

Where do all those small molecules come from? What the data reveal

Paul Peters

Director, EMEA Sales
CAS – SIIL, Netherlands

ICIC, Vienna, Austria
October 27th, 2010



CAS is a division of the American Chemical Society

ACS Vision

Improving people's lives through the transforming power of chemistry



CAS supports the mission of the ACS

ACS Mission

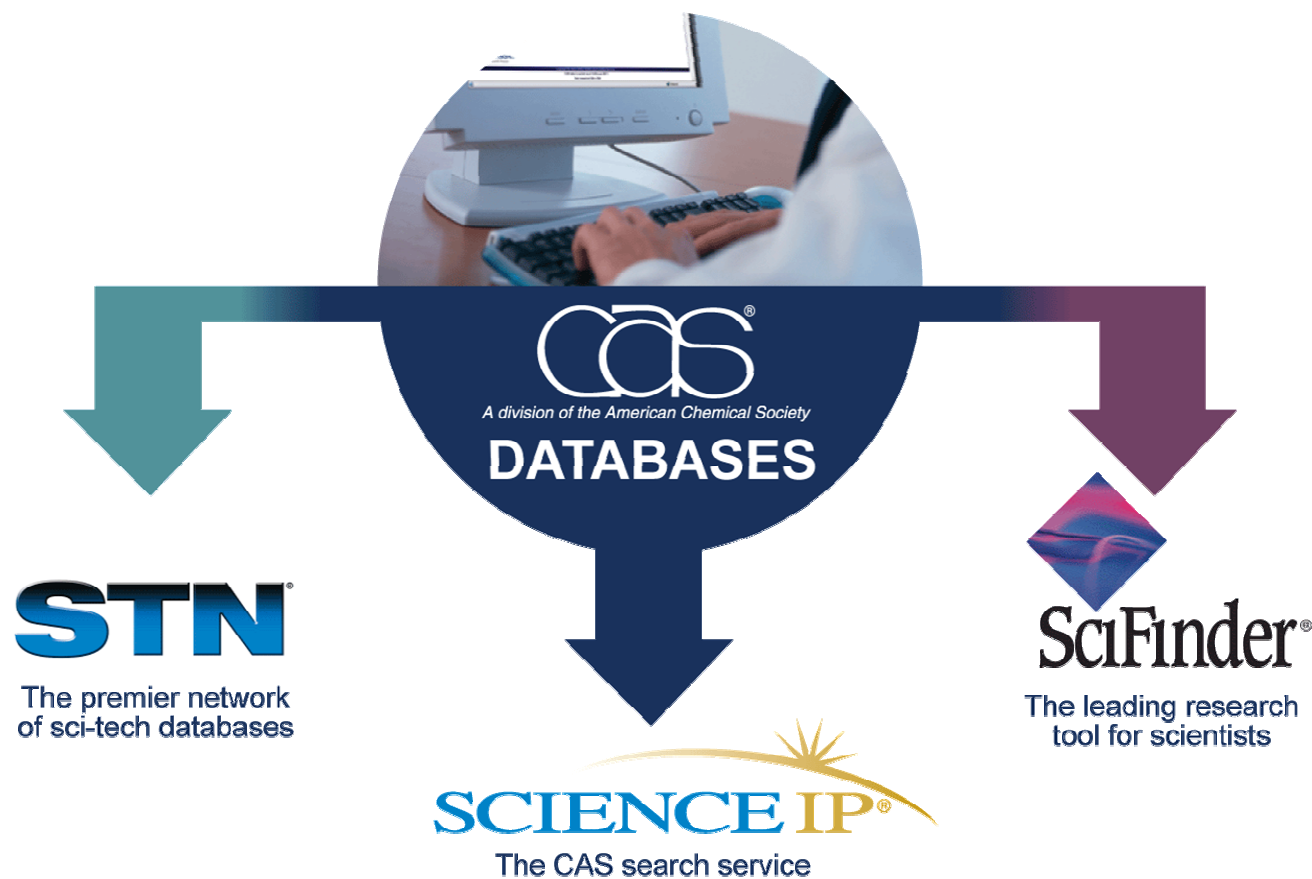
To advance the broader chemistry enterprise and its practitioners for the benefit of Earth and its people.

CAS Commitment

To support the ACS mission by providing secure access to high-quality, comprehensive chemical information.



Three major services provide access to CAS database information



CAS mission drives development of the CAS databases

CAS REGISTRYSM

>55M organic and inorganic substances

>62M sequences

Updated daily (~12K daily)

Substances reported in the literature back to 1802

Information pertaining to these substances has been enriched with experimental and predicted property data with more than 2.8 billion property values, data tags, and spectra

CASREACT[®]

>27M single- and multi-step reactions

>13M synthetic preparations

Extracted from patents and journal articles

Updated weekly (30K-50K per week)

Reactions back to 1840

Reaction conditions starting in 2003

CAS Databases

CAplusSM

>33M patent and journal article references

>10K major scientific journals covered

Patents from 61 patent offices

Updated daily (~3K daily)

Links to more than 360 publishers and 3 patent offices

Literature back to early 1800s

Cited articles from 1997 onward, currently more than 275M citations

CHEMCATS[®]

