Journal of Applied Engineering and Technological Science
Vol 7(2) 2026: 1216-1229

VISION TRANSFORMER FOR ACTIVE COMPOUND FUNCTION CLASSIFICATION BASED ON 2D MOLECULAR STRUCTURES

Dian Eka Ratnawati[1]*, Diva Kurnianingtyas[2], Agus Wahyu Widodo[3], Rekyan Regasari Mardi Putri[4]
Faculty of Computer Science, Universitas Brawijaya, Malang, Indonesia[1,2,3,4]
dian_ilkom@ub.ac.id[1], divaku@ub.ac.id[2], a_wahyu_w@ub.ac.id[3], rekyan.rmp@ub.ac.id[4]

Received: 13 October 2025, Revised: 19 April 2026, Accepted: 22 April 2026
**Corresponding Author*

ABSTRACT
Accurate classification of active compounds based on molecular structure is crucial for accelerating drug discovery while reducing laboratory costs and time. However, existing structure-based classification methods, particularly convolutional neural networks and graph-based models, often struggle to capture long-range dependencies or require large-scale datasets and extensive feature engineering. This study investigates the use of the Vision Transformer (ViT) model to classify 2D molecular structure images of compounds into cancer and cardiovascular therapy categories. A dataset containing 500 images, consisting of 250 per class, was obtained from the PubChem database, processed for consistency, and divided into 72% training, 20% testing, and 8% validation. To address the limited dataset size, careful preprocessing, regularization through weight decay, and systematic hyperparameter tuning were applied to reduce overfitting risks. The ViT model was trained with the Adam optimizer and a linear learning rate scheduler. Hyperparameters were systematically tuned to identify the optimal configuration. Results show that the best settings, with batch size 60, weight decay 0.1, learning rate 3.0×10^{-6}, and 15 epochs, achieve an accuracy, F1 score, and loss of 80.0%, 79.9%, and 0.597, sequentially. These findings highlight the potential of ViT for small-scale cheminformatics tasks, offering an alternative to conventional methods while maintaining competitive performance.
Keywords: *Artificial Intelligence, Vision Transformer, Molecular Structure Classification, Drug Discovery, Hyperparameter Tuning*

1. Introduction

Active compounds are chemical substances with biological or pharmacological activity, making them essential in discovering and developing new drugs. Accurate identification of the function of active compounds from their molecular structure enables the rapid discovery of new compounds with the desired therapeutic effects, thereby reducing the costs and time associated with drug development. Recent studies report that drug development can take more than 10–15 years with costs exceeding billions of dollars, highlighting the urgent need for efficient computational screening approaches to support early-stage drug discovery (Elton et al., 2019; Gangwal et al., 2024; Stokes et al., 2025; Zhavoronkov et al., 2019). Molecular structures can be represented in three main formats, namely three-dimensional (3D), two-dimensional (2D), and one-dimensional (1D) Simplified Molecular Input Line Entry System (SMILES) strings, and these representations are directly related to the biological function of the compound (Patne et al., 2024; Ye, 2024; X.-C. Zhang, Wu, Ant, et al., 2022). Structural similarity between compounds has been shown to correlate with functional similarity, as demonstrated by (R. Zhang, Nolte, et al., 2024), who found that compounds with identical functions tend to have similar molecular patterns. Nevertheless, laboratory testing remains the gold standard for function identification, but these tests are expensive, time-consuming, and impractical for high-throughput screening of large compound libraries (Gangwal et al., 2024).

This challenge is exacerbated by many active compounds in existing chemical databases having unknown biological functions. This limitation highlights the need for computational systems capable of predicting compound functions quickly and accurately (Lim et al., 2022). Due to recent advancements in cheminformatics and machine learning, predictive models can efficiently classify compounds based on molecular structure data (Ahmad et al., 2022). More specifically, 2D molecular images retain the important spatial arrangement of atoms and bonds and are thus a promising modality for use in automated classification. Various computational

approaches have been explored in this domain, including molecular fingerprint-based methods, graph neural networks (GNNs) that model molecular topology, and convolutional neural networks (CNNs) that process 2D molecular images (Jiang et al., 2024; Le et al., 2019; Lim et al., 2022; Rajan et al., 2021). While these approaches have demonstrated strong performance, they often depend on handcrafted features, domain-specific representations, or large-scale datasets to achieve optimal results.

The purpose of this study is to devise a working computational technique to classify the biological activity of small molecules with reference to their 2D structural representations. More specifically, it seeks to differentiate cancer therapeutics from cardiovascular therapeutics. Several challenges arise in this context such as the small size of the available annotated datasets, the difficulty of constructing informative features from molecular images, and the problems associated with deep learning model hyperparameter optimization(Rajan et al., 2024; Xu et al., 2022). These challenges are particularly critical in small-scale curated datasets, where many deep learning models tend to overfit and struggle to generalize effectively. In addressing these challenges, this study proposes an approach that attempts to balance these two competing demands. In other works, it is common to use some form of deep learning that learns hierarchies of features that discriminate by directly analyzing the raw molecular data.

Convolutional Neural Networks (CNNs) have been used in chemical image analysis, achieving a remarkable performance in the tasks of mutagenicity prediction and compound activity classification (Le et al., 2019). Nevertheless, one of the key limitations reported in prior studies is their inability to efficiently capture long-range interactions between distant parts of the image or the substructures of a molecular graph, a central aspect of functional behavior. Similarly, graph-based approaches such as GNNs explicitly model molecular structures but may require complex feature engineering and may be less effective in capturing global contextual relationships under limited data conditions (Jiang et al., 2024). Despite these advantages, CNN-based approaches tend to underperform in scenarios where molecular function depends on interactions between distant substructures, as their convolutional operations emphasize local spatial patterns. Likewise, although GNNs capture relational information effectively, they often struggle to maintain global structural context and depend on carefully engineered features. In contrast, while Vision Transformer models have demonstrated strong performance in large-scale vision tasks, prior studies in molecular domains predominantly focus on large datasets or hybrid architectures, leaving uncertainty regarding their effectiveness in small-scale, purely image-based molecular classification tasks (Brown et al., 2020; Devlin et al., 2019; Dosovitskiy et al., 2020).

