SureChEMBL@10

The challenging child with great prospects for the future

Nicolas Bosc

SureChEMBL team

EMBL-EBI

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

EMBL-EBI

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

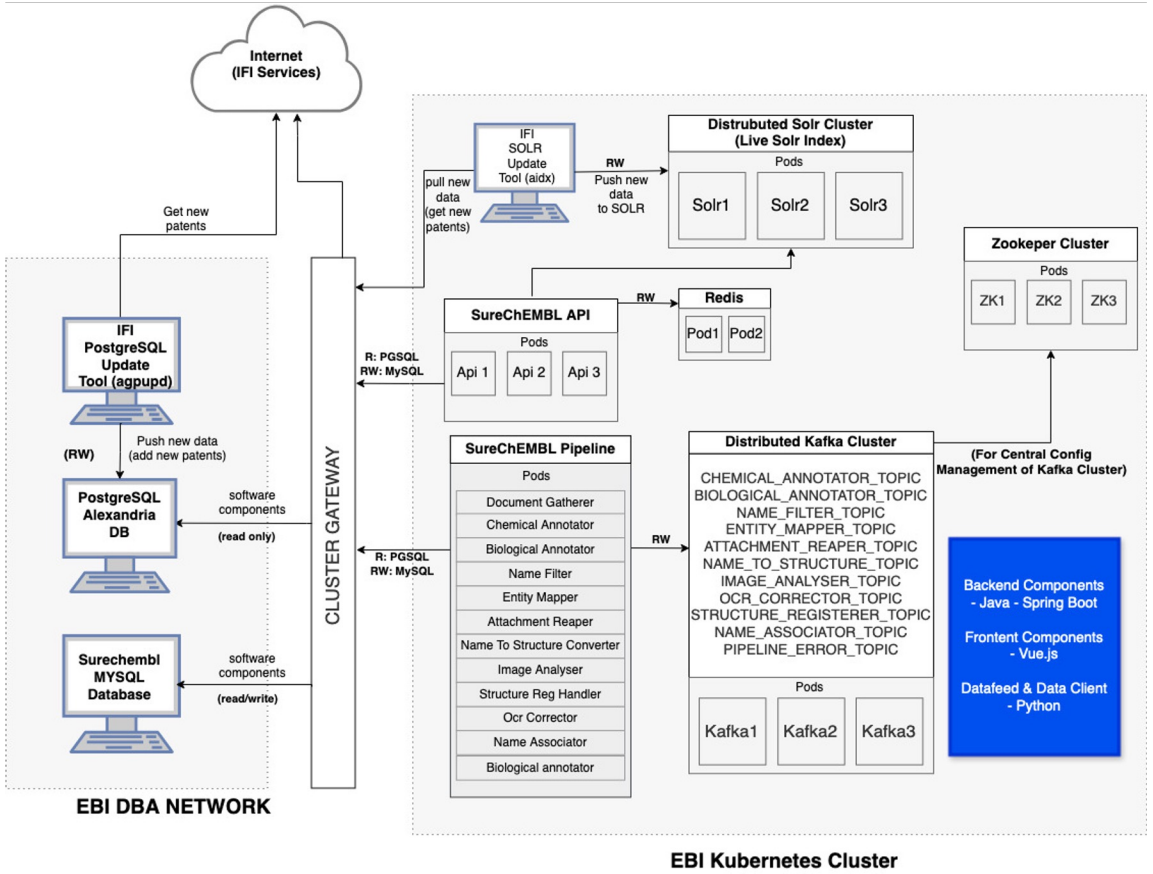
SureChEMBL: a challenging child

- 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



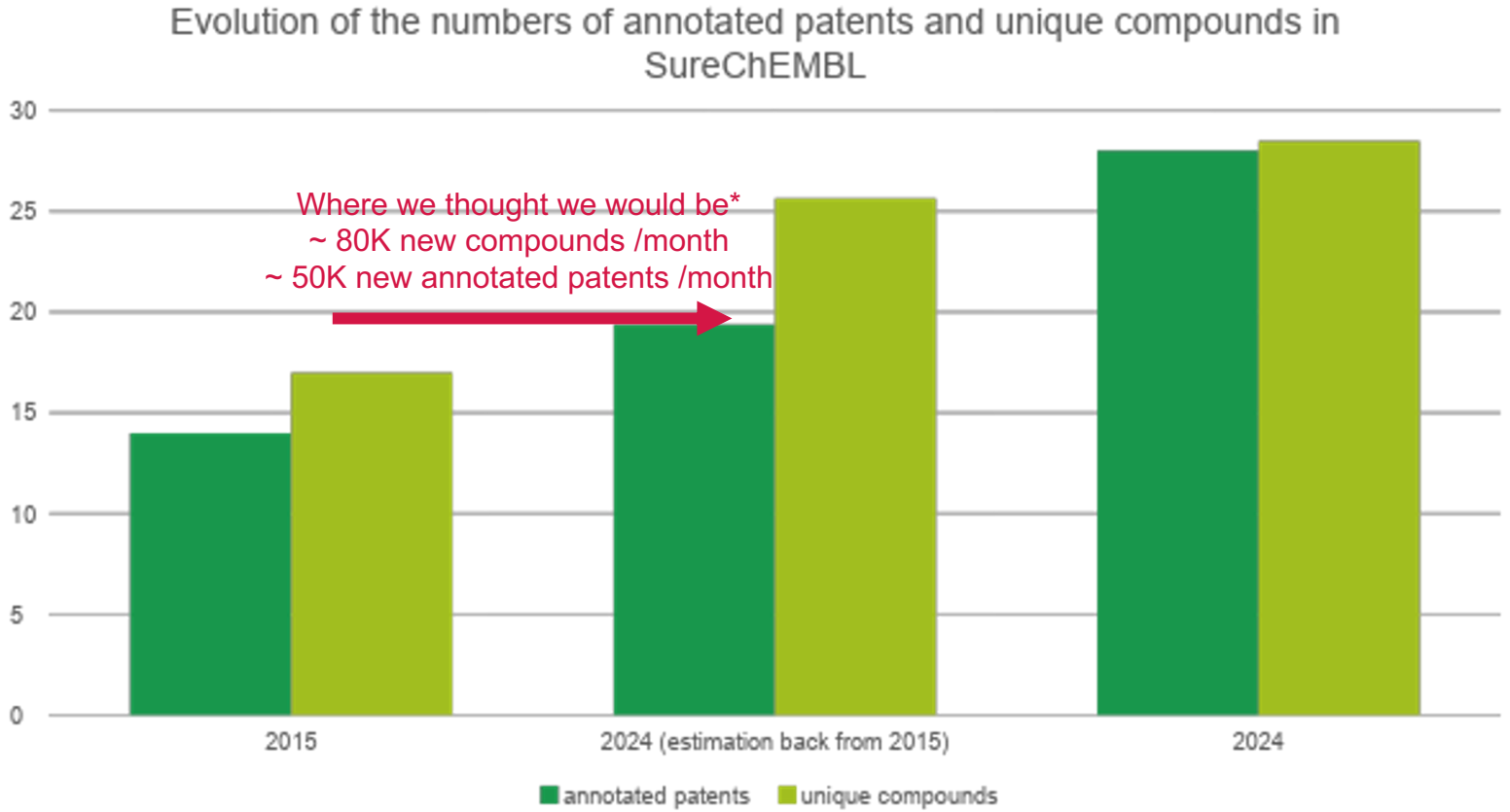
* not yet available

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	
	Granted		1980				
JPO	Applications	EN	1976	Yes (abstract)	Yes	No	
USPTO	Applications	EN	2001	Yes	Yes	Yes	
	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	

