Artificial Intelligence and Machine Learning in Translational Science Innovation

Christopher P. Austin, M.D.
Director, NCATS

The Ohio State University Center for Clinical and Translational Science
2019 Annual Scientific Meeting

December 3, 2019

All is right with the world, so we can discuss more trivial matters like AI/ML

Ohio State vs Michigan

NCAA football

NCAA football · Yesterday

Final

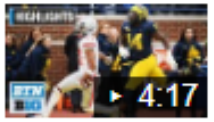
2 Ohio State Buckeyes (12 - 0)

56 - 27

17 Michigan Wolverines (9 - 3)

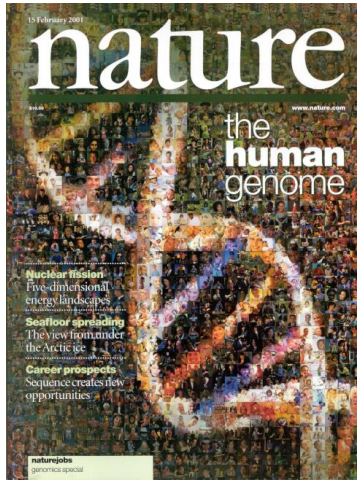
Team	1	2	3	4	T
Ohio State Buckeyes	14	14	14	14	56
Michigan Wolverines	13	3	3	8	27

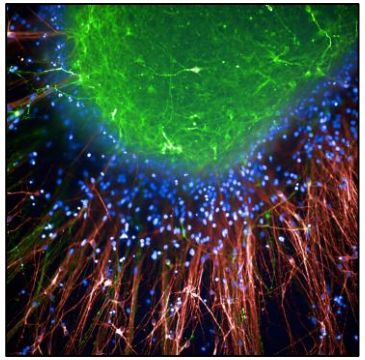
Game recap



The Best of Times, the Worst of Times

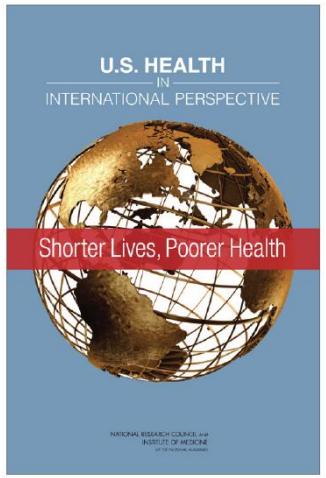
Fundamental science has seen unprecedented advances, but treatments have not





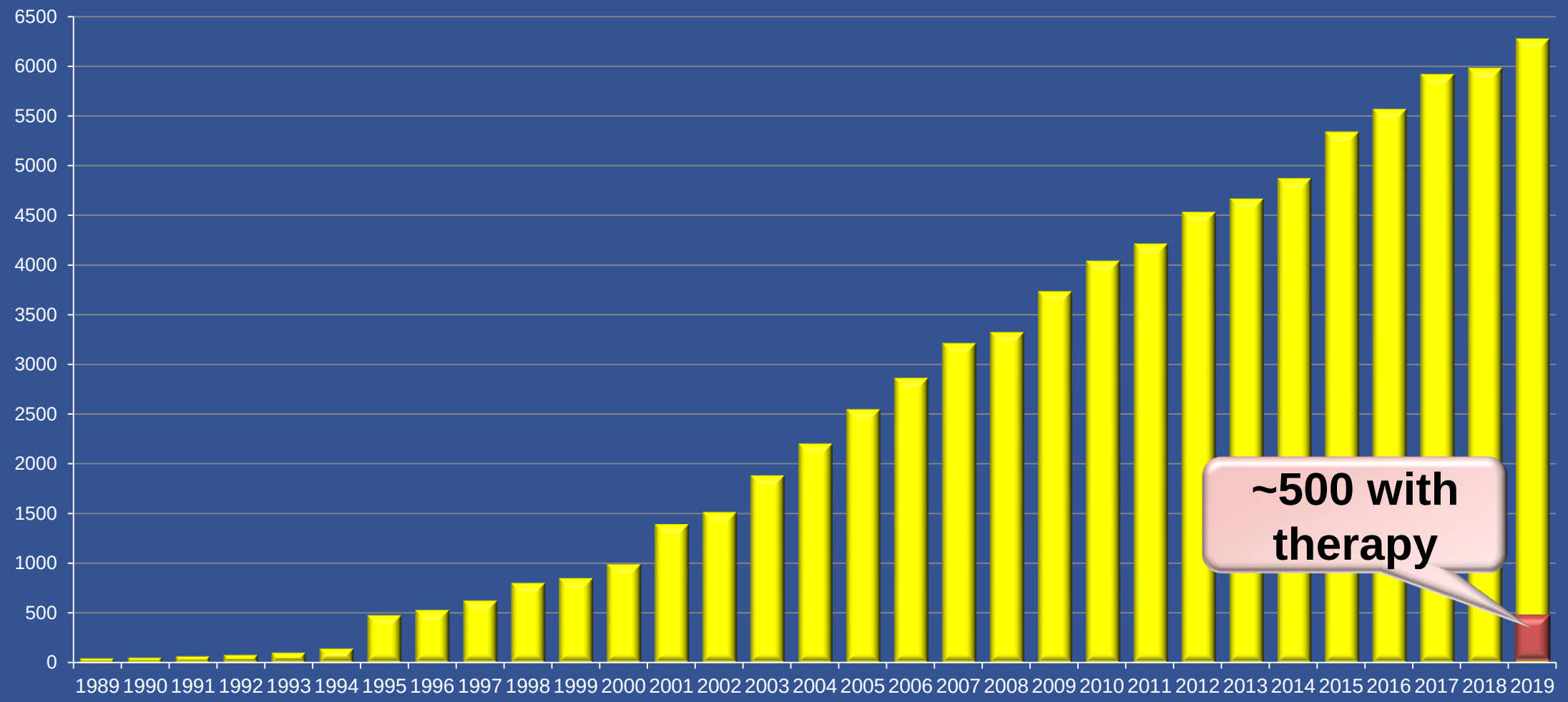
- Poor transition of basic or clinical observations into interventions that tangibly improve human health
- Intervention development failure-prone, inefficient and costly
- Poor adoption of demonstrably useful interventions

Enormous opportunity/need to deliver on the promise of science for patients



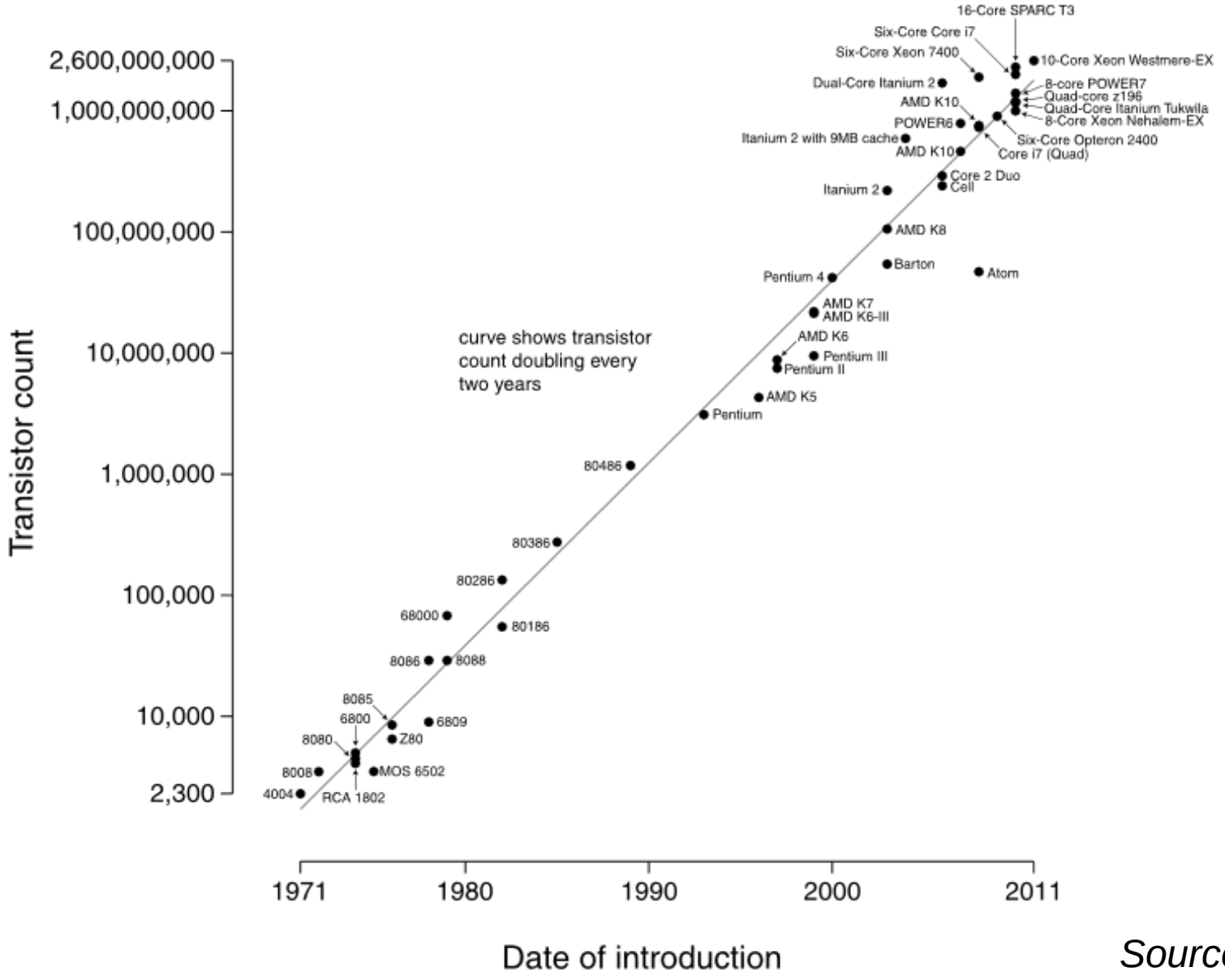
National Center for Advancing Translational Sciences

Disorders with Known Molecular Basis



Source: Online *Mendelian Inheritance in Man*, Morbid Anatomy of the Human Genome

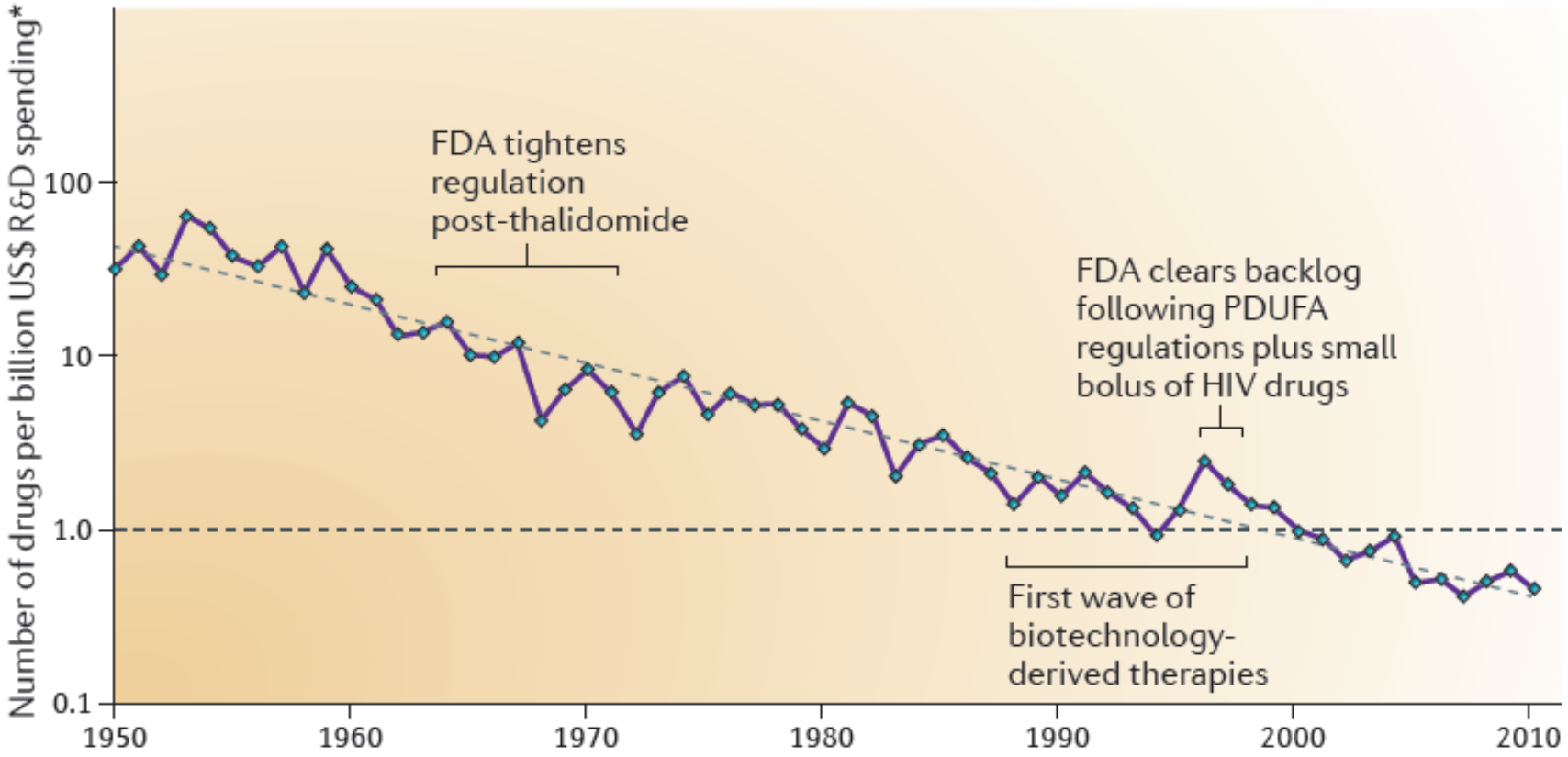
Moore's Law



Source: Wikipedia

National Center for Advancing Translational Sciences

Eroom's Law

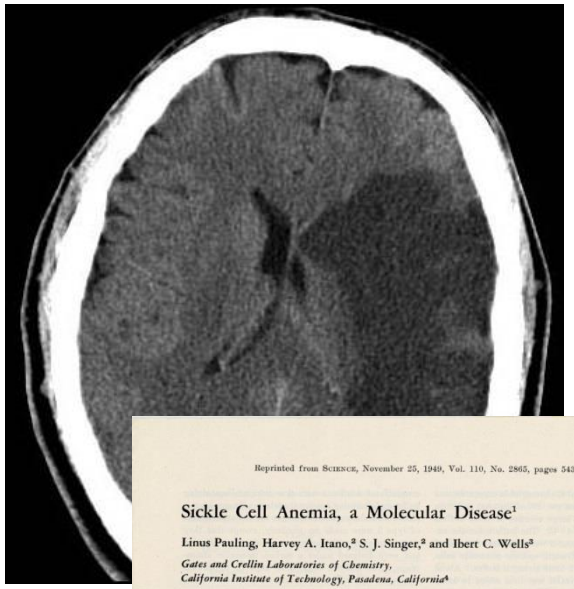


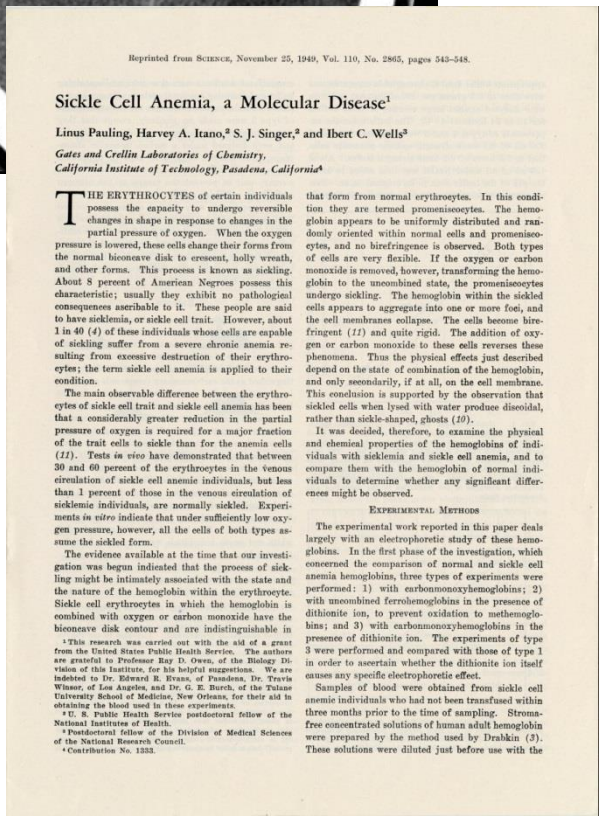
The number of new drugs approved by the FDA per billion US dollars (inflation-adjusted) spent on research and development (R&D) has halved roughly every 9 years since 1950.

