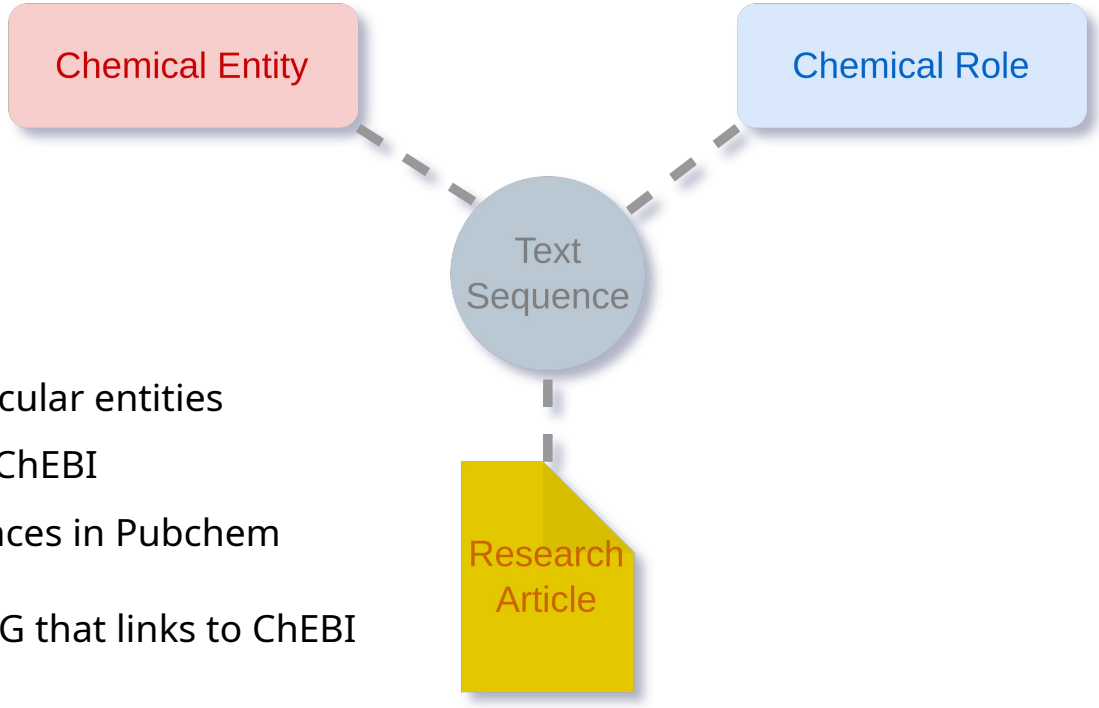
CEAR: Creating a knowledge graph of chemical entities and roles in scientific literature

Stefan **Langer**, Fabian **Neuhaus**, Andreas **Nürnberger**
Otto-von-Guericke University Magdeburg, Germany

Supported by the SmartProSys research initiative
https://www.smartprosys.ovgu.de/

Motivation



„We are drowning in information

but starved for knowledge.“ (John Naisbitt)

