Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

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

Research Article

Interpretation of multi-task clearance models from molecular images supported by experimental design

Andrés Martínez Mora[a], Mickael Mogemark[a], Vigneshwari Subramanian[a], Filip Miljković[b,*]

[a] Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Gothenburg, Sweden
[b] Medicinal Chemistry, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

ARTICLE INFO

Keywords:
Metabolic stability
Convolutional neural networks
Model interpretability
Deep learning
Pharmacokinetics

ABSTRACT

Recent methodological advances in deep learning (DL) architectures have not only improved the performance of predictive models but also enhanced their interpretability potential, thus considerably increasing their transparency. In the context of medicinal chemistry, the potential to not only accurately predict molecular properties, but also chemically interpret them, would be strongly preferred. Previously, we developed accurate multi-task convolutional neural network (CNN) and graph convolutional neural network (GCNN) models to predict a set of diverse intrinsic metabolic clearance parameters from image- and graph-based molecular representations, respectively. Herein, we introduce several model interpretability frameworks to answer whether the model explanations obtained from CNN and GCNN multi-task clearance models could be applied to predict chemical transformations associated with experimentally confirmed metabolic products. We show a strong correlation between the CNN pixel intensities and corresponding clearance predictions, as well as their robustness to different molecular orientations. Using actual case examples, we demonstrate that both CNN and GCNN interpretations frequently complement each other, suggesting their high potential for combined use in guiding medicinal chemistry design.

Introduction

Machine learning (ML), as a part of a much larger artificial intelligence (AI) methodological spectrum, continues to gain considerable attention across a myriad of medicinal chemistry and drug discovery applications [1–4]. These methods are commonly employed to derive statistical quantitative-structure property relationship (QSPR) models for a range of different molecular properties, such as bioactivity [5], animal [6] and human [7] pharmacokinetics, and toxicity [8]. Typically, such models learn to associate structural patterns of available chemical structure, often encoded as predefined molecular features (descriptors) or graph-based embeddings, with a molecular property of interest obtained through established experimental procedures [9]. If properly validated, these models could potentially extrapolate learned molecular features with great accuracy and enable prospective application to new chemical series, which plays a major role in the progression of novel candidate molecules. In this respect, chemical transformations predicted to condition desirable molecular properties could be prioritized early on, hence shortening the length of design-make-test-analyze cycle.

In addition, the potential to not only accurately predict molecular endpoints, but also chemically interpret them, would be strongly preferred in medicinal chemistry practice. Clearly, model interpretability has gathered much interest in recent years as a methodological framework which aims to identify chemical modifications that correlate with dependent variables, resulting in observed experimental measurements [10–15]. Model explanations, effectively combined with expert knowledge, have the potential to guide rational drug design more conveniently, hence encouraging medicinal chemistry practitioners to readily apply *in silico* approaches. Thus, model-assisted hypothesis generation through the application of integrated or add-on interpretability architectures could further propagate the acceptance of computational methods in pharmaceutical sciences. Contrary to acquired attention, interpretable ML has only been very little practically explored, as most of the model development focuses on improving performance accuracy, yet with only a marginal success. For a more detailed discussion on explainable ML models for property prediction, readers are referred to a recent perspective article [16].

Deep learning (DL), as a methodological component of ML, has made significant progress in biological and medical sciences, particularly for image analysis where convolutional neural networks (CNNs) have fundamentally been applied [17,18]. Unsurprisingly, attempts have been made to repurpose pre-established CNN frameworks for the prediction of different compound properties based on molecular images [19–23].

[*] Corresponding author: Pepparedsleden 1, SE-43183 Gothenburg, Sweden. Phone: +46-73-699-3320.
E-mail address: filip.miljkovic@astrazeneca.com (F. Miljković).

https://doi.org/10.1016/j.ailsci.2022.100048
Received 27 September 2022; Received in revised form 7 November 2022; Accepted 1 December 2022
Available online 2 December 2022

In this regard, molecular images are utilized for representation learning to correlate implicit chemical information encoded in pixels with a learned experimental endpoint. Highly predictive models with on par accuracy to state-of-the-art structural representations were acquired, indicating that molecular image-based models have the potential to be applied across a variety of molecular property prediction tasks. However, notable performance advantage over simpler structure-based ML models is yet to be attained. A potential advantage of CNN models is in the abundance of interpretability methods that have become increasingly available recently, particularly for medical image analysis [24]. In medical imaging, it soon became evident that without model explanations there would be very little value in applying CNN methodology solely for predicting label outputs [24]. Therefore, efforts have been made to accommodate existing and introduce novel interpretability approaches for medical image applications, thus increasing practicality and confidence in utilizing CNN models [24]. Much like clinical practitioners, medicinal chemists too require model interpretations to aid the decision-making process, but also to expose hidden data trends that would otherwise remain unnoticed.

Whereas the use of interpretability approaches in medical imaging has a longstanding tradition (readers are referred to a recent review article [24]), the application on molecular images has only recently been considered. For example, Zhong *et al.* developed CNN models to predict compound rate constants towards OH radicals, which were then interpreted using gradient-weighted class activation mappings [25]. These interpretations correctly indicated structural features responsible for molecular reactivity [25]. In another work, Iqbal *et al.* studied molecular images to distinguish between close structural analogs forming activity cliffs (ACs) and those that did not [21]. Besides establishing that molecular images could be applied to predict ACs with high accuracy, the mapping of gradient weights from convolutional layers identified characteristic structural features that contributed to these predictions [21]. In the extension of this work, the authors learned frequently occurring functional groups from the images of AC/non-AC pairs to build transfer learning models with aim to improve the accuracy and interpretability of CNN models [26]. They demonstrated that the retraining of functional group prediction models optimized weights of the convolutional layer and further enhanced the prediction accuracy. In addition, they used visualizations to discuss the rationale for the performance improvement based on the ability of CNN models to learn functional group chemistry [26].

Previously, we systematically explored the application of multi-task CNN models to predict different *in vitro* clearance endpoints from molecular image representations [23]. We confirmed that CNNs were successful at capturing implicit chemical relationships contained in images using several increasingly challenging data splitting strategies. Moreover, model benchmarking with conventional ML and DNN methods trained with structure-based representations, as well as graph convolutional neural networks (GCNN), revealed on par or increased accuracy of CNNs with clear benefit of multi-task learning. Our findings indicated that CNN models built from molecular images could successfully be used to learn multiple clearance parameters [23]. Herein, we extend our research question to answer whether CNN models built from molecular images, as well as methodologically similar GCNN models, could be used to explain chemical modifications associated with the intrinsic metabolic reactions. Understanding mechanisms of drug metabolic transformations has wide implications on the development of small molecule therapeutics including, but not limited to, drug activity *in vivo* (pharmacodynamics), distribution and elimination processes (pharmacokinetics), as well as safety liabilities which could be attributed to dosing (potential accumulation effects), off-target pharmacology, or chemical reactivity. We illustrate the potential of our interpretability models in practice where we attempt to reconcile the model-derived interpretations with the experimentally confirmed products of metabolism obtained through the respective clearance assays. Our findings are presented in the following.

Methods and materials

Intrinsic clearance data

Detailed information regarding intrinsic clearance data used for CNN and GCNN model building, evaluation, and interpretability exploration is described elsewhere [23]. In brief, experimental data for human hepatocyte intrinsic clearance (HH CLint), human liver microsome intrinsic clearance (HLM CLint), rat hepatocyte intrinsic clearance (RH CLint), and dog hepatocyte intrinsic clearance (DH CLint) were extracted and curated from AstraZeneca internal sources. In total, 139,907 unique compounds were identified for model building and evaluation, with a varying number of available data points for different clearance parameters. For example, HLM CLint and RH CLint measurements, as most commonly measured clearance parameters upon compound synthesis, were available for 101,913 and 79,043 compounds, respectively, whereas HH CLint and DH CLint (which are evaluated comparatively later) were only available for 21,616 and 6747 compounds, respectively. For the purpose of model evaluation and interpretability, same random stratified 80/20 training-test set splits were used for both CNN and GCNN models [23].

