

Automated spelling correction to improve recall rates of name-to-structure tools for chemical text mining

ChemAxon UGM Budapest, 17-18th May 2011

Sorel Muresan¹, Paul-Hongxing Xie¹, Roger Sayle²

¹ *AstraZeneca R&D Mölndal*

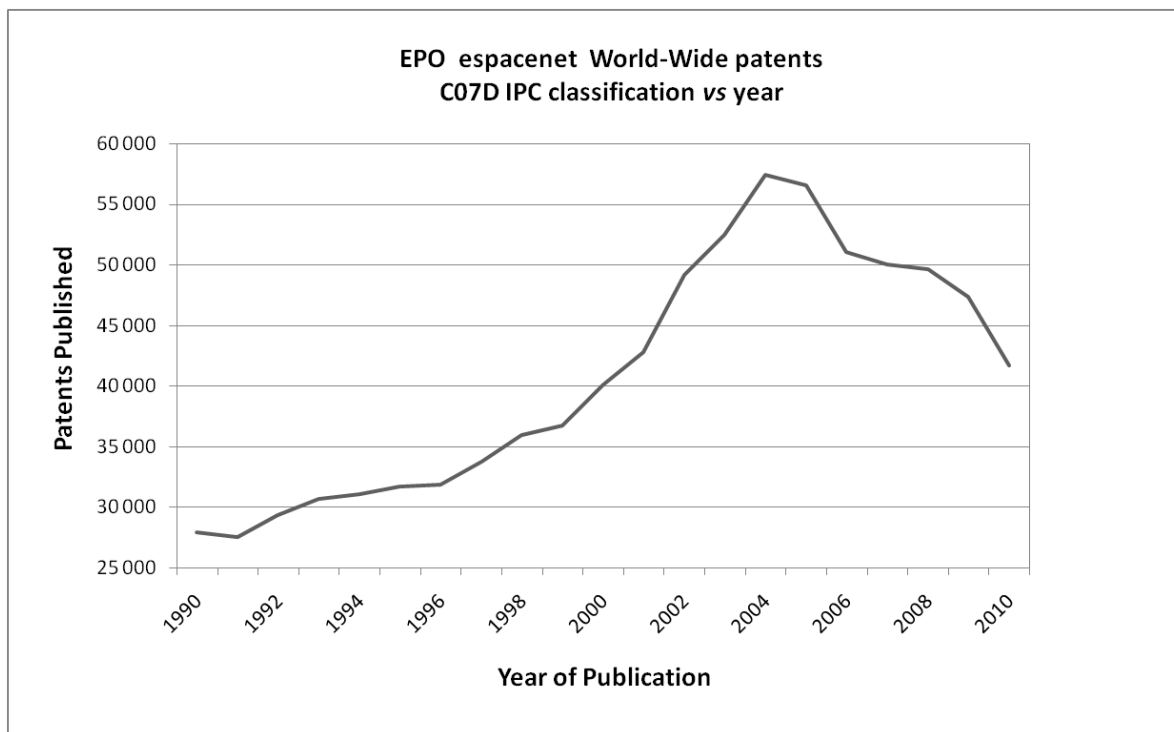
² *NextMove Software*

Unrestricted



Driver – explosion in SAR knowledgebases

- The single largest published source of *in vitro* SAR is patent applications

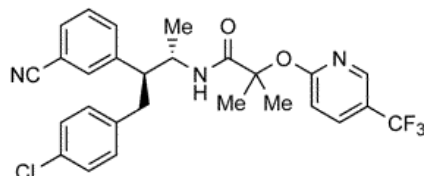


Patents as pharmaceutical data source

Complementary between journals and patents

“In certain fields, new advances are disclosed in patents long before they are published in peer-reviewed journals.” *Grubb. W.P.*

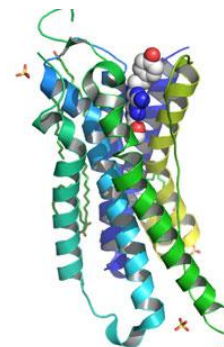
“Novel Cannabinoid-1 Receptor Inverse Agonist for the Treatment of Obesity”



modulates



CNR1



Patent application
Nov 2002

Patent publication
Mar 2004

Journal publication
Dec 2006

~18 months

2.5 years

USPTO patents (PN: US20040058820)

Journal of
Medicinal Chemistry (PMID: 17181138)

US 20040058820A1

(19) **United States**
(12) **Patent Application Publication** (10) Pub. No.: US 2004/0058820 A1
Hagmann et al. (43) Pub. Date: Mar. 25, 2004

(54) **SUBSTITUTED AMIDES** Publication Classification

(76) Inventors: William K. Hagmann, Westfield, NJ (US); Linus S. Lin, Westfield, NJ (US); Shrenik K. Shah, Menuchen, NJ (US); Ravindra N. Guthikonda, Edison, NJ (US); Hongbo Qi, Edison, NJ (US); Linda L. Chang, Wayne, NJ (US); Ping Liu, Edison, NJ (US); Helen M. Armstrong, Westfield, NJ (US); James P. Jewell, Jersey City, NJ (US); Thomas J. Lanza JR., Edison, NJ (US)

