A Deep Learning Approach to Antibiotics Discovery

2020. 05. 10.

Seongok Ryu, AITRICS

Preliminary

신약개발 분야의 큼지막한 분류

- Oncology – 항암제

  → 아직 치료하지 못한 암이 많고, 또한 기존 target에서 mutation 일어난 경우에 대한 신약 needs

  → ex) Osimertinib, Lazertinib (IND), …

- Fibrosis – 섬유화 질환, unmet needs 가 매우 많음

  → 섬유화 질환은 현재까지 irreversible 한 것으로 알려져있음. 많은 임상 study가 이루어지는 중.

  → ex) 비알코올성지방간(NASH), 다발성폐섬유화(IPF)

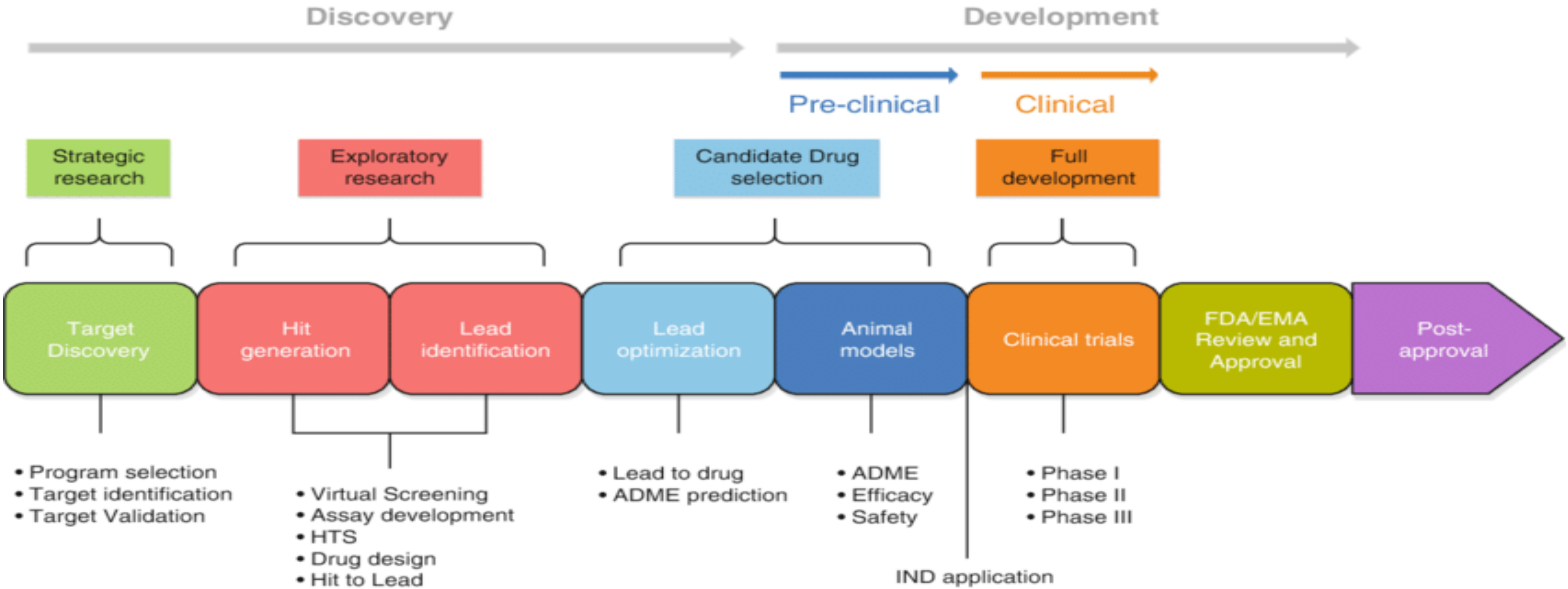
- **Anti-biotics**, Anti-viral drugs, … – Infectious disease, ex) COVID-19 치료제

  → 기존의 항생제 저항성을 나타내는 bacteria의 출현등에 의해 계속해서 새로운 항생제의 발굴 needs 가 있음.

