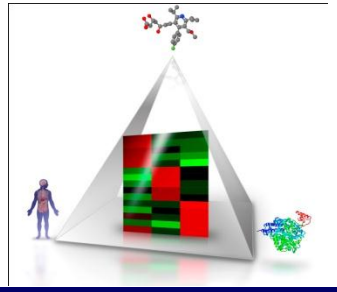
Merging Chemical, Biological and Phenotypic Data to Support Decision Making

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



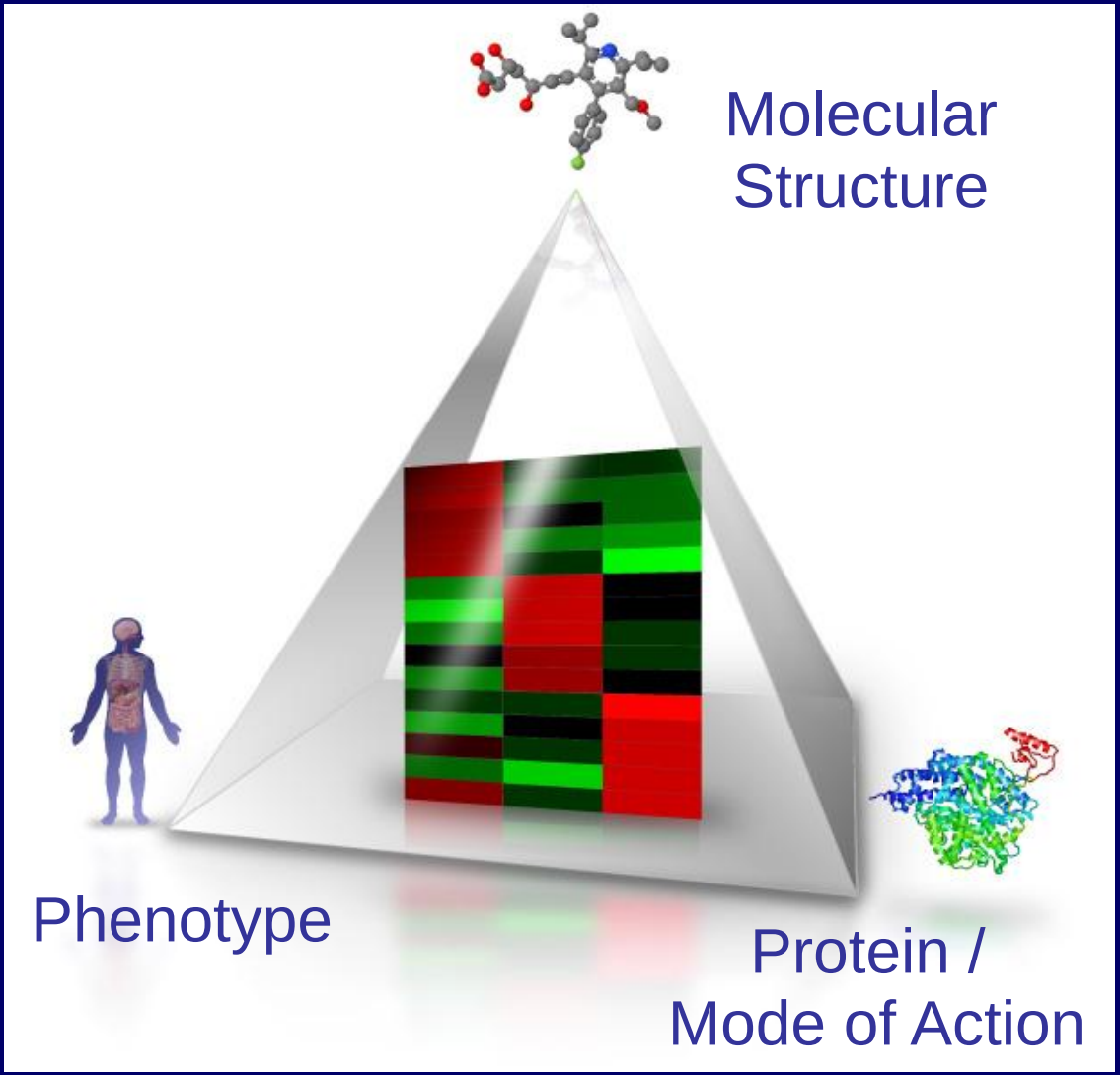
More and More Data is Available...

