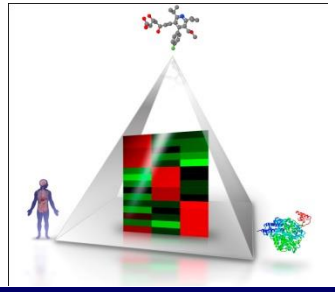
Integrating Chemical and Biological Data for Drug Discovery

Andreas Bender, PhD
Lecturer for Molecular Informatics
Unilever Centre for Molecular Science Informatics
University of Cambridge
Fellow of King's College, Cambridge

UNIVERSITY OF CAMBRIDGE



Outline

- Chemical and biological data – what is out there (and how can we use it?)

- Integrating chemical and biological data
  - Mode-of-action analysis
  - Compound sensitivity modelling in cancer
  - Using gene expression data for stem cell differentiation

- Current challenges and problems

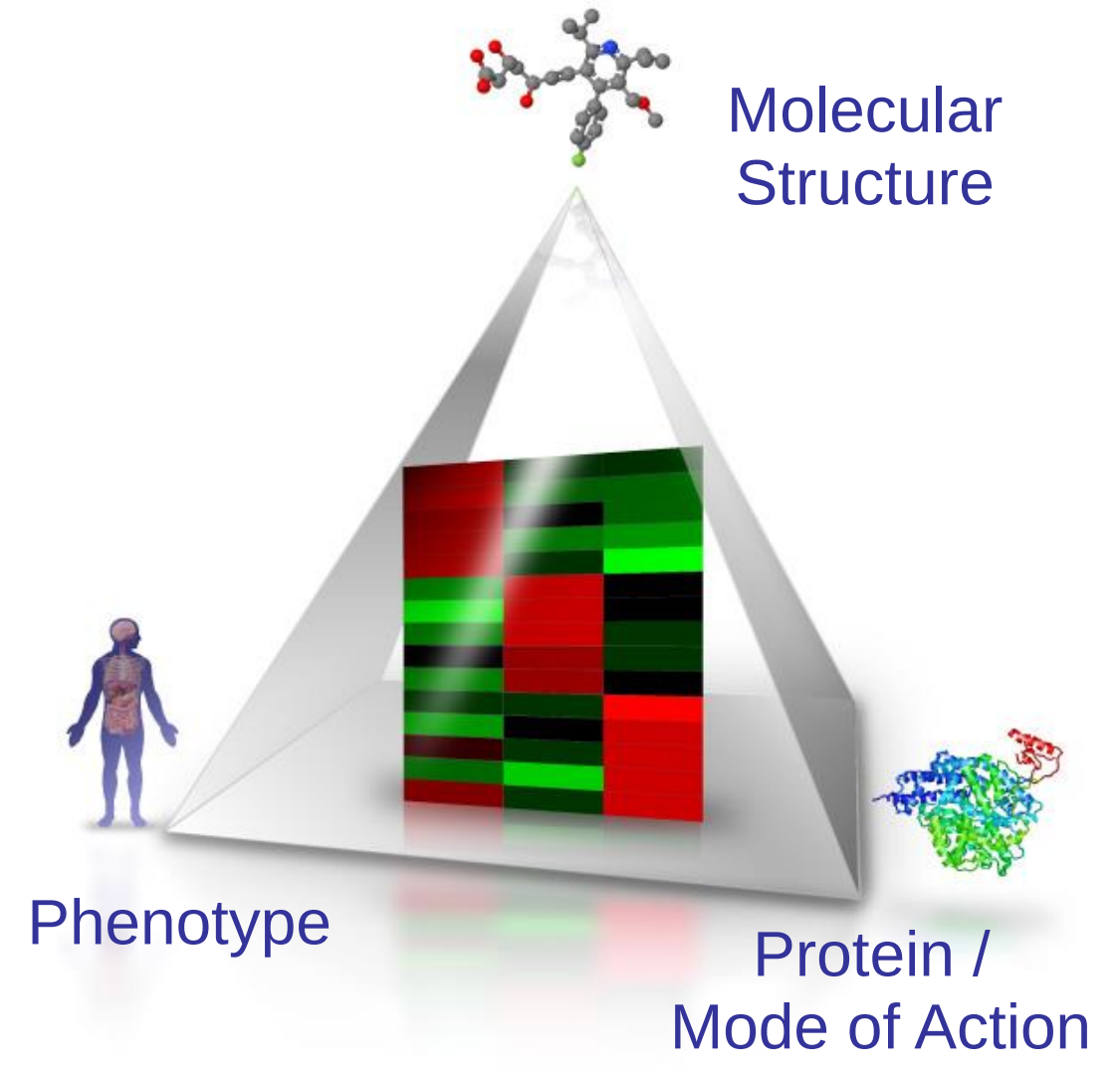
More and More Data is Available...

- But: How should we deal with it?

- Databases contain tens of millions of bioactivity data points, gene expression data, organ tox endpoint data, clinical trial data, ...
- *However*, integration – and utilization – of data is often not ideal

- This is what we aim to do in our group
  - Integrate, analyze *heterogeneous* life science data
  - Provide testable hypotheses
  - Test those hypotheses

Core Data Considered: Chemistry, Phenotype, Targets / Mode of Action



So what's the point of it all?
We would like to answer questions!

