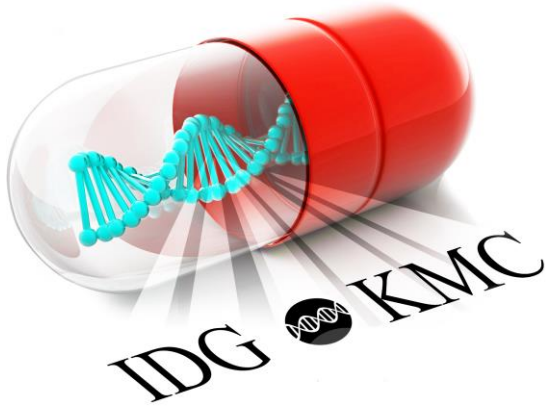
ILLUMINATING THE DRUGGABLE GENOME



and the Quest for New Drug Targets

Tudor I. Oprea

11/30/2017

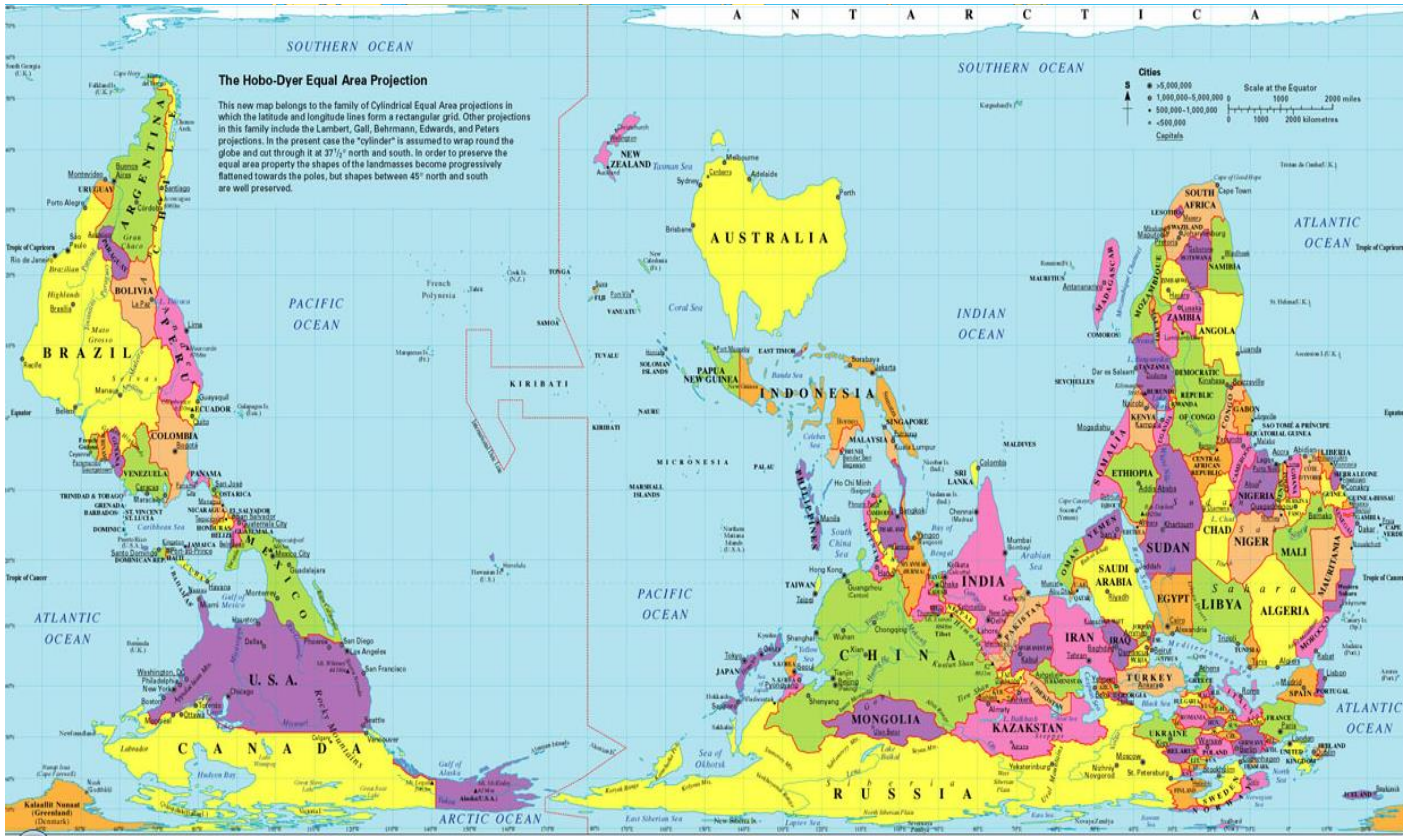
Global Engage Pharma Informatics

Lisbon, Portugal

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

UNM SCHOOL of MEDICINE

WHERE DOES KNOWLEDGE COME FROM?



- We navigate "through time and space" by defining & observing conventions
- Geographical maps are a matter of convention…
- Informatics and data science need conventions, without which there can be little progress
- **What do we know?**
- **When do we know it?**

UNM SCHOOL of MEDICINE

WHAT IS TRUTH?

- **Philosophers and scientists alike struggle with this one**

  …humans are not well equipped to handle contradictions; machines are worse

- Machine learning (binary classification models) reflects a simplified view of the Universe: *"A" or "non-A", True or False.* No alternatives
- Science exists in a world of relative truths: Stanley Cup winners change annually, gravity waves only recently confirmed,
- Humans accept political polls and weather on TV as substitutes for truth…
- We're perfectly capable to live in the murky world of half-truths and half-lies, where facts can change overnight.
- *Facts (and data) have an expiration date*
- No mathematical models can help discriminate truth from falsehood

… ALTERNATIVE FACTS IN SCIENCE

- **Confirmation bias pervades Publications. Obfuscation pervades Patents.**

  …we publish positive results, rarely bad ones, and mask "true IP" in patents

- How we write papers : *Scientist or team seeks "A", finds "A[1]" – then the paper gets written about "A[1]"*
- Serendipity & confirmation bias are rarely acknowledged, but we tend to find what we seek…
- We retro-fit new data into old models, until forced to accept "new science"
- Worse, people lie – IRL, but in science as well
- With machine learning models, we take an optimistic view – *our model would have worked if only*…
- By the time we can have meaningful predictions, the system is capable of modeling/capturing all states…

UNM SCHOOL of MEDICINE

MODELING TRUTH

- The ubiquity of *machine learning* has shifted the narrative from explanatory science (Aristotle's original intent) to *predictive science*

  …yet we can't handle Black Swans (Turkey surprise) or the fallacy of induction

- Despite considerable many-valued logic work (e.g., Gödel, Łukasiewicz), we continue to develop ML models using true/false statements.
- We should shift from binary models to four-valued logic: A, non-A, both (A, non-A) and neither A nor non-A. For example, to *drug* {confirmed} vs. *non-drug*, we should add *both drug and non-drug* (e.g., an antifungal used as pesticide), and *neither* ("don't care")
- This may improve machine learning … as we sift through computational assertions that are assigned varied degrees of truth values, it is likely that we will be better able to model (and understand) biology.
- Fast predictions further shifting the paradigm from "important" to "urgent"

TAKE HOME MESSAGE 1

DATA HAS BEEN EMBIGGENED

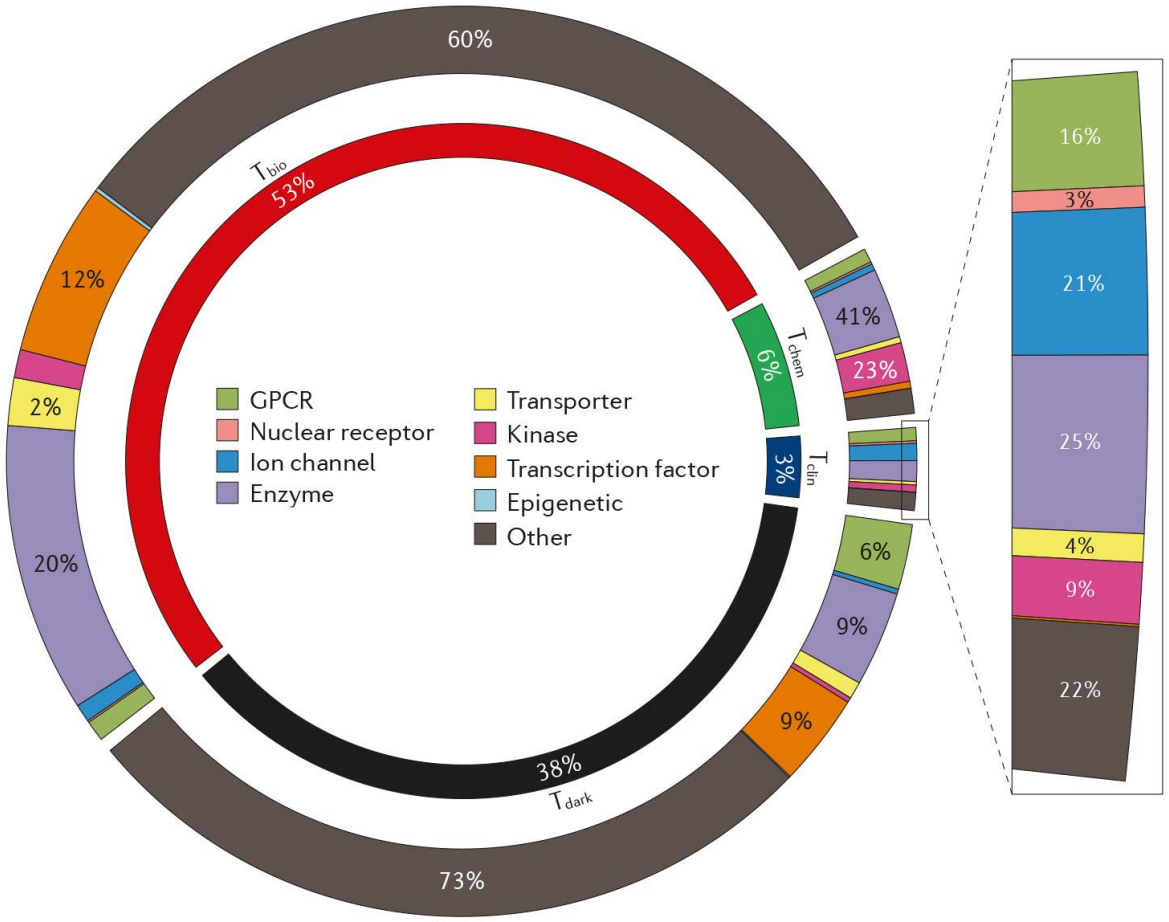
How reliable are Big Data?
Can we sift "True Data" from "False Data"?!