(51) Int. Cl.⁷ A01N 47/28; A01N 43/40; A01N 43/50; A01N 43/56; C07D 213/78; C07D 233/80; C07D 231/36

(52) U.S. Cl. 504/254; 504/260; 504/280; 504/279; 504/330; 504/336; 546/298; 548/318.1; 548/367.1; 564/48; 564/170

(57) **ABSTRACT**
Novel compounds of the structural formula (I) are antagonists and/or inverse agonists of the Cannabinoid-1 (CB1)

Discovery of *N*-[(1*S*,2*S*)-3-(4-Chlorophenyl)-2-(3-cyanophenyl)-1-methylpropyl]-2-methyl-2-[[5-(trifluoromethyl)pyridin-2-yl]oxy]propanamide (MK-0364), a Novel, Acyclic Cannabinoid-1 Receptor Inverse Agonist for the Treatment of Obesity

Linus S. Lin,^{*,†} Thomas J. Lanza, Jr.,[†] James P. Jewell,[‡] Ping Liu,[‡] Shrenik K. Shah,[‡] Hongbo Qi,[‡] Xinchun Tong,[‡] Junying Wang,[‡] Suoyu S. Xu,[‡] Tung M. Fong,[‡] Chun-Pyn Shen,[‡] Julie Lao,[‡] Jing Chen Xiao,[‡] Lauren P. Shearman,[‡] D. Sloan Stribling,[‡] Kimberly Rosko,[‡] Alison Strack,[‡] Donald J. Marsh,[‡] Yue Feng,[‡] Sanjeev Kumar,[‡] Koppara Samuel,[‡] Wenji Yin,[‡] Lex H. T. Van der Ploeg,[‡] Mark T. Goulet,[‡] and William K. Hagmann[‡]

