WAGENINGEN
UNIVERSITY & RESEARCH

How AI can provide an overview of protein quality from literature

Rutger Vlek, Hannelore Heuer, Lorijn van Rooijen, Addie van der Sluis, Aard de Jong, Jurriaan Mes
Wageningen Food & Biobased Research, Wageningen University & Research, The Netherlands

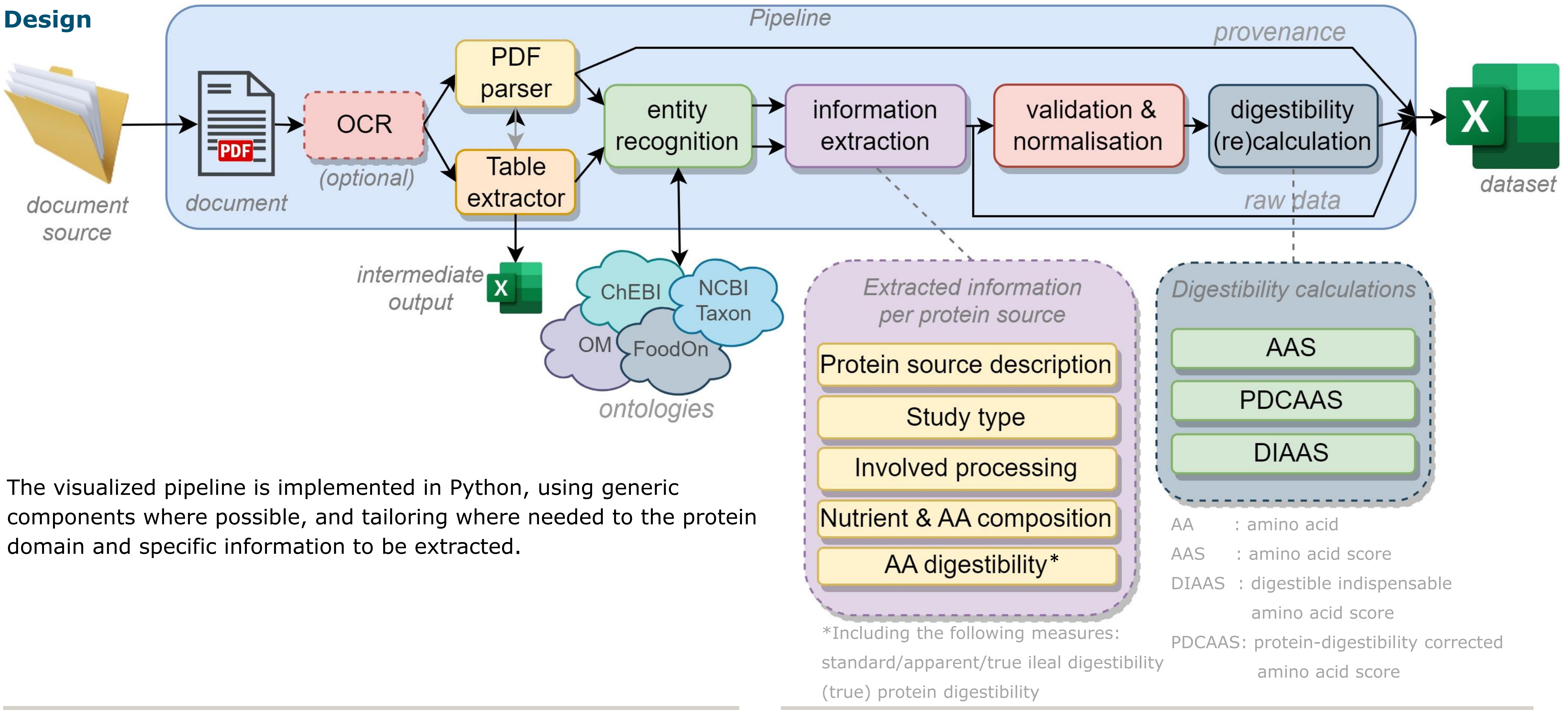
Introduction

A transition to diets with alternative, primarily plant-based, protein sources can benefit both climate and public health, but requires a better insight in their protein quality. Traditional literature studies are effortful and often incomplete, which is why this project explored the potential of artificial intelligence (AI) and natural language processing (NLP) to perform large scale, automated extraction of information on protein digestibility and protein quality from scientific literature.

Input

In order to develop the AI methodology and generate a first dataset, 463 publications on protein quality or digestibility were manually sourced. These include predominantly animal studies, but also human trials and in-vitro studies. Some of these were scanned documents without text representation, requiring optical character recognition (OCR) to make their content accessible.

Design



Pipeline

document source

document

OCR

(optional)

PDF parser

Table extractor

intermediate output

entity recognition

information extraction

validation & normalisation

digestibility (re)calculation

provenance

raw data

dataset

ChEBI

NCBI Taxon

OM

FoodOn

ontologies

Extracted information per protein source

- Protein source description
- Study type
- Involved processing
- Nutrient & AA composition
- AA digestibility*

*Including the following measures:
standard/apparent/true ileal digestibility
(true) protein digestibility

Digestibility calculations

- AAS
- PDCAAS
- DIAAS

AA : amino acid
AAS : amino acid score
DIAAS : digestible indispensable amino acid score
PDCAAS: protein-digestibility corrected amino acid score

The visualized pipeline is implemented in Python, using generic components where possible, and tailoring where needed to the protein domain and specific information to be extracted.

Generated dataset

A dataset was automatically generated by running the pipeline above on the input. This dataset contains:

- 77 unique protein sources
- 261 lines of data on protein quality and/or AA digestibility

➢The pipeline took approx. 30 seconds per document on a laptop.
➢Pipeline can be executed on more documents to expand the dataset.
➢Further validation against manual literature study is in progress.

