LAGOM: A transformer-based chemical language model for drug metabolite prediction

Artificial Intelligence in the Life Sciences 8 (2025) 100142

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

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

Research article

LAGOM: A transformer-based chemical language model for drug metabolite prediction

Sofia Larsson [a,b,1], Miranda Carlsson [a,b,1], Richard Beckmann [a], Filip Miljković [b,*], Rocío Mercado [a,*]

[a] Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, Chalmersplatsen 1, Gothenburg, 412 96, Sweden

[b] Medicinal Chemistry, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Pepparedsleden 1, Gothenburg, 431 83, Sweden

ARTICLE INFO

Dataset link: https://github.com/tsofiac/LAGOM

Keywords:
Drug metabolism
Artificial intelligence
Deep learning
Language models
Transformers
Drug discovery

ABSTRACT

Metabolite identification studies are an essential but costly and time-consuming component of drug development. Computational methods have the potential to accelerate early-stage drug discovery, particularly with recent advances in deep learning which offer new opportunities to accelerate the process of metabolite prediction. We present LAGOM (Language-model Assisted Generation Of Metabolites), a Transformer-based approach built upon the Chemformer architecture, designed to predict likely metabolic transformations of drug candidates. Our results show that LAGOM performs competitively with, and in some cases surpasses, existing state-of-the-art metabolite prediction tools, demonstrating the potential of language-model-based architectures in chemoinformatics. By integrating diverse data sources and employing data augmentation strategies, we further improve the model's generalisation and predictive accuracy. The implementation of LAGOM is publicly available at github.com/tsofiac/LAGOM.

1. Introduction

Prior to clinical development, a comprehensive characterisation of a novel drug candidate's pharmacokinetic profile is essential to ensure adequate exposure within relevant tissues and to establish robust safety margins. Early incorporation of biotransformation studies is particularly critical at this stage of drug development [1,2]. For example, identifying a compound's metabolic soft spots can guide medicinal chemists towards structurally related analogs with enhanced metabolic stability [3], while simultaneously reducing the risks of generating reactive, toxic, or drug-drug interaction-prone metabolites [4,5]. Although present experimental approaches for investigating biotransformations are capable of detecting drug-related metabolites at trace concentrations, the entire process remains both time- and resource-intensive. Thus, computational methods for predicting xenobiotic metabolism prior to compound synthesis have gathered considerable attention over the past two decades [3,6,7].

While many *in silico* approaches rely on 3D structure information, e.g., to predict how a molecule interacts with metabolic enzymes such as the cytochrome P450 family [8–10], others utilise machine learning (ML) to predict sites of metabolism (SoMs) and chemical structures of potential drug metabolites [11,12]. SoMs are the atom positions in molecules that undergo metabolic transformations [13]. Typically, ML models that propose drug metabolite structures utilise predicted SoM information to apply rule-based transformations and thus filter and rank potential metabolites [11,12,14–16]. Such a two-step approach — first predicting SoMs, then applying site-specific chemical modifications — relies on pattern-based rule matching that commonly covers general biotransformations of phase I and phase II metabolism [3]. In addition, knowledge-based systems such as Meteor Nexus [17] integrate biotransformation rules curated from the scientific literature and expert input to predict metabolites of chemical compounds that conform to the presence of such target fragments. Their primary advantage lies in offering a clear and rational foundation for each prediction, such as supporting literature references and brief descriptions outlining each biotransformation mechanism. This provides a clear benefit over the ML-based approaches that often lack an explainability component behind their predictions ("black box" character). Moreover, understanding of safety liabilities associated with certain metabolic transformations encoded by knowledge-based systems helps to avoid

* Corresponding authors.
E-mail addresses: filip.miljkovic@astrazeneca.com (F. Miljković), rocio.mercado@chalmers.se (R. Mercado).
[1] These authors contributed equally.

https://doi.org/10.1016/j.ailsci.2025.100142
Received 5 July 2025; Received in revised form 27 August 2025; Accepted 2 September 2025
Available online 17 September 2025

designing drugs carrying the potential for generating toxic metabolites [3]. Apart from their robustness and explainability, methods that depend on rule-based transformations possess several disadvantages. First, rule-based methods utilise empirically derived biotransformation rules as reported in the scientific literature, which are subsequently compiled into databases. While these databases demonstrate broad coverage of biotransformation rules — including both frequent and less common metabolic transformations — as seen in the case of SyGMa [18], they are by no means comprehensive and may lack certain reaction types, hence limiting their applicability domain. For example, GLORYx [11] extends SyGMa with additional biotransformation rules, such as glutathione conjugation reactions, which are not reported in the original SyGMa collection. Therefore, further manual curation by biotransformation experts, especially utilising proprietary sources, is required to incorporate less commonly observed metabolic processes that may have safety implications for molecular design. This may, in part, be due to a lack of suitable machine-readable formats of metabolite identification data that would enable easier extraction and curation of biotransformation rules for a more comprehensive pattern coverage [19]. Second, these models are highly dependent on SoM predictor accuracy which, if insufficient, may lead to errors being propagated to the coupled rule-based algorithms, giving rise to a high number of false positives (i.e., low model precision). Consequently, this poses a challenge in interpreting model results by domain experts, thus requiring an effective post-processing pipeline to filter and rank structures of potential drug metabolites.

General-purpose molecular language models such as Chemformer [20] and ChemBERTa [21] have demonstrated that Transformer architectures can effectively learn chemical syntax and reaction semantics directly from SMILES representations. Building on these advances, recent work has explored adapting sequence-to-sequence models for the task of metabolite prediction [22,23]. One of the principal advantages of neural machine translation methods like Transformers is their capacity for direct sequence-to-sequence prediction (e.g., translating one compound's SMILES representation to another), without the need for explicitly defining transformation templates. This enables a single-step, end-to-end approach for molecular translation, a method that has demonstrated success in areas such as *de novo* molecular generation and computer-aided synthesis planning [24]. Here, metabolite prediction tasks can be related to the latter, where a metabolite structure sequence (a product) is predicted from its original drug molecule structure sequence (a reactant). However, a key difference between the two lies in the number of experimentally conceivable products: while a chemical reaction is typically optimised to yield a single product (one-to-one), a drug molecule is metabolically transformed by a variety of enzymatic and non-enzymatic mechanisms, resulting in several distinct metabolites (one-to-many). Furthermore, for model development only several thousand drug-related metabolite transformations exist in the public domain, in contrast to the hundreds of thousands or even millions of reactions well-documented in chemical reaction databases. Collectively, these factors make the metabolite prediction task more challenging to tackle.

Working in the low data regimes for sequence-to-sequence prediction tasks can potentially be addressed using a Transformer architecture that, beyond relying on the attention layers to capture interdependencies present in the sequences, can effectively be optimised via a task-aligned pre-training strategy in combination with fine-tuning by a well-curated task-specialised dataset. While Transformer-based models for metabolite prediction have previously been published [22,23], a systematic evaluation of different pre-training approaches, task-specific data augmentation techniques, and use of ensemble models trained on differently specialised datasets originating from a rigorously curated collection of drug-focused metabolic reactions accessible in the public domain has not yet been reported.

