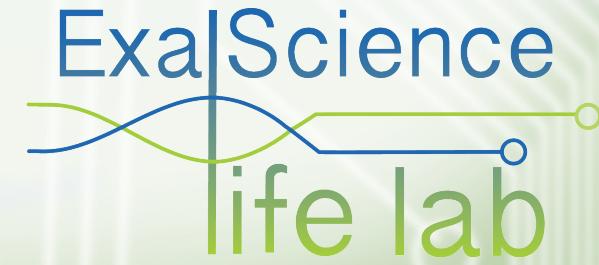
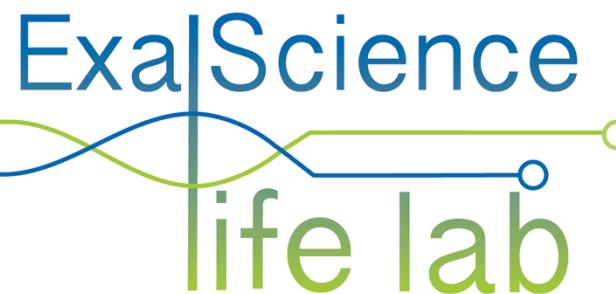


# Addressing computational bottlenecks in pharma R&D

Hugo Ceulemans, MD, PhD  
Janssen & ExaScience Life Lab



# Performance improvement of whole-genome sequencing



# Whole genome sequencing: huge market, mismatch data generation vs. data analysis

All your genetic risks are encoded in your genome

- Market in 2020: 20M genomes/y; ~10B USD/y



Since introduction Illumina HiseqX ten (2014)

- Human genome seq / 0.5h
- Cost per genome: 1000 USD

Analysis pipelines have been slower to catch up



# Whole Genome Sequencing - Analysis Pipeline

Mapping

BAM Processing

Variant Calling

FastQ

BAM

Clean BAM

VCF

BWA-cilk

elPrep

GATK

Halvade

BWA-cilk

elPrep

GATK

Halvade

# Complexity down, speed up: elPrep

## BAM Processing

Remove  
Unmapped Reads

Sort Contig  
Order

Sort Coordinate  
Order

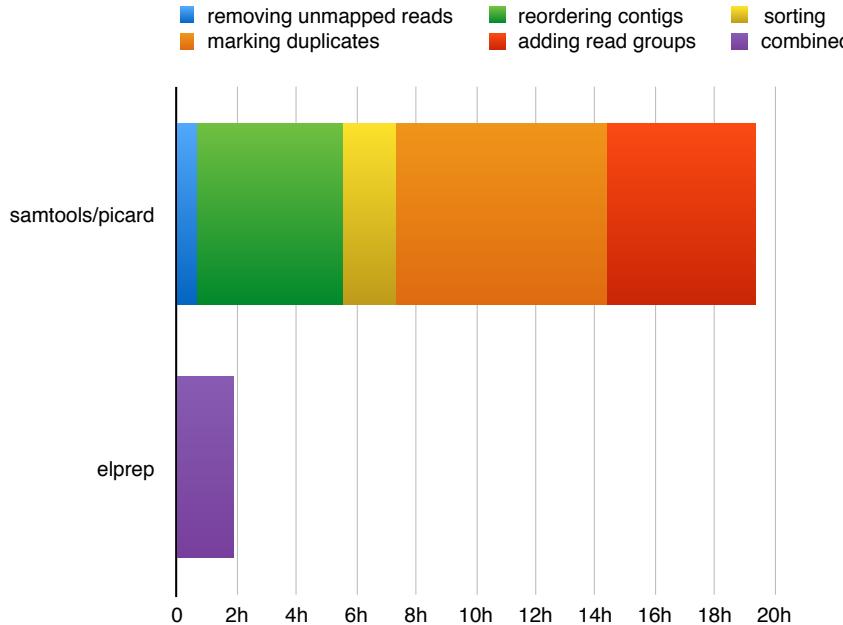
Mark  
Duplicates

Add Read  
Group Information

elPrep

# elPrep reduces running times by factor 10

- SAMtools + Picard:
  - Inefficient I/O
  - Often only single-threaded execution
- elPrep:
  - Steps as in-memory higher-order functions
  - Efficient multithreading

