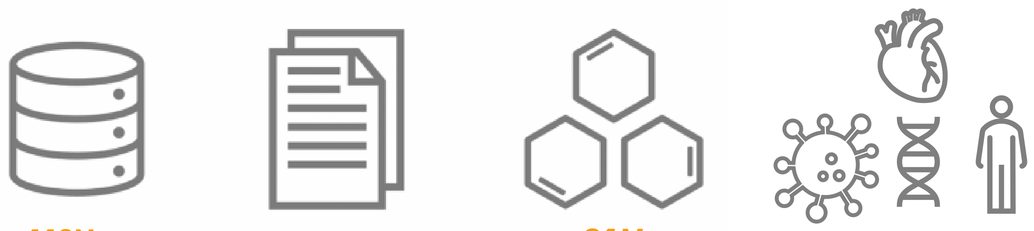
EMBL's European Bioinformatics Institute (EMBL-EBI)

SureChEMBL2.0: Evolving a Database for Annotated Patents

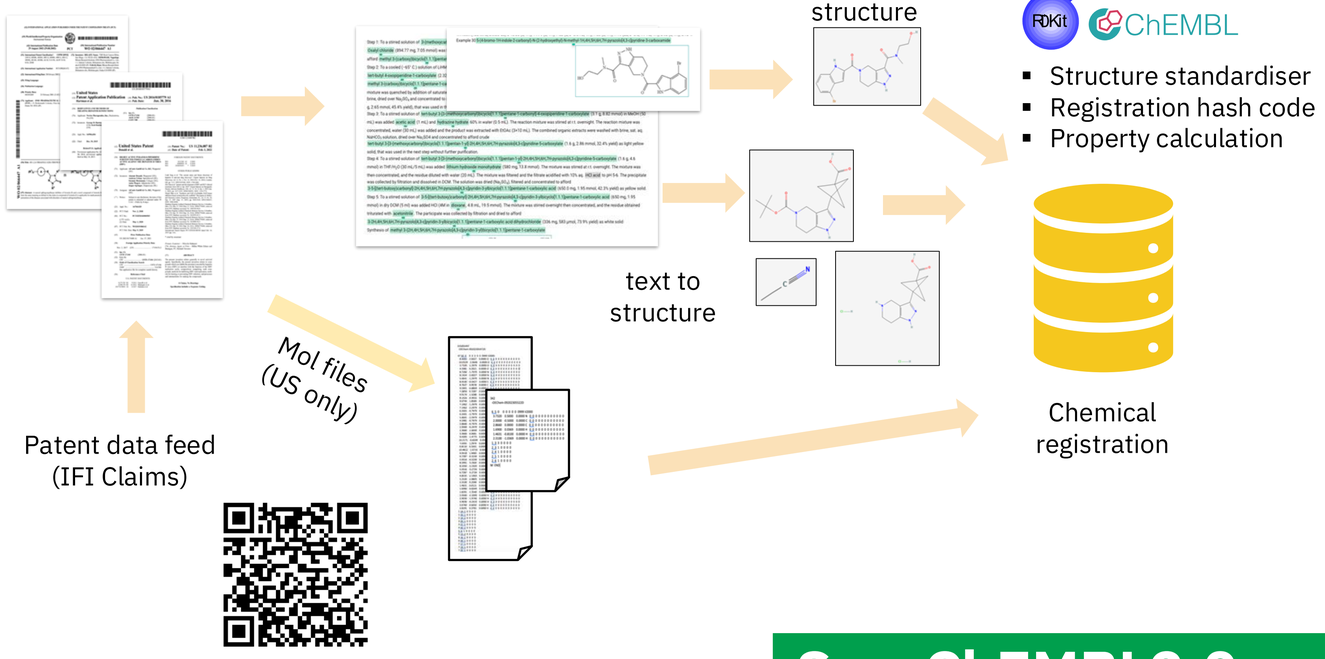
Nicolas Bosc, Tevfik Kiziloren, Eloy Felix, Carlos Moreno, Ricardo Arcila, Andrew R. Leach, Barbara Zdrazil, Noel M. O'Boyle

