

ChemSpider

How a Structure-Centric Community for Chemists Can Benefit Drug Discovery

**Antony Williams
(ChemZoo Inc)**


A Conversation of Possibilities. What if...?

- 25 million structures from vendors, patents and publications were freely available online?
- What if this collection could be filtered according to multiple targets, physchem properties, structure fragments and....
- What if all chemical vendor collections were freely available online, and updated daily?
- What if Pubmed Central were structure searchable?
- What if public domain chemistry data could be curated, tagged, enhanced in real time online?

Example Search 1



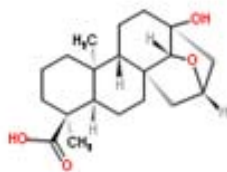
- Text Search – Identifier e.g. Quesnoin

Systematic Name, Synonym, Trade Name,
Registry Number, SMILES or InChI 

Example Search 1

INHERENT PROPERTIES, IDENTIFIERS AND REFERENCES

Quick Search: [Same Skeleton](#) [All Isomers](#)



[load](#) [save](#) [zoom](#) [jmol](#)

ChemSpider ID: 21105581
Empirical Formula: C₂₀H₃₀O₄
Molecular Weight: 334.4498
Nominal Mass: 334 Da
Average Mass: 334.4498 Da
Monoisotopic Mass: 334.214409 Da

[Support ChemSpider <](#)

Systematic Name
([OpenEye](#)):

SMILES: OC(=O)[C@@]5(C)CCC[C@]1(C)[C@@H]5CC[C@@]24C[C@H]3C[C@@](O)(CC[C@@H]12)[C@@H]4O3

InChI: InChI=1/C20H30O4/c1-17-6-3-7-18(2,16(21)22)13(17)4-8-19-10-12-11-20(23,15(19)24-12)9-5-14(17)19/h12-15,23H,3-11H2,1-2H3,(H,21,22)/t12-,13-,14-,15+,17+,18-,19+,20-/m0/s1

InChIKey: KVZUXTIZQSVUTI-VFJLVGHCBZ

DATA SOURCE(S)

Data Source	External ID(s)
Antony_Williams	N/A

NAMES AND SYNONYMS

Legend: **Validated by Experts**, [Validated by Users](#), [Non-Validated](#), [Removed by Users](#), [Redirected by Users](#), [Redirect Approved by Experts](#)

Quesnoin

(1R,4S,5S,9S,10S,13S,15S,17R)-13-Hydroxy-5,9-dimethyl-16-oxapentacyclo
[13.2.1.0~1,10~.0~4,9~.0~13,17~]octadecane-5-carboxylic acid

4H-2,3b-methanophenanthro[1,2-b]furan-6-carboxylic acid, tetradecahydro-11a-hydroxy-6,9a-dimethyl-,
(2S,3aR,3bR,5aS,6S,9aS,9bS,11aS)-

Example Search 1

⊗ PREDICTED PROPERTIES

LogP:	ACD/LogP: 3.03	# of Rule of 5 Violations:	0
ACD/LogD (pH 5.5):	2.1	ACD/LogD (pH 7.4):	0.3
ACD/BCF (pH 5.5):		ACD/BCF (pH 7.4):	
ACD/KOC (pH 5.5):		ACD/KOC (pH 7.4):	
#H bond acceptors:	4	#H bond donors:	2
#Freely Rotating Bonds:	2	Polar Surface Area:	66.76 Å ²
Index of Refraction:	1.587	Molar Refractivity:	89.31 cm ³
Molar Volume:	265.4 cm ³	Polarizability:	35.4 10 ⁻²⁴ cm ³
Surface Tension:	54.1 dyne/cm	Density:	1.25 g/cm ³
Flash Point:	175.1 Celsius	Enthalpy of Vaporization:	88.14 kJ/mol
Boiling Point:	497.4 Celsius at 760 mmHg	Vapour Pressure:	5.62E-12 mmHg at 25 Celsius

⊗ SUPPLEMENTAL INFORMATION

Description

Quesnoin, a novel unique pure organic compound, was isolated from amber discovered in the Oise River area of the Paris basin (France) and dated at 55 million years old. ¹H and ¹³C NMR indicated an unknown diterpene skeleton, quesnane. The absolute configurations of the eight chiral centers of quesnoin were determined to be 4S, 5S, 8R, 9S, 10S, 13S, 14R, and 16S. The work indicated that the climate of the Paris basin might have been tropical in the early Eocene period, 55 million years ago.

Links & References

Jossang J., Bel-Kassaoui H., Jossang A., Seuleiman M., and Nel A... Quesnoin, a Novel Pentacyclic ent-Diterpene from 55 Million Years Old Oise Amber,, *J. Org. Chem.*, 2008, 73 (2), 412 -417

[DOI: [10.1021/jo701544k](https://doi.org/10.1021/jo701544k)]

Example Search 2

• Property Search

<input checked="" type="checkbox"/> Empirical Formula:	<input type="text"/>	Exact match only
<input checked="" type="checkbox"/> Molecular Weight:	<input type="text"/> ± <input type="text"/> 1.0 (example: 123 ± 1)	<input type="checkbox"/> min/max
<input checked="" type="checkbox"/> Nominal Mass:	<input type="text"/> ± <input type="text"/> 1.0 <input type="text"/>	<input type="checkbox"/> min/max
<input checked="" type="checkbox"/> Average Mass:	<input type="text"/> ± <input type="text"/> 0.1 <input type="text"/> <input type="text"/>	<input type="checkbox"/> min/max
<input checked="" type="checkbox"/> Monoisotopic Mass:	<input type="text"/> ± <input type="text"/> 0.001 <input type="text"/> <input type="text"/>	<input type="checkbox"/> min/max

