Quantitative Comparison of Three-Dimensional Activity Landscapes of Compound Data Sets Based upon Topological Features

Javed Iqbal, Martin Vogt, and Jürgen Bajorath*

1. INTRODUCTION

Activity landscapes (ALs) are graphical representations that are designed to integrate structural and potency relationships between compounds sharing the same specific activity. AL modeling enables the visualization and graphical analysis of structure−activity relationships (SARs) for different data sets.

Over the years, a variety of AL representations of different designs and complexities have been introduced. These include two-dimensional (2D) representations such as plots of pairwise compound similarity versus potency relationships or annotated similarity-based compound networks. In addition, three-dimensional (3D) AL views such as pairwise compound activity−property−similarity distributions or maps reminding us of geographical landscapes have been studied.

These maps, in the following referred to as 3D ALs, are particularly intuitive because they can be interpreted in the same way as geographical landscapes. In 3D ALs, topological features such as mountains, plains, or valleys are associated with different SAR characteristics. For example, plains and gently sloped valleys in a 3D AL result from a series of chemical modifications (“walks” in chemical space) that are accompanied by small to moderate changes in potency, a phenotype referred to as SAR continuity. By contrast, mountainous regions and peaks in a 3D AL are a consequence of small compound modifications (small steps in chemical space) that cause large potency alterations, which represent SAR discontinuity. Most prominent peaks in discontinuous regions are termed activity cliffs and formed by pairs or groups of structural analogues with largest potency differences in compound data sets.

For SAR visualization, 3D AL models are obtained by adding compound potency values as third dimension to a 2D projection of a chemical feature (descriptor) space, for which computational approaches and parameters have been established in previous studies. From experimental potency values, a coherent surface is interpolated and color-coded by potency, representing an activity surface. The major features of the geographical landscape like topologies of different activity surfaces, as discussed above, can vary significantly, and different topological features reflect different SAR characteristics.

SAR visualization using 3D ALs represents a qualitative approach. Although 3D AL analysis can be complemented with the application of numerical SAR analysis functions to quantify SAR continuity and/or discontinuity for compound data sets, subjective graphical assessment has principal limitations when comparing different ALs. This is the case because most compound data sets exhibit SAR heterogeneity, resulting

Received: July 30, 2020
Accepted: August 27, 2020
Published: September 10, 2020
from different combinations of locally continuous and discontinuous SARs that originate from different compound subsets. As a consequence, in corresponding 3D ALs, smooth and rugged regions are interspersed. When comparing 3D AL models, it is generally difficult to judge differences in SAR information content on the basis of visual inspection. Therefore, the ability to compare 3D ALs in quantitative terms would be highly desirable to complement graphical analysis.

We have reasoned that a quantitative comparison of 3D ALs might become feasible by focusing on quantitative analysis of images of 3D ALs. The first evidence in support of this conjecture was provided by successful classification of 3D AL image variants based upon features learned using convolutional neural networks. For 3D ALs, reference representations with specifically altered topologies were generated, either significantly increasing or decreasing the proportion of valleys to peaks. From images of original and modified ALs, distinguishing features were learned and used for class label predictions using machine learning, leading to overall successful classification of these AL variants.

In light of these initial findings, we have investigated 3D AL image analysis to quantitatively compare 3D ALs and the SAR information they capture. Herein, we represent a new

---

**Figure 1.** Activity landscape views and topological features. (a) At the top, an exemplary original 3D AL (left) and the corresponding heatmap (right) are shown. The heatmap conveys a top-down view of the AL. At the bottom, topological features extracted from the heatmap are depicted. For peaks and valleys, contours are drawn covering eight threshold levels ($\pm 0.25$, $\pm 0.5$, $\pm 0.75$, and $\pm 0.9$). (b) The generation of a feature vector is illustrated that records the cumulative area of shape features for any contour threshold levels. For clarity, only individual contours with an area greater than 3 are shown and labeled with respective thresholds.
computational approach to quantitatively account for feature relationships between 3D ALs, indicating varying SAR information content and quantify similarity relationships between 3D ALs.