Convolutional neural network models

Previously, we systematically explored the application of multi-task CNN models to accurately predict four commonly investigated intrinsic metabolic clearance parameters from molecular image representations [23]. CNNs comprise a number of convolutional blocks which extract features from adjacent image pixels, followed by the non-linear mappings and pooling operators that learn these features at different scales: in the context of molecular image recognition, shallow layers focus on atom-wise features whereas deeper layers on global structural patterns. Extracted features are then processed by a fully-connected block which combines matrix multiplications and non-linear mappings to provide final prediction values. Different CNN architectures, hyperparameter combinations, and image representations were explored to build final predictive models. We subjected the training set (80% of the total data) to three-fold cross-validation with the aim of selecting an optimal hyperparameter set. Early stopping with a patience of 15 epochs (of a maximum of 150) was applied during the cross-validation (two folds were used to train the models and the remaining fold to validate them) to alleviate potential overfitting. Coefficient of determination (R^2) and root-mean-square error (RMSE) were used to assess model performance, the latter being also applied during the cross-validation stage. Best performing CNN models contained ResNet [27] 152 architecture, two fully connected layers with a dropout rate of 0.50, batch size of 128, learning rate of 2×10^{-4}, Adam optimizer, batch normalization, and were optimized using Huber loss [28]. Additionally, we investigated several molecular image representations and selected 240 × 240 pixel embeddings of two-dimensional molecular structure, with a fixed resolution of 0.03 Å/pixel and color-coded rendition of heavy atoms (e.g., red for oxygen, blue for nitrogen, etc.) as the final model input. These models showed on par or increased accuracy compared to the state-of-the-art machine learning methods trained on structure- and graph-based fingerprints, with a reasonably good performance on the random test set: HH CLint ($R^2 = 0.57 \pm 0.003$, RMSE = 0.32 ± 0.001), HLM CLint ($R^2 = 0.57 \pm 0.001$, RMSE = 0.31 ± 0.001), RH CLint ($R^2 = 0.58 \pm 0.001$, RMSE = 0.36 ± 0.001), and DH CLint ($R^2 = 0.52 \pm 0.005$, RMSE = 0.36 ± 0.002) [23].

Graph convolutional neural network models

Together with CNNs, we also reported the performance of multi-task clearance models using GCNNs [23]. Here, the directed message-passing neural network framework named Chemprop [29] was applied, which combines feature extraction from molecular graph representations (atoms encoded as nodes and bonds as edges) and prediction potential in an end-to-end manner [29]. As a method conceptually close to CNNs, GCNNs displayed comparable performance in a multi-task setting. Optimal GCNN model configuration had four message-passing layers, GCNN layer size of 900, two feedforward layers with a dropout rate of 0.15, and RMSE as a loss function. For the same random test set, the GCNN models showed performance that was comparable to CNNs: HH CLint ($R^2 = 0.57 \pm 0.02$, RMSE = 0.31 ± 0.01), HLM CLint ($R^2 = 0.53 \pm 0.03$, RMSE = 0.33 ± 0.01), RH CLint ($R^2 = 0.58 \pm 0.02$, RMSE = 0.36 ± 0.01), and DH CLint ($R^2 = 0.48 \pm 0.02$, RMSE = 0.37 ± 0.01) [23]. This further qualified CNN and GCNN clearance models for their comparative interpretability assessment.

Interpretation of image-based models

We investigated several existing interpretability frameworks commonly applied in CNN tasks:

- Class activation maps (CAMs): Interpretations are supplied by propagating learned model gradients from each predicted clearance endpoint to the last layer of the convolutional backbone, resulting in a compressed low-dimensional representation of the interaction between learned gradients and an evaluated image [30]. The learned gradients can be conceived as the contributions of each image element to the final model prediction. Since the representation from the last layer of the convolutional backbone is low-dimensional, it is required to be upsampled to its original size by an interpolation algorithm. This yields coarse-colored patches that are then projected atop the corresponding structural features of the query image, which in turn supply the information describing increased or decreased clearance contributions. CAMs were implemented using PyTorch [31].
- Shapley additive explanations (SHAP): A model-agnostic method that uses approximation of Shapley values for the interpretation of different image pixels. Originating from the game theory, SHAP approximates the contribution [32] of each image pixel to the final prediction value derived by the CNN model. These values are then approximated as the weights of a local linear model built by the perturbation of query images, which are subsequently fit to the CNN outputs [33]. SHAP approach was implemented using Captum library [34].
- Integrated gradients (IGs): The method compares the gradients that are learned by the model trained on baseline images (e.g., images consisting of 'zero' pixels, images containing white noise, etc.) to those derived from the query images. The assessment is performed gradually by generating interpolations between the baseline and the query images, which allows for an in-depth evaluation of the gradients contributing the most to the predictions [35]. Gradients displaying the greatest variation are considered to supply the most relevant information concerning analyzed image pixels. As IGs explore image contributions at a pixel level, they provide similar resolution of detail for the obtained explanations (i.e., each pixel is individually interpreted). Similar to SHAP, IG approach introduces a degree of randomness to the resulting interpretations (even with a fixed random seed), although to a lesser extent than SHAP. To increase the robustness, the method was run four times and the outputs were averaged to derive final interpretations. IG method was applied using Captum library [34].
- Occlusion: An ablation method which substitutes patches of query images with noisy patches, thus evaluating the changes in the model output. Pixels showing the largest differences in the model output are the ones interpreted to contribute the most to the predictions [36]. The generated patches were 15 × 15 pixels in size, with a window displacement of 8 × 8 pixels, which yielded blurry non-local interpretations. Captum library was used to implement the method [34].

From the methods investigated, only CAMs and IGs were proceeded to the interpretability discussion. The other two methods (SHAP and Occlusion) were discarded for different reasons: SHAP output was highly influenced by the randomness of the perturbed images (four prediction repeats were notably different to average the final output) and the Occlusion method performed significantly slower compared to others, making its application the least practical.

The global assessment of CNN interpretation methods was performed by mathematically evaluating the relationship between the CAM/IG interpretations and the prediction values obtained by the model. For each predicted compound, mean pixel intensity (given as an average of image pixel values) of the interpretation layer (for the interpretability purposes superimposed on the molecular image) derived by the two methods were extracted for each clearance endpoint. For both CAM and IG methods, mean pixel intensities were normalized in a value range between 0 and 1. As IGs only provide interpretations for the pixels constituting molecular structure, only these pixels were considered for the evaluation. Pearson correlation coefficients between mean pixel intensities and prediction values were calculated to determine interpretation method sensitivity to the actual CNN predictions. Furthermore, we briefly discussed the impact of data augmentation (i.e., different rotations of molecular structure) on the robustness of CNN model predictions and interpretations.

Interpretation of graph-based models

Chemprop GCNN framework [29] contains a built-in interpretability module which relies on the Monte Carlo Tree Search (MCTS) algorithm. For each compound, MCTS algorithm iteratively evaluates the contribution of all possible molecular substructures to the final prediction, following a random order [37]. A maximum substructure size is defined as one of the hyperparameters to alleviate computational costs. Thus, the sampled substructural motif that best fulfills the size requirement is selected as the most relevant one. All graph nodes and edges analyzed for a particular substructure are adjacent to each other, which limits the output to a single continuous substructural pattern. This is a major difference in comparison to image-based approaches, which also poses a limitation given that metabolic transformation could take place at multiple, often disconnected, substructural positions. Unlike CAMs and IGs, MCTS does not provide information regarding the level of contribution for each highlighted molecular fragment. Pseudocode for both CNN and GCNN methods is included in Fig. S1 of the Supplementary methods.

Experimentally proposed metabolic products

Liver microsomes and hepatocytes can be applied to investigate potential oxidative (microsomes and hepatocytes) and conjugative (only hepatocytes) metabolic routes which may be observed in human and different animal species. Identification of metabolic products (metabolic identities or MetIDs) via *in vitro* systems plays a crucial role in validating the use of animal species for screening and synthesis planning, in order to protect metabolic hotspots that are directly relevant to man. On occasions, the samples prepared for intrinsic clearance measurements are submitted to the qualitative liquid chromatography-mass spectrometry (LC-MS) evaluation. For the analysis, both ultraviolet (UV) chromatograms and mass spectra are acquired. The UV chromatograms are applied to quantify each peak given as the percentage (%) of the run. Subsequently, the peaks representing >5% of the UV chromatogram are obtained by mass spectra, followed by the identification of tentative MetIDs. Herein, we aimed to correlate interpretations obtained by CNN and GCNN frameworks to the metabolic transformations proposed by the experiments.

