b-it LSi
Life
Science
Informatics

LIMES
Life and Medical Sciences Bonn

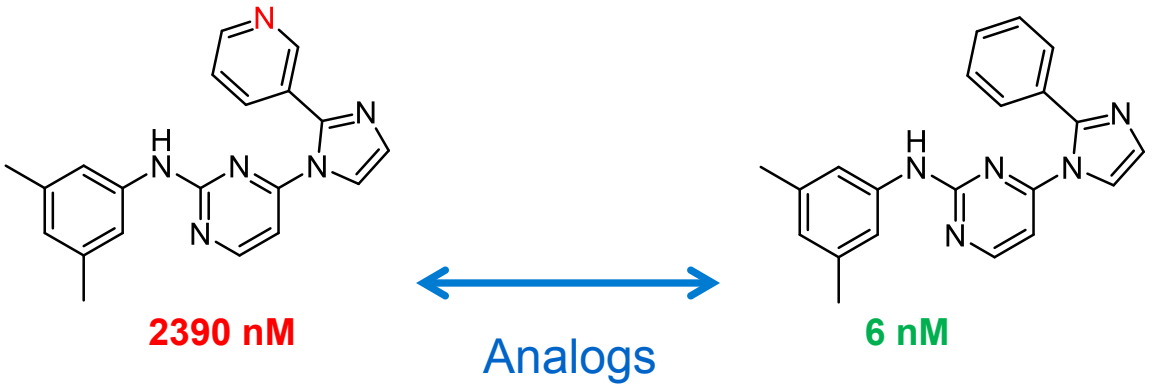
Exploring Activity Cliffs from a Chemoinformatics Perspective

Jürgen Bajorath
Life Science Informatics
University of Bonn

universitätbonn

Activity Cliff Concept

- *Activity cliff* is generally defined as a pair of structurally similar active compounds with a large difference in potency



Paradigm:

"small chemical modifications – large biological effects" ➔
high SAR information content

Activity Cliffs in Medicinal Chemistry

- Utility in SAR analysis and compound optimization
- Which compound to make next?
- Typically focused on individual compound series
- Methodological simplicity and chemical intuition are key to practical utility in med. chem.

Activity Cliffs in Chemoinformatics

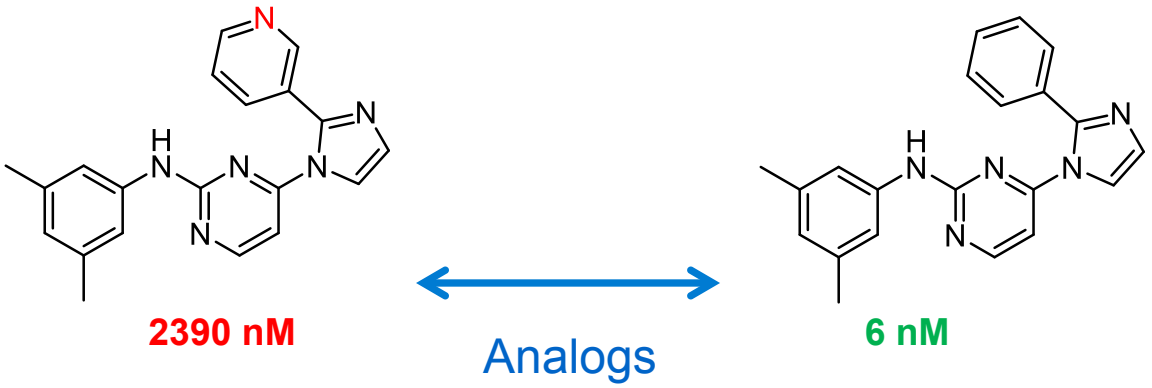
- Much stronger emphasis on methodological aspects
- Departure from individual series toward global analysis

Activity Cliffs in Chemoinformatics

- Molecular representation dependence
- Large-scale compound data mining
- Activity cliff networks
- Prediction of activity cliffs

Activity Cliffs

- *Activity cliff* is generally defined as a pair of structurally similar active compounds with a large difference in potency



Definition requires consideration of:

Similarity criterion

Potency difference criterion

Activity Cliff Definition

- **Alternative similarity criteria**

  Fingerprint Tanimoto similarity

  MACCS Tc 0.85, ECFP4 Tc 0.55

  Substructure-based similarity

  Matched molecular pairs, scaffolds

- **Potency difference criterion**

  Usually at least 1 or 2 orders of magnitude (10- or 100-fold)

1. Molecular Representations

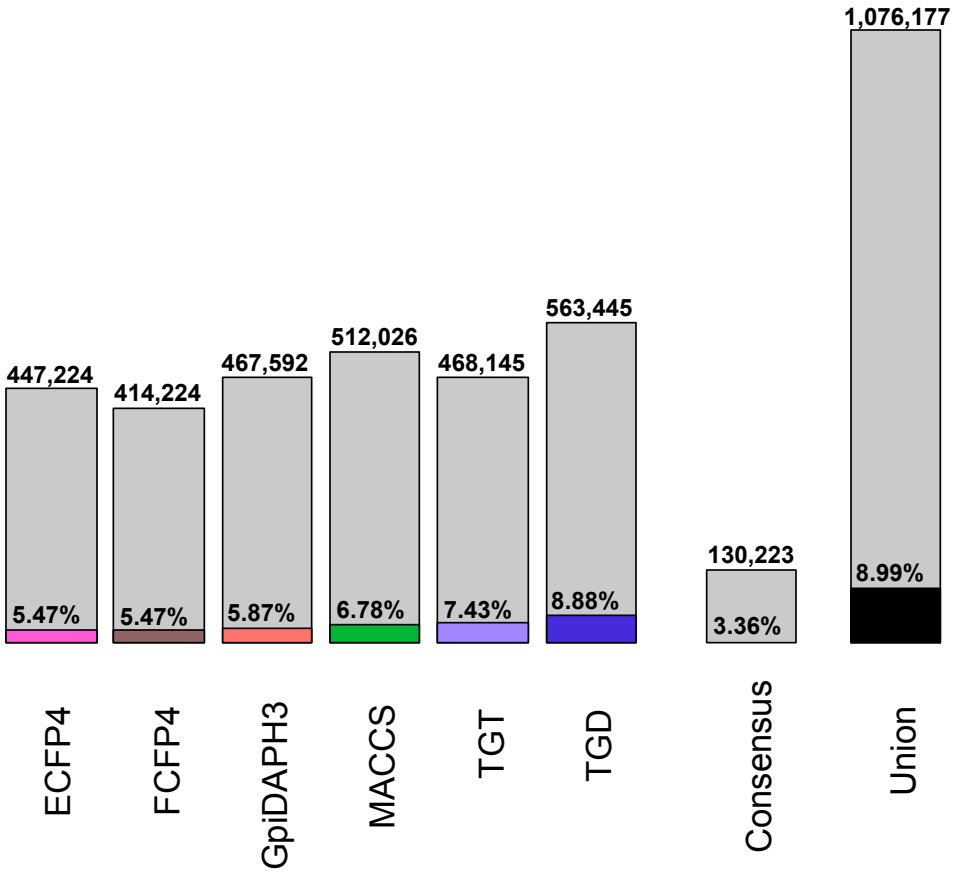
- Activity cliff distribution is strongly influenced by selected molecular representations and similarity criteria

- **Qualifying pairs** (**QP**s)
  - QPs are compound pairs exceeding a given similiarity threshold

- **Activity cliff frequency**
  - percentage of QPs with a more than 100-fold difference in potency

Molecular Representation Dependence

- QPs and activity cliff distribution for six different fingerprints
- 128 activity classes from ChEMBL with more than 100 compounds
- 35,021 unique compounds

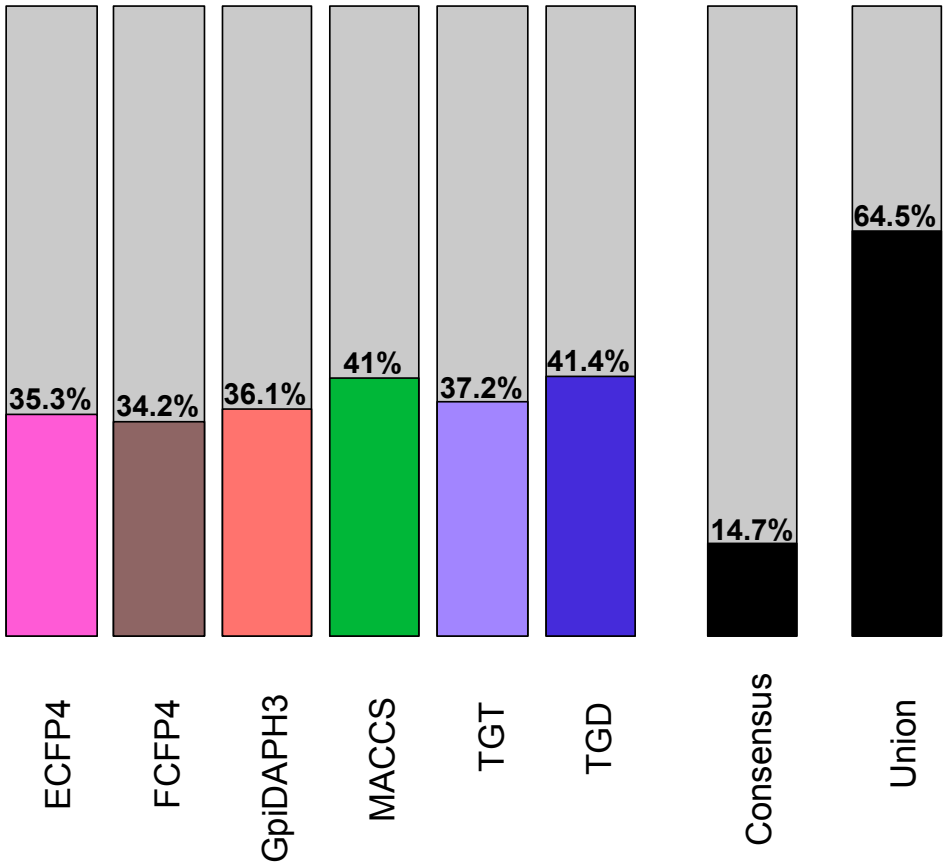


Stumpfe D, Hu Y, Dimova D & Bajorath J. J Med Chem, 57, 18 (2014)

Activity Cliff-Forming Compounds

- Percentage of compounds that form at least one activity cliff
- Union of cliff-forming compounds:

  More than 64% of all compounds form at least one cliff

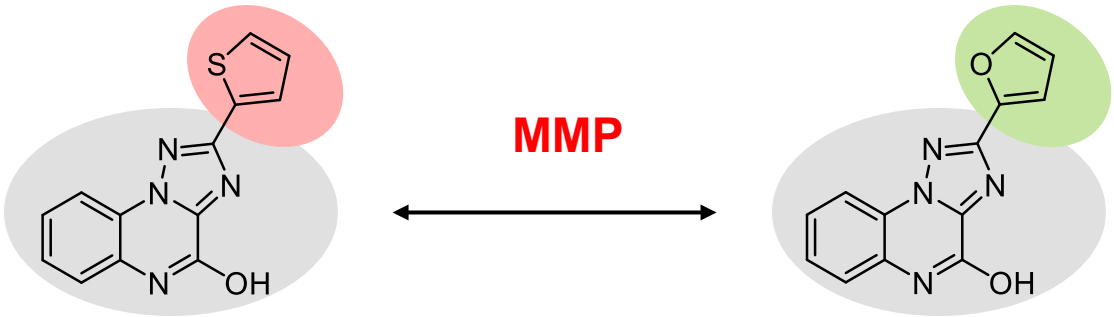


Stumpfe D, Hu Y, Dimova D & Bajorath J. J Med Chem, 57, 18 (2014)

128 activity classes (>100 cpds) from ChEMBL

MMPs as Molecular Represetation

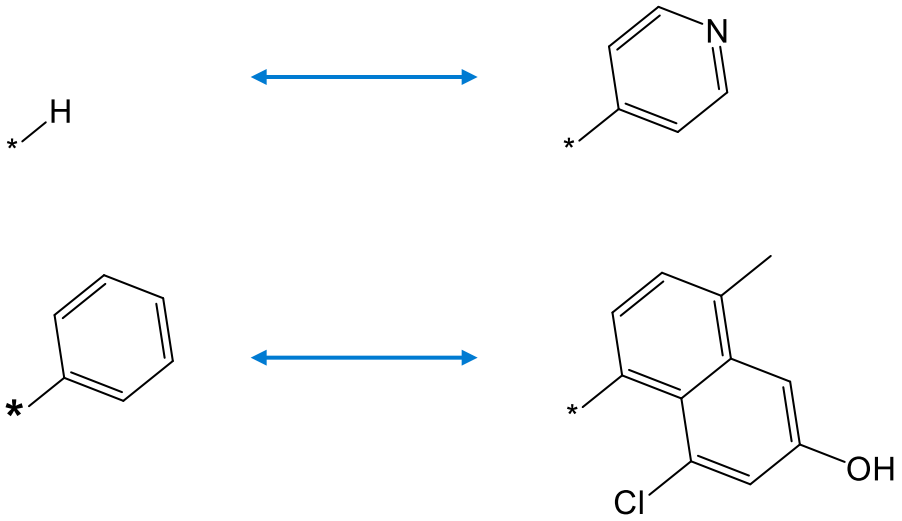
- A **M**atched **M**olecular **P**air (MMP) is formed by two structurally related compounds that
  - differ only by a small structural change at a single site
  - are related by the exchange of a substructure (termed chemical transformation)



Transformation Size Restriction

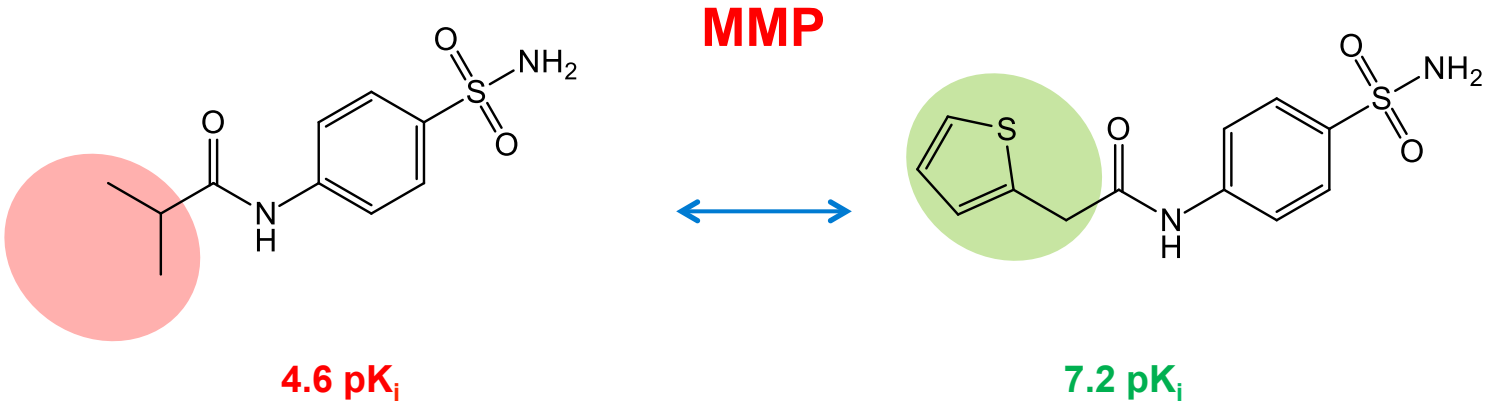
- **Transformation size-restricted MMPs** were introduced to limit transformations to small and chemically intuitive replacements