- Alzheimer, Pakinson, 유전질환, …

Preliminary

일반적인 신약개발 process

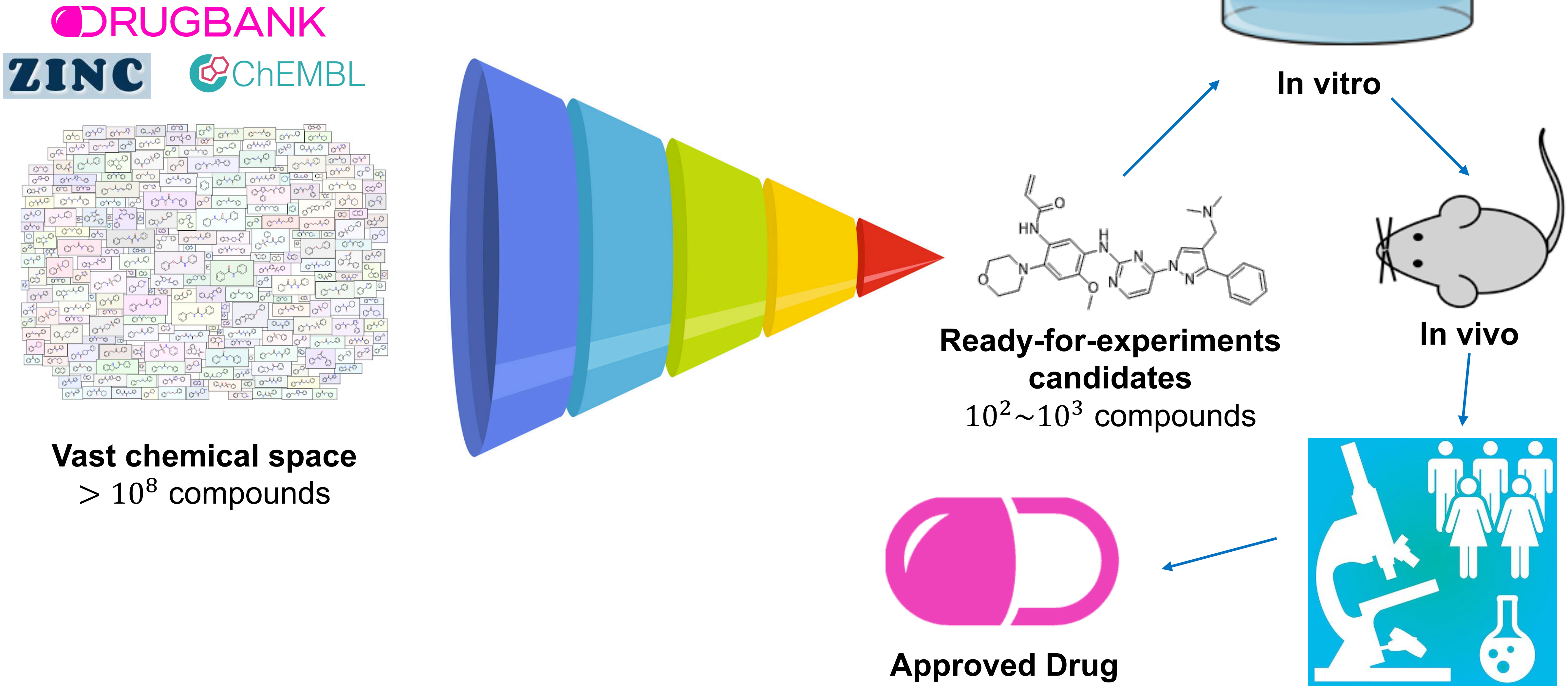


- Target identification: 어떤 protein/gene/… 을 제어할 것인가?
- ADME : bio-availability, 우리 몸에 투여해도 되는 약물인가?

