NLP for Chemistry – Introduction and Recent Advances

Camilo Thorne **Saber Akhondi**
Elsevier, Data Science, Life Sciences
c.thorne.1@elsevier.com s.akhondi@elsevier.com

Abstract

In this half-day tutorial we will be giving an introductory overview to a number of recent applications of natural language processing to a relatively underrepresented application domain: chemistry. Specifically, we will see how neural language models (transformers) can be applied (oftentimes with near-human performance) to chemical text mining, reaction extraction, or more importantly computational chemistry (forward and backward synthesis of chemical compounds). At the same time, a number of gold standards for experimentation have been made available to the research –academic and otherwise– community. Theoretical results will be, whenever possible, supported by system demonstrations in the form of Jupyter notebooks. This tutorial targets an audience interested in bioinformatics and biomedical applications, but pre-supposes no advanced knowledge of either.

Keywords: Chemical text mining, information extraction, transformer models, chemical entity formats

Introduction

Overview Chemistry was for long a *terra incognita* for natural language processing (NLP). While strong overlap with computational and statistical physics (in e.g., so-called computational chemistry) gave rise to the application of many statistical models, methods derived from NLP have only reached wide acceptance in the past twenty years (Sun et al., 2011; Akhondi et al., 2015). The aim of this tutorial is to provide a basic introduction to this emerging field, and overview some of its latest advances. Given its breath, we will focus on four fundamental use cases.

Outline This tutorial will be organized as follows:

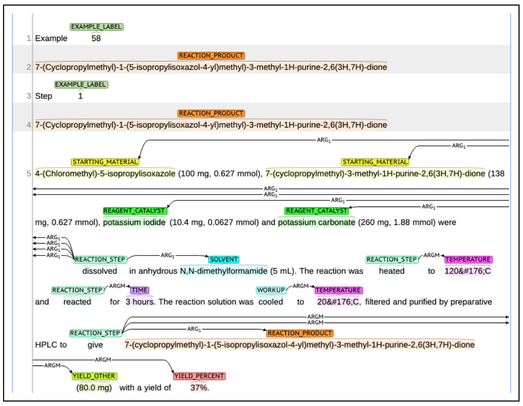
- **Block 1**. Basic chemical notions and techniques.
  50 minutes, followed by a 10 minute break.
- **Block 2**. Text mining in the chemistry domain.
  50 minutes, followed by a 10 minute break.
- **Block 3**. Distributional models for (computational) chemistry.
  50 minutes, followed by a 10 minute break.
- **Block 4**. Large language models, multimodality, applications.
  50 minutes.

For an overview of the material to be discussed in each block, please see below. The tutorial assumes no prior knowledge, with the exception to exposure to Python and natural language processing. Knowledge of chemistry is beneficial but not required.

Basic chemical notions and techniques In chemistry, the primary objects of interest are chemical compounds and reactions. A *compound* is a complex structure composed of *atoms* and *bonds*. Compounds are in turn the building blocks of *reactions*, which are relations or events wherein multiple compounds, a.k.a. *reactants*, are combined to synthesise novel compounds a.k.a *products*.

While a number of manually curated public (e.g., PubChem or SureChemBL) and commercial (e.g. Reaxys© or SciFinder©) chemical databases exist, most of the information about compounds and reactions is reported first in chemical publications, such as chemical patents and chemical journals. Their volume being so big, NLP applications have become critical in the curation and enrichment of these databases (Sun et al., 2011). A number of basic NLP tasks need to be solved for this to be possible (Sun et al., 2011; Leaman et al., 2016). **(a)** Texts need to be segmented and, crucially, tokenized. **(b)** Chemical entities need to be extracted, and normalized or disambiguated against entity identifiers in chemical databases. **(c)** Relations need to be identified. This has motivated research in this area, as well as the emergence of chemical NLP benchmarks to train machine learning models, such as e.g. the CHEMDNER (Krallinger et al., 2015) chemical named entity recognition corpus.