Scannell et al., *Nature Reviews Drug Discovery* 11:191, 2012

National Center for Advancing Translational Sciences

So…





Reprinted from SCIENCE, November 25, 1949, Vol. 110, No. 2865, pages 543–548.

Sickle Cell Anemia, a Molecular Disease[1]

Linus Pauling, Harvey A. Itano,[2] S. J. Singer,[2] and Ibert C. Wells[3]

Gates and Crellin Laboratories of Chemistry, California Institute of Technology, Pasadena, California[4]

THE ERYTHROCYTES of certain individuals possess the capacity to undergo reversible changes in shape in response to changes in the partial pressure of oxygen. When the oxygen pressure is lowered, these cells change their forms from the normal biconcave disk to crescent, holly wreath, and other forms. This process is known as sickling. About 8 percent of American Negroes possess this characteristic; usually they exhibit no pathological consequences ascribable to it. These people are said to have sicklemia, or sickle cell trait. However, about 1 in 40 (*4*) of these individuals whose cells are capable of sickling suffer from a severe chronic anemia resulting from excessive destruction of their erythrocytes; the term sickle cell anemia is applied to their condition.

The main observable difference between the erythrocytes of sickle cell trait and sickle cell anemia has been that a considerably greater reduction in the partial pressure of oxygen is required for a major fraction of the trait cells to sickle than for the anemia cells (*11*). Tests *in vivo* have demonstrated that between 30 and 60 percent of the erythrocytes in the venous circulation of sickle cell anemic individuals, but less than 1 percent of those in the venous circulation of sicklemic individuals, are normally sickled. Experiments *in vitro* indicate that under sufficiently low oxygen pressure, however, all the cells of both types assume the sickled form.

The evidence available at the time that our investigation was begun indicated that the process of sickling might be intimately associated with the state and the nature of the hemoglobin within the erythrocyte. Sickle cell erythrocytes in which the hemoglobin is combined with oxygen or carbon monoxide have the biconcave disk contour and are indistinguishable in that form from normal erythrocytes. In this condition they are termed promenisocytes. The hemoglobin appears to be uniformly distributed and randomly oriented within normal cells and promenisocytes, and no birefringence is observed. Both types of cells are very flexible. If the oxygen or carbon monoxide is removed, however, transforming the hemoglobin to the uncombined state, the promenisocytes undergo sickling. The hemoglobin within the sickled cells appears to aggregate into one or more foci, and the cell membranes collapse. The cells become birefringent (*11*) and quite rigid. The addition of oxygen or carbon monoxide to these cells reverses these phenomena. Thus the physical effects just described depend on the state of combination of the hemoglobin, and only secondarily, if at all, on the cell membrane. This conclusion is supported by the observation that sickled cells when lysed with water produce discoidal, rather than sickle-shaped, ghosts (*10*).

It was decided, therefore, to examine the physical and chemical properties of the hemoglobins of individuals with sicklemia and sickle cell anemia, and to compare them with the hemoglobin of normal individuals to determine whether any significant differences might be observed.

EXPERIMENTAL METHODS

The experimental work reported in this paper deals largely with an electrophoretic study of these hemoglobins. In the first phase of the investigation, which concerned the comparison of normal and sickle cell anemia hemoglobins, three types of experiments were performed: 1) with carbonmonoxyhemoglobins; 2) with uncombined ferrohemoglobins in the presence of dithionite ion, to prevent oxidation to methemoglobins; and 3) with carbonmonoxyhemoglobins in the presence of dithionite ion. The experiments of type 3 were performed and compared with those of type 1 in order to ascertain whether the dithionite ion itself causes any specific electrophoretic effect.

Samples of blood were obtained from sickle cell anemic individuals who had not been transfused within three months prior to the time of sampling. Stroma-free concentrated solutions of human adult hemoglobin were prepared by the method used by Drabkin (*3*). These solutions were diluted just before use with the

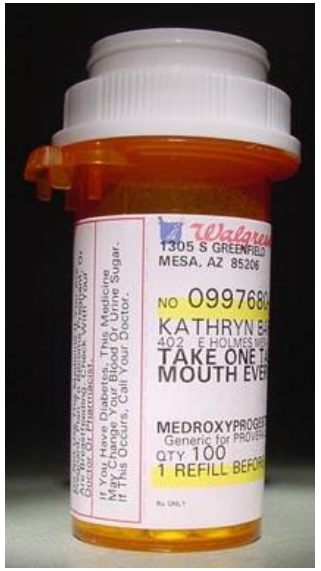
[1] This research was carried out with the aid of a grant from the United States Public Health Service. The authors are grateful to Professor Ray D. Owen, of the Biology Division of this Institute, for his helpful suggestions. We are indebted to Dr. Edward E. Evans, of Pasadena, Dr. Travis Winsor, of Los Angeles, and Dr. G. E. Burch, of the Tulane University School of Medicine, New Orleans, for their aid in obtaining the blood used in these experiments.

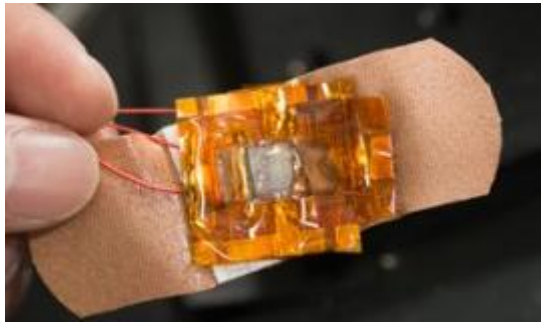
[2] U. S. Public Health Service postdoctoral fellow of the National Institutes of Health.

[3] Postdoctoral fellow of the Division of Medical Sciences of the National Research Council.

[4] Contribution No. 1333.

≠

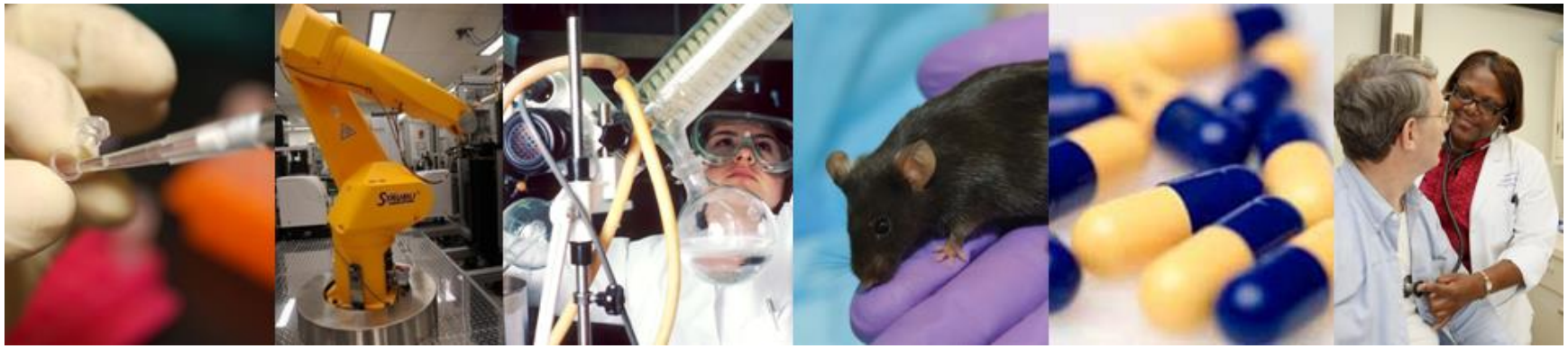




≠

How can AI/ML improve these efficiencies?

NCATS Mission



To catalyze the generation of innovative methods and technologies that will enhance the development, testing and implementation of diagnostics and therapeutics across human diseases and conditions.

What is Translation?

Translation is the process of turning observations in the laboratory, clinic, and community into interventions that improve the health of individuals and the public —from diagnostics and therapeutics to medical procedures and behavioral changes.

Translational Research endeavors to traverse a particular step of translation for a particular target or disease.

What is Translational Science?

Translational Science is the field of investigation focused on understanding the scientific and operational principles underlying each step of the translational process.

Major rate-limiting translational problems that are the focus of NCATS

- Understanding of translation
- Translational Science as a new academic discipline
- Predictive toxicology
- Predictive efficacy
- De-risking undruggable targets/untreatable diseases
- Data interoperability
- Biomarker qualification process
- Clinical trial networks
- Patient recruitment
- Electronic Health Records for research
- Harmonized IRBs
- Clinical diagnostic criteria
- Clinical outcome criteria (e.g., PROs)
- Adaptive clinical trial designs
- Shortening time of intervention adoption
- Adherence
- Methods to better measure impact on health

National Center for Advancing Translational Sciences

Major rate-limiting translational problems that are the focus of NCATS

- Understanding of translation
- Translational Science as a new academic discipline
- Predictive toxicology
- Predictive efficacy
- De-risking undruggable targets/untreatable diseases
- Data interoperability
- Biomarker qualification process
- Clinical trial networks
- Patient recruitment
- Electronic Health Records for research
- Harmonized IRBs
- Clinical diagnostic criteria
- Clinical outcome criteria (e.g., PROs)
- Adaptive clinical trial designs
- Shortening time of intervention adoption
- Adherence
- Methods to better measure impact on health

National Center for Advancing Translational Sciences

AOURP Progress

All of Us
RESEARCH PROGRAM

Eric Dishman
Director, *All of Us* Research Program
eric.dishman@nih.gov
AllofUs.nih.gov, JoinAllofUs.org, Researchallofus.org

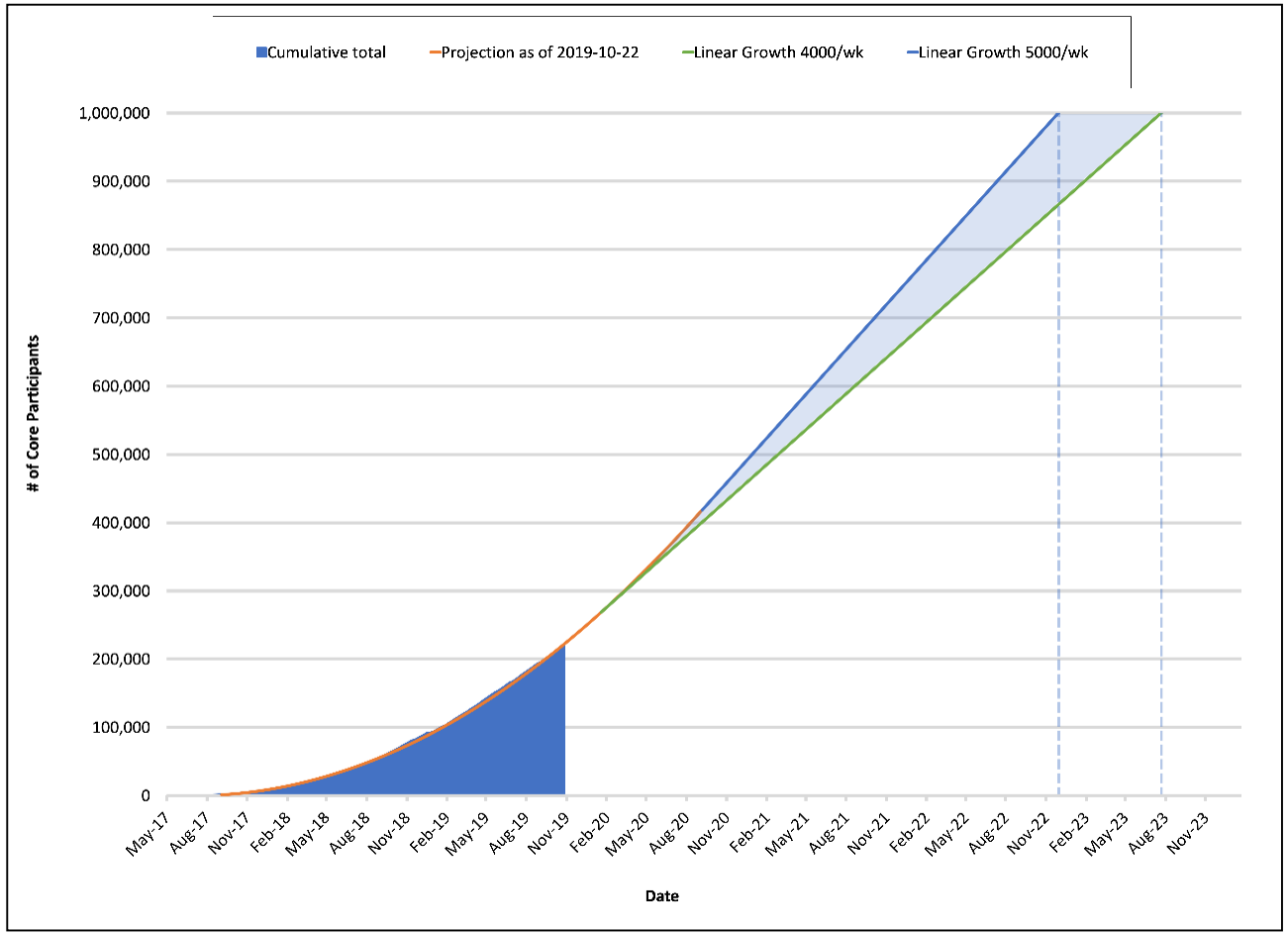
NIH
National Institutes of Health

#JoinAllofUs

Recruitment & Enrollment Progress (as of 11/6/19)

- **344k+ individuals** have started the enrollment process in some way
- **222k+ participants** have completed the initial core protocol ("core participants")
- About **3400 core participants per week** now
- **370+ enrollment sites** around the country
  - 115+ of them opened in 2019
  - ~100 more planned sites in 2020 + DV launch
- **2 mobile exhibits** traveling the country, especially to underserved areas
  - 600 days on the road since their launch
  - Engaged 55k+ people
- Very powerful network of **50+ community partners** building awareness & trust

Past and projected growth of cohort



Anticipate ramping up to an enrollment rate of 4,000+ participants/week in 2020 and expect to reach 1M total participants some time in 2023

Vision for Genomics and Return of Genomic Results in *All of Us*

Goal: Create the world's largest and most comprehensive precision medicine research platform, including genotypes and whole genome sequences on 1 million or more core participants, through a strategy that balances the need for responsible return of genomic information to participants with the scientific need for highest quality genomic results to advance precision medicine.