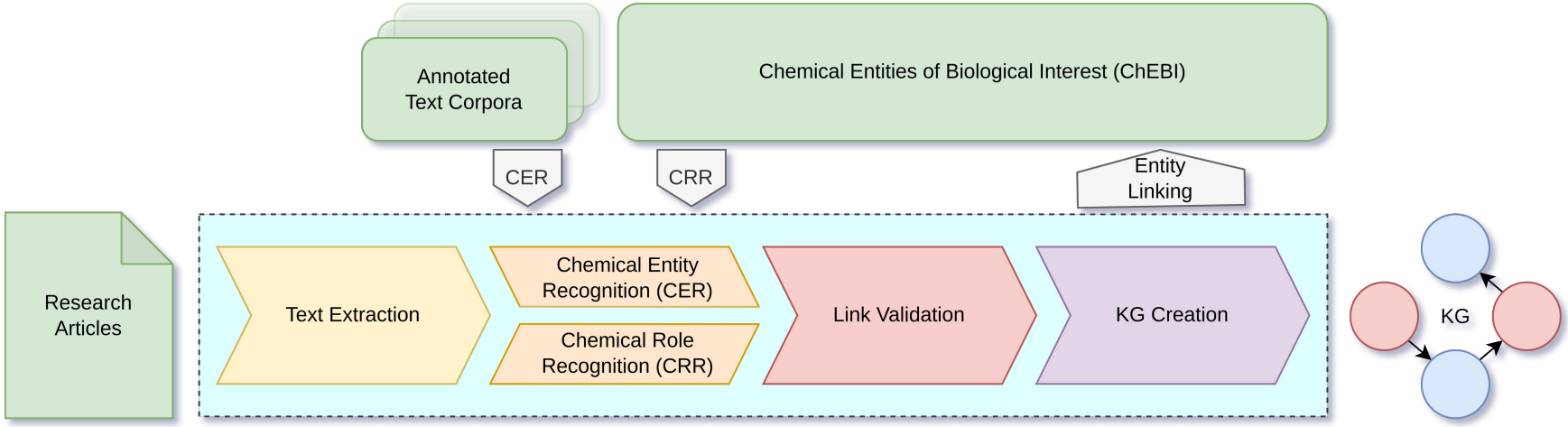
- ChEBI is an ontology of small molecular entities
- Size Challenge: 218,000 entities in ChEBI

  317 million substances in Pubchem

→ **Motivation**: automatically create KG that links to ChEBI

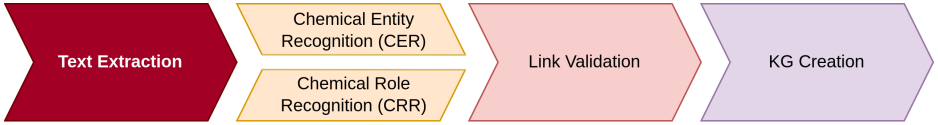
→ **Focus:** on chemical roles

Method



- Named Entity Recognition (NER) for chemicals / roles
- LLM for link validation
- Entity linking and grouping
- KG creation

Text Extraction



- Downloaded 8,000 papers from Chemrxiv
  - Agriculture and Food Chemistry
  - Organic Chemistry
  - Materials Science
  - …
- Extracted full text & page information
- Joined with Chemrxiv metadata


```
"filepath": "/local/sps-local/papers/chemrxiv_pdf/825-Sustainability_and_efficiency_assessment_of_l
"fileHash": "30d854f549b6cbd65c0ecd3ae315b6b11462582d8c6473910baf7c59e60887c4",
"contentHash": "983e40abc85b9399fd4bdece136f2091d059241b914ddae120b5cd420a0df10e",
"textHash": "ece8a03000a743fd43dc06632077b6df2c762a5f2b626796b4b031474d6a4921",
"filetype": "pdf",
"pages": [
    {
        "text": "Sustainability and efficiency assessment\nof lignin-derived phenolic synthons all
        "images": [],
        "pageNumber": 1
    },
    {
        "text": "substitutes involved epichlorohydrin. To circumvent the use of this chlorinated re
        "images": [],
        "pageNumber": 2
    },
    {
        "text": "to optimize the mechanochemical allylation of vanillin. Before studying and optimi
        "images": [
            {
                "fileHash": "4a5ec50bf08c3380d0c6f02a1d6f1a340725f8dd27e42f3e0e9163eb7858c036",
                "fileType": "png"
            },
            {
                "fileHash": "615a60efbf40d13af1f24fea52343acbbdc51831120dbf980bd0507087c77cec",
                "fileType": "png"
```

Datasets for NER

- BC5CDR
  - Chemicals, diseases and interactions from 1,500 PubMed articles
- NLM-Chem corpus
  - Chemicals from 150 full-text articles on biomedical literature
  - Chemicals which are difficult to find for NER tools
- CRAFT corpus
  - 97 full-text articles from PubMed
  - Links entities to different ontologies, including ChEBI

Datasets for NER

- BC5CDR
  - Chemicals, diseases and interactions from 1,500 PubMed articles
- NLM-Chem corpus
  - Chemicals from 150 full-text articles on biomedical literature
  - Chemicals which are difficult to find for NER tools
- CRAFT corpus
  - 97 full-text articles from PubMed
  - Links entities to different ontologies, including ChEBI

Lack of Data

- Challenge
  - Only CRAFT dataset annotates chemical roles
- Solution: Dataset augmentation on BC5CDR and NLM-Chem
  - Extract roles from ChEBI
  - Annotate in dataset using a lexical approach

Named Entity Recogn.