- "What is the reason upon treatment with A for phenotypic effect B?"
  -> *Mode of Action*
- "Which compound should I make to achieve effect C in a biological system?"
  -> *Chemistry*
- "Does patient D or patient E respond better to drug F?"
  -> *Phenotype / Phenotype Change*

Group Structure

- About 22 people (ca. 16 PhD students, plus postdocs, visitors etc.)
- Funding from ERC, BBSRC, EPSRC, CEFIC; BASF, Eli Lilly, Johnson&Johnson, AstraZeneca, Unilever, Aboca, ... Plus close to 1/3 personal scholarships
  - Public money and personal scholarships for independent method development and application
  - Company projects for ‘real’ validation of our ideas
  - Hence, both parts are crucial in my point of view

Case studies

- Mode-of-action analysis: Understanding how sleeping pills cause sleep (with Eli Lilly)

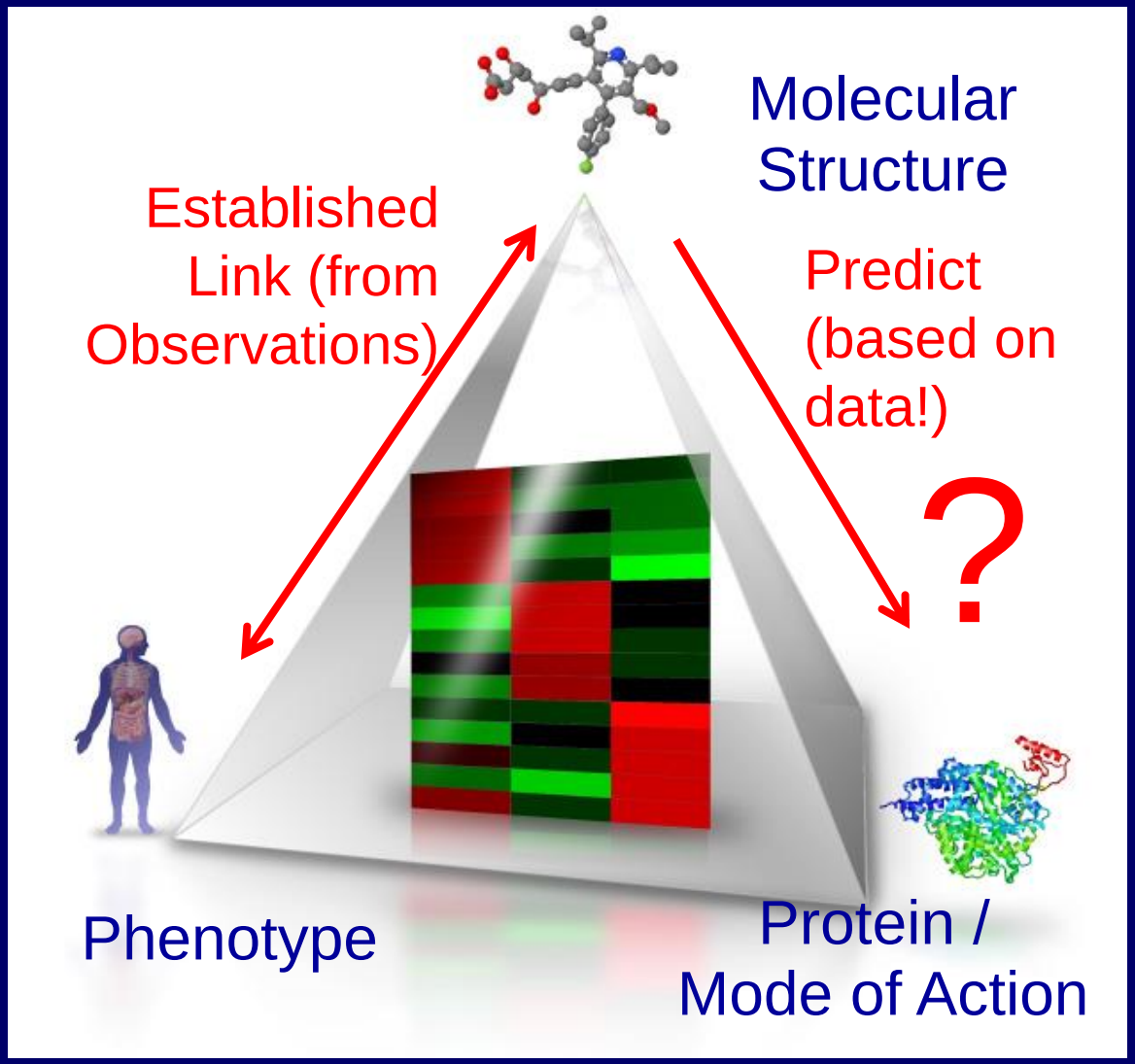
- Cell sensitivity modelling: Using gene expression data for ‘personalized medicine’ in cancer

- Using gene expression data to differentiate stem cells to cardiomyocytes

Case studies (1)

- Mode-of-action analysis: Understanding how sleeping pills cause sleep (with Eli Lilly)

Starting from *in vivo* efficacy we can predict the MoA, based on ligand chemistry