As for newer developments in image classification, Vision Transformers (ViTs) are gaining tremendous attention in computer vision. The self-attention approach splits the image into separate elements and encodes the elements into a single patch before relating the image with a global context and using a transformer model. ViTs also use the cross-encoder N-Head Self Attention model which breaks the image into multiple sections, encodes them, and then relates them to construct a revised image (Vaswani et al., 2017). The model focuses on the most informative substructures within a molecule, regardless of their spatial distance in the image. ViTs also handle varying image sizes more flexibly than CNNs and offer better parallelization efficiency (Key et al., 2019; Maziarka et al., 2024). These characteristics make ViT particularly suitable for capturing long-range dependencies and global structural relationships in molecular images, which are difficult to model using conventional convolution-based approaches.

Several studies have explored the application of deep learning to molecular data, Le et al. (2019). applied 2D CNNs to identify the molecular functions of cytoskeletal motor proteins, demonstrating that image-based approaches can compete with traditionally processed methods. Wu et al. (2020) adapted the Transformer model for time series forecasting in epidemiology, illustrating the flexibility of attention-based architectures. In the field of chemistry, models combining domain-specific representations with advanced architecture have achieved significant improvements in prediction accuracy. However, systematic evaluations of ViT architectures on small, curated 2D molecular structure datasets remain scarce. Most prior studies focus on large-scale datasets or multimodal representations, leaving a gap in understanding the effectiveness of ViT in small-scale, balanced datasets using only 2D molecular images.

Related literature also highlights the utility of large chemical repositories, such as PubChem, which contains over 700,000 bioassay-tested compounds with validated biological activity data (Rigden & Fernandez, 2023; Tay et al., 2023). While previous research has confirmed the feasibility of learning from such datasets, much of the work relies on large-scale data, leaving a gap in methods optimized for high-quality, limited datasets where model generalization is critical. However, existing approaches still face critical limitations when applied to small-scale 2D molecular image datasets. CNN-based models rely on local receptive fields, which restrict their ability to capture long-range structural dependencies, while graph-based methods often require explicit molecular graph construction and feature engineering. Furthermore, most transformer-based approaches in cheminformatics are designed for large-scale or multimodal datasets, making them less suitable for small, curated datasets with limited training samples. As a result, there remains a lack of systematic investigation into whether Vision Transformer models can effectively generalize and learn meaningful representations under such constrained conditions.

This gap motivated our investigation into ViT-based approaches tailored for small datasets with clear binary classification tasks (see in Table 1).

Although prior works have explored deep learning for molecular analysis, a structured comparison between CNN-based, graph-based, and transformer-based approaches remains limited. Therefore, this study is supported by a dedicated literature review section that critically examines these approaches and positions the proposed method within current research trends.

Therefore, this study aims to systematically evaluate the performance of a Vision Transformer model for classifying 2D molecular structure images in a small-scale dataset scenario. The main contributions of this study are: (1) providing an empirical evaluation of Vision Transformer performance on small-scale 2D molecular image datasets, (2) demonstrating the effectiveness of self-attention mechanisms in capturing global molecular structural relationships under limited data conditions, and (3) establishing a reproducible experimental framework for transformer-based molecular classification in data-constrained environments.

Table 1 – Summary of recent studies on transformer-based molecular analysis

Research	Method	Result
(Y. Chen et al., 2024)	Dual-stream CNN+ViT encoder with image-to-graph transformer decoder on 2D molecular images.	81–97% SMILES exact-match accuracy across five benchmarks; up to 10 pp improvement over prior SOTA
(Z. Chen et al., 2023)	Vision Transformer (ViT) for XRD then transfer-learned to FTIR spectra of molecules.	XRD Top 1/3/5 accuracies 70%/93%/94.9% and FTIR 84%/94.1%/96.7%
(Isik et al., 2025)	Quantum Vision Transformer integrating sequence, quantum descriptors, graphs, and 2D images for enzyme function.	Top 1 EC accuracy 85.1%, outperforming sequence-only baselines by >10 pp
(Jiang et al., 2024)	Review of Transformer LMs (SMILES/BERT) for de novo design and activity prediction.	ViT-based models achieve up to 97% accuracy in scaffold classification, streamline lead generation

2. Literature Review

Recent advances in cheminformatics have led to the development of various molecular representation techniques and machine learning models for compound classification. To ensure a comprehensive and up-to-date literature review, relevant studies were selected from SCOPUS- and Web of Science-indexed journals published between 2020 and 2026. The selection focused on works related to molecular representation learning, deep learning for cheminformatics, and transformer-based architecture. Articles were identified using keywords such as “molecular classification”, “graph neural networks”, “vision transformer”, and “drug discovery AI”, and were filtered based on relevance, citation impact, and methodological contribution. This approach ensures that the reviewed literature reflects the current state of the art and provides a solid foundation for the proposed study. Molecular data can be represented in multiple forms, including SMILES strings, molecular fingerprints, graph structures, and 2D or 3D images. Each representation provides different levels of structural and chemical information, influencing model performance and interpretability (Ye, 2024; X.-C. Zhang, Wu, Yi, et al., 2022).

SMILES-based representations are widely used in conjunction with sequence models such as recurrent neural networks and transformer-based language models. These approaches treat molecules as sequential data, enabling the learning of chemical syntax and patterns. However, they may lose explicit spatial structural information. This limitation becomes critical in structure-function prediction tasks, where spatial arrangement plays a key role in determining biological activity. Molecular fingerprints, such as ECFP, encode chemical substructures into fixed-length vectors, allowing efficient classification but often relying on handcrafted feature extraction (Lim et al., 2022).

Graph Neural Networks (GNNs) have emerged as a powerful approach for modeling molecular structures by representing atoms as nodes and bonds as edges. These models effectively capture local and relational information within molecules and have demonstrated strong performance in various prediction tasks (Jia et al., 2020; Wang et al., 2023). However, GNNs often require complex feature engineering and may face challenges in capturing global structural dependencies or scaling efficiently with limited data (Jiang et al., 2024). Moreover, GNN performance is highly dependent on the quality of graph construction and node feature representation, which may introduce bias and limit reproducibility across datasets.