EMBL-EBI

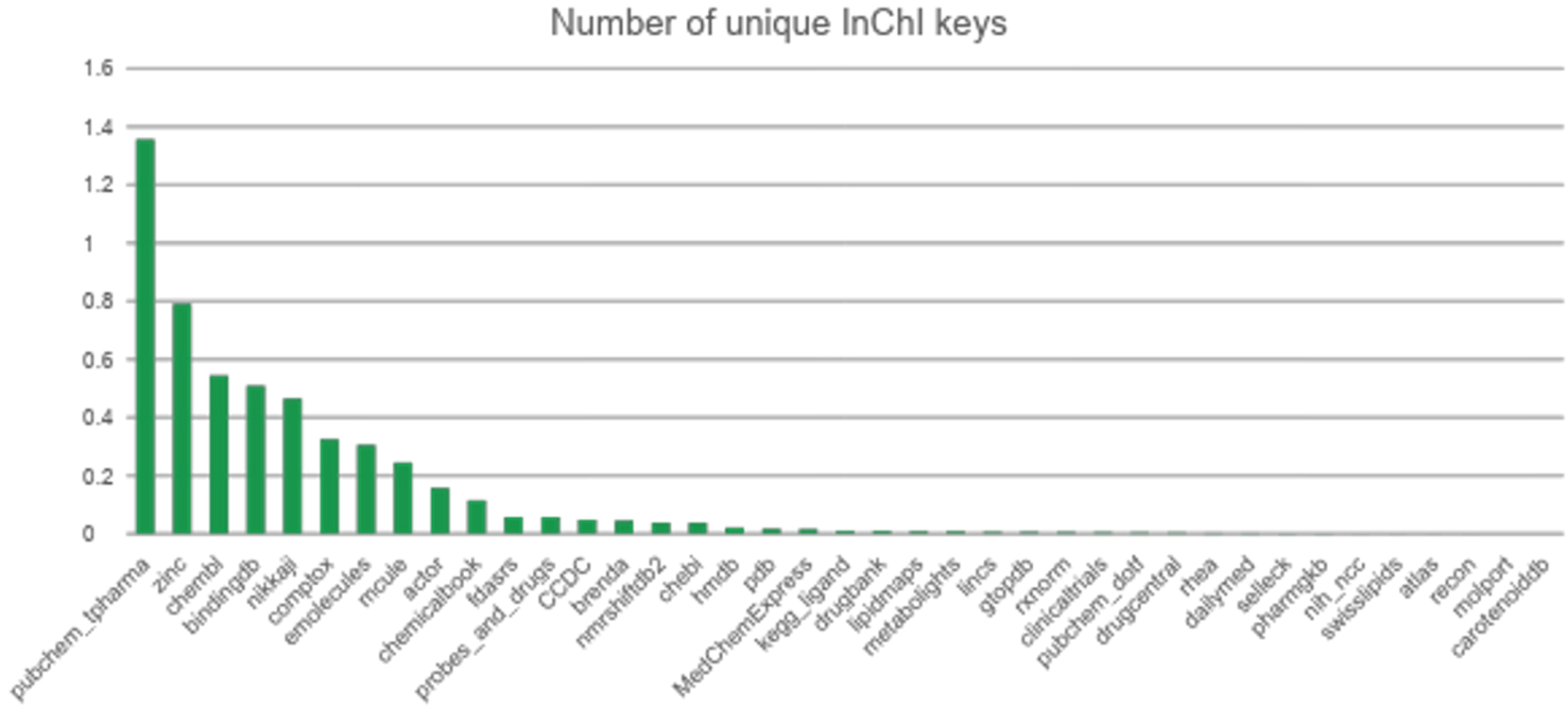
Data quantity evolution



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

EMBL-EBI

SureChEMBL chemical space vs others

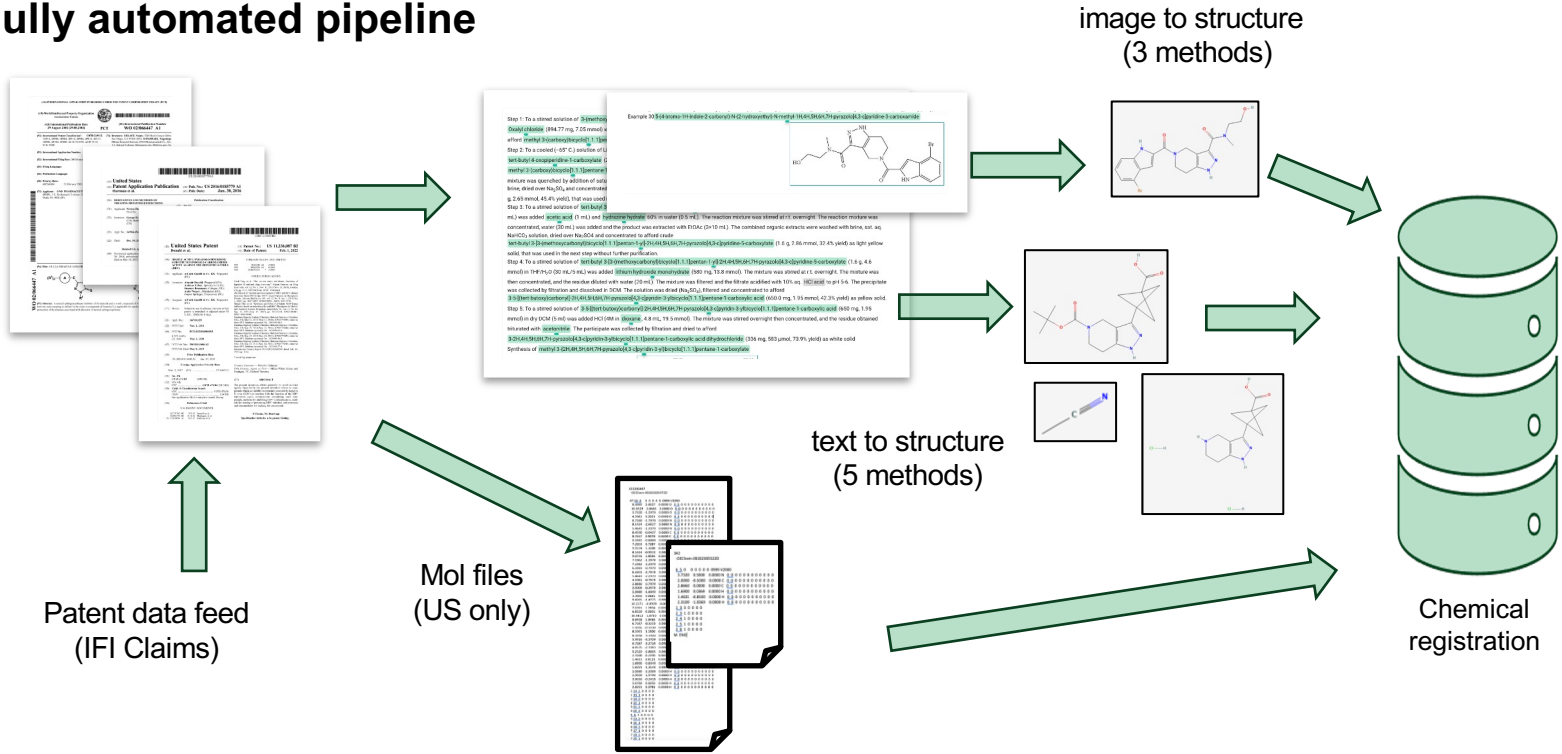


*excluding PubChem compound set

EMBL-EBI

Compound identification

Fully automated pipeline