OPTIONS

<input checked="" type="checkbox"/> ACD/LogP:	<input type="text"/> to <input type="text"/>
<input checked="" type="checkbox"/> ACD/LogD (pH 5.5):	<input type="text"/> to <input type="text"/>
<input checked="" type="checkbox"/> ACD/LogD (pH 7.4):	<input type="text"/> to <input type="text"/>
<input checked="" type="checkbox"/> Rule Of 5:	<input type="text"/> to <input type="text"/>
<input checked="" type="checkbox"/> Number of Hydrogen Bond Acceptors:	<input type="text"/> to <input type="text"/>
<input checked="" type="checkbox"/> Number of Hydrogen Bond Donors:	<input type="text"/> to <input type="text"/>
<input checked="" type="checkbox"/> Number of Freely Rotatable Bonds:	<input type="text"/> to <input type="text"/>
<input checked="" type="checkbox"/> Polar Surface Area:	<input type="text"/> to <input type="text"/>
<input checked="" type="checkbox"/> Molar Volume:	<input type="text"/> to <input type="text"/>

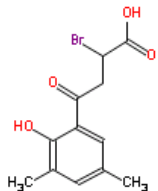
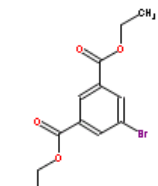
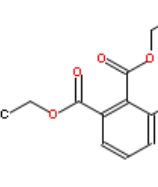
OPTIONS

Calculated Properties Search

Example Search 2

132 hits found in 2.34 seconds
Nominal_Mass >= 299 AND Nominal_Mass <= 301 AND Monoisotopic_Mass >= 299.999 and Monoisotopic_Mass <= 300.001 AND SingleComponent AND NonIsotopic

Grid
 Table
 Record
 ChemRefer
 Entrez

1 2 3 4 5 6 7							
ID	Structure	Empirical Formula	Molecular Weight	Monoisotopic Mass, Da	LogP	ACD/LogD (pH 5.5)	ACD/LogD (pH 7.4)
207470	 load save zoom jmol	C ₁₂ H ₁₃ BrO ₄	301.1332	299.999714	ACD/LogP: 3.65 XLogP: 2.50		
159022	 load save zoom jmol	C ₁₂ H ₁₃ BrO ₄	301.1332	299.999714	ACD/LogP: 4.10 XLogP: 3.50	4.1	4.1
159020	 load save zoom jmol	C ₁₂ H ₁₃ BrO ₄	301.1332	299.999714	ACD/LogP: 2.96 XLogP: 3.50	2.96	2.96

Complex Search

Search by Structure [?](#)

Search by Identifier [?](#)

Search by Elements [?](#)

Search by Properties [?](#)

Search by Calculated Properties [?](#)

Search by Data Source, Data Source Type or Focused Library [?](#)

Search by LASSO Similarity [?](#)

Input Structure

Exact
 Substructure

Search Options

Exact Match
 All Tautomers
 Same Skeleton (Including H)
 Same Skeleton (Excluding H)
 All Isomers

OPTIONS

Single/Multi-component

Search Any
 Search Single-Component Structures Only
 Search Multi-Component Structures Only

Isotopically Labeled

Search Any
 Search Isotopically Labeled Structures Only
 Disregard Isotopically Labeled Structures

Additional Filters

Filter only those having spectra associated
 Filter only those having patents link

Search Hits Limit

1000

Search Open Access Journals – ChemSpider

Match any search words (OR) Match all search words (AND)

Sources

- | | |
|--|--|
| <input checked="" type="checkbox"/> International Journal of Electrochemical Science | <input checked="" type="checkbox"/> Hindawi Publishing Corporation |
| <input checked="" type="checkbox"/> Association of Clinical Biochemists of India | <input checked="" type="checkbox"/> International Union of Crystallography |
| <input checked="" type="checkbox"/> Molecular Diversity Preservation International | <input checked="" type="checkbox"/> Medknow Publications |
| <input checked="" type="checkbox"/> PubMed Central | <input checked="" type="checkbox"/> RepositoriUM |

270 hits found in 8.73 seconds
SingleComponent AND NonIsotopic

1 2 3 4 5 6 7 8 9 10 ...

[Nitric oxide, cell death and increased **taxol** recovery \(BMC Plant Biology 2005 Volume 5 Issue Suppl 1 Page S12\) - PubMed Central Open Archives Service](#)

Don J Durzan

... -5-S1-S12 Meeting Abstract Nitric oxide, cell death and increased **taxol** recovery Durzan Don J 1 djdurzan@ucdavis.edu 1Department of Plant Sciences ...

[2'-Carbamate **taxol** \(Acta Cryst C 1995 Volume 51 Part 2 Page 295-298\) - International Union of Crystallography](#)

Q. Gao, J. Golik

... (IUCr) 2'-Carbamate **taxol** Acta Crystallographica Section C Crystal Structure Communications 0108-2701 organic compounds Volume 51 Part 2 Pages 295-298 February 1995 2 ...

[Differential partitioning of G \$\alpha\$ i1 with the cellular microtubules: a possible mechanism of development of **taxol** resistance in human ovarian carcinoma cells \(Journal Of Molecular Signaling 2006 Volume 1 Page 3\) - PubMed Central Open Archives Service](#)

Hemant K Parekh, Mahesha Adikari, Bharathi Vennapusa

... partitioning of G α i1 with the cellular microtubules: a possible mechanism of development of **taxol** resistance in human ovarian carcinoma cells Parekh Hemant K 1 hemant.parekh@ ...

[Structure of a synthetic **taxol** precursor: N-tert-butoxycarbonyl-10-deacetyl-N-debenzoyl **taxol** \(Acta Cryst C 1990 Volume 46 Part 5 Page 781-784\) - International Union of Crystallography](#)

F. Gueritte-Voegelein, D. Guenard, L. Mangatal, P. Potier, J. Guilhem, M. Cesario, C. Pascard

... (IUCr) Structure of a synthetic **taxol** precursor: N-tert-butoxycarbonyl-10-deacetyl-N-debenzoyl **taxol** Acta Crystallographica Section C Crystal Structure Communications 0108 ...