Chemical Biology Services, European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge CB10 1SD, UK

Patents describe and protect innovation, representing indirectly a rich source of information for drug development. First occurrences of compound and compound-target interaction often happen in the patent literature [1]. Despite the documents being freely accessible, data access remains challenging, and annotations are required for large-scale analysis. SureChEMBL [2] is a freely available resource that delivers extra value from patents, historically extracting chemical structures from text and images, but also more recently annotating the biomedical lexicon.
In 2025 we delivered SureChEMBL2.0, a major milestone offering a more robust pipeline for an improved data delivery. Relying on our improved system architecture, we are next going to offer improved annotation using newly implemented tools.

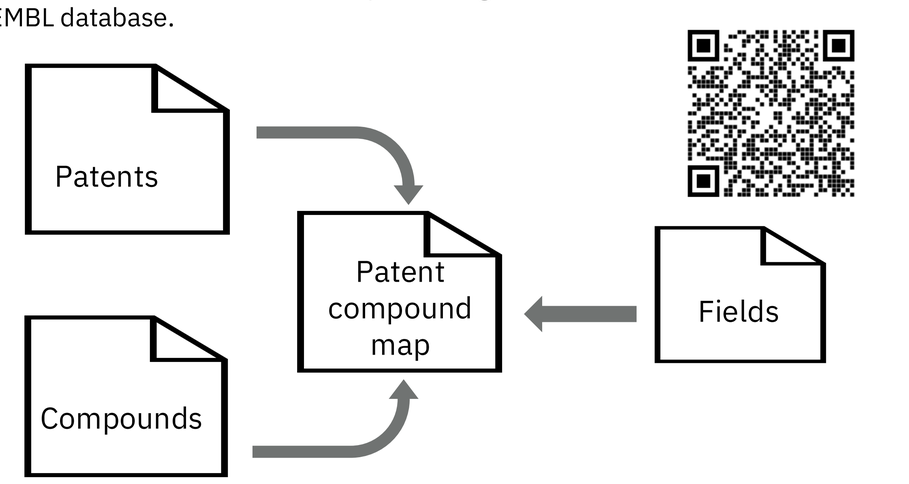


Compound identification pipeline



New export system

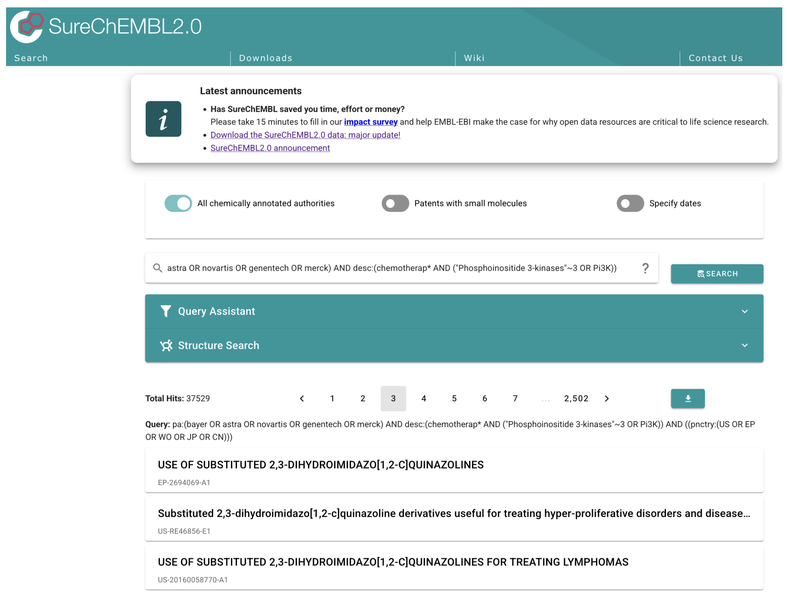
- Bulk data: new collection of core files representing the entire annotated SureChEMBL database.