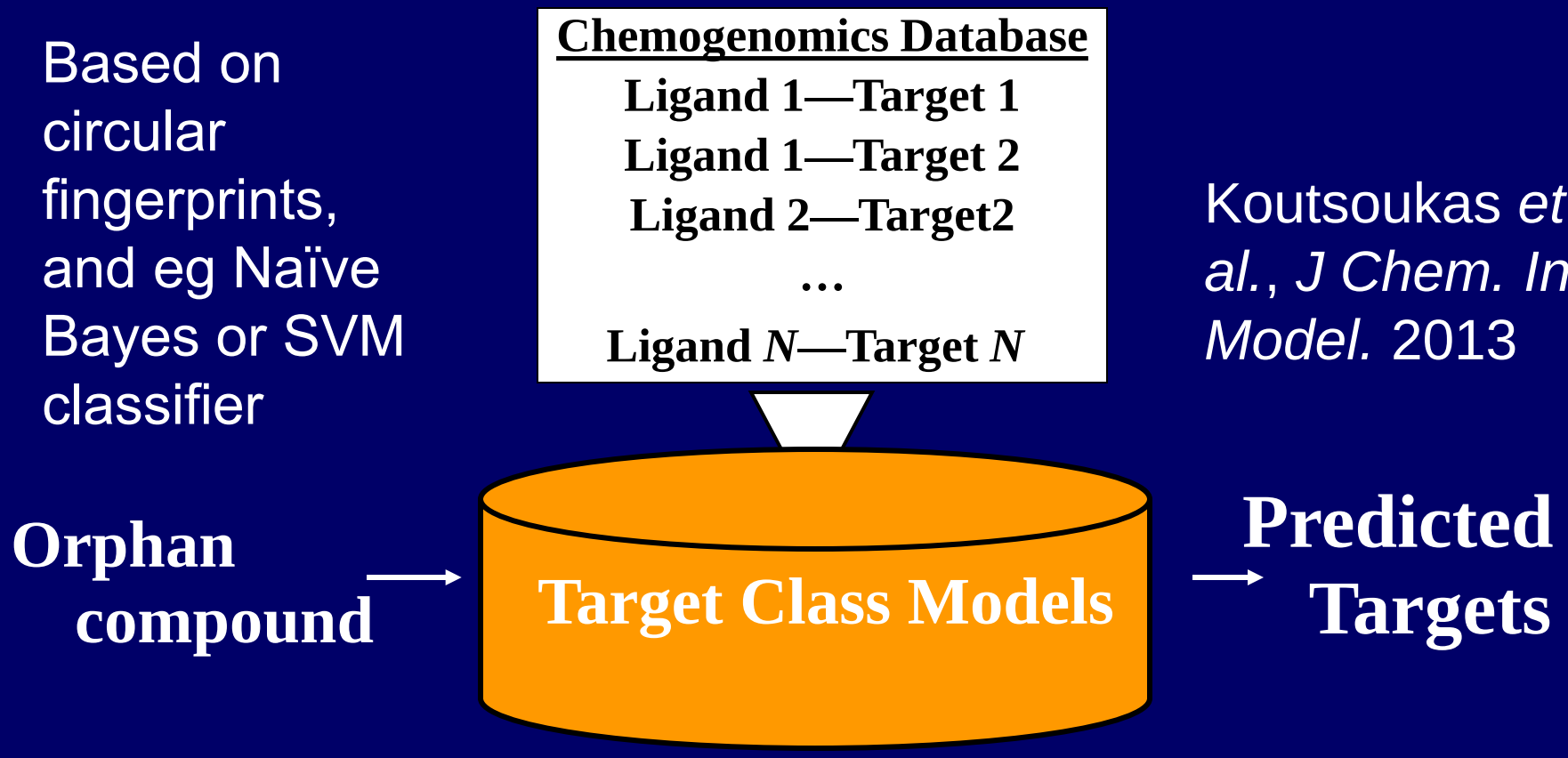


A. Koutsoukas *et* al., J Proteomics 2011 (74) 2554 – 2574.

Exploiting known bioactivity data for new decisions: Target predictions

- The models enable automated prediction of the targets or target families of orphan ligands given only their chemical structures

Based on circular fingerprints, and eg Naïve Bayes or SVM classifier



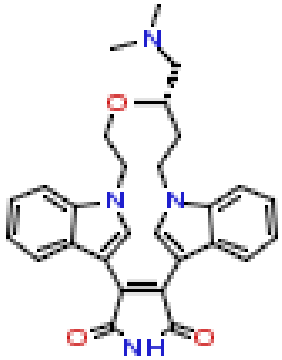
Koutsoukas *et al.*, *J Chem. Inf Model.* 2013

Prediction Examples: Gleevec, Ruboxistaurin

- Gleevec (Novartis),
  - Launched
  - Targets Bcr-Abl, c-kit, PDGFRb

- Ruboxistaurin (Lilly/Takeda),Phase III
  - PKCb

Molecule	Targets	Scores
	ABL1	46.50
	PDGFRB	28.99
	KIT	22.02
	CDK9	21.30
	BRAF	16.13
	FLT1	13.09
	PLK1	8.05
	BTK	5.44

Molecule	Targets	Scores
Chiral 	PRKCB1	95.81
	CAMK2G	87.48
	PRKCG	66.35
	PRKCA	56.99
	PRKCD	52.44
	PRKCH	51.41
	PRKCE	50.42
	PRKCZ	42.48

