Transformers and the Representation of Biomedical Background Knowledge

Oskar Wysocki[*]
Digital Experimental Cancer Medicine
Team, Cancer Biomarker Centre
CRUK Manchester Institute
University of Manchester
oskar.wysocki@manchester.ac.uk

Zili Zhou
Department of Computer Science
University of Manchester
zili.zhou@manchester.ac.uk

Paul O'Regan
Digital Experimental Cancer Medicine
Team, Cancer Biomarker Centre
CRUK Manchester Institute
University of Manchester
paul.oregan@digitalecmt.com

Deborah Ferreira
Department of Computer Science
University of Manchester
deborah.ferreira@manchester.ac.uk

Magdalena Wysocka
Digital Experimental Cancer Medicine
Team, Cancer Biomarker Centre
CRUK Manchester Institute
University of Manchester
magdalena.wysocka@digitalecmt.org

[*] Kilburn Building, Oxford Rd, Manchester M13 9PL, United Kingdom. E-mail: oskar.wysocki@manchester.ac.uk. Secondary affiliation: Department of Computer Science, University of Manchester.

[†] Other affiliations: Digital Experimental Cancer Medicine Team, Cancer Biomarker Centre, CRUK Manchester Institute, University of Manchester; Department of Computer Science, University of Manchester.

Action Editor: Byron Wallace. Submission received: 16 March 2022; revised version received: 18 August 2022; accepted for publication: 7 September 2022.

https://doi.org/10.1162/coli_a_00462

Dónal Landers
Digital Experimental Cancer Medicine
Team, Cancer Biomarker Centre
CRUK Manchester Institute
University of Manchester
donal.landers@delondraoncology.com

André Freitas[†]
Idiap Research Institute
Martigny, Switzerland
andre.freitas@manchester.ac.uk

Specialized transformers-based models (such as BioBERT and BioMegatron) are adapted for the biomedical domain based on publicly available biomedical corpora. As such, they have the potential to encode large-scale biological knowledge. We investigate the encoding and representation of biological knowledge in these models, and its potential utility to support inference in cancer precision medicine—namely, the interpretation of the clinical significance of genomic alterations. We compare the performance of different transformer baselines; we use probing to determine the consistency of encodings for distinct entities; and we use clustering methods to compare and contrast the internal properties of the embeddings for genes, variants, drugs, and diseases. We show that these models do indeed encode biological knowledge, although some of this is lost in fine-tuning for specific tasks. Finally, we analyze how the models behave with regard to biases and imbalances in the dataset.

1. Introduction

Transformers are deep learning models that are able to capture linguistic patterns at scale. By using unsupervised learning tasks that can be defined over large-scale textual corpora, these models are able to capture both linguistic and domain knowledge, which can be later specialized for specific inference tasks. The representation produced by the model is a high-dimensional linguistic space that represents words, terms, and sentences as vector projections. In Natural Language Processing, transformers are used to support natural language inference and classification tasks. The assumption is that the models can encode syntactic, semantic, commonsense, and domain-specific knowledge and use their internal representation for complex textual interpretation. While these models provided measurable improvements in many different tasks, the limited interpretability of their internal representation challenges their application in areas such as biomedicine.

In this work we elucidate a set of the internal properties of transformers in the context of a well-defined cancer precision medicine inference task, in which the domain knowledge is expressed within the biomedical literature. We focus on systematically determining the ability of these models to capture fundamental entities (gene, gene variant, drug, and disease), their relations and supporting facts, which are fundamental for supporting inference in the context of molecular cancer medicine. For example, we aim to answer the question whether these models capture biological knowledge such as the following:

- *"T790M is a gene variant"*
- *"T790M is a variant of the EGFR gene"*
- *"The T790M variant of the EGFR gene in lung cancer is associated with resistance to Erlotinib"* - well supported statement (Level A - Validated association, Confidence rating: 5 stars)
- *"The T790M variant of the EGFR gene in pancreatic cancer is associated with resistance to Osimertinib"* - less supported statement (Level C - Case study, Confidence rating: 2 stars)

