Pharma goes FAIR

Herman van Vlijmen
Janssen Pharmaceutica
Beerse, Belgium

What is FAIR?

FAIR DATA

A key enabler to achieve international-grade data stewardship is for research data and information to be published in a 'FAIR' manner.

In the FAIR Data approach, data should be:

1. Findable – Easy to find by both humans and computer systems and based on mandatory description of the metadata that allow the discovery of interesting datasets;
2. Accessible – Stored for long term such that they can be easily accessed and/or downloaded with well-defined license and access conditions (Open Access *when possible*), whether at the level of metadata, or at the level of the actual data content;
3. Interoperable – Ready to be combined with other datasets by humans as well as computer systems;
4. Reusable – Ready to be used for future research and to be processed further using computational methods.

https://www.dtls.nl/fair-data/fair-data/

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Research data

» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

Mark D. Wilkinson[1], Michel Dumontier[2], IJsbrand Jan Aalbersberg[3], Gabrielle Appleton[3], Myles Axton[4], Arie Baak[5], Niklas Blomberg[6], Jan-Willem Boiten[7], Luiz Bonino da Silva Santos[8], Philip E. Bourne[9], Jildau Bouwman[10], Anthony J. Brookes[11], Tim Clark[12], Mercè Crosas[13], Ingrid Dillo[14], Olivier Dumon[3], Scott Edmunds[15], Chris T. Evelo[16], Richard Finkers[17], Alejandra Gonzalez-Beltran[18], Alasdair J.G. Gray[19], Paul Groth[3], Carole Goble[20], Jeffrey S. Grethe[21], Jaap Heringa[22], Peter A.C. 't Hoen[23], Rob Hooft[24], Tobias Kuhn[25], Ruben Kok[22], Joost Kok[26], Scott J. Lusher[27], Maryann E. Martone[28], Albert Mons[29], Abel L. Packer[30], Bengt Persson[31], Philippe Rocca-Serra[18], Marco Roos[32], Rene van Schaik[33], Susanna-Assunta Sansone[18], Erik Schultes[34], Thierry Sengstag[35], Ted Slater[36], George Strawn[37], Morris A. Swertz[38], Mark Thompson[32], Johan van der Lei[39], Erik van Mulligen[39], Jan Velterop[40], Andra Waagmeester[41], Peter Wittenburg[42], Katherine Wolstencroft[43], Jun Zhao[44] & Barend Mons[4,5,45]

Sci Data. 2016, 3:160018

Why is FAIR important?

- Many data sets and databases are still sitting in silos with
  - Poor accessibility and/or findability of data
  - Absent or incomplete use of nomenclature standards
- The amount and diversity of scientific data is growing fast
- Most valuable analysis involves data from different domains/technologies
- Machine learning and data mining require unambiguous computer-readable data

From molecule to medicine

Basic Research/Discovery

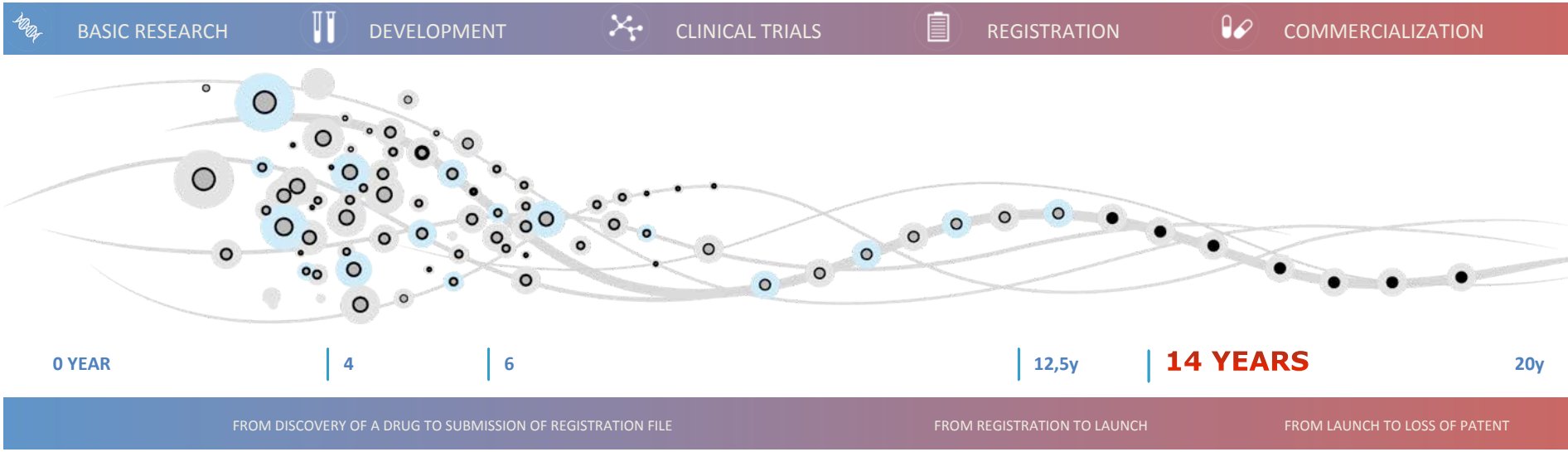
- **Select Disease**
- **Select Drug Target**
- **Identify Bioassay**
- **Find compound hits**
- **Find lead compound(s)**
- **Select clinical candidate**

Development

- **Drug metabolism and Pharmacokinetics**
- **Safety evaluation**
- **Chemical production**
- **Pharmaceutical formulation**

Clinical Trials

- **Safety (Phase I)**
- **Efficacy & Dose (Phase II)**
- **Efficacy (Phase III)**
- **Postmarketing (Phase IV)**



BEGIN: 10,000 MOLECULES

END: 1 NEW MEDICINE

Why is FAIR important to Janssen?

