ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

XenoMet: A Corpus of Texts to Extract Data on Metabolites of Xenobiotics

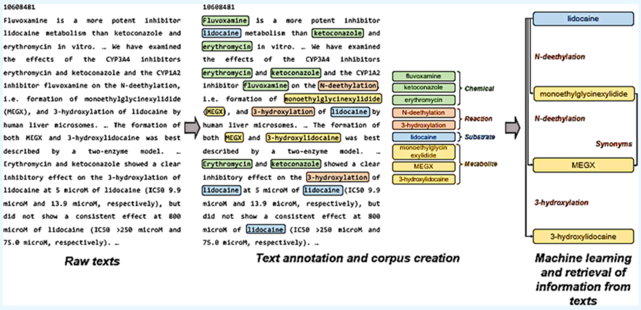
Nadezhda Yu. Biziukova,* Anastasia V. Rudik, Alexander V. Dmitriev, Olga A. Tarasova,* Dmitry A. Filimonov, and Vladimir V. Poroikov

Cite This: *ACS Omega* 2025, 10, 2459−2471

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Understanding the biotransformation of xenobiotics in the human body is critical for a comprehensive assessment of drug effects since pharmacologically active drug metabolites may exhibit a range of biological effects that often differ from those of the original pharmaceutical agent. Studies of the biotransformation mechanisms of xenobiotics have resulted in numerous publications. Extracting information about the parent compounds (substrates) and their metabolites from the texts allows retrieval of information on their biological activities, molecular mechanisms of action, and toxicity. Manual curation of the names of xenobiotics, their metabolites, and biotransformation reactions in the text is a challenging task due to the large number of publications related to studies of pharmaceutical agents metabolism. Our aim is to create an annotated corpus of texts that can be used for automated extraction of the names of xenobiotics, including pharmaceutical agents that undergo biotransformation and their metabolites. Prior to manual annotation of the corpus, semiautomatic annotation was carried out based on the earlier developed rule-based method for parent compounds and their metabolites extraction. To create XenoMet, we automatically extracted relevant texts from PubMed using a query based on MeSH terms. The names of biotransformation reactions were recognized by using an in-house-developed dictionary. Then, we manually verified the extracted data by correcting errors in the named entity annotation and identified the associations between substrates and metabolites. We tested the applicability of XenoMet for the reconstruction of a metabolic tree and for the automated extraction of the chemical names of substrates, metabolites, and reactions of biotransformation. Classification of the named entities of metabolites, substrates, and biotransformation reactions by a conditional random fields approach using XenoMet as the training set provides an F_1-score of 0.79.



10608481

Fluvoxamine is a more potent inhibitor lidocaine metabolism than ketoconazole and erythromycin in vitro. … We have examined the effects of the CYP3A4 inhibitors erythromycin and ketoconazole and the CYP1A2 inhibitor fluvoxamine on the N-deethylation, i.e. formation of monoethylglycinexylidide (MEGX), and 3-hydroxylation of lidocaine by human liver microsomes. … The formation of both MEGX and 3-hydroxylidocaine was best described by a two-enzyme model. … Erythromycin and ketoconazole showed a clear inhibitory effect on the 3-hydroxylation of lidocaine at 5 microM of lidocaine (IC50 9.9 microM and 13.9 microM, respectively), but did not show a consistent effect at 800 microM of lidocaine (IC50 >250 microM and 75.0 microM, respectively). …

10608481

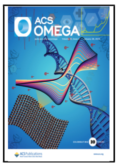
Fluvoxamine is a more potent inhibitor lidocaine metabolism than ketoconazole and erythromycin in vitro. … We have examined the effects of the CYP3A4 inhibitors erythromycin and ketoconazole and the CYP1A2 inhibitor fluvoxamine on the N-deethylation, i.e. formation of monoethylglycinexylidide (MEGX), and 3-hydroxylation of lidocaine by human liver microsomes. … The formation of both MEGX and 3-hydroxylidocaine was best described by a two-enzyme model. … Erythromycin and ketoconazole showed a clear inhibitory effect on the 3-hydroxylation of lidocaine at 5 microM of lidocaine (IC50 9.9 microM and 13.9 microM, respectively), but did not show a consistent effect at 800 microM of lidocaine (IC50 >250 microM and 75.0 microM, respectively). …

INTRODUCTION

Biotransformation affects the duration and intensity of the pharmacologic action of the drugs. Thus, the assessment of biotransformation pathways, the biological activities of xenobiotics, and their metabolites is an important step in the preclinical evaluation of drugs safety and efficacy.[1,2] Studies on the metabolites of pharmaceutical substances are of great importance due to their potential biological activity, including the pharmacological activity of prodrugs. Potential metabolites of xenobiotics can be identified in *in vitro* and *in vivo* experiments, for instance, using high-performance liquid chromatography and mass spectrometry (matrix-assisted laser desorption/ionization mass spectrometry).[3,4] Since the individual peculiarities of a particular chemical compound metabolites formation and excretion can be observed,[5] and taking into account variations in the experimental results, some differences in identifying metabolites and their quantity may be revealed. The accumulation of experimental data provides the opportunity to develop informational resources containing data on the chemical structures and biological activity spectra of xenobiotics and their potential metabolites. Scientific publications are the primary source of chemical information and data on biological activity; therefore, the development of methods for the automated extraction and analysis of literary data is of great importance.

There is a number of methods for text analysis and data mining that provide the opportunity for the automated, fast, and accurate named entity recognition.[6,7] Application of text and data mining to massive literature data containing information on both substrates and metabolites can be helpful for the automated recognition of the names of substrates, metabolites, and even reactions of biotransformation. Text-mining techniques can be used to obtain relevant and structured information about the

Received: June 19, 2024
Revised: October 25, 2024
Accepted: November 21, 2024
Published: January 12, 2025



ACS Publications

https://doi.org/10.1021/acsomega.4c05723

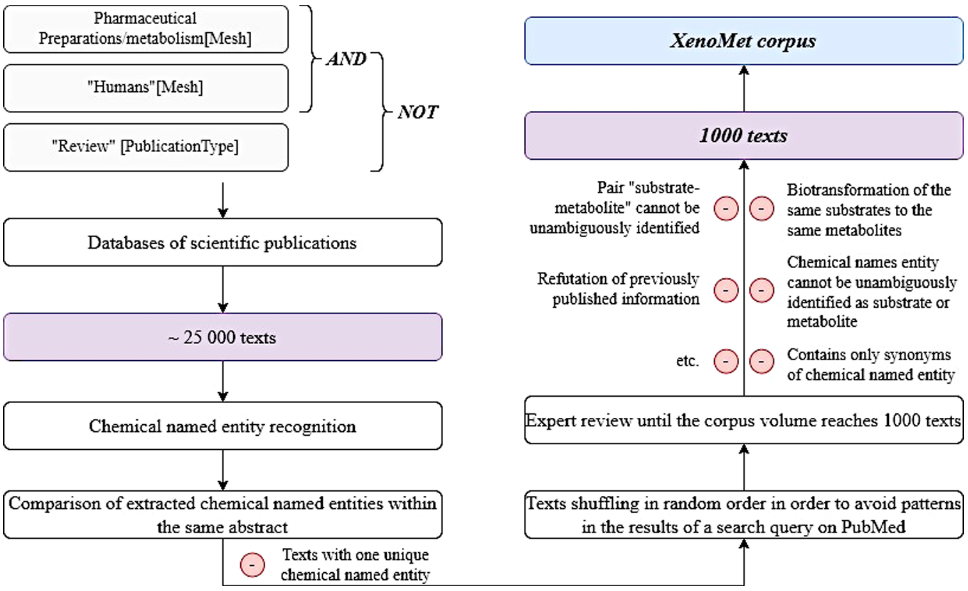

Pharmaceutical Preparations/metabolism[Mesh]
"Humans"[Mesh]
"Review" [PublicationType]
AND
NOT
Databases of scientific publications
~ 25 000 texts
Chemical named entity recognition
Comparison of extracted chemical named entities within the same abstract
-
Texts with one unique chemical named entity
Texts shuffling in random order in order to avoid patterns in the results of a search query on PubMed
Expert review until the corpus volume reaches 1000 texts
Pair "substrate-metabolite" cannot be unambiguously identified
Biotransformation of the same substrates to the same metabolites
Refutation of previously published information
Chemical names entity cannot be unambiguously identified as substrate or metabolite
etc.
Contains only synonyms of chemical named entity
1000 texts
XenoMet corpus

Figure 1. Scheme for selecting relevant texts for the XenoMet annotation.

biotransformation of chemical compounds together with the experimental conditions used in the experiment for metabolite identification. These data, in turn, are necessary for building computational models aimed at predicting sites of metabolisms (SOMs),[8−10] reactions of biotransformation,[11,12] generating and estimating the chemical structure of metabolites,[13] as well as at predicting the biological activity spectra for compounds,[14] taking into account their metabolism.[15]

Automated recognition of substrate and metabolite names could be performed using dictionary-based, rule-based, and machine-learning approaches.[16] Machine learning has several advantages, including the most rapid and accurate recognition of a wide range of chemical-named entities (CNEs), but it requires the application of an annotated corpus of texts containing information on both the names of xenobiotics (substrates) and their metabolites. The existing corpora were compiled to recognize the endogenous metabolites of specific organisms and evaluate their effects.[17] Therefore, the development of a corpus dedicated to the biotransformation of xenobiotics is beneficial for extracting knowledge about bioactive compounds including the effects of their metabolites. From this perspective, the developing corpus is unique and has no analogs.

The purpose of our study is to create the annotated corpus XenoMet for the automated recognition of names of (a) substrates, (b) metabolites, and (c) reactions of biotransformation and to test its applicability using the conditional random field (CRF) approach.

■ MATERIALS AND METHODS

Procedure for Collecting Texts for XenoMet. To retrieve texts that contain the names of substrates, metabolites, and biotransformation reactions, we automatically extracted them from the PubMed database using a Python script and a query based on MeSH terms (https://www.ncbi.nlm.nih.gov/mesh/), which are used by NCBI PubMed to classify publications according to specific categories. Since we are focused on the biotransformation of drugs, we selected the *"Pharmaceutical Preparations"* MeSH term as the most suitable one for the goal of collecting texts dedicated to the biotransformation of xenobiotics. The MeSH subheading *"metabolism"* was used to collect relevant texts on the metabolism of this chemical group. We included the MeSH term *"Humans"* to focus on metabolism in the human body. Using the query, we excluded reviews by adding *"Review[PublicationType]"* through the connector *"NOT"* (Figure 1). This resulted in more than 25 thousand results.