In the example above, the first two facts capture basic definitional knowledge (mapped respectively to an unary and binary predicate-argument relation), while the third and fourth facts capture a full scientific statement that can be mapped to a complex *n*-ary relation, and are supported by different levels of evidence in the literature. The establishment of the truth condition of facts of these types in the context of a biomedical natural language inference task is a desirable property for these models. With this motivation in mind, this work provides a critical exploration of the internal representation properties of these models, using probing and clustering methods. In summary, we aim to answer the following research questions (RQs):

RQ1 Do transformer-based models encode fundamental biomedical domain knowledge at an entity level (e.g., gene, gene variant, disease, drug) and at a relational level?

RQ2 Do these models encode complex biomedical facts/*n*-ary relations?

RQ3 Are there significant differences in how different model configurations encode domain knowledge?

RQ4 How these models cope with evidence biases in the literature (e.g., are facts more frequently expressed in the literature, elicited in the models)?

In this analysis, we used state-of-the-art transformers specialized for the biomedical domain: BioBERT (Lee et al. 2020) and BioMegatron (Shin et al. 2020). Both models are pre-trained over large biomedical text corpora (PubMed[1]). These models have been shown, in an extrinsic setting, to address complex domain-specific tasks (Wang et al. 2021), such as answering biomedical questions (Shin et al. 2020). Yet, the internal representation properties of these models are not fully characterized, a requirement for their safe and controlled application in a biomedical setting.

This article focuses on the following contributions:

- A systematic evaluation of the ability of biomedical fine-tuned transformers (BioBERT and BioMegatron) to capture entities, complex relations, and level of evidence support for biomedical facts within a

[1] www.ncbi.nlm.nih.gov/pubmed.

specific domain of inference (cancer clinical trials). Instead of focusing only on extrinsic performance (in the context of a classification task), we elicit some of the internal properties of these models with the support of clustering and probing methods.

- To the best of our knowledge, this is the first work that systematically links the evidence from a high-quality, expert-curated knowledge base with the representation of biomedical knowledge in transformers, namely, *n*-ary relations and entity types.
- We used probing methods to inspect the consistency of entities and associated types (i.e., genes, variants, drugs, diseases) contrasting pre-trained and fine-tuned models. This allowed for the evaluation of whether the model captures the fundamental biomedical/semantic categories to support interpretation. We quantified how much semantic structure is lost in fine-tuning.
- To the best of our knowledge, this is the first work that quantifies the relation of classification error to entities distribution in the dataset and evidence items in literature, emphasizing the risk of and demonstrating examples of significant errors in the cancer precision medicine inference task. We show that, despite the soundness and strength of the evidence in the biomedical literature, some well-known clinical relations can be misclassified.
- Lastly, we provided a qualitative analysis of the significant clustering patterns of the embeddings, using dimensionality reduction and unsupervised clustering methods to identify qualitative patterns expressed in the representations. This approach allowed for identification of biologically meaningful representations, for example, groups with genes from the same pathways. Additionally, by measuring homogeneity of clusters, we quantified the associations between the representations and the entity type and target labels.

The workflow of the analysis is summarized in Figure 1.

2. Methods

2.1 Motivational Scenario: Natural Language Inference in Cancer Clinical Research

Cancer precision medicine, which is the selection of a treatment for a patient based on molecular characterization of their tumor, has the potential to improve patient outcomes. For example, activating mutations in the epidermal growth factor receptor gene (EGFR) predict response to gefitinib, and amplification or overexpression of ERBB2 predicts response to anti-ERBB2 therapies such as lapatinib. Tests for these markers that guide therapy decisions are now part of the standard of care in non-small-cell lung cancer (NSCLC) and breast cancer (Good et al. 2014).

