

SureChEMBL@10

The challenging child with great prospects
for the future

Nicolas Bosc

SureChEMBL team



What is SureChEMBL

- Database of annotated patents
- First SureChem and developed by Digital Science Ltd.
- Maintained and kept as it was by EMBL-EBI from 2013
- New system introduced in 2023

167M
Patents



28M
Chemically
-annotated
documents



28.5M
Chemicals



*Biomedical
annotations
generated on
the fly*



Why is searching chemical patents useful?

- Freedom to operate
- Competitive intelligence
- State-of-the-art
- Citations and key references
- Most of the knowledge in chemical patents will never appear anywhere else
 - Compounds, scaffolds, reactions
 - Biological target, disease, indication relationships
 - Average time lag between patent and journal: 3 years

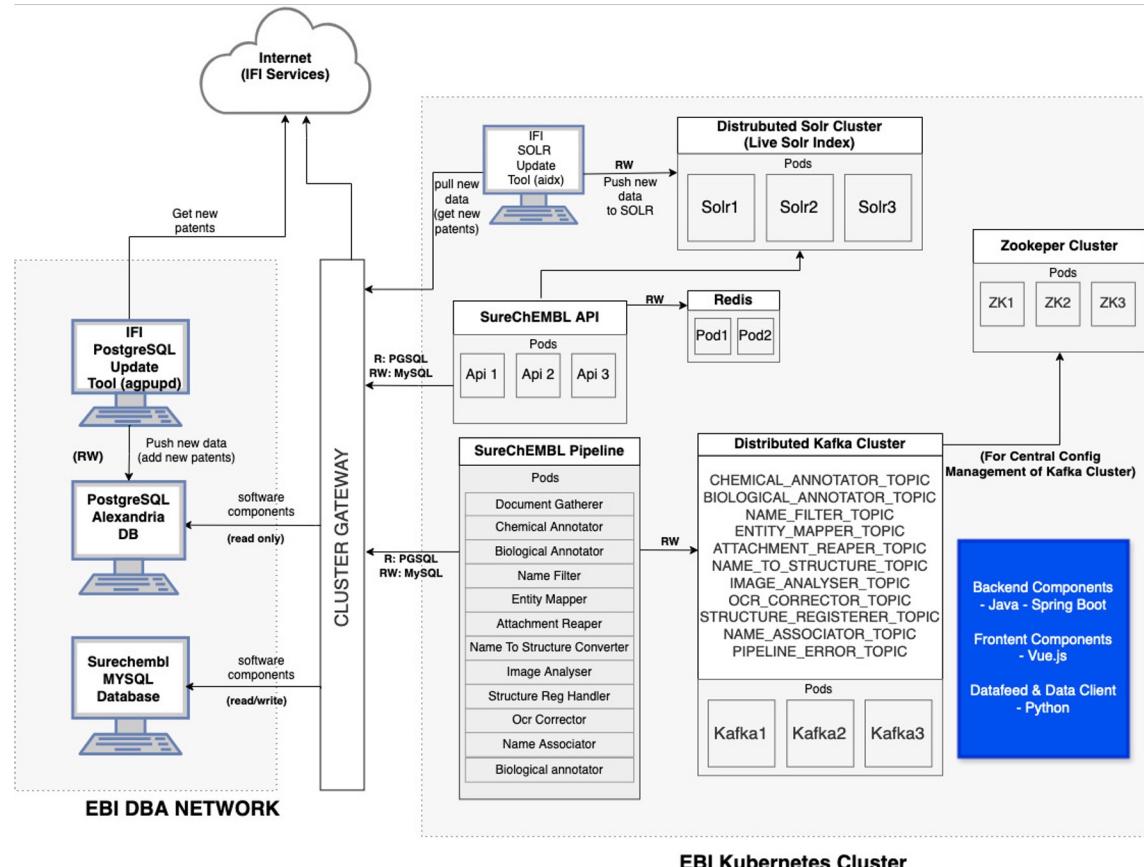
SureChEMBL: a challenging child

- Inherited system, came with its own problems...
 - Monolithic system
 - Backend: Ruby, JRuby and Perl
 - Frontend: Ruby on Rails/Jquery
 - API: Ruby on Rails
 - Old dependencies
 - Pipeline and API not fitted for big data due-to monolithic architecture
- Hard to maintain and implement new features
- Complete refactoring was required

New system architecture

Easier to develop and deliver new functionalities

- Scalable Kubernetes Microservice Architecture
- Solr Index
- Modern UI (Aligned with EBI Standards)
- Public API