AITRICS

Preliminary

Overview



DRUGBANK

ZINC

ChEMBL

Vast chemical space

$> 10^8$ compounds

Ready-for-experiments candidates

$10^2 \sim 10^3$ compounds

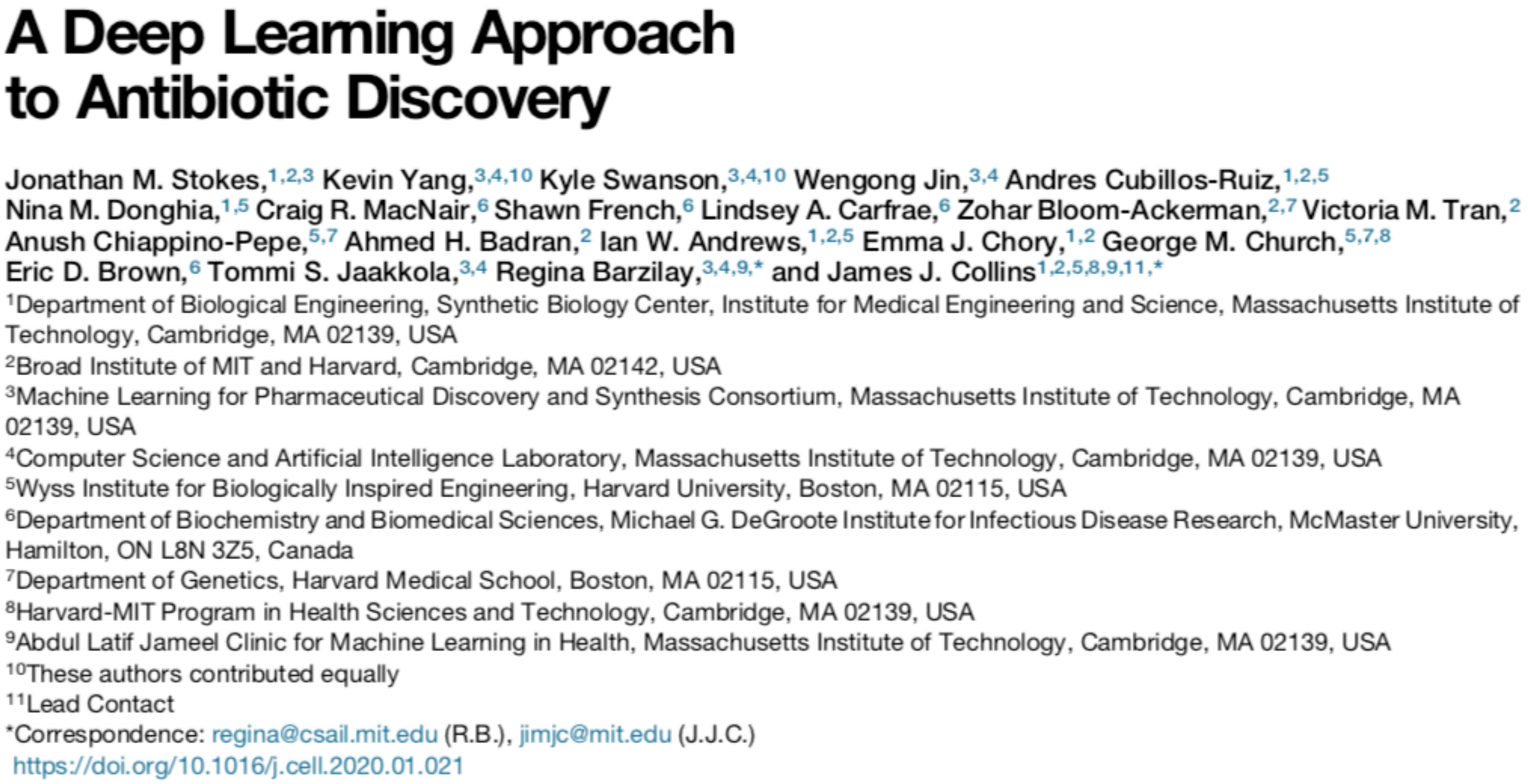
In vitro

In vivo

Clinical trials

Approved Drug

This work



A Deep Learning Approach to Antibiotic Discovery

Jonathan M. Stokes,[1,2,3] Kevin Yang,[3,4,10] Kyle Swanson,[3,4,10] Wengong Jin,[3,4] Andres Cubillos-Ruiz,[1,2,5] Nina M. Donghia,[1,5] Craig R. MacNair,[6] Shawn French,[6] Lindsey A. Carfrae,[6] Zohar Bloom-Ackerman,[2,7] Victoria M. Tran,[2] Anush Chiappino-Pepe,[5,7] Ahmed H. Badran,[2] Ian W. Andrews,[1,2,5] Emma J. Chory,[1,2] George M. Church,[5,7,8] Eric D. Brown,[6] Tommi S. Jaakkola,[3,4] Regina Barzilay,[3,4,9,*] and James J. Collins[1,2,5,8,9,11,*]

[1]Department of Biological Engineering, Synthetic Biology Center, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

[3]Machine Learning for Pharmaceutical Discovery and Synthesis Consortium, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[4]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[5]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA 02115, USA

[6]Department of Biochemistry and Biomedical Sciences, Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, ON L8N 3Z5, Canada

[7]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

[8]Harvard-MIT Program in Health Sciences and Technology, Cambridge, MA 02139, USA

[9]Abdul Latif Jameel Clinic for Machine Learning in Health, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[10]These authors contributed equally

[11]Lead Contact

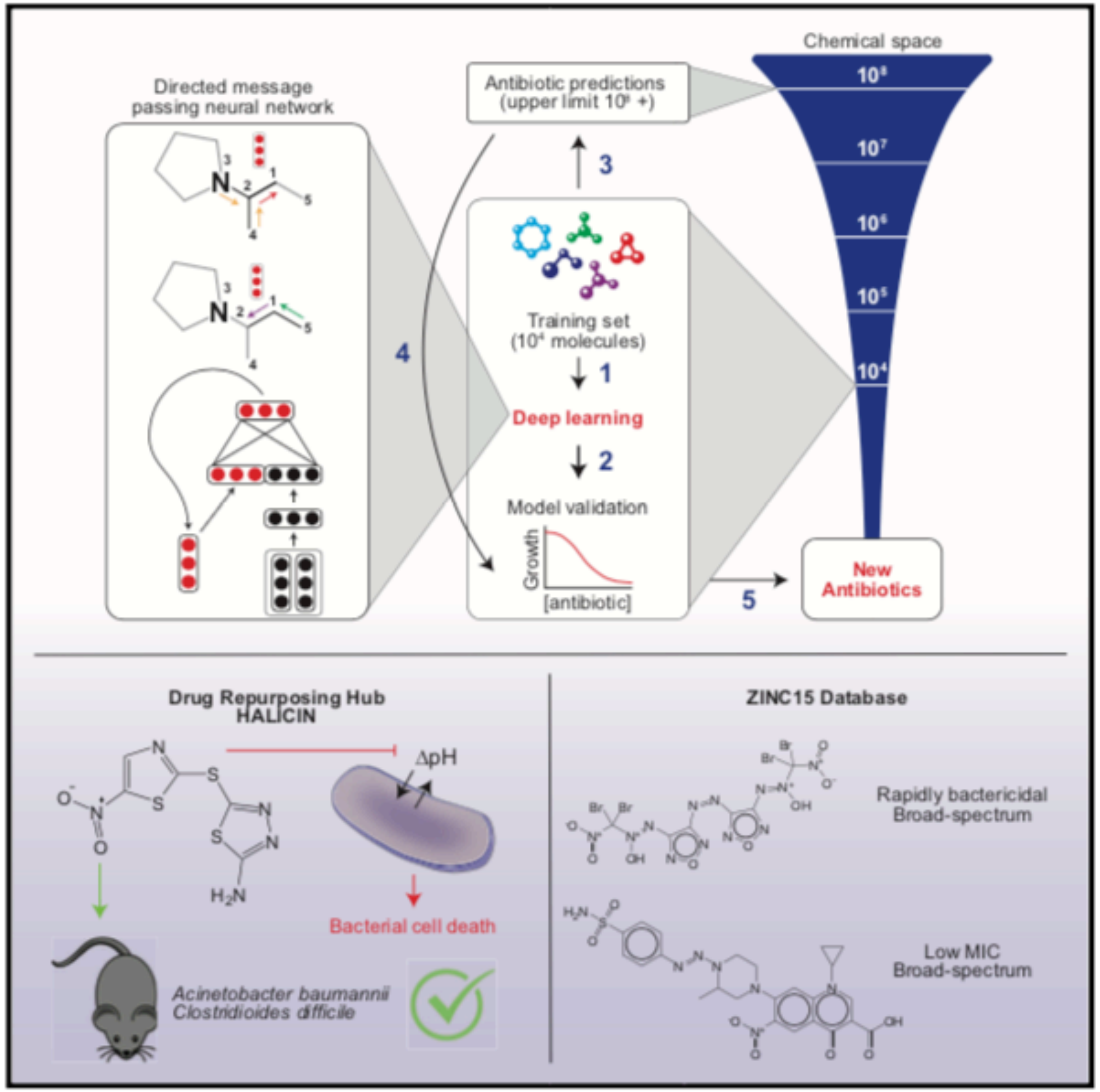
*Correspondence: regina@csail.mit.edu (R.B.), jimjc@mit.edu (J.J.C.)