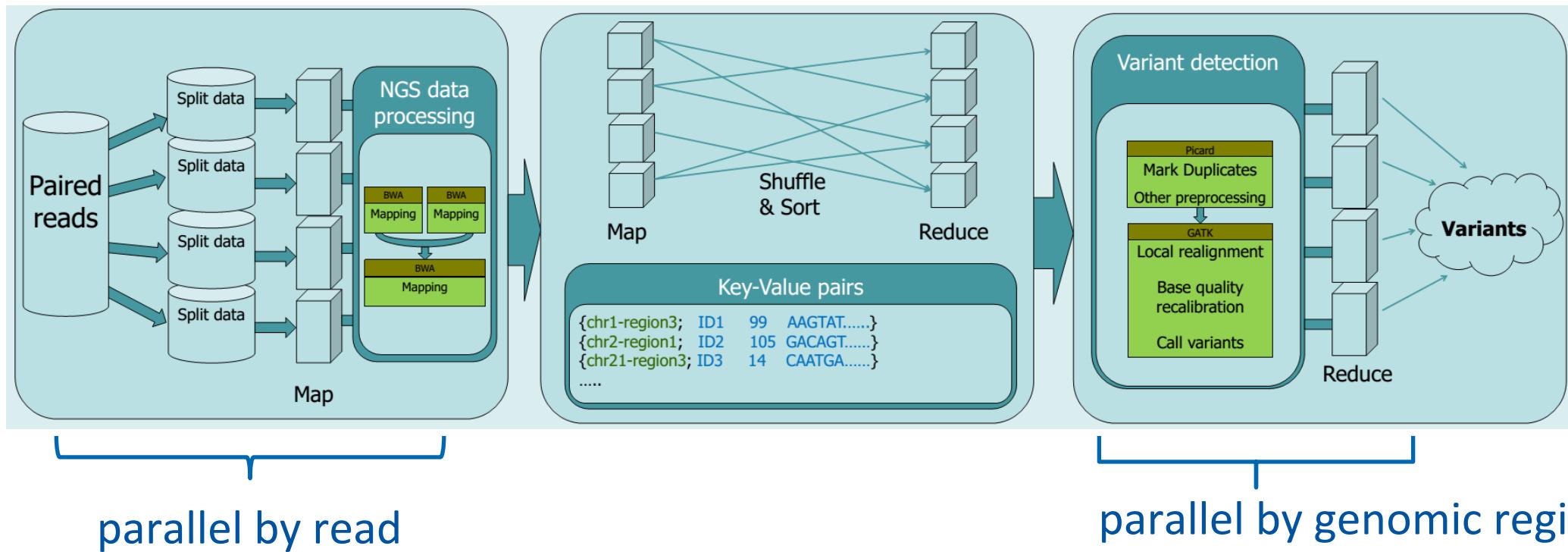


2x8 Intel Xeon E5-2680 with 256GB RAM

	NA12878	SAMtools + Picard	elPrep
Removing unmapped reads	38m12s	-	-
Sort contigs	4h55m55s	-	-
Sorting	1h46m45s	-	-
Mark duplicates	6h59m44s	-	-
Read groups	4h58m53s	-	-
Total	19h19m29s	1h54m51s	

# Moving on to multiple nodes...

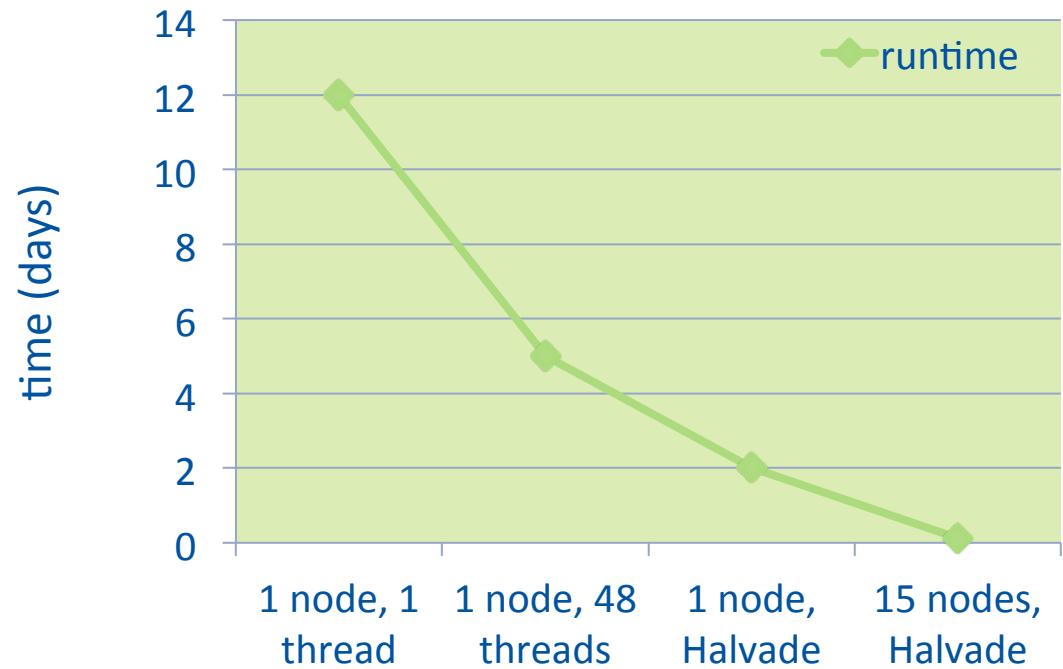
- Hadoop ALigner & VAriant DEtection = Halvade
- Map/Reduce approach



# Halvade: 1 WGS sample in under 3 hours

Mapping and alignment calling for human genome with 50x coverage:

- Rough estimate 1 node, 1 thread
  - ~288h or ~12 days
- Time on 1 node multi-threaded
  - ~120h or ~5 days
- Time on 1 node with **Halvade**
  - ~48h or ~2 days
- Time on 15 nodes with **Halvade**
  - **~3h or ~0.13 days**