2. RESULTS AND DISCUSSION

2.1. Study Concept. Images of 3D ALs provide a data source for algorithmic extraction of features that account for topological characteristics representing SARs contained in compound data sets. Our analysis was based on the premise that similarity relationships between different 3D ALs might be quantifiable if such image features could be canonicalized and formally compared. Therefore, we have recorded 3D AL images and converted them into representations, in which landscape topologies were defined by varying color pixel intensities and from which features suitable for 3D AL comparison could be systematically extracted. On the basis of the resulting feature sets, 3D AL similarity analysis was performed, and different comparisons were carried out. In the following, our computational approach is presented and applied to establish proof of concept.

2.2. Activity Landscape Image Processing. The potency surface of 3D ALs is determined by three variables capturing structural and potency relationships between compounds: distance, elevation, and color gradients. While distance accounts for structural relationships (walks in chemical space), both elevation and color gradients account for potency relationships. This inherent redundancy in representing potency relationships makes it possible to replace elevation with color profiles that capture AL topology by varying color pixel intensities. This idea is central to 3D AL image processing and comparison. Accordingly, original 3D AL images are converted into color-coded heatmaps that represent top-down views of the landscapes preserving distance and potency relationships, as illustrated in Figure 1. Furthermore, in Figure 2, an original 3D AL and the corresponding heatmap are compared in greater detail by mapping corresponding positions of exemplary active compounds participating in the formation of different activity cliffs within the same region.

2.3. Activity Landscape Features. The resulting heatmaps embed unique topological features as color profiles and color intensity-based textures. By well-defined contouring (see Section 3), shape features are defined, as shown in Figure 1a. Contours are derived based on the potency value distributions and thus capture different shape features on a relative scale, with corresponding threshold values distinguishing different contour levels. Characteristic features are then extracted and represented as AL-specific feature vectors for quantitative comparison. The generation of a feature vector accounting for the different shape features of an activity class is illustrated in Figure 1b. The feature extraction approach presented herein focuses on the detection of borders in heatmaps that encompass regions of different topologies and enclose valleys or peaks. Accordingly, different topological features yield different shapes. Feature extraction is facilitated using the marching squares algorithm (MSA)\(^{15}\) (see Section 3). Two main characteristics of shape features include the area that is covered and the color intensity range, for which thresholds are defined. Shapes representing peaks or valleys in 3D ALs are compared across different threshold levels and the
AL similarity is quantified in different ways using a suitable similarity metric.16

2.4. Exemplary Activity Landscapes. For our proof-of-concept studies, 3D ALs were generated for eight compound data sets (different activity classes) taken from release 23 of ChEMBL,17 as reported in Table 1. The compounds were active against diverse targets and covered varying potency ranges. A general cutoff value of 10 μM potency was applied to exclude borderline active compounds from 3D AL modeling. From 3D AL images, heatmaps were derived for further analysis. Figure 3 shows the heatmaps of these activity classes. The comparison reveals topological similarities and differences. All heatmaps combine mountainous and smooth regions containing peaks and valleys, respectively. These topological features mirror SAR heterogeneity. Within these variable landscapes, differences are observed. For example, the heatmap of ChEMBL data set 204 displays most prominent peaks and cliffs (characterized by sharp borders between peaks and valleys). Based upon visual inspection, the heatmap of data set 4792 appears to resemble this topological phenotype most closely among the others. Other topological similarities or differences are more difficult to judge on the basis of visual inspection. For example, it would hardly be possible to confidently predict whether the AL of set 3155 might be more similar to that of set 219 or 4550. Despite apparent data set-dependent topological differences captured by these heatmaps, it would generally be very difficult to judge (dis)similarity relationships between them, illustrating limitations of our perception. Accordingly, any consistently applicable quantitative measure of AL similarity would provide a substantial advance for AL-assisted SAR exploration.