Major steps to reach this goal:

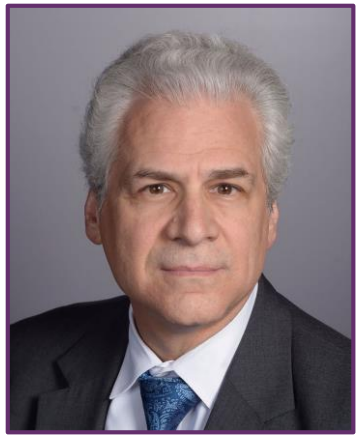
- **Short-term (2018-2019)**:
  - Develop a scalable, feasible roadmap for genomic data production
  - Deploy and integrate Genome Centers and the Genetic Counseling Resource

- **Mid-term (2020-2021):**
  - Introduce responsible return of value to participants, including non-medical results (ancestry, simple traits)
  - Pilot responsible return of medically actionable genomic results (ACMG pathogenic; potentially pharmacogenomics)
  - Evaluate performance of Genomics Platform and improve

- **Long-term (2021 and beyond):**
  - Introduce return of additional medically-actionable results (*e.g.*, polygenic risk score)
  - Shift to new technologies that enable new science (long read, and beyond)

NCATS Division of Clinical Innovation

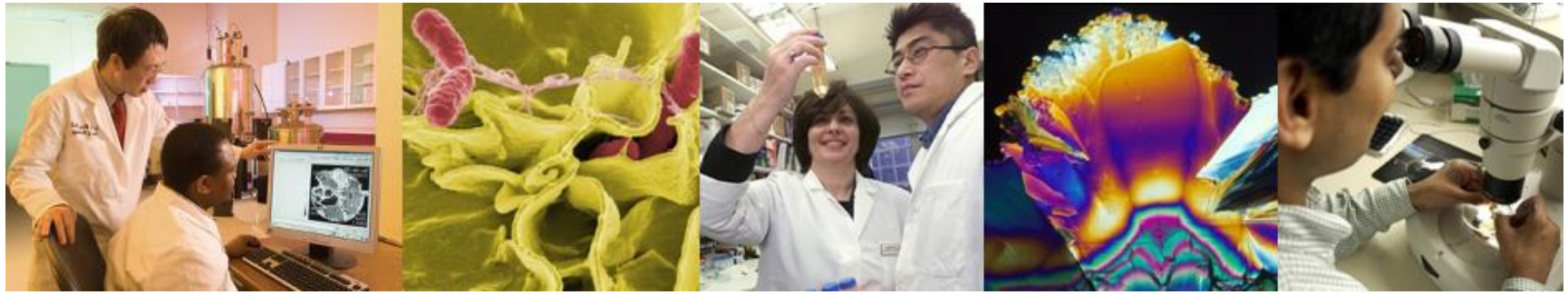
Clinical and Translational Science Awards (CTSA) Program



Mike Kurilla, M.D., Ph.D.
Director
michael.kurilla@nih.gov

- Collaboratively facilitates and accelerates translational projects locally/regionally/nationally
- Scientific and operational innovation to improve the efficiency and effectiveness of clinical translational research
- Creates, provides, and dissmeminates domain-specific translational science training
- Fosters creation of an academic discipline of translational science

Rebecca D. Jackson, M.D.
Director
OSU Center for Clinical and Translational Science



National Center for Advancing Translational Sciences

The NCATS Trial Innovation Network

TRIAL INNOVATION NETWORK

CTSA Clinical & Translational Science Awards

metrics

NETWORK PROPOSAL SUBMISSIONS		SIRB (more metrics)		STANDARD AGREEMENT (in FDP-CTSA)	
# of institutions	# of therapeutic areas	# of relying sites	# of studies	# of sites	# of studies
67	51	183	50	73	9

Search

INVESTIGATORS

submit your proposal

Hear from us within 5 business days.

TRIAL INNOVATION NETWORK

Operational innovation, excellence, and collaboration.

The Trial Innovation Network continues to accept new proposals!
Click the button below to get started.

Get Started now!

223 total proposals submitted

WELCOME!

The Trial Innovation Network is a collaborative national network that focuses on operational **innovation, excellence and collaboration** and will leverage the expertise and resources of the CTSA Program.

NETWORK LOGIN

National Center for Advancing Translational Sciences

The NCATS Accrual to Clinical Trials (ACT) Network

ACT
The ACT Network
Powered by NCATS CTSA Program

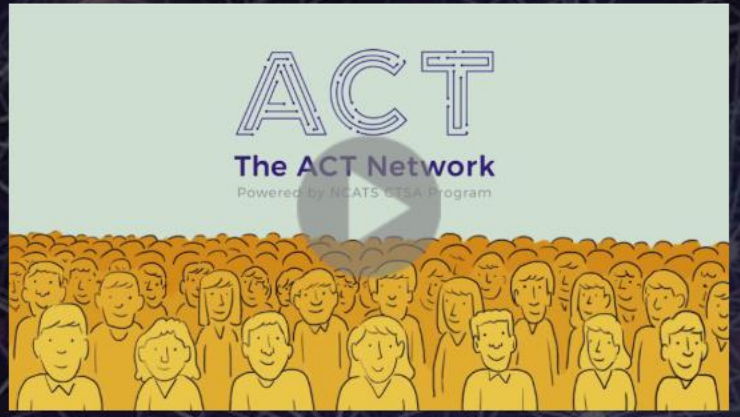
Use ACT

Register for ACT

Welcome to the ACT Network!

The ACT Network is a real-time platform allowing researchers to explore and validate feasibility for clinical studies across the NCATS Clinical and Translational Science Award (CTSA) consortium, from their desktops. ACT helps researchers design and complete clinical studies, and is secure, HIPAA-compliant and IRB-approved.

ACT was developed collaboratively by members of NCATS' Clinical and Translational Science Award (CTSA) consortium, with funding from the NIH National Center for Advancing Translational Sciences.



National Center for Advancing Translational Sciences

The ACT Network

Powered by NCATS CTSA Program

The ACT Network lets researchers explore and validate feasibility for clinical studies across the NCATS Clinical & Translational Science Award (CTSA) consortium, in real time, from their desktops.

ACT is secure and HIPAA-compliant.

125 MILLION PATIENTS
42 SITES CONNECTED
AND GROWING.

Connected to ACT:

Boston University
Children's National
Columbia University
Duke University
Emory Univ./Morehouse Univ.
Harvard University
Indiana University
Johns Hopkins University
Mayo Clinic
Medical College of Wisconsin
Medical University of South Carolina
New York University
Northwestern University
Ohio State University
Oregon Health & Science University
Pennsylvania State University
Stanford University
University of Alabama at Birmingham
U. of Arkansas for Medical Sciences
University of California, Davis
University of California, Irvine
University of California, Los Angeles
University of California, San Diego
University of California, San Francisco
Univ of Cincinnati/Cincinnati Children's
Univ of Colo/Children's Hosp. Colorado
University of Florida
University of Illinois-Chicago
University of Kansas
University of Kentucky
University of Minnesota
University of North Carolina at Chapel Hill
University of Pittsburgh
University of Southern California
UTHealth Houston
UT Health San Antonio
UT Southwestern
University of Washington
Vanderbilt University Medical Center
Virginia Commonwealth Univ.
Washington University in St. Louis
Weill Cornell Medicine

Staging for ACT:

Case Western University
Dartmouth College
Scripps Research / Scripps Health
Tufts University
University at Buffalo
University of Massachusetts
University of Miami
University of Michigan
University of New Mexico
University of Rochester
University of Texas Medical Branch
University of Utah
University of Virginia
University of Wisconsin-Madison
Wake Forest University

CONTACT

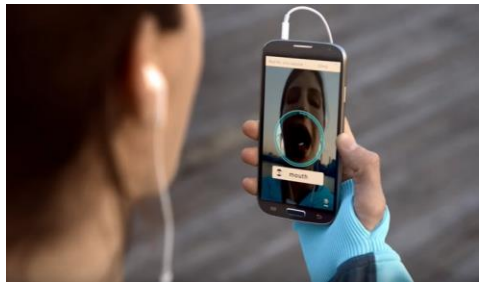
ACTNetwork@pitt.edu

ACCESS

www.ACTNetwork.us

National Center for Advancing Translational Sciences

NCATS SBIR awardee focused on AI-driven improvements in medication adherence



- Artificial intelligence smartphone application uses the patient's smartphone camera and a software algorithm to confirm the identities of the patient and the medication and verify they are taking the right medication at the right time.
- In a recent study, there was a 50 percent improvement in patient adherence in patients taking anticoagulation therapy to help prevent blood clots (*Stroke* 2016)

"The NIH support has enabled our company to attract and leverage an additional $12.25 million in financing from venture capital investors."

– Adam Hanina, M.B.A.
Co-founder and CEO

- ***Workshop July 12, 2019***
  - Co-hosted by NCATS, the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and the National Cancer Institute (NCI)
- ***Goal:***
  - Gather perspectives on and explore the issues associated with translating Machine Intelligence (MI) for applications in healthcare
- ***Issues addressed:***
  - Potential of MI in improving patient health care and outcomes
  - Barriers associated with the development and use of MI in clinical environments
  - Key issues and challenges within the primary topic areas – trustworthiness, explainability, usability, transparency, and fairness – and potential approaches to address them
- ***Follow-up:***
  - Workshop webpage with presentations, speaker bios, and Executive Summary
  - Whitepaper publication in development with workshop Co-Chairs; expected early 2020

National Center for Advancing Translational Sciences

Pressing Issues Identified by the MI Workshop

The need to…

1. Integrate monitoring over time & create a feedback loop
   - To appropriately update MI systems based on clinical developments & standard of care and promote continual monitoring of the system
2. Fund research that advances the science of healthcare
   - Including implementation
3. Promote inter-disciplinary/-sector collaboration
4. Utilize clear explanations and justifications of MI systems
   - To build trust over time and improve uptake of these systems in healthcare
5. Promote incorporation of SDoH and health outcomes
6. Emphasize transparent MI frameworks
   - To mitigate bias perpetuation and assist in interpreting internal MI system decisions

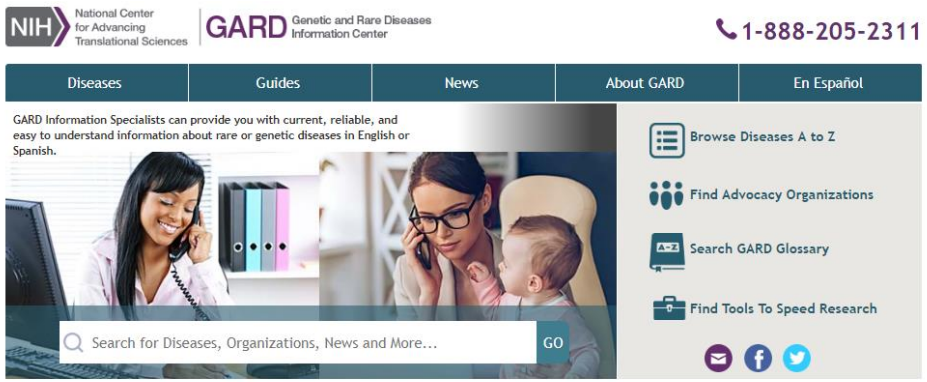
Rare Diseases Are Public Health Issue

- ~7000 diseases
  - ~80% genetic
  - ~50% onset in childhood
  - ~250 new rare diseases identified every year
- Individually rare, cumulatively common
  - Definition varies by country: US <200,000; Japan <50,000; EU <1/2,000
  - Total prevalence ~8% (US ~25 million)
- High costs in direct patient care, loss of productivity
- Accurate diagnosis often requires 5-15 yrs
- Only 5% of rare diseases have a regulatorily approved treatment
  - 2000 years before treatments for all rare diseases on current trajectory
- Solution: transition from “one disease at a time” to “many diseases at a time” approach
  - Commonalities among diseases
  - Platform technologies for diagnosis and treatment

National Center for Advancing Translational Sciences

NCATS “Many Diseases at a Time” Research Programs

Providing Information and “Big Data”



About GARD

The Genetic and Rare Diseases Information Center (GARD) is a program of the National Center for Advancing Translational Sciences (NCATS) and is funded by two parts of the National Institutes of Health (NIH): NCATS and the National Human Genome Research Institute (NHGRI). GARD provides the public with access to current, reliable, and easy-to-understand information about rare or genetic diseases in English or Spanish.