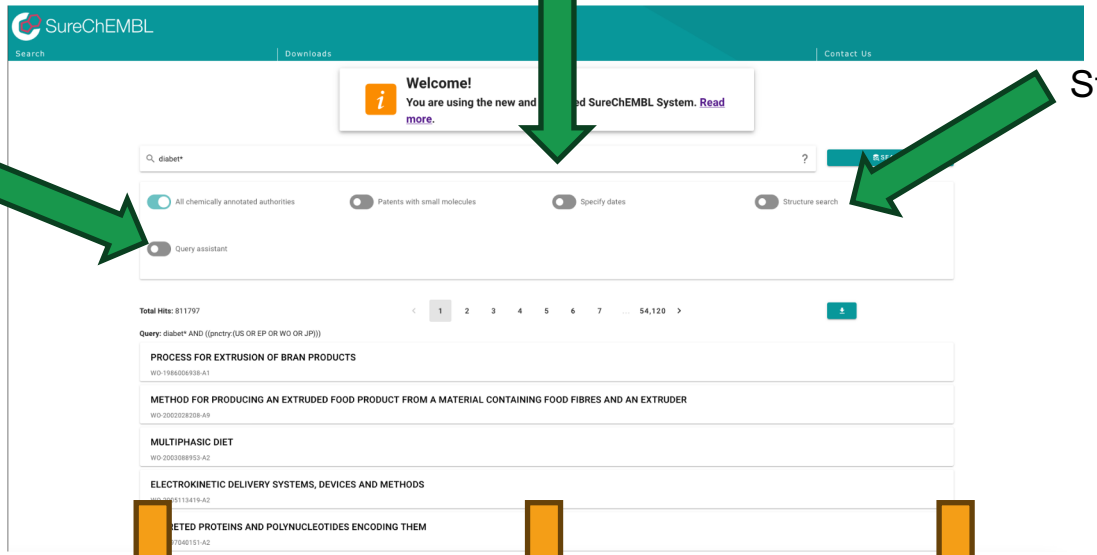
EMBL-EBI

How to access SureChEMBL data

Unique search field

Query assistant

Structure search

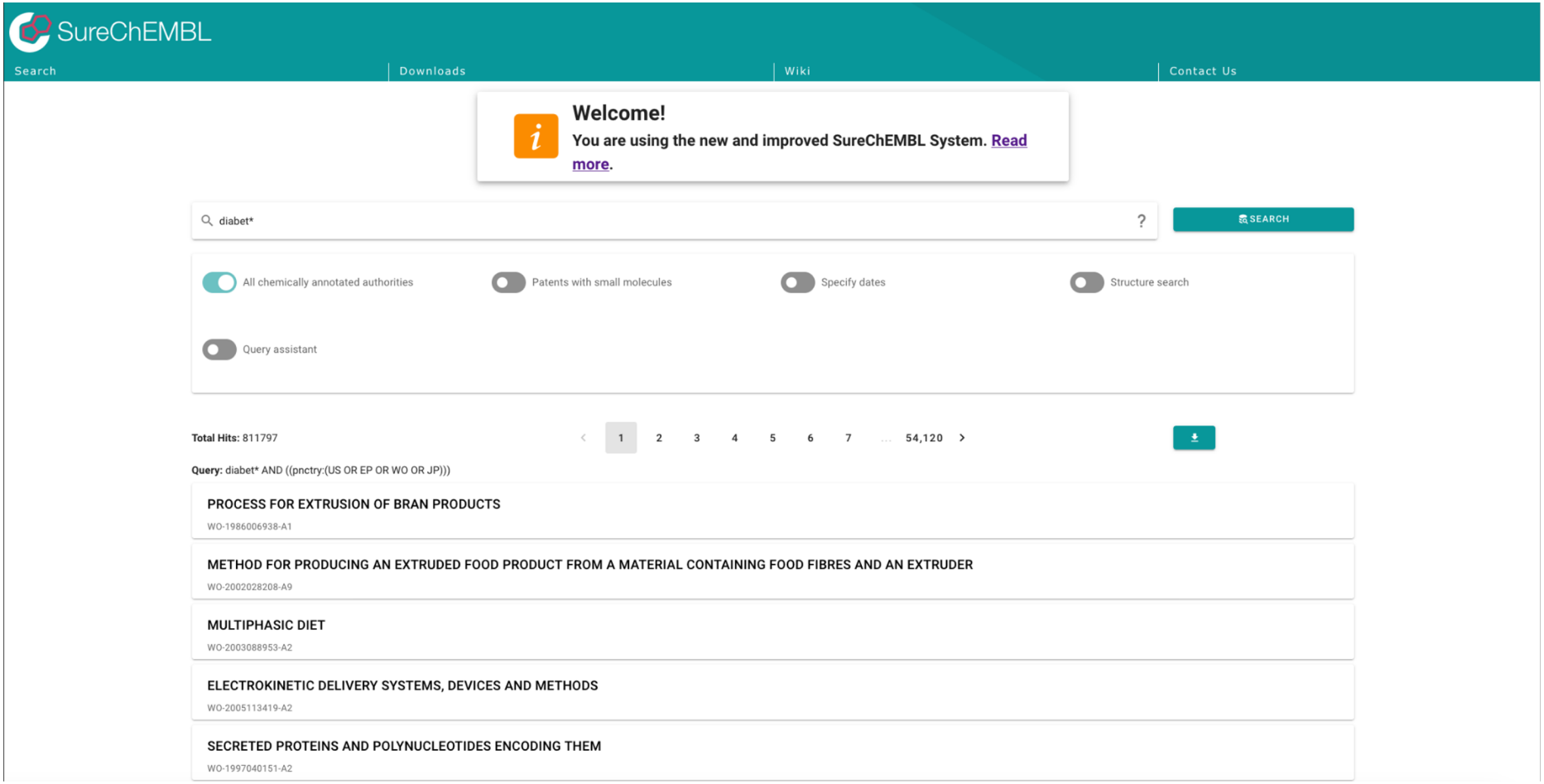


API

Data client

Downloads (Compound data dumps, MAP files)

EMBL-EBI

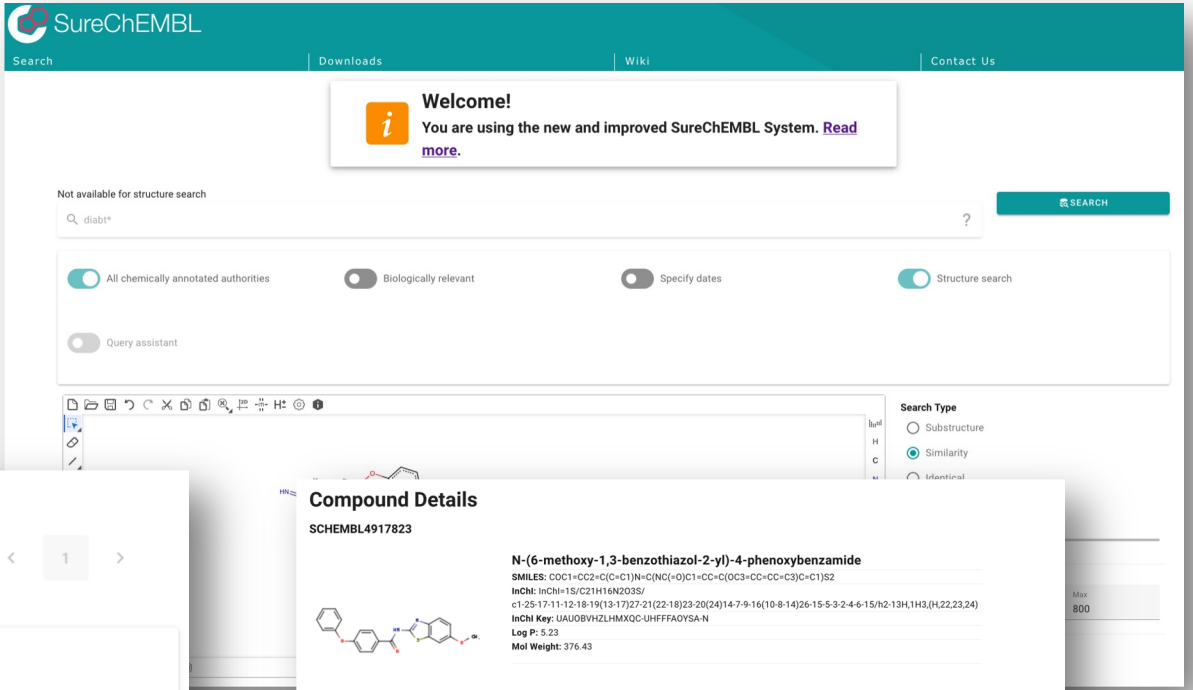

SureChEMBL
Search
Downloads
Wiki
Contact Us
Welcome!
You are using the new and improved SureChEMBL System. Read more.
diabet*
SEARCH
All chemically annotated authorities
Patents with small molecules
Specify dates
Structure search
Query assistant
Total Hits: 811797
1 2 3 4 5 6 7 ... 54,120
Query: diabet* AND ((pnctry:(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