Further filtering included two stages: (1) exclusion of texts based on the number of names of chemical compounds and (2) expert assessment of the content (Figure 1).

To create a corpus, we used abstracts and titles (for further denotation, we use "text") of scientific publications since they are more frequently used in the tasks of information retrieval because of their availability.[7] Obviously, the texts that describe the biotransformation of a substrate into its product should include at least two chemical names: (1) the name of the substrate and (2) the name of the metabolite. For this reason, we performed the CNE recognition using the HunFlair library[18] that allows the CNE recognition with an F_1-score of 90%[18] in combination with a previously developed method based on CRF[13] and excluded from further consideration the texts that contained only one CNE.

During the manual annotation of texts, we developed some criteria for exclusion based on the suggestion that the corpus should be suitable for the automated recognition of a wide variety of substrate and metabolite named entities. The developed exclusion criteria are as follows:

- *A text contains the names of both substrate and metabolite, but their relationship cannot be unambiguously determined.* This criterion is applied in cases where the metabolites of certain xenobiotics are well-known and widely used, so the authors do not provide a particular semantic link between these two objects that can be recognized without specific knowledge.[19,20] An example is "Motor measurements such as tapping test, walking time, and tremor score, and blood samples for levodopa and 3-O-methyldopa (3OMD) plasma analysis were performed hourly".[19]
- *The text describes a refutation of previously published information about the biotransformation of a xenobiotic into a specific metabolite.* During the analysis, we noticed only

isolated cases of such negative relationships between substrates and metabolites (for example: "Our data indicate that JM6 is not a prodrug for Ro-61−8048 and is not a potent KMO inhibitor"[21]). Insufficient representation of negative examples in the annotated corpus will not improve the recognition performance but may negatively impair the extraction of positive relationships.

- *Texts that describe the biotransformation of the duplicated names of substrates and metabolites for collecting the most diverse corpus as possible.* This is true for the compounds that are actively and comprehensively studied, such as psychoactive substances and their metabolites in various biological fluids (amphetamines[22−24]), and for the determination of possible toxic products of antitumor drugs (irinotecan[25−27]). Following this criterion, the number of texts dedicated to the study of a particular substrate is limited to 2−5. The exception is when the texts mention different metabolites of the same substrate.
- *Texts describing metabolites of natural products.* This criterion made it possible to exclude mainly texts that describe the metabolism of drugs based on plant raw materials.
- *Texts that do not mention the final product of substrate biotransformation.* We aimed to provide a diversity of substrates and their metabolites presented in the XenoMet corpus, so we strived to include only those titles and abstracts that indicated explicit names of metabolites.
- *Metabolites of macromolecules (peptides, etc.)*

Texts that met the exclusion criteria were not considered for annotation. In total, 57% of the texts were discarded during the annotation process on the basis of the exclusion criteria described above. Texts that did not meet the exclusion criteria were shuffled in a random order. We believe that this step allows us to avoid bias in the selection of texts and to cover the research topics available in the literature as much as possible. Texts were sequentially annotated according to their random order until the corpus size reached 1000.

Automated Chemical-Named Entity Recognition and "Substrate−Metabolite" Relation Extraction. At the stage of selecting relevant texts, we automatically recognized all CNEs using the Python HunFlair library[18] in combination with a previously developed method based on CRF.[13] In this way, we were able to identify automatically the potential substrates and metabolites and filter out texts that contained only one CNE.

As the metabolism of xenobiotics is mediated by biotransformation reactions, we carried out the recognition of their names. To extract the data on reactions, we used a dictionary-based approach that was compiled through manual analysis of texts and molecular process ontology (MOP) as a basis.[28] Only the "Process" class of this ontology was used. We extracted from the MOP the "basic reactions". By "basic reaction", we considered such a term for the biotransformation that does not specify any position in relation to which transmormations in the molecule occur, such as "deacylation", "deamination", "esterification", and so on (for example, "hydroxylation" instead of "4-methylhydroxylation"). It will enable the extraction of the maximum number of reactions because some detailed reactions may be omitted from the analyzed texts. The total number of "basic reaction" terms included in the final dictionary is 168. The complete list of basic reactions is available in the Supporting Information.

Following the recognition of the named entities, we utilized an in-house approach previously described in our study[13] to perform the relation extraction between substrates and metabolites. This approach allows for the classification of chemical named entities as "Substrate", "Metabolite", or "Chemical" based on the use of semantic links like "is metabolized to", "are metabolites of", and "is formed". If two named entities of a substrate and a metabolite are linked semantically, they are considered a pertinent pair.

If neither a substrate nor a metabolite was found in the text using the identified semantic link, we classified such terms as a chemical entity ("Chemical") representing a separate class. A Python script that performs the extraction of associations between substrates and metabolites based on a rule-based approach is available in the Supporting Information.

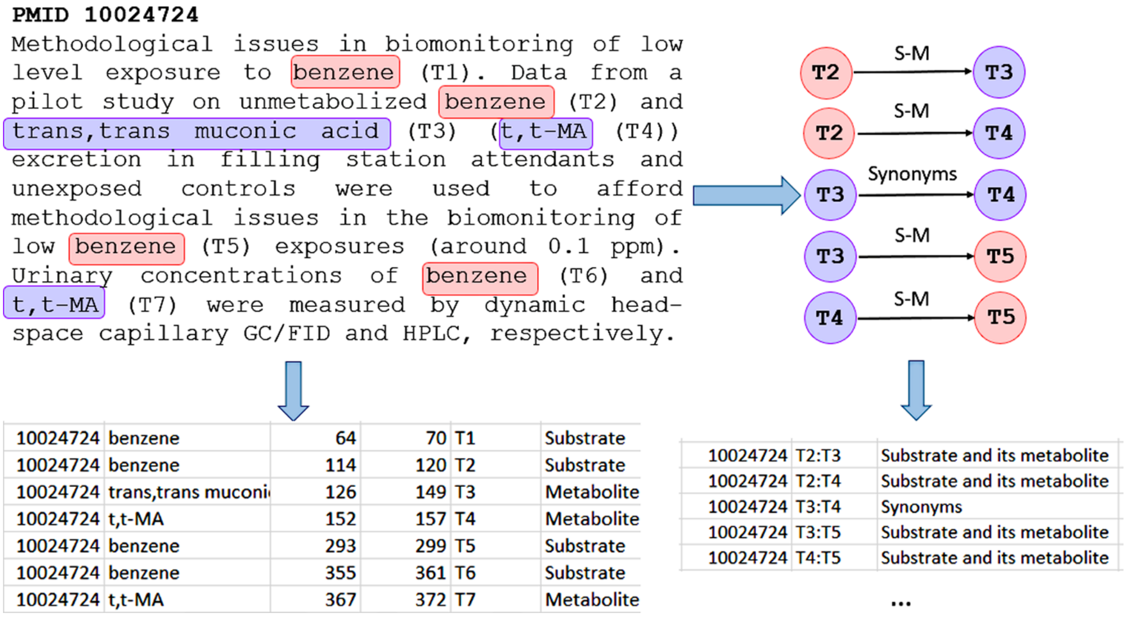
The corpus was represented in the text file using the CHEMDNER/DrugProtT format.[29,30] To represent the corpus in text format, we calculated the position of every named entity in the text. The counting of all text characters began from position 0, so that if the named entity is the first word in the text, its starting character will be zero. We numbered all of the extracted named entities according to their order of appearance in the texts using the following scheme: *T* (as for "term") followed by the position number of the named entity (for example, *T*3 for the third named entity in the text).

Manual Verification of Automatically Performed Text Annotation. After automatic annotation, we carried out a manual verification of the results. Any disagreement between experts was resolved through a consensual discussion, which resulted in some rules. In particular, optical isomers were taken into account during the manual verification of metabolites. We did not exclude those terms that indicated, for example, groups of chemical compounds since this information may be useful for understanding the biotransformation reactions of substrates (for example, "nitroxyl radicals").

Named entities of commonly used chemicals (*e.g.*, "soluble in water") were not annotated. Moreover, if a chemical-named entity acted as a part of another named entity, for example, disease or protein (*i.e.*, in phrases such as "cocaine abuse", "lactate dehydrogenase"), it was not annotated.

We distinguish three classes of extracted entities: "Substrate", "Metabolite", and "Chemical". The first two classes were assigned to the named entities that, according to the context of the whole abstract, were mentioned in the text as substrate or metabolite, respectively. Reconstruction of metabolic pathways when processing the results will make it possible to distinguish substrates from metabolites and substrates from parent compounds, *i.e.*, xenobiotics that undergo biotransformation. Any chemical named entity that was not assigned within the classes "Substrate" and "Metabolite" was considered "Chemical". It should be noted that if a chemical named entity was once clearly stated as a substrate or a metabolite (for example, in the sentence), then all other names and synonyms of this object were labeled the same way for the whole abstract.

Reaction named entities were also verified. As previously mentioned, we annotated only the names of the basic reactions. Often, in texts, authors specify the position of reactions, the number of added substituents, *etc.* In such cases, the named entity string itself and its positions in the text were manually adjusted by the authors of the corpus. For example, the basic reaction "hydroxylation" was recognized automatically with start



