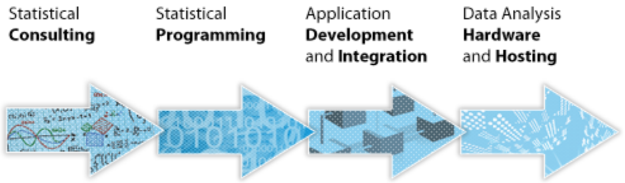
Phenotypic deconvolution: the next frontier in pharma

Marvin Steijaert (Open Analytics NV) -
Vladimir Chupakhin (Janssen Pharmaceutica) -
Hugo Ceulemans (Janssen Pharmaceutica) -
Joerg Wegner (Janssen Pharmaceutica) -

July 2, 2015

About me and Open Analytics NV

- MSc in Biomedical Engineering
- PhD in Systems Biology / Computational Biology
- Consultant at Open Analytics
- marvin.steijaert@openanalytics.eu



Statistical **Consulting**

Statistical **Programming**

Application **Development** and **Integration**

Data Analysis **Hardware** and **Hosting**

Architect

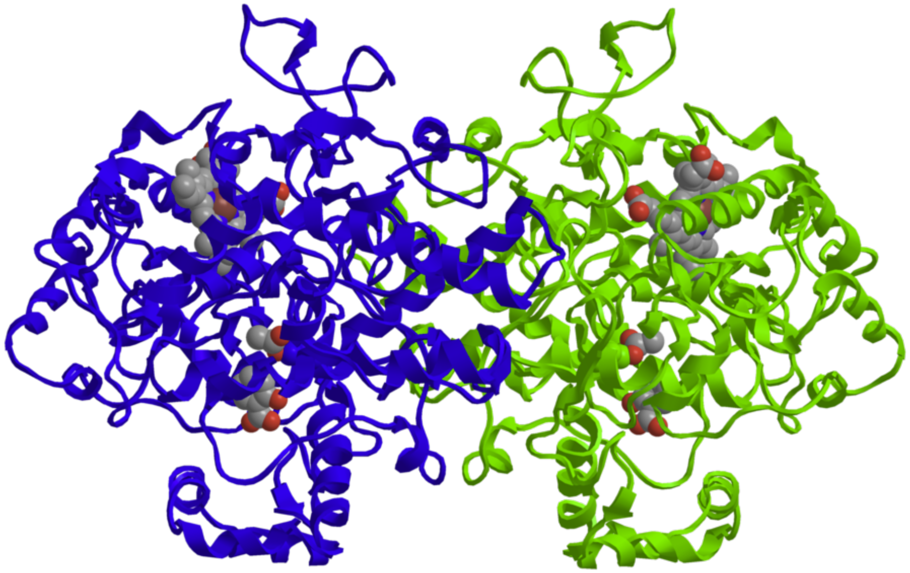
A data scientist's best friend

http://www.openanalytics.eu/architect

Pharma, the simple story

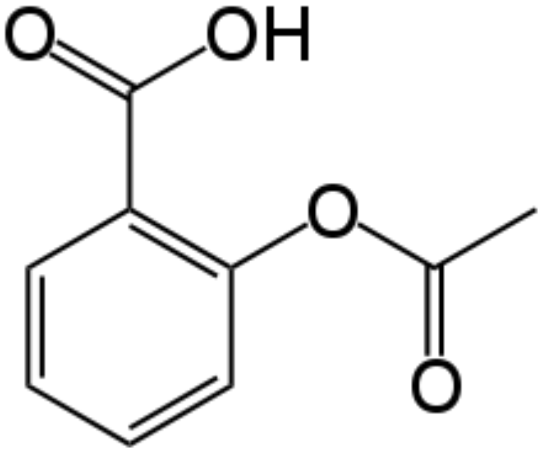
Protein target

- Focus on isolated disease related targets



Compound

- Screening for lead compounds
- Further optimization (chemical modification)
- Blockbuster drug



