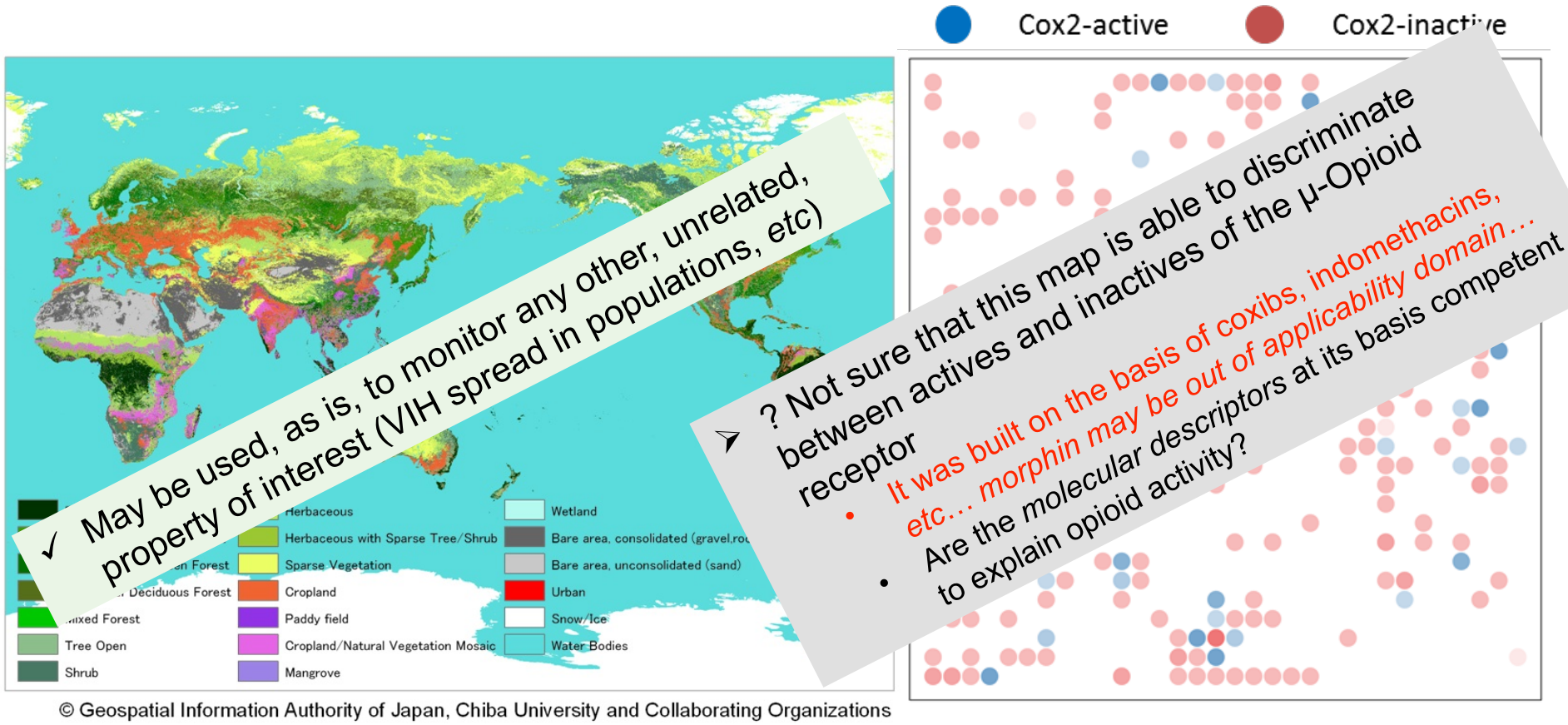
TOWARDS A UNIVERSAL MAP OF DRUG-LIKE SPACE

