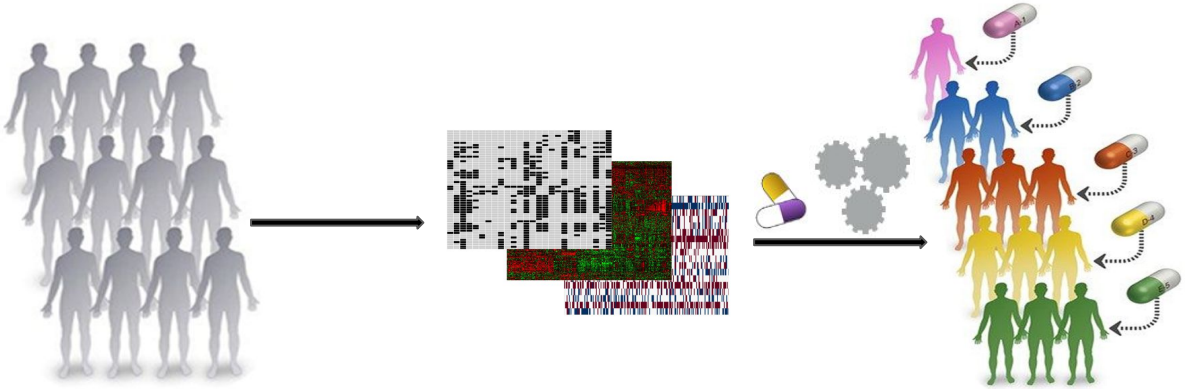
MOLI: Multi-Omics Late Integration with deep neural networks for drug response prediction

Hossein Sharifi-Noghabi, Olga Zolotareva, Colin C. Collins, and Martin Ester

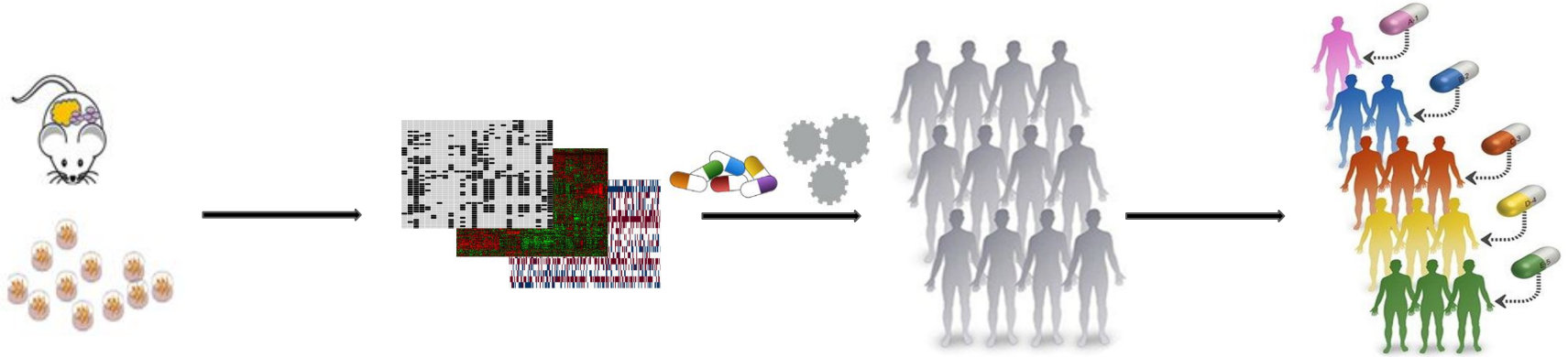
Simon Fraser University and Vancouver Prostate Centre

BASEL JULY 21–25 ISMB ECCB 2019

Motivation



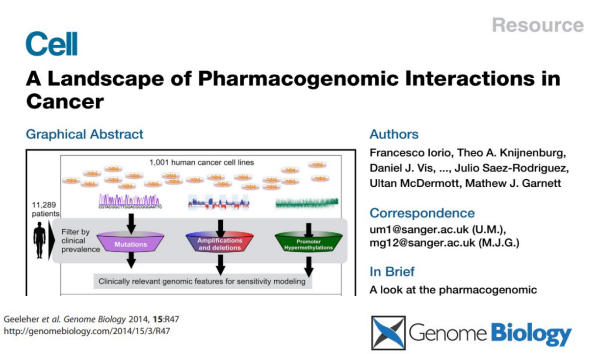
- Cannot treat patients with so many drugs
- Clinical trial data are either small or not publicly available



Motivation

- Gene expression data have been shown to be the best data type for drug response prediction.
- Recent studies suggest that adding other omics data types can improve the prediction performance.

→ How to integrate different data types???


Resource

Cell

A Landscape of Pharmacogenomic Interactions in Cancer

Graphical Abstract

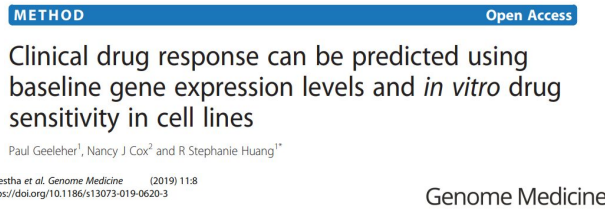
Authors
Francesco Iorio, Theo A. Knijnenburg, Daniel J. Vis, ..., Julio Saez-Rodriguez, Ultan McDermott, Mathew J. Garnett

Correspondence
um1@sanger.ac.uk (U.M.), mg12@sanger.ac.uk (M.J.G.)