>42M commercially available compounds

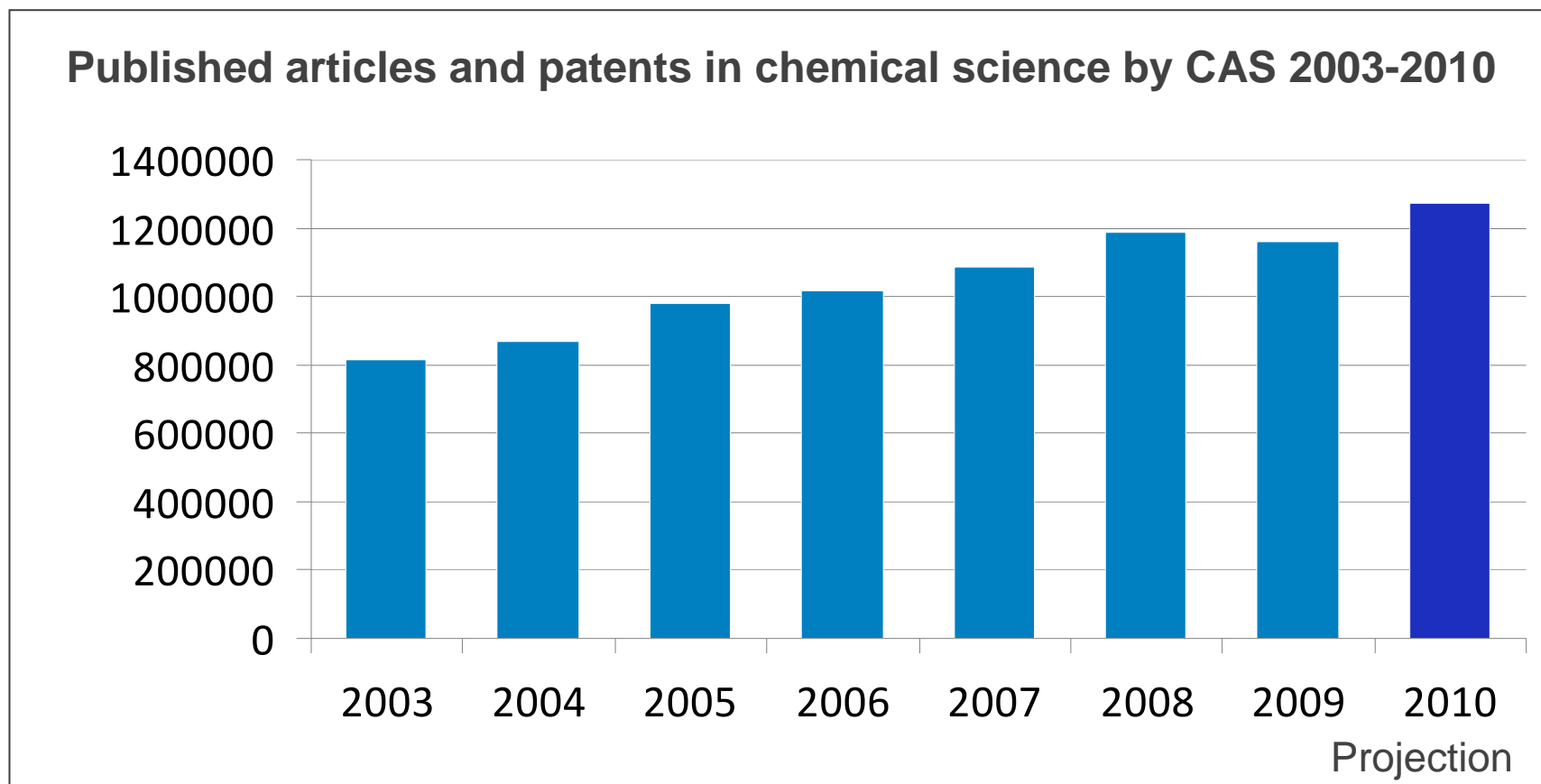
>1,100 suppliers

>1,215 chemical catalogs

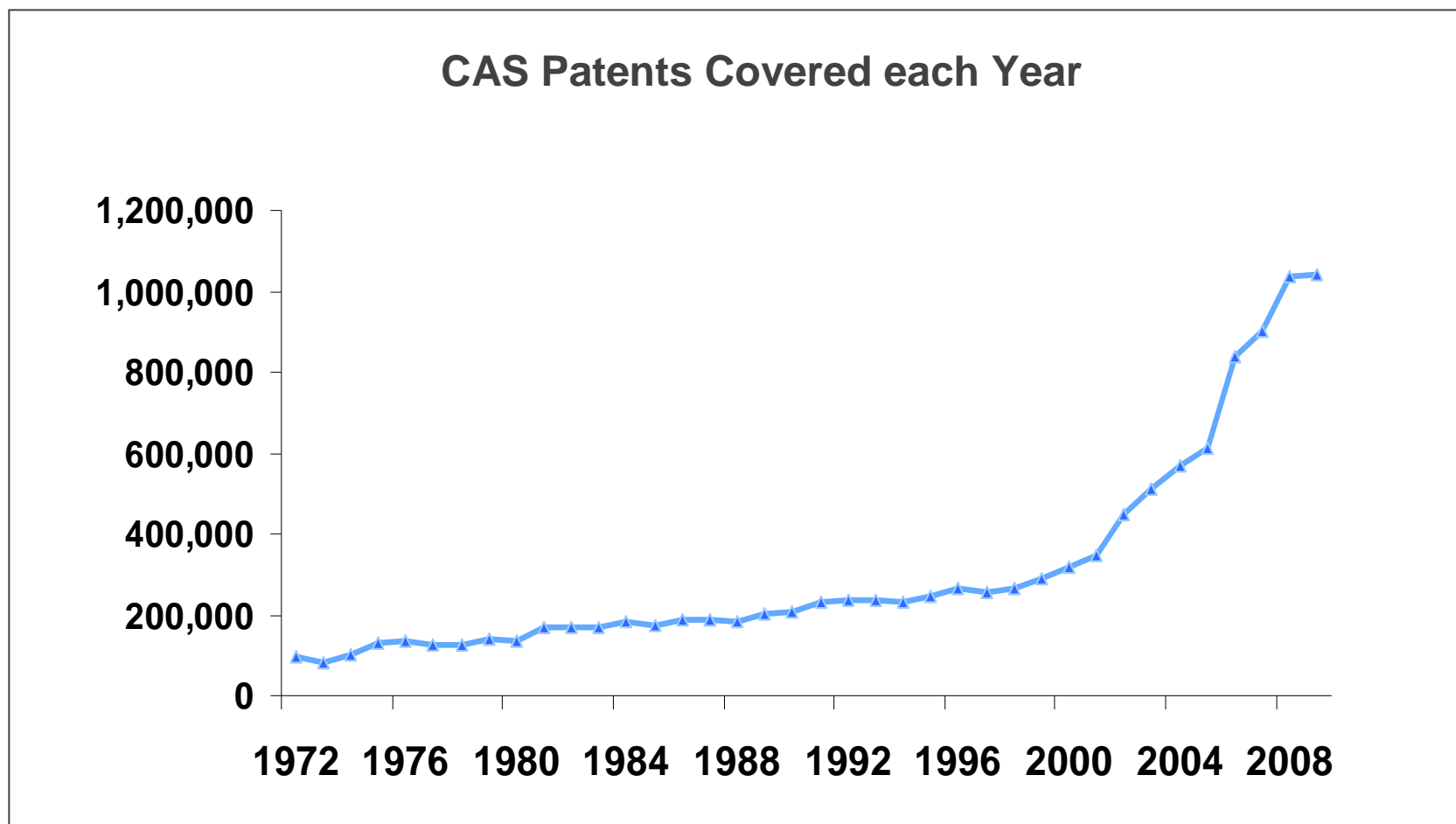
Updated when new or revised catalogs are available

Contact/ordering information including quantity and pricing (when available)

Growth in published chemistry has stayed strong in the last decade



Worldwide patenting of new chemical research continues to accelerate...

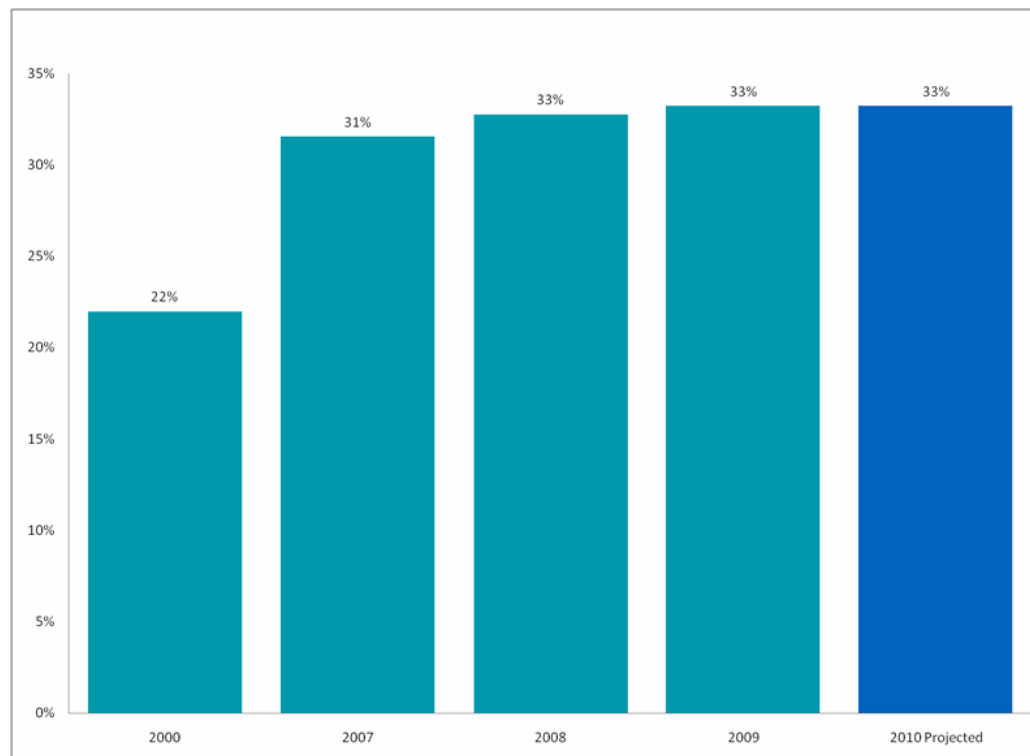


CAS analyzes global chemical information, including publications from Asia

Each year, CAS covers

- 10,000 serial journal titles and 61 patent authorities worldwide
- 2,100 Asian serial journal titles
- All major Asian patent authorities, including offices in
 - People's Republic of China
 - South Korea
 - Japan
 - India