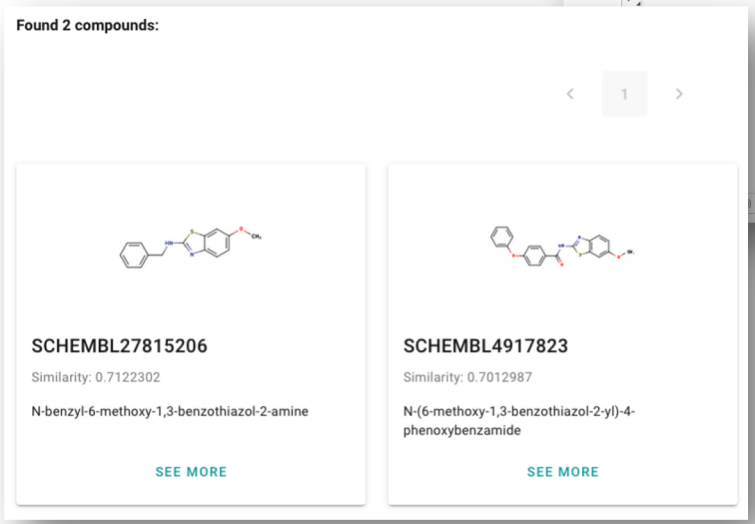
EMBL-EBI

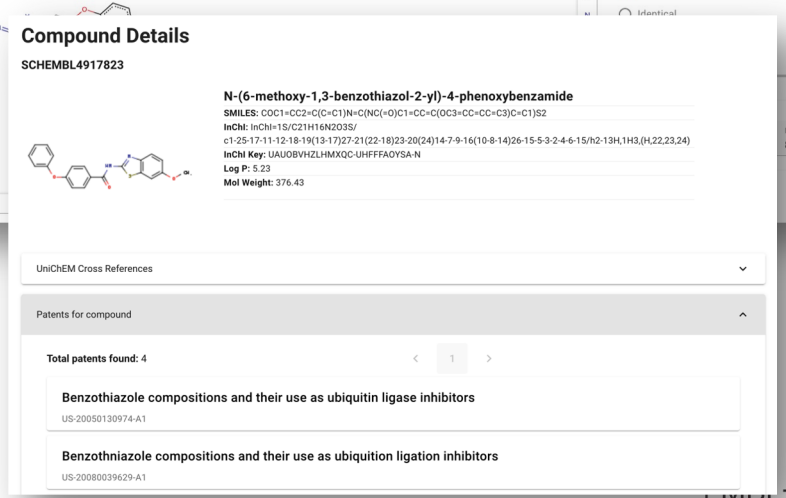
User interface

Structure search

- Substructure
- Similarity
- Identical
- Connectivity

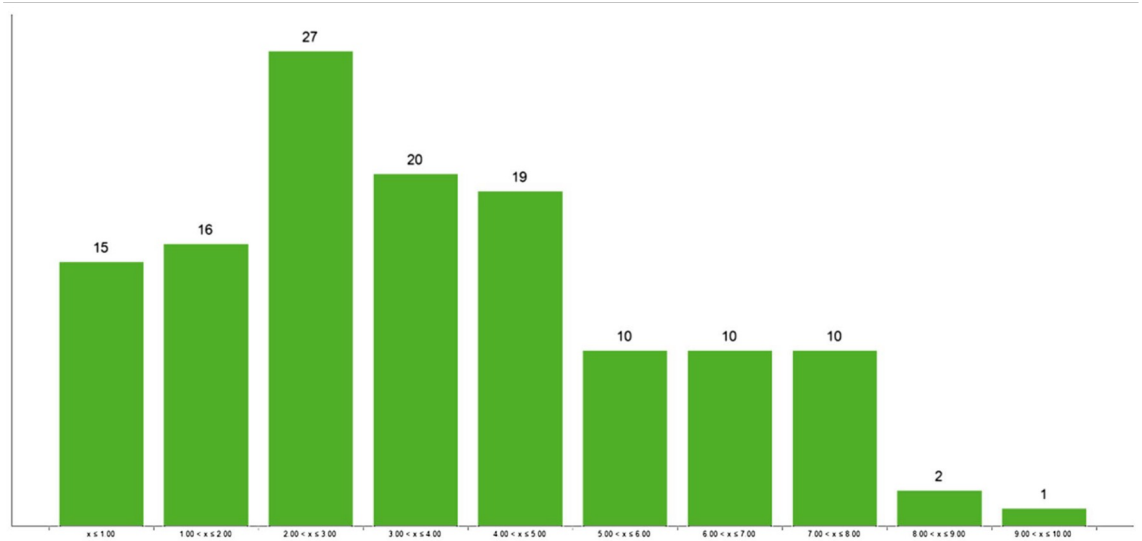






Biomedical annotation

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

Senger J Cheminform (2017) 9:26

EMBL-EBI

New biomedical annotation in SureChEMBL

- In house annotations powered by NLP
- Model trained on published Gold Standard

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

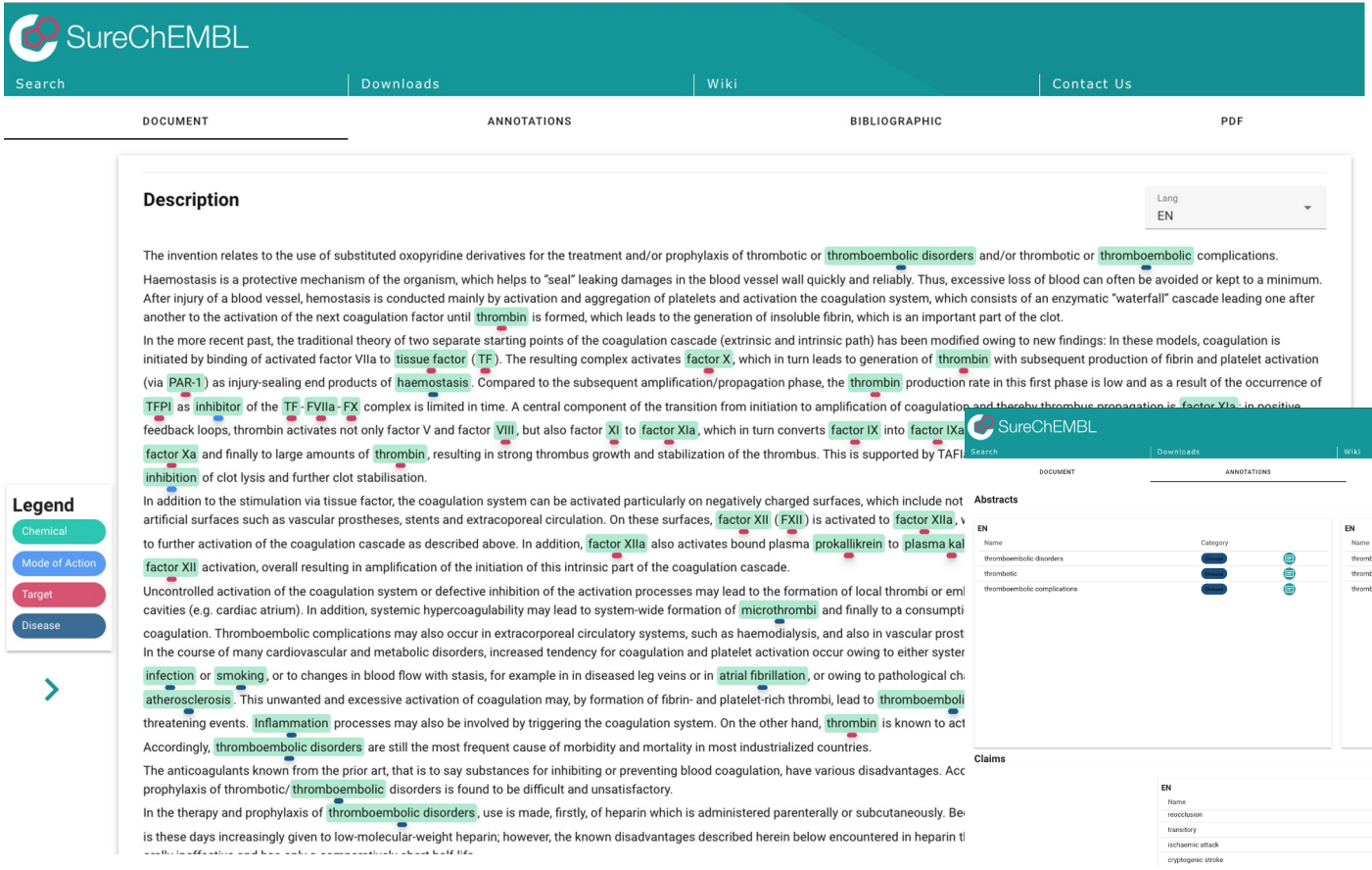
Published: September 30, 2014 • https://doi.org/10.1371/journal.pone.0107477

- 200 fully manually annotated patents
- 4 annotation types:
  - target
  - disease
  - mode of action
  - species (in-house)

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

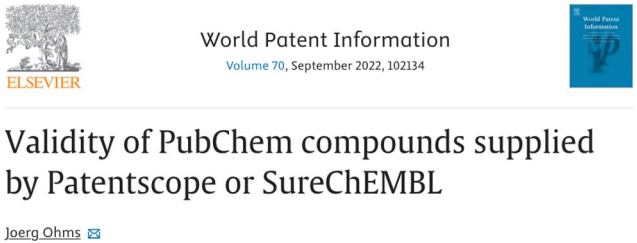
	precision	recall	F1
Disease	0.81	0.84	0.82
Mode of action	0.70	0.83	0.76