Approach: Fine-tune *google/electra-base-discriminator* model on (augmented) datasets for Named Entity Recognition (NER)

Biotransformations were performed with 0.625 µM P450 enzyme variant, 5 mM ***trans-b-methylstyrene*** (1), 5 mM ***NADH*** ***cofactor*** and 1 vol% ***isopropanol*** in reaction ***buffer***.

The ***heme*** ***cofactor*** is shown as black sticks. 5 / 16

Heme ***cofactor*** and substrate 1 are shown in sticks format, gray and cyan, respectively.

Elongation of the reaction time and application of a ***cofactor*** recycling system enabled conversion of 1 to ***phenylacetone*** with up to 4750 TTN (Fig. S12).

To demonstrate that these reactions can be performed on a preparative scale (1.0 mmol), ***ketone*** 2 was synthesized using a ***catalyst*** loading of 0.025 mol% ***ketone*** synthase (Fig. 6b).

The product was isolated with 61% yield, consuming atmospheric ***oxygen*** and ***D-glucose*** as only stoichiometric ***reagents***. Fig. 6: Application in synthesis.

Reactions were carried out using 0.625 µM KS, 5 mM of the corresponding substrate and 5 mM ***NADH*** ***cofactor***.

With this setup, the unactivated internal ***alkene*** 1 was converted to chiral ***phenylethanols*** and ***phenylethylamine*** that are important structural motifs in top-selling ***pharmaceuticals*** (Fig. S13).

CER/CRR Results

Train Corpus	Type	Eval on BC5CDR			Eval on NLM-Chem			Eval on CRAFT		
		P	R	F1	P	R	F1	P	R	F1
BC5CDR	chem	94.2	90.6	92.4	75.9	54.3	63.3	63.3	30.4	41.1
BC5CDR	role	89.5	90.7	90.1	84.7	83.1	83.9	75.4	59.1	66.3
NLM-Chem	chem	90.3	80.8	85.3	85.8	76.8	81.1	68.0	40.2	50.5
NLM-Chem	role	70.2	82.2	75.7	83.1	89.7	86.3	79.5	76.2	77.8
CRAFT	chem	85.3	67.2	75.2	65.4	44.8	53.2	93.4	85.1	89.0
CRAFT	role	65.4	63.6	64.5	81.4	77.9	79.6	93.6	92.6	93.1
NLM+BC5CDR [15]	*chem*	-	-	-	*81.0*	*71.1*	*75.7*	-	-	-
NLM+BC5CDR	chem	93.4	90.2	91.8	**85.2**	**77.5**	**81.2**	68.1	39.9	50.3
NLM+BC5CDR	role	91.5	92.0	91.7	92.3	93.9	93.1	79.5	76.2	77.8
NLM+CRAFT	chem	90.4	78.3	83.9	84.0	70.9	76.9	88.0	74.1	80.4
NLM+CRAFT	role	79.0	83.4	81.1	88.5	92.1	90.2	87.1	90.3	88.7
all corpora	chem	92.0	89.2	90.6	84.4	71.2	77.3	89.2	74.0	80.9
all corpora	role	89.8	91.6	90.7	90.5	93.7	92.1	87.3	92.2	89.7

- Blue: Evaluation on the same corpus that was used for fine-tuning
- Underlined: very low ratings for cross-dataset validation
- *Italic*: results from *
- **Bold**: our results in comparison to *

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

* R. Islamaj, R. Leaman, S. Kim, D. Kwon, C.-H. Wei, D. C. Comeau, Y. Peng, D. Cissel, C. Coss, C. Fisher, et al., Nlm-chem, a new resource for chemical entity recognition in pubmed full text literature, Scientific data 8 (2021) 91. doi:10.1038/s41597-021-00875-1.

What can we see?