Herein, we make the following three key contributions:

1. A rigorous data curation pipeline for publicly available datasets for metabolite prediction, including datasets for (i) general chemical pre-training (Virtual Analogs), (ii) metabolite-specific pre-training (MetaTrans), (iii) fine-tuning (MetXBioDB and DrugBank), and (iv) benchmarking (GLORYx).
2. A curriculum-style transfer learning pipeline leveraging the Transformer architecture. Our model, termed LAGOM (Language-model Assisted Generation Of Metabolites), achieves performance superior to traditional rule-based methods (GLORYx and SyGMa) and an earlier Transformer-based model for metabolite prediction (MetaTrans), while performing on par with the more recent MetaPredictor model, on the established GLORYx benchmark, using a single language model.
3. A comprehensive ablation-type study systematically evaluating the effectiveness of various modelling strategies for metabolite prediction. We identify beneficial practices, such as SMILES randomisation and metabolite-specific curriculum pre-training, while highlighting strategies that provided limited benefit.

2. Methods

2.1. Datasets

The datasets used herein can be categorised into three distinct types: (1) pre-training data, (2) fine-tuning data, and (3) external test data. These are summarised in Fig. 1. All metabolism/pre-training data discussed herein are structured as *parent–child pairs*, where the *parent* represents the parent compound and the *child* represents a product of a chemical transformation (e.g., a metabolite of the parent compound). These are all represented using reaction SMILES.

Pre-training data. Two major transformation datasets were used for pre-training models. First, we acquired nearly 11 million compound pairs [25] (version v1) composed of publicly disclosed bioactive molecules and their virtual matched molecular pair analogs generated through R-group decomposition and substitution using retrosynthetic fragmentation rules. The dataset was compiled along with the corresponding chemical structures, which were obtained from CHEMBL35 [26]. The compiled dataset is here-on after referred to as the Virtual Analogs (VA) dataset. Then, we acquired the MetaTrans dataset [22], which in turn is based on various databases reporting human-related metabolic transformation reactions (xenobiotic- and endogenous compound-related).

Fine-tuning. The dataset of drug-related metabolic reactions used for fine-tuning models was generated from the publicly available MetXBioDB [27] (version NORMAN-SLE-S73.0.1.7) and DrugBank [28] (version 5.1.13) databases. These contained 2130 and 3489 parent–child pairs, respectively, before pre-processing.

Hold-out test set. The well-established GLORYx dataset [11], containing a selection of parent molecules and associated metabolites from the top 100 best-selling drugs of 2018 list, was used as an external validation set. GLORYx contains 136 first-generation parent–child pairs.

2.2. Data curation

In the case of metabolic reactions occurring in multiple steps, intermediate compounds and their subsequent metabolites often have their own entries in DrugBank and MetXBioDB. This means that seemingly independent reaction pairs actually trace back to the same original drug. Therefore, to ensure appropriate stratification, each multi-step reaction in DrugBank and MetXBioDB was associated with its starting node (i.e., drug of origin), which was later used to define training/validation/test data splits (see *Data Splitting* Section 2.3). The parsed DrugBank dataset and MetXBioDB dataset were combined into one single dataset, referred to here-on after as the *LAGOM dataset*.

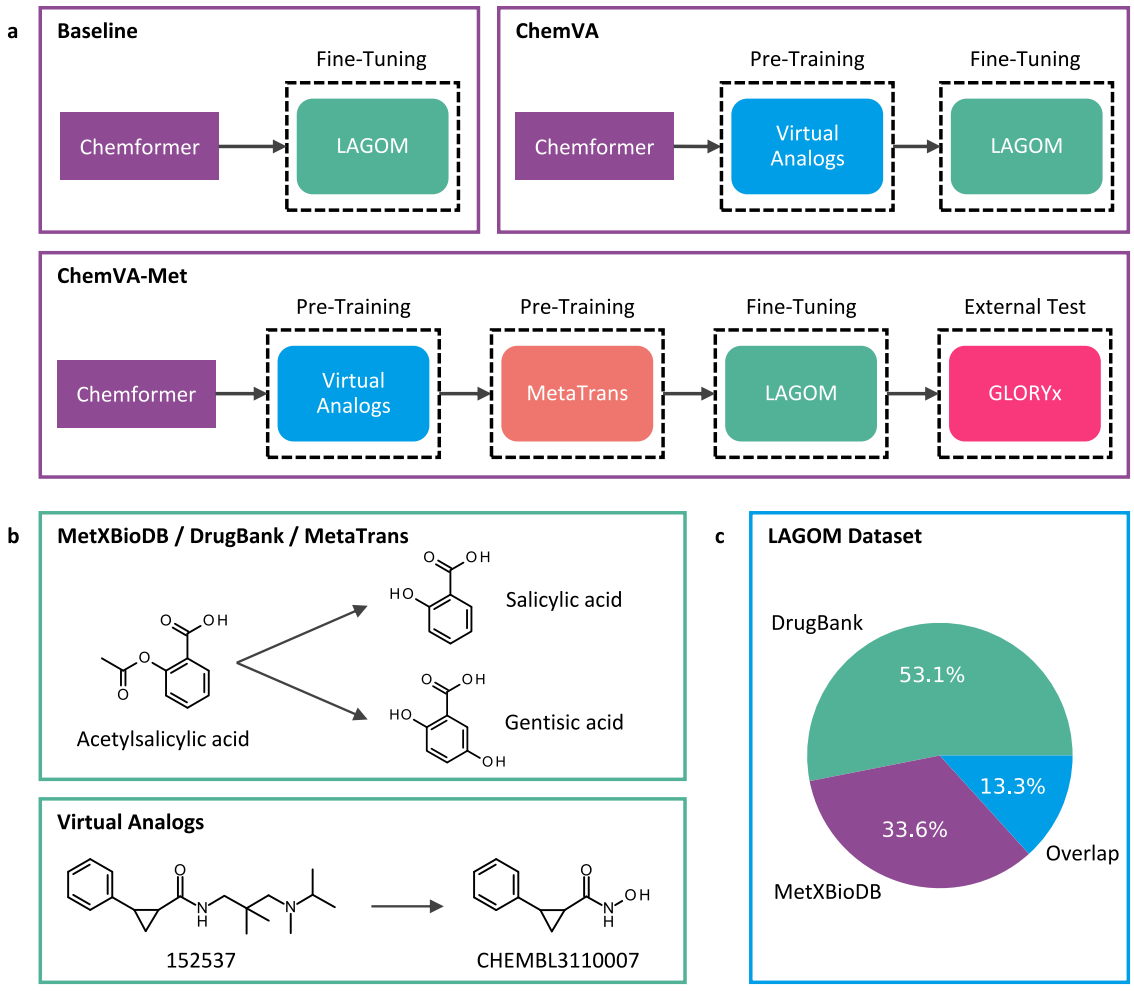

Fig. 1. Overview of the LAGOM pipeline. (a) Summary of the three main models developed in this work: the baseline Chemformer model, the ChemVA model, and the ChemVA-Met model. (b) Exemplary metabolic transformations present in the MetXBioDB/DrugBank/MetaTrans datasets and a matched molecular pair transformation available in the Virtual Analogs (VA) dataset. (c) The LAGOM dataset consists of the curated DrugBank and MetXBioDB transformations (13.3% overlap).