HELLO
happy investor
HAPPY
INVESTOR
ourself
HAPPY INVESTING

Pharma, the true story

High failure rates in clinical trials

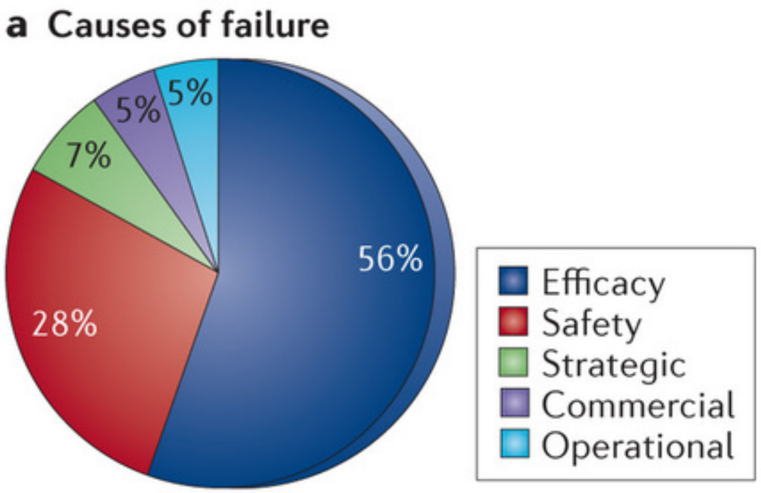
WORD
OF
MOUTH

medibank
YAHOO!
FORECLOSURE
FOR SALE
R.O.I.
Ben van Beurden
ARMS
INFORUM

Pharma, the true story

High failure rates in clinical trials

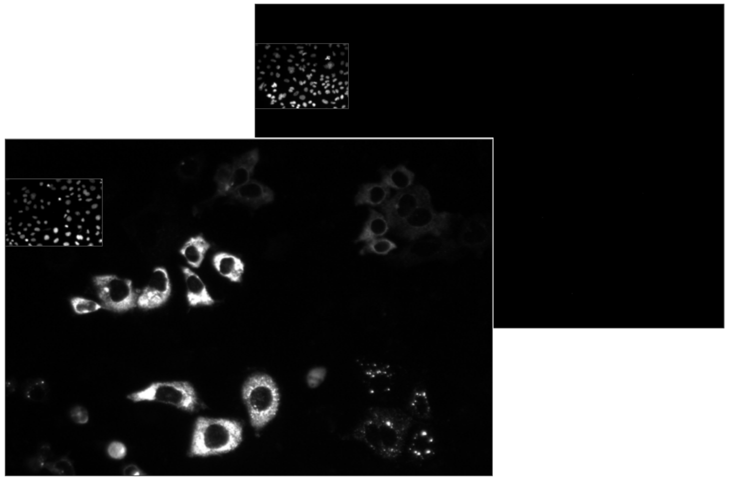
- Phase II success rates below 20%
- 84% of clinical trials fail due to efficacy and safety issues



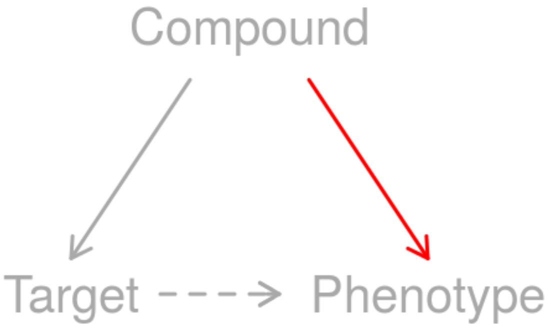
Arrowsmith, J. & Miller, P. Nat Rev Drug Discov, 2013

Phenotypic assays

- Attempt to reduce failure rates
- Compounds activity measured in different type of assay
  - Disease-relevant, multi-target, cellular context
  - Instead of classic assay: isolated target