https://en.wiktionary.org/wiki/embiggen

6/20/17 revision

IDG KMC

TARGET DEVELOPMENT LEVEL



Nature Reviews | Drug Discovery

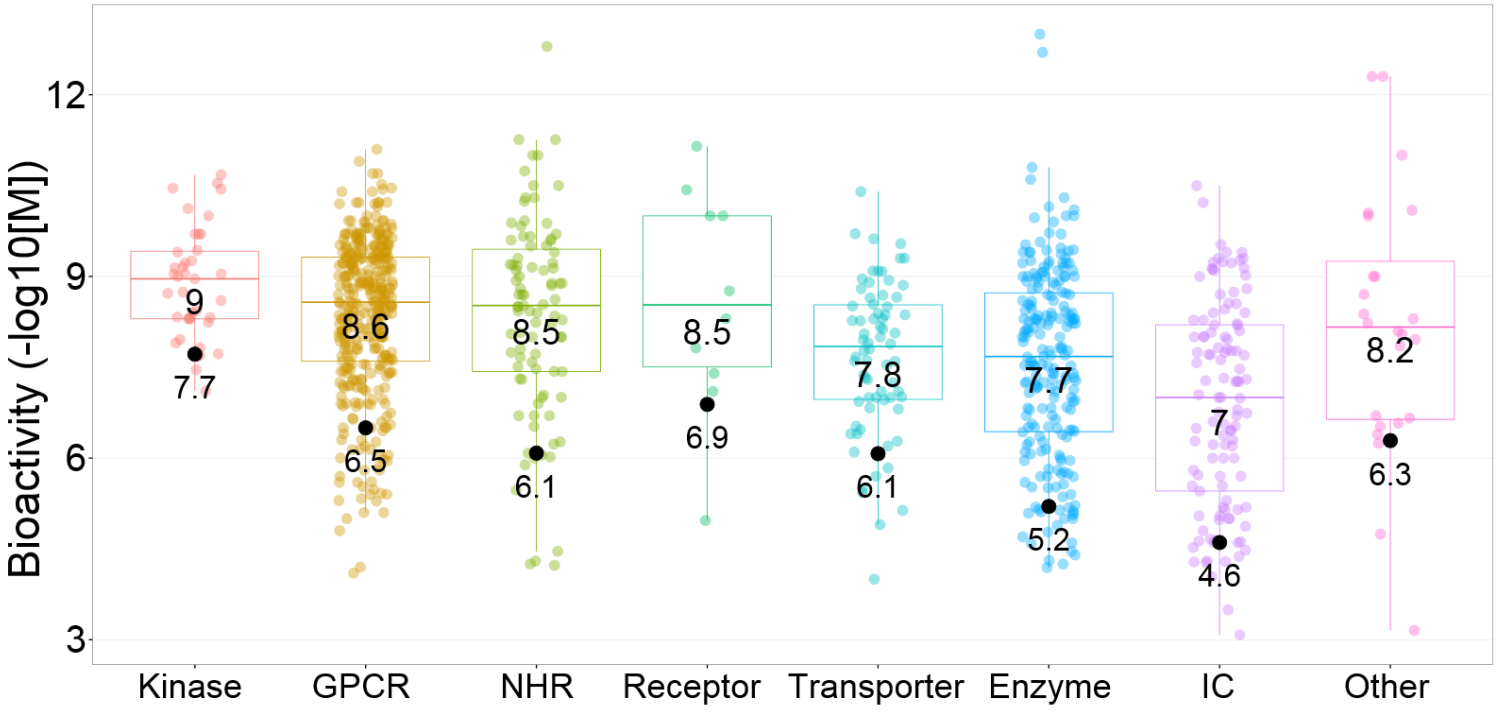
- Most protein classification schemes are based on structural and functional criteria.
- For therapeutic development, it is useful to understand how much and what types of data are available for a given protein, thereby highlighting well-studied and understudied targets.
- Proteins annotated as drug 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** ("ignorome") have little information available, and satisfy these criteria:
  - PubMed text-mining score from Jensen Lab < 5
  - <= 3 Gene RIFs
  - <= 50 Antibodies available according to antibodypedia.com