- 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 do in our group in Cambridge; integrate and analyze *heterogeneous* life science data

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



So what's the point of it all?

- "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 respond better (or with adverse effects!) to compound E or F?"
  -> *Phenotype*

Group Structure

- Close to 20 people (ca. 3 postdocs, 14 PhD students, plus visitors etc.)
- Funding from ERC, BBSRC, EPSRC, CEFIC; BASF, Eli Lilly, Johnson&Johnson, AstraZeneca, Unilever, Aboca, ... Plus close to 1/3 personal scholarships

- Idea:
  - 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

Group Research Organized in Clusters

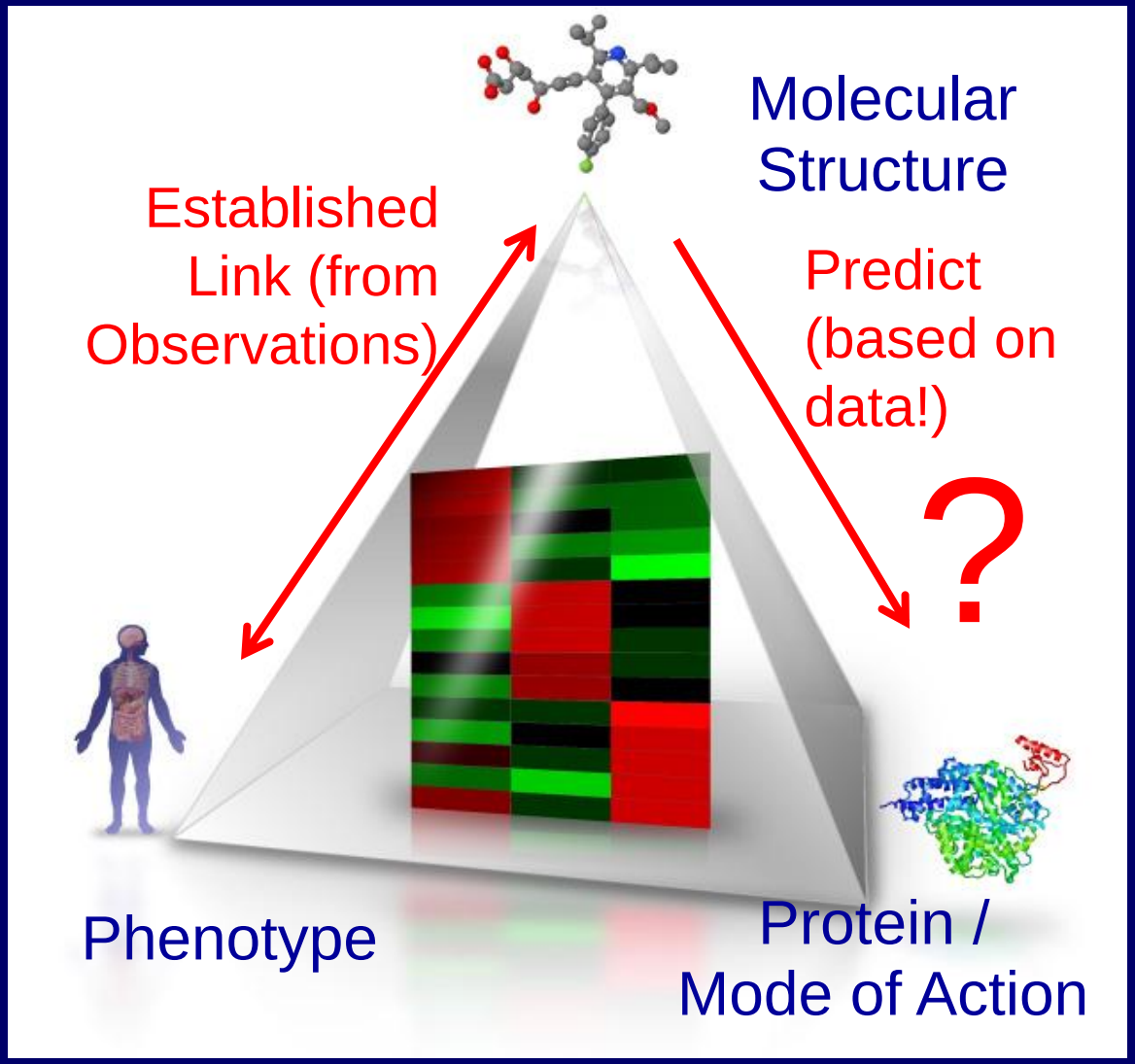
Current:

- Mode-of-action analysis (‘target prediction’); ca. 7 FTE (both methods development and applications)
- Modelling of bioactivities against target families (‘proteochemometrics’); ca. 3 FTE
- Natural products/traditional medicines; ca. 2 FTE
- Mixture modeling; ca. 2 FTE

Coming in the Future:

- Personalized medicine / pharmacogenomics / toxicogenomics; ca. 3 FTE
- Bioactive mixture modelling (for efficacy and toxicity); ca. 4 FTE (starting in 2/2014; ERC grant)

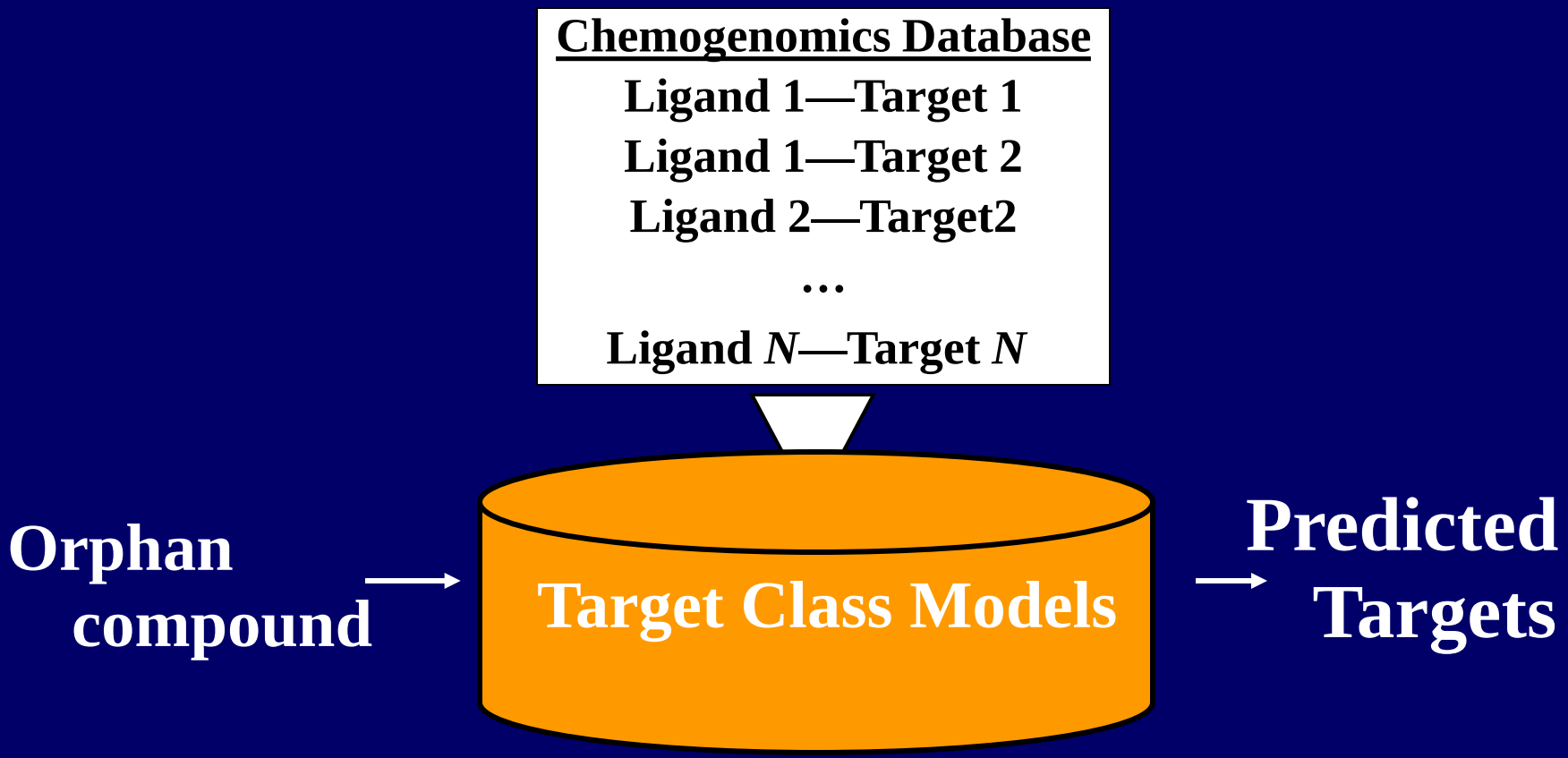
Starting from *in vivo* efficacy we can predict the MoA, based on ligand chemistry