2.5. Activity Landscape Comparison. On the basis of extracted image features, as discussed above, different types of comparisons were carried out, attempting to discern topological characteristics and relate topology to SAR information content.

First, we applied the weighted Jaccard coefficient ($J_w$) to compare feature vectors recording fractional heatmap areas contoured at different threshold levels (corresponding to different topological features). The formalism is presented in section 3.4. Importantly, the comparison of feature vectors did not depend on establishing correspondences between individual shapes. To avoid “averaging” over distinct topological features accounting for different SARs, comparisons were separately carried out for valleys (negative threshold levels) and peaks (positive thresholds), corresponding to SAR continuity and discontinuity, respectively. Feature vectors of activity classes were compared in a pairwise manner, and the ALs were then ranked separately for valleys and peaks in the order of descending similarity to set 204, which served as a reference AL. The results are reported in Table 2. As can be seen, feature vector comparison yielded meaningful rankings of landscapes with a significant spread of pairwise similarity values.

### Table 1. Compound Data Sets

| ChEMBL target ID | target name | compounds | potency ($p_K_i$) min | max |
|------------------|-------------|-----------|-----------------------|-----|
| 204              | coagulation factor II | 915       | 5                     | 12.2|
| 344              | melanin-concentrating hormone receptor 1 | 1175      | 5                     | 9.8 |
| 3155             | 5-hydroxytryptamine receptor 7 | 1094      | 5                     | 9.8 |
| 4792             | orexin receptor type 2 | 1443      | 5                     | 10.1|
| 255              | adenosine receptor A2b | 1237      | 5                     | 9.8 |
| 219              | dopamine D4 receptor | 1082      | 5                     | 10.5|
| 4550             | arachidonate 5-lipoxygenase-activating protein | 1318      | 5.6                   | 9.4 |
| 225              | serotonin receptor 2C | 1079      | 5                     | 9.7 |

*The composition of eight compound activity classes used for 3D AL analysis is summarized.*

### Table 2. Similarity-Based Ranking of Activity Landscapes

| rank | valley-based | similarity $J_{wV}$ | peak-based | similarity $J_{wP}$ |
|------|--------------|---------------------|------------|---------------------|
| 1    | 204          | 1                   | 204        | 1                   |
| 2    | 225          | 0.89                | 4792       | 0.58                |
| 3    | 219          | 0.87                | 3155       | 0.46                |
| 4    | 3155         | 0.85                | 219        | 0.31                |
| 5    | 255          | 0.78                | 255        | 0.28                |
| 6    | 344          | 0.69                | 225        | 0.15                |
| 7    | 4792         | 0.63                | 4550       | 0.14                |
| 8    | 4550         | 0.45                | 344        | 0.13                |

*3D ALs of activity classes were ranked in the order of decreasing similarity using the AL of data set 204 as a reference. IDs are used according to Table 1. Separate rankings are reported on the basis of features accounting for valleys and peaks, respectively.*

Figure 3. Heatmaps: for the eight activity classes, heatmaps derived from 3D AL images are shown. ChEMBL target IDs are reported according to Table 1.
(\(J_w\) ranges from 0 to 1). As expected, the relative ranking on the basis of valleys and peaks changed. Similarity values were generally higher for comparison of valleys than of peaks. This was expected, given the generally larger area covered by valleys in heatmaps. To the (limited) extent we were able to subjectively judge such similarity relationships, the rankings were reconcilable and intuitive. For example, in peak-based ranking, the AL of set 4792 (rank 2) was most similar to reference set 204, followed by set 3155 (rank 3), consistent with visual inspection. In valley-based ranking, the AL of set 4550 (rank 8) was most dissimilar to reference 204, which was also intuitive. On the other hand, on the basis of visual inspection, it was more difficult to understand why set 219 (rank 2) was more similar to 204 than set 225. Clearly, the similarity calculations numerically distinguished between pairwise relationships that were essentially impossible for us to judge, which we had aimed for. Successful ranking of different ALs according to topological features on the basis of calculated similarity values was considered an encouraging finding.

