

PKPDAI: Unlocking Pharmacokinetic Data from Tables with Natural Language Processing

Victoria C. Smith (1,2), Ferran Gonzalez-Hernandez (3), Palang Chotsiri (4), Thanaporn Wattanakul (4), Jose Antonio Cordero Rigo (5), Maria Rosa Ballester (5,6,7), Mario Duran Hortola (5), Mireia Poch Mascaró (5), Gill Mundin (8), Watjana Lilaonitkul (2), Frank Klopogge (9), Joseph F. Standing (1)

(1)Great Ormond Street Institute of Child Health, UCL, London, UK; (2) Institute of Health Informatics, UCL, London, UK; (3) CoMPLEX, UCL, London, UK; (4) Department of Clinical Pharmacology, Mahidol Oxford Tropical Medicine Research Unit; (5) Facultat de Ciències de la Salut Blanquerna, Universitat Ramon Llull, Barcelona, Spain; (6) Institut d'Investigació Biomèdica Sant Pau, Sant Quintí 77-79, 08041 Barcelona, Spain; (7) Centro de Investigación Biomédica en Red de Salud Mental, CIBERSAM, Madrid, Spain; (8) DMPK, Oncology R&D, AstraZeneca, Cambridge, UK; (9) Institute for Global Health, UCL, London, UK.



1. Introduction

- Pharmacokinetic (PK) **predictions** based on prior PK data, are critical in **early drug development** and **off-label dosing**.
- Existing data sets are **manually curated**, **lack study design** or **covariate** details or cover only a **handful of drugs** (e.g. DrugBank, PubChem, PK-DB).
- Rich PK data are frequently presented in **Tables in Biomedical Literature**.
- However, **20,572 papers** were published under the search term "Pharmacokinetics" in **2021 alone** on PubMed.
- Therefore, **automated approaches** are essential to efficiently extract PK data from literature at scale.

Automated Table Information Extraction comprises:

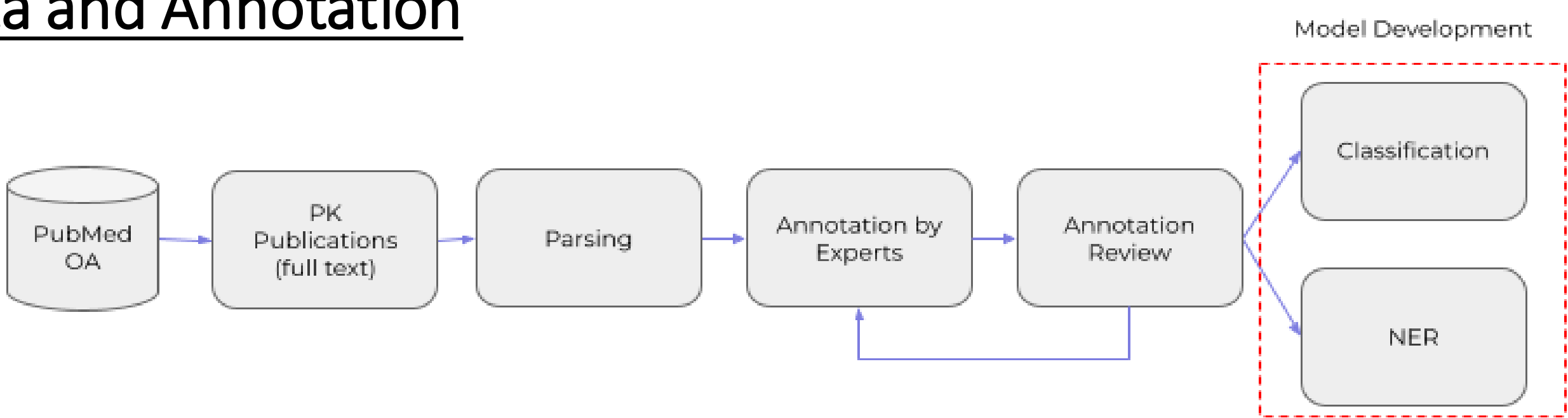
- Table Classification**.
- Named Entity Recognition (NER)** within cells in the table.
- Entity Linking** for terms within table cells
- Relation Extraction** between cells

Study objectives:

- Develop **Labelled Corpora** of PK tables and PK entities within tables cells.
- Develop **Supervised Machine Learning Pipelines** to perform steps 1 and 2.