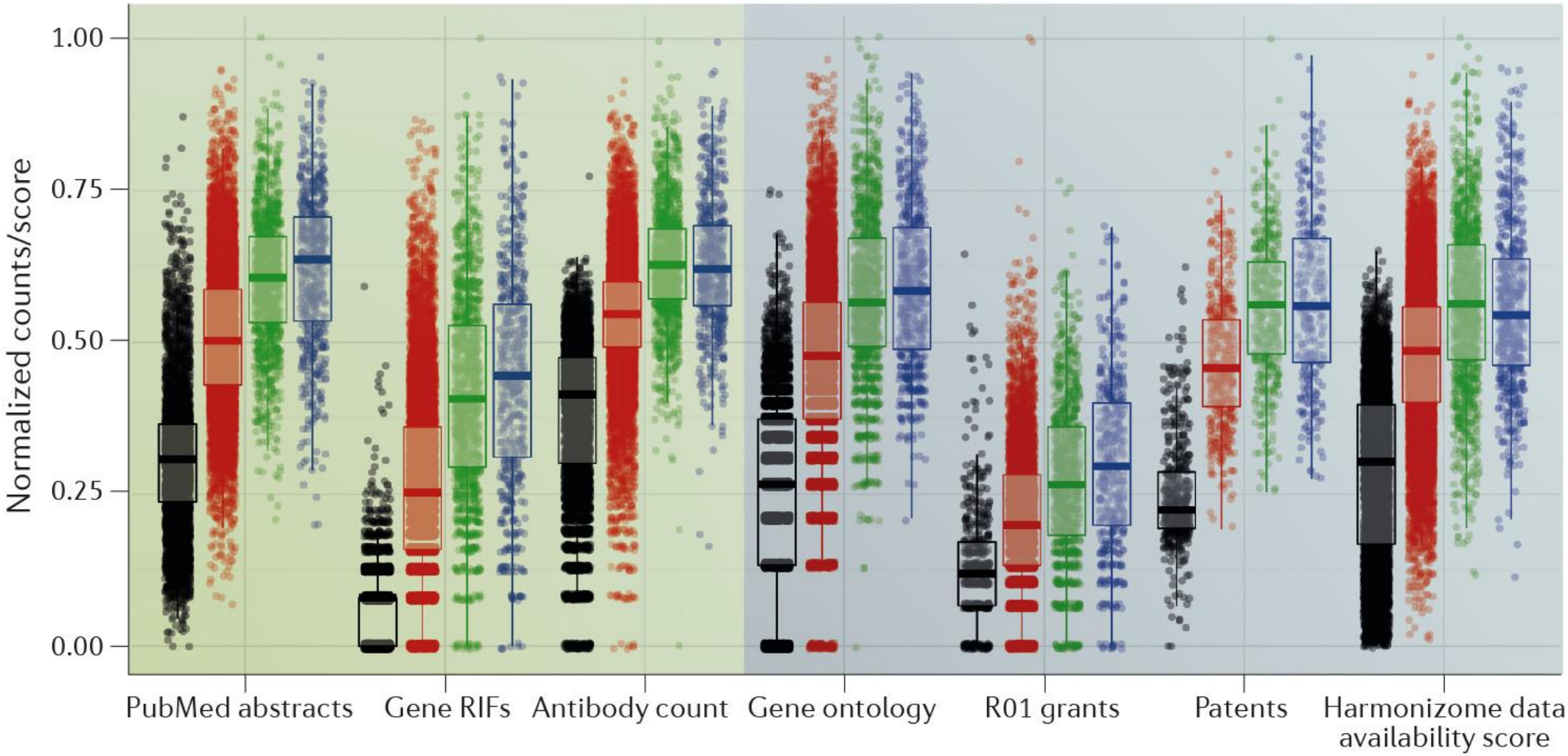
Fractional paper count

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

8/20/15 revision

IDG KMC

TDL: EXTERNAL VALIDATION



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

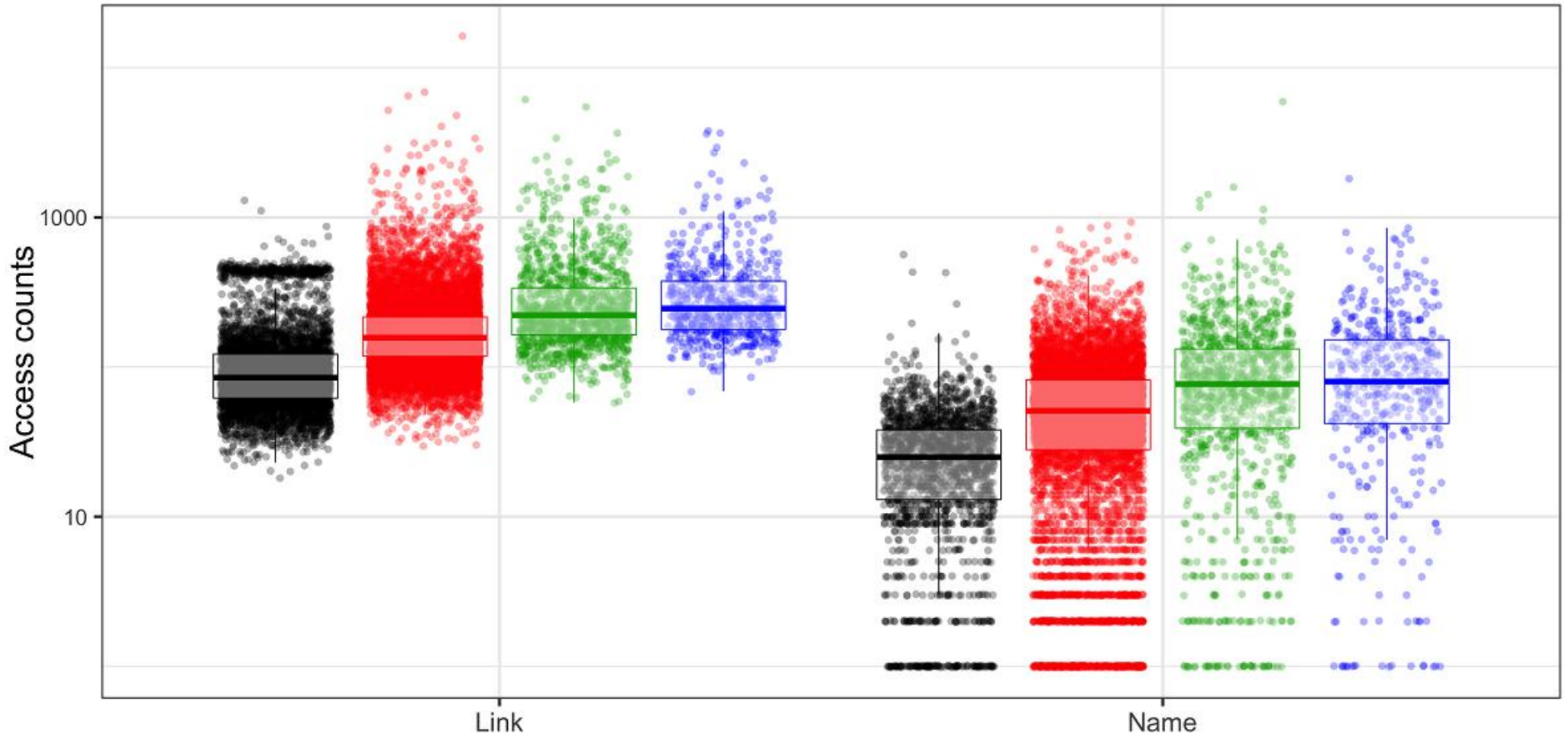
Nature Reviews | Drug Discovery

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

THE CAUSALITY DILEMMA

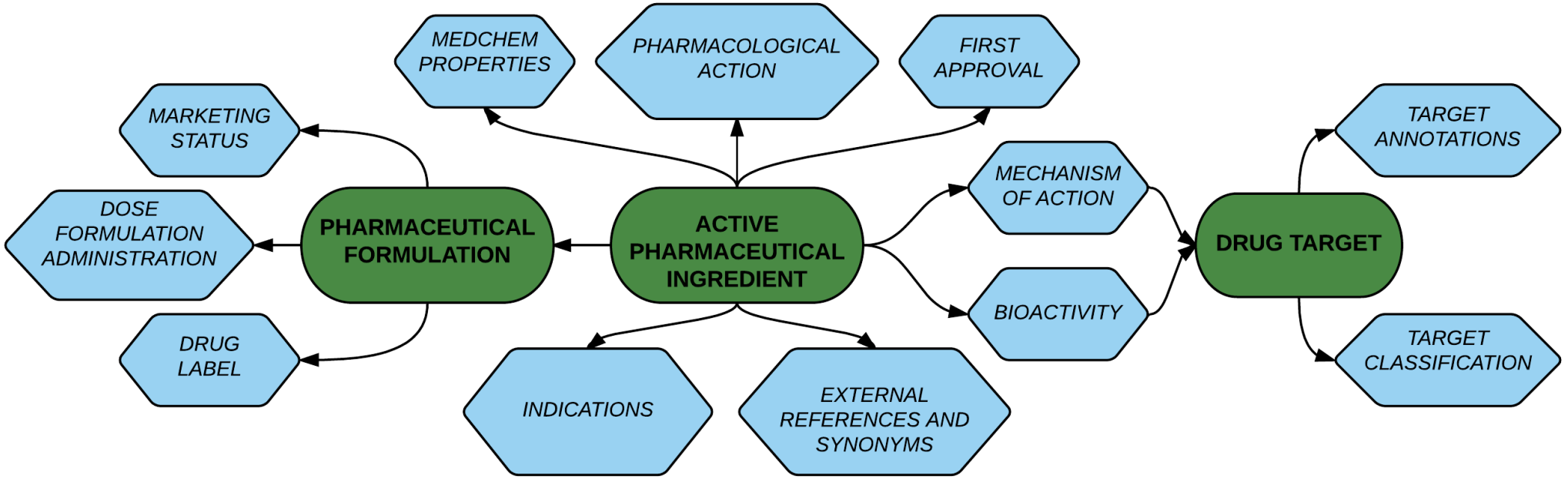
- Are Tdark proteins underfunded because there is no scientific interest in this category, or is the lack of knowledge perpetuated by lack of funding?
- It is possible that the absence of high quality, well characterized molecular probes may be a root cause for this situation.
- However, lack of tools leads to lack of interest, and lack of interest diminishes the probability of such tools being developed

WHY SHOULD ANYONE FUND TDARK?

- Leptin Receptor: Tdark in 1995, Marketed Drug in 2014
- Smoothened Receptor: Tdark in 1997, Marketed Drug in 2012
- Sphingosine 1-phosphate receptor 1: Tdark in 1997, Marketed Drug in 2010
- Orexin Receptors: Tdark in 1997, Marketed Drug in 2014
- Proprotein convertase subtilisin/kexin type 9 (PCSK9): Tdark in 1998, Marketed drugs in 2015
- Ghrelin Receptor: Tdark in 1999, successful phase 3 RCT in 2016
- TNFRSF17, aka BCMA: Tdark in 2000, phase 1 RCT in 2017

- on average it takes 15-20 years for Tdark to bear fruit

DRUGCENTRAL DATA STRUCTURE

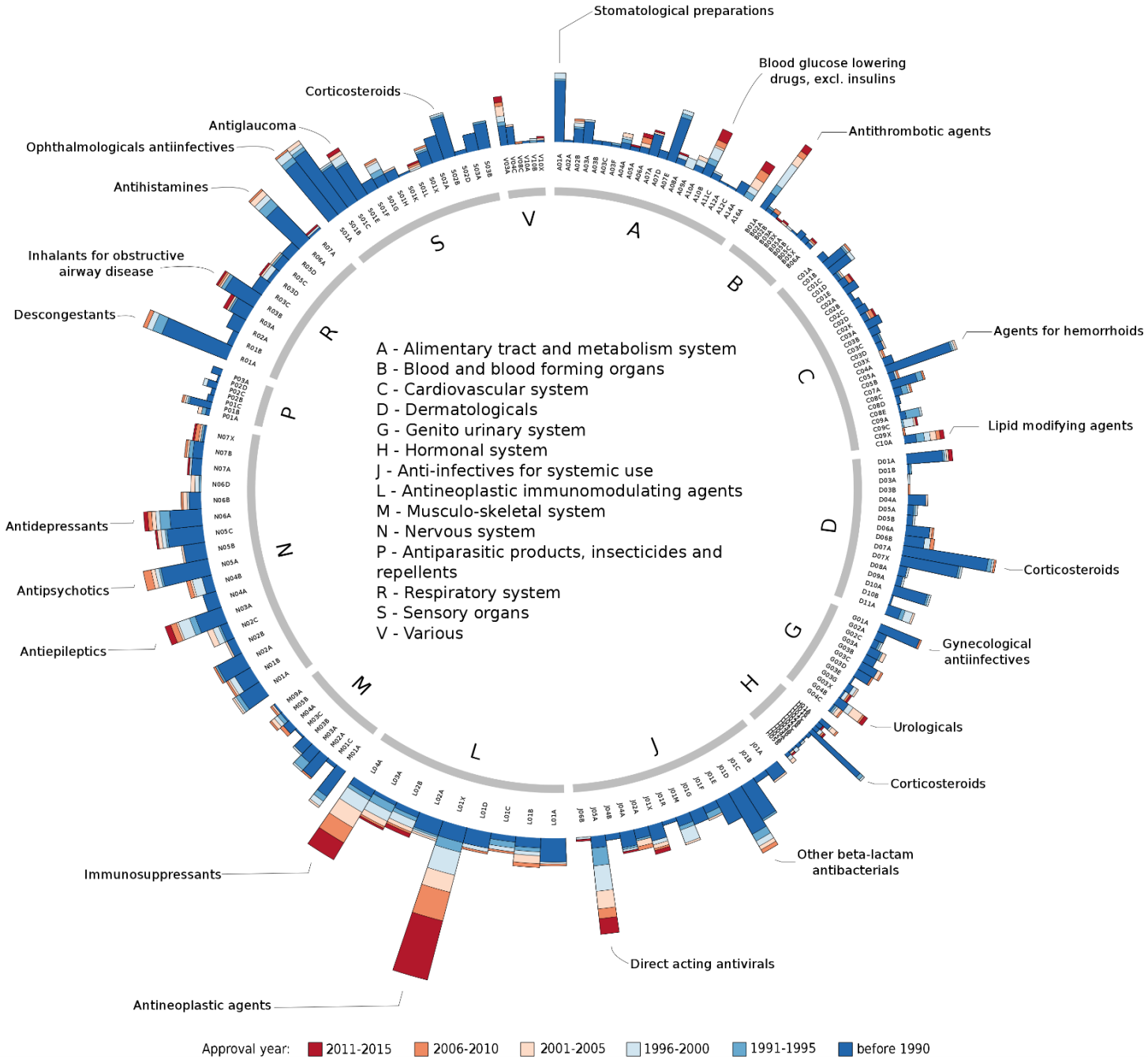


- Initially to answer "how many drugs are out there"…
- Mapped products (what patients and docs call "drugs") onto active ingredients (what scientists call "drugs")
- Also wanted to know how many drug targets there are……

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

10/20/16 revision

INNOVATION PATTERNS PER THERAPEUTIC AREA



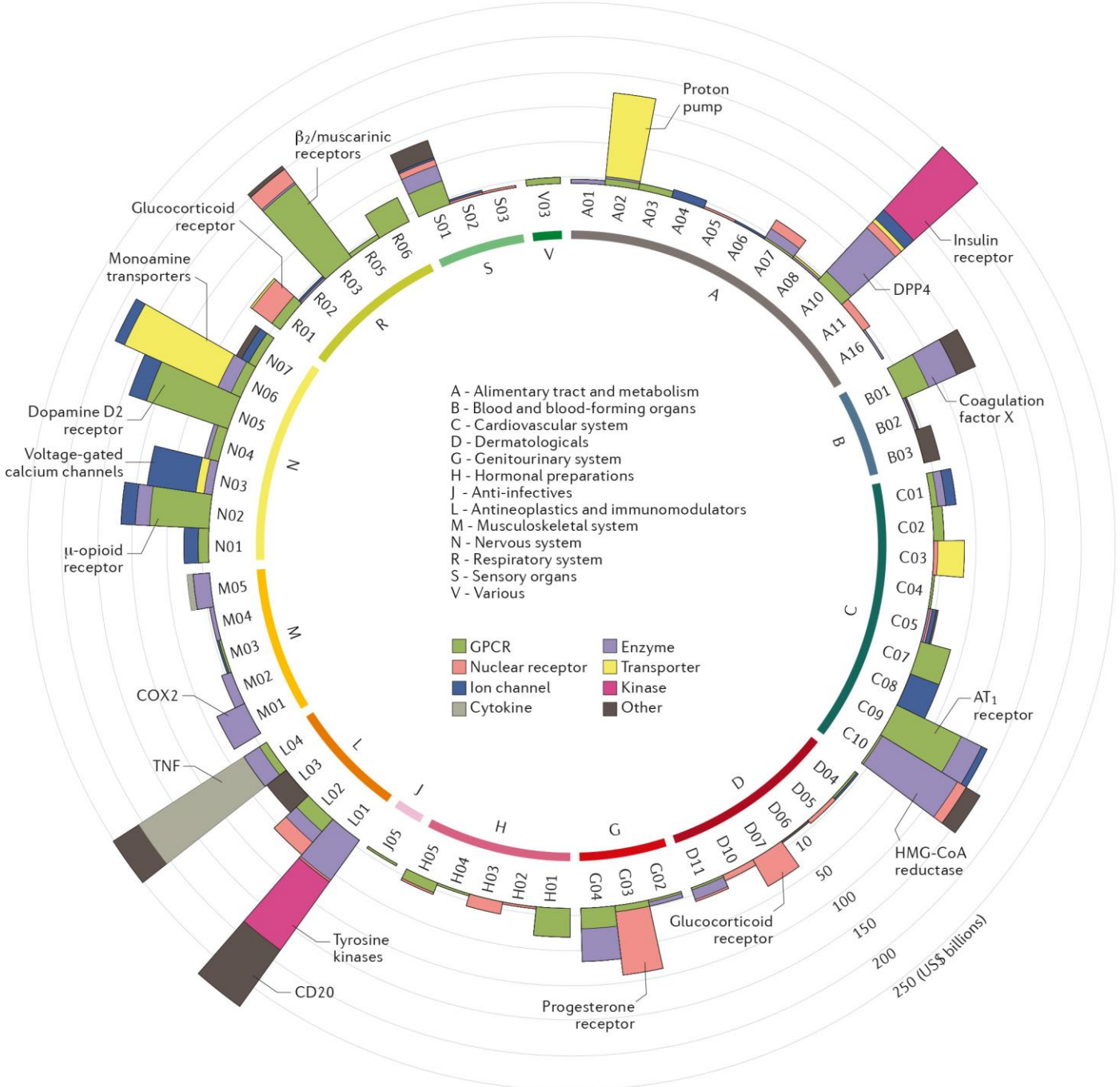
Drugs distributed by ATC codes (levels 1-2). Concentric rings indicate ATC levels. Histograms represent the number of drugs distributed per year of first approval. Maximum scale: 100.