The data curation procedure was based on the properties of the LAGOM dataset, keeping the external test set intact. The same curation procedure was then applied to the pre-training data, with an additional step of removing any overlapping reactions with the LAGOM dataset.

First, each SMILES strings was standardised using the RDKit [29] and ChEMBL Structure Pipeline [30] packages in Python. This included removal of solvent molecules and salts, neutralisation of acids and bases, removal of stereochemical information, conversion of isotopes, and canonicalisation of compound structures. Duplicate parent–child pairs were removed, including instances where a parent structure was identical to its child following the SMILES standardisation procedure. Reactions with overlapping parent molecules with the GLORYx test dataset were also removed from the dataset. This was essential to ensure an unbiased benchmarking of the results against the external test dataset.

To remove any potential outliers in the dataset, we applied additional filtering steps to the data. First, we kept compound reaction pairs that only contained the following elements: C, O, N, Cl, F, S, P, Br, and I. Atom elements that were detected in drugs and their metabolites, including their frequencies prior to curation, are displayed in Fig. 2a. Thereafter, we applied a molecular weight cut-off to remove reactions with either too small or very large parent molecules from the dataset. We also filtered the dataset based on parent–child chemical similarity. Here, a cut-off based on Tanimoto similarity using Morgan fingerprints (radius=2, nBits=1024) was applied. This score ranges from 0 to 1, with 1 indicating perfect structural similarity. Based on the distribution, reactions with a similarity score < 0.20 were excluded.

This resulted in a well-curated drug metabolism reaction dataset for fine-tuning, which was dubbed the LAGOM dataset. It consists of 4055 parent–child pairs with 2248 unique drugs listed as reaction parents. In terms of the data source composition, the majority of the data originates only from DrugBank (more than half), whereas ~13% of parent–child pairs were reported in both DrugBank and MetXBioDB (Fig. 1c) following careful curation. The LAGOM dataset contains on average 1.8 metabolites per drug molecule, with a mean Tanimoto similarity between drugs and associated metabolites of 0.61. Only a small proportion of data was filtered out after applying molecular weight and similarity threshold requirements (~5%), resulting in a highly robust dataset well-suited for developing a chemical language model for drug metabolism prediction.

The number of parent–child pairs and unique parent compounds in each of the datasets following the data curation pipeline are summarised in Table 1.

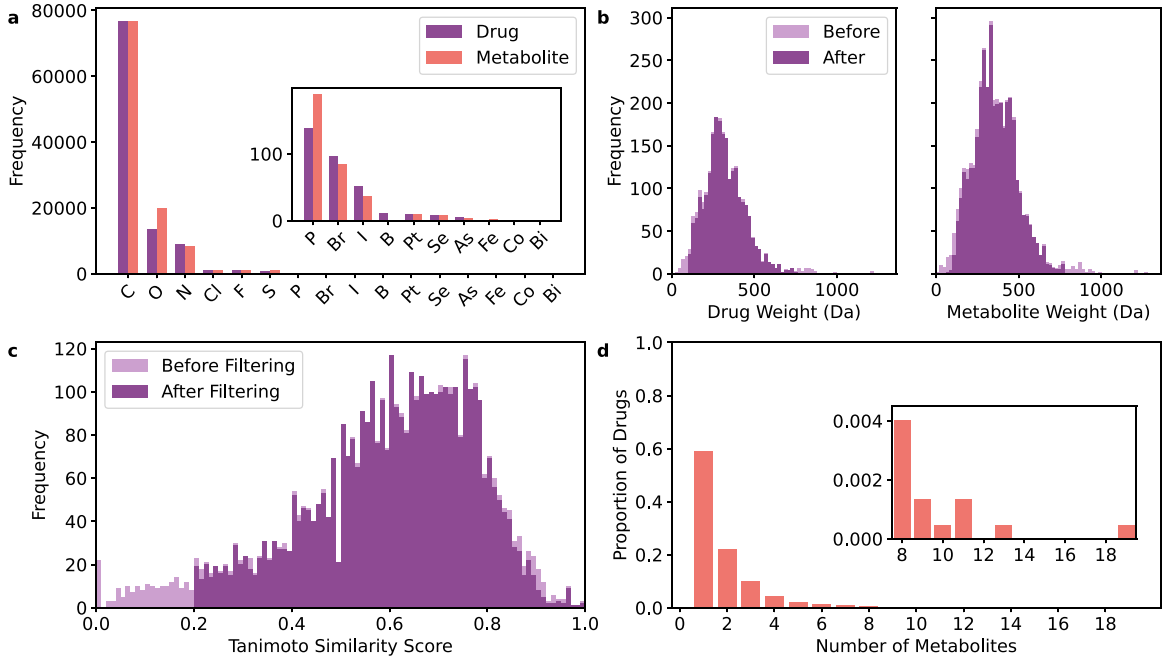

Fig. 2. LAGOM dataset summary. (a) Heavy atom distribution of the parent (drug) and child (metabolite) compounds before curation, with the inset highlighting the less common atom types. (b) Histogram of molecular weight (Dalton or Da) of drug and metabolite compounds before and after the filtering steps, highlighting how outliers with low and high molecular weights were removed. Note that there is a drug with a molecular weight > 4000 Da that is not included in the graph. (c) Histogram of Tanimoto similarities between corresponding parents and children (using Morgan fingerprints) before and after the filtering steps, displaying how metabolic transformation pairs with low structural similarity were eliminated. (d) Histogram illustrating the number of metabolites per drug following data curation, with the inset focusing on the tail end of the histogram corresponding to more metabolites.

Table 1
Summary of curated datasets used for model training and evaluation. The table reports the number of parent–child reaction pairs and the number of unique parent compounds remaining after data processing and filtering. All dataset sizes correspond to unique, non-overlapping pairs following cleaning, ensuring fair comparison across sources.

Dataset	# parent–child pairs	# unique parents
VA	10 762 115	1 251 518
MetaTrans	4243	2139
LAGOM	4055	2248
GLORYx	136	37

2.3. Data splitting

The LAGOM dataset was split into training, validation and test sets. For this, the *drug origin* for each metabolic reaction was used to split the data. The drug origin refers to the first parent compound in a multi-step reaction. In this way, we can guarantee that all reactions with the same drug origin are in the same set, thus minimising the potential of data leakage. Specifically, 85% of the data was allocated for training, 10% for validation, and 5% for testing. These sets were kept consistent for all different setups of the model in the project.

The VA dataset and MetaTrans dataset were split into a training set and a validation set at random. The ratio was 99.5% for training and 0.5% for validation, which is the same ratio as for the pre-trained Chemformer model [20]. The validation set was used to monitor the validation loss during training. No additional test set was created since the test set from the LAGOM dataset was used for testing the performance of the pre-trained model further fine-tuned by the training subset of the LAGOM dataset.