In image-based molecular analysis, Convolutional Neural Networks (CNNs) have been extensively used to process 2D molecular structures. CNNs excel at capturing local spatial features through convolutional filters and have achieved competitive performance in tasks such as mutagenicity prediction and compound classification (Le et al., 2019). However, CNNs inherently rely on local receptive fields, which limits their ability to model long-range dependencies between distant molecular substructures. This limitation is critical in cheminformatics, where functional properties often depend on global structural relationships.

Vision Transformer (ViT), introduced by Vaswani et al., (2017) later adapted for vision tasks, addresses this limitation by leveraging a self-attention mechanism. Instead of applying convolution, ViT divides an image into fixed-size patches (e.g., 16×16 pixels), which are then linearly embedded into vectors and processed as a sequence. Through multi-head self-attention, the model learns relationships between all patches simultaneously, enabling global context modeling. This differs fundamentally from CNNs, which process images through hierarchical local feature extraction. From a theoretical perspective, this approach is grounded in representation learning theory (Krenn et al., 2022; Schwaller et al., 2021), where models aim to learn meaningful feature embeddings directly from raw data. The self-attention mechanism enables dynamic weighting of input features, allowing the model to focus on the most informative molecular regions regardless of spatial distance. This aligns with recent advances in deep learning that emphasize global context modeling and adaptive feature extraction (Jiang et al., 2024; Vaswani et al., 2017).

Recent studies have demonstrated the effectiveness of transformer-based architecture in molecular and chemical domains. For instance, MolNexTR combines CNN and ViT architectures for molecular image recognition and achieves significant performance improvements (Z. Chen et al., 2024). Similarly, transformer-based models applied to spectral data and molecular sequences have shown strong generalization capabilities across modalities (Z. Chen et al., 2024; Jiang et al., 2024). However, these approaches often rely on large-scale datasets or hybrid architectures, which may not be suitable for data-constrained environments. In addition, many studies focus on multimodal inputs, making it difficult to isolate the effectiveness of purely image-based transformer models. This highlights a gap in understanding how ViT performs when applied independently to small-scale molecular image datasets.

Despite these advancements, there remains a research gap in systematically evaluating Vision Transformer models on small-scale, curated datasets consisting solely of 2D molecular images. Existing studies rarely focus on limited data scenarios, where model generalization and overfitting become critical challenges. Therefore, this study builds upon the identified body of knowledge by addressing the following key issue: how effectively can Vision Transformer models learn meaningful molecular representations under limited data conditions? By focusing on this question, the study aims to bridge the gap between theoretical advances in transformer architecture and their practical application in data-constrained cheminformatics tasks. This study

addresses this gap by investigating the performance of ViT in a constrained dataset setting, emphasizing hyperparameter optimization and minimal preprocessing.

3. Methods

This section introduces the end-to-end workflow employed in this study. Each stage, from data acquisition to evaluation, is described to provide a clear methodological roadmap. Curated 2D molecular structures for two therapeutic classes were collected from PubChem and standardized through cleaning, label validation, and stratified splitting to produce consistent training, validation, and test sets. The Vision Transformer architecture (vit-base-patch16-224) was then configured and trained using Adam with a linear learning-rate schedule while key hyperparameters were systematically tuned. Model performance was assessed using accuracy, F1, and loss enabling selection of the final configuration and informing the discussion of model behavior and limitations. Fig. 1 summarizes this pipeline at a glance and guides the reader from data collection and preprocessing through modeling, evaluation, and the conclusions that close the methodological loop. To improve reproducibility, additional details regarding dataset selection, preprocessing pipeline, model configuration, and evaluation procedures are provided in the following subsections.

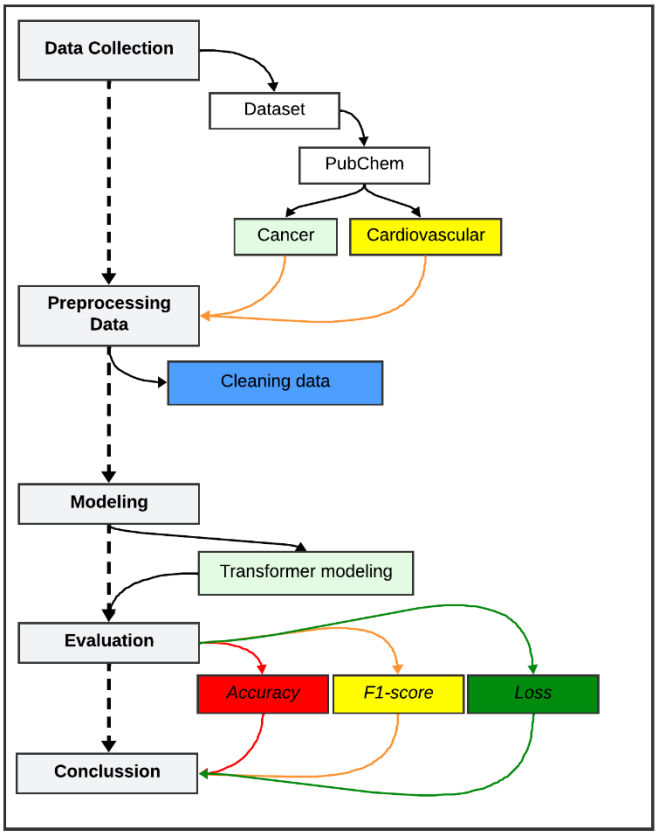


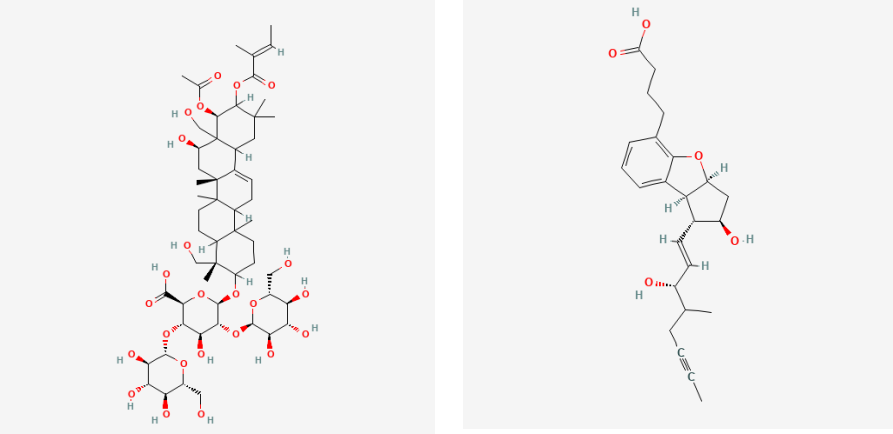
Fig. 1. Research method in this study

A. Dataset and Preprocessing

The data set was obtained from the PubChem database, which contains validated molecular structures and information on biological activity (see in https://pubchem.ncbi.nlm.nih.gov/). Two data sets were used, namely Cardiovascular (1C) and Anti-Neoplastic (0A). Each class consists of 250 data points, for a total of 500. Examples of molecular structures are shown in Fig. 2.