One particular challenge here is the syntax of vocabulary of chemical text, specially, names. While the key representation of a molecule (Sun et al., 2011) is graphical (atoms being the vertexes, and bonds the edges), a number of alternative naming conventions and textual (linear) serialization formats exist (see Figure 2), such as: **(a)** Trivial names –these are standard names for compounds. **(b)** IUPAC names –these are semi-formal names built with special characters. **(c)** SMILES strings –these are linear representations of the graph obtained by topologically ordering a spanning tree of the graph. This traditionally made tokenization a hard task, as traditional methods would break IUPAC names or SMILES (Akkasi et al., 2016). Also,



Example 58

7-(Cyclopropylmethyl)-1-(5-isopropylisoxazol-4-yl)methyl)-3-methyl-1H-purine-2,6(3H,7H)-dione

Step 1

7-(Cyclopropylmethyl)-1-(5-isopropylisoxazol-4-yl)methyl)-3-methyl-1H-purine-2,6(3H,7H)-dione

4-(Chloromethyl)-5-isopropylisoxazole (100 mg, 0.627 mmol), 7-(cyclopropylmethyl)-3-methyl-1H-purine-2,6(3H,7H)-dione (138 mg, 0.627 mmol), potassium iodide (10.4 mg, 0.0627 mmol) and potassium carbonate (260 mg, 1.88 mmol) were dissolved in anhydrous N,N-dimethylformamide (5 mL). The reaction was heated to 120&#176;C and reacted for 3 hours. The reaction solution was cooled to 20&#176;C, filtered and purified by preparative HPLC to give 7-(cyclopropylmethyl)-1-(5-isopropylisoxazol-4-yl)methyl)-3-methyl-1H-purine-2,6(3H,7H)-dione (80.0 mg) with a yield of 37%.

Figure 1: US patent snippet with reaction annotations (entities and events), in BRAT format (He et al., 2021).

even with formal representations, some degree of ambiguity seems unavoidable, stressing the need chemical name normalization at all levels (Akhondi et al., 2015).

Text mining in the chemistry domain An important contribution to this field in recent years has been the ChEMU series of shared chemical test mining tasks, organized within the CLEF 2020, 2021 and 2022 conference. In these shared tasks a novel set of chemical NLP gold sets, each constituted of 1,500 snippets of reaction texts (multi-paragraph passages describing reactions) derived from English chemical patents were made available for the research community, the main being: **(a)** A chemical named entity recognition (NER) set, with entities differentiated by the role they play in reactions (He et al., 2021). **(b)** A event extraction (EE) set, where individual reactions are annotated as events (He et al., 2021). **(c)** An anaphora resolution set, that resolves anaphors across reaction texts (Fang et al., 2021). Figure 1 illustrates the first two levels of annotations on a sample snippet. Results from the shared tasks showed that a wide variety of techniques, including symbolic, heuristic-based text processing, can achieve good results. At the same time, models derived from the BERT family of neural language models can achieve SOTA results on a par or higher than inter-annotator agreement. See Table 1 for the first two benchmarks.

Alongside this, there has also been progress on related tasks such as chemical indexing (Sun et al., 2011; Akhondi et al., 2019; Leaman et al., 2016), where the goal is to identify the most relevant chemical entities for indexing and search.

Distributional models for (computational) chemistry Multiple analogies between chemical compounds and natural or formal languages can be drawn, in particular that, like a sentence, a molecule can be understood as a (recursive) composition of atomic units or "words": base compounds and atoms. Linearized representations of chemical molecules such as SMILES strings make this analogy even more apparent (see Figure 2). SMILES strings can be tokenized (see Figure 2), and embeddings and similar deep-learning molecular representations can thus be successfully learnt via neural language models (Tshitoyan et al., 2019). Such representations can be as expressive (sometimes even more expressive) than traditional cheminformatics representations based on manually engineered chemical and physical features of molecules.

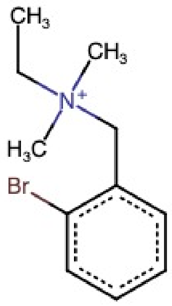
Model	NER (F1)	EE (F1)
NextMove	89.1	89.7
PubMedBERT	94.7	92.0
MelaxTech	95.7	95.3
LG-AI	97.5	92.3

Table 1: Results on the ChEMU NER and EE benchmarks (He et al., 2021; Jang et al., 2022). The latter three are based on BERT resp. encoder-only transformer models. NextMove's methods are on the other hard, based on more classical methods such as dependency parsing, grammars and transducers.

Model	Acc1	Acc2	Acc3
Dual-TF	55.3	66.7	73.0
Graph2SMILES	52.9	66.5	70.0
Chemformer	53.6	61.1	61.7
T5Chem	46.5	64.4	70.5

Table 2: SOTA (mid-2023) on USPTO-50k (Irwin et al., 2022; Sun et al., 2021; Tetko et al., 2020; Lu and Zhang, 2022a). Notice that two out of four models are text-to-text transformers (encoder-decoders).

