Lit-OTAR Framework for Extracting Biological Evidences from Literature

Santosh Tirunagari[1], Shyamasree Saha[1], Aravind Venkatesan[1], Daniel Suveges[2], Annalisa Buniello[2], David Ochoa[2], Johanna McEntyre[1], Ellen McDonagh[2], and Melissa Harrison[1,✉]

[1]Literature Services Team, European Bioinformatics Institute, European Molecular Biology Laboratory (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, CB10 1SD, Cambridge, United Kingdom
[2]Open Targets, European Bioinformatics Institute, European Molecular Biology Laboratory (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, CB10 1SD, Cambridge, United Kingdom

The lit-OTAR framework, developed through a collaboration between Europe PMC and Open Targets, leverages deep learning to revolutionise drug discovery by extracting evidence from scientific literature for drug target identification and validation. This novel framework combines Named Entity Recognition (NER) for identifying genes/proteins, diseases, organisms, and chemicals/drugs within scientific texts, and entity normalisation to map these entities to databases like Ensembl, Experimental Factor Ontology (EFO), and ChEMBL. Continuously operational, it has processed over 39 million abstracts and 4.5 million full-text articles and preprints to date, identifying more than 48.5 million unique associations that significantly help accelerate the drug discovery process and scientific research ($>$ 29.9m distinct target-disease, 11.8m distinct target-drug and 8.3m distinct disease-drug relationships). The results are made accessible through the Open Targets Platform (https://platform.opentargets.org/) as well as Europe PMC website (SciLite web app) and annotations API (https://europepmc.org/annotationsapi).

Europe PMC | Open Targets | Bioformer | Deep Learning
Correspondence: ***mharrison@ebi.ac.uk***

DRAFT

Introduction

The process of identifying drug targets is a critical aspect of drug discovery, requiring an understanding of the underlying mechanisms of diseases. Scientists rely on various sources of evidence to unravel the connections between drugs, targets and diseases, including gene expression changes, genetic variations, and clinical study data. Addressing this complexity, the Open Targets Platform (1) has been developed as a web-based tool specifically designed to integrate diverse sources of evidence to facilitate efficient identification of the most promising drug targets associated with diseases and phenotypes. The Platform combines and integrates data from more than 20 different sources to provide target–disease associations, covering evidence derived from genetic associations, somatic mutations, known drugs, differential expression, animal models, pathways and systems biology, and text mining of scientific articles. An integrated score weighs the evidence from each data source and type, which contributes to an overall score for each target–disease association. Its systematic approach involves integrating evidence of the association between drug targets and diseases, as well as providing annotation data for better understanding of targets, to harmonise this information into a coherent schema, and presenting it in a user-friendly and comprehensive manner. The Open Targets Platform can be accessed via the web browser[1], while it also provides an API, supports Google BigQuery integration, and allows users to download datasets for downstream analysis.

Extraction of assertions from scientific articles is an important aspect of this work, to this end Europe PMC (2) has played a key supporting role. As part of the PMC International Collaboration, Europe PMC[2] is a global, free, biomedical literature repository providing access to worldwide life sciences articles and preprints. It indexes an extensive collection of life science research, including over 41 million abstracts and 8.7 million full-text articles, with a growth rate of 1.7 million new articles annually. The integration of Europe PMC's text mining capabilities into the Open Targets Platform represents a strategic leveraging of literature as a unique and invaluable source of information. These text-mining techniques are particularly crucial at identifying and elucidating target–disease–drug associations within the literature, a process central to the Open Targets Platform's functionality.

The Lit-OTAR framework comprises two main components: Europe PMC text mining and the Open Targets literature component as shown in Figure 1. Europe PMC utilises deep learning techniques to identify target, disease, and drug entities within scientific documents, while Open Targets performs entity normalisation to accurately map these entities to databases like ENSEMBL[3], Experimental Factor Ontology (EFO)[4], and ChEMBL[5] and ranks the associations between target–disease–drug mentioned in these documents. The primary objective of this framework is to offer a scalable and continuous service to the scientific community for target validation.

The framework addresses a critical gap in the existing landscape of biomedical text mining. Previous efforts have mainly focused on extracting gene–disease associations from the literature (3, 4), with a few developing methods for integrated resources like DisGeNET (3), and have often fallen

[1]https://platform.opentargets.org
[2]https://www.ncbi.nlm.nih.gov/pmc/about/pmci/
[3]https://www.ensembl.org/
[4]https://www.ebi.ac.uk/efo/
[5]https://www.ebi.ac.uk/chembl/

short in terms of regular data updates and the scope of analysis, primarily focusing on abstracts.

In our 2017 study (4), we employed dictionary-based methods within a quarterly operational pipeline for data updates, utilising the Europe PMC text-mining pipeline enhanced with custom dictionaries from UniProt and EFO for annotating target and disease names. This approach, though robust, faced limitations due to its reliance on manual rules, such as abbreviation filters and blacklists of common terms, and the inherent challenges in biomedical texts like distinguishing between gene and protein names, spelling variations, and context-specific meanings of abbreviations, leading to high recall but low precision (5).

The emergence of modern natural language processing (NLP) techniques (6, 7) has revolutionised text mining by offering high efficiency and accuracy. Models like BERT (8), BioBERT (9), PubMedBERT (10), and BioFormer (11), trained on extensive biomedical corpora and fine-tuned for specific tasks, have markedly improved the accuracy of entity recognition, managing ambiguities, special characters, acronyms, and identifying synonyms and variations in expression. These advancements not only enhance recall and precision rates but also facilitate the discovery of new biological relationships from the extensive, unstructured data in the life science domain.