Empowering Patients as Research Partners

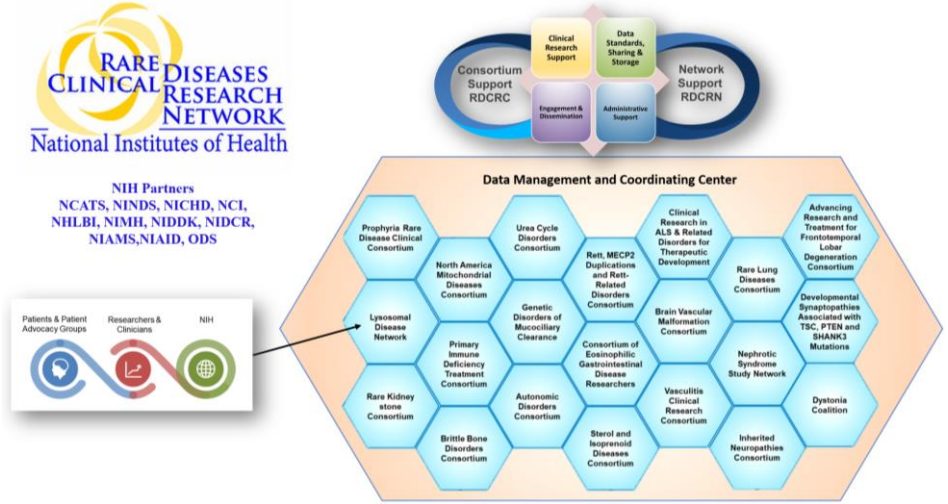
Developing Interoperable Registries

About the RaDaR Program

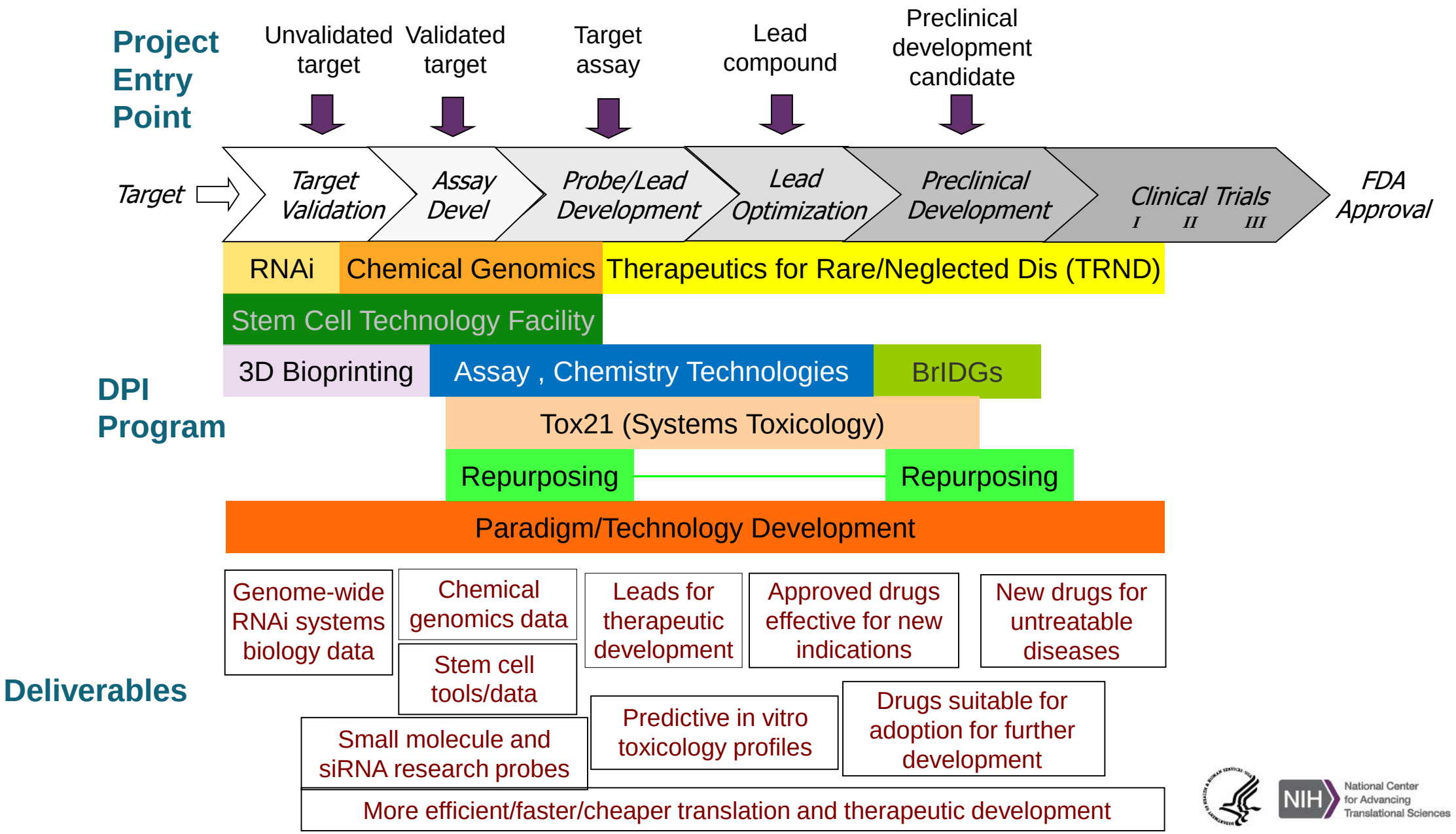
The Rare Diseases Registry (RaDaR) Program, formerly known as the Global Rare Diseases Registry Data Repository (GRDR) program, aims to provide easily accessible advice for constructing and maintaining good-quality rare disease patient registries to enable therapeutics development.

The RaDaR website is under construction. Please sign up below or check back for new information as the site is updated.

Natural History and Interventional Studies



NCATS Division of Preclinical Innovation



Use of AI in early translation

- Discovery of new probes for understudied protein targets
- Development of novel computational approaches to advance drug development processes
- Application of new techniques in computer-aided design to deliver compounds with desired property profiles

National Center for Advancing Translational Sciences

NCATS SBIR awardee focused on new AI-driven drug screening technologies

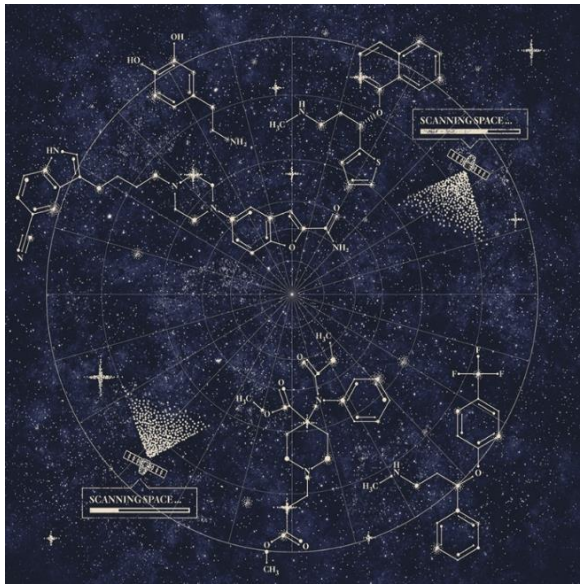
- NCATS Direct to Phase 2 SBIR grant provided support to model 2,000 genetic diseases in multiple human cell types. This approach will enable exploration of treatments for hundreds of diseases in a short period of time.
- Since the NCATS SBIR Award, Recursion has attracted more than $100 million in investments and strategic partnerships with Sanofi and Takeda
- Launched their first clinical program
- Recursion staff has grown to over 100 employees

"The SBIR award helped legitimize our project in the eyes of both investors and the pharmaceutical industry and thus was instrumental in getting the company off the ground."

– Christopher Gibson, Ph.D.
Co-founder and CEO

Defining biologically active chemical space:
A key translational challenge

- 95% of human diseases have no regulatorily approved treatment
- 90% of biological space (“targets”) is currently undrugged
- Vast chemical space: 10^{60} potential “drug-like” small molecules
  - Only 10^{7} of these have been made in the entire history of synthetic chemistry
- Current approach to exploring chemical space is inefficient



Mullard A. *Nature*, 2017 Sep 26; 549 (7673): 445-447.

National Center for Advancing Translational Sciences

Choosing from chemical space: a near-infinite number of potential molecules

Many physical screening collections contain <**1M** compounds

Is that enough?

“Did You Lose the Keys Here?”
“No, But the Light Is Much Better Here”

ChemNavigator (iResearch Library): >60 million (unique structures)

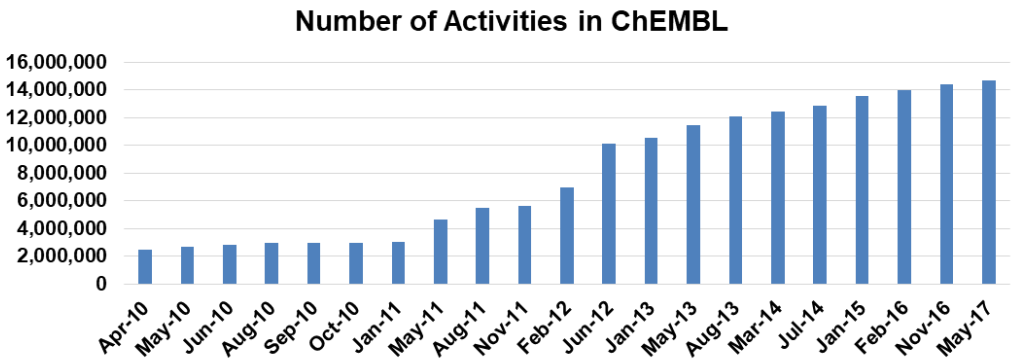
Chemical space has been estimated to be **>10^{60} molecules**[1] (for 30 or fewer heavy atoms).
There are many molecules that are **never** looked at.

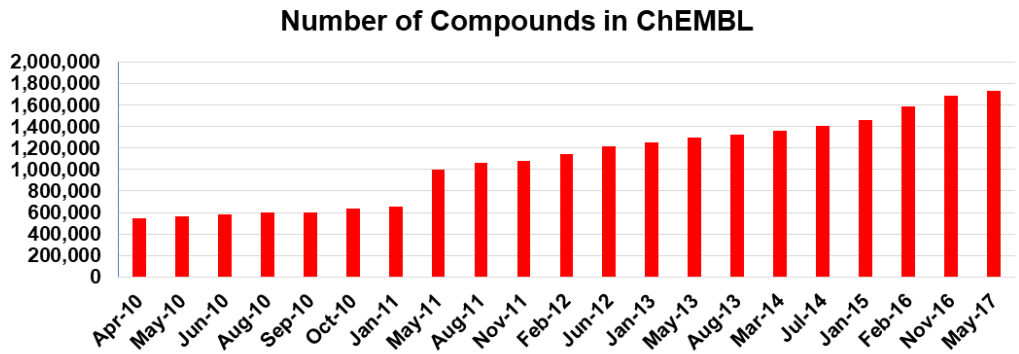
[1] Bohacek, R., McMartin, C. & Guida, W. *Med. Res. Rev*. 16, 3–50 (1996).

Problems of de novo design

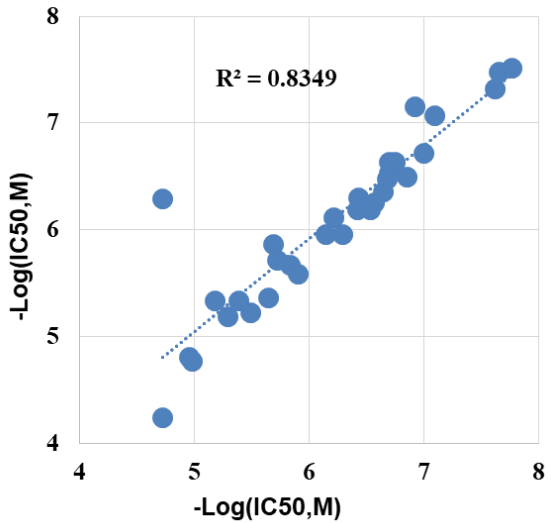
- ∉ Not in screening databases
- $$$$ Many steps and very costly
- Potentially unsafe
- FAILED Syntheses often unsuccessful

Better usage of available chemical biological data





Structures	Arginase-1	Arginase-2
NC(CCO)(CCCCB(O)O)C(=O)O	4.72	4.24
NC(CCCCB(O)O)(CCN1CCCCC1)C(=O)O	4.72	6.29
NCC(N)(CCCCB(O)O)C(=O)O	5.68	5.87
NC(CCCCB(O)O)C(=O)O	5.72	5.72
[NH3+]C(CCCCB(O)O)C(=O)[O-]	5.83	5.67
NC(CCCCB(O)O)(CCN1CCSCC1)C(=O)O	5.91	5.59
COCC1CCCN1CCC(N)(CCCCB(O)O)C(=O)O	6.22	6.12
NC(CCCCB(O)O)(CCN1CCc2ccccc2C1)C(=O)O	6.29	5.96
NC(CCCCB(O)O)(C(=O)O)C1CCNCC1	6.42	6.19
CCN(CC)CCC(N)(CCCCB(O)O)C(=O)O	6.43	6.3
NC(CCCCB(O)O)(CCN1CCC(O)CC1)C(=O)O	6.54	6.19
CCCN(C)CCC(N)(CCCCB(O)O)C(=O)O	6.85	6.49
CC(C)NCCC(N)(CCCCB(O)O)C(=O)O	7	6.72
NC(CCCCB(O)O)(C(=O)O)C1CC2CCC(C1)N2	7.1	7.07
NC(CCCCB(O)O)(C(=O)O)C1CC2CCC(C1)N2Cc1ccccc1	7.62	7.33
NC(CCCCB(O)O)(C(=O)O)C1CC2CCC(C1)N2Cc1ccc(F)c(F)c1	7.66	7.48
NC(CCCCB(O)O)(C(=O)O)C1CC2CCC(C1)N2Cc1ccc(Cl)cc1	7.77	7.52



Comprehensive mapping of chemical biology space enables the development of large-scale QSAR modeling

Aggregated results from the related biological targets could improve the quality of QSAR models.

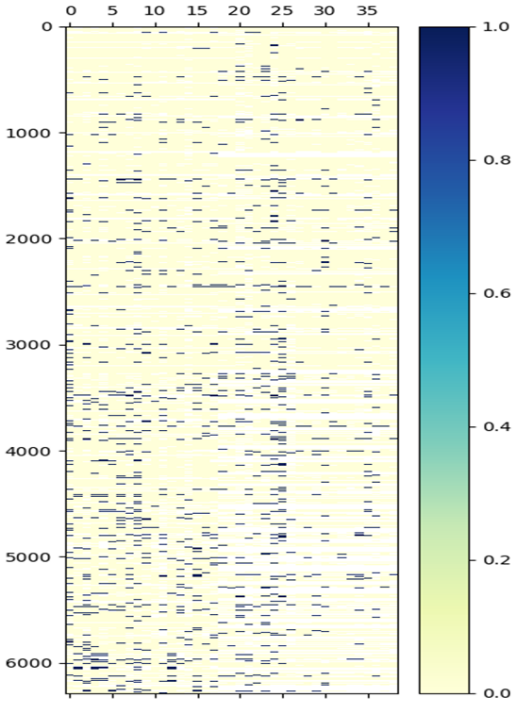
NCATS Tox21 data

7857 compounds were randomly divided into training and test set in proportion 80% and 20%

6286 compounds in the training set

1571 compounds in the test set

Aryl hydrocarbon receptor
Androgen receptor
Aromatase
Genotoxicity inducer
Estrogen receptor alpha
Thyroid receptor
Glucocorticoid receptor
Mitochondrial membrane potential disruptor
p53
Peroxisome proliferator-activated receptor gamma
Androgen receptor
Heat shock response
Nuclear factor-kappa B
Retinoic acid receptor
Retinoid-related orphan receptor gamma
Thyroid stimulating hormone receptor
Activator protein-1
Antioxidant response element
Constitutive androstane receptor
Endoplasmic reticulum stress response
Farnesoid-X-receptor
H2AX
Hypoxia-inducible factor-1
Peroxisome proliferator-activated receptor delta
Retinoid X receptor
Vitamin D receptor



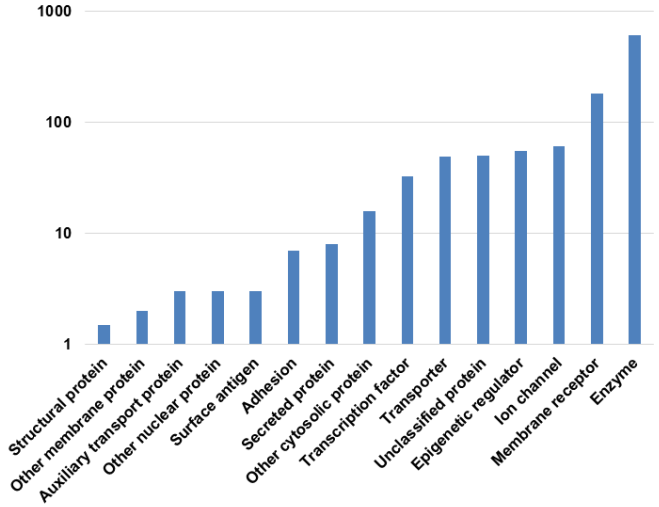
High quality data: each compounds measured 3 times, 50 concentrations.