https://doi.org/10.1016/j.cell.2020.01.021

Wengong Jin's presentation at NeurIPS 2019 Graph Representation Learning workshop
https://slideslive.com/38923996/representation-and-synthesis-of-molecular-graphs?ref=account-folder-42379-folders

This work

- 실험을 통해서 initial training data ”2,335개”를 얻음. (사족: 이것을 행할 수 있는 실행력과 준비력이 놀라움.)
- D-MPNN 이라는 GNN model 을 기본적으로 사용함.
- 여러 screening library 에 모델을 적용하여 후보군을 추려내고, emprical bio-assay 진행
- Halicin 및 그외의 두개의 물질은 broad spectrum of inhibition activity 를 보임.
- 또한 Empirical data를 model re-training 에 적용하여서 generalization ability를 높이려는 시도를 계속해서 함.



Procedure

- Initial *in vitro* screening of compounds listed as below:

  → US FDA-approved drug library (1,760) + natural products (800)

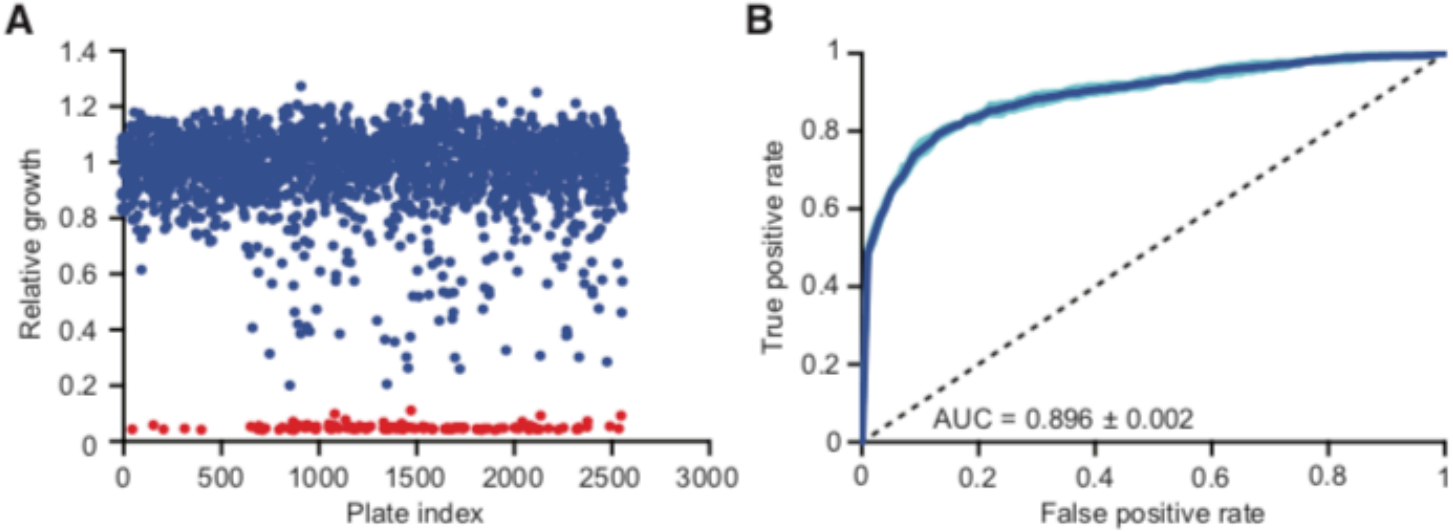
  → 2,335 unique compounds (duplicated compounds were removed)

  → 80% growth-inhibition as cut-off, 120 compounds were active and the others were inactive

- Training & Validation

  → Initial screening 으로부터 얻은 data를 이용해서 model training

  → Test set AUROC = 0.896



Procedure

- Virtual screening

  → Applying the model on “Drug Repurposing Hub” library (6,111 molecules)

