Navigating Partial UMLS Terminology: GAT Embeddings and Confidence Analysis for Multilingual Concept Linking

Notebook for the BioNNE-L Lab at CLEF 2025

Albina Burlova[1,*]

[1]National Research University Higher School of Economics, 11 Pokrovsky Boulevard, Moscow, 109028, Russian Federation

Abstract

A lightweight pipeline is presented for biomedical concept normalisation that placed 1st in the Russian track and 2nd in the bilingual track of the BioNNE-L 2025 shared task. The method combines language-aware preprocessing with multilingual GAT-based embeddings and cosine-similarity retrieval over a 4M-entry bilingual UMLS vocabulary. Without any task-specific fine-tuning, the system reaches Accuracy@1 0.72, Accuracy@5 0.83, MRR 0.76 on the hidden Russian test set and 0.68 / 0.84 / 0.75 respectively in the bilingual setting. Beyond performance, an uncertainty analysis shows that high softmax entropy reliably predicts errors under extreme partial terminology, highlighting the need for confidence-aware re-ranking and the enrichment of Russian biomedical lexicons.

Keywords

medical concept normalisation, multilingual entity linking, Russian biomedical NLP, uncertainty estimation

1. Introduction

Medical concept normalisation (MCN) - also known as *entity linking* - maps free-text mentions of biomedical entities to canonical concepts in resources such as the Unified Medical Language System (UMLS). Accurate MCN underpins evidence retrieval, pharmacovigilance, and clinical decision support, yet it remains challenging in multilingual, low-resource settings, where many concepts lack standardised lexical variants. The BioNNE-L 2025 shared task, organised within the CLEF BioASQ lab [1], targets this gap with three subtasks: English (Track 2), Russian (Track 1), and a combined bilingual setting (Track 3) [2].

The Russian subtask requires linking mentions of diseases (DISO), chemicals (CHEM), and anatomical structures (ANATOMY) to UMLS Concept Unique Identifiers (CUIs). This track presents unique challenges due to the partial Russian coverage of UMLS - with only approximately 2% of concept names available in Russian [3] - which forces systems to resolve Cyrillic mentions against English entries. Furthermore, the task includes nested entities, meaning that mentions can overlap and require joint reasoning over inner and outer spans.

Recent work has shown that graph-augmented multilingual encoders such as BERGAMOT-GAT achieve state-of-the-art performance in zero-shot biomedical concept linking across ten languages [4]. Motivated by these findings, we build a lightweight pipeline that keeps BERGAMOT-GAT frozen and integrates language-aware preprocessing, a pre-encoded bilingual UMLS vocabulary of approximately 4 million terms, and a hybrid inference strategy that prioritises exact-match lookup, leverages cached predictions, and falls back on cosine similarity search within semantic-type partitions.

Even without task-specific fine-tuning, the system ranked first in the Russian track and second in the bilingual track. In addition to strong leaderboard performance, we explore how partial terminology coverage affects predictive uncertainty. Our analysis shows that missing Russian synonyms inflate predictive entropy and error rates, whereas cosine distance offers little additional signal under full partial-terminology.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

burlova613@gmail.com (A. Burlova)

0009-0004-2662-4070 (A. Burlova)

All code and reproducible notebooks are publicly available.[1]

2. Related Work

Early approaches to biomedical concept normalisation relied on dictionary look-ups coupled with fuzzy matching and heuristic ranking rules, which proved effective in English but deteriorated sharply in cross-lingual scenarios because they presupposed the existence of target-language synonyms in the ontology [5]. The introduction of multilingual transformers shifted the field toward contextual embeddings: multilingual BERT enabled the first zero-shot pipelines, yet its alignment for specialised terms remained limited [6]. SapBERT improved cross-lingual transfer by training with synonym-contrastive objectives and achieved large gains on XL-BEL and COMETA [7, 3].

Further improvements were achieved through self-alignment objectives [8] and domain-specialised pretraining for biomedical linking tasks [9]. These models laid the foundation for modern cross-lingual medical entity linking, but they still struggle in settings with severely partial terminology.