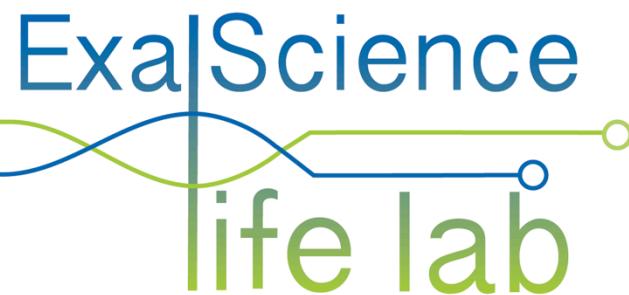


Also supported: whole exome sequencing

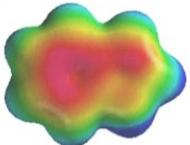
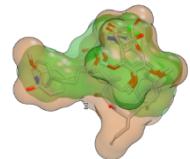
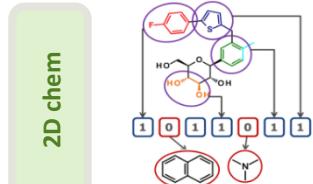
Work in progress: RNA-seq variant calling pipeline



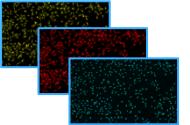
# Holistic compound activity prediction



# Using high-dimensional cpd descriptors



L1000



No fibril

Machine learning

predict cpd activities in (validated & chemically documented) assays



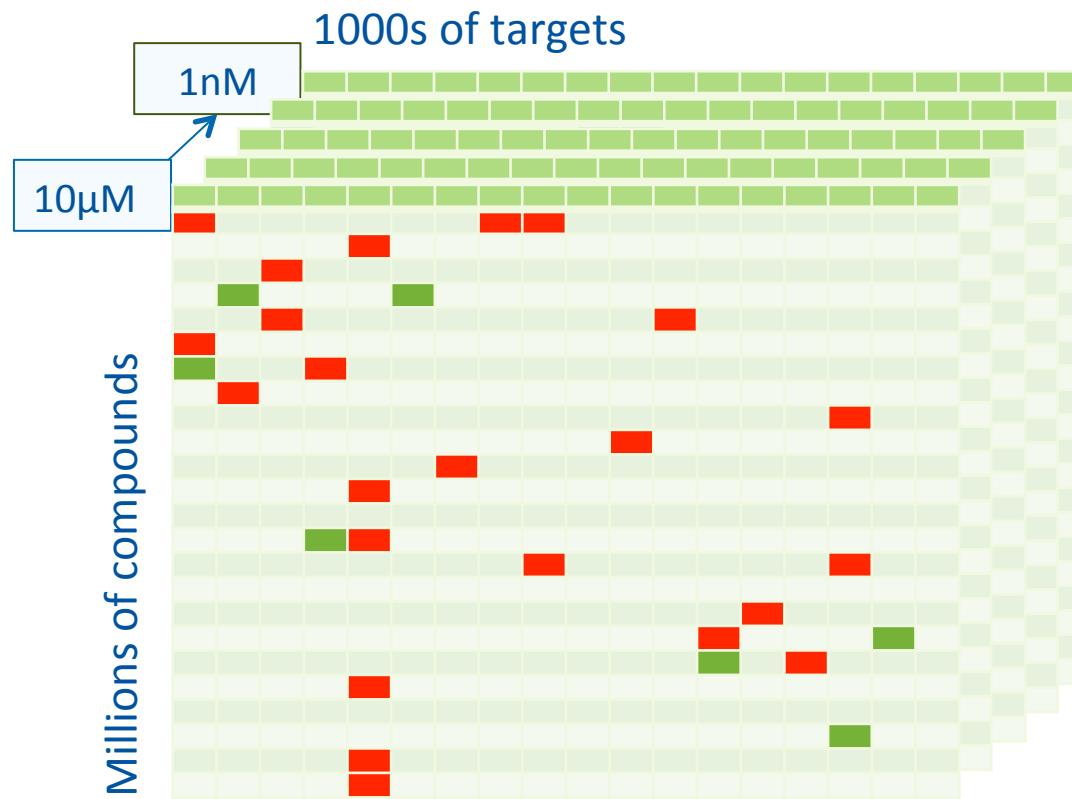
X00Ms of dose-response quality training points

# x00Ms of biochemical activities for training

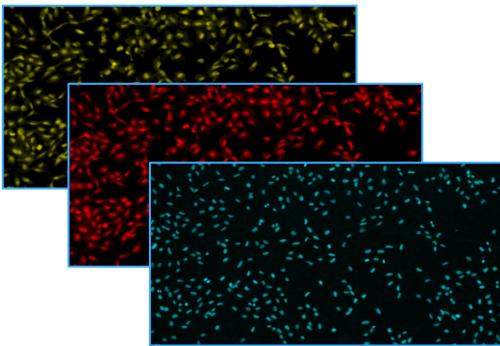
~1% can be filled up with experimental dose response data