SIDOROV Pavel, GASPAR Helena, MARCOU Gilles,
VARNEK Alexandre & HORVATH Dragos

Laboratory of Chemoinformatics,
University of Strasbourg

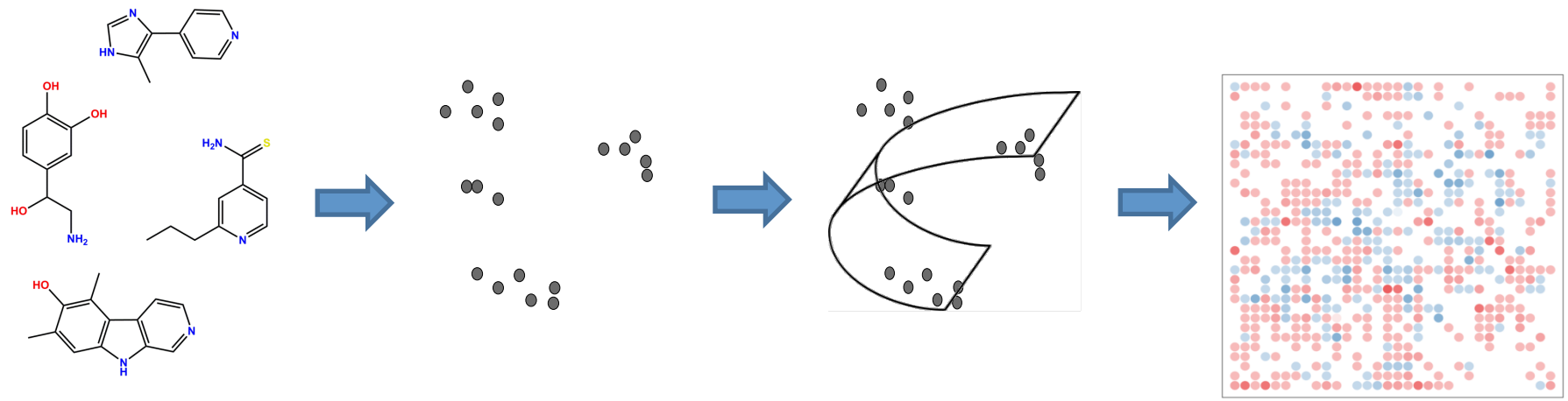
Geographic *vs.* Chemical Space (CS) Maps – the subtle difference…



- CS maps are not ‘universal’ (pluricompetent with respect to arbitrary properties) - but some are closer to this ideal…