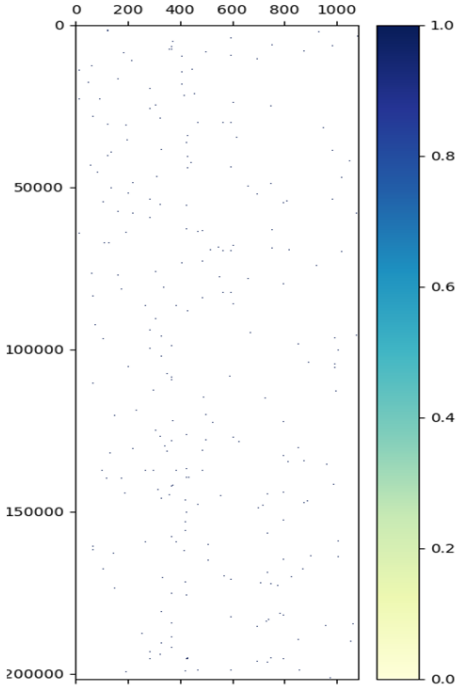
ChEMBL data

251998 compounds were randomly divided into training and test set in proportion 80% and 20%

201599 compounds in the training set

50399 compounds in the test set



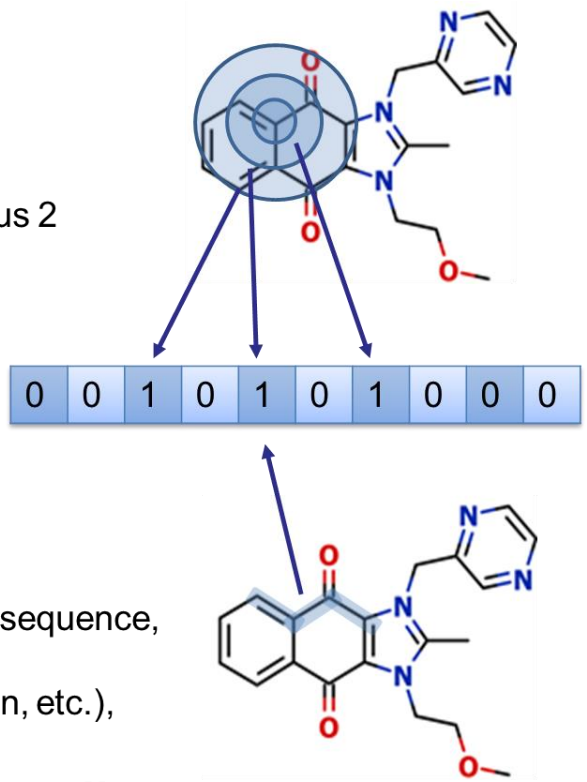


Huge target coverage: ~4000 human proteins

Molecular Representation

Descriptors:

- RDkit Morgan fingerprints
  Circular fingerprints, 1024 bit, radius 2
- RDkit Avalon fingerprints
  Path based fingerprints, 1024 bit
- RDkit AtomPair fingerprints
  Path based fingerprints, 1024 bit
- PROFEAT descriptors
  Proteins features from amino acid sequence, 14 descriptors types (Amino acid composition, Dipeptide composition, etc.), total is **1437** descriptors

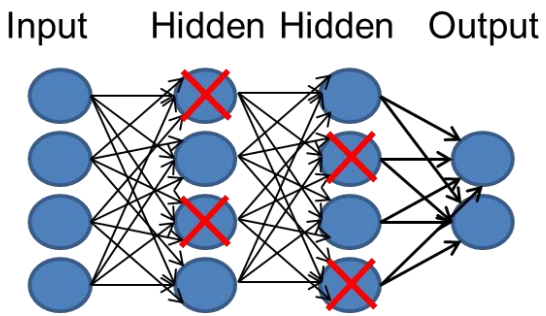


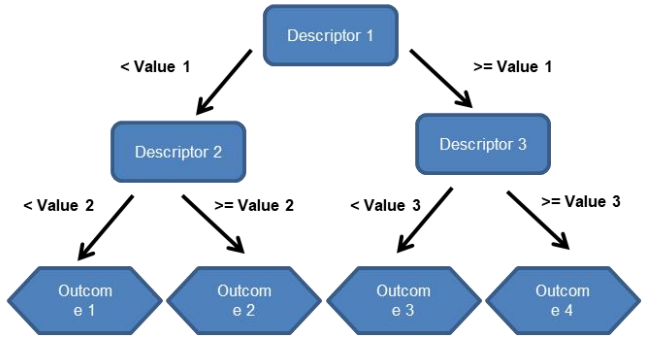
MVLEMLNPIH YNITSIVPEA
MPAATMPVLL LTGLFLLVWN

$$f(r) = \frac{Nr}{N}, r = 1, 2, \ldots 20$$

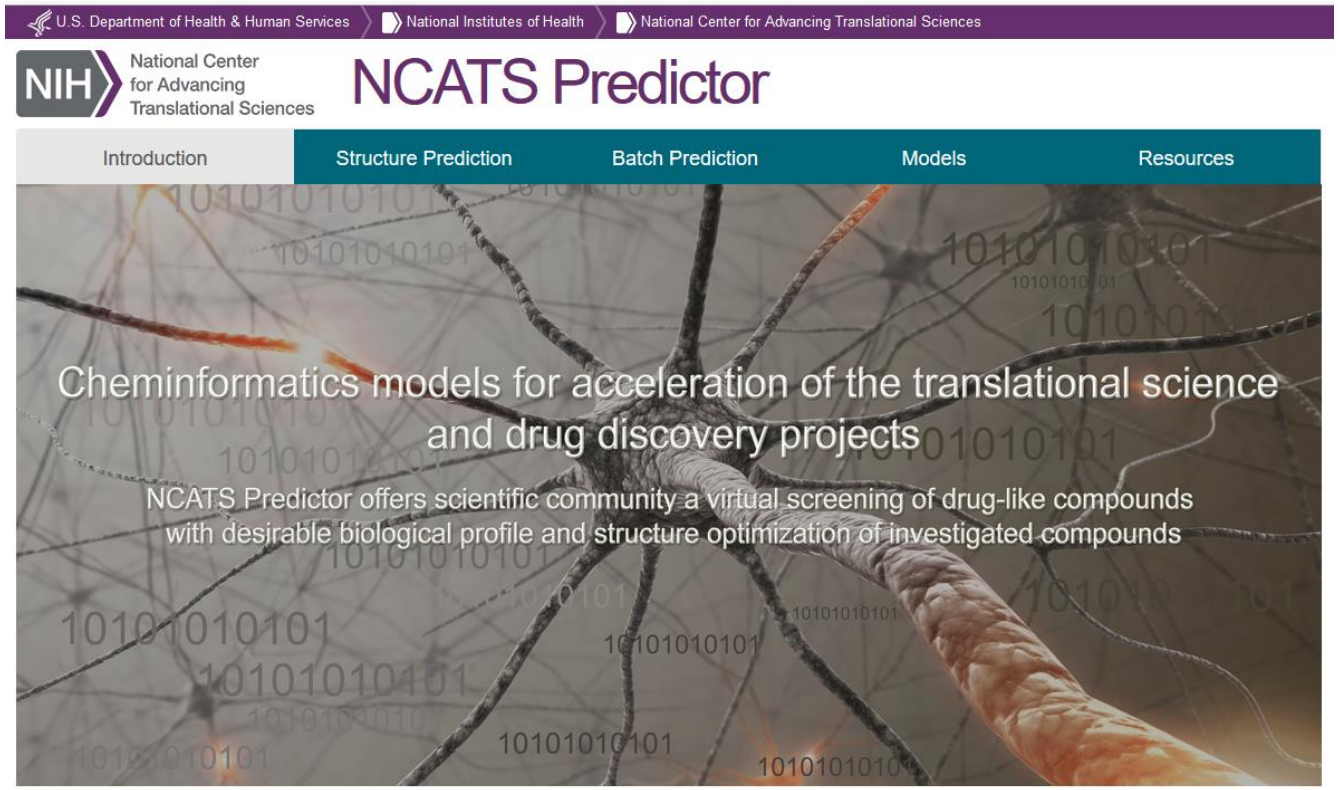
Machine Learning approaches

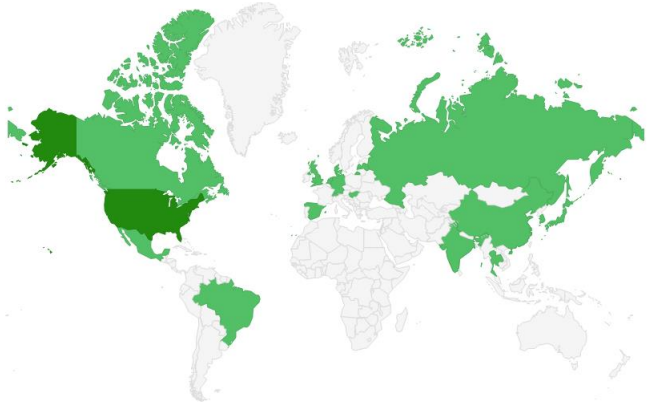
- Random Forest: 100-500 trees, 100-300 features
- Deep Learning: ReLu, 3-5 hidden layers, ADAM optimizer, Dropout, Dense layers, ConvNet

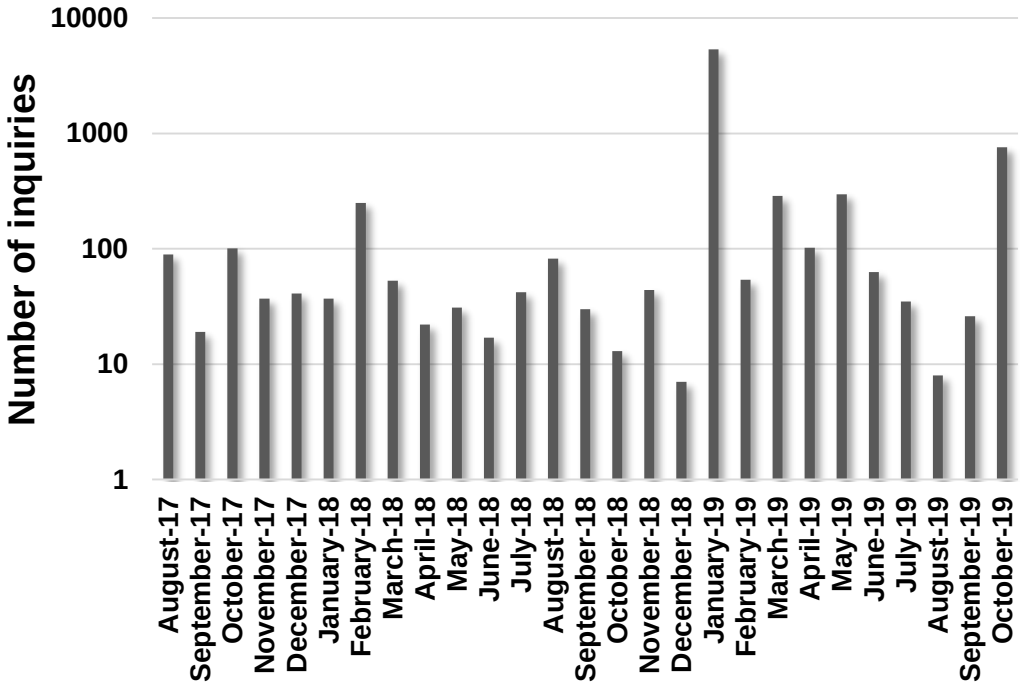




https://predictor.ncats.io/







- Predict **1121** biological activities
- Supports SMILES, drug name, images
- Allows to send the batch of compounds
- Show up neighbor activity and structure

https://doi.org/10.1021/acs.jcim.9b00526

National Center for Advancing Translational Sciences

ASPIRE

A Specialized Platform for Innovative Research Exploration

"ASPIRE aims to address two challenges of the current era in biomedical research: to harness new technologies to accelerate understanding of living systems and to fulfill the promise of science to improve the lives of the many patients with untreatable or poorly treatable diseases."

National Center for Advancing Translational Sciences

COMMENT

Mapping biologically active chemical space to accelerate drug discovery

G. Sitta Sittampalam, Dobrila D. Rudnicki, Danilo A. Tagle, Anton Simeonov and Christopher P. Austin*

A specialized platform for innovative research exploration — ASPIRE — in preclinical drug discovery could help study unexplored biologically active chemical space through integrating automated synthetic chemistry, high-throughput biology and artificial intelligence technologies.

With increasing understanding of the molecular basis of disease in the last 30 years, a major roadblock to timely translation into new therapies has been the inability to efficiently identify new areas of biologically active small-molecule chemical space[1]. Ideally, new chemical probes and drug leads that selectively modulate disease targets and pathways would be produced rapidly and inexpensively, but despite some progress in the past decade[2], the fundamental challenge of exploring chemical space to define new biology remains largely unsolved. Recently, however, advances in chemistry automation and machine learning/artificial intelligence (AI)[3] have raised the prospect of their integration with high-throughput biological screening, assay automation engineering and informatics to enable dramatically more effective, even unsupervised, exploration of biologically active chemical space.

Challenges in chemical space exploration

In its simplest terms, the goal seems straightforward: to define the set of small-molecule chemical structures needed to modulate all biological targets. However, the vast number of chemical structures in drug-like chemical space ($\sim 10^{60}$), and the smaller but still substantial number of biological targets in human and pathogen biological space ($\sim 10^{6}$), has made progress on this problem painfully slow. Currently, only ~3% of biological space is drugged and a further ~7% is tractable via small-molecule probes[1], while the percentage of drug-like chemical space that has been synthesized is miniscule[2].

The effort to define biologically active chemical space involves four main disciplines: biology, chemistry, informatics and engineering. In the last three decades, automation and parallelization have radically improved the efficiency of biological testing, informatics analysis and engineering. High-throughput screening (HTS) technologies have dramatically increased the productivity of the bench biologist such that millions of data points can be acquired in a single day. Advances in the capabilities, precision and robustness of engineering technologies at the micro- and macro-level have also enabled increasingly autonomous physiologically relevant biological screening systems. And remarkable advances in computing power and data analysis algorithms have increased the ability to analyse data by orders of magnitude in quantity and quality. These capacities have, in turn, allowed the development of data-driven principles of biological function.

By contrast, the technologies, throughput and reach of synthetic chemistry have remained relatively unchanged over the last several decades, with combinatorial chemistry, microwave synthesis and other technologies having only limited overall impact on the efficiency of chemistry to explore new chemical space (Supplementary Fig. 1). Chemistry has only recently begun adopting automation and AI technologies to facilitate existing chemistries, reaction optimization and nanoscale synthesis and library generation[3], and the general practice of chemical synthesis remains largely artisanal, with synthetic throughput of novel bioactive chemicals improved at best by tenfold over the last century. This disparate evolution of the biology and chemistry fields now limits the ability to generate novel chemical probes[4], pharmacological tools and drugs to modulate undrugged biological space, and thus contributes to translational research inefficiency.