In our current study, we have used deep learning techniques, leveraging models such as BioBERT and BioFormer, to significantly enhance our pipeline (4). This updated Lit-OTAR pipeline, has been refined for improved flexibility and modularity, effectively broadening its scope to include a novel entity category for chemicals/drugs. This enhancement has enabled the pipeline to mine for associations between drugs and targets, drugs and diseases, in addition to targets and diseases. Furthermore, we have addressed technical challenges such as sentence splitting and boundary detection in complex document structures like tables and figures. The main distinctions between the previous and the current pipeline are detailed in the Supplementary Note 1: Table 4.

Materials and Methods

As of the time of writing this article, Europe PMC hosted approximately 39 million journal and preprint abstracts and 9 million full-text journal and preprint articles. However, only a subset of these, specifically 39 million and 4.5 million, respectively, were included due to licensing restrictions (CCO and CC-BY) and their classification as original research articles[6]. This dataset and the subsequent daily addition of the data is run through our custom developed deep learning model for NER extraction (12). The generated output is formatted in JSON, with identified entities treated as matches. Moreover, when two matches or entities occur within the same sentence, they are considered as forming an association or providing evidence. We have completed a study with three experts for treating co-occurrence as association (refer Section Co-occurrence vs Association C). Subsequently, this processed data is forwarded to the Open Targets ETL for the purpose of normalisation (grounding). Disease-related entities are mapped to the Experimental Factor Ontology (EFO), chemical and drug entities to CHEMBL, and gene and protein entities to Ensembl. The resulting data is made accessible through both the Open Targets Platform and Europe PMC annotations APIs, in addition to the Scilite annotations tool (13) on the Europe PMC website.

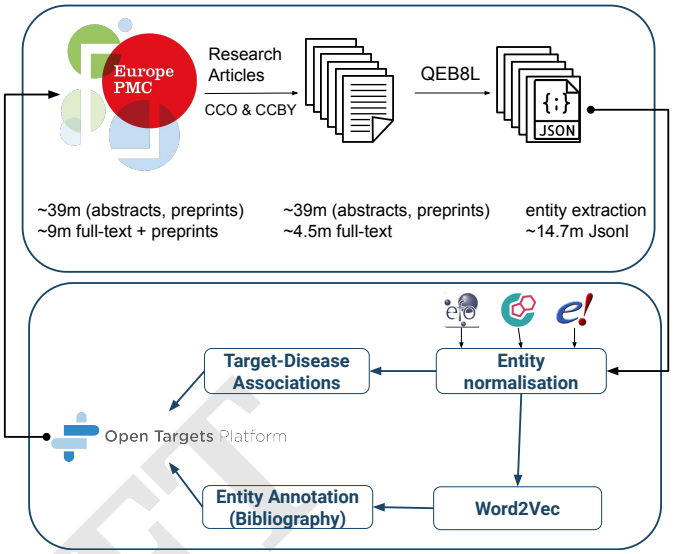

Fig. 1. Overview of the Data Selection, Processing, and Accessibility Workflow in Europe PMC and Open Targets Platforms.

A. Entity Recognition. To develop deep learning models for the Lit-OTAR framework, we utilised the Europe PMC dataset (5). Initially, this dataset did not include mentions of chemicals/drugs. To overcome this limitation, we used a publicly available chemical/drug corpus (14) to annotate the Europe PMC dataset with chemicals/drugs, preserving the human-annotated spans. The enriched dataset covers mentions of genes/proteins, diseases, chemcial/drugs, and organisms. We trained and evaluated three different models BioBERT, SpaCy, and Bioformer on this dataset, using the evaluation criteria from SemEval-2013 Task 9.1 (15) (Supplementary Note 2).

B. Entity Normalisation. The NER tagging of entities occurs at Europe PMC, while normalisation and ranking (Supplementary Note 3) take place on the Open Targets Platform (Figure 1).

The process involves matching and mapping entities to specific databases/ontologies. The pipeline uses a Word2Vec skip-gram model (16) to transform NER outputs into standardised representations. This includes mapping diseases to the EFO, chemicals to ChEMBL, and genes to Ensembl. The model generates n-dimensional vectors as word embeddings, capturing semantic similarities based on co-occurrence in the literature.

Normalisation ensures that different mentions of an entity across documents are consolidated into a single, canonical form, linked to the respective database identifier. This allows

[6] https://europepmc.org/Copyright

DRAFT

for consistent aggregation of all relevant information regarding an entity, facilitating the identification of connections between entities.

The model's calculation of similarity metrics also supports the ranking of entities, determining their relevance to the research context by analsing literature patterns. This method improves the accuracy of entity normalisation and enhances the Open Targets Platform's value for researchers by offering a comprehensive understanding of biological entity relationships and their potential therapeutic implications.

C. Co-occurrence vs Association. A curation task was conducted to annotate 252 sentences for sentiment and association, with each annotator pair assigned 168 sentences and an intentional overlap of 84 sentences between pairs to measure inter-annotator agreement. The annotation overlap was measured using Cohen's Kappa (K).