In Brief
A look at the pharmacogenomic

Geeleher et al. Genome Biology 2014, 15:R47
http://genomebiology.com/2014/15/3/R47

Genome Biology

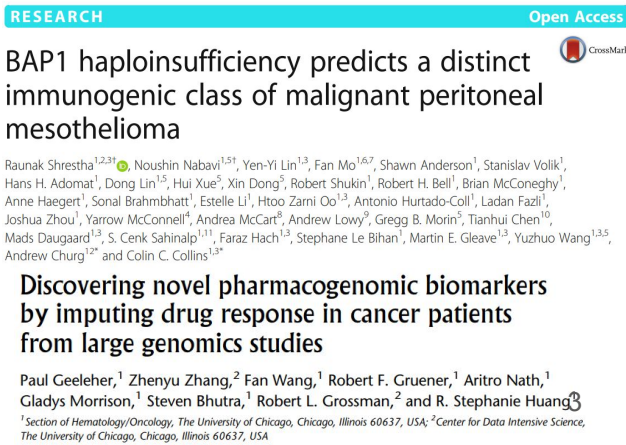

METHOD Open Access

Clinical drug response can be predicted using baseline gene expression levels and *in vitro* drug sensitivity in cell lines

Paul Geeleher[1], Nancy J Cox[2] and R Stephanie Huang[1*]

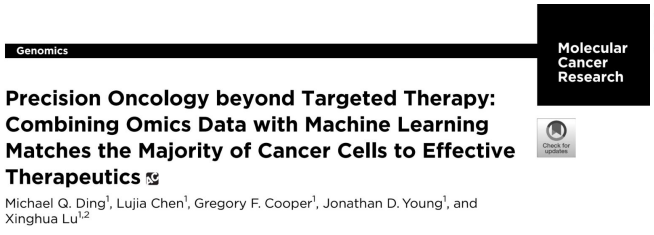
Shrestha et al. Genome Medicine (2019) 11:8
https://doi.org/10.1186/s13073-019-0620-3

Genome Medicine


RESEARCH Open Access

BAP1 haploinsufficiency predicts a distinct immunogenic class of malignant peritoneal mesothelioma

Raunak Shrestha[1,2,3†], Noushin Nabavi[1,5†], Yen-Yi Lin[1,3], Fan Mo[1,6,7], Shawn Anderson[1], Stanislav Volik[1], Hans H. Adomat[1], Dong Lin[1,5], Hui Xue[5], Xin Dong[5], Robert Shukin[1], Robert H. Bell[1], Brian McConeghy[1], Anne Haegert[1], Sonal Brahmbhatt[1], Estelle Li[1], Htoo Zarni Oo[1-3], Antonio Hurtado-Coll[1], Ladan Fazli[1], Joshua Zhou[1], Yarrow McConnell[4], Andrea McCart[8], Andrew Lowy[9], Gregg B. Morin[5], Tianhui Chen[10], Mads Daugaard[1,3], S. Cenk Sahinalp[1,11], Faraz Hach[1,3], Stephane Le Bihan[1], Martin E. Gleave[1,3], Yuzhuo Wang[1,3,5], Andrew Churg[1,2*] and Colin C. Collins[1,3*]


Genomics

Molecular Cancer Research

Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics

Michael Q. Ding[1], Lujia Chen[1], Gregory F. Cooper[1], Jonathan D. Young[1], and Xinghua Lu[1,2]

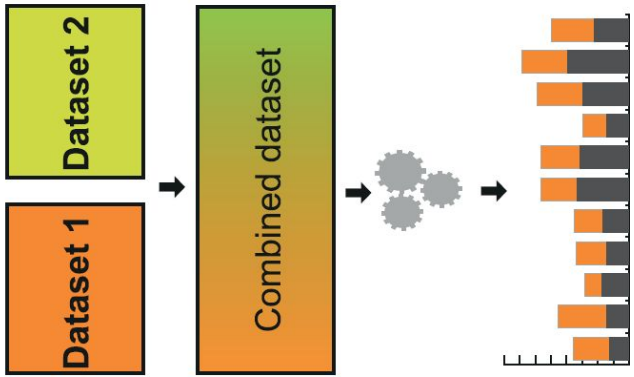
Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies

Paul Geeleher,[1] Zhenyu Zhang,[2] Fan Wang,[1] Robert F. Gruener,[1] Aritro Nath,[1] Gladys Morrison,[1] Steven Bhutra,[1] Robert L. Grossman,[2] and R. Stephanie Huang[1]

[1]Section of Hematology/Oncology, The University of Chicago, Chicago, Illinois 60637, USA; [2]Center for Data Intensive Science, The University of Chicago, Chicago, Illinois 60637, USA

Omics integration

Early integration



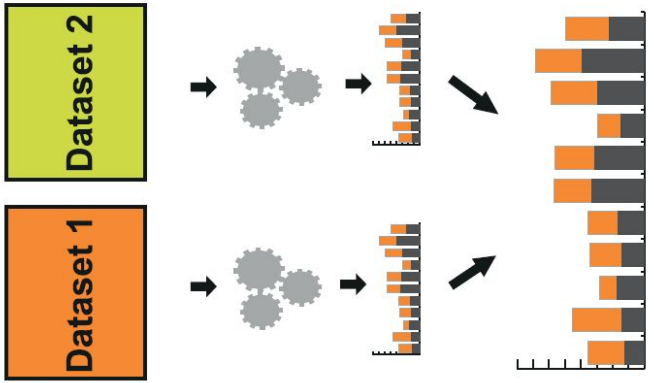
Genomics

Molecular Cancer Research

Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics

Michael Q. Ding[1], Lujia Chen[1], Gregory F. Cooper[1], Jonathan D. Young[1], and Xinghua Lu[1,2]

Late integration



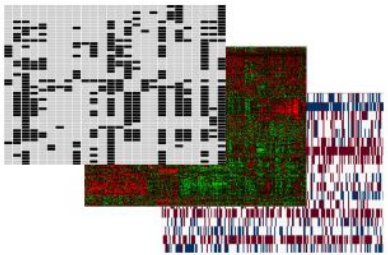
MOLI: Multi-Omics Late Integration with deep neural networks for drug response prediction

Hossein Sharifi-Noghabi[1,3], Olga Zolotareva[2], Colin C. Collins[3,4,*], and Martin Ester[1,3,*]

Figures are from Zitnik, et al. 2019 Information Fusion

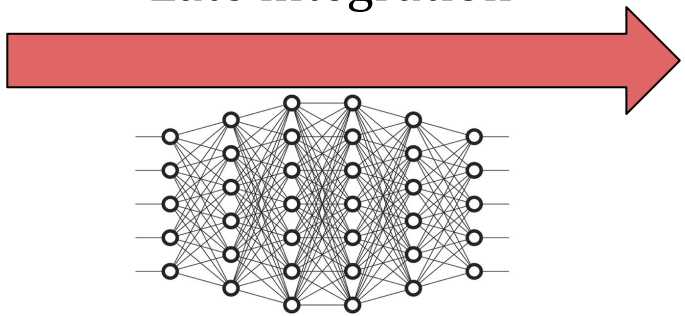
Goal

Given:

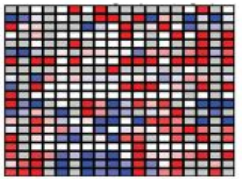


Multi-omics data

Late integration



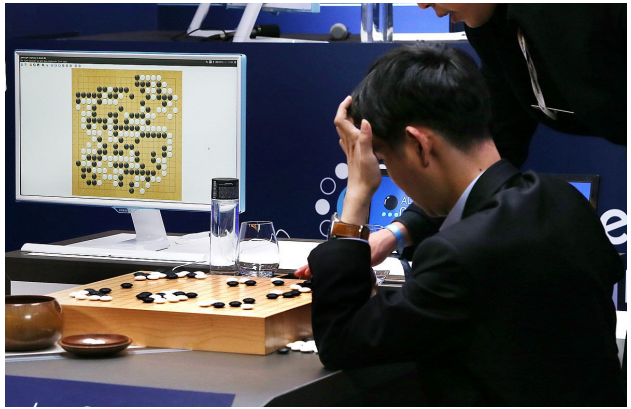
Output:



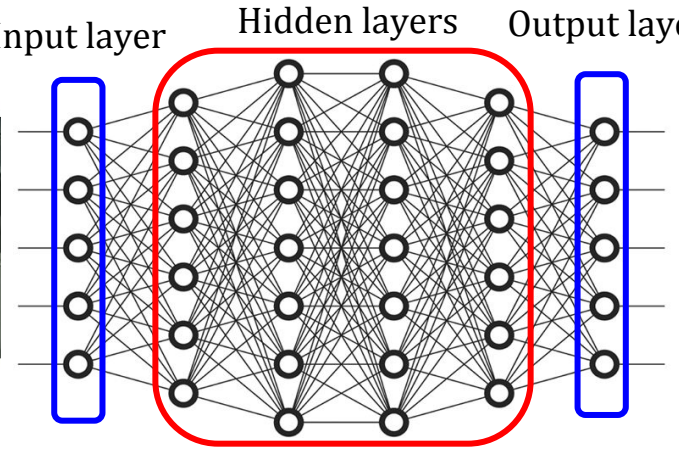
Drug response (binarized IC50)

Deep Neural Networks

- Computer vision
- Natural language processing
- Robotics
- Gaming



From:
https://www.technologyreview.com/s/604273/finding-solace-in-defeat-by-artificial-intelligence/



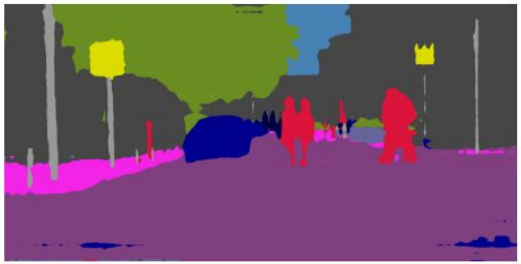
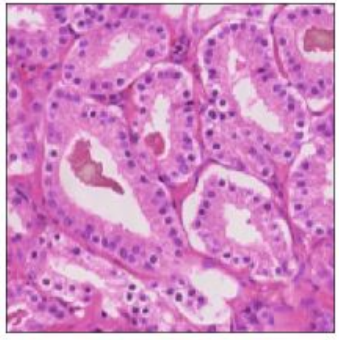
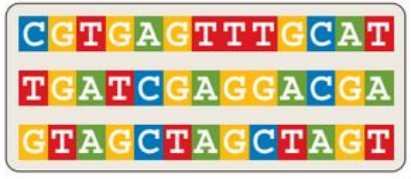
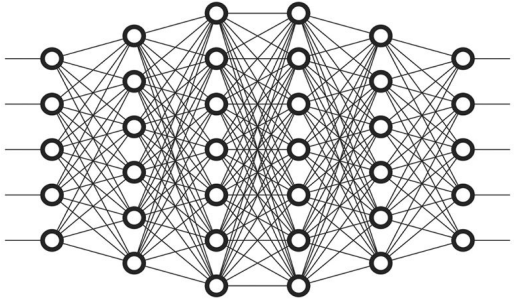


Figure is from Tzeng et al., “Adversarial Discriminative Domain Adaptation” CVPR 2017

Genomics and medicine







A new era: artificial intelligence and machine learning in prostate cancer

S. Larry Goldenberg[1], Guy Nir[1,2] and Septimiu E. Salcudean[1,2]*

Abstract | Artificial intelligence (AI) — the ability of a machine to perform cognitive tasks to achieve a particular goal based on provided data — is revolutionizing and reshaping our health-care systems. The current availability of ever-increasing computational power, highly developed pattern recognition algorithms and advanced image processing software working at very high speeds has led to the emergence of computer-based systems that are trained to perform complex tasks in bioinformatics, medical imaging and medical robotics. Accessibility to 'big data' enables the 'cognitive' computer to scan billions of bits of unstructured information, extract the relevant information and recognize complex patterns with increasing confidence. Computer-based decision-support systems based on machine learning (ML) have the potential to revolutionize medicine by performing complex tasks that are currently assigned to specialists to improve diagnostic accuracy, increase efficiency of throughputs, improve clinical workflow, decrease human resource costs and improve treatment choices. These characteristics could be especially helpful in the management of prostate cancer, with growing applications in diagnostic imaging, surgical interventions, skills training and assessment, digital pathology and genomics. Medicine must adapt to this changing world, and urologists, oncologists, radiologists and pathologists, as high-volume users of imaging and pathology, need to understand this burgeoning science and acknowledge that the development of highly accurate AI-based

REVIEWS

Deep learning: new computational modelling techniques for genomics

Gökcen Eraslan[1,2,5], Žiga Avsec[3,5], Julien Gagneur[3] and Fabian J. Theis[1,2,4]**

Abstract | As a data-driven science, genomics largely utilizes machine learning to capture dependencies in data and derive novel biological hypotheses. However, the ability to extract new insights from the exponentially increasing volume of genomics data requires more expressive machine learning models. By effectively leveraging large data sets, deep learning has transformed fields such as computer vision and natural language processing. Now, it is becoming the method of choice for many genomics modelling tasks, including predicting the impact of genetic variation on gene regulatory mechanisms such as DNA accessibility and splicing.