GVK BIO Online Structure Activity Relationship Database



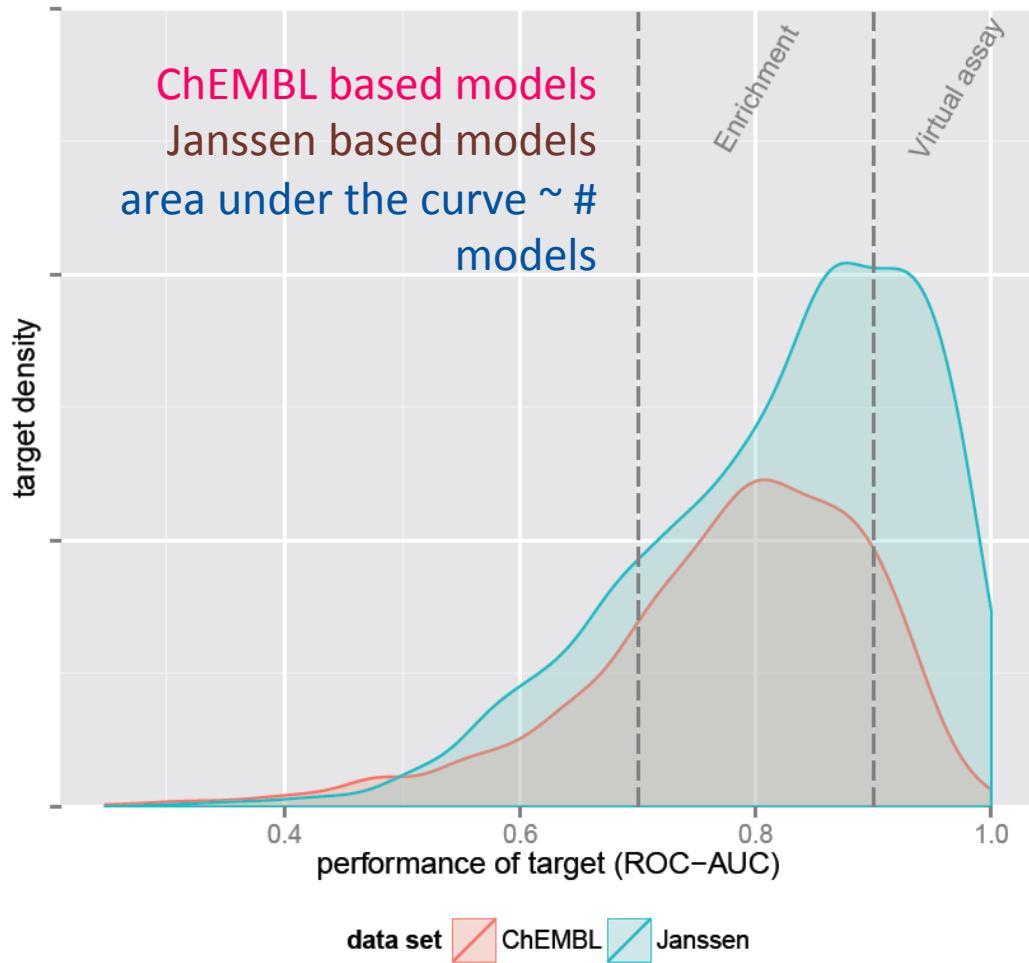
*data generated for single readout*  
>>500k cpds x 6 images = 15 Tb