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

10/20/16 revision

IDG KMC

FINANCIAL ACTIVITY PER THERAPEUTIC AREA



Human drug targets distributed by ATC codes (levels 1-2). Concentric rings indicate ATC levels. Histograms represent the global income derived from drugs that act on these targets via Mode of Action annotations. Maximum scale: 287 billion USD.

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

Nature Reviews | Drug Discovery

imshealth™

IDG KMC

NIH FUNDING: ROOM FOR IMPROVEMENT

- Text mining of all NIH grants (NIH RePORTER) for the 2000-2015 period suggests that 8858 proteis received zero funding.
- Of these, 6051 are Tdark, 2616 are Tbio. *(this is expected).*
- However, 119 are Tchem and 72 are Tclin. Possible explanations: old drug targets or funded elsewhere.
- *We are not able to track patterns of funded research in EU or other institutions. However, most of these proteins are likely to be "in the dark" given overall PubMed/Patent data*

- **Pharma and Academia could pay more attention to these 8858 underfunded proteins.**

Note: NIH is the only Funding Agency that makes available all funded research data in bulk (for text mining). Other agencies (e.g., ESF, IMI, even ERC) do not provide bulk access to funded research abstracts.

T. Oprea et al., *Nature Rev. Drug Discov,* **2017** (submitted)

5/26/17 revision

UNM SCHOOL of MEDICINE

TAKE HOME MESSAGE 2

THERE IS A KNOWLEDGE DEFICIT

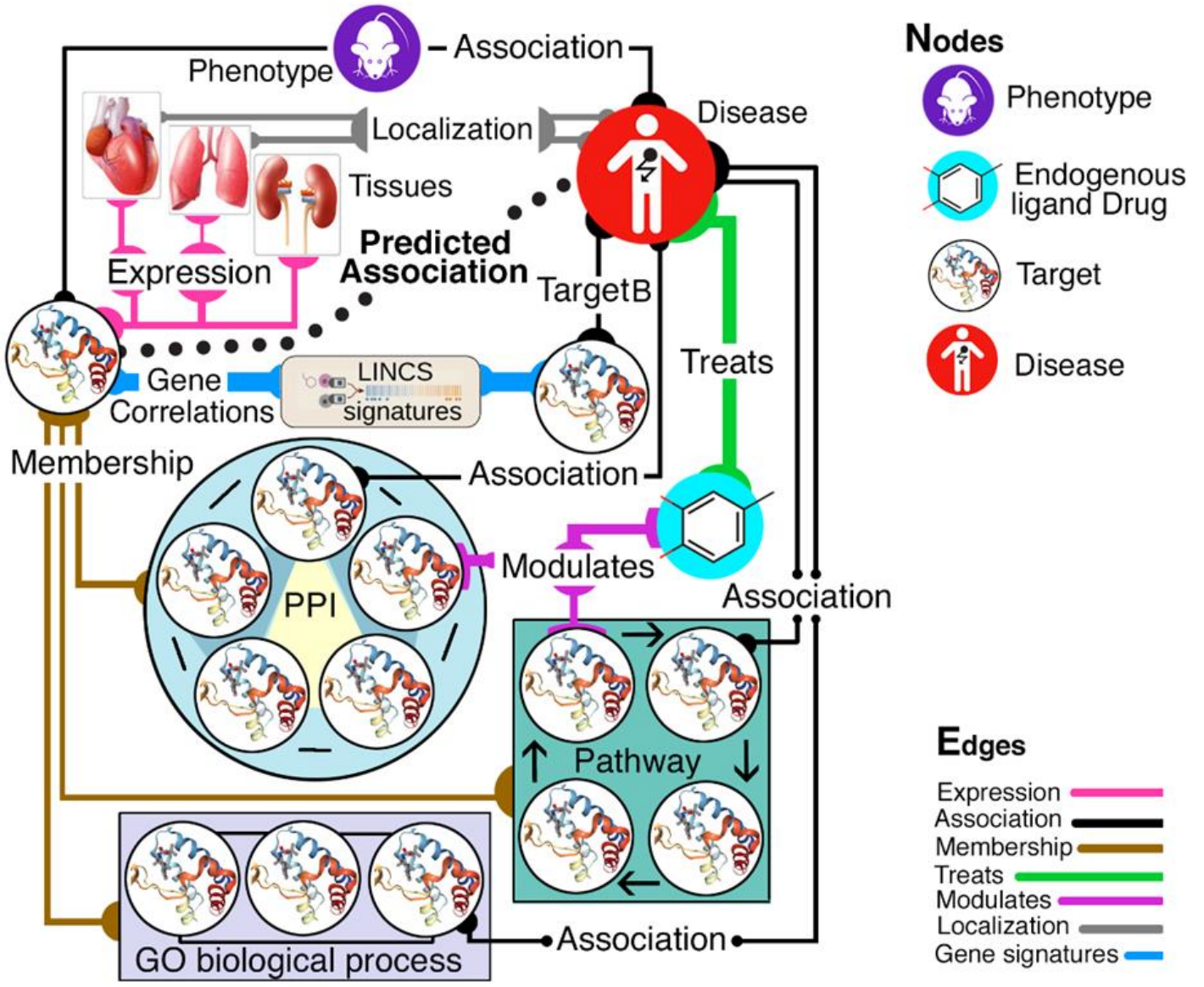
over 37% of the proteins remain
poorly described (Tdark)
~10% of the Proteome (Tclin & Tchem) can be targeted by small molecules

*We address ~15% of human diseases *) at most with therapeutic agents*

*) disease-ontology.org catalogs ~9,000 disease concepts. This resource lacks ~6,000 rare diseases. We estimate ~15,000 disease concepts; ~2500 have therapeutic agents

11/29/17 revision

PROTEIN KNOWLEDGE GRAPHS

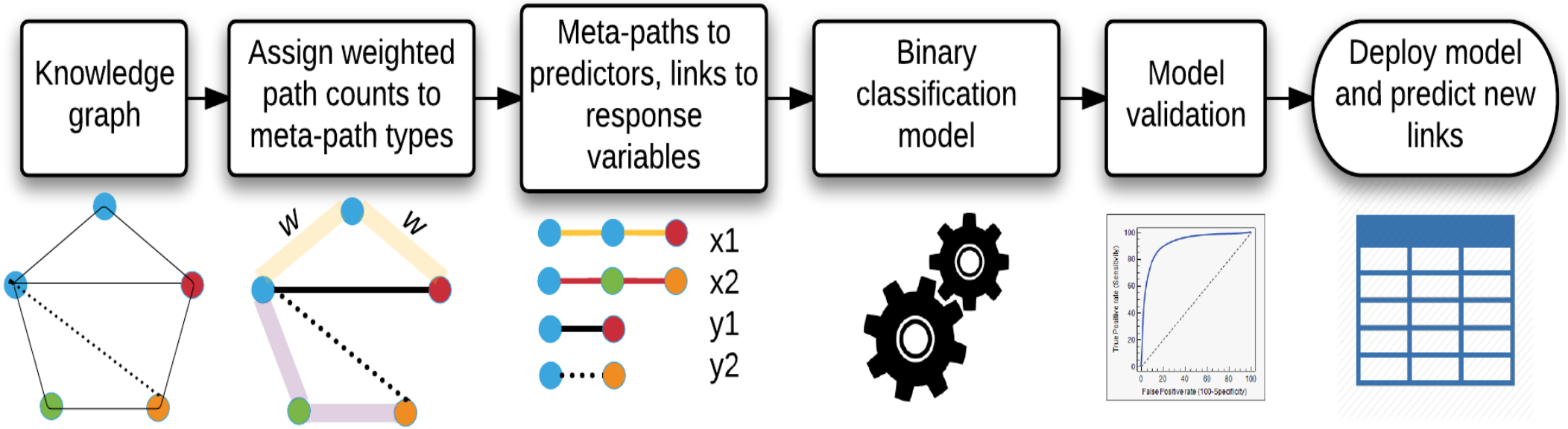


- IDG KMC2 seeks knowledge gaps across the five branches of the "knowledge tree":
- Genotype; Phenotype; Interactions & Pathways; Structure & Function; and Expression, respectively.
- We can use biological systems network modeling to infer novel relationships based on available evidence, and infer new "function" and "role in disease" data based on other layers of evidence
- Primary focus on **Tdark** & **Tbio**