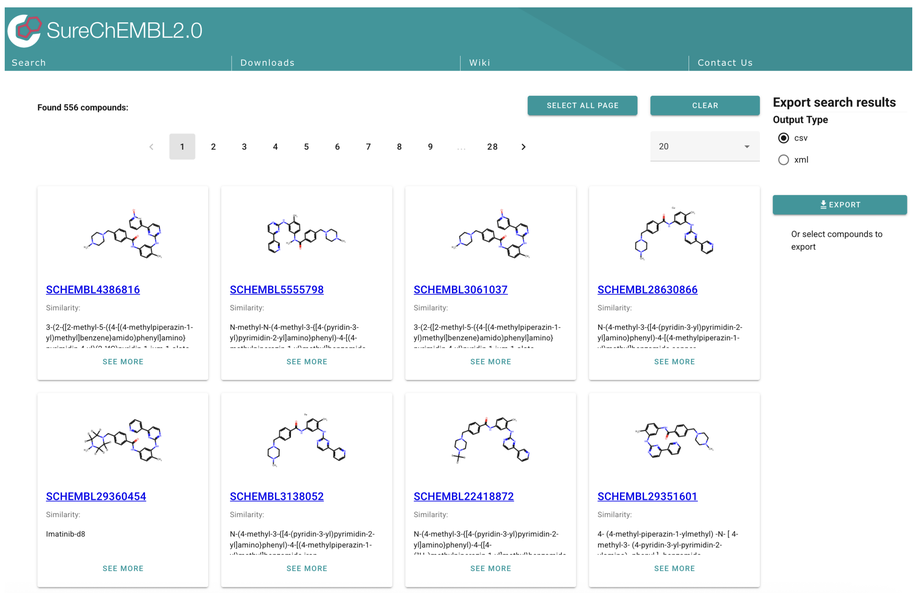
- Data updates every two weeks
- Using highly compressed but still readable parquet files (direct query or insert into internal database)

SureChEMBL2.0 user interface

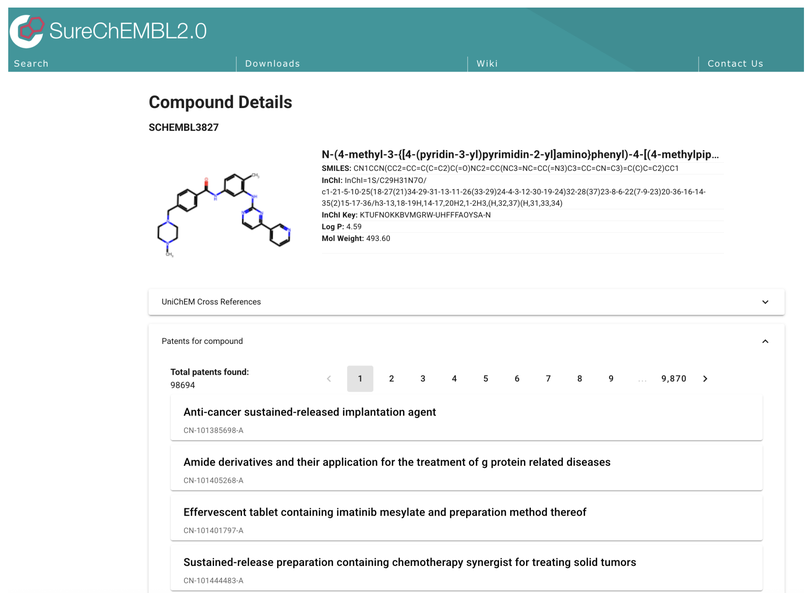
www.surechembl.org



Main page with search by keyword returning a paginated patent list matching the query