- If the data is there but nobody (re)uses it…
- Scientists at Janssen rarely use all data they have access to
  - Difficult to access multiple databases
  - Lack of awareness of databases
  - Little experience with definition of cross-domain analysis
- Data from multiple domains and sources (private, public, commercial) is needed for best possible analysis
  - Target identification and validation
  - Hit finding, H2L, Lead Optimization
  - Phenotypic screens, Omics experiments
  - Mechanistic analysis tox, side effects, drug repurposing
  - Translational analysis (cell phenotype <-> animal -> human)
  - Clinical and Real World Evidence analysis

Answering more complex questions

Scientific competency questions as the basis for semantically enriched open pharmacological space development

Kamal Azzaoui[1], Edgar Jacoby[14], Stefan Senger[2], Emiliano Cuadrado Rodríguez[3], Mabel Loza[3], Barbara Zdrazil[4], Marta Pinto[4], Antony J. Williams[5], Victor de la Torre[6], Jordi Mestres[7], Manuel Pastor[7], Olivier Taboureau[8], Matthias Rarey[9], Christine Chichester[10], Steve Pettifer[11], Niklas Blomberg[12,a], Lee Harland[13], Bryn Williams-Jones[13] and Gerhard F. Ecker[4]

Drug Discovery Today • Volume 18, Numbers 17/18 • September 2013

Open PHACTS

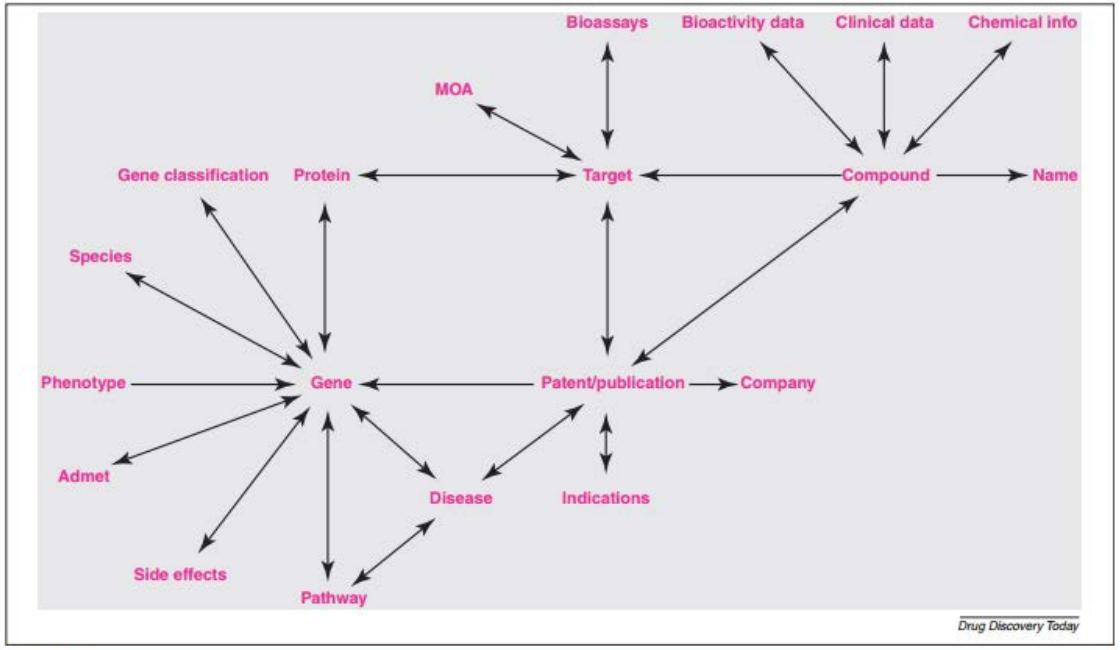


FIGURE 2

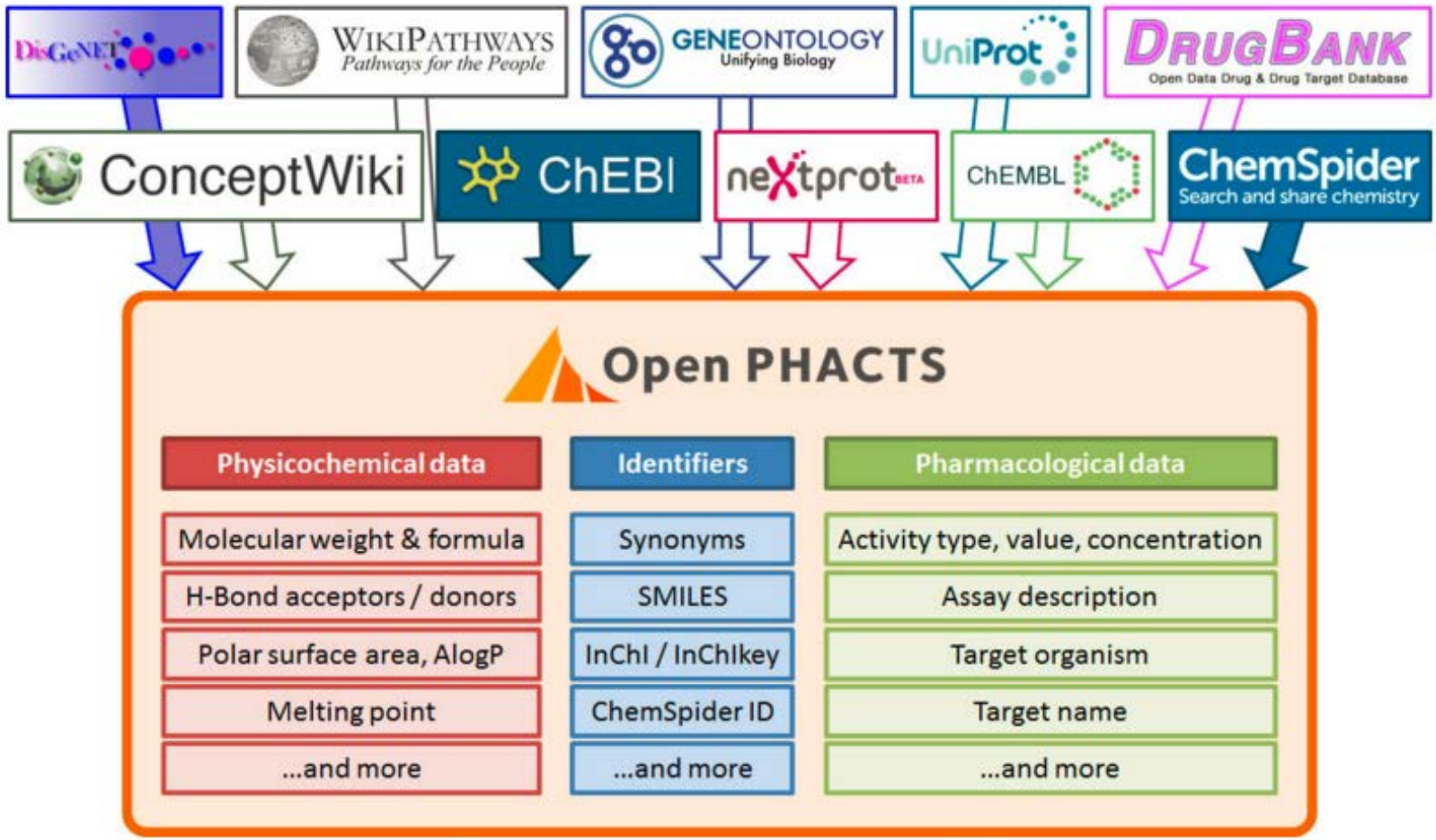
Network of data associations needed to answer the top-ranked scientific competency questions. The network reflects a cartoon that summarizes the data associations that are needed to target the top 20 research questions.