A complementary research line enriches textual embeddings with ontology structure. Graph-augmented encoders such as BERGAMOT-GAT incorporate UMLS neighbourhoods via graph attention and outperform both SapBERT and the string-matching baseline adopted in the first BioNNE task [4, 2]. Because BERGAMOT-GAT generalises well across ten languages without fine-tuning, it is adopted unchanged in our system. A related approach, CODER, combines knowledge-infused term embeddings with contrastive objectives to improve cross-lingual normalisation, demonstrating particular gains in low-resource settings [10].

In parallel, generative models for entity linking have gained traction. Autoregressive retrievers generate entity identifiers token-by-token using language models trained to mimic KB lookups [11], while biomedical variants leverage synonym-aware objectives and KB-guided pretraining to improve robustness in sparse domains [12]. Though effective, such methods typically require fine-tuning and careful control over generation constraints.

For Russian biomedical NLP, early studies focused on terminology expansion and bilingual projection, while RuCCoN introduced the first large clinical normalisation corpus and showed that SapBERT markedly outperforms rule-based methods [13]. The BioNNE-L shared-task series extends this line of work by adding nested mentions and a bilingual track, highlighting the challenges posed by partial Russian terminology [2].

Finally, although accuracy has dominated prior evaluations, the reliability and confidence calibration of multilingual normalisers remains underexplored. Sevgili et al. [14] highlight the limited attention given to predictive uncertainty, particularly in multilingual or low-resource settings. Recent work on uncertainty estimation has focused on other structured-prediction tasks - for instance, Somov and Tutubalina quantify how entropy-based confidence scores can detect errors in Text-to-SQL generation [15]. To the best of our knowledge, no study has yet investigated how the absence of target-language synonyms in UMLS modulates predictive uncertainty in biomedical concept normalisation. Our analysis in present work aims to fill this gap.

3. Task and Data

The BioNNE-L 2025 shared task targets biomedical concept normalisation (MCN) under three settings: Russian-only (Track 1), English-only (Track 2), and a combined bilingual scenario (Track 3) [16]. The annotated corpora are built on top of the NEREL-BIO dataset of Russian and English biomedical abstracts with nested entities [3, 17].

In all tracks, systems are required to map free-text entity mentions to UMLS Concept Unique Identifiers (CUIs). Mentions may be nested and overlapping, and are evaluated using Accuracy@1, Accuracy@5, and Mean Reciprocal Rank (MRR). In this work, we focus primarily on the Russian and bilingual tracks,

[1] https://github.com/AlbinaBurlova/bionn-gat-uncertainty

where partial terminology coverage and missing Russian synonyms pose significant challenges for normalization.

The organisers released dedicated training and development sets, along with a hidden test set for final evaluation. Although documents do not repeat across splits, entity mentions often recur verbatim, motivating optimisations based on mention-level caching. Statistics for the training and development data are shown in Table 1, while Table 2 summarises the test subsets.

Table 1
Training and development data. Unique counts are lowercased.

Subset	Mentions	Unique Strings
Train (EN+RU)	26,945	9,817
Dev RU	2,334	1,684
Dev EN	2,494	1,512
Dev Bilingual	4,828	3,026

Table 2
Test set sizes and unique mentions.

Subset	Mentions	Unique Strings
Test RU	6,215	3,326
Test EN	6,661	2,827
Test Bilingual	12,876	6,040

To construct the linking vocabulary, we combined the bilingual UMLS lexicon distributed by the organisers with all text–CUI pairs found in train and dev sets, including nested spans. Mentions were normalised by lowercasing and lemmatising Russian terms using `pymorphy3` [2], while English terms remained unchanged, as lemmatisation reduced accuracy.After filtering duplicates, we obtained a search space of 4.0 million rows, spanning over 1.5 million unique CUIs.

A notable fraction of mentions across both training and development sets lacked associated UMLS identifiers and were marked as CUILESS. These cases were excluded from all metric calculations and from the vocabulary used in retrieval. Table 3 reports their distribution.