Generative Topographic Mapping: a brief review

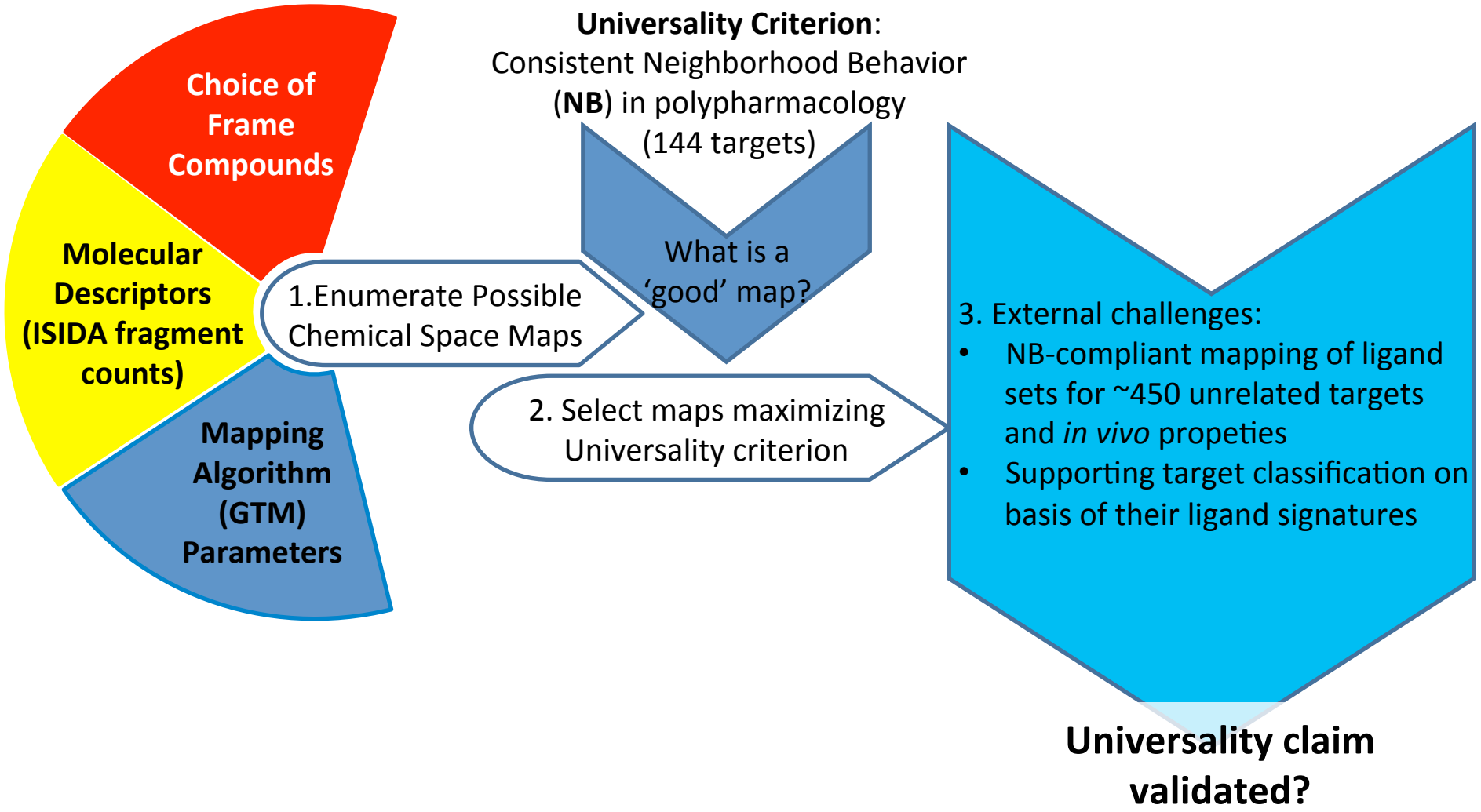
H. Gaspar et al. JCIM, 53, 12, 2013



'Frame' Compounds | *N*-dimensional Chemical Space | 2D manifold in CS | 2D map

- Molecules are represented in *n*-dimensional descriptor space
- A flexible 2D manifold is injected, molecules are projected on it
- Manifold is unbent into *KxK* square grid of nodes – 2D map
- Molecules are fuzzily associated to each node: association probabilities are called *responsibilities.*
- Similar responsibilities imply similar properties: regression & classification models