Abstract

Supporting Info

Full Text HTML

Hi-Res PDF [85 KB]

PDF w/ Links [163 KB]

Driver – improve chemical NER

- The biggest cause of missing compounds when extracting chemical entities from text is the presence of typographical errors: human errors, OCR failures, hyphenation and multiple line issues, etc.



OCR Errors: Compound Names

- 1H-ben zimidazole → 1H-benzimidazole
- triphenylposhine → triphenylphosphine
- 4- (2-ADAMANTYLCARBAMOYL) -5-TERT-BUTYL-PYRAZOL-1-YL] BENZOIC ACID →
4-(2-adamantylcarbamoyl)-5-tert-butyl-pyrazol-1-yl]benzoic acid



OCR Errors: Compound Names

- Searching full-text patents (WO, EP, US, FR, GB, DE, JP) for the term “Simvastatin” yields 9030 patents (3666 INPADOC families).

But there are 392 more patents which are not found due to typos and OCR errors:

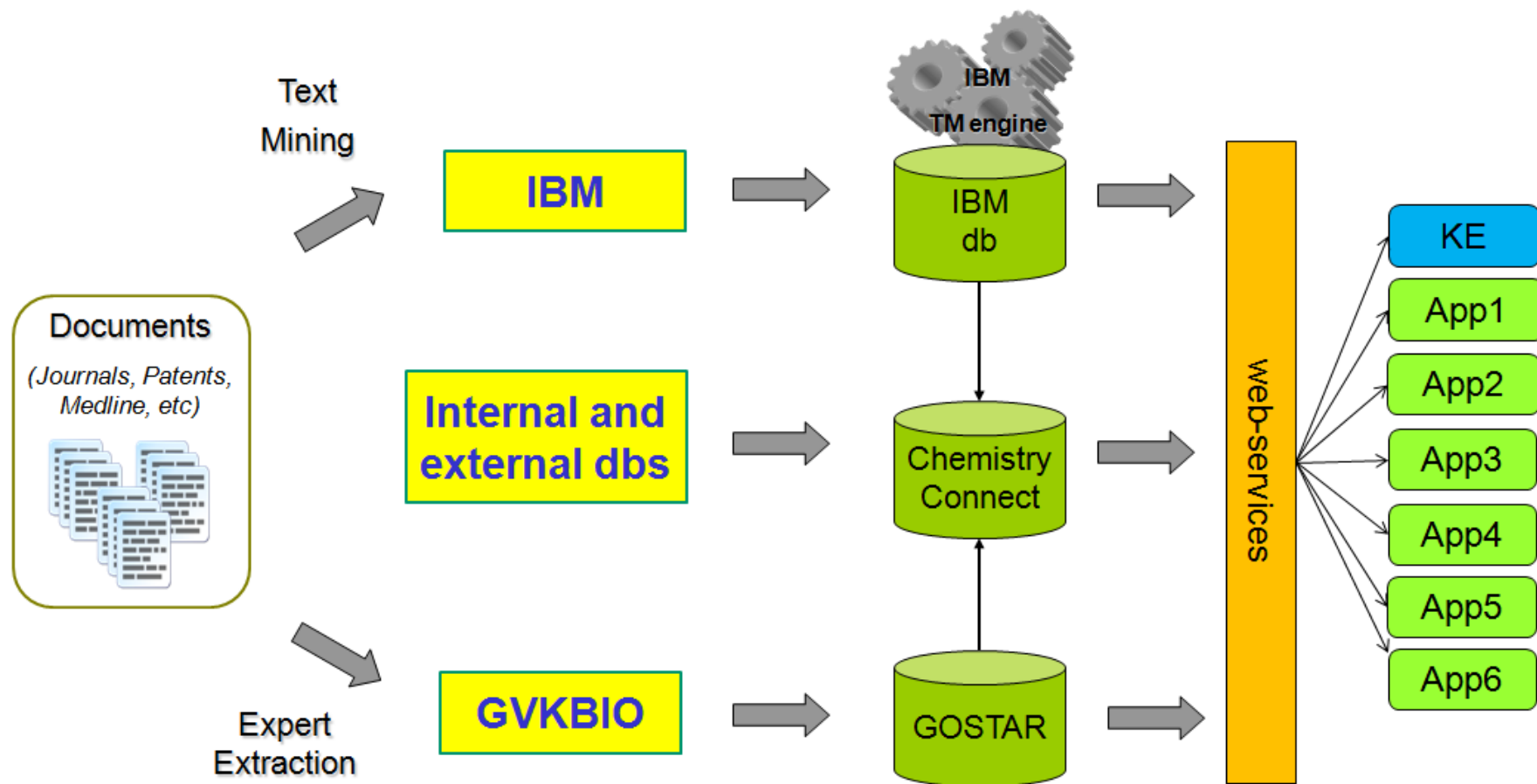
SIMVASTATINE	71
SIMVESTATIN	43
SIMVISTATIN	33
SINVASTATIN	28
SIMVARSTATIN	26
SYMVASTATIN	14
SIMAVASTATIN	13
SLMVASTATIN	9
SIMBASTATIN	8
SIMVASTSTIN	8
SIMVATATIN	7
SIMVASTATINA	6
SIMIVASTATIN	5
SIMVASTATION	4