Routine molecular characterization of patients' tumors has become feasible because of improved turnaround times and reduced costs of molecular diagnostics (Rieke et al. 2018). In England, the NHS England genomic medicine service aims to offer whole genome sequencing as part of routine care. The aim is to match people to the most

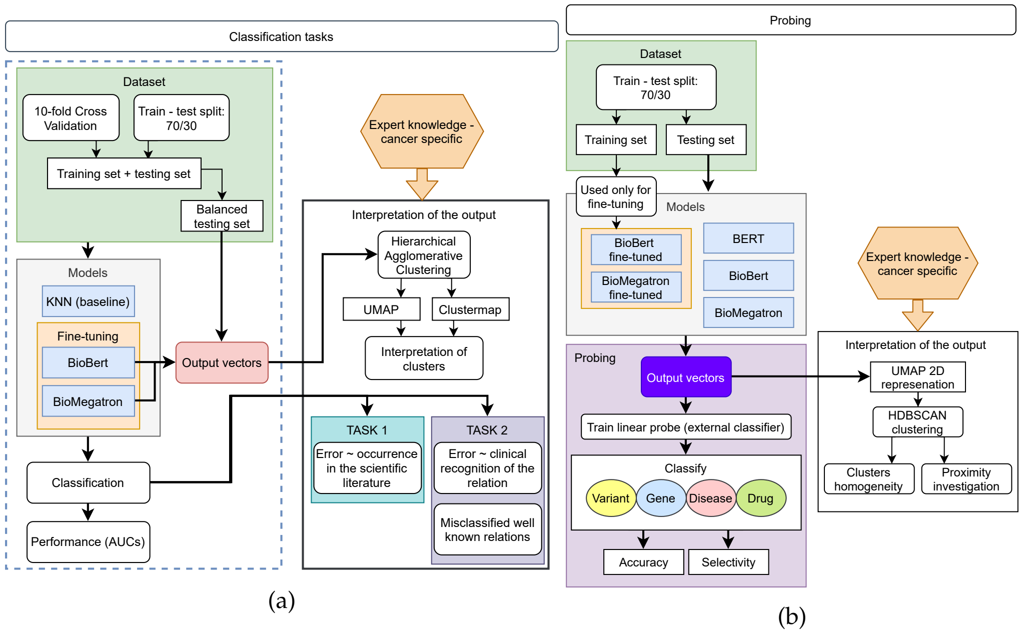


Figure 1
The workflow of the performed analysis.

effective interventions, in order to increase survival and reduce the likelihood of adverse drug reactions.[2]

Even considering only licensed treatments, the number of alternative treatments available may be very large. For example, in the United States, there are over 70 drugs approved by the US Food and Drug Administration for the treatment of NSCLC.[3] If experimental treatments are included in the decision-making process, the number of alternative treatments available is substantially increased.

Furthermore, as the breadth of molecular testing increases, so too does the volume of information available for each patient and thus the complexity of the treatment decision. Interpretation of the clinical and functional significance of the resulting data presents a substantial and growing challenge to the implementation of precision medicine in the clinical setting.

This creates a need for tools to support clinicians in the evaluation of the clinical significance of genomic alterations in order to be able to implement precision medicine. However, much of the information available to support clinicians in making treatment decisions is in the form of unstructured text, such as published literature, conference proceedings, and drug prescribing information. Natural language processing methods have the potential to scale-up the interpretation of this evidence space, which could be integrated into decision support tools. The utility of a decision support tool is expressed in providing support for individual recommendations. Despite acknowledging the inherent imperfectness of the model's overall performance, the trustworthiness and safety of such a tool would require the correct interpretation of biological facts and emerging

[2] https://www.england.nhs.uk/genomics/nhs-genomic-med-service/.
[3] https://www.cancer.gov/about-cancer/treatment/drugs/lung.

evidence. This work validates an approach of applying fine-tuned transformers to two simple NLI tasks, investigating encoded knowledge within the models together with aforementioned individual well-established clinical relations. This work contributes for the first time with two concrete cancer precision medicine inference tasks based on a high quality, manually curated dataset. For general evaluation of transformers in biomedical applications, please refer to Wang et al. (2021), Alghanmi, Espinosa Anke, and Schockaert (2021), and Jin et al. (2019), where the models are tested in multiple downstream tasks.