Target	0.82	0.86	0.84
Organism	0.97	0.96	0.97
Total	0.81	0.85	0.83

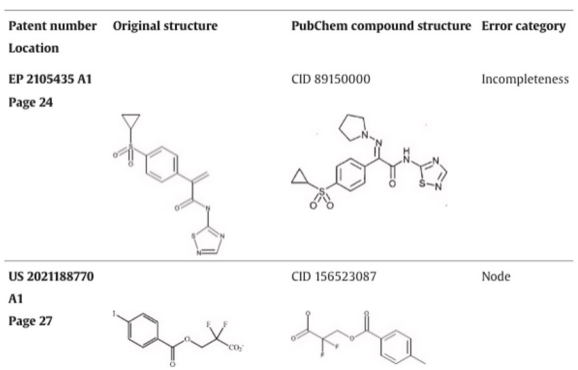
New biomedical annotation in SureChEMBL

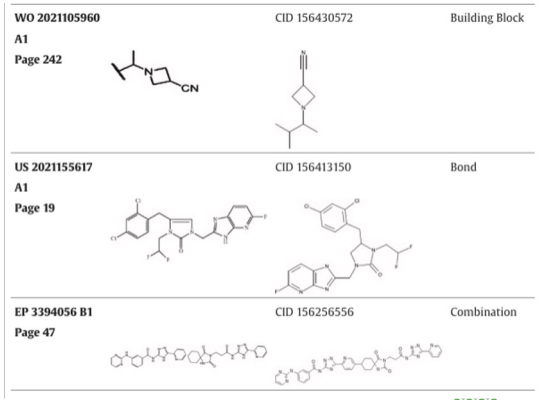


Compound structure quality improvement

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







EMBL-EBI

Compound structure quality improvement

New deep neural network models introduced recently

Home > Journal of Cheminformatics > Article

DECIMER: towards deep learning for chemical image recognition

Preliminary communication | Open access | Published: 27 October 2020

Volume 12, article number 65, (2020) Cite this article

Journal of Chemical Information and Modeling > Vol 63/Issue 7 > Article

Subscribed

Cite Share Jump to Expand

MACHINE LEARNING AND DEEP LEARNING | March 27, 2023

MolScribe: Robust Molecular Structure Recognition with Image-to-Graph Generation

Yujie Qian*, Jiang Guo, Zhengkai Tu, Zhening Li, Connor W. Coley, and Regina Barzilay*