- The bad news:
  - Incompatible datasets:
    - Some entities are annotated in CRAFT, but not in the other datasets, e. g.: „protein“, „DNA“, „RNA“, „mRNA“
    - NLM-Chem corpus and BC5CDR are closer to each other than to CRAFT
  - **→ Unavoidable: Poor out-of-distribution performance!**

- The good news:
  - We can use either dataset for their definition „chemical entity“
  - A LM fine-tuned on all datasets show acceptable result over all datasets

Link validation

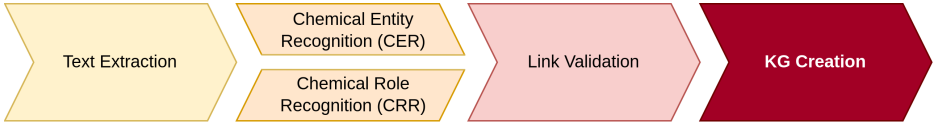
- 8,000 papers collect sentences with at least one chemical entity and one role → 115,537 sentences
- Using: *meta-llama/Llama-2-7b-chat-hf* as LM for all combinations
- Make sure to:
  - Decide based only on the provided text
  - Binary answer „***yes***“ or „***no***“

Prompts

- Prompts:
  - **System-Prompt:** *Do you agree with the question? Please answer using one word.*
  - **User-Prompt:** *In the sentence „<sentence>“: Is **<chemical>** explicitely described as a **<role>**?*
- Examples*:*
  - `answer("In order to cook the noodles, we used water as a solvent for salt.", "water", "solvent")` → „Yes.“
  - `answer("In order to cook the noodles, we used water as a solvent for salt.", "salt", "solvent")` → „No.“

→ 58,511 links were confirmed / 272,053 were rejected

KG Creation



- For all confirmed pairs or chemical entities and roles
  - Try to find in ChEBI
  - Normalize (lowercase, synonyms)
  - Count occurences of the pairs in text
  - Compare to threshold (*minRef*)

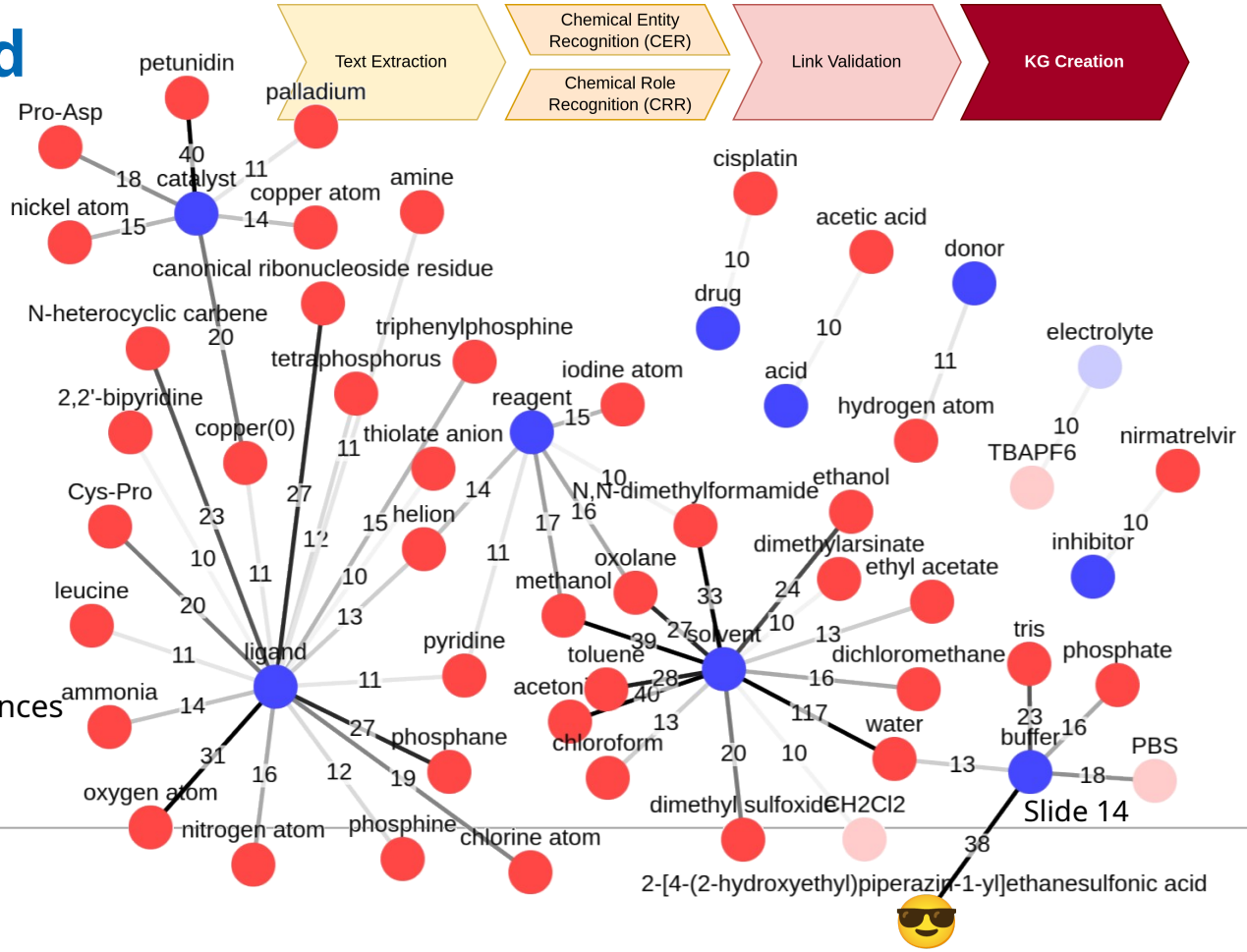
- Store KG using *Terse RDF Triple Language* (Turtle)
  - Triple: `<Chemical> :hasRole <Role>`