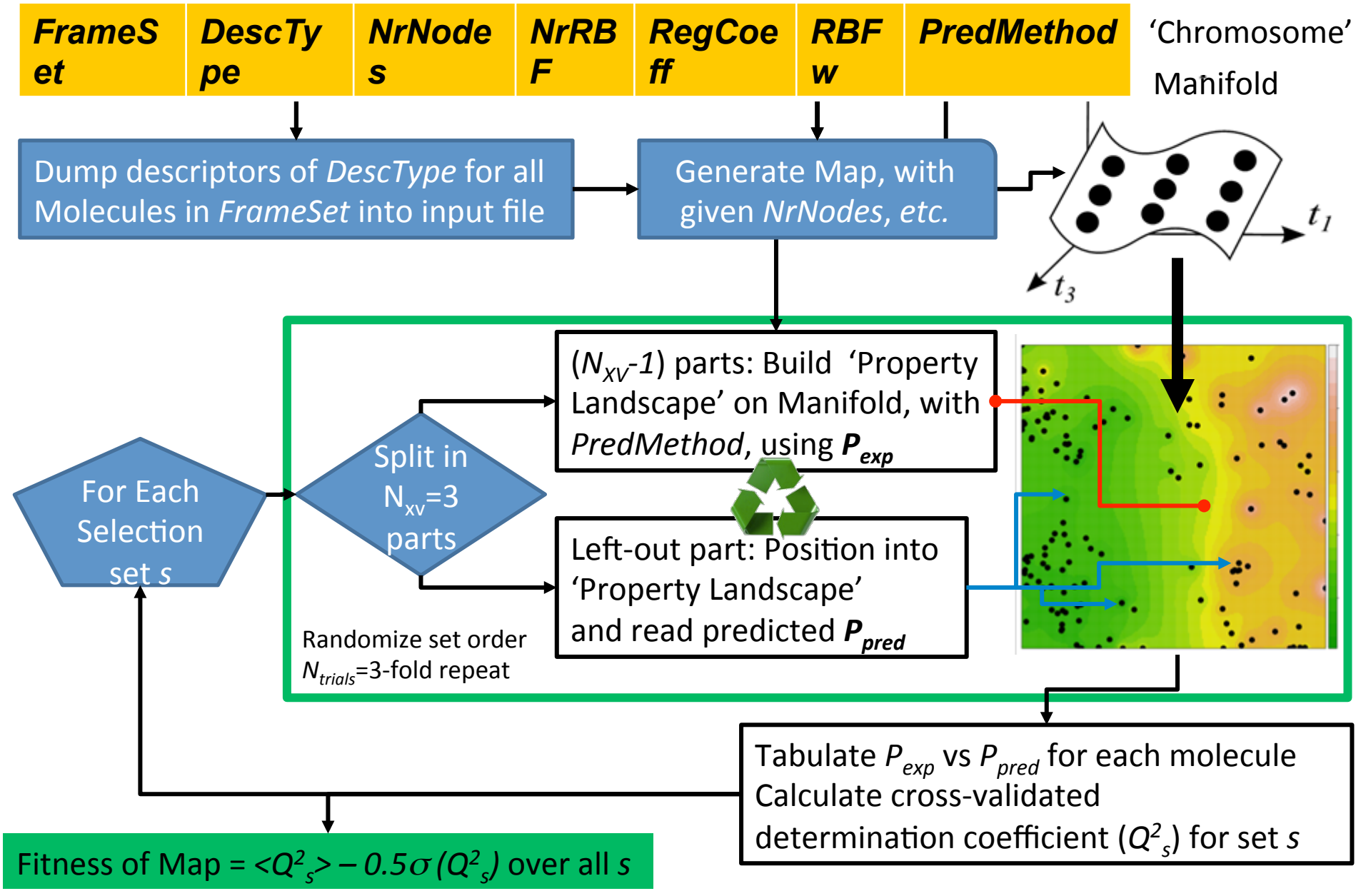
Searching for the Universal CS Map…



'Training' Data Sets…

- **Selection set**, for Universality criterion estimation:
  - 144 ChEMBL target-specific compound series, all larger than 50 compounds, curated and provided by Prof. J. Bajorath. Set members are all the compounds with reported pK_i values with respect to the associated targets (receptors, enzymes, *etc*).
  - These sets are *modelable (*robust SVM models could be obtained for each).
- **Frame sets**:
  - *Set 1*: a diverse set of 11K marketed drugs, biological reference compounds, ligands from PubChem database, as well as randomly picked ZINC compounds;
  - *Set 2*: a subset of the selection ChEMBL dataset where only *one-third* (but at least 50) of ligands of *each* target are included (9877 molecules);
  - *Set 3*: a subset of the selection ChEMBL dataset , where *half* of ligands for *half* of targets are taken (7214 molecules);
  - *Sets 4 and 5:* combinations of *Set 2* and *Set 3* with *Set 1*, e.g. fused sets labeled *Set1+2* and *Set1+3*, respectively.

Darwinian Evolution towards Universality…



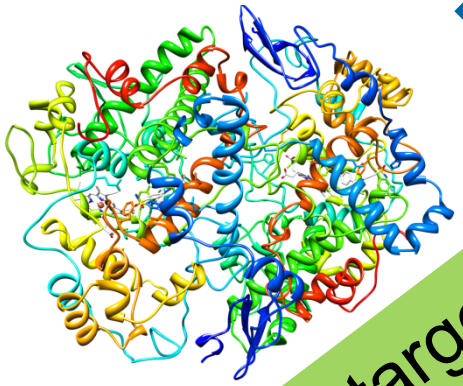
Selection of five maps with best Neighborhood Behavior (NB)

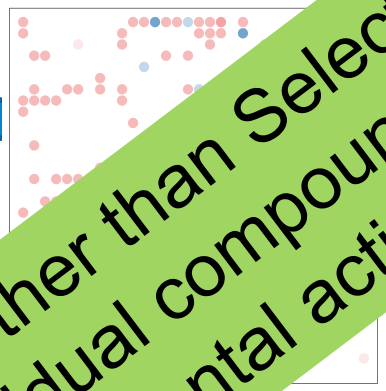
- Since selection sets feature quantitative affinity data (pK_i values), NB of maps can be expressed both in terms of regression and classification model proficiency…

Map	Descriptors & *Frame Set*
1	**IIRAB-PH-1-2**: Pharmacophore-colored atom-centered fragments, covering first and second coordination sphere; ***Set 3***
2	**IAB-FF-P-2-6:** CVFF Force-field-type-colored atom pairs at 1 to 5 bonds apart, including interposed bond information; ***Set 2***
3	**IA-FF-P-2-6:** as above, but without bond information; ***Set 3***
4	**IAB-PH-P-2-14:** Pharmacophore-colored atom pairs, at 1 to 5 bonds apart, including information on bonds nearest to terminal atoms; ***Set 2***
5	**III-PH-3-4:** Pharmacophore triplets, with edges of topological distances 3 and 4; ***Set 3***

External challenges for the five best maps…

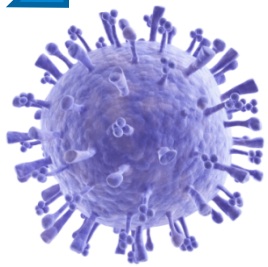
Target Binding
Discriminate Actives from (tested) Inactives, for 410 targets unrelated to Selection





emogenomics
hallenges: If targets are represented by cumulated responsi-bility vectors of their mapped ligands, would that support their classification into families?