Machine learning and other AI technologies are increasing in use and sophistication, and they learn, interpret and predict outcomes based on vast amounts of data in applications such as facial recognition and driverless vehicles. Similar technologies applied to large genomic, proteomic and clinical data sets are making in-roads into biomedical sciences. Furthermore, the development of technologies to integrate machine learning with automated chemical synthesis is currently being funded by the Defense Advanced Research Projects Agency in the "Make-It" programme, which is using both batch and flow chemistry for synthesis of on-demand pharmaceuticals in military field operations. The convergence of nascent automated chemical synthesis technologies, high-throughput biological

National Center for Advancing Translational Sciences, National Institutes of Health, Bethesda, MD, USA.

**e-mail: gurusingham.sittampalam@nih.gov*

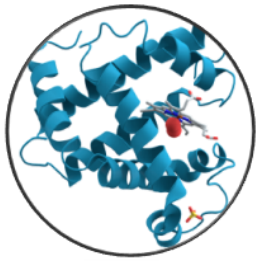
https://doi.org/10.1038/d41573-018-00007-2

ASPIRE Work Flow

Sittampalam et al., *NRDD* 2019

National Center for Advancing Translational Sciences

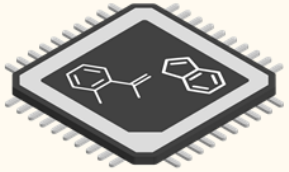
Chemistry-Focused Drug Discovery Workflow



Research Biology

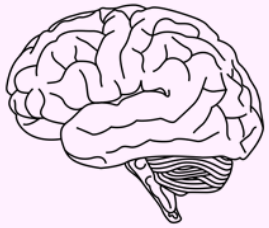
What to make?

High throughput and virtual screens



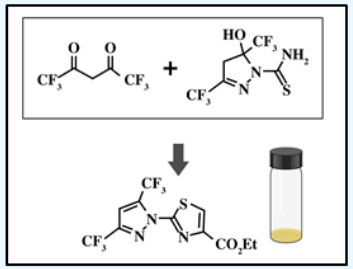
in silico predictors

How to make it?

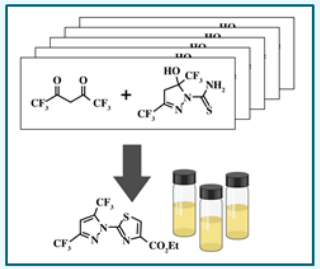


Chemistry route planning

Make



Synthesis



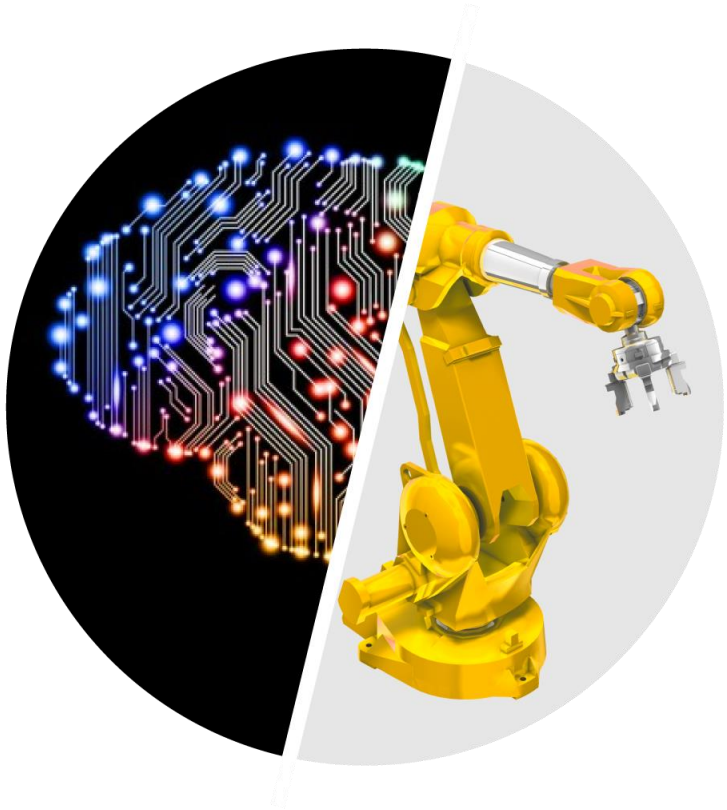
Scale up

Continued validation

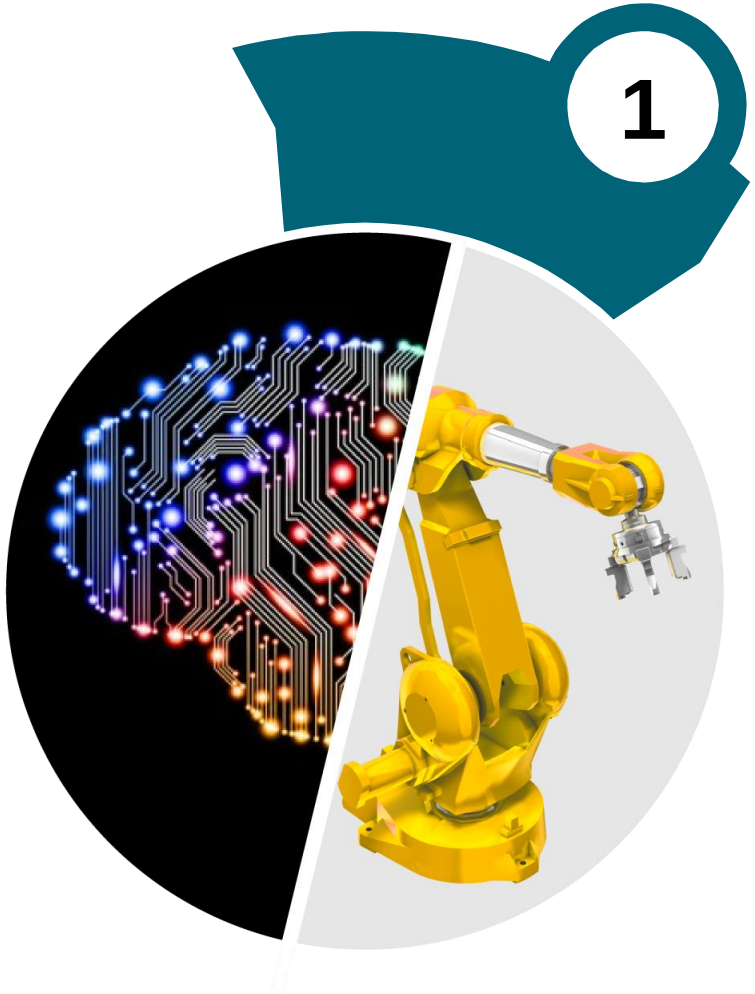
AI/ML-driven high-throughput biological screening assay optimization

- Collaboration with Kebotix
  - Startup Company in Boston that has combined AI with robotics to discover and create advanced chemicals and materials
- Utilize AI/ML to perform Design of Experiment (DOE) automatically for biological assays
- Opportunity to develop, test and implement automated biological test platform with direct interface to informatics platform(s)
- AI/ML output compared to ‘brute force’ method testing all variable combinations of assay conditions

National Center for Advancing Translational Sciences



National Center
for Advancing
Translational Sciences



Send Message

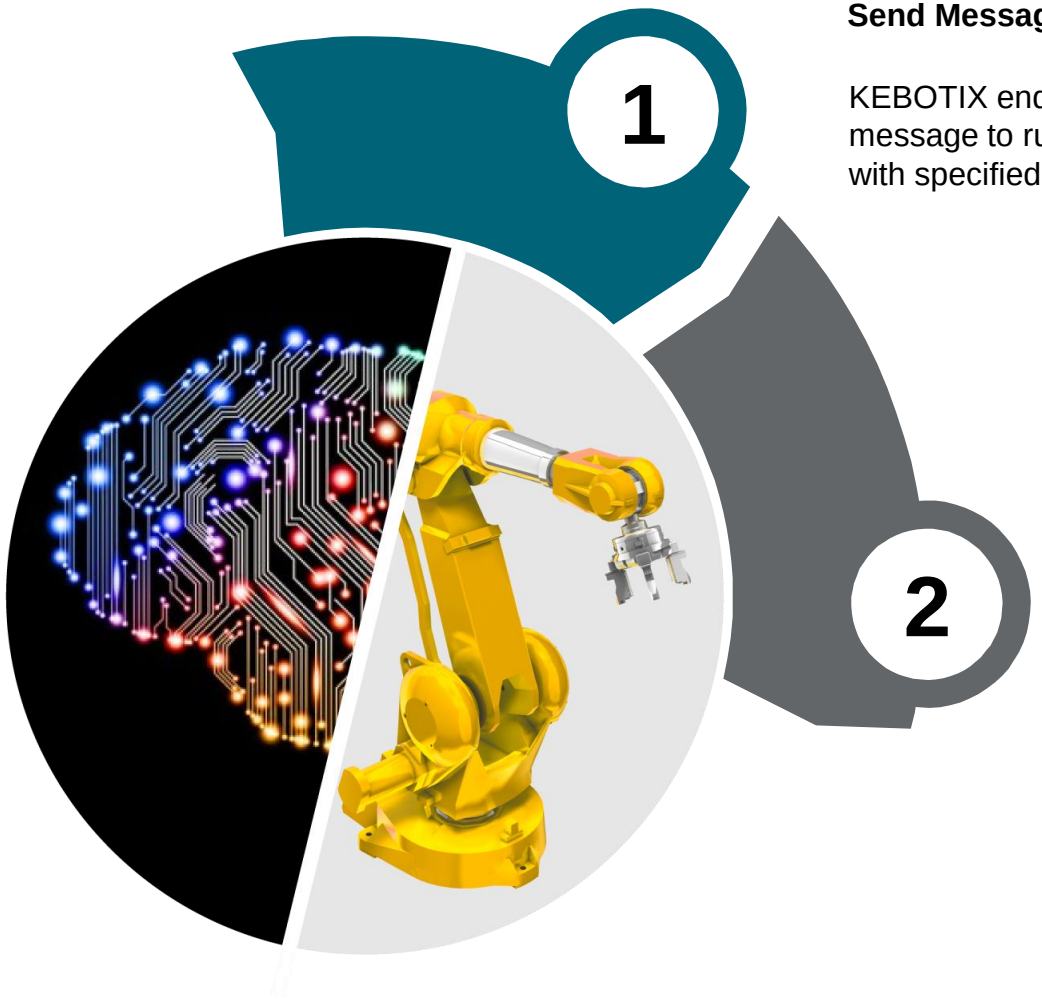
KEBOTIX enqueues a message to run an Assay with specified conditions

1.


```
{
   "Command":"Launch Auto",
   "UUID":"179051",
   "Parameters":[
      [
         "Dispense",
         "Enzyme",
         "8"
      ],
      [
         "Dispense",
         "Substrate",
         "8"
      ],
      [
         "Incubate",
         "Time",
         "2"
      ]
   ]
}
```


National Center for Advancing Translational Sciences

2.

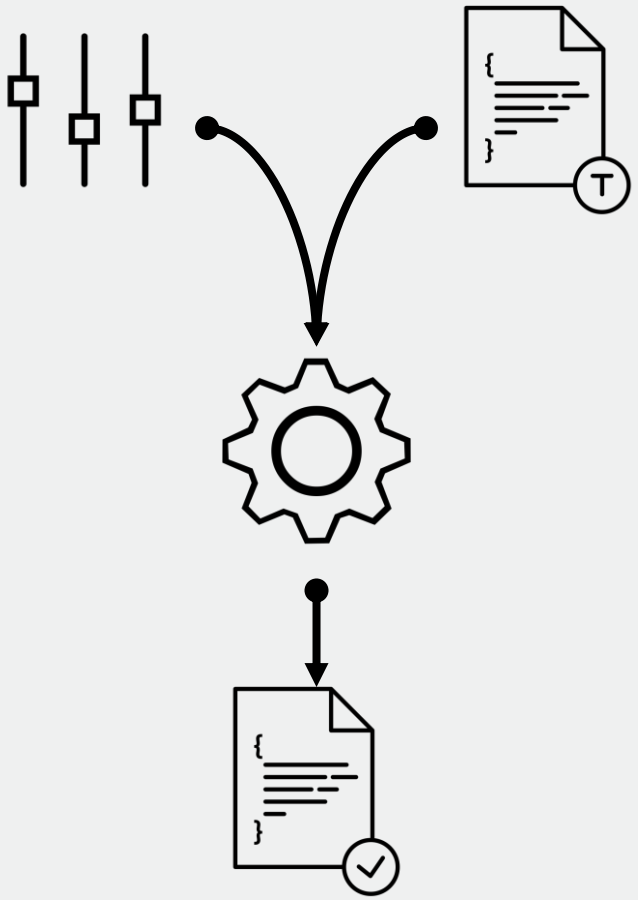


Send Message

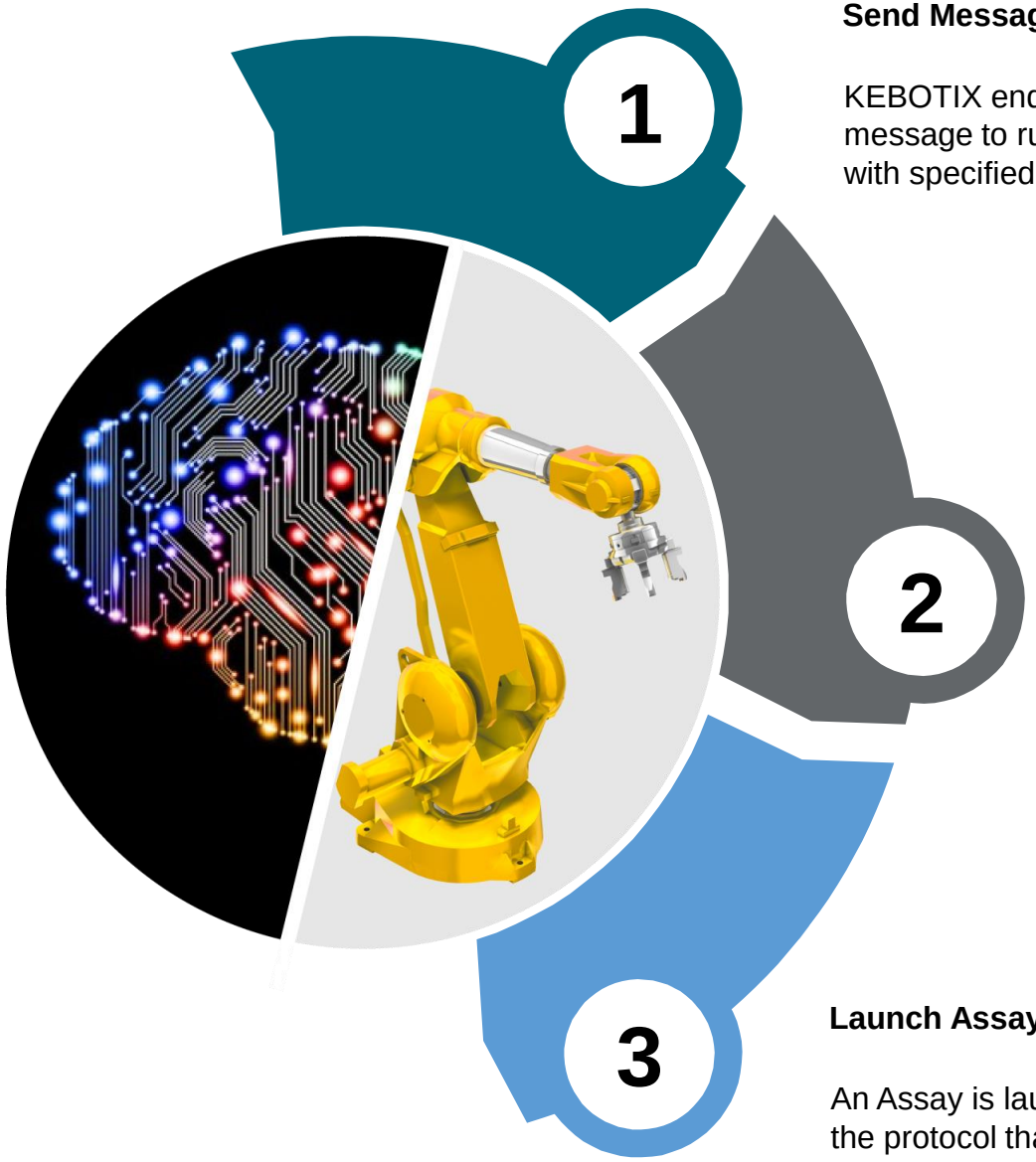
KEBOTIX enqueues a message to run an Assay with specified conditions

