Brookhaven™ National Laboratory

U.S. DEPARTMENT OF ENERGY

Informing Chemical Separations Research with ML and NLP

Carlos Soto

Computational Science Initiative, BNL

January 23, 2023

LDRD

“Utilizing artificial intelligence/machine learning and automation systems to inform and optimize chemical separations”

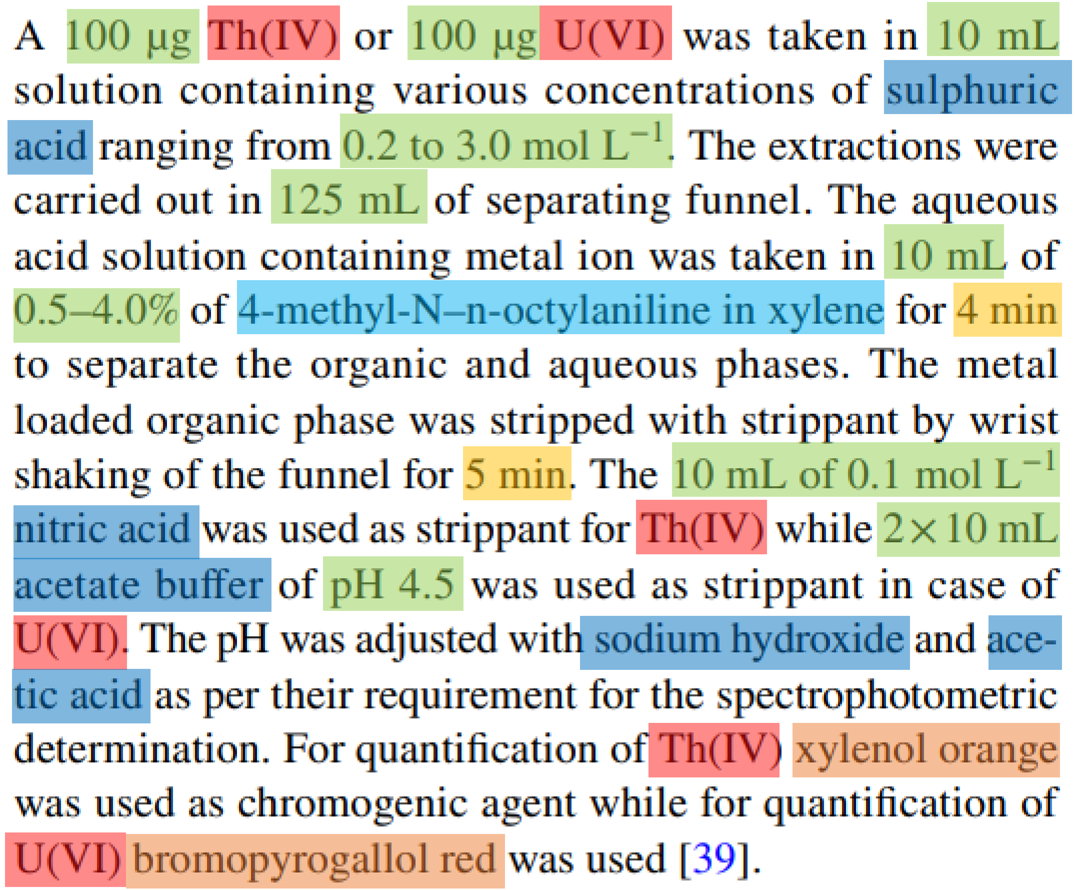
PI: Jasmine Hatcher-Lamarre (C-AD/MIRP)

In a nutshell:

- Designing new chemical separations takes months to years; high-throughput experimentation currently out of reach
- Literature contains useful properties and coordination trends of transition metals, lanthanides; so
  1. Extract these properties from literature with NLP, and use ML to generate new potential separation paths
  2. Use automation to efficiently implement and test those separations at scale

Extracting knowledge from literature

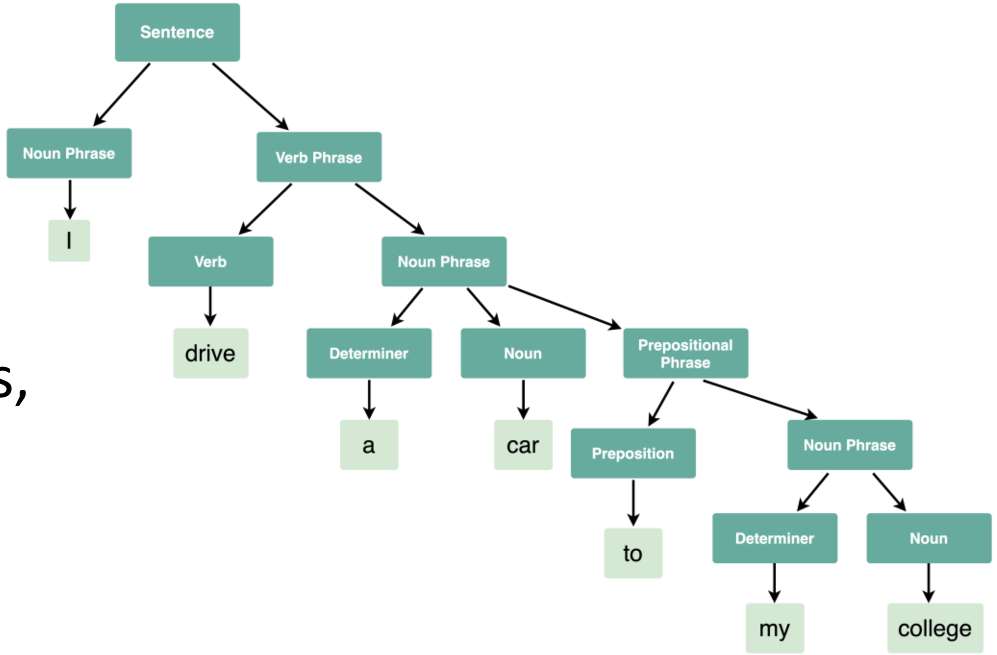
- Separations details in papers
  - Metals, extractants, contact/mixing time, flow rates, temperature, pH, K_D constant, choice of acid, functional groups, equilibrium constants, yield, etc.
- Scaling up with NLP & ML
  - Extract & organize relevant information in paper (canonicalization)
  - Compile database of separations across many papers
  - Look for emergent patterns; synthesize new extractions
- Data relevance and variety
  - Most papers won’t list all details, may cite others
  - Follow references, and broaden scope of chemical separations papers (across transition metals, lanthanides, actinides) → Trends better emerge at scale
- Building ML methods with experts in the loop
  - Include paper links and text snippets (as context) along with NLP extractions
  - Validate extractions, canonicalizations with human feedbacks
  - Iteratively improve NLP and ML models