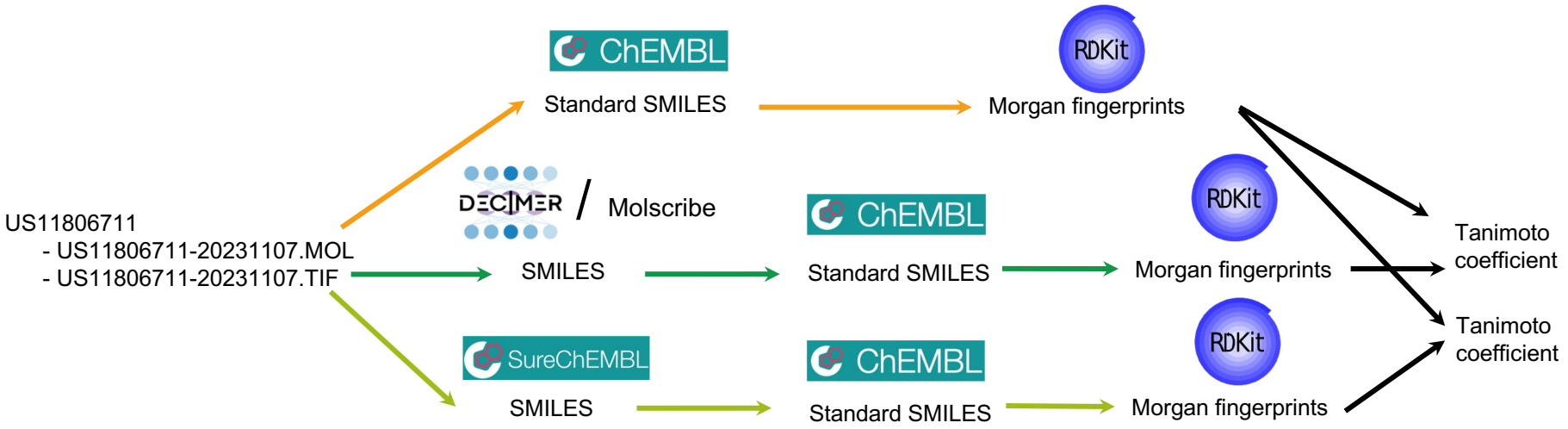
Table 2. Molecule Structure Recognition Accuracy on Synthetic, Realistic, and Perturbed Benchmarks[a]

		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
	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**

EMBL-EBI

Compound structure quality improvement

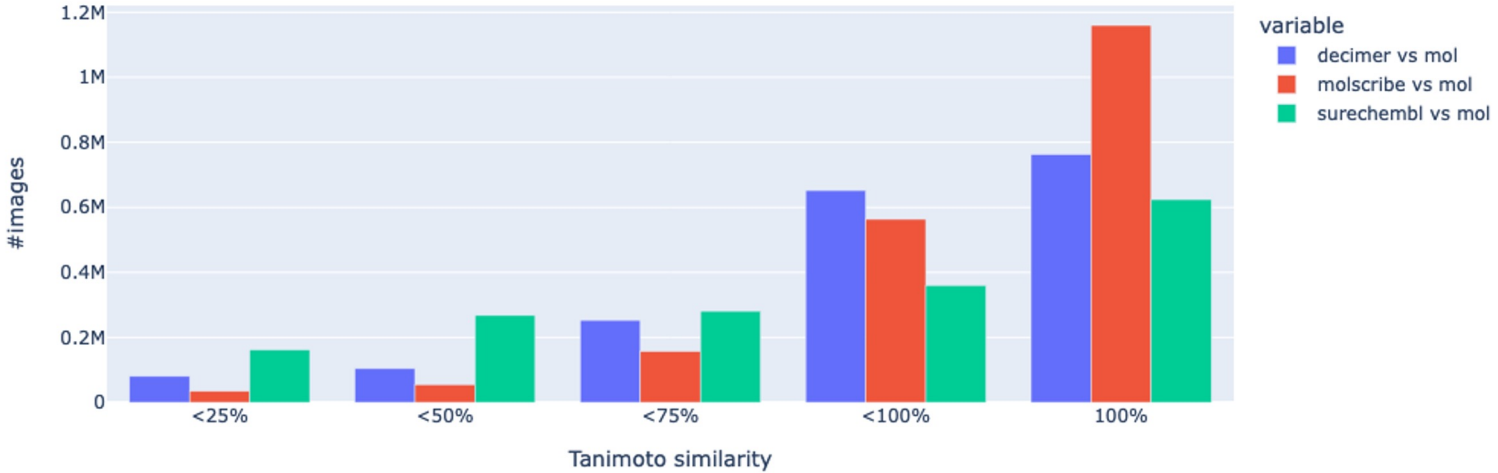
- How does it translate in practice?
- Protocol*
  - 2023 USPTO patents with MOL files
  - Both MOL and TIF files for the same compound have to be available



* new SureChEMBL paper under preparation

Compound structure quality improvement

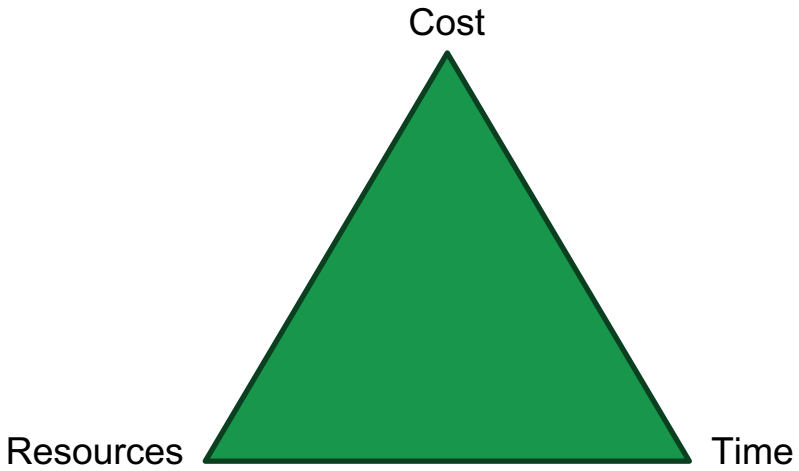
- All tools (including the ones currently used by SureChEMBL) return a majority of exact match
- Molscribe converts correctly significantly more images (and is faster)