2. Methods

Data and Annotation



Models

Table Classification

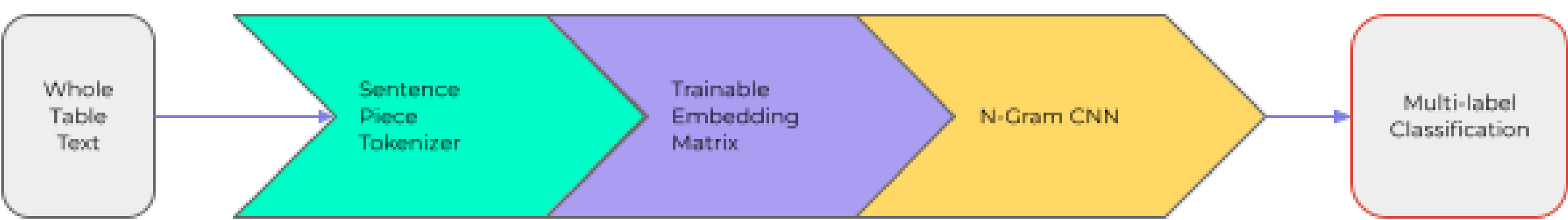
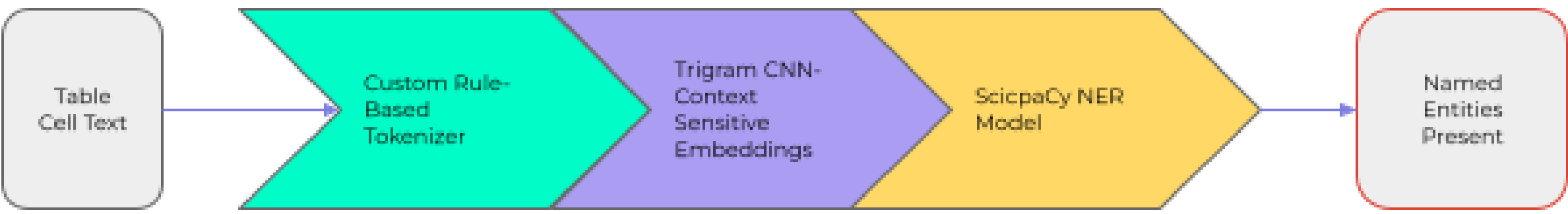
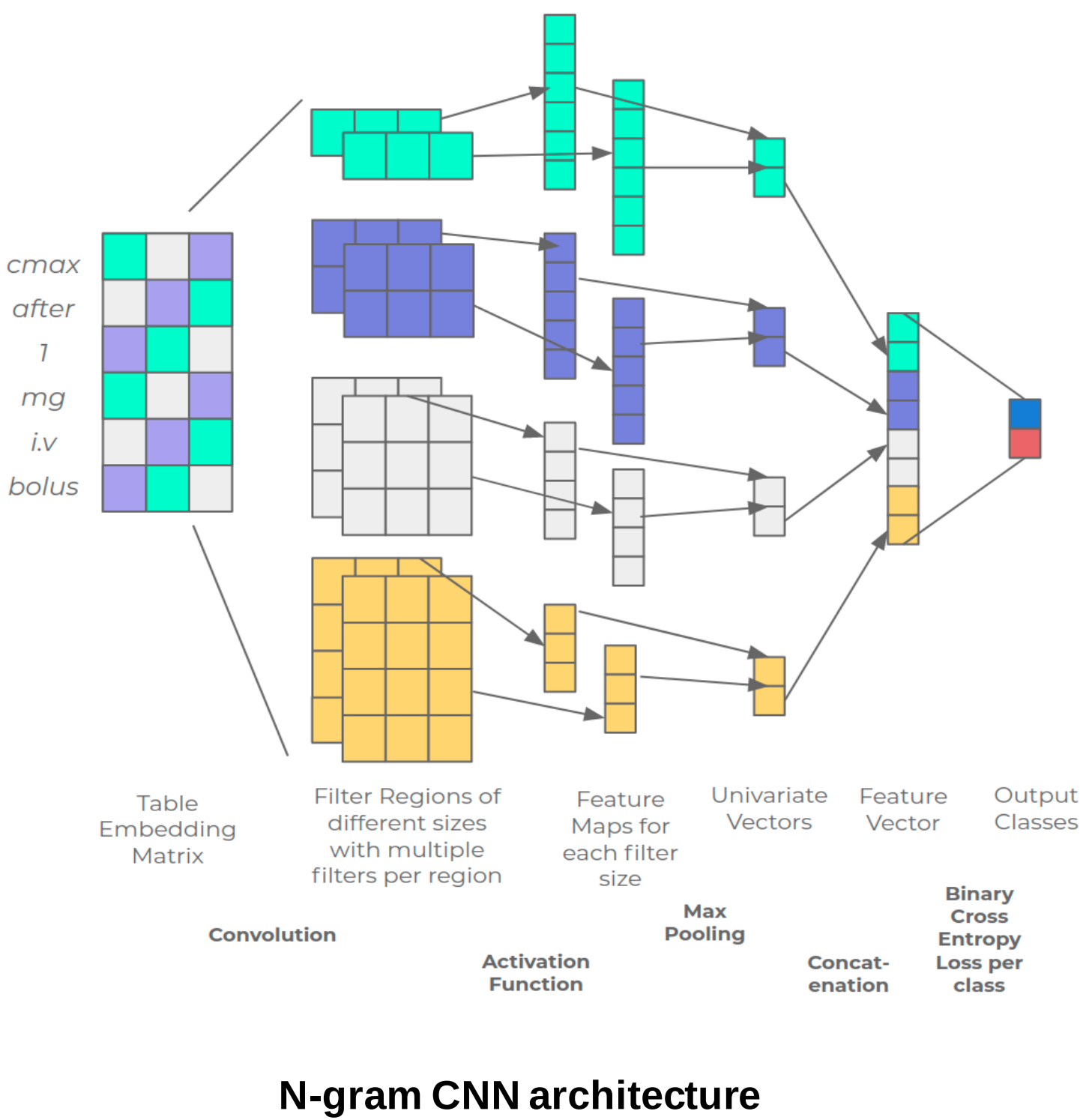