To assess the SAR information contained in shapes identified for specific thresholds, structure–activity similarity (SAS) maps\(^1,4\) were calculated for exemplary contours obtained at decreasing threshold levels, as shown in Figure 4 for a representative data set. SAS maps are scatter plots of compound pairs, in which the x-axis reports pairwise similarity and the y-axis pairwise potency differences. Here, the extended connectivity fingerprint with bond diameter 4 (ECFP4)\(^18\) was used as a molecular representation to calculate Tanimoto similarity.\(^19\) As can be seen, the contour of the highest threshold 0.75 (corresponding to a p\(K_i\) value of 8.5) comprises only very few, but very similar, compounds, with pairs formed by these compounds displaying potency differences of up to around two orders of magnitude. For the lower threshold of 0.5 (corresponding to a p\(K_i\) of 8.0), more diverse compounds are contained in the contoured region. The majority of compound pairs still share a Tanimoto similarity of 0.4 or higher, which is significant for ECFP4. Thus, given the wide spread of potency differences, this region of the 3D AL is rich in SAR information. For the threshold of 0.25 (corresponding to a p\(K_i\) of 7.5), the contour comprises many compounds with pairwise similarities of, on average, less than 0.4, which, by definition, do not yield meaningful SARs. Taken together, these findings demonstrate that by comparing only the highest threshold peak contours and lowest threshold valley contours (see Section 3), AL similarity analysis focuses on the most relevant regions with respect to SAR information content. This is the case because comparison of these features leads to the identification of similar compounds with largest potency differences in data sets that determine SARs.

**2.6. Conclusions.** Activity landscape representations are used for SAR visualization and aid in the exploration of SARs.
contained in compound data sets. Among these representations, 3D AL models are particularly intuitive. So far, comparisons of 3D ALs have been confined to the qualitative level. However, quantitative comparisons of 3D ALs would greatly help in assessing differences in SAR information content beyond of what can be appreciated on the basis of visual inspection. In this work, we have presented a computational methodology to facilitate quantitative comparisons of 3D ALs on the basis of topological features extracted from AL image data. As we have shown, numerical analysis discards similarity relationships between 3D ALs in a meaningful way and enables ranking of ALs according to relative differences in the SAR information they capture. In addition, we have demonstrated that compound subsets associated with different contoured areas representing defined topological features convey varying SAR information, as one would expect. Taken together, our findings suggest that the approach introduced complements to SAR visualization and further increases the potential of 3D ALs for large-scale SAR analysis.

3. MATERIALS AND METHODS

3.1. Three-Dimensional Activity Landscapes. Three-dimensional AL models of compound data sets were constructed as described.14 For 3D AL modeling, preferred approaches for dimensionality reduction of original chemical reference spaces, molecular representations, and similarity/distance calculations have been identified in earlier studies.5,8 Accordingly, chemical reference space was generated on the basis of ECFP416 as a molecular representation and calculation of pairwise compound Tanimoto distances.19 The 2D projection of chemical reference space was then computed using multidimensional scaling (MDS)20 applying a stress projection of chemical reference space was then computed applying a stress.

The potency gradient was applied to a limited p

channel.22 Because the original 3D AL models were created by interpolating potency values using a color gradient from red over yellow to green, without using the blue channel, the red and green (RG) channel pixel values were extracted by subtracting green channel intensity values from red channel intensity values and combined into a single intensity value ranging from −255 to 255. Hence, the least potent (brightest green), moderately potent (yellow), and highest potent (brightest red) compounds/pixels corresponded to values of −255, 0, and +255, respectively. RG pixel values were then normalized to the range of −1 to +1. The RG color model preserved more than 95% of the RGB colors, except for white regions (i.e., interpolated surface area without experimental potency backup), which was accounted for by yellow using the RG model. However, these regions only accounted for less than 5% of the surface.