**Chinese, Japanese,
and Korean language
publications account
for 33% of new CAplus
database records**



For complex chemistry, CAS chemists classify substance information and verify graphical processes and structures

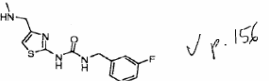
1. Review reaction and structure

WO 2009/015208 PCT/US2008/070893

Alternative process for Intermediate 4 Using Carbonyl Diimidazole:

A stirred mixture of Intermediate 1: 2-amino-4-chloromethyl-thiazole hydrochloride (27.8 g, 0.15 mol), carbonyl diimidazole (25.5 g, 0.157 mol), and anhydrous THF (0.2 L) was treated dropwise with a solution of DIPEA (26.2 mL, 0.15 mol) in THF (20 mL) at 20-30 C. After 2-3 hours stirring, a solution of 3-fluorobenzylamine (18.5 mL, 0.164 mol) in THF (40 mL) was added. The reaction was diluted with water (200 mL) and THF was evaporated under reduced pressure. The residue was extracted with DCM (2 x 200 mL). The combined extracts were dried over sodium sulfate and concentrated to leave an orange resin that was purified by silica gel chromatography (acetone/hexane) to afford Intermediate 4 as a pale yellow solid (26 g, 58% yield). (1039)

Intermediate 5: 1-(3-Fluorobenzyl)-3-(4-((methylamino)methyl)thiazol-2-yl)urea

 (1040)

Prepared by reaction of Intermediate 4 with methylamine, following the procedure described for Intermediate 3.

Alternative Process for Intermediate 5 Using N,O-Dimethylhydroxylamine:

A mixture of Intermediate 4: 2-(3-(3-fluorobenzyl)ureido)-4-chloromethyl-thiazole (40 g, 0.133 mol), N,O-dimethylhydroxylamine (80 g, 0.820 mol), sodium carbonate (40 g, 0.754 mol), and abs. EtOH (0.2 L) was stirred and heated at 60-70 C for 8-12 hours. The mixture was diluted with water (0.8 L) and cooled to 20 C with continued stirring. The



2. Create registration record

IAH	18772840Y	MMN	0480	Proc	2681
REF	98-509999	WKU	01-227156-1480	2009-02-20	1108711-86-9
TD	012198564M	Crav	jxc56	01:49:47	Code 010 / MCG0050

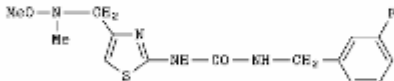
HMNO prepn. of antibacterial amide and sulfonamide substituted heterocyclic

PubNo:

MF: C14H17FN4O2S

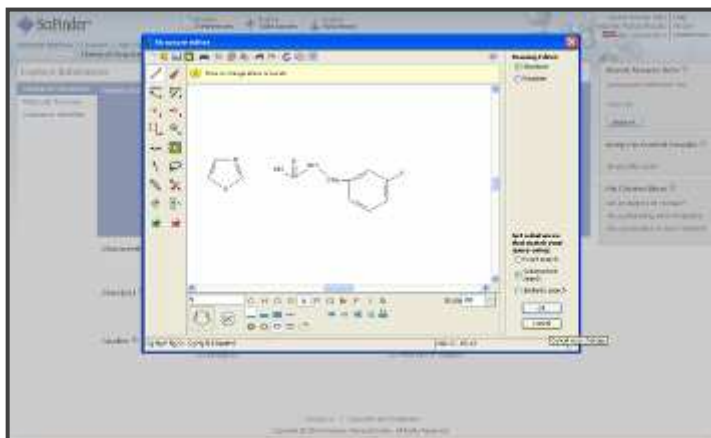
Chem:

C14H17FN4O2S
Stereo: RS



Accurate and timely, CAS substance analysis for CAS REGISTRY ensures scientists can find information

1. Scientist asks structure question



2. Finds 433 compounds and 3 publications

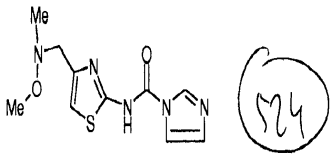
CAS is a division of the American Chemical Society.

3. Finds the PCT patent application

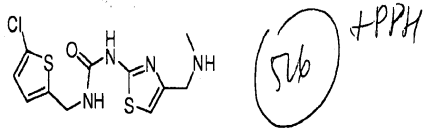
Patent No.	Kind	Date	Application No.	Date
US 200815288	A1	Jul 26, 2008	WO 2008-15288X	Jul 24, 2008
Priority Application				
US 2007063199	A	Jul 23, 2007		
US 2006217287	F	Dec 12, 2006		

CAS chemists interpret when compounds are described in terms other than singular structures or names

Intermediate 28: N-(4-((methoxy(methyl)amino)methyl)thiazol-2-yl)-1H-imidazole-1-carboxamide



Intermediate 29: 1-((5-Chlorothiophen-2-yl)methyl)-3-(4-((methylamino)methyl)thiazol-2-yl)urea



Step 1: Intermediate 28 (imidazolide) was reacted with C-(5-chlorothiophen-2-yl)methylamine using Method 7.

Step 2: The methoxyamine-urea product obtained in Step 1 was reduced with micronized zinc dust in acetic acid, following the procedure for Intermediate 5/Alternative Process/Step 2, to afford Intermediate 29.




TIN 18772844Y RUN 0527 Page 2714
 DAT 98-509999 WJH 81-227156-1527 2009-02-20 1108712-44-2 T
 TD 012198893M Chem jku55 01:49:48 Code 010 / XDD050
 UNAN prepn. of antibiobenzal oxide and sulfonamide substituted heterocyclic