T. Oprea et al., *IDG KMC Phase 2, anticipated start* **2018** (if funded)

11/29/17 revision

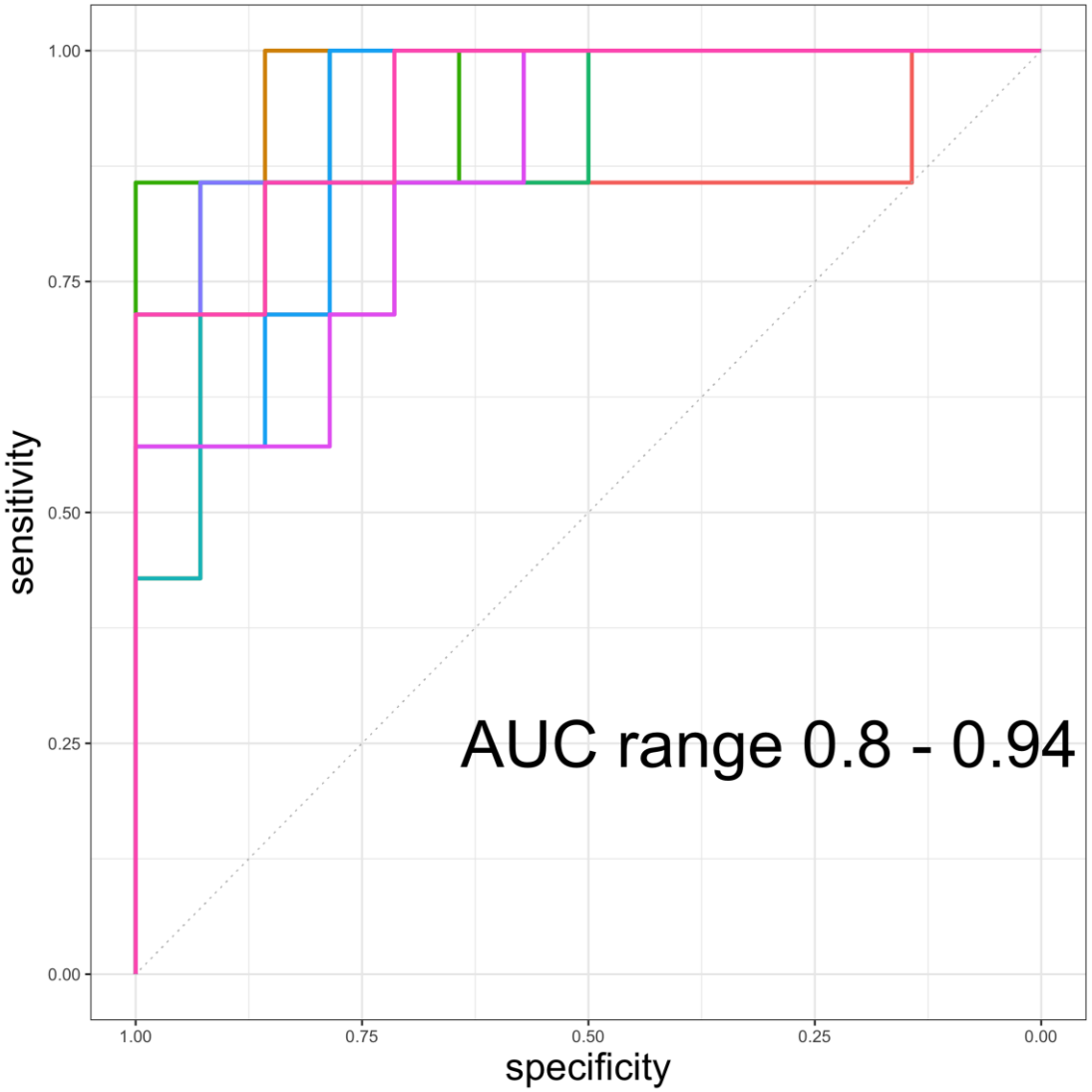
METAPATH ALGORITHM WORKFLOW



- A meta-path encodes type-specific network topology between the source node (e.g., Protein target) and the destination node (e.g., Disease or Function)
- *Target* — (member of) → PPI Network ← (member of) — Protein — (associated with) → *Disease*
- *Target* — (expressed in) → Tissue ← (localized in) — *Disease*

DILATED CARDIOMYOPATHY METAPATH MODEL

- Build data matrix from protein knowledge graph along metapaths:
  - Protein – Protein Interactions
  - Pathways
  - GO terms
  - Gene expression
  - …
- For the specific disease/phenotype "Dilated cardiomyopathy" in OMIM / TCRD
- 35 genes linked to this disease in TCRD
- Remaining genes assumed *not linked*
- Random forest binary classifier



T. Oprea et al., *IDG KMC Phase 2, anticipated start* **2018** (if funded)

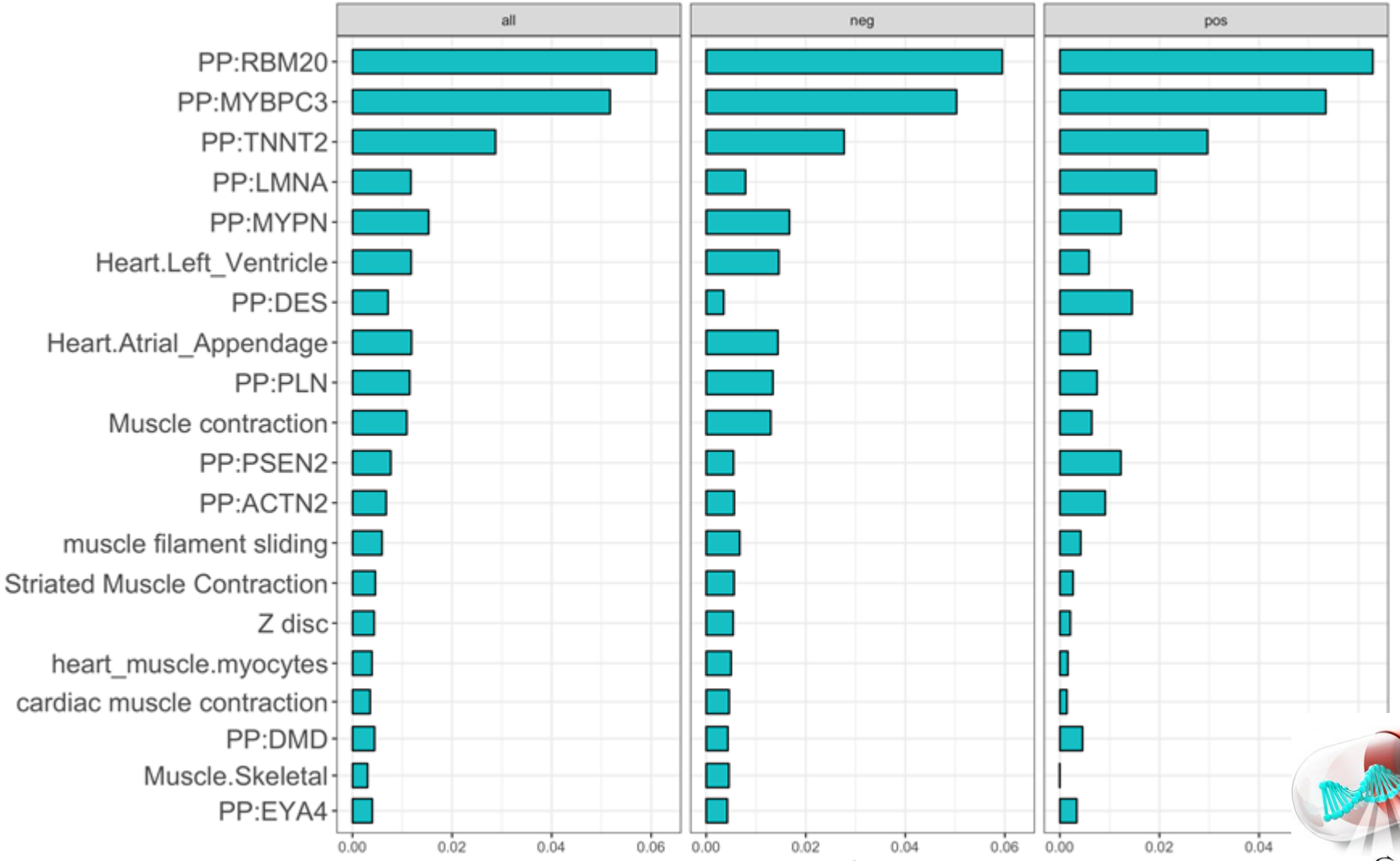
11/29/17 revision

RANDOM FOREST CLASSIFIER: VARIABLE IMPORTANCE PLOT

PP: protein – protein interaction with specified protein

VIMP higher for positive vs negative
Two are Tbio:
LMNA – *prelamin A/C*
DES - *desmin*
One is Tchem:
PSEN2 – *presenilin 2*
(for Alzheimer's)

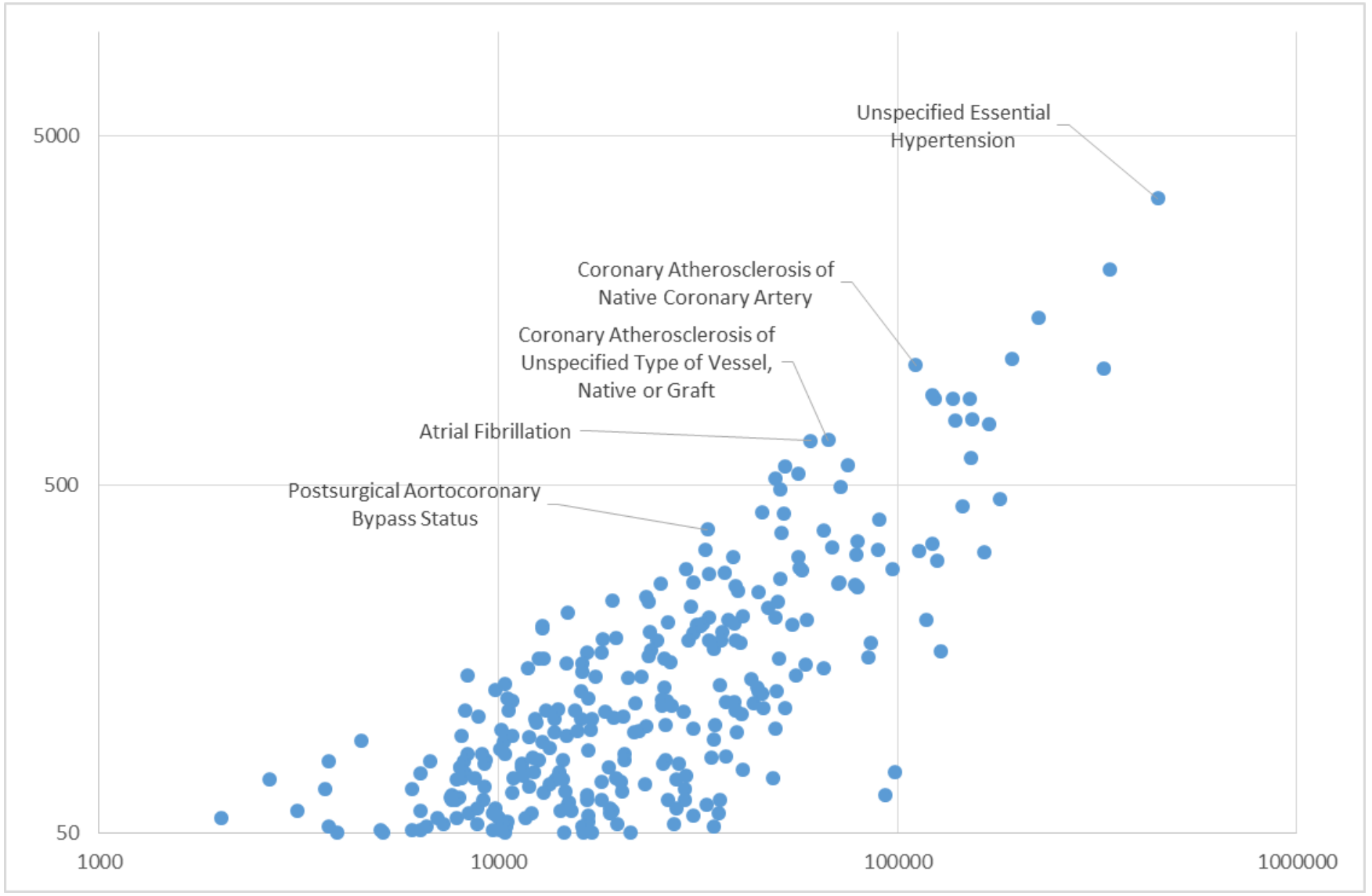
High VIMP overall:
RBM20 (Tbio)
MYBPC3 (Tbio)
TNNT2 (Tbio)