[Application of the Haller-Bauer reaction in the synthesis of **taxol**-related diterpenes: structure of the intramolecular lactam of 2-amino-5-hydroxy-4,8,11-trimethylbicyclo\[5.3.1\]undeca-3,8-diene-11-carboxylic acid \(Acta Cryst C 1991 Volume 47 Part 10 Page 2109-2112\) - International Union of Crystallography](#)

Search PubMed – ChemSpider

Search Term:

NCBI Entrez

Search

1650 hits found in 30.14 seconds

"paclitaxel"[MeSH Terms] OR taxol[Acknowledgments] OR taxol[Figure/Table Caption] OR taxol[Section Title] OR taxol[Body - All Words] OR taxol[Title] OR taxol[Abstract]

1 2 3 4 5 6 7 8 9 10 ...

ZHOU J, O'BRATE A, ZELNAK A, GIANNAKAKOU P. **Survivin Deregulation in β -Tubulin Mutant Ovarian Cancer Cells Underlies Their Compromised Mitotic Response to Taxol.** *Cancer Res.* 2004 Dec 1; **64** : 8708-8714.

PAREKH HK, ADIKARI M, VENNAPUSA B. **Differential partitioning of Gai1 with the cellular microtubules: a possible mechanism of development of Taxol resistance in human ovarian carcinoma cells.** *J Mol Signal.* 2006; **1** : 3.

BOEHMERLE W, ZHANG K, SIVULA M, HEIDRICH FM, LEE Y, JORDT SE, EHRLICH BE. **Chronic exposure to paclitaxel diminishes phosphoinositide signaling by calpain-mediated neuronal calcium sensor-1 degradation.** *Proc Natl Acad Sci U S A.* 2007 Jun 26; **104** : 11103-11108.

BOEHMERLE W, SPLITTGERBER U, LAZARUS MB, MCKENZIE KM, JOHNSTON DG, AUSTIN DJ, EHRLICH BE. **Paclitaxel induces calcium oscillations via an inositol 1,4,5-trisphosphate receptor and neuronal calcium sensor 1-dependent mechanism.** *Proc Natl Acad Sci U S A.* 2006 Nov 28; **103** : 18356-18361.

GUPTA ML JR, BODE CJ, GEORG GI, HIMES RH. **Understanding tubulin-Taxol interactions: Mutations that impart Taxol binding to yeast tubulin.** *Proc Natl Acad Sci U S A.* 2003 May 27; **100** : 6394-6397.




SHANNON KB, CANMAN JC, MOREE CB, TIRNAUER JS, SALMON ED. **Taxol-stabilized Microtubules Can Position the Cytokinetic Furrow in Mammalian Cells.** *Mol Biol Cell.* 2005 Sep; **16** : 4423-4436.

ChemSpider Data

- The database PRESENTLY contains close to 20 million compounds obtained from
 - Chemical vendors
 - Publishers
 - Commercial Database Vendors
 - US and international patents
 - Structure aggregators
 - Scraped from websites
 - Deposited by users

Quality is a Major Issue

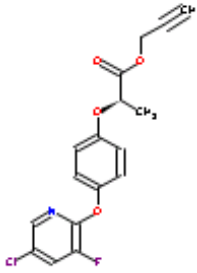
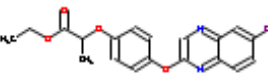
 Depositor-Supplied Synonyms: (Total: 107)

n-butanol 
1-butanol 
Butyl alcohol
n-butyl alcohol 
butanol
1-hydroxybutane
Methylolpropane
Propylcarbinol
Propylmethanol
Hemostyp

n-Butanolbutanolen
Tetrabutoxytitanium
Tyzor BP
Butanol [French]
Butyl orthotitanate
Tin tetrabutanolate
Tyzor TBT
BuOH
Tetrabutoxyzirconium
Butanolen [Dutch]
Tetrabutyl zirconate

Titanium tetrabutoxide
Titanium tetrabutylate
Butyl titanate (IV)
Zirconium tetrabutoxide
Butyl zirconate (IV)
Titanium, tetrabutoxy-
Tetrabutyl orthotitanate
Butyl alcohol (natural)
FEMA Number 2178
Normal primary butyl alcohol
Zirconic acid butyl ester
TETRABUTYL TITANATE

Curators - An Active Community

83449	 <chem>CC(=O)OCC#COC1=CC=C(Oc2nc3cc(Cl)cc(F)c3n2)C=C1</chem>	5/5/2007 1:42:21 AM	L	83449: Correct format for international common name should be clodinafop-propargyl (note propargyl with small letter p). Correct CAS for this is 105512-06-9, I don't know what the other two CAS numbers are for ? WLN is T6NJ BOR DOY1&VO2UU1& CF EG &&R Form Nanogens@aol.com
48336	 <chem>CCOC(=O)C(O)C1=CC=C(Oc2nc3cc(Cl)cc(F)c3n2)C=C1</chem>	5/5/2007 1:54:10 AM	L	48336: This is a muddle of entries for quizalofop-ethyl (note correct format for this international Common name). Quizalofop-P-ethyl (note -P- is a capital letter P in correct format) is the international common name for the chiral isomer (R Form) that replaced quizalofop-ethyl Correct CAS number for quizalofop-ethyl is 76578-14-8, Description of 89468-49-5 is not known, Correct CAS for quizalofop-P-ethyl is 100646-51-3 WLN for quizalofop-ethyl is T66 BN ENJ COR DOY1&VO2& HG WLN for quizalofop-P-ethyl is T66 BN ENJ COR DOY1&VO2& HG &&R Form Nanogens@aol.com

- Daily crowdsourced curation underway – about 40 curation emails per day, 100 identifiers per day removed, approved or added