2.3.1. Data augmentation strategies

Given the relatively small fine-tuning training set of approximately 3400 reaction pairs, augmentation techniques were employed to expand it so as to enhance model performance. To gain a clear picture of each augmentation technique's impact, each of them was explored independently.

The first data augmentation technique involved extending the training set by generating new reactions from the existing ones. Since metabolic reactions typically occur in multiple steps, every reaction can be connected to its drug of origin. New reactions were thus generated by connecting the origin drug to all of its metabolites' metabolites as new parent–child pairs, dubbed "parent-grandchild". These new reactions were then curated in a manner consistent with the original dataset to maintain coherence.

The second augmentation technique involved adding new reactions to the training set by linking each parent to itself, which is representative of a drug that does not undergo or only partially undergoes metabolism (i.e., metabolically stable drugs). This approach was dubbed "parent-parent". Given that a metabolite typically resembles its parent, this approach can enhance a model's ability to capture these similarities more effectively.

The third augmentation technique involved extending the training set by SMILES randomisation, as a single molecule can have several non-canonical SMILES representations. To form a random non-canonical SMILES string, the ordering of the atoms can be randomised. The potential advantage of SMILES randomisation is that the model does not need to learn to produce canonical SMILES strings, but rather learn the inter-relationship of atom/bond characters present in SMILES strings. During fine-tuning, each SMILES string had a 50% chance of being randomised.

Finally, while not strictly a data augmentation technique, enhancing the input data of the LAGOM dataset with descriptive annotations

can provide a model with additional information to learn from. The additional properties explored here were LogP (lipophilicity) and carbon sp^3 fraction (Csp3), as these properties are known to affect drug metabolism [31]. This was done by calculating these properties for each drug molecule, codifying property ranges as tokens, and subsequently appending the property tokens to each molecular embedding in the data.

2.4. Models

We present our three pre-trained models as well as the ensemble model approach for enhancing model performance.

2.4.1. Chemformer

As our baseline we employed Chemformer [20], an encoder--decoder Transformer model pre-trained on roughly 10^8 SMILES from ZINC-15 [32] by reconstructing the input after randomisation (50% probability) and random token-span masking (10% probability, Poisson distributed). For metabolite prediction we fine-tuned the public Chemformer using the updated Chemformer codebase in aizynthmodels v1.0.0 on the curated LAGOM dataset using the forward translation task (parent → metabolite) with teacher forcing for up to 200 epochs; the checkpoint with the best score (see Section 2.5) was used in subsequent experiments. The scoring function, together with the validation loss, was used to monitor the progress during fine-tuning.

Unless otherwise noted, hyperparameters for fine-tuning matched the original implementation: Adam optimiser ($\beta 1 = 0.9$, $\beta 2 = 0.999$), no weight decay, an initial learning rate (LR) of $1{\times}10^{-3}$ with 8000 warm-up steps followed by cosine LR decay, dropout 0.1 on all Transformer layers, an effective batch size of 512 (8*64) reaction pairs, and a validation set evaluation every epoch.

All experiments were conducted using PyTorch v2.5.1 on either a single NVIDIA A100 or V100 (with at least 128 GB) GPU.

2.4.2. ChemVA

The ChemVA model extends the chemical knowledge learned by the baseline Chemformer via additional pre-training on the VA dataset. Similar to the Chemformer pre-training, this extended pre-training was conducted over two days and employed both randomisation and masking. The same hyperparameters as for the fine-tuning were applied, except a batch size of 128 was used and the validation set was evaluated every third epoch. The final epoch of the new pre-training was used for fine-tuning on the LAGOM dataset.

2.4.3. ChemVA-Met

The ChemVA-Met model extends the ChemVA model via additional pre-training on the MetaTrans dataset, using the same settings as those applied to the previous pre-training. However, the best epoch was chosen differently due to the considerably smaller size of the dataset. The model was trained for 100 epochs, and the epoch with the lowest validation loss was selected for subsequent fine-tuning on the LAGOM dataset.

2.4.4. Ensemble models

With the aim of correctly predicting a greater range of metabolites, the concept of ensemble models was explored. An ensemble model was produced by combining four models, fine-tuned on different splits of the dataset. Three different splitting approaches were explored.

As drug molecules can potentially metabolise into several metabolic products, this poses a challenge for chemical language models to confidently predict multiple correct metabolite structures originating from a single drug molecule (one-to-many problem). Thus, one approach to split the data was to decrease the number of metabolites per drug in each model. If a drug had more than one metabolite, these were put into different splits. If it had fewer metabolites than the number of splits, the drug and its metabolites were also put in the remaining split(s). For drugs with only one metabolite, the parent–child pair was added to all splits. Consequently, all splits had at least one occurrence of each drug. This approach was named "Stratified Split".

Another approach was to split the data based on similarities between either the drug molecules or the metabolites, with the intention of creating models with different expertise. For splitting based on similarity, Morgan fingerprints and Tanimoto similarity scores between the molecules were calculated with RDKit. RDKit's Butina clustering algorithm divided the molecules into different clusters with a set distance threshold of 0.8, implying that molecules with a similarity of 0.2 or higher were clustered together. When dividing the data into splits, the clusters with the largest number of pairs were assigned first into separate splits and the remaining clusters were then used to fill up the splits to balance out their sizes. These approaches, aimed at splitting the training data for ensemble models into parent- or child-based clusters, were named "Parent Split" and "Child Split", respectively.

2.5. Evaluation

Since drugs are often metabolised to multiple metabolites, a single predicted metabolite for a given reaction does not represent a unique correct solution. To account for this, we designed a scoring function for the fine-tuning process that considers all known metabolites during evaluation. Specifically, predictions were generated using beam search (n=5) and subsequently canonicalised to ensure consistency in comparisons. The evaluation score was calculated as the fraction of true metabolites identified within these top-5 predictions relative to the total known metabolites for each drug (i.e., metabolite coverage, see Appendix A). This scoring function, alongside validation loss, guided model selection during fine-tuning. Specifically, we saved the three model checkpoints with the highest scores and subsequently selected the earliest epoch checkpoint among these three for further evaluation, provided that the validation loss had reached its minimum. Choosing the earliest epoch ensured that the selected model was closest to the minimum validation loss point, thereby mitigating potential overfitting. Appendix D illustrates how the values of the metabolite recall and the validation loss change during fine-tuning, as well as the three selected epochs with the highest metabolite coverage proceeded for model consideration.

Each fine-tuned model was tasked with predicting up to 20 metabolites per drug molecule in the test dataset. To ensure a fair and consistent evaluation, post-processing steps were applied to the predicted SMILES strings. Initially, all predictions were canonicalised, standardising their representations to match the format of the reference metabolites. During this process, invalid SMILES strings were filtered out. Subsequently, duplicate predictions were identified and removed, along with any predicted structures identical to the original drug molecule. Thus, the final set of predictions consisted solely of valid, unique, and distinct metabolic candidates.

