DRUGCENTRAL: PART OF A BIGGER PICTURE

There is a Need to Integrate Clinical Use with Active Ingredients,

Pharmaceutical Products & Associated Information at the Molecular Level

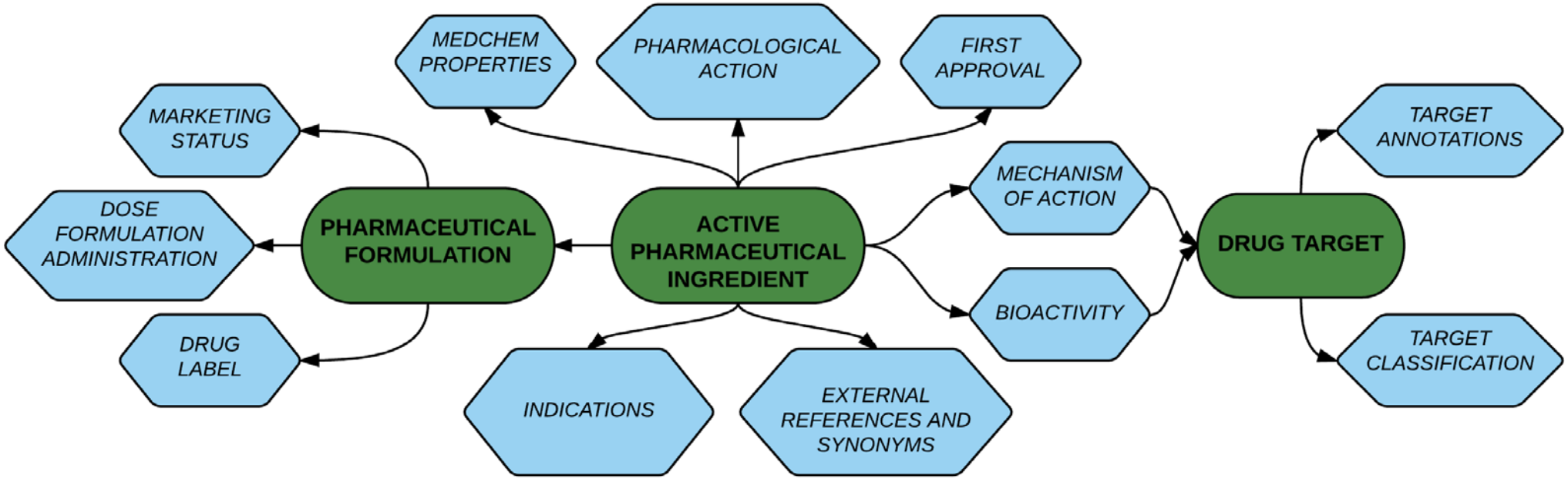
Tudor I. Oprea

8/29/2017
OHDSI
Collaborator meeting
via Webex
Albuquerque, NM

http://targetcentral.ws/
http://pharos.nih.gov
http://drugcentral.org
http://newdrugtargets.org

UNM
SCHOOL of MEDICINE

DRUGCENTRAL DATA STRUCTURE

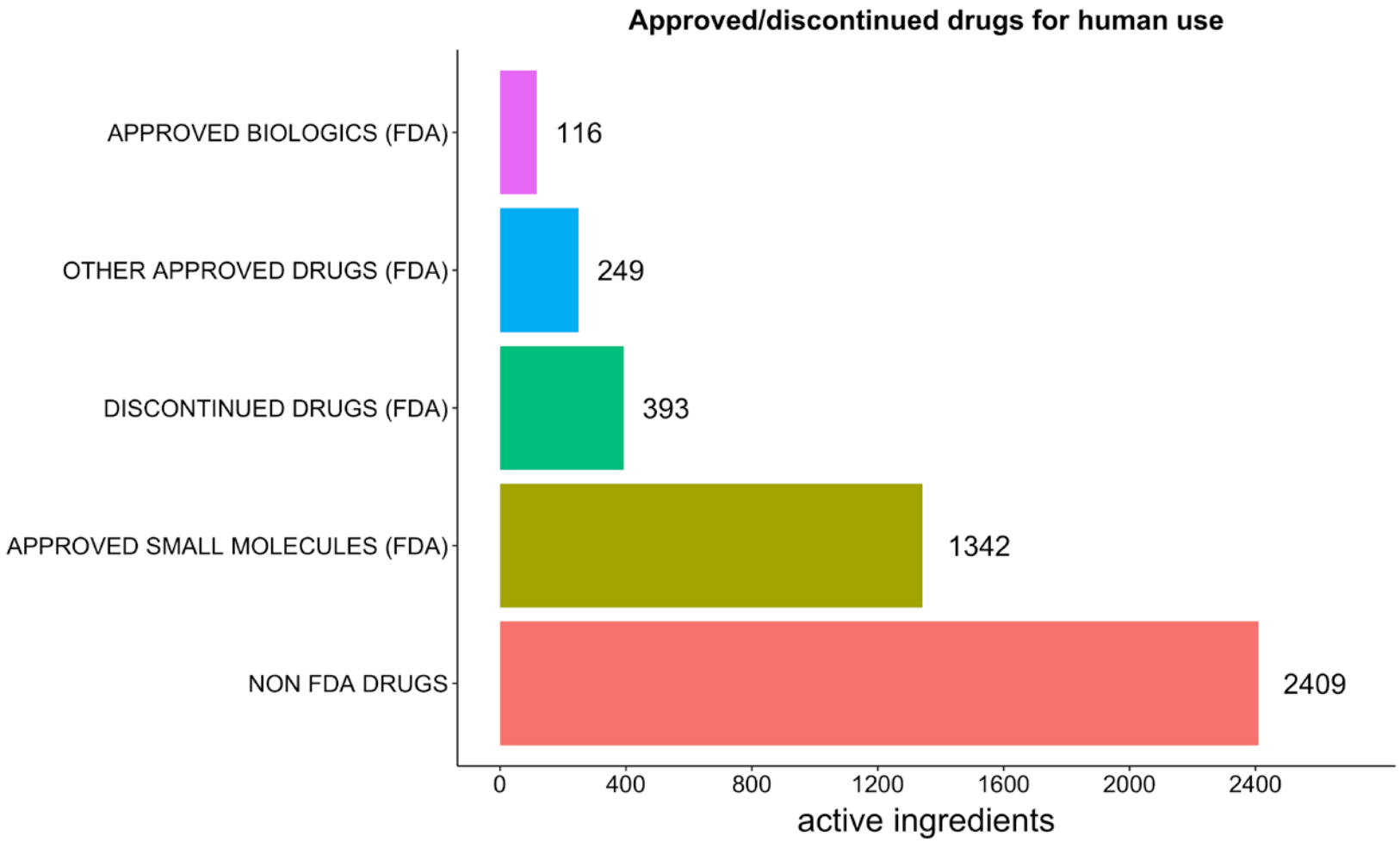


- Initially designed to answer "how many drugs are out there"…
- The Two Cultures: what patients and docs call "drugs" (products) vs. what scientists call "drugs" (active pharmaceutical ingredients)
- Also wanted to know how many drug targets there are……

Oleg Ursu et al., *Nucl Acids Res*, database issue, **2017**, doi:10.1093/nar/gkw993

8/29/17 revision

DRUGCENTRAL: API STATUS



- Total number of active ingredients: ~4500
- This includes API approved for human use worldwide, FDA approved and discontinued
- ~1500 are currently marketed and FDA approved, ~ 300 are discontinued

Oleg Ursu et al., *Nucl Acids Res*, database issue, **2017**, doi:10.1093/nar/gkw993

8/29/17 revision

MAPPING TO EXTERNAL RESOURCES

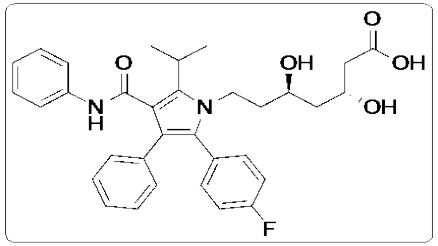
ChEBI
CHEBI:39548

KEGG
D00887

MeSH
C065179

World Health Organization
INN ID 7259

ChEMBL
CHEMBL1487



Atorvastatin

FDA
NDC 54868-4229

PubChem
CID 60823

RCSB PDB PROTEIN DATA BANK
Ligand id 117

IUPHAR/BPS Guide to PHARMACOLOGY
Ligand id 2949

RxNorm
RxCUI 83367

CAS
134523-00-5

FDA
UNII A0JWA85V8F

DRUGBANK Open Data Drug & Drug Target Database
DB01076

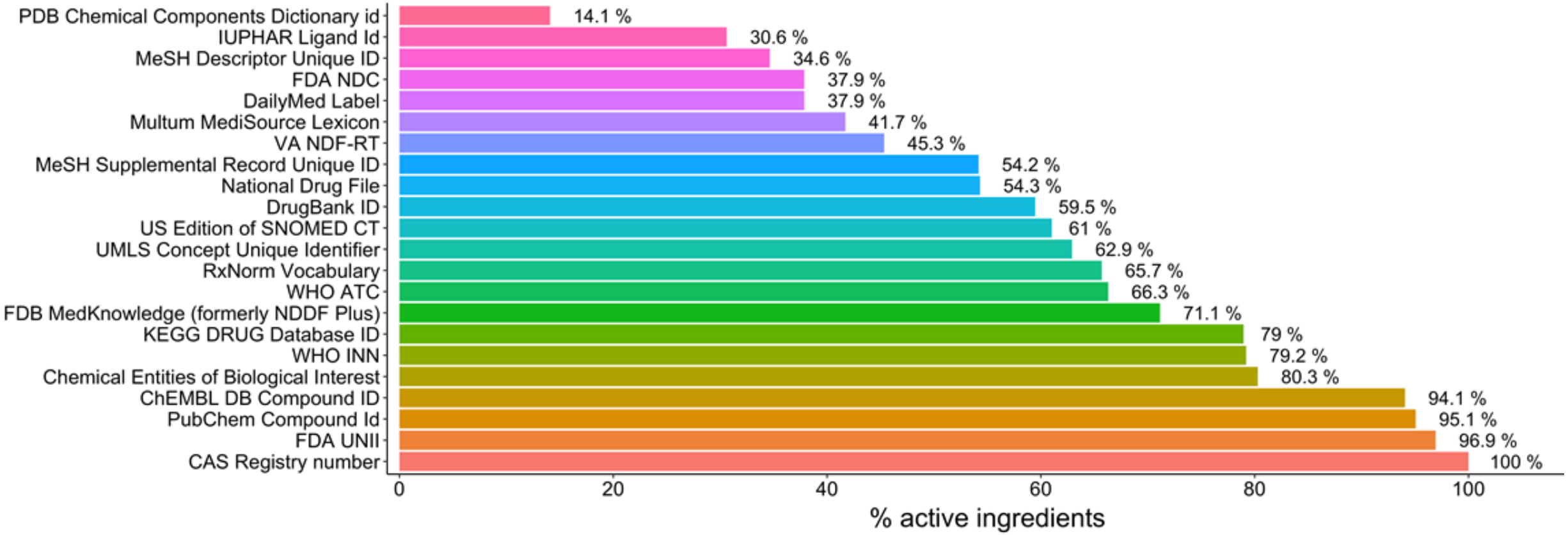
- Several online resources contain important drug information
- To facilitate data analysis we have mapping of active ingredients to most relevant drug information resources online.
- Most mappings were done using generic names and structure.
- These drug resources provide information on regulatory status, publications, pharmacology, biological activity profiles, etc.