Understanding rat sleep data

Work by Georgios Drakakis

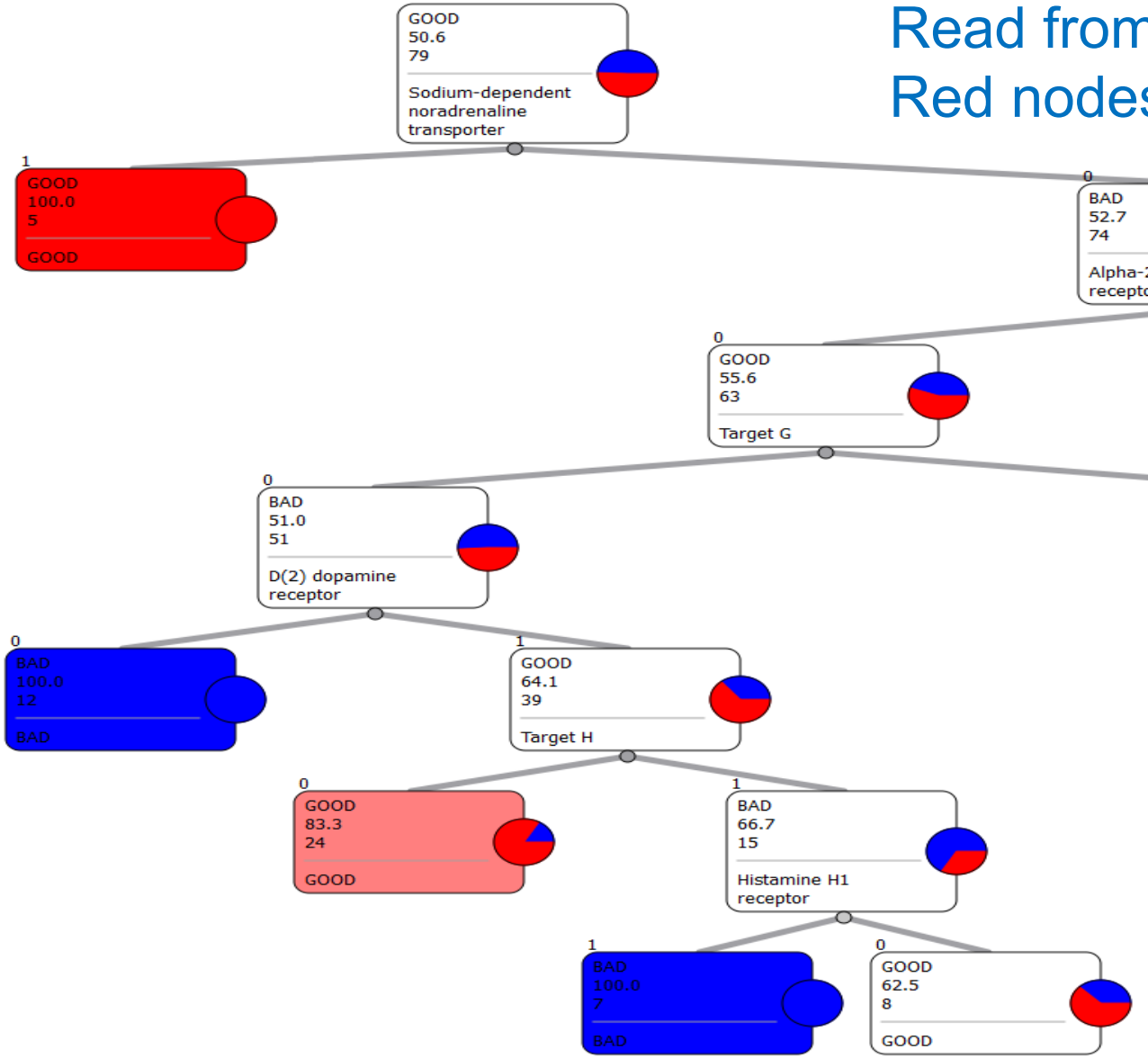
- Project with Eli Lilly
- Male Wistar rats

- Treated with ~500 sleep-inducing compounds, dozens of readouts from EEG/EMG, Abdominal Minimitter, Cage that define ‘good sleep’

- **Q: What are bioactivity profiles associated with compounds inducing good sleep?**

- Target prediction and machine learning to derive rules that make compounds ‘good sleeping pills’

Decision trees learn receptor bioactivity profiles associated with ‘good’ and ‘bad’ sleep



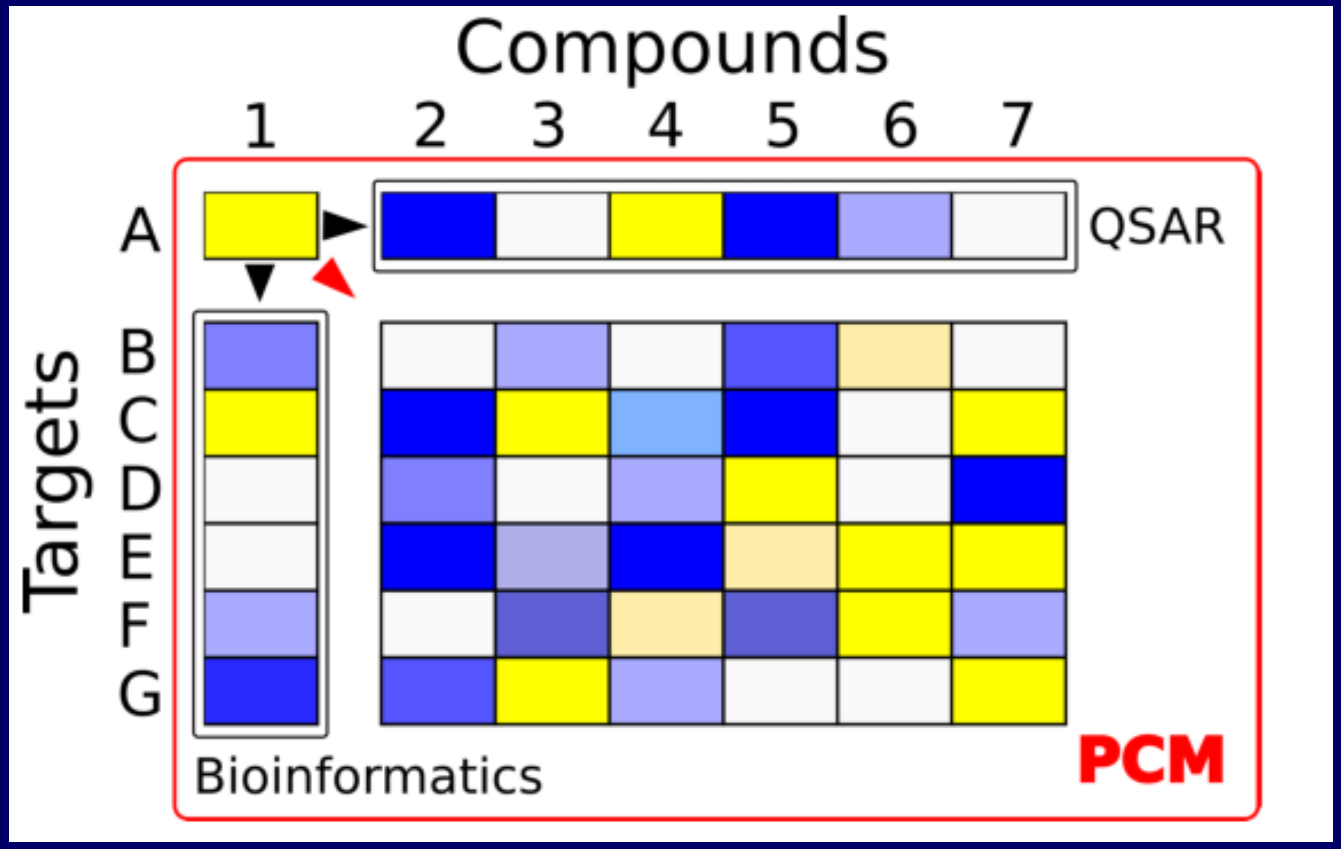
Read from top;
Red nodes = ‘good’ compounds