2.2 Reference Clinical Knowledge Base (KB)

CIViC[4] (Clinical Interpretation of Variants in Cancer) is a community-edited knowledge base (KB) of associations between genetic variations (or other alterations), drugs, and outcomes in cancer (Griffith et al. 2017). The goal of CIViC is to support the implementation of personalized medicine in cancer. Data is freely available and licensed under a Creative Commons Public Domain Dedication (CC0 1.0 Universal). The knowledge base includes a detailed curation of evidence obtained from peer-reviewed publications and meeting abstracts. The CIViC database supports the development of computational tools for the functional prediction and interpretation of the clinical significance of cancer variants. Together with OncoKB (Chakravarty et al. 2017) and My Cancer Genome,[5] it is one of the most commonly used KBs for this purpose (Borchert et al. 2021).

An evidence statement is a brief description of the clinical relevance of a variant that has been determined by an experiment, trial, or study from a published literature source. It captures a variant's impact on clinical action, which can be predictive of therapy, correlated with prognostic outcome, inform disease diagnosis (i.e., cancer type or subtype), predict predisposition to cancer in the first place, or relate to the functional impact of the variant. For each item of evidence, additional attributes are captured, including:

- *Type* - the type of clinical (or biological) association described (Predictive, Prognostic, Functional, etc.).
- *Direction* - whether the evidence supports or refutes the clinical significance of an event.
- *Level* - a measure of the robustness of the associated study, where *A - Validated association* is the strongest evidence, and *E - Inferential association* is the weakest evidence.
- *Rating* - a score (1-5 stars) reflecting the database curator's confidence in the quality of the summarized evidence.
- *Clinical Significance* - describes how the variant is related to a specific, clinically relevant property (e.g., drug sensitivity or resistance).

CIViC is programmatically accessible via API and as a full dataset and is integrated into various recent annotation tools and follows an ontology driven conceptual model.

[4] https://civicdb.org/home.
[5] https://www.mycancergenome.org/.

It allows users to transparently generate current and accurate variant interpretations because it receives monthly updates. As of October 2022, the database holds 9,302 interpretations of clinical relevance for 3,337 variants among 470 genes associated with 341 diseases and 494 drugs. Its accessibility and tabular format of the data allows for easy integration into Machine Learning pipelines, both as input data and domain knowledge incorporated in the model.

2.3 Data Preprocessing and Set-up

The process of pre-processing the CIViC data for the purpose of this study is detailed in the Appendix.

As we were interested in identifying gene variants that predict response to one or more drugs, we retained only those evidence items where *Evidence Direction* contains the value *Supports* and *Evidence type* has the value *Predictive*.

2.3.1 Task 1 - Generation of True/False Entity Pairs. The first classification task (Figure 2) was to determine whether a transformer model, pre-trained on the existing biomedical corpus and fine-tuned for the task, could correctly classify associations between pairs of entities *entity1-entity2* as true or false based on knowledge embedded from the biomedical corpus. For example, the correct classification of T790M as a variant of the EGFR gene but not of the KRAS gene.

Three types of binary relations were considered:

- drug - gene
- drug - variant
- variant - gene

Pairs of entities with genuine associations ("true pairs") were generated from the CIViC knowledge base; pairs of entities with no such association ("false pairs") were