REVIEW ARTICLE | FOCUS
https://doi.org/10.1038/s41591-018-0300-7
nature medicine

High-performance medicine: the convergence of human and artificial intelligence

Eric J. Topol

The use of artificial intelligence, and the deep-learning subtype in particular, has been enabled by the use of labeled big data, along with markedly enhanced computing power and cloud storage, across all sectors. In medicine, this is beginning to have an impact at three levels: for clinicians, predominantly via rapid, accurate image interpretation; for health systems, by improving workflow and the potential for reducing medical errors; and for patients, by enabling them to process their own data to promote health. The current limitations, including bias, privacy and security, and lack of transparency, along with the future directions of these applications will be discussed in this article. Over time, marked improvements in accuracy, productivity, and workflow will likely be actualized, but whether that will be used to improve the patient–doctor relationship or facilitate its erosion remains to be seen.

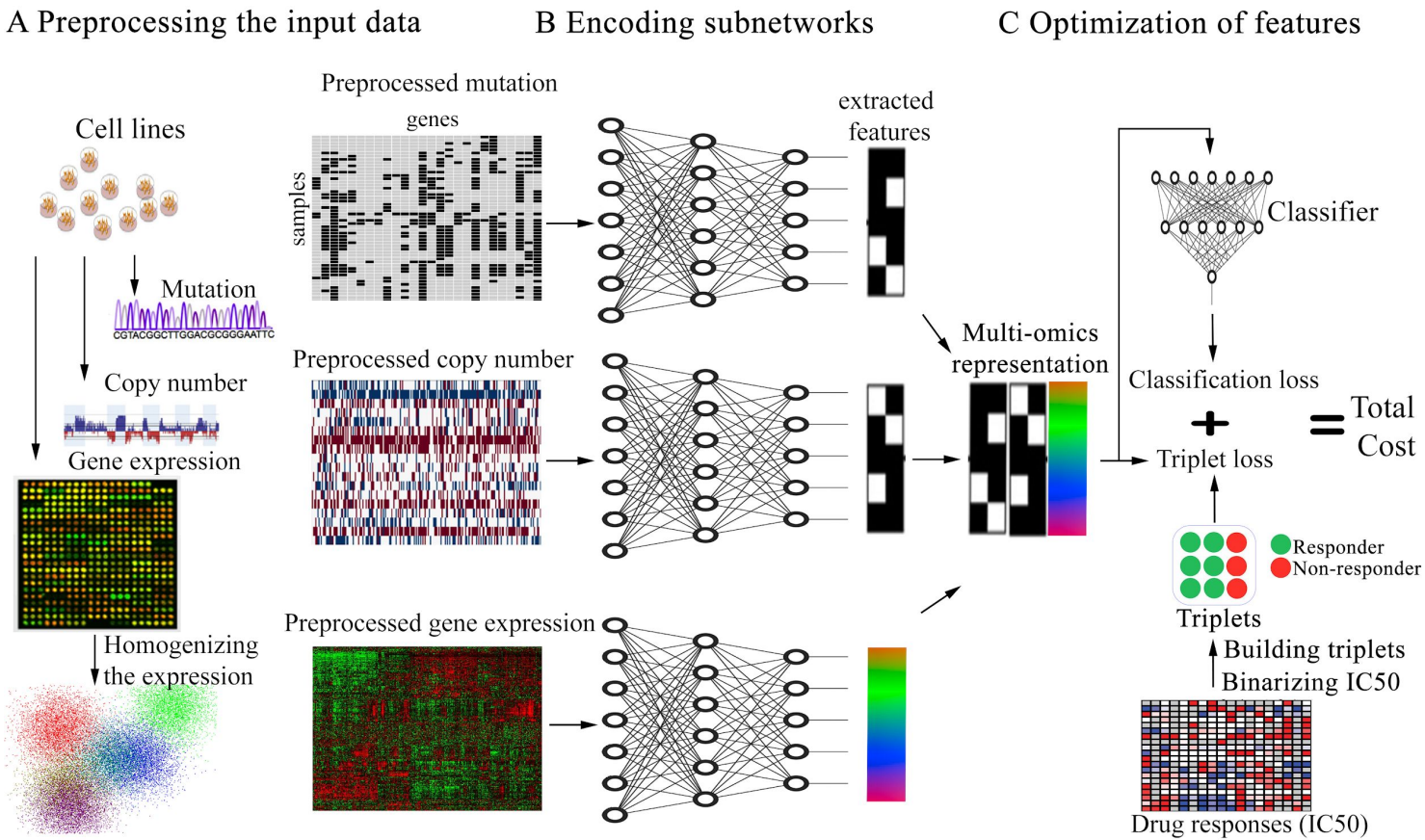
INTERFACE

rsif.royalsocietypublishing.org

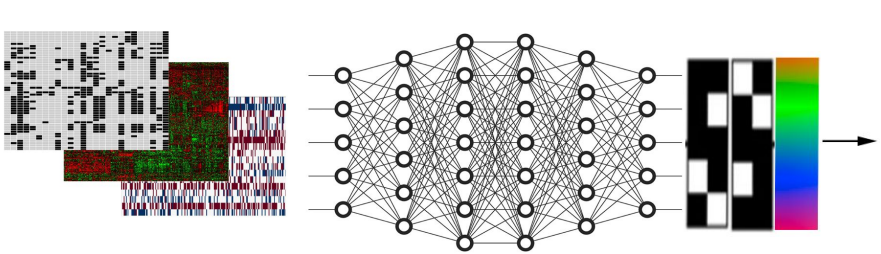
Opportunities and obstacles for deep learning in biology and medicine