S1MVASTATIN	3
SIMASTATIN	2
SIMNVASTATIN	2
SIMVASTATIV	2
SIMVASTITIN	2
SIVASTATIN	2
IMVASTATIN	1
S IMVASTATIN	1
S IMVASTATINA	1
SII MVASTATIN	1
SIM ASTATINE	1
SIMVASTACIN	1
SIMVASTAFIN	1
SIMVASTALIN	1

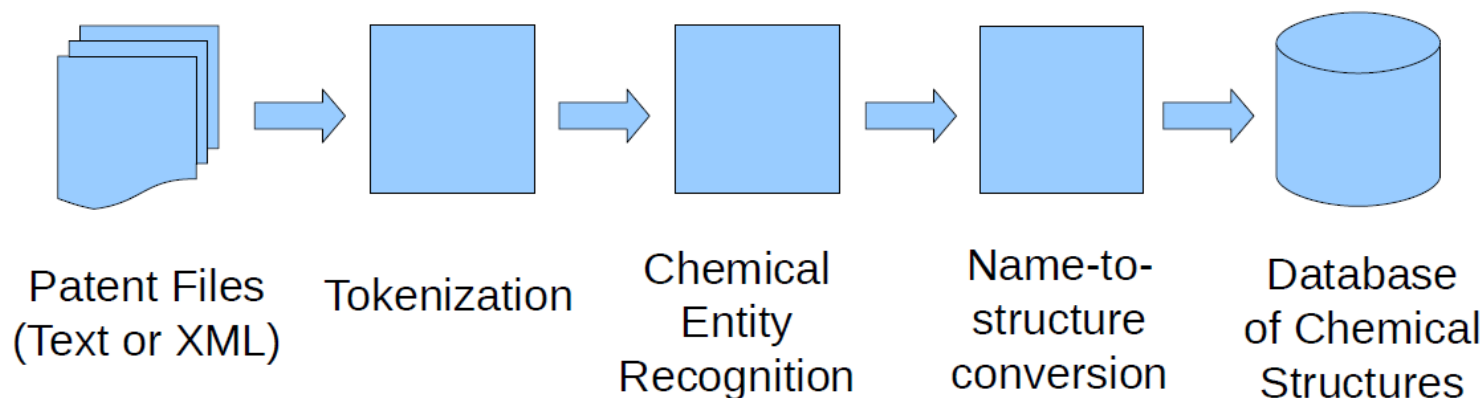
SIMVASTANINA	1
SIMVASTATI NE	1
SIMVASTATIN7	1
SIMVASTATING	1
SIMVASTATINM	1
SIMVASTATINO	1
SIMVASTATI U	1
SIMVASTATJN	1
SIMVASTATN	1
SIMVASTAT'N	1
SIRVASTATIN	1
YSIMVASTATINE	1

and more...

Chemistry Connect



Traditional text mining pipeline



- Determining the start and end of IUPAC-like names in free text is a tricky business.
- Chemical names can contain whitespace, hyphens, commas, parenthesis, brackets, braces, apostrophes, superscripts, greek characters, digits and periods.
- This is made harder still by typos, OCR errors, hyphenation, linefeeds, XML tags, line and page numbers and similar noise.