As illustrated in Figure 2, despite high expectations, the overall Cohen's Kappa value indicated an agreement below 60%, reflecting a low inter-annotator agreement that deemed the overlap unsuitable for machine learning purposes. High overlap was observed in sentiment analysis, suggesting a relative ease in identifying sentiments across annotators. The association identification presented challenges, evidenced by a low overlap. This difficulty was attributed to various factors, including short sentences lacking clear relations, long sentences with lists of multiple genes/proteins, drugs and diseases, complex sentence structures, and sentences that required additional context for accurate interpretation.

Fig. 2. Expert Annotation Score Discrepancies showing the differences in association and sentiment scores among three experts, highlighting variance in perceptions of associations [0.2 - 0.39] and sentiment analysis [0.67 - 0.7], underscoring the challenge of consistent annotation.

Following this study, we recognised that association is subjective, leading us to consider co-occurrence as a form of association itself. Consequently, we adjusted our approach to treat co-occurrence as the universal set, acknowledging that any co-occurrence might imply an association, despite the challenges in explicit identification by annotators.

Results

D. Entity Recognition. BioBERT led in precision among the models tested, achieving scores of 0.91 (CD), 0.90 (DS), 0.93 (OG), and 0.91 (GP), with similarly high recall and F1 scores, demonstrating its effectiveness in entity recognition across various categories. Given the computational demands of BioBERT, our focus shifted towards enhancing the Bioformer-8L model into the QEB8L model. By utilising ONNX for model optimisation, we significantly improved inference speeds without sacrificing performance. Further enhancements through static quantisation not only increased processing speed tenfold but also reduced the model size to approximately 77MB, all while maintaining impressive accuracy with precision scores ranging from 0.85 to 0.94 and F1 scores around 0.88 to 0.89, highlighting its balanced performance as shown in Table 1.

Table 1. The performance of different models (Dictionary, BioBERT, spaCy, QEB8L) across four categories (Chemical, Disease, Organism, Gene-Protein) in terms of Precision, Recall, and F1 Score metrics evaluated on the Gold Standard test set.

Category	Model	Precision	Recall	F1 Score
Chemical/Drug	Dictionary	0.53	0.34	0.41
	BioBERT	**0.91**	**0.92**	**0.92**
	spaCy	0.80	0.73	0.76
	QEB8L	0.85	0.90	0.88
Disease	Dictionary	0.48	0.74	0.58
	BioBERT	**0.90**	0.80	0.85
	spaCy	0.82	0.71	0.76
	QEB8L	**0.90**	**0.88**	**0.89**
Organism	Dictionary	0.68	0.90	0.78
	BioBERT	0.93	**0.86**	**0.90**
	spaCy	0.85	0.75	0.79
	QEB8L	**0.94**	0.85	0.89
Gene-Protein	Dictionary	0.48	0.74	0.58
	BioBERT	**0.91**	0.87	**0.89**
	spaCy	0.84	0.76	0.80
	QEB8L	0.90	**0.88**	**0.89**

SpaCy, recognised for its quick inference speed, presented slightly lower precision scores (0.80-0.84) and F1 scores (0.76-0.80) across categories, suggesting its practicality for production-level entity recognition tasks with its efficiency. Conversely, the Dictionary approach (our previous piepline), while serving as a baseline in the current study achieved lower precision scores (0.48-0.68) but higher recall in some instances, leading to moderate F1 scores.

Our analysis demonstrated a significant overlap between the gold standard and the QEB8L model, identifying additional entities by the QEB8L model not found in the gold standard. Entities identified by the QEB8L model and dictionaries, but absent in the gold standard, were classified as false positives. Our goal was to minimise these false positives while maximizing overlap as illustrated in Figure 3.

Moving to a deep learning approach, specifically the QEB8L model, was driven by the need to reduce false positives and improve entity coverage. The performance comparison using the gold standard test set revealed that the QEB8L model significantly outperformed our previous dictionary-based method, showcasing this advantage. The QEB8L model, trained and evaluated on the gold standard dataset,

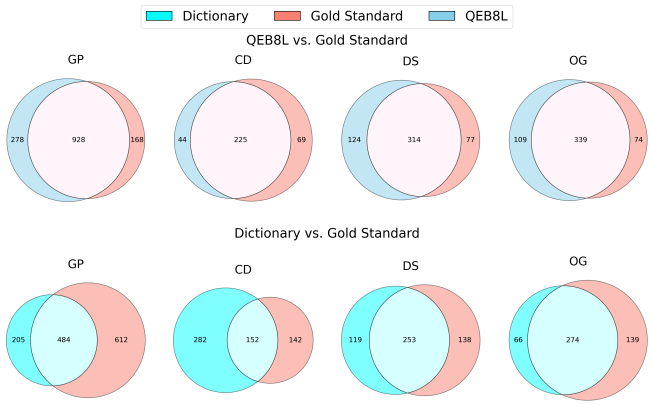

Fig. 3. Comparison of entity matches between the dictionary approach and deep learning approach (QEB8L) to the gold standard (5).

demonstrated the highest overlap with the gold standard, featuring fewer false positives and false negatives as shown in Figure 3. Some example entities are listed in Supplementary Note 4:Table 5 for the QEB8L model and Table 6 for the dictionary approach.

E. Entity Normalisation. A large proportion of the identified labels could be normalised, showing the effectiveness of our methodology in mapping biomedical entities to standardised terminologies; Disease to EFO, chemicals to ChEMBL, and Genes to Ensembl. This process is crucial for aggregating and analysing biomedical literature, facilitating the identification of relationships between diseases and potential therapeutic targets.