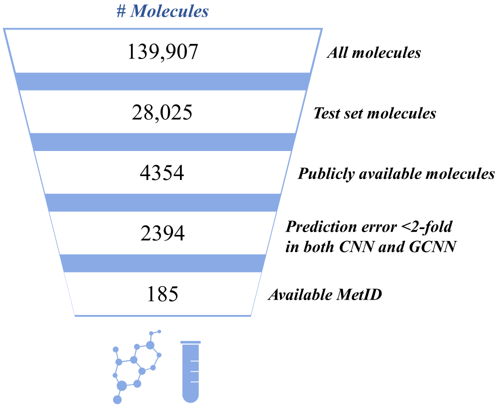


Fig. 1. Identification of molecules for interpretability assessment. A summary of the analysis is presented, as described in the text.

Results and discussion

Data for interpretability assessment

We have systematically searched the test set for publicly available compounds for which internally derived MetIDs were identified. A summary is shown in Fig. 1. As described above, both CNN and GCNN models were cross-validated and optimized using 80% of the available data (training set), whereas the remaining 20% (28,025 molecules) were used for model validation (test set) and model interpretability analysis. In addition, test set predictions were evaluated against the averaged pixel intensities for corresponding interpretations, as to evaluate the interpretation method sensitivity to the actual CNN predictions. Out of 28,025 molecules, 4354 (16%) had their chemical structure publicly disclosed. From these, we selected 2394 compounds (55%) that were predicted within a two-fold error margin for both CNN and GCNN models in at least a single common clearance endpoint. Two-fold prediction error (0.30 logarithmic units) was chosen to reflect the experimental error of both hepatocyte and microsomal clearance assay measurements [23]. Furthermore, we only examined compounds with internally available MetIDs for the corresponding clearance assays. Following these criteria, only <1% of the initial test set compounds (185 molecules) qualified for our analysis. From these, we selected several examples to discuss, one for each predicted clearance endpoint.

Performance robustness of CNN and GCNN models

To evaluate performance robustness on novel chemical series, we applied Bemis-Murcko (BM) scaffold definition to estimate the chemical overlap between randomly defined training and test sets [38]. We identified 64,393 unique BM scaffolds in our data, 6328 of which were shared between the training and the test set, which corresponded to 54,292 and 17,757 compounds, respectively. Such high overlap was expected given the stochastic nature of random splitting, which may as well limit the potential of machine learning models to confidently predict new chemistry. In that respect, we analyzed the model performance only on singleton compounds (10,268 molecules) that consituted the test set (i.e., molecules that had no other analogs according to BM scaffold definition). More challenging data splitting strategies for our models were previously examined elsewhere [23]. Fig. S2 displays the CNN and GCNN model performance for each individual clearance endpoint as evaluated on all and only singleton test set molecules. Performance differences between all and only singleton compounds were insignificant (up to 0.03, depending on the endpoint), demonstrating that the models were capable of accurately predicting compounds that belonged to unseen chemistry. Provided that none of the 185 molecules which qualified for the interpretability analysis were singletons, we could not demonstrate that the interpretations obtained through studied interpretability frameworks would correlate with their MetID experiments. However, given that the models were predictive for singleton molecules, it cannot be entirely excluded that derived interpretability models would correctly interpret novel chemistry as well.

In addition, we evaluated the prediction robustness relative to the uncertainty estimates for the compounds satisfying the above mentioned criteria. Uncertainty estimates for CNNs were obtained via rotational perturbations (different orientations of the same molecule were used as inputs), whereas for GCNNs the uncertainties were retrieved using three models with different seed values trained under the same conditions. Fig. S3 shows box plot distribution of standard deviations for CNN and GCNN predictions across different clearance endpoints. We noted relatively consistent uncertainty distributions for the discussed clearance measurements when focusing on individual modeling frameworks. Prediction uncertainties were slightly higher for GCNN models, with a wider distribution of standard deviations compared to the equivalent CNN models. Median values for both CNN and GCNN distributions were low, demonstrating that individual predictions were generally robust. Further investigation of different calibration approaches using experimental error estimates would likely be required, which is an ongoing area of interest for practical applications.

Sensitivity of CNN interpretation methods

We conducted a global sensitivity analysis of the CNN interpretation methods to the corresponding prediction values. For this purpose, normalized mean pixel intensities for each interpreted compound were correlated with model predictions obtained for each endpoint. Fig. 2 demonstrates a strong correlation across different clearance endpoints and interpretation methods, confirming that the regions higlighted by interperetation methods are sensitive to the order of magitude of predicted values. This correlation is slightly stronger in CAMs, thus making this method more sensitive. Curiously, CAMs show strong negative correlation (as opposed to IGs) across all endpoints, which could be attributed to the latent space used to build CAMs. It is thus plausible that during the CNN training the CLint values are mapped to the low latent space values, which in turn produces negative correlation in CAMs. Other potential reasons include the application of high degree interpolation algorithms to resample latent CNN dimensions with respect to the original image dimensions, as well as the color map selection where low pixel intensities are assigned to high CLint values and vice versa.

We furthermore investigated the sensitivity of CNN interpretation methods to different molecular orientations. As our models were initially built using different augmentation techniques and were robust to a number of molecular image augmentations, we reasoned that investigated CNN interpretation methods should show similar level of robustness as the models supporting them. The ability of CNN interpretations to show rotational invariability to different molecular orientations would demonstrate their resilience to perturbation attacks in a practical setting (e.g., when a molecular structure is requested to be displayed in a particular pose). Other perturbation approaches such as image blurring, application of occlusion masks, or replacement of image patches would be less meaningful to consider given that the preparation of molecular images follows a set of predefined rules (see Convolutional neural network models section), unlike in other image analysis applications where often raw and quality deprived images are considered. As shown in Figs. S4-S5 of the Supplementary material both the predicted clearance values and the structural features highlighted by the interpretations are consistent across different rotations. Clearly, the confirmation of prediction reliability and interpretation consistency establishes a strong prerequisite for the practical consideration of molecular image-based methods in medicinal chemistry, as demonstrated in the following section.