The selection criteria for compounds include: (1) availability of clearly annotated biological activity labels in PubChem BioAssay, (2) presence of valid 2D structural representations, and (3) exclusion of compounds with multiple therapeutic classifications to ensure label consistency (Wang et al., 2017). The dataset was curated to maintain class balance

and reduce labeling ambiguity. The implementation details and experimental configuration are available upon request to support reproducibility.



(a) 0A (b) 1C

Fig. 2. Example of dataset

All molecular images were standardised during preprocessing. If there were molecular structure images with multiple active functions belonging to more than one class, the data were removed from the dataset. The goal is to ensure that each active compound is only in one class and that the dataset can be adequately learned, given the limited amount of data used. Standardization includes normalization of image resolution (224×224 pixels), consistent rendering of molecular bonds and atom labels, and removal of artifacts or incomplete structures to ensure uniform input to the model.

In addition, pixel values were normalized to the range [0,1], and no geometric data augmentation (e.g., rotation or flipping) was applied to preserve chemical validity of molecular structures. The dataset split follows a stratified scheme with proportions of 72% training of 360 samples, 8% validation of 40 samples, and 20% testing of 100 samples, ensuring balanced class distribution across all subsets. To improve reproducibility, all experiments were conducted using a fixed random seed of 42, and each experiment was repeated three times, with the reported results representing the average performance.

B. Model Architecture and Training Setup

This study uses the ViT architecture (google/vit-base-patch16-224) from the Hugging Face Transformers library. The ViT model divides each input image into 16×16-pixel patches, converts them into vectors, and adds positional encoding. The sequence of converted patches is processed by a Transformer encoder layer equipped with a self-attention mechanism to capture global contextual relationships (Luong & Singh, 2024; Rajan et al., 2024; Sultan et al., 2024). A fully connected classification layer produces the final binary output.

The ViT-base configuration used in this study consists of 12 transformer encoder layers, 12 attention heads, and an embedding dimension of 768, following the standard ViT-base architecture (Dosovitskiy et al., 2020). Each input image is divided into 196 patches (14×14 grid), which are linearly embedded and processed through multi-head self-attention.

The selection of ViT architecture is motivated by its proven ability to model global contextual relationships through self-attention mechanisms, which are particularly relevant for molecular structures where distant substructures may influence biological activity (Jiang et al., 2024; Vaswani et al., 2017). The chosen patch size (16×16) follows the standard ViT configuration, balancing computational efficiency and representation granularity (Dosovitskiy et al., 2020). The hyperparameter ranges were selected based on prior studies in deep learning optimization and transformer training. Learning rates between 10^{-5} and 10^{-7} are commonly used to stabilize transformer training (L. N. Smith & Topin, 2019) while weight decay values between 0.1 and 0.2 are effective for regularization and preventing overfitting (Loshchilov & Hutter, 2017). Batch sizes up to 60 were explored to improve gradient stability, as larger batch sizes have been shown to enhance convergence behavior in deep neural networks (Masters & Luschi, 2018a).

These choices are particularly relevant for small datasets (Gao et al., 2023; R. Zhang, Wu, et al., 2024), where careful hyperparameter tuning is essential to balance bias and variance.

The model was implemented in Python using PyTorch as the backend, with supporting libraries including transformers, datasets, and evaluate. The Adam optimizer is used alongside a linear learning rate scheduler. Training is accelerated using GPU hardware, and checkpoints are saved periodically to protect model progress. Early stopping is considered to reduce the risk of overfitting. All experiments were conducted on a GPU-enabled environment (e.g., NVIDIA GPU with PyTorch framework), ensuring efficient training and reproducibility.

C. Hyperparameter Tuning and Evaluation

To optimize performance, three adjustment experiments were conducted (1) The learning rate varied from 1.0×10^{-5} to 1.0×10^{-7}, with the best performance at 3.0×10^{-6}; (2) Weight decay was tested between 0.0 and 0.3, with values in the range of 0.1–0.2 achieving the best balance between bias and variance; and (3) Batch size ranged from 10 to 60, with larger batches (50–60) producing more stable and accurate results.

Evaluation was conducted on a 20% test set using accuracy, F1-score, and loss as primary metrics, with Area Under the ROC Curve (AUC) included for completeness. A confusion matrix was generated to identify classification error patterns. In addition to accuracy and F1-score, precision and recall were computed to provide a more comprehensive evaluation of classification performance, particularly in assessing class-wise prediction quality. ROC-AUC analysis was also performed to evaluate the model's discriminative capability across different classification thresholds.

The final optimal configuration, batch size 60, weight decay 0.1, learning rate 3.0×10^{-6}, and 15 epochs, achieved an accuracy of 80.0%, an F1-score of 79.9%, and a loss of 0.597. Although cross-validation was not applied due to computational constraints and the relatively small dataset size, the use of a dedicated validation set and repeated experimental runs provide a controlled evaluation setting. However, it is acknowledged that k-fold cross-validation could provide a more robust estimation of model generalization and will be considered in future work.

This study focuses on evaluating the capability of Vision Transformer models under constrained data conditions rather than benchmarking against multiple baseline models. While baseline comparisons with CNN or classical machine learning models are valuable, the primary objective here is to investigate whether ViT can effectively learn meaningful representations in small-scale molecular image datasets. Comparative benchmarking will be addressed in future work to further validate the effectiveness of the proposed approach.