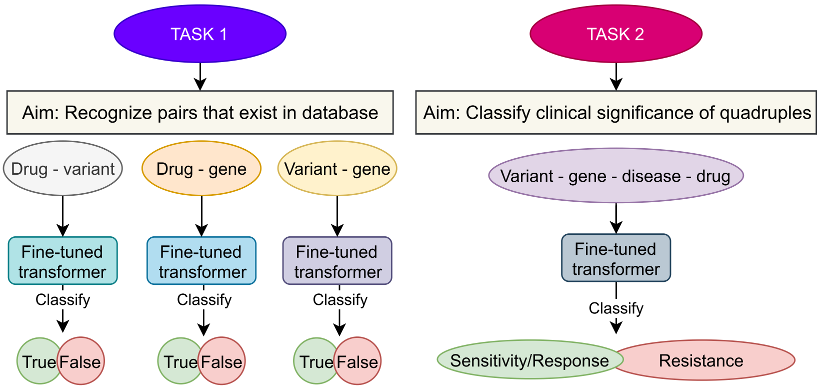


Figure 2
An overview of classification task 1 and 2. Each transformer block represents a separate model that was fine-tuned separately for each classification. Two transformers were used: BioBERT and BioMegatron.

generated by randomly selecting entities from CIViC, and excluding those that already exist (i.e., negative sampling). The dataset includes an equal number of false and true pairs. Of note, a pair can occur in multiple evidence items, that is, be duplicated in the database, but our datasets of pairs consisted of unique pairs.

2.3.2 Task 2 - Generation of Variant-Gene-Disease-Drug Quadruples. The second classification task (Figure 2) was to infer the clinical significance (CS) of a gene variant for drug treatment in a given cancer type. For example, considering examples of resistance mutations from the CIViC dataset, can the model correctly classify that the T790M variant of the EGFR gene in lung cancer confers resistance to gefitinib?

Sentences describing genuine relationships were generated using quadruples of entities extracted from CIViC, following the pattern:

"`[variant entity]` of `[gene entity]` identified in `[disease entity]` is associated with `[drug entity]`"

An evidence item in the KB contains variant, gene, disease, drug, and CS, so a quadruple can be extracted directly from the KB, and there are no false quadruples. Only unique quadruples were used to create the dataset. In the case of a combination or substitution of multiple drugs in the evidence item, we replaced `[drug entity]` with multiple entities joined with the conjunction *and* (e.g., `[drug entity1]` and `[drug entity2]` and `[drug entity3]`).

After the filtering in the pre-processing stage, 4 values for CS remained: *Resistant*, *Sensitivity/Response*, *Reduced Sensitivity*, and *Adverse Response*. Due to a negligible number of quadruples we excluded the *Adverse Response* class. The class *Reduced Sensitivity* was joined with *Sensitivity/Response*.

Multiple evidence items in CIViC can represent one quadruple. For the purpose of Task 2, only the quadruples with uniform clinical significance were selected (98% of total); that is, all evidence items for a unique quadruple describe the same relation.

2.3.3 Balancing the Test Set. In order to reduce the bias that some pairs/quadruples containing specific entities are almost always true|false or sensitive|resistant, we applied a balancing procedure (Appendix). We excluded the imbalanced pairs/quadruples from the *test set* in creating a *balanced test set*. Reducing the bias allows us to compare the test results more fairly.

2.4 Model Building

2.4.1 Baseline Model. In this article, we used a naive classification model (Nearest Neighbors Classification model [Fix and Hodges 1989]) as a baseline. The intent behind this baseline was to contrast a transformer-based model with a simple, non-pre-trained model (K-Nearest Neighbor (KNN)). This is to control for the role of the pre-training (i.e., transformer models would show better performance as a result of knowledge embedded in the model, and not due to the relations expressed in the training set). The KNN baseline is used as a control to assess the performance achieved solely due to the distribution of entities in the dataset, as KNN does not embed any distributional knowledge.

Briefly, each entity was represented as a sparse, one-hot encoded vector such that, for example, for genes, the length of the vector was equal to the total number of genes, and the element corresponding to the given gene was set to 1, while all other elements were set to 0. The model was trained and validated for each task based on subsets of the CIViC data as described below.

For Task 1, each pair of vectors (representing each pair of entities) was concatenated as an input; for Task 2, sets of 4 vectors, representing *variant*, *gene*, *disease*, and *drug* entities, were concatenated. Note that vectors for *drug* entities may contain multiple 1-values because some sentences may mention more than one drug.