CaffeineFix

- NextMove Software's CaffeineFix is intended to fill a niche opportunity as a chemical nomenclature aware automatic spell checker.
- As a pre-processing step in a pipeline, it can significantly improve the recall rates of name to structure tools in text mining applications: Lexichem, ChemAxon, ACD/Name, CambridgeSoft nam=struct, OPSIN, etc

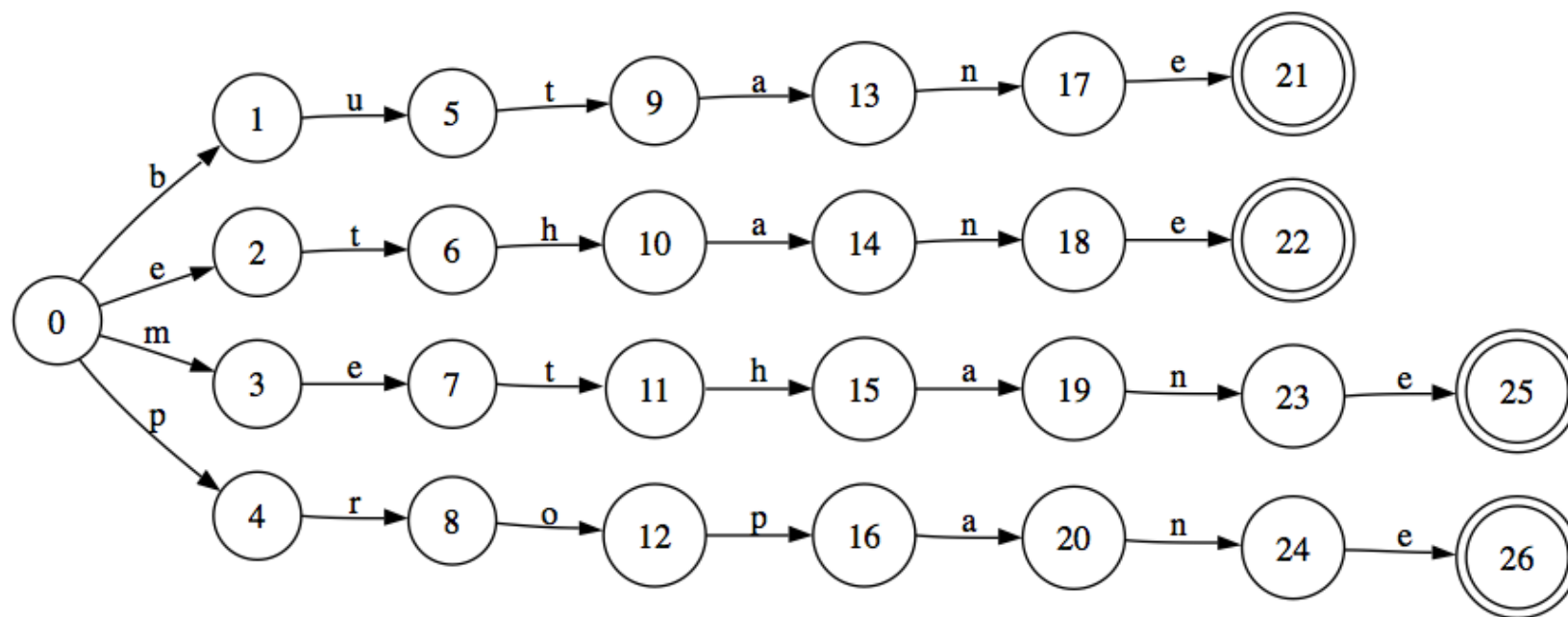


Example chemical lexicon

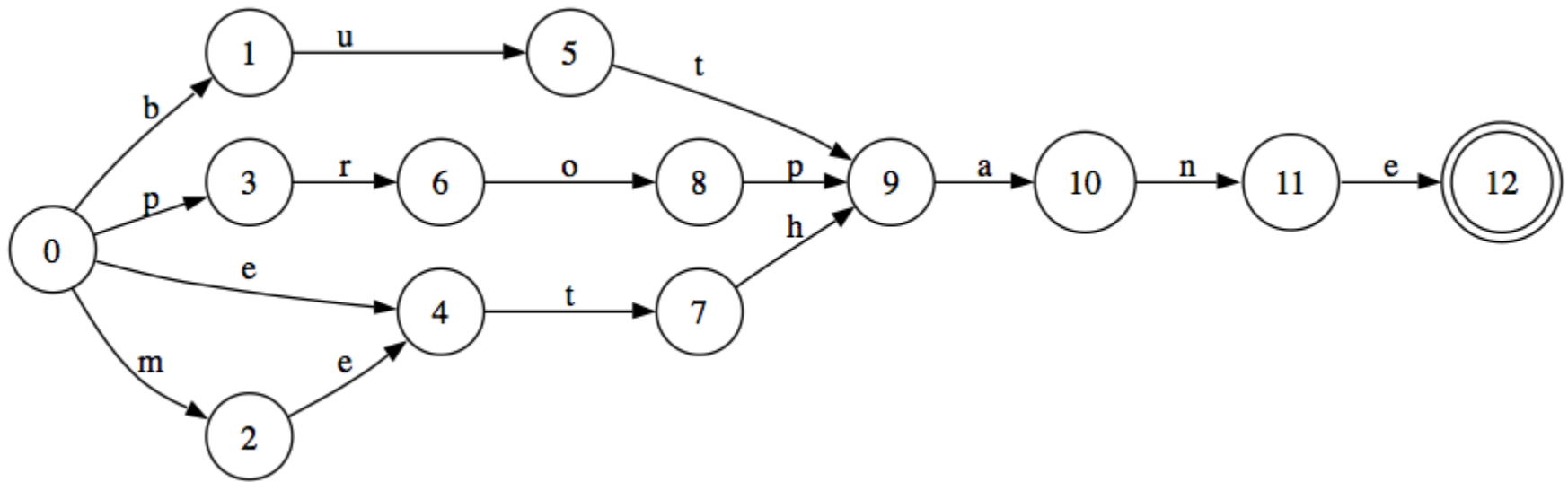
- Lower alkanes
 - Methane
 - Ethane
 - Propane
 - Butane
- Chemical NER = string matching of terms or keywords describing chemical entities
 - dictionaries
 - FSMs (finite state machines)



Representing lexicons as TRIEs



Representing lexicons as DAGs



IUPAC-like grammar example

```
locant := "#" /* any digit */
```

```
subst := "bromo" | "chloro" | "fluoro"
```

```
alk := "meth" | "eth" | "prop" | "but"
```

```
parent := alk "ane"
```

```
prefix := [ prefix "-" ] [ loc "-" ]
```

```
subst
```

```
| [ prefix ] subst
```

```
name := [ prefix [ "-" ] ] parent
```



