From Quantification of Activity Landscape Topographies to Activity Cliff Analysis

Image Analysis from a Chemical Informatics Perspective

Kumulative Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
JAVED IQBAL
aus Kalat, Pakistan

Bonn
May, 2022

Abstract

Machine learning or multivariate statistical modeling is frequently applied in the chemical informatics domain to analyze the relationships between molecular structures and their target activities/properties. Structure-activity relationship (SAR) properties and key determinants have been vastly investigated with the help of simple linear to complex 2D graphs, 2D molecular fingerprints, and 3D conformation representations. SAR characteristics can be graphically visualized and analyzed with the help of 3D activity landscape (AL) models. 3D AL models help in a straightforward interpretation of SARs by characterizing them into smooth valleys and rugged mountainous regions, reminiscing SARs as a geographical map. 3D ALs graphical nature makes them readily convertible to image data, thus amenable for image analysis. 3D AL models graphical representation of important SAR characteristics, and their readily convertibility into image formats permit SAR analysis based on image processing approaches. However, 3D ALs representation has thus far been qualitatively analyzed and limited to very few quantitative and predictive analyses. Convolutional neural networks (CNNs) are currently gaining increasing attention in chemical informatics domain to learn key structural features and derive predictions based on compound image representations. CNNs have the potential to automatically extract distinct spatial and temporal features from images, audio, and video data, which is preferable as compared to classical machine learning modeling. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) gave birth to many complex and deep CNN model architectures that have solved several other complex image analysis problems afterward. This dissertation presents contributions that advance the application of CNN modeling for SAR and activity cliff analysis and the application of recent image processing approaches to develop methodologies to quantify SAR characteristics numerically.

Acknowledgements

Contents

Chapter 1

Introduction

1.1 Structure-Activity Relationship

Structure-Activity Relationship (SAR) analysis is considered an essential medicinal chemistry task.[1] SAR exploration facilitates the precise understanding of relationship between a molecule's chemical structure and its bioactivity. SAR analysis assists in identifying fundamental structural features of compounds that can help determine compounds bioactivity. Identifying key structural features can further aid in the design of novel molecules. SAR information can be conventionally collected and analyzed with the help of an "R-group table" . The "R-group table" represents the core and analogs together in a table and is applicable for SAR analysis with a limited number of analogs. However, the technique becomes more impracticable with increasing number of analogs (thousand or even more). Developments in modern computational approaches have paved the path to analyze large-scale SARs of compound datasets. Some commonly practiced computational approaches for qualitative SAR analysis include 2D activity landscape,[2] 3D activity landscape,[3] scaffold tree,[4] SAR matrix,[5] numerical SAR index,[6] network-like similarity graph (NSG),[7] and LASSO graph.[8] In addition to qualitative SAR analysis, numerical estimation of the relationship between the molecular structure and its bioactivity is also highly important. Discovery of numerical relationship between molecular structure and its experimentally measured property or bioactivity can be efficiently estimated with the help of quantitative structure-activity relationship (QSAR) models. QSAR modeling numerically attempts to quantify any linear or non-linear relationship between numerical molecular descriptors and bioactivity in an SAR with the application of latest state-of-the-art machine learning and artificial intelligence algorithms. Precise

relationship estimation allows accurate bioactivity prediction leading to novel drug designs with the desired properties.[9]

1.2   SAR Characterization and Visualization

SAR exploration is commonly employed to understand the relationship of structural features of active compounds to their biological activity or potency.[10] SAR is commonly classified into continuous, discontinuous and, heterogeneous classes based on compound structural similarity and observed changes in bioactivity associated with structural modifications.[11] Small changes in compound structures resulting in minor to moderate potency changes form a continuous SAR. In contrast, minor structural changes that drastically alter potency values are categorized as discontinuous SARs.
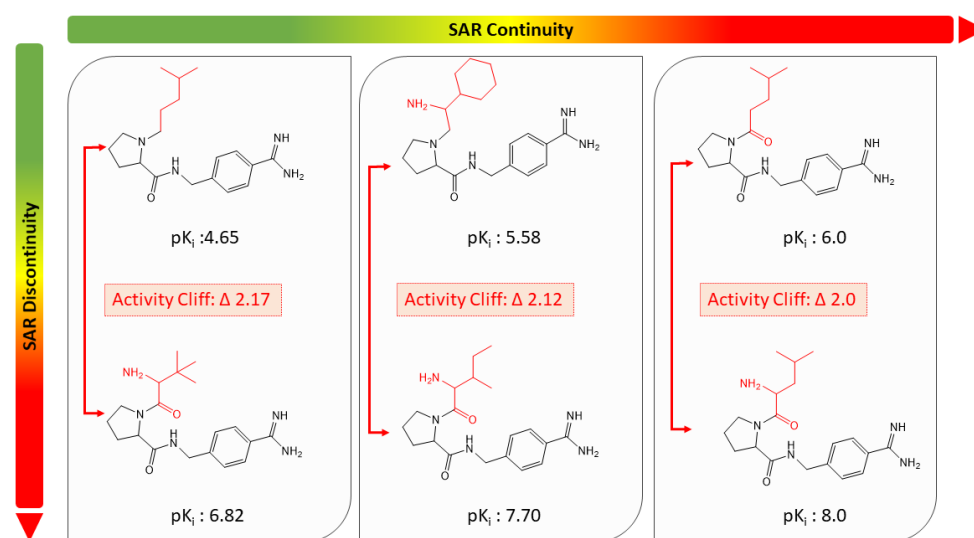
Figure 1: SAR Characterization. Exemplary SAR characterization of thrombin inhibitors with continuity (horizontal) and discontinuity (vertical) is demonstrated. Compound structures with activity values are shown and structural differences are highlighted in red.

Compound data sets sharing specific biological activity may often contain a combination of both continuous and discontinuous SAR types.[11] Accordingly, SARs with combined continuous and discontinuous components are known as heterogeneous SARs. Figure 1 demonstrates an exemplary SAR characterization of thrombin inhibitors.

1.3 Activity/Property Landscapes

Understanding structural similarity and potency relationships can help in rationalizing SAR information and identification of distinct SAR phenotypes. Activity landscape (AL) models aid in assessment of compound structural relationship with its bioactivity and facilitate SAR graphical visualization in biological target activity classes. Property landscape (PL) generalizes the AL concept by extending it with any numerically measureable experimental molecular structure-dependent property (i.e., toxicity, physicochemical properties, etc.). AL/PL of varying design and complexity are available, like two-dimensional (2D) and three-dimensional (3D) activity/property landscape representations. Structure-activity similarity (SAS) maps are commonly applied to understand relations between compound structural similarity with its bioactivity in a 2D graphical representation[12] whereas, 3D AL/PL models integrate compound structural similarity and its bioactivity in a 3D graphical illustration allowing intuitive visualization of global and local structure-activity/property relationships.[13,14] 3D ALs/PLs are generated following a generally applicable modeling protocol.[12] In the first step, compound data sets are projected within a given chemical feature space. A linear or non-linear reductional approach is applied to convert the chemical feature space into a 2D projection. Finally, the bioactivity/property values are interpolated, generating a coherent activity/property surface, colored using a continuous spectrum for the potency range in the data set. 3D ALs construct SAR characteristics in an easily interpretable manner by representing SAR characteristics in the nature of geographical maps, where the differences in landscape topographies can be viewed as smooth valleys or rugged mountains that mirror different SAR characteristics. Smooth or gently sloped regions correspond to a continuous SAR where small structural changes lead to gradual potency changes. The second prominent feature of ALs are rugged/mountainous regions which reflect SAR discontinuity, where small chemical changes cause large potency effects, resulting in the formation of activity peaks. 3D ALs representations often combine smooth and rugged/mountainous regions and form heterogeneous surfaces. Exemplary 2D SAS map and 3D AL are shown in Figure 2.

1.4 Molecular Representations

QSAR model performance strongly depends upon the provided molecular representation. Simple to advanced representations have been developed and extensively
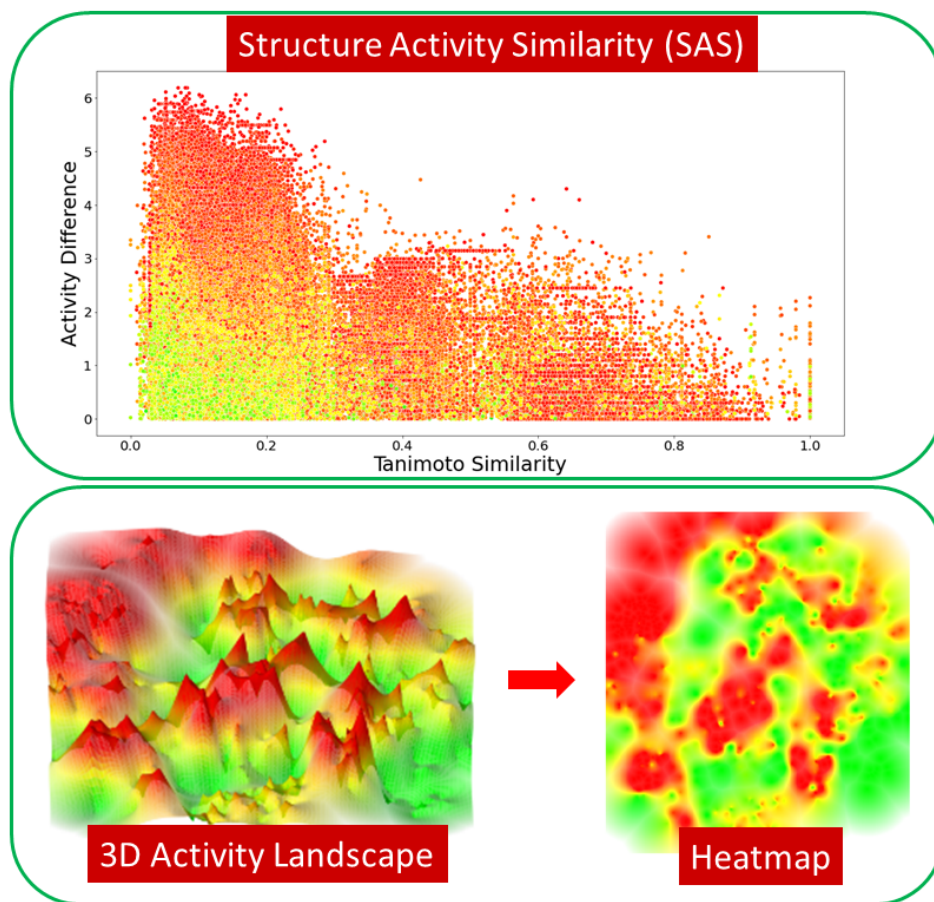
Figure 2: SAR visualization of Muscarinic acetylcholine receptor M3. SAR visualization with SAS map (top) and 3D AL topographical and corresponding heatmap (bottom) is shown. In SAS map, compound pairs are colored with the highest activity.

tested for various prediction tasks. A molecule can be generally represented in a linear or one-dimensional (1D), two-dimensional (2D), or three-dimensional (3D) data structure. The linear representation is simple and easily readable/writeable by humans or machines. The Simplified Molecular Input Line Entry System (SMILES),[15] IUPAC international chemical identifier (InChI),[16] Wiswesser Line-Formula Notation (WLN),[17] representation of structure diagram arranged linearly (ROSDAL)[18] and, Sybyl Line Notation (SLN)[19] are some examples of linear molecular structure representation. SMILES and InChI are the most widely used linear notations.[20] SMILES representation linearly encodes atom types, bond types, branching, stereochemistry, cyclic, aromaticity, charges, and many other structural properties present in a compound. Whereas InChI encodes the compound structural details in layers and sublayers with addition of tautomeric information. These linear representations are commonly used for searching and storage of compounds in databases preserv-

4

ing the graph structure of the molecule. The 2D and 3D representations encode the molecular structure in an advanced manner using graph data structures to retain a compound's structural and topology information. 2D Molecular graphs are commonly used to generate 2D fingerprints, 2D structural drawings and are favorable by latest state-of-the-art graph based machine learning models. 3D structural representation extends 2D structural representation by additionally encoding the conformational information (if any). Examples of 1D, 2D and 3D molecular representations are shown in Figure 3.

SMILES

CC1CCC(c(c)c)c(O)c1

Composite Formula

C10H20O

InChI

InChI=1S/C10H20O/c1-7(2)9-5-4-8(3)6-10(9)11/h7-11H,4-6H2,1-3H3

1D

2D Molecular Graph

2D

3D Molecular Graph

3D
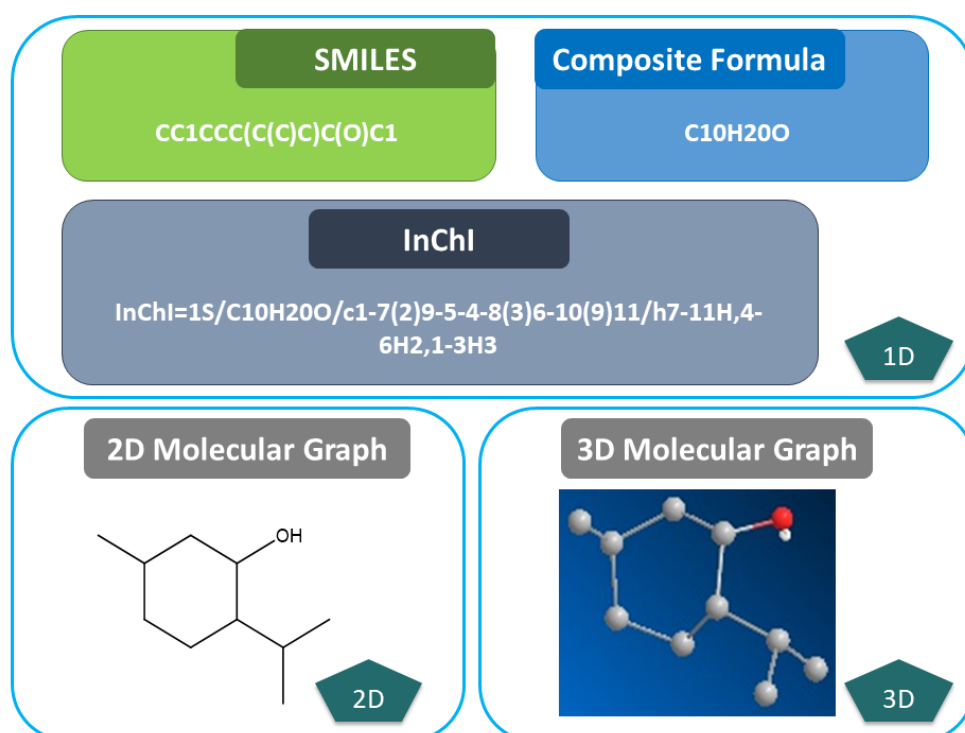
Figure 3: Molecular Representations. Exemplary 1D, 2D, and 3D molecular representations of a small molecule menthol are shown.

1.5 Molecular Fingerprints

Molecular fingerprints are special data structures that transform structural characteristics of a given compound into a vector.[21,22] In chemical informatics, fingerprints are extensively used for similarity comparisons, virtual screening, and the construction of chemical space maps.[23–26] The most commonly used molecular fingerprints are the Morgan fingerprints[27] also known as extended-connectivity fingerprints (ECFP)[28] and Molecular ACCESS System (MACCS).[29] ECFP fingerprints

derive specific circular sub-structures within a given diameter around each atom in a molecule (Figure 4). Derived circular sub-structures are then hashed into a vector representing the absence or presence of the corresponding sub-structure in off or on bits. ECFP fingerprints with large dimensions can be optionally folded into fixed-length vectors to decrease feature dimensions. In contrast MACCS fingerprints only encode the absence or presence of a predefined set of 166 structural patterns in off or on bits. Recent experiments based on modern state-of-the-art machine learning models using MACCS and ECFP fingerprints have shown superior performance of ECFP fingerprints over MACCS fingerprints suggesting preferability of ECFP fingerprints over MACCS fingerprints for several predictive tasks. [30–34]
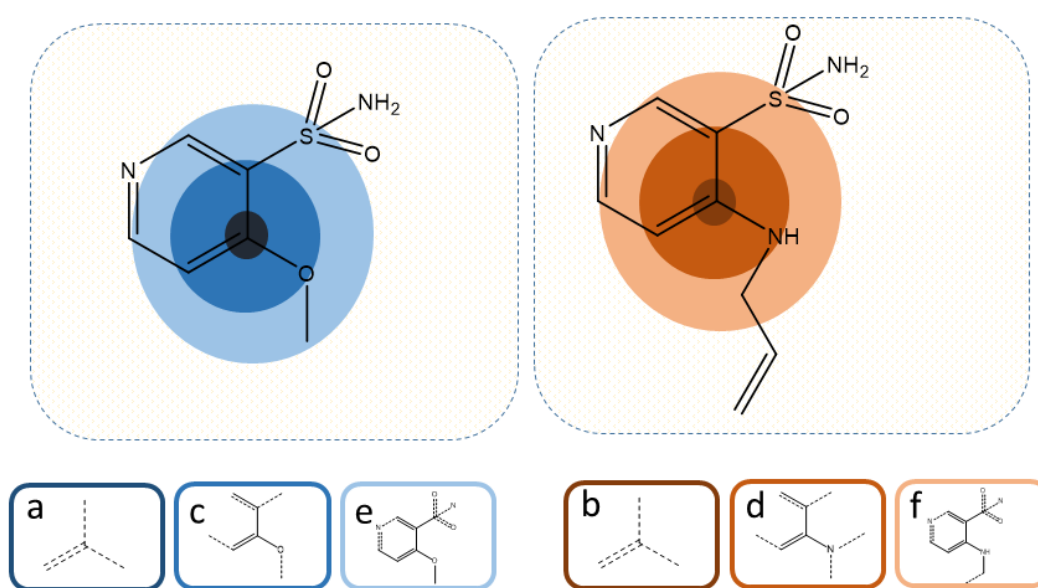
Figure 4: Circular Molecular Fingerprints. Molecular circular fingerprints along with atom environment (dashed bonds) for only a single atom (highlighted) in two analogs are shown. Circular features a-b, c-d, and e-f are generated from the radius of 0, 1, and 2, respectively.

1.6   Molecular Similarity Assessment

Molecular similarity is an important concept in medicinal chemistry and one of the utmost frequently investigated concepts in chemical informatics. Molecular similarity mainly focuses on the quantification of structural similarity (i.e., shared substructures, topologies, ring systems, etc.) between different compounds. Various computational methods have been introduced to assess the molecular similarity

6

between chemical compounds. Measurement of molecular similarity depends on the chemical representation, potential weighting of features representation, and the applied similarity coefficient/function.[35]

1.6.1 Fingerprint based Molecular Similarity

In chemical informatics, molecular similarity analysis based on fingerprints requires vectors of fingerprints to encode the molecular structural properties and a similarity measure to quantify the molecular similarity. The similarity measure is applied on fingerprints to usually yield a value between zero and one, where one indicates the highest similarity or identical fingerprints and zero indicates lowest or no-similarity.[35] Tanimoto coefficient (Tc),[36] also known as Jaccard index is one of the most frequently applied similarity measure to quantify fingerprint based molecular similarity. Tc similarity value for a given pair of compounds quantifies the percentage of structural features shared between the compounds. The Tc formulae can be defined as:

$$Tc(A,B) = \frac{c}{a+b-c}$$

Where a and b denotes the number of chemical features present in compound A and B, respectively, and c denotes the number of common features shared by both compound A and B.

1.6.2 Matched Molecular Pair Formalism

A matched molecular pair (MMP), also referred to as a single-site analog pair is defined as a pair of compounds that share a common core structure and the compounds, are only distinguished by an R-group replacement or a chemical modification at a single site (chemical transformation).[37,38] Figure 5 shows an exemplary analog pair according to the MMP formalism. The easy interpretability and intuitive nature of the MMP formalism makes the MMP concept highly attractive for medicinal chemists and increases its applicability in chemical informatics for understanding SAR key features in a chemical perspective.

1.7 Activity Cliffs

Activity cliffs (ACs) characterize the most prominent features of SAR discontinuity, giving rise to activity peaks or deep valleys in a 3D AL model. Medicinal
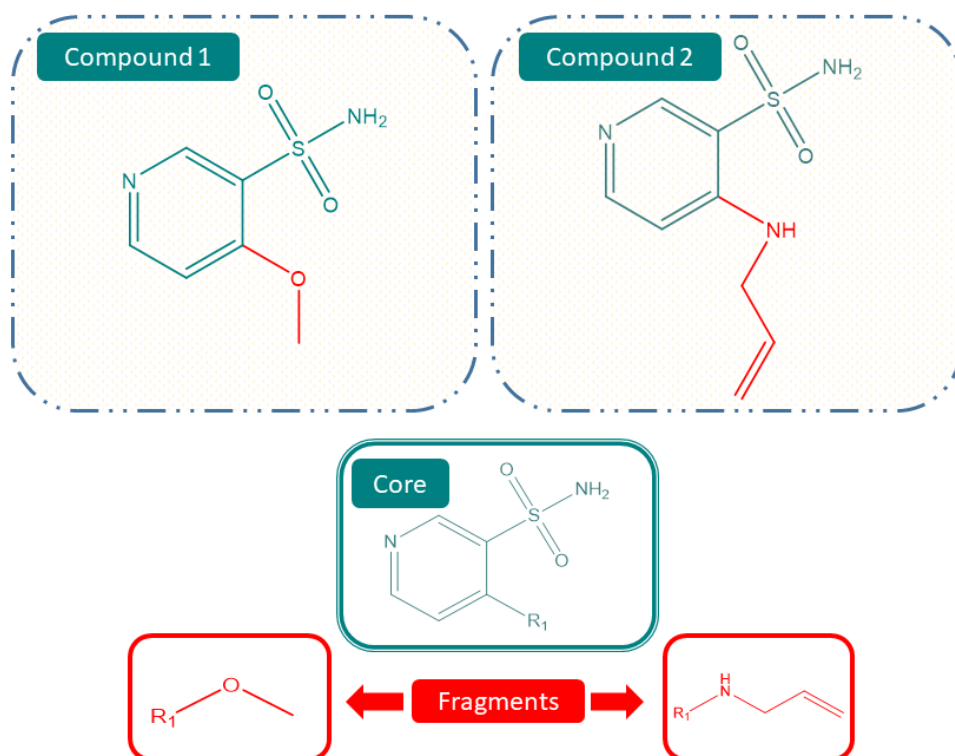
Figure 5: MMP Formalism. MMP formalism based on two compounds are shown (top). A shared core and two fragments that only differ by a chemical transformation are shown (bottom).

chemists often encounter ACs during the chemical optimization of individual compound series.[39] ACs are defined as structurally similar compound pairs with a significant change in potency values.[40,41] The identification of small structural changes that result in a significant potency difference or render compounds inactive is of high consideration. ACs frequently exist in discontinuous SARs and are considered highly explanatory for SAR analysis.[42] ACs characterization requires a clear definition of structural similarity and potency difference criteria.[39] The potency difference criterion defines a statistically significant and sufficiently large potency difference for AC formation. A compound potency difference of 100-fold change is frequently applied regardless of target activity class. In addition to the 100-fold potency difference criterion, alternative reasonable criteria based on the significance of target-dependent potency differences can also be applied for defining ACs.[43] The compound pair similarity criterion requires assessing similarity based on molecular fingerprints or, alternatively, the MMP formalism concept. Similarity criteria based on MMP formalism concept are highly suitable for defining ACs.[40] ACs that follow the MMP formalism concept and have statistically significant potency differences

are termed as MMP-cliffs.[44] The MMP formalism concept permits the limitation of MMP-cliff transformation size to obtain relatively small replacements, permitting interpretation of transformation in a chemically meaningful way.[44] MMP-cliffs have been intensely investigated for potency difference prediction using regression models and classification prediction of MMP-cliffs using fingerprint and the condensed graph of reaction formalism.[45,46]

1.8 Machine Learning in Chemical Informatics

In chemical informatics, machine learning or multivariate statistical modeling are frequently applied to estimate the relationship between descriptors encoding the chemical structure and a target property, such as molecular bioactivity. Machine learning models universally process the input data of any domain irrespective of the nature of data type.[9] For example, machine learning modeling has been successfully applied on the basis of various molecular representations such as SMILES,[47] 2D graphs,[48] 2D molecular fingerprints,[30–34,49] 3D conformations,[50] and molecular images.[51,52] The commonly applied machine learning methods in the domain of chemical informatics include support vector machines (SVM), random forests, K- nearest neighbors, and deep learning models. Application of machine learning methods has yielded robust virtual screening, accurate compound classification, and precise compound property/activity prediction compared to classical statistical methods.[30–34,49]

1.8.1 Support Vector Machines

Support Vector Machines (SVM)[53] are among the most widely applied machine learning techniques for compound classification and compound property/activity prediction tasks. SVM modeling is found to be exceptionally compatible and predictive with fingerprint-based compound representations.[30,31,34] An SVM classifier projects molecular fingerprints or features of different class labels into a high-dimensional space and attempts to find a hyperplane H that separates the binary labeled classes for the projected data points.

Figure 6 demonstrates an SVM model separating different class labels with an hyperplane. SVM can easily find infinite hyperplanes to classify the linearly separable projected data points. However, an SVM chooses the unique H that maximizes the margin between the closest points of each label (support vectors). Then, SVM
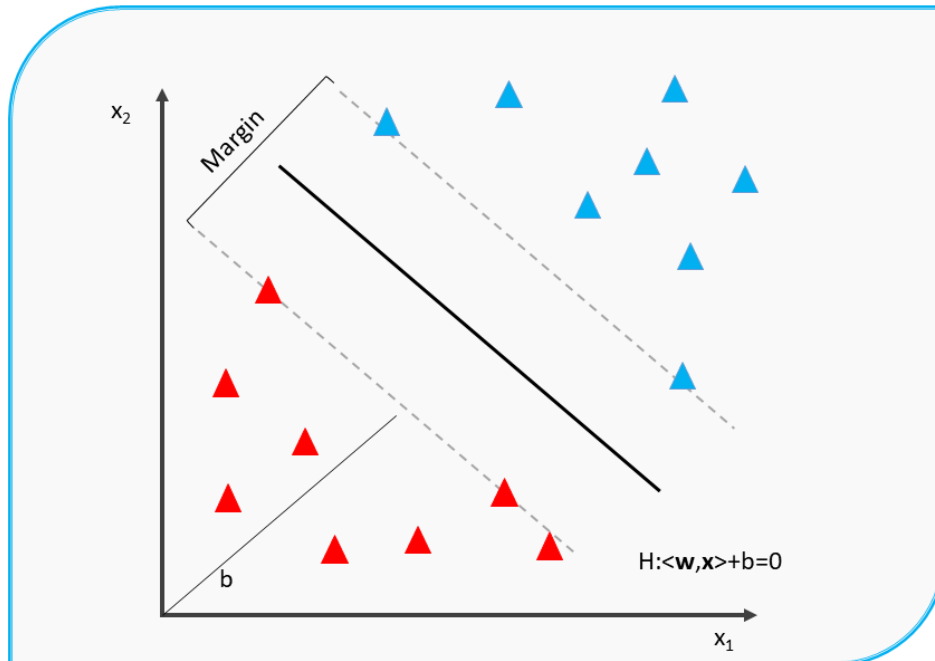
Figure 6: SVM. SVM classifer separating positive (blue triangle) and negative (red triangle) classes by finding an hyperplane H with a maximum margin is shown.

assigns labels to the test data based on which side of H the instance is present. The hyperplane H can be defined as below:

$$H = \{x | \langle w, x \rangle + b = 0\}$$

Where w is the normal vector, b is the bias and $\langle ., . \rangle$ is a scalar product. SVM requires that, the following conditions must be satisfied to ensure the correct classification of all training instances:

$$y_i(\langle x_i, w \rangle + b) \geq 1 \quad \forall_i$$

Here x_i are the training instances and $y_i \in -1, 1$ is the binary class label (negative or positive) for each training instance. The distance between the support vectors and H is given by $\frac{1}{\|w\|}$, which is maximized margin to find an optimal hyperplane. For non-linearly separable data points the "kernel trick" can be applied. A strength of SVMs is that the scalar function can be replaced with a kernel function K $\langle ., . \rangle$. In SVM,

the kernel function replaces the ordinary scalar product yielding a scalar product in an the implicit high-dimensional space improving the separability of data points.

1.8.2 Random Forest Classifier

The random forest classifier,[54] as illustrated in Figure 7, is a meta-algorithm that uses an ensemble of decision trees to to solve a classification problem using bagging and boosting approaches.
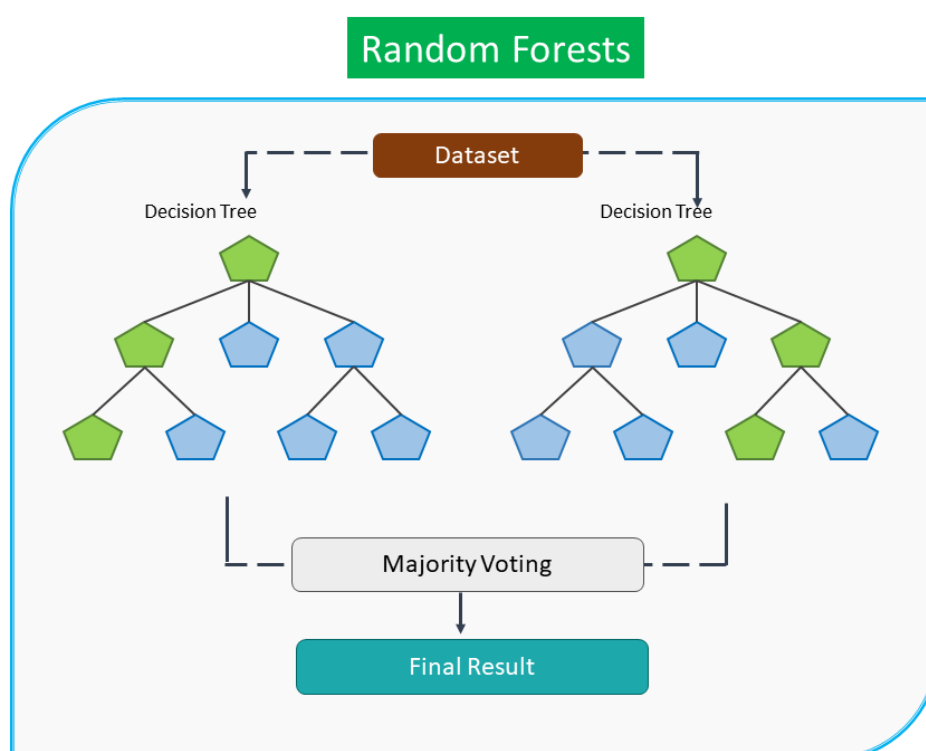
Figure 7: Random Forest Classifier. Random forest classifier processing the data using an ensemble of two decision trees is shown and the outcome of each leaf node is colored either blue or green, depending on the predicted class.

A random forest classifier consists of a number of decision trees, where each decision tree contributes a single vote to resolve the problem. As each tree may explore a different part of the solution space, the random forest classifier must ultimately determine the overall solution to the problem, which is done by considering the majority of votes. Random forest classifier has been widely applied in various fingerprint representation based classification tasks in the chemical informatics domain and achieved impressive results.[30–34]

11

1.8.3 Deep Learning

Deep learning algorithms are based on a neural network architecture. An artificial neuron or simply as neuron is a small set of computations that forms the basic unit of an artifical neural network. The artificial neuron processes information similar to a human neuron, which are the nerve cells and act as biological information processing machines. An artificial neuron is basically composed of three basic layers: an input layer, an output layer and one or multiple hidden layers. The input unit takes the input data which is then processed by hidden layers and finally output unit provides the output values of the processed data. An artificial neuron's output is usually transformed via a non-linear activation function to yield an output value. For a node i the output value Y_i is calculated as follows:

$$Y_i = g(\sum_j W_{ij} * a_j)$$

Where g is generally a nonlinear function, W_{ij} is the weight of the input node j on node i and a_j refers to input variables. Stacking of artificial neurons like multilayer perceptron (MLP) and organizing networks in multiple possible ways by placing them in a series of layers has proven to be highly flexible and extremely powerful. An example of a deep learning network involving an input layer, several hidden layers, and an output layer is shown in Figure 8.

Figure 8: Deep Neural Network. An exemplary deep neural network with an input, output and three hidden layers is demonstrated. Nodes are shown in circles and each node represent a neuron.

Deep learning has recently gained increasing attention from researchers of different domains and has been successfully applied to solve numerous real-world p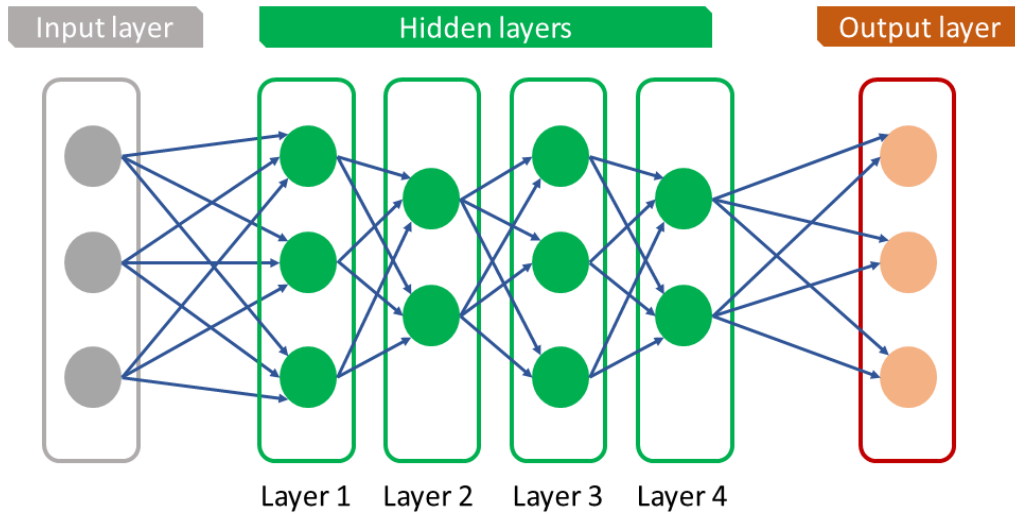roblems. Deep learning models automatically learn high-level features from data with the help of unsupervised or semi-supervised feature learning algorithms and hierarchical feature extraction, which transcends the limited feature representations of traditional machine learning. Deep learning modeling has been successfully applied in the chemical informatics domain for various analyses, including compound bioactivity prediction,[55] compound physicochemical property prediction,[56,57] multi-task property prediction,[58,59] drug design,[60–62] reaction prediction and retrosynthetic analysis,[63,64] and, ligand-protein interactions.[65]

1.8.4 Convolutional Neural Networks

Convolutional neural networks (CNNs) are deep neural networks that comprise convolutional layers to encode the locality information in the network structure.[66,67] CNNs are the most commonly applied algorithms for image-based analysis tasks.[68] The CNN structure was inspired by neurons in human and animal brains. CNNs primarily simulate the sequence of cells that forms the visual cortex in a cat's brain.[69] The commonly used CNN structure is similar to the MLP. CNN architecture commonly consists of at least one or more convolution layers, preceding sub-sampling (pooling) layers and fully connected layers as ending layers (Figure 9). In a CNN model architecture, input x of each convolutional layer is organized in height, width, and depth. Depth is the number of channels of input data. For example, an image using the RGB color model consists of red, green and blue color channels; therefore, the depth of an RGB image will be 3. Each convolutional layer can apply several filters (kernels) denoted by k with dimensions similar to the input image. The kernels form the basis to estimate the local connections between pixels, and share similar parameters (bias b^k and weight W^k) for generating k feature maps h^k. The convolution layer computes a dot product between the input and the weights and finally, by applying the nonlinearity or an activation function g to the convolution-layer output, we obtain the following:

$$h^k = g(W^k * x + b^k)$$

In order to accelerate the training process, CNNs network parameters are reduced by down-sampling every feature map into sub-sampling layers with the help of a pooling function (e.g., max or average). The fully connected layers create the high-level abstraction of mid-and low-level features and represent the last-stage layers as

13

a typical neural network. Finally, the ending layer applies an activation function to generate the scores (e.g., softmax for a binary label classification). The computed scores can be treated as probability distributions of a specific class.
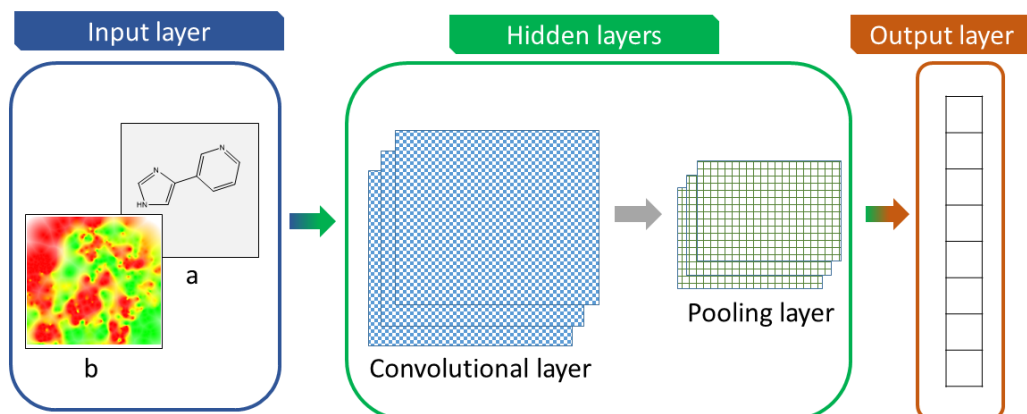
Figure 9: CNN Architecture. An exemplary CNN architecture with an input, output, and hidden layers is demonstrated.