Phenotypic high-content imaging assay

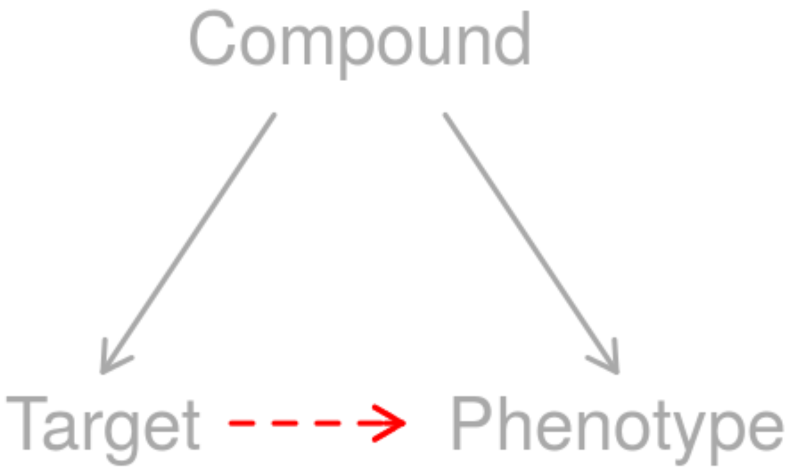


Compound

Target ---> Phenotype

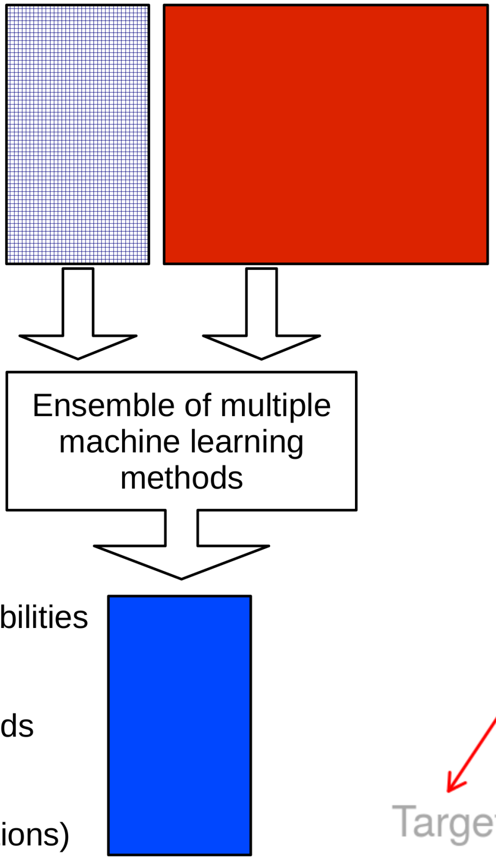
Phenotypic target identification

- Unknown mode of action in phenotypic assay
  - Required for further drug development
- Task: Identify targets that can best explain compound activity in phenotypic assay



Train data: Activity

Sparse (2%) matrix
1.5M compounds
x
2k targets
(x 5 concentrations)