Oleg Ursu et al., *Nucl Acids Res*, database issue, **2017**, doi:10.1093/nar/gkw993

8/29/17 revision

IDG KMC

EXTERNAL DATA SOURCE IDENTIFIER MAPPINGS



- Mapping of drugs to external resources ranges from 13% (PDB Ligands) to 100% (CAS registry numbers)

Oleg Ursu et al., *Nucl Acids Res*, database issue, **2017**, doi:10.1093/nar/gkw993

8/29/17 revision

DAILYMED DRUG LABELS (FDA)

NIH U.S. NATIONAL LIBRARY OF MEDICINE

DAILYMED

- Drug labels in SPL (Structured Product Label) format
- Updated Daily
- Text in sections annotated with LOINC codes
  - Summary of clinical trial results
  - Contraindications, adverse events, warnings, therapeutic dose, etc.
- Table with active/inactive ingredients, strength, route of administration
  - NDA, ANDA, UNII identifiers

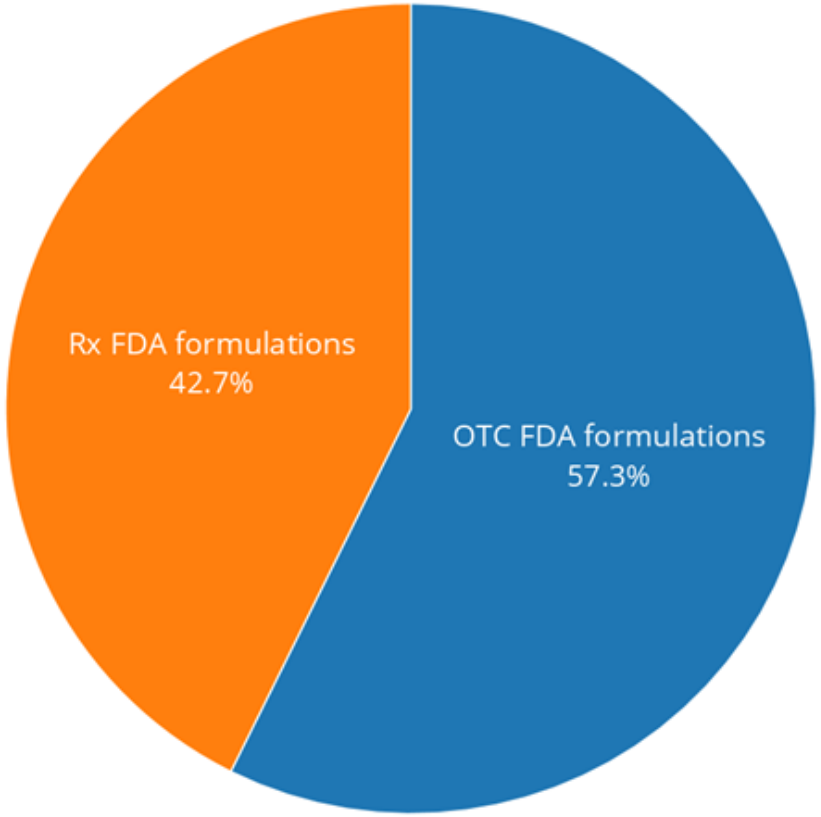
- DailyMed is the main source of information on pharmaceutical products. We use custom processing pipelines that extract text from SPL separated by LOINC sections.
- Dose, formulations and active ingredients tables are parsed and mapped to the main active ingredients table.
- Pharmaceutical formulations containing herbals, allergens, etc. products are discarded
- We do not process homeopathic labels and SPL files for devices.

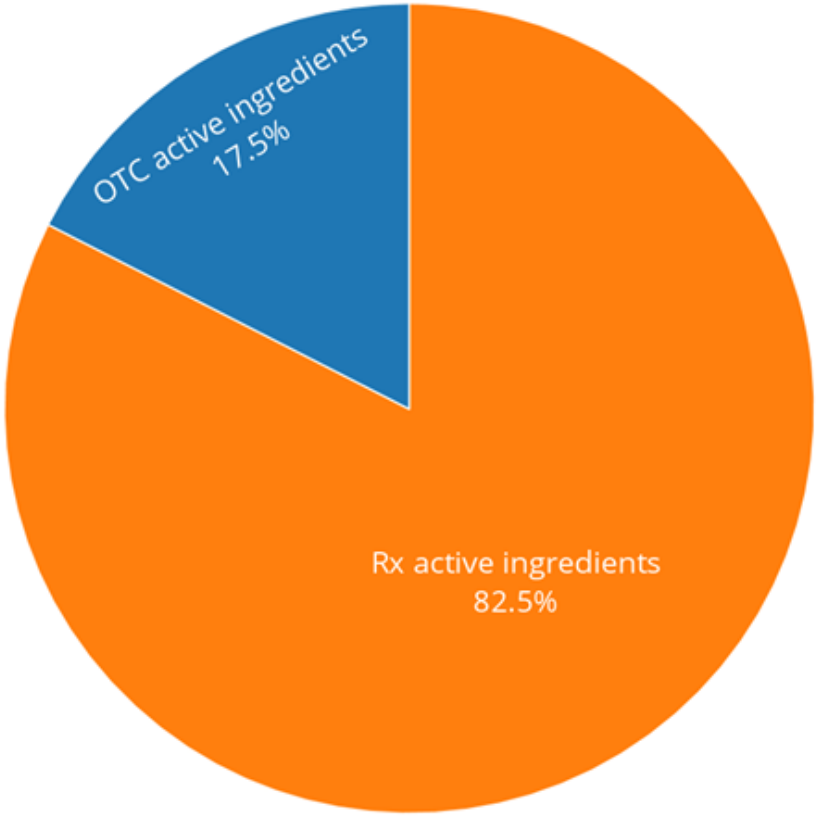
Oleg Ursu et al., *Nucl Acids Res*, database issue, **2017**, doi:10.1093/nar/gkw993

8/29/17 revision

IDG KMC

ACTIVE INGREDIETS VS PHARMACEUTICAL PRODUCTS



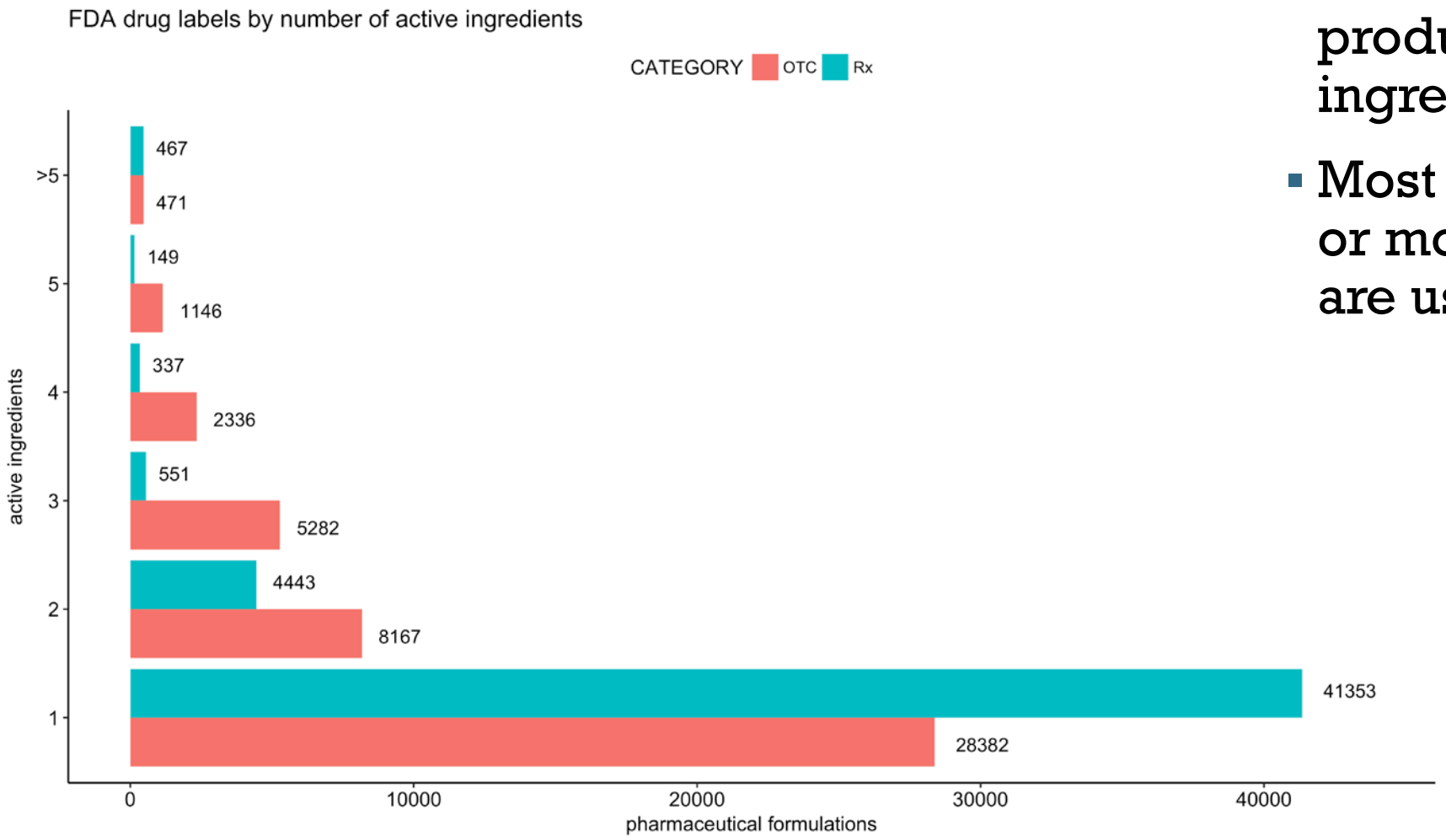


- Active ingredients in Rx products only form more than 82% of the total number of active ingredients
- However, when compared total number of pharmaceutical products OTC only active ingredients have 46% share.

Oleg Ursu et al., *Nucl Acids Res*, database issue, **2017**, doi:10.1093/nar/gkw993

8/29/17 revision

HOW MANY APIs PER PRODUCT?



- Most of the pharmaceutical products contain 1 active ingredient,
- Most of the products with 2 or more active ingredients are usually OTC.