Multi-level Curation and Approval

NAMES, DATABASE IDs AND SYNONYMS

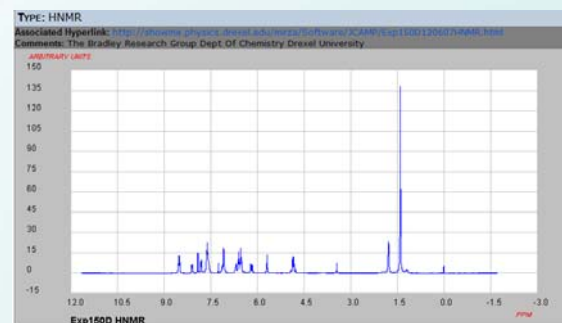
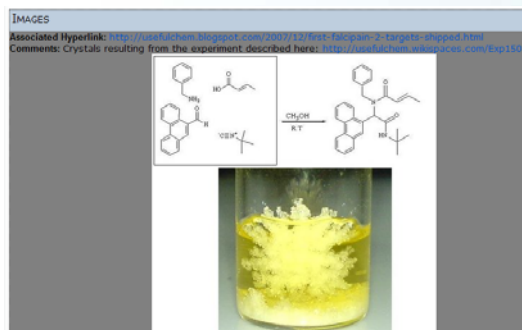
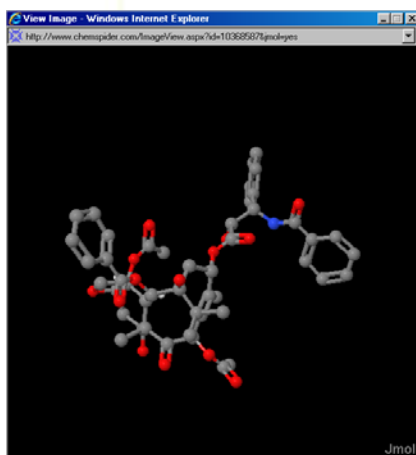
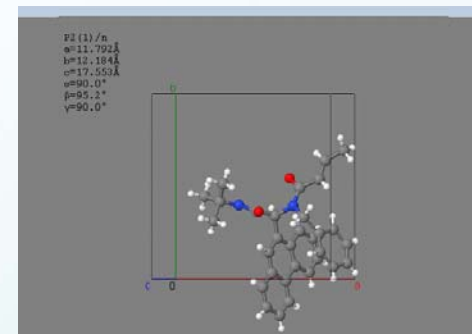
Legend: **Validated by Experts**, Validated by Users, Non-Validated, ~~Removed by Users~~, ~~Redir~~
Approved by Experts



- Galantamine** [\[Wiki\]](#)
- 3-Methoxy-11-methyl-5,6,9,10,11,12-hexahydro-4aH-[1]benzofuro[3a,3,2-ef][2]benzazepin-6-ol**
- 357-70-0** [\[RN\]](#)
- 6H-benzofuro[3a,3,2-ef][2]benzazepin-6-ol, 4a,5,9,10,11,12-hexahydro-3-methoxy-11-methyl-**
- galanthamine** [\[Wiki\]](#)
- Lycoremine**
- ~~6H-Benzofuro(3a,3,2-ef)(2)benzazepin-6-ol, 4a,5,9,10,11,12-hexahydro-3-methoxy-11-methyl-~~
- ~~6H-Benzofuro(3a,3,2-ef)(2)benzazepin-6-ol, 4a,5,9,10,11,12-hexahydro-3-methoxy-11-methyl-, (4aS-(4a.alpha.,6.beta.,8aR*))~~
- ~~6H-Benzofuro[3a,3,2-ef][2]benzazepin-6-ol, 4a,5,9,10,11,12-hexahydro-3-methoxy-11-methyl-, [4aS-(4a.alpha.,6.beta.,8aR*)]~~
- BAS 01832168
- Epigalanthamin
- Galantamin
- Jilkon
- Lycoremin
- NCI60_000004
- NSC100058
- ~~Galanthamine,(3.alpha.-~~

Post Comments, Add Data

- Ability to curate and add to the database .
 - Add structures
 - Add data (spectra, CIFs, images)
 - Add links to other pages (URLs)
 - Add publication details



ChemSpider – Research in Progress

- ChemSpider for the purpose of online virtual screening
- Applying descriptors of various types to filter a database of 20 million compounds
- In progress:
 - Utilizing SimBioSys' LASSO Descriptor
 - Collaboration based on ECCR's ChemModLab

LASSO

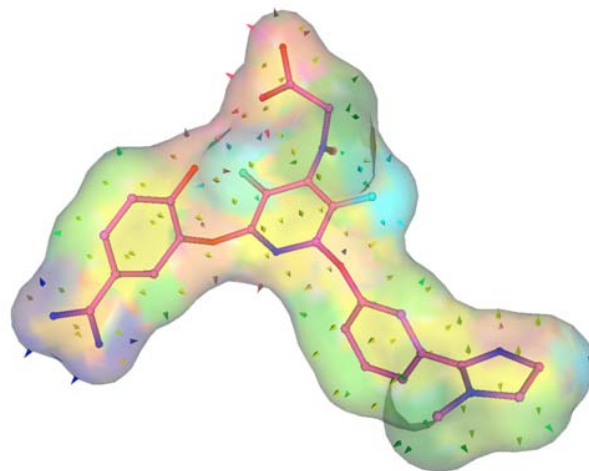
Ligand Activity by Surface Similarity Order

- LASSO uses 23 kinds of Interactive Surface Point Descriptors and
 - is conformation independent
 - screens at 1 million structures/min
 - is proven to enrich screened databases
 - provides scaffold hopping

- Hbond Donors (5 kinds)
- Acceptors (5 kinds)
- Ambivalent H donor/acceptor
- Aromatic Pi-stacking (5 kinds)
- Hydrophobic (3 kinds)
- Metal ions
- Misc (Sulfur, Halogens)

ISPT descriptor for 1FJS ligand:

0	4	0	0	1	0	4	6	1	0	0	0	8	8	0	0	23	5	2	2	0	6	0
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	----	---	---	---	---	---	---



<http://dx.doi.org/10.1007/s10822-007-9164-5>

LASSO Descriptors on ChemSpider

- 40 target receptors chosen
 - From the Database of Useful Decoys dataset
 - <http://dud.docking.org/>
 - Brian Shoichet, UCSF
- Wide range of receptor classes
- Each target family had 10s-100s of known actives
- Actives used as query files for LASSO
- LASSO similarity descriptors generated across all 40 targets and deposited on ChemSpider



LASSO Descriptors on ChemSpider

⊞ SIMBIOSys LASSO

Descriptors: 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 3, 8, 2, 0, 17, 0, 1, 1, 0, 3, 0, 0