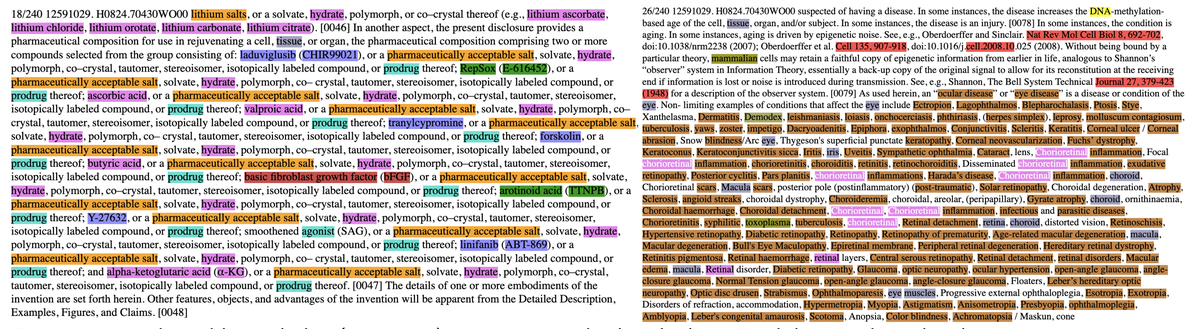
Compound search result page



Compound details page with compound structure information, UniChem cross referencing, and list of patents where the compound was found

New text annotation system

- Leadmine (NextMove Software), text mining tool for the identification and annotation of chemicals, protein targets, genes, diseases, species, named reactions, company names, cell lines, etc. [3]
- Widely adopted across industry
- Deployment in SureChEMBL:
  - Chemical annotation already integrated
- Non chemical annotation to follow later



18/240 12591029. H0824.70430WO00 lithium salts, or a solvate, hydrate, polymorph, or co–crystal thereof (e.g., lithium ascorbate, lithium chloride, lithium orotate, lithium carbonate, lithium citrate). [0046] In another aspect, the present disclosure provides a pharmaceutical composition for use in rejuvenating a cell, tissue, or organ, the pharmaceutical composition comprising two or more compounds selected from the group consisting of: laduviglusib (CHIR99021), or a pharmaceutically acceptable salt, solvate, hydrate, polymorph, co–crystal, tautomer, stereoisomer, isotopically labeled compound, or prodrug thereof; RepSox (E-616452), or a pharmaceutically acceptable salt, solvate, hydrate, polymorph, co– crystal, tautomer, stereoisomer, isotopically labeled compound, or prodrug thereof; ascorbic acid, or a pharmaceutically acceptable salt, solvate, hydrate, polymorph, co–crystal, tautomer, stereoisomer, isotopically labeled compound, or prodrug thereof; valproic acid, or a pharmaceutically acceptable salt, solvate, hydrate, polymorph, co–crystal, tautomer, stereoisomer, isotopically labeled compound, or prodrug thereof; tranylcypromine, or a pharmaceutically acceptable salt, solvate, hydrate, polymorph, co– crystal, tautomer, stereoisomer, isotopically labeled compound, or prodrug thereof; forskolin, or a pharmaceutically acceptable salt, solvate, hydrate, polymorph, co–crystal, tautomer, stereoisomer, isotopically labeled compound, or prodrug thereof; butyric acid, or a pharmaceutically acceptable salt, solvate, hydrate, polymorph, co–crystal, tautomer, stereoisomer, isotopically labeled compound, or prodrug thereof; basic fibroblast growth factor (bFGF), or a pharmaceutically acceptable salt, solvate, hydrate, polymorph, co–crystal, tautomer, stereoisomer, isotopically labeled compound, or prodrug thereof; protneoid acid (ITNPB), or a pharmaceutically acceptable salt, solvate, hydrate, polymorph, co–crystal, tautomer, stereoisomer, isotopically labeled compound, or prodrug thereof; Y-27632, or a pharmaceutically acceptable salt, solvate, hydrate, polymorph, co–crystal, tautomer, stereoisomer, isotopically labeled compound, or prodrug thereof; smoothened agonist (SAG), or a pharmaceutically acceptable salt, solvate, hydrate, polymorph, co–crystal, tautomer, stereoisomer, isotopically labeled compound, or prodrug thereof; Iminfamib (ABT-869), or a pharmaceutically acceptable salt, solvate, hydrate, polymorph, co– crystal, tautomer, stereoisomer, isotopically labeled compound, or prodrug thereof; and alpha-ketoglutaric acid (α-KG), or a pharmaceutically acceptable salt, solvate, hydrate, polymorph, co–crystal, tautomer, stereoisomer, isotopically labeled compound, or prodrug thereof. [0047] The details of one or more embodiments of the invention are set forth herein. Other features, objects, and advantages of the invention will be apparent from the Detailed Description, Examples, Figures, and Claims. [0048]