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

RDF available with different settings for minRef at:
https://wwwiti.cs.uni-magdeburg.de/iti_dke/cear/

CEAR visualized

- MinRef: 10
- Smaller set of papers
- Styling of Nodes:
  - **Dark red**: Chem in Chebi
  - Light red: Unknown Chem
  - **Dark blue**: Role in Chebi
  - Light blue: Unknown Role
- Styling of Edges:
  - The darker, the more references
  - Numbers show reference



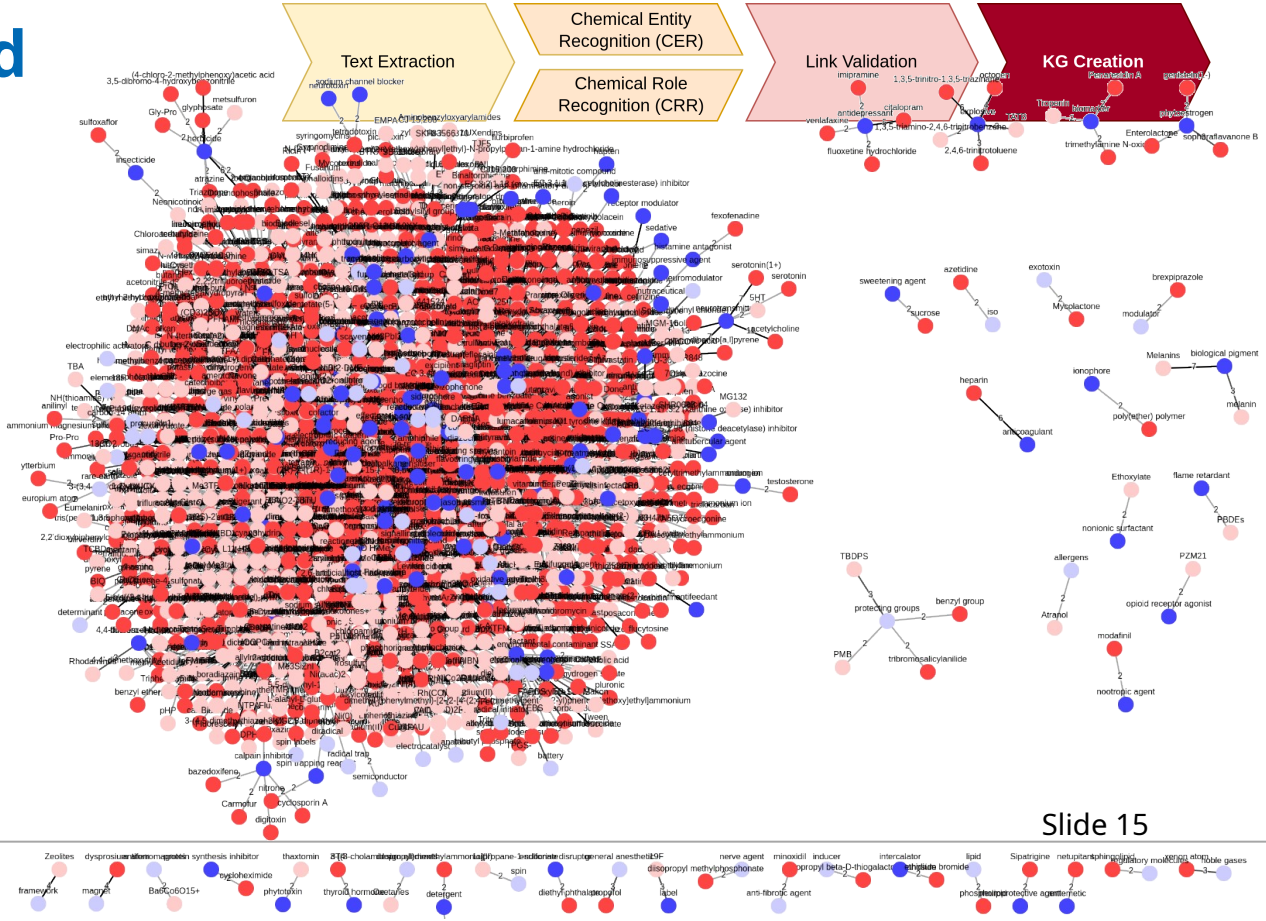
OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