A 100 µg Th(IV) or 100 µg U(VI) was taken in 10 mL solution containing various concentrations of sulphuric acid ranging from 0.2 to 3.0 mol L^{-1}. The extractions were carried out in 125 mL of separating funnel. The aqueous acid solution containing metal ion was taken in 10 mL of 0.5–4.0% of 4-methyl-N–n-octylaniline in xylene for 4 min to separate the organic and aqueous phases. The metal loaded organic phase was stripped with strippant by wrist shaking of the funnel for 5 min. The 10 mL of 0.1 mol L^{-1} nitric acid was used as strippant for Th(IV) while 2×10 mL acetate buffer of pH 4.5 was used as strippant in case of U(VI). The pH was adjusted with sodium hydroxide and acetic acid as per their requirement for the spectrophotometric determination. For quantification of Th(IV) xylenol orange was used as chromogenic agent while for quantification of U(VI) bromopyrogallol red was used [39].

P.S. More, *et al.* Extraction of Th(IV) and U(VI) with 4-methyl-N–n-octylaniline as an extracting agent. *J Radioanal Nucl Chem* **331**, 4149–4158 (2022).

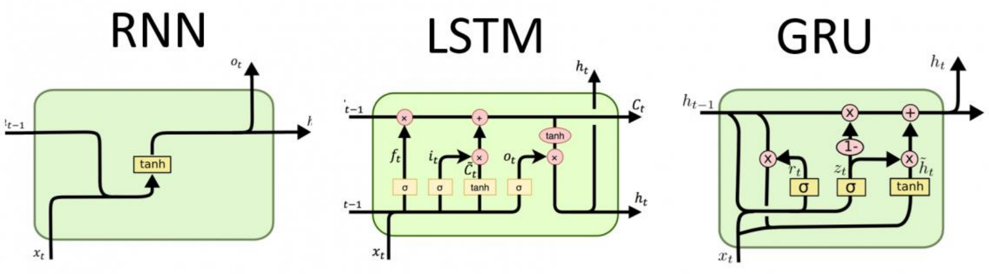
Natural Language Processing



- Classical NLP (based on rules of linguistics)
  - Basics: parsing parts-of-speech, dependencies, syntax trees, dictionaries, grammars, etc.
  - Build up with logic rules: semantics, pragmatics, natural language understanding/synthesis
  - In practice: very brittle, poor quality
- Machine Learning for NLP
  - Learn language features in data-driven manner, at vast scale
  - Embeddings, language modeling, inference and generation tasks
  - New focus: Data & Models