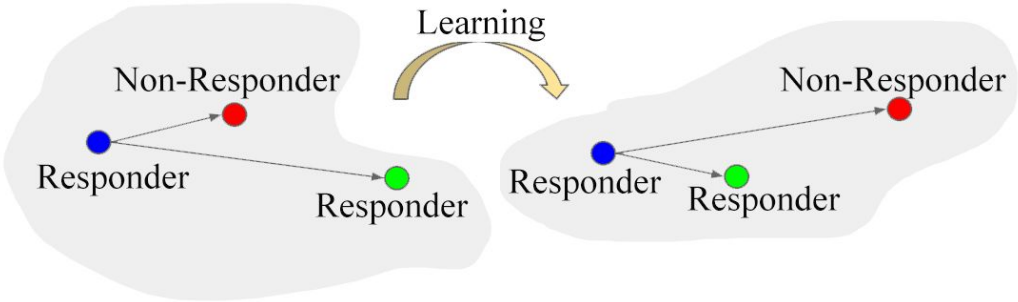
Travers Ching[1,†], Daniel S. Himmelstein[2], Brett K. Beaulieu-Jones[3], Alexandr A. Kalinin[4], Brian T. Do[5], Gregory P. Way[2], Enrico Ferrero[6],

MOLI: **M**ulti-**O**mics **L**ate **I**ntegration



MOLI: Triplet Loss function





$$d(f_{Anchor}, f_{Positive}) \leq d(f_{Anchor}, f_{Negative}),$$

$$d(f_{Anchor}, f_{Positive}) - d(f_{Anchor}, f_{Negative}) \leq 0,$$

$$d(f_{Anchor}, f_{Positive}) - d(f_{Anchor}, f_{Negative}) + \xi \leq 0,$$

$$L_{Triplet}^{\ i} = max[d(f_{Anchor}^{i}, f_{Positive}^{i}) - d(f_{Anchor}^{i}, f_{Negative}^{i}) + \xi, 0],$$

$$L_{Triplet} = \sum_{i=1}^{T} L_{Triplet}^{\ i}.$$

FaceNet: A Unified Embedding for Face Recognition and Clustering

Florian Schroff
fschroff@google.com
Google Inc.

Dmitry Kalenichenko
dkalenichenko@google.com
Google Inc.

James Philbin
jphilbin@google.com
Google Inc.

Abstract

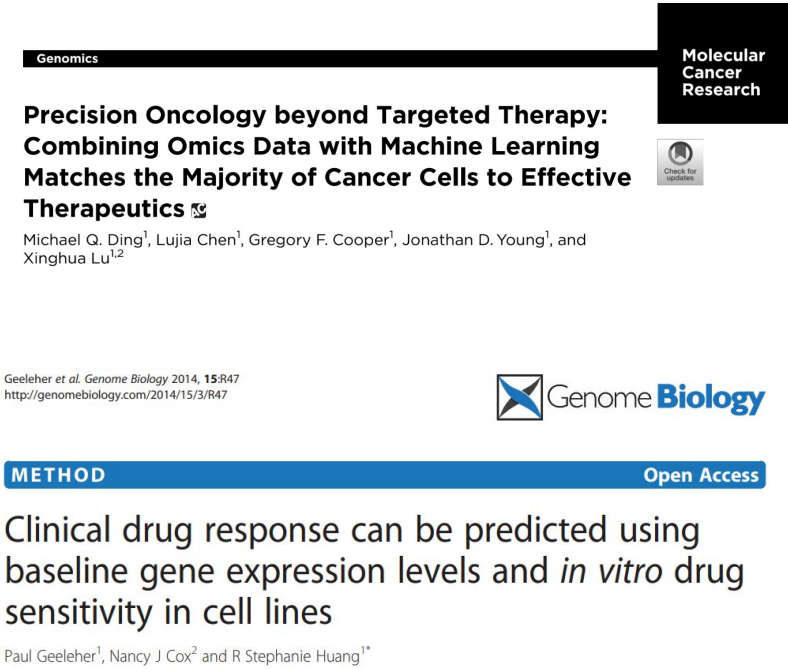
Despite significant recent advances in the field of face

Questions

1. Does MOLI outperform single-omics and early integration methods in terms of prediction AUROC?
2. Does MOLI's performance improve by including more drugs in its training data?
3. Does the response predicted by MOLI have associations with the target of a drug (for the targeted drugs)?

Baselines

- Early integration
  - Deep neural networks (Ding et al. 2018)
  - Non-negative matrix factorization
- Single-omics (gene expression)
  - Regression-based (Geeleher et al. 2014)
  - Feed forward neural network



Genomics

Molecular Cancer Research

Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics

Michael Q. Ding[1], Lujia Chen[1], Gregory F. Cooper[1], Jonathan D. Young[1], and Xinghua Lu[1,2]

Geeleher *et al. Genome Biology* 2014, **15**:R47
http://genomebiology.com/2014/15/3/R47

Genome Biology

METHOD Open Access

Clinical drug response can be predicted using baseline gene expression levels and *in vitro* drug sensitivity in cell lines

Paul Geeleher[1], Nancy J Cox[2] and R Stephanie Huang[1*]