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.

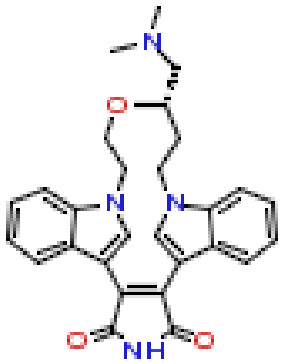


Prediction Examples: Gleevec, Ruboxistaurin

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

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

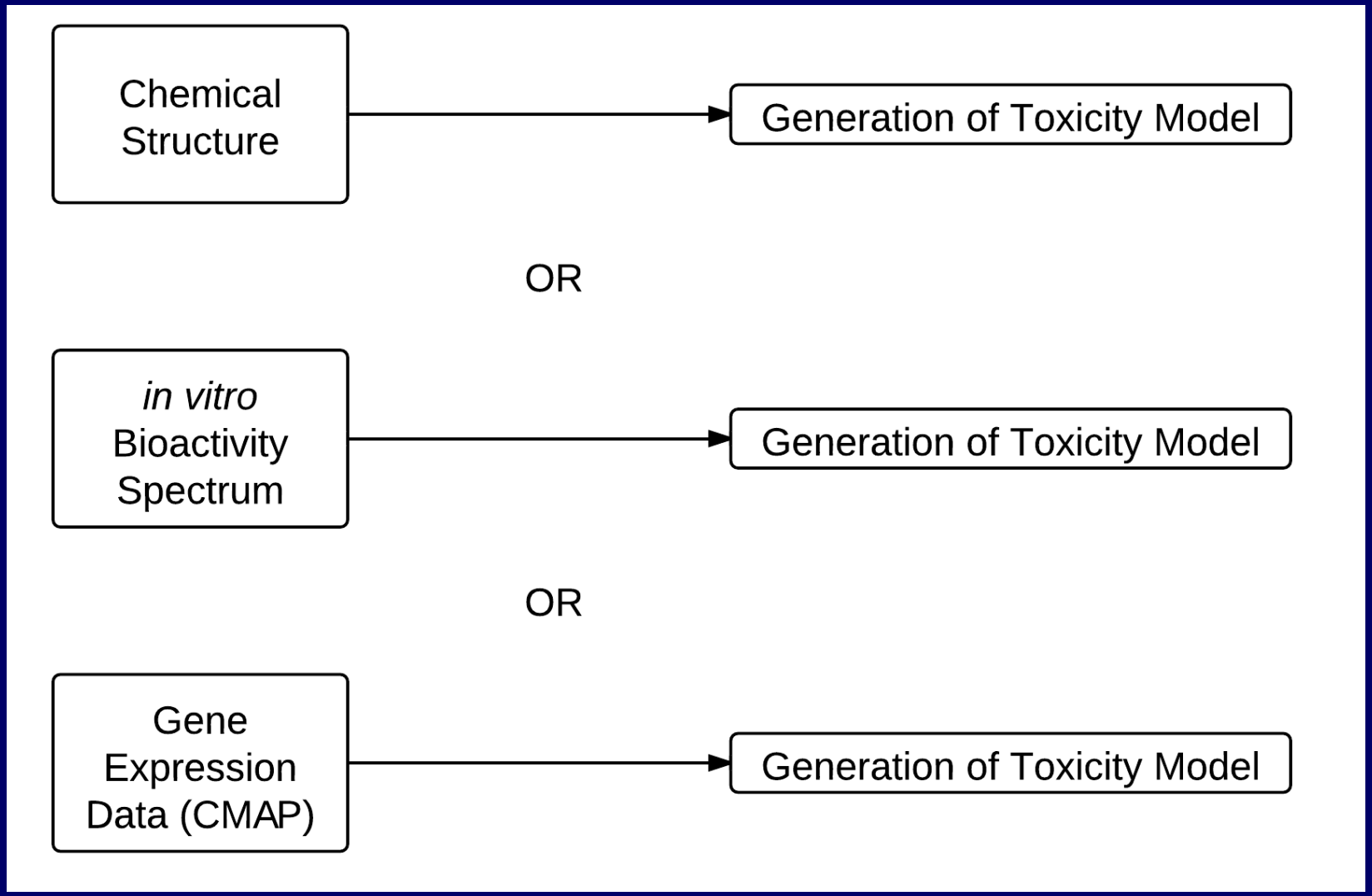
Molecule	Targets	Scores
	ABL1	46.50
	PDGFRB	28.99
	KIT	22.02
	CDK9	21.30
	BRAF	16.13
	FLT1	13.09
	PLK1	8.05
	BTK	5.44