Receive Message

The Screening System receives the message and creates a protocol to be run with the specified conditions



National Center for Advancing Translational Sciences



Send Message

KEBOTIX enqueues a message to run an Assay with specified conditions

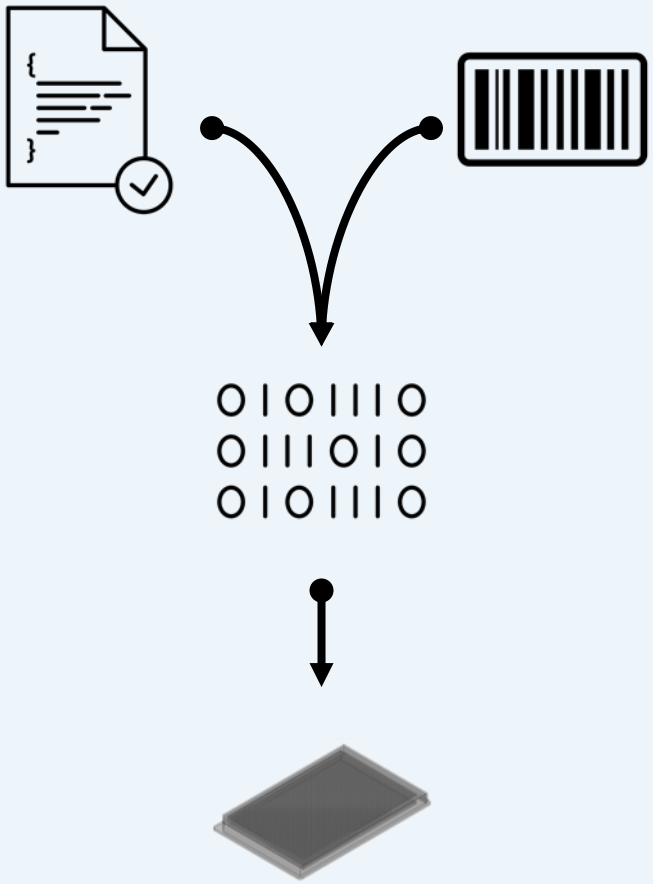
Receive Message

The Screening System receives the message and creates a protocol to be run with the specified conditions

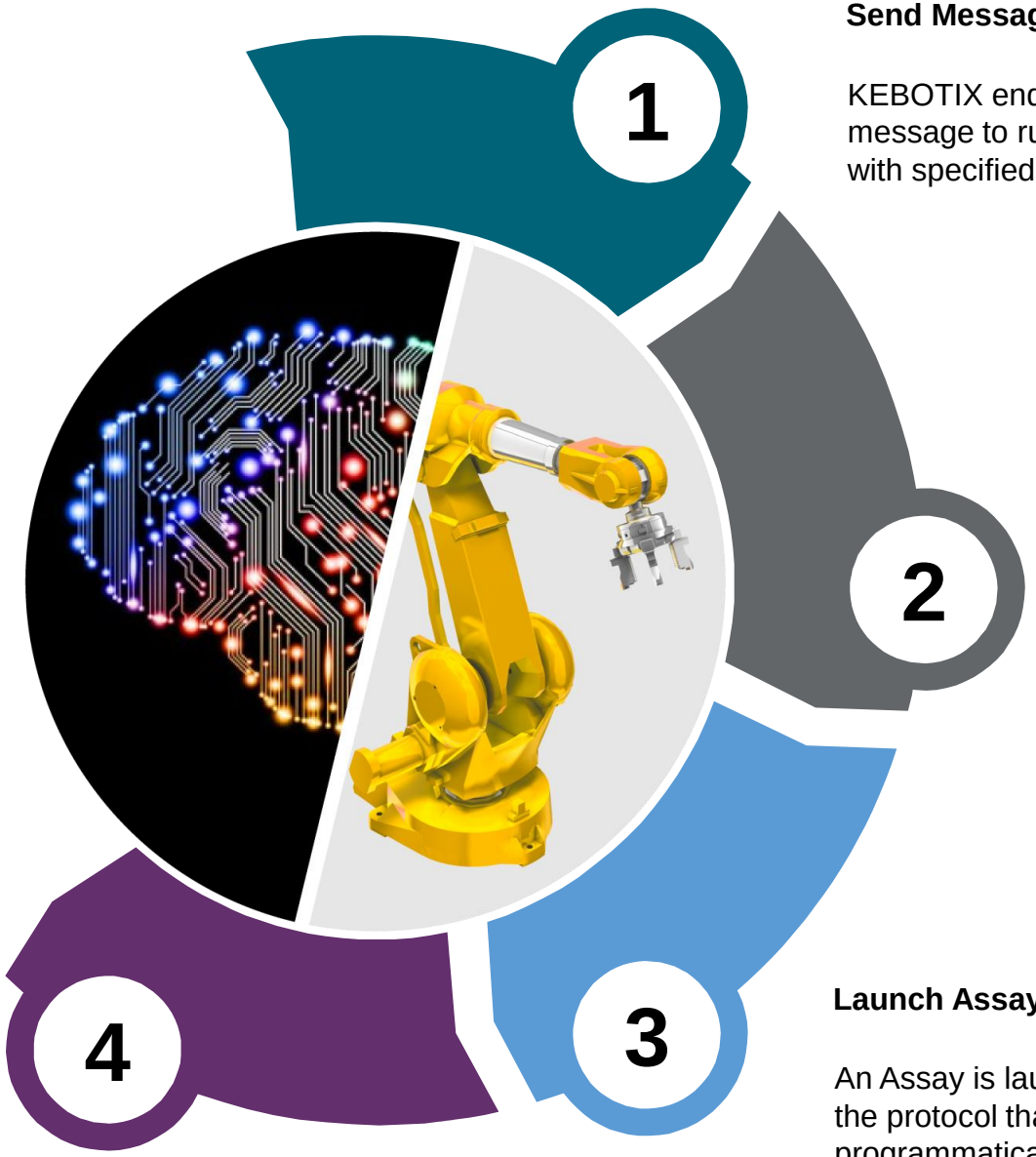
Launch Assay

An Assay is launched using the protocol that was programmatically generated

3.



National Center for Advancing Translational Sciences



Send Message

KEBOTIX enqueues a message to run an Assay with specified conditions

Receive Message

The Screening System receives the message and creates a protocol to be run with the specified conditions

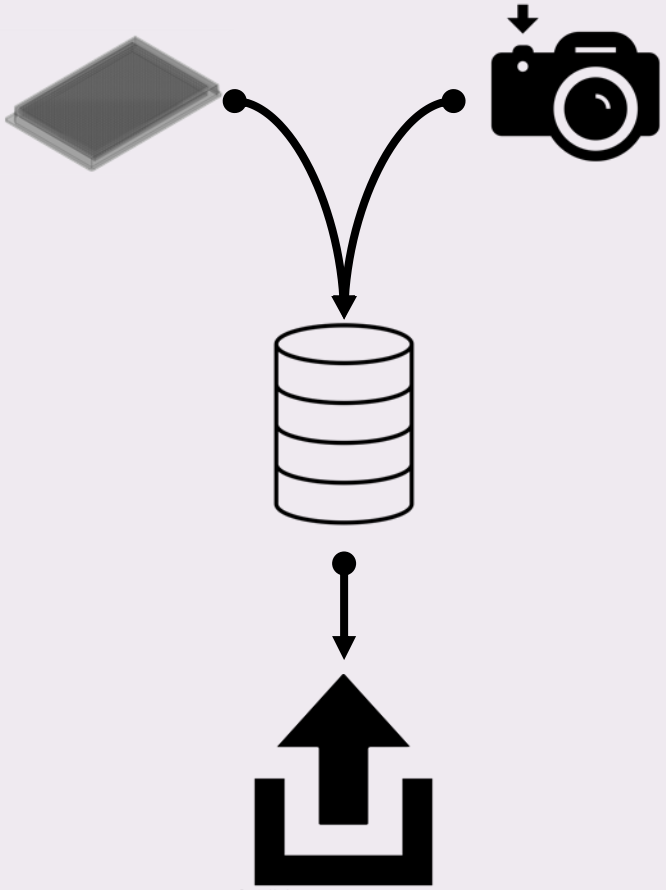
Launch Assay

An Assay is launched using the protocol that was programmatically generated

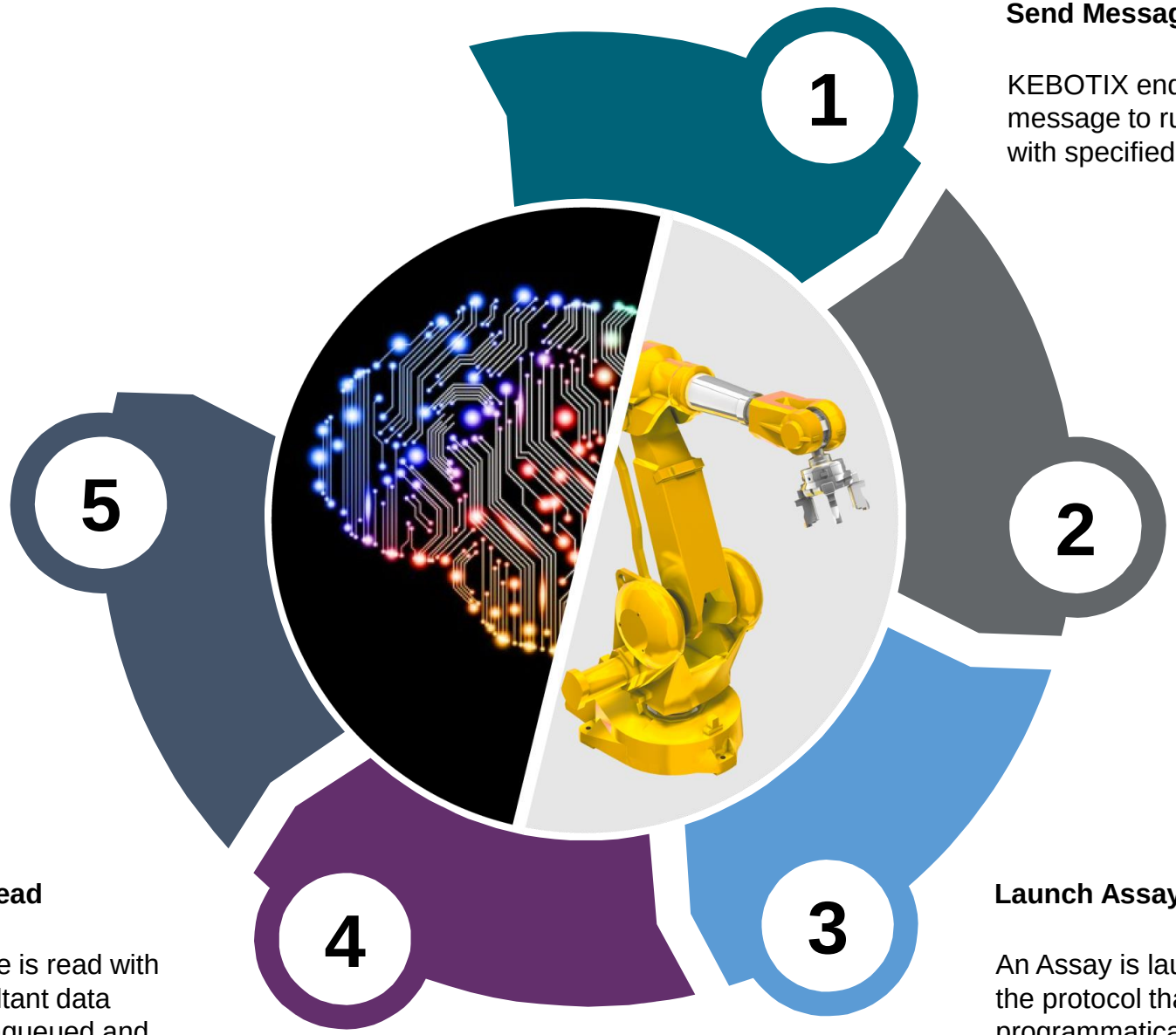
Plate Read

The plate is read with the resultant data being enqueued and sent to KEBOTIX

4.



National Center for Advancing Translational Sciences



Send Message

KEBOTIX enqueues a message to run an Assay with specified conditions

Receive Message

The Screening System receives the message and creates a protocol to be run with the specified conditions

Launch Assay

An Assay is launched using the protocol that was programmatically generated

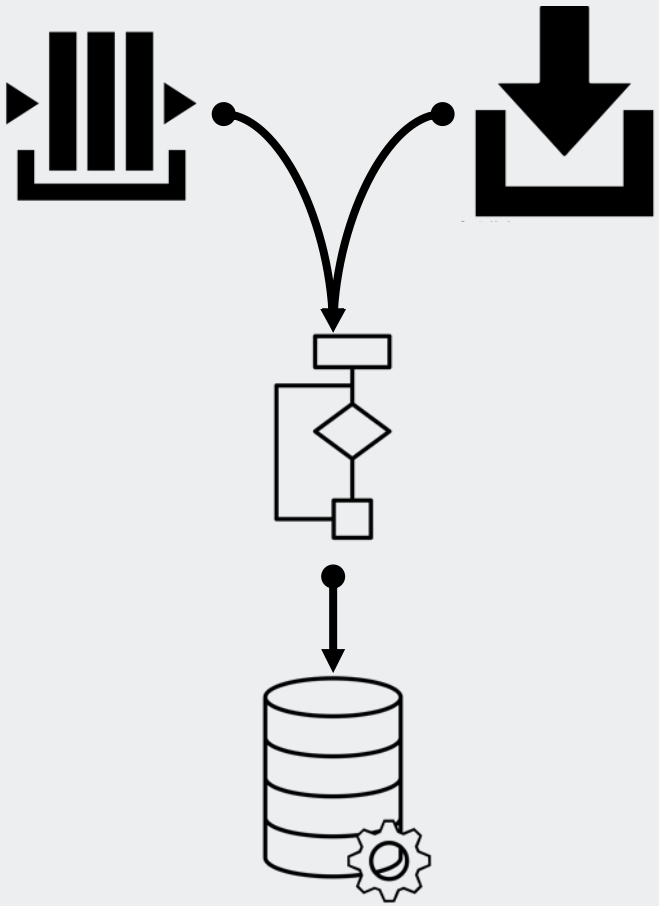
Plate Read

The plate is read with the resultant data being enqueued and sent to KEBOTIX

Data Processing

The message is received by KEBOTIX and the data is processed

5.



