Artificial Intelligence in the Life Sciences 1 (2021) 100022

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

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

Keywords:
Convolutional neural networks
Transfer learning
Molecular image analysis
Functional groups
Activity cliffs

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

Abbreviations: A, accuracy; AC, activity cliff; API, application programing interface; AUC, area under curve; BA, balanced accuracy; CGR, condensed graph of reaction; CNN, convolutional neural network; EA, multi-label example based average; E-F1, multi-label example based average F1; EMR, exact match ratio; EP, multi-label example based average precision; ER, multi-label example based average recall; FG, functional group; IV-3, Inception-V3 model; I-FG, pre-trained IV-3 with ChEMBL compounds; I-FG(F), fine-tuned I-FG; I-FG(R), re-trained I-FG; I-IN, IV-3 pre-trained with ImageNet; I-IN(F), fine-tuned I-INI-IN(R): re-trained I-IN; MCC, Matthews correlation coefficient; MMP, matched molecular pair; P, precision; R, recall; ROC, receiver operating characteristic; SCNN, simple CNN model; I-IN(F)/(R), I-IN fine-tuned/ re-trained.

☆ Given his role as Editor in Chief, Jürgen Bajorath had no involvement in the peer-review of this article and has no access to information regarding its peer-review. Full responsibility for the editorial process for this article was delegated to Mingyue Zheng.

* Corresponding author.
E-mail address: bajorath@bit.uni-bonn.de (J. Bajorath).

https://doi.org/10.1016/j.ailsci.2021.100022
Received 17 November 2021; Received in revised form 24 November 2021; Accepted 24 November 2021
Available online 27 November 2021

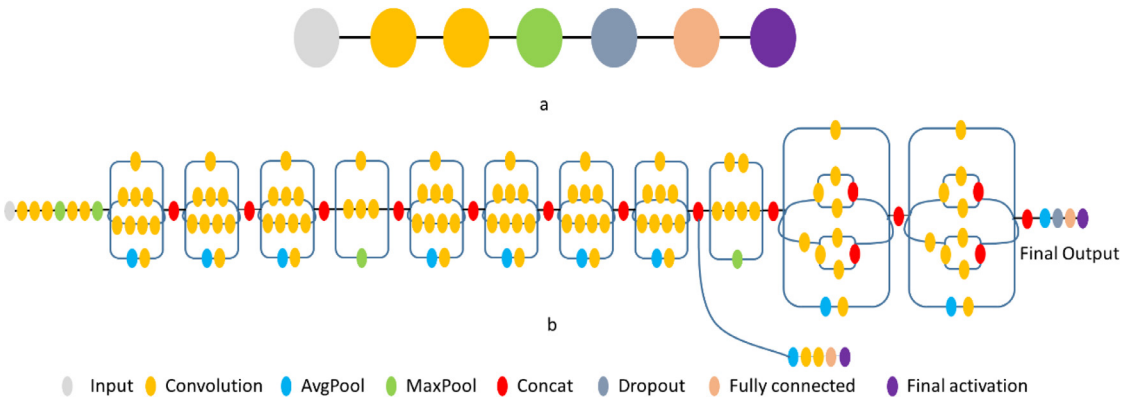

Fig. 1. Convolutional neural networks. The SCNN (top) and IV-3 (bottom) architectures are illustrated.

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 RDKit 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

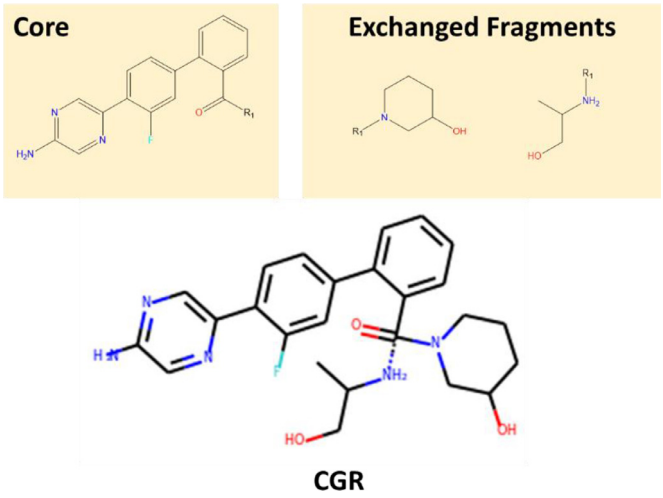

Fig. 2. Condensed graph of reaction. For an MMP consisting of a pair of compounds with a shared core and two exchanged fragments (top), the corresponding CGR representation is shown (bottom). The zero-order bond is indicated by a dashed line.