FALSE NEGATIVES IN VALIDATION SET DILATED CARDIOMYOPATHY

- Genes not annotated in OMIM with "Dilated cardiomyopathy" classified as positives:
  - NKX2-5 [https://www.ncbi.nlm.nih.gov/pubmed/25503402]
  - CACNA1C [https://www.ncbi.nlm.nih.gov/pubmed/22589547]
  - HSPB1 (HSP27) [https://www.ncbi.nlm.nih.gov/pubmed/17873025]
- These "false positives" are, in fact, involved in regulating muscle contraction
- *Given Presenilin 2 is a target for dilated cardiomyopathy, can we find a link between cardiovascular disease and Alzheimer's?*

DISEASE INTERCEPTION: 5 YEARS PRE-ALZHEIMER'S



Top Dx Prior to Alzheimer's (5 yrs or more):

Essential hypertension
Hyperlipidemia
Type 2 Diabetes mellitus
Hypercholesterolemia
Coronary atherosclerosis
Atrial Fibrillation

…

Persistent mental disorders
Memory Loss

http://rpubs.com/cbologa/alzheimers_disease

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

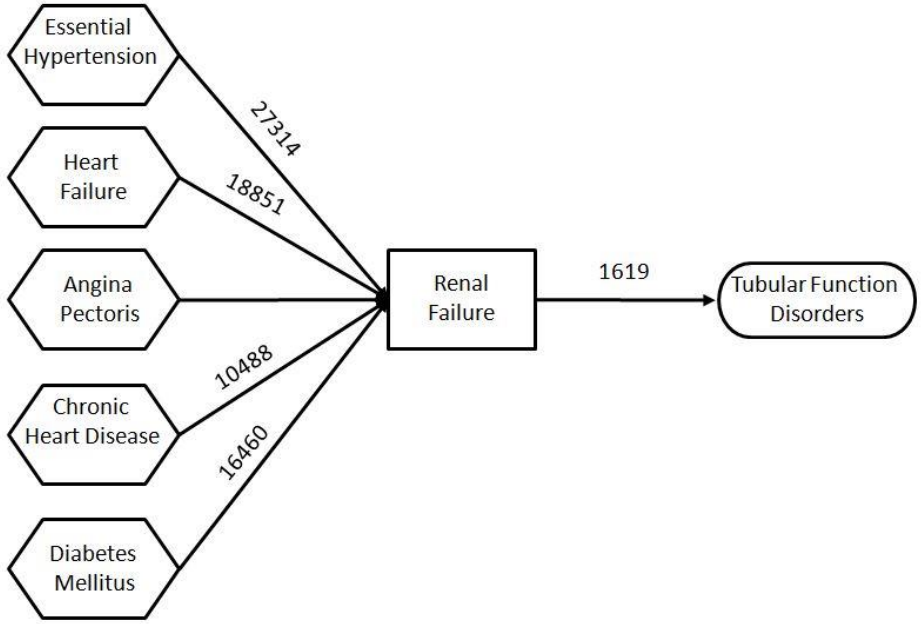
data from Cerner HealthFacts DB

11/29/17 revision

IDG KMC

DISEASE INTERCEPTION: 5 YEARS PRE-CKD

Our Cerner HealthFacts chronic kidney disease (CKD) data supports the Disease Trajectory work done on the population of Denmark (15 yrs)

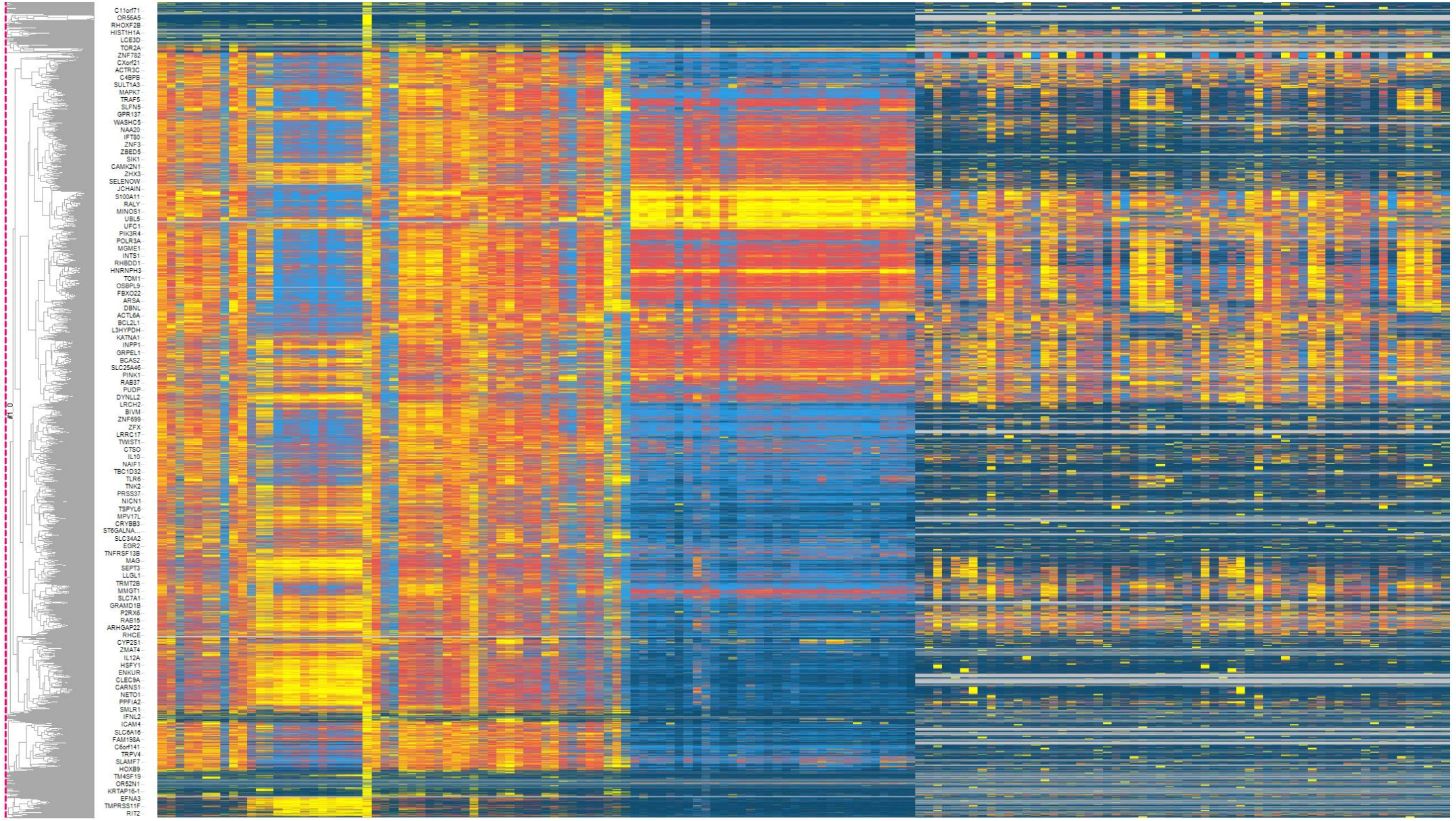


A.B. Jensen et al., *Nature Communications* **2014** 5:4022

C. Bologa, O. Ursu, J.J. Yang & T.I. Oprea (UNM) data from Cerner HealthFacts DB 10/20/16 revision/

UNM SCHOOL of MEDICINE

LARGE-SCALE EXPRESSION DATA

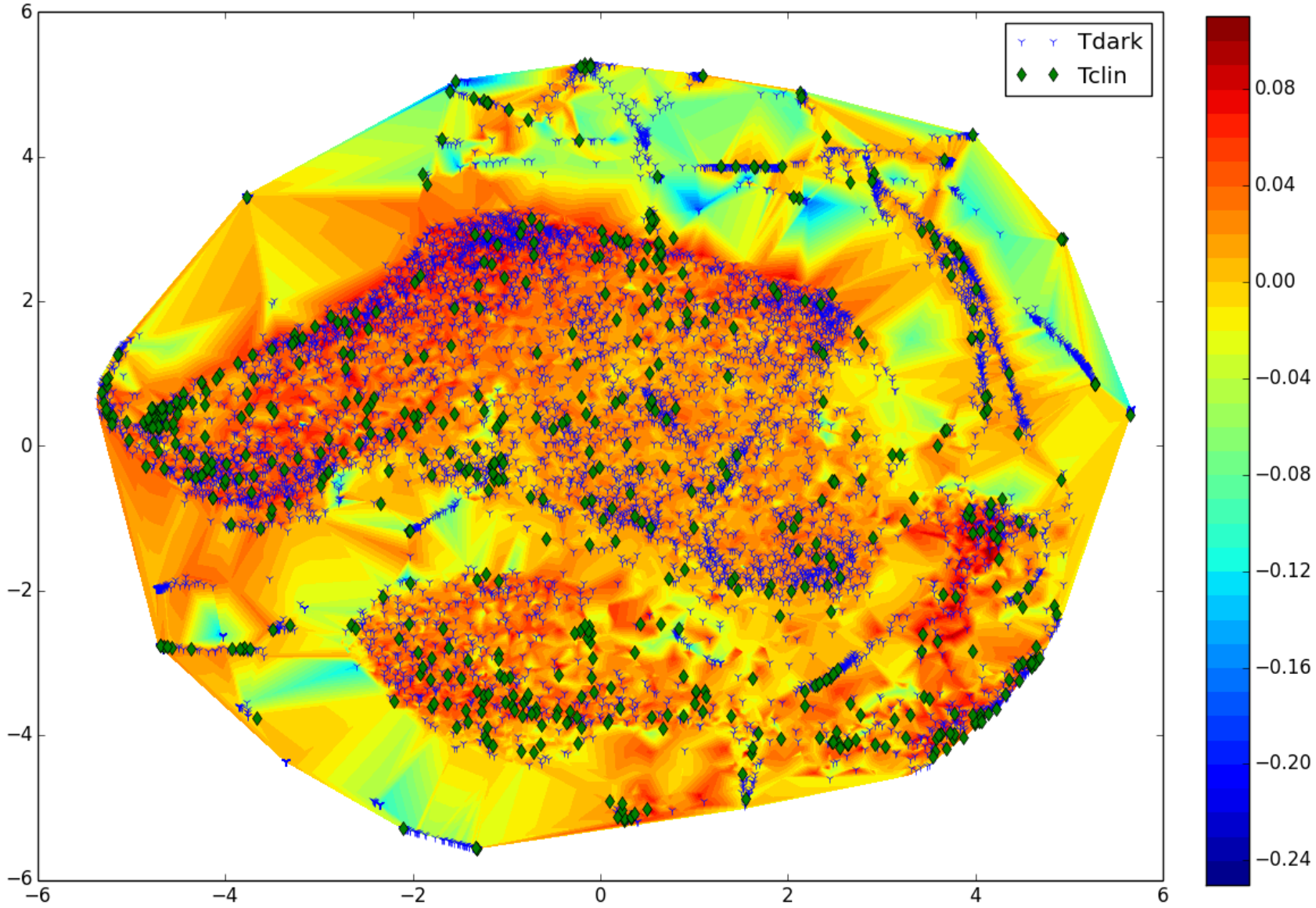


T. Oprea et al., *IDG KMC Phase 2, anticipated start* **2018** (if funded)

11/6/17 revision

IDG KMC

T-SNE DECISION FUNCTION FOR TDARK



1-class SVM model based on expression data

Tclin is "known" (everything else is "not known")