In particular, chemical transformations such as single-step retro-synthesis –predicting the reactant(s)– or its dual, forward synthesis –predicting the product(s)– can be modelled as sequence-to-sequence problems, viz., translations between the SMILES strings to the left and right of the chemical equation symbol » (see Figure 3). It can thus be solved using text-to-text transformer models from the Bart or T5 families (**?**Irwin et al., 2022; Lu and Zhang, 2022a). This is evident in Table 2, that shows the current SOTA on the main single-step chemical synthesis benchmark, the USPTO-50k gold set. This is a manually curated set of 50,000 reactions extracted from US chemistry patents. All models are deep learning models, with the first two based on the analysis of the source graphical, 2-dimensional representations of molecules, and the latter two, on neural language models and reaction SMILES.



```
SMILES          CC[N+](C)(C)Cc1ccccc1Br
tokens        CC[N+](  C[N+](C   [N+](C)
                  (C)(   C)(C   )(C)   (C)C
                 C)Cc   )Cc1   Cc1c   c1cc
              1ccc   cccc   cccc   ccc1   cc1Br
```

Figure 2: SMILES representation and tokenization of “Bretylium" (a.k.a. “N-(2-Brombenzyl)-N,N-dimethylethanaminium" in IUPAC notation) into 16 4-chargramms.

```
C.O=O.O=O » O=C=O.O.O
```

Figure 3: The combustion of methane represented in (reaction) SMILES. Dots, viz, the character ., are used as separators for the reactants, listed to the left hand side of the reaction symbol », and the products, listed to the right.

In fact, neural word embeddings, learnt from chemical corpora, seem nowadays powerful enough (Thorne and Akhondi, 2020) to learn representations of entities that correlate with molecular-based representations (see Figure 4).

Large language models, multimodality, applications The current surge in large language models (LLMs), viz., decoder-only generative transformer models with billions of parameters and trained over corpora comprising billions of words, has also reached the chemical domain. Researchers have demonstrated (Bran et al., 2023; **?**) that general-purpose models like Open-AI’s GPT-3 and GPT-4, or scholarly LLMs such as Galactica (Taylor et al., 2022) can be used as chemistry and computational chemistry assistants, even if chemistry-specific models (such as e.g. SMILES-GPT (Adilov, 2021)) still underperform. Figure 5 shows that they can be used to suggest, e.g., reactions and (even if not necessarily always factually correct) synthesis procedures, potentially helping drafting novel plans.

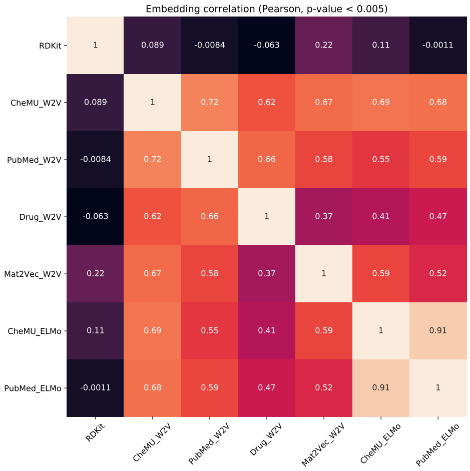
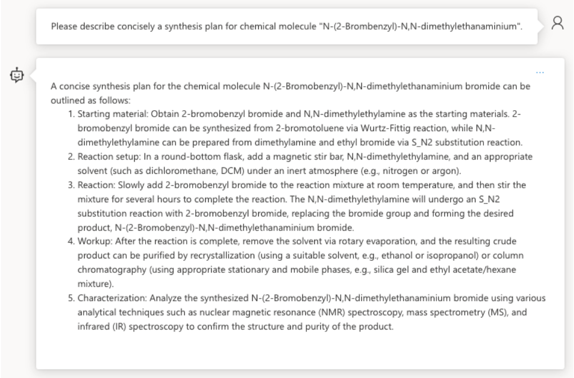


Figure 4: Correlations between chemical (word) embeddings (Thorne and Akhondi, 2020).



Please describe concisely a synthesis plan for chemical molecule "N-(2-Brombenzyl)-N,N-dimethylethanaminium".

A concise synthesis plan for the chemical molecule N-(2-Bromobenzyl)-N,N-dimethylethanaminium bromide can be outlined as follows:

1. Starting material: Obtain 2-bromobenzyl bromide and N,N-dimethylethylamine as the starting materials. 2-bromobenzyl bromide can be synthesized from 2-bromotoluene via Wurtz-Fittig reaction, while N,N-dimethylethylamine can be prepared from dimethylamine and ethyl bromide via S_N2 substitution reaction.
2. Reaction setup: In a round-bottom flask, add a magnetic stir bar, N,N-dimethylethylamine, and an appropriate solvent (such as dichloromethane, DCM) under an inert atmosphere (e.g., nitrogen or argon).
3. Reaction: Slowly add 2-bromobenzyl bromide to the reaction mixture at room temperature, and then stir the mixture for several hours to complete the reaction. The N,N-dimethylethylamine will undergo an S_N2 substitution reaction with 2-bromobenzyl bromide, replacing the bromide group and forming the desired product, N-(2-Bromobenzyl)-N,N-dimethylethanaminium bromide.
4. Workup: After the reaction is complete, remove the solvent via rotary evaporation, and the resulting crude product can be purified by recrystallization (using a suitable solvent, e.g., ethanol or isopropanol) or column chromatography (using appropriate stationary and mobile phases, e.g., silica gel and ethyl acetate/hexane mixture).
5. Characterization: Analyze the synthesized N-(2-Bromobenzyl)-N,N-dimethylethanaminium bromide using various analytical techniques such as nuclear magnetic resonance (NMR) spectroscopy, mass spectrometry (MS), and infrared (IR) spectroscopy to confirm the structure and purity of the product.

Figure 5: Asking GPT-4 (8,192-token input context version) to concisely describe a synthesis plan (sequence of reactions and reaction steps) for “N-(2-Brombenzyl)-N,N-dimethylethanaminium". We sampled with temperature $t \geq 0.7$, likelihood $p \geq 0.95$ and a 800-token stop citerion.

Another emerging field of chemical NLP research is work on multi-modality. As seen earlier, it is possible to learn neural language models on chemical texts and linearized representations of compounds and reactions, and apply them to text mining and computational chemistry tasks. However, not all chemical information is conveyed textually. A significant part is conveyed in images, structured in tables, etc. Hence the need to learn wider, more expressive representation spaces that e.g. enrich current spaces with physiochemical features and other dimensions (Soares et al., 2023; Lu and Zhang, 2022b).

Reading List and Tools

In this section we highlight the key literature pointers the audience should be aware of for a better understanding of this tutorial. We also point at some basic software tools. Readers are invited to click on the hyper-links.

Key papers While all papers cited earlier are useful, we suggest to start with (Sun et al., 2011), which covers well the problems in chemical text mining, as well as approaches that precede deep learning. Is also important to understand chemical representation formats. Regarding text mining, we suggest (He et al., 2021) and (Lu and Zhang, 2022a) for distributional models. Lastly, (Bran et al., 2023) for recent applications (large language models).

Key software tools The main open source software tool used in the cheminformatics community is perhaps RDkit, a Python library that we will be using in our demos and Jupyter notebooks. For a more extensive overview of all software tools (including tools written in languages other than Python), please check this GitHub repository. It also contains links to predictive models beyond NLP. These tools are sometimes essential for (pre)processing chemical data.

Key models Regarding word embeddings, we suggest to check out the ChELMo embeddings, pre-trained on chemical patents (even if not transformer-based) Regarding text mining models, many are closed-source. We will provide some Elsevier deep learning -based demonstration models as part of this tutorial. An open source –if dated and written in Java– starting point is ChemSpot (based on conditional random fields and manual features,). Regarding distributional models over SMILES, we recommend T5Chem.

Key chemical NLP benchmarks While the papers cited mention multiple benchmarks, we suggest to focus on the following four: **(a)** The chemical NER BioSemantics corpus. **(b)** The chemical NER CHEMDNER corpus. **(c)** The ChEMU benchmarks. **(d)** Lastly, the USPTO-50k collection of chemical reactions, the most important public benchmark for computational chemistry.

Presenters

Camilo Thorne (personal website; Google Scholar) is currently Principal Data Scientist at Elsevier. His work focuses on applying current NLP SOTA (large language models and other transformer-based NLP techniques) to the life sciences domain, and in particular to chemistry. His background spans both industry and academia. Prior to Elsevier he worked as postdoctoral fellow in biomedical NLP at the universities of Mannheim and Stuttgart, Germany, and as computational linguist at IBM, Italy. He holds a PhD in computer science from the Free University of Bozen-Bolzano, where he studied controlled natural languages and semantic web formalisms. Last, but not least, he holds extensive teaching and public speaking experience in his fields of interest.
Saber Akhondi (Google Scholar) is currently Senior Director/Head of Data Science at Elsevier He heads a group of 10+ data scientists, where he applies NLP and machine learning techniques to extract information useful for large commercial and research communities in the life sciences. He has extensive experience in the area of chemical text mining, with multiple high impact publications, and multiple international project coordination activities (ChEMU, BioSemantics). Saber Akhondi holds a PhD from Erasmus University Rotterdam, where he developed novel methods for the detection, normalization and indexing of chemical entities.