1.8.5 Deep CNN Architectures

Over the years, a variety of fundamental CNN architecture have been developed to solve complex image analysis problems. ImageNet Large Scale Visual Recognition Challenge (ILSVRC)[70] paved the path for introducing complex deep CNN architectures to successfully discriminate and annotate different objects belonging to one thousand categories in image representation. The birth of numerous CNN architectures along with award-winning deep CNN architectures AlexNet[71](2012), GoogLeNet[72] (2014), and ResNet[73] (2015), dramatically reduced the top-5 classification error rate from 26% to less than 3% and rapidly evolved the CNN architectures depth and complexity. LeNet-5[74] is one of the earliest deep CNN architectures, commonly applied to recognize handwritten digits (MNIST)[75] promoted the further development of deep and complex CNN architectures. AlexNet, the winner of ILSVRC 2012, extended LeNet-5 architecture further larger, deeper, and introduced stacking of convolutional layers directly on top of each other, instead of stacking a pooling layer on top of each convolutional layer.[71] GoogLeNet architecture, also known as Inception, developed by Christian Szegedy et al. from Google Research, reached another milestone by pushing the top-5 error rate below 7% and winning ILSVRC 2014 challenge.[72] GoogLeNet architecture was considerably deeper than AlexNet with the introduction of sub-networks termed as inception modules. The

introduction of sub-networks dramatically reduced the network parameters (from 60 million to roughly 6 million) and resulted into an optimized CNN architecture with reduced training complexity. ResNet, the winner of 2015 challenge, introduced another extremely deep CNN architecture consisting of 152 layers which further improved the top-5 error rate by pushing it to under 3.6%.[73] Many of the deep CNN architectures were improved over time to increase the model efficiency for better performance and lower computational costs. For example, Inception v2,[76] v3,[77] and v4[78] are improved versions of the GoogLeNet's initial architecture.

1.8.6 Transfer Learning

A deep learning model requires a large amount of data to understand the latent patterns of data. Hence, the deep learning model performance depends on the training data size. Training a deep learning model on large-enough data allows the pre-order layers of the model to discover and identify high-level features of the training. However, building a large-scale dataset from manual data curation with high-quality annotation is a complex and expensive task. For example, ChEMBL[79] (version 27) is a large-scale database that is manually compiled using published literature and annotates around two million distinct compounds with activity values against more than 13,000 targets. However, the limited availibility of tested compounds against some targets often presents machine learning algorithms with insufficient training data. Transfer learning attempts to solve the fundamental problem of insufficient training data by transferring the knowledge from the one domain to another domain. In addition, transfer learning significantly reduces the training time by training only a part of the complete network. Tan et al.,[80] defines transfer learning as:

"Given a learning task T_t based on D_t, and we can get the help from D_s for the learning task T_s. Transfer learning aims to improve the performance of predictive function $f_T(.)$ for learning task T_t by $T_s \neq T_t$. In addition, in the most case, the size of D_s is much larger than the size of D_t, $N_s >> N_t$. "

In various studies, deep learning has demonstrated its capabilities to learn features from one task and transfer the knowledge to many other prediction analysis tasks.[62,81]

1.9   Image Analysis in Chemical Informatics

Machine learning models take features as input and attempt to fit a mathematical model to derive insights or make predictions from data. A feature is a numerical value that represents an aspect of raw data.[82] In digital image processing, the images are treated as picture elements (pixels) organized as an array of numbers. The very basic features in an image are the image pixels, which can be extracted quickly and compared in a meaningful way. In chemical informatics, very different objects ranging from a basic molecular structure to a complex QSAR visualization (e.g., 3D AL) can be easily exported into a high-quality graphical format. Graphical formats can then be further processed and analyzed using contemporary image processing techniques. In recent years image-based modeling has gained increased attention in chemical informatics and drug discovery. For example, CNN modeling on 2D images was applied to analyze bioactivity profiles,[83,84] compound toxicity prediction,[52] and cancer cell line sensitivity and compound potency prediction.[51] An image often contains a large number of pixels and requires large computational resources. CNNs can automatically identify and extract relevant features without any human supervision.[85] However traditional machine learning models require digital image processing techniques that help in detection and extraction of features from an image. The detected features can be quickly assembled into a feature vector and fed to classical machine learning models. Image features including edge detection, shape detection, corner and boundary detection, histograms, Scale Invariant Feature Transform (SIFT),[86] or Speeded-Up Robust Features(SURF)[87] have shown remarkable milestones for various image analysis tasks using image processing approaches. As 3D AL models are reminiscent of real-world geographical landscapes, image processing approaches can be employed on 3D ALs to detect edge or topographical features for the quantification of elevation-based SAR characteristics.

1.9.1   Helping Computers Read Edge Features

The objects in an image often contain different intensities compared to their surroundings, especially the boundary of an object shows a step-change in the intensity levels. The step-change in intensity level allows humans to perceive the perimeter of an object. Edge detector operators detect step-change in the intensity levels and detect objects similar to natural human vision. Edge detection operators apply the first-order differentiation or second-order differentiation in an image to detect the edge

features.[88] Some commonly used examples of first-order differentiation operators include Roberts,[89] Prewitt,[90] Canny edge detection operator,[91] and Sobel.[92] Laplacian,[93] Zero-crossing detection and Laplacian of Gaussian are some examples of second-order differentiation operators. A step-change in intensity can be revealed by differencing adjacent points. Differentiating horizontally adjacent points will detect vertical changes in intensity and is often applied to detect horizontal edges. Whereas differencing vertically adjacent points detects horizontal changes in intensity and is often applied to detect vertical edges. Edge detection can detect important features from almost all graphical representations. For example, vertical edge detection can be applied on 3D AL images to extract elevation-based features of topographical features.

1.9.2 Helping Computers Read Topographical Features

Edges are automatically detected and require significantly less computational resources. However, edges can not take the shape information into account. The shape information is also of high importance in images. For example, in real-world geographical images topographical characteristics are localized into different shapes and contain different information based on color profiles/textures. Intensity thresholding based contouring of a real-world geographical based top-down map allows the identification of topographical characteristics (i.e., mountains, peaks, trees, rivers, valleys, etc.). 3D ALs are reminiscent to real-world geographical maps, therefore, intensity thresholding based contouring can be applied to extract topographical characteristics and transform them into a feature vector for further analysis.

1.10   Thesis Outline

This PhD thesis focuses on the applicability of state-of-the-art machine learning models and recent image processing approaches for classification prediction, and feature analysis of various SARs and ACs. The analysis attempts to investigate SARs for the existence and detection of spatial features that help machine learning models to discriminate between different SAR surfaces and attempts to extract and quantify spatial features for further analysis. Moreover, ACs, considered as the prominent features of a SAR, are investigated for classification prediction using molecular image representations. Furthermore, the influence and impact of knowledge transfer

of specific structural features (i.e., R-groups, functional groups) using deep CNN models for AC prediction analysis was also investigated.

- *Chapter 2* discusses the applicability of CNN and other machine learning modeling for successful discrimination of different SARs based on 3D AL image data.

- *Chapter 3* aims to discover, extract and numerically quantify activity profiles and textures embedded in 3D ALs image data leading to a successful comparison of various 3D ALs with varying SAR contents.

- *Chapter 4* extends *Chapter 3* by identifying 3D AL topographical features and quantifying bioactivity distribution localized in topographical regions. Moreover, the chapter discusses numerical quantification of topographical features based on spatial bioactivity distribution for similarity comparision of different SARs based on 3D AL image data.

- *Chapter 5* reviews the different designs and complexities of 3D ALs/PLs. Moreover, the chapter discusses recent advances in 3D ALs, and applicability of 3D ALs from qualitative to quantitative SAR analysis in detail.

- *Chapter 6* investigates the applicability of CNN modeling on compound images for AC prediction analysis. Furthermore, the chapter offers approaches to interpret the model predictions from a chemical perspective.

- *Chapter 7* establishes a proof of principle and investigates deep CNN modeling potential to learn functional group chemistry from compound images and transfer the acquired functional group knowledge to classify ACs successfully.

- The final chapter (*Chapter 8*) summarizes and discusses the main findings of this work.

Chapter 2

Activity Landscape Image Analysis Using Convolutional Neural Networks

Introduction

3D ALs are AL/PL models integrating compound similarity and activity data into a 3D graphical representation. They are reminiscent of geographical maps where 3D ALs preserve diverse SAR characteristics by encoding them into landscape topology. SAR attributes embedded in a 3D AL model can be effortlessly captured and exported into various image formats, which are acquiescent to image analysis and image processing approaches. This chapter aims to analyze 3D ALs through image processing approaches. The analysis investigates modern state-of-the-art machine learning models (i.e., CNNs, random forests, and SVM) to determine whether machine learning models distinguish different SAR characteristics from each other through image processing. To analyze different SAR characteristics, multiple viewpoints from 3D AL models generated from 38 different target-based compound data sets were captured and exported as high-quality PNG images. In addition to original 3D AL models that naturally tend to combine smooth and rugged surfaces, reference 3D AL models with predominantly rugged and predominantly smooth surfaces were also created. CNN and machine learning models were trained on different color-codings (i.e., RGB, monochrome, and binary) of generated 3D AL images to distinguish the original heterogeneous 3D AL models from reference 3D AL models. Moreover, the CNNs and machine learning models were trained on extracted edge features of 3D ALs image variants to compare the contribution of the topo-

graphical features in classification prediction. The obtained results of the analysis are discussed.

RESEARCH ARTICLE

Activity landscape image analysis using convolutional neural networks

Javed Iqbal, Martin Vogt and Jürgen Bajorath[*]

Abstract

Activity landscapes (ALs) are graphical representations that combine compound similarity and activity data. ALs are constructed for visualizing local and global structure–activity relationships (SARs) contained in compound data sets. Three-dimensional (3D) ALs are reminiscent of geographical maps where differences in landscape topology mirror different SAR characteristics. 3D AL models can be stored as differently formatted images and are thus amenable to image analysis approaches, which have thus far not been considered in the context of graphical SAR analysis. In this proof-of-concept study, 3D ALs were constructed for a variety of compound activity classes and 3D AL image variants of varying topology and information content were generated and classified. To these ends, convolutional neural networks (CNNs) were initially applied to images of original 3D AL models with color-coding reflecting compound potency information that were taken from different viewpoints. Images of 3D AL models were transformed into variants from which one-dimensional features were extracted. Other machine learning approaches including support vector machine (SVM) and random forest (RF) algorithms were applied to derive models on the basis of such features. In addition, SVM and RF models were trained using other features obtained from images through edge filtering. Machine learning was able to accurately distinguish between 3D AL image variants with different topology and information content. Overall, CNNs which directly learned feature representations from 3D AL images achieved highest classification accuracy. Predictive performance for CNN, SVM, and RF models was highest for image variants emphasizing topological elevation. In addition, SVM models trained on rudimentary images from edge filtering classified such images with high accuracy, which further supported the critical role of altitude-dependent topological features for image analysis and predictions. Taken together, the findings of our proof-of-concept investigation indicate that image analysis has considerable potential for graphical SAR exploration to systematically infer different SAR characteristics from topological features of 3D ALs.

Keywords: Activity landscape, Structure–activity relationships, Image processing, Image classification, Machine learning, Convolutional neural network, Landscape topology, Feature extraction

Introduction

Activity landscapes (ALs) are defined as graphical representations that integrate compound similarity and activity relationships [1, 2]. ALs graphically represent active compounds in biologically relevant chemical space, making it possible to visualize structure–activity

relationships (SARs) and identify key compounds and SAR determinants [1–8]. A variety of AL representations of different design and complexity have been introduced to visualize SARs. These include structure–activity similarity maps, other two-dimensional (2D) ALs, three-dimensional (3D) AL models, and molecular network representations [1–8]. 3D ALs can be rationalized to result from a two-dimensional (2D) projection of chemical feature space, producing a plane where compounds are separated by varying distances, to which compound potency is added as a third

*Correspondence: bajorath@bit.uni-bonn.de
Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, 53115 Bonn, Germany

dimension. From sparsely distributed potency measurements, an activity hyper-surface is interpolated [3, 8]. Compounds that are similar to each other and neighbors in chemical space but have large differences in potency form activity cliffs (ACs) [1–3], which are prominent features of 3D AL models. Such 3D ALs are reminiscent of geographical maps with varying landscape topologies [3, 8]. In 3D ALs, the activity hypersurface can be color-coded by compound potency using a color gradient, which further emphasizes different topologies. In gently sloped or smooth regions, gradual changes in chemical structure are accompanied by moderated changes in potency, which corresponds to SAR continuity [1–3]. By contrast, in rugged regions, small chemical changes lead to significant potency variations, corresponding to SAR discontinuity [1–3]. Here, ACs represent the apex of SAR discontinuity. By design, 3D ALs are descriptive in nature and are typically qualitatively analyzed. Only very few studies have thus far attempted to use AL models for compound potency predictions [4, 8].

3D AL models can be visualized and analyzed from different viewpoints and perspectives. Hence, visualization yields images with different characteristics that can be subjected to image processing methods. Thus far, however, AL visualizations have not been analyzed and compared using such approaches. Therefore, we have asked the question if 3D ALs with different topological features representing different SAR characteristics could be distinguished from each other and classified through image processing; the major topic of this study.

In recent years, deep learning has made a large impact on image processing. In particular, convolutional neural networks (CNNs) have become one of the preferred machine learning approaches for image analysis due to their ability to extract patterns from low-resolution data representations in so-called convolutional layers [9, 10]. CNNs are deep neural networks with one or more convolutional layers encoding locality information in the network structure [9, 10]. The design of CNNs renders them well-suited for processing of spatial and temporal data such as images, audio, or video signals. CNNs achieved higher performance level than other computational approaches in recognizing histopathological [11], magnetic resonance [12], medical X-ray [13], computer tomography, [14] and fundus images [15]. CNNs are also gaining increasing attention in chemical informatics and drug discovery, albeit in different contexts. For example, CNNs and random forest (RF) [16] models have been applied to predict cancer cell line sensitivity and compound potency [17] or compound toxicity [18] using 2D structural

representations in image formats. CNNs have also been employed to model bioactivity profiles using 2D images [19, 20]. Other studies investigated molecular mechanism of action based on bioactivity profiles using images from high-content screening [21–23].

In addition to CNNs, various studies have shown that other machine learning approaches like support vector machine (SVM) [24] modeling can also classify images using raw pixel intensities or extracted image features [25–30]. In addition, RF can also accurately classify high-dimensional image data [31, 32].

However, the application of CNNs or other machine learning-based image processing methods to ALs for SAR visualization and analysis has thus far not been investigated. For machine learning methods, suitable representations are required to represent data sets of varying size in a unified and standardized format to enable direct comparison. Images generated from 3D ALs are well suited because they retain the pairwise similarity relationships between compounds and account for potency values as topographical features and/or using color gradients. Images can be generated from different viewpoints ranging from top-down views of ALs to elevated or profile views where SARs become visible as peaks and valleys. Top-down views essentially yield heatmap representations if color gradients are used, as further discussed below.

In our current study, 3D AL images have been analyzed and classified using CNNs, RF, and SVM. Different projection methods and image encodings of varying resolution and information content have been generated to capture 3D AL topology in different ways and determine which factors are responsible for accurate image classification. Therefore, image variants with successively reduced information content have also been generated and investigated.

CNN, RF, and SVM models were found to be capable of correctly classifying AL image variants with different topology on the basis of structure and pixel intensity information. CNNs learning feature representations yielded overall most accurate predictions. However, RF and SVM models trained on pre-defined lower-level feature representations were also predictive. The analysis identified topological features that were of critical relevance for image classification. Taken together, our findings revealed that images of 3D ALs for SAR visualization can be distinguished through machine learning on the basis of characteristic topological features, which provides a new methodological framework for direct comparison of AL models of compound data sets of different composition and comparative SAR analysis of large data sets.

Concepts and methods

Compound activity classes

For 3D AL image analysis, 38 compound activity classes were selected from ChEMBL version 23 [33]. For each class, more than 500 compounds with pK_i potency measurements were available. Intra-class potency variations spanned several orders of magnitude. In addition, the potency value distribution of each class had an interquartile range covering at least one order of magnitude [34]. Table 1 summarizes the composition of each activity class and provides potency range statistics. Reported are final compound numbers after similarity filtering, as further described below.

Table 1 Activity classes

ChEMBL target ID	Target name	No. compounds	Potency [pK_i]		IQR (Q1–Q3)
			Min	Max	
CHEMBL204	Coagulation factor II	1099	1.00	12.19	5.30–7.47
CHEMBL205	Carbonic anhydrase 2	2701	0.60	11.10	6.24–8.04
CHEMBL214	Serotonin receptor 1A	1936	0.36	10.85	6.74–8.29
CHEMBL217	Dopamine D2 receptor	3427	2.85	10.57	6.29–7.49
CHEMBL218	Cannabinoid receptor 1	1938	3.79	10.10	5.95–7.62
CHEMBL219	Dopamine D4 receptor	1086	4.74	10.52	6.43–7.87
CHEMBL222	Sodium-dependent noradrenaline transporter	1000	2.26	9.52	5.86–7.54
CHEMBL224	Serotonin receptor 2A	1967	3.51	11.00	6.56–8.12
CHEMBL225	Serotonin receptor 2C	1085	3.51	9.70	6.23–7.70
CHEMBL226	Adenosine receptor A1	2829	4.12	12.23	5.87–7.14
CHEMBL229	Alpha-1A adrenergic receptor	594	4.04	10.44	6.90–8.40
CHEMBL233	Mu-type opioid receptor	2009	4.20	11.80	6.37–8.33
CHEMBL234	Dopamine D3 receptor	2518	4.17	10.00	6.79–8.40
CHEMBL236	Delta-type opioid receptor	1604	3.72	10.68	6.00–8.08
CHEMBL237	Kappa-type opioid receptor	1853	4.09	11.52	6.45–8.49
CHEMBL238	Sodium-dependent dopamine transporter	850	2.14	9.40	5.60–7.37
CHEMBL240	Potassium voltage-gated channel subfamily H_2	1053	3.89	9.55	5.29–6.44
CHEMBL245	Muscarinic acetylcholine receptor M3	609	4.11	10.30	6.70–9.10
CHEMBL251	Adenosine receptor A2a	3305	3.92	11.38	6.05–7.67
CHEMBL253	Cannabinoid receptor 2	2605	0.63	10.72	6.24–7.99
CHEMBL255	Adenosine receptor A2b	1265	3.37	9.80	6.30–7.82
CHEMBL256	Adenosine receptor A3	2567	1.32	11.00	6.16–7.84
CHEMBL261	Carbonic anhydrase 1	2657	0.56	11.00	5.34–7.09
CHEMBL264	Histamine H3 receptor	2323	4.07	10.60	7.21–8.70
CHEMBL344	Melanin-concentrating hormone receptor 1	1187	3.57	9.77	6.90–8.01
CHEMBL1800	Corticotropin-releasing factor receptor 1	673	4.26	9.66	6.58–8.14
CHEMBL1833	5-hydroxytryptamine receptor 2B	695	5.00	9.96	6.13–7.40
CHEMBL2014	Nociceptin receptor	839	4.40	10.70	7.09–8.52
CHEMBL3155	5-hydroxytryptamine receptor 7	1111	3.30	10.00	6.53–7.95
CHEMBL3242	Carbonic anhydrase 12	2008	3.08	9.62	6.92–8.23
CHEMBL3371	5-hydroxytryptamine receptor 6	2134	1.38	10.40	7.03–8.52
CHEMBL3594	Carbonic anhydrase 9	2347	1.34	9.92	6.61–8.04
CHEMBL3759	Histamine H4 receptor	887	2.85	10.40	5.98–7.59
CHEMBL4005	Serine/threonine protein kinase PIK3CA	882	4.65	10.52	7.01–8.46
CHEMBL4550	Arachidonate 5-lipoxygenase-activating protein	1318	5.60	9.40	6.75–8.21
CHEMBL4792	Orexin receptor type 2	1444	4.96	10.15	6.13–7.57
CHEMBL5071	Prostaglandin D2 receptor 2	794	4.48	10.00	6.49–8.41
CHEMBL5113	Orexin receptor type 1	1249	4.19	9.80	5.47–7.19

The table summarizes the composition of 38 activity classes used for 3D AL modeling. IQR represents the interquartile range of the potency value distribution of each data set

Molecular representation and similarity assessment

For similarity assessment, the extended-connectivity fingerprint with bond diameter 4 (ECFP4) [35] was calculated for each compound. ECFP4 is a topological feature set fingerprint comprising layered atom environments and represents a gold standard in the field. ECFP4 feature sets were folded into a fixed-length 1024-bit representation [35]. As a similarity metric, the Tanimoto coefficient (Tc) was used to quantify pairwise compound similarity [36]. The Tc is defined as:

$$\mathrm{Tc}(A, B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

where A, B are fingerprints of compounds A and B, respectively. Corresponding Tanimoto distance was obtained by calculating the complement $1 - \mathrm{Tc}(A, B)$.

Initially assembled activity classes were subjected to similarity filtering and only compounds were retained that possessed an ECFP4 Tc similarity of at least 0.4 to at least one other compound from the same activity class. Filtering was applied to eliminate singletons from the data sets that had no or only very weak structural relationships with other compounds (and hence did not contribute to SARs). Fingerprint and similarity calculations were performed using in-house Python scripts and the OpenEye chemistry toolkit [37].

3D activity landscapes

For generating 3D AL models, ECFP4 space was projected on a 2D plane, compound potency values were added as the third dimension and from these values, a coherent potency hyper-surface was interpolated. Different projection methods for 3D AL design have previously been investigated [8] and two methods, multi-dimensional scaling (MDS) [38] and Neuroscale [39], were found to be preferred for retaining original similarity relationships for SAR visualization. Therefore, these approaches were used herein. For projection, both MDS and Neuroscale apply stress functions based on pairwise Tanimoto distances between compounds. Neuroscale projects compounds using a radial basis function (RBF) neural network. For each Neuroscale model, the number of RBFs was optimized using sevenfold cross validation.

Hyper-surface interpolation was carried out using Gaussian process regression (GPR) [40, 41]. The resulting surface was colored according to the compound potency using a color gradient from green over yellow to red. For all images, the same color gradient was applied according to which a pK$_i$ value of 5.75 (and below) corresponded to green, the pK$_i$ range 5.76–8.74 pK$_i$ to yellow, and a pK$_i$ of 8.75 (or above) to red.
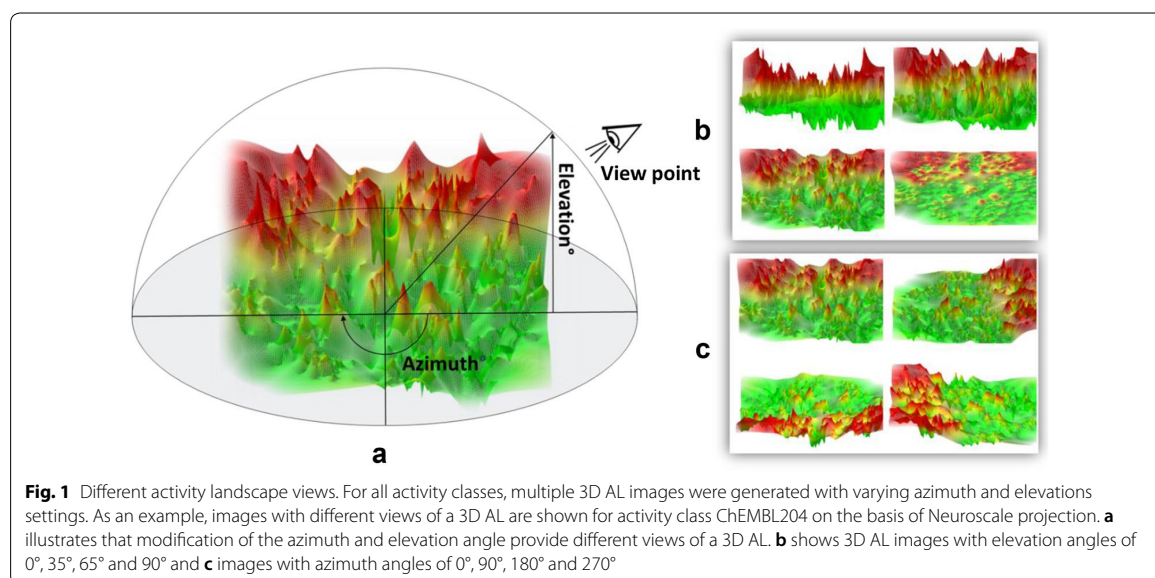
Reference landscapes

Smooth and rugged regions represent major topological features of 3D ALs that correspond to different SAR phenotypes [3]. In smooth regions, gradual changes in molecular structure are accompanied by moderate changes in potency, which represents SAR continuity. By contrast, in rugged regions, small structural changes lead to large potency variations. This corresponds to SAR discontinuity and leads to the formation of ACs. In many activity classes, continuous and discontinuous SAR components co-exist and are combined in different ways, giving rise to globally heterogeneous SARs [42, 43]. Such SAR heterogeneity is quantitatively accounted for using numerical SAR analysis functions such as the SAR Index [42]. In 3D AL models, SAR heterogeneity is represented by co-occurrence of smooth and rugged regions in different topological constellations.

To establish proof-of-concept for image classification, two reference AL models were generated for the 3D AL of each activity class in which SAR continuity/smoothness and discontinuity/ruggedness were increased, respectively, relative to the original 3D AL. Accordingly, these 3D AL variants were termed smooth and rugged reference (Ref-)ALs, respectively.

Smooth Ref-ALs were generated by selecting compounds that fell into the 2nd and 3rd quartile, i.e. the interquartile range, of the potency distribution of each activity class. Rugged Ref-ALs were obtained by considering septiles of the potency distribution and selecting compounds falling into the 1st, 3rd, 5th, and 7th septile. The resulting Ref-ALs contained about half and 4/7th the original number of compounds per class, respectively, which consistently amounted to more than 250 compounds per Ref-AL. Rugged Ref-ALs retained the potency range of the original ALs, whereas the potency of smooth Ref-ALs was reduced to the interquartile range, as reported in Table 1. It varied from ten- to 100-fold differences for most data sets while five sets had a larger than 100-fold interquartile range. As further discussed below, original 3D ALs of all 38 activity classes were generally heterogeneous in nature and were designated accordingly. Hence, for the generation of classification models, smooth and rugged Ref-ALs were distinguished from heterogeneous 3D ALs of original compound data sets, hence yielding three categories of 3D AL models for image generation.

Activity landscape images

For each original 3D AL and Ref-AL, images providing different views were generated by systematically varying azimuth (0°, 90°, 180° 270°) and elevation angles (0°, 35°, 65°,90°), as illustrated in Fig. 1. For the elevation

Iqbal *et al. J Cheminform*   (2020) 12:34

Page 5 of 15

Fig. 1 Different activity landscape views. For all activity classes, multiple 3D AL images were generated with varying azimuth and elevations settings. As an example, images with different views of a 3D AL are shown for activity class ChEMBL204 on the basis of Neuroscale projection. **a** illustrates that modification of the azimuth and elevation angle provide different views of a 3D AL. **b** shows 3D AL images with elevation angles of 0°, 35°, 65° and 90° and **c** images with azimuth angles of 0°, 90°, 180° and 270°

angle of 0°, most of the 2D projection information is lost but altitude is accounted for as a topological feature. By contrast, for the elevation angle of 90°, elevation information is only retained through potency coloring. Furthermore, original color images were converted into image variants with reduced information content including grayscale and black and white (b/w) versions as well as images generated from edge detection filters (see below). Exemplary images are shown in Fig. 2.

Convolutional neural networks

CNNs are deep neural networks characterized by one or more initial convolutional layers. CNNs are popular for image-based analysis tasks [10]. Convolutional layers only connect local neighborhoods of input neurons and perform learnable convolutions on the input data that are identical for each neuron. The output of the convolution layer is passed through a standard rectified linear unit activation (ReLU) layer. This is followed by pooling that combines outputs from local neuron clusters and reduces the dimensions and computational complexity [44]. Multiple convolutional layers can be connected to each other leading to successive reduction of layer sizes. The output of the final convolutional layer is followed by one or more fully connected neuron layers. Dropout layers that randomly deactivate a proportion of neurons are inserted between layers in order to avoid overfitting [45]. A schematic of a CNN is shown in Fig. 3.
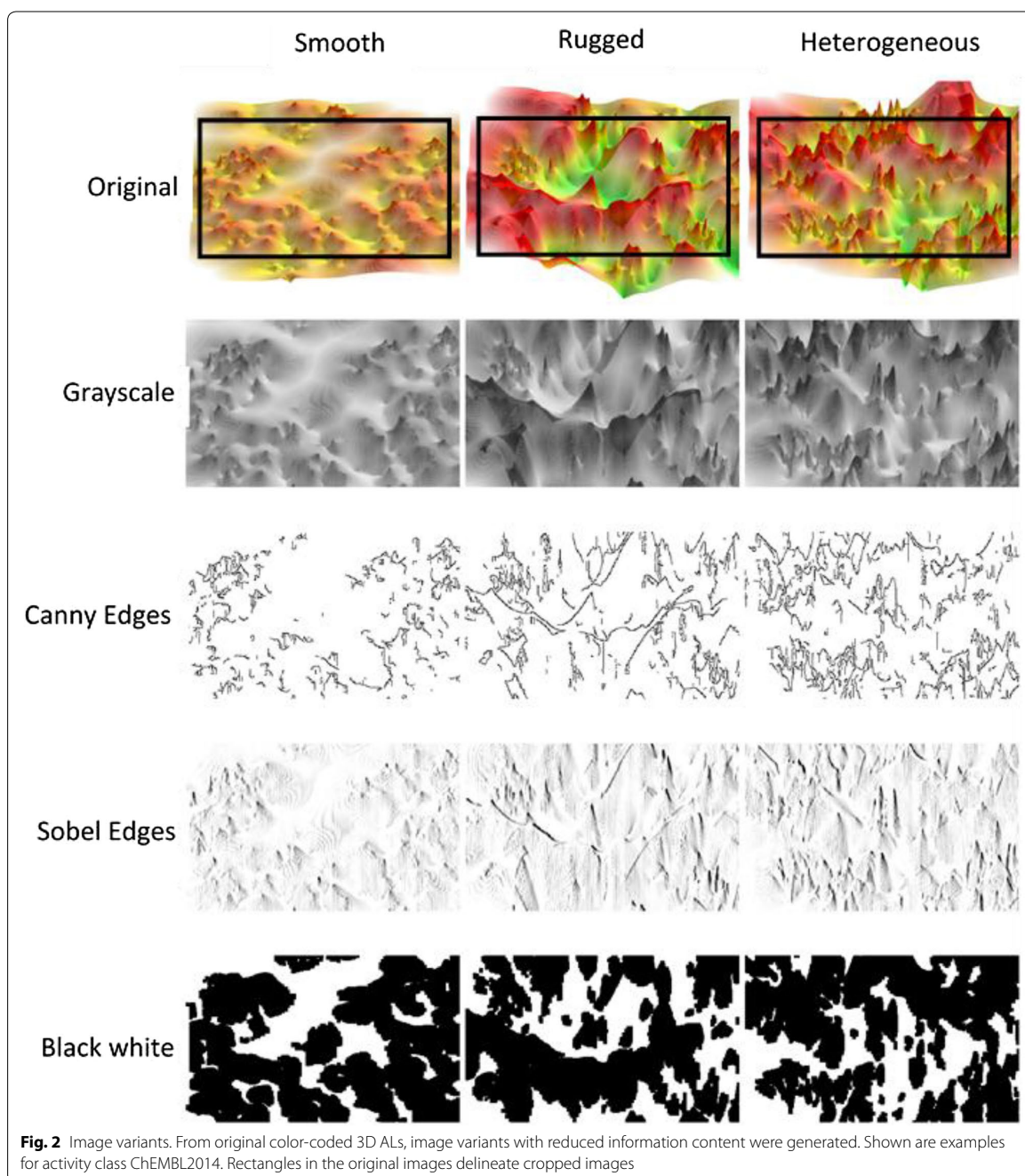
Network architecture

The CNN architecture used herein consisted of convolutional, rectified linear unit (ReLU), max-pooling, dropout, and dense layers, as illustrated in Fig. 3. Three convolutional layers with filter size of 3×3 with respect to kernel sizes of 32, 64 and 128 were added to extract image features. Each convolution layer was followed by a rectified linear unit (ReLU), a max-pooling, and a dropout layer. After "flattening" the weights, two intermediate dense layers were added followed by dropout layers. As output, a softmax layer was used to normalize learned weights as a probability distribution. CNN layers were implemented using TensorFlow (version 1.4.1) and Keras (version 2.2.4) [46, 47]. Training data were assembled from 19 randomly selected activity classes. As test sets, all images from the remaining 19 classes were used. CNN hyper-parameters were optimized using internal validation on the basis of an 80% versus 20% split of the training data. Parameter optimization included ReLU alpha over the range 0.0–0.5, dropout rates with values 0.0, 0.1, 0.3, intermediate dense layer sizes of 16, 32, 64, and 128 output neurons, and Adam optimizer learning rates of 100, 10, 1, 0.1, 0.01, 0.001, 0.005, 0.00005, and 0.000005. Each CNN model was trained until convergence was reached, which typically required ~ 20 epochs.

Alternative machine learning approaches
Support Vector Machine
Support vector machine (SVM) is a supervised machine learning algorithm that constructs a hyper-plane H in a

Iqbal *et al. J Cheminform*   (2020) 12:34

Page 6 of 15

Fig. 2 Image variants. From original color-coded 3D ALs, image variants with reduced information content were generated. Shown are examples for activity class ChEMBL2014. Rectangles in the original images delineate cropped images

given feature space to best separate different classes of objects by maximizing the distance (margin of the hyperplane) between objects having different class labels [24]. The hyper-plane for an *n*-dimensional feature space is defined as:

$$H = \{x \in \mathbb{R}^n | w, x + b = 0\}$$

Here, $w \in \mathbb{R}^n$ w is the weight vector and $b \in \mathbb{R}^n$ is the bias. If linear separation of objects with different class labels is not possible in a given feature space, the data

Fig. 3 Convolutional neural network architecture. CNN design combining convolution, rectified linear unit, max-pooling, dropout, and dense layers is schematically illustrated

Table 2 Image collections

No.	Projection	Elevation	Azimuth	Number of images
1	MDS, Neuroscale	0°, 35°, 65°, 90°	0°, 90°, 180°, 270°	3648
2	MDS	0°, 35°, 65°, 90°	0°, 90°, 180°, 270°	1824
3	Neuroscale	0°, 35°, 65°, 90°	0°, 90°, 180°, 270°	1824
4	MDS, Neuroscale	90°	0°, 90°, 180°, 270°	912
5	MDS, Neuroscale	65°	0°, 90°, 180°, 270°	912
6	MDS, Neuroscale	35°	0°, 90°, 180°, 270°	912
7	MDS, Neuroscale	0°	0°, 90°, 180°, 270°	912

Different image collections were generated to provide alternative conditions for training and testing of classification models

are projected into a higher dimensional space variant where linear separation might become feasible. Therefore, the scalar product w, x is replaced by a non-linear kernel function, thereby circumventing explicit mapping to higher dimensional space. SVM classifiers are trained using a regularization parameter that permits certain misclassification events and penalizes them with a cost factor C, which supports model generalization. For multi-class image analysis, multiple one-against-one binary SVM models were trained and the results were combined to yield a final classifier. SVM meta-parameters were optimized using tenfold cross validation including cost factor C with values of 0.01, 0.1, 1 and the kernel (linear, polynomial, or RBF). For SVM training, a total of 79,200 features extracted from images were used.

Random forest
RF is a decision tree ensemble classifier that is trained using randomized feature subsets on sub-sampled training data [16]. Herein, RF models were constructed from the subset of 79,200 image features. RF meta-parameters including the number of trees (50 or 100), minimum number of samples (2 or 5), and minimum sample leaf nodes (1 or 3) were optimized using tenfold cross validation.

Image pre-processing and feature extraction
Original 3D AL images were generated with a resolution of 1200×800 pixels. Images were cropped to reduce non-colored areas and outer boundary regions. Cropped images were resized to a resolution of 360×220. Grayscale images were obtained as the weighted sum of the red, green and blue channels using weights of 0.299, 0.587, and 0.114, respectively. These calculations were performed using the openCV library version 3 [48–51]. In addition, grayscale images were converted into b/w images by applying binary Otsu's thresholding [52]. The

pixel values of all image matrices were converted into 32-bit floating point format and normalized.

Convolution layers of neural networks can detect feature representations from given image pixel values. However, machine learning approaches such as SVM and RF are not capable of doing so. Therefore, image filters for feature extraction were applied to generate feature sets for SVM and RF calculations.

The Sobel edge operator is a convolution filter for edge detection given by the two convolution matrices:

$$G_x = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}, \quad G_y = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix}$$

It introduces an average factor for smoothing random noise of an image and extracts enhanced (thick and bright) edges [53]. Herein, the vertical improved Sobel filter G_y of Gao et al. [53] was used. In addition, the Canny edge detector was applied, representing a widely used method for edge detection [54]. The openCV implementation of the Canny edge filter was applied to obtain Canny edges [49]. The resulting row-wise flattened pixel values of edge filters were used as a feature vector. Figure 2 illustrates image variants obtained using the Sobel edge and Canny edge filters. Furthermore, two other filters were used including ORB [55] and Harris boundary features [56] that are less frequently considered for topological features.