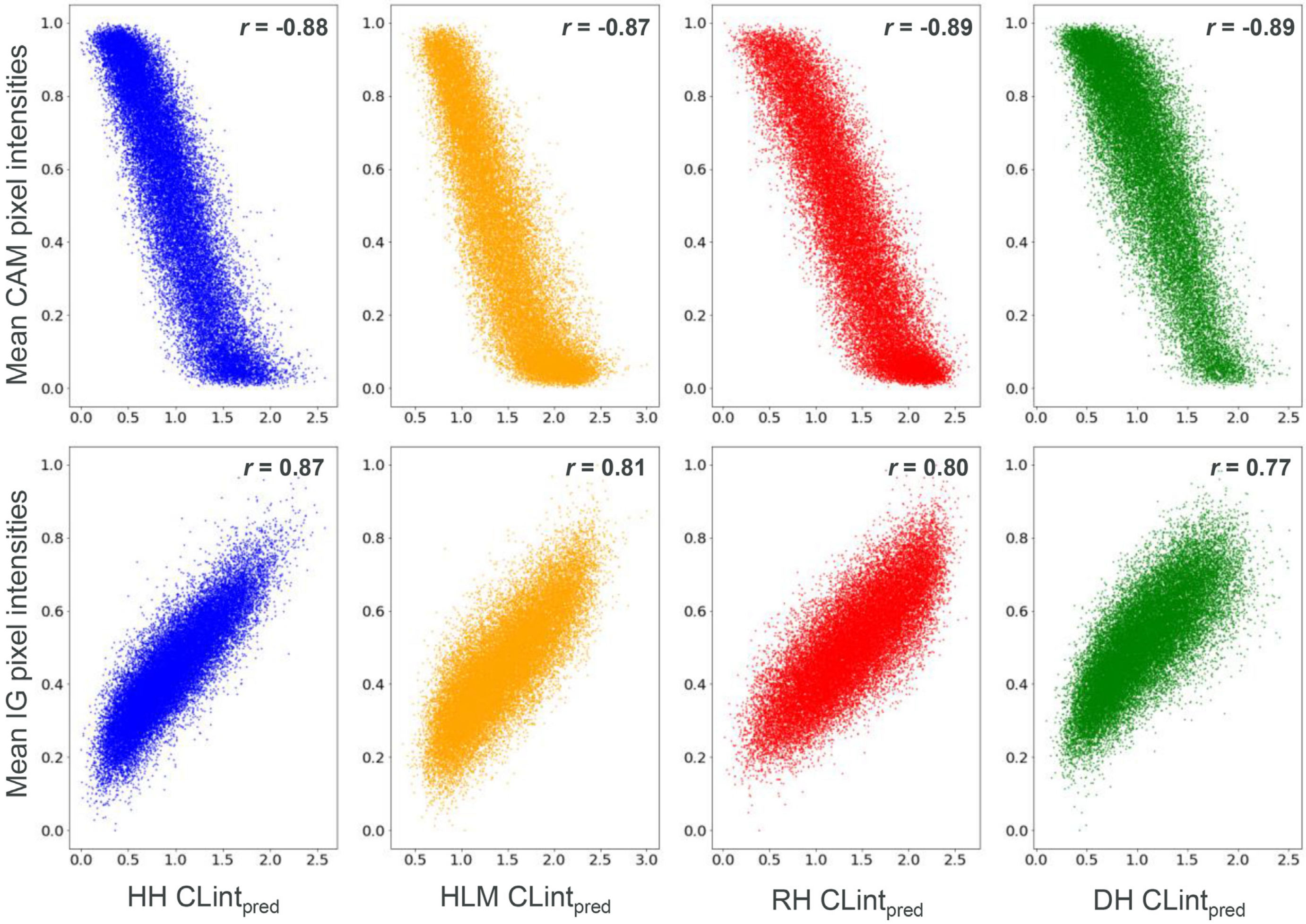


Fig. 2. Global sensitivity of CNN interpretation methods. Scatter plots show the correlation between the normalized mean pixel intensities and predicted CLint values (HH CLint, blue; HLM CLint, orange; RH CLint, red; DH CLint, green) for both CNN interpretation methods (CAMs, top; IGs, bottom). At the top right corner of each plot a Pearson's r value quantifies the correlation.

Table 1
Agreement between different interpretability methods as indicated by Spearman correlation coefficients.

Method pair	HH CLint	HLM CLint	RH CLint	DH CLint
CAM – IG	0.10	0.06	0.05	0.09
CAM – MCTS	0.43	0.56	0.05	0.40
IG – MCTS	0.02	0.04	0.02	0.01

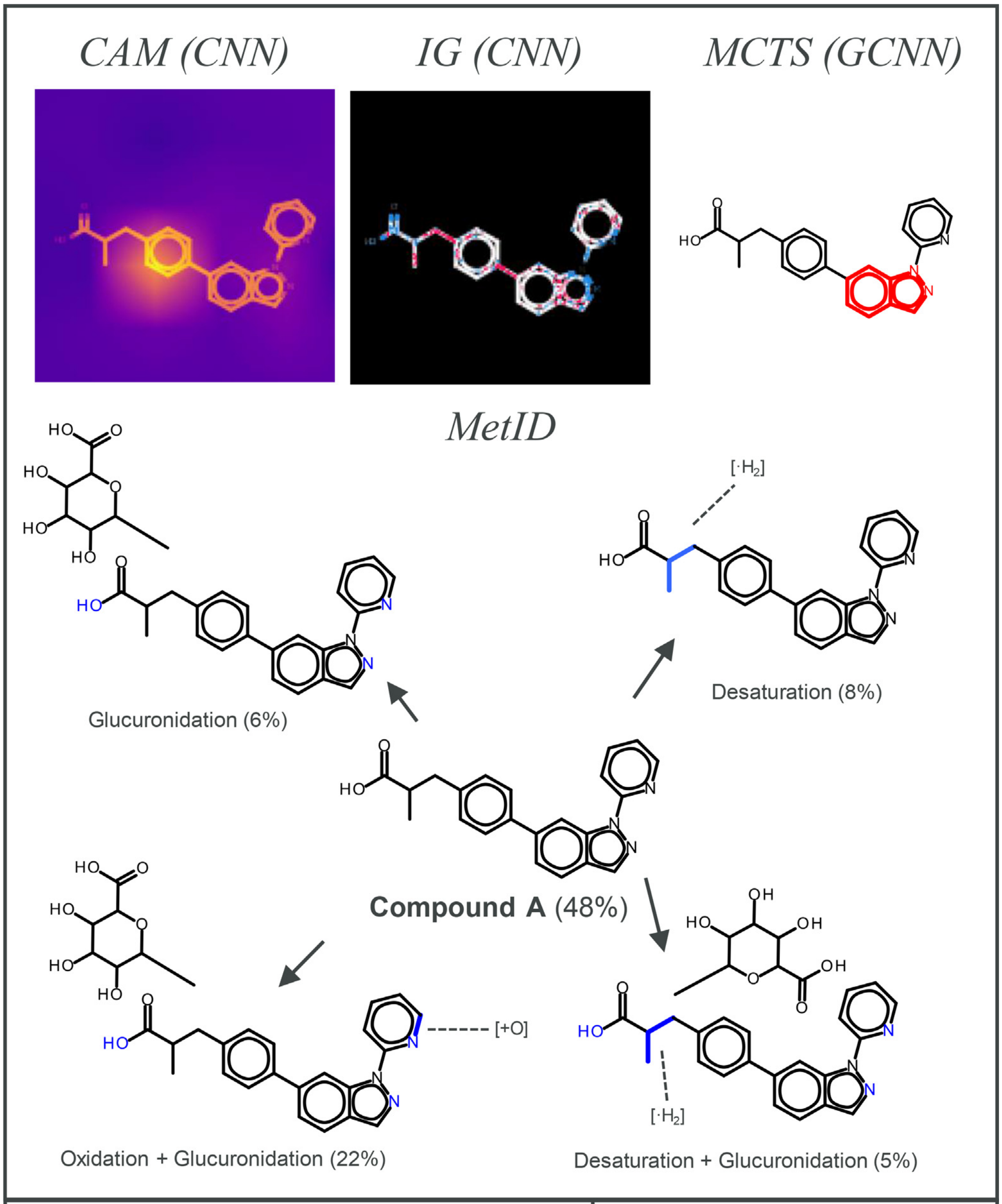
Agreement between interpretability methods

For the compounds fulfilling our criteria, we investigated the agreement between the three interpretability methods (CAMs, IGs, and MCTS) by extracting the pixel values from their interpretation maps and calculating pairwise Spearman correlation coefficients. The greater the Spearman correlation coefficient, the greater the agreement between the methods, thus demonstrating their ability to highlight same substructural motifs. Provided that MCTS method was herein applied to molecular graphs, we built an equivalent image representation for MCTS outputs where highlighted portions of molecular structure were assigned pixel values of 1, whereas the remaining (i.e., no contribution according to MCTS) were kept at 0. Table 1 summarizes Spearman correlation coefficients obtained for the pairwise method comparisons, as assessed across all four clearance endpoints. Despite being applied on different data formats and model architectures, CAMs and MCTS showed the highest correlation among all method pairs. The greatest disagreement was noted between IGs and MCTS. Endpoint-wise, RH CLint showed the greatest variability between the method pairs, scoring overall the lowest coefficient values. Altogether, Spearman correlation coefficients showed no meaningful association for the majority of methods and clearance endpoints, indicating that their interpretations could be at most complementary when applied together.