Oleg Ursu et al., *Nucl Acids Res*, database issue, **2017**, doi:10.1093/nar/gkw993

8/29/17 revision

CAPITALISM IN THE PHARMACY

Type	OTC	PRESCR
APIs	284	1,562
Drugs (”drug labels”)	46,770	43,172

- There are almost as many “OTC” as Rx drugs, but with far less APIs
- *Over 5000 drug labels contain acetaminophen (84 unique API fixed-dose combinations)*

Oleg Ursu et al., *Nucl Acids Res*, database issue, **2017**, doi:10.1093/nar/gkw993

8/29/17 revision

AUSTRALIA: TWO PRICES, ONE DRUG

Nurofen's maker misled consumers over painkillers' contents, court rules

Drug giant Reckitt Benckiser ordered to pull painkillers off Australian shelves after admitting products marketed for specific types of pain were identical

Nurofen criticised by Australian consumer watchdog over misleading claims

Reckitt Benckiser sells:

- Nurofen Back Pain,
- Nurofen Period Pain
- Nurofen Migraine Pain and
- Nurofen Tension Headache

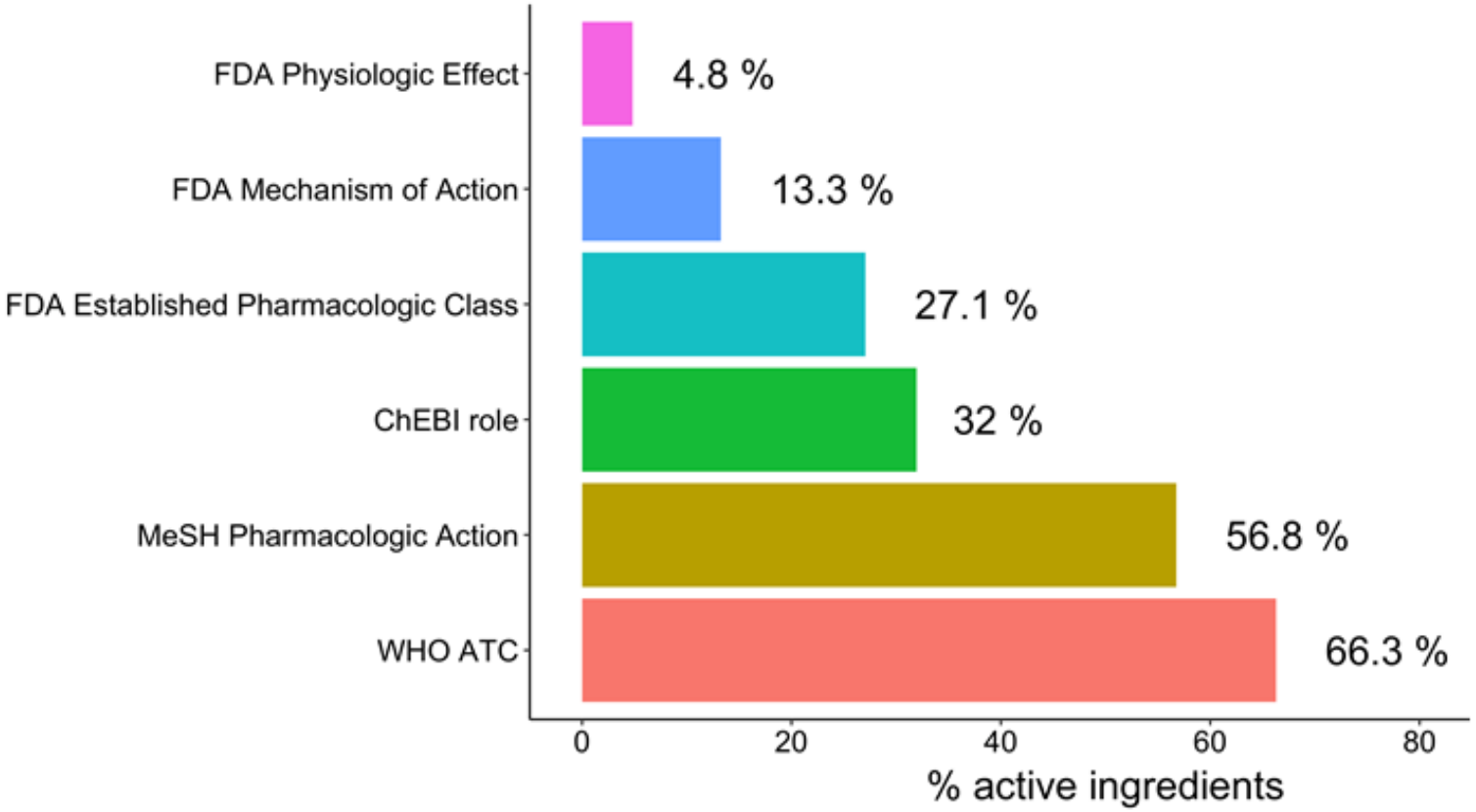
at twice the price compared to Nurofen, even though it contains exactly the same active ingredient (342mg of ibuprofen lysine, equivalent to 200mg of ibuprofen).

12/14/15 revision

TAKE HOME MESSAGE 1

PHARMACEUTICAL PRODUCTS ARE AN EQUALLY IMPORTANT COMPONENT OF DRUG RESEARCH

PHARMACOLOGIC CLASSIFICATIONS

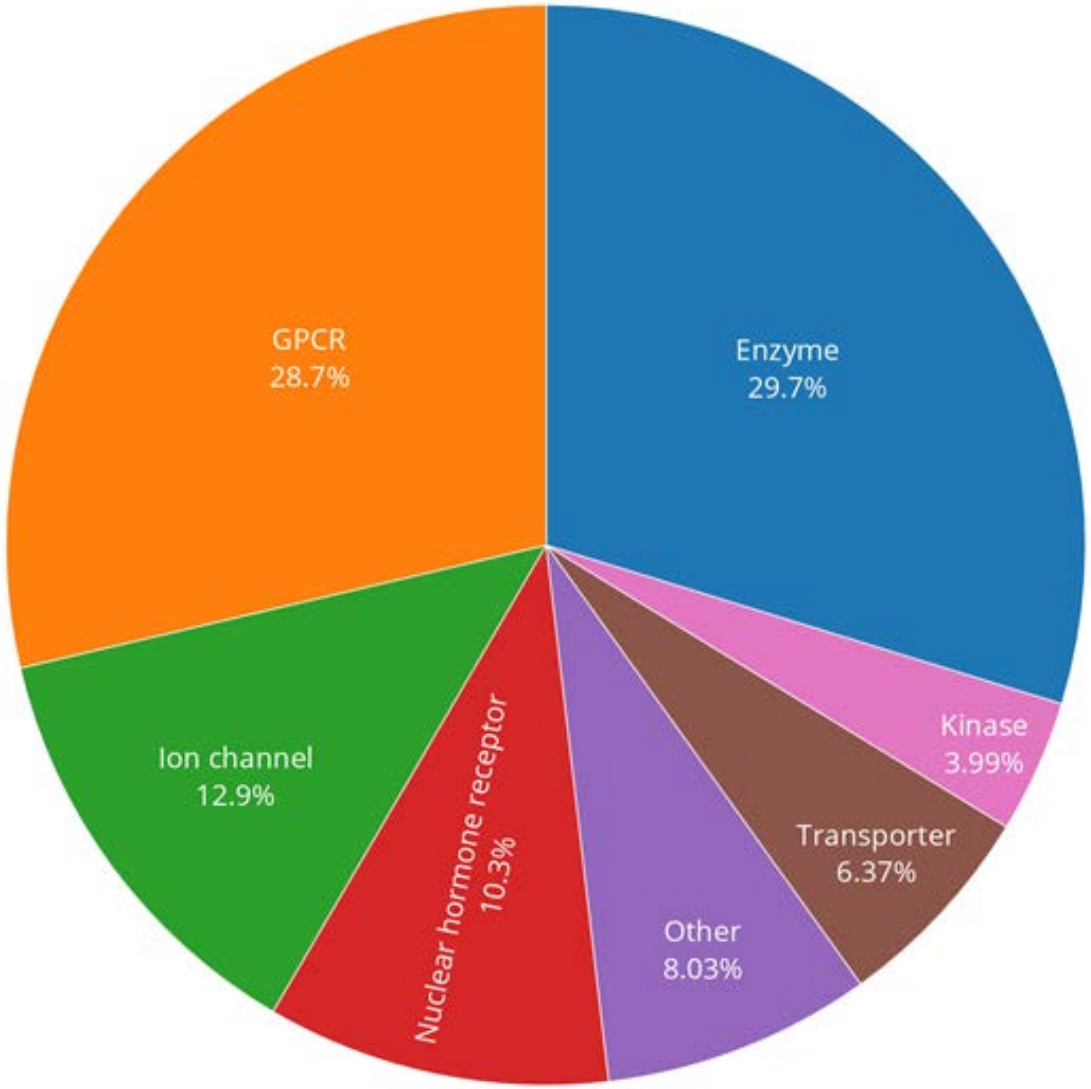


- DrugCentral integrates pharmacologic classifications from ATC, MeSH, ChEBI, and FDA
- These provide systematic groupings of drugs based on common therapeutic applications and mechanism of action

Oleg Ursu et al., *Nucl Acids Res*, database issue, **2017**, doi:10.1093/nar/gkw993

