Exploring Activity Cliffs from a Chemoinformatics Perspective

Jürgen Bajorath
Life Science Informatics
University of Bonn
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“ ➔

capacity 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 (QPs)**
  - QPs are compound pairs exceeding a given similarity 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

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


128 activity classes (>100 cpds) from ChEMBL
MMPs as Molecular Representation

- A **Matched Molecular Pair** (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:

1. \(*H\) \(\rightarrow\) \(\text{Pyridine} \)*
2. \(\text{Cyclohexane} \) \(\rightarrow\) \(\text{Heterocyclic compound} \)
3. \(\text{Cyclohexane} \) \(\rightarrow\) \(\text{Heterocyclic compound with functional groups} \)
Preferred Activity Cliff Definition

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

Activity Cliff-Forming Compounds

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

MMPs:
- 27.5%

ECFP4:
- 35.3%

FCFP4:
- 34.2%

GpiDAPH3:
- 36.1%

MACCS:
- 41%

TGT:
- 37.2%

TGD:
- 41.4%

Consensus (FP only):
- 14.7%

Consensus:
- 10.9%

Union (FP only):
- 64.5%

Union:
- 65.6%

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


414 activity classes from ChEMBL
For data set with >200 cpds, activity cliffs and cliff compounds are fairly evenly distributed among many different targets
Ligand Efficiency (LE) for MMP-Cliffs

- Changes in **LE** accompanying activity cliff formation
- Difference in **LE** between weakly and highly potent cliff partners
- **LE increase** detected for 99.1% of all activity cliffs; average $\Delta \text{LE} = 6.27$

$\text{LE} = \frac{pKi}{MW}$

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

- Changes in LipE accompanying activity cliff formation
- 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

<table>
<thead>
<tr>
<th>Cliff type</th>
<th>Isolated cliffs %</th>
<th>Coordinated cliffs %</th>
</tr>
</thead>
<tbody>
<tr>
<td>MACCS</td>
<td>1.4</td>
<td>98.6</td>
</tr>
<tr>
<td>ECFP4</td>
<td>2.2</td>
<td>97.8</td>
</tr>
<tr>
<td>MMP-cliffs</td>
<td>3.5</td>
<td>96.5</td>
</tr>
</tbody>
</table>

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

Activity Cliff Cluster Size Distribution

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

<table>
<thead>
<tr>
<th>Cluster size</th>
<th># Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-5</td>
<td>1463</td>
</tr>
<tr>
<td>6-10</td>
<td>306</td>
</tr>
<tr>
<td>10-15</td>
<td>114</td>
</tr>
<tr>
<td>15-20</td>
<td>65</td>
</tr>
<tr>
<td>21-30</td>
<td>56</td>
</tr>
<tr>
<td>31-40</td>
<td>27</td>
</tr>
<tr>
<td>41-50</td>
<td>15</td>
</tr>
<tr>
<td>51-60</td>
<td>11</td>
</tr>
<tr>
<td>61-70</td>
<td>4</td>
</tr>
<tr>
<td>71-80</td>
<td>2</td>
</tr>
<tr>
<td>81-90</td>
<td>3</td>
</tr>
<tr>
<td>91-100</td>
<td>2</td>
</tr>
<tr>
<td>101-152</td>
<td>4</td>
</tr>
</tbody>
</table>
Node Degree Distribution

- Average node degree **2.9**

- The union of all clusters follows a power law
  \[ P(k) \sim k^{-\gamma} \]
  with \( \gamma \) having a value of **2.5**, which is characteristic of **scale-free** networks

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

<table>
<thead>
<tr>
<th>Node degree</th>
<th># Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-4</td>
<td>11878</td>
</tr>
<tr>
<td>5-9</td>
<td>1552</td>
</tr>
<tr>
<td>10-14</td>
<td>341</td>
</tr>
<tr>
<td>15-20</td>
<td>155</td>
</tr>
<tr>
<td>21-30</td>
<td>85</td>
</tr>
<tr>
<td>31-40</td>
<td>17</td>
</tr>
<tr>
<td>41-50</td>
<td>9</td>
</tr>
<tr>
<td>51-60</td>
<td>4</td>
</tr>
<tr>
<td>61-70</td>
<td>3</td>
</tr>
</tbody>
</table>
Network Modification

- Deletion of all hubs with a degree $\geq 5$
  (2166 nodes, i.e. 15.4%)
Network Modification

- Deletion of all hubs with a degree $\geq 10$
  (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

Activity Cliff Cluster Topologies

- Topologies with \( \geq 3 \) instances
- Identification of 3 recurrent main topologies

<table>
<thead>
<tr>
<th>Topology</th>
<th>Diagram</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star</td>
<td><img src="image" alt="Star Diagram" /></td>
</tr>
<tr>
<td>Chain</td>
<td><img src="image" alt="Chain Diagram" /></td>
</tr>
<tr>
<td>Rectangle</td>
<td><img src="image" alt="Rectangle Diagram" /></td>
</tr>
</tbody>
</table>
Activity Cliff Cluster Topologies