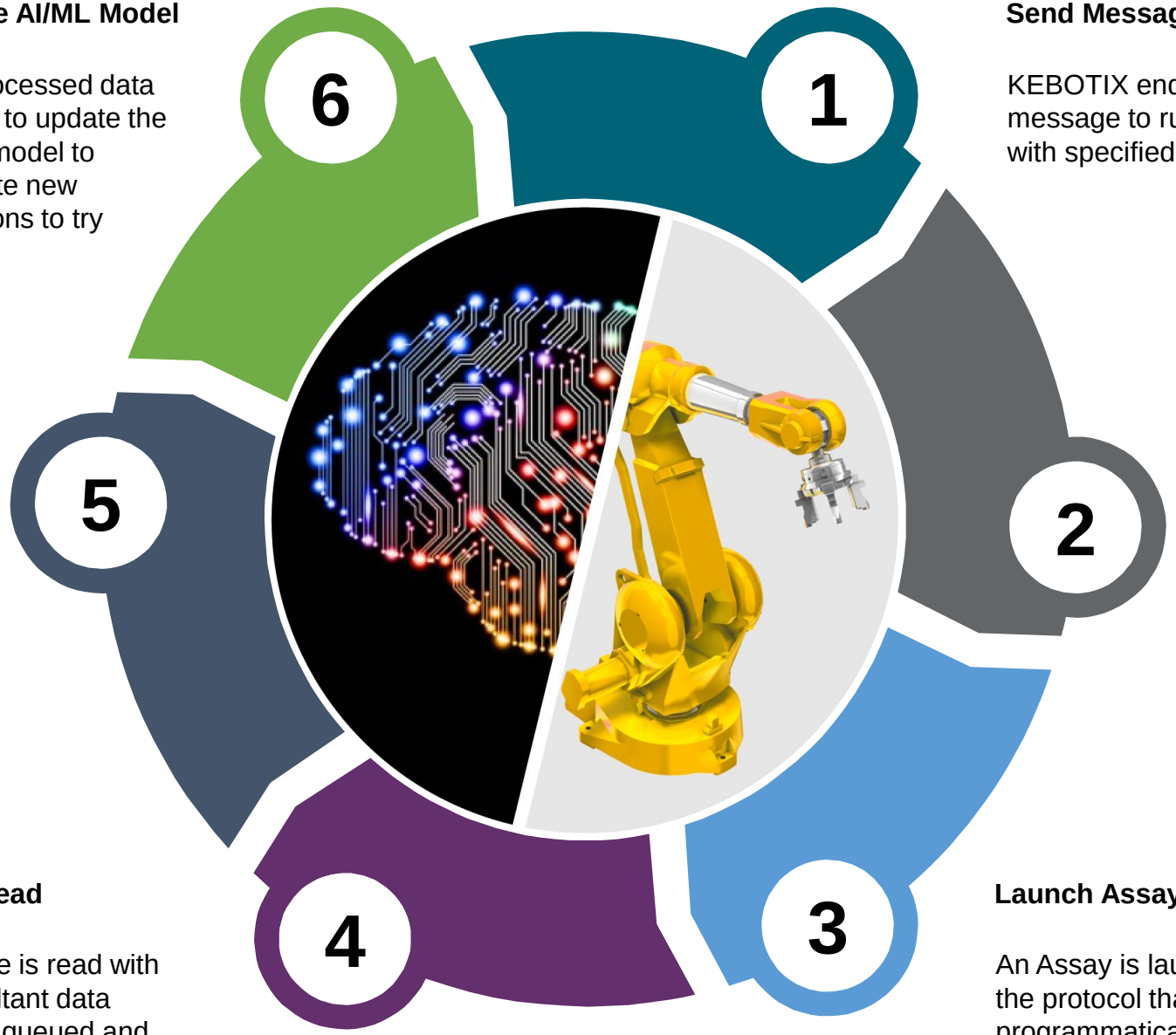
National Center for Advancing Translational Sciences

Update AI/ML Model

The processed data is used to update the AI/ML model to generate new conditions to try

Send Message

KEBOTIX enqueues a message to run an Assay with specified conditions



Data Processing

The message is received by KEBOTIX and the data is processed

Receive Message

The Screening System receives the message and creates a protocol to be run with the specified conditions

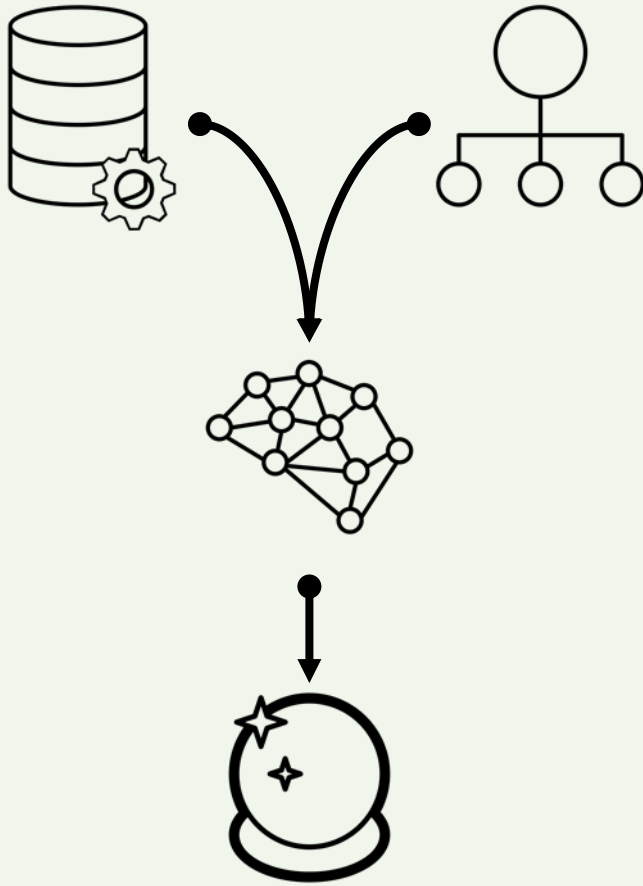
Plate Read

The plate is read with the resultant data being enqueued and sent to KEBOTIX

Launch Assay

An Assay is launched using the protocol that was programmatically generated

6.



National Center for Advancing Translational Sciences

The NCATS Biomedical Data Translator Program

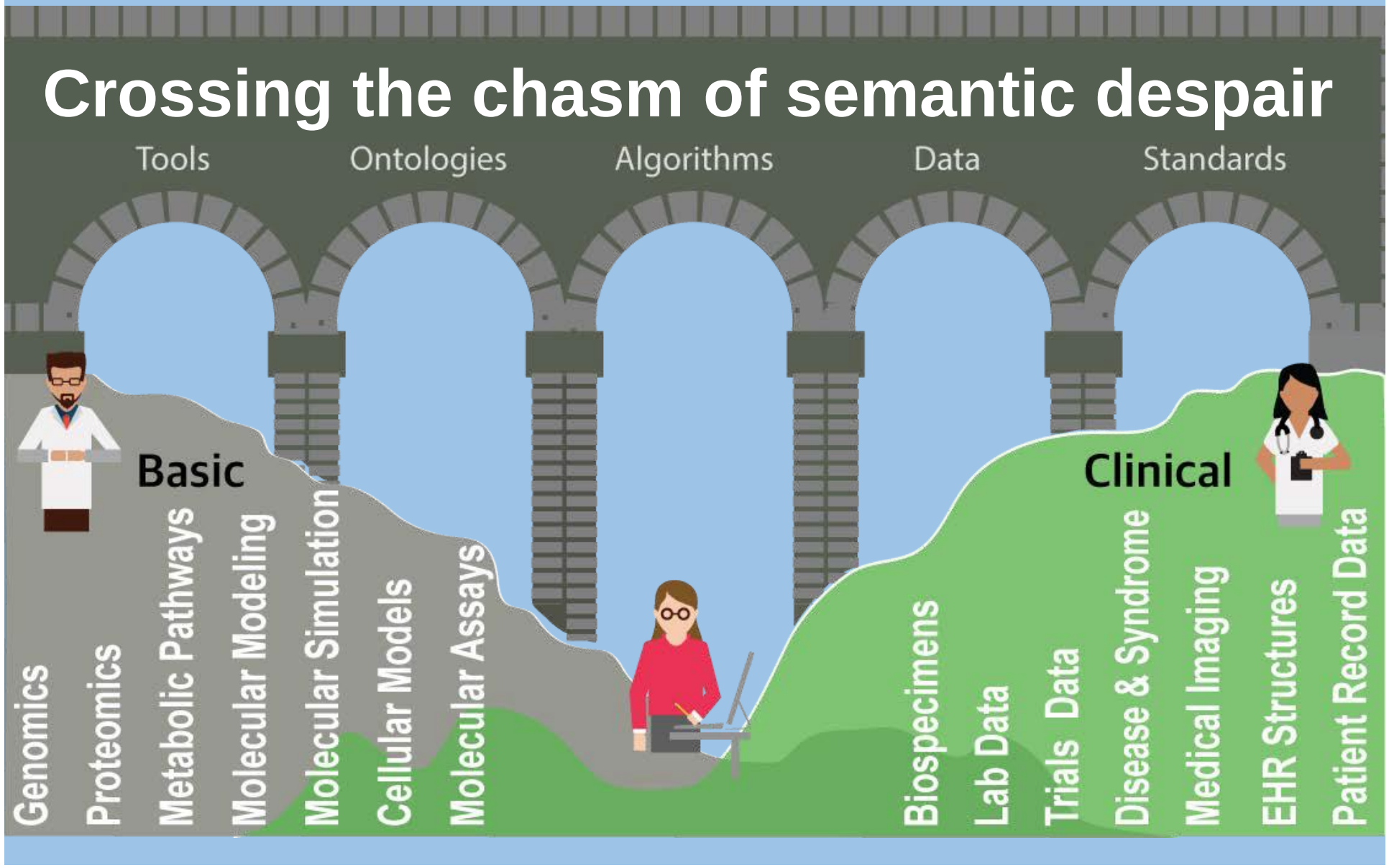


Figure courtesy of Julie McMurry & Chris Chute

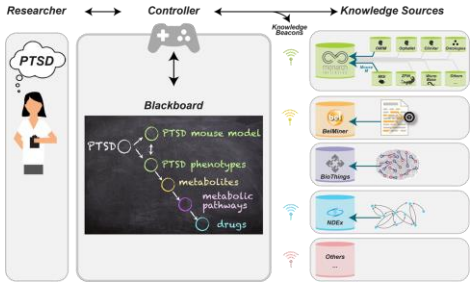
National Center for Advancing Translational Sciences

Translator Vision

Accelerate biomedical innovation by developing a **biomedical "data translator"** for the research community

- computationally-assisted exploration of knowledge
- construction of new research hypotheses

Produce tools that augment human reasoning and provide inference for understanding the pathophysiology of human disease.



Sep '16 — Notice of Awards Issued

Oct '16 — First Face-to-Face Meeting NCATS

Nov '16 — Funds Released

Jan '17 — Blackboard/ Architecture Meeting LBNL

May '17 — Adapt Knowledge Sources to Interact with Blackboard NCATS

Sep '17 — Reasoning Tool FOA Released

Oct '17 — Open Meeting/ Hackathon UNC

Jan '18 — Reasoning Tool Initial Awards issued

Hackathon UCSD

May '18 — Hackathon NCATS

Sep '18 — Hackathon OHSU

Dec '18 — CAN Review Board Meeting

Mar '19 — Hackathon UNC

Sep '19 — Open Meeting/ Hackathon ISB, Seattle

Rapid and flexible team-based development

National Center for Advancing Translational Sciences

TRANSLATOR TIDBIT 05

Finding Unanticipated Patterns in Clinical Cohorts Using Open Clinical Data

Adverse events correlated with commonly-prescribed diabetes drugs

Translator query

I know that...

Adverse events

are associated with...

Diabetes patients

who are prescribed...

Marketed drugs

For decades, it has been standard practice to treat patients based on broadly-defined groups that they fit into based on sex, racial group or other outward-facing factors. Unfortunately, this has led to drugs' often being prescribed to patients based on a relatively small amount of information known about them, with little emphasis on how the patient's own genetic or phenotypic status may affect the outcomes.

This is the case for diabetes, a metabolic disorder that affects millions of people in the U.S. and around the world. For type I diabetes, the most common prescribed medication is insulin. Patients who take insulin often give themselves injections multiple times per day, and monitor their blood sugar levels. Any relief of burden on diabetes patients from adverse events or from the prescribed medications themselves would be significant.

How might Translator help?

Translator offers the opportunity to recommend a drug of choice for a given patient by answering questions about drug-genotype or drug-phenotype interactions quickly and efficiently. In this case, the search also yielded an unexpected interaction that warrants further investigation. When this unexpected result was investigated in more depth, it was noted that a Google search for the terms produced too much noise to easily allow for the connection to be made. A PubMed search yielded 17 hits but only the 17th was relevant and that paper was from 1968 and written in Spanish, making it essentially inaccessible to a large number of researchers. Our researcher hypothesized that either (1) physicians are implicitly aware of different subcategories of diabetes and that these

Translator program products

Emerging knowledge graph standards

- https://github.com/NCATS-Tangerine/kgx

API access to biomedical knowledge graphs

- https://github.com/NCATS-Tangerine/NCATS-ReasonerStdAPI
- https://smart-api.info/registry?q=Translator

All source code on github

- https://github.com/search?q=NCATS+translator

Clin Transl Sci. 2019 Mar;12:85. doi: 10.1111/cts.125
Clin Transl Sci. 2019 Mar;12:86-90. doi: 10.1111/cts.12591

STAT Topics Opinion Podcast Video Newsletters Events Log In Subscribe TRY STAT PLUS

IN THE LAB

NIH-funded project aims to build a 'Google' for biomedical data

By RUTH HAILU @ruth_hailu_ / JULY 31, 2019

A view of the NIH campus

NIH

Every year, the National Institutes of Health spends billions of dollars for biomedical research, ranging from basic science investigations into cell processes to clinical trials. The results are published in journals, presented in academic meetings, and then — building off of their findings — researchers move on to their next project.

But what happens to the data that's collected and what more could we learn from it? If we aggregated all the data from countless years of research, might we learn something new about ourselves, the diseases that infect us, and possible treatments?

That's the hope behind the Biomedical Data Translator program, launched by the

All involved emphasized the importance of the collaborative nature of the project,
"It's basically like having a menu of great ideas from all the smartest people around the country"

https://tinyurl.com/y3pata2a

National Center for Advancing Translational Sciences

Connect With NCATS
https://ncats.nih.gov/connect

Website: ncats.nih.gov

Facebook: facebook.com/ncats.nih.gov

Twitter: twitter.com/ncats_nih_gov

YouTube: youtube.com/user/ncatsmedia

E-Newsletter: https://ncats.nih.gov/enews