  → the probability of displaying growth inhibition (final output from the ensemble model) 을 이용하여 ranking

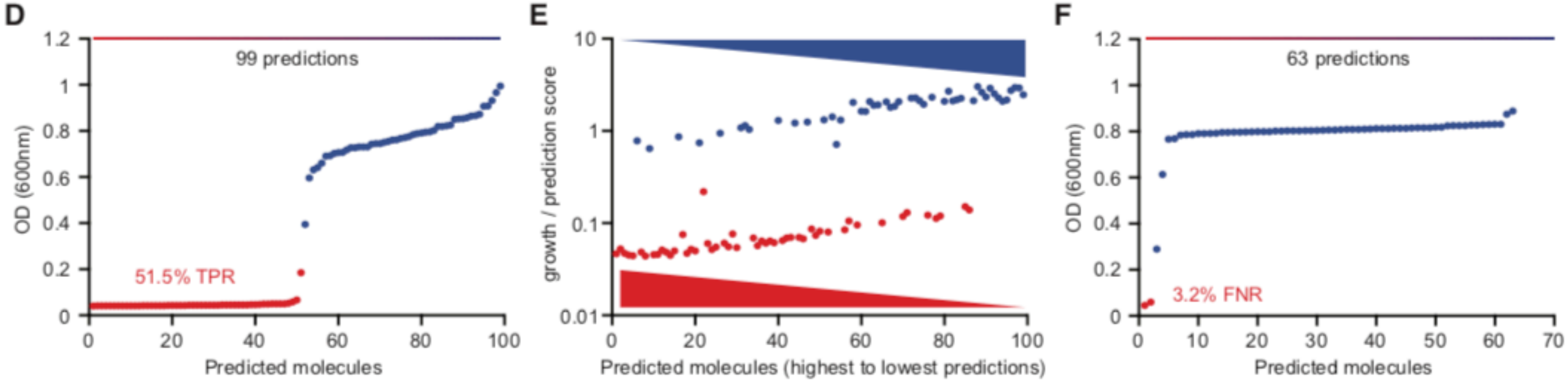
  → “99 molecules that were strongly predicted to be active” 에 대해서 empirical assay study 진행

- Empirical study

  → 99개 compounds 중 51개 compounds 가 True Positive (OD600이 0.2 이하), 나머지는 True Negative

  → Prediction score 가 높을수록 growth inhibition 이 active 일 확률이 높음

  → Lowest 63개에 대해서는 2개만이 active compounds



Procedure

- Prioritization

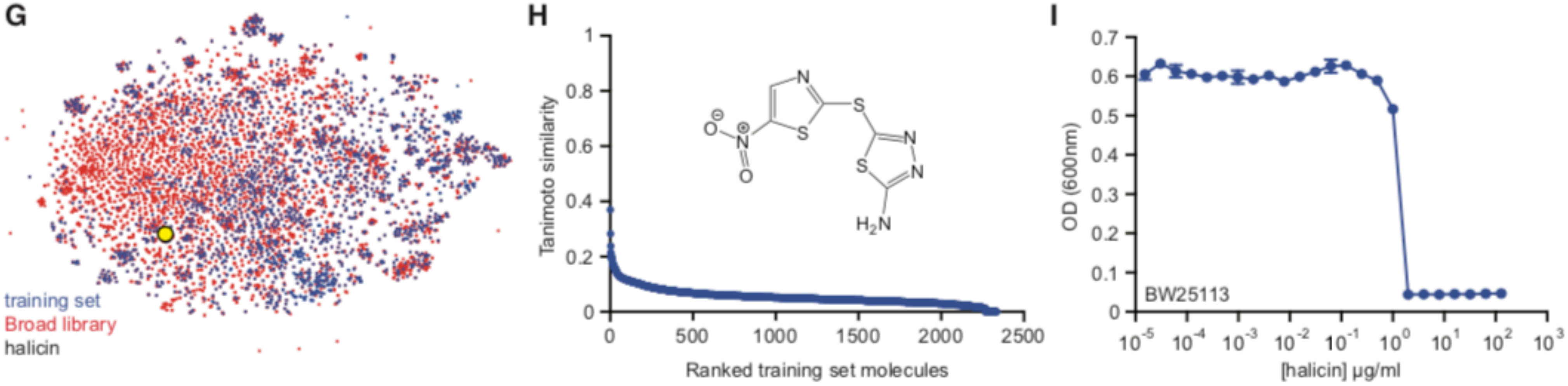
  → Clinical phase of investigation: in pre-clinical or Phase I/II/III studies

  → Structural similarity: Low structural similarity to training set molecules

  → Toxicity predicted by the model trained on the ClinTox database (to remove potentially toxic compounds)

  → 이 모든 조건을 만족하는 compound among 51 active compounds

  : c-Jun N-terminal kinase inhibitor SU3327, renamed as “Halicin”, structurally most similar to a family of nitro-containing antiparasitic compounds



Procedure

- Expand to vast chemical libraries – retrain models multiple steps with empirically observed data and infer on Wuxi library and ZINC-15 database.