NAME

MF C₁₂H₁₅ClN₄O₂S₂

INX

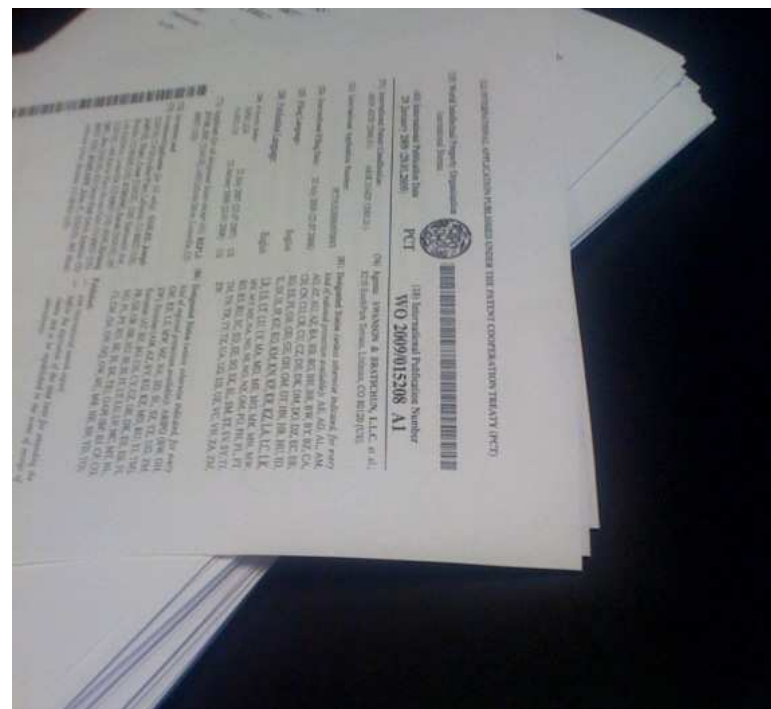
C₁₂H₁₅ClN₄O₂S₂
 Stereo: NS



From published patent to completed indexing in SciFinder®: 15 days

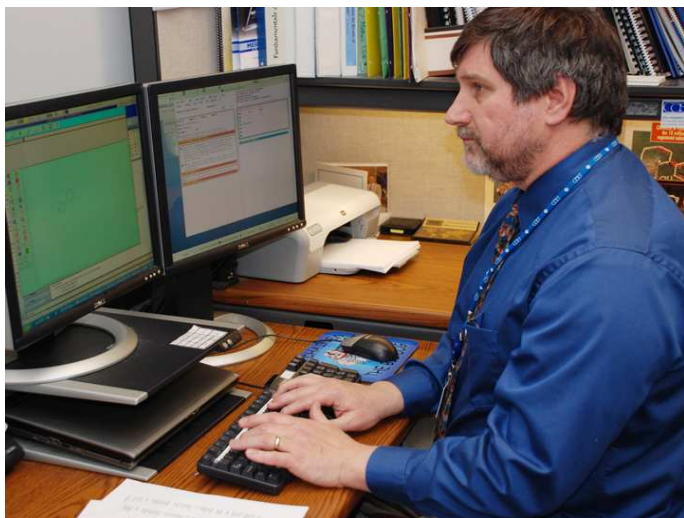
A typical chemistry patent:

- PCT application
- “A new antibacterial”
- 250 pages
- 24 claims



From published patent to completed indexing in SciFinder: 15 days

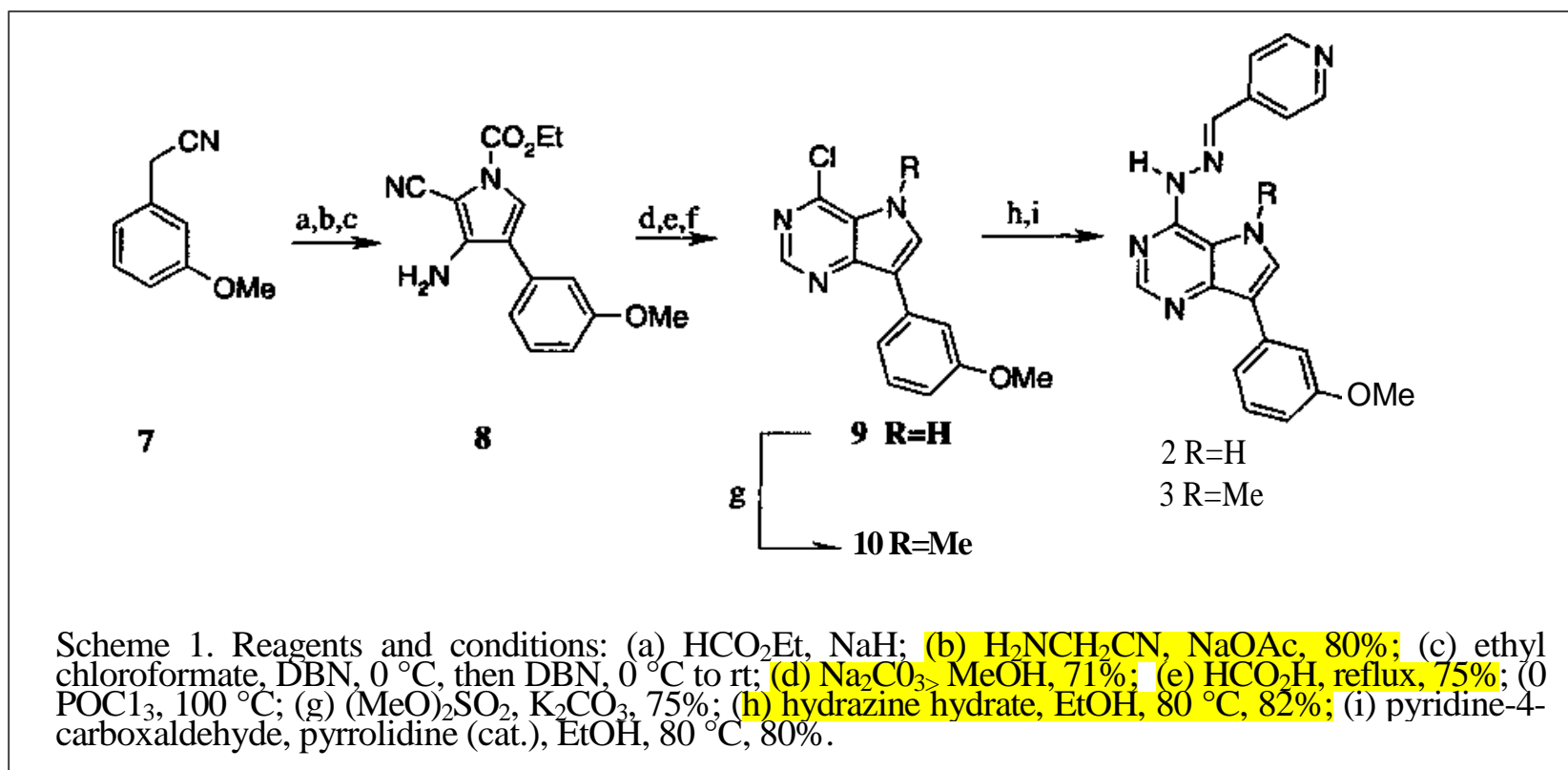
Dr. Mark Willi from CAS completes substance indexing 15 days after WIPO publishes the patent



CAS chemist analysis of single PCT application

- 917 indexed compounds from Examples and Claims
- 576 new compounds added to CAS REGISTRY
- 613 single-step reactions
- 5,394 reactions with multiple steps
- 1,029 reaction participants
- MARPAT[®] Markush structure with 2,119 substituent definitions

Challenges in substance and reaction indexing show the value of intellectual substance identification



Since products of step b, d, e, and h were characterized by their yield, CAS policy demands these are to be registered and indexed for CASREACT

CAS specialists in many fields of chemistry interpret author terminology to register compounds

Mechanism for the hydrolysis of hyaluronan oligosaccharides by bovine testicular hyaluronidase

Ikuko Kakizaki[†], Nobuyuki Ibori[†], Kaoru Kojima, Masanori Yamaguchi, Masahiko Endo