Datasets

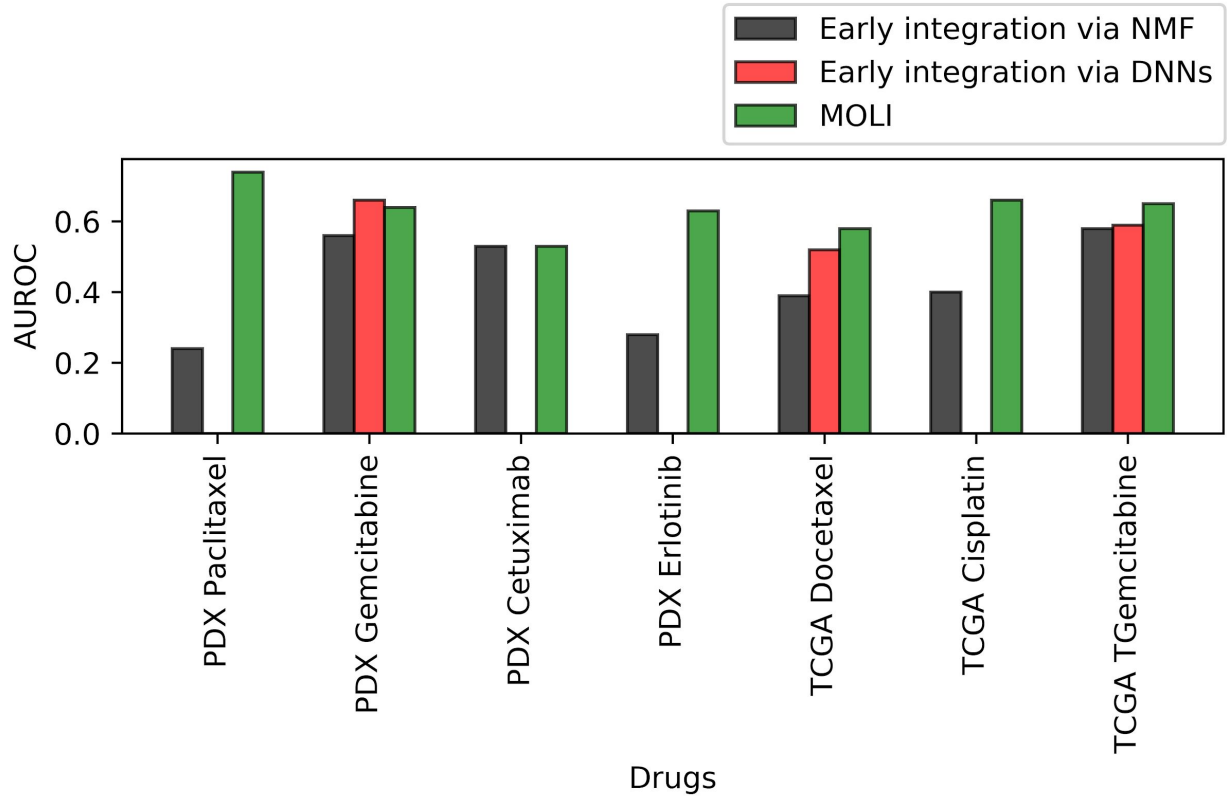
We use three main resources:

- ❖ Before treatment genomic data and after treatment response data
- Cell line data for training
  - ~1000 cell lines with multi-omics data screened with 265 drugs (Iorio et al., 2016 Cell)
- Pre-clinical data for external validation
  - ~400 PDX models with multi-omics data screened with 34 drugs (Gao et al., 2015 Nature Medicine)
- Clinical data for external validation
  - TCGA patients with the drug response available in their records

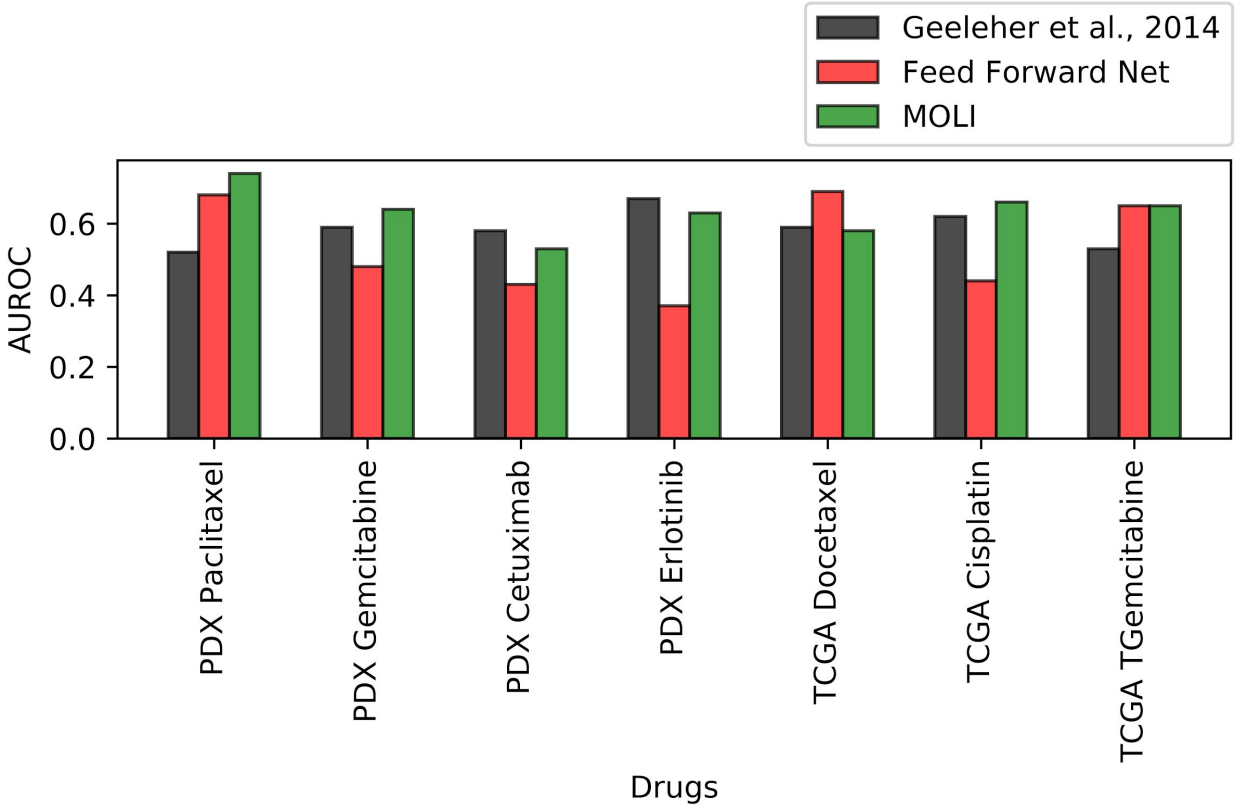
List of drugs:

- Paclitaxel
- Gemcitabine
- Erlotinib
- Cetuximab
- Cisplatin
- Docetaxel

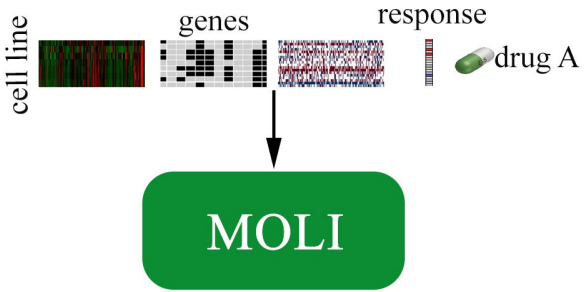
MOLI outperforms early integration baselines



