NCATS and GSRS

Danny Katzel
Tyler Peryea
Noel Southall

What is NCATS

- The National Center for Advancing Translational Sciences
- One of 27 Institutes and Centers at the National Institutes of Health (NIH)
- Established to transform the translational process so that new treatments and cures for disease can be delivered to patients faster.

NCATS and GSRS Software Development

- NCATS has 4 software developers working on GSRS everyday
- In constant collaboration with our FDA colleagues and other GSRS Stakeholders
- NCATS drives the development of non-FDA specific GSRS features
- Also helps develop FDA-specific features

NCATS and GSRS Software Development

How Does GSRS help NCATS “transform the translational process so that new treatments and cures for disease can be delivered to patients faster” ?

NCATS projects that use GSRS

Beyond providing a resource to research community that provides robust definitions of the ingredients in marketed and investigational drugs, we also use the software and data to drive the research we do at NCATS.

- Inxight Drugs
- Compound Registration
- NCATS Pharmaceutical Collection
- GARD
- Translator

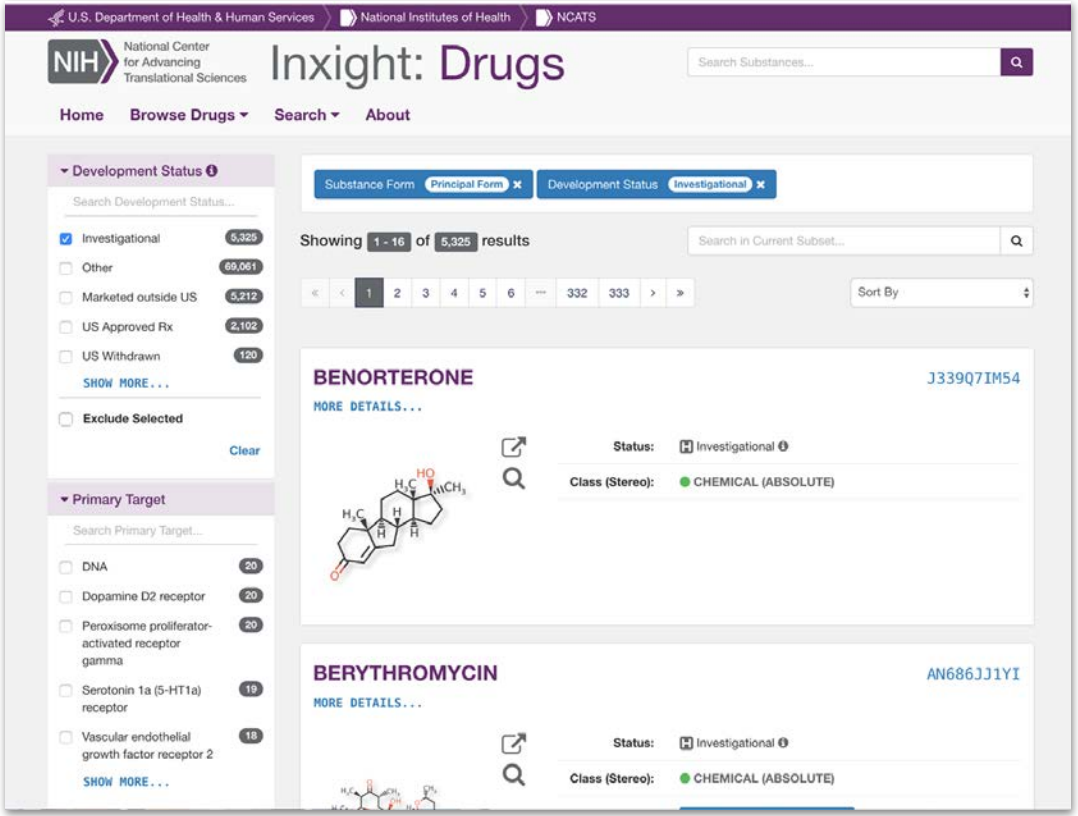
Inxight Drugs

Uses GSRS data and its web application shell to organize drug information from a research perspective at https://drugs.ncats.io/

- Under the covers, this is the same ginas software and data, but
  - customizes the presentation of GSRS information to focus on different aspects of the data
    - group salts, esters into a single ‘primary’ record
  - integrates additional data from external sources, including manually-curated data
  - includes tool substances not found in regulated medical products

drugs.ncats.io

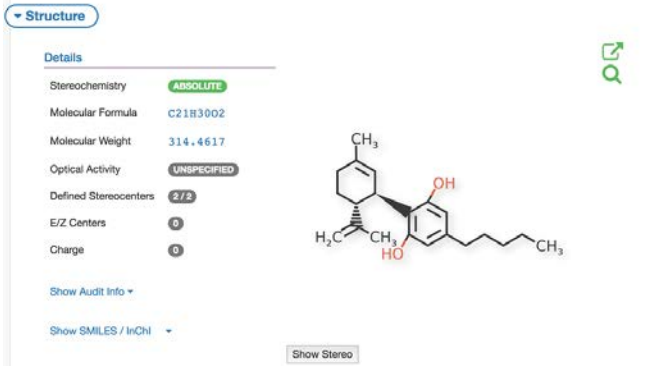
- Customizes GSRS record layout
- Integrates manually-curated data
- Groups different salt forms together into a single record

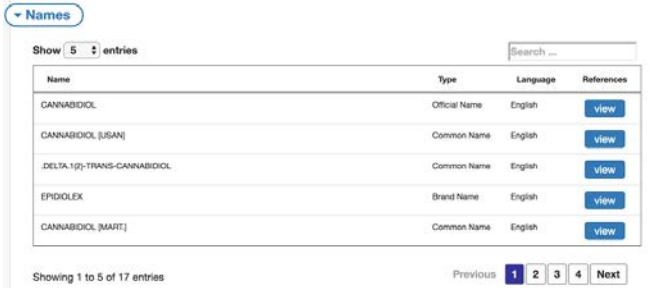


ginas.ncats.nih.gov

Regulatory perspective

- Focus on structure
- Names
- Classifications

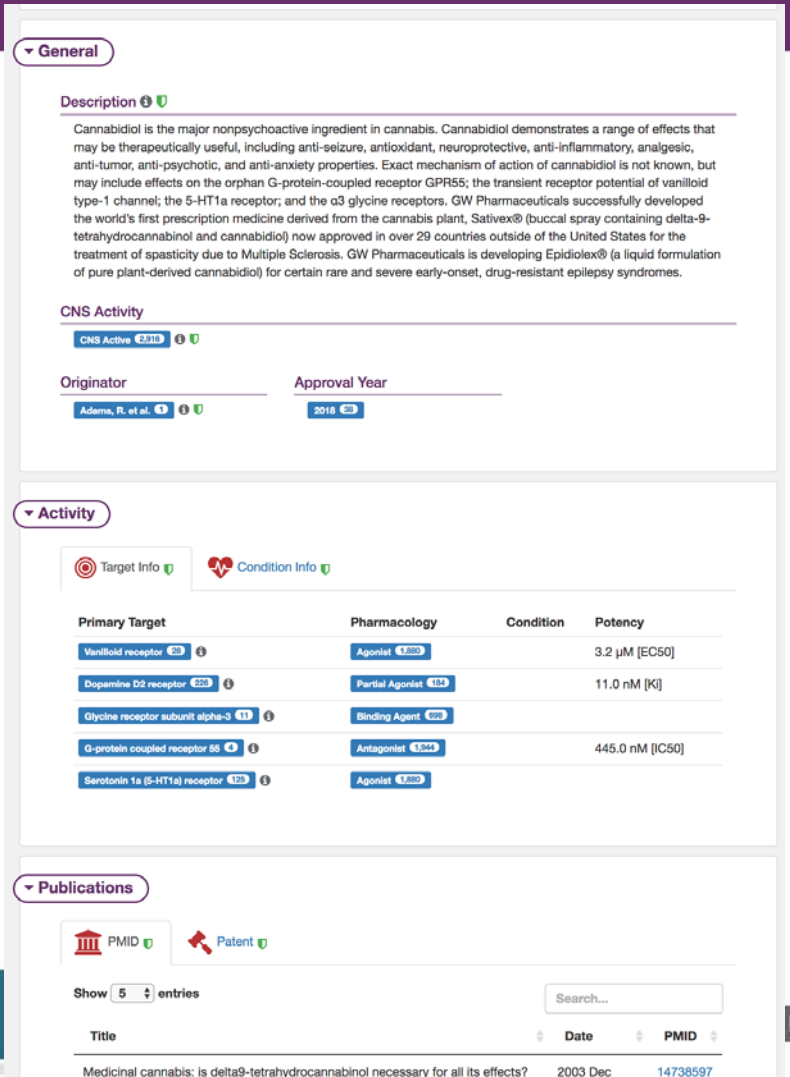




drugs.ncats.io

Focus on

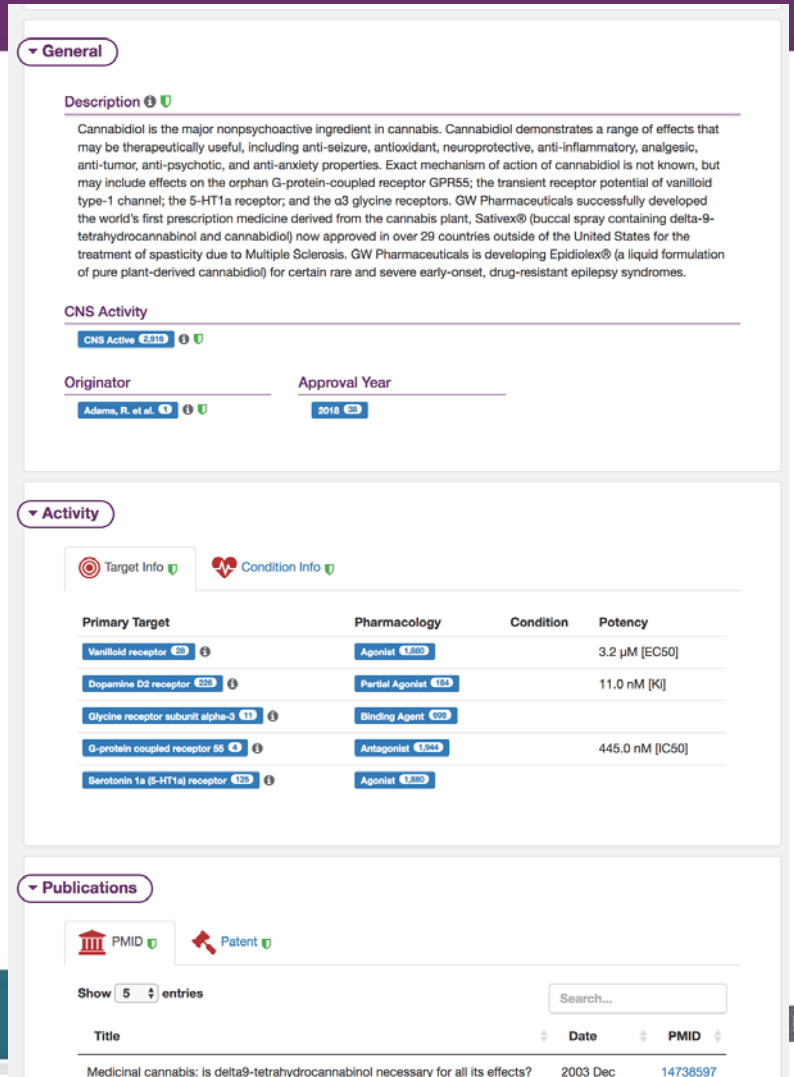
- Targets
- Conditions
- Publications
- Research Uses



drugs.ncats.io

The important thing is that we are not polluting the GSRS data store,

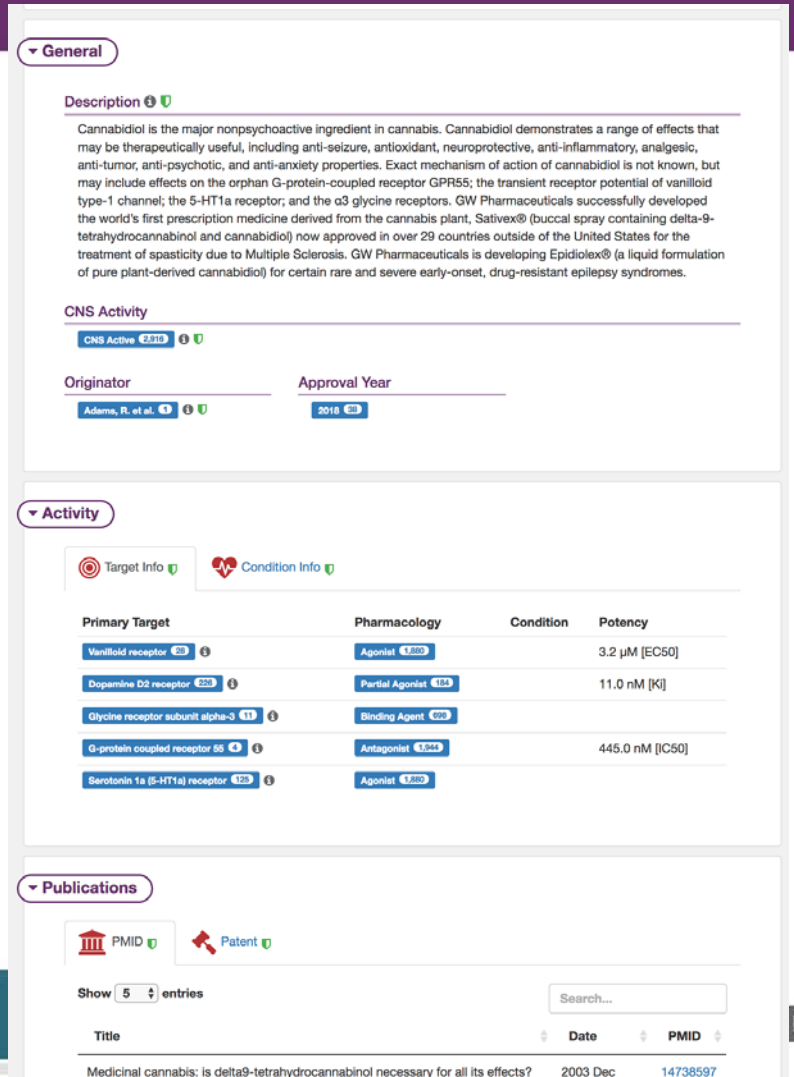
Using UNII to link out to other APIs + services and integrating that other data into this web page on the fly.