Deriving and evaluating models on image collections
Machine learning models were trained and tested on images viewed from different angles and image variants with different information content generated on the basis of MDS or Neuroscale projections. Images were grouped into different collections, as reported in Table 2. Collections 1–3 included all viewpoints and were distinguished only by the projection method. Collection 1 combined MDS and Neuroscale images while collection 2 and 3

Iqbal *et al. J Cheminform* (2020) 12:34

Page 8 of 15

only included MDS and Neuroscale images, respectively. Collections 4–7 focused on different elevation viewpoints combining MDS and Neuroscale projections. As training data, the heterogeneous (original 3D AL), smooth (Ref-AL), and rugged (Ref-AL) variants for all 38 activity classes were used yielding 114 images for each for specific viewpoint and projection. Training was performed on cropped full color, grayscale, and b/w images. Additionally, image variants from Sobel edge or Canny edge filters were used in some settings. For each elevation, four images were generated for azimuth angle of 0°, 90°, 180° and 270°. Depending on the collection, eight to 32 image variants per target were used for model derivation. Training data for all models were extracted features with normalized values obtained from pre-processed images. For training SVM and RF models, pre-processed images were represented as one-dimensional feature vectors without locality information, which was retained in CNNs via the convolutional layers. Training data were assembled from 19 of 38 randomly selected activity classes. As test sets, all images from the remaining 19 activity classes were used.

Performance evaluation

Classification performance was evaluated based on receiver-operator characteristic (ROC) curves, the area under the ROC curve (AUC), and the confusion matrix. Three standard performance measures were applied including the subset accuracy [57], Matthew's correlation coefficient (MCC) [58], and the weighted mean F1 score [59]. Subset accuracy is defined as:

$$\text{Accuracy} = \frac{1}{n} \sum_{i}^{n} [\![Z_i = Y_i]\!]$$

where n denotes the number of samples in the test set, Z_i is the predicted and Y_i is the true label for sample i and $[\![\cdot]\!]$ is the Iverson bracket taking the value of 1 for a true and 0 for a false predicate [57].

Results and discussion

Analysis concept

Our study was designed to investigate image analysis for distinguishing between 3D AL models with different topological features reflecting different SAR characteristics. Graphical SAR analysis has thus far mostly been qualitative and subjective in nature. Therefore, we reasoned that successful classification of 3D AL images according to different topological features via ML would provide a sound foundation for systematically comparing 3D ALs going beyond subjective interpretation of AL models and qualitative analysis of SAR characteristics. We emphasize that AL images do not only provide

an attractive representation for SAR visualization, but also a preferred data format for ML-based image classification. AL images are preferred because the underlying AL data matrices are difficult, if not infeasible to use for ML directly. This is the case because the AL data structure consists of an exhaustive pairwise compound similarity matrix and an array of compound potency values that must be combined. For ML, a potency-augmented similarity data matrix would need to be transformed into a fixed-format feature vector or an equivalent representation to enable direct comparison of different AL data matrices for model derivation. This is intrinsically difficult to accomplish for compound data sets of different composition and size for which ALs are usually generated. Challenging data transformations can be circumvented by using standardized images of ALs directly for ML, which also motivated ML image classification from a methodological perspective, in addition to its attractiveness for graphical SAR exploration. Standardizing images inevitably involves investigating different orientations and image views.

In order to assess how different AL features influence the classification performance of ML methods, we did not only study model performance based on different image viewpoints, but also applied two defined image processing strategies. First, for each AL, we generated reference models with increased SAR continuity/smoothness and discontinuity/ruggedness, respectively. This made it possible to determine which topological characteristics were primarily responsible for accurate image classification. Second, for each AL image, variants with successively reduced information content were generated including grayscale, b/w, and edge-filtered image variants, which were also used for training and model building. This made it possible to determine how different image encodings of topological features affect classification performance, in which form distinguishing features were detected by ML models, and which level of image information content was minimally required for classification of images capturing different AL topologies. Using images as direct encodings of ALs for classification and investigating the two image pre-processing strategies via ML represented key components of our proof-of-concept study.

Activity landscape topology

The top right image in Fig. 2 shows a representative 3D AL. For all 38 activity classes, heterogeneous ALs were obtained that combined smooth and rugged sections in different ways (further examples are provided below). Therefore, to provide topological reference states for assessing the suitability of 3D AL classification, the concept of smooth and rugged Ref-ALs was introduced. For

each original 3D AL, Ref-ALs were generated to increase either smooth or rugged AL character through consistently applied potency-based data set modification, as detailed above. For an exemplary 3D AL, the smooth and rugged Ref-AL is shown in Fig. 2 (top). The generation of

Table 3 Classification of color-coded images using convolutional neural networks

Collection	CNN	Metric
1	0.74 ± 0.01	Accuracy
	0.74 ± 0.01	F1
	0.61 ± 0.01	MCC
2	0.72 ± 0.02	Accuracy
	0.72 ± 0.02	F1
	0.58 ± 0.03	MCC
3	0.71 ± 0.02	Accuracy
	0.71 ± 0.03	F1
	0.56 ± 0.04	MCC
4	0.73 ± 0.04	Accuracy
	0.73 ± 0.04	F1
	0.60 ± 0.06	MCC
5	0.70 ± 0.04	Accuracy
	0.70 ± 0.04	F1
	0.55 ± 0.06	MCC
6	0.72 ± 0.03	Accuracy
	0.72 ± 0.03	F1
	0.58 ± 0.04	MCC
7	0.75 ± 0.03	Accuracy
	0.75 ± 0.03	F1
	0.62 ± 0.04	MCC

The table reports classification results for CNN models trained and tested on color-coded images. All values reported are averages ± standard deviations over 10 independent trials

these 3D AL variants made it possible to formulate well-defined classification tasks to distinguish heterogeneous 3D ALs from smooth and rugged AL reference states and explore features driving machine learning. Feature relevance was further assessed using other AL variants with reduced information content, as also illustrated in Fig. 2.

Classification of color-coded activity landscape images

First, 3D AL images of 38 activity classes with different combinations of projection and elevation angles and color gradients accounting for compound potency information were investigated. CNN classification models were built for all image collections according to Table 2. SVM and RF modeling were not applicable for this prediction task due to difficulties in algorithmically handling 3D color features. By contrast, CNN models preserved the dimensionality of color gradients. CNN classification performance is summarized in Table 3. CNNs reached a mean accuracy of 0.74 ± 0.1 (mean ± standard deviation) for combined projections and elevations. In addition, MCC values of ~ 0.6 or greater were obtained indicating globally accurate predictions.

When classification performance was separately considered for the different image classes, smooth Ref-ALs, rugged Ref-ALs, and heterogeneous 3D ALs from collection 1 achieved ROC AUC values of 1.00, 0.86, and 0.86, respectively, as shown in Fig. 4. In addition, the confusion matrix for all images revealed that CNNs were able to classify images of smooth, rugged and heterogeneous 3D AL variants with a true positive rate of 96%, 60% and 73%, respectively (Fig. 4), reflecting overall accurate predictions.

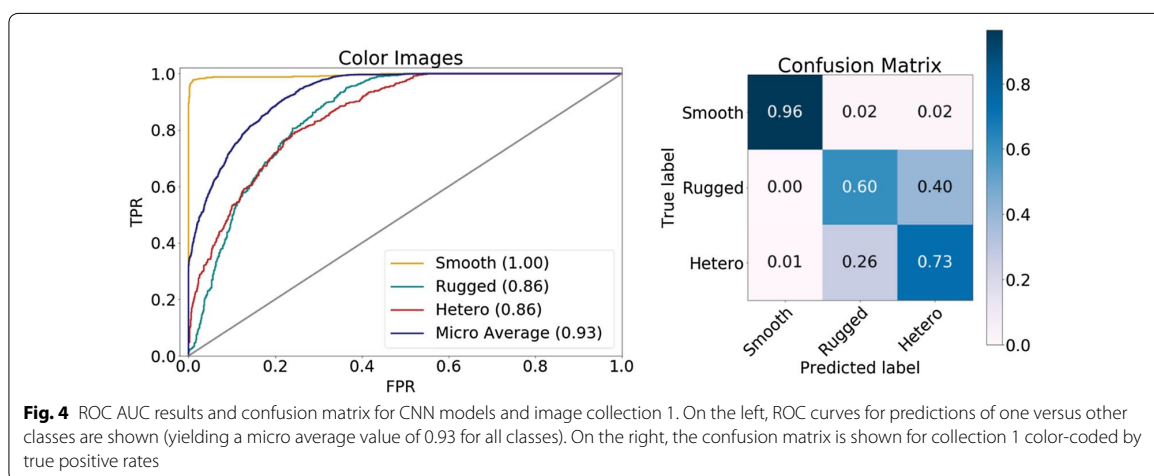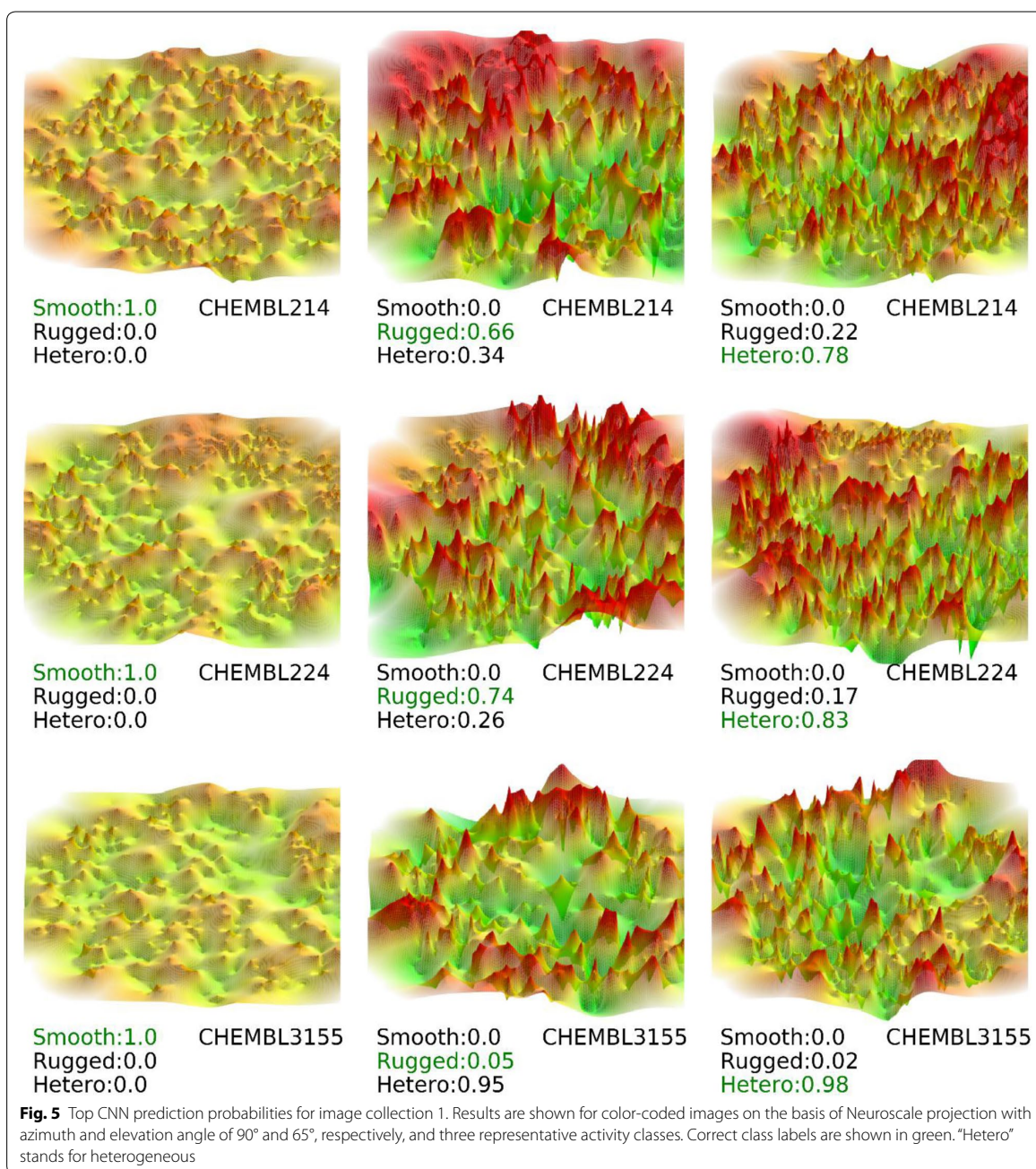Probabilities for class predictions using the best performing CNN model for collection 1 with images taken

Fig. 4 ROC AUC results and confusion matrix for CNN models and image collection 1. On the left, ROC curves for predictions of one versus other classes are shown (yielding a micro average value of 0.93 for all classes). On the right, the confusion matrix is shown for collection 1 color-coded by true positive rates

Fig. 5 Top CNN prediction probabilities for image collection 1. Results are shown for color-coded images on the basis of Neuroscale projection with azimuth and elevation angle of 90° and 65°, respectively, and three representative activity classes. Correct class labels are shown in green. "Hetero" stands for heterogeneous

from the 90° azimuth and 65° elevation angles are shown in Fig. 5.

Predicted class probabilities displayed the general trend that smooth Ref-ALs were consistently predicted with high accuracy, whereas distinguishing between heterogeneous 3D ALs and rugged Ref-ALs often represented a more challenging prediction tasks, resulting in at least slightly reduced accuracy. These observations indicated that the absence of cliffs and associated features in smooth Ref-ALs was a major determinant for correctly

Iqbal *et al. J Cheminform*   (2020) 12:34

Page 11 of 15

distinguishing them from rugged Ref-ALs and heterogeneous 3D ALs.

When comparing different projection methods (collections 1–3), only small differences in performance were observed with only a slight decrease average accuracy of 0.01–0.02 for the individual projections (collection 2 and 3) compared to the collection with combined projections (collection 1). Hence, MDS and Neuroscale projections were readily comparable for classification. CNN models trained separately on different elevations (collection 4–7) performed consistently well. Interestingly, the performance was overall best using 0° elevation angle images (collection 7), with an average accuracy of 0.75 ± 0.03. These projections only visualized altitude profiles of the 3D ALs. These findings also indicated that features associated with cliffs and their absence in smooth Ref-ALs had a major influence on the classifications. By contrast, varying image viewpoints originating from different azimuth and elevation angle combinations were not significantly affecting prediction accuracy, which alleviated the need to establish constant reference frames for 3D AL comparisons.

Collection 4 consisted of top-down views of ALs where potency differences were only distinguished by the color gradient. These views corresponded to 2D heatmap representations of ALs given in four different rotations. In this case, the accuracy of the CNN model was 0.73 ± 0.04 and thus only slightly reduced compared to the profile views of collection 7. This observation was of interest since heatmap views contained the complete information of the AL captured by the color gradient while profile views provided color information and topology. However, in contrast to lower elevation views where some topographical details might be hidden, in heatmaps, no AL features were concealed. Hence, 2D heatmaps and 3D profile views were suitable AL representations for classification of color-coded ALs. This was an encouraging finding in 3D image analysis.

Models trained on grayscale and black/white image variants

Different from color-coded 3D ALs, it was possible to train SVM and RF models on grayscale and b/w image variants, in addition to CNNs. Classification results for models trained on grayscale image variants are reported in Table 4.

As expected, for CNNs, the loss in color information slightly reduced global classification performance. However, for the combined collection 1, the reduction in accuracy from 0.74 ± 0.01 to 0.71 ± 0.02 was less than one might anticipate. Reduction in performance was largest for high elevation viewpoints (collection 4 and 5) that retained the least altitude information in their

Table 4 Classification of models trained on grayscale images

Collection	RF	SVM	CNN	Metric
1	0.57 ± 0.01	0.53 ± 0.01	0.71 ± 0.02	Accuracy
	0.57 ± 0.01	0.54 ± 0.01	0.71 ± 0.02	F1
	0.35 ± 0.01	0.30 ± 0.01	0.56 ± 0.03	MCC
2	0.54 ± 0.01	0.53 ± 0.01	0.70 ± 0.03	Accuracy
	0.55 ± 0.01	0.54 ± 0.01	0.70 ± 0.03	F1
	0.32 ± 0.02	0.29 ± 0.02	0.55 ± 0.04	MCC
3	0.55 ± 0.02	0.53 ± 0.01	0.70 ± 0.03	Accuracy
	0.56 ± 0.01	0.54 ± 0.02	0.70 ± 0.03	F1
	0.33 ± 0.02	0.30 ± 0.02	0.55 ± 0.04	MCC
4	0.54 ± 0.02	0.57 ± 0.03	0.67 ± 0.03	Accuracy
	0.54 ± 0.02	0.58 ± 0.04	0.67 ± 0.03	F1
	0.31 ± 0.03	0.36 ± 0.05	0.51 ± 0.05	MCC
5	0.55 ± 0.03	0.50 ± 0.01	0.68 ± 0.02	Accuracy
	0.56 ± 0.03	0.51 ± 0.02	0.68 ± 0.02	F1
	0.33 ± 0.04	0.25 ± 0.02	0.52 ± 0.03	MCC
6	0.58 ± 0.01	0.53 ± 0.02	0.72 ± 0.03	Accuracy
	0.58 ± 0.02	0.55 ± 0.02	0.72 ± 0.03	F1
	0.37 ± 0.02	0.30 ± 0.03	0.59 ± 0.04	MCC
7	0.69 ± 0.02	0.68 ± 0.01	0.74 ± 0.04	Accuracy
	0.69 ± 0.02	0.68 ± 0.01	0.74 ± 0.04	F1
	0.53 ± 0.03	0.52 ± 0.02	0.62 ± 0.06	MCC

The table summarizes classification performance for color-coded 3D AL and Ref-AL images using RF, SVM, and CNN models trained on grayscale images. All values reported are averages and standard deviations over 10 independent trials

projections. Thus, under these conditions, heatmap views from collection 4 were no longer a suitable AL representation, emphasizing the need for applying the color gradient for heatmaps. Moreover, observed differences in model performance between grayscale and color-coded images could be more generally explained. The color gradient used red for low, yellow for intermediate, and green for high potency values while the grayscale was determined as a weighted sum of the red, green and blue channels with weights of 0.299, 0.587, and 0.114, respectively. Thus, yellow resulting from combining red and green appeared brightest, followed by green and red, which yielded darker gray tones representing both high and low high potencies. Hence, dark gray tones did not distinguish between high and low potency values, corresponding to a loss of information. This explained why model performance reduction was largest for the top-down elevation view (0.67 ± 0.03 compared to 0.73 ± 0.04), which exclusively relied on color to differentiate topographical features. By contrast, lower elevation views profited from the presence of topographically detectable peaks and valleys that were retained in the grayscale images, thus confirming relevance of these features for ML.

Table 5 Classification of models trained on black and white images

Collection	RF	SVM	CNN	Metric
1	0.48 ± 0.01	0.44 ± 0.01	0.62 ± 0.02	Accuracy
	0.47 ± 0.01	0.45 ± 0.01	0.62 ± 0.02	F1
	0.21 ± 0.02	0.16 ± 0.01	0.43 ± 0.04	MCC
2	0.46 ± 0.01	0.43 ± 0.01	0.61 ± 0.03	Accuracy
	0.46 ± 0.01	0.44 ± 0.01	0.61 ± 0.03	F1
	0.20 ± 0.02	0.15 ± 0.02	0.42 ± 0.04	MCC
3	0.47 ± 0.01	0.46 ± 0.02	0.60 ± 0.02	Accuracy
	0.47 ± 0.01	0.46 ± 0.02	0.60 ± 0.02	F1
	0.20 ± 0.02	0.19 ± 0.03	0.41 ± 0.03	MCC
4	0.45 ± 0.02	0.47 ± 0.03	0.54 ± 0.05	Accuracy
	0.45 ± 0.02	0.48 ± 0.03	0.54 ± 0.04	F1
	0.17 ± 0.03	0.21 ± 0.04	0.32 ± 0.07	MCC
5	0.41 ± 0.03	0.39 ± 0.01	0.70 ± 0.05	Accuracy
	0.41 ± 0.03	0.39 ± 0.01	0.69 ± 0.04	F1
	0.12 ± 0.05	0.09 ± 0.02	0.54 ± 0.07	MCC
6	0.52 ± 0.03	0.51 ± 0.02	0.69 ± 0.07	Accuracy
	0.52 ± 0.04	0.51 ± 0.03	0.69 ± 0.07	F1
	0.29 ± 0.05	0.26 ± 0.03	0.53 ± 0.10	MCC
7	0.69 ± 0.02	0.68 ± 0.01	0.73 ± 0.02	Accuracy
	0.69 ± 0.02	0.68 ± 0.01	0.73 ± 0.02	F1
	0.53 ± 0.03	0.52 ± 0.02	0.59 ± 0.04	MCC

The table summarizes classification performance for color-coded 3D AL and Ref-AL images using RF, SVM, and CNN models trained on b/w images. All values reported are averages and standard deviations over 10 independent trials

Table 6 Classification of pre-processed models on the basis of edge detection

Collection	RF		SVM		Metric
	Canny	Sobel	Canny	Sobel	
1	0.48 ± 0.00	0.50 ± 0.01	0.52 ± 0.01	0.57 ± 0.01	Accuracy
	0.48 ± 0.01	0.50 ± 0.01	0.52 ± 0.01	0.58 ± 0.01	F1
	0.23 ± 0.01	0.26 ± 0.01	0.28 ± 0.02	0.36 ± 0.02	MCC
2	0.44 ± 0.01	0.50 ± 0.01	0.50 ± 0.02	0.56 ± 0.01	Accuracy
	0.45 ± 0.01	0.50 ± 0.01	0.50 ± 0.02	0.56 ± 0.01	F1
	0.16 ± 0.02	0.25 ± 0.02	0.25 ± 0.02	0.34 ± 0.02	MCC
3	0.45 ± 0.01	0.49 ± 0.02	0.51 ± 0.02	0.56 ± 0.01	Accuracy
	0.46 ± 0.01	0.49 ± 0.02	0.52 ± 0.02	0.57 ± 0.01	F1
	0.18 ± 0.01	0.24 ± 0.02	0.27 ± 0.03	0.35 ± 0.02	MCC
4	0.43 ± 0.01	0.54 ± 0.03	0.53 ± 0.04	0.60 ± 0.03	Accuracy
	0.44 ± 0.02	0.54 ± 0.03	0.54 ± 0.04	0.60 ± 0.03	F1
	0.16 ± 0.02	0.31 ± 0.05	0.30 ± 0.06	0.40 ± 0.04	MCC
5	0.44 ± 0.02	0.50 ± 0.02	0.47 ± 0.03	0.55 ± 0.02	Accuracy
	0.44 ± 0.02	0.51 ± 0.02	0.49 ± 0.03	0.57 ± 0.02	F1
	0.16 ± 0.03	0.26 ± 0.04	0.21 ± 0.05	0.33 ± 0.03	MCC
6	0.42 ± 0.01	0.50 ± 0.02	0.52 ± 0.03	0.56 ± 0.02	Accuracy
	0.42 ± 0.01	0.51 ± 0.03	0.52 ± 0.02	0.58 ± 0.02	F1
	0.13 ± 0.02	0.25 ± 0.04	0.28 ± 0.04	0.35 ± 0.03	MCC
7	0.61 ± 0.01	0.66 ± 0.03	0.73 ± 0.02	0.74 ± 0.01	Accuracy
	0.63 ± 0.01	0.66 ± 0.03	0.73 ± 0.02	0.74 ± 0.01	F1
	0.42 ± 0.02	0.50 ± 0.04	0.59 ± 0.03	0.61 ± 0.01	MCC

Furthermore, CNN model performance on collection 1 was superior to RF and SVM models. However, RF and SVM were also able to distinguish between smooth, rugged and heterogeneous 3D AL variants on the basis of grayscale encodings, with a mean prediction accuracy of 0.57 ± 0.01 and 0.53 ± 0.01, respectively. Here, random predictions would correspond to an accuracy of 0.33. CNNs outperformed SVM and RF models for the other collections, with a relative increase in accuracy of 10% or more and consistently higher F1 and MCC values. However, prediction accuracy of all methods improved significantly for the 0° elevation angle images (collection 7) where SVM and RF models reached an accuracy of 0.68 ± 0.03 and 0.69 ± 0.02, respectively, and CNNs of 0.74 ± 0.04. Taken together, the results for models trained on grayscale images revealed that (i) features learned by CNNs from 3D AL images color-coded by potency contributed to the predictions but were not essential and (ii) elevation (peak) information, as emphasized by images from collection 7, was of critical relevance for accurate classifications.

Next, SVM, RF, and CNN models trained on b/w images were investigated. As illustrated in Fig. 2, compared to original 3D AL images, b/w image variants (resulting from binarization of pixel intensities) had drastically reduced information content. Consequently, prediction accuracy of all models trained on b/w image variants was further reduced compared to models trained on grayscale images (Table 5). CNNs retained limited predictive ability for collection 1, with a mean accuracy of 0.62 ± 0.02, but mostly retrained classification performance for images with decreasing elevation angles (65°, 35°, and 0°; collection 5–7). For 0° elevation (collection 7), classification accuracy of SVM and RF models was highest, with 0.68 ± 0.01 and 0.69 ± 0.02, respectively. These observations again emphasized the critical importance of capturing 3D AL altitude information for meaningful image classification.

Edge detection in pre-processed images

Unlike CNN models, SVM and RF models cannot directly learn image feature representations from pixel values. Thus, to further evaluate the predictive ability of SVM and RF models to classify 3D AL images on the basis of topological features, Sobel operators and Canny edge filters were applied to all grayscale images. SVM and RF models were then derived using edge-filtered images from half of the activity classes and tested on edge-filtered images of the remaining half of the classes.

Iqbal *et al. J Cheminform*   (2020) 12:34

Page 13 of 15

The classification results for these SVM and RF models are reported in Table 6. For the most part, no further improvements relative to the performance of RF and SVM trained on grayscale or b/w images were observed. In addition, SVM and RF performance did not improve when applying the ORB and Harris boundary feature filters. Overall, the combination of SVM and the Sobel operator was overall preferred but confined in accuracy to 0.60; however, with a notable exception for collection 7. In this case, these SVM models achieved an accuracy of 0.73 ± 0.02 and 0.74 ± 0.01 for the Canny and Sobel filters, respectively. Interestingly, this level of classification accuracy was comparable to the one achieved by CNNs trained on original color-coded 3D AL and Ref-AL images.

Results are reported for RF and SVM models trained on edge-filtered images and applied to classify such images originating from different activity classes. All values reported are averages and standard deviations over 10 independent trials.

Importantly, edges in pre-processed images resulted from data peaks in rugged regions of 3D ALs. Hence, the classification performance of SVM model on these filtered image variants clearly indicated the critical importance of altitude-dependent topological features for image classification. The sparseness of such features in smooth Ref-ALs rationalized the ability of classification models to distinguish these image variants with very high accuracy from rugged Ref-ALs and heterogeneous 3D ALs. In these two image categories, altitude-dependent topological feature accounting for peaks in 3D ALs were prevalent. Accordingly, rugged and heterogeneous AL variants were more difficult to distinguish from each other. However, even for this classification task, overall accurate predictions were obtained, indicating that machine learning correctly detected differences in relative feature density and feature combinations.

Conclusions

In this work, we have investigated classification of 3D AL images using machine learning. The study was motivated by the need to complement SAR visualization and graphical SAR analysis with systematic computational assessment of different 3D AL representations. The study concept took into consideration that images also represented a preferred data format for machine learning using 3D AL models of compound data set of diverse composition. Therefore, for 38 different activity classes with significant compound potency variations, we have generated a variety of 3D AL image variants including Ref-ALs designed to emphasize different topological features in a consistent way. These sets of images were classified using ML on the basis of topological features accounting for different SAR characteristics. Original color-coded 3D AL models and corresponding heatmap views were accurately classified using CNN models trained on learned representations, lending credence to the use of such representations. In addition, CNN, SVM, and RF models produced meaningful classification of 3D AL images with models trained on image variants having lower information content. Furthermore, SVM models were able to accurately predict pre-processed images on the basis of edge information representing altitude-dependent features. Thus, investigating a hierarchy of AL representations with successively reduced information content revealed factors that were critical for classification. Taken together, classification of images of different design representing 3D ALs from different viewpoints revealed a pivotal role of elevation-dependent features for accurate image classification, hence providing a diagnostic for the predictions. These features were decisive for distinguishing images of smooth 3D ALs with very high accuracy from images of rugged and heterogeneous 3D ALs. In addition, images of rugged and heterogeneous 3D ALs were also differentiated with meaningful accuracy. Accordingly, on the basis of our proof-of-concept investigation, image analysis is thought to have considerable potential for distinguishing between 3D ALs with different topologies and hence for classifying them on the basis of SAR information they contain. Accordingly, future work will focus on differentiating heterogeneous 3D ALs on the basis of the relative content of SAR continuity versus discontinuity. Classification of such 3D ALs might be attempted on the basis of images capturing differential density of elevation-dependent topological features.

Iqbal *et al. J Cheminform*   (2020) 12:34

Page 14 of 15

References

1. Bajorath J, Peltason L, Wawer M, Guha R, Lajiness MS, Van Drie JH (2009) Navigating structure-activity landscapes. Drug Discov Today 14:698–705. https://doi.org/10.1016/j.drudis.2009.04.003
2. Wassermann AM, Wawer M, Bajorath J (2010) Activity landscape representations for structure-activity relationship analysis. J Med Chem 53:8209–8223. https://doi.org/10.1021/jm100933w
3. Peltason L, Iyer P, Bajorath J (2010) Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. J Chem Inf Model 50:1021–1033. https://doi.org/10.1021/ci100091e
4. Guha R (2010) The ups and downs of structure–activity landscapes. Chemoinformatics and computational chemical biology. Humana Press, Totowa, pp 101–117. https://doi.org/10.1007/978-1-60761-839-3_3
5. Bajorath J (2012) Modeling of activity landscapes for drug discovery. Expert Opin Drug Discov 7:463–473. https://doi.org/10.1517/17460441.2012.679616
6. Medina-Franco JL, Yongye AB, Fabian López-Vallejo F (2012) Consensus models of activity landscapes. Stat Model Mol Descrip QSAR/QSPR 2:307–326. https://doi.org/10.1002/9783527645121.ch11
7. Guha R, Medina-Franco JL (2014) On the validity versus utility of activity landscapes: are all activity cliffs statistically significant? J Cheminform 6:11. https://doi.org/10.1186/1758-2946-6-11
8. Miyao T, Funatsu K, Bajorath J (2019) Three-dimensional activity landscape models of different design and their application to compound mapping and potency prediction. J Chem Inf Model 59:993–1004. https://doi.org/10.1021/acs.jcim.8b00661
9. Zhang W, Itoh K, Tanida J, Ichioka Y (1990) Parallel distributed processing model with local space-invariant interconnections and its optical architecture. Appl Opt 29:4790–4797. https://doi.org/10.1364/AO.29.004790
10. Paoletti ME, Haut JM, Plaza J, Plaza A (2018) A new deep convolutional neural network for fast hyperspectral image classification. ISPRS J Photogramm Remote Sens 145:120–147. https://doi.org/10.1016/j.isprs jprs.2017.11.021
11. Hatipoglu N, Bilgin G (2017) Cell segmentation in histopathological images with deep learning algorithms by utilizing spatial relationships. Med Biol Eng Comput 55:1829–1848. https://doi.org/10.1007/s11517-017-1630-1
12. Pereira S, Pinto A, Alves V, Silva CA (2016) Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans Med Imaging 35:1240–1251. https://doi.org/10.1109/TMI.2016.2538465
13. Kallenberg M, Petersen K, Nielsen M, Ng AY, Diao P, Igel C, Vachon CM, Holland K, Winkel RR, Karssemeijer N, Lillholm M (2016) Unsupervised deep learning applied to breast density segmentation and mammographic risk Scoring. IEEE Trans Med Imaging 35:1322–1331. https://doi.org/10.1109/TMI.2016.2532122
14. Setio AAA, Ciompi F, Litjens G, Gerke P, Jacobs C, van Riel SJ, Wille MMW, Naqibullah M, Sánchez CI, van Ginneken B (2016) Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE Trans Med Imaging 35:1160–1169. https://doi.org/10.1109/TMI.2016.2536809
15. Tan JH, Acharya UR, Bhandary SV, Chua KC, Sivaprasad S (2017) Segmentation of optic disc, fovea and retinal vasculature using a single convolutional neural network. J Comput Sci 20:70–79. https://doi.org/10.1016/j.jocs.2017.02.006
16. Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
17. Cortés-Ciriano I, Bender A (2019) KekuleScope: prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images. J Cheminform 11:41. https://doi.org/10.1186/s13321-019-0364-5
18. Fernandez M, Ban F, Woo G, Hsing M, Yamazaki T, LeBlanc E, Rennie PS, Welch WJ, Cherkasov A (2018) Toxic Colors: the use of deep learning for predicting toxicity of compounds merely from their graphic images. J Chem Inf Model 58:1533–1543. https://doi.org/10.1021/acs.jcim.8b00338
19. Goh GB, Vishnu A, Siegel C, Hodas N (2018) Using rule-based labels for weak supervised learning: A ChemNet for transferable chemical property prediction. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. https://doi.org/10.1145/3219819.3219838
20. Simm J, Klambauer G, Arany A, Steijaert M, Wegner JK, Gustin E, Chupakhin V, Chong YT, Vialard J, Buijnsters P, Velter I, Vapirev A, Singh S, Carpenter AE, Wuyts R, Hochreiter S, Moreau Y, Ceulemans H (2018) Repurposing high-throughput image assays enables biological activity prediction for drug discovery. Cell Chem Biol 25:611–618.e3. https://doi.org/10.1016/j.chembiol.2018.01.015
21. Scheeder C, Heigwer F, Boutros M (2018) Machine learning and image-based profiling in drug discovery. Curr Opin Syst Biol 10:43–52. https://doi.org/10.1016/j.coisb.2018.05.004
22. Kraus OZ, Ba JL, Frey BJ (2016) Classifying and segmenting microscopy images with deep multiple instance learning. Bioinformatics 32:i52–i59. https://doi.org/10.1093/bioinformatics/btw252
23. Hofmarcher M, Rumetshofer E, Clevert D-A, Hochreiter S, Klambauer G (2019) Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks. J Chem Inf Model 59:1163–1171. https://doi.org/10.1021/acs.jcim.8b00670
24. Vapnik VN (2000) The nature of statistical learning theory, 2nd edn. Springer, New York. https://doi.org/10.1007/978-1-4757-3264-1
25. Tarabalka Y, Fauvel M, Chanussot J, Benediktsson JA (2010) SVM- and MRF-based method for accurate classification of hyperspectral images. IEEE Geosci Remote Sens Lett 7:736–740. https://doi.org/10.1109/LGRS.2010.2047711
26. Bovolo F, Bruzzone L, Carlin L (2010) A novel technique for subpixel image classification based on support vector machine. IEEE Trans Image Process 19:2983–2999. https://doi.org/10.1109/TIP.2010.2051632
27. Foody GM, Mathur A (2004) A relative evaluation of multiclass image classification by support vector machines. IEEE Trans Geosci Remote Sens 42:1335–1343. https://doi.org/10.1109/TGRS.2004.827257
28. Tzotsos A, Argialas D (2008) Support vector machine classification for object-based image analysis. In: Blaschke T, Lang S, Hay GJ (eds) Object-based image analysis: spatial concepts for knowledge-driven remote sensing applications. Springer, Berlin Heidelberg, pp 663–677. https://doi.org/10.1007/978-3-540-77058-9_36
29. Song Q, Wenjie H, Xie W (2002) Robust support vector machine with bullet hole image classification. IEEE Trans Syst Man. Cybern Part C (Applications Rev) 32:440–448. https://doi.org/10.1109/TSMCC.2002.807277
30. Chapelle O, Haffner P, Vapnik VN (1999) Support vector machines for histogram-based image classification. IEEE Trans Neural Netw 10:1055–1064. https://doi.org/10.1109/72.788646
31. Banfield RE, Hall LO, Bowyer KW, Kegelmeyer WP (2007) A comparison of decision tree ensemble creation techniques. IEEE Trans Pattern Anal Mach Intell 29:173–180. https://doi.org/10.1109/TPAMI.2007.250609
32. Xu B, Ye Y, Nie L (2012) An improved random forest classifier for image classification. In: 2012 IEEE International Conference on Information and Automation, pp 795–800. https://doi.org/10.1109/ICInfA.2012.6246927
33. Gaulton A, Hersey A, Nowotka ML, Patricia Bento A, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrian-Uhalte E, Davies M, Dedman N, Karlsson A, Magarinos MP, Overington JP, Papadatos G, Smit I, Leach AR (2017) The ChEMBL database in 2017. Nucleic Acids Res. https://doi.org/10.1093/nar/gkw1074
34. Hu H, Stumpfe D, Bajorath J (2018) Rationalizing the formation of activity cliffs in different compound data sets. ACS Omega 3:7736–7744. https://doi.org/10.1021/acsomega.8b01188
35. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754. https://doi.org/10.1021/ci100050t

36. Rogers DJ, Tanimoto TT (1960) A computer program for classifying plants. Science 80(132):1115–1118. https://doi.org/10.1126/science.132.3434.1115
37. OEChem Toolkit, version 2018.10.1, OpenEye Scientific Software: Santa Fe, NM
38. Borg I, Groenen PJF (2005) Modern multidimensional scaling: theory and applications. Springer, New York. https://doi.org/10.1007/0-387-28981-X
39. Lowe D, Tipping ME (1997) NeuroScale: novel topographic feature extraction using RBF networks. In: Mozer MC, Jordan MI, Petsche T (eds) Advances in neural information processing systems. MIT Press, Boston, pp 543–549
40. Cressie N (1993) Statistics for spatial data. Wiley, New York. https://doi.org/10.1002/9781119115151.ch1
41. Rasmussen CE (2005) Gaussian processes in machine learning. Adv Lect Mach Learn Springer, Berlin, pp 63–71. https://doi.org/10.1007/978-3-540-28650-9_4
42. Peltason L, Bajorath J (2007) SAR Index: quantifying the nature of structure-activity relationships. J Med Chem 50:5571–5578. https://doi.org/10.1021/jm0705713
43. Peltason L, Bajorath J (2009) Systematic computational analysis of structure-activity relationships: concepts, challenges and recent advances. Future Med Chem 1:451–466. https://doi.org/10.4155/fmc.09.41
44. Acharya UR, Oh SL, Hagiwara Y, Tan JH, Adeli H (2018) Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. Comput Biol Med 100:270–278. https://doi.org/10.1016/j.compbiomed.2017.09.017
45. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1929–1958
46. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, et al (2016) TensorFlow: A System for Large-Scale Machine Learning. In: 12th {USENIX} Symposium on Operating Systems Design and Implementation (OSD'16), pp 265–283
47. Chollet F et al (2015) Keras. https://keras.io/
48. Culjak I, Abram D, Pribanic T, Dzapo H, Cifrek M (2012) A brief introduction to OpenCV. In: 2012 Proceedings of the 35th International Convention MIPRO, pp 1725–1730
49. OpenCv (2014) OpenCV Library. https://opencv.org/
50. Howse J (2013) OpenCV Computer vision with Python. Packt Publishing LTD, Birmingham
51. Bradski G (2000) The OpenCV Library. https://www.drdobbs.com/open-source/the-opencv-library/184404319
52. Xu X, Xu S, Jin L, Song E (2011) Characteristic analysis of Otsu threshold and its applications. Pattern Recognit Lett 32:956–961. https://doi.org/10.1016/j.patrec.2011.01.021
53. Wenshuo G, Xiaoguang Z, Lei Y, Huizhong L (2010) An improved Sobel edge detection. In: 2010 3rd International conference on computer science and information technology, pp 67–71. https://doi.org/10.1109/ICCSIT.2010.5563693
54. Ding L, Goshtasby A (2001) On the Canny edge detector. Pattern Recognit 34:721–725. https://doi.org/10.1016/S0031-3203(00)00023-6
55. Rublee E, Rabaud V, Konolige K, Bradski G (2011) ORB: An efficient alternative to SIFT or SURF. In: 2011 International conference on computer vision, pp 2564–2571. https://doi.org/10.1109/ICCV.2011.6126544
56. Harris C, Stephens M (1988) A combined corner and edge detector. In: Proceedings of the Alvey Vision Conference. Alvety Vision Club, 23.1-23.6
57. Zhu S, Ji X, Xu W, Gong Y (2005) Multi-labelled classification using maximum entropy method. In: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, pp 274–281. https://doi.org/10.1145/1076034.1076082
58. Matthews B (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochim Biophys Acta 405:442–451. https://doi.org/10.1016/0005-2795(75)90109-9
59. Chinchor N (1992) MUC-4 evaluation metrics. In: Proceedings of the fourth message understanding conference, pp. 22–29. https://doi.org/10.3115/1072064.1072067

Publisher's Note

Summary

For the first time 3D AL models image representations with varying SAR contents were investigated using machine learning and image processing approaches for classification prediction. The overall analysis demonstrated that the 3D ALs image snapshots of multiple viewpoints of 3D ALs of different SAR characteristics were successfully classified based on color-profiles and elevation-dependent features. The CNN and machine learning models successfully distinguished original heterogeneous 3D ALs from reference predominantly smooth 3D ALs. Furthermore, the models distinguished original heterogeneous 3D ALs from reference predominantly rugged 3D ALs with meaningful accuracy. The conceptual investigation provided evidence of the contribution of elevation-based SAR topographical features in 3D AL classification analysis. Image formats obtained from 3D ALs with different topologies have considerable potential for distinguishing between 3D ALs with different SAR contents. The obtained results encourage further investigation of SARs for detection and analysis of topographical features with the application of image processing techniques based on image data.