Interpretation examples

To simplify model interpretability analysis, each interpretation method was programed to yield an interpretation layer characterized with a pre-specified coloring scheme, that would be then subsequently superposed onto the predicted molecular image. For CAMs and IGs (CNN model), continuous color scales were applied to indicate increasing propensity for metabolic transformations: from dark blue to orange for CAMs, and from blue to red for IGs. For MCTS (GCNN model), only the most contributing molecular substructure for each compound was considered and highlighted, which also posed a limitation for this method. Exemplary interpretations for different compounds and clearance endpoints are displayed in Figs. 3–6. Both the experimental and predicted clearance values shown in the figures are given in a logarithmic scale.

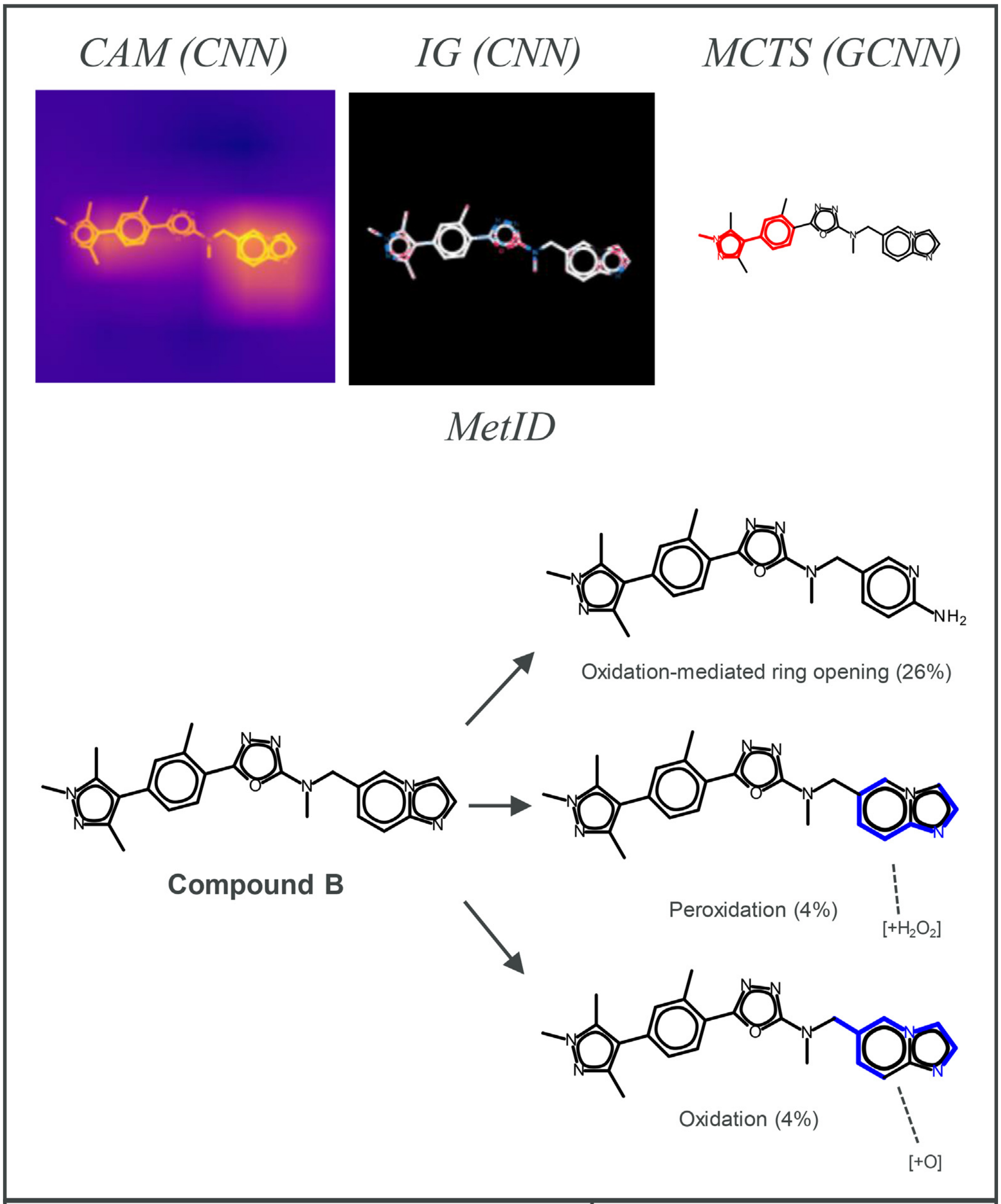
As shown in Fig. 3, compound A has a moderately high experimental HH CLint of 1.25 and was accurately predicted by both the CNN and GCNN models, displaying only a marginal error difference of 0.06 and



HH $CLint_{exp}$	1.25
HH $CLint_{pred}$ (CNN)	1.31 ± 0.09
HH $CLint_{pred}$ (GCNN)	1.09 ± 0.15

Fig. 3. Interpretation of HH CLint prediction for compound A. Shown are CNN and GCNN model interpretations obtained for the predicted HH CLint endpoint for compound A. Interpretations are complemented with the full MetID scheme, as well as the experimental and the predicted endpoint values for each model.

0.16, respectively. Experimentally obtained HH CLint MetID for compound A shows that 48% of the molecule remained unchanged (was not subjected to HH CLint metabolism), whereas 41% were metabolically transformed to plausible products. The remaining 11% could not be chemically characterized by the LC-MS procedure. Major metabolites were derived through the successive transformation of oxidation and glucuronidation (22%), only desaturation (8%), only glucuronidation (6%), and a combination of both desaturation and glucuronidation (5%). Both N- (nitrogens of indazole and pyridine) and O-glucuronidation (hydroxyl group of carboxylic acid) reactions were detected, whereas the sp^3 carbons found in the aliphatic chain were subjected to desaturation. Oxidation process was identified to take place on the pyridine ring. For the CNN model interpretations, functional groups undergoing glucuronidation show the lowest propensity for the metabolic transformation (both CAM and IGs) (Fig. 3). On another hand, area around the aliphatic chain (adjacent to the benzene ring) is characterized with greater pixel intensity by both CAM and IG methods, although this transformation was experimentally quantified to contribute the least to the metabolic output. Oxidation of the pyridine ring is highlighted as having an average intensity by CAM and a slightly higher intensity by IGs. As for