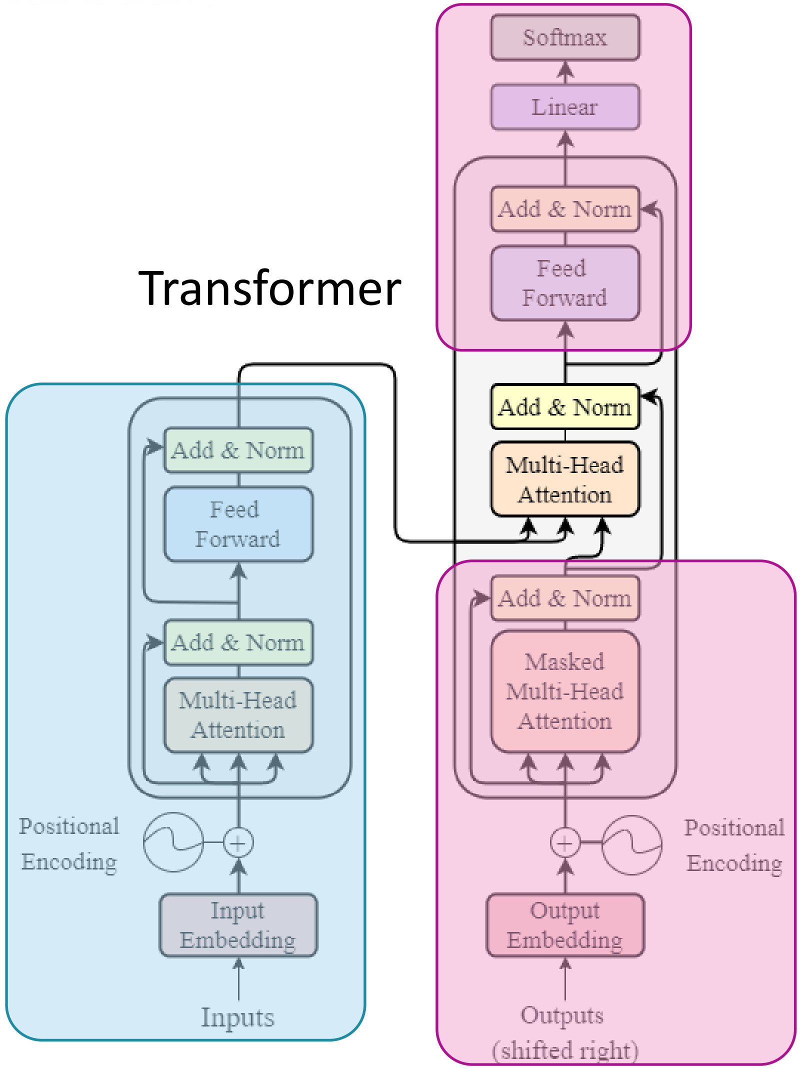
Brookhaven National Laboratory

ML Models for NLP



- Pre-2017: Recurrent NNs (LSTMs, GRUs)
- Post-2017: Transformers[1]
  - capture long-term dependencies, scale up very well, can be pre-trained
- Encoder-type models: e.g. BERT[2]
  - Useful for building "head" onto pre-trained "body"
- Decoder-type models: e.g. GPT[3]
  - Useful for generating novel text, querying

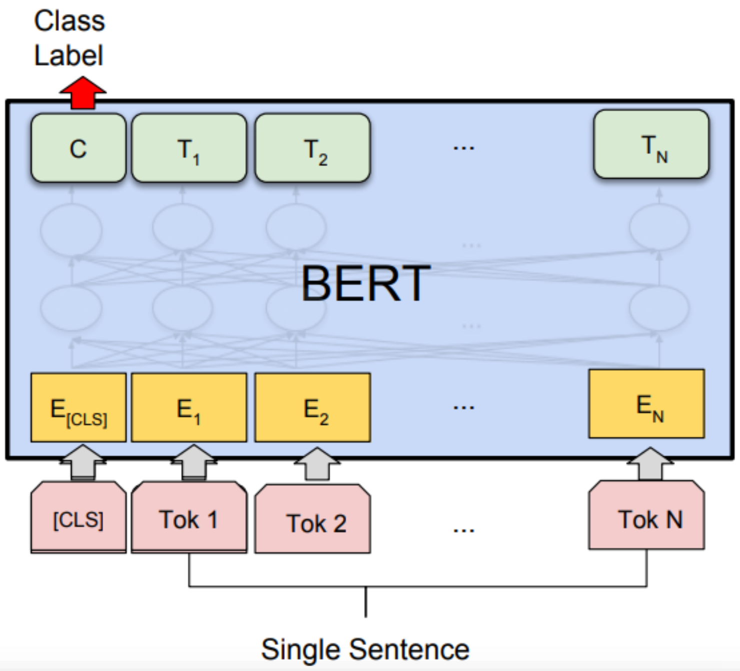
Transformer

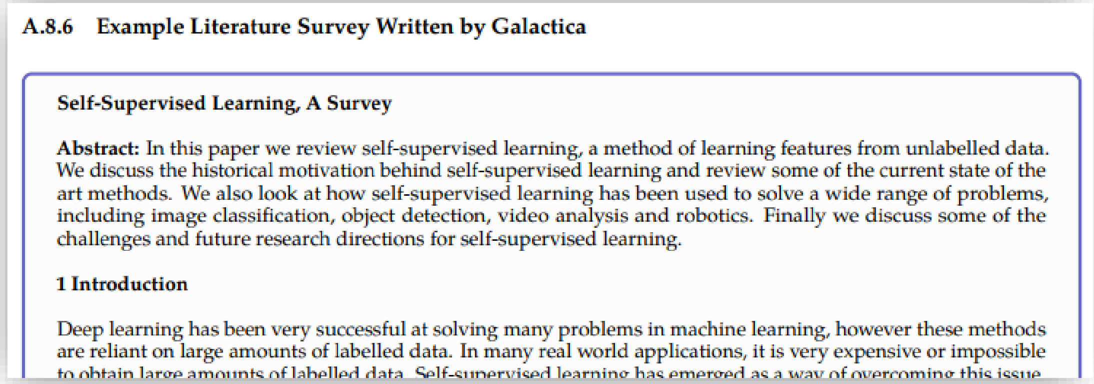


1. A. Vaswani, et al. "Attention is all you need." 2017.
2. J. Devlin, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." 2019.
3. Radford, Alec, et al. "Improving language understanding by generative pre-training." 2018.