TABLE 1

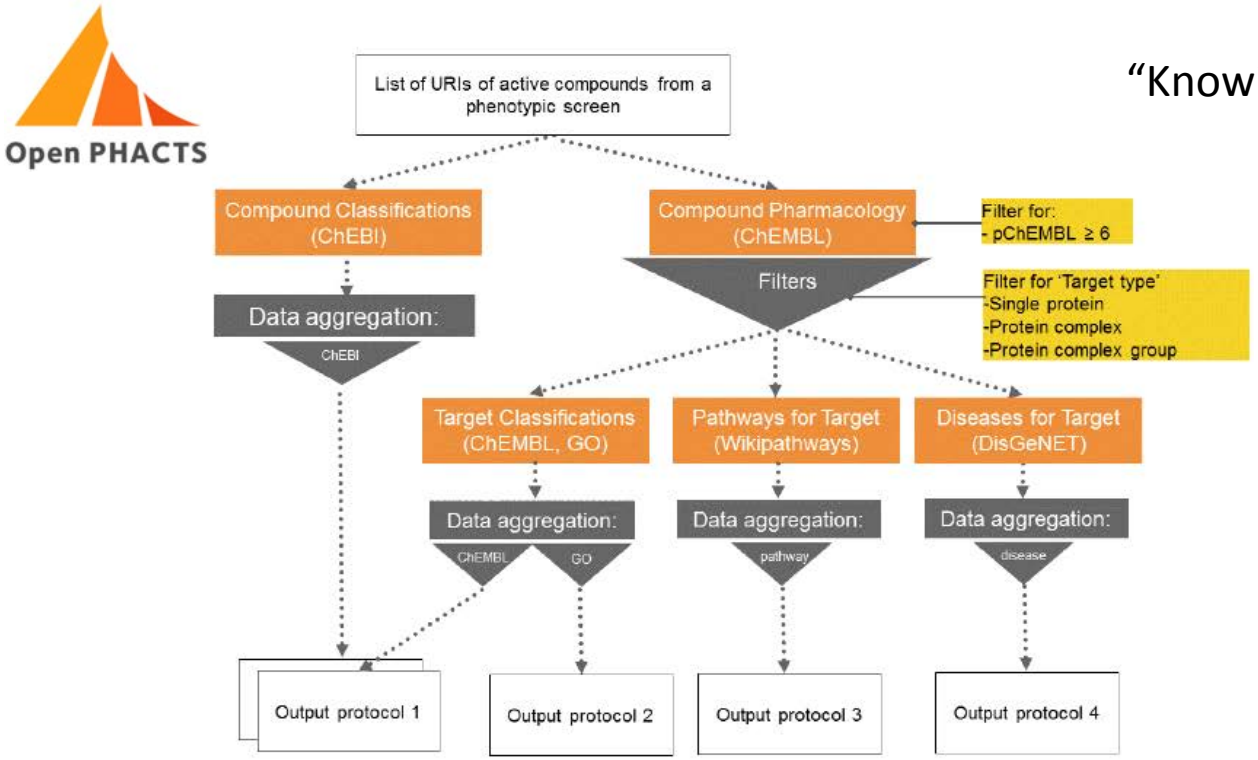
The top 20 research questions

Question number	Question
Cluster I	
Q1	Give me all oxidoreductase inhibitors active <100 nM in human and mouse
Q2	Given compound X, what is its predicted secondary pharmacology? What are the on- and off-target safety concerns for a compound? What is the evidence and how reliable is that evidence (journal impact factor, KOL) for findings associated with a compound?
Q3	Given a target, find me all actives against that target. Find/predict polypharmacology of actives. Determine ADMET profile of actives
Q4	For a given interaction profile – give me similar compounds
Q5	The current Factor Xa lead series is characterized by substructure X. Retrieve all bioactivity data in serine protease assays for molecules that contain substructure X
Q6	A project is considering protein kinase C alpha (PRKCA) as a target. What are all the compounds known to modulate the target directly? What are the compounds that could modulate the target directly? I.e. return all compounds active in assays where the resolution is at least at the level of the target family (i.e. PKC) from structured assay databases and the literature
Q7	Give me all active compounds on a given target with the relevant assay data
Q8	Identify all known protein–protein interaction inhibitors
Q9	For a given compound, give me the interaction profile with targets
Q10	For a given compound, summarize all 'similar compounds' and their activities
Q11	Retrieve all experimental and clinical data for a given list of compounds defined by their chemical structure (with options to match stereochemistry or not)
Cluster II	
Q12	For my given compound, which targets have been patented in the context of Alzheimer's disease?
Q13	Which ligands have been described for a particular target associated with transthyretin-related amyloidosis, what is their affinity for that target and how far are they advanced into preclinical/clinical phases, with links to publications/patents describing these interactions?
Q14	Target druggability: compounds directed against target X have been tested in which indications? Which new targets have appeared recently in the patent literature for a disease? Has the target been screened against in AZ before? What

Open PHACTS data sources



Phenotypic Drug Discovery Workflows



"Knowing the knowns"

BIOVIA PIPELINE PILOT

KNIME

Digles et al, MedChemComm, 7: 1237 (2016)

Open PHACTS developments: Patent Info

- Huge amount of knowledge in patent corpus, most of which will never be published elsewhere, but potentially great value to drug discovery
- SureChEMBL system (EBI) already automatically extracts compounds from these documents
- Open PHACTS consortium funded project to also extract gene/disease information (EMBL-EBI and SciBite)
- ~4 million patents in total, 260 million annotations (patent-compound, patent-gene or patent-disease associations)
- Example use cases:
  - For a given target or disease, give me all the compounds that are linked to this through patents