High-content imaging cpd descriptor @ high throughput (>500k cpds)



# The importance of data volume



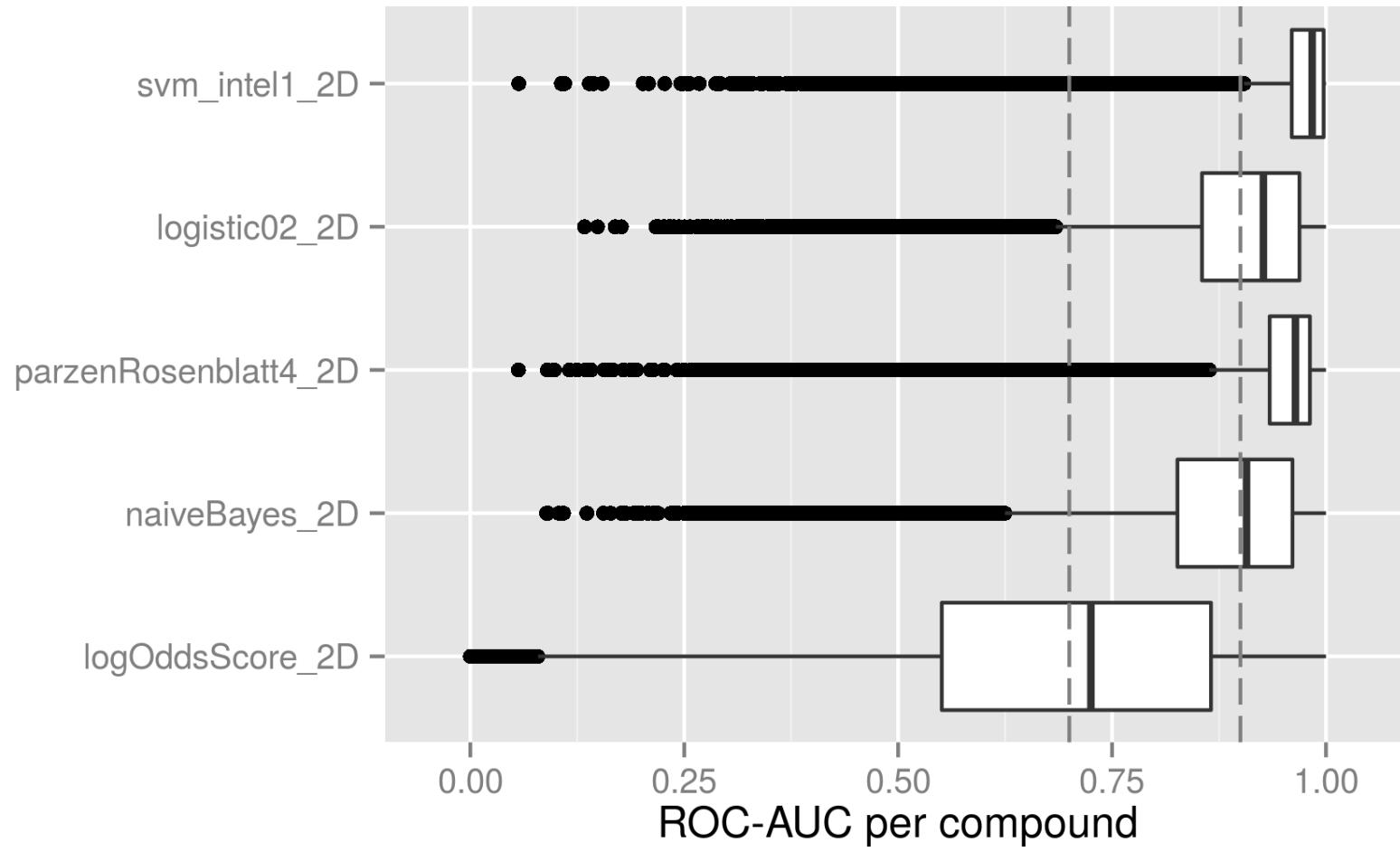
industry scale data  
⇒ models for more targets  
⇒ average quality improves