Category	Target	PDB Code	LASSO Score
Other Enzymes	PARP, poly(ADP-ribose) polymerase	1efy	0.74
Kinases	PDGFRb, platelet derived growth factor receptor kinase	N/A	0.70
Other Enzymes	COX-2, cyclooxygenase-2	1cx2	0.21
Kinases	EGFr, epidermal growth factor receptor	1m17	0.09
Nuclear Hormone Receptors	PPARg, peroxisome proliferator activated receptor	1fm9	0.08
Other Enzymes	HIVRT, HIV reverse transcriptase	1rt1	0.02
Nuclear Hormone Receptors	RXRa, retinoic X receptor R	1mvc	0.02
Kinases	P38 MAP, P38 mitogen activated protein	1kv2	0.02
Kinases	VEGFR2, vascular endothelial growth factor receptor	1vr2	0.02
Serine Proteases	FXa, factor Xa	1f0r	0.02
Metalloenzymes	PDE5, phosphodiesterase 5	1xp0	0.02

Main Page | Recent changes | Edit this page | Page history
 Printable version | Disclaimers | Privacy policy

Languages: Deutsch

Poly ADP ribose polymerase

(Redirected from **PARP**)

Poly (ADP-ribose) polymerase (PARP) is a **protein** involved in a number of c...

Contents [hide]

- Members of PARP family
- Functions
 - Role in forming polymer of ADP-ribose (PAR)
 - Role in repairing DNA nicks
 - Role of tankyrases
- External links

1efy

CRYSTAL STRUCTURE OF THE CATALYTIC FRAGMENT OF POLY (ADP-RIBOSE) POLYMERASE COMPLEXED WITH A BENZIMIDAZOLE INHIBITOR

White, A.W., Alessio, K., Calvert, A.H., Curtin, N.J., Griffin, B.J., Hostomsky, Z., Mangley, K., Newell, D.R., Srinivasan, S., Golding, B.T.

White, A.W., Alessio, K., Calvert, A.H., Curtin, N.J., Griffin, B.J., Hostomsky, Z., Mangley, K., Newell, D.R., Srinivasan, S., Golding, B.T. (2005) Resistance-modifying agents: II. Chemical and biological properties of benzimidazole inhibitors of the DNA repair enzyme poly(ADP-ribose) polymerase. *J Med Chem.* 48: 4954-4957 [Abstract]

Deposition: 2000-02-10 Release: 2001-01-17

Experimental Method: Top. X-RAY DIFFRACTION Data file

Parameters	Resolution	R-value	R-free	Space Group
2.20	0.202 (004)	0.274		P 2 ₁ 2 ₁ 2 ₁

Unit Cell	Length (Å)	Angle (°)	Volume (Å ³)	Z
a	58.36	90.00	94.20	90.98
b	90.00	90.00	90.00	90.00
c	94.20	90.00	90.00	90.00
gamma	90.98	90.00	90.00	90.00

Molecular Description: Polymer: 1. Molecule: POLY (ADP-RIBOSE) POLYMERASE. Fragment: CATALYTIC FRAGMENT. Chain: A. EC no.: 2.4.2.30

Images and Visualization: Biological Molecules

Display Options: KEGG, Jmol, MolView, PDB-Viewer, HET-Viewer, MolScribe, Chem3D, All Images

LASSO Searching Method 1 LASSO Searching Method 1

ACE, angiotensin-converting enzyme \geq 0.80
AND ALR2, aldose reductase \geq 0.65 [Add Remove](#)

OPTIONS

[About](#) LASSO

244 hits found in 4.47 seconds

LASSO_SIMILAR(ACE, angiotensin-converting enzyme desc 0.8) AND LASSO_SIMILAR(ALR2, aldose reductase desc 0.65)
AND SingleComponent AND NonIsotopic

- Example question: “What are the top 1000 molecules with LASSO descriptors similar to the actives for the Estrogen Receptor”

LASSO Searching Method 2

Find structure with LASSO Score \geq for and \leq for

Nuclear Hormone Receptors

- AR, androgen receptor
- ER, estrogen receptor; agonist
- ER, estrogen receptor; antagonist
- GR, glucocorticoid receptor
- MR, mineralocorticoid receptor
- PPARg, peroxisome proliferator activated receptor
- PR, progesterone receptor
- RXRa, retinoic X receptor R

Serine Proteases

- FXa, factor Xa
- Thrombin
- Trypsin

Folate Enzymes

- DHFR, dihydrofolate reductase

Kinases

- CDK2, cyclindependent kinase 2
- EGFr, epidermal growth factor receptor
- FGFr1, fibroblast growth factor receptor kinase
- HSP90, human heat shock protein 90
- P38 MAP, P38 mitogen activated protein
- PDGFrB, platelet derived growth factor receptor kinase
- SRC, tyrosine kinase SRC
- TK, thymidine kinase
- VEGFr2, vascular endothelial growth factor receptor

Metalloenzymes

- ACE, angiotensin-converting enzyme
- ADA, adenosine deaminase
- COMT, catechol O-methyltransferase
- PDE5, phosphodiesterase 5