Table NER



- Data Augmentation:** **Synthetic Tables** were automatically generated by (1) adding **random noise** to numeric values (2) replacing PK parameter and covariate mentions with appropriate **synonyms**.
- Active Learning:** Used to prioritise the most **informative** samples to be labelled which will have the **highest impact** on training the supervised model.



3. Results

Annotation

- A total of **2,500 tables** were annotated for table classification and **2000 tables** for NER.
- Label class distributions ranged from **25-44%** and **0.4-17%** on the annotated classification and NER datasets, respectively.

Table Classification

Data **Augmentation**, adding random noise to numeric values, was key to obtaining **optimal performance**.

Class Label	Annotator F1-score	Model F1-score
Non-Compartmental Parameters	0.92	0.94
Doses	0.92	0.89
Number of Subjects	0.91	0.88
Macro-F1	0.92	0.90

Table Cell NER

AUC0-∞ PK ($\mu\text{g}\cdot\text{h/mL}$ Units)

The AUC(0-t) PK and Cmax PK of midazolam Chem/Drug (n Num_Sub = 5 Num_Value)

PK Parameters of dADT Chem/Drug in plasma Sample_Type and blood Sample_Type following treatment with 100 Num_Value mg Units tablets Dose/Route tribendimidine Chem/Drug

intravenous infusion Dose/Route (n Num_Sub = 4 Num_Value)

Entity Label	Annotator F1-score	Model F1-score
Units	0.95	0.98
Numeric Value	0.95	0.97
PK Parameter	0.97	0.96
Number of Subjects	0.95	0.94
Measure Type	0.90	0.94
Chem/Drug	0.94	0.92
Dose/Route	0.82	0.85
Sample Type	0.87	0.83
Macro-F1	0.92	0.92

4. Conclusions

- ✓ This research presents the **first annotated corpora** to train and evaluate PK table classification and PK cell NER models.
- ✓ **Effective prediction pipelines** were developed to categorize PK tables classification and identify relevant PK parameter and covariate entities
- ✓ Uniquely valuable to **pharmacokineticists** as it will allow them to **filter relevant tables**.
- ✓ Provides the **initial 2 steps** towards **automating** PK table **information extraction**.

5. Future Work

- Publish existing work for use by the community on **pkpdai-search.com**.
- Continue to **improve existing models** by providing more **context** (more labelled data, surrounding cell information).
- Develop labelled corpora and language models for **Entity Linking** and **Relation Extraction** for PK Tables.
- Apply the Pipeline** to all PK articles from the PubMed Open Access Set.
- Investigate methods for **parsing PDF tables** to data frames to apply the method more widely.