To enhance statistical reliability, the test set was partitioned into four batches of approximately equal size (around 38 compounds each), and evaluation scores were computed separately for each batch and subsequently averaged. In contrast, the GLORYx test set was sufficiently small and thus evaluated as a single batch.

2.5.1. Evaluation metrics

Several metrics were implemented to comprehensively evaluate and benchmark model performance:

- *Validity*: The proportion of generated SMILES strings that are chemically valid.
- *Accuracy@k*: The fraction of drugs for which at least one true metabolite is predicted within the top-k predictions. Note that this is distinct from typical definitions of accuracy.

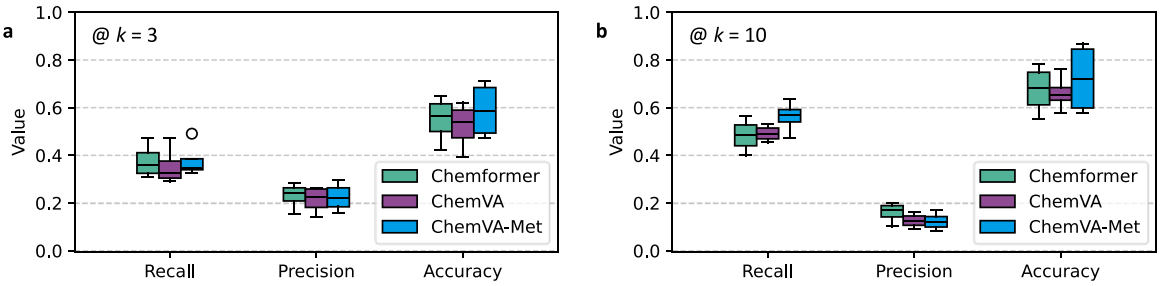

Fig. 3. Performance comparison of the different models presented in this work. (a) Metrics show recall, precision, and accuracy @ $k = 3$. (b) The same metrics @ $k = 10$. All results shown are on the held-out LAGOM test set. Note that accuracy refers to the fraction of drugs with at least one metabolite correctly predicted.

- *Precision@k*: The ratio of correctly predicted metabolites to the total number of predicted metabolites, considering the top-k predictions per drug. High precision indicates fewer incorrect predictions.
- *Recall@k*: The ratio of correctly predicted metabolites to the total number of known metabolites across all drugs, based on the top-k predictions per drug. High recall signifies effective coverage of possible metabolites.
- F_1 *score*: The harmonic mean of precision and recall. Values closer to one indicate superior performance.

The top-k predictions used in computing accuracy, precision, and recall were based only on valid and unique SMILES strings generated by the model. See Appendix A for details.

3. Results and discussion

We evaluated our models along several axes, including different pre-training strategies, ensemble approaches, and benchmark comparisons. Our findings are organised into three parts: the effect of pre-training, ensemble performance, and comparison with prior baselines.

3.1. Pre-training on metabolite-specific data improves model performance

As an initial baseline, we fine-tuned the publicly available Chemformer model on our curated LAGOM dataset. We first confirmed that SMILES randomisation during fine-tuning significantly improved model performance ($p < 0.05$; Appendix B). None of the other augmentation strategies significantly improved performance to warrant their further use in the experiments. Moreover, tokenisation of SMILES strings using Csp3 and LogP ranges, particularly the latter, worsened the performance compared to the baseline model. Therefore, only randomisation configuration was used for all subsequent experiments.

We then compared this to two additional models: ChemVA, which adds domain-relevant pre-training using the VA dataset, and ChemVA-Met, which further incorporates metabolic-specific pre-training using the MetaTrans dataset. Fig. 3 shows the comparative performance of the three pre-trained models across recall, precision, and accuracy at top-3 and top-10 prediction thresholds. We observe that pre-training with the VA dataset does not significantly improve precision and recall, whereas the addition of metabolic-specific pre-training (MetaTrans) leads to an increase in recall. For example, at $k = 10$, ChemVA-Met achieved significantly higher recall than the Chemformer baseline (with a p-value of 0.0162). On the other hand, the mean F_1 score, particularly at $k = 10$, was higher for the Chemformer model ($k = 3$: 0.28, $k = 10$: 0.24), compared to both ChemVA ($k = 3$: 0.26, $k = 10$: 0.20) and ChemVA-Met ($k = 3$: 0.28, $k = 10$: 0.20). However, at the point of model pre-training we decided to base our model selection criteria on recall, thus proceeding with the ChemVA-Met pre-training setup for further evaluation. Nevertheless, we also included the Chemformer baseline in our final model evaluation experiment using the independent GLORYx test set.

During pre-training experiments, we also evaluated the chemical validity of the SMILES predictions generated by each model, which is critical for their practical utility in drug discovery. All models consistently produced highly valid predictions, with mean validity scores of 96.6% for Chemformer, 95.4% for ChemVA, and 96.9% for ChemVA-Met. These consistently high validity rates (all above 95%) demonstrate that our fine-tuning and pre-training procedures effectively preserve the chemical correctness of the generated metabolites.

3.2. Ensemble strategies improve recall and accuracy

Next, we explored whether ensemble learning could improve metabolite prediction by combining multiple fine-tuned ChemVA-Met models trained on diverse data splits. Three splitting strategies were tested: stratified split, child-based clustering (grouping reactions by metabolite similarity), and parent-based clustering (grouping reactions by drug similarity) (Fig. 4a). The results are shown in Fig. 4b for the held-out LAGOM test set.

All ensemble models exhibited a modest increase in recall compared to the single-model baseline. However, this came at the cost of lower precision, consistent with the broader search space generated by combining predictions across models. Notably, the child-split ensemble achieved the best balance overall, offering an accuracy improvement while maintaining relatively stable precision on the LAGOM test set. As ensemble models can be considered methodologically distinct from the single model fine-tuned on all curated metabolite data, they were, irrespective of their performance on the held-out LAGOM test set, included in the final model evaluation using the external GLORYx dataset.

3.3. ChemVA-Met outperforms current benchmarks on GLORYx dataset

We benchmarked the best-performing models on the held-out GLORYx dataset and compared them to previous rule-based methods (SyGMa [18] and the original GLORYx model [11]). As this is not the first Transformer-based model to predict drug-related metabolic reactions, we also compared our results against the previously introduced MetaTrans Transformer model [22] and a more recent MetaPredictor model [33]. Results of our baseline model and best-performing model are summarised in Table 2. Additional results for the ensemble models are found in Appendix C.