The dataset highlights (as shown in Table 2) the scale and complexity of biomedical terminology, comprising numerous entities mapped from scientific publications. Diseases and syndromes alone account for over 220 million entities, with approximately 76.6% successfully normalised to known entities from EFO. The unmapped entities, indicative of the diversity and complexity in biomedical literature, underscores challenges in achieving complete normalisation. The unmapped entities are available for further study[7].

The significant diversity among unnormalised entities necessitates continuous refinement of recognition and normalisation techniques. A literature-based evidence set curated by the Uniprot team was used to benchmark the performance, focusing particularly on disease-to-target associations. This set, comprising of 969 publications with $1,038$ disease–target associations, served as a foundation for evaluating the efficiency of the lit-OTAR pipeline.

The evaluation, as detailed in Table 3, shows high match rates for target identification, indicating the Lit-OTAR framework's potential in mapping biomedical entities to standardised terminologies. On the other hand, disease recognition and normalisation presented less robust results, highlighting areas for improvement due to the complexity and variability of disease nomenclature.

[7] https://ftp.ebi.ac.uk/pub/databases/opentargets/platform/latest/output/etl/json/literature/failedCooccurrences/

These findings present the strengths and challenges of current Lit-OTAR framework, emphasising the need for advancements in handling the diversity and complexity of disease terms. Enhancing the precision of literature-based evidence collection and analysis is critical for identifying valid disease–target associations, thereby facilitating therapeutic development.

Surprisingly, the Lit-OTAR framework also facilitated unexpected achievements, such as the discovery of new disease entities and synonyms ("T2D"), and enhanced data processing capabilities, in integrating with databases like EFO and improving analyses with the FDA's Adverse Events Reporting System (FAERS). The details are presented in Supplementary Note 5.

Conclusions

The Lit-OTAR framework, collaboration between Europe PMC and Open Targets, leverages biomedical literature for drug discovery. By applying named entity recognition and entity normalisation, this framework has processed more than 39 million abstracts and 4.5 million full-text articles, identifying around 48.5 million unique associations among target–disease, target–drug, and disease–drug interactions. This study provides insights into the drug discovery process and expands scientific research. In addition, the framework demonstrates the capability to discover new entities and enrich databases and ontologies with previously unrecognised associations. The Lit-OTAR pipeline operates daily, with updates provided quarterly on both the Europe PMC and Open Targets Platforms, ensuring timely access to relevant data for researchers and supporting therapeutic research and development.

Data availability

Access the latest data version via FTP[8], GraphQL API[9], and Google BigQuery[10]. Previous versions and bulk downloads are available at Open Targets[11] and Europe PMC[12]. Further details on the Platforms are presented in the Supplementary Note 6.

Code availability

The computational frameworks and models[13] supporting this study are distributed across several repositories, maintained by Europe PMC and Open Targets, to ensure broad accessibility and facilitate collaboration.

[8] https://ftp.ebi.ac.uk/pub/databases/opentargets/platform/latest/output/etl/json/literature/

[9] https://api.platform.opentargets.org/api/v4/graphql/browser

[10] https://platform-docs.opentargets.org/data-access/google-bigquery

[11] https://ftp.ebi.ac.uk/pub/databases/opentargets/platform/

[12] https://ftp.ebi.ac.uk/pub/databases/opentargets/platform/latest/output/etl/json/epmcCooccurrences/

[13] https://www.ebi.ac.uk/biomodels/MODEL2306290002

DRAFT

Table 2. Summary of Entity Recognition and Normalisation Outcomes Across Disease/Syndrome, Chemical/Drug, and Gene/Protein Categories. Quantifies the total entities identified, the success rate of mapping these entities to standardised databases, and the diversity of unique labels encountered and successfully normalised.

Entity type	*Entity count*	*Mapped entity count*	*Unique labels*	*Unique mapped labels*	*Unique mapped entity*
Disease/ Syndrome	220,392,937	168,818,017 (76.6%)	2,196,439	166,497 (7.6%)	11,561
Chemical/Drug	122,872,756	77,826,420 (63.3%)	2,213,483	76,194 (3.4%)	10,370
Gene/Protein	347,835,641	197,124,445 (56.7%)	7,063,573	680,368 (9.6%)	28,778

Table 3. Benchmarking Entity Recognition and Normalisation Performance Using a Literature-Based Evidence Set.

	Uniprot curated evidence	**Normalised matches**
Publications	969	967 (99.8%)
Publication/target pairs	1,088	1,034 (95.0%)
Publication/disease pair	1,515	1,034 (68.3%)
Publication/disease/target triplet	1,580	748 (47.3%)
Disease/target pair	1,038	550 (53.0%)

- The QEB8L model for entity recognition is at https://github.com/ML4LitS/annotation_models.
- The Open Targets daily pipeline, under Europe PMC, is at https://gitlab.ebi.ac.uk/saha/opentargetsorchestrator and https://github.com/ML4LitS/otar-maintenance.
- Open Targets' ETL processes available at https://github.com/opentargets.

Author contributions statement

All authors played significant roles in the lit-OTAR framework's development and execution. S.T. led manuscript writing, developed and evaluated the NER models, productionalised the framework, and led experiments and analysis. S.S. set up the daily pipeline, contributed to experiments and analysis. D.S. focused on normalisation evaluation, and testing the framework for Open Targets. A.V., and A.B. assisted in manuscript writing, proofreading, and figures generation. A.V., M.H., and D.O. were key in integrating the framework within Europe PMC, with contributions to manuscript structuring and project ideation. E.M. and J.M. provided conceptual insights, ensuring a comprehensive approach.