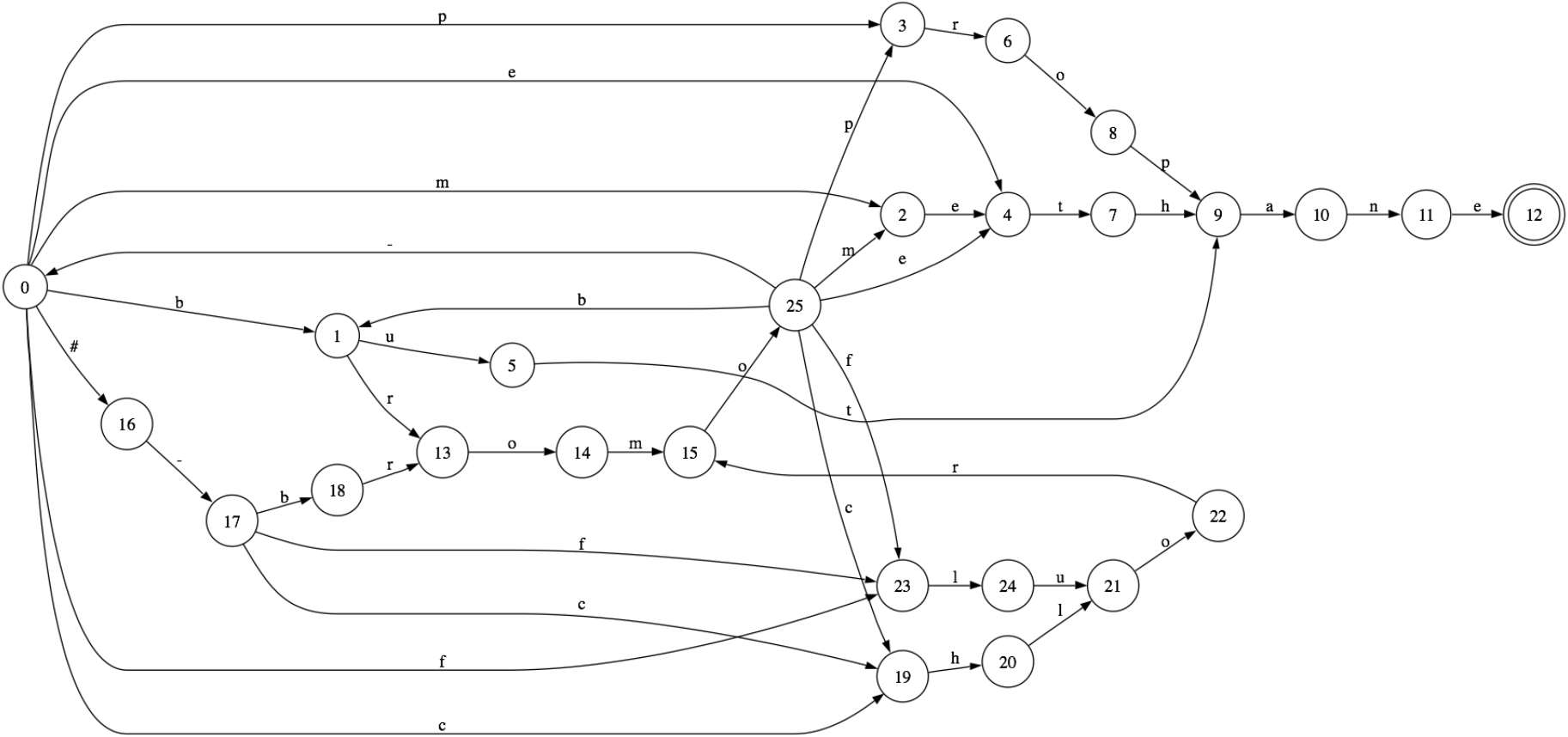
IUPAC-like grammar example

- methane
- chloroethane
- 2-bromo-propane
- chloro-bromo-methane
- 1-fluoro-2-chloro-ethane
- chlorofluoromethane

- 9-bromomethane
- 1-chloro-1-chloro-1-chloro-methane



Representing grammars as dFAs

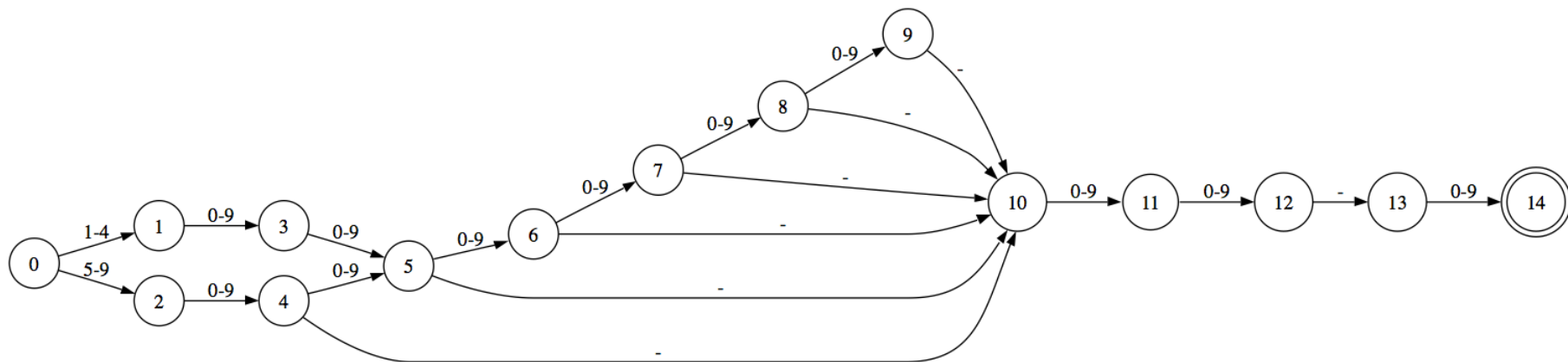


Pharmaceutical registry numbers

- Prefix: “A” | “AZ” | “BMY” | “GSK” | “LY” | ...
- Number: $\setminus d\{3-7\}$
- Suffix: (“.” $\setminus d$) | [“a” .. “z”]
- Grammar: Prefix [“ ” | “-”] Number [Suffix]



CAS registry number grammar



- Two to seven digits, followed by a hyphen, two digits, a hyphen and a final check digit
- e.g. 7732-18-5
- RegExp: $(([1-9]\{2,5\})|([5-9]\{1\}))-\{2\}-\{1\}$



Push-down automata

- Unfortunately, DFAs are not powerful enough to capture the context-sensitive grammars needed for IUPAC-like names.
- The problem is nesting of parenthesis.
- Push-down automata are variants of DFAs that maintain an additional stack.
- This allows checking that parenthesis, brackets and braces are balanced and that open and close characters are matched.