Our best model ChemVA-Met, fine-tuned with SMILES randomisation, outperformed the Chemformer baseline, fine-tuned with no augmentation, in all metrics, increasing recall from 0.37 to 0.43 and precision from 0.14 to 0.18. Additionally, it excelled in both recall and precision compared to the ensemble models, obviating the need to weigh between these two scores. Importantly, this model also achieved

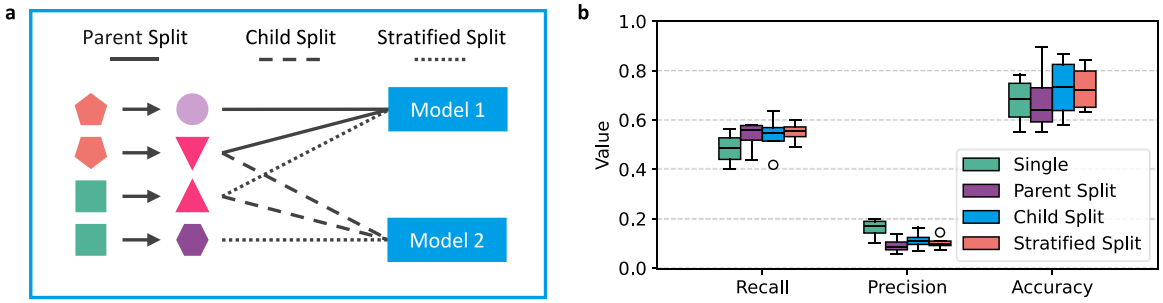

Fig. 4. (a) Overview of how the different splitting approaches for the ensemble models were performed. (b) Performance comparison of the ensemble model following different training strategies. The metrics on the single model are reported @ $k = 10$ on the held-out LAGOM test set, whereas the metrics on the ensemble models are reported so that the predictions per drug per split achieved a total value as close to 10 as possible. Note that accuracy refers to the fraction of drugs with at least one metabolite correctly predicted.

a substantially higher F_1 score (0.25) than SyGMa and GLORYx. These models achieved seemingly higher recall values, but it should be noted that the total number of predictions per molecule vastly exceeded the LAGOM number of predictions. This suggests that end-to-end, learned models can match or exceed the performance of curated rule-based tools while offering simpler and more scalable workflows.

Of particular importance was the comparison with MetaTrans, an existing Transformer model for predicting drug-related metabolic reactions. To enable direct comparison, we evaluated this model on the same, original GLORYx test dataset used to validate our models, even though some of the GLORYx parent–child pairs were included in the MetaTrans training set and could lead to an overestimation of its performance. We find, however, that our best-performing model has a comparable recall score (0.43 against 0.40) and a drastically better precision score (0.18 against 0.11), resulting in considerably better F_1 score (0.25 to 0.17). We name our model that displayed the best performance against the external hold-out set (i.e., ChemVA-Met model, fine-tuned with SMILES randomisation) the LAGOM model.

In addition, we evaluated a more recent, Transformer-based model for drug metabolite prediction, named MetaPredictor [33] using the same GLORYx hold-out set. MetaPredictor performance is comparable to the LAGOM model presented herein, with identical precision and marginally better recall, while being trained on a considerably larger number of drug-metabolite pairs (14 782 pairs) and using a SoM prompt as part of its predictive framework. Unfortunately, the exact dataset used to train MetaPredictor is not publicly available, and we cannot ascertain whether there is any overlap between the training set and the GLORYx hold-out set.

Although SyGMa and GLORYx achieve higher recall scores than LAGOM (Table 2), we emphasise that this higher recall comes at the cost of very low precision, driven by the very large number of predictions generated per molecule. This imbalance makes it difficult for end users to interpret and prioritise results as it introduces a substantial burden of false positives. Our focus here is therefore on methods that achieve a more appropriate balance between recall and precision, as reflected in the F_1 score. We believe this balance is critical for many drug discovery applications, where it is not only important to recover true metabolites but also to avoid overwhelming users with excessive, low-confidence predictions. By this standard, LAGOM substantially outperforms the rule-based baselines.

Finally, it should be stressed that the ensemble models that performed well on the LAGOM test set did not consistently outperform the single model on the GLORYx benchmark. For example, although the child-split ensemble showed promise in development, it had lower recall and F_1 score than the single ChemVA-Met model, as well as the Random Split ChemVA-Met model. On the other hand, the ensemble models were either better or equivalent to the other benchmark models when F_1 score was considered. These inconsistencies highlight the difficulty of selecting a single "best" model when held-out data and internal validation do not fully align, a common challenge in low-data domains like metabolite prediction.

Table 2
Performance comparison for the predictions of the best-performing model, *ChemVA-Met* @ $k = 10$, and the initial Chemfomer baseline model, on the GLORYx test set, against existing benchmarks of GLORYx [11], SyGMa [18], MetaTrans [22], and MetaPredictor [33]. The number of true metabolites is out of 136.

Model	Recall	Precision	F_1	True met.	Total pred.
Chemformer baseline	0.37	0.14	0.20	50	358
ChemVA-Met (LAGOM)	0.43	0.18	0.25	58	328
MetaPredictor	0.47	0.18	0.26	64	350
MetaTrans	0.35	0.15	0.21	48	316
SyGMa	0.68[a]	0.12[a]	0.20[b]	93[a]	800[a]
GLORYx	0.77[a]	0.061[a]	0.11[b]	105[a]	1724[a]

[a] Values obtained from de Bruyn Kops et al. [11].
[b] Scores calculated from values obtained according to Eqs. (1)–(3) in Appendix A.

4. Conclusions

In this work, we introduced a rigorously curated and standardised set of publicly available datasets tailored for metabolite prediction. By employing a curriculum-style transfer learning strategy with Transformer-based models, our ChemVA-Met model demonstrated superior performance compared to traditional rule-based benchmarks (SyGMa and GLORYx) and an existing Transformer-based model MetaTrans on the widely adopted GLORYx dataset. Additionally, we provide a robust and reproducible data processing pipeline suitable for future metabolite prediction tasks. Our thorough data curation ensured that these results were obtained without data leakage between pre-training, fine-tuning, and benchmark datasets.

Through a systematic ablation-type of study, we identified SMILES randomisation and metabolite-specific pre-training as particularly beneficial strategies for improving model performance. Conversely, we identified strategies that provided limited or inconsistent benefit, such as simple data augmentation methods and property annotations.

Nonetheless, the study has certain limitations. The relatively small size and chemical diversity of available metabolite datasets pose inherent constraints on model generalisability. Additionally, our results underline the difficulty in selecting optimal models based solely on internal validation, due to inconsistencies when generalising to external benchmarks. Metabolite prediction remains a fundamentally challenging task, primarily due to its one-to-many nature, data scarcity, and

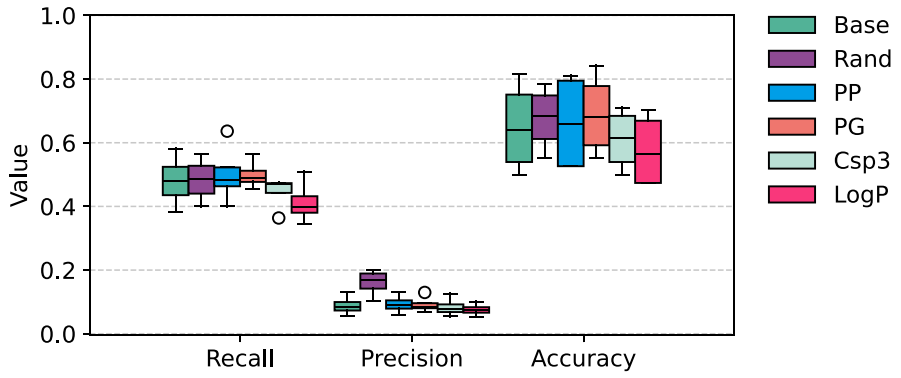

Fig. 5. Data augmentation results @ $k = 10$. Base: Baseline with no augmentation. Rand: Randomisation of SMILES strings added. PP: Parent-parent reactions added. PG: Parent-grandchild reactions added. Csp3: Annotations of Csp3 fraction added. LogP: Annotations of LogP value added. Note that accuracy refers to the fraction of drugs with at least one metabolite correctly predicted.