2.4.2 Transformers. In this work, we transfer pairs and evidence sentences into text sequences as input data of both BioBERT and BioMegatron; aggregate the outputs of transformers into one vector representation for each input sequence; and stack classification layers on top of this vector representation for our defined pairs/sentences classification tasks.

Specifically, in Task 1 when predicting the relation between a gene entity and a drug entity, we can input the following sequence into the model:
seq_{drug_gene}="[CLS] `[drug entity]` is associated with `[gene entity]` [SEP]"
Similarly, for the relationship between a variant entity and a drug entity:
$seq_{drug\_variant}$="[CLS] `[drug entity]` is associated with `[variant entity]` [SEP]"
And for a pair of gene and variant entities:
$seq_{variant\_gene}$="[CLS] `[variant entity]` is associated with `[gene entity]` [SEP]"
In Task 2, for a sentence representing a clinical significance, we define the input sequence as:
$seq_{sentence}$="[CLS] `[variant entity]` of `[gene entity]` identified in `[disease entity]` is associated with `[drug entities]`[SEP]"

Pre-trained BioBERT and BioMegatron were fine-tuned: for pairs (gene-variant, gene-drug, variant-drug true/false) classification, 5 epochs 3e-5 learning rate; for quadruple classification, 5 epochs, 1e-4 learning rate. For more details please refer to the Appendix.

2.5 Probing

This section describes the semantic probing methodology implemented in order to shed light on the obtained representations from Task 1 and Task 2. All probing experiments have been performed using the Probe-Ably[6] framework, with default configurations.

Probing is the training of an external classifier model (also called a "probe") to determine the extent to which a set of auxiliary target feature labels can be predicted from the internal model representations (Ferreira et al. 2021; Hewitt and Manning 2019; Pimentel et al. 2020). Probing is often performed as a post hoc analysis, taking a pre-trained or fine-tuned model and analyzing the obtained embeddings. For example, previous probing studies (Rives et al. 2021) have found that training language models across amino acid sequences can create embeddings that encode biological structure at multiple levels, including proteins and evolutionary homology. Knowledge of intrinsic biological properties emerges without supervision, that is, with no explicit training to capture such property.

As previously highlighted, Task 1 has three different subtasks: classifying the existence of three different pairs of entities in the dataset (drug-gene, drug-variant, and variant-gene). For each task, we obtain a fine-tuned version of BioBERT and BioMegatron. For Task 2, only one fine-tuned version is produced for each model. One crucial question is: *Do such models retain the meaning of those entities when fine-tuning the models?*

[6] `https://github.com/ai-systems/Probe-Ably/`.

One way of examining such properties is by testing if such representations can still correctly map the entities to their type (e.g., taking the representation of the word tamoxifen and correctly classifying it as a drug).

Intending to answer this question, we implement the following probing steps:

1. Generate the representations (embeddings) obtained by the fine-tuned (for Task 1 and Task 2) and non-fine-tuned models (BioBERT and BioMegatron) for each entity (drug, variant, gene, and disease) for each sentence in the test set. We also include BERT-base to the analysis in order to assess the performance of a more general model. Even though most of the entities are composed of a single word, these models depend on the WordPiece tokenizer, often breaking a word into separate pieces. For example, the word tamoxifen is tokenized as four pieces: `[Tam, ##ox, ##ife, ##n]` using the BioBERT tokenizer. To obtain a single vector for each entity, we compute the average of all the token representations composing that word. For instance, the word tamoxifen is represented as a vector containing the average of the vectors representing each of its four pieces.

2. The goal of probing is merely to find what information is already stored in the new model, not to train a new task. Thus, following standard probing guidelines (Ferreira et al. 2021), we split the representations into training, validation, and test set, using a 20/40/40 scheme. By such a split, we want to limit the number of instances seen during training and avoid overfitting over a large part of the dataset, since part of the dataset was already observed during the first task training, and the information is partly stored in the generated vectors. The model overfitting is also prevented with the use of a linear model. Each model is trained for 5 epochs, with the validation set being used to select the best performing model (in terms of accuracy).