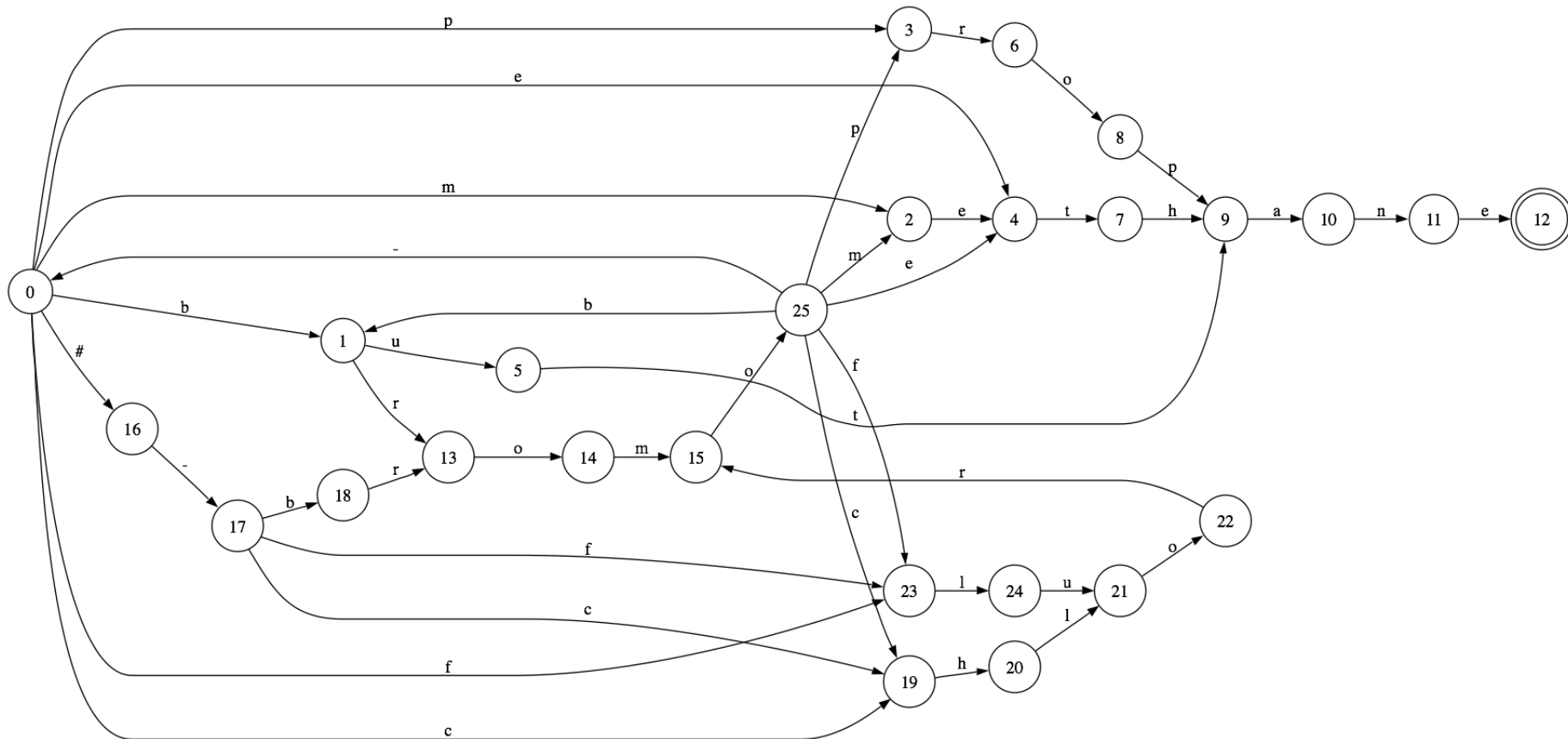


Spelling correction

- A relatively simple extension of the above exact match algorithm allows CaffeineFix's data structure to be used for automatic error correction.
- Backtracking allows consideration of substitution, insertion and deletion whilst traversing the finite state machine (FSM).
- Allows enumeration of all valid names within a specified edit-distance of a string.



Representing grammars as dFAs



2bromo-propane -> 2-bromo-propane



IBM patents

- 11 million full-text patents
 - IBM text mining & name=struct
 - CaffeineFix at D=0 and D=1
 - ChemAxon, Lexichem, OPSIN



Names vs. SMILES extracted from patents

Data Source	Chemical Names	n2s_1	n2s_2	n2s_3
IBM IP	12,831,351	4,033,247	4,072,166	4,891,063
CF (d=0)	10,311,200	4,505,685	3,829,260	3,836,953
CF (d=1)	13,523,384	5,431,587	3,993,432	4,438,586
Total	23,405,430	9,753,767	5,639,813	6,419,592



Conversion by Name Class (CF, D=0)

Class	Category	Names	ChemAxon 5.5 converts 60%	n2s_3 (%)	None (%)	
M	Molecule	7,262,798	61.4	64.8	77.1	7.8
D	Dictionary	26,876	38.1	45.1	3.5	38.5
R	Registry number	304,064	0	0	0	100
C	CAS number	47,815	0	0	0	100
E	Element	836				
P	Fragment	2,663,677	NCI/CADD Chemical Identifier Resolver converts 48%			
A	Atom fragment	96	56.6	56.5	0	6.5
Y	Polymer	295	0	44.1	22.7	36.9
G	Generic	1263	2.6	6.3	0.5	91.9
N	Noise	104	32.7	24	19.2	52.9
	Total	10,307,824	76,3	610	54.3	14.1

ChemAxon 5.5
converts 60%

NCI/CADD Chemical Identifier Resolver
converts 48%



Heatmap (Filtered Canonical SMILES, CF D=0)