Two NLP approaches

- Bottom up
  - Extract features with pre-trained encoder models (BERT[1]-like)
  - Build specialized models to recognize key elements
    - Named Entity Recognition, Text classification
- Top-down
  - Language generation with pre-trained decoder model (e.g. Galactica[2])
  - LLMs exhibit unaligned behavior
  - Prompt engineering and tuning
    - Leverage learned knowledge
    - Respond based on context from literature





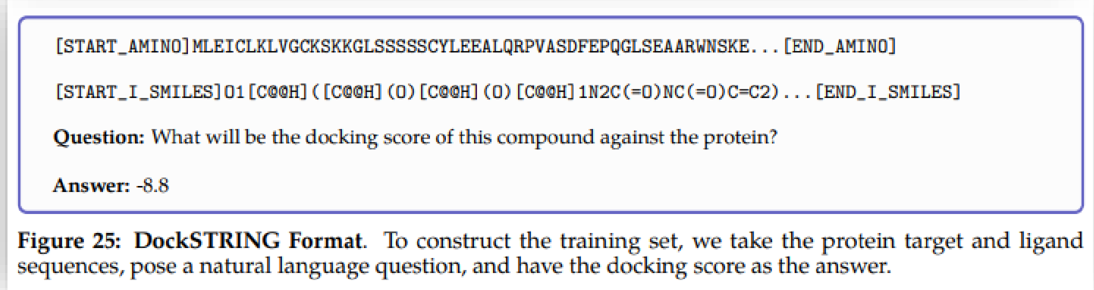
A.8.6 Example Literature Survey Written by Galactica

Self-Supervised Learning, A Survey

Abstract: In this paper we review self-supervised learning, a method of learning features from unlabelled data. We discuss the historical motivation behind self-supervised learning and review some of the current state of the art methods. We also look at how self-supervised learning has been used to solve a wide range of problems, including image classification, object detection, video analysis and robotics. Finally we discuss some of the challenges and future research directions for self-supervised learning.

1 Introduction

Deep learning has been very successful at solving many problems in machine learning, however these methods are reliant on large amounts of labelled data. In many real world applications, it is very expensive or impossible to obtain large amounts of labelled data. Self-supervised learning has emerged as a way of overcoming this issue



```
[START_AMINO]MLEICLKLVGCKSKKGLSSSSSCYLEEALQRPVASDFEPQGLSEAARWNSKE...[END_AMINO]

[START_I_SMILES]O1[C@@H]([C@@H](O)[C@@H](O)[C@@H]1N2C(=O)NC(=O)C=C2)...[END_I_SMILES]
```

Question: What will be the docking score of this compound against the protein?

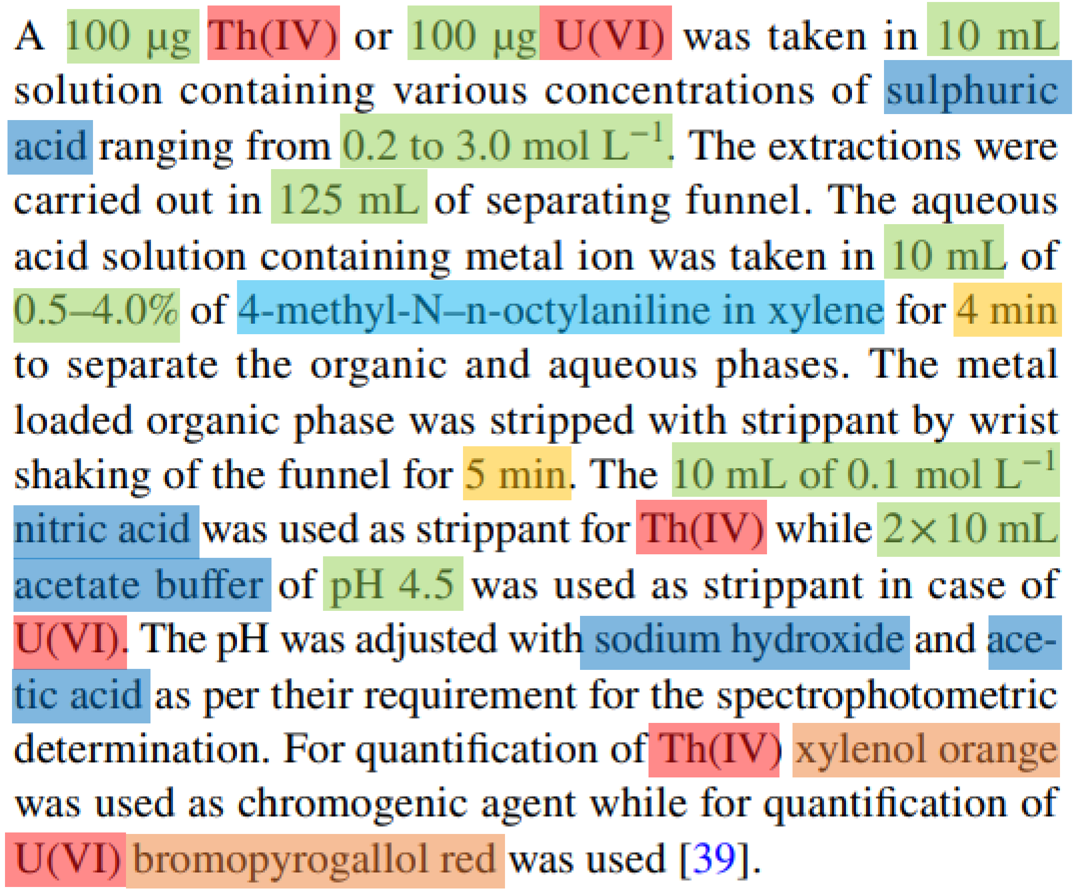
Answer: -8.8

Figure 25: DockSTRING Format. To construct the training set, we take the protein target and ligand sequences, pose a natural language question, and have the docking score as the answer.

J. Devlin, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." 2019.
R. Taylor, et al. "Galactica: A large language model for science." 2022.