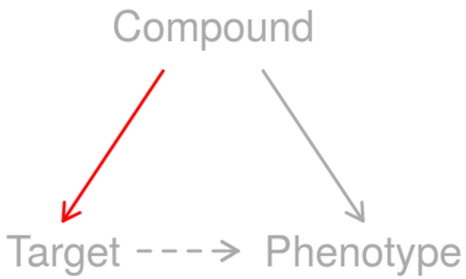
Features:
Molecular fingerprints

1.5M compounds
x
18 M substructures

Ensemble of multiple machine learning methods

Output: Probabilities

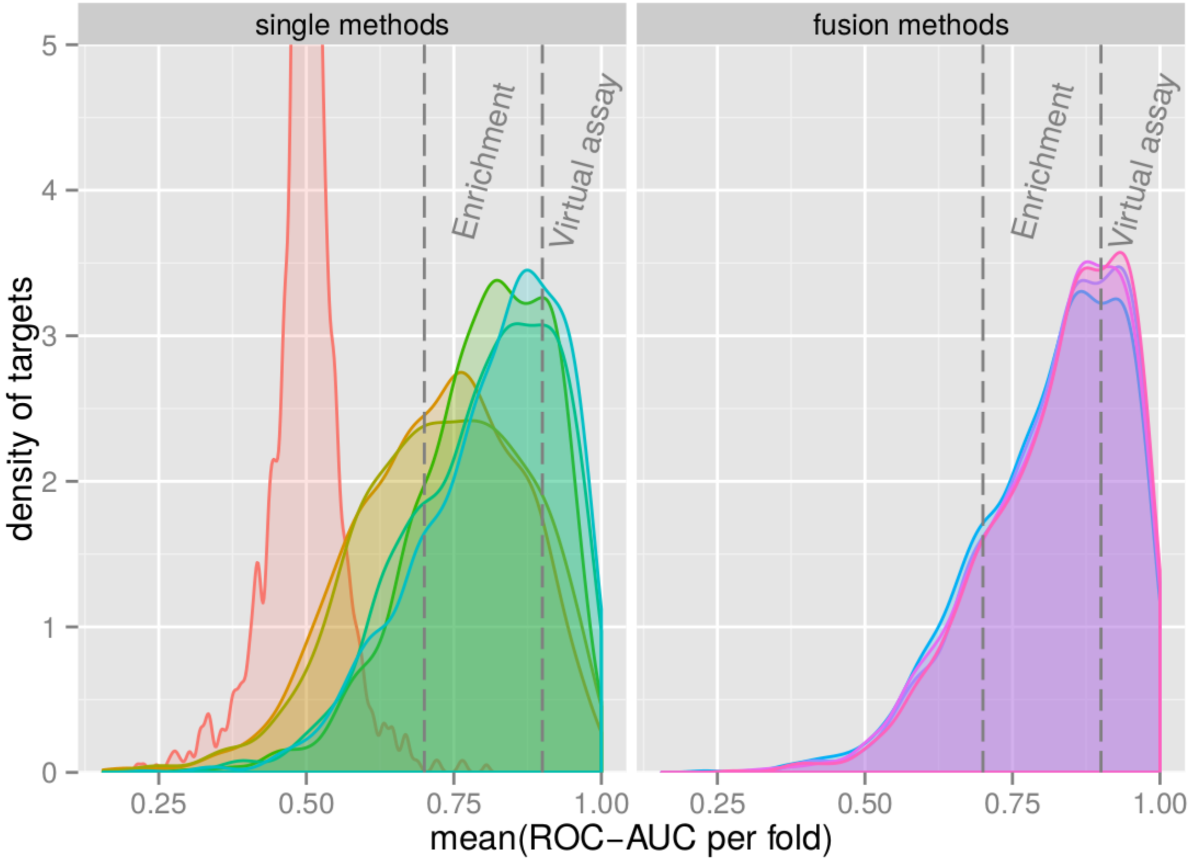
Dense matrix
1.5M compounds
x
2k targets
(x 5 concentrations)