NIH National Center for Advancing Translational Sciences

drugs.ncats.io

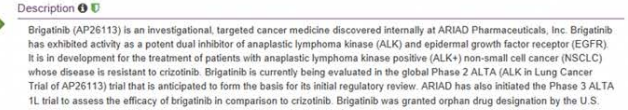
The Little Green shields are data that was manually curated by Rancho Biosciences



Rancho Curation Interface

Comprehensive Manual Curation Effort for Inxight Drugs

- Over **9000** curated substances
- **40 to 200+** total fields per substance
- Concise **descriptions**

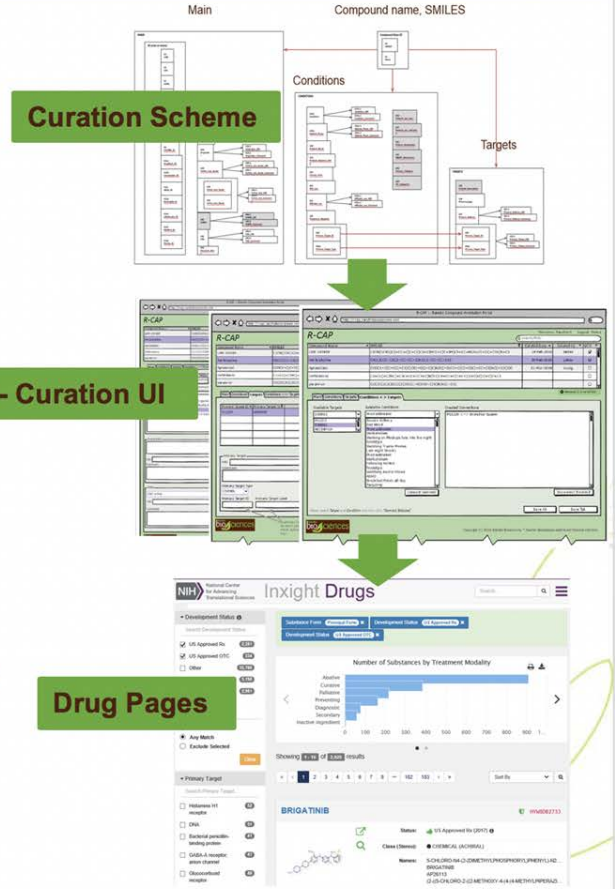

Description

Brigatinib (AP26113) is an investigational, targeted cancer medicine discovered internally at ARIAD Pharmaceuticals, Inc. Brigatinib has exhibited activity as a potent dual inhibitor of anaplastic lymphoma kinase (ALK) and epidermal growth factor receptor (EGFR). It is in development for the treatment of patients with anaplastic lymphoma kinase positive (ALK+) non-small cell cancer (NSCLC) whose disease is resistant to crizotinib. Brigatinib is currently being evaluated in the global Phase 2 ALTA (ALK in Lung Cancer Trial of AP26113) trial that is anticipated to form the basis for its initial regulatory review. ARIAD has also initiated the Phase 3 ALTA 1L trial to assess the efficacy of brigatinib in comparison to crizotinib. Brigatinib was granted orphan drug designation by the U.S.

- **Conditions / Targets**
- Development status / Approved products / Off-label indications
- **Sources** for all data
- Data source for Inxight Drugs (portal for drug development information)
- Example: https://drugs.ncats.io/ginas/app/drug/HYW8DB273J

Compounds Curation and Annotation

- More than 9,000 compounds curated
  - Manual and automated approach
  - Curation Interface
    - > 70 Fields for curation (can add new and modify)
    - Built-in ontologies
    - Built-in QC elements
    - More than 10 curators can work simultaneously
  - Curated data is included in **Inxight Drugs**
  - https://drugs.ncats.io/ginas/app



Rancho BioSciences | www.RanchoBioSciences.com

NCATS projects that use GSRS

Beyond providing a resource to research community that provides robust definitions of the ingredients in marketed and investigational drugs, we also use the software and data to drive the research we do at NCATS.

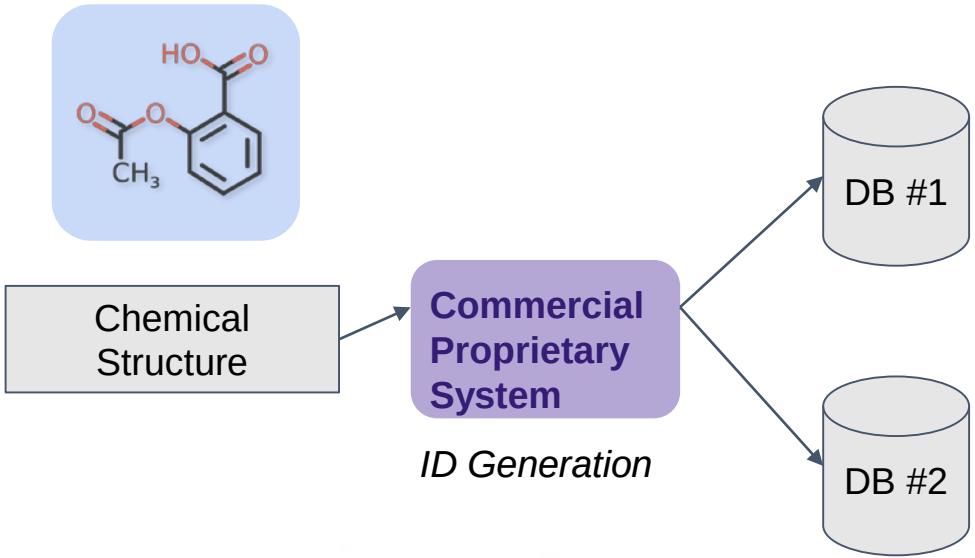
- ~~Inxight Drugs~~
- Compound Registration
- NCATS Pharmaceutical Collection
- GARD
- Translator

NCATS Compound Registration Using GSRS

NCATS has screening libraries of hundreds of thousands of molecules

We have a Compound Registration System to support the high-throughput screening and medicinal chemistry activities that we do on-site.

NCATS Registration Process

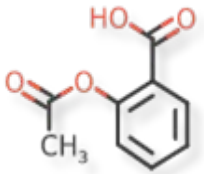
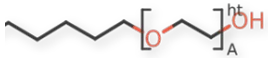


- There are multiple databases where compounds go
- The schema is the same, but are used by different tools/resources
- And have different access restrictions

Commercial Proprietary System vs GSRS

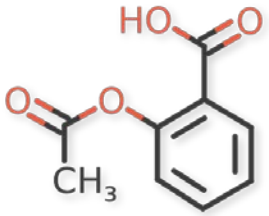
Well-supported in ***both***

Ad hoc support in **CPS**, full support in **GSRS**

Substance Type	Chemical	Polymer	Protein	Nucleic Acid	Structurally Diverse
Defined By	Chemical Structure	Structural Repeat Unit(s)	Amino Acid Sequence(s)	Nucleobase Sequence	Taxonomic Information + Part
Example			>A35X00TA2K RCPGCGQGVQAGCPGGCVEE EDGGSPAEGCAEAEGCLRRE GQECGVYTPNCAPGLQCHPP ...	>303159CVH9 TAAACGTTATAACGTT ATGACGTCAT	Organism Family: CANNABACEAE Organism Genus: CANNABIS Organism Species: SATIVA Author: L. Infraspecific Type: SUBSPECIES Infraspecific Name: SUBSP. SATIVA

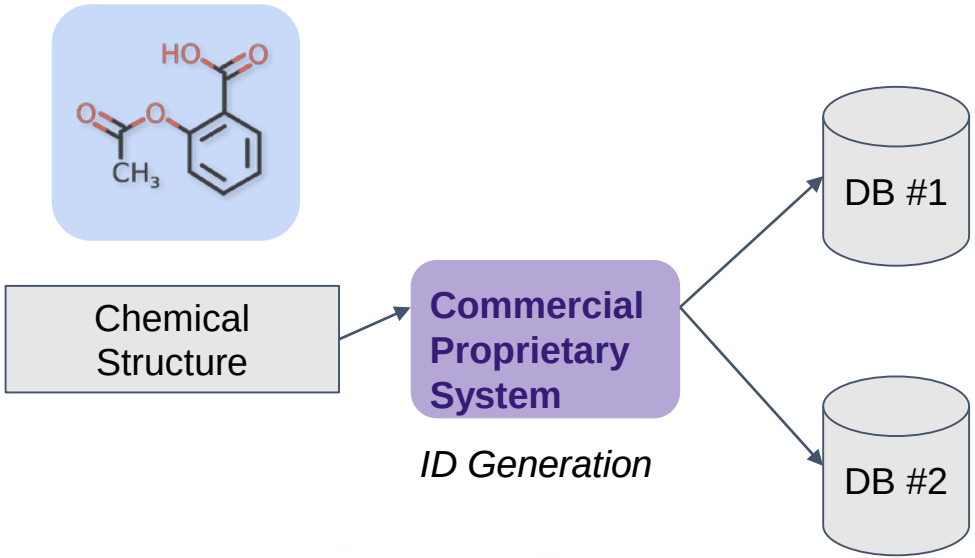
Problems with molecules in Proprietary System:

- Difficulty detecting tautomers, leads to lots of IDs
- Difficulty standardizing structures
- Harder to implement custom rules
- Harder to share best practices when we get it “right”

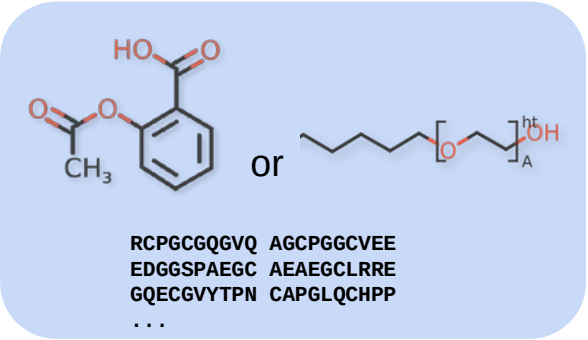


NCGC00015067
NCGC00090977
NCGC00254034
NCGC00259666
NCGC00260723

Legacy NCATS Registration Process



Future NCATS Registration Process



Substance Definition → GSRS Format → **GSRS** → DB #1, DB #2

ID Generation

NCATS projects that use GSRS

Beyond providing a resource to research community that provides robust definitions of the ingredients in marketed and investigational drugs, we also use the software and data to drive the research we do at NCATS.

- ~~Inxight Drugs~~
- ~~Compound Registration~~
- NCATS Pharmaceutical Collection
- GARD
- Translator

NCATS Pharmaceutical Collection

We've also used the GSRS information to help us build a physical collection of the approved drugs for repurposing screens.

- A comprehensive, publicly accessible collection of approved and investigational molecular entities for high-throughput screening
- provides a valuable resource for both validating new models of disease and better understanding the molecular basis of diseases and interventions.

Nearly 2,750 small molecular entities have been approved for clinical use by U.S., European Union, Japanese and Canadian authorities and also are suitable for high-throughput screening. Of these, NCATS currently has 2,500, along with about 1,000 additional investigational compounds, as part of its screening collection.

https://ncats.nih.gov/expertise/preclinical/npc

- The collection already has generated several useful probes for studying a diverse cross section of biology, including novel targets and pathways.
- NCATS provides access to the collection's content through the Therapeutics for Rare and Neglected Diseases program (TRND) and the Toxicology in the 21st Century initiative (Tox21).
- A dedicated online browser enables searching and exporting. NCATS makes regular updates to the browser to improve features and content.

The publication about this work remains the most cited paper from NCATS

How is GSRS used

GSRS data is used to see what approved drugs are missing from our screening library

Do we have the “right” structure for our substance?

How does our structure compare to what GSRS thinks is the correct structure.

NCATS projects that use GSRS

Beyond providing a resource to research community that provides robust definitions of the ingredients in marketed and investigational drugs, we also use the software and data to drive the research we do at NCATS.

- ~~Inxight Drugs~~
- ~~Compound Registration~~
- ~~NCATS Pharmaceutical Collection~~
- GARD
- Translator

Genetic and Rare Disease Information Center

NCATS' most important public website is actually not for research scientists, but for patients.
It brings 10x the traffic of all of our other resources combined.
Nearly 1 in 10 Americans has a rare disease diagnosis, and for many, their first online encounter with disease information is through the clearing houses of disease information that NIH provides.

NCATS is working with GARD to organize their information on rare diseases and treatments, using GSRS to structure their content databases.

https://rarediseases.info.nih.gov/

Chronic myeloid leukemia

Other Names: Chronic granulocytic leukemia; Chronic myelogenous leukemia; CML; See More

Categories: Blood Diseases; Rare Cancers

This disease is grouped under: Chronic myeloproliferative disorders; Myeloid leukemia

Summary

Listen

The following summary is from Orphanet, a European reference portal for information on rare diseases and orphan drugs.

orphanet Inserm

Orpha Number: 521

Disease definition
Chronic myeloid leukaemia (CML) is the most common myeloproliferative disorder accounting for 15-20% of all leukaemia cases.

Epidemiology
Its annual incidence has been estimated at between 1 and 1.5 cases per 100,000 and its prevalence at around 1 in 17,000.

Clinical description
The disease is typically triphasic with a chronic phase (CML-CP), accelerated phase (CML-AP) and blast phase (CML-BP). The majority of patients are diagnosed in the chronic phase and may be either asymptomatic (diagnosed through a routine white blood cell count) or present with fatigue, anaemia, weight loss, night sweats or splenomegaly.