Table 3
Frequency of CUILESS mentions.

Subset	Total Mentions	CUILESS Count	Share
Train (EN+RU)	26,945	2,520	9.35%
Dev RU	2,334	143	6.13%
Dev EN	2,494	99	3.97%
Dev Bilingual	4,828	242	5.01%

Although these mentions were ignored during evaluation, their high prevalence prompted an exploration of dedicated CUI-less classifiers. However, none of the tested approaches outperformed random-choice baselines, and were thus omitted from the final pipeline.

Since identical strings appear repeatedly, we implemented memoisation at inference: each unique mention is embedded once, and cached predictions are reused across documents. This optimisation reduces runtime by around 30% without affecting output quality.

[2] `pymorphy3` library: https://github.com/no-plagiarism/pymorphy3

4. Method

The system follows a modular inference pipeline based on frozen multilingual encoders and a precomputed vocabulary index. We prioritise generalisability and inference speed, avoiding any task-specific fine-tuning. The overall architecture consists of five stages: language-aware preprocessing, encoder-based embedding computation, type-partitioned vocabulary indexing, hybrid retrieval, and prediction postprocessing. Figure 1 provides a high-level overview of the system.

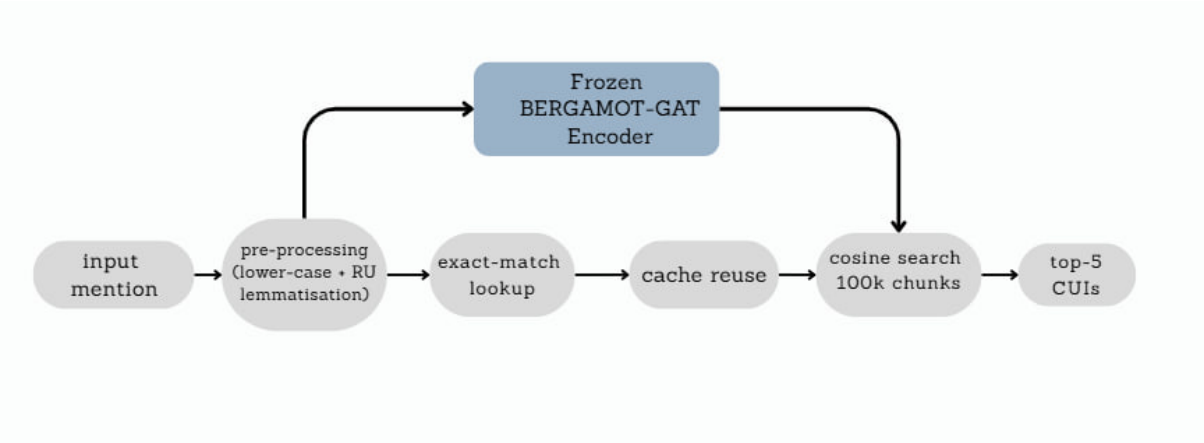

Figure 1: Overview of the concept normalisation pipeline.

4.1. Preprocessing and Language Detection

Each entity mention is lowercased and assigned a language tag via a heuristic based on Unicode script ranges. Russian strings are further lemmatised to reduce morphological sparsity. Stemming was empirically found to degrade Accuracy@1 to English mentions on the English track from 0.640 to 0.523, and thus these steps are omitted for Latin-script entries. The normalised form also serves as a key for downstream caching and matching steps.

4.2. Encoder: Frozen BERGAMOT-GAT

The publicly available model `andorei/BERGAMOT-multilingual-GAT` is used, which combines multilingual BERT with biomedical graph structure via a Graph Attention Network over the UMLS ontology [4]. The encoder is trained with multi-objective contrastive and classification losses and produces a 768-dimensional [CLS] embedding per mention. Model weights remain frozen throughout our pipeline to ensure robust cross-lingual generalisation and eliminate fine-tuning overhead.