implies a need for internalization of external methods

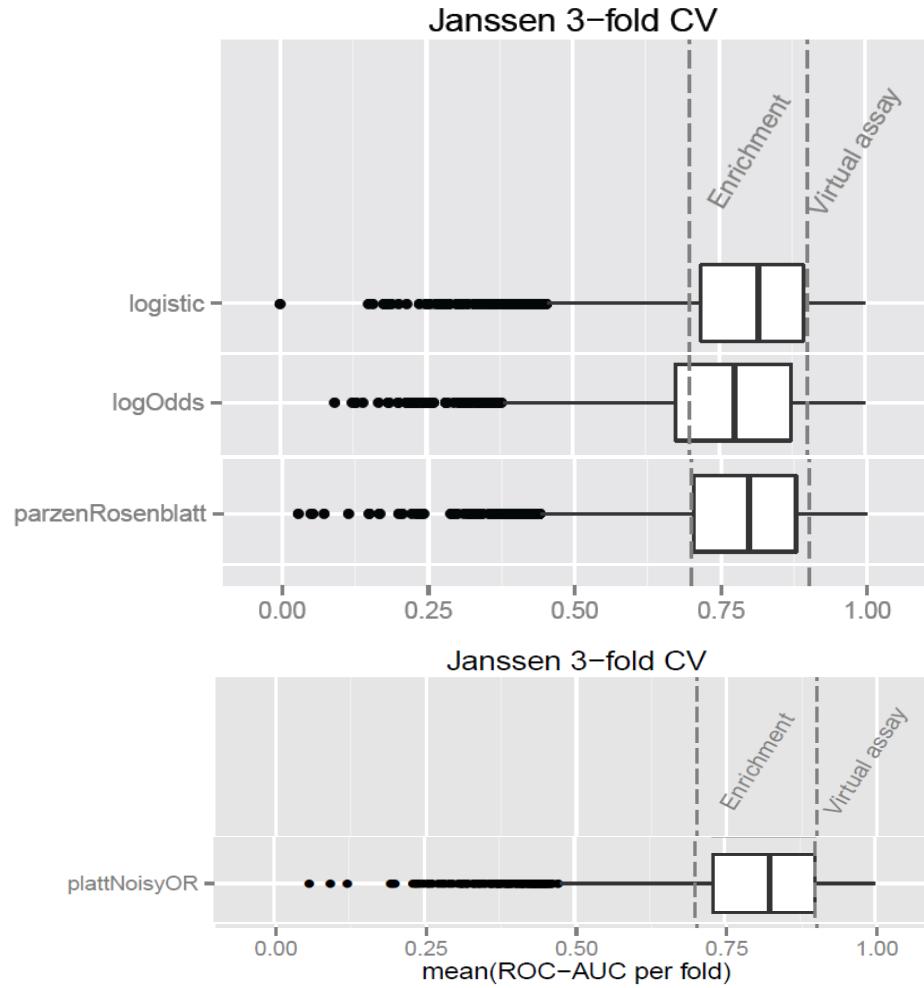




# Method choice matters for predictive performance



# Method choice matters for predictive performance



# Regression to EC<sub>50</sub>s: another Netflix challenge?



Regularized  
probabilistic  
matrix factorization