Symptoms

Listen

This table lists symptoms that people with this disease may have. For most diseases, symptoms will vary from person to person. People with the same disease may not have all the symptoms listed. This information comes from a database called the Human Phenotype Ontology (HPO). The HPO collects information on symptoms that have been described in medical resources. The HPO is updated regularly. Use the HPO ID to access more in-depth information about a symptom.

Showing 1-5 of 13 | View All

Medical Terms	Other Names	Learn More: HPO ID
100% of people have these symptoms		
Myeloproliferative disorder		0005547
30%-79% of people have these symptoms		
Abnormal basophil morphology		0001912
Fatigue	Tired [more]	0012378
Fever		0001945
Leukocytosis	Elevated white blood count [more]	0001974

Showing 1-5 of 13 | View All

Do you have more information about symptoms of this disease? We want to hear from you.

Last updated: 11/1/2018

Do you have updated information on this disease? We want to hear from you.

Diagnosis

Listen

Making a diagnosis for a genetic or rare disease can often be challenging. Healthcare professionals typically look at a person's medical history, symptoms, physical exam, and laboratory test results in order to make a diagnosis. The following resources provide information relating to diagnosis and testing for this condition. If you have questions about getting a diagnosis, you should contact a healthcare professional.

Testing Resources

- The Genetic Testing Registry (GTR) provides information about the genetic tests for this condition. The intended audience for the GTR is health care providers and researchers. Patients and consumers with specific questions about a genetic test should contact a health care provider or a genetics professional.

Treatment

Listen

FDA-Approved Treatments

The medication(s) listed below have been approved by the Food and Drug Administration (FDA) as orphan products for treatment of this condition. Learn more orphan products.

- **Bosutinib (Brand name: Bosulif)** - Manufactured by Pfizer Inc.
  FDA-approved indication: Treatment of adult patients with newly-diagnosed chronic phase Philadelphia chromosome-positive chronic myelougenous leukemia. Also for treatment of adult patients with chronic, accelerated or blast phase Philadelphia chromosome-positive (Ph+) chronic myelogenous leukemia (CML) with resistance, or intolerance to prior therapy.
  National Library of Medicine Drug Information Portal
  Medline Plus Health Information

- **Imatinib mesylate (Brand name: Gleevec®)** - Manufactured by Novartis Pharmaceuticals Corp.
  FDA-approved indication: Treatment of chronic myelogenous leukemia
  National Library of Medicine Drug Information Portal

- **Ponatinib (Brand name: Iclusig)** - Manufactured by ARIAD Pharmaceuticals Inc.
  FDA-approved indication: Treatment of adult patients with chronic phase, accelerated phase, or blast phase chronic myeloid leukemia (CML) that is resistant or intolerant to prior tyrosine kinase inhibitor therapy or Philadelphia chromosome positive acute lymphoblastic leukemia (Ph+ALL) that is resistant or intolerant to

Find a Specialist

Listen

If you need medical advice, you can look for doctors or other healthcare professionals who have experience with this disease. You may find these specialists through advocacy organizations, clinical trials, or articles published in medical journals. You may also want to contact a university or tertiary medical center in your area, because these centers tend to see more complex cases and have the latest technology and treatments.

If you can't find a specialist in your local area, try contacting national or international specialists. They may be able to refer you to someone they know through conferences or research efforts. Some specialists may be willing to consult with you or your local doctors over the phone or by email if you can't travel to them for care.

You can find more tips in our guide, How to Find a Disease Specialist. We also encourage you to explore the rest of this page to find resources that can help you find specialists.

Healthcare Resources

- To find a medical professional who specializes in genetics, you can ask your doctor for a referral or you can search for one yourself. Online directories are provided by the American College of Medical Genetics and the National Society of Genetic Counselors. If you need additional help, contact a GARD Information Specialist. You can also learn more about genetic consultations from Genetics Home Reference.

Research

Listen

Research helps us better understand diseases and can lead to advances in diagnosis and treatment. This section provides resources to help you learn about medical research and ways to get involved.

Clinical Research Resources

- ClinicalTrials.gov lists trials that are related to Chronic myeloid leukemia. Click on the link to go to ClinicalTrials.gov to read descriptions of these studies.

  Please note: *Studies listed on the ClinicalTrials.gov website are listed for informational purposes only; being listed does not reflect an endorsement by GARD or the NIH. We strongly recommend that you talk with a trusted healthcare provider before choosing to participate in any clinical study.*

Right now most of the GARD data is manually entered

Using GSRS for its standardized data to help make the system updates more automated /programmatic

NCATS projects that use GSRS

Beyond providing a resource to research community that provides robust definitions of the ingredients in marketed and investigational drugs, we also use the software and data to drive the research we do at NCATS.

- ~~Inxight Drugs~~
- ~~Compound Registration~~
- ~~NCATS Pharmaceutical Collection~~
- ~~GARD~~
- Translator

NCATS Biomedical Data Translator

A research program that is designed to reduce significant barriers between research discovery and clinical trials, in part by organizing the copious amounts of research data currently available.

Working with more than 19 research labs around the country that specialize in biomedical data infrastructure, we are defining a system for helping research scientists work together with data. This initiative is currently in a feasibility testing phase - defining three aspects of what a future system must do.

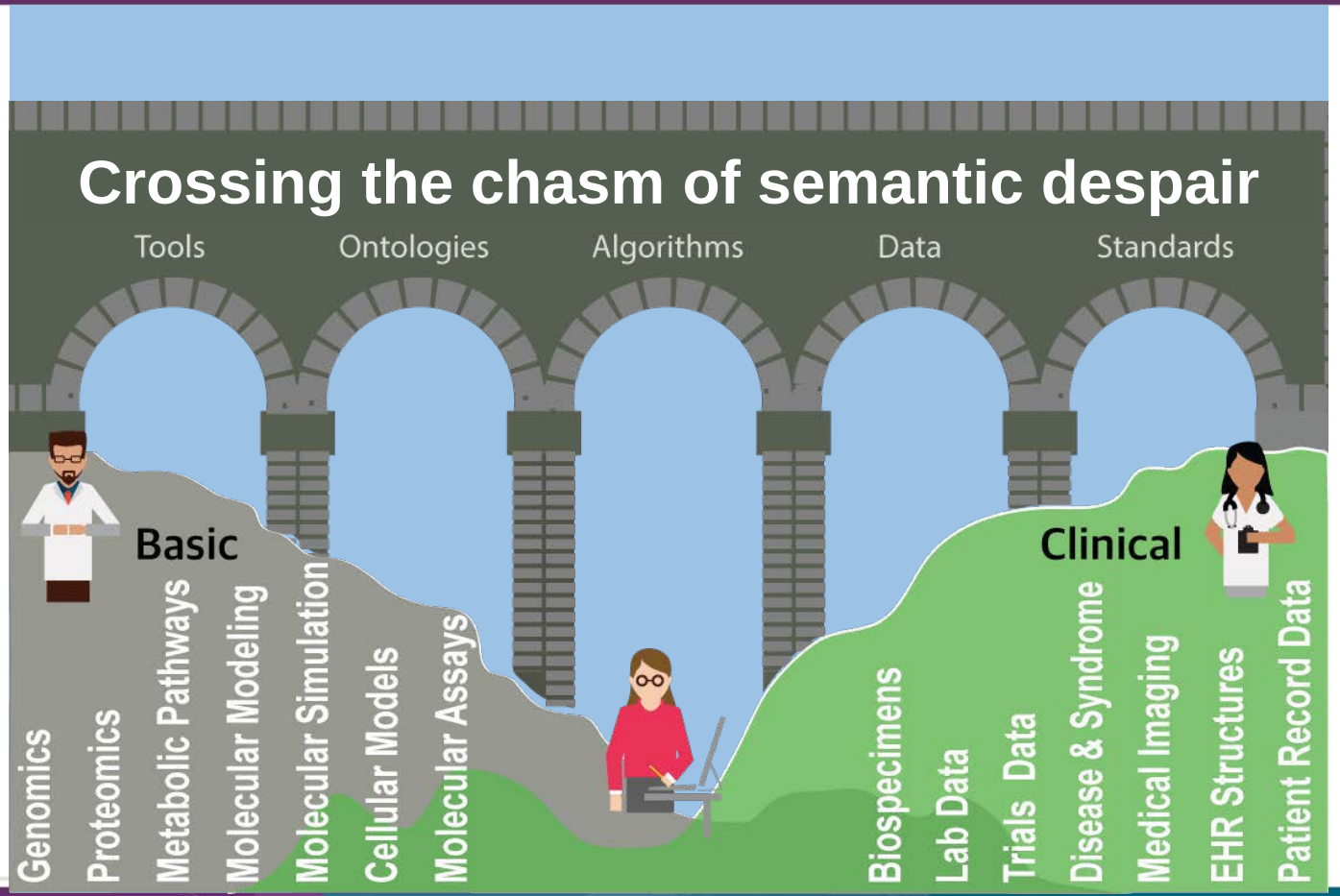


Figure courtesy of Julie McMurry & Chris Mungall

NIH National Center for Advancing Translational Sciences

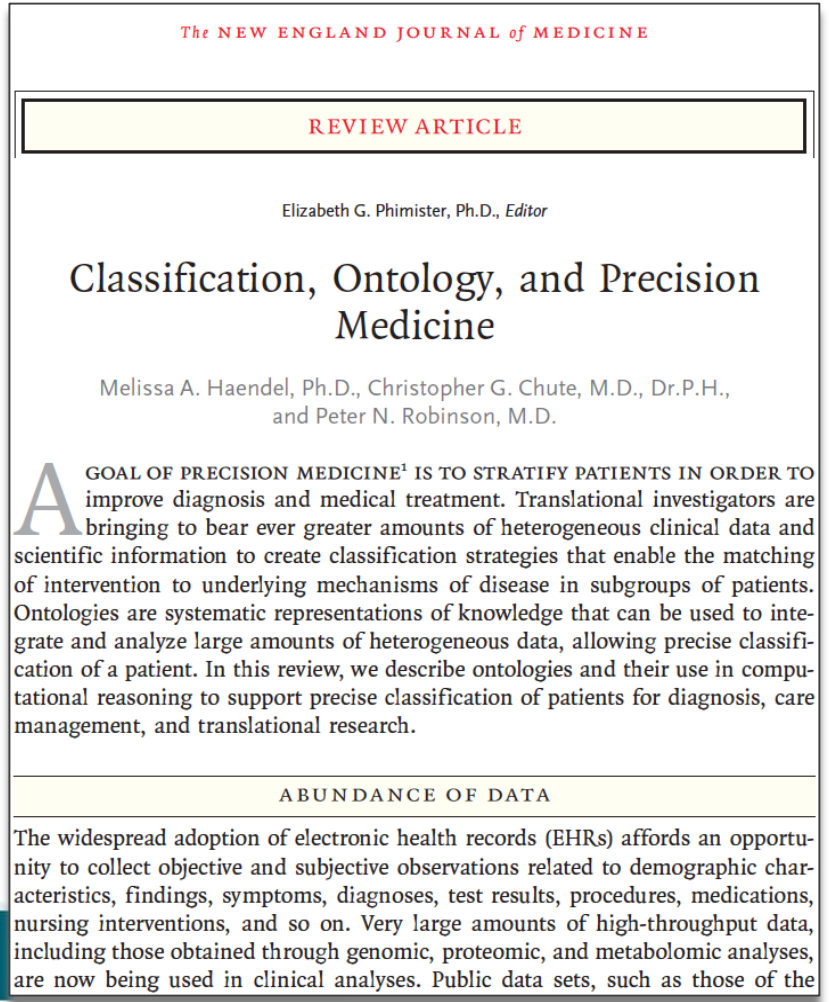
Biomedical Data Translator

- Novel data integration
  - Environment/patient health
  - Molecular/phenotype/disease
  - Drugs/adverse events

- Novel analytic capability
  - Data-driven modeling
  - Probabilistic graphical modeling

Biomedical Data Translator

- Novel data integration
  - Environment/patient health
  - Molecular/phenotype/disease
  - Drugs/adverse events
- Novel analytic capability
  - Data-driven modeling
  - Probabilistic graphical modeling



The NEW ENGLAND JOURNAL of MEDICINE

REVIEW ARTICLE

Elizabeth G. Phimister, Ph.D., *Editor*

Classification, Ontology, and Precision Medicine

Melissa A. Haendel, Ph.D., Christopher G. Chute, M.D., Dr.P.H., and Peter N. Robinson, M.D.

A GOAL OF PRECISION MEDICINE[1] IS TO STRATIFY PATIENTS IN ORDER TO improve diagnosis and medical treatment. Translational investigators are bringing to bear ever greater amounts of heterogeneous clinical data and scientific information to create classification strategies that enable the matching of intervention to underlying mechanisms of disease in subgroups of patients. Ontologies are systematic representations of knowledge that can be used to integrate and analyze large amounts of heterogeneous data, allowing precise classification of a patient. In this review, we describe ontologies and their use in computational reasoning to support precise classification of patients for diagnosis, care management, and translational research.

ABUNDANCE OF DATA

The widespread adoption of electronic health records (EHRs) affords an opportunity to collect objective and subjective observations related to demographic characteristics, findings, symptoms, diagnoses, test results, procedures, medications, nursing interventions, and so on. Very large amounts of high-throughput data, including those obtained through genomic, proteomic, and metabolomic analyses, are now being used in clinical analyses. Public data sets, such as those of the

October 11, 2018
N Engl J Med 2018; 379:1452-1462

Biomedical Data Translator

- Novel data integration
  - Environment/patient health
  - Molecular/phenotype/disease
  - Drugs/adverse events
- Novel analytic capability
  - Data-driven modeling
  - Probabilistic graphical modeling

Citation: Clin Transl Sci (2018) XX, 1; doi:10.1111/cts.12595

EDITORIAL

Deconstructing the Translational Tower of Babel

Christopher P. Austin*, Christine M. Colvis and Noel T. Southall

A principal stumbling block in translation is the compartmentalized nature of data—from biomedical research, disease classifications, health records, clinical trials, and adverse event reports—across diseases and disciplines. These silos impede discovery of commonalities across diseases, and the distinct languages that each discipline uses impede the cross-discipline understanding that is required for efficient translation from basic to clinical to public health science.