- Gives us *multiple*, *multi-target* mode-of-action hypotheses
- Can be used eg for compound selection

Case studies (2)

- Cell sensitivity modelling: Using gene expression data for ‘personalized medicine’ in cancer

Integrated Modelling of Chemical and Biological Data



Isidro Cortes-Ciriano, Qurrat Ul Ain,*et al.* *MedChemComm* 2015

Large-scale prediction of growth inhibition on NCI60 cancer cell-lines

- Some cancers (and their drivers) are well understood, others less so
- Amount of biological information on genomic and proteomic (and metabolomic etc.) level that can be generated is huge

- Question is, *which pieces of biological information tell us which compound is (*selectively) cytotoxic in a given cell line (or, ideally, patient-derived cells)?

Isidro Cortes-Ciriano *et al.*, *Bioinformatics*, *under revision*

Different Biological Information Considered Gives Different Results

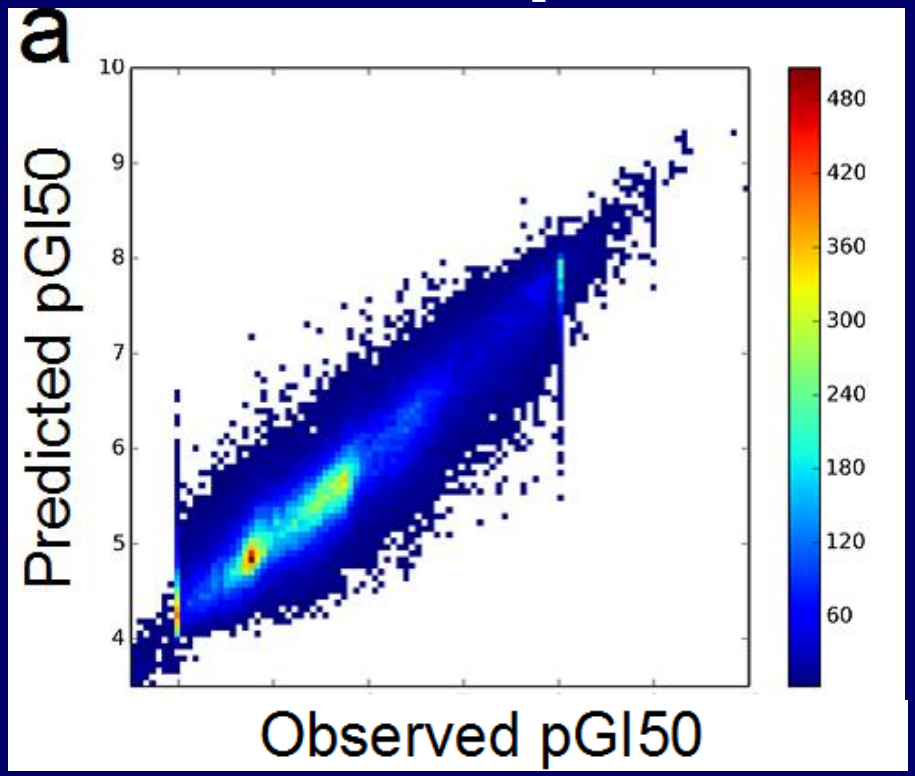
Data types compiled for all (59) NCI-60 cell lines:

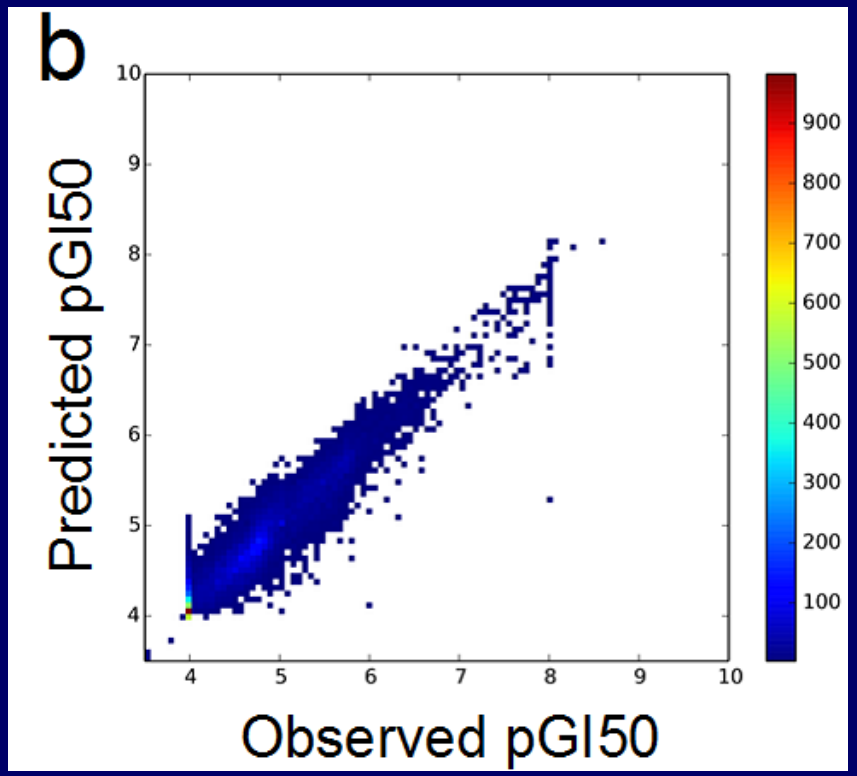
- Gene transcript levels
- miRNA expression
- DNA copy-number variation
- Whole exome sequencing
- Cell-line fingerprints
- Protein abundance for 89 proteins
- Protein levels of the global proteome

17,142 distinct compounds and 941,831 data-points from the NCI-60 screens

Isidro Cortes-Ciriano *et al.*, *Bioinformatics*, *under revision*

Leave-One-Cell-Line Out and Leave-One-Compound Out Validation





- Cell sensitivity models outperformed previous approaches
- Now being tested prospectively in pancreatic cancer (and other cancers) with Addenbrooke's, Ohio Cancer Centre

Case studies (3)

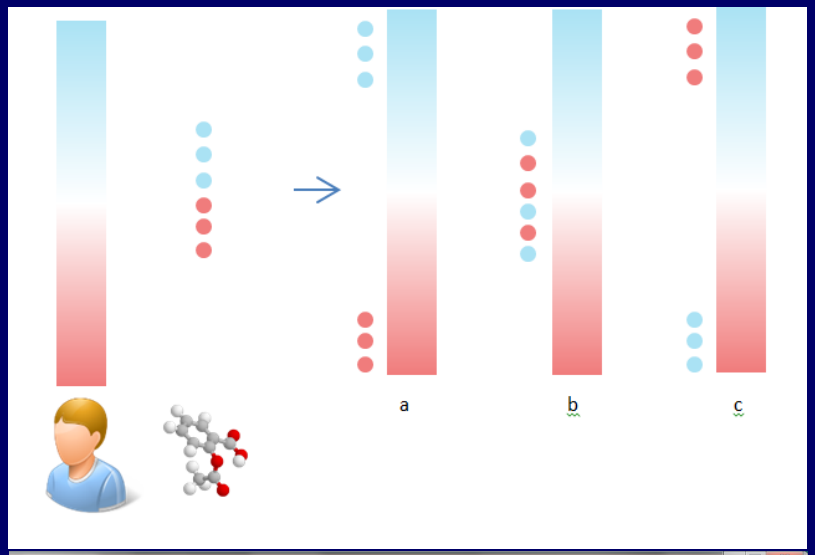
- Using gene expression data to differentiate stem cells to cardiomyocytes

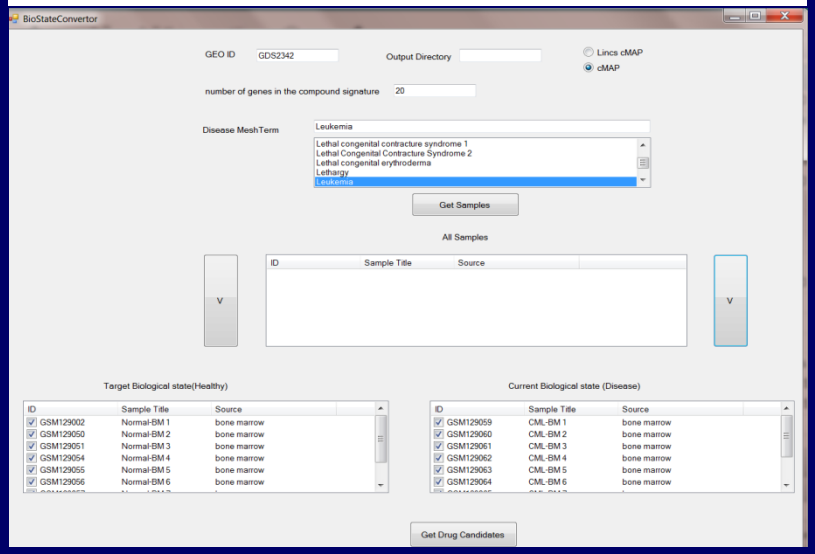
“BioStateConverter” (work of Yasaman KalantarMotamedi)

- Compound-Disease mapping *via* gene expression data

- Connectivity Map Compound Expression Data, to 250+ diseases/ biological states from GEO

- Disease data extracted via text mining (rather tedious process)