Other Enzymes

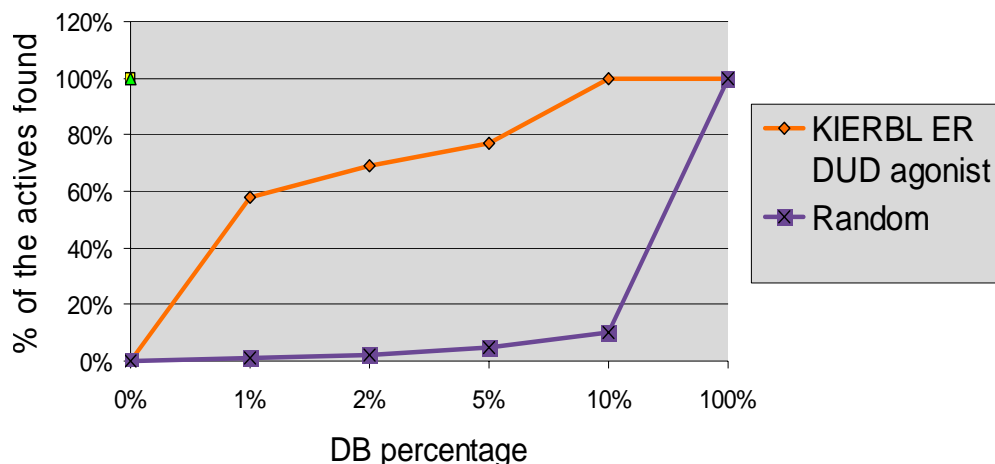
- AChE, acetylcholinesterase

Example for ER enrichment

- Remember : DUD ER training set
- KIERBL Dataset (EPA's DSSTox)
 - Estrogen Receptor Binding K_i values for 50 compounds of environmental relevance: Laws *et. al.* Toxicol Sci. 2006 Nov; 94(1):46-56. Epub 2006 Aug 29
- 15 “binders”: 3-5x weaker the natural ligand 17-beta-estradiol
- 14 million structure subset of LASSO descriptors
- Are known actives recovered?

Enrichment Plot

Enrichment plot
LASSO & ChemSpider tested with ER agonist



- 60% of actives were recovered in the top 1% of the database
- “Environmental binders” are weak binders!
- Top ranked compounds might be active ER binders
- Candidates for experimental investigation?

Data are 1 Week Old

Work to be Done Yet To Validate Further

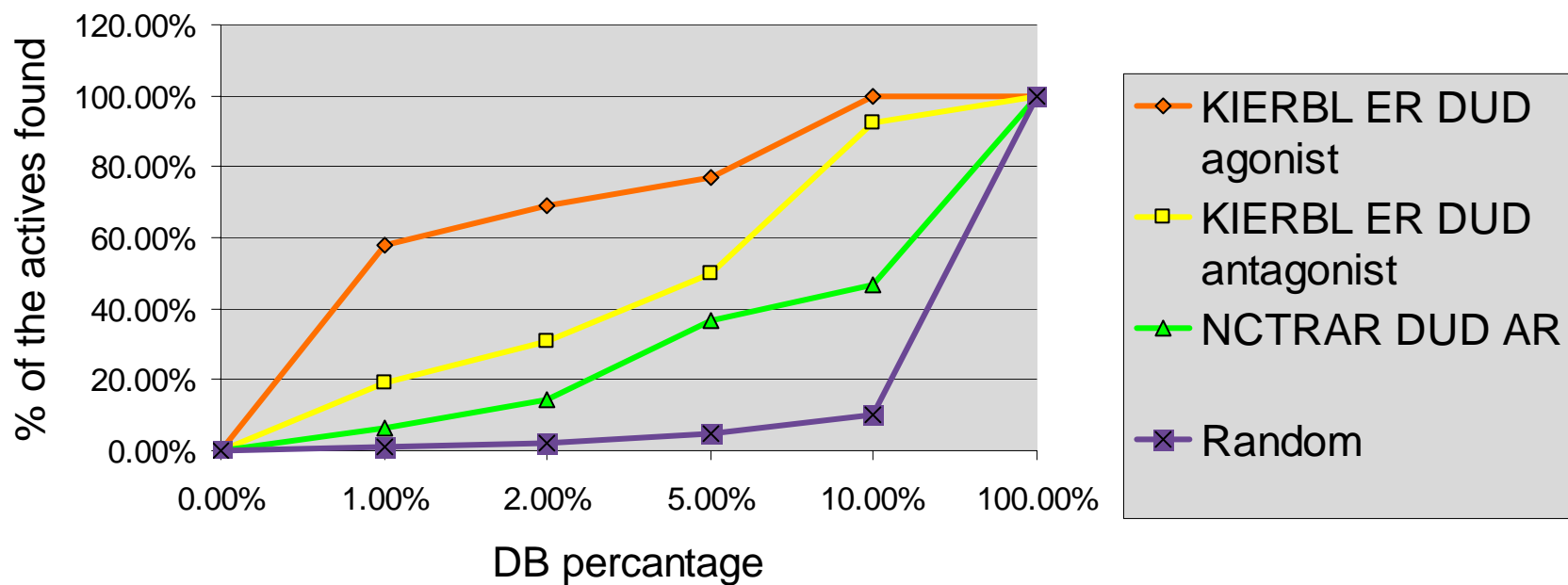
- Run LASSO descriptors on remaining members of database
- Use PhysChem filters at time of Searching (already pre-calculated and in properties)
- Use Structure filters at time of searching
- Use Patent filters at time of searching
- Validate on **real examples** from drug discovery

FDA's NCTR AR (Androgen Receptor)