Latent  
compound  
variables

U

Bayesian  
probabilistic  
matrix factorization

$\Lambda_u, \mu_u$

hyperpriors  
priors

Latent target  
variables

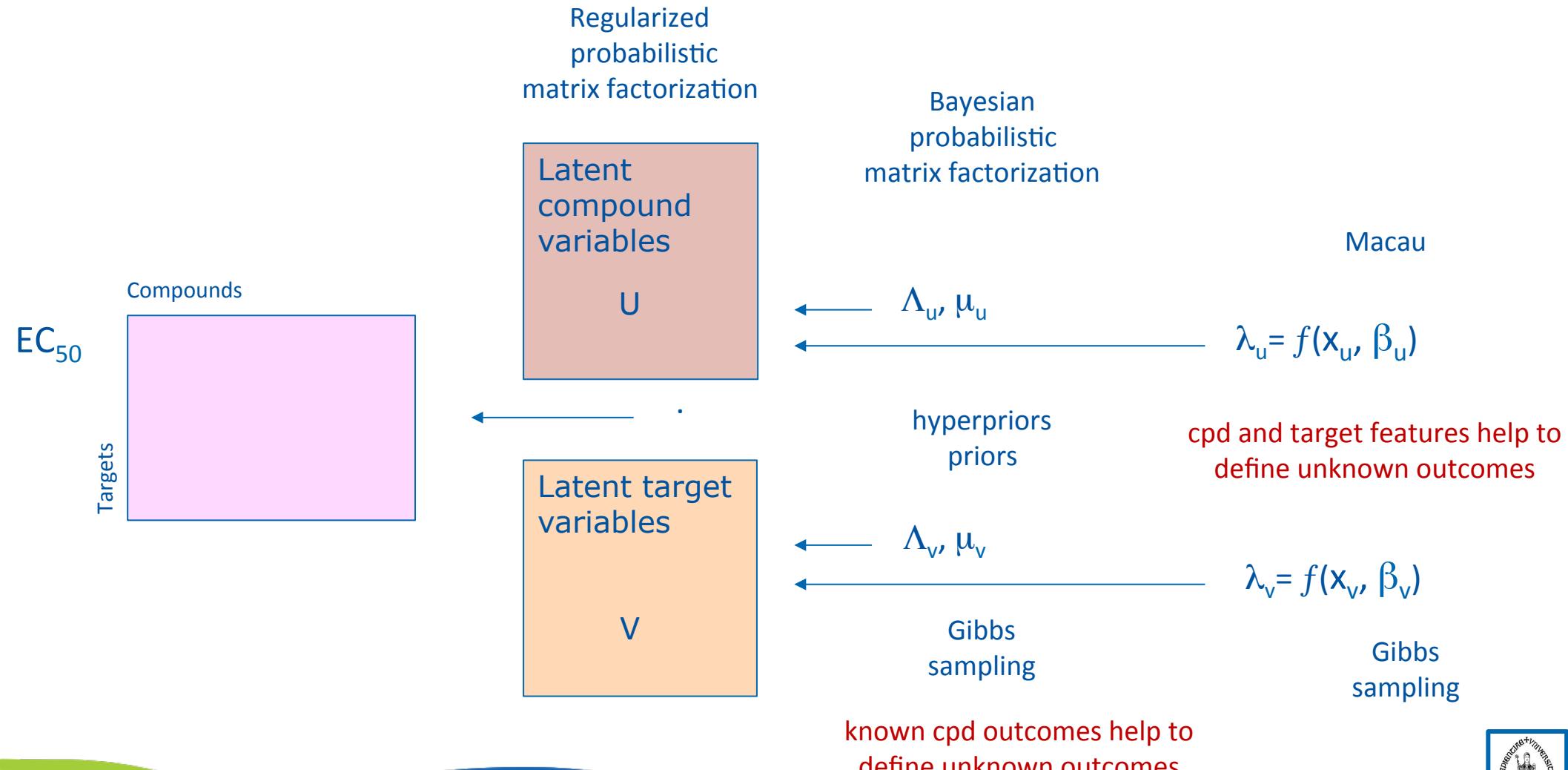
V

$\Lambda_v, \mu_v$

Gibbs  
sampling

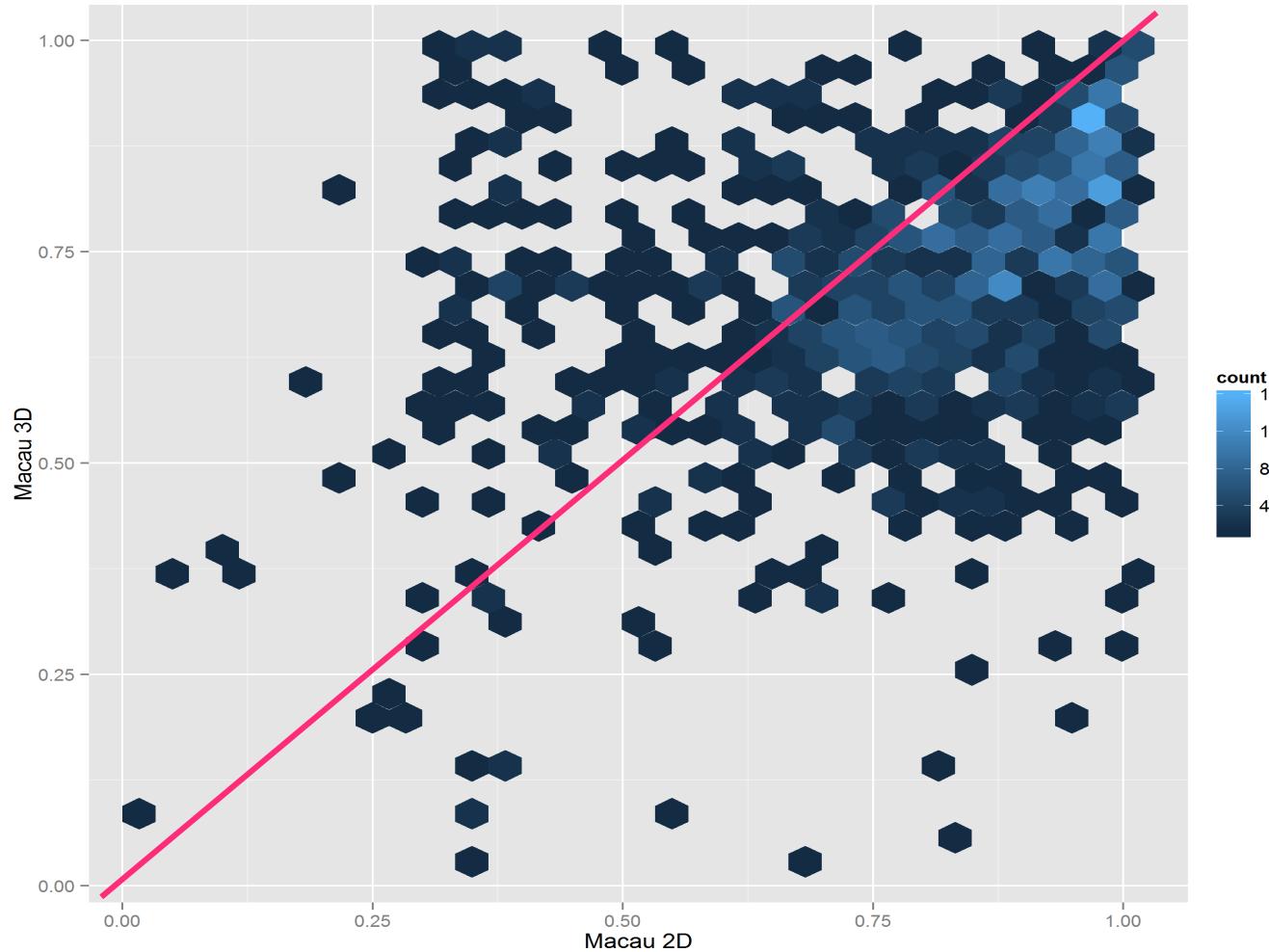
known cpd outcomes help to  
define unknown outcomes