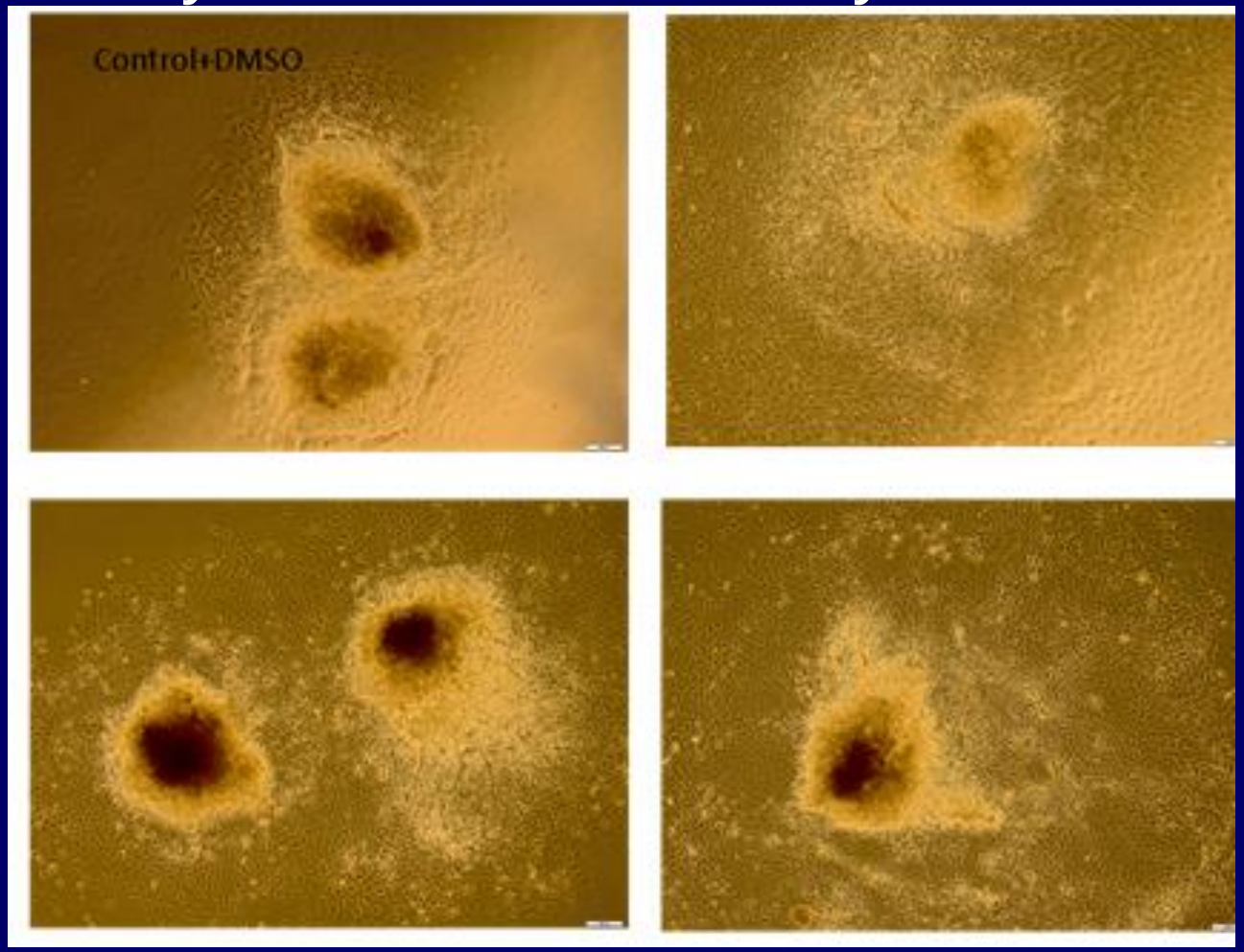


Selected compound induces differentiation of stem cells into cardiac myocytes (by RT-PCR; work with Dr Nasr, Royan Institute, Isfahan)