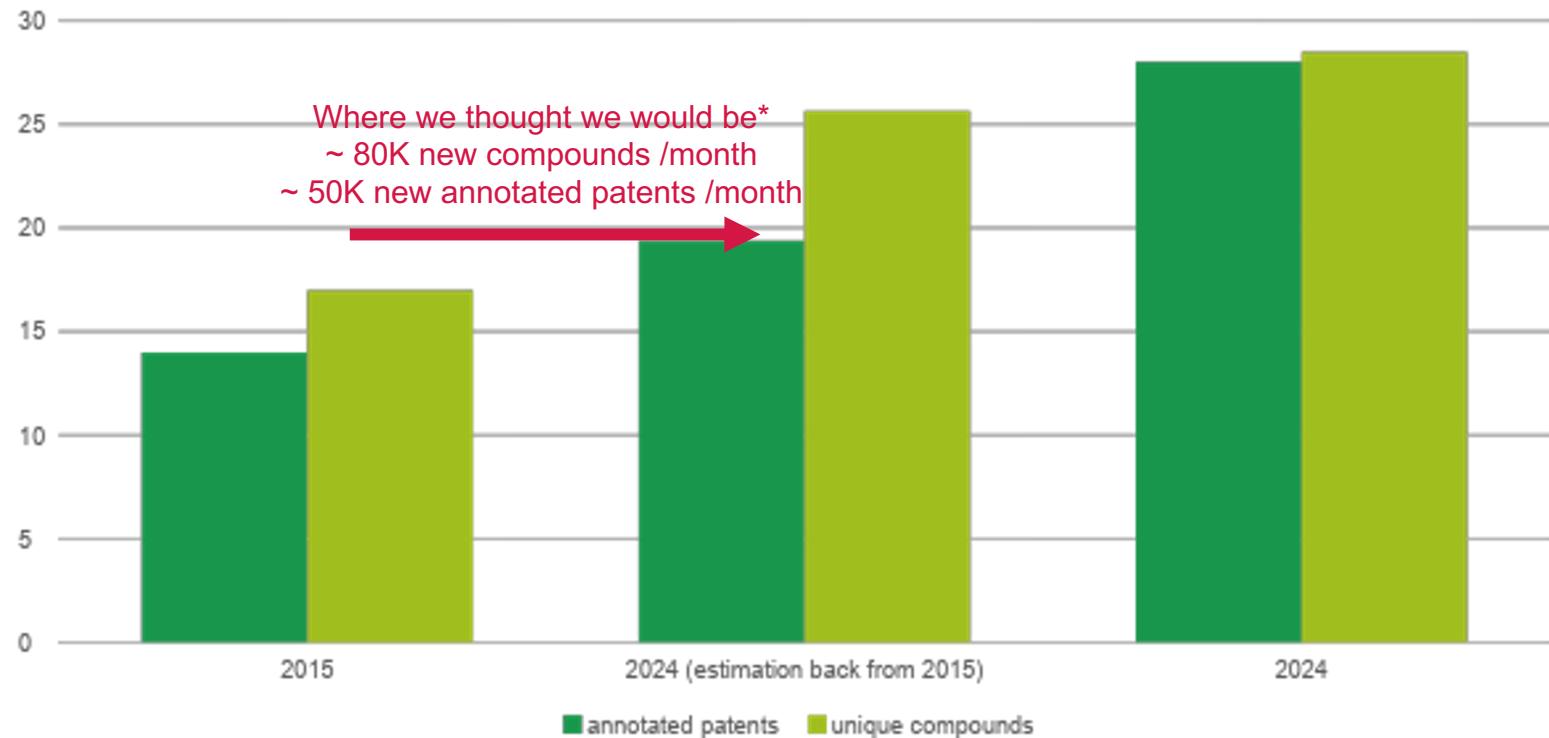


Patent coverage

Authorities	Kind	Language	From	Full text	Biblio. data	Attachments	Annotated in SureChEMBL	
CNIPA*	Applications	EN	1985	Yes (English translation)	Yes	No	2-7 days after receipt	
	Granted							
EPO	Applications	DE, EN, FR	1978	Yes	Yes	Yes	2-7 days after receipt	
	Granted		1980					
JPO	Applications	EN	1976	Yes (abstract)	Yes	No		
USPTO	Applications	EN	2001	Yes	Yes	Yes	2-7 days after receipt	
	Granted		1920-1949	Yes (abstract)	Yes	Yes (PDF)		
			1950-1975	Yes (abstract & claims)	Yes	Yes (PDF)		
			1976	Yes	Yes	Yes		
WIPO	Applications	EN, FR	1978	Yes	Yes	Yes		