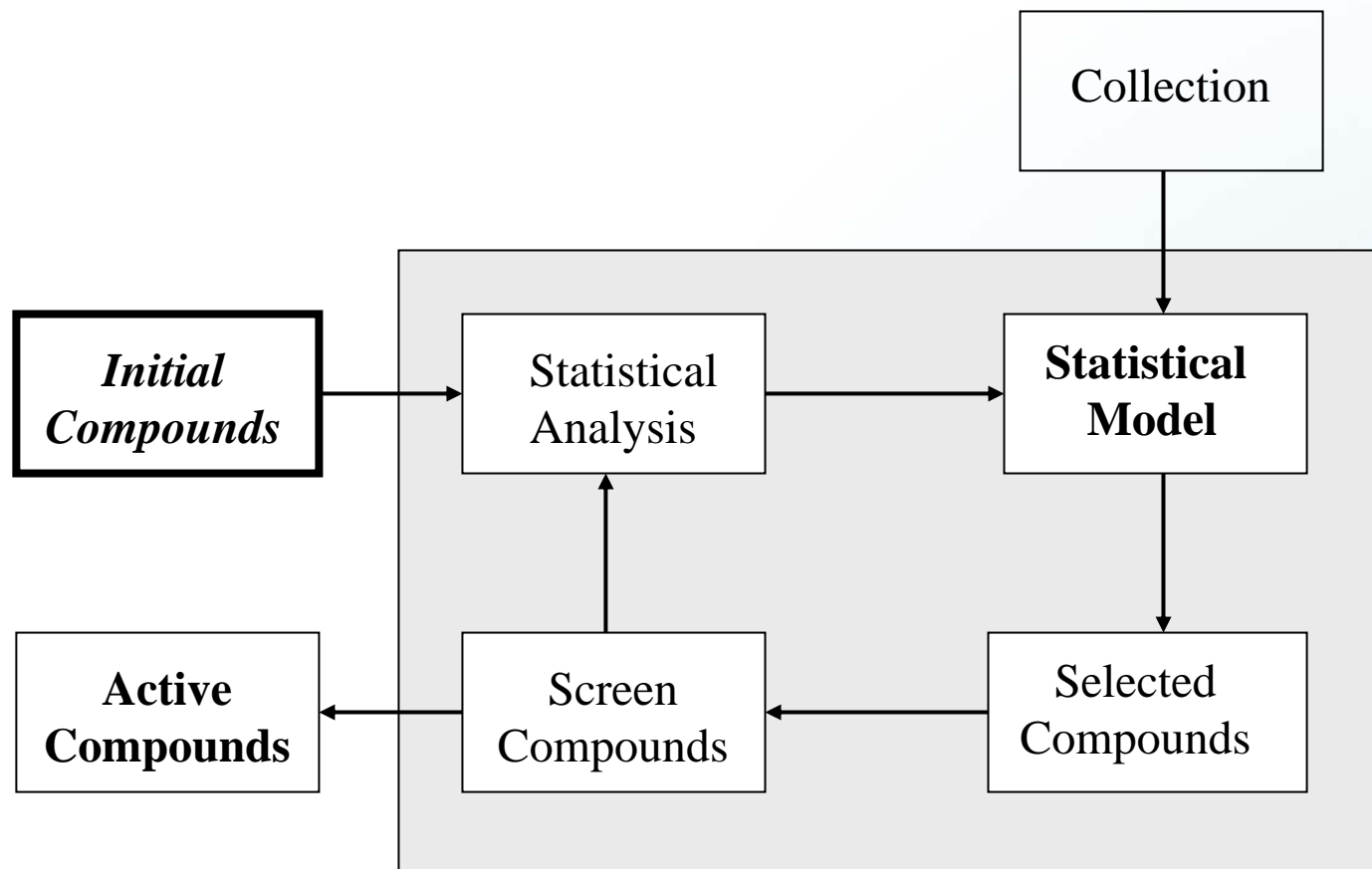
- 203 chemicals with relative binding affinities and activity threshold classes
 - (a) strong
 - (b) moderate
 - (c) weak
 - (d) inactive/non-binding ligands

General Trends?

Enrichment plots: LASSO & ChemSpider tested
with AR + ER



ChemModLab + ChemSpider ECCR at North Carolina State



ChemModLab

Methods include:

- Trees: randomForest, rpart, tree
- Neural networks
- k-nearest neighbors
- Support vector machines
- Partial least squares
- Partial least squares with linear discriminant analysis
- Least angle regression
- Ridge regression
- Elastic net
- Principal components regression
- *Family ensemble of k-nearest neighbors, using 70% selection*
- *Family ensemble of tree, using 70% selection*
- *Family ensemble of rpart, using 70% selection*
- *randomForest using 70% selection*

AID460

The Penn Center for Molecular Discovery

Cathepsin L

57,821 compounds tested

100 nominally active compounds

48 / 100 retested active

Results

2M commercial compounds taken from ChemSpider.

Batches of 50k compounds processed.

42 batches on 12 processors takes about 12 minutes.

Top 1000 predicted compounds selected

Best model was Random Forest

PowerMV, 1000 predicted actives

PowerMV V0.72

File View Analysis Tools Window Help

New Open Table SDF List Text Explorer Home

Explorer 1000_sdf.NoHydrogen(1000) AID460 SD Active noH(100) 729558(10)

Layout Task

Layout

Cell Size 128,128

Column 4

Display (none)

Rotate 0

Search

In (name)

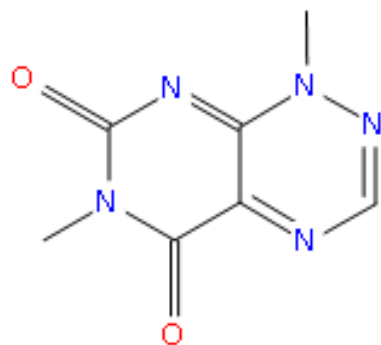
Find Next Find All

High Quality
 Show Attributes
 Show Hydrogen
 Show Title

5307224	1	3240892	2	3245798	3	729563	4
3240265	5	672046	6	5308596	7	934363	8
3369197	9	3000014	10	1220770	11	5308248	12