Examples of largest permitted transformations



Preferred Activity Cliff Definition

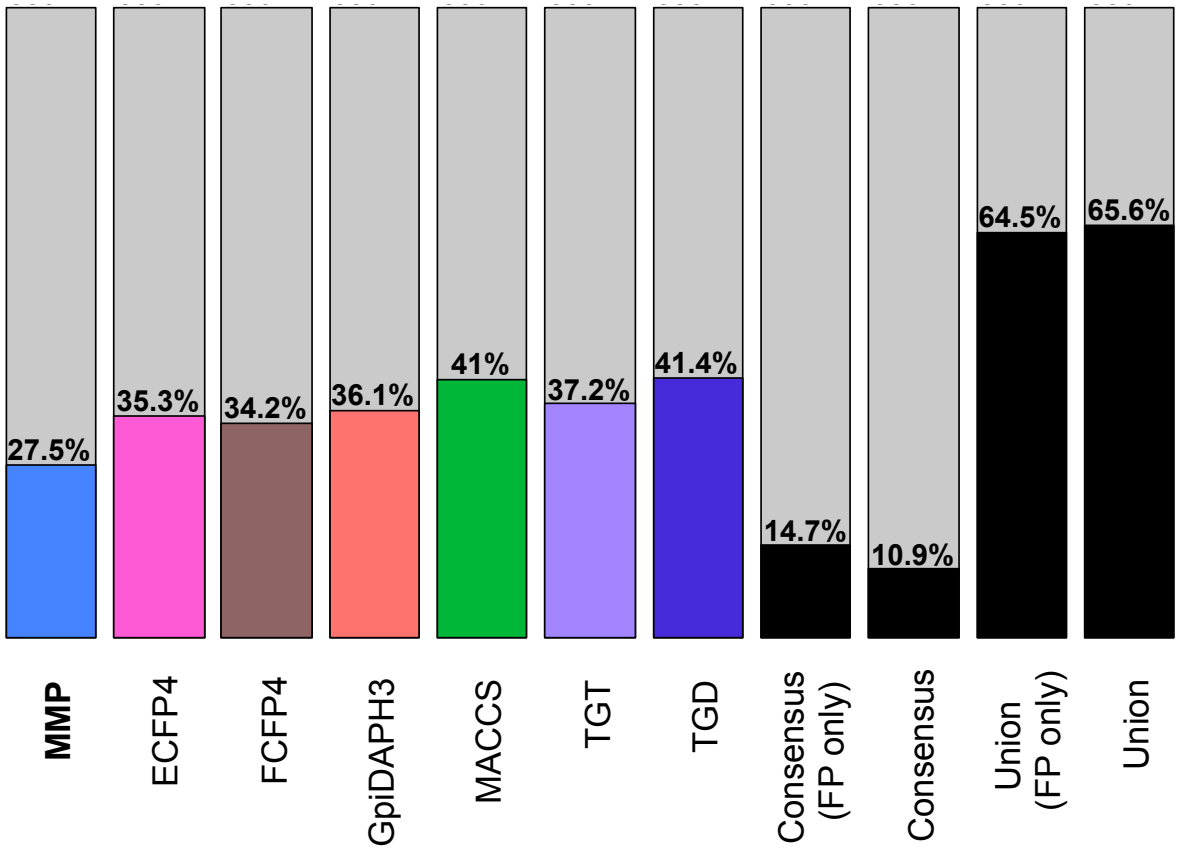
- Transformation size-restricted MMPs
  - substructure-based similarity assessment (med. chem. focus)
- At least 100-fold difference in potency
- Equilibrium constants (K_i)



Stumpfe D & Bajorath J. J Chem Inf Model 52, 2348 (2012)

Activity Cliff-Forming Compounds

- MMPs and six fingerprint representations
- MMPs yield smallest percentage of cliff compounds



Stumpfe D, Hu Y, Dimova D & Bajorath J. J Med Chem, 57, 18 (2014)

128 activity classes (>100 cpds) from ChEMBL

2. Large-Scale Data Mining

Proportion of bioactive compounds forming activity cliffs ?

Percentage of all bioactive compounds involved in the formation of activity cliffs (ChEMBL survey):

31.7% (ECFP4/Tanimoto-based cliffs)

22.8% (**MMP-cliffs**)

Large-Scale Data Mining

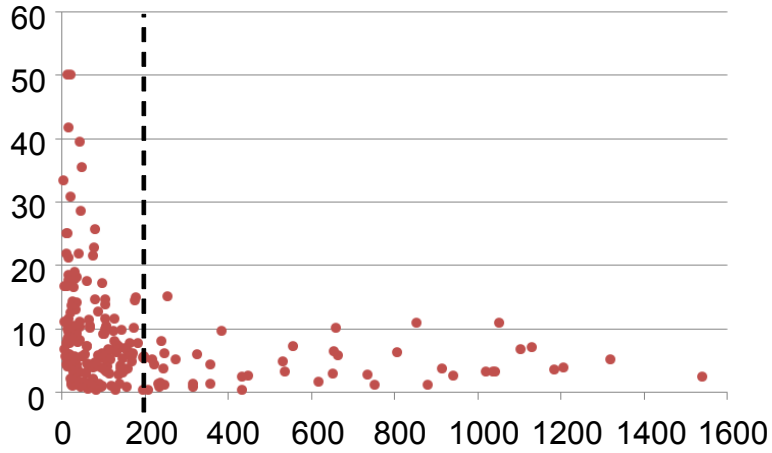
Currently available high-confidence activity cliffs ?

(ChEMBL version 17)

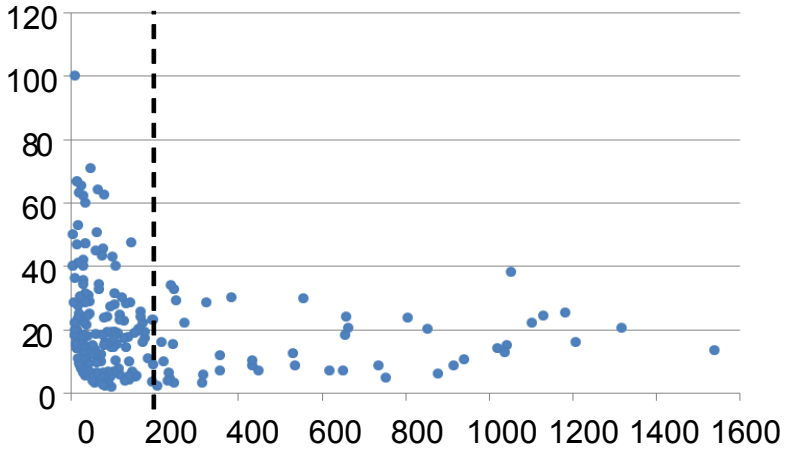
20,080 MMP-cliffs detected for 293 targets involving 11,783 unique active compounds

Target Distribution

% MMP-cliffs



% Cliff-forming compounds



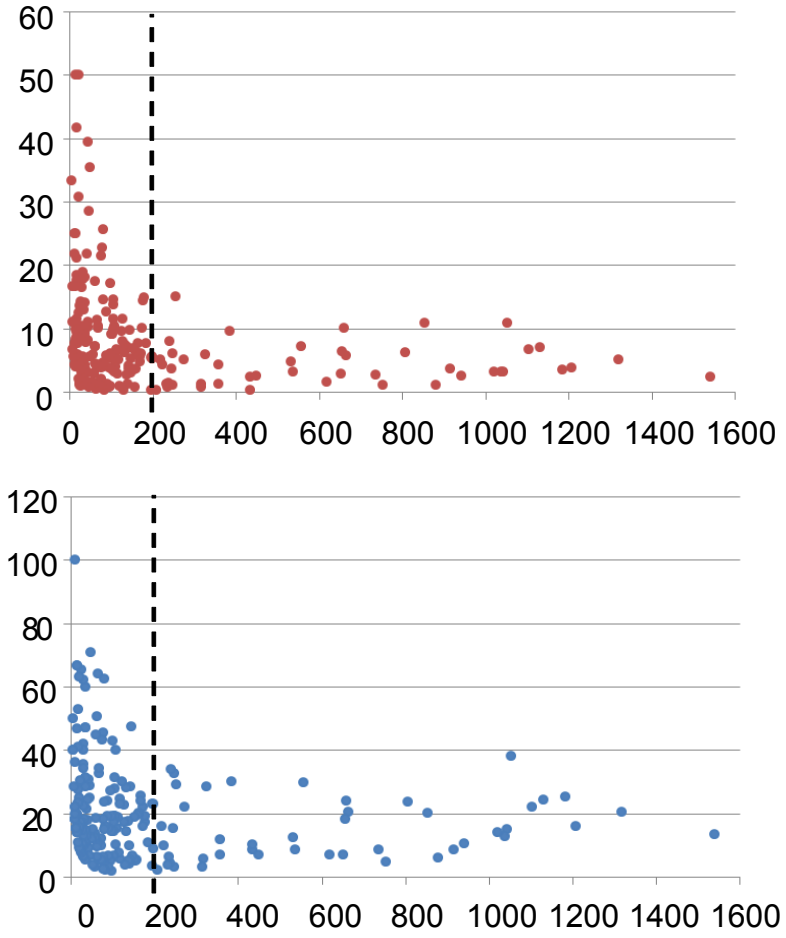
 # Compounds

Hu Y, Stumpfe D, Bajorath J. F1000Research 2, 199 (2013)

414 activity classes from ChEMBL

Target Distribution

For data set with >200 cpds, activity cliffs and cliff compounds are fairly evenly distributed among many different targets



414 activity classes from ChEMBL

Ligand Efficiency (LE) for MMP-Cliffs

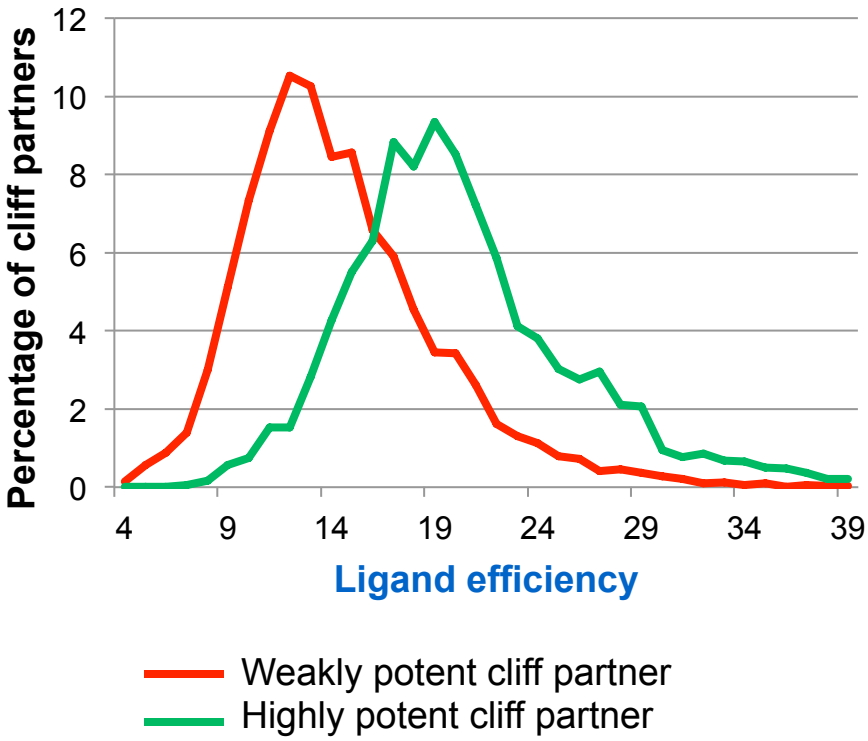
- Changes in **LE** accompanying activity cliff formation

LE = pKi / MW

- Difference in **LE** between weakly and highly potent cliff partners

- **LE increase** detected for **99.1%** of all activity cliffs;

  average **Δ LE = 6.27**



de la Vega de Leon A & Bajorath J. AAPS J 16, 335 (2014)

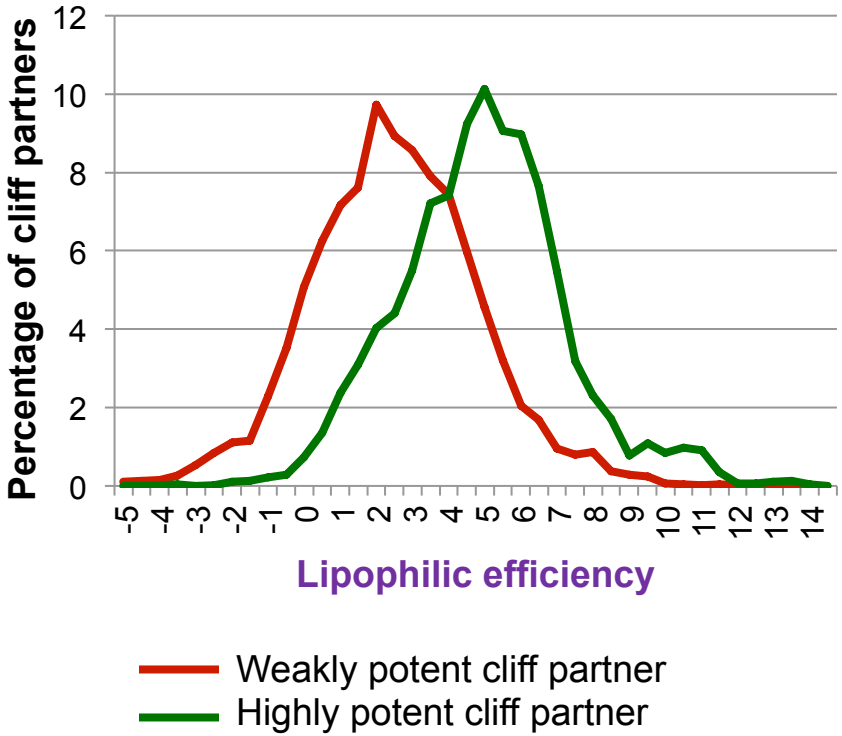
Lipophilic Efficiency (LipE)