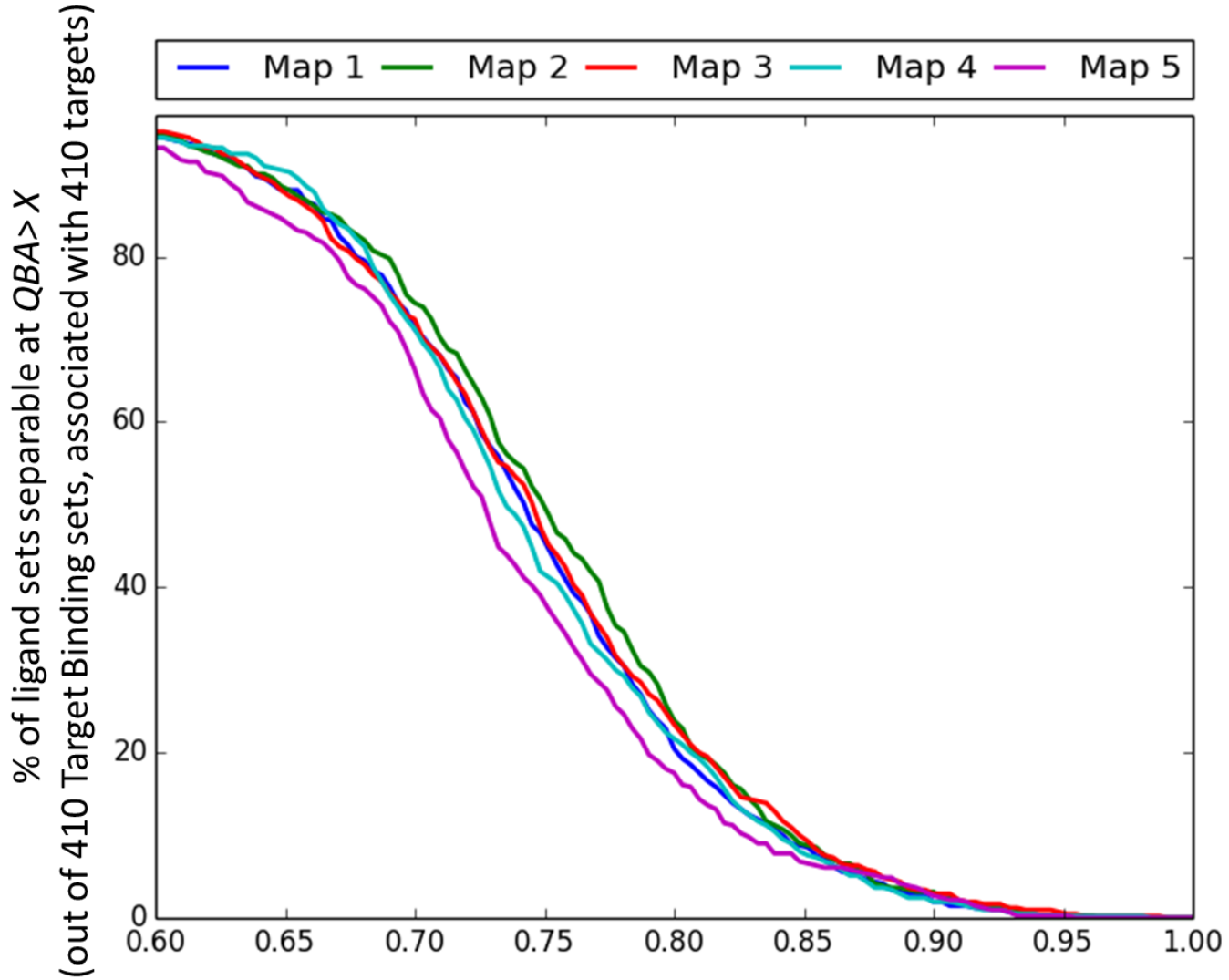
- > 450 targets (other than Selection targets),
- ~150,000 individual compound structures
- ~300,000 experimental activity values