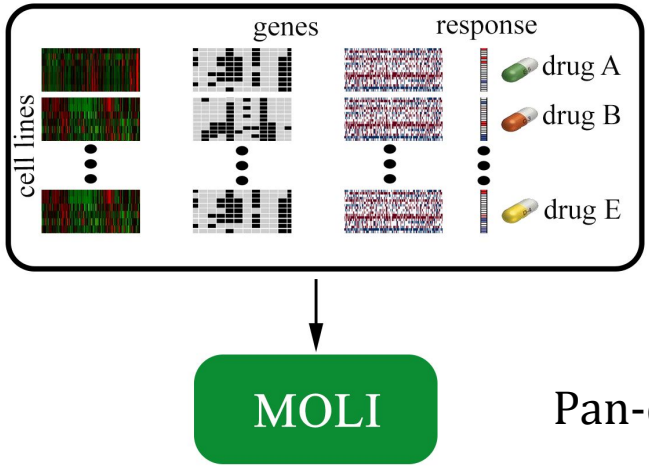
MOLI outperforms single-omics baselines



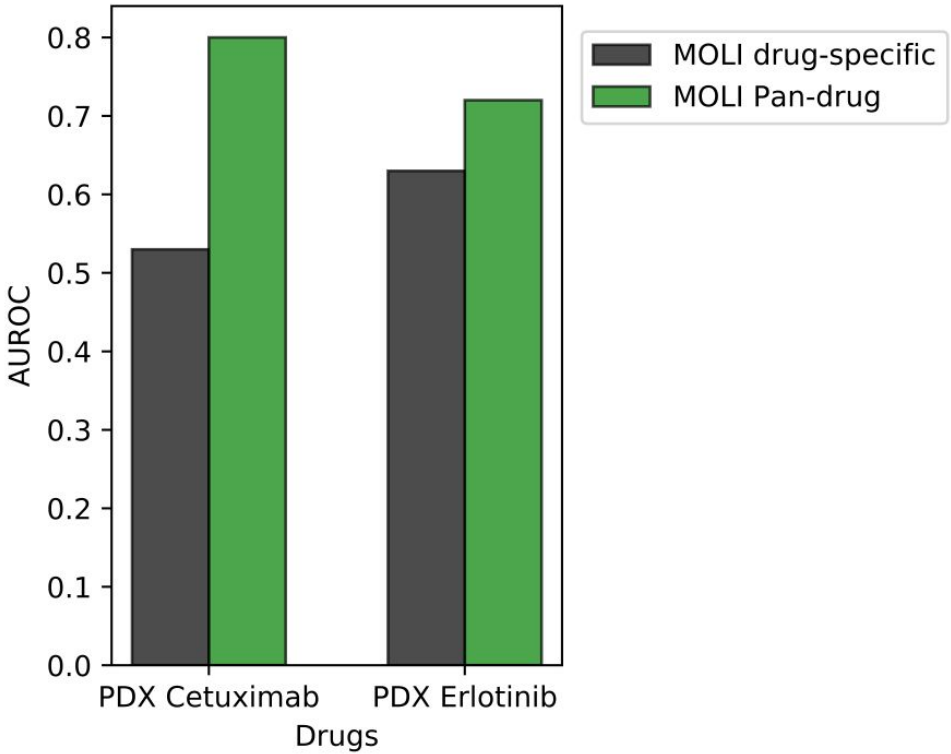
Pan-drug training data outperforms drug-specific



Drug specific

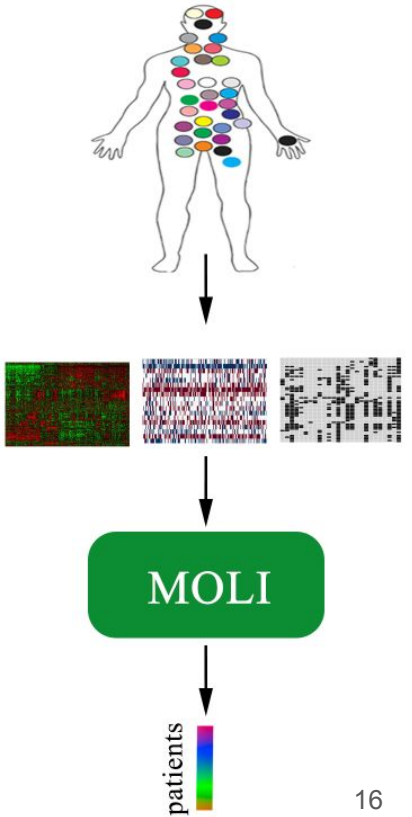


Pan-drug



Predict response for TCGA patients with no treatment

- **Step 1**: Apply the trained MOLI to TCGA cohorts separately and get the predicted responses
- **Step 2**: Select EGFR genes from REACTOME
- **Step 3**: Fit a multiple linear regression to the level of the expressions of the EGFR genes and the responses predicted by MOLI
- We found significant associations in:
  - Prostate cancer
  - Kidney cancer
  - Breast cancer
  - Lung cancer (treatment for 70% of the patients in US, Li et al., 2019 Plos one)



Summary

- Proposed the first method for multi-omics late integration with deep neural networks
- Employed triplet loss function in multi-omics late integration for drug response prediction
- Introduced pan-drug training data based on transfer learning for the targeted drugs
- Obtained better performance compared to the state-of-the-art methods

Github: https://github.com/hosseinshn/MOLI

Future direction

- Domain adaptation between cell lines, PDX, and patients data
- Incorporating domain expert/biological knowledge

CAMDA Thursday, July 25th

12:20 PM-12:40 PM

Proceedings Presentation: PRECISE: A domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors

Soufiane Mourragui, Delft University of Technology and the Netherlands Cancer Institute, Netherlands
Marco Loog, TU Delft and University of Copenhagen, Netherlands
Mark van de Wiel, VUmc Amsterdam, Netherlands
Marcel Reinders, TU Delft and Leiden University Medical Center, Netherlands
Lodewyk Wessels, The Netherlands Cancer Institute, Netherlands