- Changes in **LipE** accompanying activity cliff formation

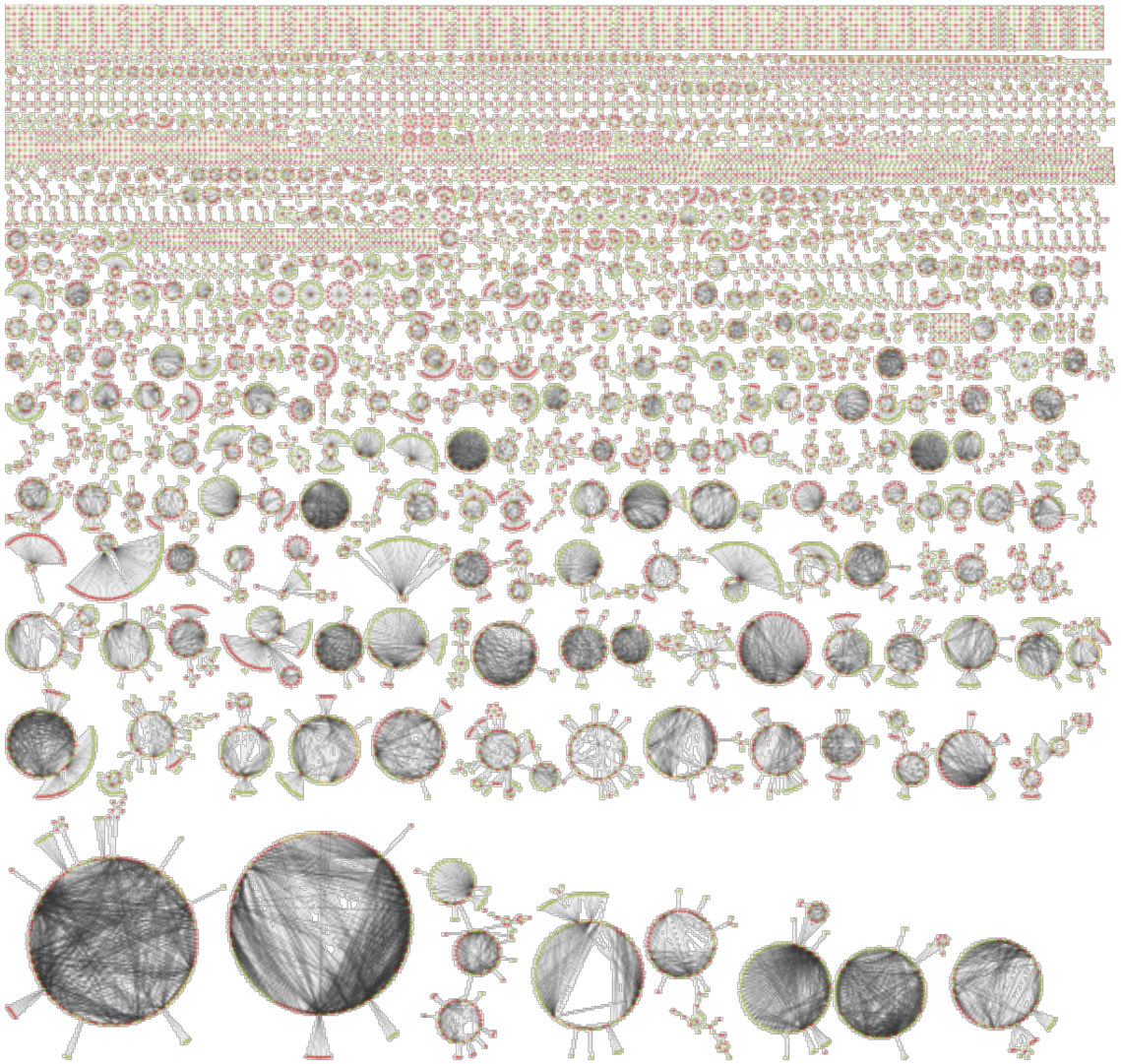
$$\textbf{LipE} = pK_i - \textbf{cLogP}$$

- Difference in **LipE** between weakly and highly potent cliff partners

- **LipE increase** detected for **96.7%** of all activity cliffs; average **Δ LipE = 2.42**



3. Activity Cliff Network Analysis



Isolated vs. Coordinated Cliffs

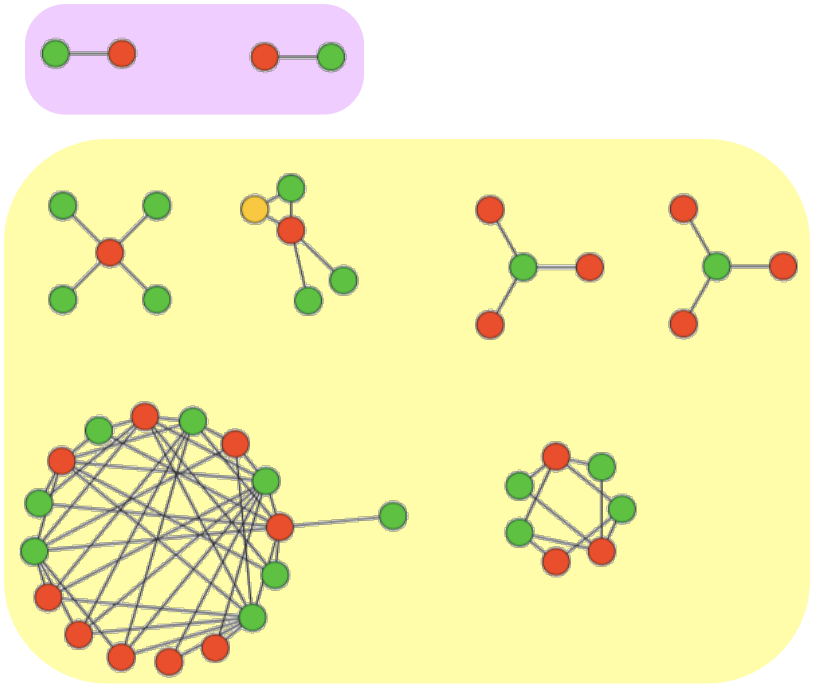
- ‘Isolated‘ cliffs: cliff partners are only involved in a single activity cliff
- ‘Coordinated‘ cliffs: cliff partners are involved in multiple and overlapping activity cliffs

Cliff type	Isolated cliffs %	Coordinated cliffs %
MACCS	1.4	98.6
ECFP4	2.2	97.8
MMP-cliffs	3.5	96.5

128 activity classes (>100 cpds) from ChEMBL

Isolated vs. Coordinated Cliffs

- MMP-cliff network for serotonin 1d receptor ligands



46 compounds (nodes)

69 MMP-cliffs (edges)

2 isolated cliffs

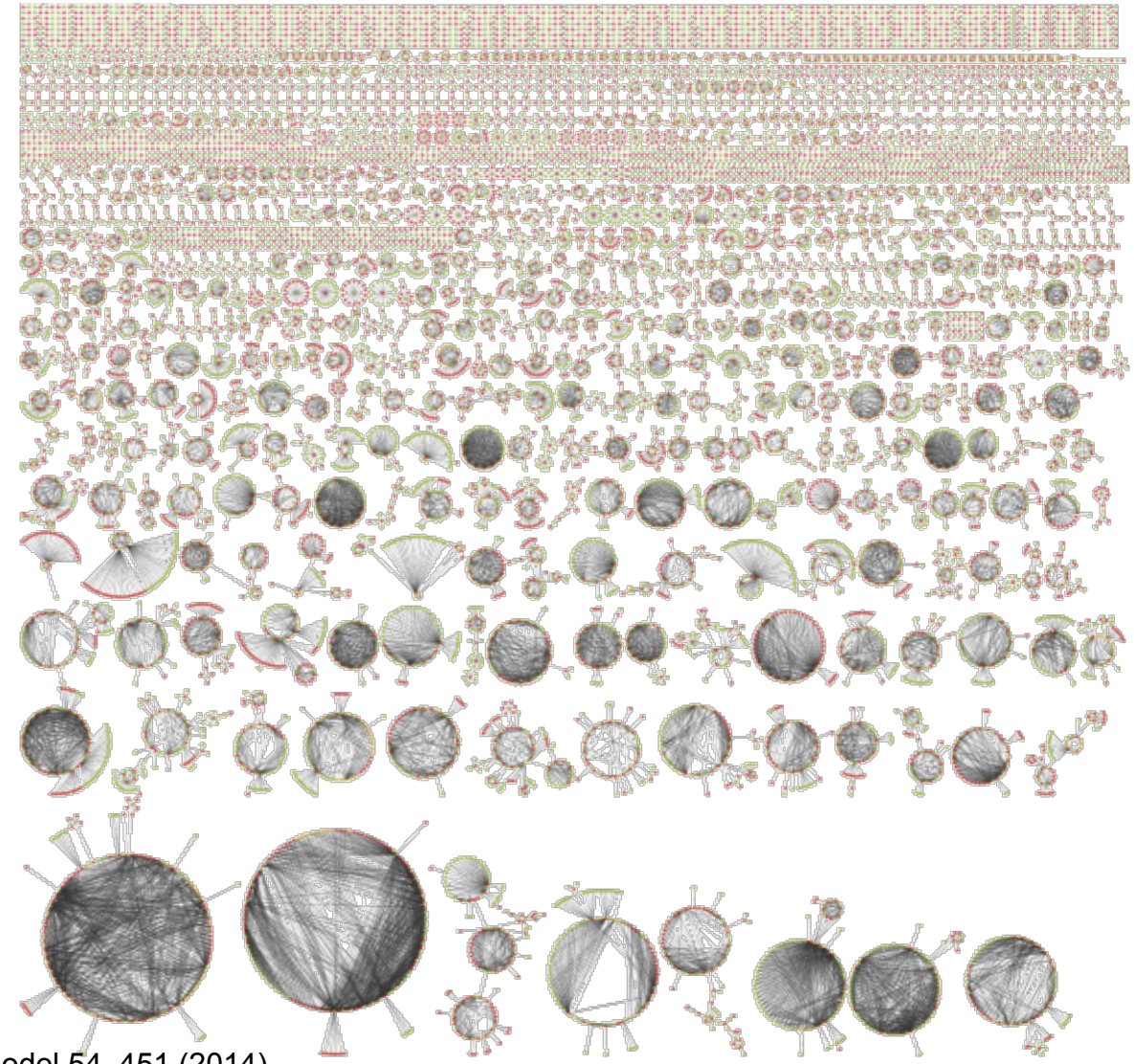
67 coordinated cliffs

- highly potent cliff partner
- weakly potent cliff partner
- both highly and weakly potent cliff partner

Global MMP-Cliff Network

- ChEMBL 17
- **14,044** nodes (compounds)
- **20,080** edges (MMP-cliffs)

- Many separate components
- **2072 clusters**



Stumpfe D et al. & Bajorath J. J Chem Inf Model 54, 451 (2014)

Activity Cliff Cluster Size Distibution

- 769 isolated cliffs
- 1303 coordinated cliff cluster
- 26 clusters with > 50 compounds
- 420 clusters comprising six to 15 compounds

Cluster size	# Cluster
1-5	1463
6-10	306
10-15	114
15-20	65
21-30	56
31-40	27
41-50	15
51-60	11
61-70	4
71-80	2
81-90	3
91-100	2
101-152	4

Node Degree Distribution

- Average node degree 2.9
- The union of all clusters follows a power law

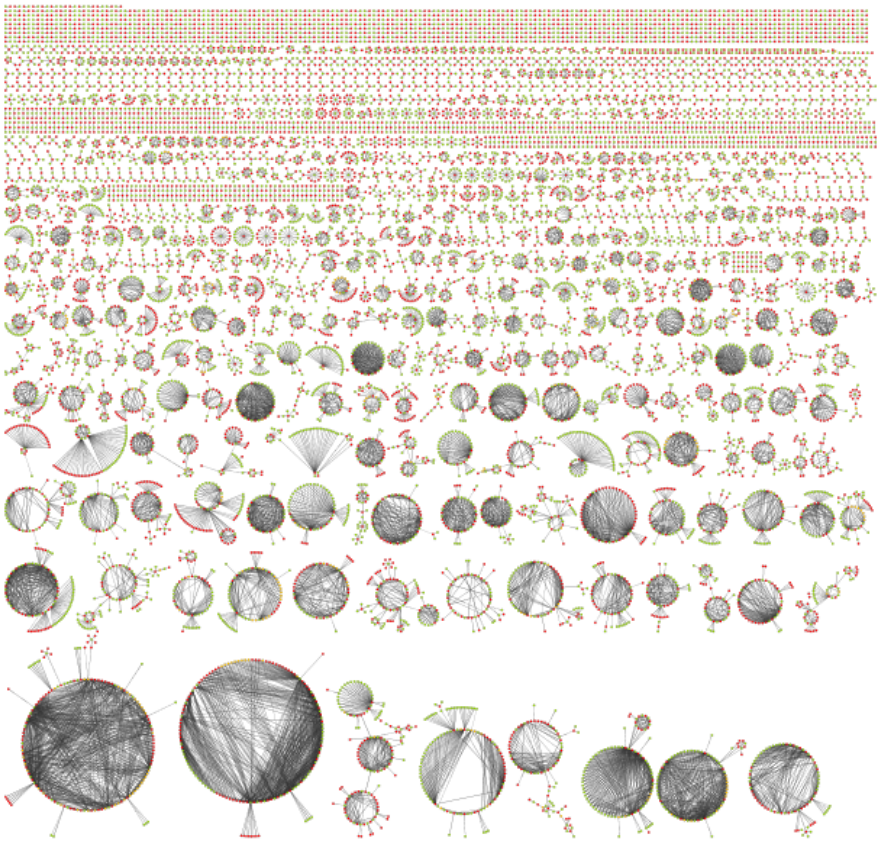
$$P(k) \sim k^{-\gamma}$$

with γ having a value of 2.5, which is characteristic of *scale-free* networks

- Many densely connected nodes: **activity cliff hubs**

Node degree	# Nodes
1-4	11878
5-9	1552
10-14	341
15-20	155
21-30	85
31-40	17
41-50	9
51-60	4
61-70	3