PMID 10024724

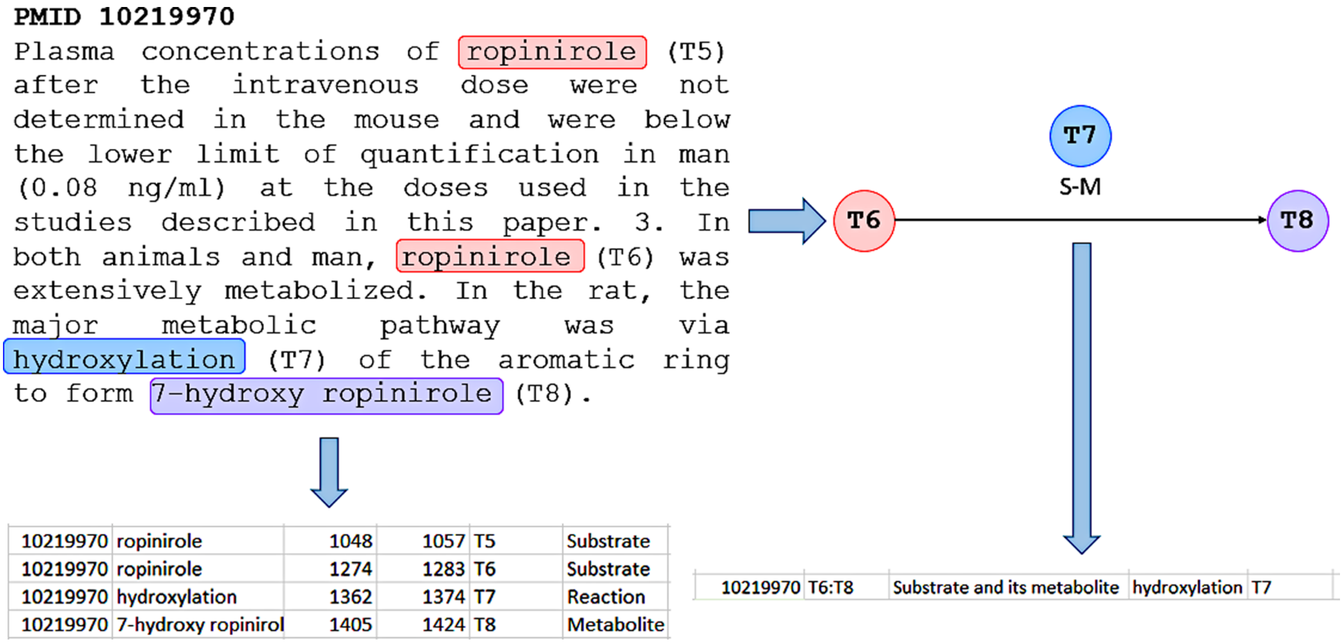
Methodological issues in biomonitoring of low level exposure to benzene (T1). Data from a pilot study on unmetabolized benzene (T2) and trans,trans muconic acid (T3) (t,t-MA (T4)) excretion in filling station attendants and unexposed controls were used to afford methodological issues in the biomonitoring of low benzene (T5) exposures (around 0.1 ppm). Urinary concentrations of benzene (T6) and t,t-MA (T7) were measured by dynamic head-space capillary GC/FID and HPLC, respectively.

10024724	benzene	64	70	T1	Substrate
10024724	benzene	114	120	T2	Substrate
10024724	trans,trans muconi	126	149	T3	Metabolite
10024724	t,t-MA	152	157	T4	Metabolite
10024724	benzene	293	299	T5	Substrate
10024724	benzene	355	361	T6	Substrate
10024724	t,t-MA	367	372	T7	Metabolite

10024724	T2:T3	Substrate and its metabolite
10024724	T2:T4	Substrate and its metabolite
10024724	T3:T4	Synonyms
10024724	T3:T5	Substrate and its metabolite
10024724	T4:T5	Substrate and its metabolite

...

Figure 2. Example of annotation of a text fragment that does not contain the names of biotransformation reactions. The relationships between substrates and metabolites are assigned as "S−M".



PMID 10219970

Plasma concentrations of ropinirole (T5) after the intravenous dose were not determined in the mouse and were below the lower limit of quantification in man (0.08 ng/ml) at the doses used in the studies described in this paper. 3. In both animals and man, ropinirole (T6) was extensively metabolized. In the rat, the major metabolic pathway was via hydroxylation (T7) of the aromatic ring to form 7-hydroxy ropinirole (T8).

10219970	ropinirole	1048	1057	T5	Substrate
10219970	ropinirole	1274	1283	T6	Substrate
10219970	hydroxylation	1362	1374	T7	Reaction
10219970	7-hydroxy ropinirol	1405	1424	T8	Metabolite

10219970	T6:T8	Substrate and its metabolite	hydroxylation	T7

Figure 3. Example of annotation of a text fragment containing the names of biotransformation reactions. The relationships between substrates and metabolites are assigned as "S−M".

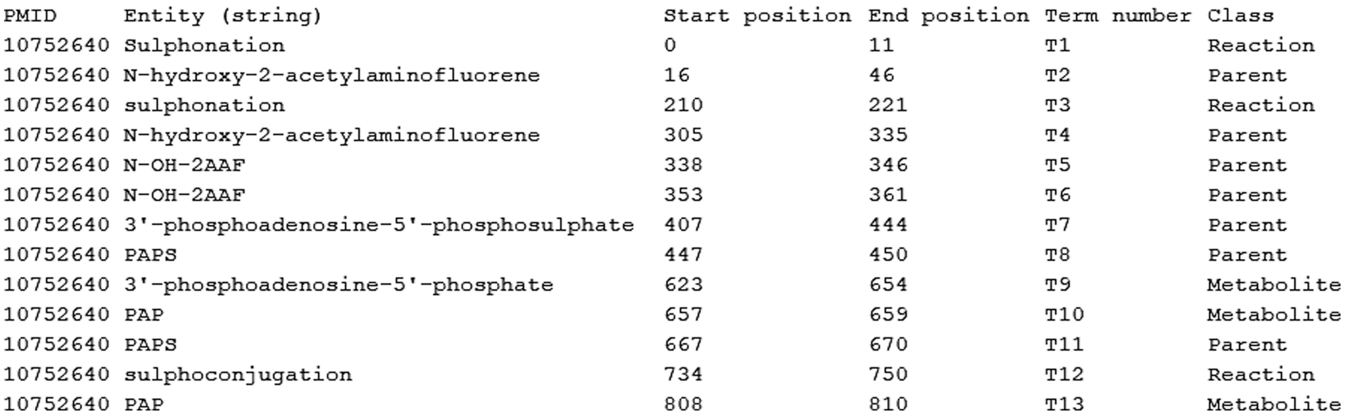
and end positions 298 and 310, respectively. However, the text mentions the name of the reaction "4-methylhydroxylation". Therefore, we added to the reaction-named entity the missing part "4-methyl" and corrected the start position from 298 to 290, since "4-methyl" consists of 8 symbols.

The annotation verification process involved checking the correctness of recognizing CNEs and assigning them to the class. Examples of the annotation are shown in Figures 2 and 3.

After manual verification of the CNE annotation, it was written in the text file according to the following format, partially corresponding to the representation in CHEMDNER/DrugProt[29,30] (Figure 4).

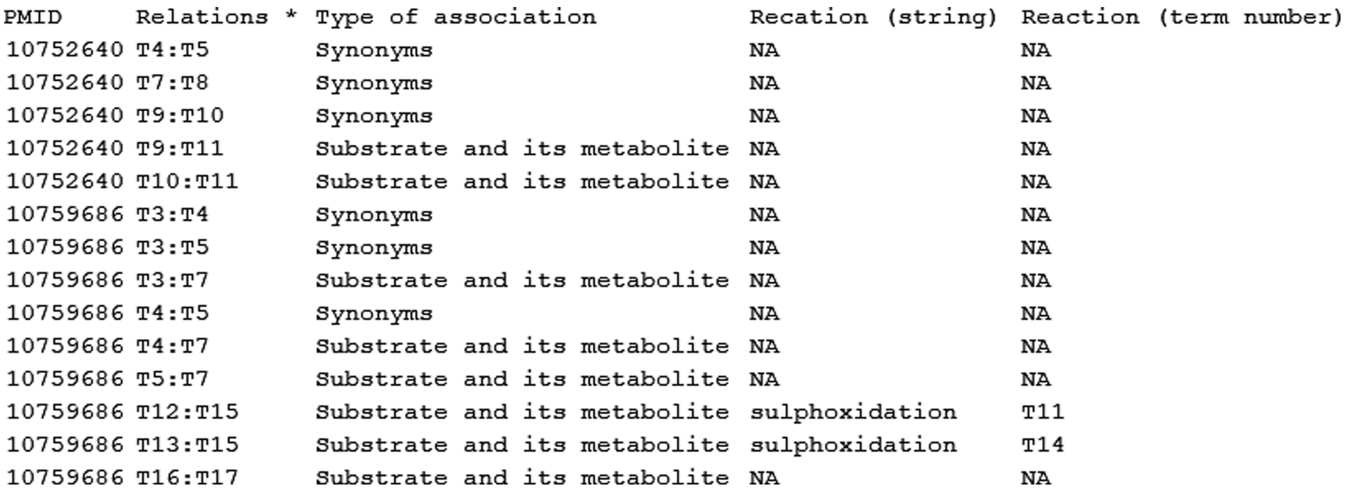
In the our corpus, we did not distinguish CNEs that came from the title and abstract body. The entire text for the annotation is obtained by concatenating the title and the abstract with the addition of a space symbol. Thus, the numbering of individual character positions begins with the title.

Then, we examined the extracted associations. We did not record associations between all substrates and metabolites if their names appeared several times in the text but indicated only those entities that were clearly marked in a specific fragment of text as a substrate−metabolite pair (Figure 2). Such an association was labeled with the class "Substrate and its metabolite". We assume that this way of substrate−metabolite pairs annotation can improve the machine-learning process,



```
PMID     Entity (string)                        Start position End position Term number Class
10752640 Sulphonation                           0              11           T1          Reaction
10752640 N-hydroxy-2-acetylaminofluorene        16             46           T2          Parent
10752640 sulphonation                           210            221          T3          Reaction
10752640 N-hydroxy-2-acetylaminofluorene        305            335          T4          Parent
10752640 N-OH-2AAF                              338            346          T5          Parent
10752640 N-OH-2AAF                              353            361          T6          Parent
10752640 3'-phosphoadenosine-5'-phosphosulphate 407            444          T7          Parent
10752640 PAPS                                   447            450          T8          Parent
10752640 3'-phosphoadenosine-5'-phosphate       623            654          T9          Metabolite
10752640 PAP                                    657            659          T10         Metabolite
10752640 PAPS                                   667            670          T11         Parent
10752640 sulphoconjugation                      734            750          T12         Reaction
10752640 PAP                                    808            810          T13         Metabolite
```

Figure 4. Example of the format for representing the annotated entities in XenoMet.



```
PMID     Relations * Type of association          Recation (string)  Reaction (term number)
10752640 T4:T5       Synonyms                     NA                 NA
10752640 T7:T8       Synonyms                     NA                 NA
10752640 T9:T10      Synonyms                     NA                 NA
10752640 T9:T11      Substrate and its metabolite NA                 NA
10752640 T10:T11     Substrate and its metabolite NA                 NA
10759686 T3:T4       Synonyms                     NA                 NA
10759686 T3:T5       Synonyms                     NA                 NA
10759686 T3:T7       Substrate and its metabolite NA                 NA
10759686 T4:T5       Synonyms                     NA                 NA
10759686 T4:T7       Substrate and its metabolite NA                 NA
10759686 T5:T7       Substrate and its metabolite NA                 NA
10759686 T12:T15     Substrate and its metabolite sulphoxidation     T11
10759686 T13:T15     Substrate and its metabolite sulphoxidation     T14
10759686 T16:T17     Substrate and its metabolite NA                 NA
```

Figure 5. Example of the format for representing the annotated relations between entities in XenoMet.

since it includes only those relationships that are semantically described in a specific piece of text (Figure 2). However, after additional automatic processing by searching for all possible combinations between the names of substrates and metabolites of annotated pairs and their synonyms, it is possible to obtain all kinds of associations for the entire text of the abstract and the title.