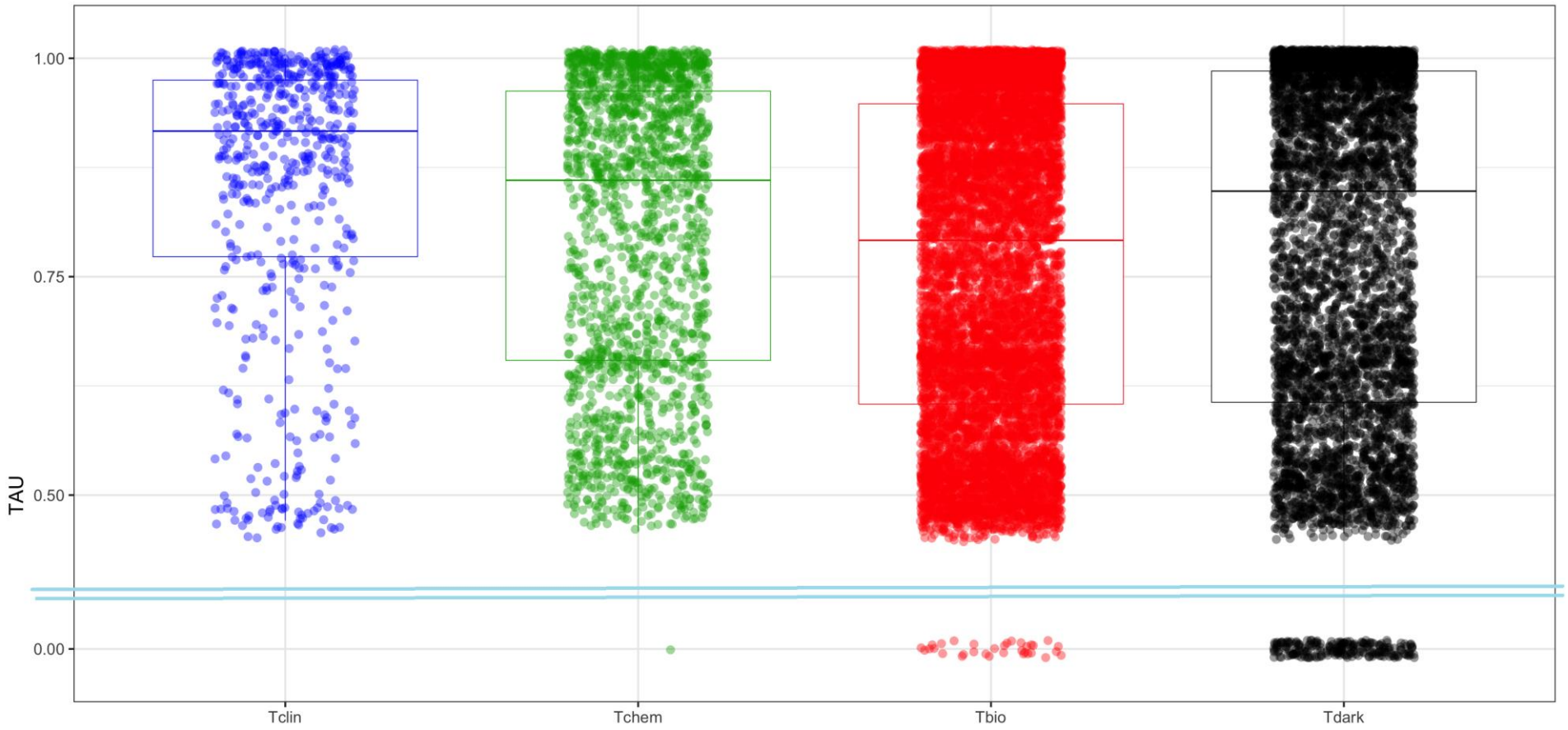
Goal: Identify Tdark proteins that share similar expression patterns to Tclin

PJ Kindermans, KR Müller et al.

2/10/16 revision

IDG KMC

IT'S NOT JUST TISSUE SPECIFICITY:



$$\tau = \frac{\sum_{i=1}^{N}(1 - x_i)}{N - 1}$$

- N is the number of tissues & x_i is the expression profile component, normalized by the maximal component value.

Data from GTEx v7

- $N = 53$
- All non-selective genes have $\tau = 0.43$;
- Selective genes, $\tau = 1$;
- No data, $\tau = 0$.

TSI, τ, provides enrichment for disease & tissue specific genes for drug targets (GSK paper from V. Kumar et al., *Sci Rep* **2016**, 6: 36205, PMC5099936)
Except… TSI does not discriminate well Tclin among TDL classes…

TSI formula from Yanai et al., Bioinformatics **2005,** 21:650-659 bti042

11/6/17 revision

IDG KMC

IMPC: HIGH THROUGHPUT PHENOTYPING

ARTICLE

doi:10.1038/nature19356

High-throughput discovery of novel developmental phenotypes

Mary E. Dickinson[1]*, Ann M. Flenniken[2,3]*, Xiao Ji[4]*, Lydia Teboul[5]*, Michael D. Wong[2,6]*, Jacqueline K. White[7], Terrence F. Meehan[8], Wolfgang J. Weninger[9], Henrik Westerberg[5], Hibret Adissu[2,10], Candice N. Baker[11], Lynette Bower[12], Ann-Marie Mallon[5], R. Mark Henkelman[2,6], Steve D. M. Brown[5], David J. Adams[7], K. C. Kent Lloyd[12], Colin McKerlie[2,10], Arthur L. Beaudet[17], Maja Bućan[26] & Stephen A. Murray[11]

Approximately one-third of all mammalian genes are essential for life. Phenotypes resulting from knockouts of these genes in mice have provided tremendous insight into gene function and congenital disorders. As part of the International Mouse Phenotyping Consortium effort to generate and phenotypically characterize 5,000 knockout mouse lines, here we identify 410 lethal genes during the production of the first 1,751 unique gene knockouts. Using a standardized phenotyping platform that incorporates high-resolution 3D imaging, we identify phenotypes at multiple time points for previously uncharacterized genes and additional phenotypes for genes with previously reported mutant phenotypes. Unexpectedly, our analysis reveals that incomplete penetrance and variable expressivity are common even on a defined genetic background. In addition, we show that human disease genes are enriched for essential genes, thus providing a dataset that facilitates the prioritization and validation of mutations identified in clinical sequencing efforts.

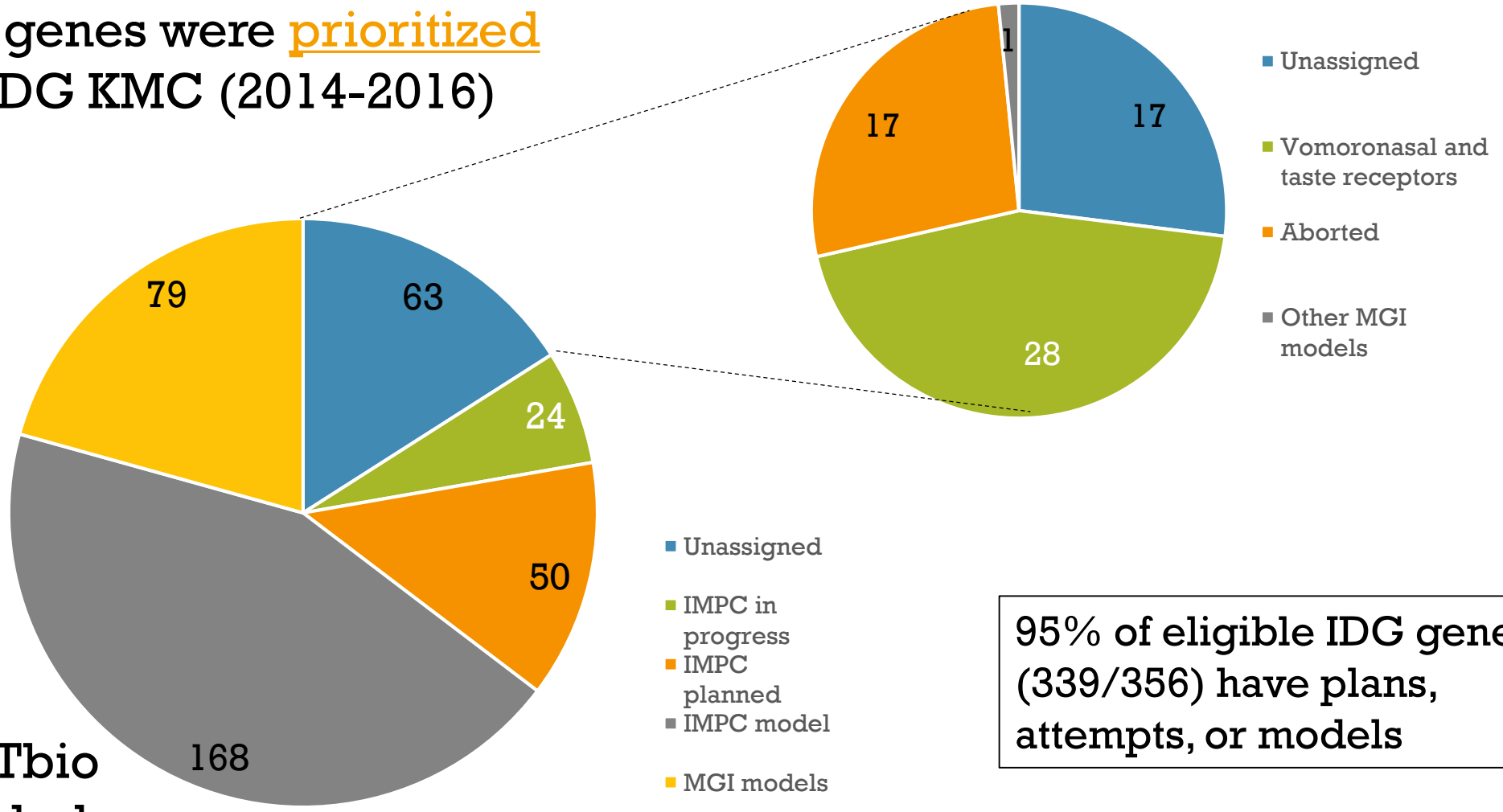
IMPC is currently composed of 18 research institutions and 5 national funders