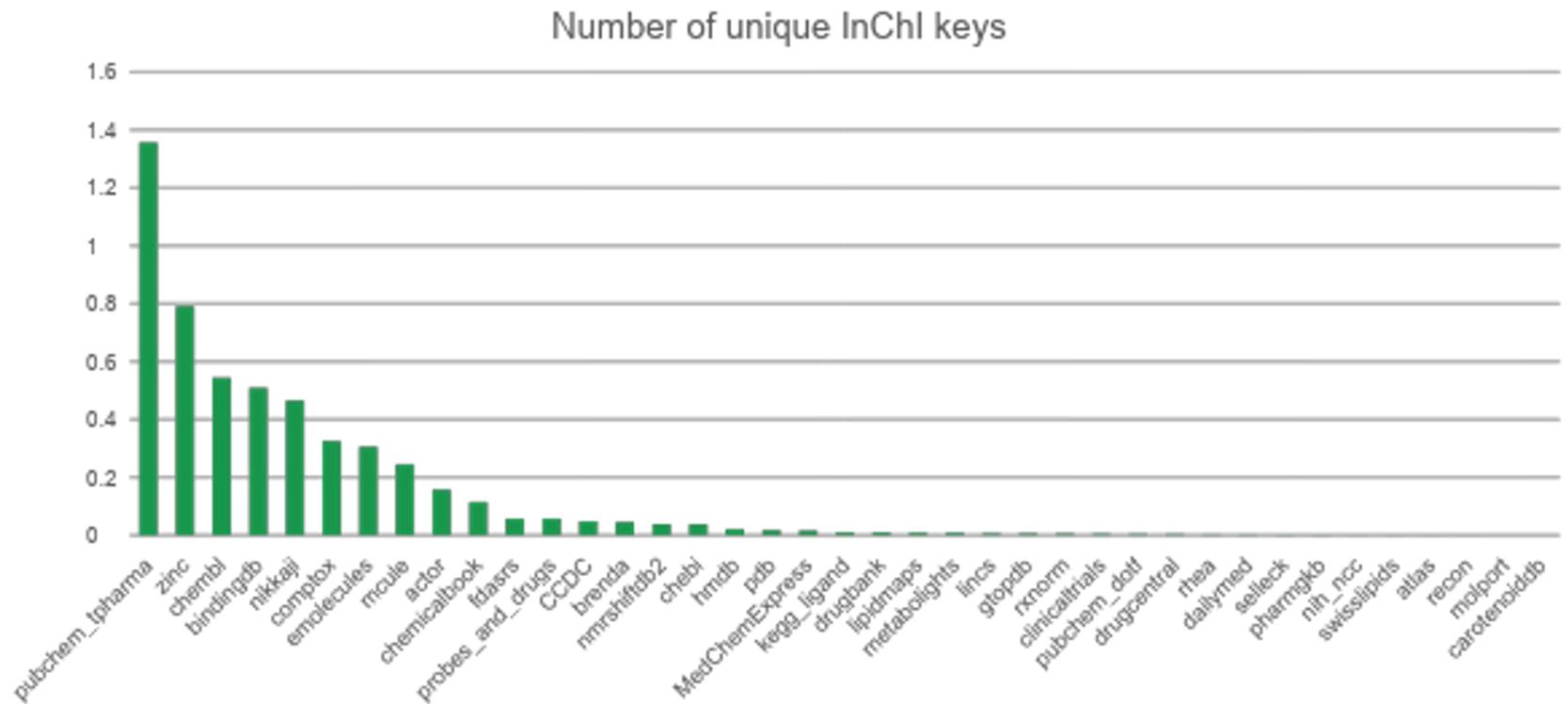
Data quantity evolution

Evolution of the numbers of annotated patents and unique compounds in SureChEMBL



*SureChEMBL paper - NAR, 2016, Vol. 44, D1220–D1228

SureChEMBL chemical space vs others



*excluding PubChem compound set

Compound identification

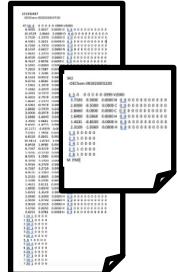
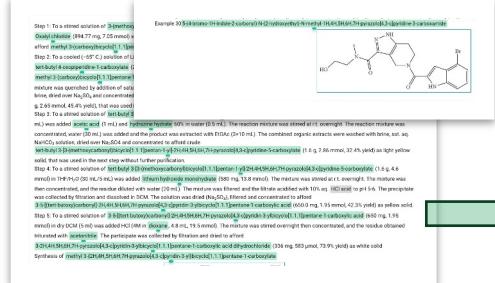
Fully automated pipeline



Patent data feed
(IFI Claims)



Mol files
(US only)



text to structure
(5 methods)