Chapter 3

Computational Method for Quantitative Comparison of Activity Landscapes on the Basis of Image Data

Introduction

SAR types represented in a 3D AL model resemble real-world geographical topologies. For example, SAR characteristics in 3D ALs can be visually perceived as the combination of mountains and valleys where discontinuous surfaces tend to form mountains or peaks and continuous surfaces can be viewed as valleys. *Chapter 2* establishes proof of concept where machine learning models can discriminate between different 3D AL models accurately and successfully classify original 3D AL surfaces from pre-dominantly smooth and rugged surface types. However, the findings in chapter 2 are limited to 3D AL variants and reference states with intentionally modified topologies. Therefore, it remained unclear whether image analysis might also be applicable to discern between actual 3D ALs with original data sets. Accordingly, the current chapter investigates 3D ALs for quantitative comparison of SARs based on actual unmodified data sets. In this conceptual analysis, 3D AL models were generated for four target activity classes and the heatmap (top-down) representations were quantitatively compared to discern the SAR (dis)-similarity in 3D ALs. Furthermore, the current chapter introduces a numerical methodology to computationally extract distributed potency features from 3D ALs images corresponding to

different SAR topological features and provides a robust 3D AL comparison based on the feature vector that reflects varying SAR information content.

Computational Method for Quantitative Comparison of Activity Landscapes on the Basis of Image Data

Javed Iqbal, Martin Vogt and **Jürgen Bajorath** *

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry,
Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, D-53115 Bonn, Germany;
jiqbal@bit.uni-bonn.de (J.I.); martin.vogt@bit.uni-bonn.de (M.V.)
* Correspondence: bajorath@bit.uni-bonn.de; Tel.: +49-228-7369-100

Abstract: Activity landscape (AL) models are used for visualizing and interpreting structure–activity relationships (SARs) in compound datasets. Therefore, ALs are designed to present chemical similarity and compound potency information in context. Different two- or three-dimensional (2D or 3D) AL representations have been introduced. For SAR analysis, 3D AL models are particularly intuitive. In these models, an interpolated potency surface is added as a third dimension to a 2D projection of chemical space. Accordingly, AL topology can be associated with characteristic SAR features. Going beyond visualization and a qualitative assessment of SARs, it would be very helpful to compare 3D ALs of different datasets in more quantitative terms. However, quantitative AL analysis is still in its infancy. Recently, it has been shown that 3D AL models with pre-defined topologies can be correctly classified using machine learning. Classification was facilitated on the basis of AL image feature representations learned with convolutional neural networks. Therefore, we have further investigated image analysis for quantitative comparison of 3D ALs and devised an approach to determine (dis)similarity relationships for ALs representing different compound datasets. Herein, we report this approach and demonstrate proof-of-principle. The methodology makes it possible to computationally compare 3D ALs and quantify topological differences reflecting varying SAR information content. For SAR exploration in drug design, this adds a quantitative measure of AL (dis)similarity to graphical analysis.

Keywords: active compounds; three-dimensional activity landscapes; topological features; structure–activity relationships; image analysis; grid representations; landscape comparison; similarity measures

1. Introduction

Graphical representations are desirable to support the analysis of structure–activity relationships (SARs), especially when large sets of active compounds are investigated [1,2]. SARs are determined by chemical similarity and potency relationships between compounds active against a given target [3]. If sequential structural modifications of compounds lead to small or moderate changes in potency, SARs are continuous in nature. By contrast, if small structural changes cause potency alterations of large magnitude, SARs are discontinuous. Activity cliffs (ACs), i.e., pairs of structural analogs with large potency differences, are centers of SAR discontinuity in datasets [3,4]. Activity landscape (AL) representations were introduced to combine the analysis of similarity and potency information in various ways [5–7]. AL representations differ substantially in their design and complexity [5]. They range from plots and graphs such as the structure–activity similarity map [8,9], SAR map [10], or ligand-target differentiation map [11] and annotated molecular networks such as network-like similarity graphs [12] to multi-dimensional representations [13]. Formally, in an n-dimensional AL

model, n-1 dimensions represent a chemical descriptor/feature space and the nth dimension represents activity space [13]. In other words, a multi-dimensional AL can be rationalized as a chemical feature space containing a biological activity hypersurface [5]. For SAR visualization and interpretation, three-dimensional (3D) AL views are particularly attractive because they are akin to geographical maps, with topologies accounting for the presence of characteristic local SARs [13–15]. Accordingly, smooth regions in 3D ALs mirror SAR continuity, whereas rugged regions are indicative of SAR discontinuity and contain ACs. For different compound datasets, 3D AL models can be constructed by combining a two-dimensional (2D) projection of (n-1)-dimensional chemical space with compound potency values subsequently added as a third dimension [14,15]. From distributed potency values, a coherent potency surface is interpolated and color-coded by potency, resulting in a 3D view reminiscent of a geographical map [14,15]. In many compound datasets, both continuous and discontinuous SARs are found to coexist that are formed by different compound subsets [16]. The coexistence of locally continuous and discontinuous SARs gives rise to global SAR heterogeneity [16] and 3D ALs containing both smooth and rugged regions, termed variable ALs [14,15]. Of note, a 3D AL principally represents a non-linear quantitative SAR (QSAR) model, given its interpolated potency hypersurface to which descriptor coordinates of test compounds can be mapped. As such, the 3D AL is suitable for mapping of active compounds to regions of high or low potency, but not for actual potency prediction in lieu of machine learning [15]. This typically is a consequence of intrinsic overfitting of a 3D AL model to a given dataset, which yields a high-resolution SAR visualization, but prohibits generalization of the model for the prediction of numerical potency values [15]. Hence, while numerical SAR analysis functions are applicable to account for SAR continuity [16], discontinuity [16–18], or heterogeneity [16] in datasets, 3D AL models have thus far only been qualitatively analyzed and compared [19].

While SAR visualization is a key task of 3D AL modeling, one would clearly benefit from a more quantitative comparison of 3D ALs. Qualitative analysis of 3D ALs typically aims at relating topological features to SAR characteristics such as the relative content of continuous vs. discontinuous SAR components. One of the key tasks in SAR exploration of compound datasets is revealing differences in SAR information between different sets [5]. For example, for practical compound optimization, one often would like to assess which sets of compounds with activity against related targets have similar SAR characteristics and are rich in SAR discontinuity. Such datasets would be preferentially selected as compound source to guide optimization efforts. On the other hand, for computational SAR modeling and QSAR, one would like to give preference to datasets that contain more SAR continuity. Estimating and comparing relative SAR information content goes beyond the opportunities of SAR visualization and qualitative AL comparison. For example, large-scale SAR analysis would greatly benefit from identifying datasets that have similar SAR characteristics to a given compound set of interest, which is impossible on the basis of visual inspection. Hence, the ability to systematically relate topological differences to varying SAR information content and quantify SAR similarity of different datasets would complement SAR visualization and further extend the utility of 3D ALs, beyond intuitive analysis and comparison. Quantitative assessment would also aid in differentiating between datasets with heterogeneous 3D ALs displaying subtle topological differences that are difficult to appreciate on the basis of visual inspection. This is of practical relevance when evaluating the potential of further advancing SARs for compounds with activity against related targets. In such cases, one would favor focusing on compound sources capturing more SAR discontinuity than others, similar to the application scenario described above. Hence, there are multiple reasons motivating the development of methods for quantitative comparison of ALs.

However, from a computational point of view, the development of quantitative 3D AL methods is far from being a trivial task. Recently, it has been attempted to classify 3D ALs using machine learning (ML) based on features extracted from color-coded AL images using convolutional neural networks [20]. For a given 3D AL, variants with altered topologies were generated by either increasing the smoothness (continuity) or ruggedness (discontinuity) of the original AL. These topologically modified reference states were then distinguished from original 3D ALs by binary class label prediction

using various ML approaches including deep learning [20]. These studies provided first evidence that 3D AL models with different topological features can be correctly classified on the basis of image data. However, the findings were limited to 3D AL variants and reference states with deliberately modified topologies. Accordingly, it remained unclear whether image processing might also be applicable to differentiate between original heterogeneous 3D ALs. Therefore, we have investigated a conceptually different image-based computational approach to determine (dis)similarity relationships between original 3D ALs of different compound datasets.

2. Results and Discussion

2.1. Activity Landscape Images

Images generated from 3D ALs preserve pairwise compound similarity relationships and their potency data as topographical features. Topology and color codes account for different SAR characteristics of compound datasets. Of note, SARs are determined by potency differences between compounds with varying degrees of similarity and are thus largely independent of absolute potency values. The interpolated potency surface of a 3D AL yields color gradients that can be represented in heatmaps without significant information loss (see Methods) [20]. Such heatmaps represent a top-down view of the 3D AL and the encoded color gradients implicitly—but comprehensively—account for the spatial distribution of topological features. Hence, 3D ALs and corresponding heatmaps are in principle well suited for image analysis. Given the aim of our method development effort, we have reasoned that comparing AL image features in a well-defined way should have the potential to discriminate between different 3D ALs in quantitative terms.

2.2. Image Similarity Analysis

Three-dimensional AL images embed topological features and color profiles, which are characterized by different color gradients resulting from potency value and compound similarity distributions (the combination of which determines AL topology). In heatmaps derived from 3D ALs, topological features and ensuing color gradients are encoded by color pixel intensities that can be algorithmically extracted. The basic premise underlying similarity-based comparison of 3D AL images, as introduced herein, is that scaled color pixel intensities can be quantitatively compared across different heatmaps. To this end, a common grid representation of heatmaps plays a central role. Using an evenly spaced grid, the heatmap is divided into a constant number of cells, which are assigned to different categories based upon color intensity threshold values. The distribution of cells over different categories is then quantitatively compared as a measure of AL (dis)similarity. Figure 1 illustrates the approach. Methodological and calculation details are provided in the Methods Section.

2.3. Heatmaps and Grid Representations

Conversion of 3D AL images into heatmaps established a reference frame for quantitative AL comparison. The heatmap corresponded to a top-down view of the color-coded 3D AL. Heatmaps were mapped onto an evenly spaced grid of dimensionality 56×60. Accordingly, each heatmap was divided into total 3360 cells. Figure 2 shows a 3D AL representation for a set of 673 corticotropin-releasing factor receptor 1 ligands, the corresponding heatmap, and its grid representation.
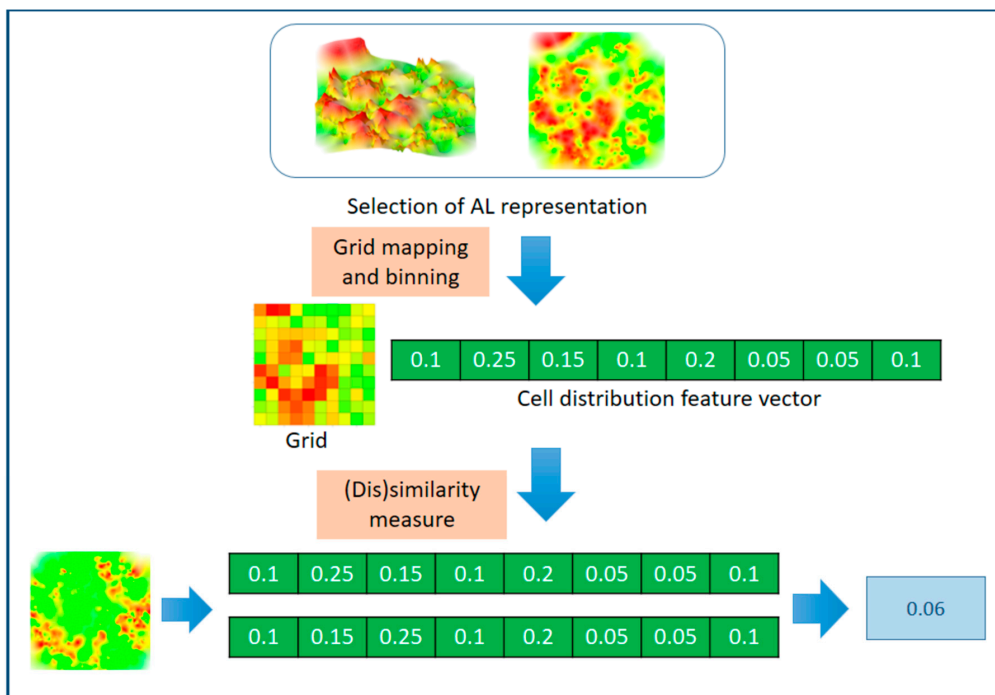
Figure 1. Three-dimensional activity landscape (AL) image-based similarity assessment. The schematic illustrates similarity analysis involving the conversion of 3D AL images into heatmaps, grid-based classification of heatmap cells according to different color intensity thresholds, and quantitative comparison of the resulting cell distribution feature vectors as a measure of AL (dis)similarity.
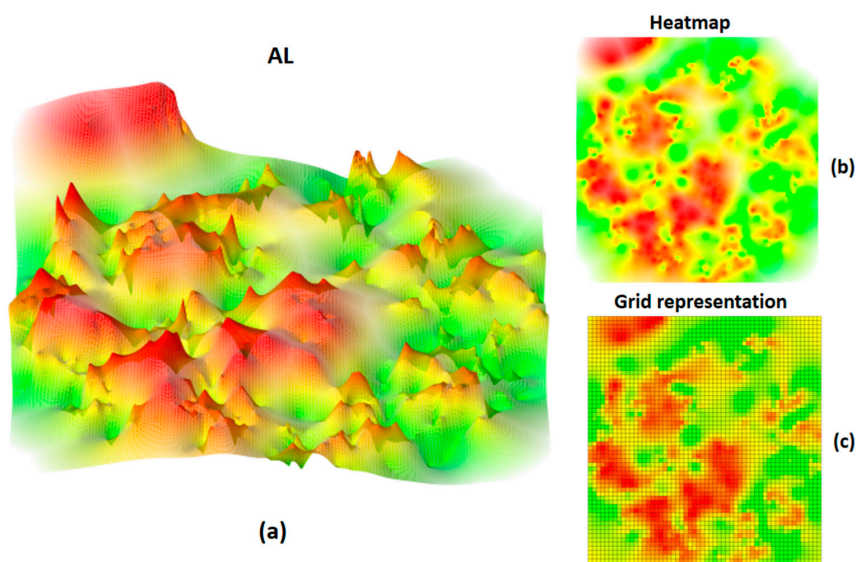

Figure 2. Activity landscape and heatmap representations. In (**a**), the 3D AL of a set of corticotropin-releasing factor receptor 1 ligands taken from ChEMBL version 23 [21] is shown (generated as detailed in the Methods Section). The surface is color-coded according to compound potency using a continuous spectrum ranking from red (high potency) over yellow to green (low potency). In (**b**), the corresponding heatmap is displayed. In (**c**), the heatmap is represented on an evenly spaced grid.

In Figure 3 below, the heatmap of the of corticotropin-releasing factor receptor 1 ligands is enlarged, and positions of exemplary weakly or highly potent compounds are mapped. These compounds originated from two different analog series and occupy distant regions in the heatmap. The weakly potent compounds are found in a green region (corresponding to a valley) and the highly potent in a red region (formed by peaks). The representation illustrates color intensity-based encoding of 3D AL topology resulting from different compound potency levels.
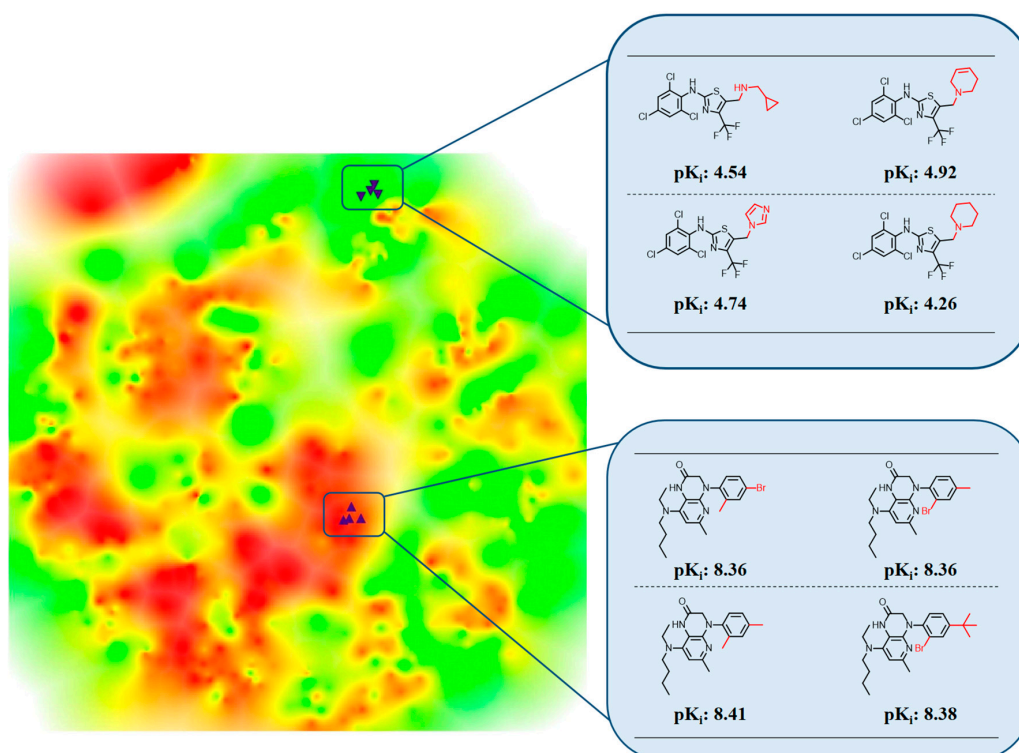

Figure 3. Heatmap with compound positions. The heatmap from Figure 2b is enlarged and positions of different compounds are indicated using black triangles. Exemplary weakly potent (top) and highly potent compounds (bottom) belong to two different analog series and map to green regions (valleys) and red regions, respectively. For each compound, its potency value is reported, and structural modifications distinguishing analogs from each series are highlighted in red.

For heatmaps, red and green channel pixel intensity values were combined into a single intensity value ranging from -1 to 1 (see Methods). To identify peaks using color intensities, positive threshold value intervals of (0, 0.25), (0.25, 0.5), (0.5, 0.75), and (0.75, 1.0) were applied. To identify smooth regions (valleys), negative threshold intervals of (0, −0.25), (−0.25, −0.5), (−0.50, −0.75), and (−0.75, −1.0) were used. Thus, different threshold intervals represented highest elevations (peaks), deepest valleys, and intermediate peak-to-valley and valley-to-peak regions. It should be noted that pixel intensities do not only encode potencies of individual molecules. Because intensities are obtained by interpolating a color gradient reflecting potencies of neighboring compounds, pixel intensities also implicitly account for locality information.

2.4. Grid-Based Similarity Analysis

Heatmap cells were assigned to eight different categories on the basis of the threshold value intervals specified above. The assignment yielded an AL-dependent distribution of categorized cells. Grid-based partitioning of a heatmap and categorization of the resulting cell population

49

provided two fundamental advantages for subsequent similarity analysis. First, a constant number of cells was obtained; second, the distribution of cells over different threshold categories was image orientation-invariant. While grid locality information is not preserved in the global cell distribution, SAR information from neighboring molecules is retained by interpolated intensities and hence implicitly included in the comparison. As a measure of AL (dis)similarity, cell distributions of different heatmaps were quantitatively compared by calculating symmetric relational entropy and cosine distances (see Methods).

2.5. Activity Landscape Comparison

Figure 4 shows heatmaps for 3D ALs of four exemplary compound datasets from ChEMBL version 23 [21] that are reported in Table 1. The datasets consisted of 673–887 compounds with activity against different targets covering different potency ranges. All four sets were characterized by SAR heterogeneity, i.e., their 3D ALs contained both smooth and rugged regions, corresponding to SAR continuity and discontinuity, respectively. However, on the basis of visual inspection, there also were apparent differences between these ALs, reflecting varying SAR information content. For example, the heatmap of compound dataset CHEMBL1800 (C1800) contained more and more widely distributed peak regions than the others, and C1800 and CHEMBL238 (C238) appeared to be overall the most dissimilar pair. Other relationships involving CHEMBL3759 (C3759) CHEMBL1833 (C1833) were difficult to judge, illustrating the limitations of visual inspection.
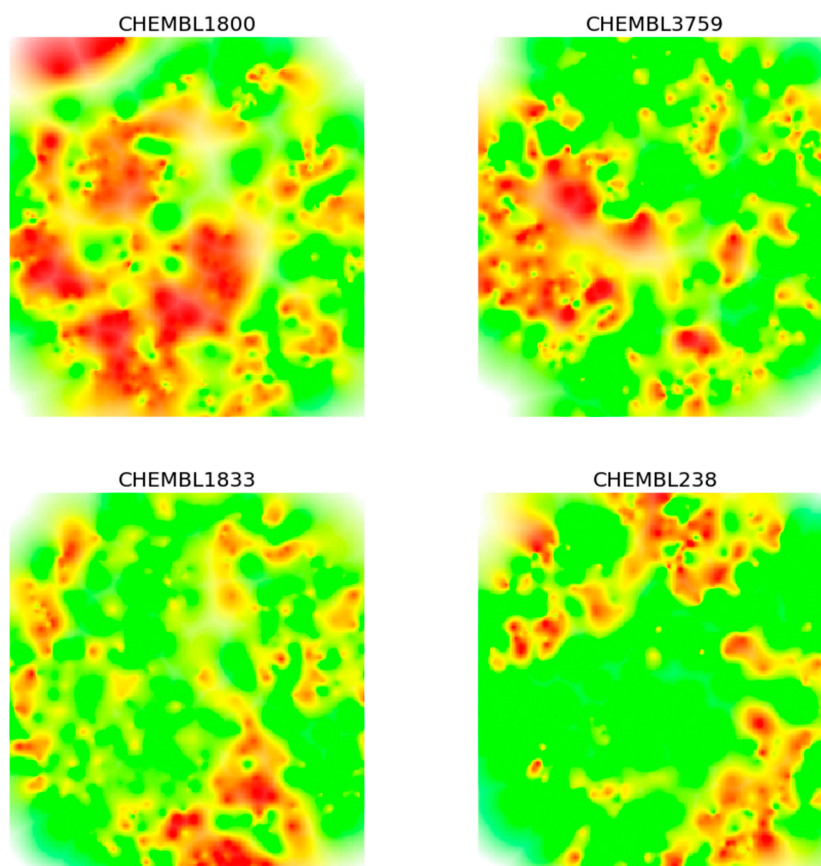

Figure 4. Heatmaps from different activity landscapes. Heatmaps derived from 3D ALs of four different compound datasets according to Table 1 are compared.

Table 1. Datasets. The table summarizes the composition of four exemplary compound datasets (different activity classes) used for 3D AL analysis.

ChEMBL Target ID	Target Name	Number of Compounds	Potency (pK_i)	
			Min	Max
1800	Corticotropin-Releasing Factor Receptor 1	673	4.3	9.7
3759	Histamine H4 receptor	887	2.9	10.4
1833	5-hydroxytryptamine receptor 2B	695	5.0	10.0
238	Sodium-dependent dopamine transporter	850	2.1	9.4

For the heatmaps, we then determined the grid-based cell intensity distributions over the eight threshold intervals. Figure 5 compares these distributions.
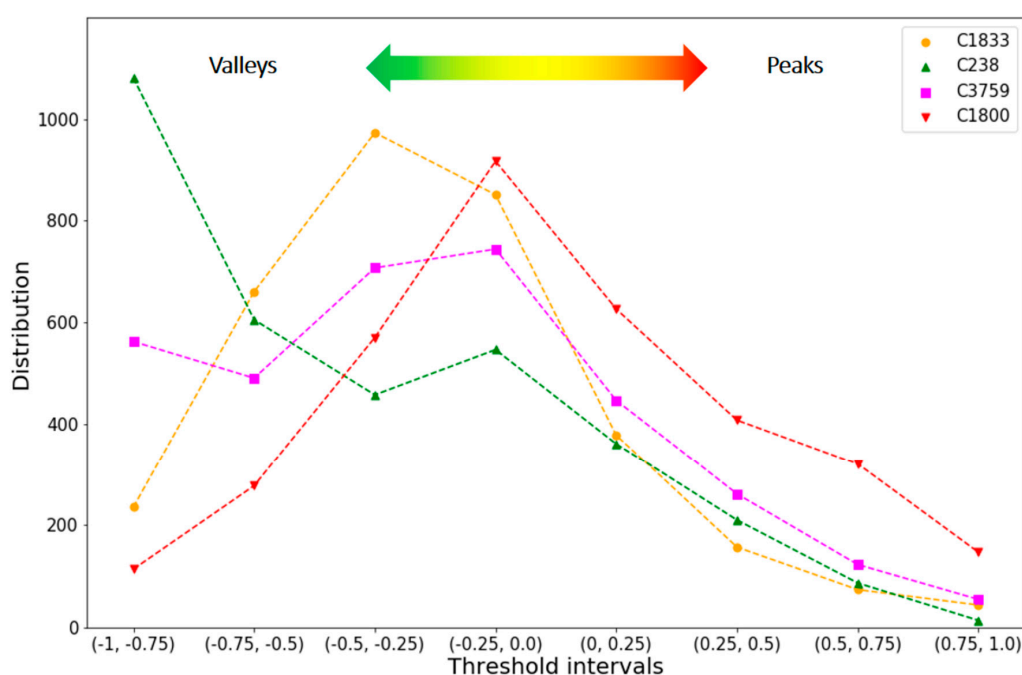

Figure 5. Cell intensity distributions. For the four compound datasets, cell distributions over eight threshold intervals are reported and color-coded as indicated. From left to right, valleys (interval 1–3), intermediate regions (interval 4,5), and peaks (interval 6–8) are accounted for.

Large differences between these distributions were found in smooth regions. Here, C238 (green curve in Figure 5) had by far the largest number of cells accounting for valleys and C1800 (red) the smallest. Furthermore, C1800 and C238 had the largest and smallest number of cells covering intermediate regions, respectively. In peak regions, the distributions of three of the four datasets were similar, except C1800, which had a larger number of cells accounting for peaks than the others. These findings were consistent with conclusions that could be drawn from visual inspection. Going beyond what could be concluded on the basis of visual inspection, the profiles of C1833 (orange) and C3759 (magenta) were found to be overall similar. While C1833 had more cells accounting for smooth regions than C3759, the traces of the distributions closely followed each other in intermediate and peak regions. Taken together, comparison of cell intensity distributions revealed quantifiable differences between AL images of different datasets and thus provided a sound basis for (dis)similarity analysis.

To quantify differences between cell distributions in a pairwise manner and provide a numerical measure of AL (dis)similarity, relative entropy (RE) was determined by calculating the Kullback–Leibler

divergences (KLD) [22] between feature vectors of cell distributions (see Methods). In addition, cosine distance (CD) values were determined for pairwise comparison of distribution feature vectors, a standard dissimilarity measure [23]. Increasing RE and CD values indicate increasing dissimilarity between ALs. Table 2 reports the results of pairwise comparisons of the cell intensity profiles.

Table 2. Similarity calculations. Relative entropy (RE) and cosine distance (CD) values for comparison of cell distribution feature vectors are reported.

AL Comparison	RE	CD
C1800/C3759	0.19	0.10
C1833/C238	0.28	0.24
C1800/C1833	0.22	0.12
C1800/C238	0.53	0.33
C3759/C1833	0.08	0.05
C3759/C238	0.09	0.09

As expected, largest RE (0.53) and CD (0.33) values were obtained for the C1800/C238 comparison, confirming that the ALs of these datasets were most dissimilar. By contrast, smallest RE and CD values were calculated for the C3759/C1833 (0.08 and 0.05, respectively) and the C3759/C238 comparison (0.09 in both instances). As discussed above, C3759 and C1833 yielded the overall most similar cell distributions. Furthermore, for the C1833/C238 comparison, intermediate RE (0.28) and CD (0.24) values were obtained, which were also reconcilable on the basis of the observed distribution traces. The comparisons revealed that RE and CD calculations were suitable for comparing cell distribution feature vectors. Since RE values covered a larger value range for the reported comparisons than CD values, we would assign preference to the former, at least in these cases. Regardless, the calculations reported herein are generally applicable and provide a first quantitative measure of 3D AL (dis)similarity.

2.6. Conclusions

The AL concept was introduced for graphical analysis of SARs contained in compound datasets. For SAR visualization, 3D AL representations are particularly intuitive since they are akin to geographical maps and their topological features mirror SAR characteristics. Three-dimensional ALs of most compound datasets are variable in nature, reflecting different degrees of SAR heterogeneity. Going beyond visual inspection and qualitative comparison of 3D ALs, the ability to quantitatively account for topological differences between 3D ALs would provide substantial support for SAR exploration of compound datasets and various practical applications. In this work, we have introduced a computational methodology to quantify (dis)similarity relationships between 3D ALs on the basis of image data. Three-dimensional AL images can be converted into heatmaps representing a top-down view of the ALs with very little loss in information such that color intensities and textures represent topological features. Heatmaps are then mapped onto evenly spaced grids with constant dimensions, which yields a constant number of cells, providing a basis for AL comparison. These cells are then categorized on the basis of color intensities, which implicitly account for the spatial distribution of corresponding topological features they represent. Differences in the distribution of cells over different threshold intervals are then quantified as a measure of 3D AL (dis)similarity. Importantly, cell-based comparison of ALs is image-orientation invariant and thus generally applicable. As shown in our proof-of-concept investigation, comparison of categorized cell distributions provides a meaningful quantitative readout for comparison of 3D ALs and, thus, further extends the utility of AL representations for SAR exploration.

3. Methods

3.1. Three-Dimensional Activity Landscapes

Three-dimensional AL models of compound datasets (Figure 2) were generated following the protocol reported in [20]. Briefly, chemical reference space was generated on the basis of extended connectivity fingerprint with bond diameter 4 (ECFP4) [24] Tanimoto distances [25] for pairwise compound comparisons. The 2D projection of chemical reference space was computed using multi-dimensional scaling (MDS) [26], applying a stress function based on pairwise Tanimoto distances. The potency surface was interpolated via Gaussian process regression (GPR) [27] and color-coded by compound potency applying a continuous color gradient from red (highest potency in a dataset) over yellow to green (lowest potency). Intermediate potency values were computed using a Gaussian process based on prior covariances of experimental potency values. The "Sum of Matern and White" kernel [27] was used assuming a mean of zero to derive relationships between experimental data points (potency values). Gaussian noise factors were applied to permit minor variations of z-values for points on the x-y plane and optimize the global fit of the surface to experimental data points. Noise factors were adjusted for each target activity class by optimizing the kernel's alpha parameter between 10^{-1} and 10^{-7} over 10 iterations. The gradient was applied to a limited pK_i range from 3.72 (green) over 5.75 (yellow) to 8.75 (red). Potencies outside this range were assigned to green (less than 3.72) or red (larger than 8.75).

3.2. Image Processing and Analysis

For each 3D AL, a heatmap was initially obtained using the RGB color model of OpenCV version 3.0 with eight bits per channel [28,29]. Heatmaps were cropped to dimensions of 280×300 pixels (starting from the original 600×400 pixels including white excess areas). Because 3D AL models were created by interpolating potency values using the 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. Accordingly, the dataset compound with lowest potency (brightest green pixels), intermediate (yellow pixels), and highest potency (brightest red 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 shades of white (i.e., interpolated surface area without experimental potency), which were accounted for by yellow hues using the RG model.

3.3. Grid Representation

Each heatmap was mapped to an evenly spaced grid of dimensionality 56×60, forming total 3360 square cells. Color intensity values were divided into eight different threshold intervals (categories), as specified above. Average pixel intensity values from the 25 pixels of each cell were assigned to the cell, and the distribution of cells over the eight threshold intervals was determined. For comparison, cell intensity distributions were encoded as individual feature vectors.