3 days 5 days

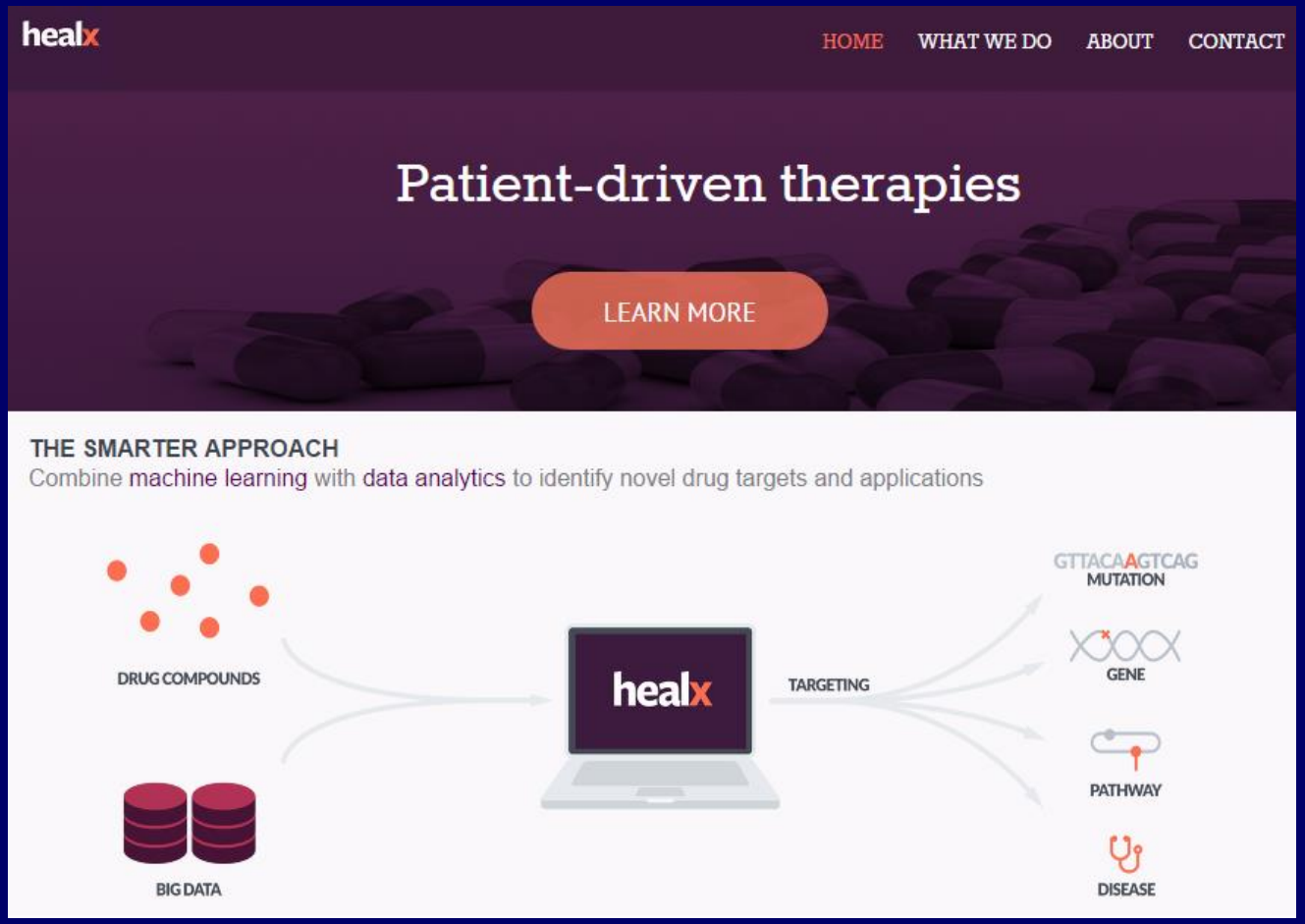
Control

Compound



Startup ‘Healx’ founded, for ‘data-driven drug repurposing in rare diseases’

www.healx.io



- Emphasis on patient groups
- CEO Tim Guilliams, funded by Amadeus and others
- CUE ‘Life Science Startup of the Year’ 2015

Current problems

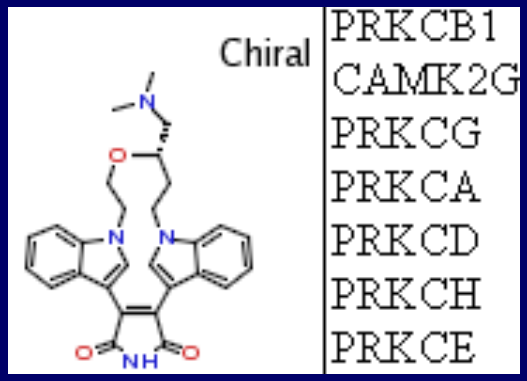
- Inconsistent annotations (eg protein target identifiers, *etc.*) of both chemical and biological data
- Chemical data often proprietary (in companies, commercial databases, etc.)
- Insufficient understanding of biological readouts (what do eg gene expression data ‘mean’? Where can they be used? *Etc.*)
- Unclear disease relevance of model systems (and hence unclear relevance of much of the data we have available)
- Etc.

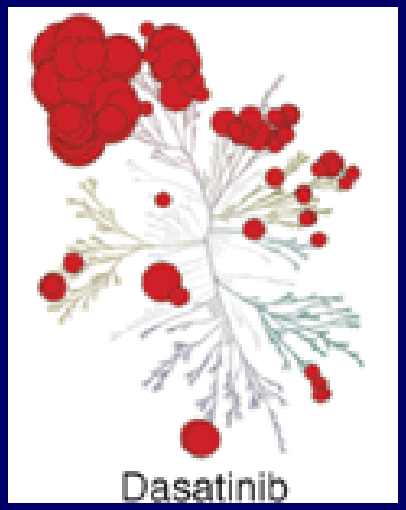
So how can we integrate data to help drug discovery?

- Data relating to drug discovery is *diverse, distributed, and often inconsistently annotated*

- *However,* on all levels more (and better structured) information is becoming publicly available

- *In our group we use this to*
  - *(a) select bioactive compounds*
  - *(b) understand compound action, and*
  - *(c) predict compound action ('personalized medicine')*





Acknowledgments

UNIVERSITY OF CAMBRIDGE
Department of Chemistry

Aakash Ravindranath
Ain Qurrat
Alexios Koutsoukas
Avid Afzal
Bobby Glen
Chad Allen
Daniel Murrell
David Marcus
Fatima Baldo
Fazlin Mohd Fauzi
Georgios Drakakis
Krishna Bulusu
Lewis Mervin
Oscar Mendez Lucio
Richard Lewis
Rucha Chiddarwar
Shardul Paricharak
Salundi Basappa
Sharif Siam
Siti Zuraidah Sobir
Sonia Liggi
Sudeshna Guha Neogi
Yasaman Motamedi

Universiteit Leiden
The Netherlands

Ad P. IJzerman
Bart Leidselink
Gerard van Westen

BASF
The Chemical Company

Sebastian Rohrer
Stefan Tresch
Antje Wolf
Klaus-Juergen Schleifer

AstraZeneca

Ola Engkvist
Thierry Kogej

MERCK MILLIPORE

Martin Augustin
Tom Klenka

Institut Pasteur

Therese Malliavin
Isidro Cortes

erc

Unilever

Ian Stott

Johnson & Johnson
PHARMACEUTICAL RESEARCH & DEVELOPMENT, L.L.C.

Hinrich Goehlmann
Herman van Vlijmen
Joerg K. Wegener

Lilly

Mike Bodkin
David Evans
Suzanne Brewerton

Aboca

Massimo Mercato
Anna Maidecchi

UNIVERSITI TEKNOLOGI MARA

CONACYT

ISLAMIC DEVELOPMENT BANK - BANQUE ISLAMIQUE DE DEVELOPPEMENT

EPSRC
Engineering and Physical Sciences Research Council

NWO

cefic

BBSRC
bioscience for the future

IWT