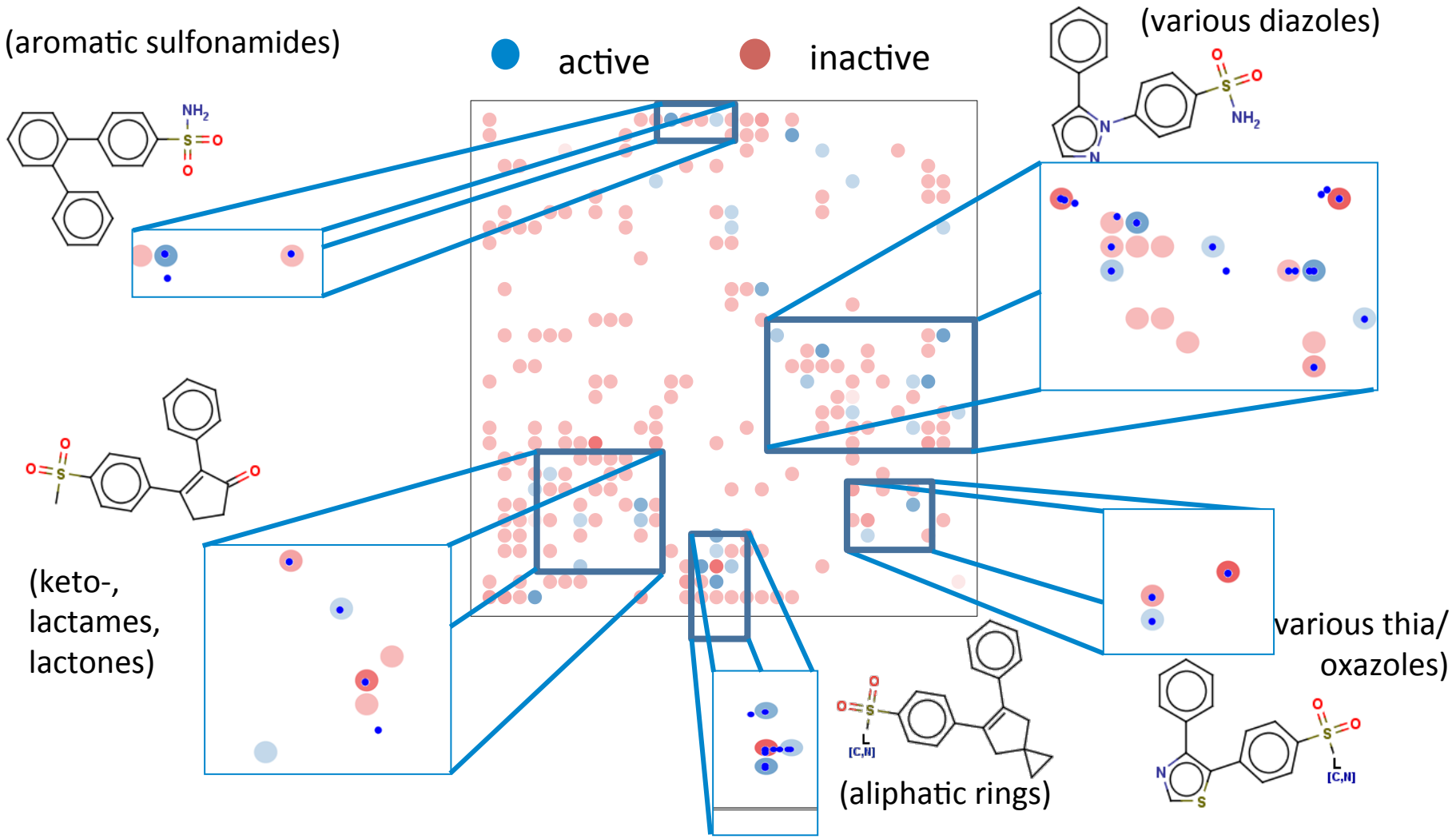
Antimala
Discriminate A om (tested) Inactiv or 17 antimalarial bioassays

Antiviral activity (7 virus families)
Distinguish the set of compounds acting on a given family from all the other antivirals

External challenge – Target binding..