3.4. Similarity Analysis

To quantify (dis)similarity between any two 3D ALs images based upon their heatmaps, relative information entropy and cosine distances were calculated for cell distribution feature vectors. Relative entropy (RE), also known as the Kullback–Leibler divergence (KLD), is calculated between two probability distributions $P(x)$, $P(y)$. These were obtained from the feature vectors x and y by converting the distributions to relative frequencies. KLD is defined as [22]:

$$KLD(P(x)\|P(y)) = \sum P(x)\, log\left(\frac{P(x)}{P(y)}\right) \tag{1}$$

53

Given the intrinsic asymmetry of KLD, symmetric relative entropy values for comparison of feature vector probability distributions $P(x)$ and $P(y)$ were obtained by taking the average of $\text{KLD}(P(x)\|P(y))$ and $\text{KLD}(P(y)\|P(x))$:

$$RE(P(x), P(y)) \; = \; \frac{\sum \; P(x) \, log\left(\frac{P(x)}{P(y)}\right) + \sum \; P(y) \, log\left(\frac{P(y)}{P(x)}\right)}{2} \tag{2}$$

In addition, cosine distances between feature vectors were calculated. The cosine coefficient is widely used to measure the relationship between any two given feature vectors by calculating the cosine of the angle between the two vectors [23]. It is defined as the inner product of two vectors divided by the product of their lengths. The cosine distance CD is obtained by subtracting the cosine similarity value from 1 and given by:

$$CD(x, y) = 1 - \left(\frac{x \cdot y}{\|x\|\|y\|}\right) \tag{3}$$

References

1. Stumpfe, D.; Bajorath, J. Methods for SAR Visualization. *RSC Adv.* **2012**, *2*, 369–378. [CrossRef]
2. Medina-Franco, J.L.; Martinez-Mayorga, K.; Giulianotti, M.A.; Houghten, R.A.; Pinilla, C. Visualization of the Chemical Space in Drug Discovery. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 322–333. [CrossRef]
3. 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. [CrossRef] [PubMed]
4. Cruz-Monteagudo, M.; Medina-Franco, J.L.; Perez-Castillo, Y.; Nicolotti, O.; Cordeiro, M.N.D.; Borges, F. Activity Cliffs in Drug Discovery: Dr. Jekyll or Mr. Hyde? *Drug Discov. Today* **2014**, *19*, 1069–1080. [CrossRef] [PubMed]
5. Wassermann, A.M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure-Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53*, 8209–8223. [CrossRef]
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. [CrossRef] [PubMed]
7. Vogt, M. Progress with Modeling Activity Landscapes in Drug Design. *Expert Opin. Drug Discov.* **2018**, *13*, 605–615. [CrossRef]
8. Shanmugasundaram, V.; Maggiora, G.M. Characterizing Property and Activity Landscapes Using an Information-Theoretic Approach. In Proceedings of the 222nd American Chemical Society National Meeting, Division of Chemical Information, Chicago, IL, USA, 26–30 August 2001; American Chemical Society: Washington, DC, USA,, 2001. Abstract no. 77.
9. Yongye, A.B.; Byler, K.; Santos, R.; Martínez-Mayorga, K.; Maggiora, G.M.; Medina-Franco, J.L. Consensus Models of Activity Landscapes with Multiple Chemical, Conformer, and Property Representations. *J. Chem. Inf. Model.* **2011**, *51*, 2427–2439. [CrossRef]
10. Agrafiotis, D.; Shemanarev, M.; Connolly, P.; Farnum, M.; Lobanov, V. SAR Maps: A New SAR Visualization Technique for Medicinal Chemists. *J. Med. Chem.* **2007**, *50*, 5926–5937. [CrossRef]

11. 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. [CrossRef]

12. 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. [CrossRef] [PubMed]

13. Maggiora, G.M. On Outliers and Activity Cliffs—Why QSAR often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535. [CrossRef] [PubMed]

14. 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. [CrossRef] [PubMed]

15. 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. [CrossRef] [PubMed]

16. Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure-Activity Relationships. *J. Med. Chem.* **2007**, *50*, 5571–5578. [CrossRef] [PubMed]

17. Guha, R.; Van Drie, J.H. Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646–658. [CrossRef]

18. Guha, R.; Van Drie, J.H. Assessing How Well a Modeling Protocol Captures a Structure-Activity Landscape. *J. Chem. Inf. Model.* **2008**, *48*, 1716–1728. [CrossRef]

19. Stumpfe, D.; Bajorath, J. Recent Developments in SAR Visualization. *Med. Chem. Commun.* **2016**, *7*, 1045–1055. [CrossRef]

20. Iqbal, J.; Vogt, M.; Bajorath, J. Activity Landscape Image Analysis Using Convolutional Neural Networks. *J. Cheminform.* **2000**, *12*, e34. [CrossRef]

21. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A.P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L.J.; Cibrián-Uhalte, E.; et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954. [CrossRef]

22. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]

23. Coz, J.J.D.; Díez, J.; Bahamonde, A.; Sañudo, C.; Alfonso, M.; Berge, P.; Dransfield, E.; Stamataris, C.; Zygoyiannis, D.; Valdimarsdottir, T.; et al. *Advances in Data Mining: Applications in Medicine, Web Mining, Marketing, Image and Signal Mining*; Perner, P., Ed.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 297–309.

24. Rogers, D.; Hahn, M. Extended-connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754. [CrossRef] [PubMed]

25. Rogers, D.J.; Tanimoto, T.T. A Computer Program for Classifying Plants. *Science* **1960**, *132*, 1115–1118. [CrossRef] [PubMed]

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

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

28. Culjak, I.; Abram, D.; Pribanic, T.; Dzapo, H.; Cifrek, M. A Brief Introduction to OpenCV. In Proceedings of the 35th International Convention MIPRO, Opatija, Croatia, 21–25 May 2012; pp. 1725–1730.

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

Summary

The analysis goes beyond the visual inspection and qualitative comparison of 3D ALs by assessing the ability to compare topological differences between 3D ALs quantitatively. A novel computational methodology is introduced to quantify (dis)similarity relationships between 3D ALs of varying SAR contents. The methodology maps the heatmap image representations of 3D ALs of target compound data sets onto evenly spaced grids with constant dimensions, which yields a constant number of cells. The features from each cell are then extracted based on color intensities and categorized on the basis of activity thresholds. The color intensities directly reflect the corresponding potency distribution that implicitly accounts for the spatial distribution of corresponding topographical features. Extracted features from 3D AL heatmaps can be quantified to measure (dis)similarity between different 3D ALs. The results obtained in the proof-of-concept investigating the comparison of categorized cell distributions based on 3D ALs of four target activity classes provided conceivable and meaningful quantitative dis(similarity) comparisons of 3D ALs with varying SARs . The methodology allows the numerical feature extraction of complex 3D ALs and further extends the utility of AL image representations for quantitative SAR exploration. 3D ALs successful quantification and conceivable quantitative comparison based on image data highly motivate the further investigation of 3D ALs to detect SAR topographical features.

Chapter 4

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

Introduction

3D ALs are readily convertible to heatmaps representing a top-down view with color profiles and topographical features conservation. Therefore, 3D AL heatmap representations are particularly intuitive for image-based SAR analysis. SAR topological features in 3D ALs resemble real-world geographical features. For example, small structural changes with small or gradual potency changes give rise to the formation of a smooth valley, whereas small structural changes with high potency differences form peaks or mountains. 3D ALs of most compound datasets are heterogeneous; therefore, smooth valleys and rugged regions often co-occur in 3D ALs that represent a heterogeneous SAR with the combination of continuous and discontinuous regions. The chapter presents a novel methodology to analyze 3D ALs based on topographical characteristics. The methodology seeks to identify different SAR characteristics based on 3D ALs topographical features (i.e., valleys, peaks) and quantifies the activity distribution localized within the identified topographical characteristics. The methodology, proof of concept and findings are discussed.

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

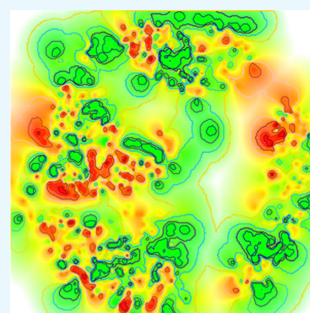Javed Iqbal, Martin Vogt, and Jürgen Bajorath*

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Visualization of structure−activity relationships (SARs) in compound data sets substantially contributes to their systematic analysis. For SAR visualization, different types of activity landscape (AL) representations have been introduced. Three-dimensional (3D) AL models in which an activity hypersurface is constructed in chemical space are particularly intuitive because these 3D ALs are reminiscent of "true" (geographical) landscapes. Accordingly, the topologies of 3D AL representations can be immediately associated with different SAR characteristics of compound data sets. However, the comparison of 3D ALs has thus far been confined to visual inspection and qualitative analysis. We have focused on image analysis as a possible approach to facilitate a quantitative comparison of 3D ALs, which would further increase their utility for SAR exploration. Herein, we introduce a new computational methodology for quantifying topological relationships between 3D ALs. Images of color-coded 3D ALs were converted into top-down views of these ALs. From transformed images, different categories of shape features were systematically extracted, and multilevel shape correspondence was determined as a measure of AL similarity. This made it possible to differentiate between 3D ALs in quantitative terms.

1. INTRODUCTION

Activity landscapes (ALs) are graphical representations that are designed to integrate structural and potency relationships between compounds sharing the same specific activity.[1,2] AL modeling enables the visualization and graphical analysis of structure−activity relationships (SARs) for different data sets.[2]

Over the years, a variety of AL representations of different designs and complexities have been introduced.[1−8] These include two-dimensional (2D) representations such as plots of pairwise compound similarity versus potency relationships[1,4] or annotated similarity-based compound networks.[2,9] In addition, three-dimensional (3D) AL views such as pairwise compound activity−property−similarity distributions[4] or maps reminding us of geographical landscapes[3,8,10] 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.[2] 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.[2] Most prominent peaks in discontinuous regions are termed activity cliffs[10,11] and formed by pairs or

groups of structural analogues with largest potency differences in compound data sets.[11]

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.[3,8] From experimental potency values, a coherent surface is interpolated and color-coded by potency, representing an activity surface.[3,8] 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,[12,13] subjective graphical assessment has principal limitations when comparing different ALs. This is the case because most compound data sets exhibit SAR heterogeneity, resulting
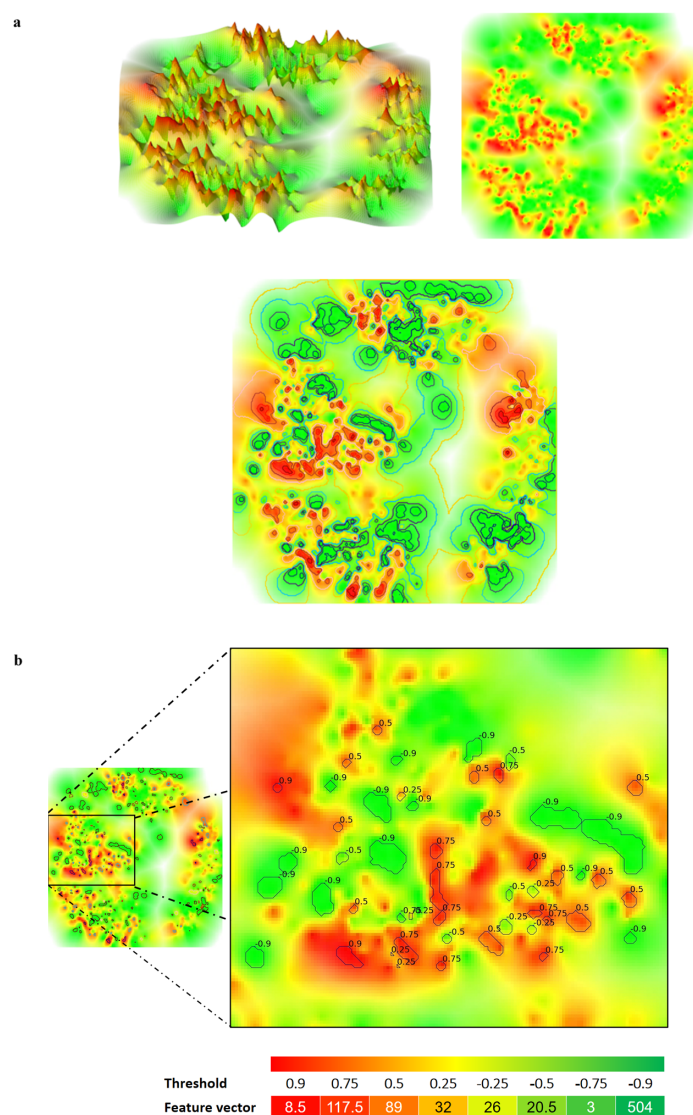
24111
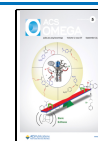
Figure 1. Activity landscape views and topological features. (a) At the top, an exemplary original 3D AL (left) and the corresponding heatmap (right) are shown. The heatmap conveys a top-down view of the AL. At the bottom, topological features extracted from the heatmap are depicted. For peaks and valleys, contours are drawn covering eight threshold levels (± 0.25, ± 0.5, ± 0.75, and ± 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.

from different combinations of locally continuous and discontinuous SARs that originate from different compound subsets.[12] 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.[14] 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.[14]

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
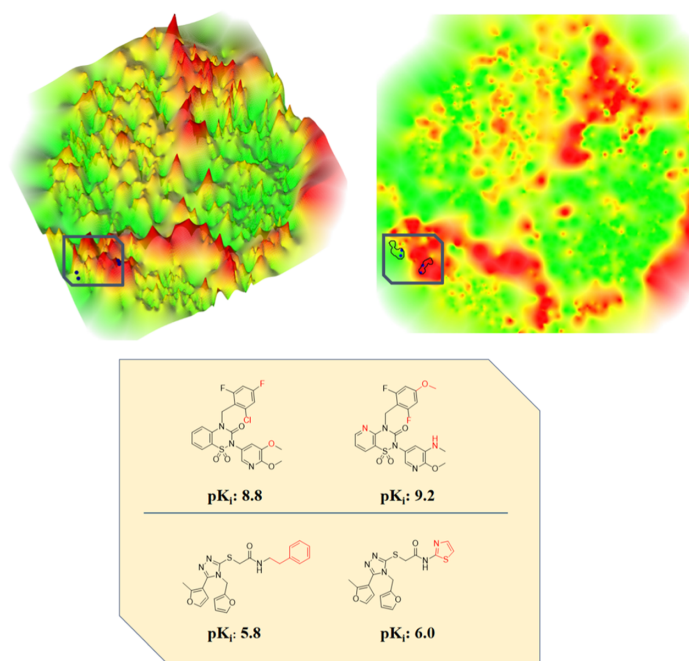
Figure 2. Compound mapping on activity landscape representations. Corresponding positions of highly and weakly potent orexin receptor type 2 antagonists belonging to different analogue series are mapped on the original 3D AL of the data set (top left) and the heatmap representation (top right), respectively. The pair of highly potent analogues and the pair of weakly potent analogues contribute to the formation of different activity cliffs within the same region of the AL. At the bottom, the structures of these compounds are shown with their logarithmic potency values. Chemical modifications distinguishing highly potent analogues (top) and weakly potent analogues (bottom) are highlighted (red).

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

63

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

Table 1. Compound Data Sets[a]

ChEMBL target ID	target name	compounds	potency [pK_i] min	potency [pK_i] 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	10
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

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

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

Table 2. Similarity-Based Ranking of Activity Landscapes[a]

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

[a]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.

seen, feature vector comparison yielded meaningful rankings of landscapes with a significant spread of pairwise similarity values
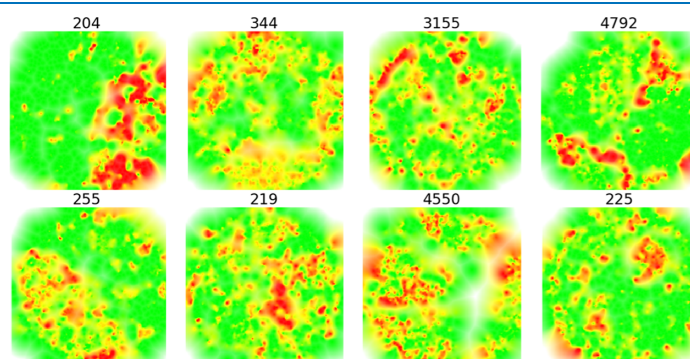

Figure 3. Heatmaps: for the eight activity classes, heatmaps derived from 3D AL images are shown. ChEMBL target IDs are reported according to Table 1.
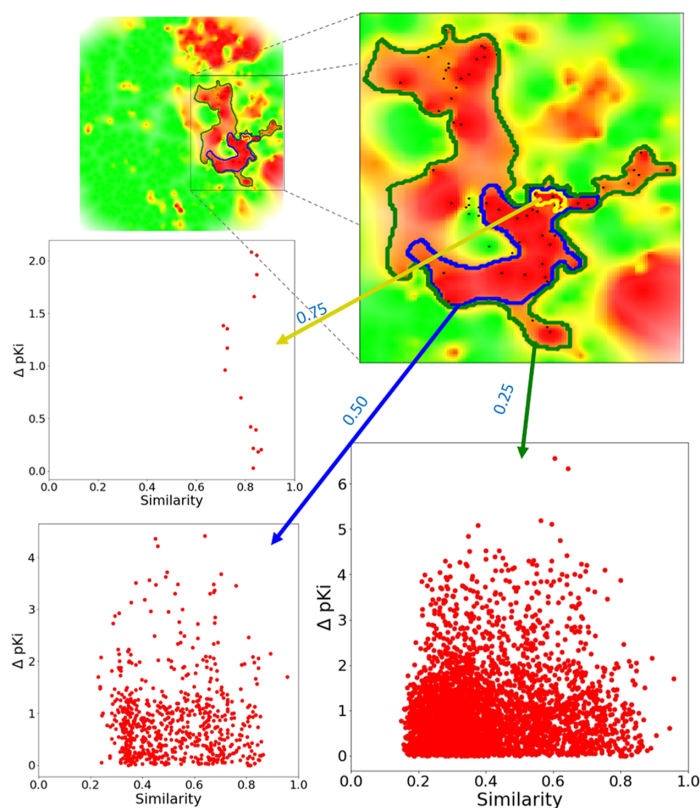
Figure 4. Feature-based structure−activity relationship information. In the heatmap of activity class 204, an exemplary peak region is contoured at different threshold levels of 0.75 (yellow), 0.50 (blue), and 0.25 (green) and a close-up view is shown. For compounds associated with the shapes detected at each contour level, structure−activity similarity (SAS) maps are shown capturing associated SAR information.

(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 pK_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 pK_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 pK_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 discerns 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.[3,8] Accordingly, chemical reference space was generated on the basis of ECFP4[18] 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 function based upon pairwise Tanimoto distances. MDS was previously found to be a preferred dimensionality reduction approach for retaining compound distances in original chemical reference (fingerprint) spaces. For the data sets used herein, the 2D projections preserved original ECFP4 Tanimoto distance relationships between compounds, with correlation coefficients between Tanimoto distances in fingerprint space and Euclidean distances in 2D projections of ∼0.7. The potency surface was interpolated via Gaussian process regression.[21] This approach interpolates intermediate values by a Gaussian process based upon prior covariances of experimental potency values. The "Sum of Matern and White" kernel[21] was used assuming a mean of zero to derive relationships between experimental data points (potency values), and Gaussian noise factors were applied to permit minor variations of z-values for points on the $x-y$ plane and optimize the global fit of the surface to experimental data points. Noise factors were regularized by optimizing the kernel's α parameter between 10^{-1} and 10^{-7} over 10 iterations. The potency gradient was applied to a limited pK_i range from 5.0 (green) over 7.0 (yellow) to 9.0 (red). Potency values larger than 9.0 were assigned to red.

3.2. Image Preprocessing. For each 3D AL, a heatmap was initially computed using the red, green, and blue (RGB) color model of openCV version 3.0 with eight bits per 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 MSA[15,23] 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 similarity[16]

$$J_w = \frac{\sum \min(a_i, b_i)}{\sum \max(a_i, b_i)}$$

The resulting coefficients for peaks and valleys were termed J_{wP} (A,B) and J_{wV} (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; 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

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.

■ 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. Consensus 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*, 5571−5578.

(13) Guha, R.; Van Drie, 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*, No. 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). *Biológia* **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 Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(19) Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants. *Science* **1960**, *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*, No. e453.

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

Summary

3D ALs visualize SAR characteristics as easily-comparable topographical features and benefit from their intuitive nature. The chapter introduced a novel methodology for a fast and robust quantitative comparison of 3D ALs based on topographical SAR characteristics. The methodology attempts to identify topographical features with the help of recent image processing approaches and demonstrates a successful comparison of a template 3D AL to a set of reference 3D ALs and provides a similarity ranking among them. The methodology detects the topographical features in 3D ALs that characterize varying SAR contents with the help of intensity-based thresholding and contouring of the thresholded image pixels. Furthermore, the methodology quantifies the localization of bioactivity distribution in identified topographical features numerically. The numerical analysis captures relative differences in the SAR information and discerns the similarity relationship between 3D ALs in a meaningful way. The finding complements quantitative SAR analysis and further increases the potential of 3D ALs for large-scale numerical SAR analysis.

Chapter 5

From Qualitative to Quantitative Analysis of Activity and Property Landscapes

Introduction

3D AL models provide an attractive way to visualize and explore SAR information in large chemical compound data sets. A variety of AL/PL representations are available for SAR analysis. 3D ALs were initially confined to qualitative SAR assessment, focusing on a better understanding of SAR characteristics and identification of key active compounds. However, in *Chapters 2-4* of this thesis, new analysis concepts have been introduced that focus on a quantitative SAR analysis. This chapter reviews the different design and complexity of ALs/PLs, the qualitative progression in analyzing ALs/PLs, and recent quantitative advancements in AL/PL analysis. *Chapters 2-4* of this thesis contributes to the review and discusses recent quantitative advancements focusing on numerical similarity quantification and comparison of distinct SARs based on 3D AL image data.

Viewpoint

From Qualitative to Quantitative Analysis of Activity and Property Landscapes

Gerald Maggiora, José L. Medina-Franco, Javed Iqbal, Martin Vogt, and Jürgen Bajorath*

Read Online

ACCESS | 📊 Metrics & More | 📖 Article Recommendations

ABSTRACT: Activity or, more generally, property landscapes (PLs) have been considered as an attractive way to visualize and explore structure−property relationships (SPRs) contained in large data sets of chemical compounds. For graphical analysis, three-dimensional representations reminiscent of natural landscapes are particularly intuitive. So far, the use of such landscape models has essentially been confined to qualitative assessment. We describe recent efforts to analyze PLs in a more quantitative manner, which make it possible to calculate topographical similarity values for comparison of landscape models as a measure of relative SPR information content.
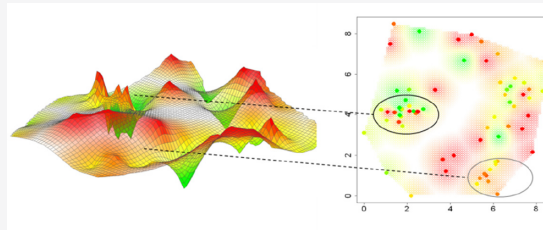
■ INTRODUCTION

Activity landscape (AL) models were originally introduced to aid in the analysis of structure−activity relationships (SARs) in compound data sets. AL representations complement the assessment of structural relationships with potency information and facilitate SAR visualization. As a compound structure-dependent feature, any numerically quantifiable experimental molecular property (such as, for example, physicochemical properties, metabolic stability, or toxicity) can be considered in lieu of biological activity. Accordingly, although ALs were introduced first, the underlying concept can be generalized and extended to property landscapes (PLs) of any type. In 1991, Lajiness, Maggiora, and colleagues proposed three-dimensional (3D) AL views reminiscent of geographical maps in which structural relationships between active compounds were represented as distances in a plane and potency values were added as a third dimension.[1,2] In 2001, Shanmugasundaram and Maggiora introduced a 2D representation comparing structure and activity relationships, the structure-activity similarity (SAS) map, which spawned a number of further AL/PL developments.[3] Then, in a seminal contribution published in 2006, Maggiora discussed multidimensional ALs to rationalize activity cliffs:[4]

A typical N-dimensional activity landscape is composed of an (N − 1)-dimensional chemical space; each dimension is described by a coordinate, which is generally defined by a single molecular descriptor or combination of descriptors. The Nth dimension is defined by the activity space that is derived from the measured activity of each of the assayed compounds. In three dimensions activity landscapes are closely akin to Nature's landscapes. For many years it has been assumed that similar molecules tend to have similar activities, leading to activity landscapes comparable to the gently rolling hills found on the Kansas prairie. Mounting evidence suggests, however, that this picture is not as universal as once thought but is in many cases rather more like the rugged landscapes of Utah's Bryce Canyon. This new topographical metaphor clearly implies that very similar molecules may in some cases possess very different activities leading to what can be called activity cliffs.

Maggiora

Such intuitive 3D ALs can be calculated for any given compound data sets employing a protocol involving chemical space projection and hypersurface interpolation from sparsely distributed compound potency values,[5] as further discussed below. Spatial features of 3D ALs, i.e., their topography or topology reflect SAR characteristics of compound data sets.

To describe AL features, the word topology is often formally incorrectly used, where the correct term is topography. Although topography describes geographical landscapes, it is a metaphor for other entities, such as the ALs and PLs described in this work. By contrast, topology is a field of mathematics (sometimes called "rubber sheet" geometry) that is concerned with the spatial properties of objects that are preserved under continuous deformations. Hence, in geography, topographically distinct landscapes may in many cases be topologically identical. In other fields such as landscape architecture, the term topology is used to describe spatial features.

Considering alternative design concepts, ALs were generally defined "... as any representation that integrates the analysis of the structural similarity of and potency differences between compounds sharing the same biological activity"[6] and this definition also applies to PLs accounting for other molecular properties. Recently, AL/PL analysis has been further extended through machine learning and image analysis. In the following, the evolution of the AL/PL concept and recent developments are discussed and key aspects of AL/PL modeling are highlighted.

■ DISCUSSION

Activity/Property Landscapes of Different Design and Complexity. A variety of AL/PL representations have been introduced including distinct 2D and 3D representations, as illustrated in Figures 1−3.

In the SAS map in Figure 1, structural similarity typically represents conventional (fingerprint-based) Tanimoto similarity while activity similarity between compound i and j is defined as follows:
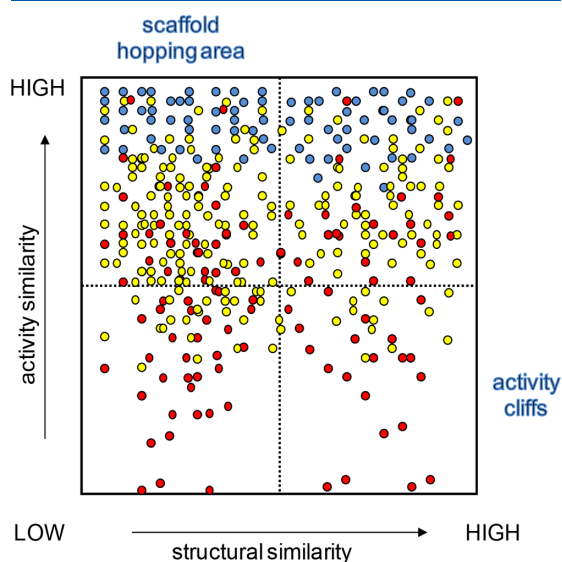

Figure 1. Alternative activity landscape views: SAS map. Structural and activity similarity are compared in a pairwise manner. Each dot corresponds to a pairwise compound comparison. Dots are color-coded to indicate whether the more active compound in a pair is highly (blue), intermediately (yellow), or weakly (red) potent [Reproduced from ref 6. Copyright 2010 American Chemical Society].

$$\text{sim}_{\text{act}}(i, j) = 1 - \frac{|P_i - P_j|}{P_{\text{max}} - P_{\text{min}}}$$

P_i and P_j give the potency of compounds i and j, respectively, and the term $P_{\text{max}} - P_{\text{min}}$ represents the difference between the maximum and minimum potency of the compound data set.

The upper-left section of the map contains compound pairs with low structural and high activity similarity (hence delineating a scaffold hopping region), whereas the upper-right section contains compound pairs with high structural and activity similarity (representing structural analogs conveying little SAR information). Similarly, compound pairs in the lower-left section share low structural and low activity similarity and are thus not SAR-informative. By contrast, compound pairs in the lower-right section have high structural similarity and low activity similarity (the latter corresponds to large potency differences); thus, they represent activity cliffs. The SAS map has been proven useful for many applications. Notably, Medina-Franco and colleagues have designed various extensions of the SAS map,[7−10] generalized the format for PL modeling,[8,9] for example, by generating SAS maps for exploring structure−cell permeability relationships,[8] and provided publicly available implementations.[10] Among others, specialized extensions of the SAS map include molecular representation (descriptor) invariant consensus AL models[7] and dual-activity difference (DAD) maps, in which differences for two specific activities are plotted and data points representing compound pairs are color-coded by similarity or target selectivity.[10] In addition, a 3D extension was introduced where similarity relationships for another molecular property was added as a third dimension to the SAS map.[7] Furthermore, Medina-Franco et al. also integrated AL analysis with other molecular modeling techniques to rationalize SARs.[11]

Figure 2 shows an annotated similarity-based compound network providing an alternative 2D AL model that is focused on individual compounds rather than pairs. This AL representation is useful for visualizing both global and local SAR characteristics of compound data sets. As illustrated, local SARs are centered on compound subsets that emerge as clusters in network-based AL models. Their analysis enables the identification and selection of SAR-informative compounds. SAR information content can be quantified globally and locally (on a per-compound basis) through the inclusion of numerical SAR analysis functions.[6] For example, Guha and Van Drie developed the structure−activity landscape index (SALI) and integrated the formalism with AL modeling.[12,13] SALI is a compound pair-based score defined as:

$$\text{SALI}(i, j) = \frac{P_i - P_j}{1 - \text{sim}(i, j)}$$

P_i and P_j represent the potency of compound i and j, respectively, and $\text{sim}(i, j)$ is the Tanimoto similarity of these compounds. The SALI score was designed to account for SAR discontinuity and identify activity cliffs at given score threshold levels. It can be used to construct an alternative network-based AL view in which compounds are represented as nodes that are connected by a directed edge (indicating increasing compound potency) if their pairwise SALI score exceeds a pre-defined threshold.[12,13] By increasing SALI threshold levels, the network is successively reduced and only limited numbers of edges remain that ultimately focus on largest-magnitude activity cliffs.

Maps of compound distributions in chemical features spaces can also be generated using generative topographic mapping
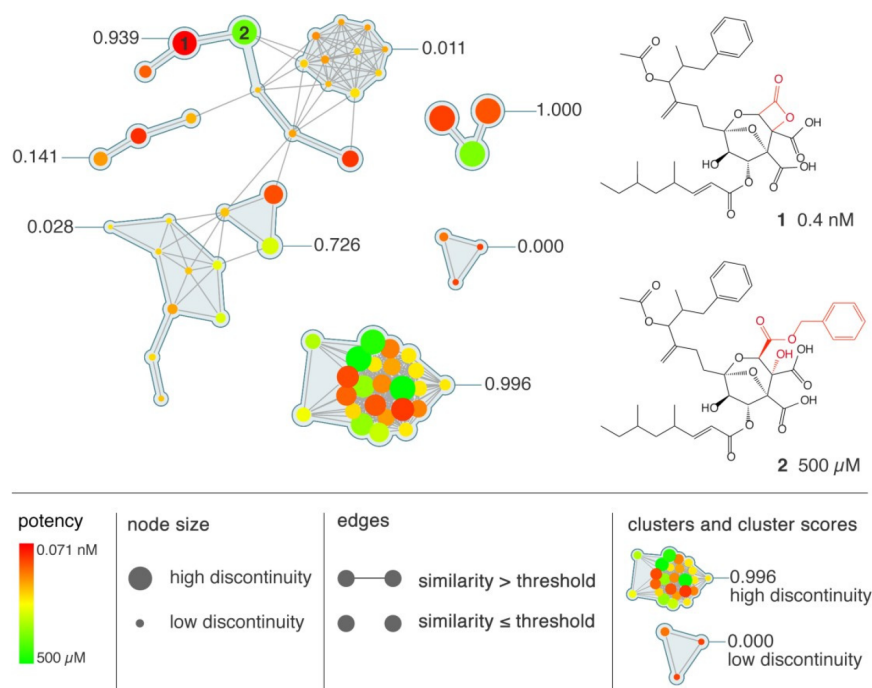
Figure 2. Alternative activity landscape views: SAR network. Annotated similarity-based compound network in which nodes represent individual compounds and edges pairwise similarity relationships (with respect to a Tanimoto coefficient threshold). Nodes are color-coded by compound potency and scaled in size according to a numerical SAR discontinuity score to account for the degree of local SAR discontinuity introduced by a compound, i.e., the discontinuity score component of the SAR index (SARI).[6] Accordingly, pairs of large red and green nodes represent activity cliffs. Distances between compounds and network components (clusters) are determined by a graphical layout algorithm (and thus have no chemical meaning) [Reproduced from ref 6. Copyright 2010 American Chemical Society].

(GTM), a dimension reduction technique related to self-organizing maps, which has been intensely investigated by Varnek and colleagues.[14,15] Such maps can be annotated with compound activity information or other molecular properties, yielding another variant of 2D PLs.[15] GTM-derived ALs can be generated for exceedingly large compound data sets, providing global low-resolution views of biologically relevant chemical space.

"Geographical" 3D ALs illustrated in Figure 3 are generated on the basis of a generally applicable modeling protocol.[5] Accordingly, compound data sets are placed within a given chemical feature space that is then subjected to dimensionality reduction. This yields a 2D projection in which inter-compound distances represent dissimilarities. Compound potency values are then added as a third dimension. A coherent activity surface is obtained by interpolating between these sparsely distributed values. The resulting surface is colored by potency using a continuous spectrum accounting for the potency range in the data set. Areas of the interpolated surface that lack compounds and associated potency values remain white. In the 3D AL model, smooth and gently sloped regions correspond to SAR continuity. Hence, in these regions, chemical modifications (i.e., "walks" in chemical space) lead to gradual potency changes, rendering corresponding compound subsets suitable for QSAR modeling and predictions. By contrast, rugged regions reflect SAR continuity where small chemical changes cause large potency effects, resulting in the formation of activity cliffs. While all of the AL/PL variants discussed above are informative, as well

as others not discussed herein, the especially intuitive nature of 3D models akin to natural landscapes renders them particularly attractive for SAR visualization and exploration, as so well expressed by Maggiora's "topographical metaphor" quoted above. Therefore, current studies have increasingly focused on 3D AL models, resulting in different types of applications and new analysis concepts.

Comparison of Corresponding Property Landscapes. The intuitive access to 3D models can be exploited, for example, by comparing different PLs, as illustrated in Figure 4. Landscapes accounting for various molecular properties can be generated on the basis of the same chemical space projection, thus enabling direct comparison. In the example presented below, a 3D AL is shown for a data set combining known ligase inhibitors with virtual candidate compounds having predicted potency values and compared to a corresponding 3D PL model with calculated ClogP values as an indicator of hydrophobicity.[16] The topography of the 3D AL is heterogeneous and reveals a number of activity cliffs. By contrast, the ClogP surface is overall smooth and reflects generally high compound hydrophobicity (as is often observed). However, small regions of low hydrophobicity are also detected. Corresponding regions in these models can be screened for candidate compounds having high potency and comparably low hydrophobicity.

For example, the compound displayed in Figure 4 maps to an activity cliff region and is predicted to have high potency. It also maps to a small cavity of low hydrophobicity in the 3D PL model on the right-hand side of the figure. Hence, this compound
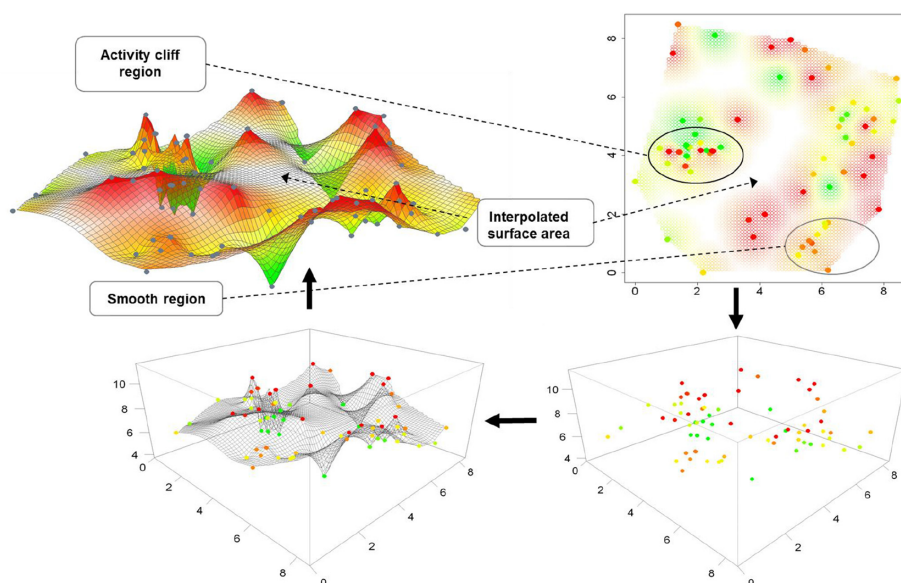
Figure 3. Alternative activity landscape views: three-dimensional representation. An exemplary 3D AL is shown (top left) and its generation is illustrated. Compounds are mapped into a chemical feature space and a 2D projection is generated (top right) in which decreasing Euclidean or Tanimoto distances between compounds (dots color-coded by potency) indicate increasing similarity. Compound potency values are added as a third dimension to the projection (bottom right). From sparsely distributed "potency points", a coherent "activity surface" is interpolated (bottom left) and color-coded according to compound potency values to yield the final 3D AL model. White surface areas are algorithmically interpolated in regions that do not contain compounds and hence lack potency information.
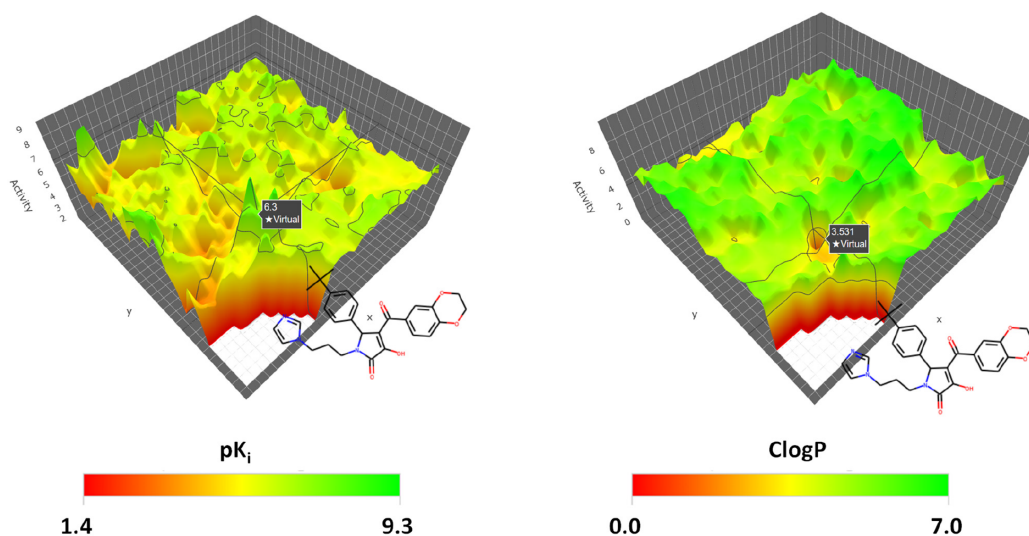

Figure 4. Application of 3D AL/PL models: different property landscapes. On the left, a 3D AL model is shown for known E3 ubiquitin-protein ligase inhibitors and newly designed candidate compounds. On the right, a corresponding ClogP landscape model is displayed. The position of a candidate compound is mapped [Reproduced from ref 16. Copyright 2019 American Chemical Society].

would represent an attractive candidate for identifying new inhibitors. Generating PL representations for combined sets of known active and virtual compounds also illustrates benefits of analyzing and prioritizing virtual candidates within the PL context of existing active compounds, for which 3D PLs provide a nearly ideal framework.