3.3. Feature Extraction: Contours and Shapes. Feature extraction was performed on the basis of heatmaps of original size 543 × 543 pixels that were rescaled to 300 × 300 pixels. This representation corresponded to a top-down view of the original 3D AL color-coded as described above. Extraction of features proceeded in two steps. First, contour lines (i.e., lines of equal intensity in the image) were used to identify regions encompassing valleys and peaks in 3D ALs. The canonical heatmap representation was thus segmented into different regions using contour lines. The scikit-image implementation of MSA15,23 was applied to extract the contours. MSA represents the 2D version of the marching cubes algorithm,15 which creates a contour line segment by mapping an image onto a square grid.

For contour extraction, each heatmap was initially binarized using threshold values of 0.25, 0.5, 0.75, and 0.9, respectively, to delineate shapes representing peaks, while inversely binarized negative thresholds of −0.25, −0.5, −0.75, and −0.9 were used to identify smooth regions (valleys). The following MSA parameter settings were used: high connectivity, high positive orientation, and iso-line of level 7. The resulting contour lines represented nonintersecting closed curves. Shapes were subsequently characterized as groups of contour levels of increasing threshold magnitude, with higher-threshold contours being enclosed by lower-threshold contour lines. Contour areas were calculated on the basis of Green’s theorem using computed image moments.24 Threshold contours identified individual peaks of the AL for highest positive thresholds and valleys for lowest negative thresholds of individual shapes. Peak and valley contours were only considered if they were contained in at least one contour of a lower-magnitude threshold. Each peak and valley was then defined by its area and its threshold level.

3.4. Activity Landscape Similarity Analysis. To quantify the similarity of two 3D ALs images based upon their heatmaps, peaks and valleys were generated and compared. For each AL and threshold, the total areas of the peak and valley contours corresponding to the given threshold were determined, resulting in a feature vector comprising eight values

\[ a = (a_{-0.9}, a_{-0.75}, a_{-0.5}, a_{-0.25}, a_{0.25}, a_{0.5}, a_{0.75}, a_{0.9}) \]

This vector represents the total area of the peaks/valleys at each threshold level. For each pair of landscapes A, B (with corresponding feature vectors \( a, b \)), a similarity coefficient was established by calculating the weighted Jaccard index or Ružička similarity16

\[ J_d = \frac{\sum \min(a_i, b_i)}{\sum \max(a_i, b_i)} \]
The resulting coefficients for peaks and valleys were termed $J_{uv}$ \((A,B)\) and $J_{uv}$ \((A,B)\), respectively.

### AUTHOR INFORMATION

**Corresponding Author**

Jürgen Bajorath — Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, D-53115 Bonn, Germany; [orcid.org/0000-0002-0557-5714](https://orcid.org/0000-0002-0557-5714); Phone: 49-228-7369-100; Email: bajorath@bit.uni-bonn.de

**Authors**

Javed Iqbal — Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, D-53115 Bonn, Germany

Martin Vogt — Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, D-53115 Bonn, Germany; [orcid.org/0000-0002-3931-9516](https://orcid.org/0000-0002-3931-9516)

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.0c03659

**Author Contributions**

The study was carried out and the manuscript written with contributions of all authors. All authors have approved the final version of the manuscript.

**Notes**

The authors declare no competing financial interest.

### ACKNOWLEDGMENTS

J.I. is supported by a Ph.D. fellowship from the German Academic Exchange Service (DAAD) in collaboration with the Higher Education Commission (HEC) of Pakistan.

### REFERENCES

1. Shanmugasundaram, V.; Maggiora, G. M. In Characterizing Property and Activity Landscapes Using an Information-Theoretic Approach, Proceedings of 222nd American Chemical Society National Meeting, Division of Chemical Information, Chicago, IL, Aug 26–30, 2001; American Chemical Society: Washington, D.C., 2001; Abstract no. 77.

2. Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure-Activity Relationship Analysis. *J. Med. Chem.* 2010, 53, 8209–8223.