ACKNOWLEDGEMENTS
We thank the curators at Molecular Connections for their biocuration efforts. We would also extend our thanks to Zunaira Shafique and all the development team from Europe PMC and Open Targets Platform for their contributions to the Europe PMC Platform and Open Targets Platform. This work was supported by: the European Molecular Biology Laboratory-European Bioinformatics Institute (S.T, A.V, MH); and the OpenTargets grant 2056 (S.T, S.S).

Competing interests

No competing interest is declared.

Bibliography

1. David Ochoa, Andrew Hercules, Miguel Carmona, Daniel Suveges, Jarrod Baker, Cinzia Malangone, Irene Lopez, Alfredo Miranda, Carlos Cruz-Castillo, Luca Fumis, et al. The next-generation open targets platform: reimagined, redesigned, rebuilt. *Nucleic acids research*, 51(D1):D1353–D1359, 2023.
2. Summer Rosonovski, Maria Levchenko, Rajat Bhatnagar, Umamageswari Chandrasekaran, Lynne Faulk, Islam Hassan, Matt Jeffryes, Syed Irtaza Mubashar, Maaly Nassar, Madhumiethaa Jayaprabha Palanisamy, et al. Europe pmc in 2023. *Nucleic Acids Research*, 52(D1): D1668–D1676, 2024.
3. Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855, 11 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz1021.
4. Şenay Kafkas, Ian Dunham, and Johanna McEntyre. Literature evidence in open targets-a target validation platform. *Journal of Biomedical Semantics*, 8:1–9, 2017.
5. Xiao Yang, Shyamasree Saha, Aravind Venkatesan, Santosh Tirunagari, Vid Vartak, and Johanna McEntyre. Europe pmc annotated full-text corpus for gene/proteins, diseases and organisms. *bioRxiv*, pages 2023–02, 2023.
6. Thanh Hai Dang, Hoang-Quynh Le, Trang M Nguyen, and Sinh T Vu. D3NER: biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information. *Bioinformatics*, 34(20):3539–3546, 04 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty356.
7. Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70, 2020.
8. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
9. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
10. Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.
11. Li Fang, Qingyu Chen, Chih-Hsuan Wei, Zhiyong Lu, and Kai Wang. Bioformer: an efficient transformer language model for biomedical text mining. *arXiv preprint arXiv:2302.01588*, 2023.
12. S. Tirunagari and M. Harisson. (2023). *Accelerating Biomedical Named Entity Recognition with Quantised EPMCA Bioformer-8L (QEB8L) Model (Version 0. 0*, 0., 2023.
13. Aravind Venkatesan, Jee-Hyub Kim, Francesco Talo, Michele Ide-Smith, Julien Gobeill, Jacob Carter, Riza Batista-Navarro, Sophia Ananiadou, Patrick Ruch, and Johanna McEntyre. Scilite: a platform for displaying text-mined annotations as a means to link research articles with biological data. *Wellcome open research*, 1, 2016.
14. Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):1–17, 2015.
15. Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. Lessons learnt from the ddiextraction-2013 shared task. *Journal of biomedical informatics*, 51:152–164, 2014.
16. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

DRAFT

Supplementary Note 1: Old pipeline vs new Lit-OTAR pipeline

Table 4. Comparative Analysis of the Old Pipeline (4) and the Current Pipeline Across Various Aspects

Aspect	Old Pipeline	New Lit-OTAR Pipeline
Data Source	Europe PMC (PubMed and PubMed Central) CCO & CC-BY Original Research articles	Europe PMC (PubMed and PubMed Central, and Preprints). CCO & CC-BY Original Research articles
Data Size	26 million abstracts, 1.2 million full-text articles	39 million journal article and preprint abstracts and 4.5 million full-text articles and counting
Approach	Dictionary-based	Deep learning (Bioformer-8L)
Types	Genes/protein and Disease	Genes/protein, Disease, Organisms, and Chemical/Drug
Accuracy	High recall but low precision	Improved precision and recall
Evidences	Gene–Disease	Gene–Disease Gene–Drug Disease–Drug
Article Scoring	Confidence scores based on location	Confidence scores based on location (similar to old pipeline (4))
Operational	Quarterly (terminated on 04/2021)	Daily since 04/2021
Benchmarking	None	Benchmarking of NER methods
Notes	Manual rules, abbreviation filter with heuristic rules, limited completeness, false positives	Improved accuracy, normalisation, broader scope, improved sentence splitter, reduced limitations
Performance	Precision: 0.54 Recall: 0.67 F-score: 0.58	Precision: 0.90 Recall: 0.88 F-score: 0.89

DRAFT

Supplementary Note 2: Evaluation Criteria

The NER was evaluated using the evaluation criteria used by SemEval-2013 Task 9.1 (15), which allows assessment of the system's performance based on four levels of strictness: strict, exact, partial, and type. These levels consider the match of entity boundaries and types. Using these metrics to evaluate the performance of the NER go beyond simple strict classification and take into account partial matching. To compare the differences between the output of the NER system and the correct annotations, two factors were considered: the exact string and the type of the entity. However, because there can be overlapping entities from different categories and data formats, each system per category was evaluated. This means that in certain cases, the counts in the "Strict" and "Exact" cells become equal. Similarly, this applies to the values in the cells that correspond to partial matching and incorrect matching. After evaluating the NER system using the metrics discussed above, the precision, recall, and F1 score was calculated for benchmarking.