From Qualitative to Quantitative Assessment. Landscape modeling was intensely investigated about a decade ago, but over the past few years, only a limited number of new developments have been reported including various SAS map variants and GTM-based chemical space views, as discussed above. Numerical SAR functions have been applied in the context of AL analysis and used to design activity cliff-centric
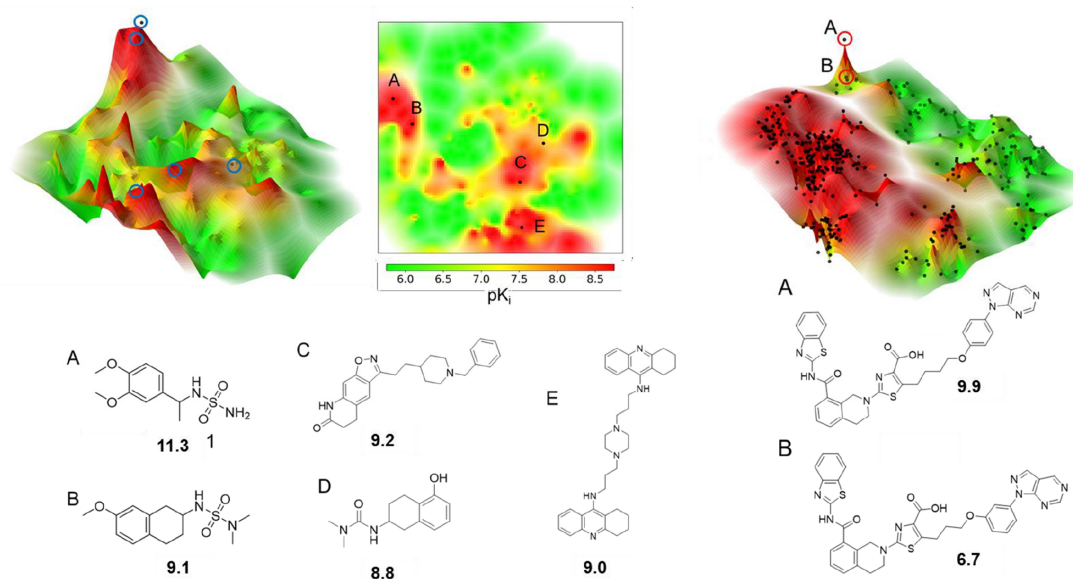
Figure 5. Application of 3D AL models: compound mapping. On the left and center of the figure, a 3D AL and its 2D projection are shown for a data set of acetylcholinesterase inhibitors to which five potent test inhibitors (A−E) were mapped. The positions of these inhibitors are indicated on the 2D map in the center of the figure and shown as blue circles on the 3D AL on the left-hand side. On the right-hand side of the figure, a 3D AL of Bcl-2 inhibitors is displayed. Many highly and weakly potent test compounds (indicated as black dots) were accurately mapped to corresponding landscape regions. In addition, two exemplary training set inhibitors (A and B) are shown that form a prominent activity cliff. Their positions in the 3D AL are indicated using red circles.

representations. On the other hand, 3DAL/PL representationshave mostly been analyzed and compared in a qualitative manner, the primary aim of which is gaining a better understanding of the characteristic features of data sets and the identification of key active compounds.[6] Recently, new analysis concepts have been introduced that provide a more quantitative assessment of 3D ALs.

Activity Landscapes and QSAR. Although not often considered, by design a 3D AL represents a nonlinear QSAR model for a given data set. In principle, this model should enable compound potency predictions. However, if a 3D AL representation is highly resolved, the model is, by its very nature, strongly overfit to the "training" data set it represents, calling into question whether 3D ALs could be used for potency predictions of external test compounds. Recently, the potential of 3D ALs for potency predictions has been investigated.[17] In this study, different dimensionality reduction and interpolation methods were evaluated for their ability to reproduce compound distances in the original fingerprint-based feature space in 2D projections of the original space and to accurately account for the potency distributions of the data set on the basis of the interpolated activity surface. The best overall results were obtained for 2D projections generated with Neuroscale, a neural network-based adaptation of multidimensional scaling, in combination with Gaussian process regression for surface interpolation.[17] The resulting 3D ALs had well-resolved topographies. Exemplary landscapes are shown in Figure 5.

Projection methods including explicit mapping functions such as Neuroscale generate compound coordinates in the 2D representations. This makes it possible to map test compounds from the original features space onto an interpolated surface. Moreover, Gaussian process regression predicts potency values

represented by the surface from coordinates in the 2D projections, providing a framework for QSAR. Accordingly, the mapping of compounds at different potency levels onto 3D ALs, representing a semi-quantitative exercise, and actual compound potency predictions were both investigated. Three-dimensional AL-based compound potency predictions were found to be, at best, comparable in overall accuracy to conventional regression models, but also failed in some instances, depending on the projection methods and data sets. These findings underscored the effect that intrinsic activity surface overfitting in 3D ALs with highly resolved topographies had on predicted activities. However, both highly and weakly potent test compounds were generally accurately mapped to the corresponding 3D AL regions of high and low potency, respectively,[17] as illustrated in Figure 5.

Thus, rather promising predictions were obtained at the semi-quantitative level. In practical applications, test compounds can be mapped to 3D ALs, and it can then be determined whether they fall into regions of high or low potency, smooth areas or activity cliff regions, thus integrating SAR visualization and qualitative assessment with comparison and prioritization of candidate compounds. Semi-quantitative compound mapping further extends the utility of 3D ALs. Although these pilot studies were confined to ALs, from a methodological point of view, the analysis can be readily extended to 3D PLs accounting for a variety of molecular properties, provided available experimental data are of sufficient quality.

Conceptually, 3D ALs are low-dimensional projections from an abstract high-dimensional chemical space that depend on the chosen feature representation and similarity metric. To further improve the predictive capabilities of 3D ALs, molecular feature representations should be optimized.[18] For instance, feature

selection methods have been successfully used in ligand-based virtual screening.[19] However, the typically heterogeneous nature of 3D ALs due to the presence of smooth as well as rugged regions including steep activity cliffs indicates that the underlying feature space for assessing chemical similarity accommodates potency effects in rather different ways, depending on the local neighborhood. Thus, while efforts to optimize molecular representations of 3D ALs are principally meaningful, the presence of SAR discontinuities and activity cliffs is a reality that cannot be avoided and fundamentally limits the capability of quantitative non-local models for potency predictions. Finding representations that render molecules dissimilar by increasing the resolution of features has a "smoothing" effect on 3D AL models due to a loss of SAR discontinuity information, thus leading to a departure from "SAR reality".

Landscape Image Analysis. Another recently introduced approach to 3D AL analysis involves the quantification of differences in topography between landscapes as a measure of varying SAR information content, which goes beyond of what can be learned by visual inspection. A key question is how a more quantitative assessment of topographical differences might be facilitated; a non-trivial task. We have reasoned that quantifying these differences might become feasible on the basis of AL image data. In principle, color-coded 3D ALs should be well suited for image processing and analysis. A proof-of-concept was established in an initial study where ensembles of 3D AL variants with defined topographical differences were generated, i.e., reference landscapes with either increased smoothness or ruggedness compared to an original model.[20] From these different variants, AL images with constant resolution (pixel density) were recorded from different perspectives (i.e., by varying azimuth and elevation angles). From standardized images, training and test sets of 3D ALs belonging to different topographical categories (i.e., original landscapes, smooth variants, or rugged variants) were assembled. Image feature representations were learned from pixel intensity values using a convolutional neural network (CNN) architecture with multiple convolutional layers for feature extraction. The CNN models preserved the dimensionality of color gradients and the extracted feature representations were found to yield generally accurate classifications of test images of the 3D AL variants.[20]

Topography-based classification of the 3D ALs was carried out using a form of numerical similarity analysis. Importantly, the topographical characteristics of 3D ALs are comprehensively accounted for by three variables including *distance* between points (i.e., compounds), *elevation*, and *color gradients*. Distance captures structural relationships and elevation as well as color gradients account for potency relationships. While elevation information, i.e., the third dimension of 3D ALs, is critical for their intuitive nature and interpretability, landscape topography is numerically captured in images by varying pixel intensities of color gradients, the presence of which plays into the strengths of image analysis. Moreover, pixel intensities also implicitly account for locality information because they result from the interpolation of color gradients accounting for potency variations among compounds that are close to each other in chemical space. Given this intrinsic redundancy in representing potency relationships, original 3D AL images were transformed into standardized heatmaps representing a top-down view of the 3D AL capturing color gradients, without detectable loss in topographical information These heatmaps provided the basis for different feature extraction approaches enabling the

generation of feature vectors for numerical similarity calculations,[21,22] as summarized in Figure 6.
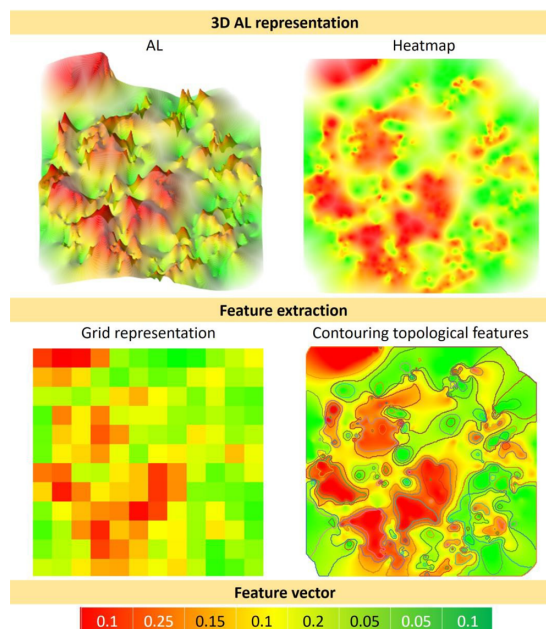

Figure 6. Application of 3D AL/PL models: image analysis and feature extraction. At the top, an image of an original 3D AL for a set of corticotropin-releasing factor receptor 1 ligands and the corresponding heatmap representing a top-down view of a landscape are shown. From the heatmap, different types of features are extracted including grid-based color intensity features or contour-based shape features, which are encoded in feature vectors for use in similarity calculations.

Color intensity features were extracted after superimposing heatmaps on a canonical grid (with 25 pixels per cell).[21] On the basis of a red−green (RG) color model accounting for the potency color code, cells were categorized based upon color intensity threshold intervals reflecting the spatial distribution of landscape features. For a given heatmap, mean pixel intensities per cell were determined and cells were assigned to different color intensity intervals. The distribution of cells over eight intensity intervals was then captured in feature vectors reporting the count of cells per category.[21] Of note, cell-based comparison of ALs images is orientation invariant (because neighborhood information is required in this case).

Alternatively, color shape features were extracted from contoured heatmaps following a two-step approach.[22] First, the canonical heatmap image was segmented into different regions by contouring lines of equal color intensity according to different threshold levels covering the entire range of lowest elevation ("valleys") to highest elevation ("peaks"). These contour lines represented non-intersecting closed curves that enclosed discrete sections (areas) of the landscape. Hence, higher-threshold contours were fully enclosed by lower-threshold contours. For feature definition, the heatmap was partitioned using positive threshold values representing peaks and numerically corresponding negative thresholds representing valleys. *Shapes* were then defined as groups of contours at increasing threshold levels. Accordingly, each shape represented

a threshold-specific contoured area. These shapes were then encoded feature vector representing different threshold intervals.[22]

Cell- or shape-based feature vectors were compared using different similarity measures. These calculations enabled the assessment of pairwise 3D AL similarity[21] or similarity-based ranking of 3D ALs relative to reference landscapes.[22] Similarity calculations on the basis of feature vectors produced generally meaningful results in instances where topographical differences could be well appreciated on the basis of visual inspection.[21,22] Of course, topographical similarity calculations become particularly informative when differences are increasingly difficult to appreciate.

CONCLUSIONS

Herein, we have briefly reviewed the AL/PL concept and its evolution and discussed recent scientific activities in this area. Over the years, much of the work has been focused on ALs, SAR visualization, and qualitative SAR assessment. However, at least a few studies have generalized these approaches to the study of PLs, taking different compound properties into consideration. Among alternative representations, 3D ALs/PLs reminiscent of natural landscapes are probably most intuitive and suitable for SAR visualization. While new developments in the area of PL design and analysis have been scarce for a number of years, recent research activities have concentrated on integrating PL assessment with other computational modeling methods and on applying or comparing 3D ALs in a more quantitative manner. For example, semi-quantitative mapping of active compounds to 3D ALs complements QSAR modeling and adds a visualization component. Furthermore, on the basis of image analysis, topographical similarity of 3D AL models can be quantified as an indicator of relative SAR content. The latter approach is albeit intellectually stimulating but still in its infancy. While a proof-of-concept has been established, practical utility still needs to be explored on a larger scale. For example, ranking large ensembles of 3D ALs according to their topographical similarity to reference landscapes and identifying compound data sets with similar SAR characteristics would create a bridge between SAR/SPR visualization and numerical analysis that would further extend the applicability domain and utility of AL/PL modeling.

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, Bonn D-53115, Germany;* ⊙ orcid.org/0000-0002-0557-5714; Phone: +49-228-2699-306; Email: bajorath@bit.uni-bonn.de; Fax: +49-228-2699-341

Authors

Gerald Maggiora − *University of Arizona BIOS Institute, Tucson, Arizona 85721-0240, United States*

José L. Medina-Franco − *Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico*

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

Martin Vogt − *Department of Life Science Informatics, B-IT LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Bonn D-53115, Germany;* ⊙ orcid.org/0000-0002-3931-9516

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jcim.0c01249

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Lajiness, M. Evaluation of the Performance of Dissimilarity Selection Methodology. In *QSAR: Rational Approaches to the Design of Bioactive Compounds*; Silipo, C., Vittoria, A., Eds.; Elsevier: Amsterdam, Netherlands, 1991; pp 201−204.

(2) Maggiora, G. M.; Shanmugasundaram, V.; Lajiness, M. S.; Doman, T. N.; Schulz, M. W. A Practical Strategy for Directed Compound Acquisition. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 317−332.

(3) Shanmugasundaram, V.; Maggiora, G. M. Characterizing Property and Activity Landscapes Using an Information-Theoretic Approach. *Proceedings of 222nd American Chemical Society National Meeting*; Division of Chemical Information, Chicago, IL, August 26−30, 2001; American Chemical Society: Washington, D.C., 2001; abstract no. 77.

(4) Maggiora, G. M. On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535−1535.

(5) 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.

(6) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations For Structure-Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53*, 8209−8223.

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

(8) Rojas-Aguirre, Y.; Medina-Franco, J. L. Analysis of Structure-CACO-2 Permeability Relationships Using a Property Landscape Approach. *Mol. Diversity* **2014**, *18*, 599−610.

(9) Medina-Franco, J.; 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.

(10) González-Medina, M.; Méndez-Lucio, O.; Medina-Franco, J. L. Activity Landscape Plotter: A Web-based Application for the Analysis of Structure-Activity Relationships. *J. Chem. Inf. Model.* **2017**, *57*, 397−402.

(11) López-López, E.; Rabal, O.; Oyarzabal, J.; Medina-Franco, J. L. Towards the Understanding of the Activity of G9a Inhibitors: An Activity Landscape and Molecular Modeling Approach. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 659−669.

(12) Guha, R.; Van Drie, J. H. Assessing How Well a Modeling Protocol Captures a Structure-Activity Landscape. *J. Chem. Inf. Model.* **2008**, *48*, 1716−1728.

(13) Guha, R. Exploring Structure-Activity Data Using the Landscape Paradigm. *WIREs Comput. Mol. Sci.* **2012**, *2*, 829−841.

(14) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. Stargate GTM: Bridging Descriptor and Activity Spaces. *J. Chem. Inf. Model.* **2015**, *55*, 2403−2410.

(15) Klimenko, K.; Marcou, G.; Horvath, D.; Varnek, A. Chemical Space Mapping and Structure-Activity Analysis of the ChEMBL Antiviral Compound Set. *J. Chem. Inf. Model.* **2016**, *56*, 1438−1454.

(16) Yoshimori, A.; Tanoue, T.; Bajorath, J. Integrating the Structure-Activity Relationship Matrix Method with Molecular Grid Maps and Activity Landscape Models for Medicinal Chemistry Applications. *ACS Omega* **2019**, *4*, 7061−7069.

(17) 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.

(18) Vogt, M. Progress with modeling activity landscapes in drug discovery. *Expert Opin. Drug Discovery* **2018**, *13*, 605−615.

(19) Vogt, M.; Bajorath, J. Similarity searching for potent compounds using feature selection. *J. Chem. Inf. Model.* **2013**, *53*, 1613−1619.

(20) Iqbal, J.; Vogt, M.; Bajorath, J. Activity Landscape Image Analysis Using Convolutional Neural Networks. *J. Cheminf.* **2020**, *12*, No. e34.

(21) Iqbal, J.; Vogt, M.; Bajorath, J. Computational Method for Quantitative Comparison of Activity Landscapes on the Basis of Image Data. *Molecules* **2020**, *25*, No. e3952.

(22) Iqbal, J.; Vogt, M.; Bajorath, J. Quantitative Comparison of Three-dimensional Activity Landscapes of Compound Data Sets Based upon Topological Features. *ACS Omega* **2020**, *5*, 24111−24117.

Summary

3D AL/PL models of heterogeneous SARs resemble real-world geographical landscapes with a combination of smooth and rugged regions. The intuitive nature of 3D AL/PL models makes them highly suitable for SAR visualization, and therefore 3D ALs have primarily been analyzed and compared qualitatively to analyze different SARs for the identification of key active compounds. The semi-quantitative investigation of 3D ALs for the prediction of potency values by mapping test compounds onto an interpolated 3D AL surface provided the first insights into 3D AL-based bioactivity prediction. In 3D AL models, continuous regions can be viewed as smooth and gently sloped regions, and discontinuous regions form as rugged regions. ACs are prominent features of SARs, and their presence in a SAR gives rise to the formation of steep activity peaks or deep activity valleys in a 3D AL model. In *Chapter 2* the successful discrimination of original 3D ALs from referenced ALs variants with increased smoothness and ruggedness provided the first clue to detecting unique elevation-based features from 3D AL image data. *Chapters 3-4* introduced further image analysis concepts that extract features from 3D AL based on texture and topographical features for the assessment of numerical similarity/dis-similarity between different 3D AL models. The applicability of new analysis concepts pushes the AL analysis beyond the visual inspection in a more robust and quantitatively meaningful way.

Having characterized activity landscape models, the focus of the subsequent chapters is on activity cliffs, which represent the most informative and striking features of ALs.

Chapter 6

Prediction of Activity Cliffs on the basis of Images Using Convolutional Neural Networks

Introduction

ACs are the prominent features of SARs. ACs tend to form deep valleys, steep peaks or rugged mountains in 3D ALs mirroring discontinuous SAR topological characteristics. *Chapter 3-4* have introduced novel methods to quantify the different SAR characteristics based on 3D AL image data. However, the application of image processing to learn structural context to distinguish compound pairs forming MMP-cliffs from non-ACs is still undefined. Therefore, the current chapter investigates applicability of image processing and CNNs to learn structural features from compound pairs forming MMP-cliffs and distinguish MMP-cliffs from non-ACs. Furthermore, the chapter discusses CNN modeling applicability for MMP-cliff analysis based on combined compound-pair image representation. Finally, the methodological details and analysis results are discussed.

Prediction of activity cliffs on the basis of images using convolutional neural networks

Javed Iqbal[1] · Martin Vogt[1] · Jürgen Bajorath[1]

Abstract

An activity cliff (AC) is formed by a pair of structurally similar compounds with a large difference in potency. Accordingly, ACs reveal structure–activity relationship (SAR) discontinuity and provide SAR information for compound optimization. Herein, we have investigated the question if ACs could be predicted from image data. Therefore, pairs of structural analogs were extracted from different compound activity classes that formed or did not form ACs. From these compound pairs, consistently formatted images were generated. Image sets were used to train and test convolutional neural network (CNN) models to systematically distinguish between ACs and non-ACs. The CNN models were found to predict ACs with overall high accuracy, as assessed using alternative performance measures, hence establishing proof-of-principle. Moreover, gradient weights from convolutional layers were mapped to test compounds and identified characteristic structural features that contributed to successful predictions. Weight-based feature visualization revealed the ability of CNN models to learn chemistry from images at a high level of resolution and aided in the interpretation of model decisions with intrinsic black box character.

Keywords Activity cliffs · Matched molecular pairs · Image analysis · Convolutional neural networks · Convolutional feature visualization

Introduction

In recent years, convolutional neural networks (CNNs) have gained increasing attention in chemical informatics and pharmaceutical research. For example, two-dimensional (2D) images of molecular graphs [1–5] and three-dimensional (3D) images of activity landscapes [6] have been used for deriving CNN models and extracting specific features from image data. For example, the Inception-ResNet v2 architecture was used to train CNN models on images from a large data set comprising 1.7 million compounds and predict physicochemical properties such as logP [1]. In addition, quantitative property predictions on the basis of compound images were reported using Chemception [2] and ChemNet [3]. Furthermore, the Toxic Colors approach [4] added atom labels, colored dots, and partial charge maps to image

representations for compound toxicity predictions while Kekulescope [5] only used Kekulé structures as input for compound potency and cell line toxicity predictions. Taken together, these investigations have indicated the potential of various CNN architectures to extract specified molecular features from 2D image representations and use these features for property predictions. Different from molecular structure-based approaches, 3D images of activity landscape variants were used for feature extraction and classification of landscape models according to structure–activity relationship (SAR) characteristics of the corresponding compound data sets [6].

While CNNs have thus far mostly been trained on 2D compound images, to our knowledge, they have not been used to process images of pairs of closely related compounds and predict differences in properties at the level of pairs. Activity cliffs (ACs) represent a prominent paradigm for compound pair-encoded property differences [7]. ACs are defined as pairs or groups of similar compounds or structural analogs with large differences in activity (potency) [7, 8]. Accordingly, ACs embody the pinnacle of SAR discontinuity, i.e., small chemical modifications leading to large potency alterations, and

✉ Jürgen Bajorath
bajorath@bit.uni-bonn.de

1   Department of Life Science Informatics, B-IT, LIMES
    Program Unit Chemical Biology and Medicinal
    Chemistry, Rheinische Friedrich-Wilhelms-Universität,
    Friedrich-Hirzebruch-Allee 6, 53115 Bonn, Germany

are a major source of SAR information [8]. An elegant formalism for the systematic identification of pairs of structural analogs is the matched molecular pair (MMP) concept and its algorithmic implementation [9]. An MMP is defined as a pair of compounds that share a common core structure and are only distinguished by a chemical modification at a single site (termed a chemical transformation) [9]. As such, MMPs are well suited for representing ACs, which has led to the introduction of MMP-cliffs [10]. An MMP-cliff is defined as an MMP formed by two compounds that are active against the same target and have a statistically significant difference in potency [10].

As a consistent molecular representation, MMP-cliffs have been used for predicting ACs at different levels. First, MMP-cliffs have been systematically distinguished from MMPs with only small or no potency differences using support vector machine classification on the basis of fingerprint representations and specialized compound pair-based kernel functions [11]. Subsequently, MMP-cliffs have also been successfully predicted in a methodologically simpler manner applying the condensed graph of reaction formalism [12]. In addition, potency differences encoded by MMPs have been quantitatively predicted using support vector regression [13]. To aid in the interpretation of machine learning models, fingerprint features determining correct AC predictions have been mapped back to the original compounds to delineate critically important substructures distinguishing MMP-cliffs from other MMPs [11].

Herein, we have attempted to predict MMP-cliffs from image data using CNNs. In addition to assessing classification performance, we have made use of recent developments in convolutional layer visualization [14–17] to identify and display key features contributing to correct AC predictions. Our proof-of-concept investigation further extends the current spectrum of molecular image-based modeling in chemical informatics.

Material and methods

Compound activity classes

From ChEMBL (version 26) [18], three compound activity classes with available high-confidence activity data were extracted. Compounds were tested against single human targets in direct interaction assays at highest assay confidence (ChEMBL confidence score 9). As potency measurements, assay-independent equilibrium constants (pK_i values) were required. Multiple measurements for the same compound were averaged, provided all values fell within the same order of magnitude; otherwise, the compound was disregarded. Table 1 reports the targets and composition of these activity classes.

Matched molecular pairs and activity cliffs

For activity classes, all possible MMPs were generated by systematically fragmenting individual exocyclic single bonds and sampling core structures and substituents in index tables [9]. For substituents, size restrictions were applied to limit MMP formation to typically observed structural analogs [10]. Accordingly, a substituent was permitted to contain at most 13 non-hydrogen atoms and the core had to be at least twice as large as the substituent. Additionally, for MMP compounds, the maximum difference in non-hydrogen atoms between the substituents was set to eight, yielding transformation size-restricted MMPs [10].

An MMP qualified as an MMP-cliff if the two structural analogs had an at least 100-fold difference in potency ($\Delta pK_i \geq 2.0$) [10]. To avoid potency difference-dependent boundary effects in AC prediction, compounds forming a non-AC MMP were restricted to an at most tenfold difference in potency. Furthermore, to balance structural heterogeneity of large activity classes originating from different sources, MMPs were only retained if their compounds and core structures were found in multiple MMPs. Table 1 reports MMP and MMP-cliff statistics for the activity classes.

Table 1 Activity classes

Target name	ChEMBL target ID	MMP-cliffs			Non-AC MMPs		
		MMPs	Unique cores	Unique substituents	MMPs	Unique cores	Unique substituents
Thrombin	204	456	61	168	3595	554	567
Tyrosine kinase Abl	1862	1122	37	251	6143	322	419
Mu opioid receptor	233	466	114	286	9712	1230	959

For each activity class, the total number of compounds, MMP-cliffs, non-AC MMPs, unique core structures, and substituents are reported

Molecular image representations

Each MMP core and the associated substituents were treated as separate molecular objects using the RDKit application programming interface (API) [19]. For each unique core and substituent, high-resolution portable network graphics (PNG) compound images with 500×500 pixels were generated using the RDkit Chem.Draw package (version 2020.03.5) [19]. In images, substituent attachment sites were replaced with an asterisk symbol. To represent an MMP core and the two substituents defining the transformation in combined form, core and substituent images were resized to 300×300 pixels and then horizontally concatenated in a single image of dimensions $300 \times 900 \times 3$ (height $\times$ width $\times$ color-channels). Figure 1 illustrates MMP image generation. The pixel values of all image matrices were converted into 32-bit floating point format and normalized. Images were processed using openCV (version 4.4.0) [20–22].

Convolutional neural network architecture

Figure 2 shows the CNN architecture designed for image analysis, consisting of convolutional, pooling, dropout, and dense layers. Two convolutional layers with kernel size of 32 and respective filter sizes of 3×3 and 5×5 were used to extract key features from MMP images. The convolutional layers were followed by a pooling, dropout, and dense layer. Max-pooling was used as pooling layer to compute the maximum value in each patch of each convolved feature map. A dropout layer was added to avoid overfitting. After 'flattening' the weights, the softmax function was applied to normalize learned weights and yield a probability distribution. CNN layers were implemented using TensorFlow (version 2.2.0) [23] and Keras (version 2.2.4) [24].

Performance measures

CNN models were trained to systematically distinguish between MMP-cliffs and corresponding non-AC MMPs. The classification performance of CNN models was evaluated using receiver-operator characteristic (ROC) curves
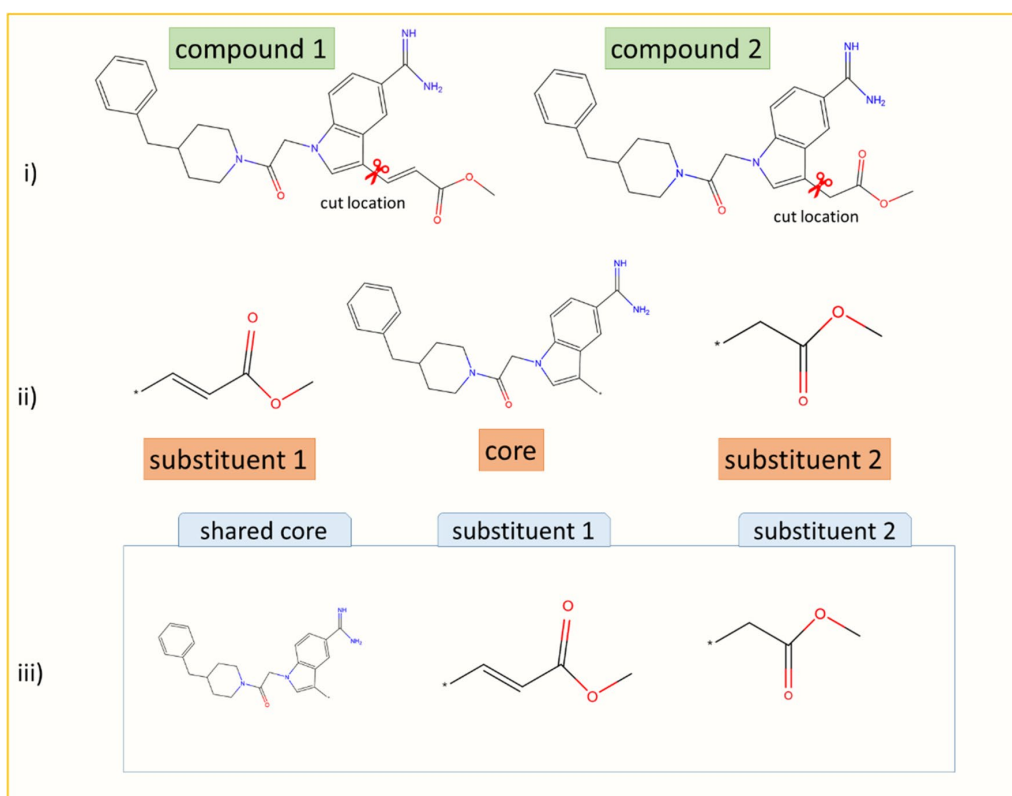

Fig. 1 Generation of MMP images. **i** Two compounds forming an MMP are shown. **ii** The common core and the two substituents defining the chemical transformation are displayed. **iii** Separate core and substituent images are horizontally concatenated yielding a single image
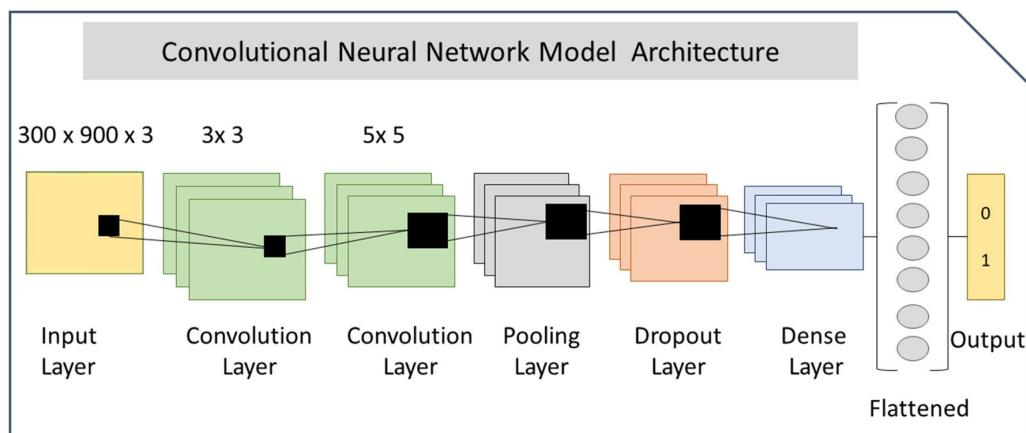
Fig. 2 The CNN model architecture used for MMP image analysis is shown

and the area under the ROC curve (AUC). In addition, model performance was assessed with four performance measures including overall accuracy (A), balanced accuracy (BA), weighted mean F1 score [25], and Mathews correlation coefficient (MCC) [26], defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$F1 = 2 \times \frac{TP}{2TP + FP + FN}$$

$$ROCAUC = \frac{1}{2} - \frac{1}{2}\frac{FP}{FP + TN} + \frac{1}{2}\frac{TP}{TP + FN}$$

$$A = \frac{TP + FN}{TP + TN + FP + FN}$$

$$BA = \frac{\left(\frac{TP}{TP+FN}\right) + \left(\frac{TN}{TN+FP}\right)}{2}$$

TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives respectively.

Convolutional layer feature visualization

Spatial information from the convolutional layers of trained models was extracted using the Grad-Cam algorithm [17]. Channel-based mean values of the resulting convolutional feature map activation weights were mapped to the original image for feature visualization.

Results and discussion

Convolutional neural network models

CNN models were derived to distinguish between MMP-cliffs and non-AC MMPs on the basis of molecular images generated for three distinct activity classes including thrombin inhibitors (target/activity class ID 204), tyrosine kinase Abl inhibitors (class 1862), and mu opioid receptor ligands (class 233). As shown in Fig. 1, MMP images for CNNs combined the shared core structure with the pair of substituents representing the chemical transformation. Images of the three structures constituting an MMP were concatenated horizontally to obtain a single image. In contrast to displaying two compounds forming an MMP side-by-side, this image format contained no redundant substructure (duplicated core).

CNN models were separately trained in 10 independent trials on a set of 4050–10,178 images, dependent on the activity class. Training images were obtained by randomly selecting half of the MMP-cliffs per class (228–561; Table 1) and half of the non-AC MMPs (1797–4856). The resulting models were then tested on the remaining half of the MMP-cliff and non-AC MMP images. ROC curves for the best performing individual classification models are shown in Fig. 3. These CNN models yielded accurate AC predictions, with ROC-AUC values of 0.97 (204), 0.93 (233), and 0.92 (1862). In addition, Table 2 reports the mean prediction accuracy of the CNN models for each activity class on the basis of alternative performance measures. Although training and test sets were imbalanced, i.e., they containing many more non-AC
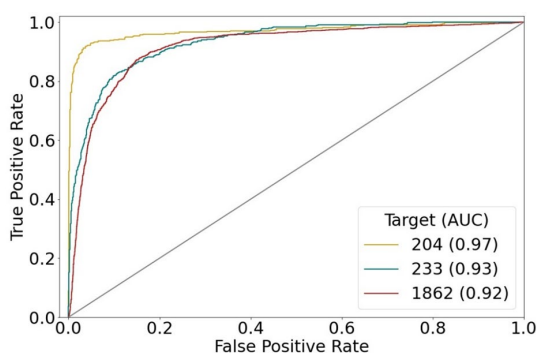
Fig. 3 ROC curves. The performance of the best CNN prediction models is monitored in ROC curves. For each curve (activity class, indicated by target ID), the resulting AUC is reported

MMPs than MMP-cliffs, the predictions were generally stable (i.e., yielding very low standard deviations) and consistently successful on the basis of all performance measures. Overall, CNN classification accuracy was highest for thrombin inhibitors (class 204), with mean AUC of 0.97 (AUC = 0.97), F1 = 0.85, MCC = 0.83, A = 0.97, and BA = 0.90, followed by class 1862 and 233. Although AUC and A values were also high for class 233 (0.92 and 0.96, respectively), predictions for this class yielded lowest F1 = 0.36, MCC = 0.39, and BA = 0.63 values, indicating that the majority class (non-AC MMPs) was predicted here more accurately than the minority class (MMP-cliffs). In this case, only < 5% of all MMPs represented MMP-cliffs. Thus, these results were expected. The use of balanced training sets would likely further increase prediction accuracy, which is meaningful from a machine learning perspective. However, for AC predictions, balancing MMP-cliff and non-AC MMP training sets would represent an unrealistic scenario because ACs are generally rare among qualifying compound pairs [8]. Regardless, even in the presence of class label imbalance, image-based classification of MMP-cliffs vs. non-AC MMPs was overall surprisingly accurate, more so than we anticipated.

Image feature visualization

Convolutional features naturally retain spatial information, which is lost in fully-connected layers. Therefore, the Grad-Cam algorithm was applied to visualize convolutional layer activation weights [17]. Figures 4, 5 and 6 show examples of original images onto which channel-based mean values of activation weights of the corresponding convolutional feature map were superimposed. All MMPs shown in Figs. 4 and 5 were correctly predicted while Fig. 6 also shows a false positive prediction. Visualization of convolutional layers revealed that most of the key image features were captured by the first convolutional layer. However, in a number of instances, the second convolutional layer was also capable of extracting and emphasizing key features, as shown in Fig. 7. Accordingly, addition of the second convolution layer typically further improved classification accuracy.