Article first published online: 24 FEB 2010

DOI: 10.1111/j.1742-4658.2010.07600.x

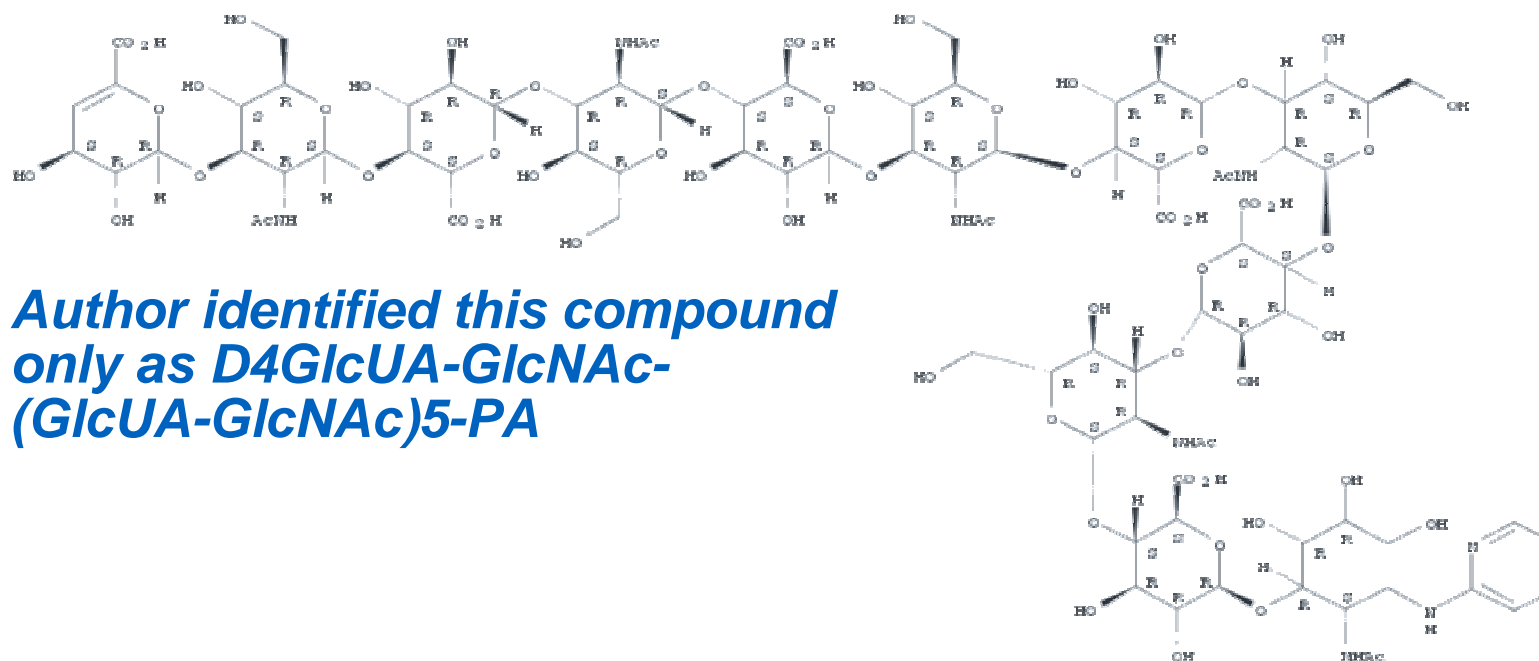
© 2010 The Authors. Journal compilation © 2010 FEBS

Issue



FEBS Journal

Volume 277, Issue 7, pages 1776–1786, April 2010

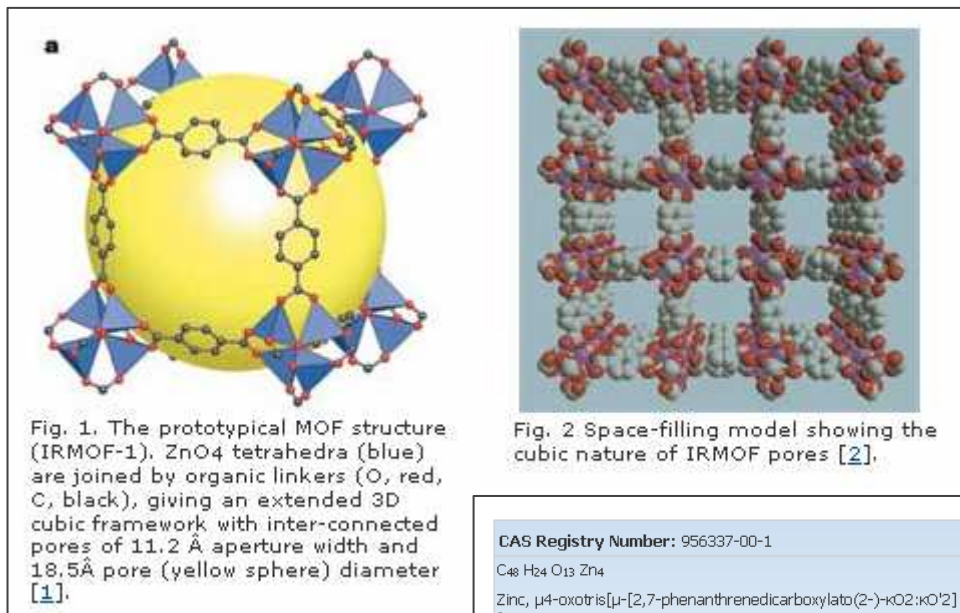


Patents regularly describe substances in ambiguous ways: In WO 2007089907 this “desired product” is fully characterized

Compound 34: Diisopropyl azodicarboxylate (DIAD) (1.20 mL, 6.08 mmol) was added to triphenylphosphine (1.60 g, 6.08 mmol) in THF (100 mL) at 0 °C. and was stirred for half an hour during which time the yellow solution became a paste.

Compound 14 (2.58 g, 4.06 mmol) and *p*-nitrobenzoic acid (0.81 g, 4.87 mmol) were dissolved in THF (50 mL) and added to the paste. The resulted mixture was stirred at ambient temperature overnight. Water (100 mL) was added and the mixture was made slightly basic by adding NaHCO₃ solution followed by extraction with EtOAc (3x50 mL). The combined extracts were washed with brine once and dried over anhydrous Na₂SO₄. The desired product (2.72 g, 85% yield) was obtained as white powder after SiO₂ chromatography (Et₂O/hexanes 1:2). m.p. 207-209 °C.; IR (KBr) 3434, 3056, 2940, 2868, 1722, 1608, 1529, 1489, 1448, 1345 cm⁻¹; ¹H NMR (CDCl₃, 300 MHz) δ 8.30-8.26 (m, 2 H), 8.21-8.16 (m, 2 H), 7.46-7.42 (m, 6 H), 7.31-7.18 (m, 9 H) 5.33 (bs, 1 H), 4.02 (bs, 1 H), 3.90 (bs, 1 H), 3.09-2.97 (m, 2 H), 2.68 (td, J=14.95, 2.56 Hz, 1 H), 2.29-2.19 (m, 1 H), 2.07-1.06 (series of multiplets, 24 H), 1.01 (s, 3 H), 0.98 (d, J=6.6 Hz, 3 H), 0.70 (s, 3 H); ¹³C NMR (CDCl₃, 75 MHz) δ 164.21, 150.56, 144.70, 136.79, 130.77, 128.88, 127.86, 126.98, 123.70, 86.47, 73.24, 73.00, 68.70, 64.22, 47.79, 46.79, 42.15, 39.76, 37.47, 35.52, 35.34, 34.23, 33.79, 32.46, 31.12, 28.74, 27.71, 26.85, 26.30, 25.16, 23.41, 17.98, 12.77; HRFAB-MS (thioglycerol+Na⁺ matrix) m/e: ([M+Na]⁺) 808.4203 (53.8%), calcd. 808.4189.

(Relatively) new substance classes can be registered

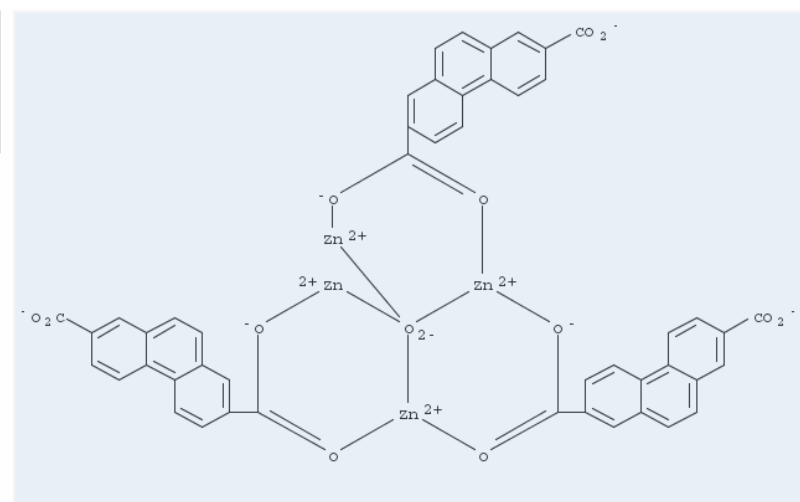


Metal-organic frameworks show great potential for capture of H₂ or CO₂ or in other gas separation processes

CAS Registry Number: 956337-00-1

C₄₈H₂₄O₁₃Zn₄

Zinc, μ₄-oxotris[μ-[2,7-phenanthrenedicarboxylato(2-)-κO2:κO'2]]tetra-
Coordination Compound



Source of Registration: CA

~3 References

By the end of 2010, CAS REGISTRY will contain more than 56 million compounds



On Sept. 7, CAS scientists recorded the 50-millionth chemical substance into the CAS Registry. It received a unique identifier, the CAS Registry Number (CAS RN), and was associated with its authoritative source, in this case a World Intellectual Property Organization application, WO2009/097695, published on Aug. 13, 2009. The substance comes from the examples section of a 199-page patent document and is (5Z)-5-[(5-fluoro-2-hydroxyphenyl)methylene]-2-(4-methyl-1-piperazinyl)-4(5H)-thiazolone, CAS RN 1181081-51-5.