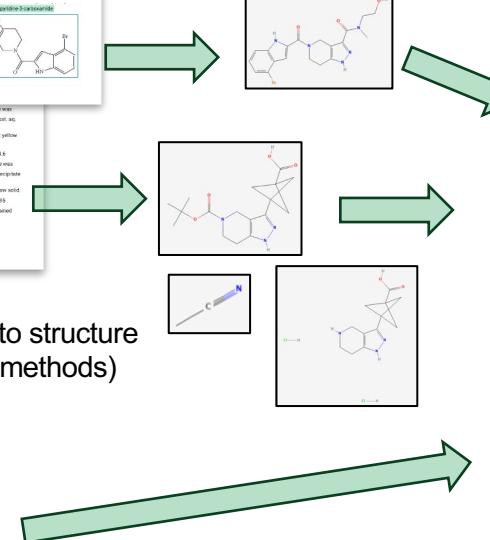
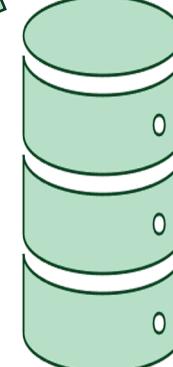
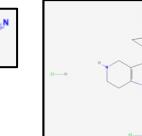
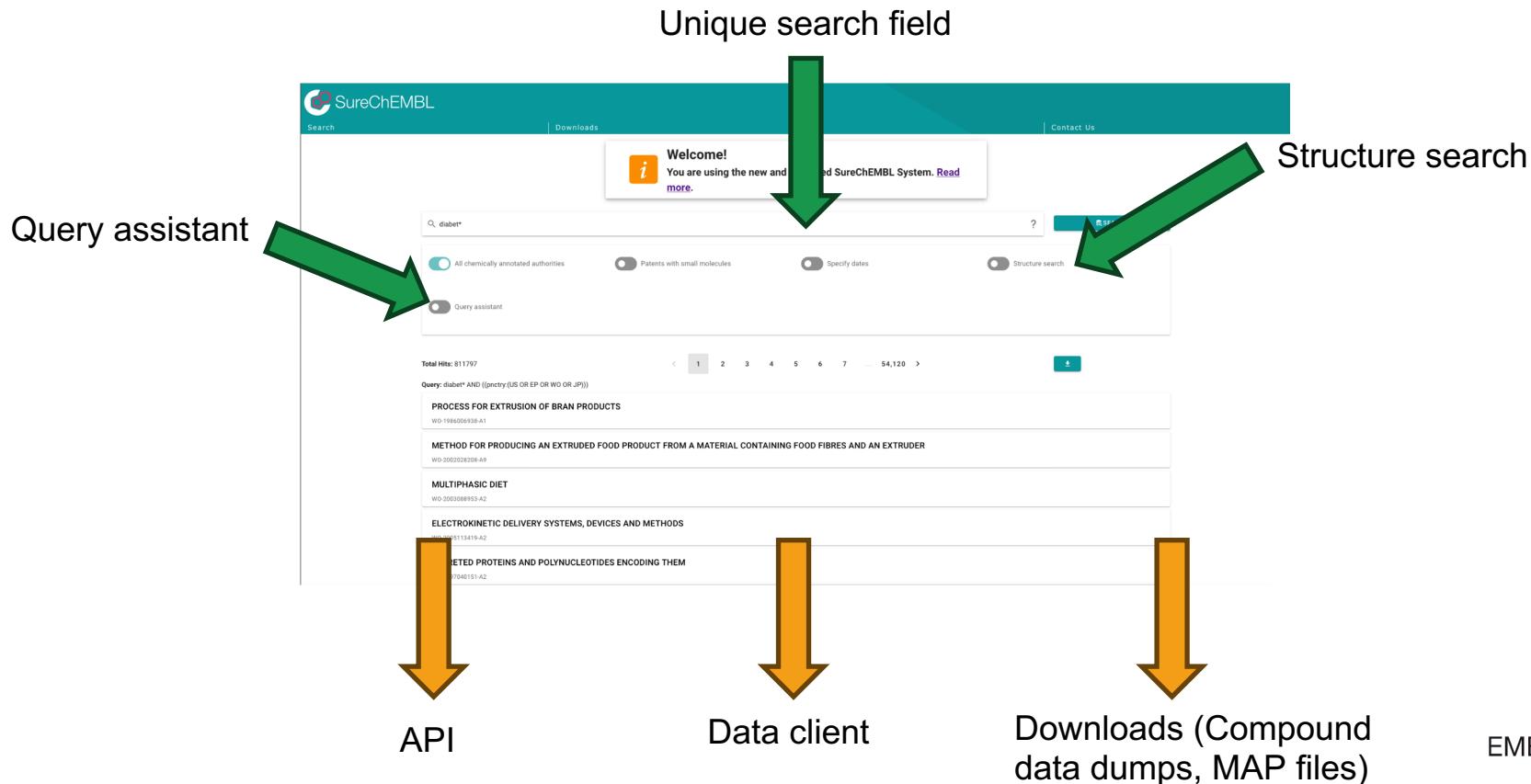


image to structure
(3 methods)



Chemical
registration

How to access SureChEMBL data





Welcome!

You are using the new and improved SureChEMBL System. Read [more](#).

 diabet*[?](#)

SEARCH

 All chemically annotated authorities Patents with small molecules Specify dates Structure search Query assistant

Total Hits: 811797

[<](#) [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) ... [54,120](#) [>](#)

Query: diabet* AND ((pntry:(US OR EP OR WO OR JP)))

PROCESS FOR EXTRUSION OF BRAN PRODUCTS

WO-1986006938-A1

METHOD FOR PRODUCING AN EXTRUDED FOOD PRODUCT FROM A MATERIAL CONTAINING FOOD FIBRES AND AN EXTRUDER

WO-2002028208-A9

MULTIPHASIC DIET

WO-2003088953-A2

ELECTROKINETIC DELIVERY SYSTEMS, DEVICES AND METHODS

WO-2005113419-A2

SECRETED PROTEINS AND POLYNUCLEOTIDES ENCODING THEM

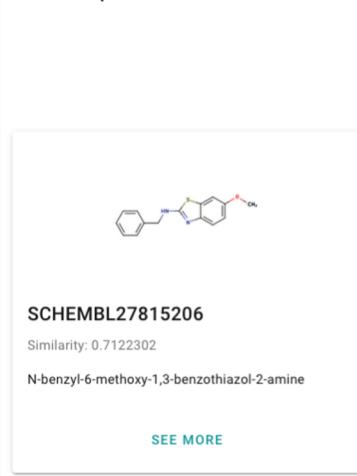
WO-1997040151-A2

User interface

Structure search

- Substructure
- Similarity
- Identical
- Connectivity

Found 2 compounds:



Welcome!
You are using the new and improved SureChEMBL System. [Read more.](#)

Not available for structure search

SEARCH

All chemically annotated authorities

Biologically relevant

Specify dates

Query assistant

Structure search

Search Type

Substructure

Similarity

Identical

Compound Details

SCHEMBL4917823

N-(6-methoxy-1,3-benzothiazol-2-yl)-4-phenoxybenzamide

SMILES: CC(=O)C1=C(C=C1)N=C(NC(=O)C1=CC=C(C=C1)C=C1)S(=O)(=O)c1c2ccccc2c3cc(O)cc4c3cc(O)cc2c1