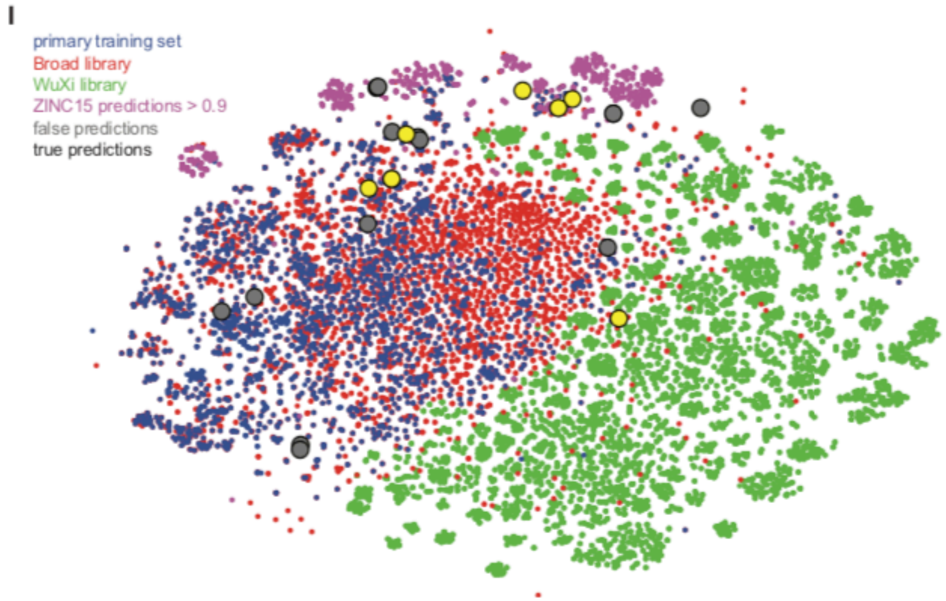
  → 9,997 molecules from Wuxi Anti-tuberculosis (결핵) library at Broad Institute 에 model을 적용

  → Highest prediction score = 0.37 (Drug Repurposing Hub에 적용했을 때 가장 큰 score = 0.97)

  → Top 200 highest score compounds & Top 100 lowest score compounds 에 대해 in vitro assay를 진행