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.

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 (±0.03)	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).

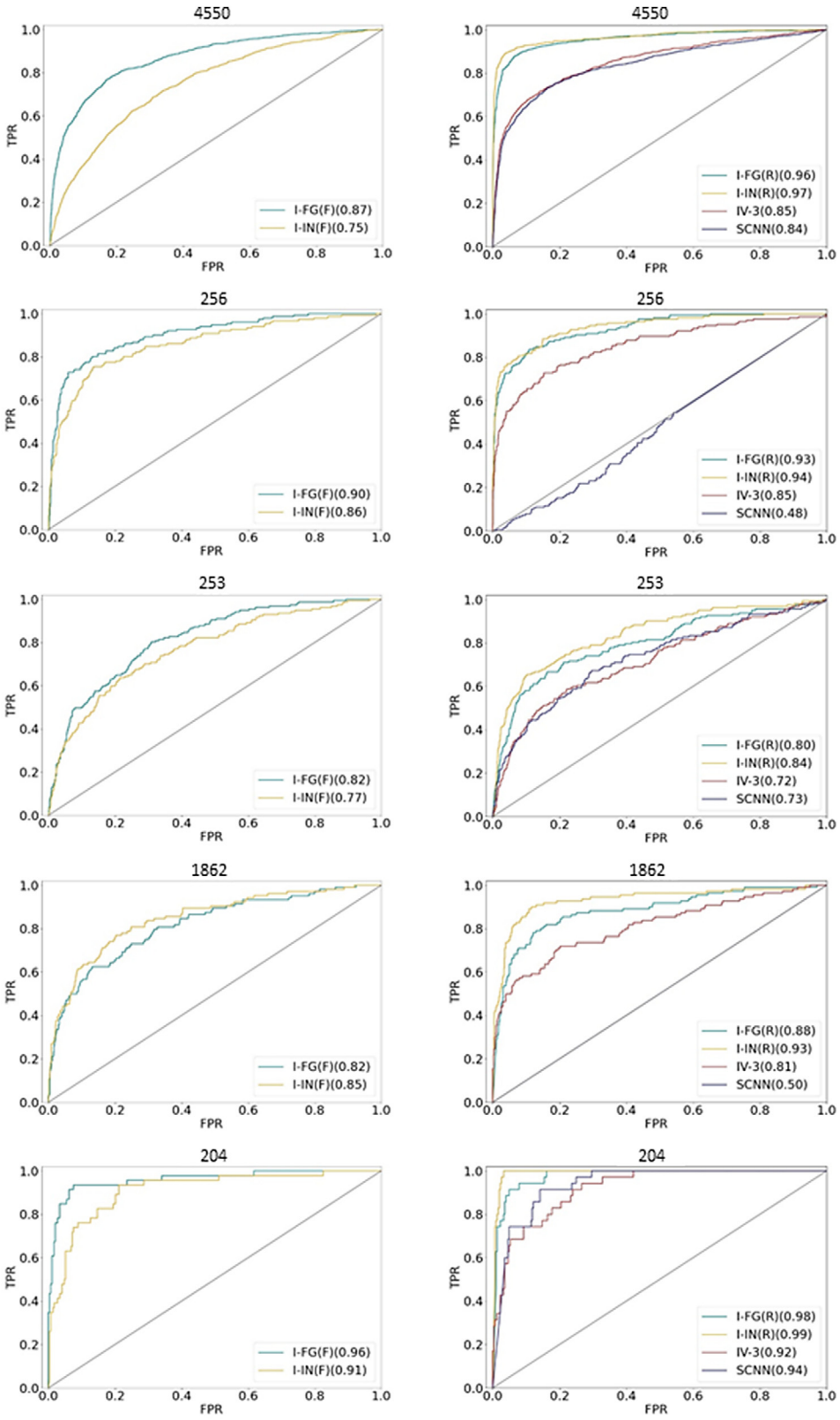

Fig. 3. Receiver operating characteristic curves. The performance of the best CNN models from an individual AC prediction trial is monitored in ROC curves for the different activity classes after fine-tuning (left column) and re-training (right). For each curve, AUC values are reported in parentheses.