3. After obtaining all representations for each model and respective entity types, we train a total of 50 linear probes to classify each representation into the correct entity label. The number 50 is a default configuration and recommended value from the Probe-Ably framework. These different 50 models are contrasted using a measure of complexity. When using models containing a large number of parameters, there is a possibility that the probing training will reshape the representation to fit the new task, leading to inconclusive results; therefore, we opt for a simpler linear model to avoid this phenomena. We follow previous research in probing (Pimentel et al. 2020), measuring the complexity of a linear model $\hat{y} = W\mathbf{x} + \mathbf{b}$ by using the nuclear norm of the weight matrix W, computed as:

$$||\mathbf{W}||_* = \sum_{i=1}^{min(|\mathcal{T}|,d)} \sigma_i(\mathbf{W})$$

where $\sigma_i(W)$ is the i-th singular value of W, $|\mathcal{T}|$ is the number of targets (e.g., number of possible entities), and d is the number of dimensions in the representation (e.g., 768 dimensions for BERT-base).

The nuclear norm is then included in the loss (weighted by a parameter λ)

$$-\sum_{i=1}^{n} \log p(t^{(i)} \mid \mathbf{h}^{(i)}) + \lambda \cdot ||\mathbf{W}||$$

and is thus regulated in the training loop, where t is a single value of $\mathcal{T}$. In order to obtain 50 different models, we randomly initialize the dropout and λ parameter. As suggested in Pimentel et al. (2020), we show the results across all the different initializations in Figure 5. Having models with different complexity allows us to see if the results are consistent across different complexities, with the best performance usually being obtained by the more complex models.

4. For each trained probe, we also train an equivalent control probe. The control probe is a model trained for the same task as the original probe, however, the training is performed using random labels, instead of the correct ones. Having a control task can been seen as an analogy to having a study with placebo medication. When the performance on the probing task is better than the control task, it is known that the probe model is capturing more than random noise.

5. The performance of the probes is measured in terms of *Accuracy* and *Selectivity* for the test set. The selectivity score, namely, the difference in accuracy between the representational probe and a control probing task with randomized labels, indicates that the probe architectures used are not expressive enough to "memorize" unstructured labels. Ensuring that there is no drop-off in selectivity increases the confidence that we are not falsely attributing strong accuracy scores to the representational structure where over-parameterized probes (i.e., probes that contain several learnable parameters) could have explained them.

2.6 Clustering

In addition to the evaluation of models' performance in a probing setting, we investigated with the support of clustering methods whether the output vectors can identify potential relationships between entity pairs and/or quadruples.

For clustering the output in Tasks 1 and 2 we used hierarchical agglomerative clustering (HAC) with Ward variance minimization algorithm (ward linkage) and Euclidean distance as distance metric on both the rows (output dimensions) and the columns (vector representations of true pairs). Then we identified clusters using a distance threshold defined pragmatically after visual investigation of the clustermap and dendrogram. For clustering the output used in Probing, we used HDBSCAN (McInnes, Healy, and Astels 2017; McInnes and Healy 2017), with parameter min cluster size = 120, while the remaining parameters kept their default values.

We applied Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) (McInnes et al. 2018) to compare patterns observable after dimensionality reduction into 2 dimensions with clusters obtained via HAC. UMAP parameters: default (n components = 2, n neighbors = 15)

The UMAP representation constitutes multiple distinct groups that contain various entity types or target labels. To quantify that, the HDBSCAN algorithm was used, which identifies clusters of densely distributed points. We used homogeneity metric as a measure of proportion of various labels in one cluster. It can be defined as the ratio of the count of the most common label in the cluster and the total count in the cluster, for example, if a cluster contains 40 drugs and 10 genes, homogeneity equals 0.8. Ideally, all clusters would score 1.