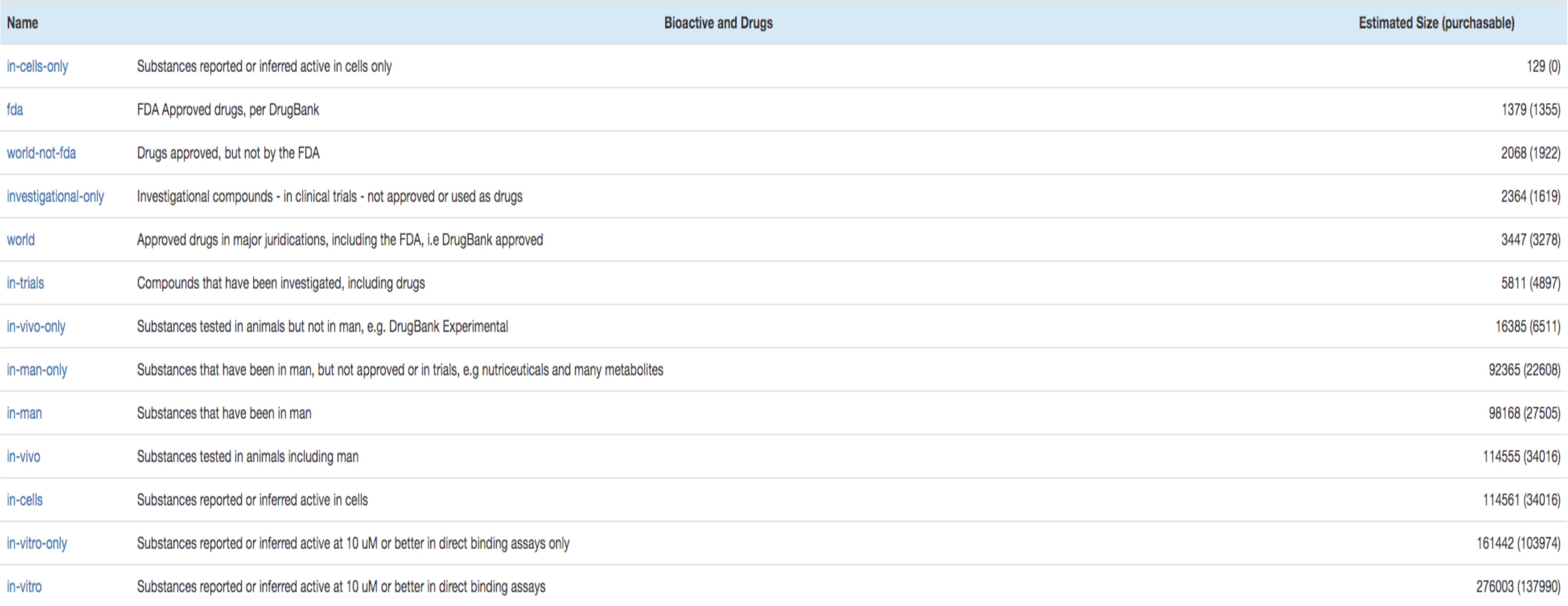
  → 전부다 inactive compounds

  → 추가적으로 확보된 실험데이터들 (Drug Repurposing Hub 99개, Wuxi 300개) 이용해서 model re-training



Procedure

- ZINC-15 database: https://zinc.docking.org/substances/subsets/



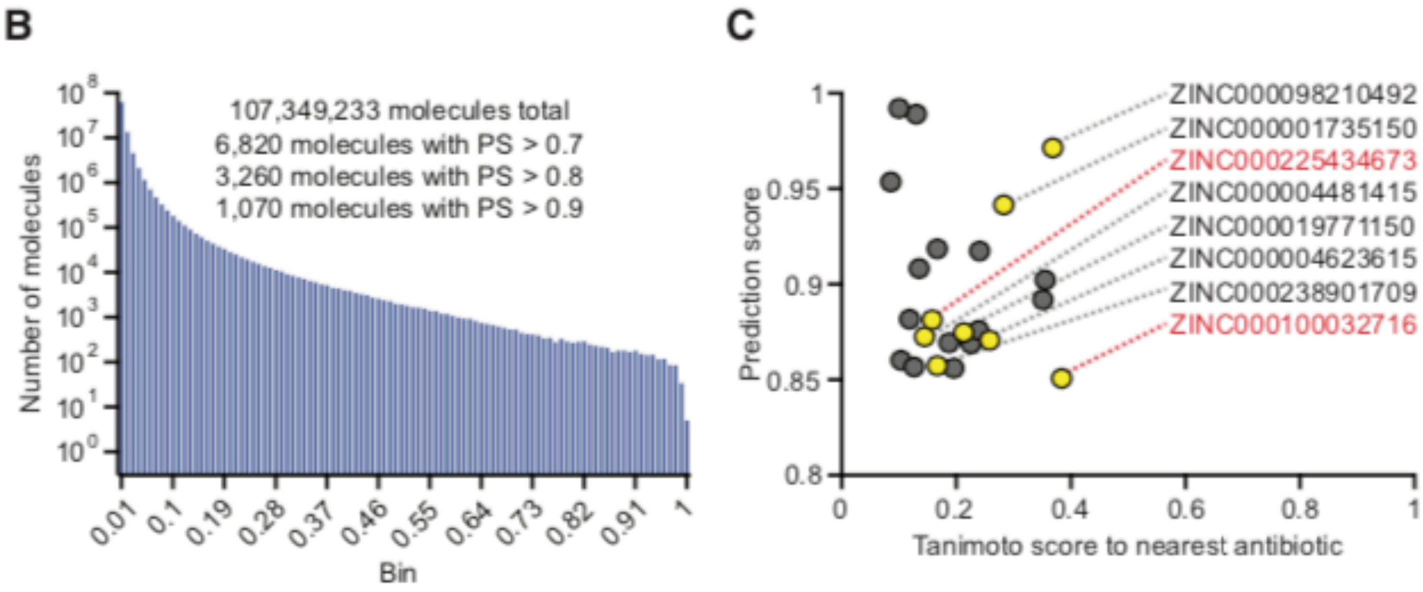
Name	Bioactive and Drugs	Estimated Size (purchasable)
in-cells-only	Substances reported or inferred active in cells only	129 (0)
fda	FDA Approved drugs, per DrugBank	1379 (1355)
world-not-fda	Drugs approved, but not by the FDA	2068 (1922)
investigational-only	Investigational compounds - in clinical trials - not approved or used as drugs	2364 (1619)
world	Approved drugs in major juridications, including the FDA, i.e DrugBank approved	3447 (3278)
in-trials	Compounds that have been investigated, including drugs	5811 (4897)
in-vivo-only	Substances tested in animals but not in man, e.g. DrugBank Experimental	16385 (6511)
in-man-only	Substances that have been in man, but not approved or in trials, e.g nutriceuticals and many metabolites	92365 (22608)
in-man	Substances that have been in man	98168 (27505)
in-vivo	Substances tested in animals including man	114555 (34016)
in-cells	Substances reported or inferred active in cells	114561 (34016)
in-vitro-only	Substances reported or inferred active at 10 uM or better in direct binding assays only	161442 (103974)
in-vitro	Substances reported or inferred active at 10 uM or better in direct binding assays	276003 (137990)

Procedure

- Expand to vast chemical libraries – retrain models multiple steps with empirically observed data and infer on Wuxi library and ZINC-15 database.

→ ZINC-15 database 의 “antibiotics-like” tranche 에 대해 re-trained model 을 적용

→ “# of score > 0.7 = 6,820”, “# of score > 0.8 = 3,260”, “# of score > 0.9 = 1,070”