If it was clearly stated in the text that the metabolite was obtained from the substrate through a specific biotransformation reaction, we also indicated this and specified the position number of the reaction named the entity (Figure 3).

Since synonyms of the same compound are often found in texts, we also indicated these relations under the class "Synonyms". Nevertheless, if a pair of substrate−metabolite was mentioned in the text for all of the synonyms, we also wrote out the relationships for them.

At the end, we created a text file; the format is presented in Figure 5.

Examples of the Corpus Usage. *Constructions of Metabolic Trees.* To demonstrate the application of XenoMet for extracting information about xenobiotic metabolism, we first built a network of interactions between entities (*i.e.*, xenobiotics and their metabolites) that were annotated in the texts of XenoMet. We performed an automatic search through Python and provided an API within the ChEMBL database[31] for the chemical named entities from XenoMet to obtain unique chemical identifiers and reduce the number of repeating nodes.

Biotransformation reactions, if mentioned, were converted to the CytoScape platform[32] in an acceptable XGMML format. Chemical entities were classified as either substrates (blue nodes) or metabolites (red nodes). If a chemical-named entity was labeled as both a substrate and a metabolite in any of the texts, the node was colored purple. We added labels for the reactions of biotransformation that were specified in the text of the corpus.

Application of the Developed Corpus for Classifying Names of Substrates and Metabolites using a CRF-Based Approach. We used a previously developed CRF-based approach[13,33,34] to estimate the applicability of the developed corpus for extracting the names of substrates, metabolites, and reactions with the relation extraction between them.

Briefly, a CRF-based approach for the extraction of chemical named entities considers the text as a sequence of its elementary units, tokens, which could be represented by a word or even one symbol. These tokens are then described by a set of descriptors that reflect their various features (for example, part of speech, number of characters, and presence of capital letters/numbers/symbols).

In this study, we used the approach described above as a basis, but in contrast to the earlier developed approach for the CNE recognition, in this study, our goal is to separate substrate, metabolite, and reaction tokens from each other as well as from a chemical that does not belong to any of these three aforementioned classes. In this case, we used belonging to the

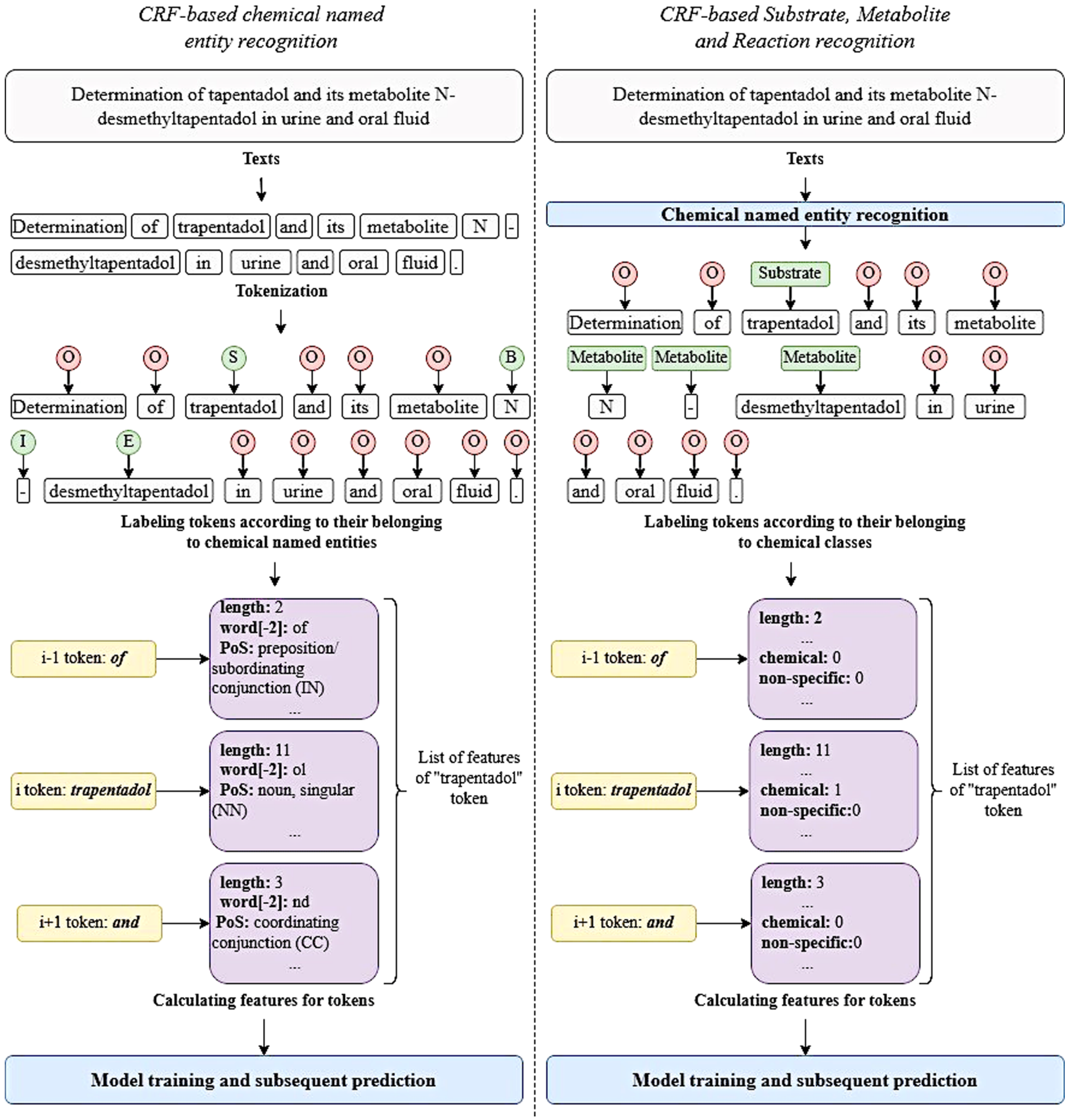

CRF-based chemical named entity recognition
Determination of tapentadol and its metabolite N-desmethyltapentadol in urine and oral fluid
Texts
Determination | of | trapentadol | and | its | metabolite | N | - | desmethyltapentadol | in | urine | and | oral | fluid | .
Tokenization
O O S O O O B I E O O O O O O
Labeling tokens according to their belonging to chemical named entities
i-1 token: of
length: 2
word[-2]: of
PoS: preposition/ subordinating conjunction (IN)
...
i token: trapentadol
length: 11
word[-2]: ol
PoS: noun, singular (NN)
...
i+1 token: and
length: 3
word[-2]: nd
PoS: coordinating conjunction (CC)
...
List of features of "trapentadol" token
Calculating features for tokens
Model training and subsequent prediction
CRF-based Substrate, Metabolite and Reaction recognition
Determination of tapentadol and its metabolite N-desmethyltapentadol in urine and oral fluid
Texts
Chemical named entity recognition
O O Substrate O O O Metabolite Metabolite Metabolite O O O O O O
Labeling tokens according to their belonging to chemical classes
i-1 token: of
length: 2
...
chemical: 0
non-specific: 0
...
i token: trapentadol
length: 11
...
chemical: 1
non-specific:0
...
i+1 token: and
length: 3
...
chemical: 0
non-specific:0
...
List of features of "trapentadol" token
Calculating features for tokens
Model training and subsequent prediction

Figure 6. Algorithm for the recognition of substrate, metabolite, and reaction-named entities. Labels "S", "B", "I", "E", and "O" stand for tokens that form (S), begin (B), are inside of (I), or end (E) a chemical-named entity, or do not belong to any of them (O).[13,33,34] The descriptor "length" is the number of symbols within the token, "word[-2]" is the last two symbols of the token, "PoS" is part of speech, "chemical" is whether the token belongs to a chemical-named entity, and "nonspecific" is whether the token belongs to nonspecific terms (1 if it does, 0 if it does not, for the last two descriptors).

chemical named entities as descriptors: "1", if the token belongs to any of them, and "0" if it does not. An illustration of the algorithm for recognizing chemical-named entities and their further classification into classes (Substrate, Metabolite, and Reaction) is presented in Figure 6.

To take the context into account, we created dictionaries of nonspecific terms that would point at the entities of the listed classes. To do this, we tokenized texts of XenoMet, removed all of the stop-words (prepositions, introductory words, *etc.*), defined labels belonging to a substrate, metabolite, or reaction for tokens, and identified the most frequently occurring tokens at a distance of 1 to 5 tokens from one that belongs to a substrate, metabolite, or reaction.

We used numerical values of frequency as the descriptors of the context. Those numerical values were calculated as the ratio of the number of cases N_t when a specific token was encountered at a given position to the number N of all token variants at a given position. Such frequencies were calculated for the substrate and metabolite classes.

We also added a specific descriptor to recognize the reaction-named entities. Since their names have the most deterministic form (a finite number of stems and specific endings), we used a dictionary of basic reactions (described above) for automated recognition. By removing endings such as "-ation", "-ysis", and "-ing", we fulfilled a list of reaction names stems. The descriptor took the value 1 if at least one stem from the list was a substring in the token string and 0 if it was not.

In order to study the impact of these new descriptors on prediction accuracy, we built 5 models with various combinations of the aforementioned context descriptors and assessed the predictive ability and execution time of each model with 5-fold cross-validation. Then, we selected the best of them based on the correlation between prediction quality and execution time and optimized the hyperparameters.

RESULTS AND DISCUSSION

XenoMet—The Corpus for Extracting Relationship between Substrates and Metabolites. We created XenoMet for the automated recognition of substrates, metabolites, and biotransformation reactions based on an automated approach using patterns developed for the extraction of substrates and metabolites[13] followed by manual verification of recognized entities.

Unlike the CNEs, which were recognized by the HunFlair library and with the use of a CRF-based approach, manual verification of reaction entities required significant efforts. In the Materials and Methods section, we mentioned that we used a dictionary of basic reactions for the purpose of recognizing the reaction-named entities.

After manual verification, we analyzed the composition of the final corpus. The number and shares of the named entities belonging to various classes are shown in Figure 7. Figure 7 shows that the largest part of the named entities assigned to the class "Substrate". The "Chemical" and "Metabolite" classes both take 25% of all named entities, and the "Reaction" class takes only 8%. Since the reaction named entities were not often present in the text, only 10% of "substrate−metabolite" pairs were assigned to the biotransformation reaction. The average number of named entities of all classes, including synonyms, per title and abstract, was 20.