NLP Task Examples

- Encoder approach
  - Classify text snippets according to contents
  - Recognize relevant chemical separation entities of defined classes
  - Infer relationships between extracted entities
- Decoder approach
  - Directly query model with paper text as context, e.g.:
    - *"{context} Does this paragraph describe a chemical separation?"*
    - *"{cumulative-context} What extractant is used in this paper?"*
  - Relevance and quality of outputs will vary widely by default
    - Build and train re-ranking proxy reward functions



A 100 µg Th(IV) or 100 µg U(VI) was taken in 10 mL solution containing various concentrations of sulphuric acid ranging from 0.2 to 3.0 mol L^{-1}. The extractions were carried out in 125 mL of separating funnel. The aqueous acid solution containing metal ion was taken in 10 mL of 0.5–4.0% of 4-methyl-N–n-octylaniline in xylene for 4 min to separate the organic and aqueous phases. The metal loaded organic phase was stripped with strippant by wrist shaking of the funnel for 5 min. The 10 mL of 0.1 mol L^{-1} nitric acid was used as strippant for Th(IV) while 2×10 mL acetate buffer of pH 4.5 was used as strippant in case of U(VI). The pH was adjusted with sodium hydroxide and acetic acid as per their requirement for the spectrophotometric determination. For quantification of Th(IV) xylenol orange was used as chromogenic agent while for quantification of U(VI) bromopyrogallol red was used [39].

P.S. More, *et al.* Extraction of Th(IV) and U(VI) with 4-methyl-N–n-octylaniline as an extracting agent. *J Radioanal Nucl Chem* **331**, 4149–4158 (2022).

Data, Caveats, and Warnings

- Compiling first-round list of journals to construct training and validation data corpus
    - Emphasize open access journals (API-accessible, machine-parseable)
    - ML tools for parsing old PDFs and scanned papers, if necessary
- Bias and variety
    - e.g. underrepresented negative results
    - LLMs often pre-trained on general language, require domain adaptation
    - Data augmentation techniques may be necessary, depending on text volume
- Open questions and challenges
    - Maximizing value of expert annotation and validation effort
    - Verifying generated separations are practical

Additional ML Opportunities

- Extracting knowledge from charts and tables
  - Useful for both recent and old papers

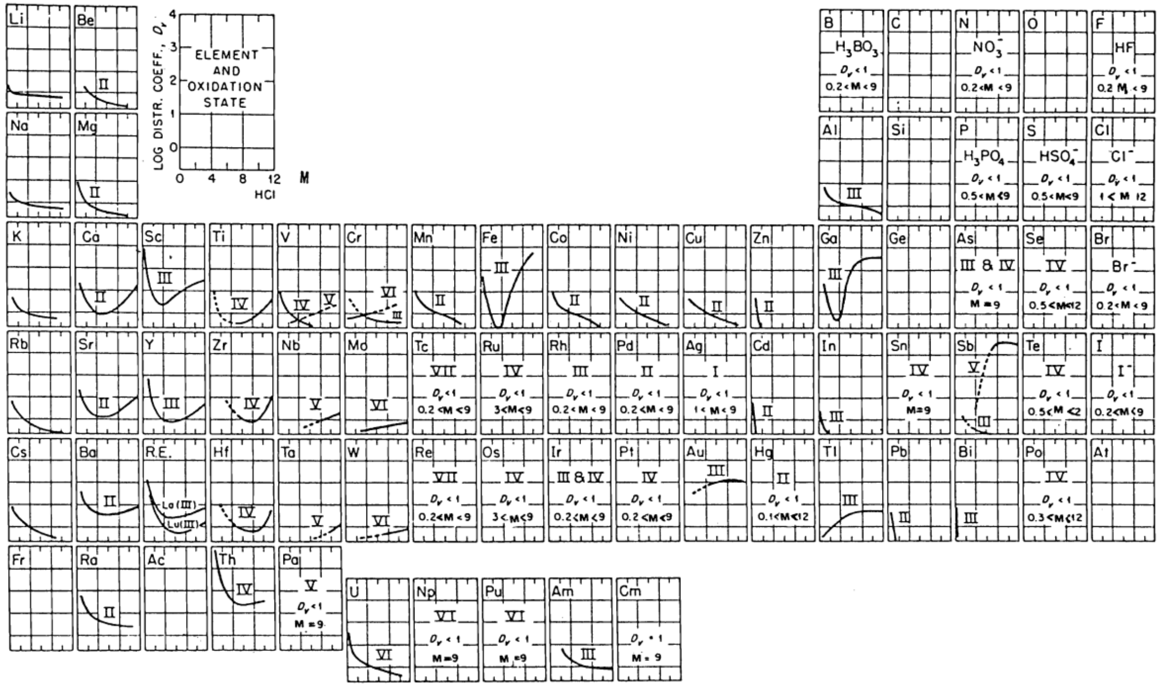


Fig. 1. Cation exchange distribution coefficients in HCl solutions.(Dowex 50-X4) (Ref.1)