- Important to find new extraction tools to continue this annotation and make available at EBI

A broad set of use cases can be addressed using a linked data system

Some examples:

Target identification and validation

- Give me all direct and indirect supporting evidence linking a gene and disease
- Are there examples of compounds targeting any member of this target family?
- What are the relevant indirect links between a gene and a phenotypic assay?

Lead identification and optimization

- What compounds bind to this target or related targets (family, 3D similarity)?
- What bioactivities and pathways are associated to a compound?
- Show the activity of these compounds on all kinases involved in this pathway
- What are potential side-effects of hitting similar binding sites to our target?
- What side effects have similar compounds ?

Biomarker discovery

- What secreted proteins in a particular tissue are associated with this cellular pathway and might be biomarkers?
- New biomarkers: for which indirect biomarker-disease links there is no direct reported association, and which ones have the strongest level of data support?

Querying higher level research questions

A comparison of the queries that are done today versus what will be possible

TODAY	FUTURE
What are the Janssen compounds active in this Janssen assay?	Give me all **internal/commercial/public** data on compounds that are active on my target **and other closely related** targets.
What is the difference in gene expression profile between tumor and normal tissue?	Given the differences in expression profiles between these tissues, **give me the compounds with biochemical activity profiles that resemble the difference profile most**
Search PubMed for potential target-disease association: “bcl2 schizophrenia”	Show me **all possible direct and indirect** links between bcl2 and schizophrenia, **ranked by level of scientific data support**
I have a CDK7 lead compound. Is there anything known in PubMed on toxicity of CDK7 inhibitors?	Given my CDK7 lead compound, what are the **most likely mechanisms** by which this **compound class** could cause toxicity

Internal efforts in Discovery: **Chem^3**

- **Semantic graph database** of internal and external data linked to chemistry
  - Compound Activity:
    - ABCD, Athena, PIRlab, CAPE (all internal Janssen)
    - ChEMBL, PubChem, GOSTAR, Clarivate
    - Pending:
      - SureChEMBL
      - Etc. based on user needs
  - Platform: Virtuoso
  - Fast chemical cartridge (internal)
- **Interface** in 3DX, Pipeline Pilot, R

Examples of Linked Data challenges in Pharma

STANDARD_TYPE	UNIT_COUNT
AC50	7
Activity	421
EC50	39
IC50	46
ID50	42
Ki	23
Log IC50	4
Log Ki	7
Potency	11
log IC50	0

>5000 types

STANDARD_TYPE	STANDARD_UNITS	COUNT(*)
IC50	nM	829448
IC50	ug.mL-1	41000
IC50		38521
IC50	ug/ml	2038
IC50	ug ml-1	509
IC50	mg kg-1	295
IC50	molar ratio	178
IC50	ug	117
IC50	%	113
IC50	uM well-1	52
IC50	p.p.m.	51
IC50	ppm	36
IC50	uM-1	25
IC50	nM kg-1	25
IC50	milliequivalent	22
IC50	kJ m-2	20