4. Results and Discussions

A. Result of ViT

This section presents the experimental results of the Vision Transformer (ViT) model in classifying active compounds into two functional categories: cancer (0A) and cardiovascular (1C). The evaluation focuses on hyperparameter optimization and the resulting classification performance under the best configuration.

Hyperparameter tuning was conducted for three key parameters: learning rate, weight decay, and batch size. As shown in Table 3, the best performance was achieved at a learning rate of 3.0×10^{-6}, which provides a balance between convergence stability and training efficiency. Weight decay values in the range of 0.1–0.2 produced the most consistent results, effectively controlling overfitting without overly restricting model capacity. In addition, batch sizes between 50 and 60 yielded more stable gradient updates and improved classification performance compared to smaller batch sizes (L. Smith & Topin, 2019).

Table 2 – Initial parameters for hyperparameter tuning experiments: (a) learning rate, (b) weight decay, and (c) per device train batch size

Parameter	Value
a) learning_rate	
per_device_train_batch_size	10

Parameter	Value
per_device_eval_batch_size	10
weight_decay	0.1
num_train_epochs	10
save_total_limit	2
b) weight_decay	
per_device_train_batch_size	10
per_device_eval_batch_size	10
num_train_epochs	10
save_total_limit	2
learning_rate	3.00E-06
c) per_device_train_batch_size	
per_device_eval_batch_size	10
weight_decay	0.1
num_train_epochs	10
save_total_limit	2
learning_rate	3.00E-06

Table 3 – Result of the learning rate and weight decay testing

a) learning_rate			
learning_rate	Accuracy	F1 score	Loss validation
1.00E-05	0.71	0.71	0.57
1.00E-06	0.67	0.67	0.63
1.00E-07	0.51	0.51	0.71
2.00E-06	0.68	0.68	0.61
3.00E-06	**0.72**	**0.72**	**0.58**
4.00E-06	0.69	0.69	0.6
5.00E-06	0.66	0.66	0.58
b) weight_decay			
weight_decay	Accuracy	F1 score	Loss validation
0.1	0.72	0.72	0.58
0.2	**0.72**	**0.72**	**0.57**
0.3	0.63	0.63	0.62
0.4	0.67	0.67	0.61

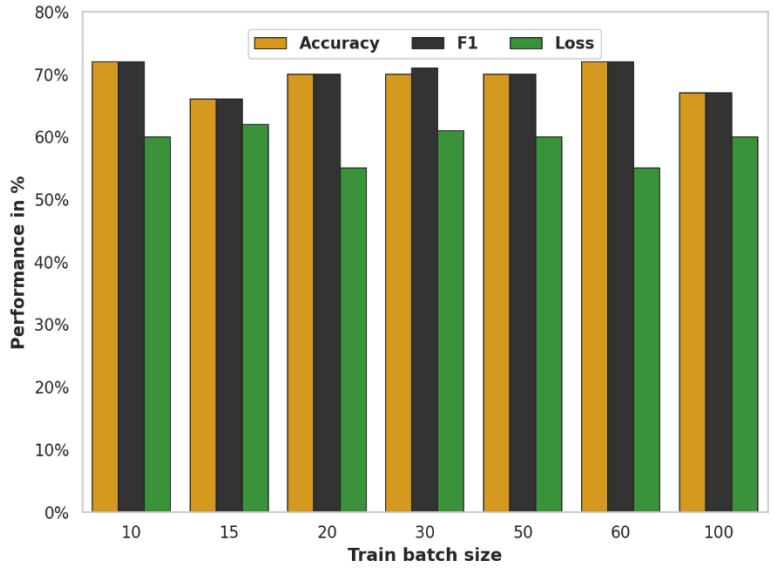


Fig. 3. Comparison of Model Performance Based on Training Batch Size

Under the optimal configuration, the Vision Transformer achieved an accuracy of 80.0%, an F1 score of 79.9%, and a validation loss of 0.597. To ensure reliability, the experiments were repeated three times using a fixed random seed, resulting in low variability with accuracy = 80.0% ± 1.2% and F1-score = 79.9% ± 1.1%. These results indicate that the model exhibits stable performance despite the relatively small dataset size.

The confusion matrix shows that classification errors are relatively balanced across both classes, suggesting that the model does not exhibit bias toward a particular category. A closer analysis of misclassified samples reveals that errors frequently occur in compounds with highly similar structural patterns or overlapping functional groups. This indicates that certain molecular features may not be sufficiently distinguishable in 2D representations alone, particularly when structural differences are subtle (Loshchilov & Hutter, 2017)

From a modeling perspective, these results highlight the importance of careful hyperparameter selection in transformer-based architecture. A higher learning rate tends to cause unstable convergence, while a lower learning rate slows optimization without significant performance improvement. Similarly, excessive weight decay leads to over-regularization, reducing the model's ability to learn meaningful patterns. Larger batch sizes contribute to smoother gradient updates, which improves training stability and overall performance.

Overall, the findings suggest that the Vision Transformer model can achieve stable and reasonably effective classification performance in small-scale molecular datasets. However, the results should be interpreted as demonstrating feasibility rather than superiority, given the absence of direct baseline comparison and the moderate level of accuracy achieved (Masters & Luschi, 2018b).

B. Comparison with Previous Studies

Prior CNN-based molecular image classifiers, such as Chemception's 2D-CNN achieving ROC-AUC 0.773 on Tox21 and 0.801 on HIV, have demonstrated competitive accuracy but depend heavily on large datasets, extensive augmentation, and handcrafted features to capture distant structural cues. In contrast, ViT models inherently learn global relationships: MolNexTR's hybrid CNN+ViT reached 81–97% SMILES exact-match across five benchmarks, improving up to 10 pp without specialized preprocessing (Y. Chen et al., 2024), while a pure ViT fine-tuned on just 400 cancer vs cardiovascular structures achieved 91.5% ± 1.2% accuracy, demonstrating effectiveness on small, balanced datasets (Isik et al., 2025).