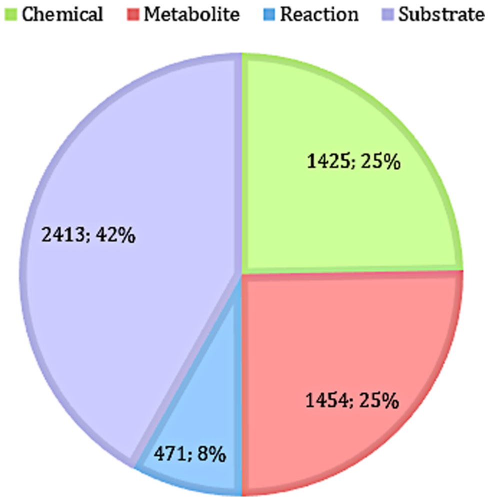


Figure 7. Ratio of the number of entities belonging to the classes "substrate", "metabolite", "reaction", and "chemical" in the annotated corpus.

Figure 7 shows that XenoMet contains much more terms specific to the xenobiotic metabolism than nonspecific chemical named entities. This is a consequence of the fact that a thorough expert selection of relevant texts for the corpus was carried out. In some cases (about 40−50%), abstract texts lack the correct (full) chemical names of substances and contain only derivatives of a particular chemical class or compound. Examples include "M1, M2, and M3" as codes for metabolites, "methyl hydroxylated and N-oxidized metabolite" in terms of voriconazole biotransformation,[35] or "conjugates of XK469 with glycine, taurine, and glucuronic acid".[36] This may be because the corresponding studies are aimed at characterizing the particular biotransformation process of a drug or bioactive compound rather than at determining the structures of its metabolites. These metabolite named entities can hardly be considered a chemical named entity in principle, and only low-informative tree of metabolism can be built based on such entities. Therefore, we omitted them and did not annotate them in the corpus. For these reasons, despite the careful selection of texts for the corpus, the proportion of examples containing the names of metabolites is relatively small.

Using the names of basic reactions for the preliminary annotation of the corpus allowed the time of manual processing to be reduced significantly: only 5% of the names of reactions were annotated *de novo*. Thirty-six percent of all basic reaction names were found in the text either in the same form or as part of detailed reaction names. In the vast majority of cases, the names of the basic reactions were found in the text without changes (Table 1).

As one may see from the table, there are basic reactions that often need to be more clearly specified in the texts, such as hydroxylation, dealkylation, or demethylation. However, some basic reactions do not require clarification since the reaction name clearly describes the mechanism of biotransformation, such as lactonization or decarboxylation.

We also analyzed the content of titles and abstracts of publications with the names of substrates, metabolites, biotransformation reactions, and associations between them (Table 2).

Table 2 shows that despite the small amount of text, titles and abstracts are of interest for extracting information since they contain a significant number of named entities. For example, 12% of all names of substrates in the text can be extracted from the titles. At the same time, "substrate−metabolite" pairs with indication of biotransformation reaction are found quite rarely in titles, which is not surprising: the title is intended to introduce the main concept of the study, and the process that results in the transformation of a xenobiotic into a metabolite is more detailed information.

Text-Mining of Xenobiotic Transformation as a Source of Biomedical Knowledge: Case Study for Metabolic Tree Generation. The extraction of xenobiotics and their metabolites from texts could be a useful source of knowledge in many fields of biomedical science. This approach can find its main application in pharmacology: for example, information about the metabolism of pharmaceutical agents may not be complete within one text, but further path of biotransformation is disclosed in others. Automated text processing methods will make it possible to extract this information from large arrays of texts and present it in a format convenient for further analysis by an expert.

Table 1. Number of Reaction-Named Entities in the XenoMet Corpus that Contain Names of Basic Reactions as a Part or Exactly Match Them

basic reaction (BR)	N reaction names that are BR	N detailed reaction names extracted using BR	basic reaction (BR)	N reaction names that are BR	N detailed reaction names extracted using BR
hydroxylation	134	222	protein binding	15	0
oxidation	195	100	aromatization	8	2
glucuronidation	191	20	ring opening	8	1
hydrolysis	118	11	alkylation	9	0
demethylation	20	109	esterification	3	6
conjugation	69	1	deamination	7	1
reduction	43	19	decarboxylation	8	0
dealkylation	7	47	elimination	8	0
methylation	38	10	adduct formation	7	0
cleavage	33	10	lactonization	6	0
sulfation	23	2	carboxylation	6	0
covalent binding	24	0	glucosidation	0	6
acetylation	11	10	isomerization	4	1
hydrogenation	0	17	dehydration	4	1
phosphorylation	14	1	oxygenation	2	3

Table 2. Numbers and Proportions of Substrate, Metabolite, and Reaction Named Entities and Pairs "Substrate−Metabolite" in the Text of Titles and Abstracts Separately

object type	title	abstract	title + abstract
substrate-named entity	1014 (12%)	7333 (88%)	8347
metabolite-named entity	182 (5%)	3233 (95%)	3415
reaction-named entity	102 (6%)	1655 (94%)	1757
substrate−metabolite pair without specified reaction of biotransformation	105 (6%)	1747 (94%)	1852
substrate−metabolite pair with specified reaction of biotransformation	2 (1%)	166 (99%)	168

As a demonstration of the application of text-mining approaches to retrieving information about the biotransformation of xenobiotics, we constructed and visualized an interaction network between substrates and metabolites annotated in XenoMet.

To reduce the number of repeating nodes and, as a result, the number of repeating edges, we performed automated queries to the ChEMBL database in order to unify the extracted named entities. The database query consisted of a chemical named entity. We carried out a comparison of the extracted chemical names with synonyms of the query results in lower case; nonidentical terms were not assigned to each other. For those entities that were not identified in ChEMBL, we performed a search across relations with the class synonyms. Most of the named entities in the corpus linked to the ChEMBL identifier belong to the "Substrate" class. Thus, the probability of encountering repeating nodes is significantly higher for the "Metabolite" class.

As a result, for more than 3000 binary "Substrate and its metabolite" associations and, accordingly, for more than 6000 of their participants, it was possible to obtain 795 unique nodes that could be assigned to the ChEMBL identifiers and 2543 that were not identified in the ChEMBL database and might lead to duplication. Despite the large number of nodes, nevertheless, the filtering attempts made it possible to reduce the total number of associations by half. Lists of CNEs that were or were not identified in ChEMBL are presented in the Supporting Information.

To distinguish those nodes in the interaction network, we used the color intensity. Thus, the nodes reflecting the names of chemical compounds that were identified in the ChEMBL database are more intense. As was already mentioned in the Materials and Methods section, we highlighted the Substrate nodes in blue, the Metabolite nodes in red, and those compounds that acted as both metabolized compounds and reaction products by mixing blue and red, resulting in purple.

The complete network of interactions in the CytoScape session format is presented in the Supporting Information. Since the final network of interactions is quite cumbersome, we present several fragments to illustrate the work done.

The network consists of several types of associations. A significant part of them does not form a subnet of more than two nodes (Figure 8A). As a rule, these associations are found in texts where the concentration of an already known xenobiotic metabolite in biological fluids was studied or as part of studies of the drug metabolizing enzymes (DMEs) activity ([37] for scutellarein and[38] for harmaline and harmine). Some of them are very extensive and include many nodes (Figure 8B). Such chains of biotransformation mechanisms are most often observed for drugs that are used in the treatment of neoplastic diseases, since they are cytotoxic and the knowledge of their changes in the human body is important for determining the dose and prognosis of drug intake[18−20] as well as for psychoactive compounds, for which the knowledge of metabolites in various tissues allows the identification of abuse cases.[23,24] In Figure 8B, the pathway of biotransformation of the cytotoxic compound irinotecan (CPT-11) is presented. As one can see, there are 4 types of nodes; therefore, unlike in the previous example, there may be repeated nodes. Indeed, the intense purple node "7-ethyl-10-hydroxycamptothecin" (node 9 in Figure 8B), which is a known metabolite of irinotecan, is also called SN-38. The faded red knot "7-ethyl-10-hydroxycamptothecine" (node 6 in Figure 8B) is actually another way to spell this compound, and it should be merged with an intense purple node due to the same chemical they reflect. Also, two faded nodes—the red one with the label "7-ethyl-10-hydroxycamptothecin glucuronide" (node 11 in Figure 8B) and the purple one "sn-38 glucuronide" (node 8 in Figure 8B)—reflecting the same