Diversity Considerations

This topic contributes to topic diversity by introducing an underrepresented application domain of natural language processing (and machine learning): computational chemistry. It will be of particular interest to researchers in the biomedical and bioinformatics domain, and more generally, to researchers of cross-disciplinary life sciences and data science backgrounds.

Other Information

Presenters This tutorial will be given by two persons, who will alternate each other for the different blocks.

Course infrastructure The presenters will try to illustrate practically the methods described with Jupyter notebooks whenever possible. Slides, notebooks and announcements will be distributed and managed through a public GitHub repository (or a public website) and Google Colab, accessible to all participants. For the tutorial, we request only a room sufficiently large for all registered attendants, with good internet connection and a projector.

Ethics Statement

Methods will be demonstrated using datasets and platforms that are freely accessible for research purposes.

Bibliographical References

Sanjar Adilov. 2021. Generative pre-training from molecules. *ChemRxiv*.

Saber A. Akhondi, Sorel Muresan, Antony J. Williams, and Jan A. Kors. 2015. Ambiguity of non-systematic chemical identifiers within and between small-molecule databases. *J. Cheminformatics*, 7:54:1–54:10.

Saber A. Akhondi, Hinnerk Rey, Markus Schwörer, Michael Maier, John P. Toomey, Heike Nau, Gabriele Ilchmann, Mark Sheehan, Matthias Irmer, Claudia Bobach, Marius A. Doornenbal, Michelle Gregory, and Jan A. Kors. 2019.

Automatic identification of relevant chemical compounds from patents. *Database J. Biol. Databases Curation*, page baz001.

Abbas Akkasi, Ekrem Varoglu, and Nazife Dimililer. 2016. Chemtok: A new rule based tokenizer for chemical named entity recognition. *BioMed Research International*.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *CoRR*.

Biaoyan Fang, Christian Druckenbrodt, Saber A. Akhondi, Jiayuan He, Timothy Baldwin, and Karin Verspoor. 2021. Chemu-ref: A corpus for modeling anaphora resolution in the chemical domain. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1362–1375. Association for Computational Linguistics.

Jiayuan He, Dat Quoc Nguyen, Saber A. Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Zubair Afzal, Zenan Zhai, Biaoyan Fang, Hiyori Yoshikawa, Ameer Albahem, Lawrence Cavedon, Trevor Cohn, Timothy Baldwin, and Karin Verspoor. 2021. Chemu 2020: Natural language processing methods are effective for information extraction from chemical patents. *Frontiers Res. Metrics Anal.*, 6:654438.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. 2022. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022.

Youngrok Jang, Hosung Song, Junho Lee, Gyeonghun Kim, Yireun Kim, Stanley Jungkyu Choi, Honglak Lee, and Kyunghoon Bae. 2022. Context aware named entity recognition and relation extraction with domain-specific language model. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 782–796. CEUR-WS.org.

Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. 2015. CHEMDNER: the drugs and chemical names extraction challenge. *J. Cheminformatics*, 7(S-1):S1.

Robert Leaman, Chih-Hsuan Wei, Cherry Zou, and Zhiyong Lu. 2016. Mining chemical patents with an ensemble of open systems. *Database J. Biol. Databases Curation*.

Jieyu Lu and Yingkai Zhang. 2022a. Unified deep learning model for multitask reaction predictions with explanation. *Journal of Chemical Information and Modeling*.

Jieyu Lu and Yingkai Zhang. 2022b. Unified deep learning model for multitask reaction predictions with explanation. *J. Chem. Inf. Model.*, 62(6):1376–1387.

Eduardo Soares, Emilio Vital Brazil, Karen Fiorela Aquino Gutierrez, Renato Cerqueira, Dan Sanders, Kristin Schmidt, and Dmitry Zubarev. 2023. Beyond chemical language: A multimodal approach to enhance molecular property prediction. *CoRR*.

Bingjun Sun, Prasenjit Mitra, C. Lee Giles, and Karl T. Mueller. 2011. Identifying, indexing, and ranking chemical formulae and chemical names in digital documents. *ACM Trans. Inf. Syst.*, 29(2).

Ruoxi Sun, Hanjun Dai, Li Li, Steven Kearnes, and Bo Dai. 2021. Towards understanding retrosynthesis by energy-based models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 10186–10194.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *CoRR*.

Igor V. Tetko, Pavel Karpov, Ruud van Deursen, and Guillaume Godin. 2020. Augmented transformer achieves 97% and 85% for top5 prediction of direct and classical retro-synthesis. *CoRR*.

Camilo Thorne and Saber A. Akhondi. 2020. Word embeddings for chemical patent natural language processing. *CoRR*.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nat.*, 571(7763):95–98.