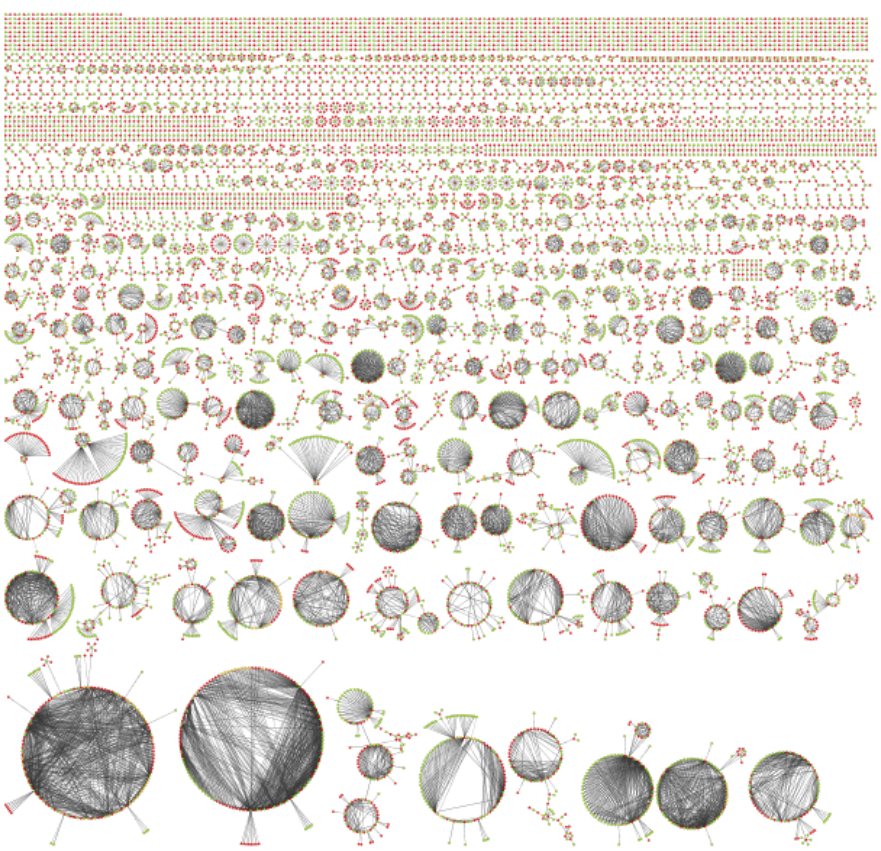
Network Modification

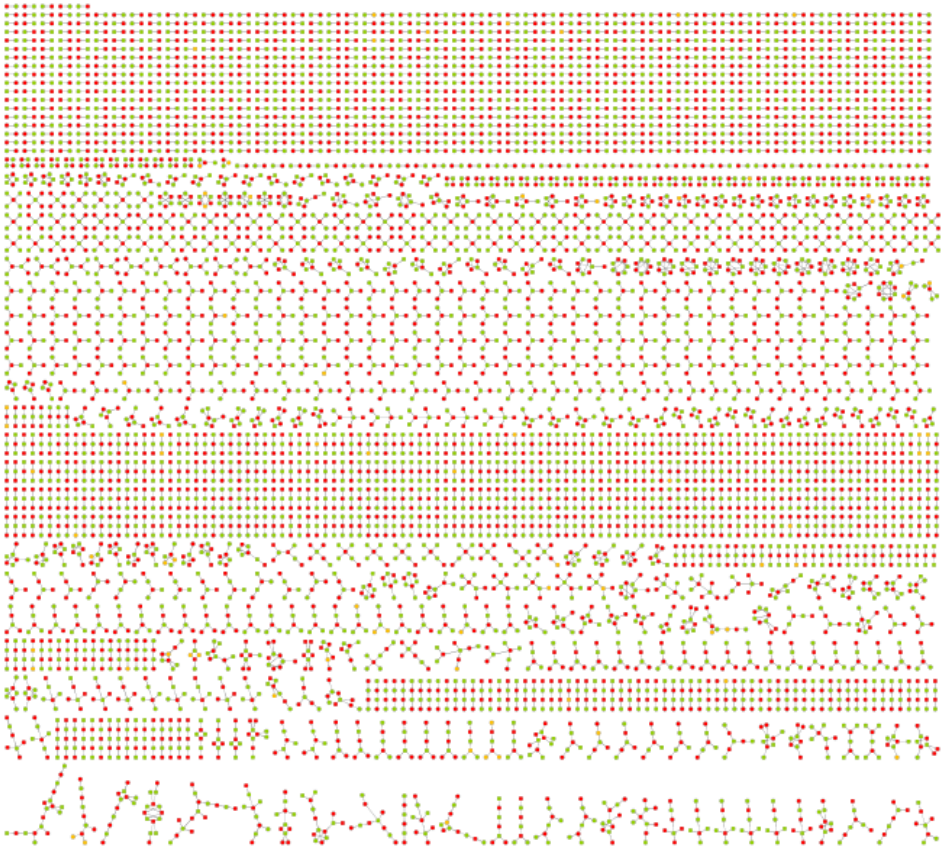


- Deletion of all hubs with a degree ≥ 5

  (2166 nodes, i.e. 15.4%)

Network Modification





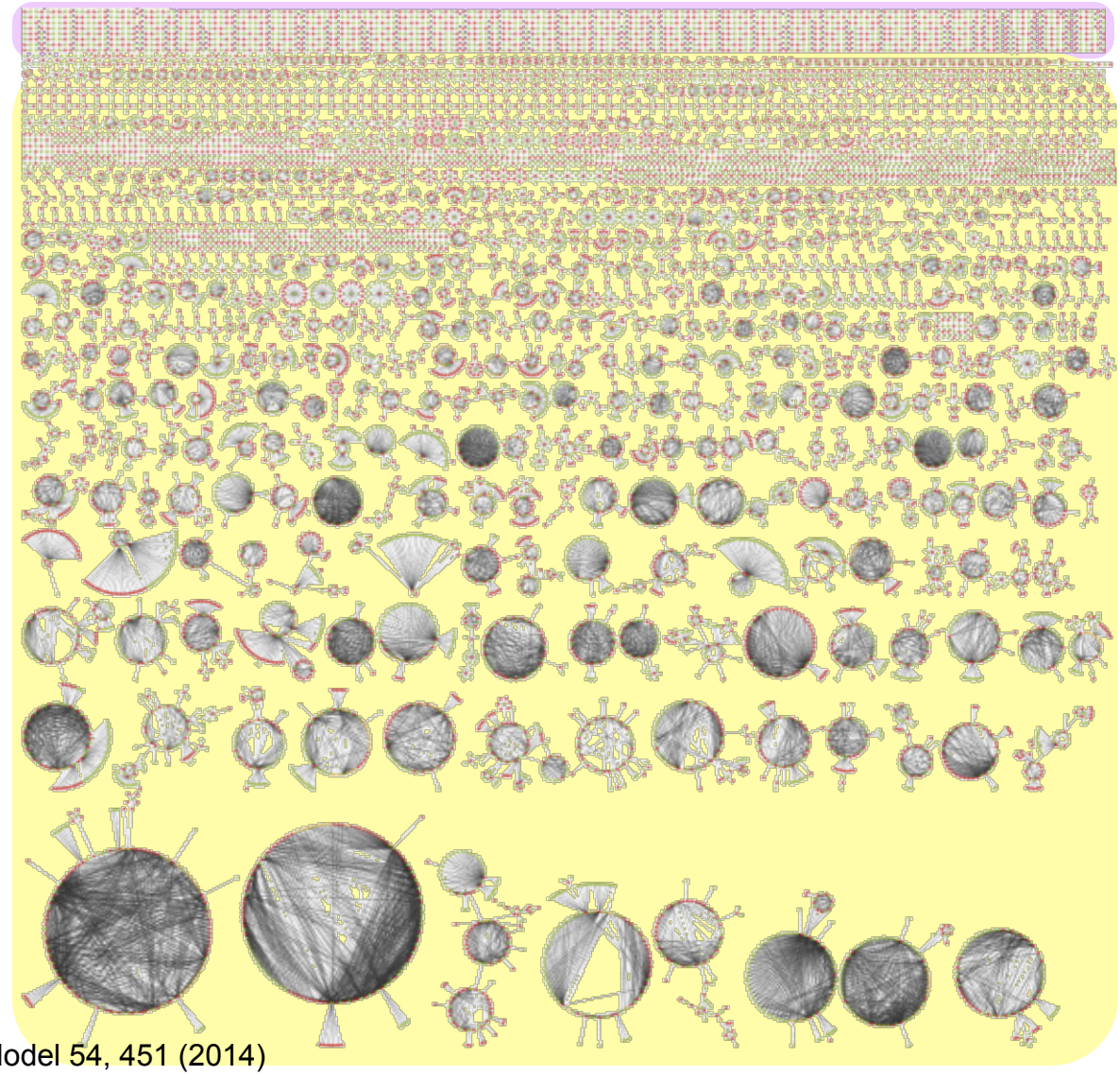
- Deletion of all hubs with a degree ≥ 10

  (614 nodes, i.e. 4.4%)

Global MMP-Cliff Network

- 2072 clusters
- 769 isolated cliffs
- **19,311** coordinated cliffs in **1303** clusters

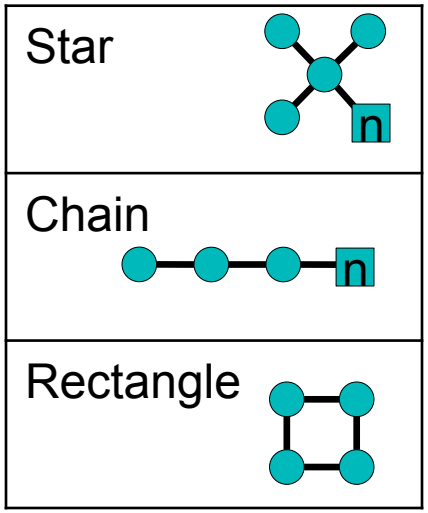
- **450 cluster topologies** with 1 to 769 instances

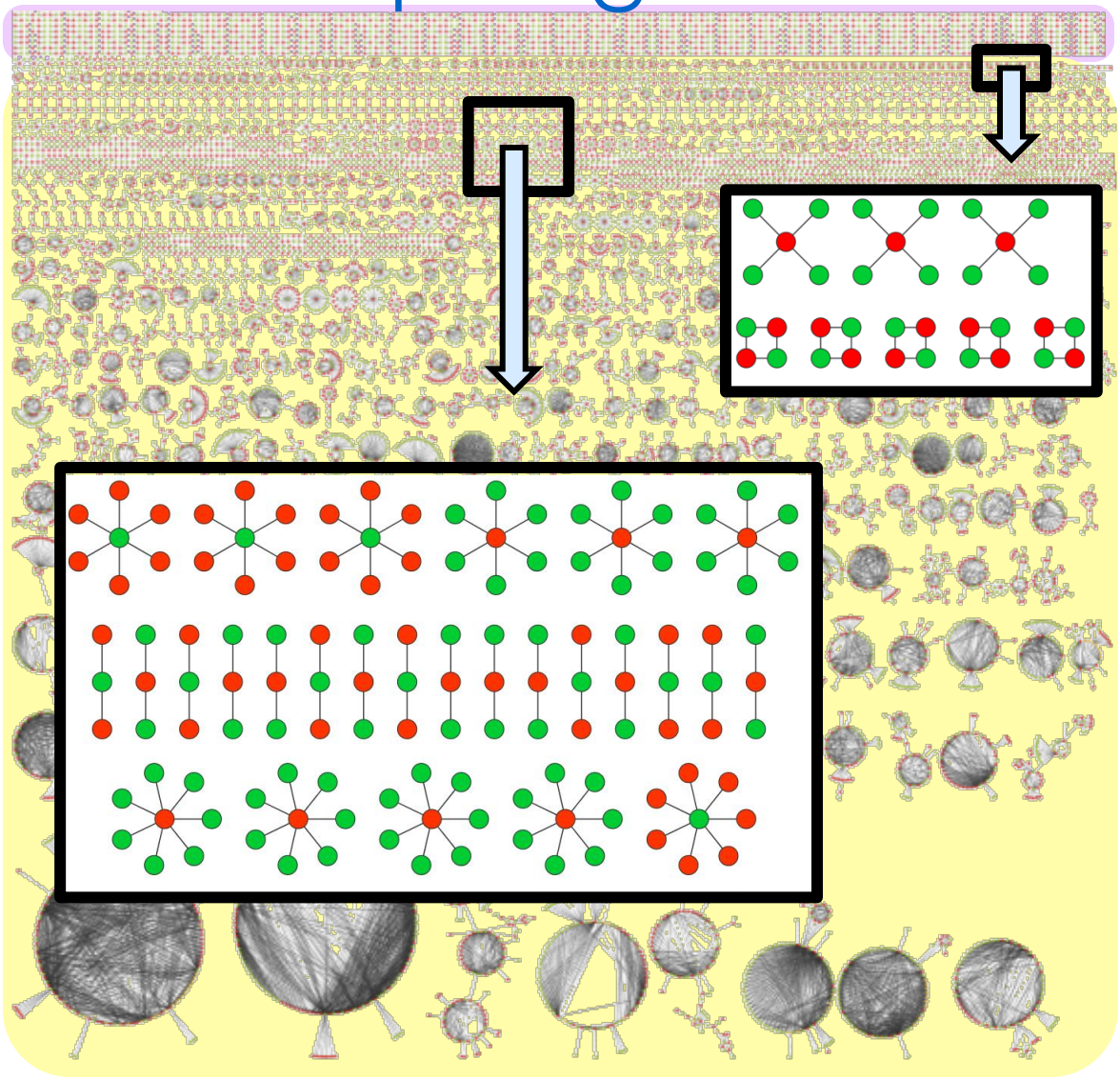


Stumpfe D et al. & Bajorath J. J Chem Inf Model 54, 451 (2014)

Activity Cliff Cluster Topologies

- Topologies with ≥ 3 instances
- Identification of **3 recurrent** main topologies

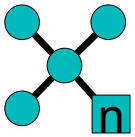
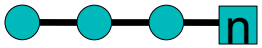
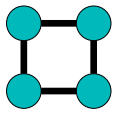


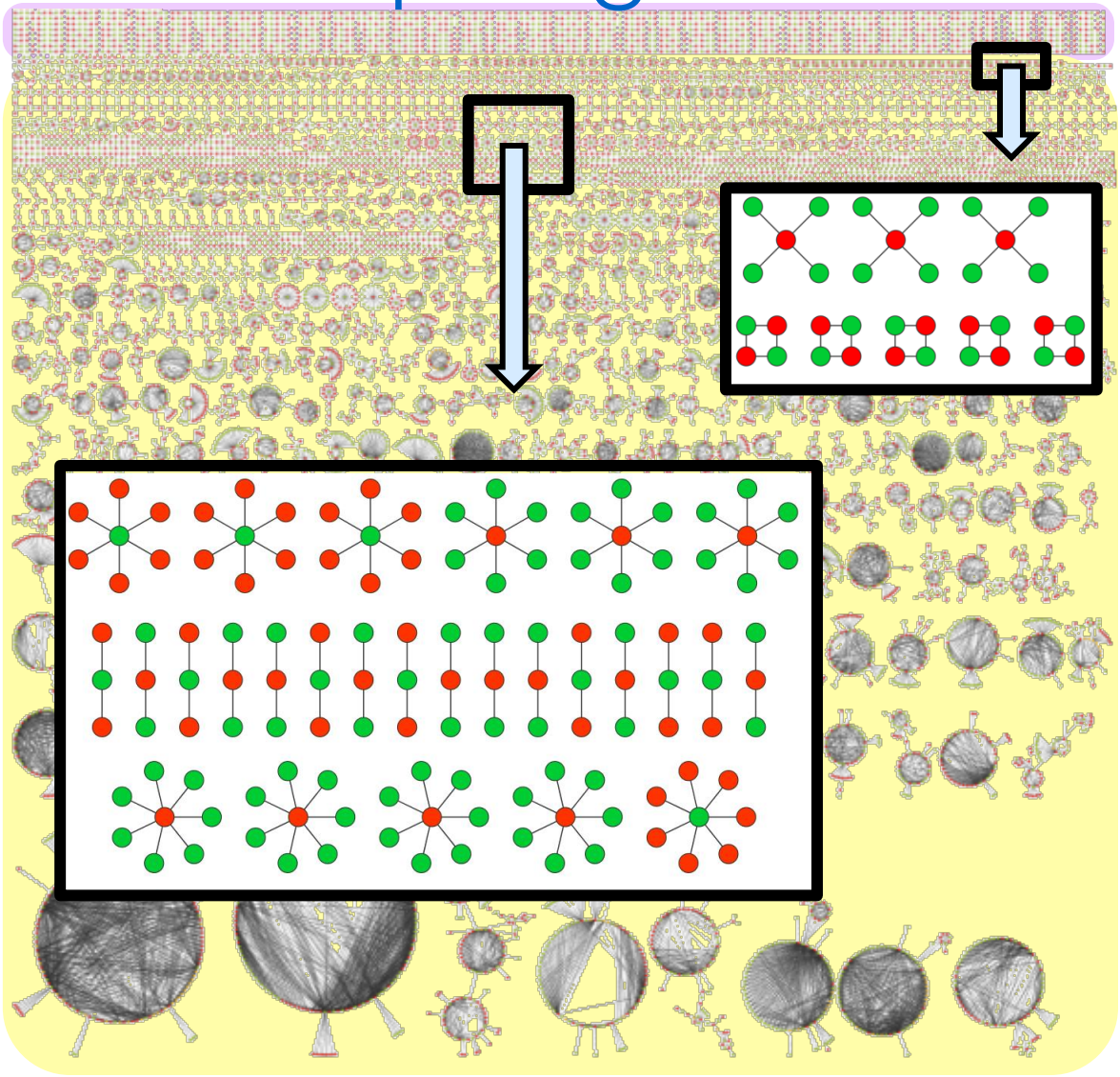


Activity Cliff Cluster Topologies

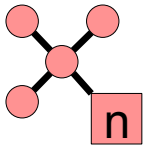
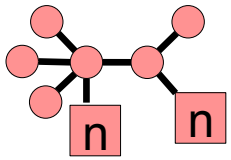
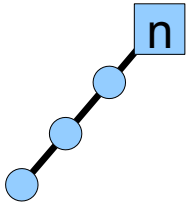
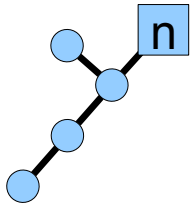
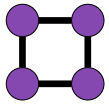
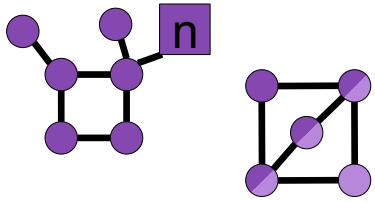
- Topologies with ≥ 3 instances