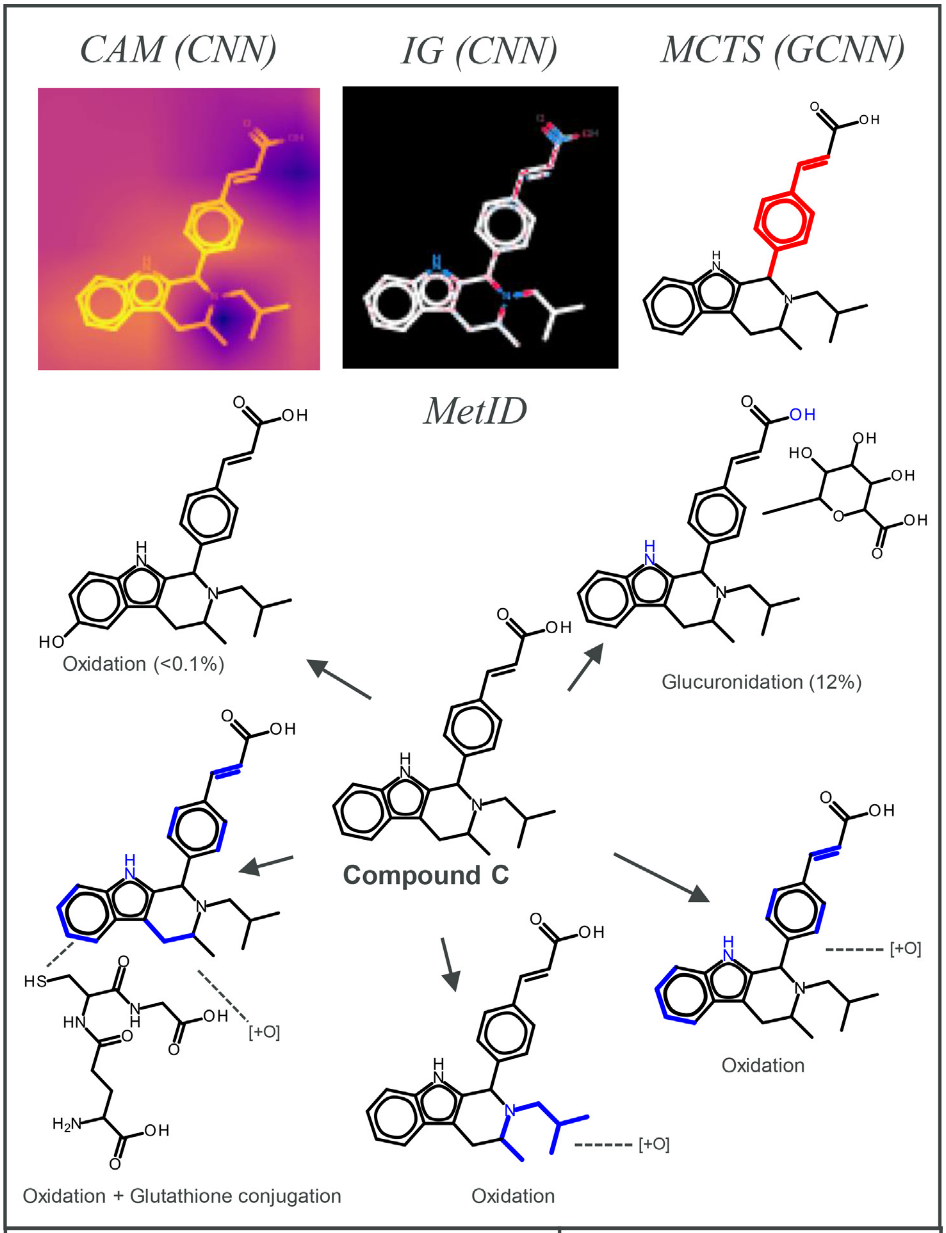


HLM CLint$_{exp}$	2.03
HLM CLint$_{pred}$ (CNN)	1.97 ± 0.02
HLM CLint$_{pred}$ (GCNN)	1.91 ± 0.03

Fig. 4. Interpretation of HLM CLint prediction for compound B. Shown are CNN and GCNN model interpretations obtained for the predicted HLM CLint endpoint for compound B. Interpretations are complemented with the full MetID scheme, as well as the experimental and the predicted endpoint values for each model.

the GCNNs, MCTS algorithm highlights indazole ring as the most contributing substructure to the prediction which, according to the available MetID, was conjugated via N-glucuronidation. Carboxylic acid alcohol, suggested by MetID to undergo glucuronidation, has not been highlighted by the interpretability methods. However, LC-MS does not distinguish between the potential sites of transformation or whether O- or N-glucuronidation takes place. In this example, all three interpretability approaches were complementary to each other, with not a single method highlighting all experimentally detected metabolic transformations. For HH CLint, this compound shared the same BM scaffold with another 16 compounds, 12 of which were in the training set.

Fig. 4 displays compound B [39] which has a high experimental HLM Clint (HLM CLint$_{exp}$ = 2.03) and comparably predicted CNN (HLM CLint$_{pred}$ = 1.97) and GCNN (HLM CLint$_{pred}$ = 1.91) model values. For this compound, HLM CLint MetID shows that only 34% of the metabolic output is chemically characterized. This can further be disseminated into

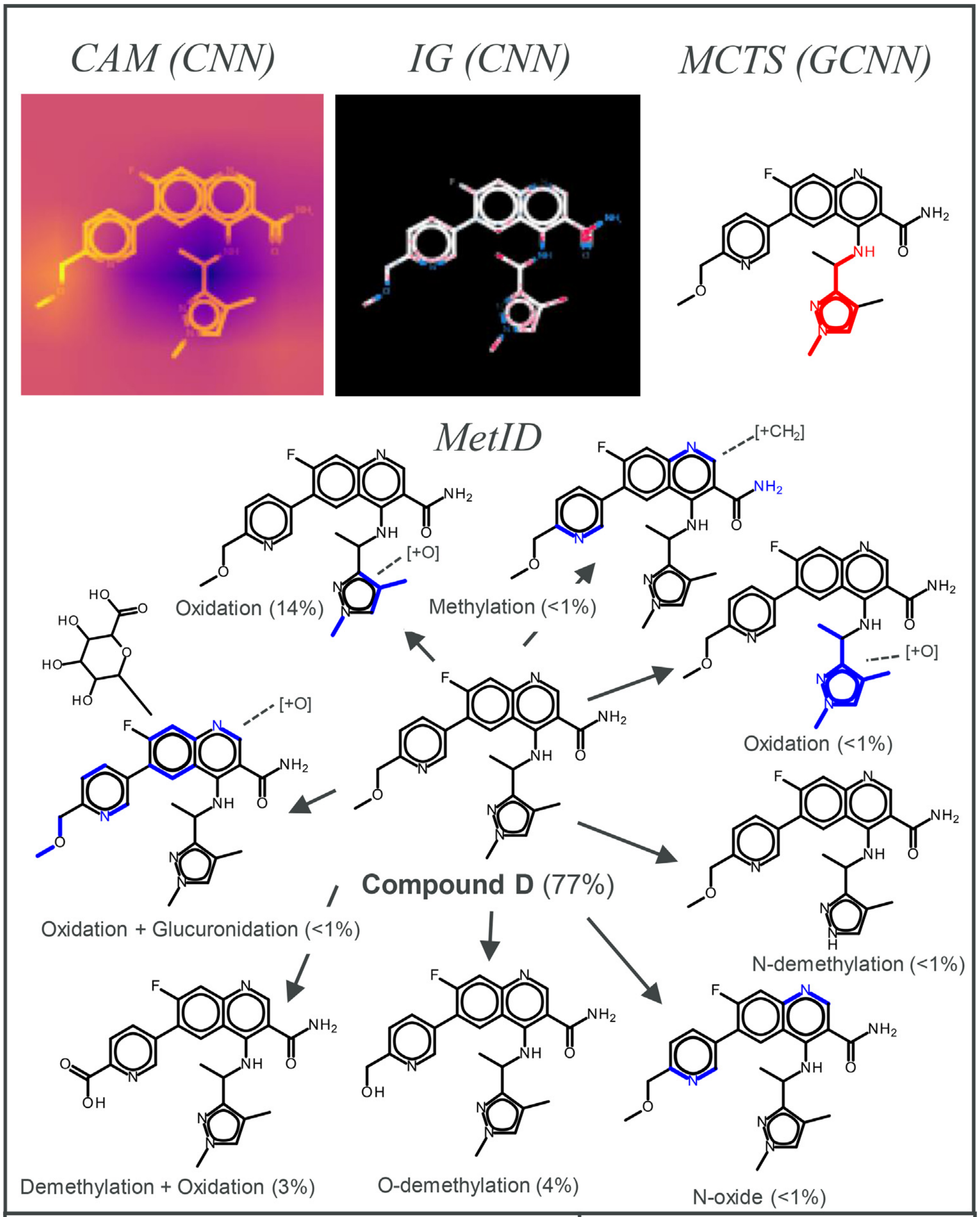


RH $CLint_{exp}$	0.99
RH $CLint_{pred}$ (CNN)	1.23 ± 0.01
RH $CLint_{pred}$ (GCNN)	1.29 ± 0.05

Fig. 5. Interpretation of RH CLint prediction for compound C. Shown are CNN and GCNN model interpretations obtained for the predicted RH CLint endpoint for compound C. Interpretations are complemented with the full MetID scheme, as well as the experimental and the predicted endpoint values for each model.

the oxidation-mediated ring opening (26%), free radical peroxidation (4%), and oxidation (4%) mechanisms, all detected for the imidazopyridine ring. Indeed, oxidative opening of the imidazopyridine ring was predictable since structurally similar indoles undergo similar oxidative cleavage via the P450-mediated oxidation [40]. Both CAM and IGs correctly highlight imidazopyridine moiety (the latter focusing on the imidazole portion) as a metabolic hotspot, which accounts for all three identified metabolic products. In addition, IGs predict pyrazole and oxadiazole as additional structural motifs undergoing metabolism. On another hand, MCTS (GCNN model) suggests that phenylpyrazole moiety is the major contributing feature to the compound B metabolism. However, it cannot be entirely excluded that among the metabolites unidentified by LC-MS occurs a byproduct (or several) generated via these substructures. For HLM CLint, the BM scaffold of this compound was contained in another 29 compounds, 18 of which were part of the training set.