By contrast, imagine a world in which researchers had a way to easily access and interrelate these data and languages. Such a tool would accelerate hypotheses about, e.g., which drugs have the potential to treat diseases, the impact of environmental exposures on the onset or worsening of disease; what might be causing illness in patients for whom existing approaches have failed to identify the origin of their symptoms; and better understand the relationships between rare and common diseases. This is the vision of the Biomedical Data Translator: to bridge the current symptom-based diagnosis of disease with research-based molecular and cellular characterizations through an informatics platform that enables interrogation of relationships across the full spectrum of data types, from disease names, clinical signs and symptoms, to organ and cell pathology, genomics, and drug effects.

When we committed to this vision in 2016, we were well aware of its ambitious scope. We, therefore, designed the program to be different in virtually every way from how National Institutes of Health (NIH) research projects are typically competed, supported, and managed, and have taken an explicitly flexible and staged approach to its construction. For the last 24 months, the National Center for Advancing Translational Sciences (NCATS) has been funding a feasibility assessment phase of the Translator, focused on identifying data integration and inclusion barriers and exploring inferential or predictive models that would provide new insights into biology, health, and disease. It was assumed that we did not understand all requirements or needed capabilities when we started, and the platform is being built in an agile way with frequent modifications driven by data from pressure testing using research questions that have been difficult to address by other means. Operationally, the NCATS supports the insights into diseases and possible treatments, and is able to make inferences and predictions even when data are missing. The early results are in, and they are encouraging, as you will read in the articles from the investigators.[1,2]

Two hundred years ago, chemists created a comprehensive enumeration of the elements and systematic relationships among them. This Periodic Table transformed chemistry by placing it on firm scientific footing. We envision the Translator doing the same for translational science.

Funding. No funding was received for this work.

Conflict of Interest. The authors declared no competing interests for this work.

1. Fecho, A. & Ahalt, S. Toward a universal biomedical data translator. *Clin. Transl. Sci.* (in print).
2. Fecho, K. & Clemons, P. The Biomedical Data Translator program: conception, culture, and community. *Clin. Transl. Sci.* (in print).

November 9, 2018
Clin Transl Sci 10.1111/cts.12595


YOUR
DATA
HERE

Conclusion

Our collaboration with FDA on implementing ISO 11238 for regulatory **and research** use has been a fantastic success for NCATS, in advancing our research mission!

… and we hope it will be helpful advancing your own organizational goals as well.

Acknowledgements

NCATS

Tyler Peryea
Noel Southall
Dac-Trung Nguyen
Ivan Grishagin
Dammika Amugoda
Jorge Neyra
Niko Anderson
Mark Williams
Tim Sheils
Chris LeClair
Paul Shinn
Qian Zhu
Christine Colvis
Mark Williams

FDA

Larry Callahan
Frank Switzer
Yulia Borodina
Ramez Ghazzaoui
Elaine Johansen
Ta-Jen Chen
Archana Newatia
Mary-Ann Slack
Alex Welsch
Sarah Stemann
Yoshiyuki Tokiwa
Lavanya Balabhadra

Rancho Biosciences

Laura Brovold
Yulia Skovpen

Medicines Evaluation Board

Herman Diederik
Marcel Hoefnagel
Joris Kampmeijer
Ciska Matai
Burt Kroes

US Pharmacopeial Convention

Fouad Atouf
Andrej Wilk
Tina Morris

Uppsala Montintoring Centre/ WHO

Malin Jakobsson

Health Canada

Vikesh Srivastava

Federal Institute for Drugs and Medical Devices(Germany)

Thomas Balzer

SwissMedic

Philipp Weyermann

European Medicines Agency

Paolo Alcini
Sabine Brosch
Telonis Pangiotis
Llaria Del Seppia

Dow Corning

Katherine Ulman

Royal Botanic Gardens, Kew (UK)

Bob Allkin
Elizabeth Dauncey

EXTRA SLIDES

GSRS Features

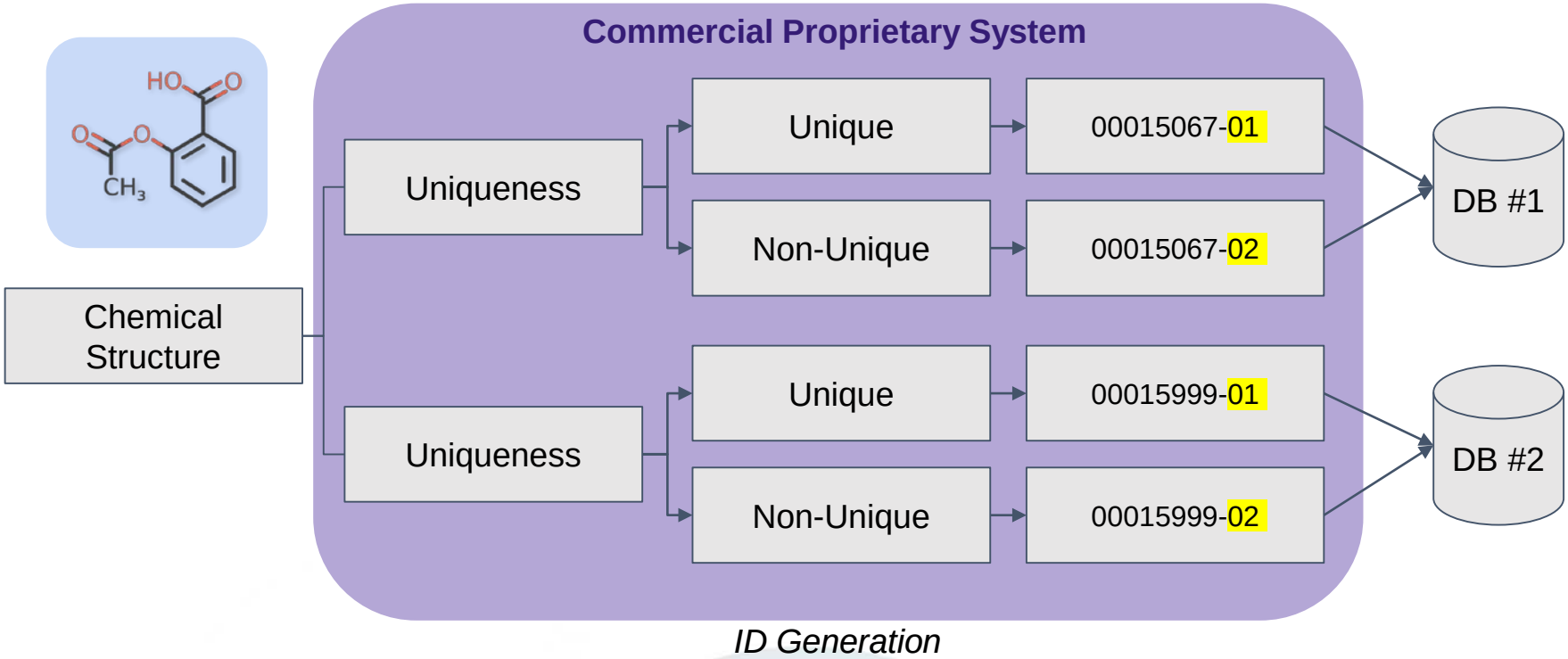
- Assign Unique Identifiers for Substances
- Maintain list of synonyms for each substance across different languages
- Store other text metadata descriptions, alternate IDs and codes
- Store chemical structure

Find Data

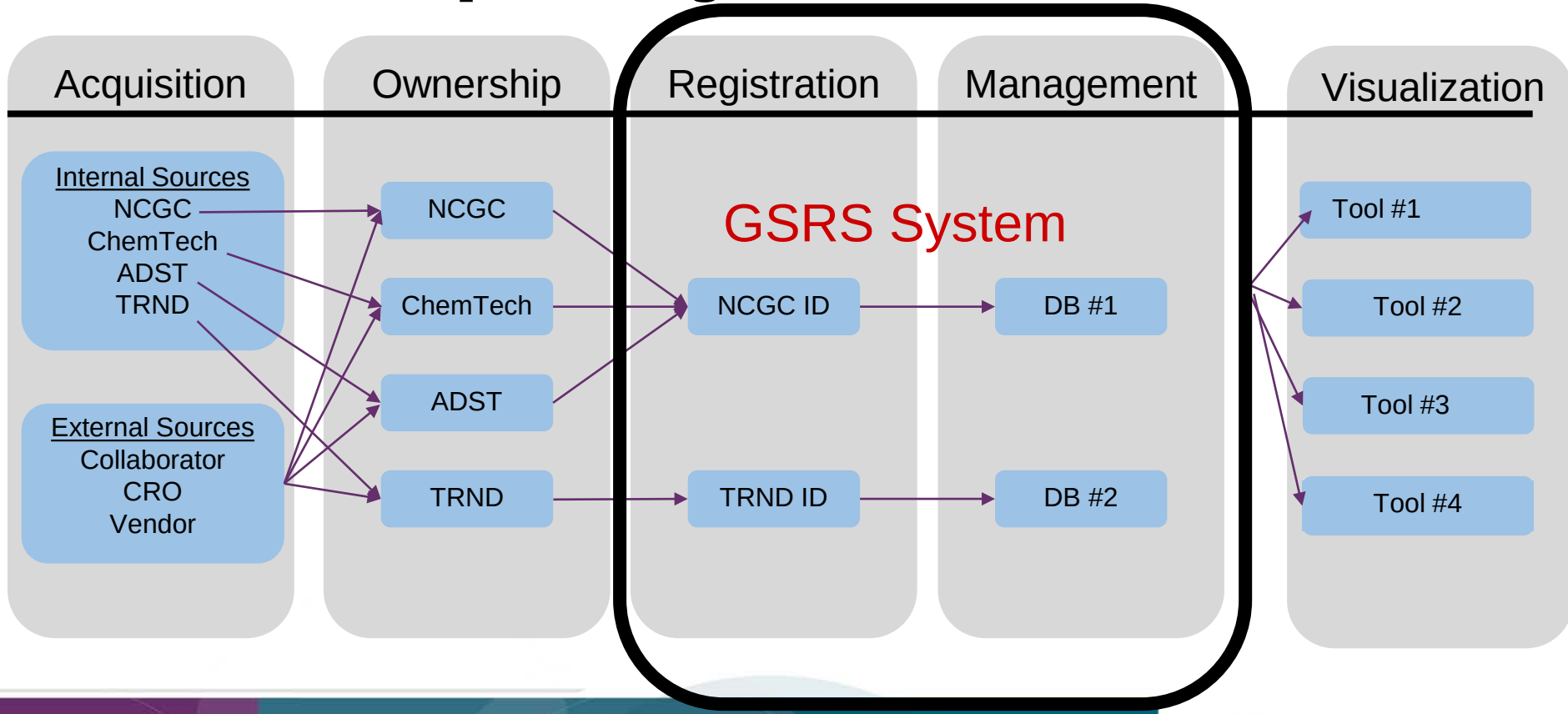
- Support complex searches by names and/or metadata
- Support structure searches
- Support sequence similarity searches

Support Regulation Lifecycle

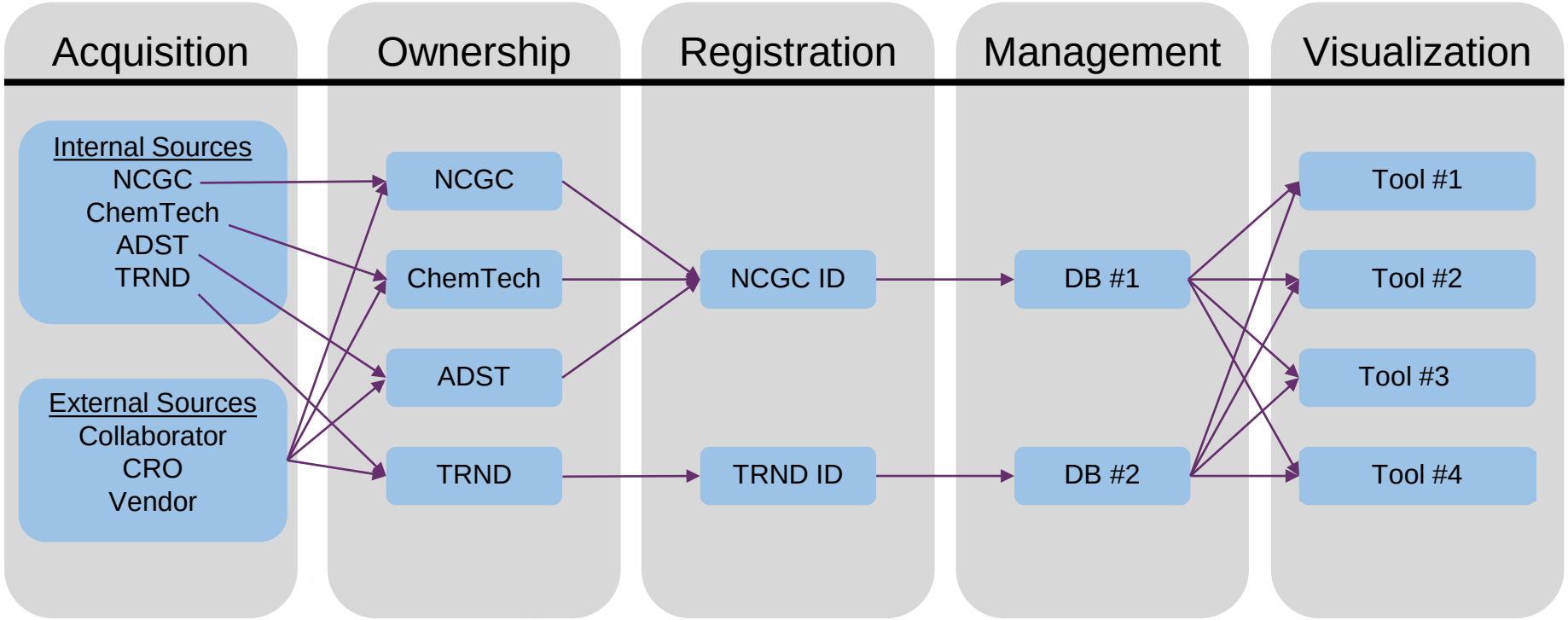
NCATS Registration Process



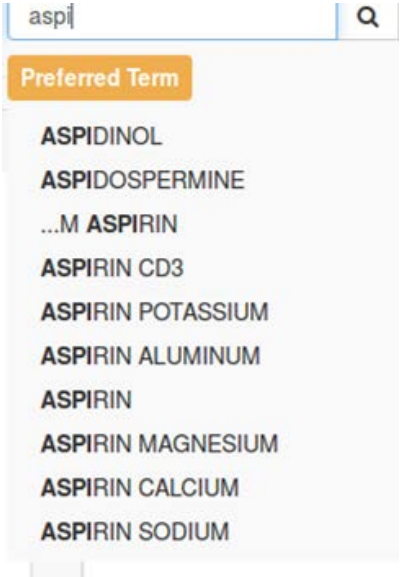
Compound Registration Workflow



Compound Registration Workflow



GSRS Text Search



Structure Search | Sequence Search | Search ...