# Macau: extending BPMF with features

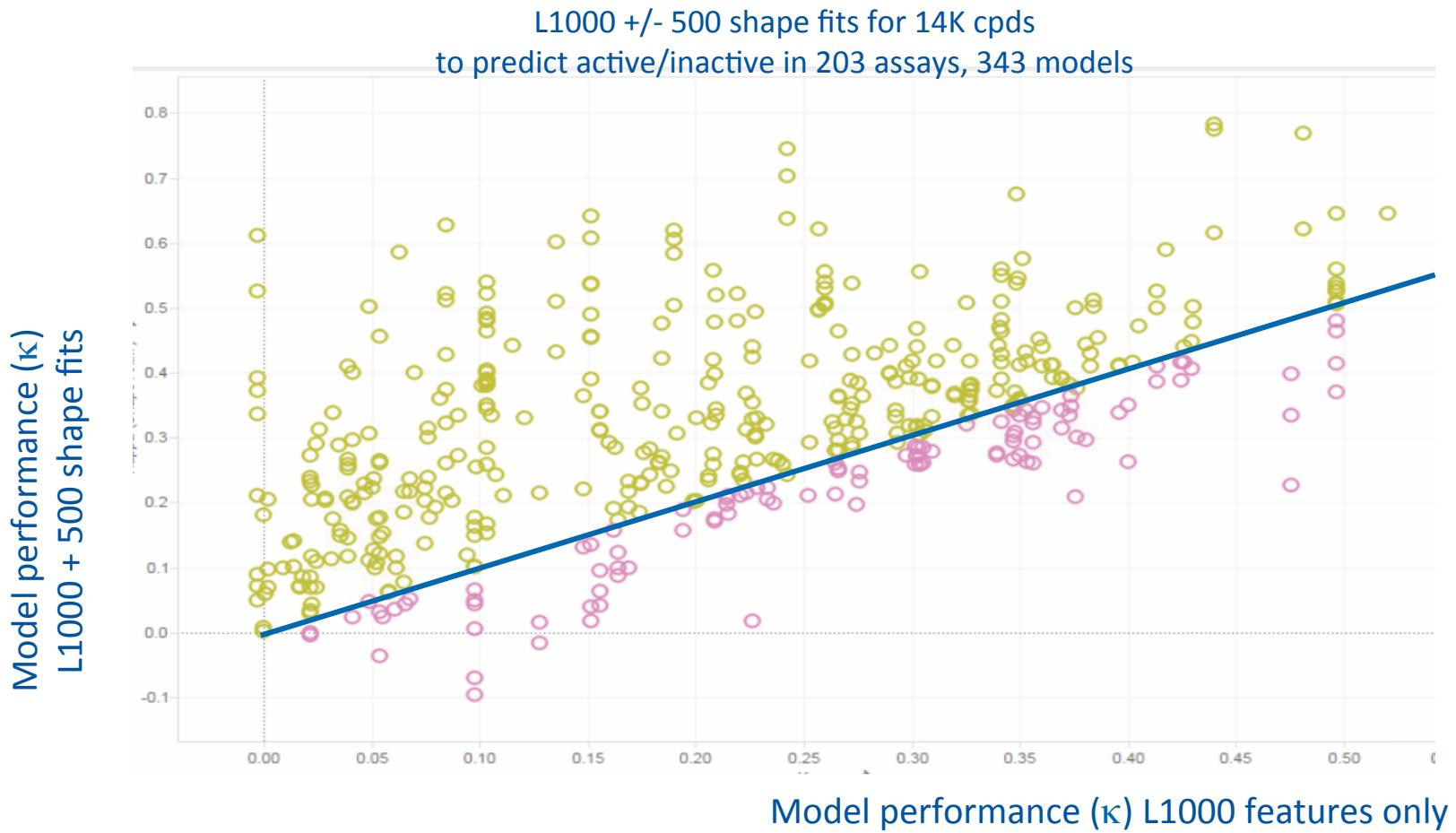


# Input choice matters for predictive performance

BMPF/Macau - active/inactive @ 10 nM – 2D vs. 3D



# Input combination boosts predictive performance



# Examples indicative of compute need

Wildcat cluster: 32 nodes x 2 CPUs x 18 cores = 1152 cores 22.3 GHz,  
256 GB memory, 2x 400 GB SSD, 2x 4 TB HDD

- Model build on full dataset 2D, “Classic classifiers” active/inactive single concentration level:

- Parzen-Rosenblatt
- Logistic regression
- Naive Bayes
- LogOdds score

⇒ 33h on single node

- Model build on full dataset 2D “Macau/BPMF” regression to EC<sub>50</sub>s:

⇒ 7d on 15 nodes (optimized from 80d)

- Feature extraction from HTS image screen

⇒ 10d on 32 nodes



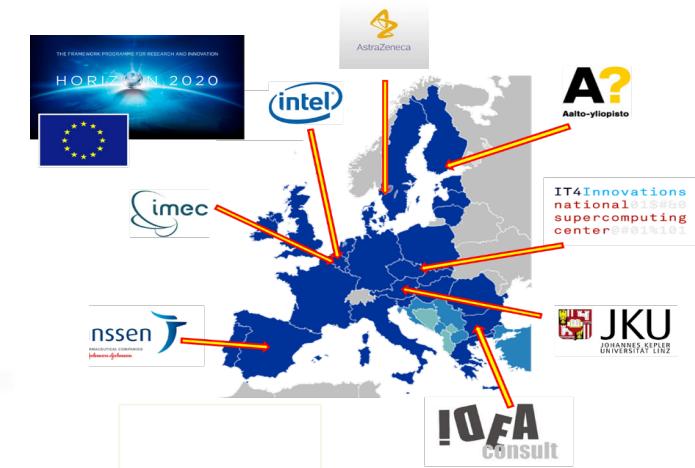
# Collaborative development of compute solutions



ExaScience  
Regional funding  
Flemish region, Belgium



Exascience Exa2CT  
EU funding via FP7



Exascience ExCAPE  
EU funding via H2020



# Acknowledgements

## University of Linz

- Prof. Sepp Hochreiter
- Andreas Mayr
- Andreas Mitterecker
- Günther Klambauer

## University of Leuven

- Prof. Yves Moreaux
- Prof. Karl Van Meerberg
- Jaak Simms
- Adam Arany
- Albert-Jan Yzelman
- Dries Harnie

## Free university of Brussels

- Prof. Wolfgang De Meuter
- Dries Harnie

## Univeristy of Ghent

- Prof. Jan Fostier
- Dries Decap

## BROAD Institute

- Prof. Anne Carpenter
- Mark Bray
- Vebjorn Llosa
- Shantanu Singh

## IMEC

- Wilfried Verachtert
- Tom Ashby
- Roel Wuyts
- Tom Vander Aa
- Imen Chakroun
- Charlotte Herzeele

## Arcadia

- Alexander Shevin
- Andrey Gedich

## OpenAnalytics

- Tobias Verbeke
- Marvin Steijaert
- Laure Cougnaud

## Janssen

- Jörg Wegner
- Vladimir Chupakhin
- Alexander Vapirev
- Hugo Ceulemans
- Emmanuel Gustin
- An De Bondt
- Hinrich Goehlmann
- Pieter Peeters

## Intel

- Luc Provoost
- Sergei Osokhin
- Pascal Constanza