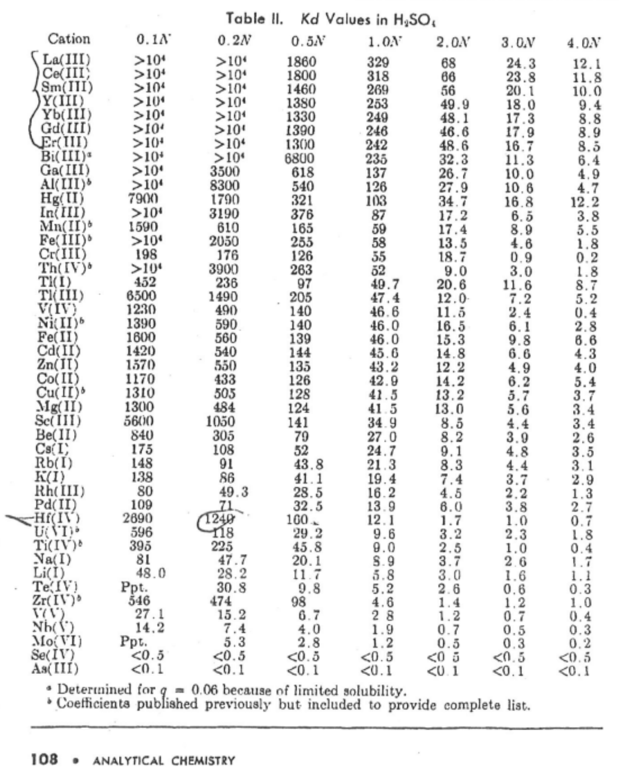


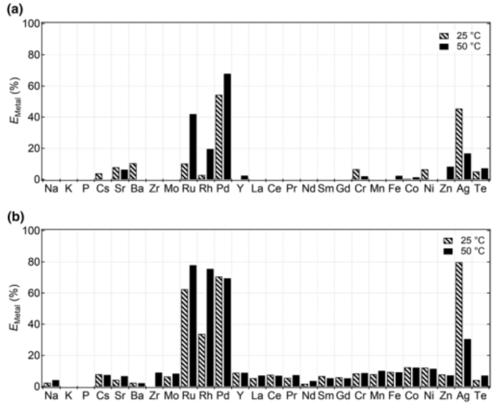
Table II. *Kd* Values in H_2SO_4

Cation	0.1*N*	0.2*N*	0.5*N*	1.0*N*	2.0*N*	3.0*N*	4.0*N*
La(III)	$>10^4$	$>10^4$	1860	329	68	24.3	12.1
Ce(III)	$>10^4$	$>10^4$	1800	318	66	23.8	11.8
Sm(III)	$>10^4$	$>10^4$	1460	269	56	20.1	10.0
Y(III)	$>10^4$	$>10^4$	1380	253	49.9	18.0	9.4
Yb(III)	$>10^4$	$>10^4$	1330	249	48.1	17.3	8.8
Gd(III)	$>10^4$	$>10^4$	1390	246	46.6	17.9	8.9
Er(III)	$>10^4$	$>10^4$	1300	242	48.6	16.7	8.5
Bi(III)[a]	$>10^4$	$>10^4$	6800	235	32.3	11.3	6.4
Ga(III)	$>10^4$	3500	618	137	26.7	10.0	4.9
Al(III)[b]	$>10^4$	8300	540	126	27.9	10.6	4.7
Hg(II)	7900	1790	321	103	34.7	16.8	12.2
In(III)	$>10^4$	3190	376	87	17.2	6.5	3.8
Mn(II)[b]	1590	610	165	59	17.4	8.9	5.5
Fe(III)[b]	$>10^4$	2050	255	58	13.5	4.6	1.8
Cr(III)	198	176	126	55	18.7	0.9	0.2
Th(IV)[b]	$>10^4$	3900	263	52	9.0	3.0	1.8
Tl(I)	452	236	97	49.7	20.6	11.6	8.7
Tl(III)	6500	1490	205	47.4	12.0	7.2	5.2
V(IV)	1230	490	140	46.6	11.5	2.4	0.4
Ni(II)[b]	1390	590	140	46.0	16.5	6.1	2.8
Fe(II)	1600	560	139	46.0	15.3	9.8	6.6
Cd(II)	1420	540	144	45.6	14.8	6.6	4.3
Zn(II)	1570	550	135	43.2	12.2	4.9	4.0
Co(II)	1170	433	126	42.9	14.2	6.2	5.4
Cu(II)[b]	1310	505	128	41.5	13.2	5.7	3.7
Mg(II)	1300	484	124	41.5	13.0	5.6	3.4
Sc(III)	5600	1050	141	34.9	8.5	4.4	3.4
Be(II)	840	305	79	27.0	8.2	3.9	2.6
Cs(I)	175	108	52	24.7	9.1	4.8	3.5
Rb(I)	148	91	43.8	21.3	8.3	4.4	3.1
K(I)	138	86	41.1	19.4	7.4	3.7	2.9
Rh(III)	80	49.3	28.5	16.2	4.5	2.2	1.3
Pd(II)	109	71	32.5	13.9	6.0	3.8	2.7
Hf(IV)	2690	1240	160	12.1	1.7	1.0	0.7
U(VI)[b]	596	118	29.2	9.6	3.2	2.3	1.8
Ti(IV)[b]	395	225	45.8	9.0	2.5	1.0	0.4
Na(I)	81	47.7	20.1	8.9	3.7	2.6	1.7
Li(I)	48.0	28.2	11.7	5.8	3.0	1.6	1.1
Te(IV)	Ppt.	30.8	9.8	5.2	2.6	0.6	0.3
Zr(IV)[b]	546	474	98	4.6	1.4	1.2	1.0
V(V)	27.1	15.2	6.7	2.8	1.2	0.7	0.4
Nb(V)	14.2	7.4	4.0	1.9	0.7	0.5	0.3
Mo(VI)	Ppt.	5.3	2.8	1.2	0.5	0.3	0.2
Se(IV)	<0.5	<0.5	<0.5	<0.5	<0.5	<0.5	<0.5
As(III)	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1

[a] Determined for $q = 0.06$ because of limited solubility.
[b] Coefficients published previously but included to provide complete list.