InChI: InChI-1S/C21H16N2O3S/c1-25-17-11-18-19(13-17)27-21(22-18)23-20(24)14-7-9-16(10-8-14)26-15-5-3-2-4-6-15/h2-13H,1H3,(H,22,23,24)

InChI Key: UAUOBVHZLHMXQC-UHFFFAOYSA-N

Log P: 5.23

Mol Weight: 376.43

UniChem Cross References

Patents for compound

Total patents found: 4

Benzothiazole compositions and their use as ubiquitin ligase inhibitors
US-20050130974-A1

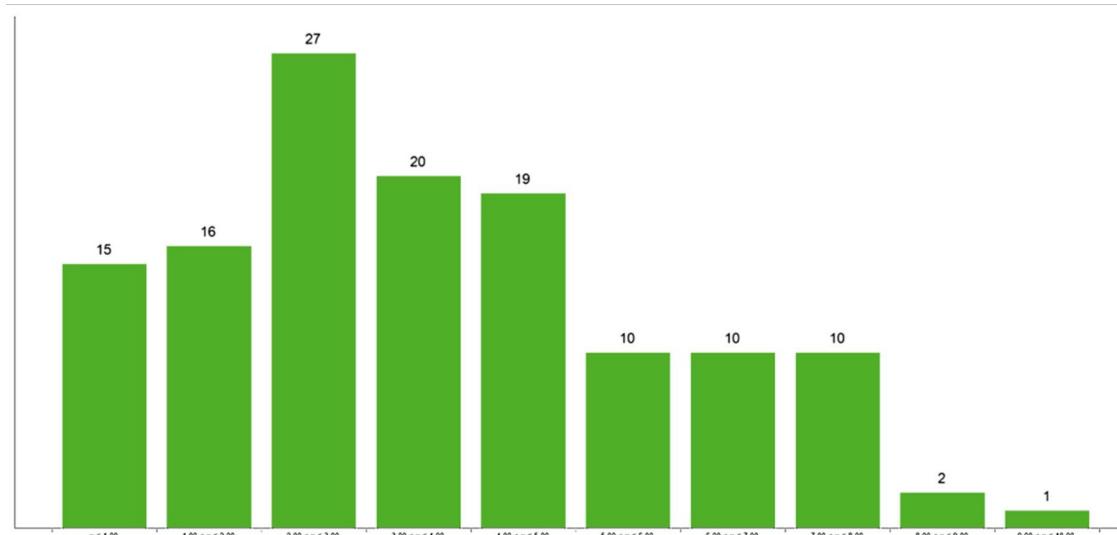
Benzothiazole compositions and their use as ubiquitin ligation inhibitors
US-20080039629-A1

Max 800

EBI

Biomedical annotation

Biological target – disease – compound relationships often mentioned first in patent literature



Binned delay between publication in the scientific literature after appearing in a patent in years
For each bin the number of compound-target interactions pairs is given

New biomedical annotation in SureChEMBL

- In house annotations powered by NLP
- Model trained on published Gold Standard
 - 200 fully manually annotated patents
 - 4 annotation types:
 - target
 - disease
 - mode of action
 - species (in-house)

PLOS ONE

OPEN ACCESS

PEER-REVIEWED

RESEARCH ARTICLE

Annotated Chemical Patent Corpus: A Gold Standard for Text Mining

Saber A. Akhondi, Alexander G. Klenner, Christian Tyrchan, Anil K. Manchala, Kiran Boppana, Daniel Lowe, Marc Zimmermann, Sarma A. R. P. Jagarlapudi, Roger Sayle, Jan A. Kors, Sorel Muresan

Published: September 30, 2014 • <https://doi.org/10.1371/journal.pone.0107477>

- Retrained bioformer-8L model (lightweight BERT)
- Already available
- Not yet in the downloads
 - Ontology linking – in development



	precision	recall	F1
Disease	0.81	0.84	0.82
Mode of action	0.70	0.83	0.76
Target	0.82	0.86	0.84
Organism	0.97	0.96	0.97
Total	0.81	0.85	0.83

New biomedical annotation in SureChEMBL

SureChEMBL

Search | Downloads | Wiki | Contact Us

DOCUMENT ANNOTATIONS BIBLIOGRAPHIC PDF

Description

The invention relates to the use of substituted oxypyridine derivatives for the treatment and/or prophylaxis of thrombotic or **thromboembolic disorders** and/or thrombotic or **thromboembolic complications**. Haemostasis is a protective mechanism of the organism, which helps to "seal" leaking damages in the blood vessel wall quickly and reliably. Thus, excessive loss of blood can often be avoided or kept to a minimum. After injury of a blood vessel, hemostasis is conducted mainly by activation and aggregation of platelets and activation of the coagulation system, which consists of an enzymatic "waterfall" cascade leading one after another to the activation of the next coagulation factor until **thrombin** is formed, which leads to the generation of insoluble fibrin, which is an important part of the clot.