For the English-only evaluation track (Track 1), which prohibits multilingual encoders, we used a separate model `andorei/gebert_eng_gat` pre-trained on English biomedical data. Despite tuning parameters (`max_length = 64`, `batch_size = 512`), this model achieved third place on the leaderboard with Accuracy@1 $= 0.64$, highlighting the difficulty of competing against fine-tuned English baselines. Nonetheless, the primary focus remained on the bilingual and Russian tracks.

4.3. Vocabulary Embeddings and Type Indexing

The concept vocabulary combines all unique mention–CUI pairs from the train and dev sets with the official bilingual UMLS lexicon. After deduplication and lemmatisation of Russian entries, we obtain approximately 4 million entries covering over 1.5 million unique CUIs. Each entry is embedded once using the frozen encoder and stored on CPU. To reduce memory usage and speed up cosine retrieval, the vocabulary is split into three semantic-type partitions: DISO, CHEM, and ANATOMY.

Table 4
Performance across tracks and preprocessing/model variants.

Setting	Accuracy@1	Accuracy@5	MRR
Track 2: Bilingual			
+ Lemmatisation, no lang filter	0.669	0.810	0.728
+ Lemmatisation, + lang filter	0.631	0.761	0.685
No lemmatisation, no filter	0.609	0.802	0.691
Track 1: Russian			
+ Lemmatisation, no lang filter	0.716	0.828	0.761
Lowercase only, no filter	0.689	0.752	0.716
+ Lemmatisation, + lang filter	0.687	0.746	0.712
Track 1: English			
`gebert_eng_gat`, lowercase	0.640	0.833	0.719
`sapbert`, lowercase	0.639	0.838	0.721
`gebert_eng_gat`, stemming	0.523	0.524	0.523

4.4. Inference Pipeline

During inference, the system applies a three-step hybrid retrieval strategy:

1. **Exact match**: if a mention string exactly matches a vocabulary entry (after normalisation), the corresponding CUI is returned.
2. **Cache lookup**: repeated mentions are resolved via a key–value cache that maps (text, type) pairs to previously retrieved predictions.
3. **Cosine search**: for unmatched cases, the mention is encoded and compared against 100k-sized chunks of the corresponding type-specific partition using cosine similarity. Top-100 scores per chunk are retained and merged before deduplication.

The final output consists of the top-k predictions ($k = 5$). If fewer than five candidates survive deduplication, the last one is repeated to meet format constraints.

4.5. Efficiency and Hardware Setup

All experiments were conducted on a single NVIDIA A100-40GB GPU in Google Colab. The offline embedding stage is CPU-bound, while inference uses GPU acceleration. Our key hyperparameters are:

- `batch_size` = 512
- `max_length` = 48 (increased to 64 for English track)
- cosine index chunk size = 100,000 embeddings

The full bilingual test set (12,876 mentions) is processed in under two hours, including I/O.

5. Results

We evaluate our system on all three tracks of the BioNNE-L 2025 shared task. Table 4 summarises the performance under various settings. All runs use the same frozen encoder and retrieval logic, varying only in preprocessing, language filters, and the underlying encoder model (for Track 1: English).

5.1. Track 2: Bilingual

Our best bilingual result is obtained without language filtering: all mentions — regardless of script — search against the full bilingual vocabulary. This setting combines Russian lemmatisation with aggressive lowercasing, yielding Accuracy@1 of 66.9% and MRR of 72.7. Restricting the vocabulary by language degrades performance by over 3 pp, confirming that Russian mentions often resolve better against English entries due to partial Russian coverage.

5.2. Track 1: Russian

The Russian track mirrors this pattern: using the full bilingual vocabulary outperforms language-constrained alternatives by a margin of nearly 3 pp in Accuracy@1. Lemmatisation consistently improves retrieval accuracy compared to plain lowercasing. Our system achieves Accuracy@1 of 71.6%, ranking first among all participating systems.

It is hypothesised that this gain partially stems from transliterated mentions in the Russian test set — e.g., Latin-script tokens referring to Russian concepts. In such cases, allowing retrieval from the English subgraph helps bridge the gap in Russian lexical coverage.