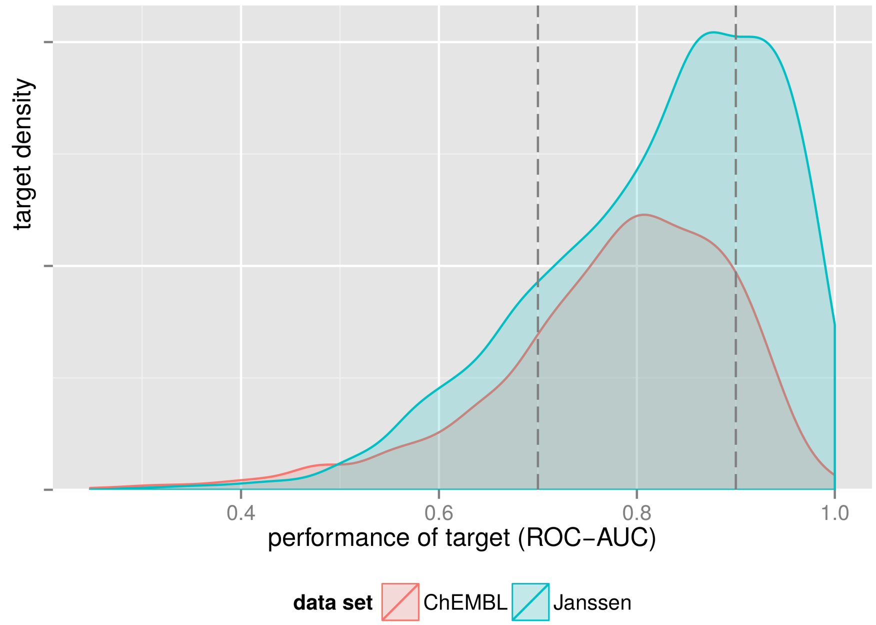
Compound

Target ---> Phenotype

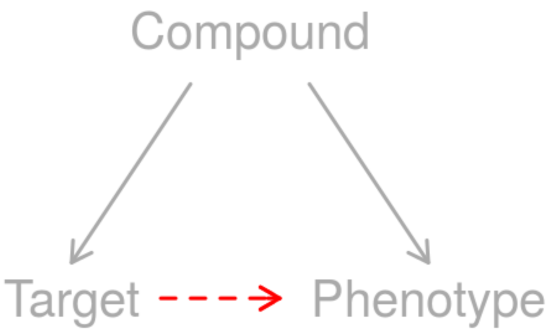
Ensemble of best classifiers



- Literature in chemogenomics field: only public data
- We use public + internal + commercial data
  - More (and better quality) train data
  - Better prediction quality
- Not all available methods were up to the task



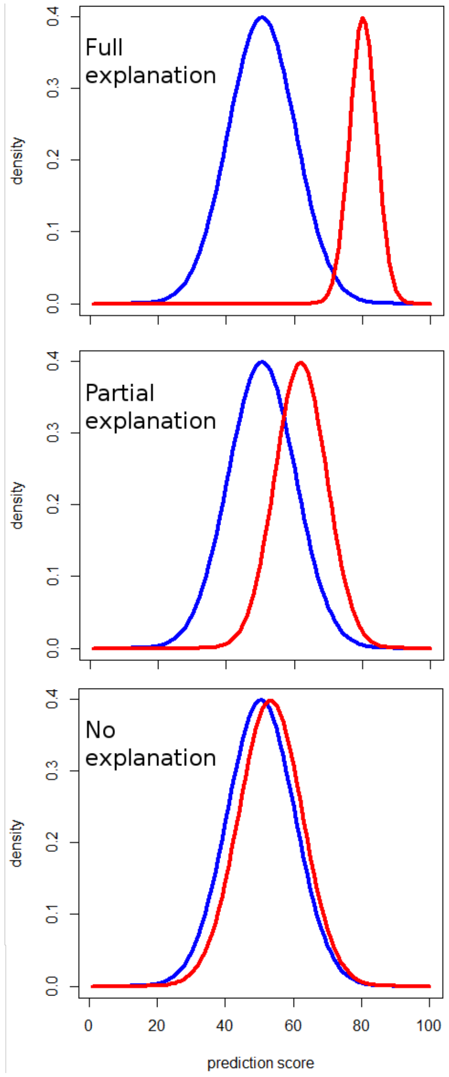
- How well does a target explain the phenotype?
  - ROC-AUC
- How well do multiple targets explain the phenotype?
  - Elastic net logistic regression (glmnet)
  - Random Forest (Boruta)



Compound

Target

Phenotype



Next steps

- Scientific:
  - Ongoing improvements of methods
  - Repeat benchmark with additional sampling
  - Publication(s)
- Operational:
  - Application in various disease areas
  - Now: experimental follow up
  - Iterative cycle of wet lab experiments and modeling

happy investor
HAPPY INVESTOR
ourself
HAPPY INVESTING

Acknowledgements

Contributors

- University of Linz
  - Sepp Hochreiter
- University of Leuven
  - Yves Moreau
- IMEC
  - Roel Wuyts
- Arcadia
- Intel Corporation NV
- Janssen Pharmaceutica
  - R&D Discovery Sciences, Computational Sciences
- Open Analytics

Sponsors

- ChemBioBrige IWT
- Exascience IWT

Tak for din opmærksomhed!

Bonus material

Under the hood…

- **C++** (heavy work on "large matrix")
  - **Boost**
  - **TBB**
  - **JCompoundMapper**
- **Spark** on **YARN** (distributed runs for nested cross-validation)
- **R** (everything else, including analysis and postprocessing)
  - Faster, simpler and more elegant code : **data.table**
  - Target identification : **glmnet**, **Boruta**, **randomForest**
  - Reporting : **rmarkdown**