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

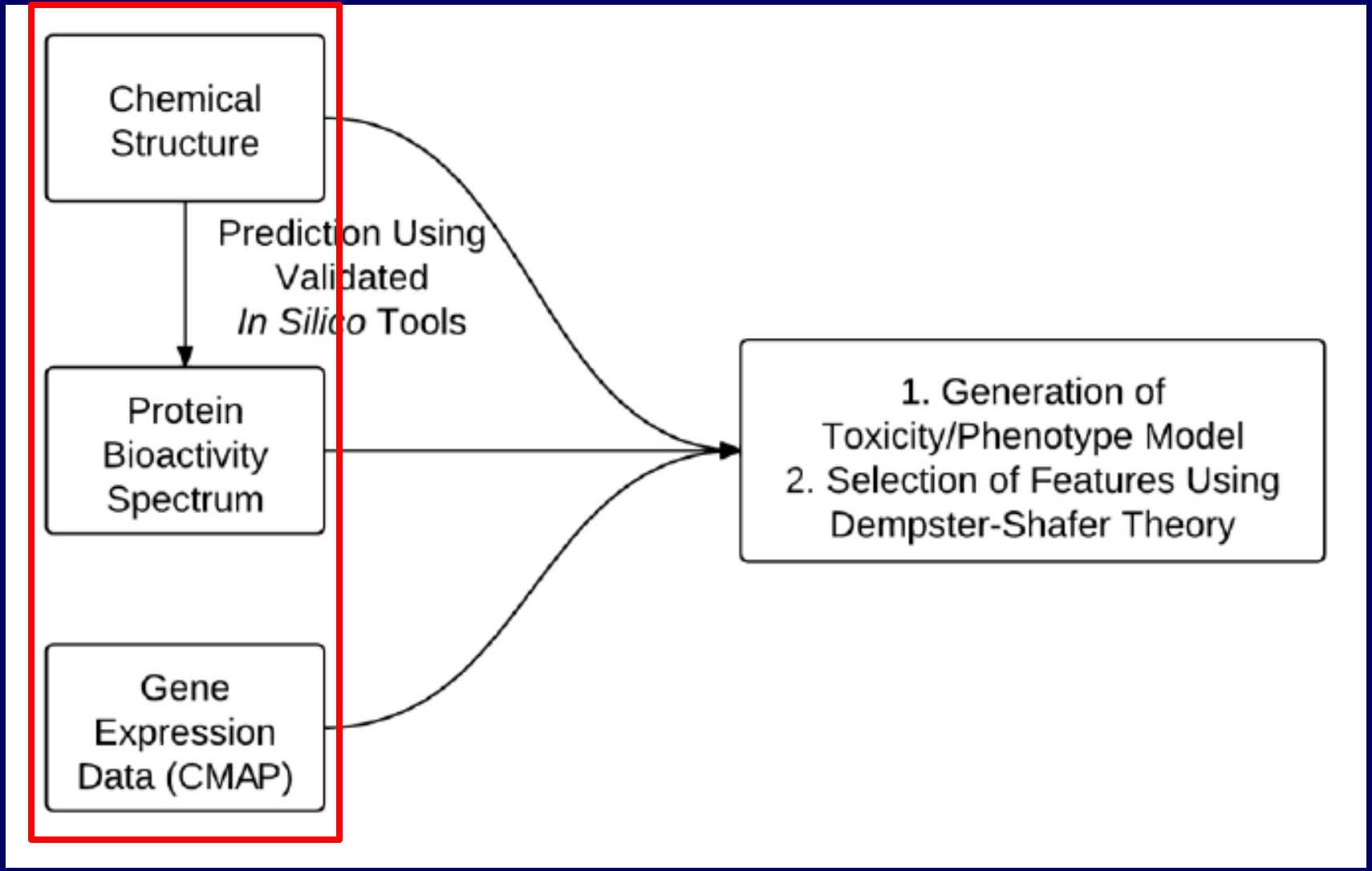
Where we can we apply this concept? To many phenotypes, such as ...