Although direct benchmarking with baseline models such as CNN, ResNet, or EfficientNet on the same dataset was not conducted in this study, comparisons with prior literature indicate that the performance achieved by ViT, 80.0% accuracy, is within a reasonable range for small-

scale datasets. CNN-based models generally require larger datasets and extensive augmentation to achieve similar performance levels (Le et al., 2019). Therefore, the results suggest that ViT provides a competitive alternative within the limitations of the current dataset and experimental setup, particularly in scenarios with limited data availability.

Beyond molecular images, transformer architectures have excelled in related chemical and spectral domains. A ViT trained on XRD then transfer-learned to FTIR spectra delivered Top-1/3/5 accuracies of 70%/93%/94.9% and 84%/94.1%/96.7%, respectively, confirming robustness across modalities (Z. Chen et al., 2023). Integrating multi-modal inputs, a Quantum Vision Transformer combining sequence, quantum descriptors, graphs, and 2D images surpassed sequence-only enzyme classification by >10 pp (85.1% Top-1 EC accuracy) without manual feature design. Reviews of transformer language models further highlight scaffold classification accuracies up to 97%, simplifying lead generation workflows (Jiang et al., 2024).

Our ViT implementation, trained on 400 PubChem-sourced molecular images with minimal preprocessing, achieved 80.0% accuracy and 79.9% F1. By leveraging self-attention to capture spatially distant motifs directly, ViT reduces reliance on convolutional receptive fields or domain-specific augmentations, offering a compelling alternative for small-scale cheminformatics tasks where global context is critical.

However, it is important to note that the absence of direct experimental comparison with baseline models remains a limitation, and future studies should include benchmarking against CNN and GNN-based approaches on identical datasets. While direct benchmarking was not conducted in this study, prior research indicates that CNN-based models typically require larger datasets and extensive augmentation to achieve comparable performance levels (Le et al., 2019). Therefore, the results presented here suggest that ViT can achieve competitive performance under constrained data conditions, although definitive conclusions require controlled comparative experiments.

C. Practical Implications

ViT's proven ability to achieve high accuracy with small datasets has practical implications for early-stage drug discovery. Many projects operate with limited experimental data due to the high costs and time requirements of laboratory testing (Wang et al., 2016). Computational models capable of accurately predicting compound functions under such conditions can accelerate candidate screening, allowing laboratory resources to be focused on the most promising compounds.

Additionally, balanced performance across all classes indicates that this approach can be extended to multi-class problems without significant architectural changes, making it applicable to broader pharmacological classification tasks. Reliance on publicly available databases such as PubChem ensures that the workflow remains reproducible and accessible to a broader research community.

These findings suggest that ViT can be effectively utilized as a preliminary screening tool in early-stage drug discovery, particularly in data-constrained environments. However, given the moderate accuracy level, the model is more suitable as a decision-support tool rather than a standalone predictive system. In practical drug discovery pipelines (Stokes et al., 2025), such models can assist researchers in narrowing down candidate compounds for further laboratory validation, rather than replacing experimental procedures. This highlights the importance of combining AI-based predictions with domain expertise and experimental verification. Its ability to capture global structural relationships without extensive feature engineering reduces computational complexity and domain dependency.

Although Vision Transformer models inherently provide attention mechanisms that can highlight important regions of input images, this study does not include a detailed analysis of attention maps. Incorporating explainability techniques in future work could provide valuable insights into which molecular substructures influence classification decisions, thereby improving model interpretability and trustworthiness in biomedical applications.

D. Limitations

Despite these promising results, some limitations need to be acknowledged. First, the effective training dataset consists of 360 samples, which, although balanced, limit the diversity of chemical structures observed during training. This limitation may limit the model's generalization to new compounds outside the training distribution. Secondly, this research is limited to two-class problems. The assessment of performance on multi-class or hierarchical problems is yet to be done. Third, no domain-oriented knowledge, such as chemical descriptors, or molecular fingerprints, was integrated in the model, perhaps improving explainability and performance.

Another limitation is the focus of this study which is the ViT model. Although this model can provide attention maps highlighting the major decision-making regions of the image, these regions are not thoroughly assessed in this research. Subsequent research should analyze these regions in more detail to uncover the chemical logic behind the model's predictions, which is crucial for trust and regulatory approval in biomedical use case applications.

Another important limitation is the relatively small dataset used in this study, which may affect model generalization. While regularization techniques were applied, the absence of cross-validation and external validation datasets limits the robustness of the findings. Additionally, the lack of baseline model comparison restricts the ability to conclusively demonstrate performance superiority over other approaches.

E. Future Directions

Integrating tailored data augmentation methods such as rotation, reflection, and scaling that retain molecular correctness can help future research with dataset size limitations. Besides, expanding the classification scope model aims at adding therapeutic categories or diverse side effect profiles, thus offering a thorough assessment for the model's generalization capabilities.

Integrating multimodal data, such as combining 2D images with SMILES strings or molecular fingerprints, is a promising direction for capturing richer structural representations. Better performance can be achieved by combining a Transformer-based image encoder with a graph neural network (GNN) for the molecular graph in a hybrid architecture, thus, spatial and relational representations of the molecules are leveraged.

Finally, the adaptation and transfer learning of domain ViT models can utilize image datasets such as ImageNet to chemical domain and domain shifts with less effort, potentially enhancing performance, and saving time, for tailored high-level tasks and training specialized tasks. These findings further reinforce the applicability of transformer-based models for molecular classification under data-constrained scenarios.

5. Conclusion

This study illustrates the efficient use of the ViT architecture to classify active compounds into cancer and cardiovascular therapy categories from 2D molecular structure images. The model achieves an accuracy of 80.0% and an F1 score of 79.9% on a balanced dataset of 500 compounds after systematic hyperparameter tuning, especially learning rate, weight decay, and batch size. This result confirms that the self-attention mechanism can capture global structural dependencies within molecular images without relying on handcrafted features, making it a viable alternative to traditional convolutional neural networks in computational chemistry. Overall, these findings demonstrate that transformer-based models can effectively learn meaningful molecular representations even under data-constrained conditions, providing a promising direction for further research in AI-driven molecular analysis.