- Cover **861** of **1303** clusters main topologies

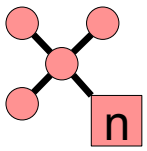
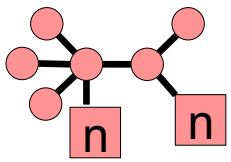
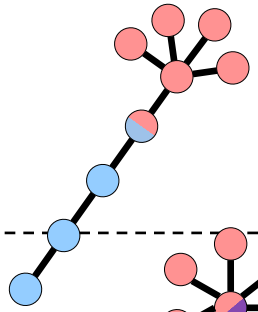
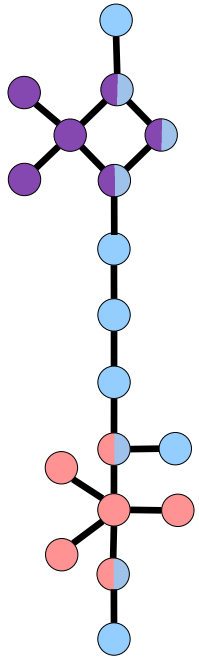
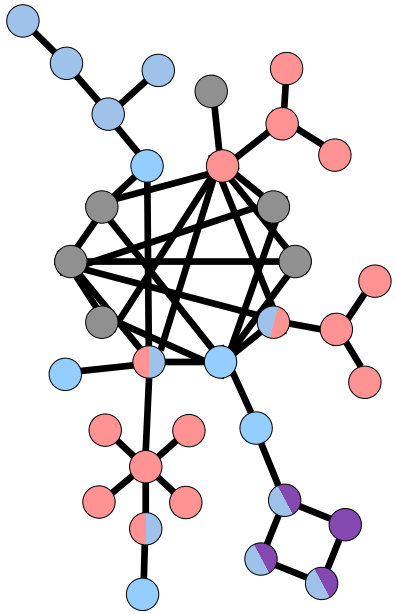
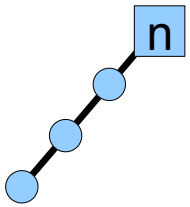
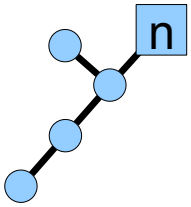
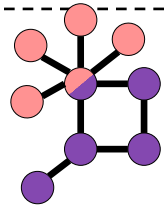
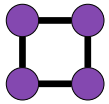
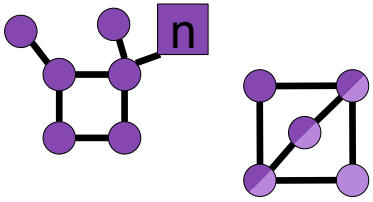
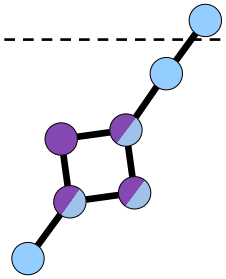
Star	
Chain	
Rectangle	



Main Topologies and Extensions

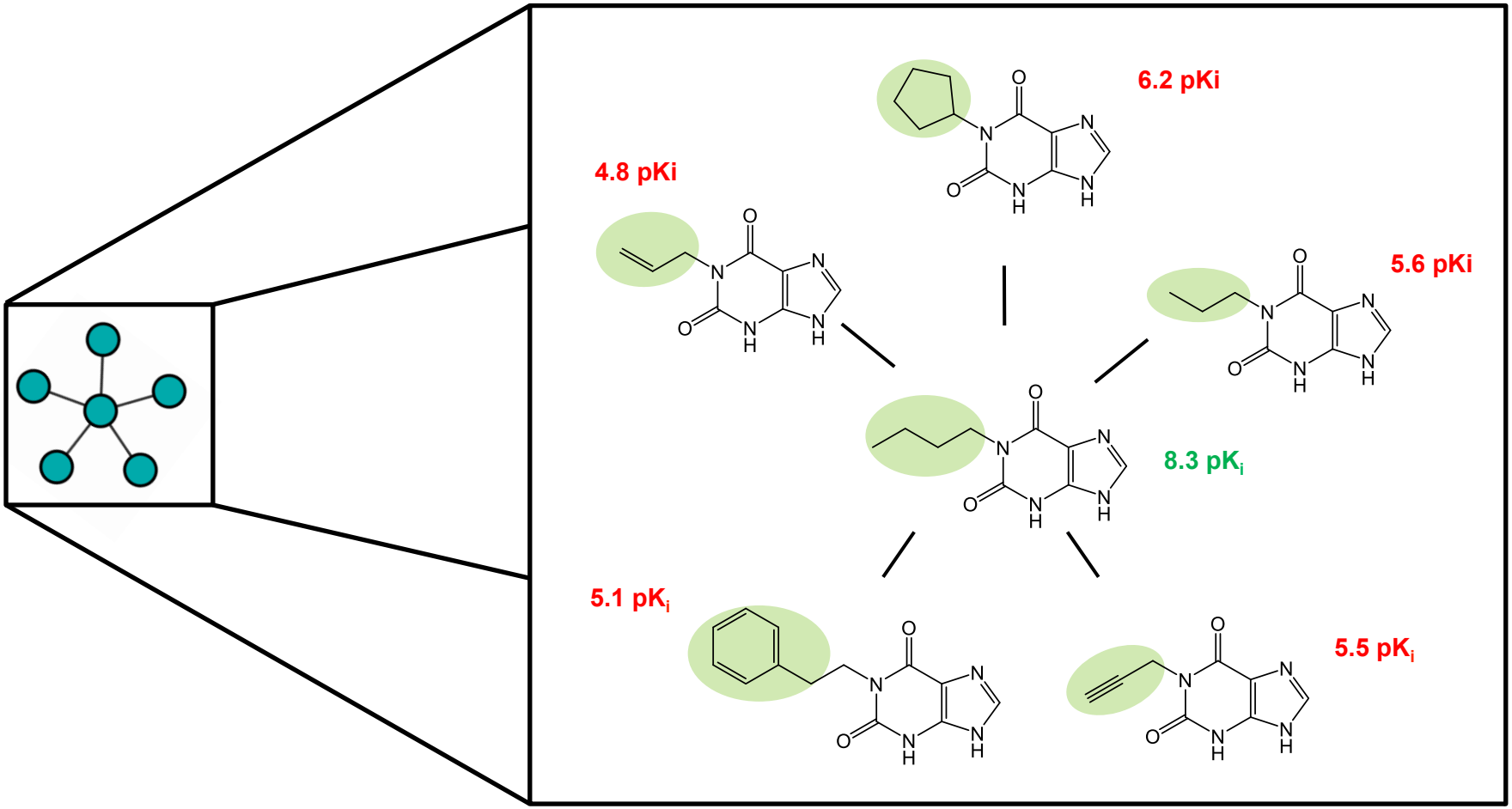
Main topology	Extensions of main topology
Star 	Twin Star 
Chain 	Modified Chain 
Rectangle 	Modified Rectangle 

Main Topologies and Extensions

Main topology	Extensions of main topology	Hybrid topologies	Irregular topologies
Star 			 
Chain 			
Rectangle 			