- Gene Expression Data (with Johnson&Johnson)
- Malaria Cellular Hits (NWO)
- Modes of Actions of Agrochemicals (BASF)
- Traditional Chinese and Indian Medicines
- Phenotypic Screening Data (AstraZeneca)
- Rat Sleep Data (Eli Lilly)
- 3D Kidney Toxicity Readouts (Leiden University)
- New Approaches to Toxicity Modeling (CEFIC)
- ‘Legal Highs’ (EMCDDA)
- Adverse Drug Reactions (Internal)
- Xenopus Developmental Assays (UEA)
- ...

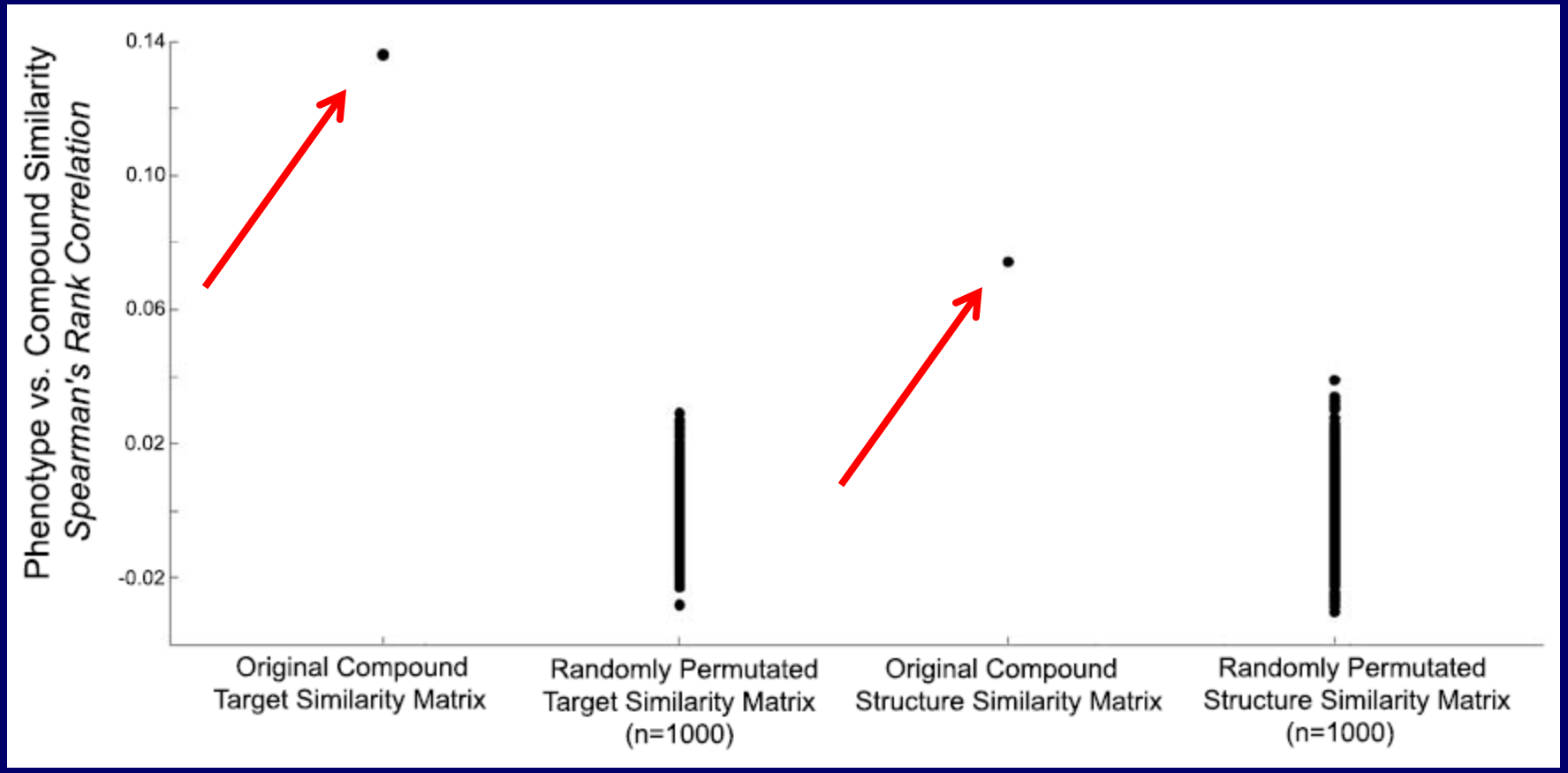
What Has Been Done *Previously*? Toxicity Models Are/Were Usually Based on *Single-Domain* Data