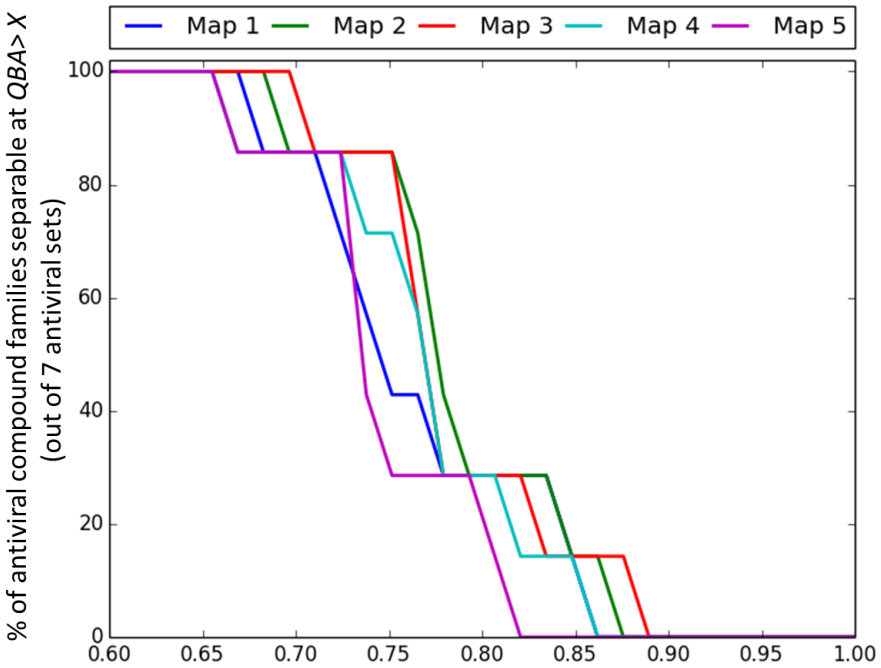


External challenge – Cox-2 ligands

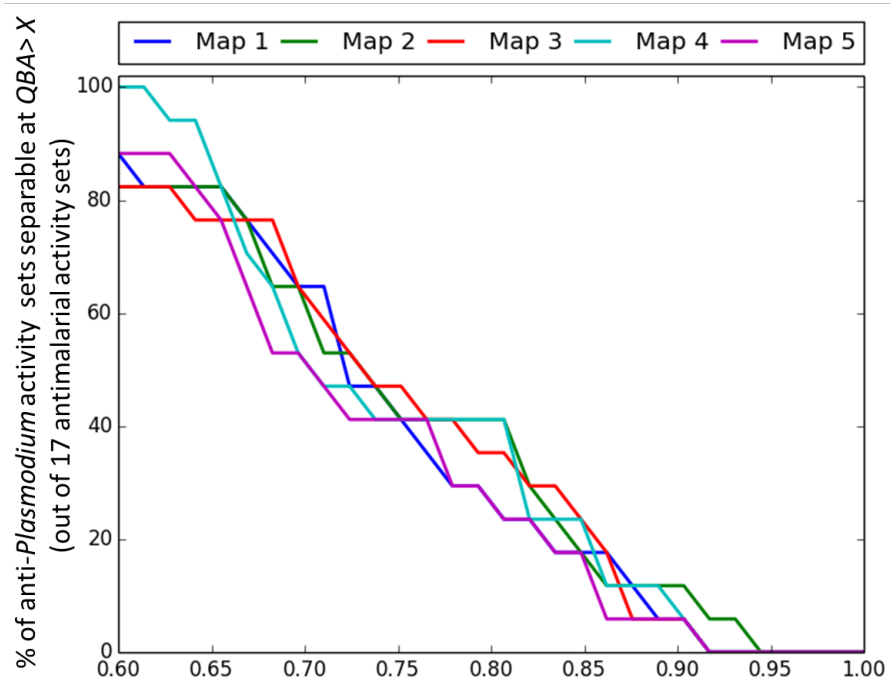


Map of ***Cox-2*** (CHEMBL230) ligands, cross-validated balanced accuracy = 0.7.
In the zoomed-in portions – common substructures for coxib-like ligands only

External challenge – Antiviral and Antimalarial compound recognition..



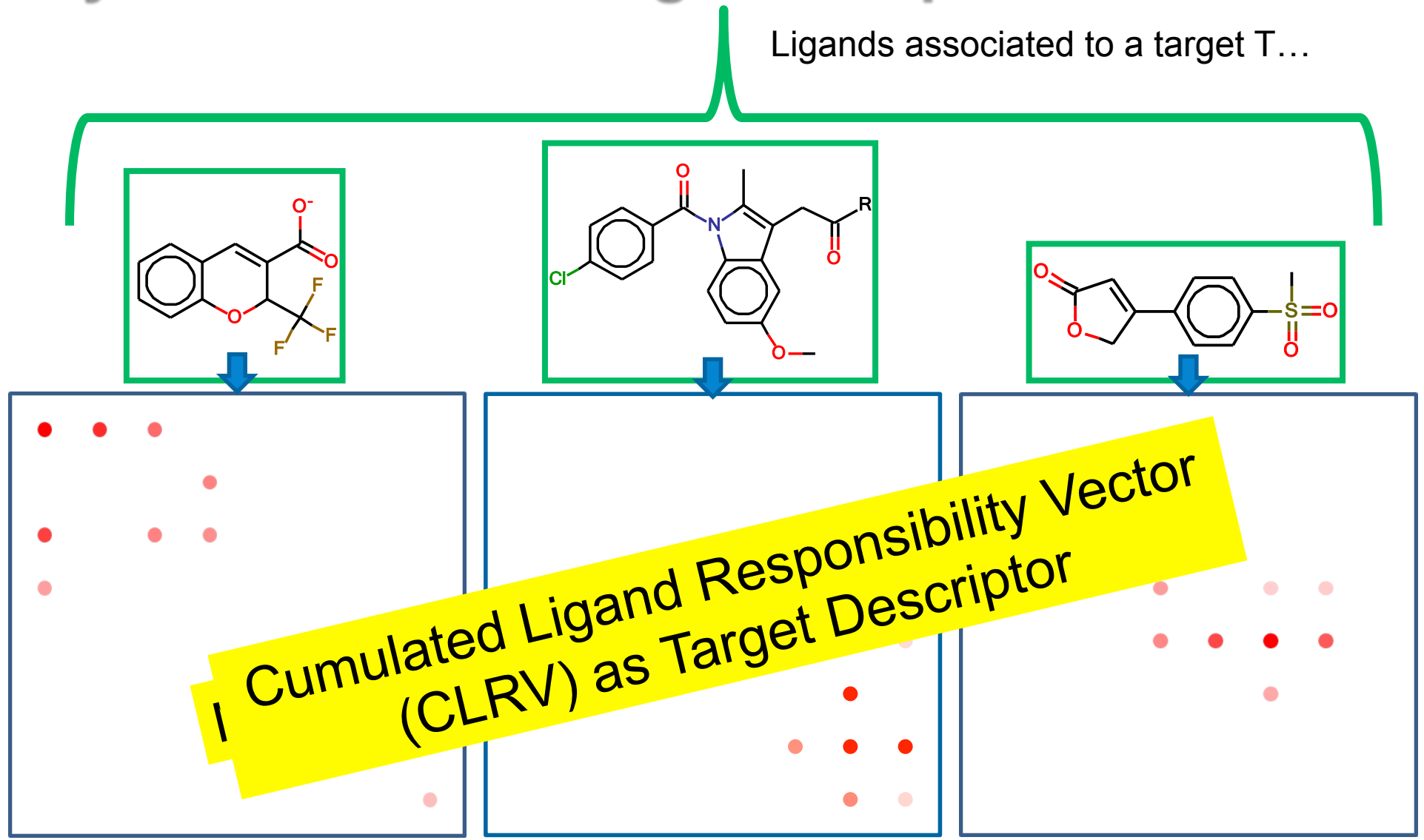
QBA in leave-1/3-out cross-validated activity class prediction, per ligand set