Table 3
Performance of the ensemble models on the GLORYx test set. The number of true metabolites is out of 136.

Model	Recall	Precision	F_1	True met.	Total pred.
ChemVA-Met Ensemble Stratified Split	0.43	0.15	0.22	58	380
ChemVA-Met Ensemble Child Split	0.35	0.15	0.21	48	330
ChemVA-Met Ensemble Parent Split	0.38	0.16	0.23	52	326

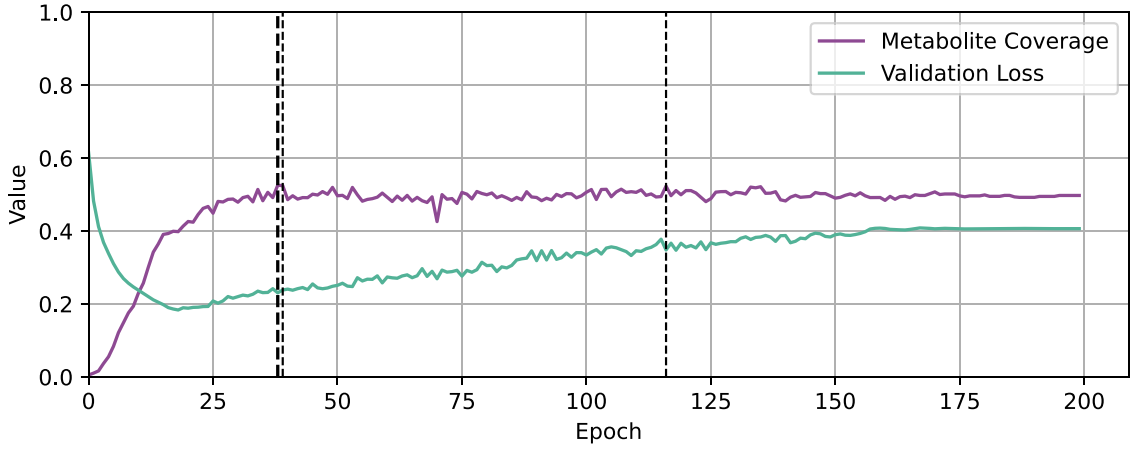

Fig. 6. Metabolite recall score and validation loss during fine-tuning on the LAGOM dataset with no augmentation on Chemformer. The dashed lines mark the three epochs (epochs 38, 39 and 116) with the highest metabolite coverage, of which epoch 38 (thicker dashed line) is closest to the validation loss minimum and was therefore proceeded with for model development.

high chemical diversity. This complexity was particularly evident in the limited generalisation performance of the child-split ensemble model, which, despite promising initial results, underperformed on external benchmarks. Future work may further explore evaluation schemes that explicitly account for user burden in metabolite prediction, as well as systematic evaluations of coverage across different biotransformation classes to better understand when our model succeeds and when it fails.

Future research directions include expanding curated datasets with richer, experimentally validated metabolic transformations collected under the same conditions, potentially in collaboration with industry partners, as well as the development of improved model selection and validation strategies. These steps could significantly enhance the robustness, accuracy, and practical applicability of Transformer-based models in metabolite prediction tasks.

CRediT authorship contribution statement

Sofia Larsson: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Miranda Carlsson:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Richard Beckmann:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Formal analysis. **Filip Miljković:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Rocío Mercado:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Filip Miljković is an employee of AstraZeneca and may own AstraZeneca shares. During the course of the project, Sofia Larsson and Miranda Carlsson were conducting a Master's thesis at Chalmers University of Technology, supported by AstraZeneca.

Acknowledgements

The authors thank Annie Westerlund for help and advice with modifying the Chemformer codebase, and Amanda Dehlén and Pär Aronsson for preliminary results and initial code upon which this project was based. RM acknowledges the funding provided by the Wallenberg AI, Autonomous Systems, and Software Program (WASP), supported by the Knut and Alice Wallenberg Foundation, Sweden. FM acknowledges funding support by AstraZeneca. RB acknowledges funding provided by the Intel and Merck AWASES program. The computations and data storage were partially enabled by resources provided by (1) Chalmers e-Commons, (2) AstraZeneca compute resources, and (3) the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

Appendix A. Metrics

For evaluating the model performance, precision and recall scores are used, defined as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{1}$$

$$Recall = \frac{TP}{TP + FN}, \tag{2}$$

where *TP* denotes *true positives*, i.e., the correctly predicted metabolites, *FP* denotes *false positives*, i.e., the valid but incorrect predictions, and *FN* denotes *false negatives*, i.e., the true metabolites not identified in the predictions.

To balance the metrics of precision and recall the F_1 score is used. It is defined as the harmonic mean between precision and recall according to the following equation:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{3}$$

Appendix B. Data augmentation

In Fig. 5 we illustrate the results of data augmentation in the Chemformer model. We observed that randomisation performed best, especially for precision, and proceeded to use those settings for the rest of the analyses in this work. To verify that randomisation was significantly better, a t-test was performed on the precision metric. The randomised model showed a *p*-value below 0.05 compared to all other models, which verified the decision of proceeding this these settings.

Appendix C. Ensemble models

Table 3 summarises the performance of the ensemble models on the GLORYx test set. We can observe that, although the child-split ensemble showed promise in development, it displays the lowest recall and F_1 score of all the utilised splitting approaches. Additionally, the stratified split approach shows the same recall score as our best-performing model.

Appendix D. Additional fine-tuning details

Fig. 6 illustrates scoring of the different epochs during the initial fine-tuning. The three epochs, illustrated with dashed lines, with the highest metabolite coverage, i.e., the value of the scoring function described in Section 2.5, were saved. Of these, epoch 38 was selected for further evaluation as it was the epoch closest to the validation loss minimum.

Data and code availability

Data and code is available on GitHub at https://github.com/tsofiac/LAGOM.