5.3. Track 1: English

The monolingual English track required a separate model. We tested `gebert_eng_gat` and `sapbert`, both fine-tuned on English biomedical data. The best result (Accuracy@1 of 64.0%) comes from `gebert_eng_gat` with minimal preprocessing. Stemming, by contrast, leads to a dramatic drop of 11 pp, highlighting the fragility of surface-level changes in monolingual settings.

6. Uncertainty and Multilingual Behaviour

Because gold CUIs for the hidden test set are not publicly available, uncertainty is evaluated on the bilingual development split `bi_dev`. After deduplicating mentions by their lower-cased, lemmatised string, we obtain a clean evaluation pool of 2,571 unique mentions. For each mention, we store (i) the cosine similarity to its gold embedding, (ii) softmax entropy of the top-20 similarity scores, and (iii) the language of the surface form at rank 1. These variables are then correlated with prediction accuracy to test our uncertainty hypotheses.

Spearman's ρ between cosine distance and entropy is 0.015 ($p = 0.54$), providing *no evidence* for H1/H2. Cosine distances are tightly clustered in the 0.63–0.71 range, reflecting the fully partial-terminology scenario in which all Russian mentions must be mapped to English concepts. This low variance effectively saturates the distance signal and makes it uninformative for uncertainty estimation.

In contrast, entropy shows a strong link with model confidence. Table 5 bins the 2,571 mentions into entropy quintiles. Accuracy@1 falls from 0.87 in the lowest-entropy bin to 0.10 in the highest, while MRR drops four-fold-strongly supporting H5. The high-entropy quintile (Q5) contains only 49 correct predictions out of 514; by contrast, Q1 contains 448 correct answers. Setting an entropy threshold at the 80th percentile would remove 59% of all errors at a cost of just 13% of correct predictions. This highlights entropy as a reliable proxy for confidence despite its narrow numerical range (2.98–2.99).

Table 5
Performance across entropy quintiles on `bi_dev`. Entropy is computed over the top-20 softmax-normalised similarities.

Entropy quintile	n	Acc@1	MRR
Q_1 (1.61–2.99)	515	0.870	0.412
Q_2	514	0.872	0.409
Q_3	514	0.839	0.383
Q_4	514	0.595	0.212
Q_5 (2.99–3.00)	514	0.095	0.040

Uncertainty also varies by mention length and semantic type. One-token mentions achieve Acc@1 = 0.78 and mean entropy of 2.98, whereas seven-token or longer mentions fall to Acc@1 = 0.20 and reach the highest entropy (≈ 2.99). Although such long entities represent only 11% of the data, they account for 38% of the lowest-ranked predictions. In terms of semantic category, chemical entities are normalised with slightly higher accuracy and lower entropy than diseases and anatomy, partially confirming H6.

For every mention in `bi_dev`, the top-ranked surface form is in English. As no gold CUI in this slice has a Russian synonym, the variable `top1_lang` is constant, rendering H3/H4 untestable. Importantly, this absence of Russian forms does not result in higher entropy per se - the bottleneck appears to be terminological, not linguistic.

In summary: (i) cosine distance fails to reflect uncertainty under extreme partial-terminology; (ii) entropy is a robust predictor of error and supports entropy-aware filtering; (iii) longer and more complex entities are harder to normalise; (iv) enriching UMLS with Russian surface variants remains crucial. These findings directly motivate two future directions to be outlined: automatic expansion of Russian UMLS terminology and entropy-based abstention for safe biomedical deployment.

7. Conclusion

We introduced a compact, fine-tuning-free pipeline for biomedical concept normalisation that couples script-aware preprocessing, a frozen BERGAMOT-GAT encoder and a lightweight retrieval scheme over a 4 M-entry bilingual UMLS vocabulary. Despite its simplicity, the approach achieved competitive results in the BioNNE-L 2025 shared task, placing first in the Russian track and second in the bilingual track (0.72 / 0.83 / 0.76 and 0.68 / 0.84 / 0.75 for Acc@1, Acc@5 and MRR, respectively).