The 50 million publicly disclosed substances represent a consequential milestone. The CAS Registry has been continuously operated for the purposes of uniquely identifying chemical substances since its inception, now more than 40 years ago. Surprisingly, it took CAS only nine months to register the last 10 million substances. In those nine months, CAS has registered at least 25 unique substances per minute.

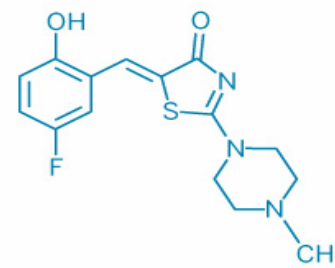
It took 33 years for CAS to encounter the first 10 million substances in the published literature. As CAS forecast two years ago, the pace of discovery in chemistry, especially of small molecules, is increasing. And so, although in 2008 CAS registered a then-record total of 8.5 million substances, that record has already been shattered.

Related Stories

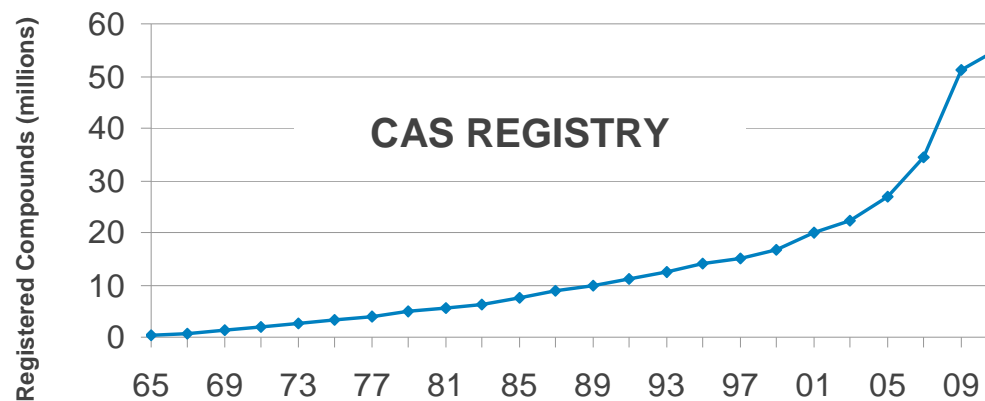
- [CAS Registers 50 Millionth Compound](#)

Topics Covered

[CAS Registry Number](#), [WIPO](#)



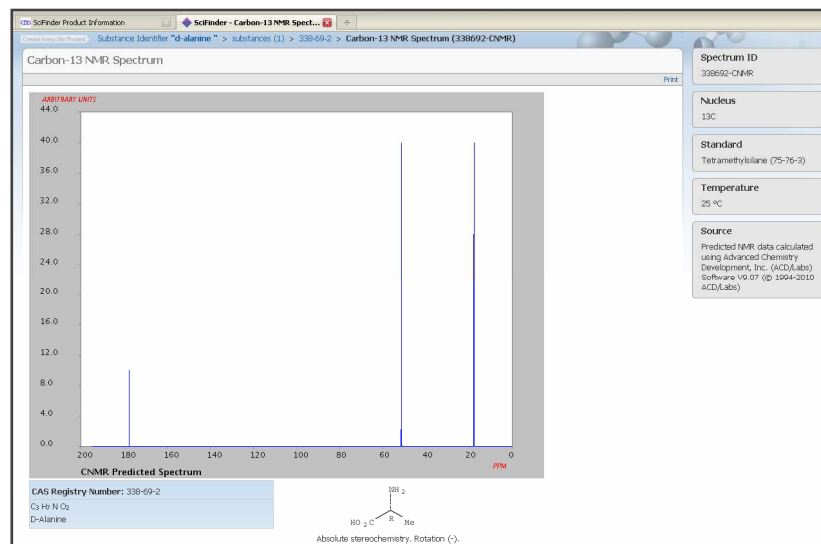
Total Small Molecules



CAS REGISTRY substances are enriched with spectra, numeric properties, tags, and published sources

Spectra

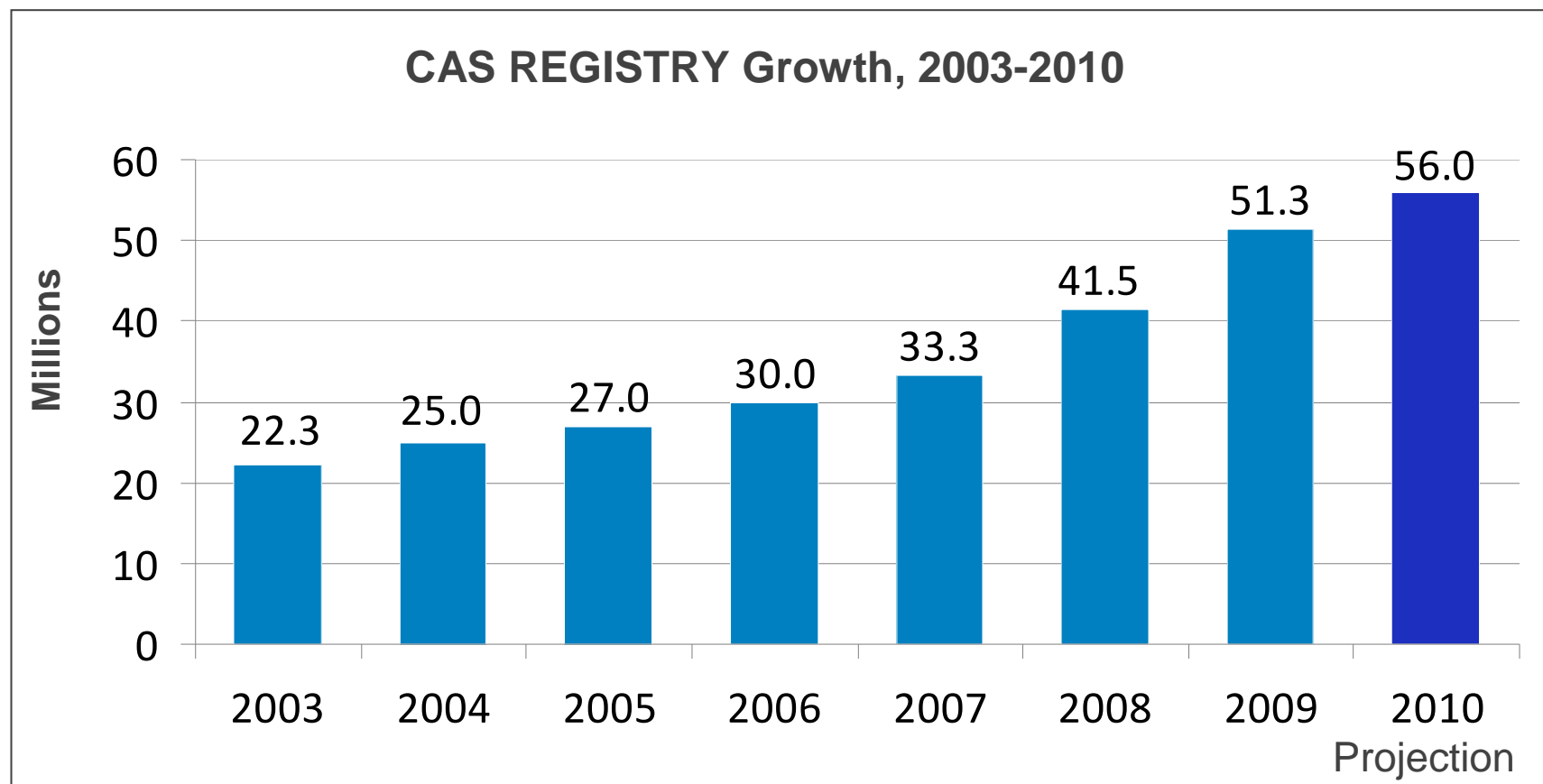
- More than 42M calculated NMR spectra (H1, C13, hetero), with 26M added in 2009
- More than 700,000 experimental spectra (MS, NMR, IR, Raman), with another 300,000 newly acquired MS, IR, and NMR to be added in 2010