In Fig. 5, compound C shows a relatively low experimental RH CLint value (RH $CLint_{exp}$ = 0.99), which was slightly overpredicted by both CNN (RH $CLint_{pred}$ = 1.23) and GCNN (RH $CLint_{pred}$ = 1.29) models. Here, MetID clearly displays both N- and O-glucuronidation as the major quantifiable metabolites (12%) whereas the oxidation of multiple



DH CLint$_{exp}$	0.78
DH CLint$_{pred}$ (CNN)	0.73 ± 0.02
DH CLint$_{pred}$ (GCNN)	0.88 ± 0.14

Fig. 6. Interpretation of DH CLint prediction for compound D. Shown are CNN and GCNN model interpretations obtained for the predicted DH CLint endpoint for compound D. Interpretations are complemented with the full MetID scheme, as well as the experimental and the predicted endpoint values for each model.

functional groups and aromatic rings was detected but not quantified by LC-MS. Additionally, the transformation coupling oxidation and glutathione conjugation was noted as well. CAM correctly highlights a large part of the tricyclic ring as the potential site of metabolism, which was confirmed by the MetID to undergo oxidation, glucuronidation, and glutathione conjugation. On another hand, IGs characterizes tricyclic nitrogens with a low propensity for metabolic transformations, whereas the area around the tertiary amine of the tetrahydroisoquinoline ring is highlighted as a region of high metabolic propensity. In the latter, the activated Cα is a well-known metabolic soft spot for compound classes such as the displayed tetrahydroisoquinoline and other fused aryl piperidines. These are particularly vulnerable to oxidative metabolism (as indicated by the MetID), which may lead to the formation of highly reactive and toxic iminium ions [41]. In addition, CAM correctly highlights the benzene ring (oxidation according to the MetID), that is only in part detected by IGs. For CAM interpretation, an average pixel intensity is shown for the isobutyl group and the sp^2 carbons of the Michael acceptor, all undergoing oxidation. The oxidation of the sp^2 carbons was also confirmed by IGs. Interestingly, unlike in Fig. 3, hydroxyl group of the carboxylic acid which participates in the glucuronide conjugation is correctly highlighted by IGs, whereas in CAM it is shown to have only a minor propensity for metabolic transformation. MCTS algorithm highlights only 4-methylstyrene as the most contributing molecular substructure to the GCNN prediction, which was suggested by MetID to undergo oxidative metabolism. The majority of metabolic transformations were detected by CAM, which was complemented only to a minor extent by IGs. For RH CLint, this compound shared its BM scaffold with another 451 compounds, 359 of which were present in the training set.

Fig. 6 discusses compound D which has a low DH CLint of 0.78 and accurately predicted values by both CNN (DH $CLint_{pred}$ = 0.73) and GCNN (DH $CLint_{pred}$ = 0.88). A major portion of the available parent compound remained relatively unchanged (77%). Main metabolic transformations are reported for the pyrazole ring (oxidation, 14%) and ether functional group (demethylation, 4%), the latter being additionally transformed by oxidation to a carboxylic acid moiety (3%). Each of the remaining reported metabolic reactions accounted for <1% of presence (methylation, oxidation, N-demethylation, N-oxide formation, and oxidation/glucuronidation). The highest pixel intensity in CAM is reported for the ether group, which undergoes O-demethylation, oxidation, and glucuronidation. This is only insignificantly highlighted by IGs. However, IGs correctly interpret pyrazole ring as the key metabolic hotspot (oxidation and N-demethylation), which is by CAM highlighted with only an average intensity. In MCTS, a structurally extended pyrazole pattern is highlighted as the most contributing molecular feature, in accordance with the major reported metabolite. N-methylation of pyridine, quinoline, and amide nitrogen are not captured by either CAM or IGs. Similar observations are made for aryl oxidation and N-oxide formation. Here, IGs and MCTS correctly predict the location of the main metabolic transformation, whereas CAM complement interpretation findings with the detected modifications on the ether. In addition to DH CLint MetID, compound D was also characterized with MetIDs for HH CLint and RH CLint experiments, which are provided in the Supplementary material (Figs. S6-S7). In short, similar metabolic transformations (yet in different ratios) were present for both HH CLint (Fig. S6) and RH CLint (Fig. S7), with 86% of the parent compound D preserved in HH CLint and 47% in RH CLint. However, we noticed that the same CNN interpretation methods highlighted different substructural features for all three MetID-containing clearance endpoints (Figs. 6, S6-S7). For HH CLint (Fig. S6), CAM method assigns the greatest pixel intensity to the quinoline ring, although it accounted only for 1% of the detected compounds. No clear substructural features were prioritized for IGs. On the other hand, MCTS correctly recognizes pyrazole as the major metabolic hotspot (oxidation and methylation). Interestingly, for RH CLint (Fig. S7) all three methods correctly predict pyrazole ring as the major substructural motif undergoing RH metabolism, orchestrated via numerous transformative mechanisms. Besides this compound, we detected another two molecules with DH CLint measurements sharing the same BM scaffold, both present in the training set.

Conclusions

Herein, we have explored several model interpretability frameworks which build upon the existing multi-task CNN and GCNN models that learn from in-house intrinsic clearance data. For this purpose, we applied CAMs and IGs to interpret CNN predictions and a built-in MCTS algorithm to explain Chemprop GCNN outputs. For an adequate interpretability assessment, only publicly disclosed compounds that were accurately predicted by both CNN and GCNN models and were internally reported with chemically characterized MetIDs were considered for the analysis. This permitted us to draw relatively clear conclusions on the applicability of CNN- and GCNN-based interpretations to explain metabolism-related chemical modifications obtained through the experimental design. Despite being partially trained and tested on the common substructures, the models proved to be robust on unseen chemistry. Only a slight increase in RMSE values was detected when compared to all test set predictions.

Provided that MetID experiments are only rarely performed (largely for few project candidates) and that only a handful of examples could be efficiently interpreted, large scale interpretability assessment which considers in-house-generated MetIDs would offer only a limited insight into the general interpretation capabilities of such models. For a more comprehensive analysis, a more frequent generation of MetID experiments (often limited by costs), or a collaborative resource framework bringing together large pharmaceutical companies, would be required. However, we have still demonstrated that mean pixel intensities associated with interpretation layers were highly correlated with CNN predictions. In addition, there was only a little overlap between substructures highlighted by different methods, suggesting that these interpretability frameworks could potentially generate complementary information for the metabolic hotspot assessment. Furthermore, these methods displayed strong interpretation robustness to different orientations of molecular images, supported by consistently highlighted substructural features in these images. On a case-by-case basis, no clear preference could be given to any of the evaluated interpretability methods, as it was frequently observed that both CNN and GCNN interpretations were complementary to each other. Interestingly, CAM and IG algorithms often provided different interpretations, which could be attributed to their global- and local-enabled foci, respectively. In addition, model interpretations frequently highlighted structural features that were not identified by MetIDs. However, provided that LC-MS outputs are often not chemically characterized in their entirety, it cannot be altogether excluded that a part of the unidentified metabolic output contains products that result from model-associated substructural motifs. For these reasons, the application of these methods could be envisioned in concert to guide medicinal chemists' intuition towards or away from plausible metabolic transformations. For example, once the predicted clearance values are obtained, a(n) (experienced) medicinal chemist may begin to build confidence in a particular interpretability method which, according to the chemist's experience, provides plausible explanations for an investigated drug chemotype. This would also prompt the application to other highlighted areas of chemical structure that are less straightforward to deconvolute but may lead, for instance, to rarely observed metabolic transformation associated with safety liabilities. Such suspicions may later be reconciled by an extensive literature search. In cases where less knowledge about particular chemical series exists, conjoined application of two or more interpretability frameworks, either through weighted scoring (where overlapping features would receive higher scores) or a simple combination of highlighted substructures would be recommended. Provided that metabolic modifications could simultaneously take place at different, often disconnected substructural motifs, and that some interpretability modules frequently highlight only a single structural locus, combined use of different interpretation methods would secure better coverage of potential transformation sites. These are just some of the potential application scenarios which merit our further investigation and interest. That being said, potential future directions extending this work include, but are not limited to: a) improvement of model performance with respect to experimental error of measurement, b) further exploration of interpretability methods and the development of custom-made algorithms that focus on molecular image representations, c) detailed guidance on use and best practices, including when and how to apply interpretation methods alone or in combination, and d) continuous application and prospective validation on ongoing medicinal chemistry projects.