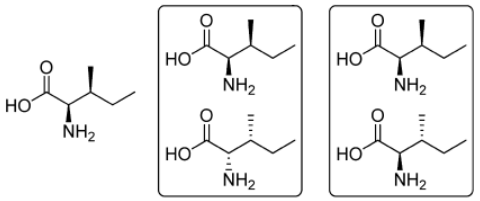
Implemented using the Quantities, Dimension, Units, Types Ontology (http://www.qudt.org/)

~ 100 units

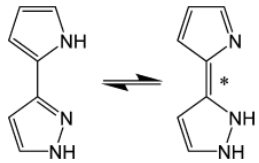
Data types and units for pharmacological activity in ChEMBL

Names & Taxonomy

Protein names	*Recommended name:* **Tyrosine-protein kinase BTK** (EC:2.7.10.2) *Alternative name(s):* • Agammaglobulinemia tyrosine kinase ▪ *Short name:*ATK • B-cell progenitor kinase ▪ *Short name:*BPK • Bruton tyrosine kinase
Gene names	*Name:*BTK Synonyms:AGMX1, ATK, BPK



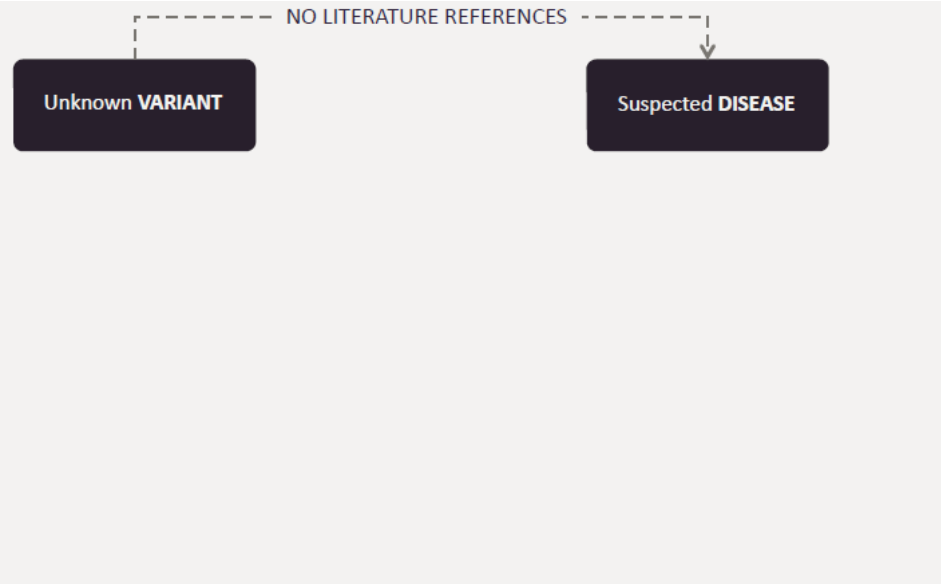
Stereochemistry



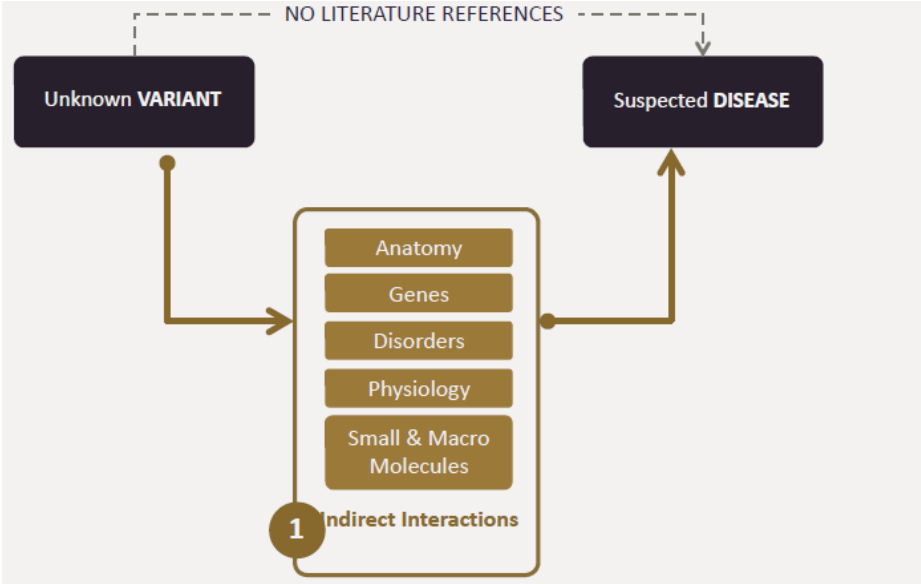
Tautomerism

Lee and Gobbi. J. Chem. Inf. Model. 2012, 52, 285−292

Internal efforts in Discovery: Exploring use of Euretos Knowledge Platform for TI/TV