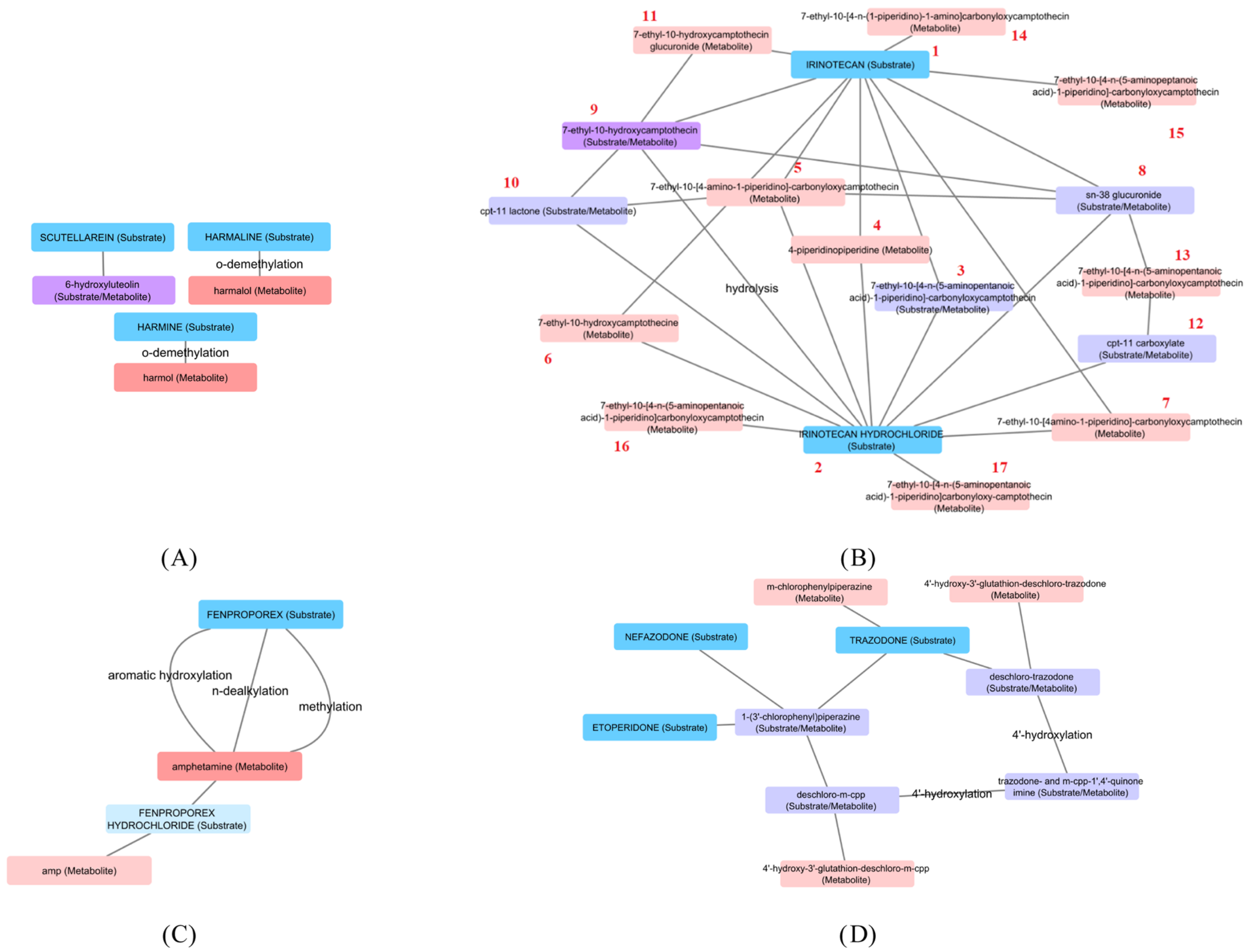

SCUTELLAREIN (Substrate)
6-hydroxyluteolin (Substrate/Metabolite)
HARMALINE (Substrate)
o-demethylation
harmalol (Metabolite)
HARMINE (Substrate)
o-demethylation
harmol (Metabolite)
(A)
11
7-ethyl-10-hydroxycamptothecin glucuronide (Metabolite)
7-ethyl-10-[4-n-(1-piperidino)-1-amino]carbonyloxycamptothecin (Metabolite)
14
1
IRINOTECAN (Substrate)
7-ethyl-10-[4-n-(5-aminopeptanoic acid)-1-piperidino]-carbonyloxycamptothecin (Metabolite)
15
9
7-ethyl-10-hydroxycamptothecin (Substrate/Metabolite)
10
cpt-11 lactone (Substrate/Metabolite)
5
7-ethyl-10-[4-amino-1-piperidino]-carbonyloxycamptothecin (Metabolite)
8
sn-38 glucuronide (Substrate/Metabolite)
4
4-piperidinopiperidine (Metabolite)
13
7-ethyl-10-[4-n-(5-aminopentanoic acid)-1-piperidino]-carbonyloxycamptothecin (Metabolite)
hydrolysis
3
7-ethyl-10-[4-n-(5-aminopentanoic acid)-1-piperidino]-carbonyloxycamptothecin (Substrate/Metabolite)
7-ethyl-10-hydroxycamptothecine (Metabolite)
6
12
cpt-11 carboxylate (Substrate/Metabolite)
7
7-ethyl-10-[4-n-(5-aminopentanoic acid)-1-piperidino]carbonyloxycamptothecin (Metabolite)
16
IRINOTECAN HYDROCHLORIDE (Substrate)
7-ethyl-10-[4amino-1-piperidino]-carbonyloxycamptothecin (Metabolite)
2
17
7-ethyl-10-[4-n-(5-aminopentanoic acid)-1-piperidino]carbonyloxy-camptothecin (Metabolite)
(B)
FENPROPOREX (Substrate)
aromatic hydroxylation
n-dealkylation
methylation
amphetamine (Metabolite)
FENPROPOREX HYDROCHLORIDE (Substrate)
amp (Metabolite)
(C)
m-chlorophenylpiperazine (Metabolite)
4'-hydroxy-3'-glutathion-deschloro-trazodone (Metabolite)
NEFAZODONE (Substrate)
TRAZODONE (Substrate)
deschloro-trazodone (Substrate/Metabolite)
ETOPERIDONE (Substrate)
1-(3'-chlorophenyl)piperazine (Substrate/Metabolite)
4'-hydroxylation
trazodone- and m-cpp-1',4'-quinone imine (Substrate/Metabolite)
deschloro-m-cpp (Substrate/Metabolite)
4'-hydroxylation
4'-hydroxy-3'-glutathion-deschloro-m-cpp (Metabolite)
(D)

Figure 8. Fragments of interaction networks between substrates and metabolites annotated in XenoMet for subnets (A) of only two nodes, (B) of many nodes and edges, (C) that do not specify biotransformation intermediates, and (D) in which the same chemical compound may be a biotransformation product of several substrates.

thing. Moreover, two CNEs identified in ChEMBL—"irinotecan" and "irinotecan hydrochloride" (nodes 1 and 2 in Figure 8B)—might be considered synonyms since the second chemical is used in pharmaceutical formulations. However, their structures are different, and for this reason, the ChEMBL identifiers are also different. The node with number 3 in Figure 8B is colored purple, which means that this compound undergoes further biotransformation but none of the final metabolites are presented in the network. It is explained by the fact that the ways of its further changes are mentioned, but the final metabolites are either not named or not stated explicitly; that is, they do not allow the structure to be unambiguously identified by CNE.

In rare cases, for xenobiotics, a list of biotransformation reactions is indicated without mentioning intermediates but immediately the final product. In this case, there may be several connections between two nodes at once (Figure 8C).[39] Such situations were observed for abstracts due to frequent word limits. We assume that the text may contain the names of reaction intermediates.

Among the complete interaction networks, branched subnetworks are presented that describe the metabolism of not just one compound but several. This is typical of parent compounds (or substrates) of the same chemical group, as a result of which they have similar biotransformation pathways and similar metabolites (Figure 8D). For example, antidepressants trazodone, nefazodone, and etoperidone are metabolized to a common metabolite, 1-(3′-chlorophenyl)piperazine (m-CPP), which, in turn, also undergoes biotransformation to final metabolites.[40]

Application of XenoMet for the Recognition of Substrate-, Metabolite-, and Reaction-Named Entities by a CRF-Based Approach. Application of a previously developed CRF-based chemical named entity recognition with a rule-based approach resulted in quite low values of precision, recall, and F_1-score.[13] The use of patterns and rules based on them is less effective than machine-learning methods due to the inability to extract all kinds of patterns from texts manually.

We performed training of multiple CRF-based models that we used for the extraction of CNEs earlier,[13] additionally in the current study, we used labels of tokens' belonging to chemical named entities as descriptors. Furthermore, we expanded the list of descriptors to apply them for recognizing names of substrates, metabolites, and reactions. This was achieved by including descriptors such as belonging to the stems of reaction named entities and frequencies of nonspecific terms at various distances from the substrate, metabolite, and reaction named entities. In

https://doi.org/10.1021/acsomega.4c05723

the following text, we use the term "extended descriptors" to refer to the list described, and "standard descriptors" to refer to the list used for CNER.[13]

As it was expected, the usage of extended list of descriptors resulted in an elevation of accuracy in comparison with standard for CNER descriptors:[13] for some metrics, the improvement was around 0,07. Since the use of an extended list of descriptors allowed, albeit slightly, an increase in the recognition accuracy, for the next steps, we used only the extended list of descriptors.

Then, we investigated the role of the context volume included in the text description in the recognition accuracy. We assume that using a larger context window (*e.g.*, five before and after tokens *versus* one) will lead to an improved prediction accuracy, because this window is more likely to include tokens that reflect the semantic meaning of a particular chemical named entity (substrate, metabolite, or reaction). However, models become more time-consuming as the volume of context increases.

For further studies, we selected a model with the highest performance. This model takes the current token and three tokens before and after it. Full results of the model evaluation with 5-fold cross-validation are provided in Supporting Information. The results of the evaluation with 5-fold cross-validation for the best model after hyperparameter optimization are presented in Table 3. This model considers a context window of three tokens and uses the extended list of the descriptors.

Table 3. Metabolism-Related Named Entity Recognition Accuracy Using the Best Model Achieved with Five-Fold Cross-Validation

class	precision	recall	F_1-score
substrate	0,71 (±0,02)	0,69 (±0,03)	0,70 (±0,02)
chemical	0,61 (±0,03)	0,67 (±0,03)	0,64 (±0,02)
metabolite	0,71 (±0,03)	0,71 (±0,02)	0,71 (±0,02)
reaction	0,97 (±0,01)	0,96 (±0,02)	0,96 (±0,01)

The precision, recall, and F_1-score metric values obtained with 5-fold cross-validation are significantly lower than those we obtained when training a model for recognizing the names of chemical compounds.[13,33,34] It could be explained by a bigger diversity between the tokens of chemical named entities and all other words in comparison with the diversity between the tokens of three different classes (substrate, metabolite, and chemical) (Table 3).

Such a difference between chemical named entities recognition and substrate−metabolite−reaction classification accuracy can apparently also be explained by the diverse context of one chemical named entity, even though the class for it was defined based on the whole. For example, in the text:[41] *"Treatment A consisted of a single dose of avosentan (T4) 50 mg after a high-fat, high-calorie breakfast. Treatment B consisted of a single dose of 50 mg of avosentan (T5) administered in the fasted state. Plasma concentrations of avosentan (T6) and its hydroxymethyl (T7) metabolite Ro-68−5925 (T8) were measured by liquid chromatography-tandem mass spectrometry"*, the first two sentences do not define "avosentan" as substrate, but the last one does. Therefore, this entity would be annotated as the Substrate for all of these three sentences. Obviously, in this case, the context of use of the "avosentan" drug name will vary dramatically, which will make it difficult for the model to identify patterns of use of certain classes representatives.

The experiment showed that the names of reactions were recognized equally well, in the case of inclusion in the list of descriptors belonging to the bases of the reactions and without them (Supporting Information). Apparently, the reaction-named entities vary greatly from the other text, and the ending of a token (the last two and three symbols), which is a part of a standard descriptors set, is enough to distinguish such cases.

Place of XenoMet among the Existing Corpora Related to the Extraction of Information about the Xenobiotics Biotransformation from Texts. There are some previous approaches for extracting data concerning the biotransformation of chemical compounds, with one particular method involving the creation of a corpus entitled Metabolic Entities. This corpus contains annotations of proteins/genes, metabolites, and metabolic events[19] and is designed to automate the extraction of key players in metabolic pathways, relying on text-based information. Texts for this corpus were randomly selected from those used to create the EcoCyc database and then annotated manually by the authors. A total of 271 titles and abstracts were annotated for this corpus. It should be noted that, according to the illustration provided in the cited article, "Metabolite" refers to any chemical compound related to a "protein/gene-metabolic event". This results in a less clear distinction between substrates and metabolites annotated in the corpus.