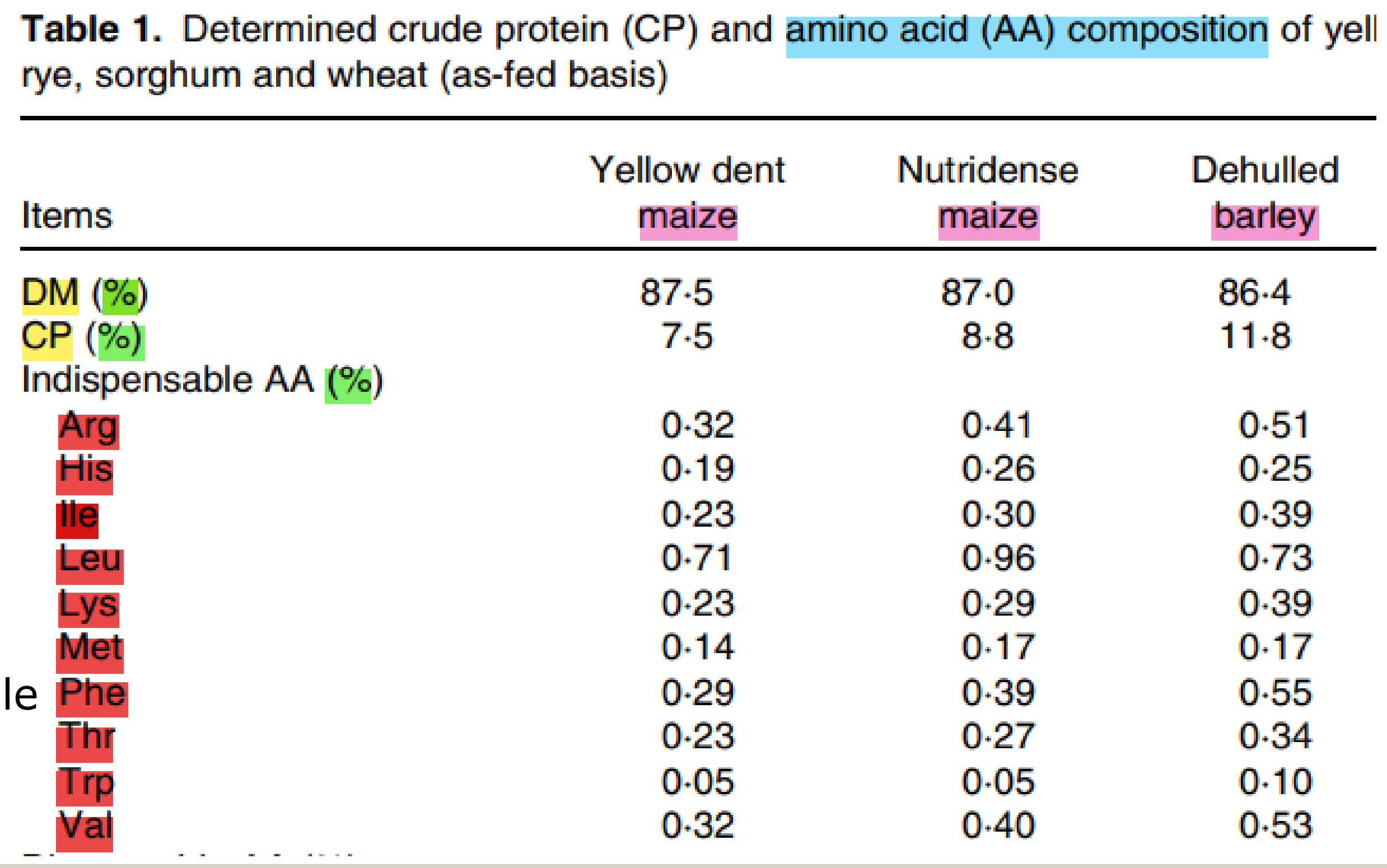


Table 1. Determined crude protein (CP) and amino acid (AA) composition of yell
rye, sorghum and wheat (as-fed basis)

Items	Yellow dent maize	Nutridense maize	Dehulled barley
DM (%)	87·5	87·0	86·4
CP (%)	7·5	8·8	11·8
Indispensable AA (%)			
Arg	0·32	0·41	0·51
His	0·19	0·26	0·25
Ile	0·23	0·30	0·39
Leu	0·71	0·96	0·73
Lys	0·23	0·29	0·39
Met	0·14	0·17	0·17
Phe	0·29	0·39	0·55
Thr	0·23	0·27	0·34
Trp	0·05	0·05	0·10
Val	0·32	0·40	0·53

Figure 1. Example of entity recognition on table caption and headers.

Insights

- Use of domain ontologies replaced the need for laborious training data labelling, while yielding good performance for recognition of relevant entities, and offering information standards as well as rich meta-data.
- Despite near 100% performance on entity recognition, capturing and interpreting the overarching information patterns proved challenging.
- Large variations in how digestibility is measured and quantified hamper unified interpretation.
- Limited data available in literature on AA digestibility, a lot more on composition.
- Methodology highly reusable for other information extraction problems.

Conclusions

- AI holds great potential to reduce human effort in literature surveys.
- Further improvement and validation of the methodology is suggested.
- Dataset provides a single point of access to information on protein quality that was previously scattered.
- Dataset offers insight in variation within/between protein sources.
- Dataset helps identifying knowledge gaps on alternative protein sources.

Wageningen University & Research
Contact: rutger.vlek@wur.nl
www.wur.nl/wfbr

This research was funded by: Unilever, ADM, Oatly & Kellogg's