Browse

Substance Type
- Chemical 36

Molecular Weight
- 0:200 30
- 200:400 19
- 400:600 5

Source Tag
- WARNING 29
- MI 21
- WHO-DD 21
- MART. 11
- INN 8
- USP-RS 7
- VANDF 7
- HSDB 6
- INCI 6
- PH. EUR 5

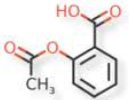
Relationships
- SALT/SOLVATE of PARENT 13
- PARENT of SALT/SOLVATE 10

There is one exact (name or code) match for *"ASPIRIN"*

ASPIRIN

R16CO5Y76E

ACHIRAL



Names: DURLAZA
CLOPIDOGREL/ACETYLSALICYLIC ACID COM...
ASPIRIN COMPONENT OF AXOTAL
ASPIRIN COMPONENT OF EXCEDRIN
ASPIRIN COMPONENT OF TALWIN COMPOUND

Codes: CAS: *50-78-2*

WHO-ATC: *C10BX02* *B01AC06* *C10BX05* *B01AC56* *N02BA71* *C10BX04* *A01AD05* *N02BA01* *N02BA51* *M01BA03* *C10BX01*

CFR: *21 CFR 343.12* *21 CFR 343.13*

DRUG BANK: *DB00945*

Relationships: 27

Formula: C9H8O4

Mol Weight: 180.16

Show All Records Matching Search

Browse

Structure Search | Sequence Search | Search ...

Substance Type
- Chemical 36

Molecular Weight
- 0:200 30
- 200:400 19
- 400:600 5

Source Tag
- WARNING 29
- MI 21
- WHO-DD 21
- MART. 11
- INN 8
- USP-RS 7
- VANDF 7
- HSDB 6
- INCI 6
- PH. EUR 5

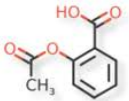
Relationships
- SALT/SOLVATE of PARENT 13
- PARENT of SALT/SOLVATE 10

There is one exact (name or code) match for *"ASPIRIN"*

ASPIRIN

R16CO5Y76E

ACHIRAL



Names: DURLAZA
CLOPIDOGREL/ACETYLSALICYLIC ACID COM...
ASPIRIN COMPONENT OF AXOTAL
ASPIRIN COMPONENT OF EXCEDRIN
ASPIRIN COMPONENT OF TALWIN COMPOUND

Codes: CAS: *50-78-2*

WHO-ATC: *C10BX02 B01AC06 C10BX05 B01AC56 N02BA71 C10BX04 A01AD05 N02BA01 N02BA51 M01BA03 C10BX01*

CFR: *21 CFR 343.12 21 CFR 343.13*

DRUG BANK: *DB00945*

Relationships: 27

Facets

Formula: C9H8O4

Mol Weight: 180.16

Show All Records Matching Search

Sequence Searches

Sequence Query: `ATKAVCVLKGDGPVQ...` (edit search)

2

Sort By: Sort By

SUDISMASE

`OZ9YA0932I`

PROTEIN

Names: SUDISMASE

Codes: CAS: *110294-55-8*
INN: *6193*

Relationships: 1

Subunits: 1

Subunit 1 `6e2b8560-6ce3-467b-a0ab-fb0729579dbf`

```
identity = 1.000
local    = 1.000
sub      = 1.000
matched  = 153
ATKAVCVLKGDGPVQGIINFEQKESNGPVKVWGSIKGLTEGLHGFHVHEFGDNTAGCTSAGPHFNPLSRKHGGPKDEERHVGDLGNVTADKDGVADVSIEDS\
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ATKAVCVLKGDGPVQGIINFEQKESNGPVKVWGSIKGLTEGLHGFHVHEFGDNTAGCTSAGPHFNPLSRKHGGPKDEERHVGDLGNVTADKDGVADVSIEDS\
Target Sites:1_1-1_153
```

LEDISMASE

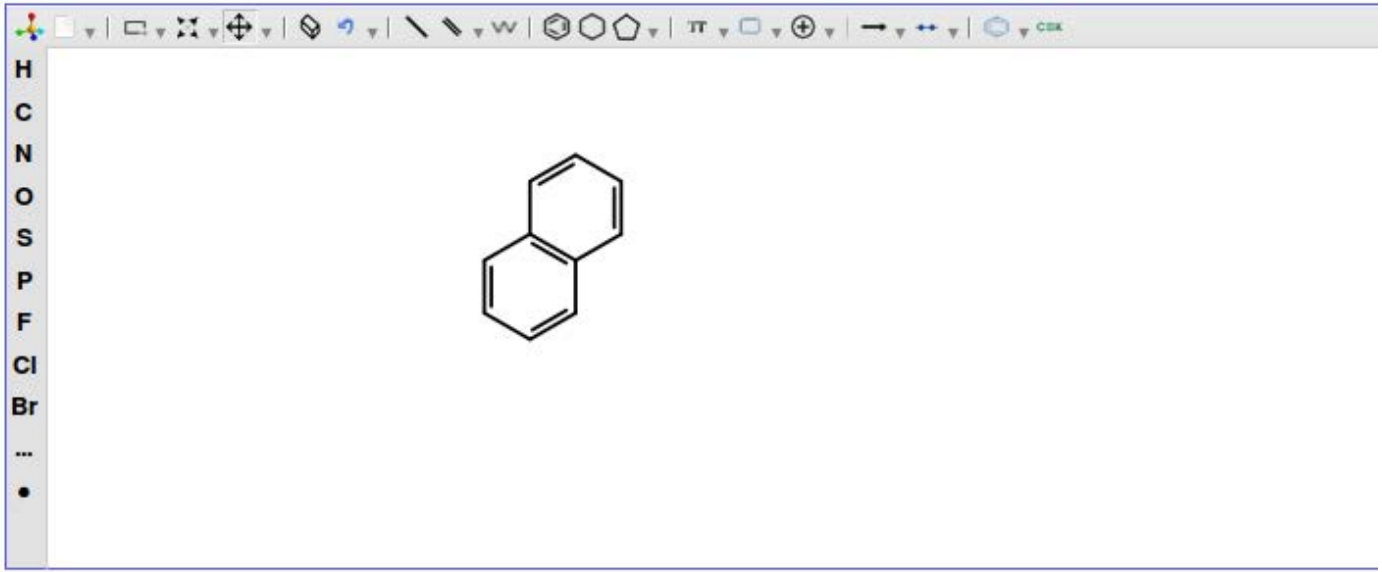
`0786H5RV0T`

PROTEIN

Names: ledismasum
SUPEROXIDE DISMUTASE HUMAN (E. COLI)
SUPEROXIDE DISMUTASE HUMAN (NON-GLY…

Substructure Search

Draw Structure



Search | Substructure

Get Structure From Name

Resolve Name

Substructure Query: `c1ccc2ccccc2c1` (edit search)

Substance Type

- Chemical 1627

Molecular Weight

- 0:200 555
- 200:400 1009
- 400:600 331
- 600:800 141
- 800:1000 89
- >1000 67

Source Tag

- WARNING 734
- MI 327
- WHO-DD 173
- HSDB 143
- INN 141
- USAN 108
- INCI 91
- MART. 89
- PH. EUR 59
- VANDF 50

Relationships

- SALT/SOLVATE of PARENT 195

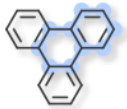
1627 « ‹ 1 2 3 4 5 6 7 8 ··· 101 102 › »

Sort By: Sort By

TRIPHENYLENE

18WX3373I0

ACHIRAL



Similarity `0.947`

Names: TRIPHENYLENE

Codes: CAS: *217-59-4*
ECHA (EC/EINECS): *205-922-9*
MERCK INDEX: *M11183*
MESH: *C009590*

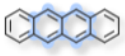
Formula: C18H12

Mol Weight: 228.29

TETRACENE

QYJ5Z6712R

ACHIRAL



Similarity `0.947`

Names: TETRACENE
NAPHTHACENE

Codes: CAS: *92-24-0*
ECHA (EC/EINECS): *202-138-9*
MERCK INDEX: *M7724*
WIKIPEDIA: *TETRACENE*

Formula: C18H12

Browser Add-on Demo

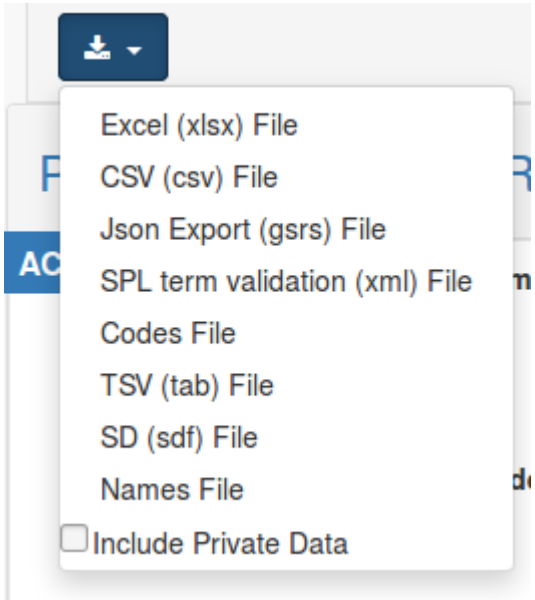
GSRS Challenges

- No Unit Tests in 2015
- Working with FDA on their version
  - FDA has extra proprietary data and additional Databases we don’t have access too
  - FDA wants to integrate their searches against this extra data
  - Can’t Test Data ourselves
  - FDA developers work remotely
- FDA Release and Deploy
  - FDA not only adds their own code to interact with their systems, but they change our code as well.
  - Code is in separate Git Branch - Code was merged a few times a year causing “merge hell” that took days sometimes weeks to resolve.
  - Manually built code on developer’s Windows machine sent to unix server team to deploy.
  - Sometimes manually built code contained files that weren’t committed!

Challenge Accepted

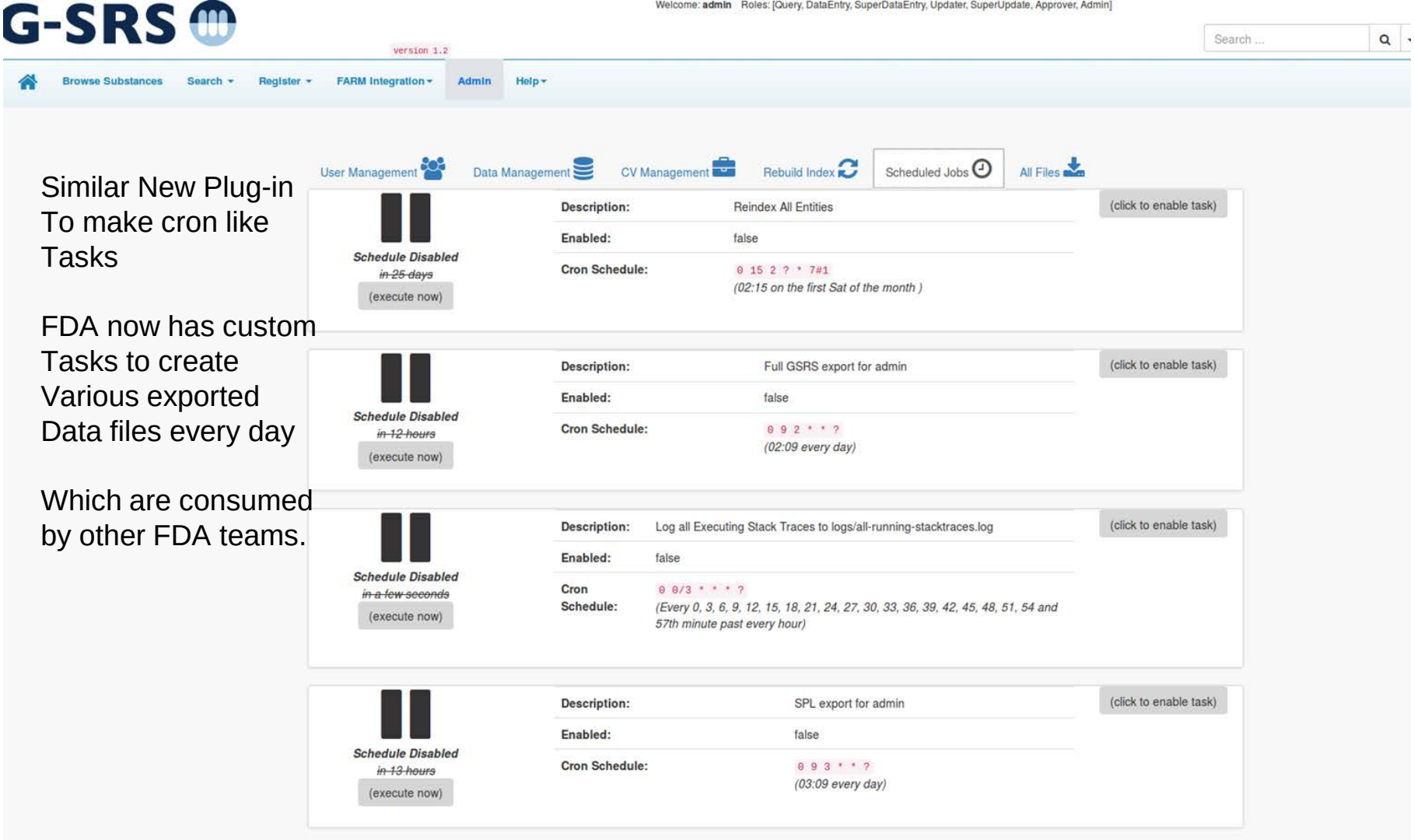
- Automated Testing
  - Wrote 1100+ automated tests (and counting)
  - Set up Jenkins Continuous Integration (CI) to test and deploy instances
- FDA Merge Problems
  - A Junior FDA developer works at NCATS with us 2x/week
  - New Merge Branch where both NCATS and FDA developers push changes to nearly every day. This has eliminated “merge Hell”
  - FDA now builds and deploys directly from git code checkout.
  - **Refactored code to change class and interface designs to a Plugin-in architecture so that FDA changes would touch NCATS code as little as possible.**

Plugable Exports to save search information



- Simple Exporter interface to implement with 1 method
- Corresponding Factory class to create instances and provide parameter options
- Add class names to config file for ginas to find it

FDA DEVELOPERS HAVE WRITTEN THEIR OWN CUSTOM EXPORTERS



Similar New Plug-in
To make cron like
Tasks

FDA now has custom
Tasks to create
Various exported
Data files every day

Which are consumed
by other FDA teams.