8/29/17 revision

DRUG TARGETS – MECHANISM OF ACTION

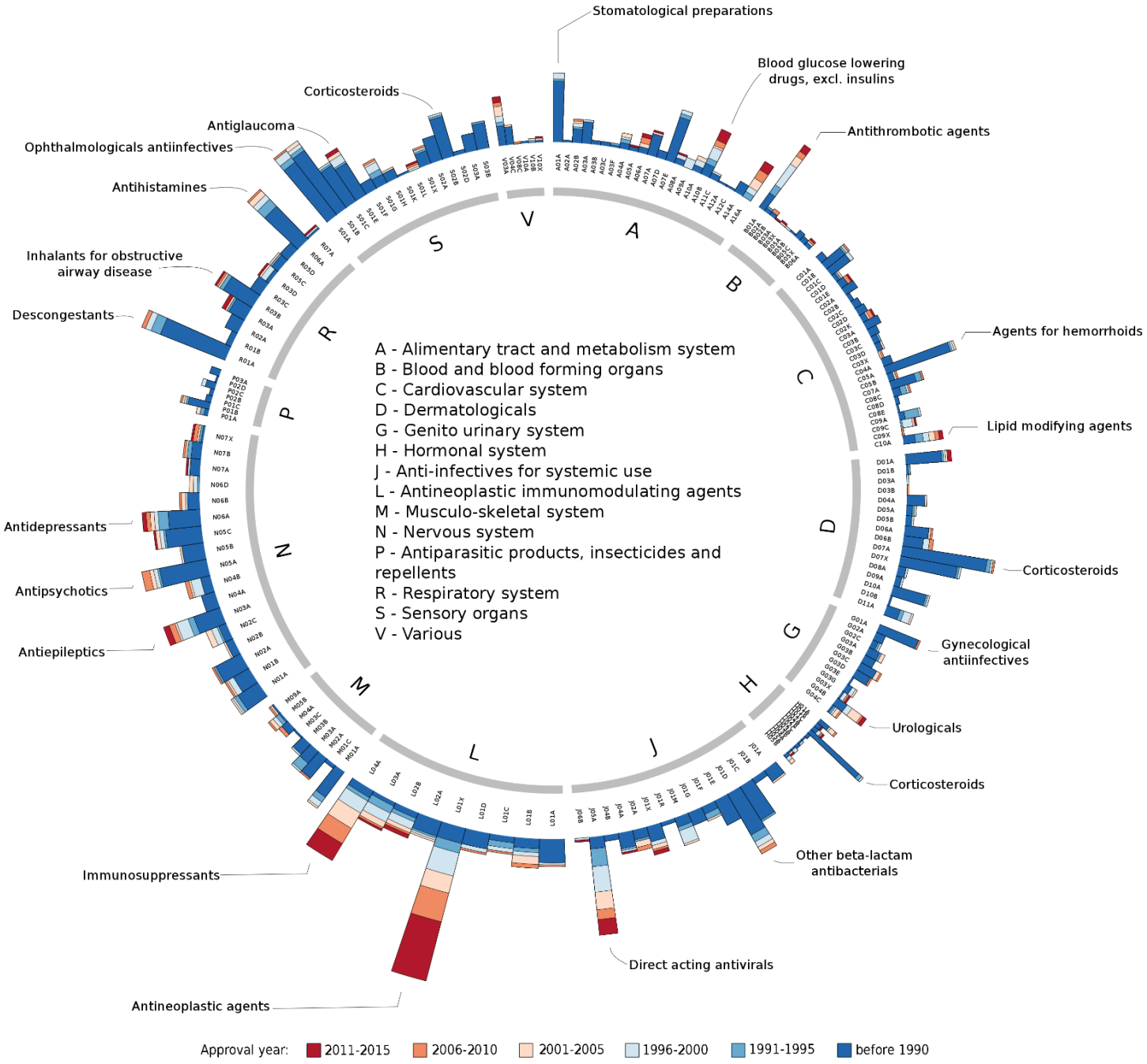


- Because most of the drugs Mechanism of Action is mediated by protein targets, DrugCentral collects and combines data on biological activity profile from multiple sources
- The ChEMBL database is the primary source of MoA data.
- Median target binding data shows that drugs targeting GPCR, NR, and Kinases are among the most potent drugs with potency in low nM range.

R. Santos & O. Ursu et al., *Nature Rev. Drug Discov*, **2017**, doi:10.1038/nrd.2016.230

10/20/16 revision

INNOVATION PATTERNS PER THERAPEUTIC AREA



Drugs distributed by ATC codes (levels 1-2). Concentric rings indicate ATC levels. Histograms represent the number of drugs distributed per year of first approval. Maximum scale: 100.

R. Santos & O. Ursu et al., *Nature Rev. Drug Discov*, **2017**, doi:10.1038/nrd.2016.230

10/20/16 revision

IDG KMC

COMMERCIAL IMPACT OF TARGET CLASSES

Target Class	Targets	APIs	Sales (B USD)	Market Share
GPCR	72	372	889.17	27.42%
Enzyme	88	234	683.14	21.06%
Nuclear receptor	16	111	340.13	10.49%
Transporter	18	82	323.99	9.99%
Ion channel	23	167	281.11	8.67%
Kinase	43	49	240.46	7.41%
Cytokine	9	12	184.29	5.68%
Other	43	68	300.83	9.28%

What are the most lucrative targets from a therapeutic perspective? We evaluated data from **IMS Health** on drug sales from 75 countries, including Europe, North America and Japan, aggregated over a 5-year period (2011–2015). After excluding botanicals, traditional Chinese and homeopathic medicines and drugs perturbing non-human targets, we identified 51,095 unique products. These were mapped to 1,069 active pharmaceutical ingredients from DrugCentral, corrected by number of APIs per product, then by number of Tclin targets per API.

T. Oprea et al., *Nature Rev. Drug Discov. poster,* Jan **2017**

11/15/16 revision

imshealth™

IDG KMC

TOP 20 DRUG TARGETS BY REVENUE

Gene	Protein Target	Action	Sales (B USD)
TNF	Tumor necrosis factor	Immunosuppressants	**163.39**
INSR	Insulin receptor	Hypoglycemic agents	**143.55**
NR3C1	Glucocorticoid receptor	Anti-inflammatory	**142.75**
HMGCR	3-hydroxy-3-methylglutaryl-coenzyme A reductase	Hypolipidemic agents	**122.55**
ATP4A/ ATP4B	Proton Pump (K^+ ATP-ase)	Anti-ulcer agents	**118.16**
AGTR1	Type-1 angiotensin II receptor	Antihypertensive agents	**99.98**
ADRB2	Beta-2 adrenergic receptor	Bronchodilators	**90.02**
OPRM1	Mu-type opioid receptor	Analgesics	**87.97**
PTGS2	Cyclooxygenase-2	Analgesics	**84.04**
DRD2	D2 dopamine receptor	Antipsychotic agents	**74.91**
CHRM[1-5]	Muscarinic acetylcholine receptor	Anticholinergics	**64.16**
SLC6A4	Sodium-dependent serotonin transporter	Antidepressants	**59.18**

Gene	Protein Target	Action	Sales (B USD)
HTR2A	5-hydroxytryptamine receptor 2A	Antipsychotics	**57.58**
CACNA1S/ CACNA1C/ CACNA1D/ CACNA1F	L-type calcium channel	Antihypertensive agents	**55.97**
SLC6A2	Sodium-dependent noradrenaline transporter	antidepressants & psychostimulants	**55.72**
VEGFA	Vascular endothelial growth factor A	antineovascularisation agents	**55.15**
HRH1	Histamine H1 receptor	antihistamines	**53.55**
IFNAR1/IFNAR2	Type I interferon receptor	immunostimulants	**51.40**
SCN[1,2,3,4,5,7,8,9,10,11]A	Voltage-gated sodium channel	antiarrhythmics & antiepileptics	**50.64**
ESR1	Estrogen receptor	contraceptives / estrogen agonists	**50.35**

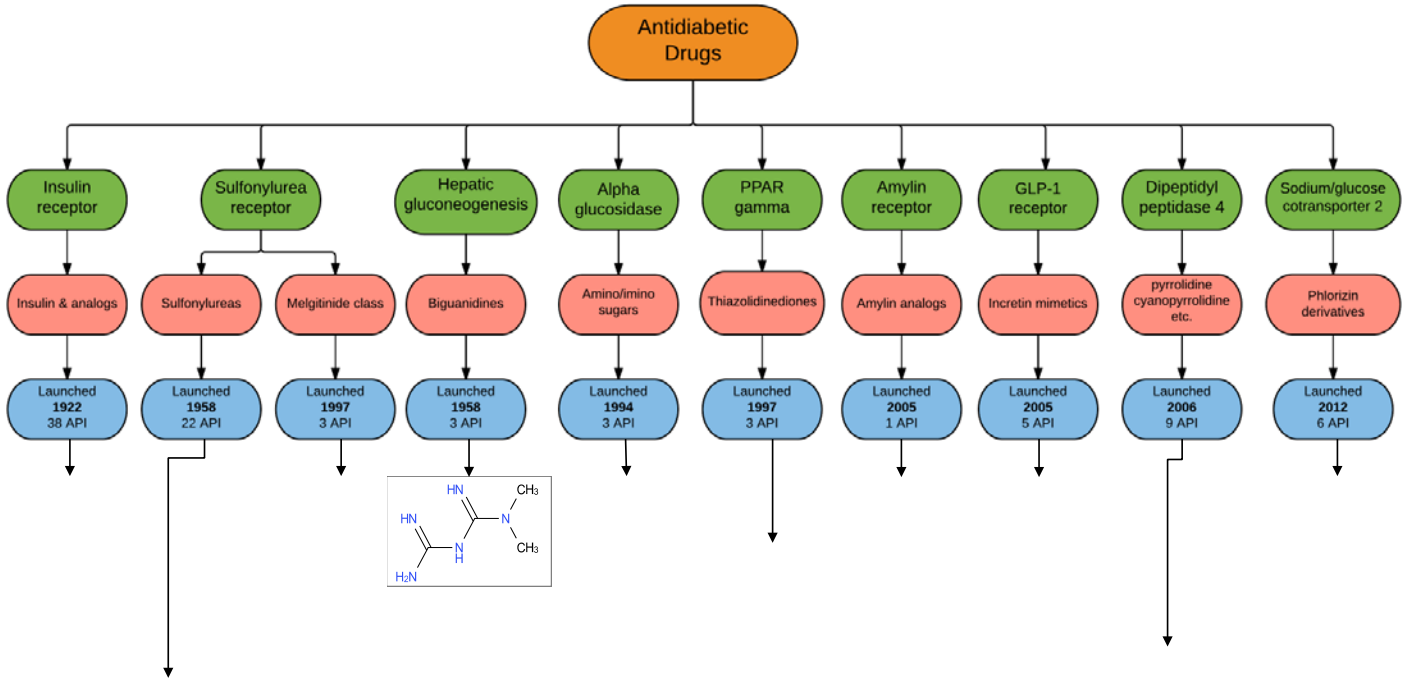
T. Oprea et al., *Nature Rev. Drug Discov. poster,* Jan **2017**

11/15/16 revision

imshealth™

IDG KMC

DRUG INDICATIONS: ANTIDIABETICS

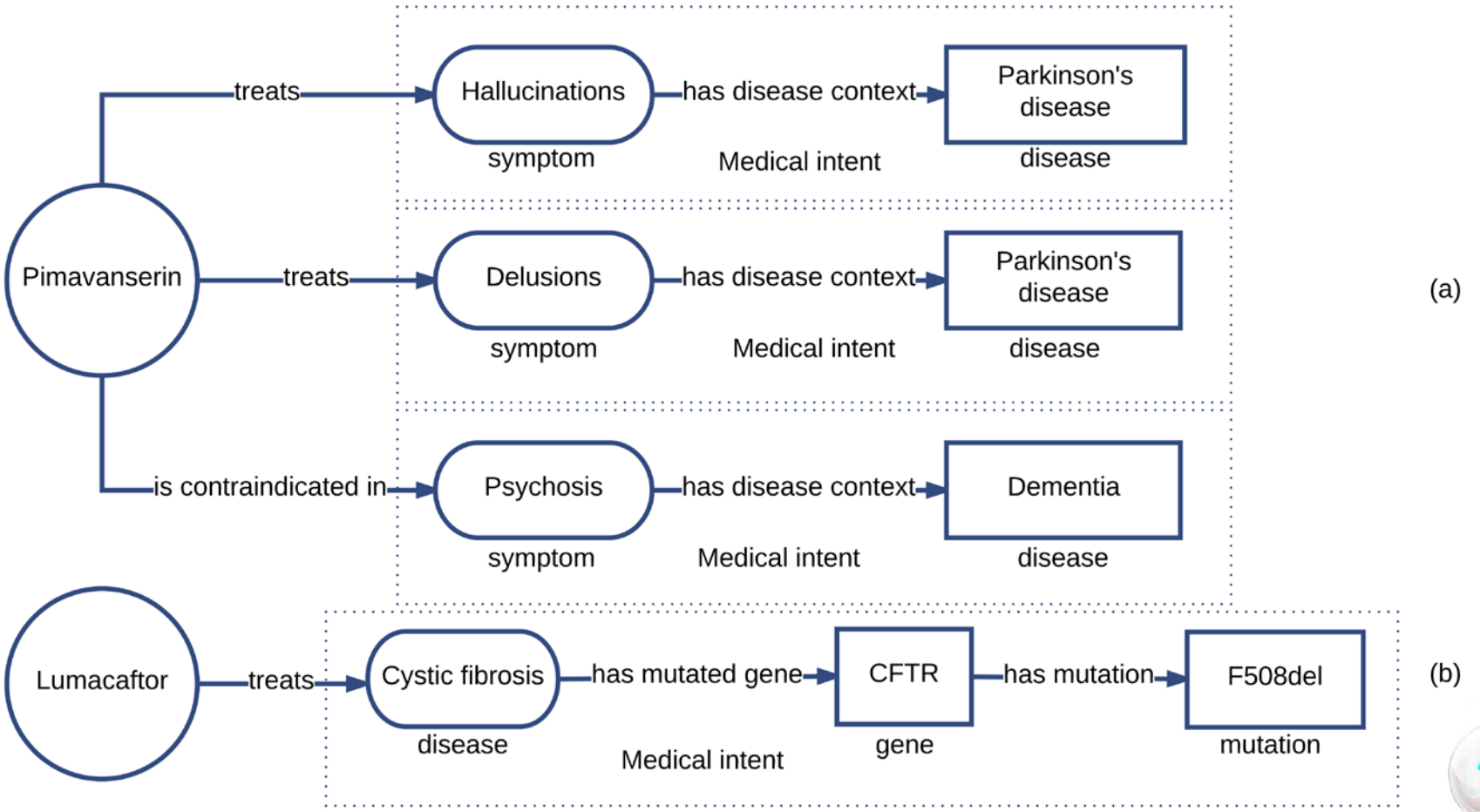


- By combining information for drug indications, targets, pharmacologic class, and structures, it is possible to get a quick overview for different areas of therapeutic interest, as an example drugs for diabetes