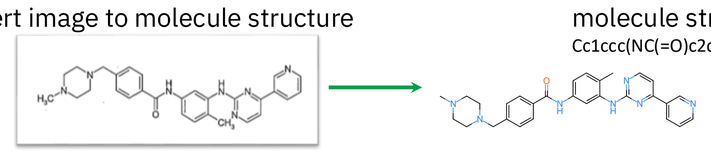
26/240 12591029. H0824.70430WO00 suspected of having a disease. In some instances, the disease increases the DNA-methylation-based age of the cell, tissue, organ, and/or subject. In some instances, the disease is an injury. [0078] In some instances, the condition is aging. In some instances, aging is driven by epigenetic noise. See, e.g., Oberdoerffer and Sinclair. Nat Rev Mol Cell Biol 8, 692-702, doi:10.1038/nrm2238 (2007); Oberdoerffer et al. Cell 135, 907-918, doi:10.1016/j.cell.2008.10.025 (2008). Without being bound by a particular theory, mammalian cells may retain a faithful copy of epigenetic information from earlier in life, analogous to Shannon's "observer" system in Information Theory, essentially a back-up copy of the original signal to allow for its reconstitution at the receiving end if information is lost or noise is introduced during transmission. See, e.g., Shannon, The Bell System Technical Journal 27, 379-423 (1948) for a description of the observer system. [0079] As used herein, an "ocular disease" or "eye disease" is a disease or condition of the eye. Non- limiting examples of conditions that affect the eye include Ectropion, Lagophthalmos, Blepharochalasis, Ptosis, Stye, Xanthelasma, Dermatitis, Demodex, leishmaniasis, loiasis, onchocerciasis, phthiriasis, (herpes simplex), leprosy, molluscum contagiosum, tuberculosis, yaws, zoster, impetigo, Dacryoadenitis, Epiphora, exophthalmos, Conjunctivitis, Scleritis, Keratitis, Corneal ulcer / Corneal abrasion, Snow blindness/Arc eye, Thygeson's superficial punctate keratopathy, Corneal neovascularization, Fuchs' dystrophy, Keratoconus, Keratoconjunctivitis sicca, Iritis, iris, Uveitis, Sympathetic ophthalmia, Cataract, lens, Chorioretinitis, inflammation, Focal Spotty retinal, inflammation, chorioretinitis, choroiditis, retinitis, retinochoroiditis, Disseminated Spotty retinal, inflammation, exudative retinopathy, Posterior cyclitis, Pars planitis, Anterior uveitis, inflammations, Harada's disease, Sympathetic, inflammation, choroid, Chorioretinal scars, Macula scars, posterior pole (postinflammatory) (post-traumatic), Solar retinopathy, Choroidal degeneration, Atrophy, Sclerosis, angioid streaks, choroidal dystrophy, Choroideremia, choroidal, areolar, (peripapillary), Gyrate atrophy, choroid, ornithinaemia, Choroidal haemorrhage, Choroidal detachment, Chorioretinal, Chorioretinal, inflammation, infectious and parasitic diseases, Chorioretinitis, syphilitic, toxoplasma, tuberculosis, Chorioretinal, Retinal detachment, retina, choroid, distorted vision, Retinoschisis, Hypertensive retinopathy, Diabetic retinopathy, Retinopathy, Retinopathy of prematurity, Age-related macular degeneration, macula, Macular degeneration, Bull's Eye Maculopathy, Epiretinal membrane, Peripheral retinal degeneration, Hereditary retinal dystrophy, Retinitis pigmentosa, Retinal haemorrhage, retinal layers, Central serous retinopathy, Retinal detachment, retinal disorders, Macular edema, macula, Retinal disorder, Diabetic retinopathy, Glaucoma, optic neuropathy, ocular hypertension, open-angle glaucoma, angle-closure glaucoma, Normal Tension glaucoma, open-angle glaucoma, angle-closure glaucoma, Floaters, Leber's hereditary optic neuropathy, Optic disc drusen, Strabismus, Ophthalmoparesis, eye muscles, Progressive external ophthalplegia, Esotropia, Exotropia, Disorders of refraction, accommodation, Hypermetropia, Myopia, Astigmatism, Anisometropia, Presbyopia, ophthalmoplegia, Amblyopia, Leber's congenital amaurosis, Scotoma, Anopsia, Color blindness, Achromatopsia / Maskun, cone