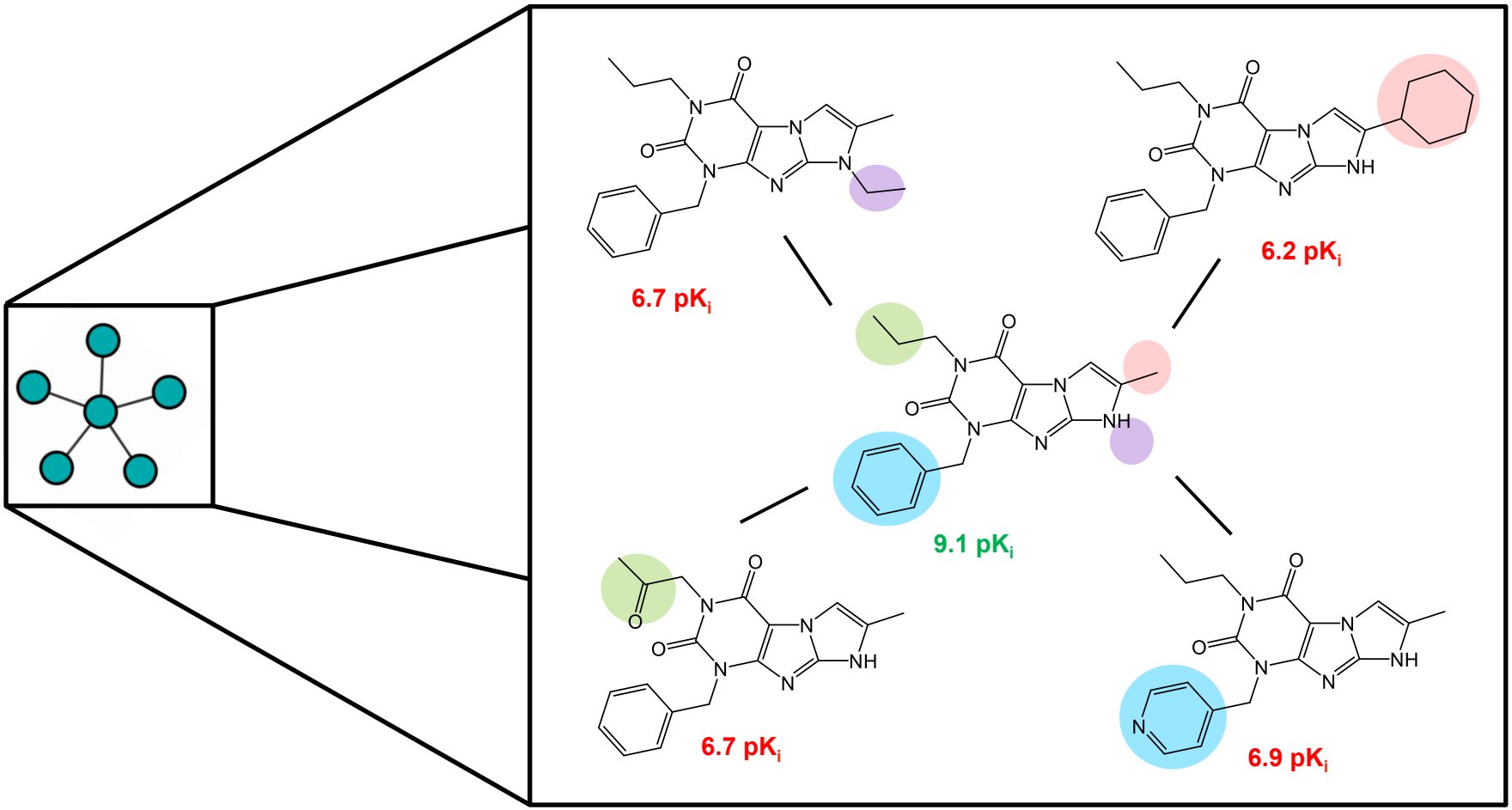
Star Topology Example

- Adenosine A3 receptor ligands



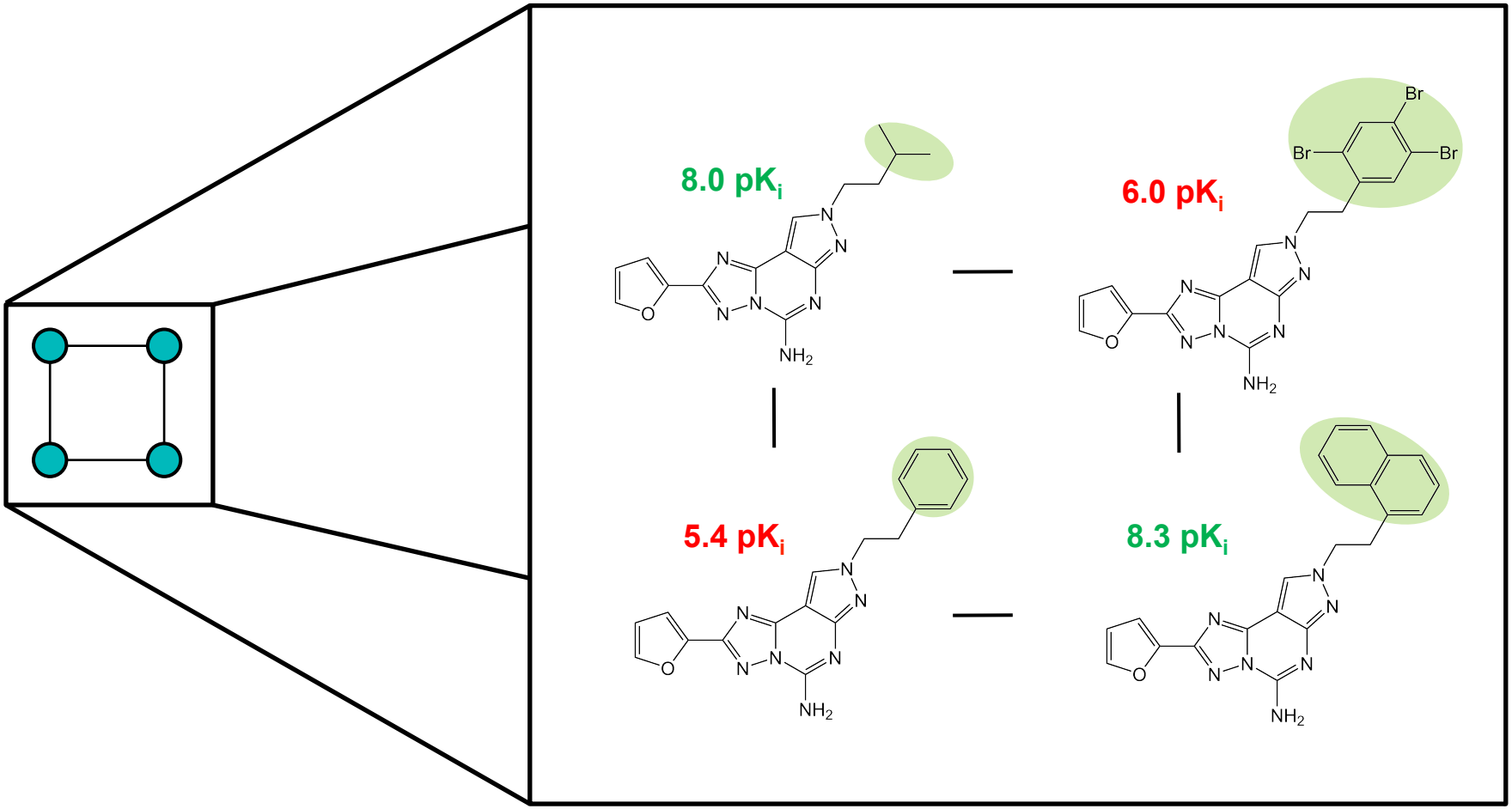
Star Topology Example

- Adenosine A3 receptor ligands



Rectangle Topology Example

- Adenosine A2b receptor ligands

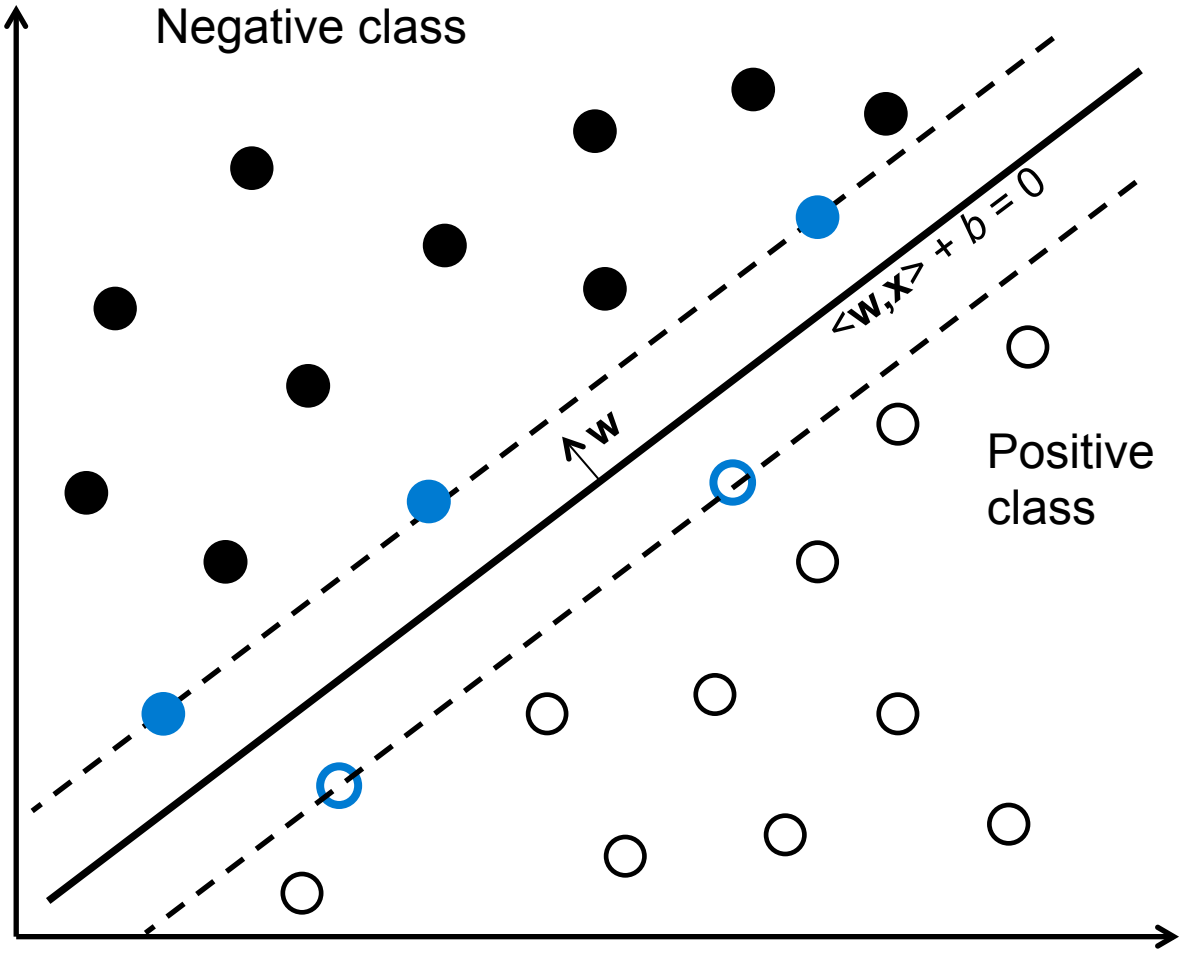


4. Can We Predict Activity Cliffs?

- Support vector machines for prediction of activity cliffs in compound data sets
- Non-trivial problem: compound pairs (with different potency) need to be predicted
- Design of compound pair-based kernel functions

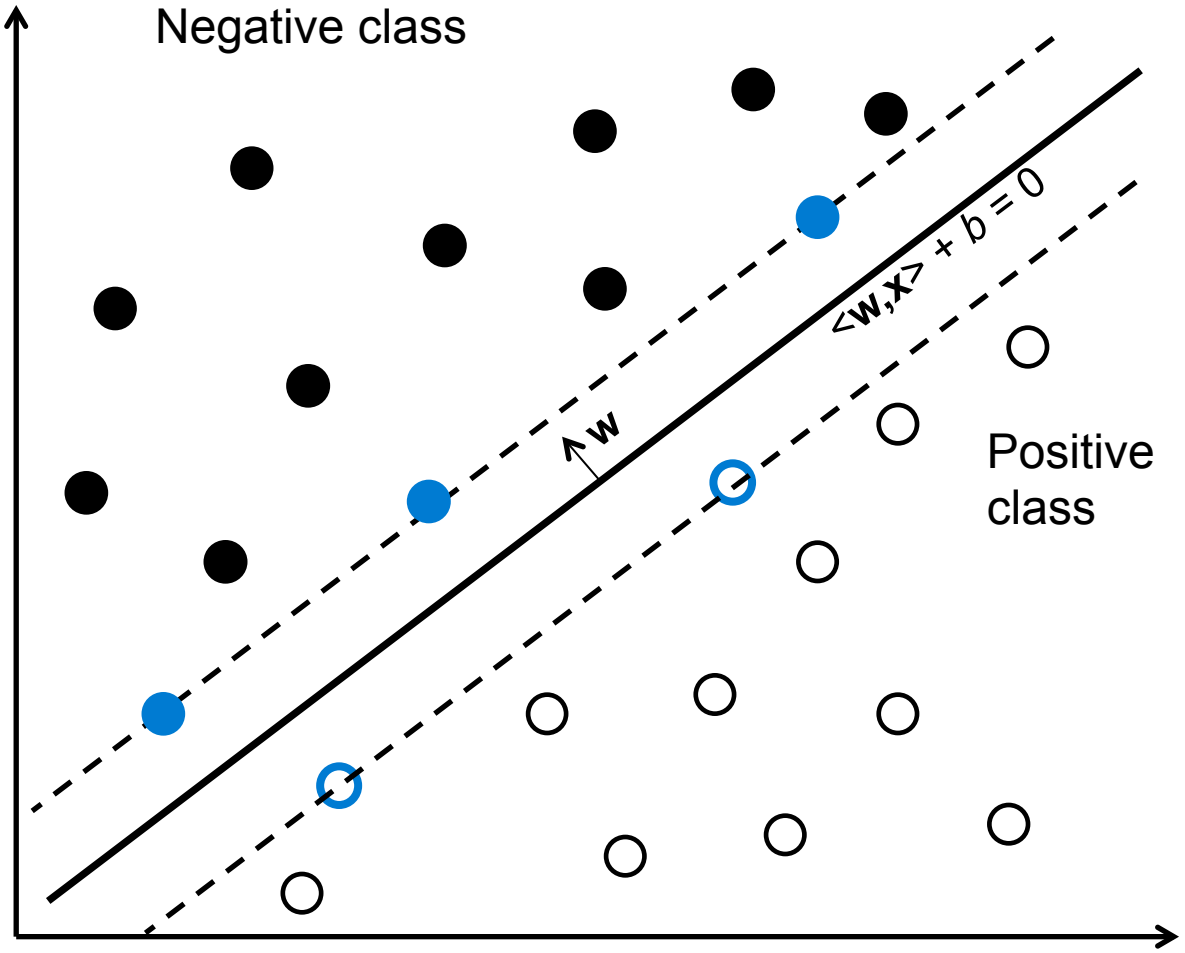
Support Vector Machines (SVMs)

- Derivation of a separating hyperplane in chemical space between positive and negative training compounds
- If no linear separation is possible data are projected into higher dimensional spaces through the use of kernel functions



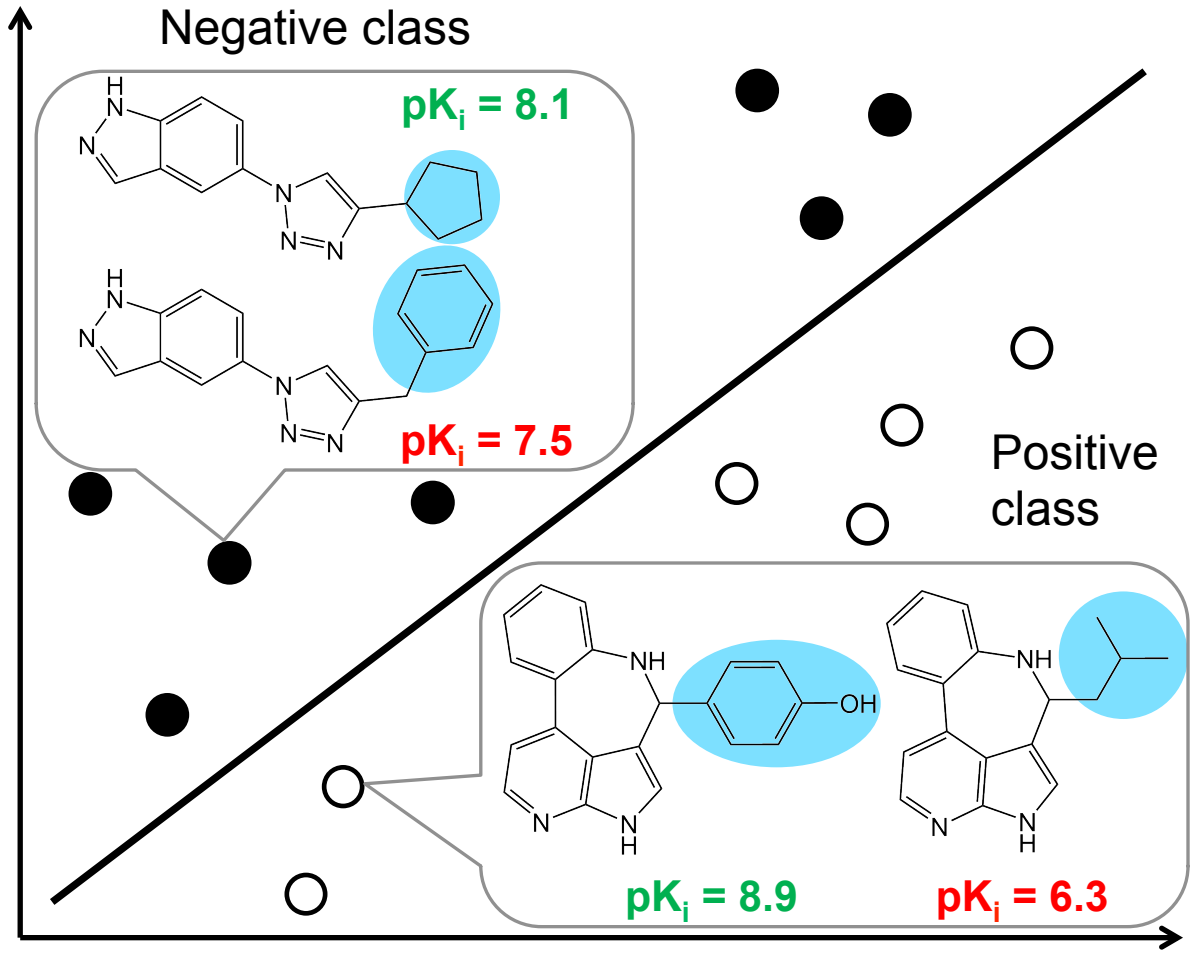
Support Vector Machines (SVMs)

- Binary classification of test compounds depending on which side of the hyperplane they fall
- Ranking of test compounds based on their (positive or negative) distance from the hyperplane



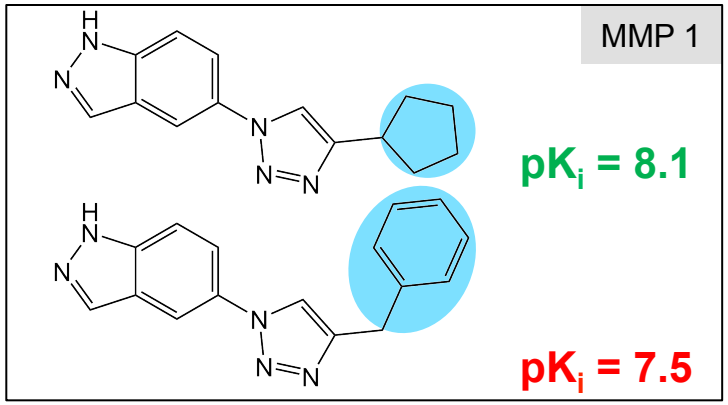
SVMs in Compound Pair Space

- Data points are compound pairs (MMPs)
- Negative class:
  - MMPs **not** forming activity cliffs
- Positive class:
  - MMPs forming activity cliffs
- Reference space: compound pair space