In the more recent past, the traditional theory of two separate starting points of the coagulation cascade (extrinsic and intrinsic path) has been modified owing to new findings: In these models, coagulation is initiated by binding of activated factor VIIa to **tissue factor (TF)**. The resulting complex activates **factor X**, which in turn leads to generation of **thrombin** with subsequent production of fibrin and platelet activation (via **PAR-1**) as injury-sealing end products of **haemostasis**. Compared to the subsequent amplification/propagation phase, the **thrombin** production rate in this first phase is low and as a result of the occurrence of **TFPI** as inhibitor of the **TF-FVIIa-FX** complex is limited in time. A central component of the transition from initiation to amplification of coagulation and thereby thrombus propagation is **factor XIa**, in positive feedback loops, thrombin activates not only factor V and factor **VIII**, but also factor **XI** to **factor XIa**, which in turn converts **factor IX** into **factor IXa**, **factor XIa** and finally to large amounts of **thrombin**, resulting in strong thrombus growth and stabilization of the thrombus. This is supported by TAFI inhibition of clot lysis and further clot stabilisation.

In addition to the stimulation via tissue factor, the coagulation system can be activated particularly on negatively charged surfaces, which include not artificial surfaces such as vascular prostheses, stents and extracorporeal circulation. On these surfaces, **factor XII (FXII)** is activated to **factor XIIa**, which in turn activates **prothrombin** to **thrombin**, overall resulting in amplification of the initiation of this intrinsic part of the coagulation cascade.

Uncontrolled activation of the coagulation system or defective inhibition of the activation processes may lead to the formation of local thrombi or emphysema (e.g. cardiac atrium). In addition, systemic hypercoagulability may lead to system-wide formation of **microthrombi** and finally to a consumptive coagulopathy. Thromboembolic complications may also occur in extracorporeal circulatory systems, such as haemodialysis, and also in vascular prostheses. In the course of many cardiovascular and metabolic disorders, increased tendency for coagulation and platelet activation occur owing to either systemic **infection** or **smoking**, or to changes in blood flow with stasis, for example in diseased leg veins or in **atrial fibrillation**, or owing to pathological changes such as **atherosclerosis**. This unwanted and excessive activation of coagulation may, by formation of fibrin- and platelet-rich thrombi, lead to **thromboembolic threatening events**. **Inflammation** processes may also be involved by triggering the coagulation system. On the other hand, **thrombin** is known to act as a physiological anticoagulant. Accordingly, **thromboembolic disorders** are still the most frequent cause of morbidity and mortality in most industrialized countries.

The anticoagulants known from the prior art, that is to say substances for inhibiting or preventing blood coagulation, have various disadvantages. Accordingly, prophylaxis of thrombotic or **thromboembolic disorders** is found to be difficult and unsatisfactory.

In the therapy and prophylaxis of **thromboembolic disorders**, use is made, firstly, of heparin which is administered parenterally or subcutaneously. Bemiparin is these days increasingly given to low-molecular-weight heparin; however, the known disadvantages described herein below encountered in heparin therapy are often in effective and have to be taken into account.

Lang EN

SureChEMBL Search Downloads Wiki Contact Us DOCUMENT ANNOTATIONS BIBLIOGRAPHIC PDF

Abstracts

EN	Name	Category
EN	thromboembolic disorders	blue circle
EN	thrombotic	green circle
EN	thromboembolic complications	blue circle

EN	Name	Category
EN	thromboembolic disorders	blue circle
EN	thrombotic	green circle
EN	thromboembolic complications	blue circle

Claims

EN	Name	Category
EN	reocclusion	blue circle
EN	transitory	green circle
EN	ischemic attack	blue circle
EN	cryptogenic stroke	green circle

Compound structure quality improvement

- SureChEMBL pipeline automatically converts images into structures
 - CLiDE (no longer used)
 - OSRA
 - Imago
- No manual curation or verification
 - Error prone

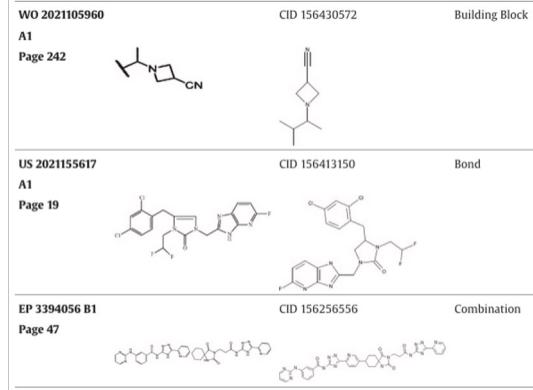
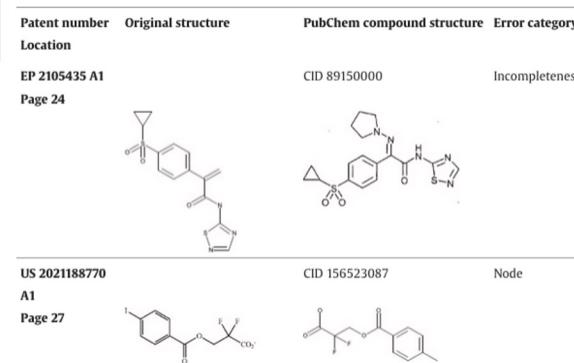


World Patent Information
Volume 70, September 2022, 102134



Validity of PubChem compounds supplied
by Patentscope or SureChEMBL

[Joerg Ohms](#)