Not ‘OR’ – Domains Have Different Information Contents, Hence *AND*!



Higher Correlation of Predicted Targets Than Structures With Phenotypes



Young, Bender et al., Nature Chem. Biol. 2008, 4, 59 - 68

First evaluation: So does it work? (Work of Chad Allen, funded by LRI Award)

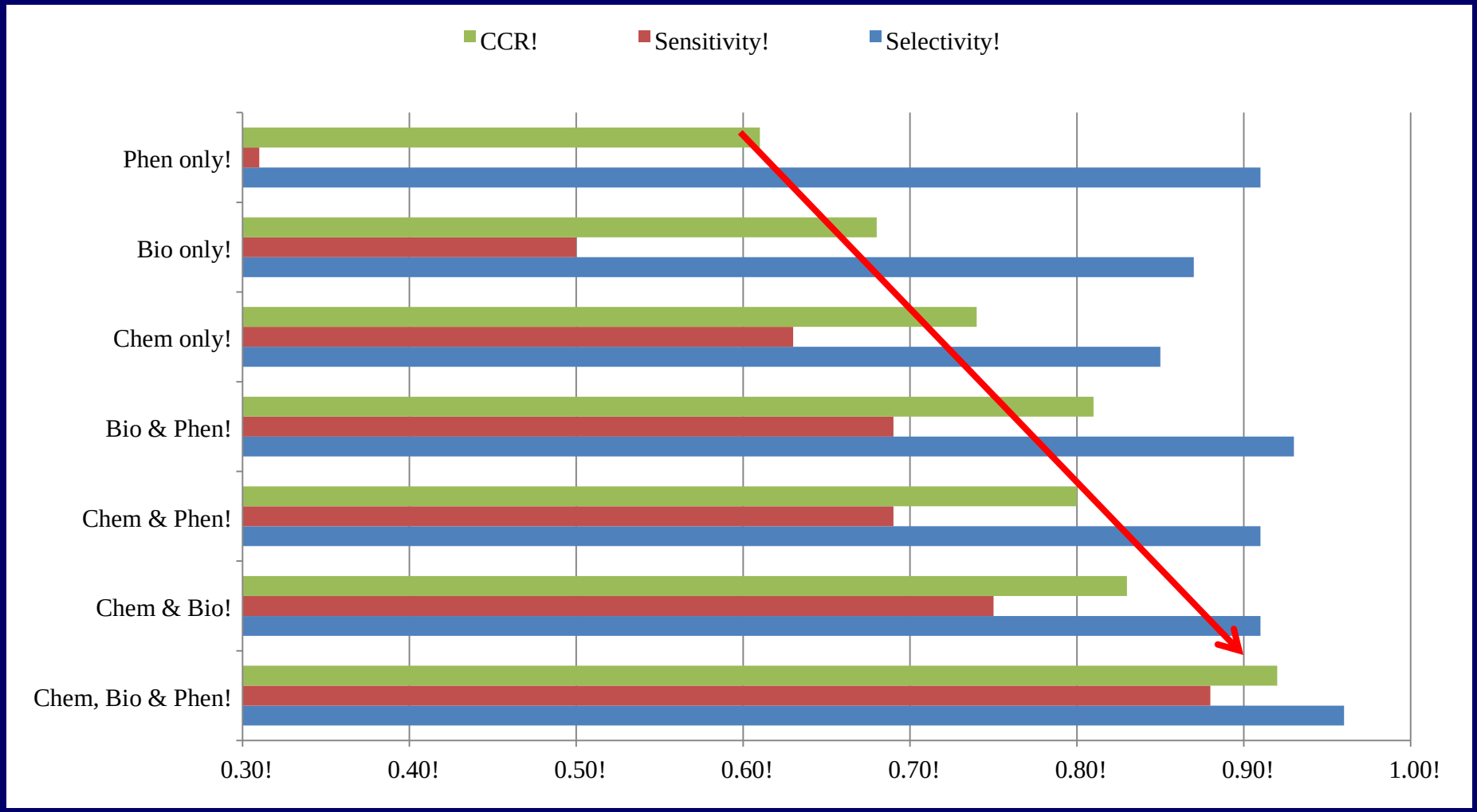
- Dataset of 362 compounds, classified as either toxic or nontoxic using rat LD50 data.
- Tripartite descriptor set:
  - 2D and physiochemical molecular descriptors (from PaDEL-Descriptor)
  - *Protein-affinity target scores against panel of 477 human proteins*
  - PCA of HTS concentration-response cell viability curves for 13 cell lines (originally from PubChem, pre-processed to remove noise by collaborators from UNC)*

* A. Sedykh, H. Zhu, H. Tang, L. Zhang, A. Richard, I. Rusyn, and A. Tropsha, Environ. Health Perspect., 2011, 119, 364–370.

Performance comparison (Chem/Bio/Phen = Chemical/Biological/Phenotypic Descriptors)

	Selectivity	*Sensitivity*	*CCR*
Chem, Bio & Phen	0.96	0.88	0.92
Chem & Bio	0.91	0.75	0.83
Chem & Phen	0.91	0.69	0.80
Bio & Phen	0.93	0.69	0.81
Chem only	0.85	0.63	0.74
Bio only	0.87	0.50	0.68
Phen only	0.91	0.31	0.61

Inclusion of Descriptor Types *Continuously Increase Performance* (CCR = Correct Classification Rate)



Take-home message

- We have lots of data on compounds and their properties – and we should make use of it!

- Including ‘knowledge-based’ bioactivity spectra as descriptors significantly increases correct classification rate on toxicity dataset used

- This is relevant both from the statistical/performance angle, as well as regarding biological interpretation (protein targets *have a meaning*!)

- Future work will use additional datasets; establish ‘Applicability Domain’ of model; consider dose and PK/PD properties; …

Acknowledgments

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

Gerard J. P. van Westen
Ad P. IJzerman
Bart Leidselink

Sebastian Rohrer
Klaus-Juergen Schleifer

Ola Engkvist

Ian Stott

Hinrich Goehlmann
Herman van Vlijmen
Joerg K. Wegener

Mike Bodkin
David Evans
Suzanne Brewerton

Martin Augustin
Tom Klenka

Therese Malliavin
Isidro Cortes