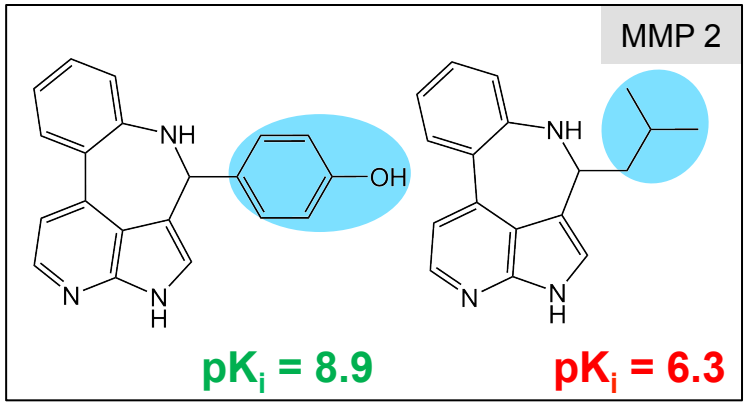


Heikamp K et al. & Bajorath J. J Chem Inf Model 52, 2354 (2012)

Design of a Transformation Kernel



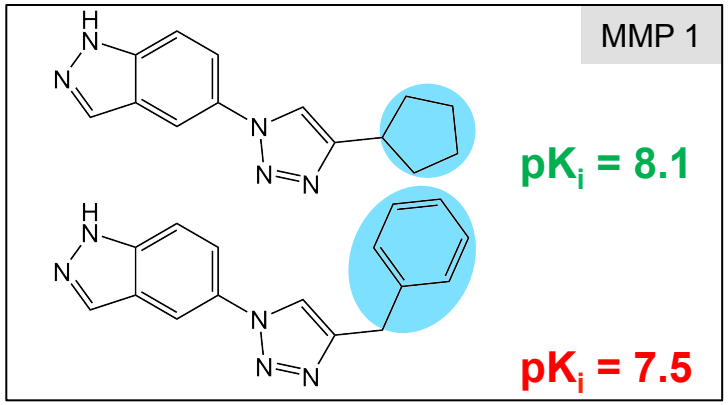
non-Cliff

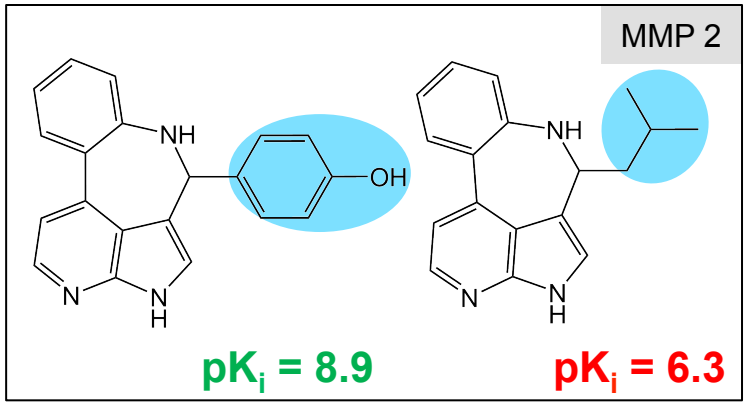


Activity Cliff

Design principle:
-encode activity cliff transformations and compare them with transformations from non-cliffs

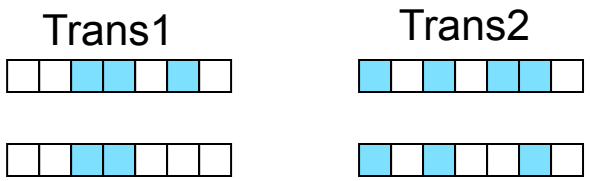
Transformation Kernel





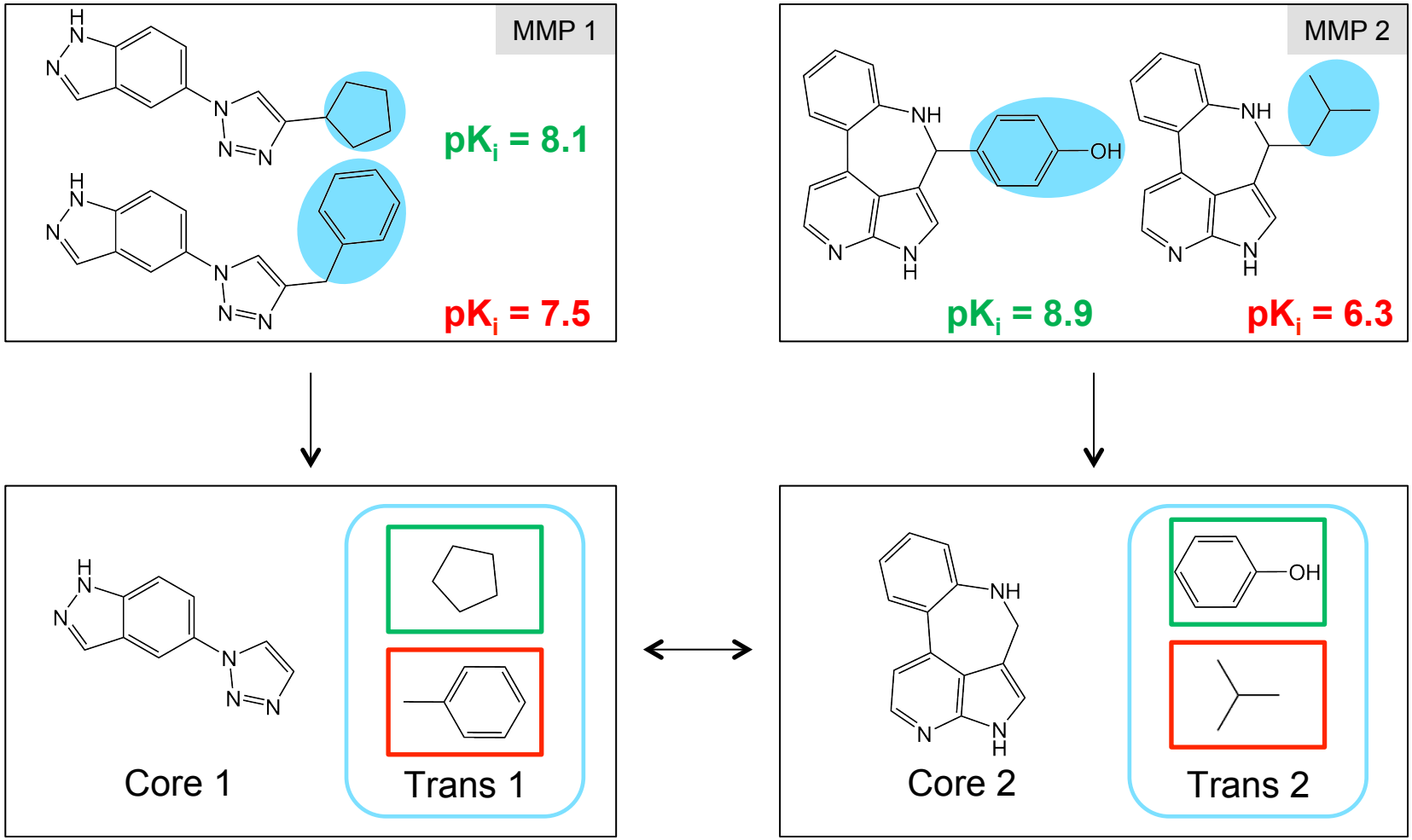
Transformation space

Step 1:
Fingerprint representation of transformation substructures

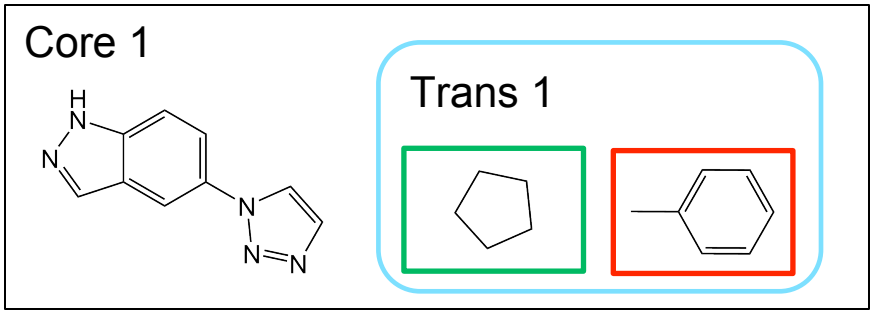


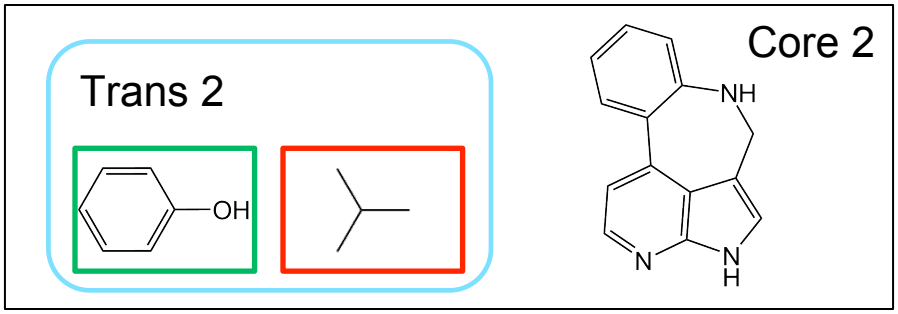
(fingerprints: structural keys or atom pairs)

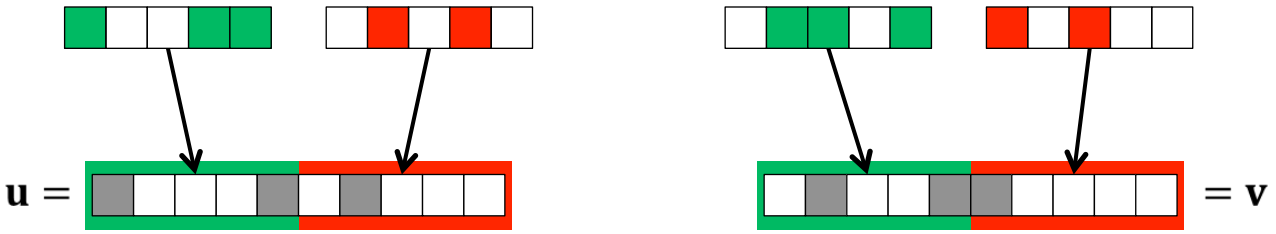
Kernel for Compound Pairs



Transformation Kernel





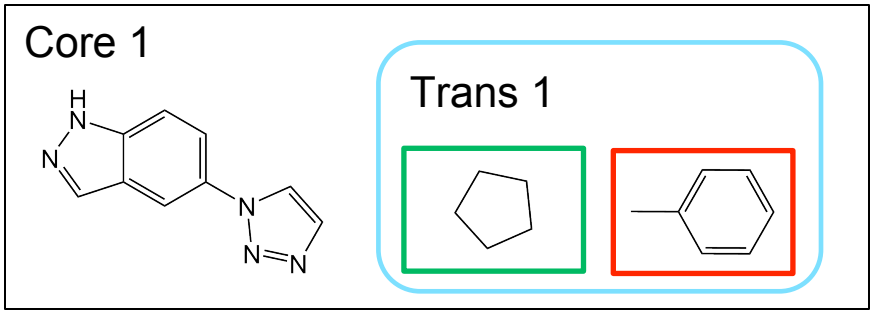


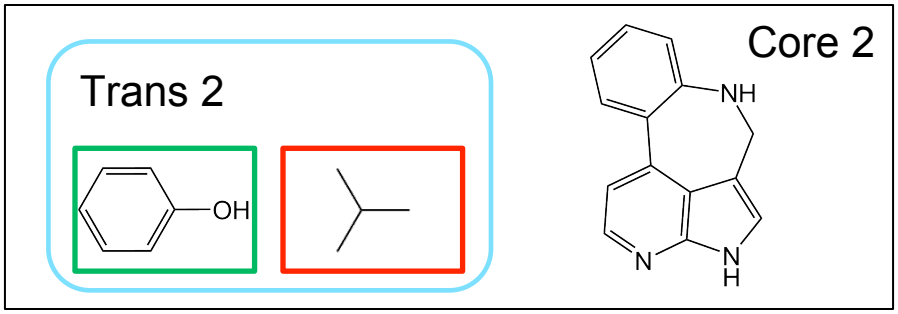
Step 2:
Substructure difference vector (size 2n)
from transformation mini-fingerprints (1n)

(each pair of transformations
yields feature vectors u and v)

Substructure difference vector contains all features that distinguish the transformation substructures

Transformation Kernel





$\mathbf{u} =$   $= \mathbf{v}$

Transformation space

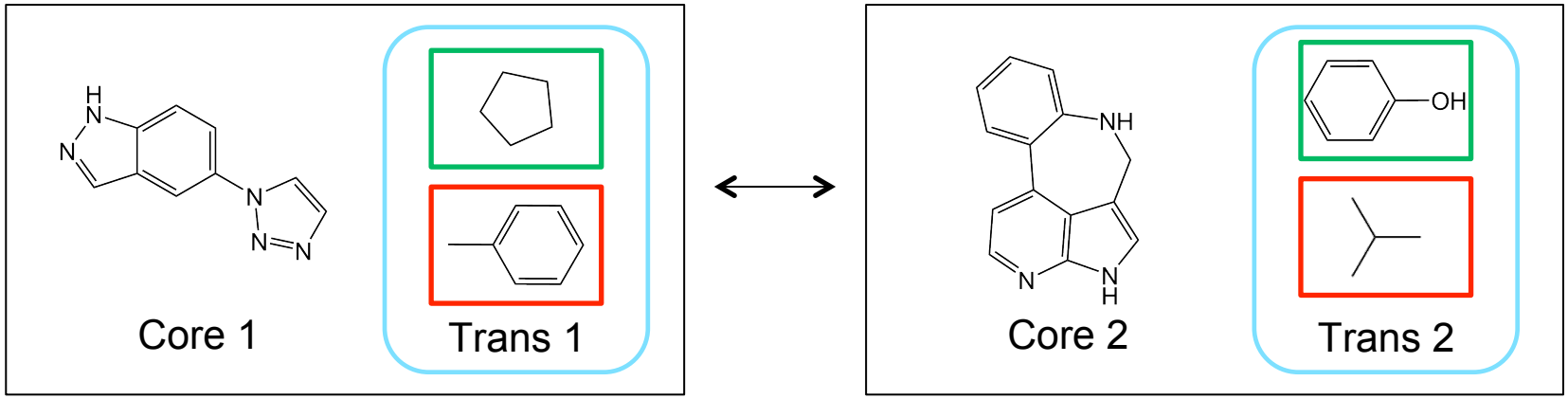
Step 3:
Kernel from substructure difference feature vectors

$K_{\text{transformation}}(u, v) = K(u, v)$

(calculate Tanimoto kernel from feature vectors)

$$K_{Tanimoto}(\mathbf{u}, \mathbf{v}) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle}$$