Supplementary Note 3: Article Scoring and Ranking

The article scoring method in the lit-OTAR framework follows the approach detailed in (4). The algorithm scores scientific articles on their relevance to target–disease associations, helping to rank/prioritise articles by their relevance. The algorithm uses a weighting system that assigns different values to article sections, from full-text articles to abstracts, based on their ability to highlight key entities. For instance, the "Title" section gets the highest weight as it summarises the study's findings, while the "Introduction" is weighted least, given its focus is on known information. In abstracts, weight is given based on sentence position, with the analysis of 360 MEDLINE abstracts (4) showing that the last sentence, usually detailing results, is considered most significant.

Supplementary Note 4: Examples: False Positives and False Negatives

Supplementary Note 5: Other Achievements

The additional entities not found in dictionaries or the gold standard test set (Figure 3), which have been discovered through context learning in deep learning, further facilitate the addition of new entities to databases/ontologies. In one such scenario, the framework has aided in identifying diseases previously unlinked to any specific disease entity within the EFO [14]. Through

[14] https://github.com/opentargets/issues/issues/1555?ref=blog.opentargets.org?ref=blog.opentargets.org

Table 5. Example Entities Found through QEB8L but Missed in Gold Standard vs. Entities Found in Gold Standard but Missed through QEB8L.

Entity Type	Found in QEB8L but missed in Gold Standard	Found in Gold Standard but missed in QEB8L
Gene-Protein (GP)	['CPZ', 'Flp', 'APP180', 'hALK', 'CT4']	['YALI0D20108g', 'YALI0E32901g', 'LDH4', 'YALI0A9470g', 'MSC1']
Chemical-Disease (CD)	['No916601429', 'phycoerythrin', 'thymidine', 'butyrate', 'crystal']	['YALIOB19470', '11192732', 'pentose phosphate', 'QAPP67', 'CS36962']
Disease (DS)	['ischemia', 'CUMS', 'CIS', 'facial angiofibromas', 'Rhizoma']	['aneurysms', 'hallucinations', 'Postherpetic Neuralgia', 'lymphadenopathy', 'pertussis']
Organism (OG)	['Platyhelminthes', 'protozoans', 'merozoites', 'SZ', 'kids']	['Murine', 'Methanobrevibactr', 'wisent', 'proviruses', 'hermaphrodite']

Table 6. Example Entities Found through Dictionary Approach but Missed in Gold Standard vs. Entities Found in Gold Standard but Missed Through Dictionary Approach.

Entity Type	Found in Dictionary approach but missed in Gold Standard	Found in Gold Standard but missed in Dictionary approach
Gene-Protein (GP)	['nodal', 'Calc', 'MPI', 'LPS', 'NHLT']	['mTau', 'At1g61795', 'eIF2', 'proton / Pi symporters', 'collagen type IV']
Chemical-Disease (CD)	['3At', 'silver', 'sec', 'Peptide', 'Lipopolysaccharide']	['nucleotide', 'YALIOB19470', 'carboxylates1617', 'serine', '11192732']
Disease (DS)	['Trauma', 'ischemia', 'facial angiofibromas', 'bluetongue', 'hermaphrodite']	['CHD', 'SZ', 'RR - MS', 'B - NHL', 'memory deficits']
Organism (OG)	['Euglenozoa', 'Platyhelminthes', 'cotton', 'Białowieża', 'Gibbon']	['Gram - positive cocci', 'bulls', 'Euglenezoa', 'Methanobrevibactr', 'rodent']

a preliminary analysis of frequently occurring non-grounded labels, we identified new synonyms for existing diseases, notably recognising "T2D" as a synonym for Type II Diabetes Mellitus ($EFO_0001360$). This discovery alone added 281,184 matched labels across $29,040$ unique PubMed identifiers (PMIDs), significantly enriching the dataset and enhancing the accuracy of disease-related data mapping. These advancements underscore the Lit-OTAR framework's ability to refine and augment biomedical databases and analytical processes, providing valuable insights into drug safety and efficacy.
In another scenario, it enhanced the processing of data from the FDA's Adverse Events Reporting System (FAERS). This system, which compiles reports of adverse events and medication errors submitted to the FDA, presents unique challenges, such as distinguishing between a drug's adverse events and its indications. To address this, an increase in EFO cross-references to MedDRA was necessary in order to find out whether excluding reports where the adverse event matches the drug indication could improve the analytical outcomes' power and effectiveness. The normalisation pipeline developed for the Lit-OTAR, was able to map a significant portion of MedDRA[15] labels associated with adverse reactions to their corresponding EFO terms. This effort resulted in a cross-reference list containing approximately 10,000 mappings, which are assessed to be highly reliable.

Supplementary Note 6: Data Platforms

The datasets generated are made available through both Open Targets Platform and Europe PMC. While Open Targets Platform provides a web interface for data exploration and as bulk download, in Europe PMC the datasets are accessible both via the Annotations API and the website.