108 • ANALYTICAL CHEMISTRY

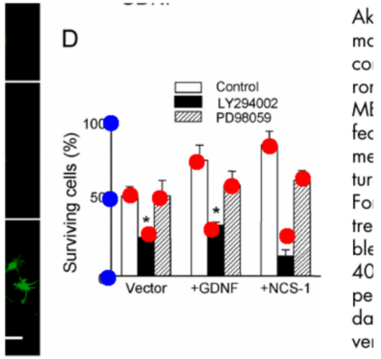
Fig. 10 Extraction selectivity of metal ions from the simulated HLLW using the ionic liquids ([Bmim][NTf_2]: (**a**), [Bmim][NfO]: (**b**)) containing MPE-TDGA. ([MPE-TDGA]: 50 mM, Mixing time: 2 h, $V_{0,Aq} = V_{IL}$: 1 cm^3)

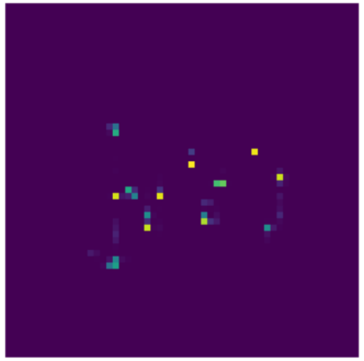


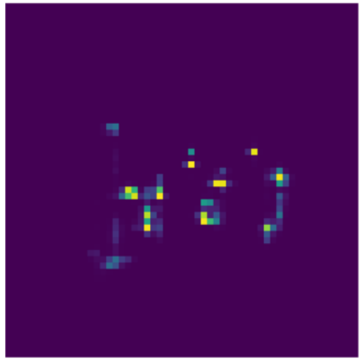
F.Nelson,T.Murase and K.A.Kraus,J.Chromatog.,13,503(1964).

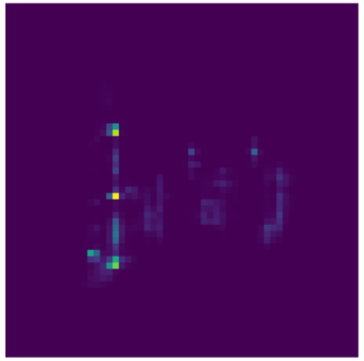
H. Oosugi, et al. "Extraction behaviors of platinum group metals from an aqueous HNO3 solution using ionic liquids containing a novel thiodiglycolamide-type extractant." *Journal of Radioanalytical and Nuclear Chemistry* 331.11 (2022): 4577-4585.