Adding Additional Data to GSRS

- Rancho Biosciences has been manually curating additional data about NCATS compounds
- Data includes:
  - Description
  - PubMed Ids
  - Who created it (the Originator)
  - Highest Phase
  - Synonyms
  - Mesh, Chembl, DrugBank Ids etc
  - **Condition information**
  - **Targets**

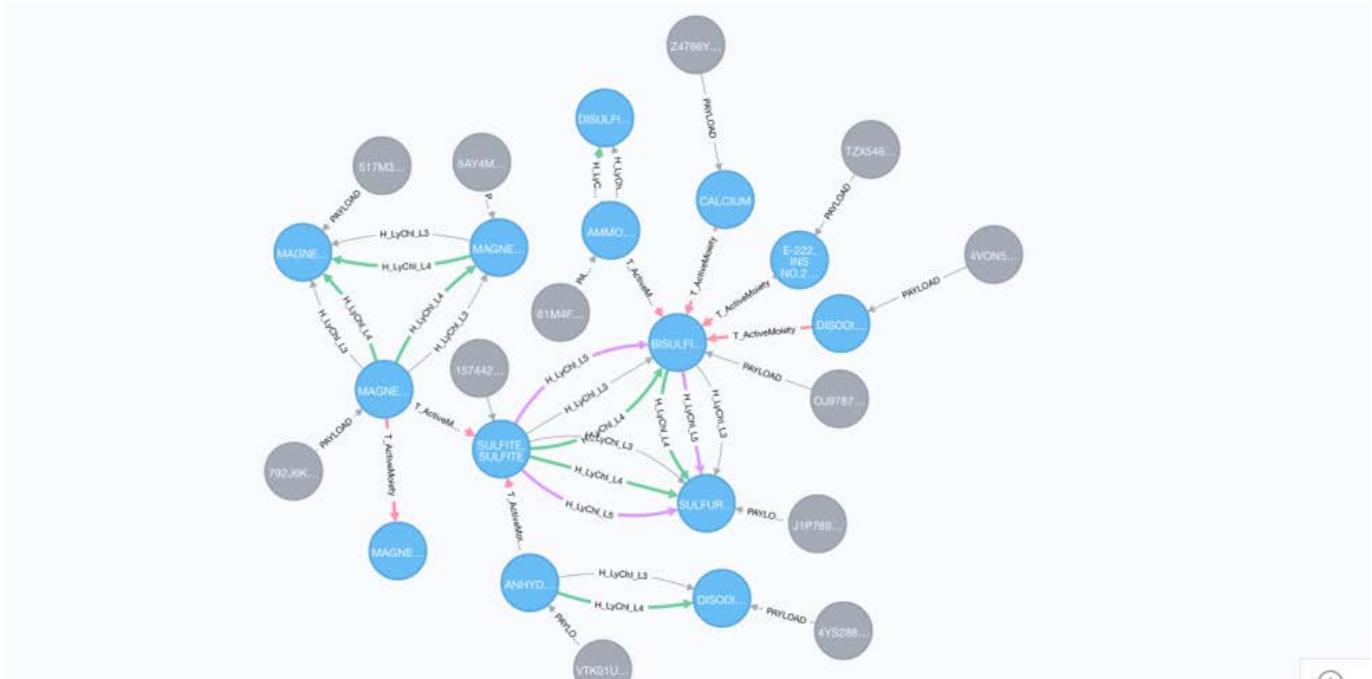
Rancho Curated Compound Data

Over 6,000 manually curated so far

	A	B	C	D	E	
1	CompoundName	CompoundSmiles	Originator	OriginatorUri	OriginatorComment	CompoundDescription
3	BENZOYLECGONINE	CN1[C@H]2CC[C@@H]1[C@H]([C@	Unknown	Unknown		Benzoylecgonine can be found in medical products as
14	BENZOYLECGONINE	CN1[C@H]2CC[C@@H]1[C@H]([C@	Unknown	Unknown		Benzoylecgonine can be found in medical products as
71	ESONARIMOD	CC(=O)SCC(CC(=O)C1=CC=C(C)C=	Taisho Pharmaceutical	http://adisinsight.springer.com/drugs/800002341	Esonarimod, a propio	Esonarimod (KE-298), a derivative of propionic acid d
89	METHENAMINE	C1N2CN3CN1CN(C2)C3	A. Butlerov	http://www.seipub.org/acer/Download.aspx?ID=2657		Methenamine is an antibacterial agent for preventing r
90	METHENAMINE	C1N2CN3CN1CN(C2)C3	A. Butlerov	http://www.seipub.org/acer/Download.aspx?ID=2657		Methenamine is an antibacterial agent for preventing r
92	GANCICLOVIR	NC1=NC2=C(N=CN2COC(CO)CO)C(	Syntex Research	http://shodhganga.inflibnet.ac.in/bitstream/10603/2450/15/15_chapter%2	The first synthesis of	Ganciclovir is a synthetic acyclic nucleoside analogue
93	GANCICLOVIR	NC1=NC2=C(N=CN2COC(CO)CO)C(	Syntex Research	http://shodhganga.inflibnet.ac.in/bitstream/10603/2450/15/15_chapter%2	The first synthesis of	Ganciclovir is a synthetic acyclic nucleoside analogue
58	CANDESARTAN	CCOC1=NC2=C(N1CC3=CC=C(C=C	Takeda, Osaka, Japan	https://www.ncbi.nlm.nih.gov/pubmed/8205603	TCV-116	Candesartan is classified as an angiotensin II recepto
59	CANDESARTAN	CCOC1=NC2=C(N1CC3=CC=C(C=C	Takeda, Osaka, Japan	https://www.ncbi.nlm.nih.gov/pubmed/8205603	TCV-116	Candesartan is classified as an angiotensin II recepto
92	METHAPYRILENE	CN(C)CCN(CC1=CC=CS1)C2=CC=C	Unknown	Unknown		Methapyrilene is an antihistamine and anticholinergic
93	METHAPYRILENE	CN(C)CCN(CC1=CC=CS1)C2=CC=C	Unknown	Unknown		Methapyrilene is an antihistamine and anticholinergic
85	NALIDIXIC ACID	CCN1C=C(C(O)=O)C(=O)C2=C1N=C	George Lesher	http://www.ncbi.nlm.nih.gov/pubmed/15942877		Nalidixic acid is a quinolone antibacterial indicated for
13	BLEOMYCIN	CC[C@@H]1OC(OC(C(NC(=O)C2=C	Umezawa et al (Institute of Microbial	http://onlinelibrary.wiley.com/doi/10.1002/1097-0142(1967)20:5%3C891::AID-CNCR2820200555		Bleomycin is a mixture of cytotoxic glycopeptide antib
14	BLEOMYCIN	CC[C@@H]1OC(OC(C(NC(=O)C2=C	Umezawa et al (Institute of Microbial	http://onlinelibrary.wiley.com/doi/10.1002/1097-0142(1967)20:5%3C891::AID-CNCR2820200555		Bleomycin is a mixture of cytotoxic glycopeptide antib
15	BLEOMYCIN	CC[C@@H]1OC(OC(C(NC(=O)C2=C	Umezawa et al (Institute of Microbial	http://onlinelibrary.wiley.com/doi/10.1002/1097-0142(1967)20:5%3C891::AID-CNCR2820200555		Bleomycin is a mixture of cytotoxic glycopeptide antib
16	BLEOMYCIN	CC[C@@H]1OC(OC(C(NC(=O)C2=C	Umezawa et al (Institute of Microbial	http://onlinelibrary.wiley.com/doi/10.1002/1097-0142(1967)20:5%3C891::AID-CNCR2820200555		Bleomycin is a mixture of cytotoxic glycopeptide antib
17	BLEOMYCIN	CC[C@@H]1OC(OC(C(NC(=O)C2=C	Umezawa et al (Institute of Microbial	http://onlinelibrary.wiley.com/doi/10.1002/1097-0142(1967)20:5%3C891::AID-CNCR2820200555		Bleomycin is a mixture of cytotoxic glycopeptide antib
18	BLEOMYCIN	CC[C@@H]1OC(OC(C(NC(=O)C2=C	Umezawa et al (Institute of Microbial	http://onlinelibrary.wiley.com/doi/10.1002/1097-0142(1967)20:5%3C891::AID-CNCR2820200555		Bleomycin is a mixture of cytotoxic glycopeptide antib
19	BLEOMYCIN	CC[C@@H]1OC(OC(C(NC(=O)C2=C	Umezawa et al (Institute of Microbial	http://onlinelibrary.wiley.com/doi/10.1002/1097-0142(1967)20:5%3C891::AID-CNCR2820200555		Bleomycin is a mixture of cytotoxic glycopeptide antib
20	BLEOMYCIN	CC[C@@H]1OC(OC(C(NC(=O)C2=C	Umezawa et al (Institute of Microbial	http://onlinelibrary.wiley.com/doi/10.1002/1097-0142(1967)20:5%3C891::AID-CNCR2820200555		Bleomycin is a mixture of cytotoxic glycopeptide antib
21	BLEOMYCIN	CC[C@@H]1OC(OC(C(NC(=O)C2=C	Umezawa et al (Institute of Microbial	http://onlinelibrary.wiley.com/doi/10.1002/1097-0142(1967)20:5%3C891::AID-CNCR2820200555		Bleomycin is a mixture of cytotoxic glycopeptide antib
22	BENZYL ALCOHOL	OCC1=CC=CC=C1	Macht	http://www.the-dermatologist.com/article/7191		Benzyl Alcohol is an aromatic alcohol used in a wide v
23	BENZYL ALCOHOL	OCC1=CC=CC=C1	Macht	http://www.the-dermatologist.com/article/7191		Benzyl Alcohol is an aromatic alcohol used in a wide v
49	CEFODIZIME	[H][C@]12SCC(CSC3=NC(C)=C(CC	Unknown	Unknown		Cefodizime is a third generation cephalosporin with a
50	CEFODIZIME	[H][C@]12SCC(CSC3=NC(C)=C(CC	Unknown	Unknown		Cefodizime is a third generation cephalosporin with a
51	CEFODIZIME	[H][C@]12SCC(CSC3=NC(C)=C(CC	Unknown	Unknown		Cefodizime is a third generation cephalosporin with a
52	VITAMIN D	[H][C@@]1(CC[C@@]2([H])C(CCC[	Unknown	Unknown		Cholecalciferol (/ˌkoʊləkælˈsɪfərɒl/) (vitamin D3) is or
53	VITAMIN D	[H][C@@]1(CC[C@@]2([H])C(CCC[	Unknown	Unknown		Cholecalciferol (/ˌkoʊləkælˈsɪfərɒl/) (vitamin D3) is or
61	BENZOYL PEROXIDE	O=C(OOC(=O)C1=CC=CC=C1)C2=C	The Borden company, Limited	http://www.ncbi.nlm.nih.gov/pubmed/14328040		Benzoyl peroxide (BPO) is an organic compound in th
62	BENZOYL PEROXIDE	O=C(OOC(=O)C1=CC=CC=C1)C2=C	The Borden company, Limited	http://www.ncbi.nlm.nih.gov/pubmed/14328040		Benzoyl peroxide (BPO) is an organic compound in th
63	LINEZOLID	CC(=O)NC[C@H]1CN(C(=O)O1)C2=	Pharmacia and Upjohn Company	http://www.google.com/patents/WO1995007271A1?cl=en	Linezolid was discov	Linezolid is an antibiotic used for the treatment of infe
64	LINEZOLID	CC(=O)NC[C@H]1CN(C(=O)O1)C2=	Pharmacia and Upjohn Company	http://www.google.com/patents/WO1995007271A1?cl=en	Linezolid was discov	Linezolid is an antibiotic used for the treatment of infe
65	LINEZOLID	CC(=O)NC[C@H]1CN(C(=O)O1)C2=	Pharmacia and Upjohn Company	http://www.google.com/patents/WO1995007271A1?cl=en	Linezolid was discov	Linezolid is an antibiotic used for the treatment of infe
66	LINEZOLID	CC(=O)NC[C@H]1CN(C(=O)O1)C2=	Pharmacia and Upjohn Company	http://www.google.com/patents/WO1995007271A1?cl=en	Linezolid was discov	Linezolid is an antibiotic used for the treatment of infe
89	ACETYLSULFAMETHOXAZOLE	CC(=O)NC1=CC=C(C=C1)S(=O)(=O	Unknown	Unknown		N-acetyl Sulfamethoxazole is a metabolite of sulfame
90	ACETYLSULFAMETHOXAZOLE	CC(=O)NC1=CC=C(C=C1)S(=O)(=O	Unknown	Unknown		N-acetyl Sulfamethoxazole is a metabolite of sulfame
07	LAMIVUDINE	NC1=NC(=O)N(C=C1)[C@@H]2CS[	Biochem Pharma Inc.	http://www.thepharmaletter.com/article/lamivudine-positive-effects-in-her	http://www.google.co	Lamivudine is a nucleoside reverse transcriptase inhil

All of this Data is stitched Together

Stitcher Code uses a Neo4j graph database and custom algorithms written By Trung to cluster all the datapoints from disparate data sources



GSRS Support for External Additional Data

Work In Progress - New Plugin to support extra data from external datasources not in the GSRS Database

Currently only supports spreadsheet mappings

New Data is Processed by Lucene Indexes and the UI to:

- Enhance user experience
- Find new insights that weren’t apparent before
- UI geared more towards Researchers than Regulators

Structure Search | Sequence Search | Search ...