- **Topologies with** $\geq 3$ instances

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

<table>
<thead>
<tr>
<th>Topology</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star</td>
<td><img src="image" alt="Star Diagram" /></td>
</tr>
<tr>
<td>Chain</td>
<td><img src="image" alt="Chain Diagram" /></td>
</tr>
<tr>
<td>Rectangle</td>
<td><img src="image" alt="Rectangle Diagram" /></td>
</tr>
</tbody>
</table>
# Main Topologies and Extensions

<table>
<thead>
<tr>
<th>Main topology</th>
<th>Extensions of main topology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star</td>
<td>Twin Star</td>
</tr>
<tr>
<td>Chain</td>
<td>Modified Chain</td>
</tr>
<tr>
<td>Rectangle</td>
<td>Modified Rectangle</td>
</tr>
</tbody>
</table>

![Diagram showing main topologies and their extensions](image)
# Main Topologies and Extensions

<table>
<thead>
<tr>
<th>Main topology</th>
<th>Extensions of main topology</th>
<th>Hybrid topologies</th>
<th>Irregular topologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star</td>
<td><img src="image1" alt="Star Extension" /></td>
<td><img src="image2" alt="Hybrid Star" /></td>
<td><img src="image3" alt="Irregular Star" /></td>
</tr>
<tr>
<td>Chain</td>
<td><img src="image4" alt="Chain Extension" /></td>
<td><img src="image5" alt="Hybrid Chain" /></td>
<td><img src="image6" alt="Irregular Chain" /></td>
</tr>
<tr>
<td>Rectangle</td>
<td><img src="image7" alt="Rectangle Extension" /></td>
<td><img src="image8" alt="Hybrid Rectangle" /></td>
<td><img src="image9" alt="Irregular Rectangle" /></td>
</tr>
</tbody>
</table>
Star Topology Example

- Adenosine A3 receptor ligands

---

The diagram illustrates the star topology example with various molecules and their pKi values. The molecules are connected in a star-like pattern, with each arm representing a different ligand. The pKi values for each ligand are indicated as follows:

- 6.2 pKi
- 5.6 pKi
- 8.3 pKi
- 4.8 pKi
- 5.1 pKi
- 5.5 pKi
Star Topology Example

- Adenosine A3 receptor ligands
Rectangle Topology Example

- Adenosine A2b receptor ligands

5.4 pKᵢ

8.0 pKᵢ

6.0 pKᵢ

8.3 pKᵢ
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

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

non-Cliff

Activity Cliff

\[
\begin{align*}
\text{MMP 1} & \\
pK_i &= 8.1 \\
pK_i &= 7.5 \\
\end{align*}
\]

\[
\begin{align*}
\text{MMP 2} & \\
pK_i &= 8.9 \\
pK_i &= 6.3 \\
\end{align*}
\]
Transformation Kernel

Step 1:
Fingerprint representation of transformation substructures

(fingerprints: structural keys or atom pairs)
Kernel for Compound Pairs

MMP 1

Core 1

Trans 1

MMP 2

Core 2

Trans 2

$\text{pK}_i = 8.1$

$\text{pK}_i = 7.5$

$\text{pK}_i = 8.9$

$\text{pK}_i = 6.3$
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

Step 3:
Kernel from substructure difference feature vectors

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

\[ K_{\text{Tanimoto}}(u, v) = \frac{\langle u, v \rangle}{\langle u, u \rangle + \langle v, v \rangle - \langle u, v \rangle} \]
Design of an MMP Kernel

Design principle:
- combine core structure and transformation information
- add core structure representation to substructure difference vectors
MMP Kernel

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

Combining core and transformation feature vectors yields a kernel product.

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

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

\[ K_{\text{MMP}} = K_{\text{Tanimoto}}(c_1, c_2) \times K_{\text{Tanimoto}}(t_1, t_2) \]
## Accurate Prediction of Activity Cliffs

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

### Table

<table>
<thead>
<tr>
<th>Target</th>
<th>Transformation kernel</th>
<th>MMP kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TPR</td>
<td>TNR</td>
</tr>
<tr>
<td>*fxa</td>
<td>72.69</td>
<td>91.79</td>
</tr>
<tr>
<td>*mcr4</td>
<td>78.15</td>
<td>96.72</td>
</tr>
<tr>
<td>*kor</td>
<td>66.87</td>
<td>88.92</td>
</tr>
<tr>
<td>*thr</td>
<td>81.15</td>
<td>90.23</td>
</tr>
<tr>
<td>*aa3</td>
<td>71.95</td>
<td>88.46</td>
</tr>
<tr>
<td>cal2</td>
<td>97.44</td>
<td>95.40</td>
</tr>
<tr>
<td>catb</td>
<td>88.33</td>
<td>97.56</td>
</tr>
<tr>
<td>dpp8</td>
<td>99.29</td>
<td>100.00</td>
</tr>
<tr>
<td>jak2</td>
<td>92.73</td>
<td>88.42</td>
</tr>
</tbody>
</table>

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

### Definitions:
- **TPR**: True positive rate
- **TNR**: True negative rate
- **P**: Precision
- **F-score**: \( \frac{2 \cdot \text{TPR} \cdot P}{\text{TPR} + P} \)
Accurate Prediction of Activity Cliffs

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

Important structural patterns

- Green circles: highly potent compounds
- Red circles: weakly potent compounds

Calpain 2 inhibitors
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