CEAR visualized

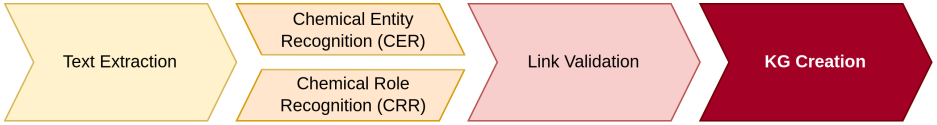
- MinRef:

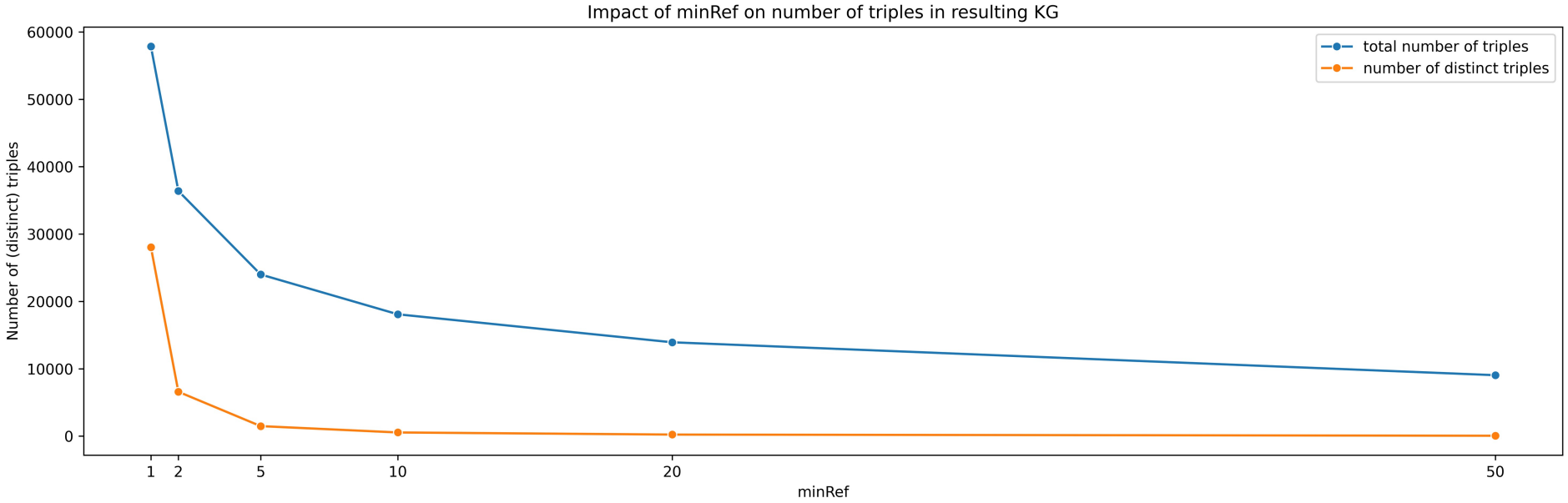
  2 on 8,000 papers

- Higher minRef:
  - Higher confidence
  - Less novelty

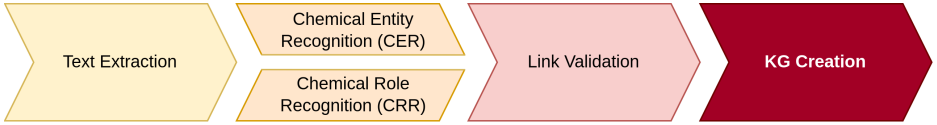


Impact of minRef





Most / Less frequent



Most frequent triples

Least frequent triples

source	chemical entity	source	chemical role	count
ChEBI	water	ChEBI	solvent	1,085
ChEBI	methanol	ChEBI	solvent	551
ChEBI	dimethyl sulfoxide	ChEBI	solvent	438
ChEBI	N,N-dimethylformamide	ChEBI	solvent	402
ChEBI	oxolane	ChEBI	solvent	398
ChEBI	acetonitrile	ChEBI	solvent	388
ChEBI	2-[4-(2-hydroxyethyl)piperazin-1-yl]e...	ChEBI	buffer	375
ChEBI	tris	ChEBI	buffer	271
ChEBI	ethanol	ChEBI	solvent	268
ChEBI	toluene	ChEBI	solvent	268
CEAR	PBS	ChEBI	buffer	249
CEAR	1-propionyl-d-lysergic acid diethylam...	ChEBI	drug	1
CEAR	tetracetate	ChEBI	ligand	1
CEAR	peroxysulfate(2-)	ChEBI	oxidising agent	1