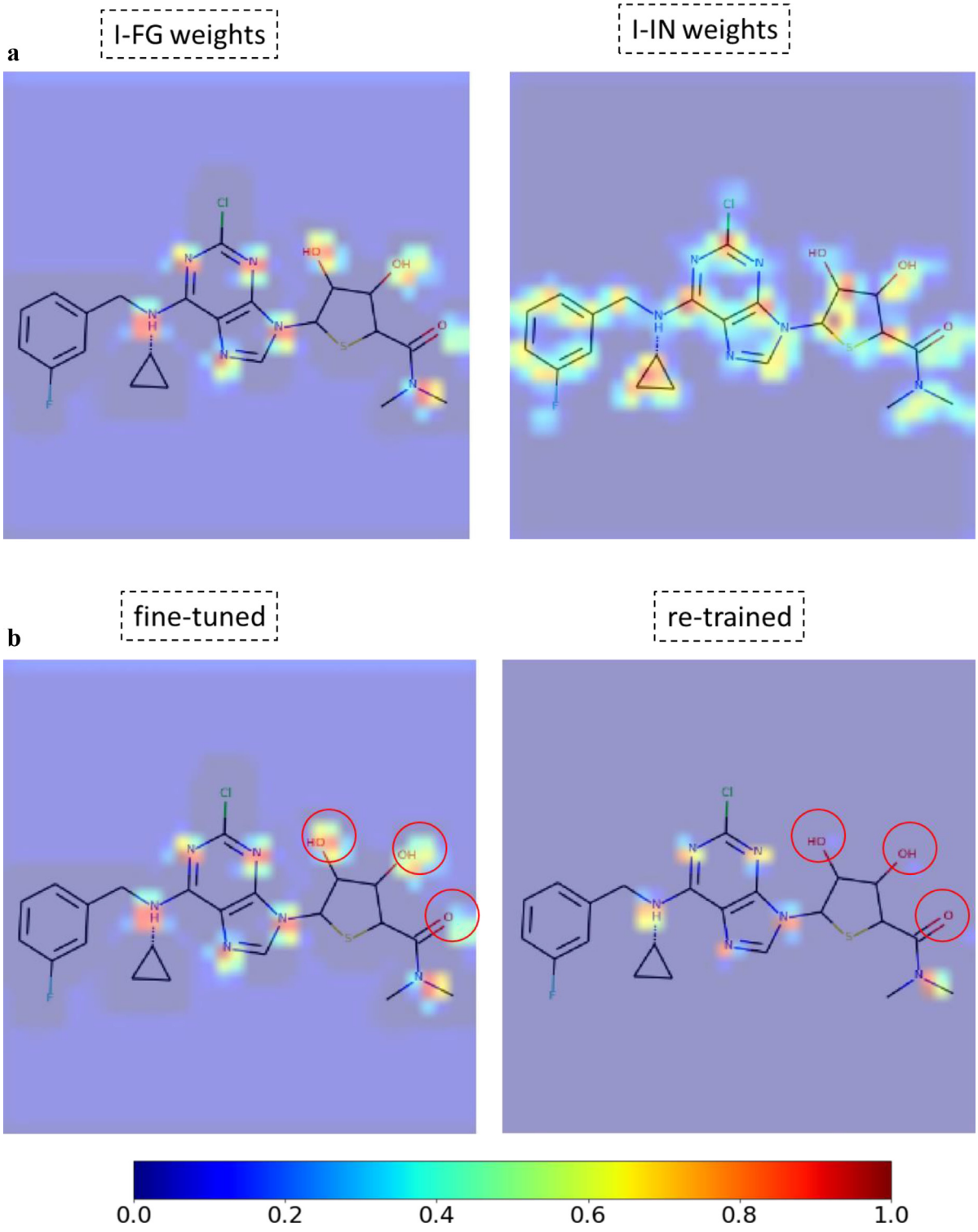

Fig. 4. Mapping of convolutional layer activation weights. Mean gradient weights of the fifth convolutional layer of different models are mapped to a CGR image of an exemplary MMP and displayed. The continuous color code from blue (0) over green (0.5) to red (1) indicates weights from 0 to 1. In (a), 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 (±0.02), I-IN: 0.81 (±0.17), IV-3: 0.74 (±0.02), and SCNN: 0.72 (±0.08) were obtained and mean MCC values of 0.77 (±0.02), 0.65 (±0.35), 0.56 (±0.02), and 0.47 (±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 (±0.04), I-IN: 0.70 (±0.03), IV-3: 0.56 (±0.02), and SCNN: 0.54 (±0.05) were obtained and mean MCC values of 0.38 (±0.03), 0.43 (±0.03), 0.19 (±0.05), and 0.12 (±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.