A. Europe PMC Annotations API. The Europe PMC Annotations API[16] is one of the main methods of accessing texted mined outputs (also called annotations) hosted by Europe PMC. Derived from both abstracts and open access full-text articles, these annotations are an invaluable resource for researchers needing programmatic access to the vast repository. One of the motivations of this is to make text-mined annotations available to the larger scientific community. To this end, the annotations are modelled based on the W3C Web Annotation Data Model[17]. This has ensured the annotations are standardised and, most importantly, FAIRified for wider consumption. The annotations as made available under the Apache License Version 2.0[18].
The API's RESTful architecture offers a modular structure, facilitating various functionalities essential for fetching specific annotations based on article IDs, entities, providers, relationships, or article sections. This flexibility is crucial for researchers

[15] https://www.meddra.org/?ref=blog.opentargets.org
[16] www.europepmc.org/AnnotationsApi
[17] https://www.w3.org/TR/annotation-model/
[18] https://www.apache.org/licenses/LICENSE-2.0

provider	OpenTargets	**Provider of the annotations that the user is interested in.**	query	string
filter	1 (default)	If the parameter is equal to 1, for each article only annotations of the specific provider will be retrieved. If the parameter is equal to 0, all the annotations will be retrieved for articles which also contain annotations of the specific provider. For example, if you search for annotations of the provider 'Europe PMC', you would get an overview of all annotations for each article, together with the annotations of the provider 'Europe PMC'	query	integer
format	JSON (default)	Output format of the response: • JSON will produce a JSON representation of the articles and relative annotations • XML will produce a XML representation of the articles and relative annotations • JSON-LD will produce a JSON linked Data representation of the annotations. To see details about JSON-LD go to http://europepmc.org/AnnotationsApi#jsonLD • ID_LIST will produce a list of articles identifiers including pmcid if available	query	string
cursorMark		CursorMark for pagination of the result list. For the first request you can omit the parameter or use the default value 0.0. For every following page use the value of the returned nextCursorMark element	query	double
pageSize	4	Number of articles the user wishes to retrieve in each page. The value must be between 1 and 8	query	integer

Fig. 4. Accessing annotations by OpenTargets on the Europe PMC annotations API

aiming to extract detailed and targeted information from the literature. The functionality varies from fetching annotations by article, entity name and provider (e.g. OpenTargets) to relationships such as gene-disease relationships (see Figure 4).
The API delivers results in various formats, including JSON, XML, and ID_LIST (for article identifiers), catering to different user preferences and requirements. Moreover, the annotations are available in JSON-LD format, providing a graph representation that enhances data interoperability and integration.
This extensive accessibility to annotated biomedical literature through the Europe PMC Annotations API significantly empowers researchers, facilitating the extraction and analysis of rich datasets for advancing scientific discoveries.

B. Visualisations on Europe PMC Website. SciLite (13) is an application developed as part of Europe PMC to display annotations hosted by the Platform directly on research articles and, those annotated entities are linked to the corresponding database or ontology resources, so the website user can get to the underlying data in a straightforward way. This application is embedded in the website which makes API requests using the Annotations API, to fetch all relevant annotations for a given article (see Figure 6. A detailed description of the design behind SciLite can be found here[19]

[19] https://europepmc.github.io/techblog/algorithm/2018/07/04/locating-text-html-pages.html

Curl

```
curl -X GET --header 'Accept: application/json' 'https://www.ebi.ac.uk/europepmc/annotations_api/annotationsByAr
```

Request URL

```
https://www.ebi.ac.uk/europepmc/annotations_api/annotationsByArticleIds?articleIds=MED%3A20877624&type=Gene%20Di
```

Response Body


```
[
  {
    "id": "http://europepmc.org/article/PMC/PMC2943476#opentargets-2915b48a600716541193101b4d
    "type": "Gene Disease Relationship",
    "creator": "Open Targets Platform",
    "body": {
      "source": "http://europepmc.org/article/PMC/PMC2943476",
      "isPartOf": "Discussion",
      "type": "TextualBody",
      "value": "HIV infection studies showed p-values above the p GT correction threshold; ho
    },
    "target": {
      "type": "List",
      "items": [
        {
          "id": "http://purl.uniprot.org/uniprot/P09417",
          "label": "PR) "
```


Fig. 5. The screenshot of the Open Target annotation (in JSON-LD format) retrieved from the Europe PMC Annotations API for the article with PMID 20877624. The annotation URI for the Open Target annotation is highlighted.

C. Making annotations FAIR and Citable. One of the main aims of establishing Europe PMC Annotation API is to make text mined annotations FAIR, where users are able cite the text mined entities and relationships. This allows the consumer of the content to understand and trace the source of information. The Web Annotation Model specification allows annotations to be uniquely identified using URIs, offering a mechanism to cite annotations across multiple Platforms. To this end, for all Open Targets annotations we have minted resolvable annotation URIs. For instance, for a given article ID the corresponding Open Target annotation can be retrieved in JSON-LD format, that will contain the annotation URI (see Figure 5).

D. Results in the Open Targets Platform. The Europe PMC dataset is utilised in two ways: The first is to use the dataset to extract evidence for the association of targets and diseases. Co-occurrences of target and disease entities are considered evidence for the association of those entities. In detail, when a target and a disease are mentioned in the same sentence within a publication, this constitutes one piece of Europe PMC evidence for the association of that target and that disease [Figure 7 (a)]. The second involves using the dataset to provide context to the Platform entities. Users can browse the available literature for the entity of their choice through the Bibliography widget, for example all the papers linked to cystic fibrosis [Figure 7(b)]. For more details refer article (1).