Oleg Ursu et al., *Nucl Acids Res*, database issue, **2017**, doi:10.1093/nar/gkw993

8/29/17 revision

ONTOLOGY-BASED CAPTURE OF THERAPEUTIC INTENT FROM DRUG INDICATIONS

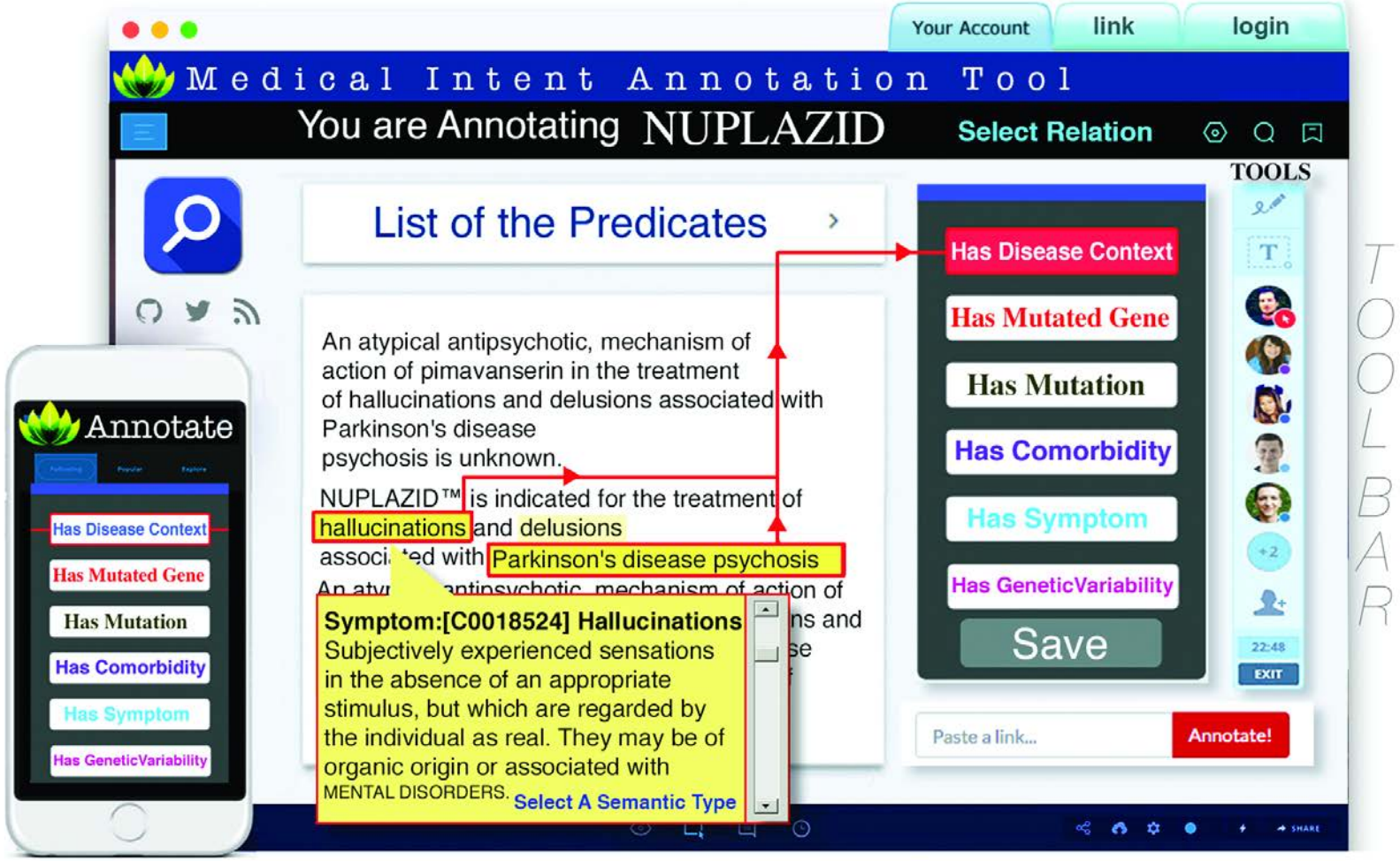


S.J. Nelson, T.I. Oprea et al., *J. Am. Med. Inform. Assoc.* **2017** (otx064)

5/25/17 revision

IDG KMC

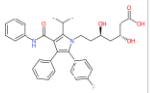
CURATION TOOL FOR ANNOTATING DRUG INDICATIONS



TOOLBAR

1/25/17 revision

DRUGCENTRAL.ORG

Molecule	Description
 Molfile Inchi **Synonyms:** • CI-981 • CI981 • atorvastatin calcium hydrate • atorvastatin calcium anhydrous • atorvastatin	A pyrrole and heptanoic acid derivative,HYDROXYMETHYLGLUTARYL-COA REDUCTASE INHIBITOR (statin), and ANTICHOLESTEREMIC AGENT that is used to reduce serum levels of LDL-CHOLESTEROL; APOLIPOPROTEIN B; AND TRIGLYCERIDES and to increase serum levels of HDL-CHOLESTEROL in the treatment of HYPERLIPIDEMIAS and prevention of CARDIOVASCULAR DISEASES in patients with multiple risk factors. • Molecular weight: 558.64 • Formula: C33H35FN2O5 • CLOGP: 4.46 • LIPINSKI: 1 • HAC: 5 • HDO: 4 • TPSA: 111.79 • ALOGS: -5.95 • RINGS: 4 • ROTB: 12

(a)

Dose	Unit	Route
20	mg	O

Date	Agency	Company	Orphan
Dec. 17, 1996	FDA	PFIZER	

Source	Code	Description
ATC	C10AA05	CARDIOVASCULAR SYSTEM LIPID MODIFYING AGENTS LIPID MODIFYING AGENTS, PLAIN HMG CoA reductase inhibitors

(b)

Disease	Relation	SNOMED_ID	DOID
Hypercholesterolemia	Indication	13644009	
Hypertension	Indication	38341003	DOID:10763
Arteriosclerotic Vascular Disease	Indication	72092001	DOID:2349
Disease of Liver	Contraindication	235856003	DOID:409
Rhabdomyolysis	Contraindication	240131006	
Pregnancy	Contraindication	289908002	

(c)

ID	Source
DB01076	DRUGBANK_ID
D000069059	MESH_DESCRIPTOR_UI
134523-03-8	SECONDARY_CAS_RN
D00887	KEGG_DRUG
7259	INN_ID
C0GEJ5QCSO	UNII

(d)

Target	Class	Swissprot	Action	Type	Activity value (- logIM)	Mechanism action	Bioact source	MoA source
3-hydroxy-3-methylglutaryl-coenzyme A reductase	Enzyme	HMDH_HUMAN	INHIBITOR	IC50	8	☑	WOMBAT-PK	CHEMBL
Cytochrome P450 3A4	Enzyme	CP3A4_HUMAN		IC50	5.29		CHEMBL	
3-hydroxy-3-methylglutaryl-coenzyme A reductase	Enzyme	HMDH_RAT		IC50	8.42		CHEMBL	

(e)

Product	Category	Ingredients	NDC	Form	Quantity	Route	Marketing	Label
Caduet	HUMAN PRESCRIPTION DRUG LABEL	2	0069-2150	TABLET, FILM COATED	10 mg	ORAL	NDA	19 sections
Lipitor	HUMAN PRESCRIPTION DRUG LABEL	1	0071-0157	TABLET, FILM COATED	40 mg	ORAL	NDA	19 sections
Atorvastatin Calcium	HUMAN PRESCRIPTION DRUG LABEL	1	0378-2015	TABLET, FILM COATED	10 mg	ORAL	ANDA	19 sections
Amlodipine besylate and atorvastatin calcium	HUMAN PRESCRIPTION DRUG LABEL	2	0378-4511	TABLET, FILM COATED	20 mg	ORAL	ANDA	18 sections

(f)

- Live presentation should follow (Oleg Ursu)

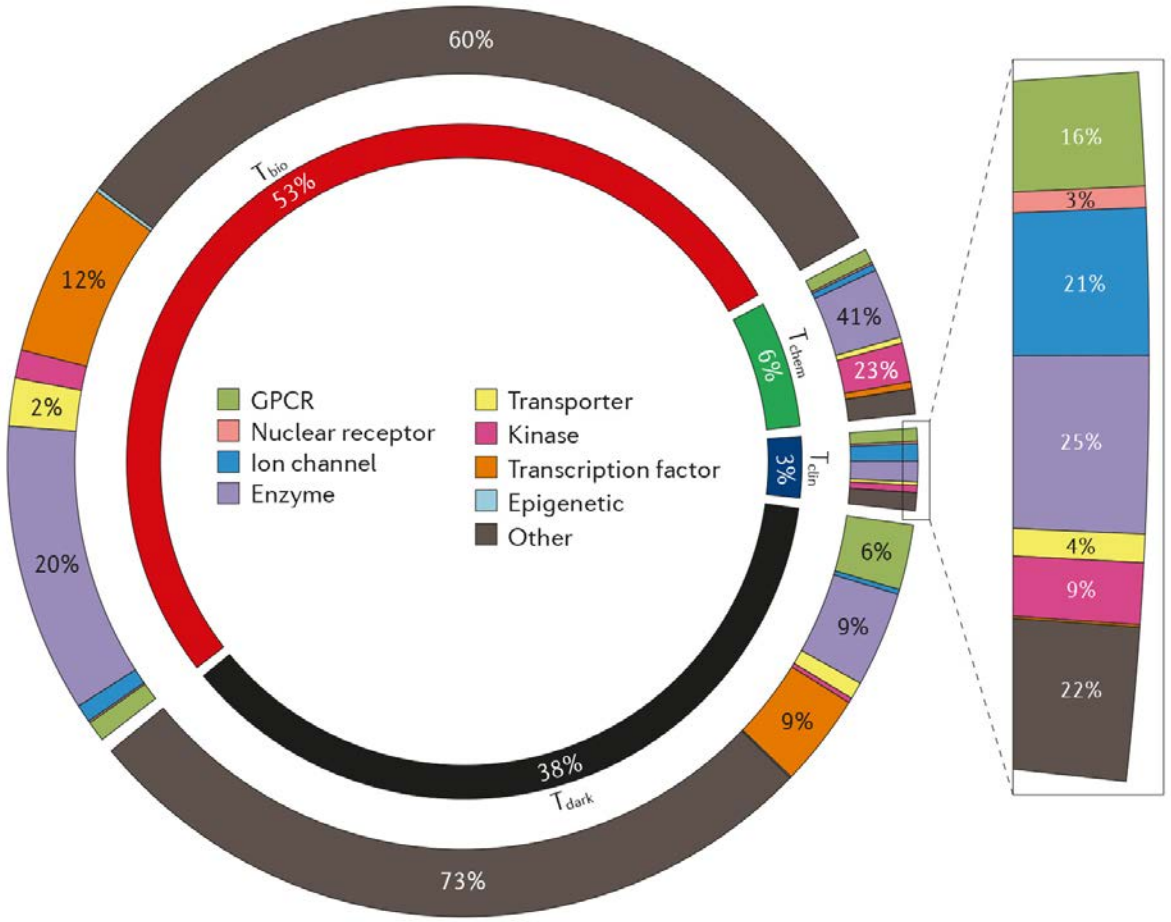
Oleg Ursu et al., *Nucl Acids Res*, database issue, **2017**, doi:10.1093/nar/gkw993