EMBL-EBI

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

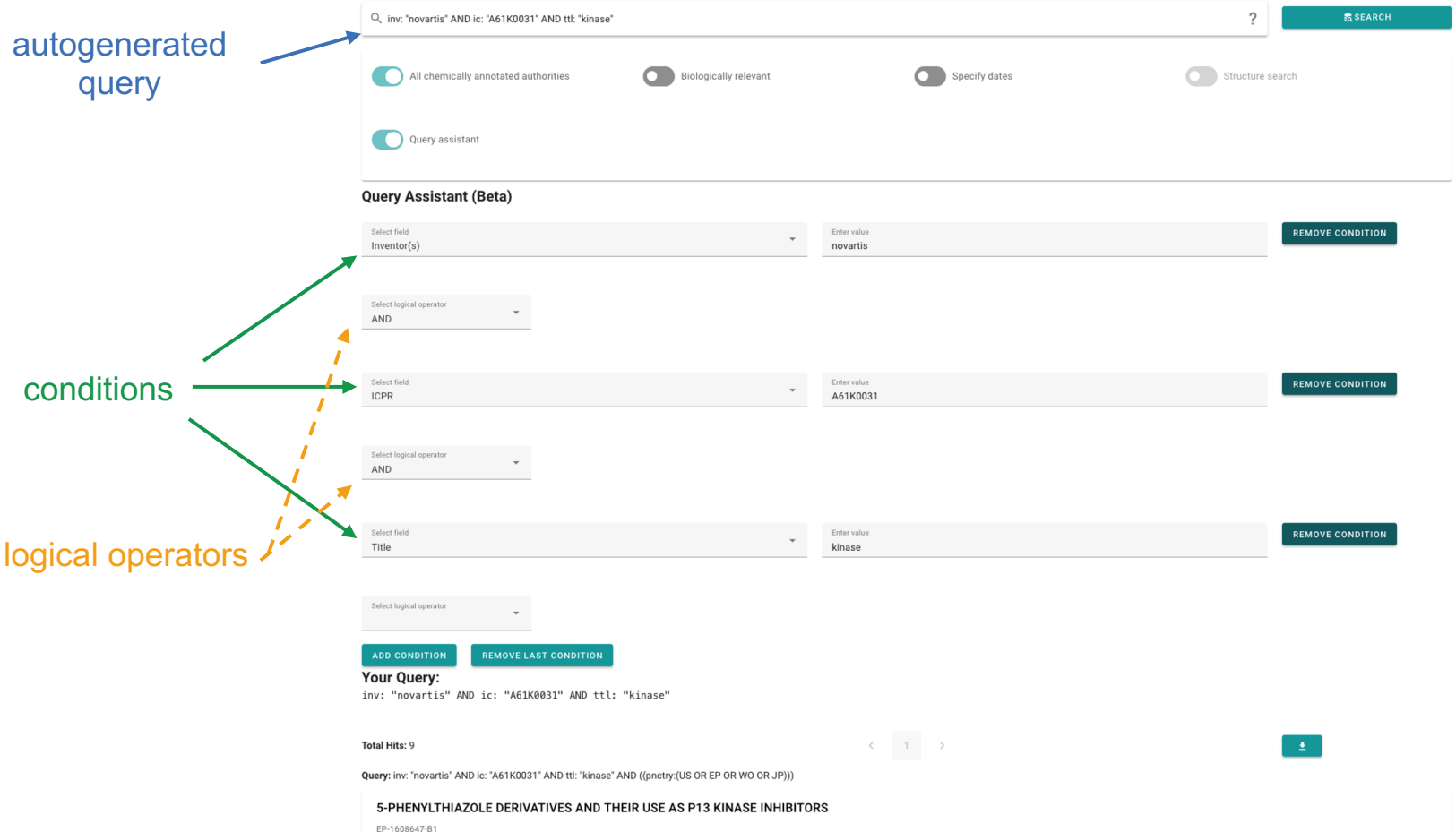
Acknowledgments

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

OPSIN

- Funders

User interface: Query assistant (Beta)



EMBL-EBI

Data availability – Swift delivery

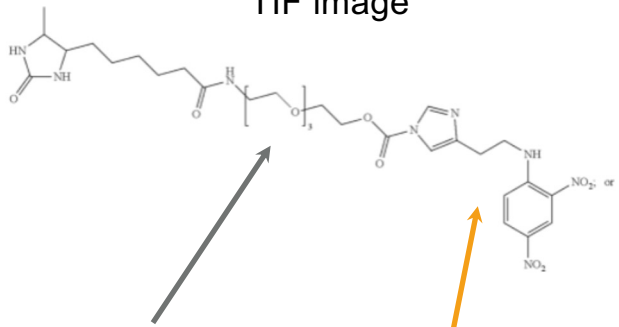
Source	Day of Source Data Availability	Delay from Patent Publication Data (original language)	Availability IFI CLAIMS Global DB (original language)	Translation availability (EN)	Availability in SureChEMBL	Annotated
EP	Wednesday	Same day	Same day	1 day later	Same as for IFI DB	2-7 days later
JP Grants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

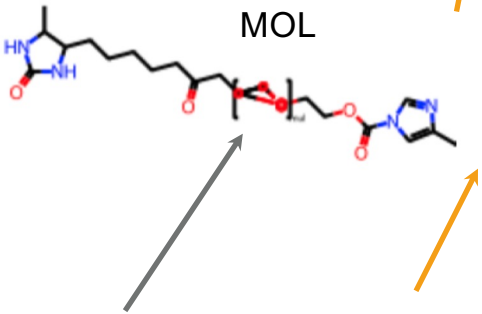
- Already good results but likely underestimate reality
- Structure in MOL file can be quite different from structure in the image…

PATENT

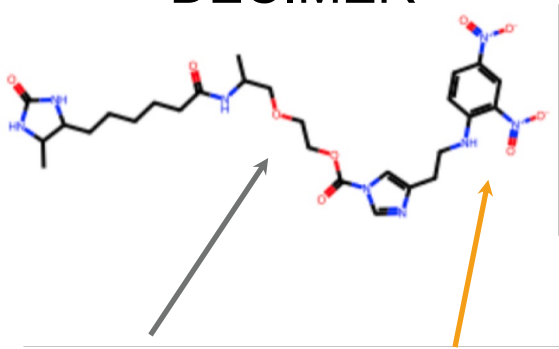
TIF image

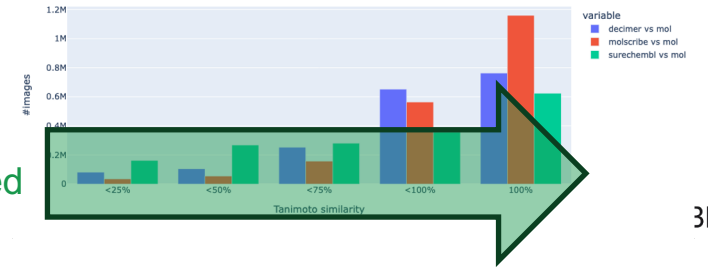


MOL



DECIMER





performance shift expected