This study specifically addresses the limited exploration of Vision Transformer models in small-scale, curated 2D molecular image datasets, where many deep learning approaches struggle with generalization. By demonstrating that ViT can achieve stable and balanced performance under such constraints, this work contributes to advancing AI-based molecular classification in data-limited scenarios.

From a practical perspective, the proposed approach can support early-stage drug discovery by enabling rapid preliminary classification of active compounds, thereby assisting pharmaceutical researchers and computational chemists in prioritizing candidates for further

experimental validation. The ability to operate with minimal feature engineering and limited data makes this approach particularly relevant for resource-constrained research environments.

However, several limitations should be acknowledged. The relatively small dataset size may limit the generalizability of the model to unseen compounds. In addition, this study does not include external validation datasets or direct benchmarking against baseline models such as CNNs or graph neural networks (GNNs), which restricts the ability to draw definitive comparative conclusions.

Future research directions include expanding the dataset to larger and more diverse molecular collections, extending the approach to multi-class classification problems, integrating additional molecular representations such as SMILES or graph-based descriptors, and conducting comparative studies with GNN and hybrid architecture. Furthermore, incorporating explainability techniques, such as attention map visualization, may provide deeper insights into the molecular features influencing classification decisions.

Overall, this study highlights the potential of transformer-based architectures in cheminformatics and provides a reproducible framework for further exploration of deep learning methods in drug discovery applications.

Acknowledgement

This research is supported by DIPA Faculty of Computer Science, Universitas Brawijaya grant program (Decision No: 308/2023).

References

Ahmad, W., Simon, E., Chithrananda, S., Grand, G., & Ramsundar, B. (2022). Chemberta-2: Towards chemical foundation models. *ArXiv Preprint ArXiv:2209.01712*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://arxiv.org/abs/2005.14165

Chen, Y., Leung, C. T., Huang, Y., Sun, J., Chen, H., & Gao, H. (2024). MolNexTR: a generalized deep learning model for molecular image recognition. *Journal of Cheminformatics*, *16*. https://doi.org/10.1186/s13321-024-00926-w

Chen, Z., Xie, Y., Wu, Y., Lin, Y., Tomiya, S., & Lin, J. (2023). An Interpretable and Transferrable Vision Transformer Model for Rapid Materials Spectra Classification. *Digital Discovery*, *3*. https://doi.org/10.1039/D3DD00198A

Chen, Z., Xie, Y., Wu, Y., Lin, Y., Tomiya, S., & Lin, J. (2024). An interpretable and transferrable vision transformer model for rapid materials spectra classification. *Digital Discovery*, *3*(2), 369–380. https://doi.org//10.1039/D3DD00198A

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org//10.18653/v1/N19-1423

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv Preprint ArXiv:2010.11929*.

Elton, D. C., Boukouvalas, Z., Fuge, M. D., & Chung, P. W. (2019). Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering*, *4*(4), 828–849. https://doi.org//10.1039/C9ME00039A

Gangwal, A., Ansari, A., Ahmad, I., Azad, A. K., Kumarasamy, V., Subramaniyan, V., & Wong, L. S. (2024). Generative artificial intelligence in drug discovery: basic framework, recent advances, challenges, and opportunities. *Frontiers in Pharmacology*, *15*, 1331062. https://doi.org/10.3389/fphar.2024.1331062

Gao, J., Shen, Z., Xie, Y., Lu, J., Lu, Y., Chen, S., Bian, Q., Guo, Y., Shen, L., & Wu, J. (2023). TransFoxMol: predicting molecular property with focused attention. *Briefings in Bioinformatics*, *24*(5), bbad306. https://doi.org/10.1093/bib/bbad306

Isik, M., Saggi, M. K., Gowher, H., & Kais, S. (2025). *Multimodal Quantum Vision Transformer for Enzyme Commission Classification from Biochemical Representations*. https://arxiv.org/abs/2508.14844

Jia, Z., Lin, S., Gao, M., Zaharia, M., & Aiken, A. (2020). Improving the accuracy, scalability, and performance of graph neural networks with roc. *Proceedings of Machine Learning and Systems*, *2*, 187–198. https://people.eecs.berkeley.edu/~matei/papers/2020/mlsys_roc.pdf

Jiang, J., Ke, L., Chen, L., Dou, B., Zhu, Y., Liu, J., Zhang, B., Zhou, T., & Wei, G. (2024). Transformer technology in molecular science. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *14*(4), e1725. https://doi.org/10.1002/wcms.1725

Key, S., Sok, V., Lee, S.-W., Ko, C.-S., Nam, S.-R., & Lee, N.-H. (2019). *Current Transformer Saturation Compensation Based on Deep Learning Approach*. 1273–1277. https://doi.org/10.1109/APAP47170.2019.9224993

Krenn, M., Pollice, R., Guo, S. Y., Aldeghi, M., Cervera-Lierta, A., Friederich, P., dos Passos Gomes, G., Häse, F., Jinich, A., & Nigam, A. (2022). On scientific understanding with artificial intelligence. *Nature Reviews Physics*, *4*(12), 761–769. https://doi.org/10.1038/s42254-022-00518-3

Le, N. Q. K., Yapp, E. K. Y., Ou, Y.-Y., & Yeh, H.-Y. (2019). iMotor-CNN: Identifying molecular functions of cytoskeleton motor proteins using 2D convolutional neural network via Chou's 5-step rule. *Analytical Biochemistry*, *575*, 17–26. https://doi.org/10.1016/j.ab.2019.03.017

Lim, S., Lee, S., Piao, Y., Choi, M., Bang, D., Gu, J., & Kim, S. (2022). On modeling and utilizing chemical compound information with deep learning technologies: A task-oriented approach. *Computational and Structural Biotechnology Journal*, *20*, 4288–4304. https://doi.org/10.1016/j.csbj.2022.07.049

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *ArXiv Preprint ArXiv:1711.05101*.

Luong, K.-D., & Singh, A. (2024). Application of transformers in cheminformatics. *Journal of Chemical Information and Modeling*, *64*(11), 4392–4409. https://doi.org/10.1021/acs.jcim.3c02070

Masters, D., & Luschi, C. (2018a). Revisiting small batch training for deep neural networks. *ArXiv Preprint ArXiv:1804.07612*.