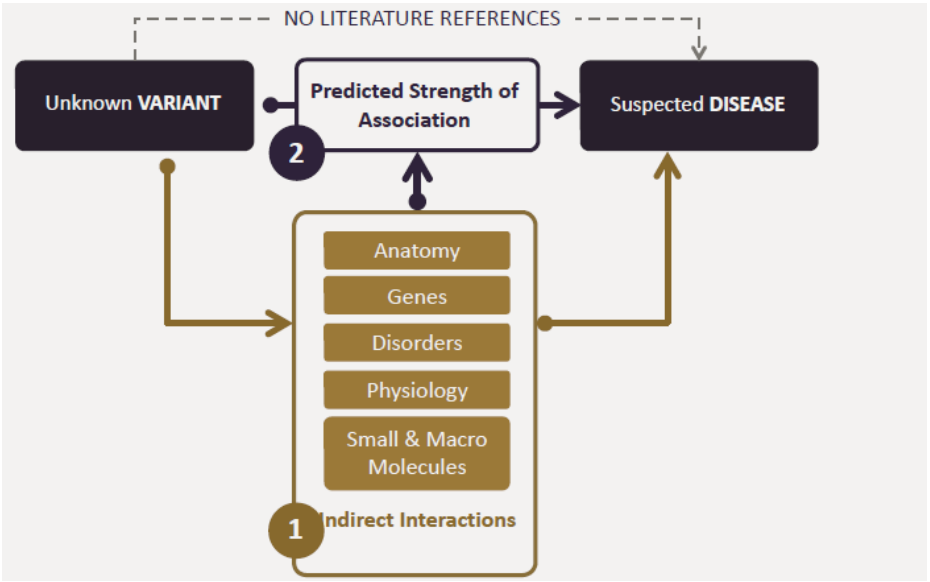
Step 1



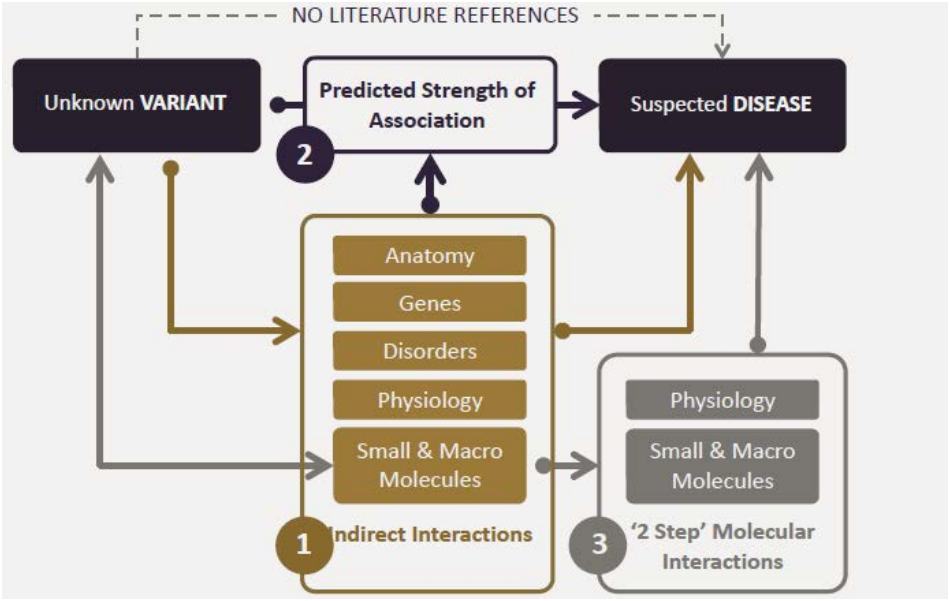
Step 2

www.euretos.com

Internal efforts in Discovery: Exploring use of Euretos Knowledge Platform for TI/TV



Step 3



Step 4

www.euretos.com

IMI2

12th Call for proposals

Topic 2: FAIRification of IMI and EFPIA data

- Call launched in July 2017
- Candidate consortia are currently being evaluated
- Likely start project Q3 2018, duration 3 years
- Budget: The financial contribution from IMI2 is a maximum of EUR 4M
- Pharma partners: Janssen, AZ, Bayer, Boehringer Ingelheim, Eli Lilly, GSK, Novartis

Summary of FAIRification Proposal

- Select data sets and databases from finished and ongoing IMI projects, based on:
  - Scientific value of making this data accessible and interoperable
  - Complexity of making the data available
- Select databases at individual EFPIA companies
  - Selection based on value for companies
  - Consolidation to limited set of data domains
- FAIRify these data sets to enable the sustainable use of the data in answering research questions
  - Work sessions with data owners and FAIRification experts, including data domain experts (vocabularies, ontologies, use cases) and IT experts (conversion of data, database implementation)
  - Implementation of sustainable solution for storage and maintenance of FAIRified IMI databases
  - Identification of sustainable solution for storage and maintenance of FAIRified EFPIA databases

IMI Projects

- ABIRISK
Anti-Biopharmaceutical Immunization: Prediction and Analysis of Clinical Relevance to Minimize the Risk

- ADVANCE
Accelerated development of vaccine benefit-risk collaboration in Europe

- AETIONOMY
Organising mechanistic knowledge about neurodegenerative diseases for the improvement of drug development and therapy

- BioVacSafe
Biomarkers for Enhanced Vaccine Immunosafety

- BTCure
Be The Cure

- CHEM21
Chemical manufacturing methods for the 21st century pharmaceutical industries

- COMBACTE
Combatting Bacterial Resistance in Europe

- COMBACTE-CARE
Combatting Bacterial Resistance in Europe - Carbapenem Resistance

- COMBACTE-MAGNET
Combatting bacterial resistance in Europe - molecules against Gram negative infections

- COMPACT
Collaboration on the optimisation of macromolecular pharmaceutical access to cellular targets