CEAR	2-[4-(2-hydroxyethyl)piperazin-1-yl]et...	**CEAR**	buffers	1
ChEBI	5-fluorouracil	ChEBI	antineoplastic agent	1
CEAR	SiCl4 + 4SO2 + 4MeCl (10) Thionyl chl...	ChEBI	reagent	1
ChEBI	phenylacetonitrile	ChEBI	nucleophilic agent	1
CEAR	α-chloroamide	ChEBI	cofactor	1
CEAR	Cu–t-Bu-BDPP	**ChEBI**	catalyst	1

Future Work

- Use new dataset EnzChemRED which has been published in April 2024
- Enhance:
  - Text extraction
  - Link validation
- Add text references to the KG using RDF-star
- Use more papers for KG creation
- Deploy exploration system for chemistry

Stefan **Langer**, Fabian **Neuhaus**, Andreas **Nürnberger**
Otto-von-Guericke University Magdeburg, Germany

Summary

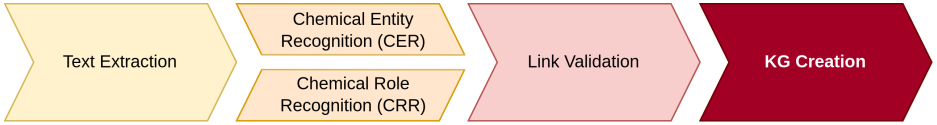
- We presented an algorithm to automatically create a KG from research papers
- Challenge: lack of annotator agreement / different definitions
- Adjustable: selected subcategories of chemistry / roles
- Can be used to enhance ChEBI ontology