Patent annotation with Leadmine (prototype), orange: generic chemical name, pink: generic molecule, grey: anatomy, violet: molecule dictionary, turquoise: mechanism, green: PubChem dictionary, dark red: gene, yellow: polymer, light red: journal, khaki: organism, dark orange: disease

Molecule structure detection with AI

- Convert image to molecule structure

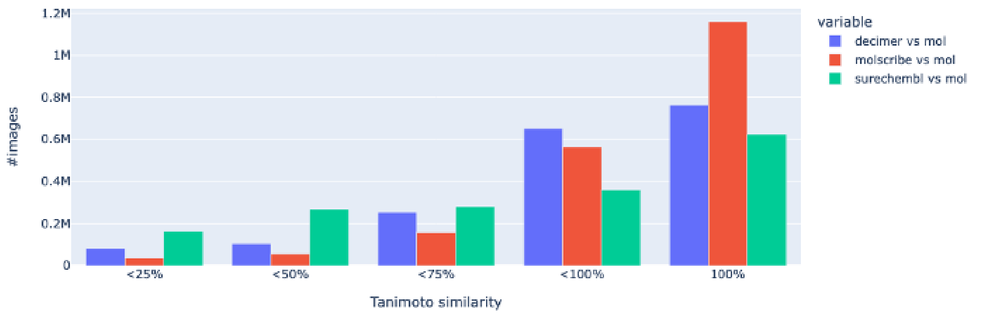
patent image



molecule structure once converted

Cc1ccc(NC(=O)c2ccc(CN3CCN(C)CC3)cc2)cc1Nc1nccc(-c2cccnc2)n1

- New generation OCR tools were evaluated on 2023 US patent dataset: DECIMER [4] / Molscribe [5]
- Compare structures from MOL files with structures extracted from associated images
- AI-based model, in particular MolScribe, gives improved structure accuracy from image extraction
- Now on-going incorporation of MolScribe into SureChEMBL2.0



[1] Senger, S. (2017). Assessment of the significance of patent-derived information for the early identification of compound–target interaction hypotheses. J. Cheminform. *9*, 26.
[2] Papadatos, G., Davies, M., Dedman, N., Chambers, J., Gaulton, A., Siddle, J., Koks, R., Irvine, S.A., Pettersson, J., Goncharoff, N., et al. (2016). SureChEMBL: a large-scale, chemically annotated patent document database. Nucleic Acids Res. *44*, D1220–D1228
[3] Lowe, D., Sayle, R. (2015). LeadMine: a grammar and dictionary driven approach to entity recognition. J. Cheminform. *7*, S1, S5.
[4] Rajan, K., Brinkhaus, H.O., Agea, M.I., Zielesny, A., and Steinbeck, C. (2023). DECIMER.ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. Nat. Commun. *14*, 5045.
[5] Qian, Y., Guo, J., Tu, Z., Li, Z., Coley, C.W., and Barzilay, R. (2023). MolScribe: Robust Molecular Structure Recognition with Image-to-Graph Generation. J. Chem. Inf. Model. *63*, 1925–1934

www.ebi.ac.uk

www.ebi.ac.uk/about/teams/chemical-biology-services/

EMBL's European Bioinformatics Institute (EMBL-EBI)
Wellcome Genome Campus, Hinxton,
Cambridgeshire. CB10 1SD. UK.
T +44(0)1223 494 444