Current Ginas Release

Substance Type
- Chemical 36

Molecular Weight
- 0:200 30
- 200:400 19
- 400:600 5

Source Tag
- WARNING 29
- MI 21
- WHO-DD 21
- MART. 11
- INN 8
- USP-RS 7
- VANDF 7
- HSDB 6
- INCI 6
- PH. EUR 5

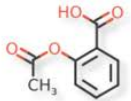
Relationships
- SALT/SOLVATE of PARENT 13
- PARENT of SALT/SOLVATE 10

There is one exact (name or code) match for *"ASPIRIN"*

ASPIRIN

R16CO5Y76E

ACHIRAL



Names: DURLAZA
CLOPIDOGREL/ACETYLSALICYLIC ACID COM...
ASPIRIN COMPONENT OF AXOTAL
ASPIRIN COMPONENT OF EXCEDRIN
ASPIRIN COMPONENT OF TALWIN COMPOUND

Codes: CAS: *50-78-2*

WHO-ATC: *C10BX02* *B01AC06* *C10BX05* *B01AC56* *N02BA71* *C10BX04* *A01AD05* *N02BA01* *N02BA51* *M01BA03* *C10BX01*

CFR: *21 CFR 343.12* *21 CFR 343.13*

DRUG BANK: *DB00945*

Relationships: 27

Formula: C9H8O4

Mol Weight: 180.16

Show All Records Matching Search

Structure Search | Sequence Search | Search ...

Current Ginas Release

Substance Type
- Chemical 36

Molecular Weight
- 0:200 30
- 200:400 19
- 400:600 5

Source Tag
- WARNING 29
- MI 21
- WHO-DD 21
- MART. 11
- INN 8
- USP-RS 7
- VANDF 7
- HSDB 6
- INCI 6
- PH. EUR 5

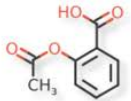
Relationships
- SALT/SOLVATE of PARENT 13
- PARENT of SALT/SOLVATE 10

There is one exact (name or code) match for *"ASPIRIN"*

ASPIRIN

R16CO5Y76E

ACHIRAL



Names: DURLAZA
CLOPIDOGREL/ACETYLSALICYLIC ACID COM...
ASPIRIN COMPONENT OF AXOTAL
ASPIRIN COMPONENT OF EXCEDRIN
ASPIRIN COMPONENT OF TALWIN COMPOUND

Codes: CAS: *50-78-2*

WHO-ATC: *C10BX02 B01AC06 C10BX05 B01AC56 N02BA71 C10BX04 A01AD05 N02BA01 N02BA51 M01BA03 C10BX01*

CFR: *21 CFR 343.12 21 CFR 343.13*

DRUG BANK: *DB00945*

Relationships: 27

Formula: C9H8O4

Mol Weight: 180.16

Facets

Show All Records Matching Search

Just basic chemical info +IDs
No additional data

Collapse Filters

Stitcher Type

- Stitcher Parent 42085
- Stitcher Child 37863

Rule of Five (JCHEM)

Bioavailability (JCHEM)

Primary Target

- Histamine H1 receptor 171
- Bacterial penicillin-binding protein 166
- Dopamine D2 receptor 141
- Glucocorticoid receptor 138
- Bacterial 70S ribosome 123

More ...

Highest Phase

- Approved 4431
- Approved (off-label) 1465
- Phase II 733
- Phase III 532
- Preclinical 508

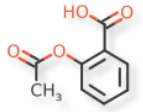
85092 « ‹ 1 2 3 4 5 6 7 8 ··· 5318 5319 › »

Sort By: Sort By

ASPIRIN

R16C05Y76E

ACHIRAL



Alternative Definitions: 1

Names: DURLAZA
CLOPIDOGREL/ACETYLSALICYLIC ACID COM...
ASPIRIN COMPONENT OF AXOTAL
ASPIRIN COMPONENT OF EXCEDRIN
ASPIRIN COMPONENT OF TALWIN COMPOUND

Codes: CAS: *50-78-2*

WHO-ATC: *C10BX02 B01AC06 C10BX05 B01AC56 N02BA71 C10BX04 A01AD05 N02BA01 N02BA51 M01BA03 C10BX01*

CFR: *21 CFR 343.12 21 CFR 343.13*

DRUG BANK: *DB00945*

Relationships: 28

Formula: C9H8O4

Mol Weight: 180.16

Description

Aspirin is a nonsteroidal anti-inflammatory drug. Aspirin is unique in this class of drugs because it irreversibly inhibits both COX-1 and COX-2 activity by acetylating a serine residue (Ser529 and Ser516, respectively) positioned in the arachidonic acid-binding channel, thus inhibiting the synthesis of prostaglandins and reducing the inflammatory response. The drug is used either alone or in combination with other compounds for the treatment of pain, headache, as well as for reducing the risk of stroke and heart attacks in patients with brain ischemia and cardiovascular diseases.

Ginas with
Rancho Stitched Data

Collapse Filters

Stitcher Type

- Stitcher Parent 42085
- Stitcher Child 37863

Rule of Five (JCHEM)

Bioavailability (JCHEM)

Primary Target

- Histamine H1 receptor 171
- Bacterial penicillin-binding protein 166
- Dopamine D2 receptor 141
- Glucocorticoid receptor 138
- Bacterial 70S ribosome 123

More ...

Highest Phase

- Approved 4431
- Approved (off-label) 1465
- Phase II 733
- Phase III 532
- Preclinical 508

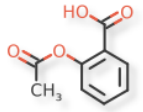
85092 « ‹ 1 2 3 4 5 6 7 8 ··· 5318 5319 › »

Sort By: Sort By

ASPIRIN

R16C05Y76E

ACHIRAL



Alternative Definitions: 1

Names: DURLAZA
CLOPIDOGREL/ACETYLSALICYLIC ACID COM...
ASPIRIN COMPONENT OF AXOTAL
ASPIRIN COMPONENT OF EXCEDRIN
ASPIRIN COMPONENT OF TALWIN COMPOUND

Codes: CAS: *50-78-2*

WHO-ATC: *C10BX02 B01AC06 C10BX05 B01AC56 N02BA71 C10BX04 A01AD05 N02BA01 N02BA51 M01BA03 C10BX01*

CFR: *21 CFR 343.12 21 CFR 343.13*

DRUG BANK: *DB00945*

Relationships: 28

RanchoTargets

Formula: C9H8O4

Weight: 180.16

Rancho Phase

Description

Aspirin is a nonsteroidal anti-inflammatory drug. Aspirin is unique in this class of drugs because it irreversibly inhibits both COX-1 and COX-2 activity by acetylating a serine residue (Ser529 and Ser516, respectively) positioned in the arachidonic acid-binding channel, thus inhibiting the synthesis of prostaglandins and reducing the inflammatory response. The drug is used either alone or in combination with other compounds for the treatment of pain, headache, as well as for reducing the risk of stroke and heart attacks in patients with brain ischemia and cardiovascular diseases.

Ginas with
Rancho Stitched Data

GSRS Technical Information

- Java Application using Play 2.3 Framework
- Lucene Indexes for Text Searches and Faceting
- Relational Database using Ebean for ORM
- Front End
  - Play Scala Templates (Twirl)
  - Javascript - JQuery/ Angular 1/ Angular 2
- Back End
  - Oracle / MySql Relational Database (~ 12 GB of data for 85K Substances)
  - Lucene Indexes (4.7 GB)

Compound Standardization

Structure Standardization algorithm

1. Tidy up structure (e.g., fix valence, kekulize)
2. Break structure into individual components
3. For each component
   a) Apply “business” rules based on a combination of InChI & FDA guidelines (e.g., neutralize)
   b) Generate a canonical tautomeric form based on a tautomer “force field”
   c) Check if component is a salt or solvent via a lookup table; if yes and salt removal flag is on, remove component.
4. Fuse components
5. Postprocessing
   a) (de-) protonation
   b) mobile charges
   c) stereo perception

Hash key generation

- Multiple resolution similar to InChI's and CACTVS/NCI's hash keys
- Four layers
    1. Topology (more coarse than InChI's /c layer)
    2. Topology + atom labels (tautomer insensitive)
    3. Topology + atom labels + bond order
    4. Full structure
- Unlike InChI, hashes for the layers are not independent; they are chained together.
- Final hash is encoded as a string over the 32-character alphabet ({A-Z} ∨ {1-9}) \ {E,I,O}
- Lexicographically meaningful

Grouping compound bioassay results

1. Compounds are grouped by parent (sumatriptan, sumatriptan succinate)
    1. Compounds with more than one parent are left alone (mixtures)
    2. Link back to compound/sample structure preserved

Name Standardization approach

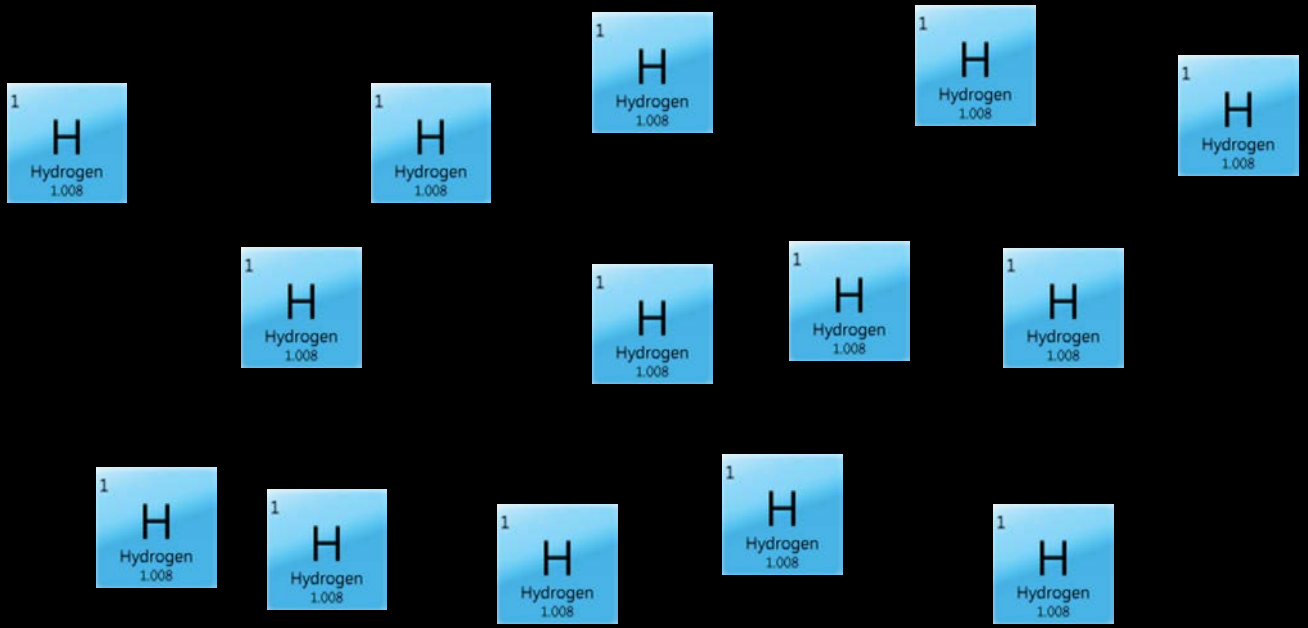
1. Tidy up name (e.g., CAPS, greek, hyphens/spaces)
2. Allow only one compound per name
   a) Rank by depositor (~lowest CID)
3. Choosing Preferred Name; from lowest unsalted CID
   a) Nonproprietary names preferred
   b) CAS / IUPAC allowed if have to
   c) Internal codes recognized and used only as last report
4. Name-Structure conflicts
   a) Assume name is ‘correct’ -> map *sample* to current standard structure
   b) Send list of conflicts back to depositor for clarification