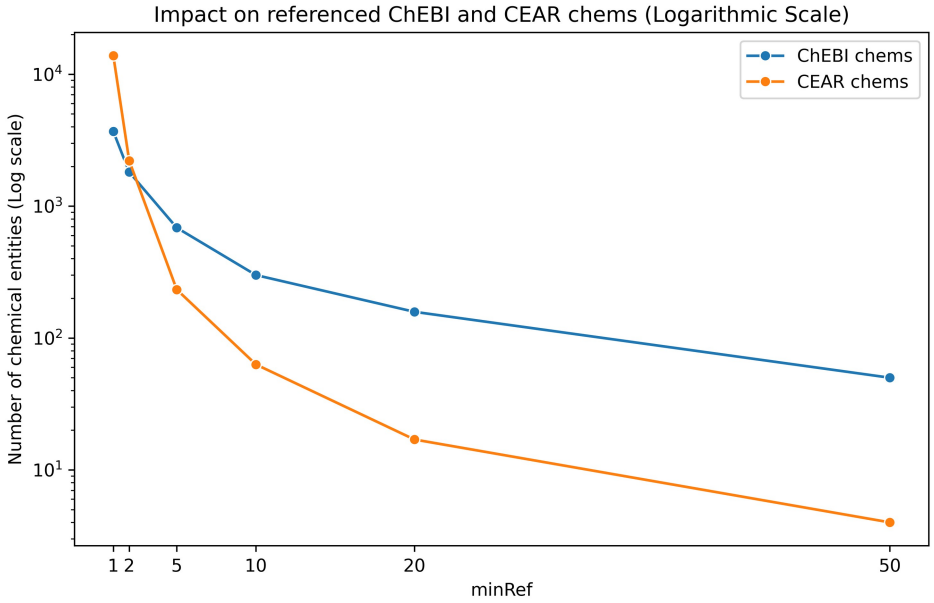
Stefan **Langer**, Fabian **Neuhaus**, Andreas **Nürnberger**
Otto-von-Guericke University Magdeburg, Germany

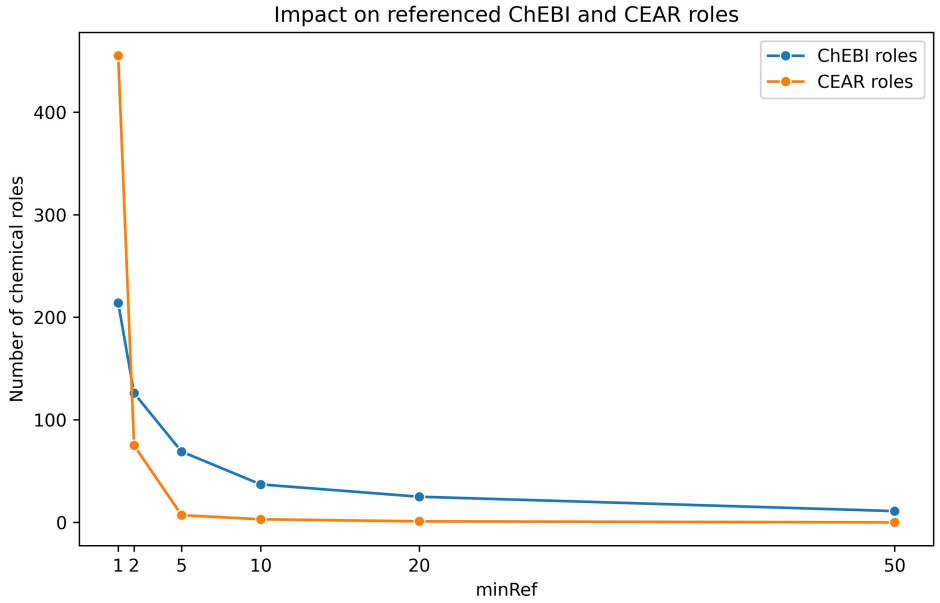
CEAR: Creating a knowledge graph of chemical entities and roles in scientific literature

Stefan **Langer**, Fabian **Neuhaus**, Andreas **Nürnberger**
Otto-von-Guericke University Magdeburg, Germany

Chem/Role frequencies







OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

What can we see?

- Other Ideas to work with incompatible datasets (Slide 10)
  - Use voting system during NER step (AND / OR / ...)
  - Create own dataset
    - Label all entities recognized by the different models
    - Apply manual work by experts
    - Use to fine-tune new model
  - Create different KGs and apply manual work to join them

→ Problem still remains