- DDMoRe
Drug Disease Model Resources

- DIRECT
Diabetes research on patient stratification

- DRIVE-AB
Driving re-investment in R&D and responsible antibiotic use

- EBiSC
European Bank for induced pluripotent Stem Cells

- Ebola+
Ebola and other filoviral haemorrhagic fevers

- EHR4CR
Electronic Health Records Systems for Clinical Research

- ELF
European Lead Factory

- EMIF
European Medical Information Framework

- EMTRAIN
European Medicines Research Training Network

- ENABLE
European Gram-negative Antibacterial Engine

- EPAD
European prevention of Alzheimer's dementia consortium

- eTOX
Integrating bioinformatics and chemoinformatics approaches for the development of Expert systems allowing the in silico prediction of toxicities

- eTRIKS
Delivering European Translational Information & Knowledge Management Services

- EU-AIMS
European Autism Interventions - a Multicentre Study for Developing New Medications

- Eu2P
European programme in Pharmacovigilance and Pharmacoepidemiology

- EUPATI
European Patients' Academy on Therapeutic Innovation

- EUROPAIN
Understanding chronic pain and improving its treatment

- FLUCOP
Standardization and development of assays for assessment of influenza vaccines correlates of protection

- GETREAL
Incorporating real-life clinical data into drug development

- iABC
Inhaled antibiotics in bronchiectasis and cystic fibrosis

- IMIDIA
Improving beta-cell function and identification of diagnostic biomarkers For treatment monitoring in diabetes

- iPiE
Intelligent Assessment of Pharmaceuticals in the Environment

- K4DD
Kinetics for Drug Discovery

- MARCAR
Biomarkers and molecular tumour classification for non-genotoxic carcinogenesis

- MIP-DILI
Mechanism-Based Integrated Systems for the Prediction of Drug-Induced Liver Injury

- ND4BB
New Drugs for Bad Bugs

- NEWMEDS
Novel methods leading to new medications in depression and schizophrenia

- Onco Track
Methods for systematic next generation oncology biomarker development

- Open PHACTS
The Open Pharmacological Concepts Triple Store

- ORBITO
Oral biopharmaceutics tools

- Pharma-Cog
Prediction of cognitive properties of new drug candidates for neurodegenerative diseases in early clinical development

- Pharmatrain
Pharmaceutical Medicine Training Programme

- PRECISESADS
Molecular reclassification to find clinically useful biomarkers for systemic autoimmune diseases

- Predect
New models for preclinical evaluation of drug efficacy in common solid tumours

- PreDiCT-TB
Model-based preclinical development of anti-tuberculosis drug combinations

- PRO-active
Physical Activity as a Crucial Patient Reported Outcome in COPD

- PROTECT
Pharmacoepidemiological research on outcomes of therapeutics by a European consortium

- Quic-Concept
Quantative imaging in cancer:connecting cellular process with therapy

- RAPP-ID
Development of rapid point-of-care test platforms for infectious diseases

- SAFE-T
Safer and Faster Evidence-based Translation

- SafeSciMET
European Modular Education and Training Programme in Safety Sciences for Medicines

- SPRINTT
Sarcopenia and physical frailty in older people: multi-component treatment strategies

- STEMBANCC
Stem cells for biological assays of novel drugs and predictive toxicology

- SUMMIT
Surrogate markers for micro- and macro-vascular hard endpoints for innovative diabetes tools

- TRANSLOCATION
Molecular basis of the bacterial cell wall permeability

- U-BIOPRED
Unbiased biomarkers for the prediction of respiratory disease outcomes

- WEB-RADR
Recognising Adverse Drug Reactions

Preliminary analysis of IMI projects for availability of FAIRifiable data sets (partial list)
Analysis done by Anthony Rowe (Janssen) and Colm Carroll (IMI)