3. Peltason, L.; Iyer, P.; Bajorath, J. Rationalizing Three-dimensional Activity Landscapes and the Influence of Molecular Representations on Landscape Topology and the Formation of Activity Cliffs. *J. Chem. Inf. Model.* 2010, 50, 1021–1033.

4. Yongye, A. B.; Byler, K.; Santos, R.; Martínez-Mayorga, K.; Maggiora, G. M.; Medina-Franco, J. L. Censsion Models of Activity Landscapes with Multiple Chemical, Conformer, and Property Representations. *J. Chem. Inf. Model.* 2011, 51, 1259–1270.

5. Iyer, P.; Dimova, D.; Vogt, M.; Bajorath, J. Navigating High-Dimensional Activity Landscapes: Design and Application of the Ligand-Target Differentiation Map. *J. Chem. Inf. Model.* 2012, 52, 1962–1969.

6. Medina-Franco, J. P.; Navarrete-Vázquez, G.; Méndez-Lucio, O. Activity and Property Landscape Modeling is at the Interface of Chemoinformatics and Medicinal Chemistry. *Future Med. Chem.* 2015, 7, 1197–1211.

7. Vogt, M. Progress with Modeling Activity Landscapes in Drug Design. *Expert Opin. Drug Discovery* 2018, 13, 605–615.

8. Miyao, T.; Funatsu, K.; Bajorath, J. Three-dimensional Activity Landscape Models of Different Design and Their Application to Compound Mapping and Potency Prediction. *J. Chem. Inf. Model.* 2019, 59, 993–1004.

9. Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure–Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure–Activity Relationship Indices. *J. Med. Chem.* 2008, 51, 6075–6084.

10. Maggiora, G. M. On Outliers and Activity Cliffs — Why QSAR often Disappoints. *J. Chem. Inf. Model.* 2006, 46, 1535.

11. Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and their Utility in Medicinal Chemistry. *J. Med. Chem.* 2014, 57, 18–28.

12. Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure-Activity Relationships. *J. Med. Chem.* 2007, 50, S571–S578.

13. Guha, R.; Van Driessche, J. H. Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* 2008, 48, 646–658.

14. Iqbal, J.; Vogt, M.; Bajorath, J. Activity Landscape Image Analysis Using Convolutional Neural Networks. *J. Cheminf.* 2020, 12, e34.

15. Lorensen, W. E.; Cline, H. E. In *Marching Cubes: A High Resolution 3D Surface Construction Algorithm*, Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques; Association for Computing Machinery: New York, NY, 1987; pp 163–169.

16. Ružička, M. Anwendung mathematisch-statistischer Methoden in der Geobotanik (Synthetische Bearbeitung von Aufnahmen). *Biologie* 1958, 13, 647–661.

17. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. *ChEMBL: A Large-scale Bioactivity Database for Drug Discovery*. *Nucleic Acids Res.* 2012, 40, D1100–D1107.

18. Rogers, D.; Hahn, M. Extended-connectivity Fingerprint. *J. Chem. Inf. Model.* 2010, 50, 742–754.

19. Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants. *Science* 1966, 132, 1115–1118.

20. Borg, I.; Groenen, P. J. F. *Modern Multidimensional Scaling: Theory and Applications*; Springer: New York, 2005.

21. Rasmussen, C. E. Gaussian Processes in Machine Learning. In *Summer School on Machine Learning*; Springer: Berlin, Heidelberg, 2003; pp 63–71.

22. Bradski, G.; Kaehler, A. *Learning OpenCV: Computer Vision with the OpenCV Library*; O’Reilly Media, Inc.: Sebastopol, CA, USA, 2008.

23. Van der Walt, S.; Schönberger, J.; Nunez-Iglesias, J.; Boulogne, F.; Warner, J.; Yager, N.; Gouillart, E.; Yu, T. The Scikit-Image Contributors. *Scikit-Image: Image Processing in Python*. *PeerJ* 2014, 2, e453.

24. Edwards, C. H. *Calculus and Analytic Geometry*; Prentice-Hall College Division: Upper Saddle River, 1990.