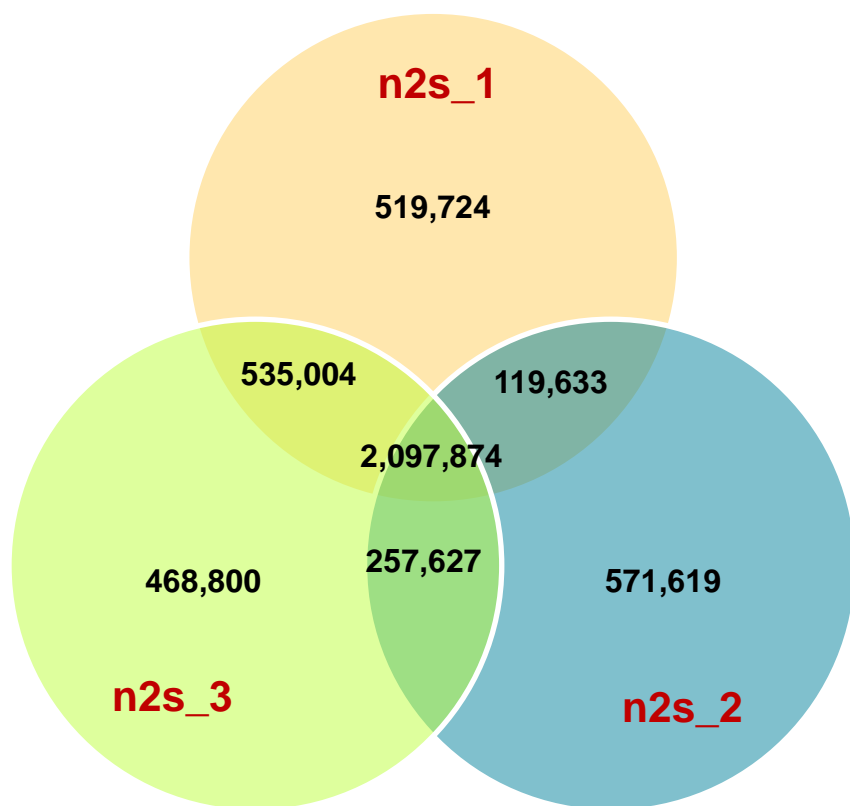
n2s_1	3,272,235
n2s_2	3,046,753
n2s_3	3,359,305

	n2s_1	ns2_2	n2s_3
n2s_1	1.00	0.73	0.78
n2s_2	0.68	1.00	0.70
n2s_3	0.80	0.77	1.00

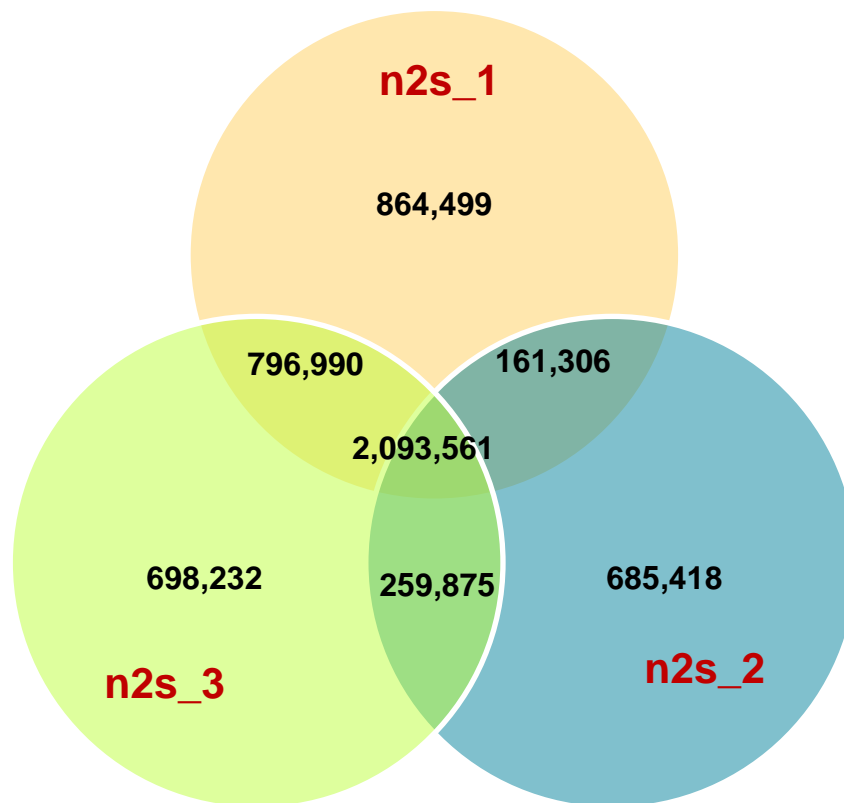


Venn diagrams (Filtered Canonical SMILES, CF)

D=0 (4,570,281)



D=1 (5,559,881)



Unique Canonical SMILES

Data Source	SMILES
IBM IP (CS)	5,148,087
IBM IP (CS+L+C+O)	6,643,120
CF (L+C+O) (D=0)	4,570,281
CF (L+C+O) (D=1)	5,559,881
Total	8,750,382



Summary

- Unique chemistry from patents via text mining (12% out of 47M parent structures in Chemistry Connect)
- CaffeineFix significantly improves extraction rates (22% increase from D=0 to D=1 for the filtered set of SMILES)
- name2structure software are complementary (40% of the structures come from single n2s contributions)



Acknowledgements

- Plamen Petrov
- Thierry Kogej
- Ithipol Suriyawongkul

- Markus Sitzmann
- Daniel Lowe
- Daniel Bonniot



Thank you!