Compound structure quality improvement

New deep neural network models introduced recently

[Home](#) > [Journal of Cheminformatics](#) > Article

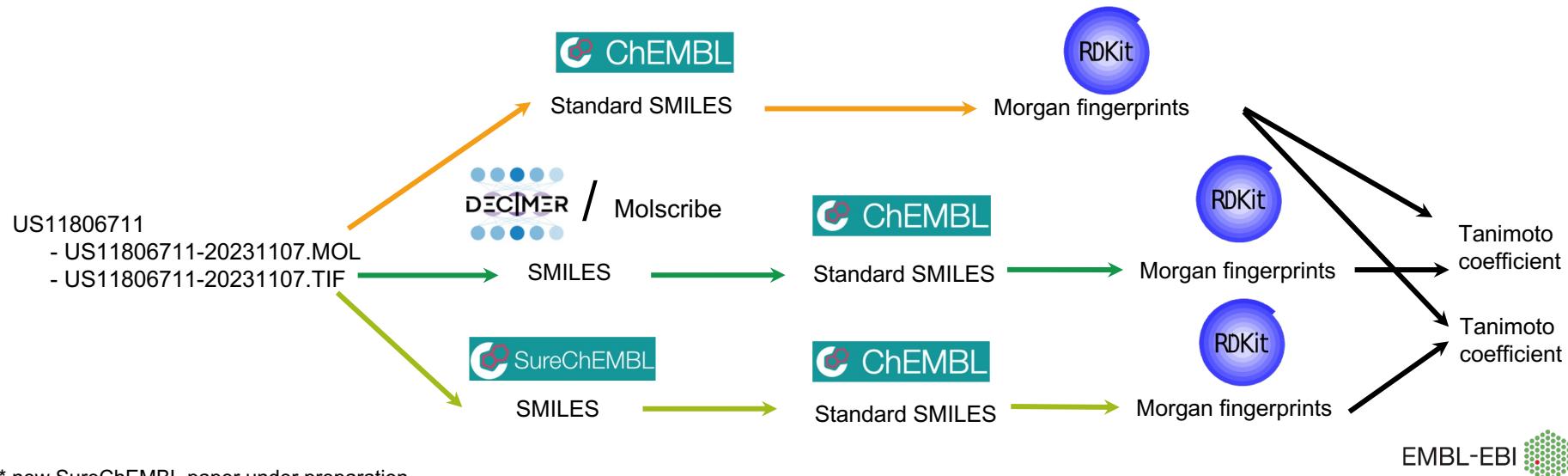
DECIMER: towards deep learning for chemical image recognition

Table 2. Molecule Structure Recognition Accuracy on Synthetic, Realistic, and Perturbed Benchmarks

		Synthetic			Realistic					
Models		Indigo	ChemDraw	CLEF	JPO	UOB	USPTO	Staker	ACS	
Rule-based	MolVec	95.4	87.9	82.8	67.8	80.6	88.4	0.8	47.4	
	OSRA	95.0	87.3	84.6	55.3	78.5	87.4	0.0	55.3	
Machine learning-based	Img2Mol ^c	58.9	46.4	18.3	16.4	68.7	26.3	17.0	23.0	
	DECIMER	69.6	86.1	62.7	55.2	88.2	41.1	40.8	46.5	
Graph Generation	SwinOCSR ^d	74.0	79.6	30.0	13.8	44.9	27.9	-	27.5	
	MSE-DUDL ^b	-	-	-	-	-	-	77.0	-	
	ChemGrapher ^b	-	-	-	-	70.6	-	-	-	
	Image2Graph ^b	-	-	51.7	50.3	82.9	55.1	-	-	
	Ours	Baseline	94.1	92.2	87.4	74.8	88.2	91.5	86.1	59.8
		MolScribe	97.5	93.8	88.9	76.2	87.9	92.6	86.9	71.9