To better understand the system's behaviour, we carried out an uncertainty analysis on a deduplicated development subset. While cosine distance to the gold concept proved uninformative under full partial terminology, soft-max entropy over the top-20 candidates emerged as a reliable confidence signal: the most entropic quintile contained the vast majority of errors. We also observed higher uncertainty for long nested mentions and for anatomy/disease entities, whereas chemical terms were linked more confidently.

Future work. These findings point to two directions we plan to explore: (i) automatic expansion of Russian surface forms in UMLS to reduce the terminology gap, and (ii) integration of entropy-aware re-ranking or abstention mechanisms to improve reliability in downstream applications.

Acknowledgments

This work is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

Declaration on Generative AI

During the preparation of this paper, the author used the AI-based tool Grammarly for grammar and spelling checking. No generative AI tools were used for content generation or idea development. The author has carefully reviewed and edited all content and takes full responsibility for the integrity and originality of the publication.

References

[1] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, F. N. Maria Di Nunzio, Giorgio, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: L. P. A. G. S. d. H. J. M. F. P. P. R. D. S. G. F. N. F. Jorge Carrillo-de Albornoz, Julio Gonzalo (Ed.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[2] V. Davydova, N. Loukachevitch, E. Tutubalina, Overview of the bionne task on biomedical nested named entity recognition at bioasq 2024, in: CLEF Working Notes, 2024.
[3] N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, Biomedical concept normalisation over nested entities with partial umls terminology in russian, in: Proceedings of LREC–COLING, 2024.
[4] A. Sakhovskiy, N. Semenova, A. Kadurin, E. Tutubalina, Biomedical entity representation with graph-augmented multi-objective transformer, in: Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 4626–4643. doi:10.18653/v1/2024.findings-naacl.288.
[5] R. Leaman, R. I. Doğan, Z. Lu, Dnorm: Disease name normalization with pairwise learning-to-rank, Bioinformatics 29 (2013) 2909–2917.
[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019.
[7] F. Liu, et al., Self-aligned pre-training for biomedical entity representations, in: Proceedings of EMNLP, 2021.
[8] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pre-training for biomedical entity representations, in: Proc. NAACL, 2021. URL: https://aclanthology.org/2021.naacl-main.334.
[9] F. Liu, I. Vulić, A. Korhonen, N. Collier, Learning domain-specialised representations for cross-lingual biomedical entity linking, in: Proc. ACL (short), 2021. URL: https://aclanthology.org/2021.acl-short.72.
[10] Z. Yuan, Z. Zhao, H. Sun, J. Li, F. Wang, S. Yu, CODER: Knowledge-infused cross-lingual medical term embedding for term normalization, Journal of Biomedical Informatics 126 (2022) 103983. doi:10.1016/j.jbi.2021.103983.
[11] N. D. Cao, G. Izacard, S. Riedel, F. Petroni, Autoregressive entity retrieval, in: Proc. ICLR, 2021. URL: https://openreview.net/forum?id=5k8F6UU39V.
[12] H. Yuan, Z. Yuan, S. Yu, Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning, in: Proc. NAACL, 2022. URL: https://aclanthology.org/2022.naacl-main.296.
[13] V. Nesterov, et al., Ruccon: Clinical concept normalization in russian, in: Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 239–245.
[14] O. Sevgili, A. Shelmanov, M. Arkhipov, A. Panchenko, C. Biemann, Neural entity linking: A survey of models based on deep learning, Semantic Web 13 (2022) 527–570. doi:10.3233/SW-222986.
[15] O. Somov, E. Tutubalina, Confidence estimation for error detection in text-to-sql systems, in: Proceedings of AAAI, 2025.
[16] A. Sakhovskiy, N. Loukachevitch, E. Tutubalina, Overview of the BioASQ BioNNE-L Task on Biomedical Nested Entity Linking in CLEF 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[17] N. Loukachevitch, et al., Nerel-bio: A dataset of biomedical abstracts annotated with nested named entities, Bioinformatics (2023). doi:10.1093/bioinformatics/btad161.