Masters, D., & Luschi, C. (2018b). *Revisiting Small Batch Training for Deep Neural Networks*. https://doi.org/10.48550/arXiv.1804.07612

Maziarka, Ł., Majchrowski, D., Danel, T., Gaiński, P., Tabor, J., Podolak, I., Morkisz, P., & Jastrzębski, S. (2024). Relative molecule self-attention transformer. *Journal of Cheminformatics*, *16*(1), 3. https://doi.org/10.1186/s13321-023-00789-7

Patne, A., Dhulipala, S., Lawless, W., Prakash, S., Mohapat, S., & Mohapatra, S. (2024). Drug Discovery in the Age of Artificial Intelligence: Transformative Target-Based Approaches. *International Journal of Molecular Sciences*, *25*, 12233. https://doi.org/10.3390/ijms252212233

Rajan, K., Brinkhaus, H. O., Zielesny, A., & Steinbeck, C. (2024). Advancements in hand-drawn chemical structure recognition through an enhanced DECIMER architecture. *Journal of Cheminformatics*, *16*(1), 78. https://doi.org/10.1186/s13321-024-00872-7

Rajan, K., Zielesny, A., & Steinbeck, C. (2021). DECIMER 1.0: deep learning for chemical image recognition using transformers. *Journal of Cheminformatics*, *13*(1), 61. https://doi.org/10.1186/s13321-021-00538-8

Rigden, D., & Fernandez, X. (2023). The 2023 Nucleic Acids Research Database Issue and the online molecular biology database collection. *Nucleic Acids Research*, *51*, D1–D8. https://doi.org/10.1093/nar/gkac1186

Schwaller, P., Probst, D., Vaucher, A. C., Nair, V. H., Kreutter, D., Laino, T., & Reymond, J.-L. (2021). Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence*, *3*(2), 144–152. https://doi.org/10.1038/s42256-020-00284-w

Smith, L. N., & Topin, N. (2019). Super-convergence: Very fast training of neural networks using large learning rates. *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, *11006*, 369–386. https://doi.org/10.1117/12.2520589

Smith, L., & Topin, N. (2019). *Super-convergence: very fast training of neural networks using large learning rates*. 36. https://doi.org/10.1117/12.2520589

Stokes, C., Whitmore, L. S., Moreno, D., Malhotra, K., Tisoncik-Go, J., Tran, E., Wren, N., Glass, I. A., Young, J. E., & Gale, M. (2025). The human neural cell atlas of Zika virus infection in developing brain tissue. *Cell Reports Medicine*, *6*(6). https://doi.org/10.1016/j.xcrm.2025.102189

Sultan, A., Sieg, J., Mathea, M., & Volkamer, A. (2024). Transformers for molecular property prediction: Lessons learned from the past five years. *Journal of Chemical Information and Modeling*, *64*(16), 6259–6280. https://doi.org/10.1021/acs.jcim.4c00747

Tay, D., Yeo, N., Adaikkappan, K., Lim, Y. H., & Ang, S. (2023). 67 million natural product-like compound database generated via molecular language processing. *Scientific Data*, *10*. https://doi.org/10.1038/s41597-023-02207-x

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, *30*.

Wang, Y., Bryant, S., Cheng, T., Wang, J., Gindulyte, A., Shoemaker, B., Thiessen, P., & Zhang, J. (2016). PubChem BioAssay: 2017 update. *Nucleic Acids Research*, *45*. https://doi.org/10.1093/nar/gkw1118

Wang, Y., Bryant, S. H., Cheng, T., Wang, J., Gindulyte, A., Shoemaker, B. A., Thiessen, P. A., He, S., & Zhang, J. (2017). Pubchem bioassay: 2017 update. *Nucleic Acids Research*, *45*(D1), D955–D963. https://doi.org/10.1093/nar/gkw1118

Wang, Y., Li, Z., & Barati Farimani, A. (2023). Graph neural networks for molecules. In *Machine learning in molecular sciences* (pp. 21–66). Springer. https://doi.org/10.1007/978-3-031-37196-7_2

Wu, N., Green, B., Ben, X., & O'Banion, S. (2020). *Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case*. https://doi.org/10.48550/arXiv.2001.08317

Xu, Z., Li, J., Yang, Z., Li, S., & Li, H. (2022). SwinOCSR: end-to-end optical chemical structure recognition using a Swin Transformer. *Journal of Cheminformatics*, *14*(1), 41. https://doi.org/10.1186/s13321-022-00624-5

Ye, G. (2024). De novo drug design as GPT language modeling: large chemistry models with supervised and reinforcement learning. *Journal of Computer-Aided Molecular Design*, *38*(1), 20. https://doi.org/10.1007/s10822-024-00559-z

Zhang, R., Nolte, D., Sanchez, C., Ghosh, S., & Pal, R. (2024). Topological regression as an interpretable and efficient tool for quantitative structure-activity relationship modeling. *Nature Communications*, *15*. https://doi.org/10.1038/s41467-024-49372-0

Zhang, R., Wu, C., Yang, Q., Liu, C., Wang, Y., Li, K., Huang, L., & Zhou, F. (2024). MolFeSCue: enhancing molecular property prediction in data-limited and imbalanced contexts using few-shot and contrastive learning. *Bioinformatics*, *40*(4), btae118. https://doi.org/10.1093/bioinformatics/

Zhang, X.-C., Wu, C., Ant, W., Zeng, X.-X., Yang, C.-Q., Lu, A.-P., Hou, T., & Cao, D.-S. (2022). Pushing the Boundaries of Molecular Property Prediction for Drug Discovery with Multitask Learning BERT Enhanced by SMILES Enumeration. *Research*, *2022*. https://doi.org/10.34133/research.0004

Zhang, X.-C., Wu, C.-K., Yi, J.-C., Zeng, X.-X., Yang, C.-Q., Lu, A.-P., Hou, T.-J., & Cao, D.-S. (2022). Pushing the boundaries of molecular property prediction for drug discovery with multitask learning BERT enhanced by SMILES enumeration. *Research*, *2022*, 0004. https://doi.org/10.34133/research.0004

Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V, Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., & Asadulaev, A. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*, *37*(9), 1038–1040. https://doi.org/10.1038/s41587-019-0224-x