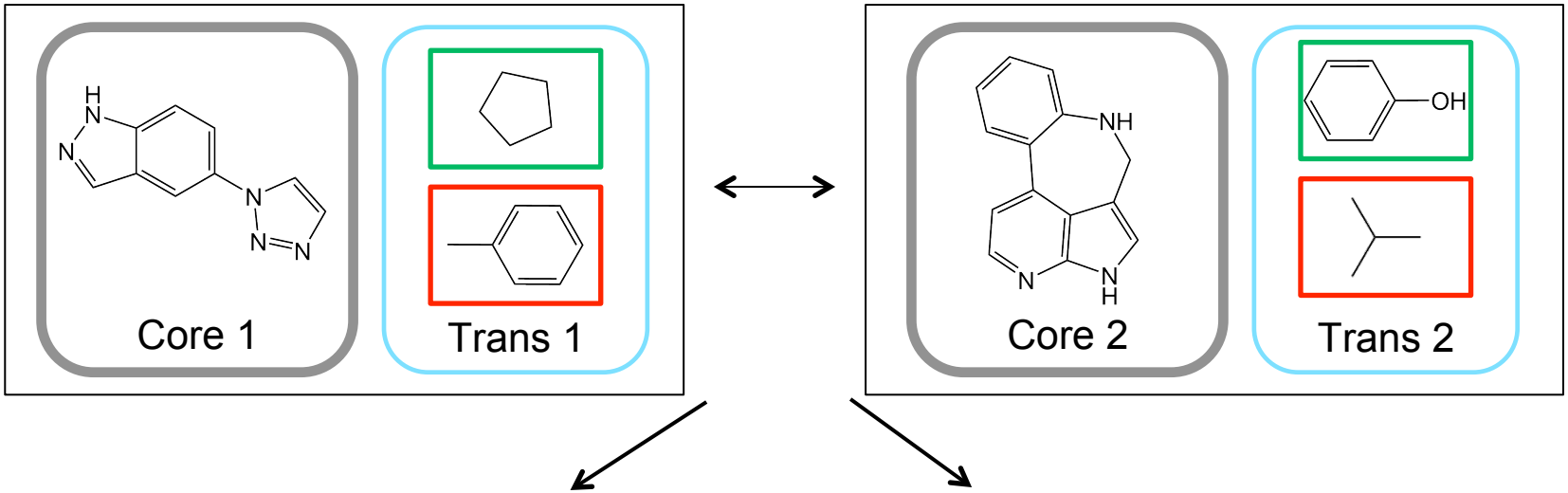
Design of an MMP Kernel



Design principle:

- combine core structure and transformation information
- add core structure representation to substructure difference vectors

MMP Kernel



Core space

Core1 $\mathbf{c}_1$

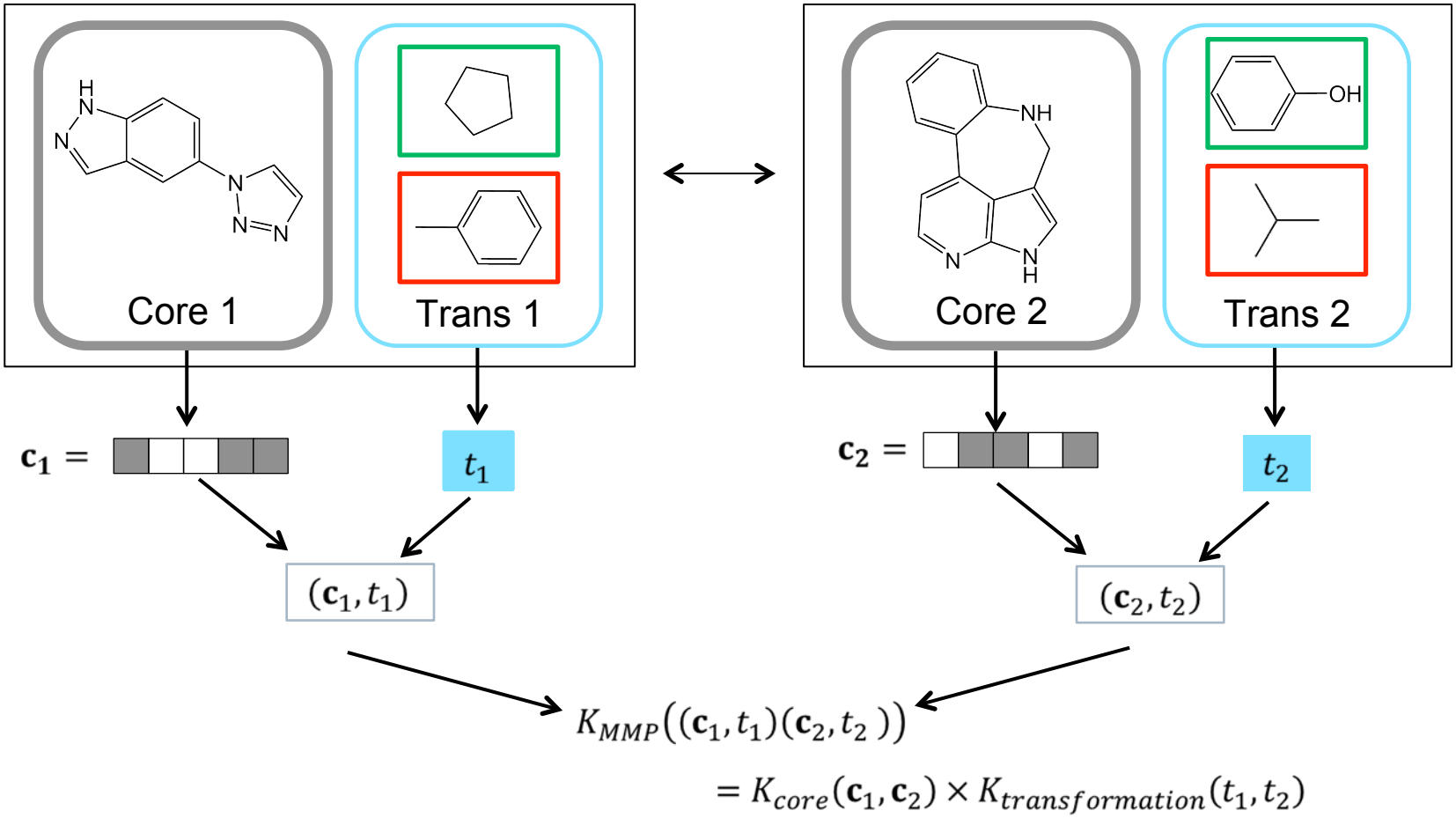
Core2 $\mathbf{c}_2$

Transformation space

$K_{\text{transformation}}(u, v) = K(u, v)$

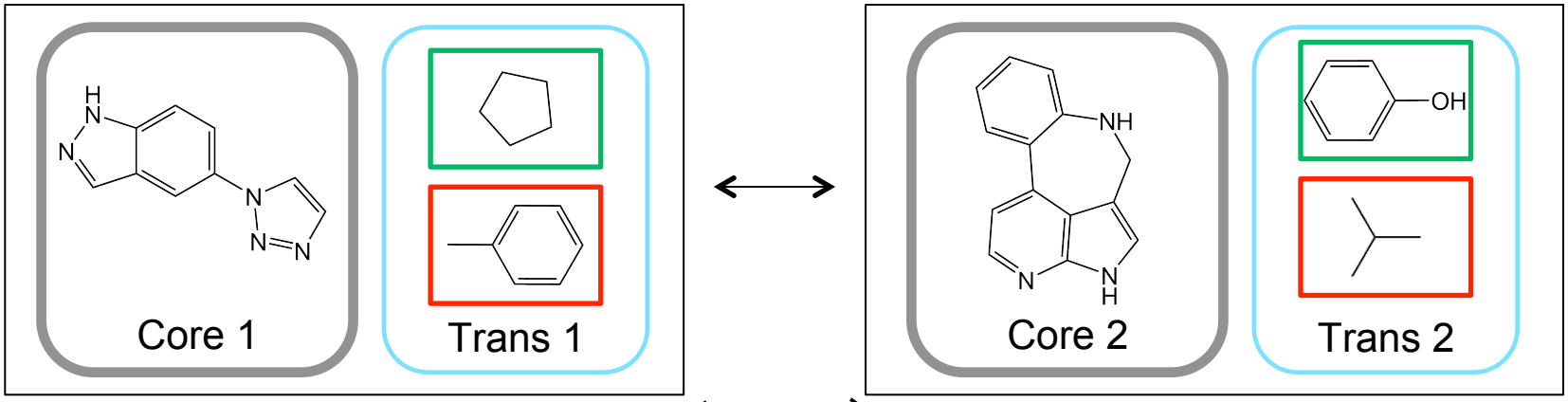
(fingerprint representation of core:
structural fragments or atom environments)

MMP Kernel



Combining core and transformation feature vectors yields a kernel product

MMP Kernel



Core space

Transformation space

$$K_{MMP} = K_{core}(c_1, c_2) \times K_{transformation}(t_1, t_2)$$

$$K_{MMP} = K_{Tanimoto}(c_1, c_2) \times K_{Tanimoto}(t_1, t_2)$$

Accurate Prediction of Activity Cliffs

Target	Transformation kernel				MMP kernel			
	TPR	TNR	P	F-score	TPR	TNR	P	F-score
*fxa	72.69	91.79	50.45	59.51	82.17	96.11	70.92	76.03
*mcr4	78.15	96.72	53.02	63.01	83.05	99.06	80.82	81.82
*kor	66.87	88.92	35.44	46.26	72.58	96.87	67.88	70.04
*thr	81.15	90.23	58.99	68.29	84.05	95.43	76.20	79.85
*aa3	71.95	88.46	38.63	50.19	74.45	97.20	73.23	73.57
cal2	97.44	95.40	92.14	94.65	97.69	97.63	95.79	96.70
catb	88.33	97.56	91.10	89.39	90.83	98.67	95.43	92.76
dpp8	99.29	100.00	100.00	99.63	99.29	100.00	100.00	99.63
jak2	92.73	88.42	82.82	87.30	91.82	90.53	85.55	88.28

* Unbalanced composition: ratio of non-cliffs to cliffs between 6 and 21

Parameters:

- MACCS for transformation substructure representation
- Molprint2D for core structure representation
- Tanimoto/transformation kernel, Tanimoto/MMP kernel

TPR	True positive rate
TNR	True negative rate
P	Precision
F-score	$2 \cdot TPR \cdot P/(TPR + P)$

Accurate Prediction of Activity Cliffs

Target	Transformation kernel				MMP kernel			
	TPR	TNR	P	F-score	TPR	TNR	P	F-score
fxa	72.69	91.79	50.45	59.51	82.17	96.11	70.92	76.03
mcr4	78.15	96.72	53.02	63.01	83.05	99.06	80.82	81.82
kor	66.87	88.92	35.44	46.26	72.58	96.87	67.88	70.04
thr	81.15	90.23	58.99	68.29	84.05	95.43	76.20	79.85
aa3	71.95	88.46	38.63	50.19	74.45	97.20	73.23	73.57
cal2	97.44	95.40	92.14	94.65	97.69	97.63	95.79	96.70
catb	88.33	97.56	91.10	89.39	90.83	98.67	95.43	92.76
dpp8	99.29	100.00	100.00	99.63	99.29	100.00	100.00	99.63
jak2	92.73	88.42	82.82	87.30	91.82	90.53	85.55	88.28

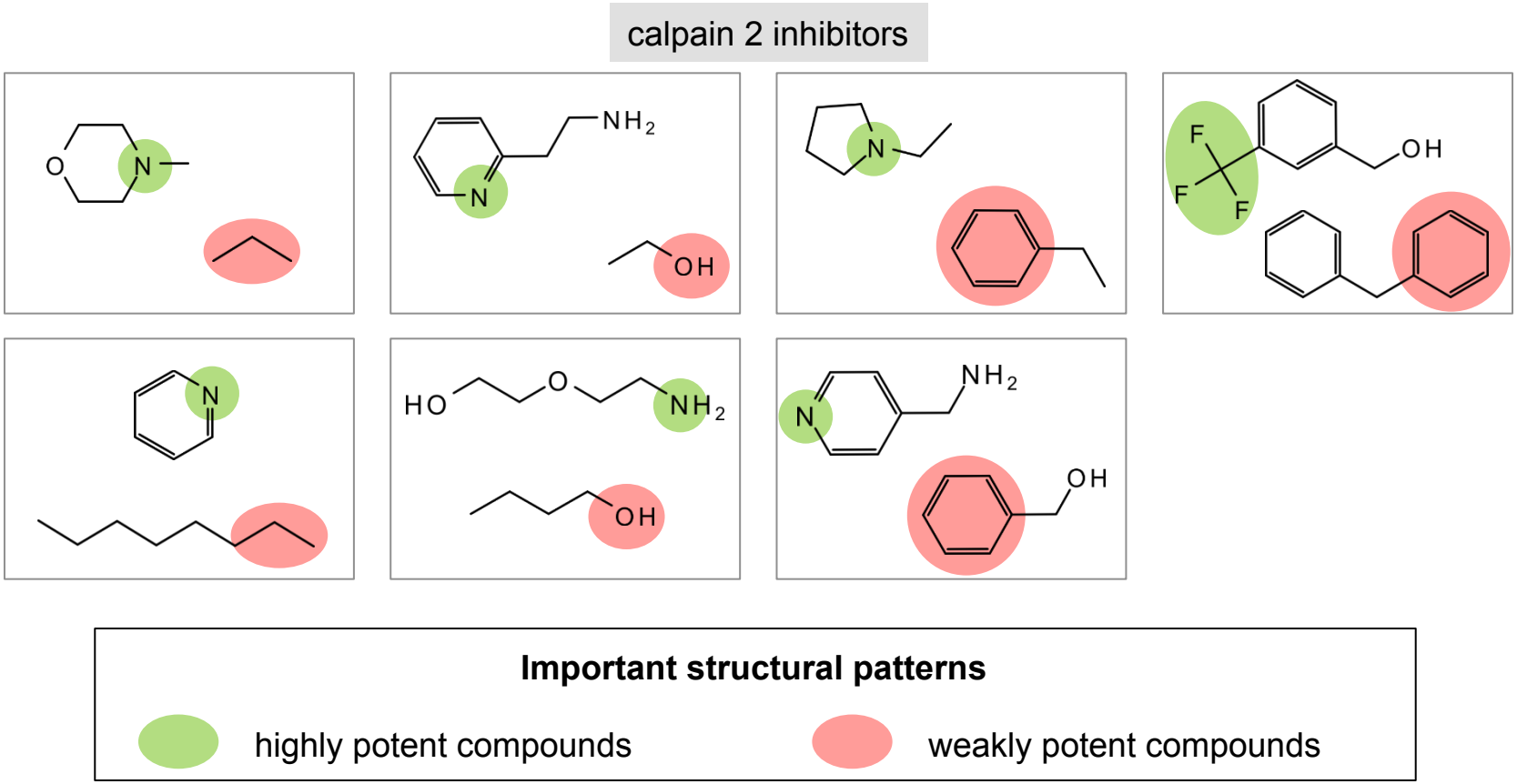
Results:

- Both methods can accurately predict activity cliffs in different data sets
- Prediction accuracy is further improved when core structure information is added to transformation information (MMP kernel)

Heikamp K et al. & Bajorath J. J Chem Inf Model 52, 2354 (2012)

Activity Cliff Transformations

- Identification of characteristic cliff transformations leading to highly potent compounds



Activity Cliff Summary

- Similarity / potency difference criteria are critical
- Cliffs can be represented in different ways
- Preference for MMP-cliffs
- Bioactive compounds frequently form activity cliffs
- Similar distribution over different targets
- Most cliffs are formed in a coordinated manner
- Global activity cliff network: scale-free
- Activity cliff clusters with recurrent topology
- Prediction of activity cliffs via SVM / MMP kernels