Mitochondrial genes and previously published studies

We further examined NEMPs that were previously reported as cellular gene products required for HIV-infection in screens using siRNAs [16], [17], [18], mRNA expression [21], or proteomics [19], [20] for SNPs associated with AIDS-1987 (Table S4). In our analysis of progression to AIDS, no SNPs within the 151 NEMP genes that were identified by the HIV infection studies [16], [17], [18] showed *p*-values above the p_{GT} correction threshold; however, fifty-nine genetic associations from twenty genes produce unadjusted $p \leq 0.01$ with the lowest *p*-value (0.0009) found in the gene for quinoid dihydropteridine reductase (QDPR) (rs2535228) for time to AIDS-1987 (HR = 0.7); six other SNPs in this region showed *p*-values from 0.004–0.01 (Table S5). SNPs within three of the gene fifteen genes replicated in two or more studies were associated with accelerated progression to AIDS-1987 in the current study: *NADH Dehydrogenase (Ubiquinone) 1 Beta Subcomplex, 7* (*NDUFB7*), *Isocitrate Dehydrogenase 1* (*IDH1*) , and *Isocitrate Dehydrogenase 3 (NAD+) Alpha* (*IDH3A*)] (NDUFB7 rs6511939 HR = 1.6, *p* = 0.008; IDH1 rs7580715 HR = 2.1, *p* = 0.009, IDH3A rs11855354, rs8032618 and rs12903696 HR = 1.6, p = 0.007–0.009).

Fig. 6. SciLite annotation tool highlighting the gene-disease association sentence using the LinkBack call. The LinkBack feature is based on the LinkBack API (which accepts the unique 'code' in an annotation ID) and text-annotator.

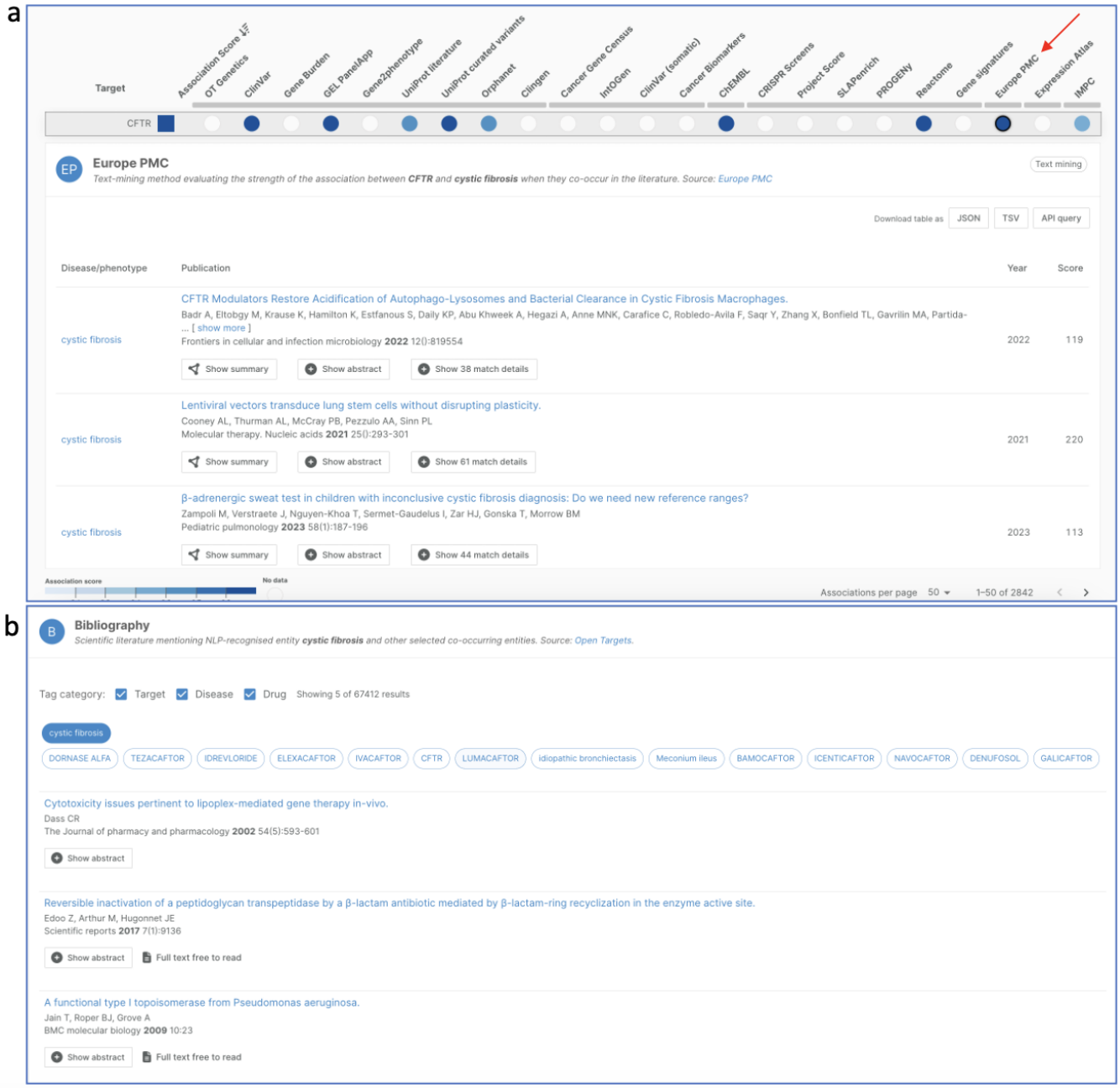

Fig. 7. Summary of how the lit-OTAR results are utilised and visualised in the Open Targets Platform. a. Europe PMC (red arrow) as data source for evidence of target–disease associations; b. Bibliography widget from a disease profile profile page.