8/29/17 revision

TAKE HOME MESSAGE 2

LINKING DRUGS TO TARGETS AND INDICATIONS GUIDES FURTHER RESEARCH

TARGET DEVELOPMENT LEVEL



Nature Reviews | Drug Discovery

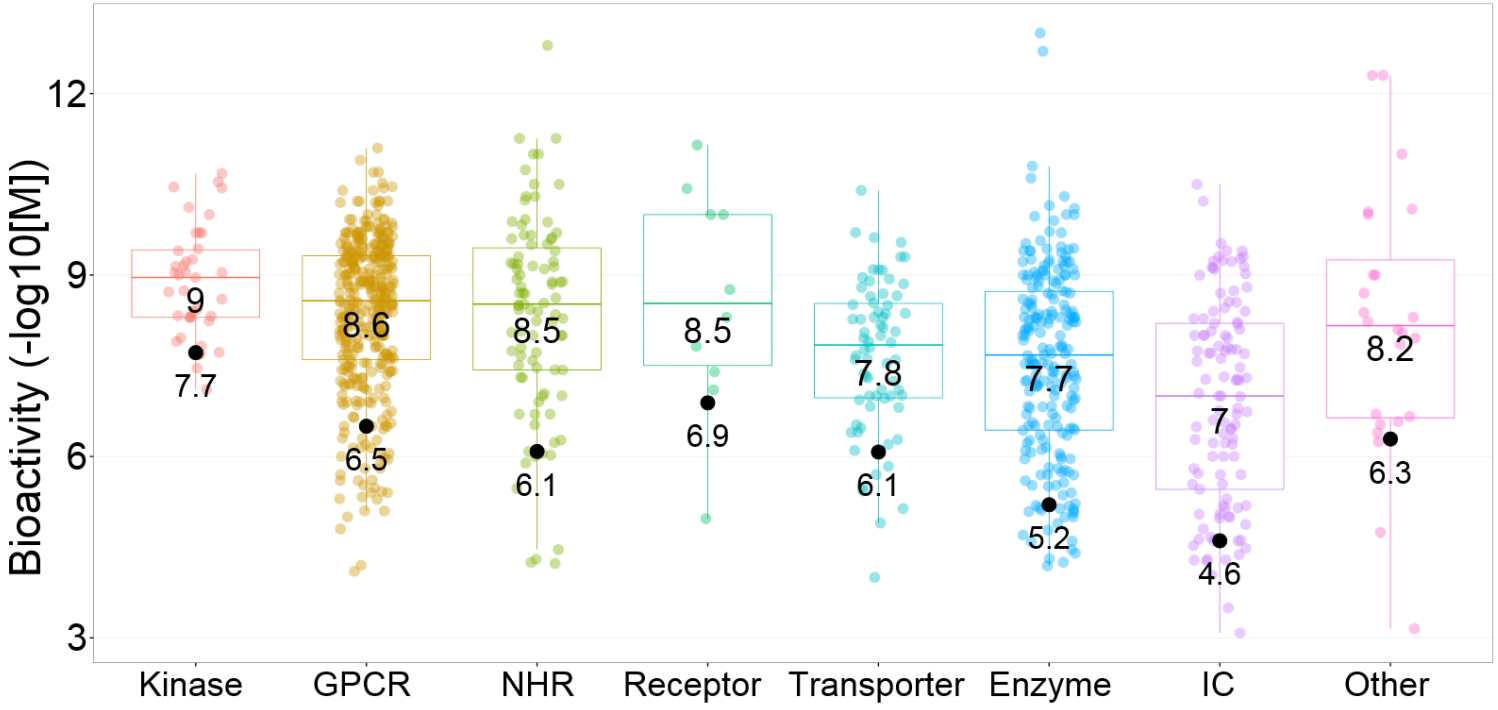
- Most protein classification schemes are based on structural and functional criteria.
- For therapeutic development, it is useful to understand how much and what types of data are available for a given protein, thereby highlighting well-studied and understudied targets.
- Proteins annotated as drug targets are **Tclin**
- Proteins for which *potent* small molecules are known are **Tchem**
- Proteins for which biology is better understood are **Tbio**
- Proteins that lack antibodies, publications or Gene RIFs are **Tdark**

T. Oprea et al., *Nature Rev. Drug Discov. poster,* Jan **2017**

11/09/16 revision

IDG KMC

D-T DEVELOPMENT LEVEL 1



Bioactivities of approved drugs (by Target class)

- **Tclin** proteins are associated with drug Mechanism of Action (MoA)
- **Tchem** proteins have bioactivitis in ChEMBL and DrugCentral, + human curation for some targets
  - Kinases: <= 30nM
  - GPCRs: <= 100nM
  - Nuclear Receptors: <= 100nM
  - Ion Channels: <= 10μM
  - Non-IDG Family Targets: <= 1μM

ChEMBL: database of bioactive chemicals
https://www.ebi.ac.uk/chembl/
DrugCentral: online drug compendium
http://drugcentral.org/

10/19/16 revision

D-T DEVELOPMENT LEVEL 2

- **Tbio** proteins lack small molecule annotation cf. Tchem criteria, and satisfy one of these criteria:
  - protein is above the cutoff criteria for **Tdark**
  - protein is annotated with a GO Molecular Function or Biological Process leaf term(s) with an Experimental Evidence code
  - protein has confirmed OMIM phenotype(s)
- **Tdark** ("understudied proteins") have little information available, and satisfy these criteria:
  - PubMed score (text mining) from Jensen Lab < 5
  - <= 3 Gene RIFs
  - <= 50 Antibodies available according to antibodypedia.com