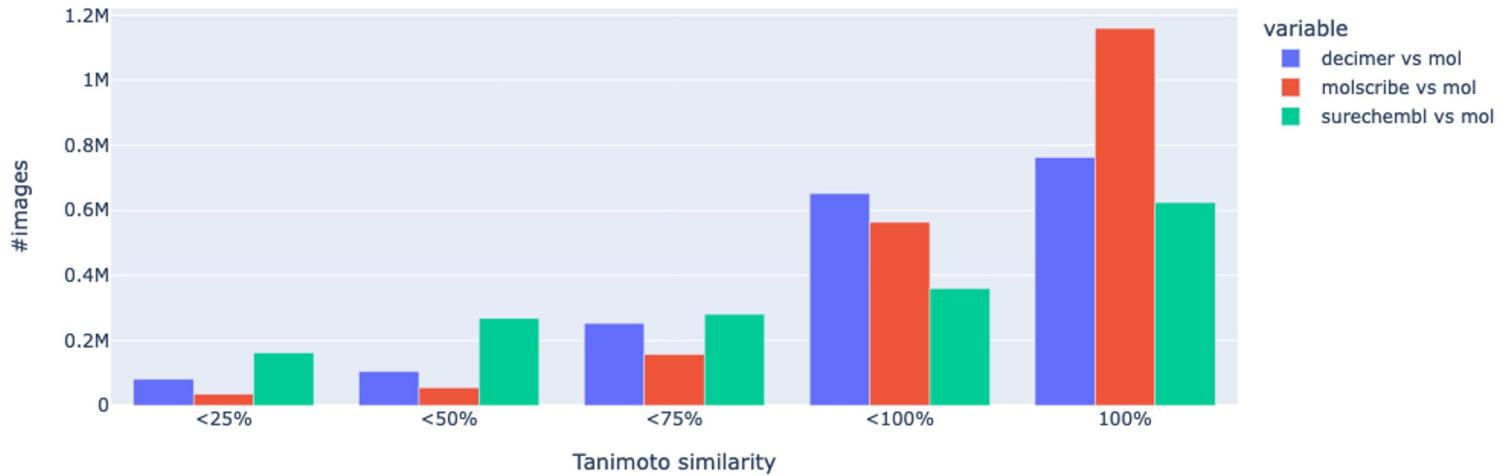
Compound structure quality improvement

- How does it translate in practice?
- Protocol*
 - 2023 USPTO patents with MOL files
 - Both MOL and TIF files for the same compound have to be available



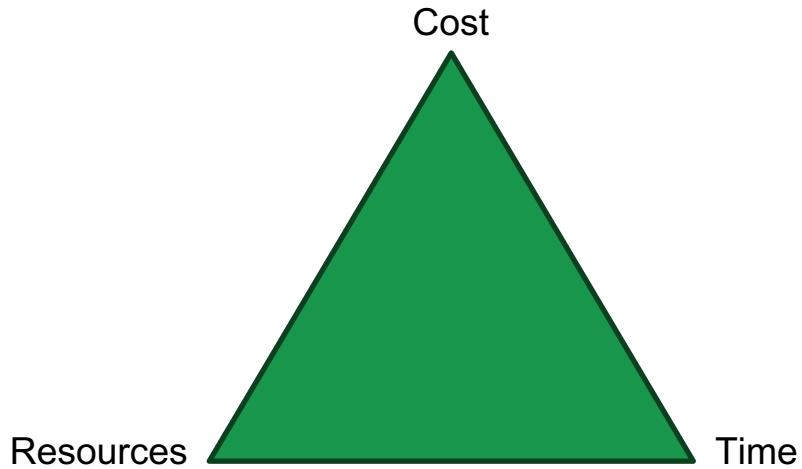
Compound structure quality improvement

- All tools (including the ones currently used by SureChEMBL) return a majority of exact match
- Molscribe converts correctly significantly more images (and is faster)



Incoming improvements

- New image2structure protocol
- Chinese patents
- RDKit integration
- Biomedical annotation downloads
- Core Chemical Structure Integration
- Metadata in the MAP files
- API improvement



Great prospects for the future

- Robust and customizable system infrastructure
- Efficient and modern UI for improving the user experience
- More accurate chemical and biological annotation in a timely manner
- Data available from various ways
- Continuous development bringing regular fixes and/or functionalities

Acknowledgments

- SureChEMBL team
 - Tevfik Kiziloren
 - Ricardo Arcila
 - Maria J Falaguera
 - Eloy Felix
 - Barbara Zdrazil
 - Andrew Leach
- Chemical biology group @ EBI
- Service providers



- Funders



User interface: Query assistant (Beta)

autogenerated
query

The screenshot shows the top search bar with the query "inv: \"novartis\" AND ic: \"A61K0031\" AND ttl: \"kinase\"". Below the search bar are several toggle switches: "All chemically annotated authorities" (on), "Biologically relevant" (off), "Specify dates" (off), and "Structure search" (off). A blue arrow points from the text "autogenerated query" to the search bar.

The screenshot shows the "Query Assistant (Beta)" interface with three condition rows. Each row has a "Select field" dropdown, a "Select logical operator" dropdown (set to "AND"), and an "Enter value" input field with a "REMOVE CONDITION" button. The first row has "Inventor(s)" selected, "novartis" as the value, and "A61K0031" as the value for the second condition. The second row has "ICPR" selected, "A61K0031" as the value, and "Title" as the value for the third condition. The third row has "Title" selected and "kinase" as the value. A green arrow labeled "conditions" points to the first two rows, and an orange dashed arrow labeled "logical operators" points to the second and third rows. A green arrow also points from the "conditions" label to the "Select field" dropdown of the first row.

Your Query:
inv: "novartis" AND ic: "A61K0031" AND ttl: "kinase"

Total Hits: 9

< 1 >



Query: inv: "novartis" AND ic: "A61K0031" AND ttl: "kinase" AND ((pnctry:(US OR EP OR WO OR JP)))

5-PHENYLTHIAZOLE DERIVATIVES AND THEIR USE AS P13 KINASE INHIBITORS

EP-1608647-B1

Data availability – Swift delivery

Source	Day of Source Data Availability	Delay from Patent Publication Data (original language)	Availability IFI CLAIMS Global DB (original language)	Translation availability (EN)	Availability in SureChEMBL	Annotated
EP	Wednesday	Same day	Same day	1 day later	Same as for IFI DB	2-7 days later
JP Grants	Wednesday	2-3 days	2-3 days after publication	1 day later		
JP Applications	Thursday	2-3 days	2-3 days after publication	1 day later		
US Grants	Tuesday	Same day	Same day			
US Applications	Thursday	Same day	Same day			
WO	Thursday	Same day	Same day	1 day later		

Compound structure quality improvement

- Already good results but likely underestimate reality
- Structure in MOL file can be quite different from structure in the image...

PATENT