References

[1] Zhang Donglu, Luo Gang, Ding Xinxin, Lu Chuang. Preclinical experimental models of drug metabolism and disposition in drug discovery and development. Acta Pharm Sin B 2012;2(6):549–61.
[2] Shu Yue-Zhong, Johnson Benjamin M, Yang Tian J. Role of biotransformation studies in minimizing metabolism-related liabilities in drug discovery. AAPS J 2008;10:178–92.
[3] Kirchmair Johannes, Göller Andreas H, Lang Dieter, Kunze Jens, Testa Bernard, Wilson Ian D, Glen Robert C, Schneider Gisbert. Predicting drug metabolism: experiment and/or computation? Nat Rev Drug Discov 2015;14(6):387–404.
[4] Lin Jiunn H, Lu Anthony YH. Role of pharmacokinetics and metabolism in drug discovery and development. Pharmacol Rev 1997;49(4):403–49.
[5] Thompson Richard A, Isin Emre M, Ogese Monday O, Mettetal Jerome T, Williams Dominic P. Reactive metabolites: current and emerging risk and hazard assessments. Chem Res Toxicol 2016;29(4):505–33.
[6] Tran Thi Tuyet Van, Tayara Hilal, Chong Kil To. Artificial intelligence in drug metabolism and excretion prediction: recent advances, challenges, and future perspectives. Pharmaceutics 2023;15(4):1260.
[7] Zhai Jingchen, Man Viet Hoang, Ji Beihong, Cai Lianjin, Wang Junmei. Comparison and summary of in silico prediction tools for CYP450-mediated drug metabolism. Drug Discov Today 2023;28(10):103728.
[8] Afzelius Lovisa, Hasselgren Arnby Catrin, Broo Anders, Carlsson Lars, Isaksson Christine, Jurva Ulrik, Kjellander Britta, Kolmodin Karin, Nilsson Kristina, Raubacher Florian, et al. State-of-the-art tools for computational site of metabolism predictions: comparative analysis, mechanistical insights, and future applications. Drug Metab Rev 2007;39(1):61–86.
[9] Li Jianing, Schneebeli Severin T, Bylund Joseph, Farid Ramy, Friesner Richard A. IDSite: an accurate approach to predict P450-mediated drug metabolism. J Chem Theory Comput 2011;7(11):3829–45.
[10] Cruciani Gabriele, Carosati Emanuele, De Boeck Benoit, Ethirajulu Kantharaj, Mackie Claire, Howe Trevor, Vianello Riccardo. MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. J Med Chem 2005;48(22):6970–9.
[11] de Bruyn Kops Christina, Sícho Martin, Mazzolari Angelica, Kirchmair Johannes. GLORYx: prediction of the metabolites resulting from phase 1 and phase 2 biotransformations of xenobiotics. Chem Res Toxicol 2020;34(2):286–99.
[12] Flynn Noah R, Dang Na Le, Ward Michael D, Swamidass S Joshua. XenoNet: inference and likelihood of intermediate metabolite formation. J Chem Inf Model 2020;60(7):3431–49.
[13] Sícho Martin, Stork Conrad, Mazzolari Angelica, de Bruyn Kops Christina, Pedretti Alessandro, Testa Bernard, Vistoli Giulio, Svozil Daniel, Kirchmair Johannes. FAME 3: predicting the sites of metabolism in synthetic compounds and natural products for phase 1 and phase 2 metabolic enzymes. J Chem Inf Model 2019;59(8):3400–12.
[14] Flynn Noah R, Swamidass S Joshua. Message passing neural networks improve prediction of metabolite authenticity. J Chem Inf Model 2023;63(6):1675–94.
[15] Oeren Mario, Walton Peter J, Suri James, Ponting David J, Hunt Peter A, Segall Matthew D. Predicting regioselectivity of AO, CYP, FMO, and UGT metabolism using quantum mechanical simulations and machine learning. J Med Chem 2022;65(20):14066–81.
[16] Öeren Mario, Kaempf Sylvia C, Ponting David J, Hunt Peter A, Segall Matthew D. Predicting regioselectivity of cytosolic sulfotransferase metabolism for drugs. J Chem Inf Model 2023;63(11):3340–9.
[17] Marchant Carol A, Briggs Katharine A, Long Anthony. In silico tools for sharing data and knowledge on toxicity and metabolism: Derek for windows, meteor, and vitic. Toxicol Mech Methods 2008;18(2–3):177–87.
[18] Ridder Lars, Wagener Markus. SyGMa: combining expert knowledge and empirical scoring in the prediction of metabolites. ChemMedChem: Chem Enabling Drug Discov 2008;3(5):821–32.
[19] Ahlqvist Marie, Karlsson Isabella Bonner, Ekdahl Anja, Ericsson Cecilia, Jurva Ulrik, Miljković Filip, Chen Ya, Winiwarter Susanne. Metabolite identification data in drug discovery: Data generation and trend analysis. 2025, ChemRxiv.
[20] Irwin Ross, Dimitriadis Spyridon, He Jiazhen, Bjerrum Esben Jannik. Chemformer: a pre-trained transformer for computational chemistry. Mach Learn: Sci Technol 2022;3(1):015022.
[21] Ahmad Walid, Simon Elana, Chithrananda Seyone, Grand Gabriel, Ramsundar Bharath. ChemBERTa-2: Towards chemical foundation models. 2022, arXiv preprint arXiv:2209.01712.
[22] Litsa Eleni E, Das Payel, Kavraki Lydia E. Prediction of drug metabolites using neural machine translation. Chem Sci 2020;11(47):12777–88.

[23] Multari Silvia, Özçelik Rıza, Mazzolari Angelica, Nobile Marco Salvatore, Grisoni Francesca. Predicting metabolic reactions with a molecular transformer for drug design optimization. In: 2024 IEEE conference on computational intelligence in bioinformatics and computational biology. IEEE; 2024, p. 1–8.
[24] Miljković Filip, Rodríguez-Pérez Raquel, Bajorath Jürgen. Impact of artificial intelligence on compound discovery, design, and synthesis. ACS Omega 2021;6(49):33293–9.
[25] Dimova Dilyana, Bajorath Jürgen. Systematic design of analogs of active compounds covering more than 1000 targets. Zenodo; 2016.
[26] ChEMBL FTP directory. 2024.
[27] Djoumbou-Feunang Yannick, Fiamoncini Jarlei, Gil-de-la Fuente Alberto, Greiner Russell, Manach Claudine, Wishart David S. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. J Cheminformatics 2019;11:1–25.
[28] Wishart David S, Feunang Yannick D, Guo An C, Lo Elvis J, Marcu Ana, Grant Jason R, Sajed Tanvir, Johnson Daniel, Li Carin, Sayeeda Zinat, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46(D1):D1074–82.
[29] rdkitorg. Getting started with the rdkit in python. 2025, [Accessed 30 January 2025].
[30] Bento A Patrícia, Hersey Anne, Félix Eloy, Landrum Greg, Gaulton Anna, Atkinson Francis, Bellis Louisa J, De Veij Marleen, Leach Andrew R. An open source chemical structure curation pipeline using rdkit. J Cheminformatics 2020;12:1–16.
[31] Grogan Sean, Preuss Charles V. Pharmacokinetics. U.S. National Library of Medicine; 2023, [Accessed 03 November 2025].
[32] Sterling Teague, Irwin John J. ZINC 15–ligand discovery for everyone. J Chem Inf Model 2015;55(11):2324–37.
[33] Zhu Keyun, Huang Mengting, Wang Yimeng, Gu Yaxin, Li Weihua, Liu Guixia, Tang Yun. MetaPredictor: in silico prediction of drug metabolites based on deep language models with prompt engineering. Brief Bioinform 2024;25(5).