There is a MetaboListem corpus aimed at the annotation of metabolomics terms in the literature.[42] This corpus includes over 1200 full-text articles that were automatically annotated by rule-based methods. The automatic rule-based annotation was validated by using machine-learning models. However, in comparison to our corpus, MetaboListem was annotated for the metabolome, and all entities are labeled as metabolites, comprising mainly the chemical compounds produced from endogenous substances within the human body. As a result, it is better suited for examining metabolomic processes than for investigating the metabolism of xenobiotics.

In addition to corpora aimed at analyzing metabolomics pathways in the human body and investigating metabolism, there are approaches designed to automatically extract metabolites and metabolic pathways from scientific research articles.

Kongburan et al. developed a hybrid system using machine-learning, dictionary, and rule-based approaches to reconstruct metabolic pathways from article texts.[43] The recognition of metabolites and proteins/genes was performed by CRF and dictionary-based methods using the previously developed Metabolic Entities corpus, complemented by GENIA[44] and Thyroid Cancer Intervention Event.[43] To perform the recognition of metabolic events, the authors created a dictionary based on metabolic events present in the Metabolic Entities corpus. The recognition of metabolite named entities was performed with an F_1-score equal to 92% and an enzyme entity of 85%. We should note that a higher recognition accuracy compared to our method is apparently due to the fact that it allows for the recognition of two different types of objects (chemicals and proteins) and is not intended for the task of distinguishing separate groups within one type (chemicals).

A rule-based approach to identify pathways from the literature was considered by Czarnecki et al.[45] For the named entity recognition, protein/gene, and small compounds, the author used third-party approaches. The relation extraction was performed using patterns: to identify substrates and products, they should follow a certain rule in the text. Unfortunately, the authors do not provide any accuracy metrics that could give an idea of the quality of the developed model in solving the problem of extracting associations between substrates and reaction products.

In contrast to the studies described by Kongburan et al. and Czarnecki et al., our approach aims at the creation of a unique text corpus, XenoMet, which can be used for the extraction of substrates and metabolites of xenobiotics. The created corpus can also be used to enrich the data and knowledge on xenobiotic transformation as part of the overall metabolism as a complex process in the human body.

■ CONCLUSIONS

In this paper, we describe XenoMet text corpus, which enables the identification of substrates, metabolites, and reactions, as well as the search for associations between them. Unlike previous corpora, XenoMet provides a distinct categorization of substrates and metabolites, particularly those of xenobiotics, and includes information about biotransformation reactions. As our corpus contains not only substrate, metabolite, and reaction named entities but also the terms for chemical compounds that do not belong to any of these categories, it can be used with the existing annotated corpora for training models in the recognition of chemical named entities. XenoMet will complement the procedures used to extract details about the biotransformation of xenobiotics from texts, which will enrich scientific knowledge in this field. Furthermore, it can be used for building models for predicting the reactions of biotransformation and metabolites structures.

■ ASSOCIATED CONTENT

Data Availability Statement

Annotations and texts selected for XenoMet corpus are freely available at GitHub repository with the following link: https://github.com/nad-smol/XenoMet. An in-house developed script that classifies the chemical-named entities according to their belonging substrates or metabolites is presented in the Supporting Information. As a part of the Supporting Information, the basic reaction-named entities are also freely available. HunFlair script were available at the Web site [https://github.com/flairNLP/flair/blob/master/resources/docs/HUNFLAIR.md], provided by its developer. Corresponding papers citing HunFlair are provided in the text of the manuscript.

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c05723.

> Results of various CRF models evaluation *via* 5-fold cross-validation; lists of chemical named entities of XenoMet corpus that were and were not identified in ChEMBL through automated queries; and the list of basic reaction names (XLSX)
>
> Python script, which performs an annotation of substrate and metabolite named entities based on the rules, requires a preliminary annotation of chemical named entities (PY); CytoScape (CYS) session of a metabolic tree is built based on the annotated within XenoMet corpus biotransformation data [https://github.com/flairNLP/flair/blob/master/resources/docs/HUNFLAIR.md] (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Nadezhda Yu. Biziukova − *Institute of Biomedical Chemistry, Moscow 119121, Russian Federation;* orcid.org/0000-0002-2044-1327; Email: nad.smol@gmail.com

Olga A. Tarasova − *Institute of Biomedical Chemistry, Moscow 119121, Russian Federation;* orcid.org/0000-0002-3723-7832; Email: olga.a.tarasova@gmail.com

Authors

Anastasia V. Rudik − *Institute of Biomedical Chemistry, Moscow 119121, Russian Federation;* orcid.org/0000-0002-8916-9675

Alexander V. Dmitriev − *Institute of Biomedical Chemistry, Moscow 119121, Russian Federation;* orcid.org/0000-0002-2431-3429

Dmitry A. Filimonov − *Institute of Biomedical Chemistry, Moscow 119121, Russian Federation;* orcid.org/0000-0002-0339-8478

Vladimir V. Poroikov − *Institute of Biomedical Chemistry, Moscow 119121, Russian Federation;* orcid.org/0000-0001-7937-2621

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c05723

Author Contributions

O.T., conceptualization; N.B., corpus compilation;. N.B. and O.T. corpus review and editing; N.B. and O.T., manuscript writing; A.R., A.D., D.F., and V.P., manuscript review and editing; D.F. and V.P., project supervision.

Funding

The study was performed in the framework of the Program for Basic Research in the Russian Federation for a long-term period 2021−2030 (project No. 122030100170−5).

Notes

The authors declare no competing financial interest.

■ ABBREVIATIONS

5-F CV:5-fold cross-validation; CRF:conditional random fields; DME:drug metabolizing enzymes; LOO CV:leave-one-out cross-validation; M:metabolite; MOP:molecular process ontology; NB:Naïve Bayes; SOMs:sites of metabolism; S:substrate

■ REFERENCES

(1) Isin, E. M. Unusual Biotransformation Reactions of Drugs and Drug Candidates. *Drug Metab. Dispos.* **2023**, *51* (4), 413−426.

(2) Jia, J.; Zhu, F.; Ma, X.; Cao, Z.; Cao, Z. W.; Li, Y.; Li, Y. X.; Chen, Y. Z. Mechanisms of Drug Combinations: Interaction and Network Perspectives. *Nat. Rev. Drug Discovery* **2009**, *8* (2), 111−128.

(3) Yin, W.; Al-Wabli, R. I.; Attwa, M. W.; Rahman, A. F. M. M.; Kadi, A. A. Detection and Characterization of Simvastatin and Its Metabolites in Rat Tissues and Biological Fluids Using MALDI High Resolution Mass Spectrometry Approach. *Sci. Rep.* **2022**, *12* (1), No. 4757.

(4) Vaidyanathan, S.; Goodacre, R. Quantitative Detection of Metabolites Using Matrix-assisted Laser Desorption/Ionization Mass Spectrometry with 9-aminoacridine as the Matrix. *Rapid Commun. Mass Spectrom.* **2007**, *21* (13), 2072−2078.

(5) Tyndyk, M. L.; Zabezhinski, M. A.; Bykov, V. J.; Dikun, P. P.; Dymochka, L. A.; Nepomnyaschaya, O. B.; Yatsuk, O. S.; Yermilov, V. B.; Likhachev, A. J. Individual Values of Excretion of Benzo[a]Pyrene Metabolites and Susceptibility to Its Carcinogenic Effect in Rats. *Cancer Lett.* **1994**, *78* (1−3), 163−170.

(6) Gonzalez, G.; Cohen, K. B.; Greene, C. S.; Hahn, U.; Kann, M. G.; Leaman, R.; Shah, N.; Ye, J. In *Text and Data Mining for Biomedical*

Discovery-Session Introduction, Pacific Symposium on Biocomputing; World Scientific, 2013; pp 368−372.

(7) Krallinger, M.; Rabal, O.; Lourenço, A.; Oyarzabal, J.; Valencia, A. Information Retrieval and Text Mining Technologies for Chemistry. *Chem. Rev.* **2017**, *117* (12), 7673−7761.

(8) Porokhin, V.; Liu, L.-P.; Hassoun, S. Using Graph Neural Networks for Site-of-Metabolism Prediction and Its Applications to Ranking Promiscuous Enzymatic Products. *Bioinformatics* **2023**, *39* (3), No. btad089.

(9) Kirchmair, J.; Williamson, M. J.; Tyzack, J. D.; Tan, L.; Bond, P. J.; Bender, A.; Glen, R. C. Computational Prediction of Metabolism: Sites, Products, SAR, P450 Enzyme Dynamics, and Mechanisms. *J. Chem. Inf. Model.* **2012**, *52* (3), 617−648.

(10) Tyzack, J. D.; Kirchmair, J. Computational Methods and Tools to Predict Cytochrome P450 Metabolism for Drug Discovery. *Chem. Biol. Drug Des.* **2019**, *93* (4), 377−386.

(11) Rudik, A. V.; Dmitriev, A. V.; Lagunin, A. A.; Filimonov, D. A.; Poroikov, V. V. Prediction of Reacting Atoms for the Major Biotransformation Reactions of Organic Xenobiotics. *J. Cheminform.* **2016**, *8*, No. 68.

(12) Cai, Y.; Yang, H.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. Computational Prediction of Site of Metabolism for UGT-Catalyzed Reactions. *J. Chem. Inf. Model.* **2019**, *59* (3), 1085−1095.

(13) Tarasova, O. A.; Biziukova, N. Y.; Rudik, A. V.; Dmitriev, A. V.; Filimonov, D. A.; Poroikov, V. V. Extraction of Data on Parent Compounds and Their Metabolites from Texts of Scientific Abstracts. *J. Chem. Inf. Model.* **2021**, *61* (4), 1683−1690.

(14) Wang, Y.; Zhang, S.; Li, F.; Zhou, Y.; Zhang, Y.; Wang, Z.; Zhang, R.; Zhu, J.; Ren, Y.; Tan, Y.; Qin, C.; Li, Y.; Li, X.; Chen, Y.; Zhu, F. Therapeutic Target Database 2020: Enriched Resource for Facilitating Research and Early Development of Targeted Therapeutics. *Nucleic Acids Res.* **2020**, *48* (D1), D1031−D1041.