Presentation Overview: Show

PharmacoDB

MLSCB Wednesday, July 24th

12:20 PM-12:40 PM

DrugCell: A visible neural network to guide precision medicine

Samson Fong, University of California San Diego, United States
Trey Ideker, Department of Medicine, University of California, San Diego, United States
Brent Kuenzi, University of California San Diego, United States
Jisoo Park, University of California San Diego, United States
Jason Kreisberg, University of California San Diego, United States

Presentation Overview: Show

Tutorial PM5: Biomarker discovery and machine learning in large pharmacogenomics datasets

Room: Kairo 1/2 (Ground Floor)

Sunday, July 21, 2:00 pm - 6:00 pm

Presenters

Arvind Singh Mer, Princess Margaret Cancer Center, University of Toronto, Canada
Zhaleh Safikhani, Princess Margaret Cancer Center, University of Toronto, Canada
Petr Smirnov, Princess Margaret Cancer Center,Vector Institute, University of Toronto, Canada
Benjamin Haibe-Kains, Princess Margaret Cancer Center,Vector Institute,Ontario Institute for Cancer Research, University of Toronto, Canada

We are hiring!

Our lab at SFU is looking for highly motivated and curious postdocs interested in method development for different biological problems!

Please contact Prof. Martin Ester:

Email: ester@sfu.ca

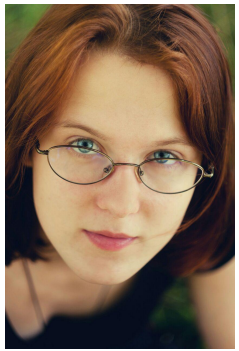
Acknowledgement

Dr. Ester's lab (SFU)
Martin Ester
Sahand Khakabi
Mehrdad Mansouri
Raquel Aoki
Oliver Snow
Shuman Peng
Qingyuan Feng
Ali Arab
Jialin Lu

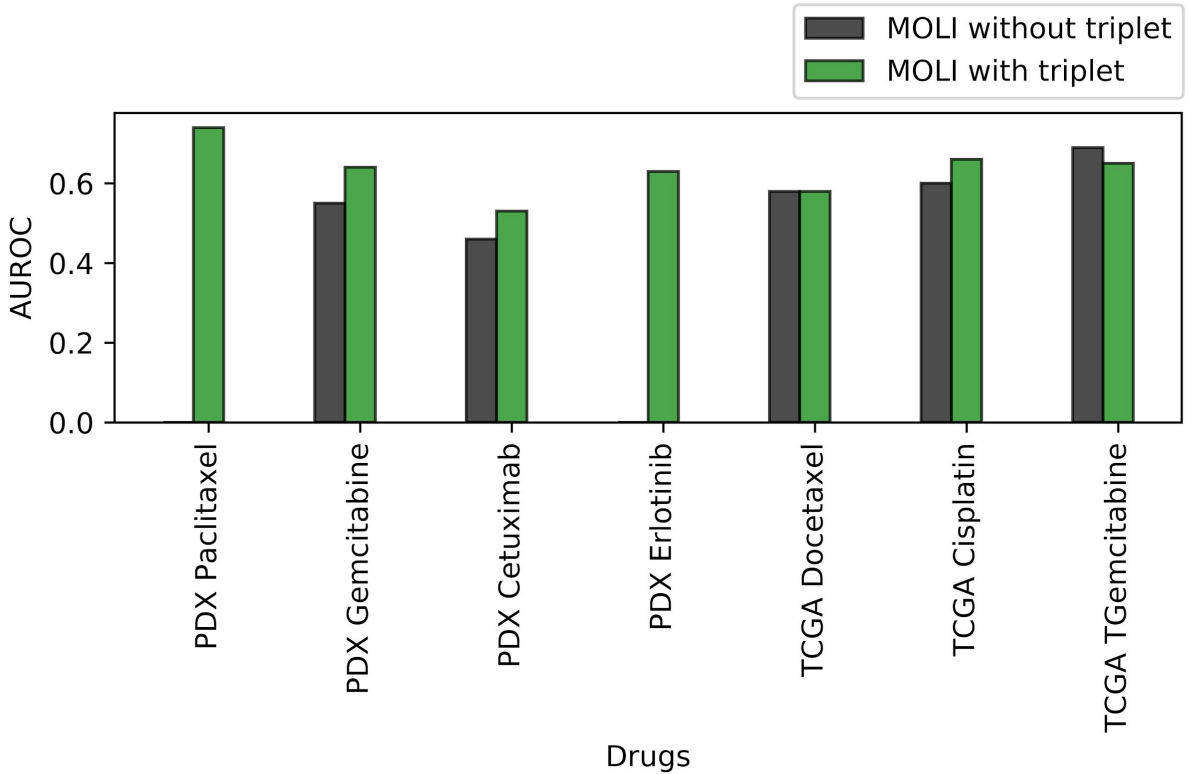
Dr. Collin's lab (VPC-UBC)
Colin C. Collins
Stephane Le Bihan
Stanislav Volik
Yen-Yi Lin
Raunak Shrestha
Shawn Anderson
Anne Haegert
Robert Bell

Dr. Hach's lab (VPC-UBC)
Hossein Asghari
Baraa Orabi

University of Bielefeld
Olga Zolotareva



Triplet loss performance



Precision-Recall

Table S4 Performance comparison in terms of the Area Under Precision-Recall Curve

Method/Drug	PDX-Paclitaxel	PDX-Gemcitabine	PDX-Cetuximab	PDX-Erlotinib	TCGA-Docetaxel	TCGA-Cisplatin	TCGA-Gemcitabine
Random classifier P/(P+N)	0.12	0.28	0.08	0.14	0.5	0.91	0.37
Geeleher et al. 2014	0.1	0.28	0.06	0.11	**0.51**	0.85	0.38
Early Integration via NMF	0.21	0.35	0.07	0.28	**0.51**	**0.93**	0.37
Early Integration via DNNs	NSC	0.35	NSC	NSC	0.45	NSC	**0.46**
MOLI complete	**0.24**	**0.49**	0.11	**0.33**	0.49	**0.93**	0.45
MOLI pan-drug	NA	NA	**0.2**	0.28	NA	NA	NA

P: number of positive cases; N: number of negative cases; NMF: non-negative matrix factorization; DNNs: Deep Neural Networks