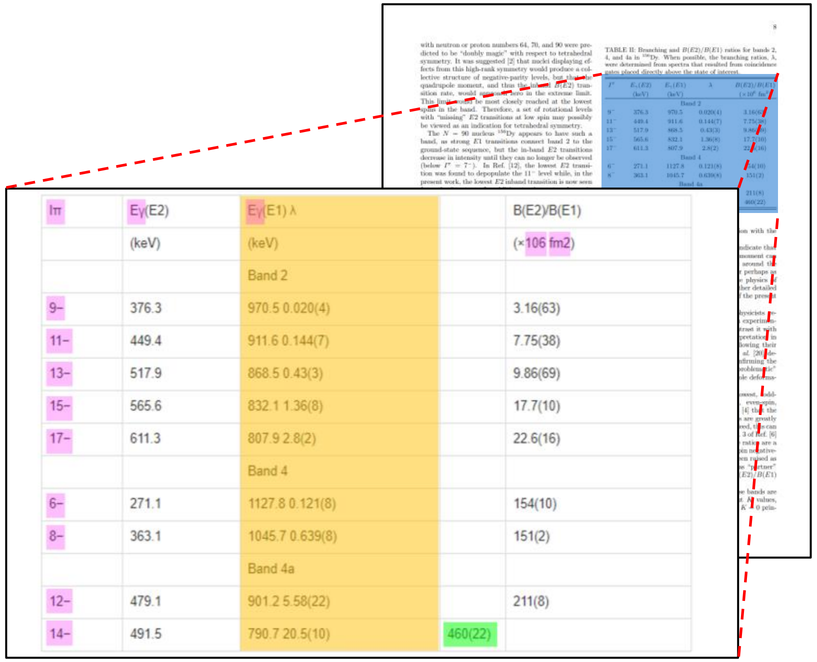
Additional ML Opportunities



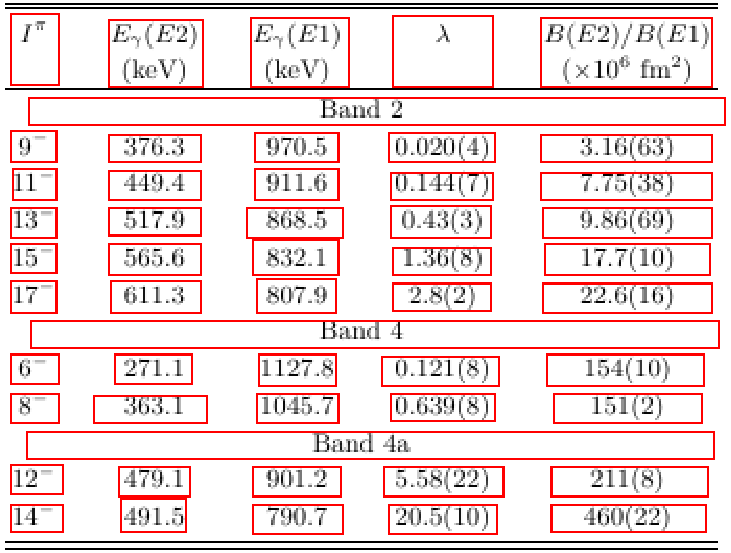




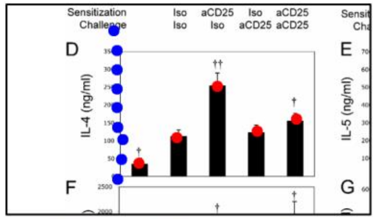


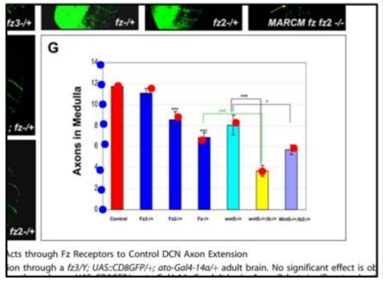


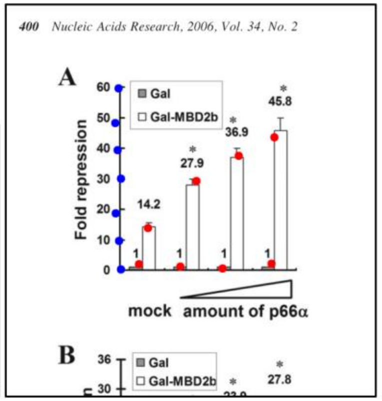
Iπ	Eγ(E2)	Eγ(E1) λ		B(E2)/B(E1)
	(keV)	(keV)		(×106 fm2)
		Band 2		
9–	376.3	970.5 0.020(4)		3.16(63)
11–	449.4	911.6 0.144(7)		7.75(38)
13–	517.9	868.5 0.43(3)		9.86(69)
15–	565.6	832.1 1.36(8)		17.7(10)
17–	611.3	807.9 2.8(2)		22.6(16)
		Band 4		
6–	271.1	1127.8 0.121(8)		154(10)
8–	363.1	1045.7 0.639(8)		151(2)
		Band 4a		
12–	479.1	901.2 5.58(22)		211(8)
14–	491.5	790.7 20.5(10)	460(22)	

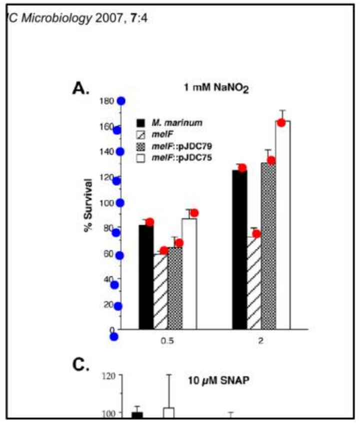


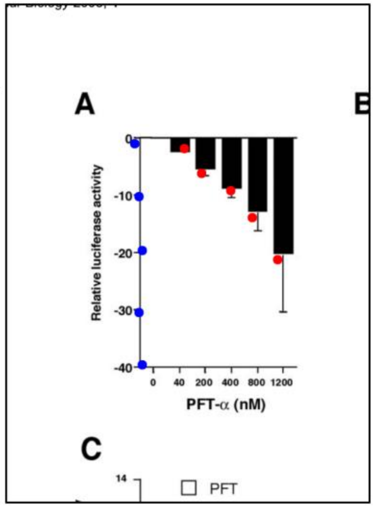
I^π	$E_\gamma(E2)$ (keV)	$E_\gamma(E1)$ (keV)	λ	$B(E2)/B(E1)$ ($\times 10^6$ fm^2)
		Band 2		
9^-	376.3	970.5	0.020(4)	3.16(63)
11^-	449.4	911.6	0.144(7)	7.75(38)
13^-	517.9	868.5	0.43(3)	9.86(69)
15^-	565.6	832.1	1.36(8)	17.7(10)
17^-	611.3	807.9	2.8(2)	22.6(16)
		Band 4		
6^-	271.1	1127.8	0.121(8)	154(10)
8^-	363.1	1045.7	0.639(8)	151(2)
		Band 4a		
12^-	479.1	901.2	5.58(22)	211(8)
14^-	491.5	790.7	20.5(10)	460(22)


Sensitization
Challenge
Iso
Iso
aCD25
Iso
Iso
aCD25
aCD25
aCD25
Sensit
Cha
D
E
IL-4 (ng/ml)
IL-5 (ng/ml)
F
G


fz2-/+
fz2-/+
MARCM fz fz2 -/-
G
Axons in Medulla
Control
cts through Fz Receptors to Control DCN Axon Extension
ion through a fz3/Y; UAS::CD8GFP/+; ato-Gal4-14a/+ adult brain. No significant effect is ob


400 Nucleic Acids Research, 2006, Vol. 34, No. 2
A
Gal
Gal-MBD2b
Fold repression
mock
amount of p66α
B
Gal
Gal-MBD2b


C Microbiology 2007, 7:4
A.
1 mM NaNO2
M. marinum
melF
melF::pJDC79
melF::pJDC75
% Survival
C.
10 μM SNAP


A
B
Relative luciferase activity
PFT-α (nM)
C
PFT