Learning structural features from compound images

The convolutional layer weights of the best performing CNN model (class 204) for MMP images from the test set were systematically extracted and visualized. A compelling observation was that weights from the CNN models detected specific structural features in MMP images. For example, convolutional layers were capable of recognizing primary, secondary, and tertiary amines as well as various ring structures. Moreover, the model was able to differentiate between substituents with different structures. In Fig. 4, the model distinguished between ring and aliphatic substituents, which is clearly evident by comparing mapped convolutional layer weights. Different weight distributions led to accurate predictions of MMP-cliffs and non-AC MMPs with very high probabilities of at least 94%. Furthermore, the model learned to differentiate between alternative cyclic structures, hence accounting for molecular topology.

In Fig. 5, the CNN model assigned high weights to secondary and tertiary amines in rings of substituents of correctly predicted MMP-cliffs and non-AC MMPs. Notably, the presence of different amines was a characteristic feature of all MMPs originating from class 204. However, by comparing the MMP-cliff and non-AC MMP in Fig. 5b and c, respectively, it becomes clear that detecting a tertiary amine
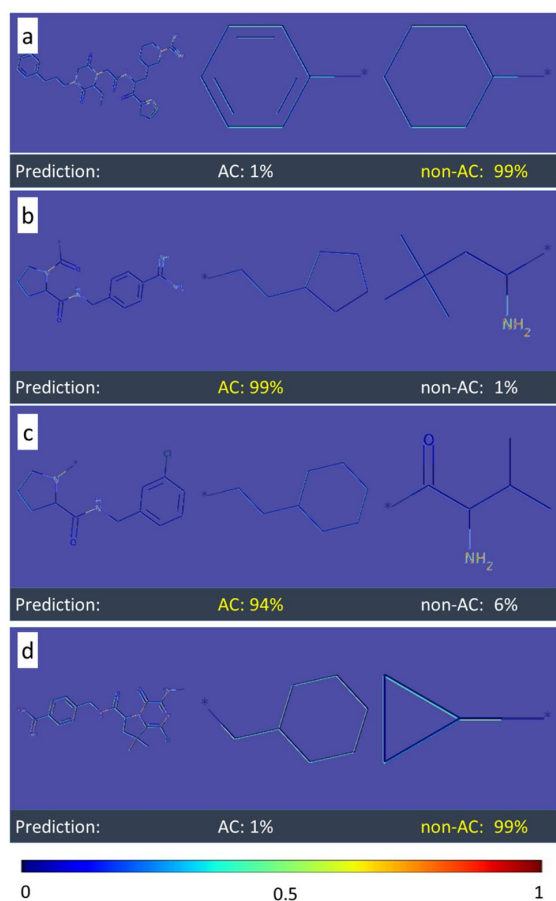
Table 2 Mean prediction accuracy

Target	AUC	F1	MCC	Accuracy	
				A	BA
204	0.97 ± 0	0.85 ± 0.02	0.83 ± 0.02	0.97 ± 0	0.90 ± 0.02
1862	0.92 ± 0.01	0.54 ± 0.08	0.50 ± 0.07	0.88 ± 0.01	0.70 ± 0.05
233	0.92 ± 0.02	0.36 ± 0.10	0.39 ± 0.08	0.96 ± 0	0.63 ± 0.05

For MMP-cliff/non-AC MMP classification models, the mean AUC, F1, MCC, global accuracy (A) and balanced accuracy (BA) values ± standard deviations over 10 independent trials are reported

Fig. 4 Mapping of activation weights. For four exemplary MMPs from class 204, mean gradient weights of the first convolutional layer are displayed on the respective structures and color-coded according to the given continuous color spectrum. Classification probabilities for each class (AC, non-AC) are given (%) and the correct class label of each MMP is colored in yellow. Shown are **a** a non-AC MMP with phenyl and cyclohexyl substituents, **b** and **c** MMP-cliffs with similar core structures and substituents, and **d** a non-AC MMP with different aliphatic ring substituents

Fig. 5 Mapping of activation weights. For three similar MMPs from class 204, mean gradient weights of the first convolutional layer are displayed. The representation is according to Fig. 4. **a** and **b** show MMP-cliffs and **c** shows a non-AC MMP. Highly weighted secondary or tertiary amines are encircled

Fig. 6 Mapping of activation weights. For three MMPs from class 204, mean gradient weights of the first convolutional layer are displayed. The representation is according to Fig. 4. **a** and **b** show a correctly predicted MMP-cliff and non-AC MMP, respectively. Highly weighted primary or secondary amines are encircled. **c** shows a false positive MMP-cliff prediction. Highly weighted primary and secondary amines are shared by the substituents. The distinguishing single and double bonds were detected with medium weights and are encircled

alone was not sufficient to distinguish between the MMP-cliff and non-AC MMP because this feature was shared by both. In this case, the core structures of these MMPs were distinct and different core features were detected and assigned high weights, hence illustrating core contributions to accurate predictions. In the MMPs shown in Fig. 6a and b, primary and secondary amines were also detected as distinguishing features in aliphatic substructures. Furthermore, Fig. 6c reports a false positive MMP-cliff prediction. On the basis of the MMP alone, this prediction error cannot be rationalized. To these ends, weights in similar MMPs with different class labels must be compared, as illustrated in Fig. 5. Nonetheless, this example was interesting because
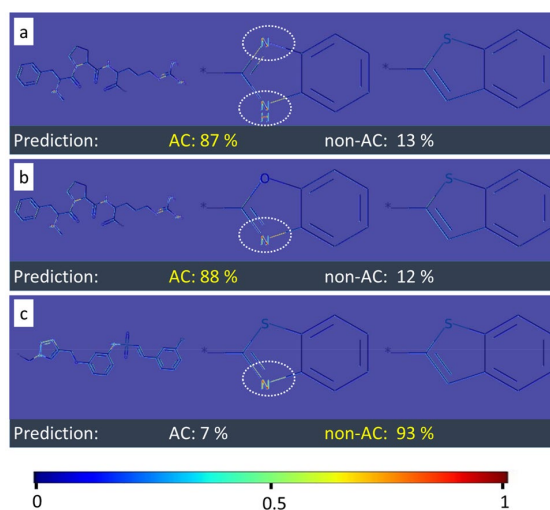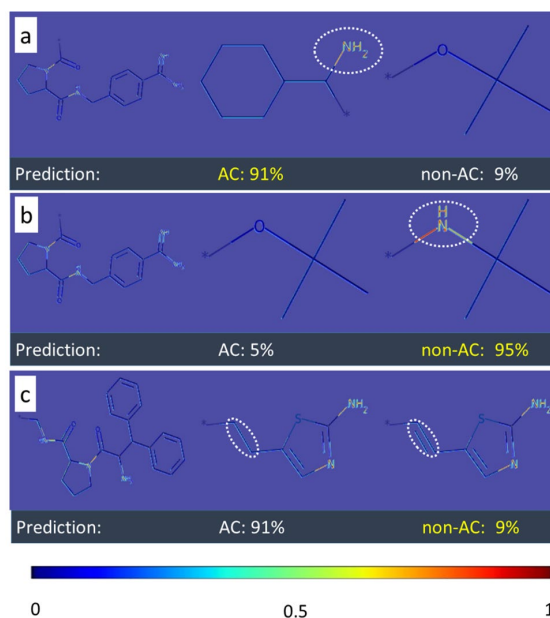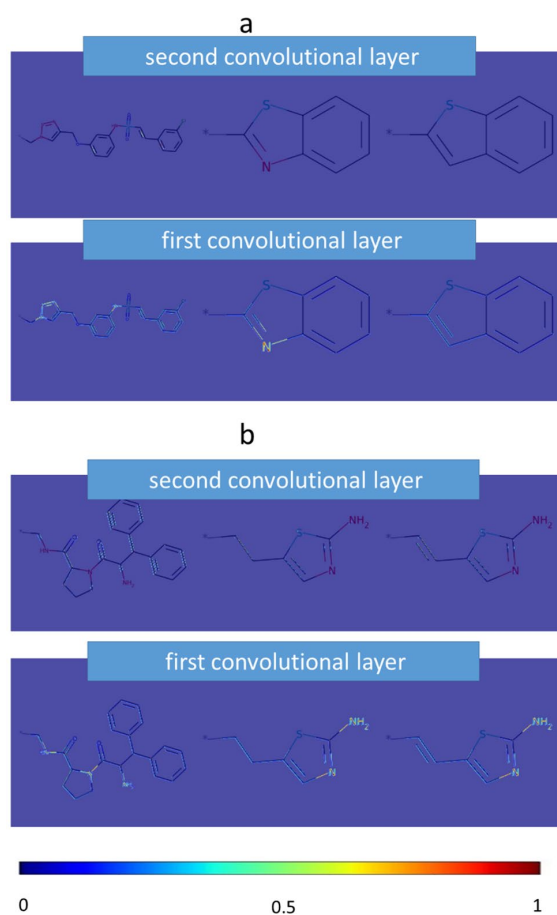
Fig. 7 Mapping of first and second layer activation weights. **a** and **b** show the MMPs from Figs. 5c and 6c, respectively, with mean gradient weights of the first and second convolutional layer. Here, weights from the second convolution layer reinforced structural features detected by first layer weights and also identified additional features

the replacement of a single bond with a double bond, i.e., a change in bond order representing a minute chemical modification at the level of images, was detected with medium weights as a distinguishing substituent feature.

Taken together, these convolutional layer weight-based visualizations demonstrated the capacity of the CNN model to detect signature features of compounds from a given activity class (such as the presence of various amines) as well as specific chemical features that distinguished cores and/or substituents of MMPs, including different ring structures, individual functional groups, or bond orders. Mapping weights from different convolutional layers often further emphasized such features or identified additional ones, as illustrated in Fig. 7. The correct detection of specific features distinguishing MMPs with different class labels

provided a rationale for the overall accuracy of the AC predictions. Differences between substituents detected by the CNN model can be analyzed at the level of individual MMP images, while understanding differently weighted core features requires comparisons of multiple MMPs. Visualization of key features in MMP cores and substituents aids in the interpretation of CNN model decisions that typically have black box character, hence improving model accessibility.

Conclusion

In the work, we have attempted the prediction of MMP-cliffs, which are an intuitive AC representation, on the basis of MMP image data using CNN models. To our knowledge, these are the first molecular image-based property predictions at the level of compound pairs. In our proof-of-concept investigation, encouraging accuracy was achieved in systematically distinguishing between MMP-cliffs and non-AC MMPs. While ACs were successfully predicted before using other machine learning approaches, we have been particularly interested in the question whether CNNs are capable of extracting chemical features and small feature differences from images of pairs of structural analogs that correctly distinguish between SAR continuity (embodied by non-ACs) and discontinuity (ACs). Mapping of convolutional layer weights to test compounds and visualizing corresponding structural features put the analysis on a level beyond statistical assessment of prediction accuracy. Visualization revealed the ability of CNN models to detect specific chemical features including distinct substructures and individual functional groups that distinguished structural analogs or MMPs with different properties. Thus, the models were capable to learn chemistry from MMP images, which resulted in successful AC predictions.

References

1. Szegedy C, Ioffe S, Vanhoucke V, Alemi A (2016) Inception-v4, Inception-ResNet and the impact of residual connections on learning. arXiv:1602.07261

2. Goh GB, Siegel C, Vishnu A, Hodas NO, Baker N (2017) Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. arXiv:1706.06689

3. Goh GB, Vishnu A, Siegel C, Hodas N (2018) Using rule-based labels for weak supervised learning: a ChemNet for transferable chemical property prediction. In: Proceedings of the ACM SIG-KDD international conference on knowledge discovery and data mining

4. Fernandez M, Ban F, Woo G, Hsing M, Yamazaki T, LeBlanc E, Rennie PS, Welch WJ, Cherkasov A (2018) Toxic Colors: the use of deep learning for predicting toxicity of compounds merely from their graphic images. J Chem Inf Model 58:1533–1543

5. Cortés-Ciriano I, Bender A (2019) KekuleScope: prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images. J Cheminform 11:e41

6. Iqbal J, Vogt M, Bajorath J (2020) Activity landscape image analysis using convolutional neural networks. J Cheminform 12:e34

7. Maggiora GM (2006) On outliers and activity cliffs—why QSAR often disappoints. J Chem Inf Model 46:1535–1535

8. Stumpfe D, Bajorath J (2012) Exploring activity cliffs in medicinal chemistry. J Med Chem 55:2932–2942

9. Hussain J, Rea C (2010) Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. J Chem Inf Model 50:339–348

10. Hu X, Hu Y, Vogt M, Stumpfe D, Bajorath J (2012) MMP-Cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs. J Chem Inf Model 52:1138–1145

11. Heikamp K, Hu X, Yan A, Bajorath J (2012) Prediction of activity cliffs using support vector machines. J Chem Inf Model 52:2354–2365

12. Horvath D, Marcou G, Varnek A, Kayastha S, de la Vega de León A, Bajorath J (2016) Prediction of activity cliffs using condensed graphs of reaction representations, descriptor recombination, support vector machine classification, and support vector regression. J Chem Inf Model 56:1631–1640

13. de la Vega de León A, Bajorath J (2014) Prediction of compound potency changes in matched molecular pairs using support vector regression. J Chem Inf Model 54:2654–2663

14. Griffin G, Perona P (2008) Learning and using taxonomies for fast visual categorization. In: 2008 IEEE conference on computer vision and pattern recognition. pp 1–8

15. Mahendran A, Vedaldi A (2016) Visualizing deep convolutional neural networks using natural pre-images. Int J Comput Vis 120:233–255

16. Nguyen A, Yosinski J, Clune J (2016) Multifaceted feature visualization: uncovering the different types of features learned by each neuron in deep neural networks. arXiv:1602.03616

17. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 ieee international conference on computer vision (ICCV). pp 618–626

18. Gaulton A, Hersey A, Nowotka ML, Patricia Bento A, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrian-Uhalte E, Davies M, Dedman N, Karlsson A, Magarinos MP, Overington JP, Papadatos G, Smit I, Leach AR (2017) The ChEMBL database in 2017. Nucleic Acids Res 45:D945–D954

19. Landrum G RDKit: open-source cheminformatics. https://www.rdkit.org. Accessed 19 Jan 2021

20. Culjak I, Abram D, Pribanic T, Dzapo H, Cifrek M (2012) A brief introduction to OpenCV. In: MIPRO 2012—35th international convention on information and communication technology, electronics and microelectronics—proceedings. pp 1725–1730

21. OpenCv (2014) OpenCV library. https://opencv.org. Accessed 19 Jan 2021

22. Bradski G (2000) The OpenCV library. Dr Dobb's J Softw Tools 25:120–125

23. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X. (2016) TensorFlow: a system for large-scale machine learning. In: 12th USENIX Symposium on operating systems design and implementation (OSDI 16), Savannah, GA

24. Chollet F (2015) Keras. https://github.com/keras-team/keras. Accessed 19 Jan 2021

25. Chinchor N (1992) MUC-4 evaluation metrics. In: Proceedings of the 4th conference on message understanding. Association for Computational Linguistics, USA. pp 22–29

26. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405:442–451

Summary

The CNN modeling applicability for the MMP-cliff classification prediction based on compound-pair image data was investigated. The CNN modeling potential to automatically learn essential chemical structural features from pairs of structural analogs (MMP-cliffs) and the correct classification prediction of MMP-cliffs based on chemical structural features was analyzed. CNN models demonstrated their capability to learn chemical structural features from compound images and successfully classify MMP-cliffs with encouraging accuracy. Furthermore, the mapping analysis of the convolutional layer weights onto test compounds confirmed the ability of CNN models to detect specific chemical features. CNN models revealed their potential to learn distinct substructures and individual functional groups to distinguish MMP-cliffs from non-ACs. The discrimination of MMP-cliffs from non-ACs based on image representations indicated CNNs classification prediction capability by learning chemistry from compound images.

Chapter 7

Learning Functional Group Chemistry from Molecular Images Leads to Accurate Prediction of Activity Cliffs

Introduction

Successful learning of specific chemical structural information from images, as discussed in *Chapter 7*, was an encouraging finding. Therefore this chapter further investigates CNN models to discover whether they have the potential to learn functional group chemistry from compound images and discriminate between MMP-cliffs from non-ACs with the help of functional group knowledge. Deep CNN models often require large amounts of data in order to learn structural features and patterns, whereas compound target data sets lack the availability of information of untested compounds; hence often encounter the problem of data insufficiency. The current chapter applies the transfer learning approach to functional group chemistry for MMP-cliff classification prediction. CNN models were trained using condensed graph of reaction (CGR) image representations of compound pairs. Calculations and results are discussed.

Research Article

Learning functional group chemistry from molecular images leads to accurate prediction of activity cliffs☆

Javed Iqbal, Martin Vogt, Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Friedrich-Hirzebruch-Allee 6, D-53115 Bonn, Germany

ARTICLE INFO

ABSTRACT

Advances in image analysis through deep learning have catalyzed the recent use of molecular images in chemoinformatics and drug design for predictive modeling of compound properties and other applications. For image analysis and representation learning from molecular graphs, convolutional neural networks (CNNs) represent a preferred computational architecture. In this work, we have investigated the questions whether functional groups (FGs) and their distinguishing chemical features can be learned from compound images using CNNs of different complexity and whether such knowledge might be transferable to other prediction tasks. We have shown that frequently occurring FGs were comprehensively learned, leading to highly accurate multi-label FG predictions. Furthermore, we have determined that the FG knowledge acquired by CNNs was sufficient for accurate prediction of compound activity cliffs (ACs) via transfer learning. Re-training of FG prediction models on AC data optimized convolutional layer weights and further improved prediction accuracy. Through feature weight analysis and visualization, a rationale was provided for the ability of CNNs to learn FG chemistry and transfer this knowledge for effective AC prediction.

Introduction

Deep learning has significantly advanced image analysis in different areas including biology and medicine [1–3]. In addition to natural language processing, image analysis has in recent years been one of the growth areas for deep learning, which has contributed much to its increasing popularity across different fields. Among deep learning architectures used for image processing and classification as well as learning from graph representations, convolutional neural networks (CNNs) play a major role [1,4-6]. Progress in image and graph analysis has also begun to impact chemistry, where molecular images have recently been used for representation learning and the prediction of various compound properties [7–13]. Although it remains to be determined whether image-based approaches might further improve the performance level of machine learning based upon conventional chemical descriptors, proof-of-

concept has been established for the use of molecular image data and some promising results have been obtained. In image-based chemical applications as well as representation learning from molecular graphs, CNN architectures have preferentially been used.

In another proof-of-concept study, we have recently predicted activity cliffs (ACs), which are formed by pairs of active structural analogues with significant potency differences [14], on the basis of image data [15]. ACs are of particular interest in medicinal chemistry because they capture small chemical modifications having large biological effects and are thus rich in structure-activity relationship information [14]. For image-based AC prediction, a CNN architecture was also used [15]. AC prediction represents a special task because in this case, test instances are compound pairs, rather than individual molecules. Accordingly, in AC prediction, the negative class consists of pairs of active structural analogs with small or no differences in potency. ACs were first correctly
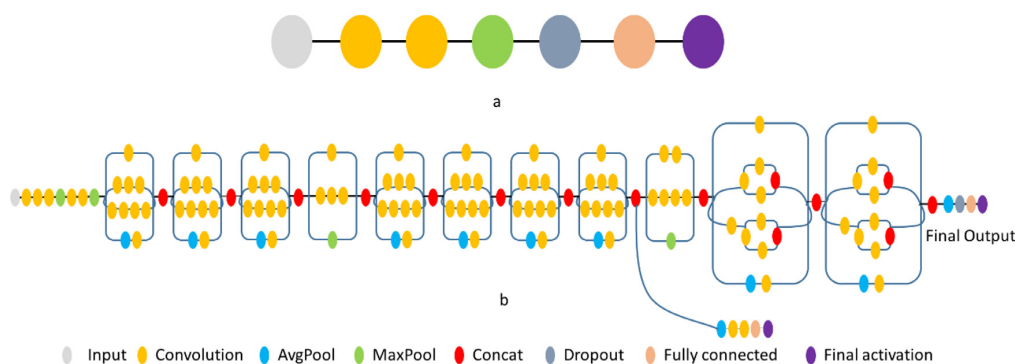

Fig. 1. Convolutional neural networks. The SCNN (top) and IV-3 (bottom) architectures are illustrated.

predicted using support vector machines with newly designed kernel functions for descriptor vectors representing compound pairs [16] and image-based predictions yielded comparable accuracy [15].

In this work, we use AC prediction as a test case to investigate whether functional group chemistry can be learned from molecular images and transferred to other prediction tasks. Therefore, a two-step learning approach was devised, as detailed in the following.

Study concept

Compounds forming an AC as defined herein (see below) share their core structure and are distinguished by the replacement of a substituent (R-group, functional group). Accordingly, if it were possible to learn different functional groups (FGs) from images and recognize them in compounds, FGs distinguishing compounds in ACs and non-AC compound pairs might be detected, hence providing a basis for AC prediction. For our analysis, frequently occurring FGs were identified. It was first attempted to learn FGs from molecular images using CNNs and then to apply this knowledge for predicting ACs via transfer learning. Hence, following this analysis scheme, successful prediction of ACs would confirm the ability to learn FG chemistry from compound images.

Transfer learning [17,18] refers to the process of learning a new task by transfer of knowledge from a related task for which models have already been derived. Transfer learning is applied in machine learning to derive models in the presence of related predictions tasks, in particular, when limited amounts of training data are available for individual tasks.

To facilitate the analysis, a deep CNN architecture was pre-trained and fine-tuned in different ways. Pre-training was carried out using compound images to learn FG chemistry and accurately classify compounds based on the presence or absence of FGs. Only the final layers of these CNN models were then fine-tuned for the complex task of AC prediction based on transfer learning. Alternatively, all layers of CNN models pre-trained on compound images or general image data were re-trained using AC data. As a control, models initialized with randomized weights were trained only on AC data.

Methods and materials

Definition of functional groups

Substructures representing commonly observed FGs were extracted from compounds using the FG identification algorithm introduced by Ertl [19]. The method identifies all heteroatoms in a molecule together with carbon atoms connected by non-aromatic double or triple bonds to other carbons or any heteroatom, acetal carbons, and oxirane, aziridine and thiirane rings, and combines subsets of connected marked atoms into FGs [19]. FGs are then extracted together with atom environment information (i.e., bonded carbon or hydrogen atoms) and assigned to

different classes following a generalization scheme associated with the FG identification algorithm [19].

Molecular image representations

Each compound was represented as a molecular object using the RD-Kit application programming interface (API) [20]. For each compound, high-resolution portable network graphics images with 500×500 pixels were generated using the RDkit Chem.Draw package (version 2020.03.5) [20]. The images were re-sized to 300×300 pixels and each pixel value from each channel was subtracted from the maximum pixel value of 255 to invert the colors and generate a black background. Supplementary Figure S1 shows exemplary molecular image representations. Pixel values of all image matrices were converted into 32-bit floating point format and normalized to the range of 0–1. Images were processed using openCV (version 4.5.0) [21–23].

Convolutional neural network architectures

Two distinct CNN architectures were assembled for multi-label classification of FGs and prediction of ACs and compared.

Simple architecture

A basic/simple CNN architecture (termed SCNN) shown in Fig. 1a was previously used for proof-of-concept image-based AC prediction [15]. This CNN architecture comprised two convolutional layers with 32 kernels and respective filter sizes of 3×3 and 5×5 to extract key image features. The convolutional layers were followed by a pooling, dropout, and dense layer. Max-pooling was used as pooling layer to compute the maximum value in each patch of each convolved feature map. A dropout layer was added to avoid overfitting. To train the SCNN model on compound images or condensed graph of reaction (CGR) representations (see below), the input layer was modified to accept images with 300×300 resolution. As final layer activation functions, *sigmoid* and *softmax* were used for FG multi-label classification and AC prediction, respectively. Models were trained using the Adam optimizer to minimize binary cross entropy loss with initial learning rates of 10^{-3} and 10^{-5} for FG multi-label classification and AC prediction, respectively. CNN layers were implemented using TensorFlow (version 2.2.0) [24] and Keras (version 2.4.3) [25].

Complex architecture

As a complex CNN architecture, Inception-V3 (IV-3) [26] with a depth of 42 layers was used, as shown in Fig. 2b. IV-3 represents a further improved version of the GoogleNet [27] architecture. The IV-3 model was previously used for classification analysis of ImageNet's Large Visual Recognition Challenge data [28]. To train the IV-3 model on the compound and CGR representations, the receptive field for the input layer was modified to accept images with 300×300 resolution and
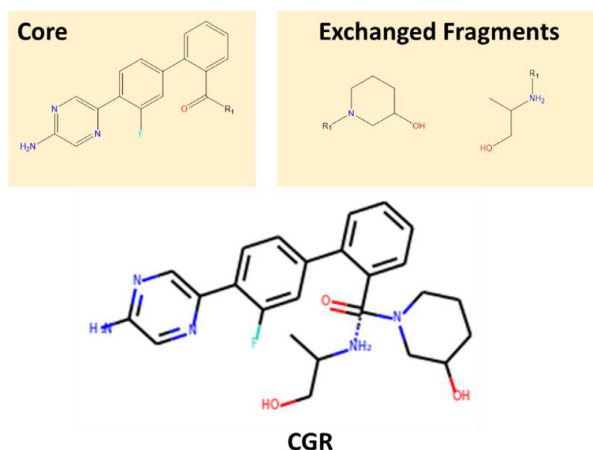
Fig. 2. Condensed graph of reaction. For an MMP consisting of a pair of compounds with a shared core and two exchanged fragments (top), the corresponding CGR representation is shown (bottom). The zero-order bond is indicated by a dashed line.

stride 2. For transfer learning, the IV-3 architecture was slightly modified by replacing the last fully connected layer with three fully connected layers of output dimensions of 500, 1000, and 2000 neurons, respectively. As final layer activation functions, *sigmoid* and *softmax* were used for FG multi-label classification and AC prediction, respectively. Models were trained using the Adam optimizer to minimize binary cross entropy loss with initial learning rates of 10^{-3} and 10^{-5} for FG multi-label classification and AC prediction, respectively. The model architecture was implemented using TensorFlow (version 2.2.0) and Keras (version 2.4.3).

Matched molecular pairs and activity cliffs

A matched molecular pair (MMP) is defined as a pair of compounds that are only distinguished by a chemical modification at a single site [29]. As such, MMPs are highly suitable for the representation of ACs [30]. From compounds, MMPs were generated by systematically fragmenting individual exocyclic single bonds and organizing core structures and substituents in index tables [29]. For substituents, size restrictions for distinguishing fragments were applied to limit MMPs to pairs of typical structural analogs [30]. Accordingly, a substituent was permitted to contain at most 13 non-hydrogen atoms and the core had to be at least twice as large as the substituent. Additionally, for MMP compounds, the maximum difference in non-hydrogen atoms between the substituents was set to eight [30]. Furthermore, MMPs from a compound activity class were only retained if their core structures were found in multiple MMPs.

An MMP formed by two compounds sharing the same activity was classified as an AC if the two structural analogs had an at least 100-fold difference in potency ($\Delta pK_i \geq 2.0$) [30]. To avoid potency difference-dependent boundary effects in AC prediction, compounds forming a non-AC MMP were permitted to have an at most 10-fold difference in potency.

Condensed graph of reaction image representations

MMPs can be represented in a single graph using the condensed graph of reaction (CGR) approach [31]. The CGR formalism was originally conceived to combine reactants and products graphs based upon a superposition of invariant parts [31]. The resulting CGR is a completely connected graph in which each node represents an atom and each edge a bond. In a CGR, the shared core of an MMP and the two exchanged

substituent fragments are represented as a single pseudo-molecule. MMP CGRs were generated using an in-house Python script and converted into a pseudo-molecule using the RDKit API. The larger fragment was connected with the core via a single bond and the smaller fragment a hypothetical zero-order bond [32]. For each pseudo-molecule, high-resolution portable network graphics images with 500 × 500 pixels were generated using the RDkit Chem.Draw package (version 2020.03.5) [20] as illustrated in Fig. 2.

The images were re-sized to 300 × 300 pixels. Each pixel value was subtracted from the maximum pixel value of 255 to invert the colors and convert the white to a black background. Pixel values of all image matrices were converted into 32-bit floating point format and normalized to the range of 0–1. CGR images were processed using openCV (version 4.5.0) [21–23]. MMP CGR images were generated using 12 CGR rotations with differences of 30° including default (0°), ±30°, ±60°, ±90°, ±120°, ±150°, and 180° Image rotations are illustrated in Supplementary Figure S2.

Feature visualization

The Grad-Cam algorithm [33] was used to extract spatial information from the convolutional layers of the trained CNN models. Channel-based mean pixel values of the resulting convolutional feature map activation weights were mapped to the original image for visualization.

Performance measures

Functional group multi-label classification
FG multi-label classification performance of CNN models was evaluated using five different performance measures including multi-label example based average accuracy (EA), exact match ratio (EMR), average precision (EP), average recall (ER), and average F1 score(E-F1) [34,35]. Definitions are provided as Supplementary Methods.

Activity cliff prediction
CNN models were trained to systematically distinguish between ACs and non-AC MMPs. The classification performance of CNN models was evaluated using receiver-operator characteristic (ROC) curves and the area under the ROC curve (AUC). In addition, model performance was assessed with six performance measures including overall accuracy (A), balanced accuracy (BA), precision (P), recall (R), weighted mean F1 score [36], and Matthews correlation coefficient (MCC) [37]. Definitions are provided as Supplementary Methods.

Compound activity classes

From ChEMBL (version 26) [38], five compound activity classes with available high-confidence activity data were extracted. Compounds were tested against single human targets in direct interaction assays at highest assay confidence (ChEMBL confidence score 9). As potency measurements, assay-independent equilibrium constants (pK_i values) were required. Multiple measurements for the same compound were averaged, provided all values fell within one order of magnitude; otherwise, the compound was disregarded. Table 1 reports the compounds activity classes and MMP/AC statistics.

Identification of functional groups

From 80,641 unique ChEMBL compounds belonging to 992 activity classes with available pK_i values for human targets, 46,671 compounds with a size range of 25–35 non-hydrogen atoms were selected such that images of these compounds generated for the subsequent analysis were of comparable size. From these compounds, a total of 257,663 FGs were algorithmically extracted [19] and the 100 most frequently occurring FGs were selected. Only 110 compounds did not contain any of these FGs, leaving 46,561 unique compounds for subsequent modeling.

3

Table 1

Compound activity classes, matched molecular pairs and activity cliffs.

Target name	Compounds	ACs			Non-AC MMPs		
		MMPs	Unique cores	Unique substituents	MMPs	Unique cores	Unique substituents
Thrombin (CHEMBL204)	332	149	13	65	979	96	251
Tyrosine kinase ABL (CHEMBL1862)	345	378	21	152	2655	106	332
5-Lipoxygenase activating protein (CHEMBL4550)	883	4227	57	556	23,720	216	783
Adenosine A3 receptor (CHEMBL256)	1378	531	83	246	5116	497	596
Cannabinoid CB2 receptor (CHEMBL253)	1698	518	124	302	5425	528	800

The composition of target-based compound activity classes is summarized and MMP/AC statistics are provided for each class. Target names are reported and ChEMBL target IDs are given in parentheses.

Table 2

Multi-label functional group classification.

Metric	IV-3 model	SCNN model
EA	0.96 (± 0.01)	0.44 (± 0.02)
EMR	0.89 (± 0.01)	0.07 (± 0.01)
EP	0.97 (± 0.01)	0.51 (± 0.01)
ER	0.99 (± 0.01)	0.68 (± 0.02)
E-F1	0.98 (± 0.01)	0.56 (± 0.02)

SMILES representations of the top 100 FGs are provided as Supplementary Methods.

Results and discussion

Multi-label classification of functional groups

SCNN and IV-3 models were trained to learn the 100 most frequent FGs from images of the 46,561 compounds. The models were designed to learn key features from each compound and map the output to a 100-dimensional FG vector, in which each value represented a probability distribution for the presence of the corresponding FG according to its rank. For model training and testing, the compounds and their images were divided with FG multi-label stratification using a previously reported technique [39,40] into training (70%) and test (30%) sets for three independent trials. Table 2 reports test results as mean values and standard deviations of different performance measures. IV-3 models yielded an EA of 0.96, EMR 0.89, EP 0.97, ER 0.99, and E-F1 0.98, whereas SCNN models were much less accurate, with an EA of 0.44, EMR 0.07, EP 0.51, ER 0.68, and E-F1 0.56.

For the IV-3 and SCNN model, mean values of different performance measures and standard deviations (in parentheses) over three independent prediction trials are reported.

The EMR value demonstrated that IV-3 models comprehensively learned the 100 FGs and accurately classified 89% of the test compounds. By contrast, SCNN model only classified 7% of the compounds correctly. Hence, only the complex IV-3 models were able to learn key image features corresponding to the most frequent 100 FGs with high precision. Therefore, the IV-3 model with the highest EMR value was selected with FG weights from pre-training with ChEMBL compounds for transfer learning (termed I-FG model).

Prediction of activity cliffs via transfer learning

To investigate the transferability of the acquired FG knowledge, the I-FG model with pre-trained weights was used to predict ACs. For comparison, the IV-3 model with pre-calculated weights for ImageNet [28] was also used (termed I-IN model). For each transfer learning model, all pre-trained layer weights were kept constant, with the exception of the last three fully connected layers, which were allowed to optimize weights during fine-tuning. For AC prediction, the models were fine-tuned using CGR image representations of ACs identified in five different compound activity classes according to Table 1. Fine-tuning was carried out over 10 independent trials by randomly sampling 70% of the AC and non-AC MMP images from the sets of 149 – 4227 ACs and 979 – 23,720 non-AC MMPs, depending upon the activity class. The resulting models were then tested on the remaining 30% of AC and non-AC MMP images. The performance of the I-FG and I-IN models is summarized in Table 3 and Fig. 3 (left column). Both models were found to be predictive, but I-FG with specifically derived FG weights was consistently more accurate than I-IN with general image weights. The I-FG models reached mean BA values of 0.63–0.86, MCC values of 0.31–0.73, and F1 values of 0.35–0.76 for the different activity classes. In addition, I-FG yielded ROC AUC values of 0.87, 0.90, 0.82, 0.82, and 0.96 for the different classes, while I-IN produced values of 0.75, 0.86, 0.77, 0.85 and 0.91, respectively. Furthermore, transfer learning using fine-tuned I-FG models was stable for all activity classes, as indicated by very low standard deviations. Taken together, these results reflected overall suc-

Table 3

Transfer learning performance.

Target	Model	A	BA	MCC	F1	P	R
4550	I-FG	0.88 (±0.01)	0.73 (±0.03)	0.49 (±0.02)	0.55 (±0.04)	0.63 (±0.07)	0.51 (±0.09)
	I-IN	0.80 (±0.04)	0.65 (±0.05)	0.28 (±0.03)	0.38 (±0.08)	0.40 (±0.09)	0.43 (±0.17)
256	I-FG	0.92 (±0.01)	0.76 (±0.05)	0.52 (±0.05)	0.55 (±0.05)	0.57 (±0.08)	0.57 (±0.12)
	I-IN	0.90 (±0.04)	0.70 (±0.08)	0.42 (±0.06)	0.43 (±0.10)	0.58 (±0.19)	0.44 (±0.21)
253	I-FG	0.91 (±0.01)	0.63 (±0.04)	0.31 (±0.04)	0.35 (±0.05)	0.44 (±0.08)	0.31 (±0.09)
	I-IN	0.88 (±0.03)	0.62 (±0.05)	0.24 (±0.05)	0.28 (±0.07)	0.34 (±0.08)	0.30 (±0.14)
204	I-FG	0.94 (±0.01)	0.86 (±0.04)	0.73 (±0.04)	0.76 (±0.04)	0.77 (±0.06)	0.76 (±0.07)
	I-IN	0.91 (±0.02)	0.75 (±0.05)	0.57 (±0.06)	0.60 (±0.06)	0.72 (±0.12)	0.54 (±0.12)
1862	I-FG	0.88 (±0.01)	0.70 (±0.03)	0.42 (±0.04)	0.48 (±0.04)	0.53 (±0.07)	0.45 (±0.07)
	I-IN	0.87 (±0.03)	0.66 (±0.11)	0.34 (±0.15)	0.36 (±0.20)	0.56 (±0.18)	0.37 (±0.27)

For each model, mean A, BA, MCC, F1, P, and R values and standard deviations (in parentheses) over 10 independent trials are reported for the different activity classes (identified by CHEMBL target IDs according to Table 1).