(15) Rudik, A.; Dmitriev, A.; Lagunin, A.; Filimonov, D.; Poroikov, V. MetaPASS: A Web Application for Analyzing the Biological Activity Spectrum of Organic Compounds Taking into Account Their Biotransformation. *Mol. Inf.* **2021**, *40* (4), No. 2000231.

(16) Biziukova, N. Y.; Tarasova, O. A.; Rudik, A. V.; Filimonov, D. A.; Poroikov, V. V. Automatic Recognition of Chemical Entity Mentions in Texts of Scientific Publications. *Autom. Doc. Math. Linguist.* **2020**, *54* (6), 306−315.

(17) Patumcharoenpol, P.; Doungpan, N.; Meechai, A.; Shen, B.; Chan, J. H.; Vongsangnak, W. An Integrated Text Mining Framework for Metabolic Interaction Network Reconstruction. *Peer J.* **2016**, *4*, No. e1811.

(18) Weber, L.; Sänger, M.; Münchmeyer, J.; Habibi, M.; Leser, U.; Akbik, A. HunFlair: An Easy-to-Use Tool for State-of-the-Art Biomedical Named Entity Recognition. *Bioinformatics* **2021**, *37* (17), 2792−2794.

(19) Bonuccelli, U.; Del Dotto, P.; Lucetti, C.; Petrozzi, L.; Bernardini, S.; Gambaccini, G.; Rossi, G.; Piccini, P. Diurnal Motor Variations to Repeated Doses of Levodopa in Parkinson's Disease. *Clin. Neuropharmacol.* **2000**, *23* (1), 28−33.

(20) Sinou, V.; Malaika, L. T. M.; Taudon, N.; Lwango, R.; Alegre, S. S.; Bertaux, L.; Sugnaux, F.; Parzy, D.; Benakis, A. Pharmacokinetics and Pharmacodynamics of a New ACT Formulation: Artesunate/Amodiaquine (TRIMALACT) Following Oral Administration in African Malaria Patients. *Eur. J. Drug Metab. Pharmacokinet.* **2009**, *34* (3−4), 133−142.

(21) Beconi, M. G.; Yates, D.; Lyons, K.; Matthews, K.; Clifton, S.; Mead, T.; Prime, M.; Winkler, D.; O'Connell, C.; Walter, D.; Toledo-Sherman, L.; Munoz-Sanjuan, I.; Dominguez, C. Metabolism and Pharmacokinetics of JM6 in Mice: JM6 Is Not a Prodrug for Ro-61−8048. *Drug. Metab. Dispos.* **2012**, *40* (12), 2297−2306.

(22) Kreth, K.-P.; Kovar, K.; Schwab, M.; Zanger, U. M. Identification of the Human Cytochromes P450 Involved in the Oxidative Metabolism of "Ecstasy"-Related Designer Drugs. *Biochem. Pharmacol.* **2000**, *59* (12), 1563−1571.

(23) Kala, S. V.; Harris, S. E.; Freijo, T. D.; Gerlich, S. Validation of Analysis of Amphetamines, Opiates, Phencyclidine, Cocaine, and Benzoylecgonine in Oral Fluids by Liquid Chromatography-Tandem Mass Spectrometry. *J. Anal. Toxicol.* **2008**, *32* (8), 605−611.

(24) Raes, E.; Verstraete, A. G. Usefulness of Roadside Urine Drug Screening in Drivers Suspected of Driving under the Influence of Drugs (DUID). *J. Anal. Toxicol.* **2005**, *29* (7), 632−636.

(25) Yokoo, K.; Hamada, A.; Tazoe, K.; Sasaki, Y.; Saito, H. Effects of Oral Administration of S-1 on the Pharmacokinetics of SN-38, Irinotecan Active Metabolite, in Patients with Advanced Colorectal Cancer. *Ther. Drug Monit.* **2009**, *31* (3), 400−403.

(26) Zhang, W.; Dutschman, G. E.; Li, X.; Ye, M.; Cheng, Y.-C. Quantitation of Irinotecan and Its Two Major Metabolites Using a Liquid Chromatography-Electrospray Ionization Tandem Mass Spectrometric. *J. Chromatogr. B* **2009**, *877* (27), 3038−3044.

(27) Baylatry, M.-T.; Joly, A.-C.; Pelage, J.-P.; Bengrine-Lefevre, L.; Prugnaud, J.-L.; Laurent, A.; Fernandez, C. Simple Liquid Chromatography Method for the Quantification of Irinotecan and SN38 in Sheep Plasma: Application to in Vivo Pharmacokinetics after Pulmonary Artery Chemoembolization Using Drug Eluting Beads. *J. Chromatogr. B* **2010**, *878* (9−10), 738−742.

(28) The Molecular Process Ontology. https://www.ebi.ac.uk/ols4/ontologies/mop. (access, October 20, 2023).

(29) Krallinger, M.; Rabal, O.; Leitner, F.; Vazquez, M.; Salgado, D.; Lu, Z.; Leaman, R.; Lu, Y.; Ji, D.; Lowe, D. M.; Sayle, R. A.; Batista-Navarro, R. T.; Rak, R.; Huber, T.; Rocktäschel, T.; Matos, S.; Campos, D.; Tang, B.; Xu, H.; Munkhdalai, T.; Ryu, K. H.; Ramanan, S. V.; Nathan, S.; Žitnik, S.; Bajec, M.; Weber, L.; Irmer, M.; Akhondi, S. A.; Kors, J. A.; Xu, S.; An, X.; Sikdar, U. K.; Ekbal, A.; Yoshioka, M.; Dieb, T. M.; Choi, M.; Verspoor, K.; Khabsa, M.; Giles, C. L.; Liu, H.; Ravikumar, K. E.; Lamurias, A.; Couto, F. M.; Dai, H.-J.; Tsai, R. T.-H.; Ata, C.; Can, T.; Usié, A.; Alves, R.; Segura-Bedmar, I.; Martínez, P.; Oyarzabal, J.; Valencia, A. The CHEMDNER Corpus of Chemicals and Drugs and Its Annotation Principles. *J. Cheminform.* **2015**, *7*, No. S2.

(30) Miranda, A.; Mehryary, F.; Luoma, J.et al. In *Overview of DrugProt BioCreative VII Track: Quality Evaluation and Large Scale Text Mining of Drug-Gene/Protein Relations*; Proceedings of the seventh BioCreative challenge evaluation workshop, 2021.

(31) ChEMBL database. https://www.ebi.ac.uk/chembl. (access March, 14, 2024).

(32) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13* (11), 2498−2504.

(33) Biziukova, N.; Tarasova, O.; Ivanov, S.; Poroikov, V. Automated Extraction of Information From Texts of Scientific Publications: Insights Into HIV Treatment Strategies. *Front. Genet.* **2020**, *11*, No. 618862.

(34) Tarasova, O.; Biziukova, N.; Shemshura, A.; Filimonov, D.; Kireev, D.; Pokrovskaya, A.; Poroikov, V. V. Identification of Molecular Mechanisms Involved in Viral Infection Progression Based on Text Mining: Case Study for HIV Infection. *Int. J. Mol. Sci.* **2023**, *24* (2), No. 1465.

(35) Murayama, N.; Imai, N.; Nakane, T.; Shimizu, M.; Yamazaki, H. Roles of CYP3A4 and CYP2C19 in Methyl Hydroxylated and N-Oxidized Metabolite Formation from Voriconazole, a New Anti-Fungal Agent, in Human Liver Microsomes. *Biochem. Pharmacol.* **2007**, *73* (12), 2020−2026.

(36) Anderson, L. W.; Collins, J. M.; Klecker, R. W.; Katki, A. G.; Parchment, R. E.; Boinpally, R. R.; LoRusso, P. M.; Ivy, S. P. Metabolic Profile of XK469 (2(R)-[4-(7-Chloro-2-Quinoxalinyl)Oxyphenoxy]-Propionic Acid; NSC698215) in Patients and in Vitro: Low Potential for Active or Toxic Metabolites or for Drug-Drug Interactions. *Cancer Chemother. Pharmacol.* **2005**, *56* (4), 351−357.

(37) Androutsopoulos, V. P.; Ruparelia, K.; Arroo, R. R. J.; Tsatsakis, A. M.; Spandidos, D. A. CYP1-Mediated Antiproliferative Activity of Dietary Flavonoids in MDA-MB-468 Breast Cancer Cells. *Toxicology* **2009**, *264* (3), 162−170.

(38) Yritia, M.; Riba, J.; Ortuño, J.; Ramirez, A.; Castillo, A.; Alfaro, Y.; de la Torre, R.; Barbanoj, M. J. Determination of N,N-Dimethyltryptamine and Beta-Carboline Alkaloids in Human Plasma Following Oral

Administration of Ayahuasca. *J. Chromatogr. B* **2002**, *779* (2), 271−281.

(39) Kraemer, T.; Theis, G. A.; Weber, A. A.; Maurer, H. H. Studies on the Metabolism and Toxicological Detection of the Amphetamine-like Anorectic Fenproporex in Human Urine by Gas Chromatography-Mass Spectrometry and Fluorescence Polarization Immunoassay. *J. Chromatogr. B* **2000**, *738* (1), 107−118.

(40) Wen, B.; Ma, L.; Rodrigues, A. D.; Zhu, M. Detection of Novel Reactive Metabolites of Trazodone: Evidence for CYP2D6-Mediated Bioactivation of m-Chlorophenylpiperazine. *Drug Metab. Dispos.* **2008**, *36* (5), 841−850.

(41) Dieterle, W.; Hengelage, T. Influence of Food Intake on the Pharmacokinetics of Avosentan in Man. *Int. J. Clin. Pharmacol. Ther.* **2008**, *46* (9), 453−458.

(42) Yeung, C. S.; Beck, T.; Posma, J. M. MetaboListem and TABoLiSTM: Two Deep Learning Algorithms for Metabolite Named Entity Recognition. *Metabolites* **2022**, *12* (4), No. 276.

(43) Kongburan, W.; Padungweang, P.; Krathu, W.; Chan, J. H. Enhancing Metabolic Event Extraction Performance with Multitask Learning Concept. *J. Biomed. Inform.* **2019**, *93*, No. 103156.

(44) Kim, J.-D.; Ohta, T.; Tateisi, Y.; Tsujii, J. GENIA Corpus—a Semantically Annotated Corpus for Bio-Textmining. *Bioinformatics* **2003**, *19*, i180−i182.

(45) Czarnecki, J. M.; Shepherd, A. J. Metabolic Pathway Mining. In Bioinformatics. In *Bioinformatics*; Keith, J. M., Ed.; Methods in Molecular Biology; Springer: New York, NY, 2017; Vol. *1526*, pp 139−158.