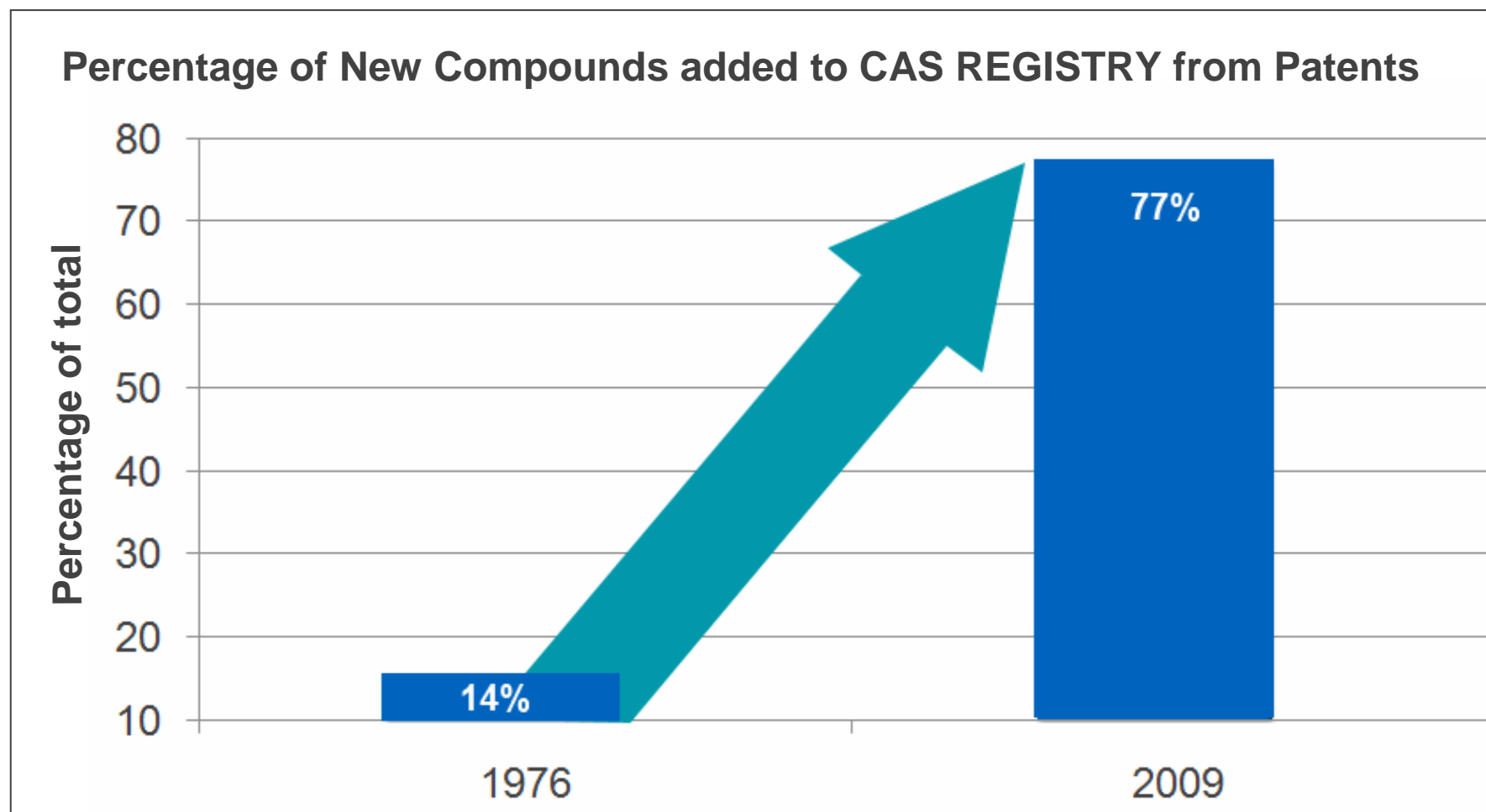
Numeric

- More than 4.5M experimental property values (m.p., b.p., optical rotary power, etc.)
- 7.5M data tags linked to indexed documents
- 2.68B calculated bio-“evaluation” metrics (bio-concentration, Log P, Lipinski, etc.)

CAS continues to register new small molecules in significant numbers



Patenting has increasingly "monetized" new chemical information



*CA Database annual average is 23% patents

CAS has indexed more than 10.1M exemplified prophetic substances from patent documents since 1993

j) 2-[2-Bromo-5-(3-methoxy-propoxy)-phenyl]-ethanol

The mixture of 1 mmol of 4-bromo-3-(2-hydroxy-ethyl)-phenol [319473-28-4] in 5 ml of acetone is stirred with 2 mmol of K_2CO_3 and 1.1 mmol of 1-bromo-3-methoxy-propane [36865-41-5] at reflux temperature over 22 h. The mixture is poured onto ice/ H_2O and extracted with TBME (2x). The combined organic layers are washed with brine, dried over Na_2SO_4 and concentrated in vacuo. Purification by flash chromatography (SiO_2 60F) affords the title compound, which is identified based on the R_f value.

According to the procedures described in example 4, the following compound(s) is(are) prepared in an analogous manner:

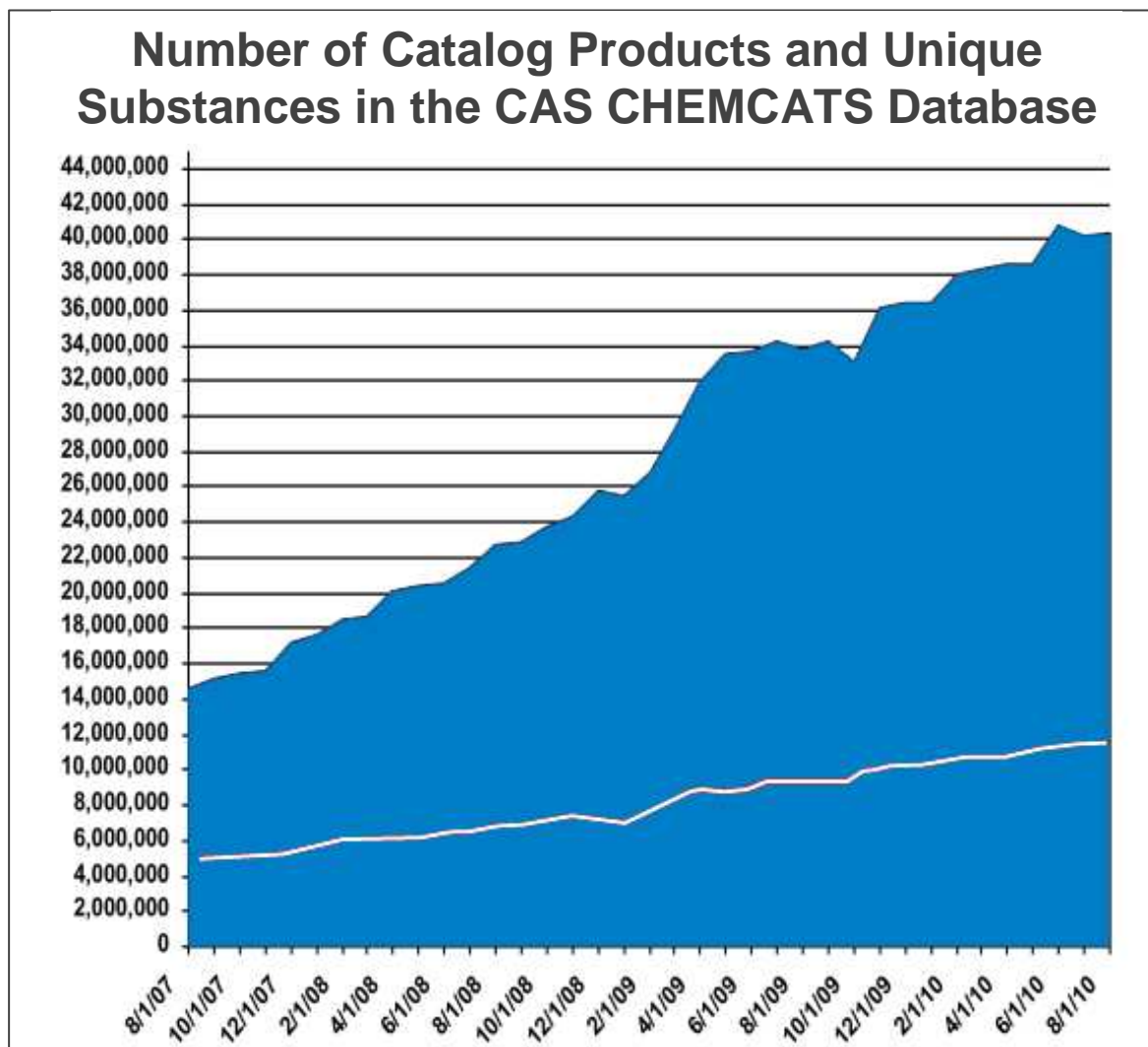
8 (1S,3'S)-5-(3-Methoxypropoxy)-3'-{[4-(3-methoxypropyl)-3,4-dihydro-2H-1,4-benzoxazin-6-yl]methoxy}-3H-spiro[2-benzofuran-1,4'-piperidine]

using 4-bromo-3-hydroxymethyl-phenol [2737-20-4] instead of 4-bromo-3-(2-hydroxy-ethyl)-phenol [319473-28-4] in step j.

12 (1S,3'S)-7-(3-Methoxypropoxy)-3'-{[4-(3-methoxypropyl)-3,4-dihydro-2H-1,4-benzoxazin-6-yl]methoxy}-4,5-dihydro-3H-spiro[2-benzoxepine-1,4'-piperidine]

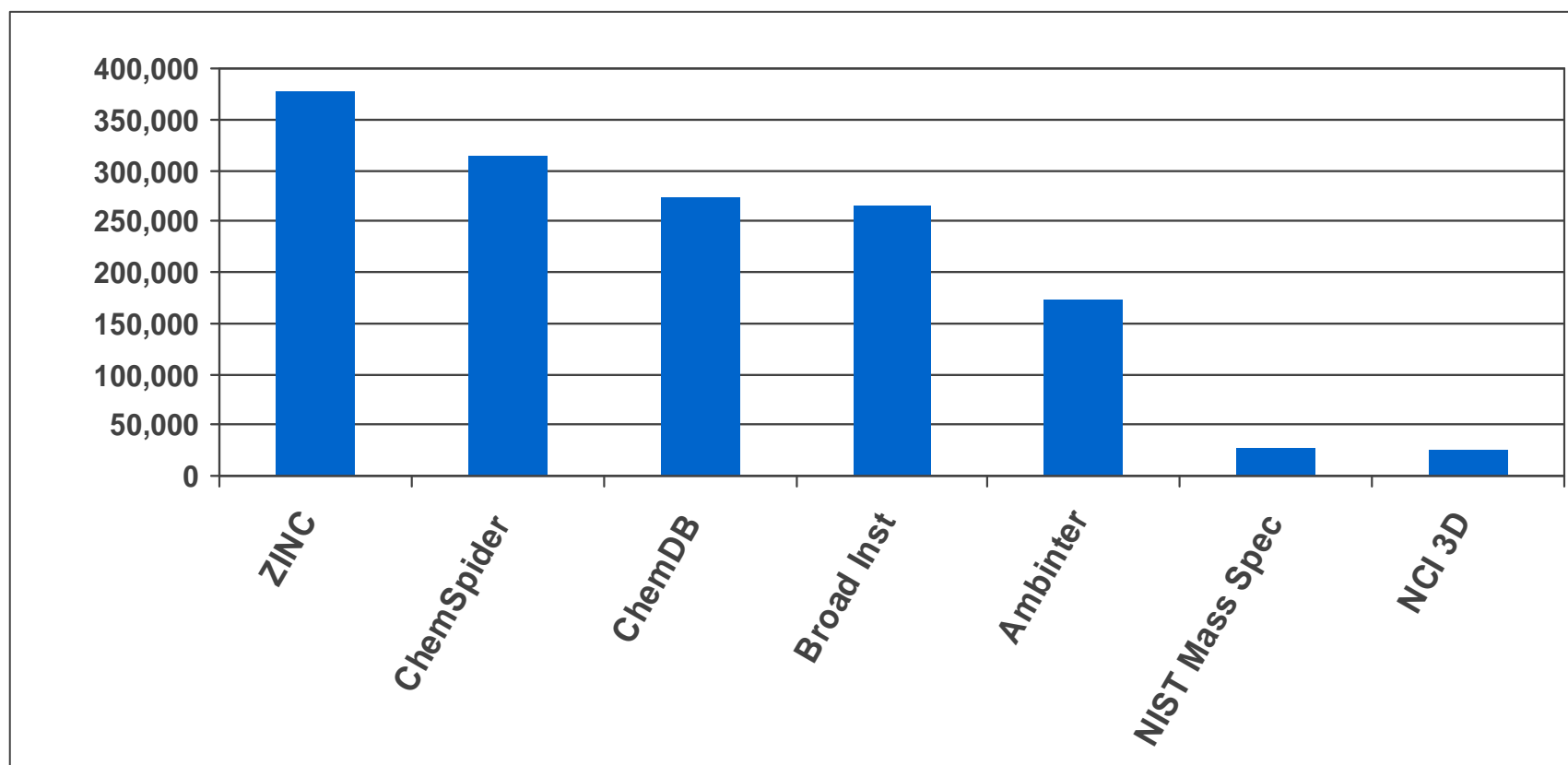
using 4-bromo-3-(3-hydroxy-propyl)-phenol instead of 4-bromo-3-(2-hydroxy-ethyl)-phenol [319473-28-4] in step j.

CHEMCATS continues to grow and remains a source of new small molecules



Chemical substances from web-based sources provide a moderate addition to the small molecules in REGISTRY

- 1.6M substances have been captured from Internet substance collections (They must be from a reputable source and contain characterizing data)
- The following chart illustrates some of the larger collections:



Conclusions: What the data reveal about small molecules

- Information professionals and scientists from around the world rely on CAS to provide the most authoritative, current, and complete collection of substances
- Substances are mainly from the journals and patents covered by CAS databases and indexed by CAS scientists
- Intellectual processing of substance identification provides superior content
- Unique substances are found in chemical catalogs and chemical libraries as long as there is proof they exist
- Internet sources provide some complementary pointers to disclosed substance information