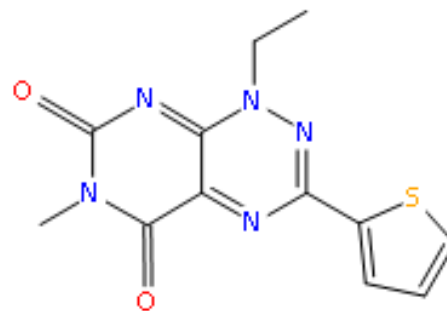
Similarity Set 1

66541



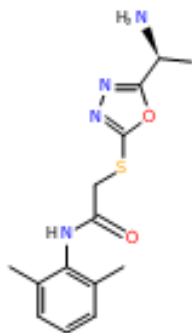
0.000

653297



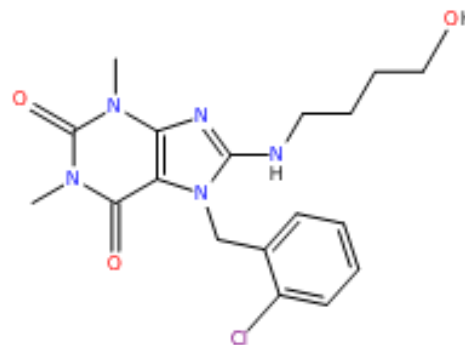
0.528

932345



0.675

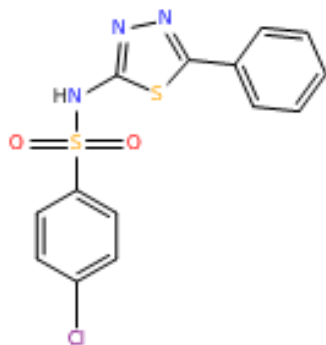
2962700



0.682

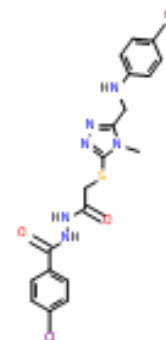
Similarity Set 2

1013432



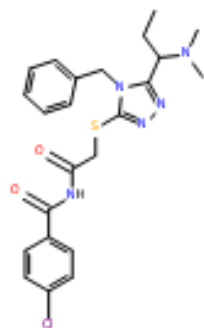
0.000

1147226



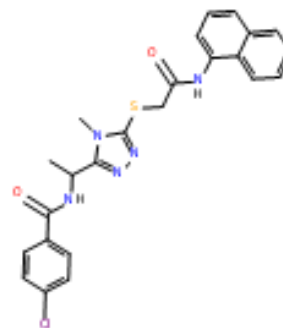
0.500

4882495



0.515

3655137



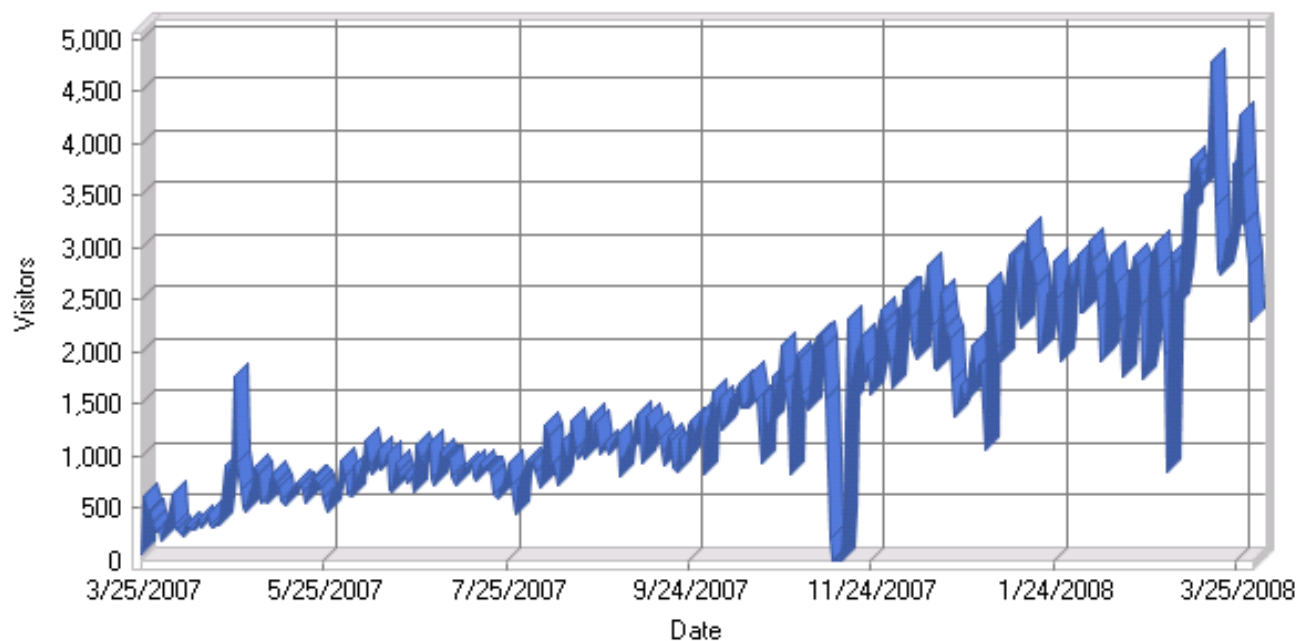
0.519

ChemSpider and ChemModLab

- ChemSpider is a rich compound collection
 - Integrated to patents
 - Integrated to literature
 - Integrated to chemical vendors
- ChemModLab is an online QSAR modeling service
- Mashing together offers great opportunities
- Similar opportunities with UNC-Chapel Hill's Carolina Cheminformatics WorkBench

How Many People Visit ChemSpider?

- 1 year online >4000 unique visitors per day
- >3000 web service calls per day – increasing as more vendors integrate



A Conversation of Possibilities. What if...?

- 20 million structures from vendors, patents and publications **ARE** freely available online?
- This collection **CAN** be filtered according to multiple targets, physchem properties, structure fragments
- Chemical vendor collections **ARE** freely available online, and updated **regularly**?
- Pubmed Central **will soon be** structure searchable?
- Public domain chemistry data **CAN BE** curated, tagged, enhanced in real time online?

Acknowledgments

- LASSO: Rocky Goldsmith and Aniko Simon
- NISS: Stan Young and Jackie Hughes-Oliver
- The ChemSpider team
- The Advisory Group (30 people)
- Commercial Partners (SureChem, OpenEye, ChemAxon, SimBioSys, ACD/Labs, Microsoft)
- Open Source community (Jmol, JSpecView)
- Depositors, Curators and Users

Where to from here? Short term

- Integrated text and structure/substructure searching of the Open Access literature and Pubmed is in development (close)
- Batch-deposition system for large structure files and associated data (close)
- Web-based scraping of structure-based information in development
- Enhanced web services layer for vendors and companies to integrate searches
- Open Notebook Science support
- Deposit latest SureChem Patent Database

Where to from here? Mid-term

- Spidering for Chemistry – extract data from articles, webpages and data sources and stay within copyright
- Deeper integration to text-based searching and conversion of chemical names to structures for online structure searching:
 - Improved integration with NCBI Entrez system
 - Integration to Google Scholar (?)
 - Integration to Microsoft Live Academic (?)