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.2022.100048.

References

[1] Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. Drug Discov. Today 2015;20:318–41.

[2] Lo Y, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. Drug Discov. Today 2018;23:1538–46.

[3] Miljković F, Rodríguez-Pérez R, Bajorath J. Impact of artificial intelligence on compound discovery, design, and synthesis. ACS Omega 2021;6:33293–9.

[4] Rodríguez-Pérez R, Miljković F, Bajorath J. Machine learning in chemoinformatics and medicinal chemistry. Annu. Rev. Biomed. Data Sci. 2022;5:43–65.

[5] Rodríguez-Pérez R, Bajorath J. Multitask machine learning for classifying highly and weakly potent kinase inhibitors. ACS Omega 2019;4:4367–75.

[6] Obrezanova O, Martinsson A, Whitehead T, Mahmoud S, Bender A, Miljković F, Grabowski P, Irwin B, Oprisiu I, Conduit G, Segall M, Smith GF, Williamson B, Winiwarter S, Greene N. Prediction of in vivo pharmacokinetic parameters and time–exposure curves in rats using machine learning from the chemical structure. Mol. Pharm. 2022;19:1488–504.

[7] Miljković F, Martinsson A, Obrezanova O, Williamson B, Johnson M, Sykes A, Bender A, Greene N. Machine learning models for human in vivo pharmacokinetic parameters with in-house validation. Mol. Pharm. 2021;18:4520–30.

[8] Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: toxicity prediction using deep learning. Front. Environ. Sci. 2016;3:e80.

[9] David L, Thakkar A, Mercado R, Engkvist O. Molecular representations in AI-driven drug discovery: a review and practical guide. J. Cheminform. 2020;12:e56.

[10] Hansen K, Baehrens D, Schroeter T, Rupp M, Müller K-R. Visual interpretation of kernel-based prediction models. Mol. Inform. 2011;30:817–26.

[11] Balfer J, Bajorath J. Introduction of a methodology for visualization and graphical interpretation of Bayesian classification models. J. Chem. Inf. Model. 2014;54:2451–68.

[12] Balfer J, Bajorath J. Visualization and interpretation of support vector machine activity predictions. J. Chem. Inf. Model. 2015;55:1136–47.

[13] Polishchuk P. Interpretation of quantitative-structure activity relationship models: past, present, and future. J. Chem. Inf. Model. 2017;57:2618–39.

[14] Rodríguez-Pérez R, Bajorath J. Interpretation of compound activity predictions from complex machine learning models using local approximations and Shapley values. J. Med. Chem. 2020;63:8761–77.

[15] Rodríguez-Pérez R, Bajorath J. Interpretation of machine learning models using Shapley values: application to compound potency and multi-target activity predictions. J. Comput.-Aided Mol. Des. 2020;34:1013–26.

[16] Rodríguez-Pérez R, Bajorath J. Explainable machine learning for property predictions in compound optimization. J. Med. Chem. 2021;64:17744–52.

[17] Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Ann. Rev. Biomed. Eng. 2017;19:221–48.

[18] Moen E, Bannon D, Kudo T, Graf W, Covert M, Van Halen D. Deep learning for cellular image analysis. Nature Meth 2019;16:1233–46.

[19] 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.

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

[21] 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;35:1157–64.

[22] Yoshimori A. Prediction of molecular properties using molecular topographic map. Molecules 2021;26:e4475.

[23] Martínez Mora A, Subramanian V, Miljković F. Multi-task convolutional neural networks for predicting in vitro clearance endpoints from molecular images. J. Comput.-Aided Mol. Des. 2022;36:443–57.

[24] Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: a review of interpretability methods. Comput. Biol. Med. 2022;140:e105111.

[25] Zhong S, Hu J, Yu X, Zhang H. Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: transfer learning, data augmentation and model interpretation. Chem. Eng. J. 2021;408:e127998.

[26] Iqbal J, Vogt M, Bajorath J. Learning functional group chemistry from molecular images leads to accurate prediction of activity cliffs. Artif. Intell. Life. Sci. 2021;1:e100004.

[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, arXiv (2015). https://doi.org/10.48550/arXiv.1512.03385.

[28] Huber PJ. Robust estimation of a location parameter. Ann. Math. Statist. 1964;35:73–101.

[29] Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M, Palmer A, Settels V, Jaakkola T, Jensen K, Barzilay R. Analyzing learned molecular representations for property prediction. J. Chem. Inf. Model. 2019;59:3370–88.

[30] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proc. IEEE conference on computer vision and pattern recognition; 2016. p. 2921–9. doi:10.1109/CVPR.2016.319.

[31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A.K. Xamla, E. Yang, Z. Devito, M. Raison Nabla, A. Tejani, S. Chilamkurthy, Q. Ai, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: an imperative style, high-performance deep learning library, arXiv (2019). https://doi.org/10.48550/arXiv.1912.01703.

[32] Shapley LS. A value for n-person games. In: Kuhn H, Tucker A, editors. Contributions to the theory of games ii. Princeton: Princeton University Press; 1953. p. 307–17.

[33] S. Lundberg, S.-.I. Lee, A unified approach to interpreting model predictions, arXiv (2017). https://doi.org/10.48550/arXiv.1705.07874.

[34] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, O. Reblitz-Richardson. Captum: a unified and generic model interpretability library for PyTorch, arXiv (2020). https://doi.org/10.48550/arXiv.2009.07896.

[35] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, arXiv (2017). https://doi.org/10.48550/arXiv.1703.01365.

[36] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, arXiv (2013). https://doi.org/10.48550/arXiv.1311.2901.

[37] H. Yuan, H. Yu, J. Wang, K. Li, S. Ji, On explainability of graph neural networks via subgraph explorations, arXiv (2021). https://doi.org/10.48550/arXiv.2102.05152.

[38] Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. J. Med. Chem. 1996;39:2887–93.

[39] McCoull W, Boyd S, Brown MR, Coen M, Collingwood O, Davies NL, Doherty A, Fairley G, Goldberg K, Hardaker E, He G, Hennessy EJ, Hopcroft P, Hodgson G, Jackson A, Jiang X, Karmokar A, Lainé A-L, Lindsay N, Mao Y, Markandu R, McMurray L, McLean N, Mooney L, Musgrove H, Nissink JMW, Pflug A, Pilla Reddy V, Rawlins PB, Rivers E, Schimpl M, Smith GF, Tentarelli S, Travers J, Troup RI, Walton J, Wang C, Wilkinson S, Williamson B, Winter-Holt J, Yang D, Zheng Y, Zhu Q, Smith PD. Optimization of an imidazo[1,2-*a*]pyridine series to afford highly selective type I1/2 dual Mer/Axl kinase inhibitors with in vivo efficacy. J. Med. Chem. 2021;64:13524–39.

[40] Dalvie DK, Kalgutkar AS, Khojasteh-Bakht C, Obach RS, O'Donnell JP. Biotransformation reactions of five-membered aromatic heterocyclic rings. Chem. Res. Toxicol. 2002;15:269–99.

[41] Kalgutkar AS, Gardner I, Obach RS, Shaffer CL, Callegari E, Henne KR, Mutlib AE, Dalvie DK, Lee JS, Nakai Y, O'Donnell JP, Boer J, Harriman SP. A comprehensive listing of bioactivation pathways of organic functional groups. Curr. Drug. Metab. 2005;6:161–225.