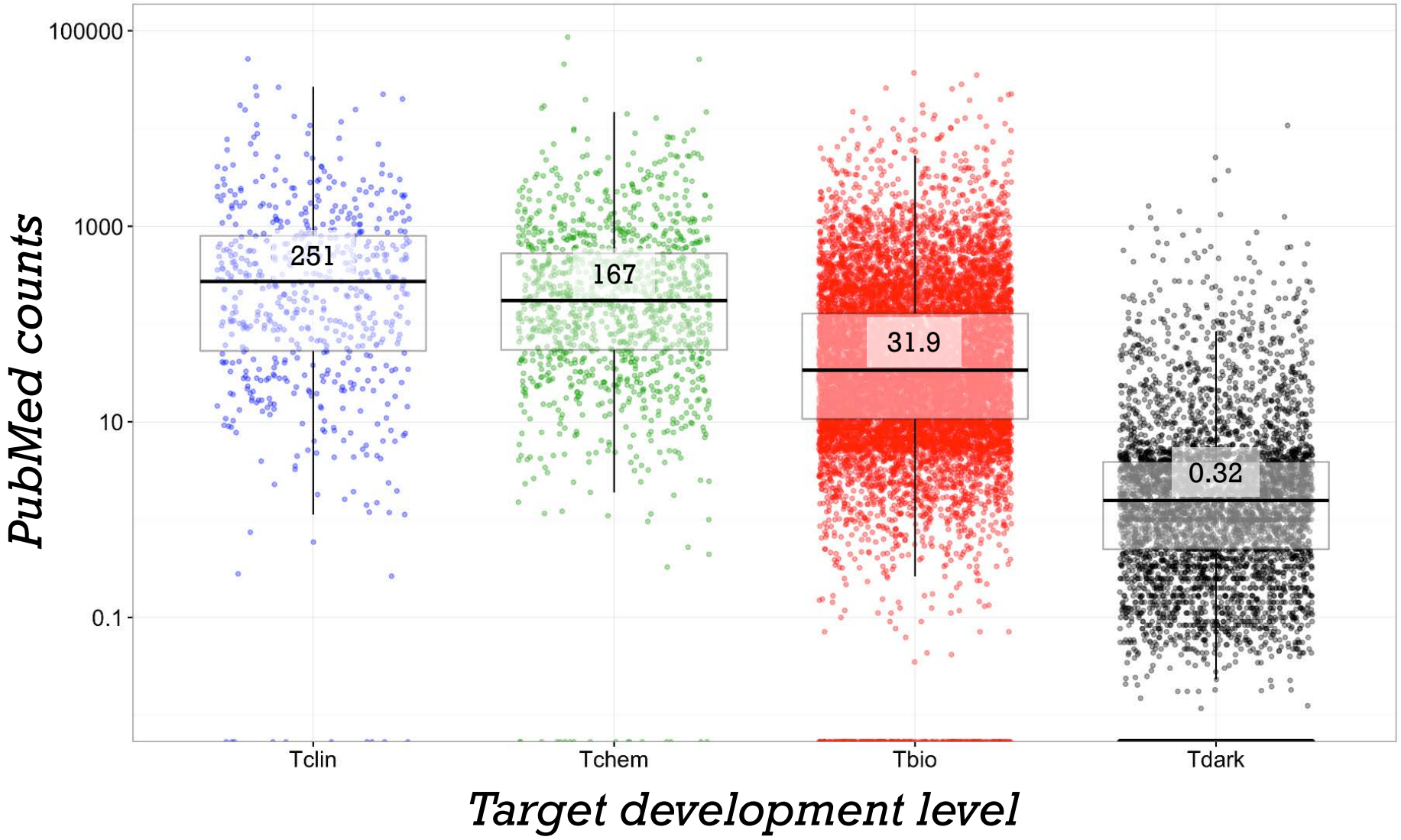
Fractional paper count

$$PubMed\ score = \sum_{j \in D} \frac{n_{ij}}{n_{.j}}$$

8/20/15 revision

IDG KMC

TDL VS PUBLICATIONS

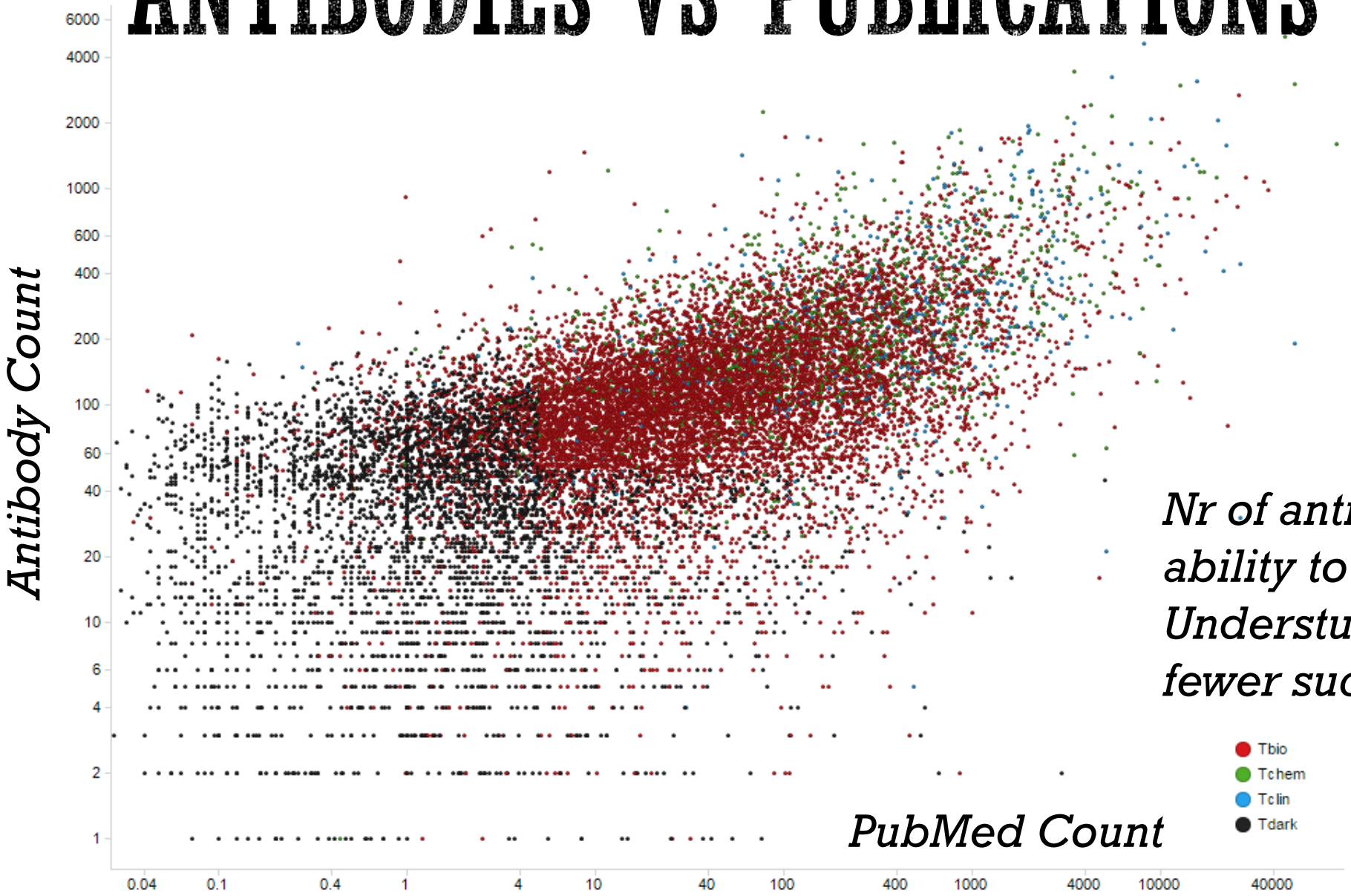


We used name entity recognition software (from LJ Jensen's lab) to evaluate nearly ~27 million abstracts (including ~2M full text articles) to derive a publication score per protein

P.F. Sune, et all, *Methods* **2105**, 74:83-89

11/14/16 revision

ANTIBODIES VS PUBLICATIONS



Nr of antibodies reflects our ability to characterize proteins. Understudied proteins have fewer such tools.

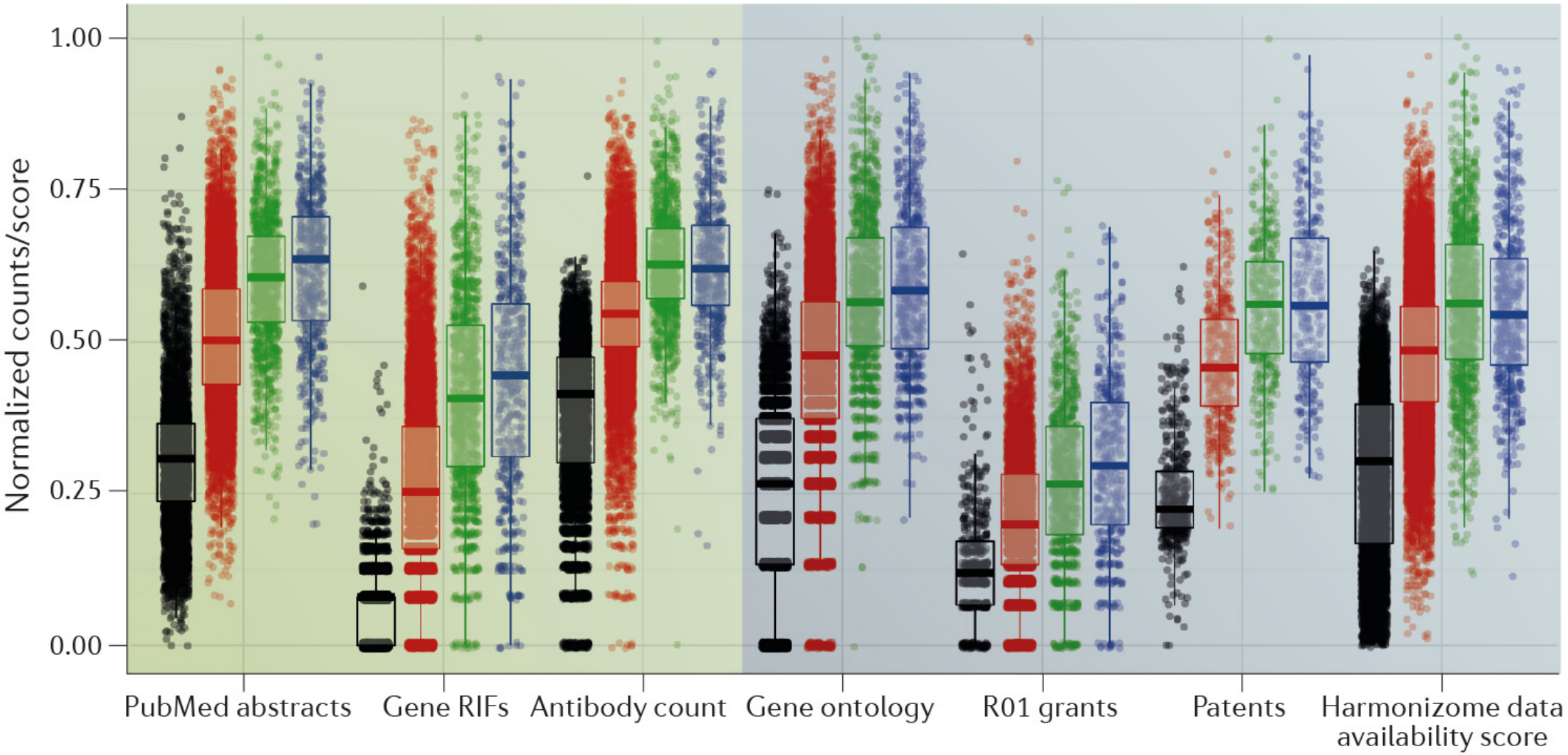
Human proteome (20,186 proteins). Spearman R = 0.68. Axes in log scale.

Antibodypedia.com

8/31/16 revision

IDG KMC

TDL: EXTERNAL VALIDATION



Tdark parameters differ from the other TDLs across the 4 external metrics cf. Kruskal-Wallis post-hoc pairwise Dunn tests

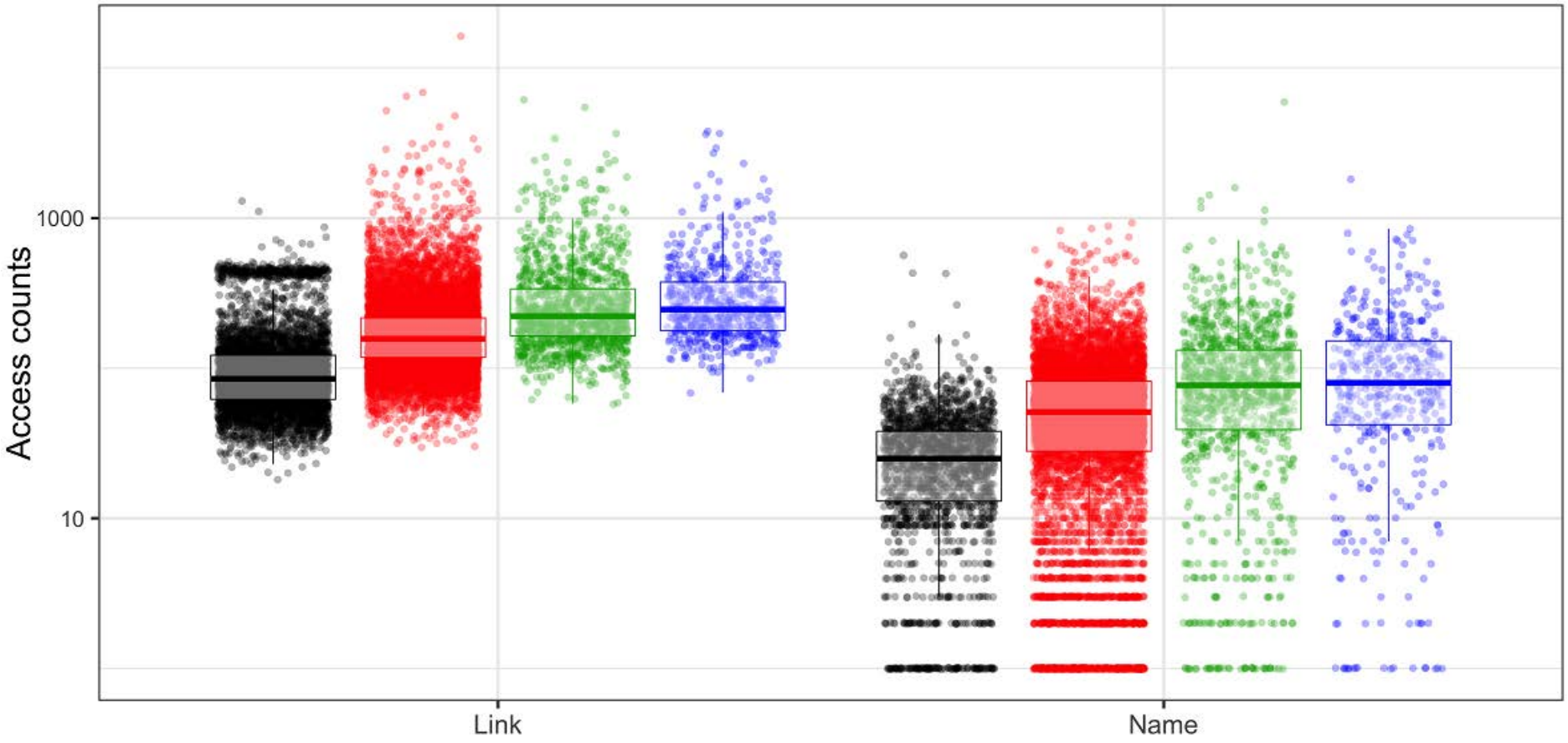
Nature Reviews | Drug Discovery

T. Oprea et al., *Nature Rev. Drug Discov. poster,* Jan **2017**

8/31/16 revision

IDG KMC

PATTERNS OF CURIOSITY



"Counts by Name" == users accessing the STRING website and typing in a gene symbol.
"Counts by Link" == users accessing the network for a gene in STRING by linking to it from another resource