4
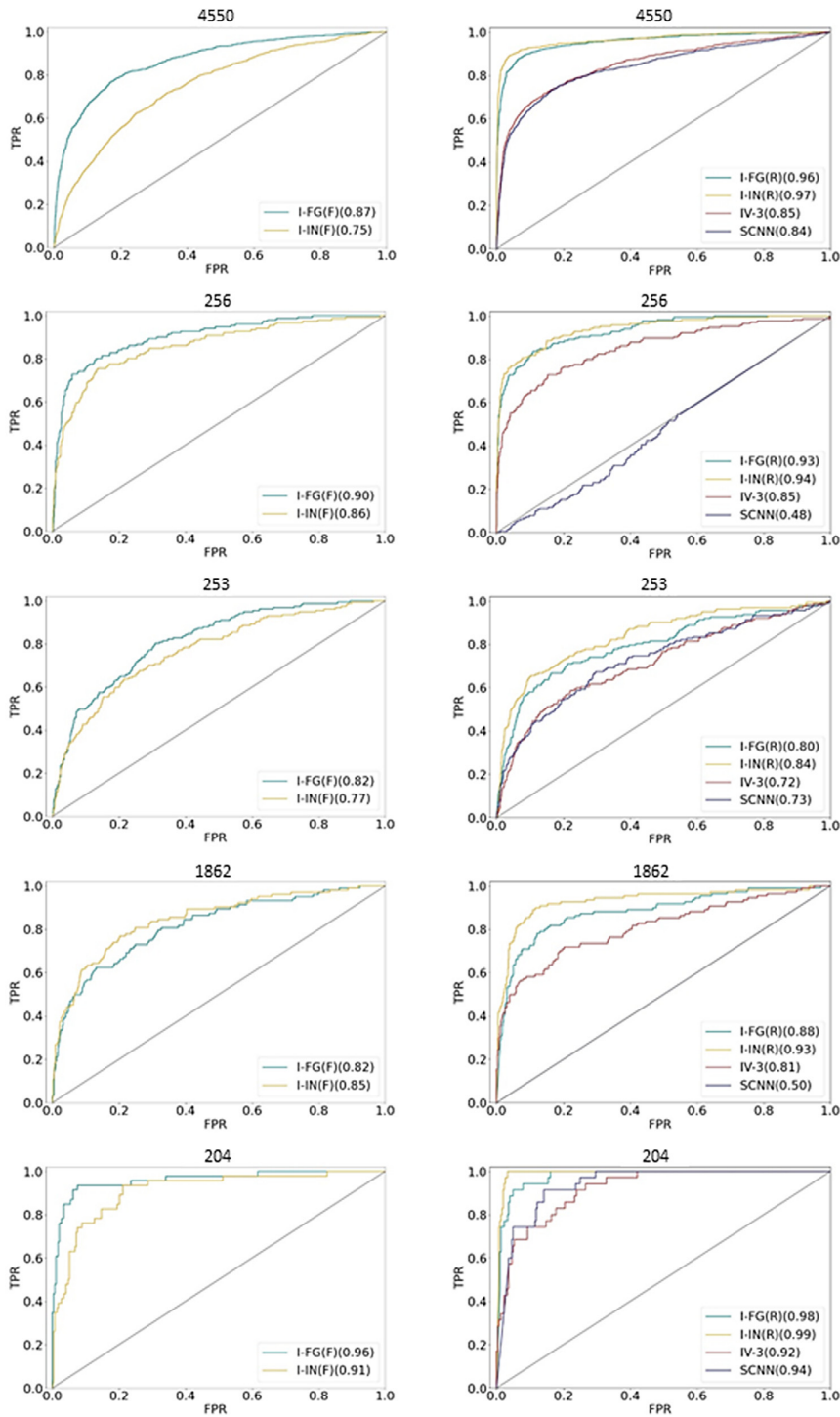
Fig. 3. Receiver operating characteristic curves. The performance of the best CNN models from an individual AC prediction trial is monitored in ROC curves for the different activity classes after fine-tuning (left column) and re-training (right). For each curve, AUC values are reported in parentheses.

5
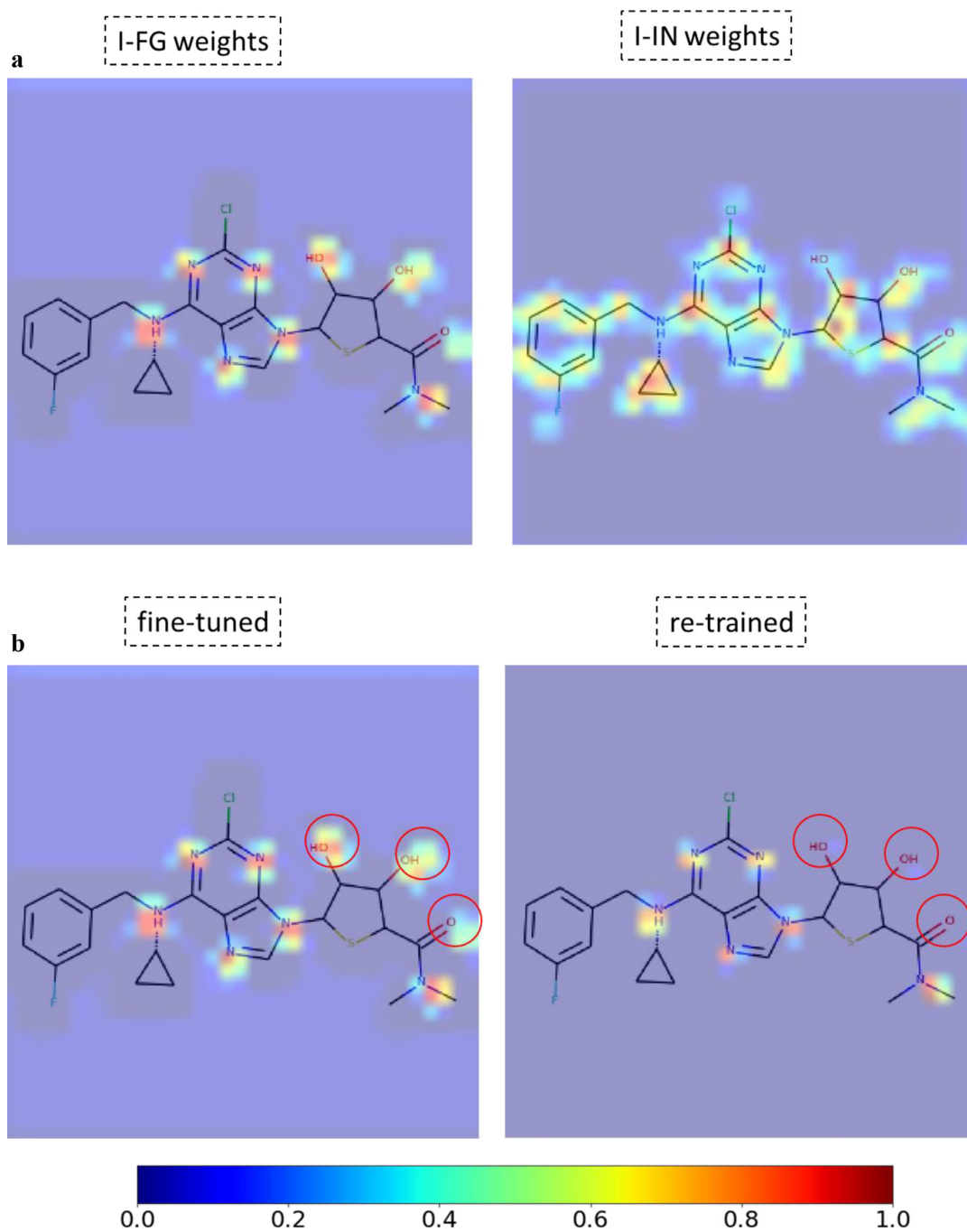
Fig. 4. Mapping of convolutional layer activation weights. Mean gradient weights of the fifth convolutional layer of different models are mapped to a CGR image of an exemplary MMP and displayed. The continuous color code from blue (0) over green (0.5) to red (1) indicates weights from 0 to 1. In (a), weights from the fine-tuned I-FG and I-IN AC prediction models are compared. In (b), the fine-tuned and re-trained I-FG model are compared. Observed changes in the optimized convolutional layer weights are circled red.

cessful AC predictions, hence providing proof-of-concept for the ability to learn FG chemistry and the transfer learning approach.

Activity cliff prediction with re-trained models

In addition to transfer learning, the I-FG and I-IN models pre-trained with different weight initializations were also re-trained. During re-training on 70% of the AC and non-AC MMP images, weights were optimized across all CNN layers. In addition to I-FG and I-IN, the IV-3 and SCNN architectures initialized with random weights were also re-trained. The four models were then used for AC prediction and compared. Re-training produced predictive models in most cases, as reported in Supplementary Table S1 and shown in Fig. 3 (right column). After re-training the mean performance of I-FG and I-IN was comparable or slightly superior for I-IN, but I-FG predictions were more stable than I-IN predictions. Compared to I-FG models, the performance of I-IN models was partly improved based upon re-training with AC data. During re-training of I-FG and I-IN, the models learned specific local structural features and further optimized pre-trained weights accordingly, which slightly improved classification performance. Furthermore, prediction accuracy of the IV-3 and SCNN with random weight initialization was generally lower. For example, for activity class 4550, mean BA values of I-FG: 0.88 ($\pm$0.02), I-IN: 0.81 ($\pm$0.17), IV-3: 0.74 ($\pm$0.02), and SCNN: 0.72 ($\pm$0.08) were obtained and mean MCC values of 0.77 ($\pm$0.02), 0.65 ($\pm$0.35), 0.56 ($\pm$0.02), and 0.47 ($\pm$0.17), respectively. For activity class 253, prediction accuracy of the models generally decreased and approached the level of random predictions for the IV-3 and SCNN models. In this case, mean BA values of I-FG: 0.67 ($\pm$0.04), I-IN: 0.70 ($\pm$0.03), IV-3: 0.56 ($\pm$0.02), and SCNN: 0.54 ($\pm$0.05) were obtained and mean MCC values of 0.38 ($\pm$0.03), 0.43 ($\pm$0.03), 0.19 ($\pm$0.05), and 0.12 ($\pm$0.13), respectively. Overall, re-training of I-FG and I-IN on AC and non-AC MMP images also yielded meaningful AC predictions.

Model performance with rotated image variants

FGs are easily recognizable by a chemist in an image regardless of the orientation of the molecule. However, for image analysis using CNNs, orientation-independence is not ensured and should be evaluated. Therefore, to assess the influence of the molecular orientation on model performance, a test set of rotated CGR images of ACs and non-AC MMPs was generated. The fine-tuned and re-trained I-FG, I-IN, IV-3, and SCNN models were tested again using 11 rotated image variants of each MMP. For each MMP CGR representation, 11 different rotations were generated in increments of 30° as described in the Methods and materials section and illustrated in Supplementary Figure S2. The results are summarized in Supplementary Figure S3. Interestingly, both fine-tuned and re-trained I-FG models essentially retained their original performance in a rotation-invariant manner with only slight reduction for three activity classes (1862, 204 and 256), as shown in Figure S3a. By contrast, I-IN models displayed significant reductions in performance or failed for rotational image variants and the IV-3 and SCNN models consistently failed. For the two remaining activity classes (4550 and 253), all models consistently failed (Figure S3b). Hence, learning of FG chemistry by I-FG led to image rotation invariance in AC prediction for some (but not all) compound classes.

Rationalization of learned functional group chemistry

The analysis of feature weights of convolutional layers of different CNN models might help to better understand how FG chemistry was learned. Therefore, convolutional layer weights were extracted from different AC prediction models for MMP CGR images. The weights were then mapped on the images and visualized. Fig. 4 shows a representative example. In Fig. 4a, feature weights from the fine-tuned I-FG and I-IN models are compared. A compelling observation was that the fine-tuned I-FG model detected heteroatoms as specific chemical features

that were involved in the formation of most FGs. On the other hand, the I-IN model detected more general chemical features covering both the core structure and FGs. These observations provided a rationale for successful transfer learning by the I-FG model, given its ability to specifically recognize FGs that distinguished between compound forming ACs and non-AC MMPs.

Fig. 4b compares feature weights from the fine-tuned and re-trained I-FG models, which were initially subjected to the same FG learning process. Re-training based on AC and non-AC MMP images then optimized the weights, leading to slightly improved accuracy of the re-trained model in AC prediction, as discussed above. The comparison shows that re-training in this case selectively optimized weights on nitrogen atoms and de-prioritized oxygen atoms. Hence, the model clearly distinguished between specific chemical features that were initially learned from compound images and frequently occurring FGs.

Conclusion

In this work, we have addressed the question whether FGs can be learned from compound images using different CNN architectures and if this knowledge would be sufficient for compound pair-based prediction of ACs, which largely relies on detecting FG replacements leading to varying compound potency differences. Therefore, a transfer learning scheme was investigated, which confirming that CNN models learned sufficient R-group chemistry from images of individual compounds to accurately predict ACs on the basis of CGR images. We also demonstrated that transfer learning could be replaced by re-training CNN models on AC and non-AC MMP images. Re-training optimized convolutional layer weights from FG-oriented pre-training, leading to further improved AC prediction accuracy. Finally, a rationale for learning FG chemistry and transferring this knowledge was provided by the analysis and visualization of CNN model-internal feature weights. Taken together, the results of our analysis confirm the ability of CNNs to learn FG chemistry from compound images and further expand the methodological framework for AC predictions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2021.100022.

References

[1] Rawat W, Wang Z. Deep convolutional neural networks for image classification: a comprehensive review. Neur Comput 2017;29:2352–449.
[2] Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Ann Rev Biomed Eng 2017;19:221–48.
[3] Moen E, Bannon D, Kudo T, Graf W, Covert M, Van Valen D. Deep learning for cellular image analysis. Nature Meth 2019;16:1233–46.
[4] Sun Y, Xue B, Zhang M, Yen GG. Evolving deep convolutional neural networks for image classification. IEEE Trans Evol Comput 2019;24:394–407.
[5] Zhang S, Tong H, Xu J, Maciejewski R. Graph convolutional networks: a comprehensive review. Comput Soc Netw 2019;6:1–23.
[6] Chen M, Wei Z, Huang Z, Ding B, Li Y. Simple and deep graph convolutional networks. Proc Mach Learn Res 2020;119:1725–35.
[7] Gomes J, Ramsundar B, Feinberg EN, Pande VS. Atomic convolutional networks for predicting protein-ligand binding affinity. arXiv preprint 2017.
[8] Chuang KV, Gunsalus LM, Keiser MJ. Learning molecular representations for medicinal chemistry. J Med Chem 2020;63:8705–22.
[9] Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. arXiv preprint 2016.
[10] Goh GB, Siegel C, Vishnu A, Hodas NO, Baker N. Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. arXiv preprint 2017.

[11] Goh GB, Siegel C, Vishnu A, Hodas N. Using rule-based labels for weak supervised learning: A ChemNet for transferable chemical property prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018. p. 302–10.

[12] Fernandez M, Ban F, Woo G, Hsing M, Yamazaki T, LeBlanc E, Rennie PS, Welch WJ, Cherkasov A. Toxic Colors: the use of deep learning for predicting toxicity of compounds merely from their graphic images. J Chem Inf Model 2018;58:1533–43.

[13] Cortés-Ciriano I, Bender A. KekuleScope: prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images. J Cheminf 2019;1:41.

[14] Stumpfe D, Bajorath J. Exploring activity cliffs in medicinal chemistry. J Med Chem 2012;55:2932–42.

[15] Iqbal J, Vogt M, Bajorath J. Prediction of activity cliffs on the basis of images using convolutional neural networks. J Comput Aided Mol Des 2021 in press. doi:10.1007/s10822-021-00380-y.

[16] Heikamp K, Hu X, Yan A, Bajorath J. Prediction of activity cliffs using support vector machines. J Chem Inf Model 2012;52:2354–65.

[17] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2009;22:1345–59.

[18] Yang Q, Zhang Y, Dai W, Pan SJ. Transfer learning. Cambridge, UK: Cambridge University Press; 2020.

[19] Ertl P. An algorithm to identify functional groups in organic molecules. J Cheminform 2017;9:36.

[20] Landrum G. RDKit: open-source cheminformatics 2021. https://www.rdkit.org.

[21] Culjak I, Abram D, Pribanic T, Dzapo H, Cifrek M. A brief introduction to OpenCV. In: 2012 Proceedings of the 35th International Convention MIPRO; 2012. p. 1725–30.

[22] OpenCv. OpenCV library 2014. https://www.opencv.org.

[23] Bradski G. The OpenCV library. Dr Dobb's. J Softw Tools 2000;25:120–5.

[24] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu YZX. TensorFlow: a system for large-scale machine learning. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Savannah, GA; 2016.

[25] Chollet F. Keras 2021. https://github.com/keras-team-keras.

[26] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision 2015. https://arxiv.org/abs/1512.00567.

[27] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. Cornel University Library; 2014. http://arxiv.org/abs/1409.4842

[28] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 2015;115:211–52.

[29] Hussain J, Rea C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. J Chem Inf Model 2010;50:339–48.

[30] Hu X, Hu Y, Vogt M, Stumpfe D, Bajorath J. MMP-cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs. J Chem Inf Model 2012;52:1138–45.

[31] A Fourches D, Hoonakker F, Solov'ev VP. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. J Comput Aided Mol Des 2005;19:693–703.

[32] Clark AM. Accurate specification of molecular structures: the case for zero-order bonds and explicit hydrogen counting. J Chem Inf Model 2011;51:3149–57.

[33] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vision, Springer Sci Bus Media LLC 2019;128:336–59.

[34] Godbole S, Sarawagi S. Discriminative Methods for Multi-labeled Classification. In: Dai H, Srikant R, Zhang C, editors. Advances in knowledge discovery and data mining. pakdd 2004. Berlin, Heidelberg: Springer; 2004. p. 22–30.

[35] Sorower MS. A literature survey on algorithms for multi-label learning, 18. Corvallis: Oregon State University; 2010. p. 1–25.

[36] Chinchor N. MUC-4 evaluation metrics. Proceeding of the 4th Conferenc on Message Understanding. Assoc Comput Linguist USA 1992:22–9.

[37] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–51.

[38] Gaulton A, Hersey A, Nowotka ML, Patricia Bento A, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrian-Uhalte E, Davies M, Dedman N, Karlsson A, Magarinos MP, Overington JP, Papadatos G, Smit ILA. The ChEMBL database in 2017. Nucleic Acids Res 2017;45:D945–54.

[39] Sechidis K, Tsoumakas G, Vlahavas I. On the Stratification of Multi-label Data. In: Gunopulos D, Hofmann T, Malerba D, Vazirgiannis M, editors. Machine learning and knowledge discovery in databases. ecml pkdd 2011. Berlin, Heidelberg: Springer; 2011. p. 145–58.

[40] Szymański P, Kajdanowicz T. A network perspective on stratification of multi-label data. In: Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications. Proccedings of Machine Learning Research, 74; 2017. p. 22–35.

8

Summary

This chapter investigated the application of different CNN architectures to learn functional group chemistry from compound image representations, and estimated the impact of acquired functional group chemistry knowledge for ACs classification prediction. The transfer learning scheme confirmed the potential of deep CNN models to learn sufficient functional group chemistry from images of individual compounds leading to accurate ACs prediction based on CGR images. Furthermore, CNN models were allowed to learn structural features from CGR images in re-training process and compared. Finally, the convolutional feature analysis was used to identify significant features in original images, which confirmed the potential of CNNs to learn functional groups chemistry. Overall, CNN modeling performance confirmed the ability of CNNs to learn functional group chemistry from compound images and thus further expands the methodological framework for AC predictions.

Chapter 8

Conclusion

Recent image processing approaches and state-of-the-art machine learning models have rapidly progressed in multiple scientific domains. Predictive image analysis of different SAR contents from different compound datasets based on 3D ALs image representations was carried out in *Chapter 2* for the first time. CNNs and machine learning models have the potential to distinguish different 3D AL surfaces based on elevation-dependent topographical features that mirror SAR characteristics. Image processing approaches can be applied to detect, segment, and extract 3D AL topographical features to compare different SARs. 3D AL images in 8-bit RGB color models with better color contents improve the classification performance, whereas the standard conversion of the image from RGB model to two-dimensional monochrome and binary colors reduces the model performance. Binary images are preferable for learning elevation-based topographical features from 3D AL images, which helps classify different SARs on elevation-based SAR topologies. 3D AL heatmaps are appropriate to capture the complete top-down view of the AL/PL by preserving spatial features, color profiles, and topographical characters. Heatmap representations in image formats allow extraction and quantification of activity/property distributions as 3D ALs/PLs topographical features (i.e., valleys and mountains). The introduction of a novel methodology in *Chapter 3* that evenly divides 3D AL heatmaps into fixed-sized cells successfully extracted and quantified the color features that directly reflect compound activity values. Extracted color features quantifying different topographies obtained by thresholding color values can be represented as a single feature vector for further comparison. 3D AL comparisons based on extracted feature vectors with standard metrics provided conceivable and robust results and confirmed that 3D AL topographical features characterizing SAR

contents are quantifiable by applying image processing approaches. Topographical characteristics (i.e., valleys, peaks, or mountains) in a 3D AL are detectable, segment-able, and quantifiable. The novel methodology introduced in *Chapter 4* quantifies prominent topographical features of 3D ALs and has the potential to segment activity peaks and deep valleys from 3D AL heatmap representations. Each segmented region allows the quantification of activity distributions and makes it possible to discern a similarity between different 3D ALs on topographical characteristics. Furthermore, the methodology allows numerical comparison and established successful similarity ranking of different 3D ALs to a template 3D AL based on topographical features using image representations. CNN modeling and image analysis help in the identification of textures, color profiles, and topographical features from 3D ALs, allowing AL/PL quantitative comparison in a meaningful way.

The successful application of image processing approaches in 3D ALs quantitative analysis discussed in *Chapters 2-4* provided evidence that image processing can be applied to analyze 3D ALs in image-representable data structures. Therefore, follow up *Chapters 6-7* investigated CNN modeling to analyze prominent features of SARs known as ACs. CNN modeling and image analysis can be employed to further investigate ACs, which localize mainly in rugged regions. Pair of compounds in molecular graph data structures can be converted into 2D images and, therefore, amenable for image analysis. In *Chapter 6* CNN modeling was found to successfully discriminate MMP-cliffs from non-ACs by learning structural features systematically from pair of compound image representations (MMPs). In addition, CNN modeling learned unique structural representations, such as nitrogen and nitrogen-related functional groups, which helped accurate classification prediction. Classification of MMP-cliffs and non-ACs based on learning structural chemistry aided the model prediction interpretation from a chemistry perspective. The potential of CNNs to learn structural features precisely to distinguish MMP-cliffs from non-ACs in *Chapter 6* was an encouraging finding. However, it was unclear whether CNNs can learn the functional group chemistry from compound images and distinguish ACs from non-ACs based on acquired functional group knowledge. Successful learning of functional groups and precise classification of MMP-cliffs and non-ACs, as discussed in *Chapter 7*, clearly indicates the power of deep CNN models such as Inception to learn and distinguish between MMP-cliffs and non-ACs by transferring the functional group knowledge. Furthermore, visualization of convolutional features demonstrated CNNs consistently focused heavy-atoms by giving them high weightage. The focus on heavy-atoms confirms the CNN modeling ability to learn

functional groups and successful discrimination of ACs from non-ACs by transferring the learned functional group chemistry knowledge.

From 3D AL analysis (*Chapters 2-4*) to AC analysis(*Chapters 6-7*) based on image representations, the CNN modeling and incorporation of image analysis concepts has proven to be very successful and thus highly applicable for image analysis in the chemical informatics domain. The application of image processing approaches has the potential to learn structural features and patterns from various data structures representable as image formats (i.a., ALs, molecular graphs, etc.). Image-based deep CNN modeling can be employed to learn structural chemistry for a particular prediction task using molecular image representations. The prediction analysis of tasks based on the learned chemistry could help in the interpretation of the results from a chemical perspective.

Future work includes the application of image processing approaches to investigate the applicability of image vision approaches to interpret CNN models in a chemically meaningful way.

Bibliography

[1] Stumpfe, D.; Bajorath, J. Recent developments in SAR visualization. *Med. Chem. Commun.* **2016**, *7*, 1045–1055.

[2] Méndez-Lucio, O.; Pérez-Villanueva, J.; Castillo, R.; Medina-Franco, J. L. Identifying Activity Cliff Generators of PPAR Ligands Using SAS Maps. *Molecular informatics* **2012**, *31*, 837–846.

[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. *Journal of Chemical Information and Modeling* **2010**, *50*, 1021–1033.

[4] Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree - Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *Journal of Chemical Information and Modeling* **2007**, *47*, 47–58.

[5] Wassermann, A. M.; Haebel, P.; Weskamp, N.; Bajorath, J. SAR Matrices: Automated Extraction of Information-Rich SAR Tables from Large Compound Data Sets. *Journal of Chemical Information and Modeling* **2012**, *52*, 1769–1776.

[6] Peltason, L.; Bajorath, J. SAR Index: Quantifying the nature of structure-activity relationships. *Journal of Medicinal Chemistry* **2007**, *50*, 5571–5578.

[7] 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. *Journal of Medicinal Chemistry* **2008**, *51*, 6075–6084.

[8] Gupta-Ostermann, D.; Hu, Y.; Bajorath, J. Introducing the LASSO Graph for Compound Data Set Representation and Structure-Activity Relationship Analysis. *Journal of Medicinal Chemistry* **2012**, *55*, 5546–5553.

[9] Muratov, E. N. et al. QSAR without borders. *Chemical Society Reviews* **2020**, *49*, 3525–3564.

[10] Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today* **2007**, *12*, 225–233.

[11] Peltason, L.; Bajorath, J. Systematic computational analysis of structure-activity relationships: Concepts, challenges and recent advances. *Future Medicinal Chemistry* **2009**, *1*, 451–466.

[12] Shanmugasundaram, V.; Maggiora, G. M. Characterizing property and activity landscapes using an information-theoretic approach. Proceedings of 222nd American Chemical Society National Meeting, Division of Chemical Information. 2001; p 77.

[13] Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating structure-activity landscapes. *Drug Discovery Today* **2009**, *14*, 698–705.

[14] Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity landscape representations for structure-activity relationship analysis. *Journal of Medicinal Chemistry* **2010**, *53*, 8209–8223.

[15] Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.

[16] Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *Journal of Chemical Information and Modeling* **2015**, *7*, 1–34.

[17] Wiswesser, W. J. How the WLN Began in 1949 and How It Might Be in 1999. *Journal of Chemical Information and Computer Sciences* **1982**, *22*, 88–93.

[18] Gasteiger, J.; Engel, T. *Chemoinformatics: A Textbook*; John Wiley & Sons, 2006.

[19] Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A versatile language for chemical structure representation. *Journal of Chemical Information and Computer Sciences* **1997**, *37*, 71–79.

[20] O'Boyle, N. M. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *Journal of Chemical Information and Modeling* **2012**, *4*, 1–14.

[21] Bajusz, D.; Rácz, A.; Héberger, K. Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching In: Reference Module in Chemistry, Molecular Sciences and Chemical Engineering. *Comprehensive Medicinal Chemistry III* **2017**, *3-8*, 329–378.

[22] Chen, H.; Kogej, T.; Engkvist, O. Cheminformatics in Drug Discovery, an Industrial Perspective. *Mol Inform* **2018**, *37*, 1800041.

[23] Awale, M.; Visini, R.; Probst, D.; Arús-Pous, J.; Reymond, J. L. Chemical space: Big data challenge for molecular diversity. *Chimia* **2017**, *71*, 661–666.

[24] Naveja, J. J.; Medina-Franco, J. L. ChemMaps: Towards an approach for visualizing the chemical space based on adaptive satellite compounds. *F1000Research* **2017**, *6*.

[25] Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing pitfalls in virtual screening: A critical review. *Journal of Chemical Information and Modeling* **2012**, *52*, 867–881.

[26] Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.

[27] Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures - A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **1965**, *5*, 107–113.

[28] Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **2010**, *50*, 742–754.

[29] MACCS Structural Keys; Symyx Software:San Ramon, CA. **2005**,

[30] Feldmann, C.; Bajorath, J. Machine learning reveals that structural features distinguishing promiscuous and non-promiscuous compounds depend on target combinations. *Scientific Reports* **2021**, *11*, 7863.

[31] Jasial, S.; Gilberg, E.; Blaschke, T.; Bajorath, J. Machine Learning Distinguishes with High Accuracy between Pan-Assay Interference Compounds That Are Promiscuous or Represent Dark Chemical Matter. *Journal of Medicinal Chemistry* **2018**, *61*, 10255–10264.

[32] Rodríguez-Pérez, R.; Bajorath, J. Prediction of Compound Profiling Matrices, Part II: Relative Performance of Multitask Deep Learning and Random Forest Classification on the Basis of Varying Amounts of Training Data. *ACS Omega* **2018**, *3*, 12033–12040.

[33] Rodríguez-Pérez, R.; Miljković, F.; Bajorath, J. Assessing the information content of structural and protein-ligand interaction representations for the classification of kinase inhibitor binding modes via machine learning and active learning. *Journal of Chemical Information and Modeling* **2020**, *12*, 36.

[34] Rodríguez-Pérez, R.; Miyao, T.; Jasial, S.; Vogt, M.; Bajorath, J. Prediction of Compound Profiling Matrices Using Machine Learning. *ACS Omega* **2018**, *3*, 4713–4723.

[35] Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry* **2014**, *57*, 3186–3204.

[36] Rogers, D. J.; Tanimoto, T. T. A computer program for classifying plants. *Science* **1960**, *132*, 1115–1118.

[37] Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched molecular pairs as a medicinal chemistry tool. *Journal of Medicinal Chemistry* **2011**, *54*, 7739–7750.

[38] Hussain, J.; Rea, C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *Journal of Chemical Information and Modeling* **2010**, *50*, 339–348.

[39] Hu, Y.; Stumpfe, D.; Bajorath, J. Advancing the activity cliff concept. *F1000Research* **2013**, *2*.

[40] Maggiora, G. M. On outliers and activity cliffs - Why QSAR often disappoints. *Journal of Chemical Information and Modeling* **2006**, *46*, 1535.

[41] Stumpfe, D.; Bajorath, J. Exploring activity cliffs in medicinal chemistry. *Journal of Medicinal Chemistry* **2012**, *55*, 2932–2942.

[42] Hu, Y.; Maggiora, G. M.; Bajorath, J. Activity cliffs in PubChem confirmatory bioassays taking inactive compounds into account. *Journal of Computer-Aided Molecular Design* **2013**, *27*, 115–124.

[43] Hu, H.; Stumpfe, D.; Bajorath, J. Rationalizing the Formation of Activity Cliffs in Different Compound Data Sets. *ACS Omega* **2018**, *3*, 7736–7744.

[44] Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-cliffs: Systematic identification of activity cliffs on the basis of matched molecular pairs. *Journal of Chemical Information and Modeling* **2012**, *52*, 1138–1145.

[45] Heikamp, K.; Hu, X.; Yan, A.; Bajorath, J. Prediction of activity cliffs using support vector machines. *Journal of Chemical Information and Modeling* **2012**, *52*, 2354–2365.

[46] Horvath, D.; Marcou, G.; Varnek, A.; Kayastha, S.; De La Vega De León, A.; Bajorath, J. Prediction of activity cliffs using condensed graphs of reaction representations, descriptor recombination, support vector machine classification, and support vector regression. *Journal of Chemical Information and Modeling* **2016**, *56*, 1631–1640.

[47] Hirohara, M.; Saito, Y.; Koda, Y.; Sato, K.; Sakakibara, Y. Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC Bioinformatics* **2018**, *19*, 526.

[48] Chen, J.; Si, Y. W.; Un, C. W.; Siu, S. W. Chemical toxicity prediction based on semi-supervised learning and graph convolutional neural network. *Journal of Chemical Information and Modeling* **2021**, *13*, 93.

[49] Blaschke, T.; Miljković, F.; Bajorath, J. Prediction of Different Classes of Promiscuous and Nonpromiscuous Compounds Using Machine Learning and Nearest Neighbor Analysis. *ACS Omega* **2019**, *4*, 6883–6890.

[50] Krivák, R.; Hoksza, D. Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features. *Journal of Chemical Information and Modeling* **2015**, *7*, 12.

[51] Cortés-Ciriano, I.; Bender, A. Kekulescope: Prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images. *Journal of Chemical Information and Modeling* **2019**, *11*, 41.

[52] Fernandez, M.; Ban, F.; Woo, G.; Hsing, M.; Yamazaki, T.; Leblanc, E.; Rennie, P. S.; Welch, W. J.; Cherkasov, A. Toxic Colors: The Use of Deep Learning for Predicting Toxicity of Compounds Merely from Their Graphic Images. *Journal of Chemical Information and Modeling* **2018**, *58*, 1533–1543.

[53] Shoesmith, E.; Vapnik, V.; Kotz, S. *The Statistician*; Springer Science & Business Media, 1984; Vol. 33; p 324.

[54] Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.

[55] Rodríguez-Pérez, R.; Bajorath, J. Evaluation of multi-target deep neural network models for compound potency prediction under increasingly challenging test conditions. *Journal of Computer-Aided Molecular Design* **2021**, *35*, 285–295.

[56] Aliper, A.; Plis, S.; Artemov, A.; Ulloa, A.; Mamoshina, P.; Zhavoronkov, A. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Molecular pharmaceutics* **2016**, *13*, 2524–2530.

[57] Lusci, A.; Pollastri, G.; Baldi, P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *Journal of chemical information and modeling* **2013**, *53*, 1563–1575.

[58] Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-task Neural Networks for QSAR Predictions. *arXiv preprint arXiv:1406.1231* **2014**,

[59] Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity prediction using deep learning. *Frontiers in Environmental Science* **2016**, *3*, 80.

[60] Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of Generative Autoencoder in De Novo Molecular Design. *Molecular Informatics* **2018**, *37*, 1700123.

[61] Jaques, N.; Gu, S.; Bahdanau, D.; Hernández-Lobato, J. M.; Turner, R. E.; Eck, D. Sequence tutor: Conservative fine-tuning of sequence generation models with KL-control. 34th International Conference on Machine Learning, ICML 2017. 2017; pp 2587–2596.

[62] Segler, M. H.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science* **2018**, *4*, 120–131.

[63] Jin, W.; Coley, C.; Barzilay, R.; Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. Advances in Neural Information Processing Systems. 2017.

[64] Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Central Science* **2017**, *3*, 1103–1113.

[65] Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *Journal of Chemical Information and Modeling* **2017**, *57*, 942–957.

[66] Paoletti, M. E.; Haut, J. M.; Plaza, J.; Plaza, A. A new deep convolutional neural network for fast hyperspectral image classification. *ISPRS Journal of Photogrammetry and Remote Sensing* **2018**, *145*, 120–147.

[67] Zhang, W.; Itoh, K.; Tanida, J.; Ichioka, Y. Parallel distributed processing model with local space-invariant interconnections and its optical architecture. *Applied Optics* **1990**, *29*, 4790–4797.

[68] Alzubaidi, L.; Zhang, J.; Humaidi, A. J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M. A.; Al-Amidie, M.; Farhan, L. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data* **2021**, *8*, 1–74.

117

[69] Hubel, D. H.; Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology* **1962**, *160*, 106–154.

[70] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* **2015**, *115*, 211–252.

[71] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* **2017**, *60*, 84–90.

[72] Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper With Convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015.

[73] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016; pp 770–778.

[74] LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. 1998; pp 2278–2324.

[75] Dong, J. X.; Krzyzak, A.; Suen, C. Y. A multi-net local learning framework for pattern recognition. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. 2001; pp 328–332.

[76] Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on Machine Learning. 2015; pp 448–456.

[77] Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 2016; pp 2818–2826.

[78] Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. A. Inception-v4, inception-ResNet and the impact of residual connections on learning. 31st AAAI Conference on Artificial Intelligence, AAAI 2017. 2017; pp 4278–4284.

[79] Gaulton, A. et al. The ChEMBL database in 2017. *Nucleic Acids Research* **2017**, *45*, D945–D954.

[80] Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. A survey on deep transfer learning. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2018; pp 270–279.

[81] Zhang, W.; Li, R.; Zeng, T.; Sun, Q.; Kumar, S.; Ye, J.; Ji, S. Deep Model Based Transfer and Multi-Task Learning for Biological Image Analysis. *IEEE Transactions on Big Data* **2020**, *6*, 322–333.

[82] Zheng, A.; Casari, A. *Feature Engineering for Machine Learning*; O'Reilly Media Inc, 2018; p 218.

[83] Goh, G. B.; Vishnu, A.; Siegel, C.; Hodas, N. Using rule-based labels for weak supervised learning: A ChemNet for transferable chemical property prediction. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2018; pp 302–310.

[84] Simm, J. et al. Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery. *Cell Chemical Biology* **2018**, *25*, 611–618.e3.

[85] Zhang, Q.; Zhang, M.; Chen, T.; Sun, Z.; Ma, Y.; Yu, B. Recent advances in convolutional neural network acceleration. *Neurocomputing* **2019**, *323*, 37–51.

[86] Lowe, D. G. Object recognition from local scale-invariant features. Proceedings of the IEEE International Conference on Computer Vision. 1999; pp 1150–1157.

[87] Bay, H.; Tuytelaars, T.; Van Gool, L. SURF: Speeded up robust features. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Berlin, Heidelberg, 2006; pp 404–417.

[88] Vega-Rodriguez, M. A. Review: Feature Extraction and Image Processing. *The Computer Journal* **2004**, *47*, 271–272.

[89] Zhao, D.; Yang, L.; Wu, X.; Wang, N.; Li, H. *Lecture Notes in Electrical Engineering*; Springer, 2012; Vol. 136 LNEE; pp 299–305.

[90] Prewitt, J. Object enhancement and extraction. *Picture processing and Psychopictorics* **1970**, *10*, 15–19.

[91] Canny, J. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1986**, *PAMI-8*, 679–698.

[92] Gao, W.; Yang, L.; Zhang, X.; Liu, H. An improved Sobel edge detection. Proceedings - 2010 3rd IEEE International Conference on Computer Science and Information Technology, ICCSIT 2010. 2010; pp 67–71.

[93] Berzins, V. Accuracy of Laplacian edge detectors. *Computer Vision, Graphics, & Image Processing* **1984**, *27*, 195–210.