QBA in leave-1/3-out cross-validated activity class prediction, per ligand set

Chemogenomics Challenge: describe targets by their Cumulated Ligand Responsibilities

Ligands associated to a target T…



IF

Cumulated Ligand Responsibility Vectors (CLRV) are valid target descriptors

THEN

Targets within a family have, *on the average*, more similar CLRV than extra-family targets.

WHERE

Inter-target distance is the complement of CLRV Tanimoto scores.

Cohesion/Separation = mean intra/inter-family distances.

Super-family	Family	Family size	Shortest inter-target distance in family	Cohesion & StdDev in family	Shortest inter-target distance to others	Separation & StdDev to others	p value	
gpcr	Adenosine	4	0.182	0.442 ± 0.165	0.936	0.993 ± 0.011	1.00E-09	✔
kin	TK	35	0.035	0.408 ± 0.148	0.043	0.490 ± 0.237	1.00E-09	!?
gpcr	Serotonin	8	0.38	0.816 ± 0.164	0.54	0.977 ± 0.050	6.00E-07	✔
gpcr	Opioid	4	0.202	0.481 ± 0.263	0.917	0.991 ± 0.013	2.40E-05	✔
gpcr	Melano-cortin	4	0.322	0.644 ± 0.186	0.813	0.989 ± 0.022	5.60E-05	✔
gpcr	Prostanoid	8	0.099	0.823 ± 0.203	0.472	0.981 ± 0.046	7.40E-05	✔
gpcr	Dopamine	5	0.253	0.688 ± 0.266	0.54	0.972 ± 0.056	0.0021	✔
gpcr	EDG	4	0.378	0.713 ± 0.195	0.719	0.964 ± 0.053	0.0062	✔
gpcr	Nucleotide-like	6	0.182	0.774 ± 0.290	0.472	0.990 ±0.033	0.0074	✔
gpcr	Somato-statin	4	0.115	0.606 ± 0.308	0.502	0.977 ± 0.045	0.0086	✔
gpcr	Adrenergic	7	0.035	0.798 ± 0.314	0.804	0.984 ± 0.023	0.0091	✔
gpcr	Histamine	4	0.644	0.863 ± 0.112	0.756	0.965 ± 0.046	0.042	!
kin	CMGC	8	0.171	0.401 ± 0.138	0.067	0.448 ± 0.216	0.094	?
kin	AGC	12	0.102	0.570 ± 0.295	0.036	0.525 ± 0.272	0.23	?
kin	Src	6	0.17	0.412 ± 0.156	0.035	0.454 ± 0.217	0.33	?
kin	CAMK	9	0.053	0.438 ± 0.345	0.034	0.469 ± 0.286	0.6	?

Conclusions

- A strategy to choose parameters (both internal parameters of the method and meta-parameters like descriptor type and frame sets) has been applied to the GTM algorithm, in order to build generally applicable, universal – that is, polypharmacologically competent – maps.
- Selected maps were challenged to coherently separate actives from inactives, in projections of novel target-specific ligands, or *in vivo* tested compounds. Albeit nor the ligands, neither the targets were represented at map selection stage, the challenges were largely successful.
- Furthermore, Cumulated Ligand Responsibility Vectors produced when projecting target-specific ligand collections on these maps are coherent target descriptors, as they were seen to agree with accepted target classification schemes.
- Thus, the maps are surely perfectible, but general and robust 'Universal' representations of CS… as reflected in today's ChEMBL database.

Datasets:
***Prof. J. Bajorath**, University of Bonn*
***B. Viira**, University of Tartu*
***K. Klimenko**, University of Strasbourg*
& A.V. Bogatsky Physico-chemical Institute
***T. Gimadiev**, Kazan Federal University*
& University of Strasbourg

Number Crunching:
High Perfomance Computing Centers
*at the **Universities of Strasbourg** (EL) and **Cluj** (RO)*

Thank you for your attention