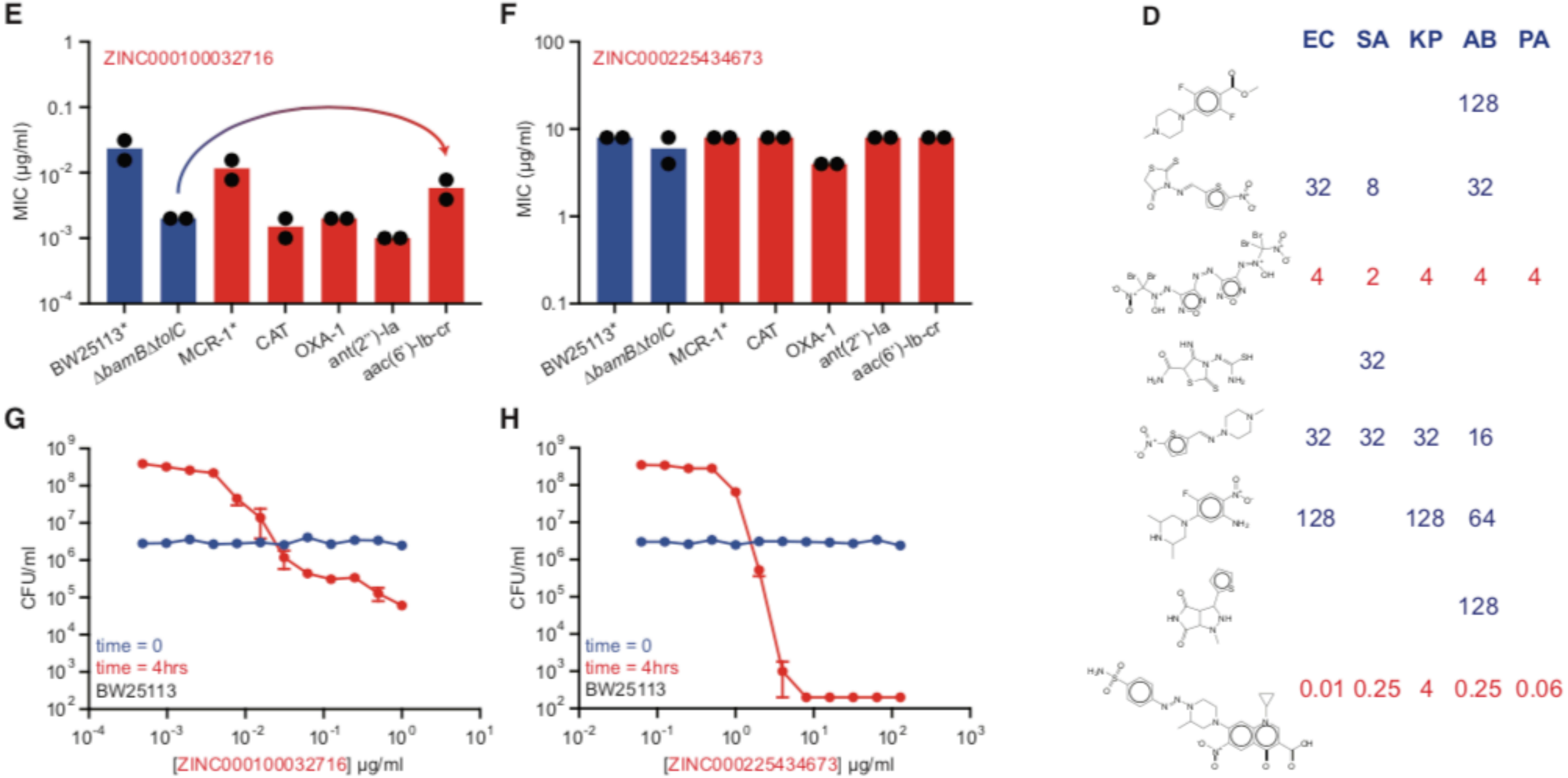
AITRICS

Procedure

- Expand to vast chemical libraries – retrain models multiple steps with empirically observed data and infer on Wuxi library and ZINC-15 database.

→ 앞에서 소개한 바와 같은 prioritization 을 진행

→ 최종적으로 23개 compounds 에 대해서 여러 세포주에 대한 growth inhibition assay 진행.



Authors’ message

"Machine learning is imperfect. Therefore, the success of DNN model-guided antibiotic discovery rests heavily on the coupling of these approaches to appropriate experimental designs"

- The assay design for training:

  ”what is the biological outcome that is desired after cells are exposed to compounds?”

  → They selected **growth inhibition as the biological property** on which they would gather training data.

  → Where their screen was largely mechanism-of-action agnostic, future **applications could incorporate phenotypic screening conditions** that enrich for molecules against **specific biological targets**.

Authors’ message

"Machine learning is imperfect. Therefore, the success of DNN model-guided antibiotic discovery rests heavily on the coupling of these approaches to appropriate experimental designs"

- The composition of the training data itself

  “what chemistry should be the model be trained?”

  → For *in vivo* application, training data must be sufficiently diverse

  → **Broadest structural variation** possible in the training phase **to maximize the probability of successful generalization in new chemical spaces**.

Authors’ message

"Machine learning is imperfect. Therefore, the success of DNN model-guided antibiotic discovery rests heavily on the coupling of these approaches to appropriate experimental designs"

- Prediction prioritization:

  “what is the most appropriate approach to selecting tens of molecules for follow-up investigation from thounds of strongly predicted compounds?”

  i) Given a high prediction score → Using the ensemble of models

  ii) Structurally unique relative to clinical antibiotics → Based on Tanimoto similarity analyses

  iii) Unlikely to display toxicity → Using the toxicity prediction models trained with the ClinTox dataset.

Authors' message

"Machine learning is imperfect. Therefore, the success of DNN model-guided antibiotic discovery rests heavily on the coupling of these approaches to appropriate experimental designs"

- Comments on using generative models for drug-discovery

  → Using generative model aims to find novel molecules that are even not included in screening libraries.

  → For experimental validation, most of them may be chemically synthesized.

  → Combining generative models with retrosynthesis ML models would be powerful.