M.E. Dickinson et al., *Nature* **2016**, 537:508-514

IMPC

IMPC USES IDG KMC TARGET PRIORITIZATION

384 genes were prioritized by IDG KMC (2014-2016)



95% of eligible IDG genes (339/356) have plans, attempts, or models

306 genes are Tbio
90 genes are Tdark
42 genes are Tchem

Slide from Steve Murray, Jackson Lab

11/29/17 revision

IMPC

TAKE HOME MESSAGE 3

THE TRUTH IS ALREADY HERE, IT'S JUST NOT EVENLY DISTRIBUTED

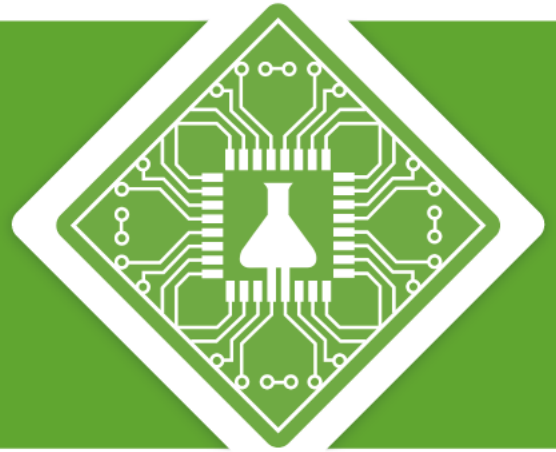
the age of informatics-driven
drug/target discovery is here

Accurate Information is the Foundation of *in silico* Methods

PHARMA.AI

MEDICINAL CHEMISTRY REIMAGINED

Insilico Medicine, Inc
Emerging Technology Centers
Johns Hopkins University
B301, 1101 33rd Street
Baltimore, MD, 21218



INSILICO MEDICINE

- **Drug Discovery**
- **Drug Repurposing**
- **Pathway Activation Analysis**
- **Biomarker Development**
- **Clin. Trials Predictors**
- **Aging Research**

Project conceived by
ALEX ZHAVORONKOV, PHD
alex@insilico.com

www.insilico.com

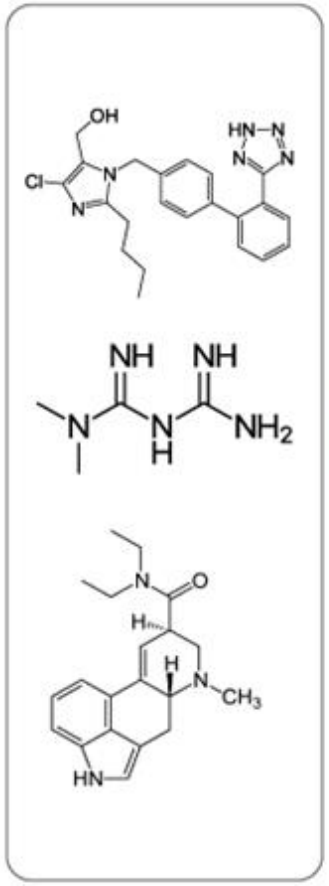
11/29/17 revision

CHEMISTRY ASSOCIATED EMOTIONAL RESPONSES: BUILDING PHARMA.AI

APPROVED DRUGS

MOLECULES FAILED IN CLINICAL TRIALS

MIX OF MOLECULES WITH MULTIPLE NUMERICAL PROPERTIES DISPLAYED TO THE CHEMIST



EMOTIONAL RESPONSE

LIKE

DISLIKE

11/29/17 revision

IDG KMC

PHARMA.AI – WHEN WILL IT PASS THE TURING TEST?

TARGET SELECTION IS PRECOMPETITIVE KNOWLEDGE

- The drug industry reward system is based on patents
- These are awarded for **drugs**, not targets
- Clinically validated targets lead to the me-too phenomenon
- Time to pool resources together on Target Selection (which includes target identification, target validation and all –omic research)

- Industry should lead a Target Selection Consortium: Partner with academia and co-sponsor "double blind" target finding (avoid the reproducibility crisis)
- @John Baldoni / GSK: we'd like to team up with OpenTargets & ATOM

THE UNREALIZED PROMISE: AI FOR DD

- We somehow expect that in the near future, artificial intelligence (AI; machine learning; *singularity*) will be able to process all data related to biomolecules, their biologic endpoints, ex vivo, in vivo and clinical data, and discover drugs
- The problem with that: GIGO
- Garbage in, Garbage out.

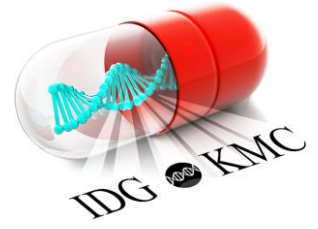
https://www.linkedin.com/pulse/why-ai-ready-drug-discovery-tudor-oprea

5/20/17 revision

IDG KMC

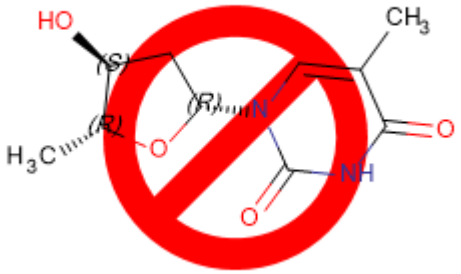
PEOPLE ARE PART OF THE PROBLEM

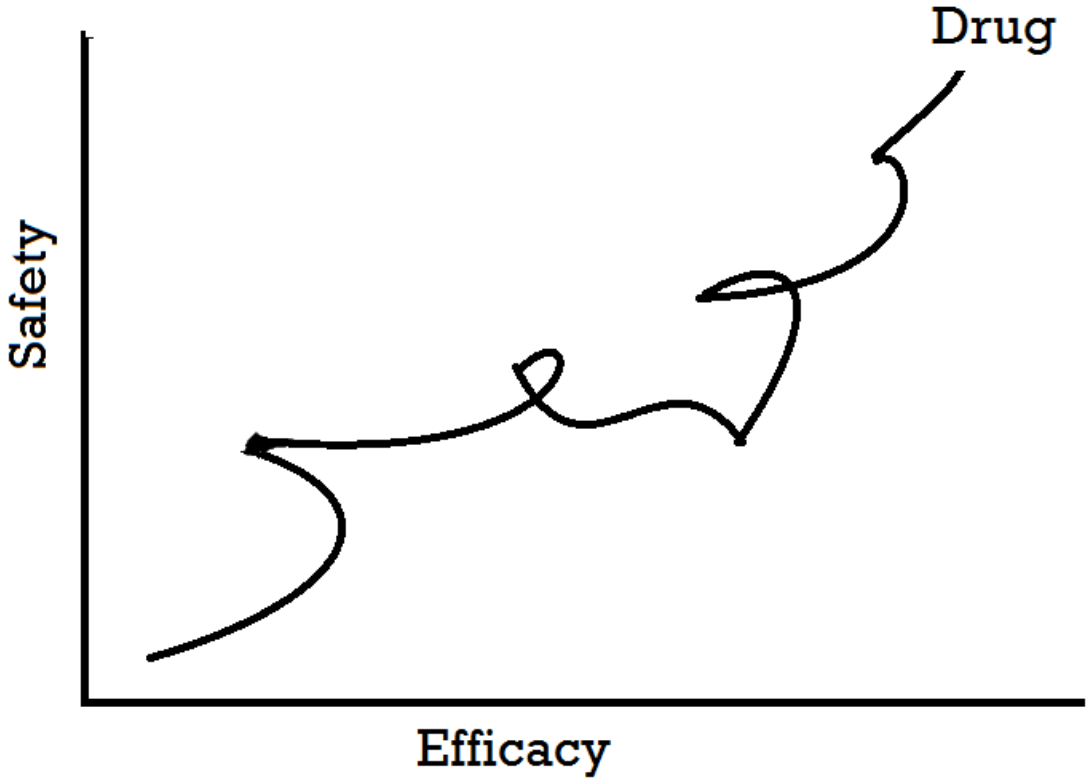
- An epistemological question: *Can AI discover new knowledge?* To date, no credible evidence of this has been provided. Chatbots, winning at chess, GO and *Jeopardy!* does not count.
- We live in the world of *alternative facts* when it comes to research (not just politics). *People lie.* See work by JP Ioannidis, but also notes from Bayer (Asadullah et al) and Amgen (Begley et al). As long as AI gets false data, we cannot provide what's needed.
- As of now, patents cannot name AI as inventors. Patent offices world-wide grant patents to *people*. What happens when AI does become the inventor? Will this happen?!
- Will it matter?



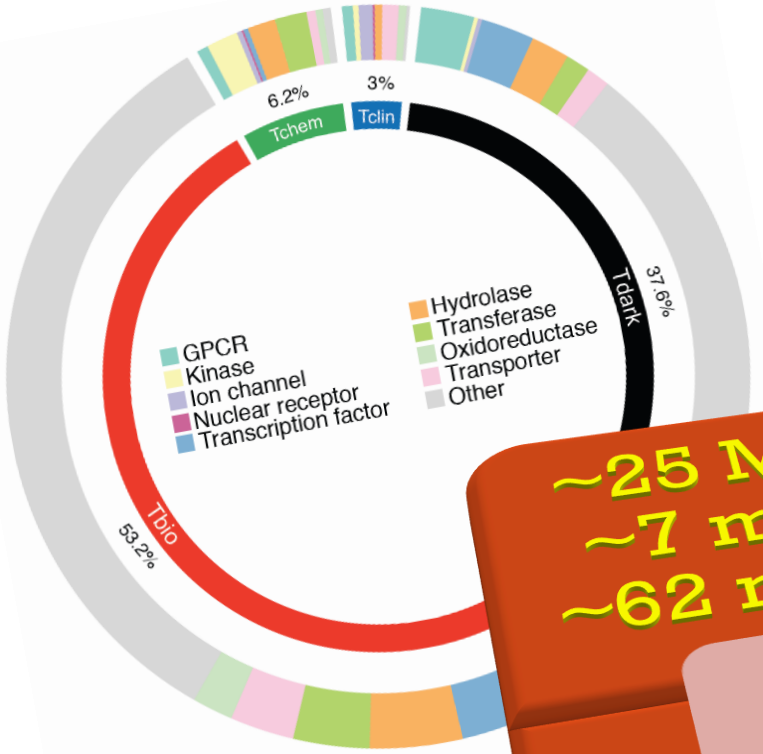
BUT THE SCIENTIFIC ISSUE IS DEEPER

- The unrealized promise relates to our (in)ability to explain the two pillars of clinical drug effectiveness: Safety and Efficacy.
- In short, be it Drug Safety or Patient Safety, or indeed Clinical Efficacy, *we remain unable to model these processes as function of molecular structure.*





Illuminating the Druggable Genome Knowledge Management Center



~25 Million Papers
~7 million Patents
~62 million Patients

~20,000 Proteins

Seeking New Knowledge

~15,000 Diseases

~4,500 Drugs

IDG KMC