Program	C	Project	Project Acronym	Project Title	Start Da	End Dat	FAIRification	legal Coordinator Short Name (A
IMI1	1	115001	MARCAR	bioMARkers and molecular tumor c	1/01/2010	30/06/2015	Yes	Novartis
IMI1	1	115002	eTOX	Integrating bioinformatics and chem	1/01/2010	31/12/2016	Yes	Novartis
IMI1	1	115003	SAFE-T	SAFER AND FASTER EVIDENCE-	15/06/2009	14/06/2015	Yes	Novartis
IMI1	1	115004	PROTECT	Pharmacoepidemiolocal Research (	1/09/2009	30/04/2015	Yes	EMA
IMI1	1	115005	IMIDIA	Improving beta-cell function and ide	1/02/2010	30/09/2015	Yes	SAD
IMI1	1	115006	SUMMIT	SUrrogate markers for vascular Mic	1/11/2009	31/10/2015	Yes	Boehringer Ingelheim
IMI1	1	115007	EUROPAIN	Understanding chronic pain and imp	1/10/2009	30/09/2015	Yes	HLU
IMI1	1	115008	NEWMEDS	Novel Methods leading to New Med	1/09/2009	28/02/2015	Yes	HLU
IMI1	1	115009	PHARMA-COG	Prediction of cognitive properties of	1/01/2010	31/12/2015	Yes	GSK
IMI1	1	115010	U-BIOPRED	Unbiased Biomarkers for the Predic	1/10/2009	30/09/2015	Yes	NOV
IMI1	1	115011	PRO- Active	Physical Activity as a Crucial Patie	1/09/2009	31/05/2016	Yes	Chiesi
IMI1	1	115012	SafeSciMET	European Modular Education and T	1/01/2010	31/08/2016	No	ROCHE
IMI1	1	115013	PharmaTrain	Pharmaceutical Medicine Training F	1/05/2009	30/04/2014	No	PharmaTrain Federation
IMI1	1	115014	EU2P	European programme in Pharmaco	1/09/2009	30/06/2016	Yes	ROCHE
IMI1	1	115015	EMTRAIN	European Medicines Research Trai	1/10/2009	30/09/2016	No	AstraZeneca
IMI1	2	115142	BTCURE	BeTheCuRE	1/04/2011	31/03/2017	Yes	UCB
IMI1	2	115151	QUIC-CONCEPT	QUantitative Imaging in Cancer: CC	1/09/2011	31/08/2016	Yes	EORTC
IMI1	2	115153	RAPP-ID	Development of RApid Point-of-Care	1/04/2011	30/09/2016	Yes	JNJ
IMI1	2	115156	DDMoRe	Drug Disease Model Resources	1/03/2011	31/08/2016	?	Pfizer
IMI1	2	115188	PREDECT	New Models for Preclinical Evaluati	1/02/2011	30/04/2016	Yes	Servier
IMI1	2	115189	EHR4CR	Electronic Health Record systems	1/03/2011	29/02/2016	Blank	AZ
IMI1	2	115191	Open PHACTS	The Open Pharmacological Concep	1/03/2011	29/02/2016	Yes	GSK
IMI1	2	115234	OncoTrack	OncoTrack - Methods for systema	1/01/2011	31/12/2016	Yes	BHP
IMI1	3	115300	EU-AIMS	European Autism Interventions - A	1/04/2012	31/03/2017	Yes	ROCHE
IMI1	3	115303	ABIRISK	Anti-Biopharmaceutical Immunizati	1/03/2012	28/02/2017	Yes	GSK
IMI1	3	115308	BioVacSafe	Biomarkers For Enhanced Vaccine	1/03/2012	28/02/2017	Yes	GSK Vaccines Srl
IMI1	3	115317	DIRECT	Diabetes REsearCh on patient sTra	1/02/2012	31/01/2017	Yes	SAD
IMI1	3	115334	EUPATI	European Patients' Academy on Th	1/02/2012	31/01/2017	No	VFA
IMI1	3	115336	MIP-DILI	Mechanism-Based Integrated Syste	1/02/2012	31/01/2017	Yes	AstraZeneca
IMI1	3	115337	PreDiCT-TB	Model-based preclinical developme	1/05/2012	30/10/2017	Yes	GSK
IMI1	4	115360	CHEM21	Chemical Manufacturing Methods f	1/10/2012	30/09/2016	?	GSK
IMI1	4	115363	Compact	Collaboration on the Optimisation o	1/11/2012	31/10/2017	?	SAD
IMI1	4	115366	K4DD	Kinetics for Drug Discovery (K4DD)	1/11/2012	31/10/2017	Yes	Bayer
IMI1	4	115369	OrBiTo	Oral biopharmaceutics tools	1/10/2012	30/09/2017	Yes	AstraZeneca
IMI1	4	115372	EMIF	European Medical Information Fram	1/01/2013	31/12/2017	Yes	Janssen
IMI1	4	115439	StemBANCC	Stem cells for Biological Assays of	1/10/2012	30/09/2017	Yes	ROCHE
IMI1	4	115446	eTRIKS	Delivering European Translational I	1/10/2012	30/09/2017	Yes	AstraZeneca
IMI1	5	115489	EUC²LID	European Lead Factory	1/01/2013	31/12/2017	?	Bayer
IMI1	6	115523	COMBACTE	Combatting Bacterial Resistance in	1/01/2013	31/12/2019	?	AstraZeneca
IMI1	6	115525	Translocation	Molecular basis of the outer membr	1/01/2013	31/12/2017	Yes	GSK
IMI1	7	115546	GetReal	Incorporating real-life clinical data in	1/10/2013	31/12/2016	Yes	GSK
IMI1	7	115557	ADVANCE	Accelerated Development of Vaccin	1/10/2013	30/09/2018	Yes	EMC
IMI1	8	115565	PRECISESADS	Molecular Reclassification to Find (	1/02/2014	31/01/2019	Yes	UCB
IMI1	8	115568	AETIONOMY	Aetionomy – Organising Mechanist	1/01/2014	31/12/2018	Yes	UCB
IMI1	8	115582	EBiSC	European Bank for induced pluripot	1/01/2014	31/12/2016	Yes	Pfizer

Expected Impact

- The scientific community can maximally leverage data from legacy and current IMI projects
- Strengthening the capacity of creation, curation, and stewardship of FAIR databases within IT communities of academia, SME, and pharma
- Better understanding of the complexity, structure, and breadth of pharmaceutical data: Allow the SME community to make their data, analysis tools and services better connected and aligned to pharma
- A long-lasting value-adding impact on effective scientific data usage

Topic 4: European Health Data Network (EHDN)

Building on IMI1 programmes, e.g.
EHR4CR (Electronic Health Records for Clinical Research)
EMIF (European Medical Information Framework)
ADVANCE (Data framework for vaccine risk/benefit analysis)
GetReal (RWE collection and synthesis)
and other FP7 programmes, we will be utilising the outputs of those PPPs, the relationships and their contributions
(Big Data for Better Outcomes BD4BO)

- **Work Package 1: Methodological Research**
  - Federated network creation
  - Data harmonisation, evaluation and quality benchmark
  - Evidential linkage with Regulators, HTAs
  - *From technology to RWE*

Conclusions

- Pharma has strong interest and need for implementing FAIR data principles
- Expertise is usually lagging due to complex legacy data systems and limited IT resources
- Scientific acceptance of FAIRification in Pharma requires strong use case examples
- Collaboration with academia and SME community catalyzes expertise and acceptance of FAIR data