Beginning of “stuff” presentation

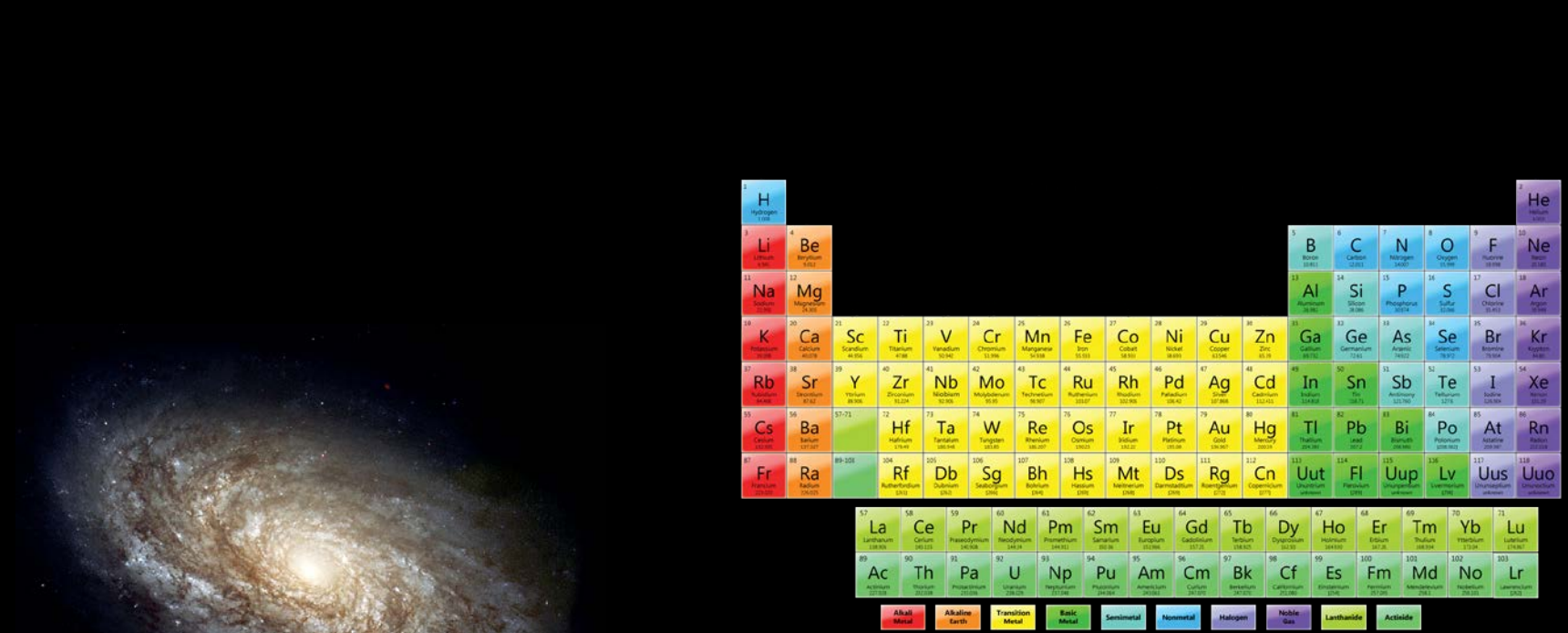
A Brief History of Stuff

13.8 Billion Years ago

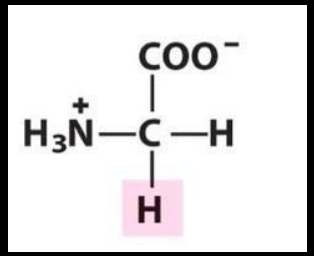
13.8 Billion Years ago
(no stuff)

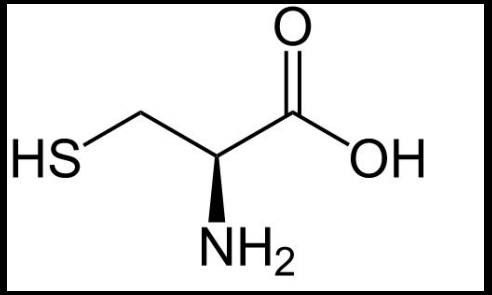


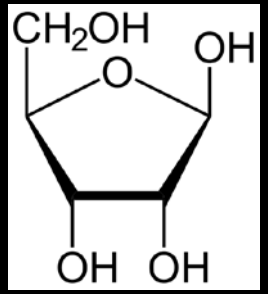
13.7 Billion Years ago
(lots of stuff, but it’s all the same)



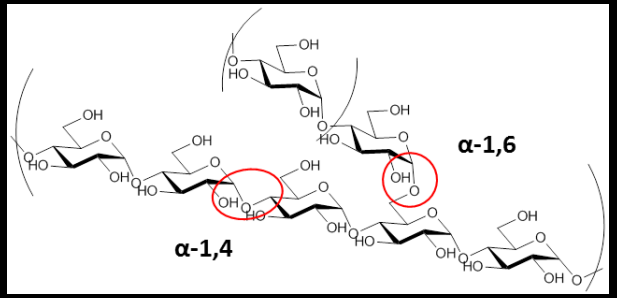
~10 Billion Years ago
(small amounts of *very* simple stuff)

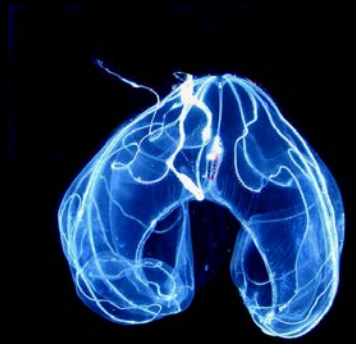




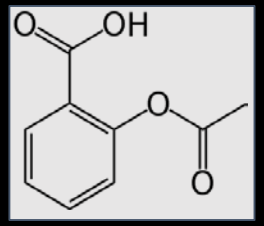


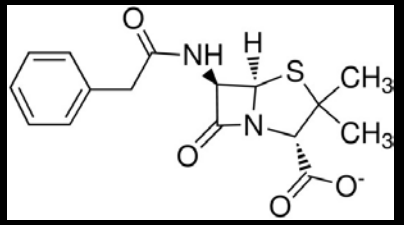
~4 Billion Years ago
(small amounts of simple stuff for *life*)

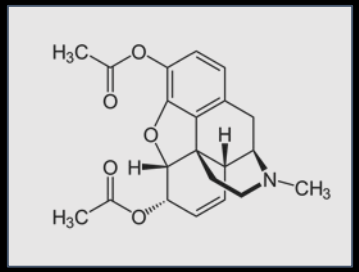




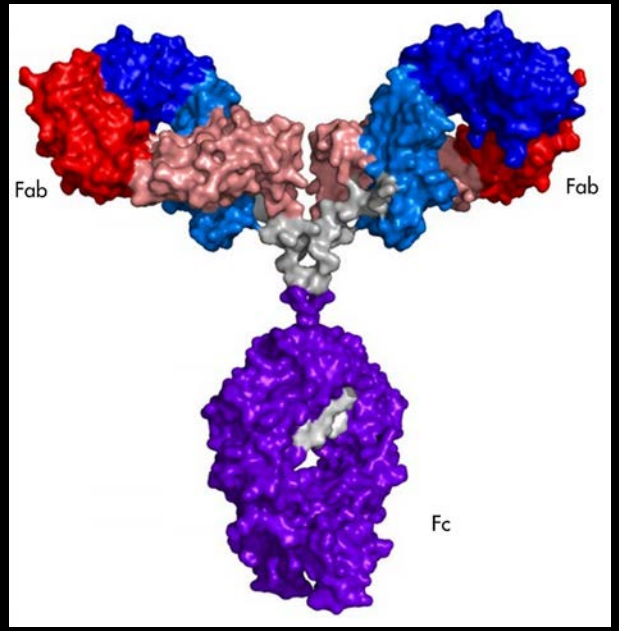
~1 Billion Years ago
(lots of new complicated stuff)

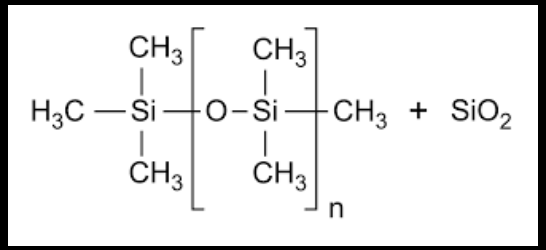


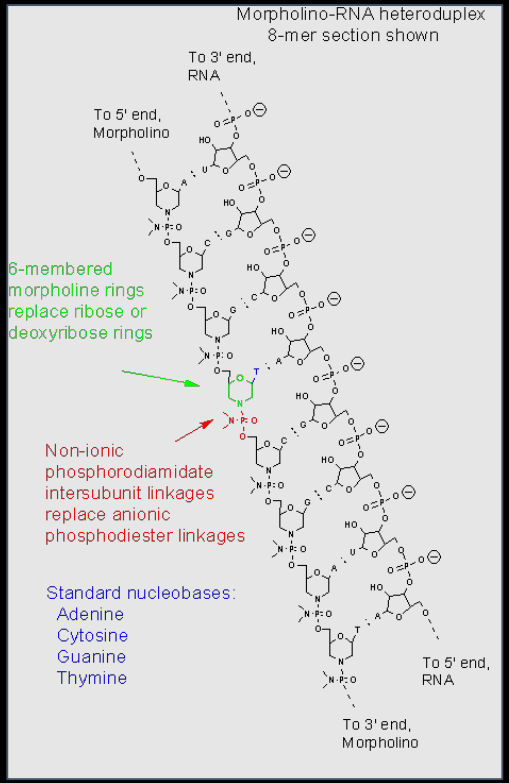




~100 years ago
(we start making lots of new simple stuff)







~25 years ago
(we start making lots of new *complicated* stuff)

Today

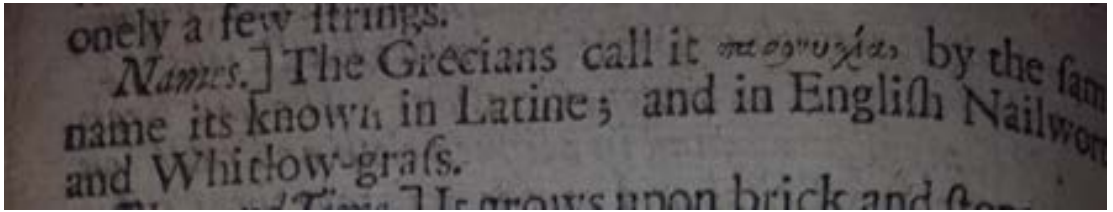
Today
(we’ve got to get our stuff together)

How would you make a catalog all the stuff?

- Researchers and Regulators need to know
  - What stuff exists
  - What it is
  - Where it’s used
  - What it does
  - **Whether it does anything else**

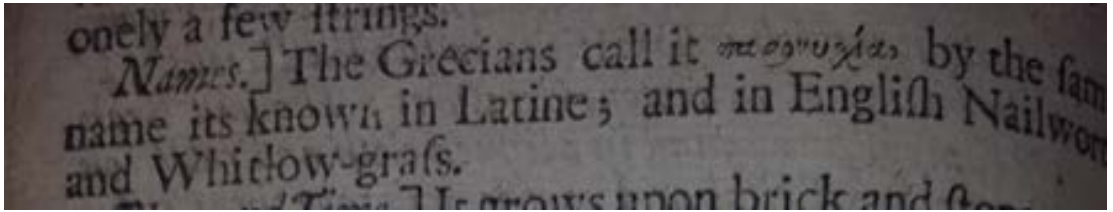
How do you deal with it all?

1. “Adam” Approach:
    - Find everything and name it
        - Bestiaries
        - Early Pharmacopoeias

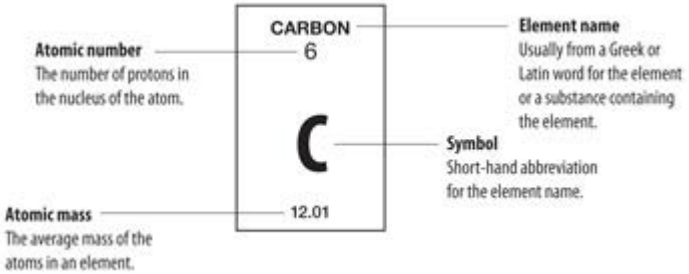


How do you deal with it all?

1. “Adam” Approach:
   - Find everything and name it
     - Bestiaries
     - Early Pharmacopoeias



1. “Atom” Approach:
   - Find what’s important, and systematically describe it
     - The Periodic Table
     - International Phonetic Alphabet



Substances at FDA

~1980: “Adam”-based database (Ingredient Dictionary)

2005: *Moderately* “Atom”-based system (SRS)

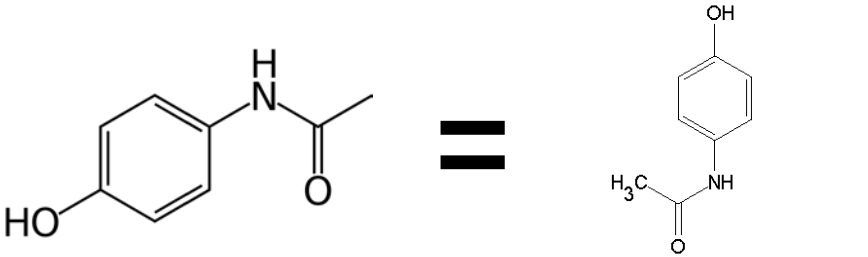
2017: *Extended* “Atom”-based system (GSRS)

Adam Approach: FDA CDER Ingredient Dictionary

- Started in the 1980s
- Registered active/inactive ingredients found in drugs
- Every ingredient had
  - An identifier (BDNUM)
  - A *Preferred Term*
  - A list of synonyms
- Some also had
  - CAS registry number
  - Wisswesser Line Notation
- ~20k ingredients
- No *rigorous* definitions (structure optional)
- Many duplicates
- Couldn’t do substructure search / similarity

From “Adam” to “Atom”

- Small molecules are pretty simple to define rigorously



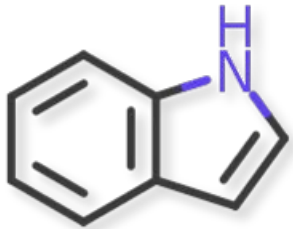
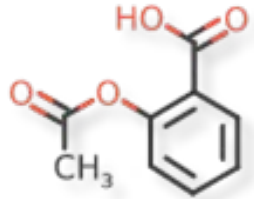
What does matters:

- What atoms you have
- How they connect to each other

What doesn’t matter:

- How you draw it
- What color it is
- How it smells
- How it’s used
- What it does

A word about storing Chemical Structures

Chemical Structure	WLN	Smiles	InChI
	1O9	COCCCCCCCCC	InChI=1S/C10H22O/c1-3-4-5-6-7-8-9-10-11-2/h3-10H2,1-2H3
	T56 BMJ	N1C=CC2=C1C=CC=C2	InChI=1S/C8H7N/c1-2-4-8-7(3-1)5-6-9-8/h1-6,9H
	1VOR BVQ	CC(=O)OC1=C(C=CC=C1)C(O)=O	InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)

Maybe a note about Lychi?