Data courtesy of Christian von Mehring, STRING-db

3/28/17 revision

IDG KMC

TAKE HOME MESSAGE 3

THERE IS A KNOWLEDGE DEFICIT

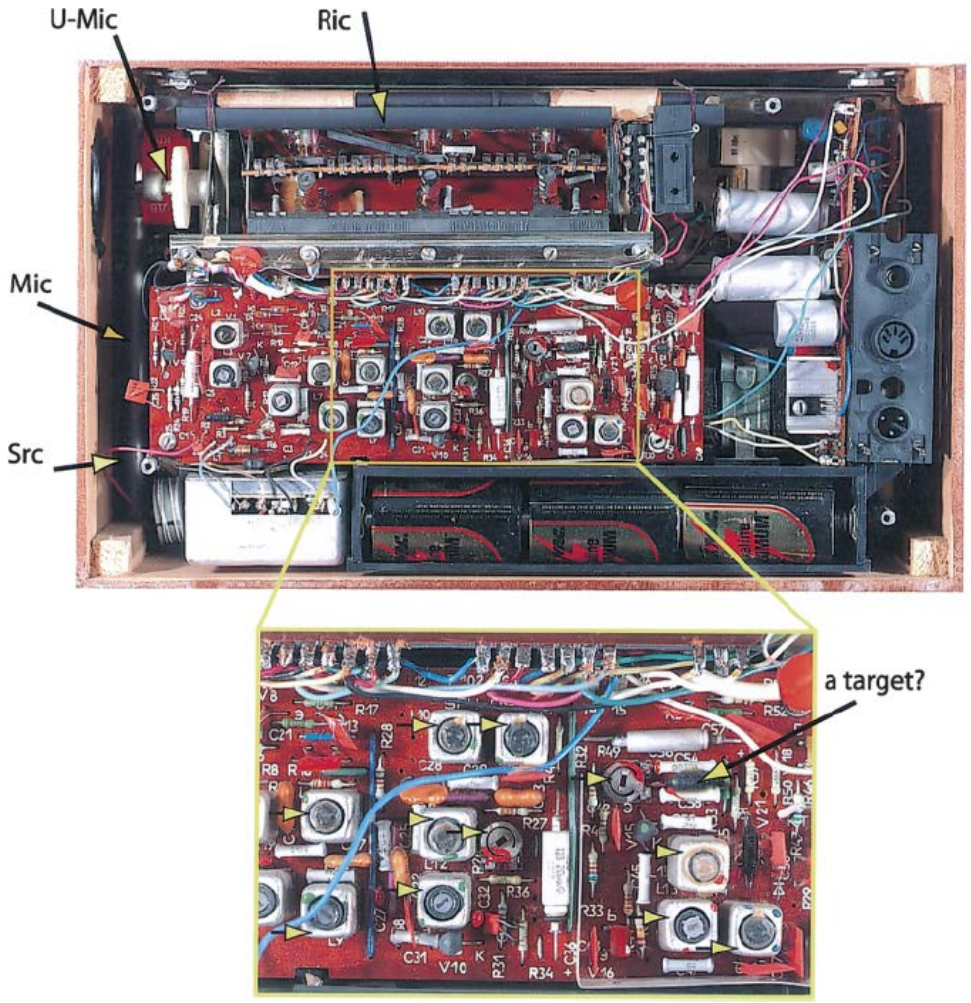
over 37% of the proteins remain understudied (Tdark)

~10% of the Proteome (Tclin & Tchem) can be targeted by small molecules

3/28/17 revision

IDG KMC

BIOLOGY AND ALTERNATIVE FACTS



The absence of a quantitative language *"is the flaw of biological research"* or "The more facts we learn the less we understand".

A biologist describing a Radio:

Src: Serendipitously Recovered Component
(wire connecting to the antenna, which is)
Mic: Most Important Component
but you really need
Ric: Really Important component *(rectifier)*
and U-Mic (Undoubtedly Most Important Component) *[the switch]*

When little is known, don't expect knowledge to accumulate quickly

Y. Lazebnik, Cancer Cell **2002**,2:179-182

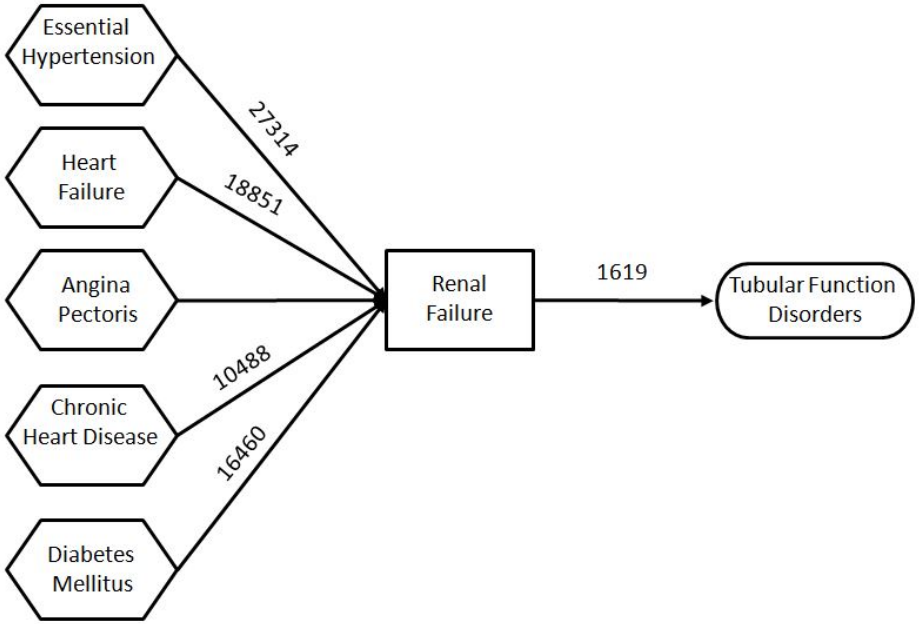
CONCEPTUAL FALLACY: SEPARATION BY ORGAN/CELL

- Medicine maintains this separation for necessity: by organ (e.g., cardiology, ophthalmology), by disease category (e.g., oncology, infection)

- NIH Institutes are organized in a similar way.
- Many pharma companies are organized by Therapy Area.
- … yet genes / proteins / pathways do not observe such separation
- **The impact of this "mental divide" in science has yet to be understood.**

IDG KMC

PRE- CHRONIC KIDNEY DISEASE (5 YEARS)



A.B. Jensen et al., *Nature Communications* **2014** 5:4022

C. Bologa, O. Ursu, J.J. Yang & T.I. Oprea (UNM)

data from Cerner HealthFacts DB

10/20/16 revision

UNM SCHOOL of MEDICINE

DISEASES ARE CONCEPTS

- Diseases lack physical manifestation outside patients.
- **The search for cures has to be patient centered**
- …Animal models should be combined with patient data mining
- Remember David Horrobin's papers…

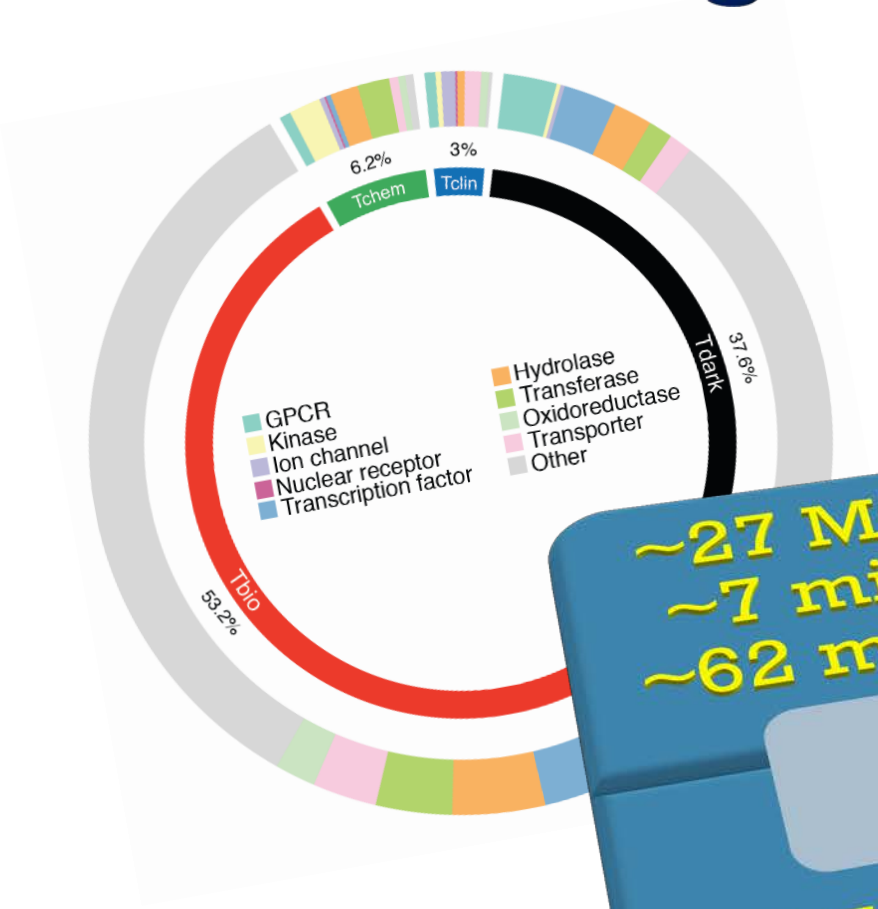
OPINION

Modern biomedical research:
an internally self-consistent universe
with little contact with medical reality?

David F. Horrobin

IDG KMC

Illuminating the Druggable Genome Knowledge Management Center



~27 Million Papers
~7 million Patents
~62 million Patients

~20,000 Proteins

Seeking New Knowledge

~15,000 Diseases

~4,500 Drugs

IDG KMC

DRUGCENTRAL IS PART OF OUR TRANSLATIONAL INFORMATICS DIVISION

Oleg Ursu et al., *Nucl Acids Res*, database issue, **2017**, doi:10.1093/nar/gkw993

8/29/17 revision

IDG KMC