6QC T3TSU
C04AD03



t-SNE visualization of embeddings of chemical classes



Thank you for listening



PICROSS – product information cross search

- Enables both lexicographic and semantic search of product information

PICROSS

Product information cross search

SmPC

PL

Database info

Context of use

Text

Section

All

Search type

☒ Case insensitive

☐ Case sensitive

☐ Semantic

Max results

20

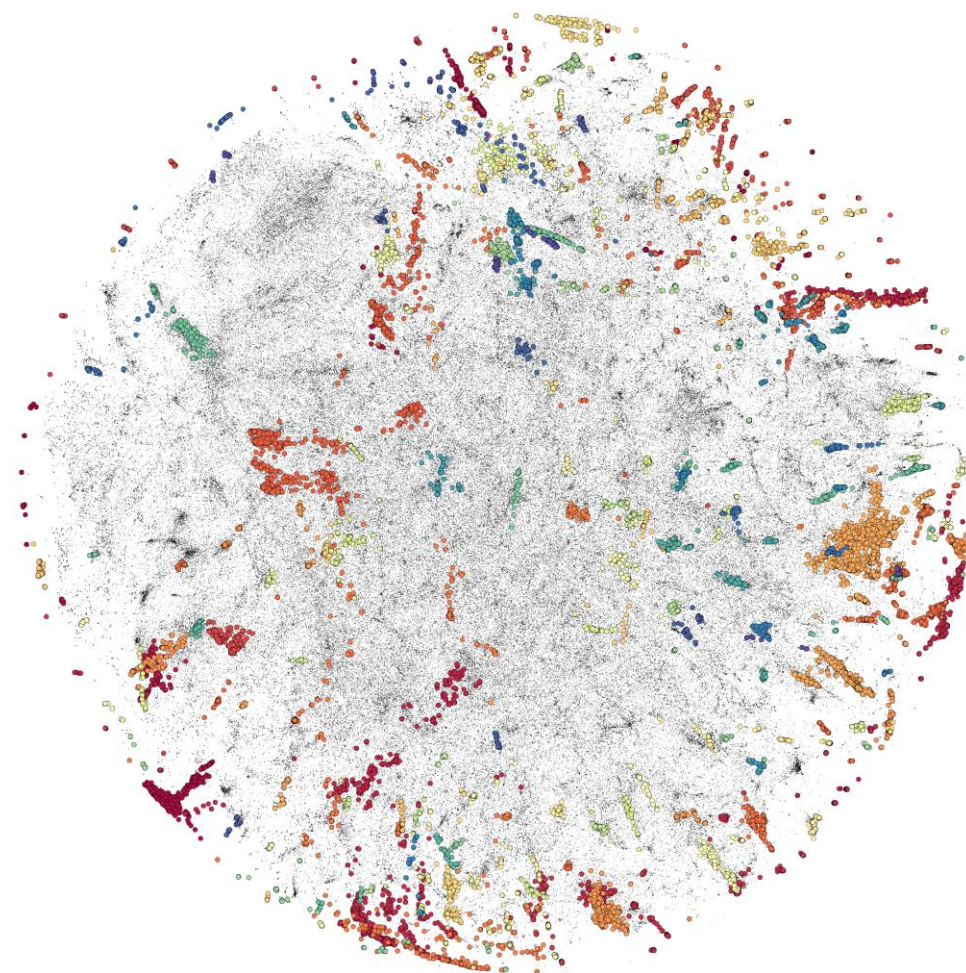
Search info

Search

Download results

Product information embedding space

Cluster id	Spread	Number of sentences	Number of unique sentences
10	0,46	1316	425
(n=281) Do not take a double dose to make up for a forgotten dose. (n=89) Do not take a double dose to make up for a forgotten tablet. (n=34) Do not use a double dose to make up for a forgotten dose. (n=24) If you miss a dose, take it as soon as you remember. (n=18) If you forget to take a dose, take it as soon as you remember. (n=17) Do not inject a double dose to make up for a forgotten dose. (n=16) Then take your next dose at the usual time. (n=15) Do not take a double dose to make up for forgotten individual doses. (n=15) Then take the next dose as usual. (n=15) If you forget for more than 12 hours, simply take the next single dose at the usual time.			



RADAR

AI aided support

Make a KM plot separating treatment groups by color coding and show numbers at risk



Create a Kaplan-Meier plot using the survival and survminer packages

The variables included in the prompt are:

- **AVAL** : This variable contains the time to event or censoring, which is essential for creating a Kaplan-Meier plot.
- **CNSR** : This variable indicates whether a row has been censored (CNSR > 0) or not (CNSR = 0), which is necessary for the survival analysis.
- **TRT01A** : This variable represents the treatment group for each subject, which will be used to separate the treatment groups by color coding in the Kaplan-Meier plot.

The `survival` and `survminer` packages are suggested for use in the prompt because they provide functions for survival analysis, including the creation of Kaplan-Meier plots. The `ggplot2` package, which is included in `survminer`, will be used to create the plot and display the numbers at risk.

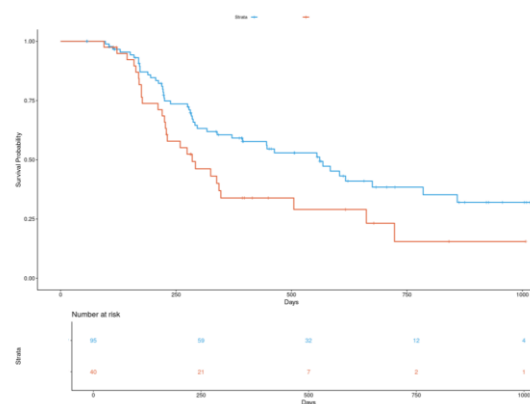
Code:

```
library(survival)
library(survminer)

# Create survival object
surv_obj <- Surv(time = df$AVAL, event = 1 - df$CNSR)

# Fit Kaplan-Meier model
km_fit <- survfit(surv_obj ~ TRT01A, data = df)

# Create Kaplan-Meier plot with risk table
ggsurvplot(km_fit,
            data = df,
```



Summary of R Code

This code performs a **Kaplan-Meier survival analysis** using the `survival` and `survminer` packages in R. It creates a survival object from the `AVAL` (time-to-event) and `CNSR` (censoring indicator) variables, then fits a Kaplan-Meier model stratified by treatment group (`TRT01A`).

The resulting plot displays:

- **Survival probability over time**, separated by treatment groups with distinct colors
- **Numbers at risk** below the plot
- **Censoring marks** on the survival curves
- **Customized legend**, axis labels, and title

Key Steps:

1. **Data Preparation:**
 - `Surv()` creates a survival object where `event = 1 - CNSR` (0 for censored, 1 for event)
2. **Model Fitting:**
 - `survfit()` fits the Kaplan-Meier model by treatment group (`TRT01A`)
3. **Visualization:**
 - `ggsurvplot()` generates the plot with risk table, custom colors, and labels

Output:

A Kaplan-Meier survival curve showing two treatment arms: " " (blue) and " " (orange), along with risk counts at each time point.