3. Results

3.1 Can Transformers Recognize Existing Relations/Associations? - Task 1

3.1.1 Distribution of Entities in Pairs. A total of 8,032 entity pairs were included in this analysis: 5,320 (66%) in the training set, 2,412 in the imbalanced test set, and 1,090 in the balanced test set (Table 1).

Table 1
Statistics about the datasets used in Task 1: Number of unique pairs and entities.

	Pairs (both True and False) (n)				Unique (n)			Unique in balanced test set (n)		
	Total	Train set	Test set	Balanced test set (% of test set)	Genes	Variants	Drugs	Genes	Variants	Drugs
drug – variant	3,676	2,272	1,104	418 (38%)	–	897	242	–	321	134
drug - gene	2,480	1,736	744	396 (53%)	302	–	432	235	–	193
variant – gene	1,876	1,312	564	276 (49%)	125	910	–	72	235	–

Entities in the dataset were distributed non-uniformly, resembling a Pareto distribution. For drug-gene pairs, the majority of pairs involving the most common genes and drugs were true (Figure A.1a). A similar pattern was observed for drug-variant pairs (Figure A.1b). In contrast, for variant-gene pairs, the majority of pairs involving the most common variant entities were false (Figure A.1c).

3.1.2 Performance. We evaluated the classification performance both on the test set and balanced test set using area under the Receiver Operator Characteristic curve (AUC, Table 2).

In all cases, performance was superior for the imbalanced dataset compared with the balanced dataset. As the usage of the balanced test set is to adjust the analysis for frequent pairs with consistent labels (almost all true or all false), the drop in performance suggests that the fine-tuned models are sensitive to the distribution bias in the training set and learn statistical regularities. They favor more frequent pairs and disfavor less frequent ones, which aligns with previous research (Nadeem, Bethke, and Reddy 2021; Gehman et al. 2020; McCoy, Pavlick, and Linzen 2019; Zhong, Friedman, and Chen 2021; Gururangan et al. 2018; Min et al. 2020).

Performance of the transformers was superior to the baseline model in all cases, except for drug-gene classification against the imbalanced dataset. For the drug-gene scenario, the AUC is close to 0.5, which means that classification resembles random

Table 2
AUC in classification task 1.

Pairs + Model	**Imbalanced**		**Balanced**	
	Test set	**10fold CV (sd)**	**Test set**	**10fold CV (sd)**
Drug-Variant				
KNN (baseline)	0.771	.821 (.023)	0.486	.444 (.044)
BioBERT	0.834	.856 (.027)	0.590	.569 (.033)
BioMegatron	0.847	.850 (.022)	0.642	.580 (.070)
Drug-Gene				
KNN (baseline)	0.705	.770 (.025)	0.492	.425 (.037)
BioBERT	0.743	.762 (.024)	0.544	.506 (.048)
BioMegatron	0.722	.755 (.045)	0.572	.512 (.055)
Variant-Gene				
KNN (baseline)	0.683	.778 (0.022)	0.434	.413 (.056)
BioBERT	0.826	.855 (.033)	0.677	.669 (0.62)
BioMegatron	0.828	.813 (.078)	0.671	.627 (.104)

Table 3
Number of evidence items related to the type of pair in the dataset.

Pair	Number of evidence items				
	1	**>1**	**>2**	**≥10**	**≥20**
gene-drug ($n = 1{,}240$)	795 (64.1%)	445 (35.9%)	267 (21.5%)	73 (5.9%)	41 (3.3%)
variant-gene ($n = 938$)	596 (63.5%)	342 (36.5%)	215 (22.9%)	41 (4.4%)	17 (1.8%)
variant-drug ($n = 1{,}838$)	1,347 (73.3%)	491 (26.7%)	230 (12.5%)	20 (1.1%)	1 (3.02%)

guessing and is very limited, if any biological knowledge is utilized (RQ1). Considering only the performance in Task 1, there is no significant difference between BioBERT and BioMegatron, establishing an equivalence of both representations in the context of this task (RQ3).

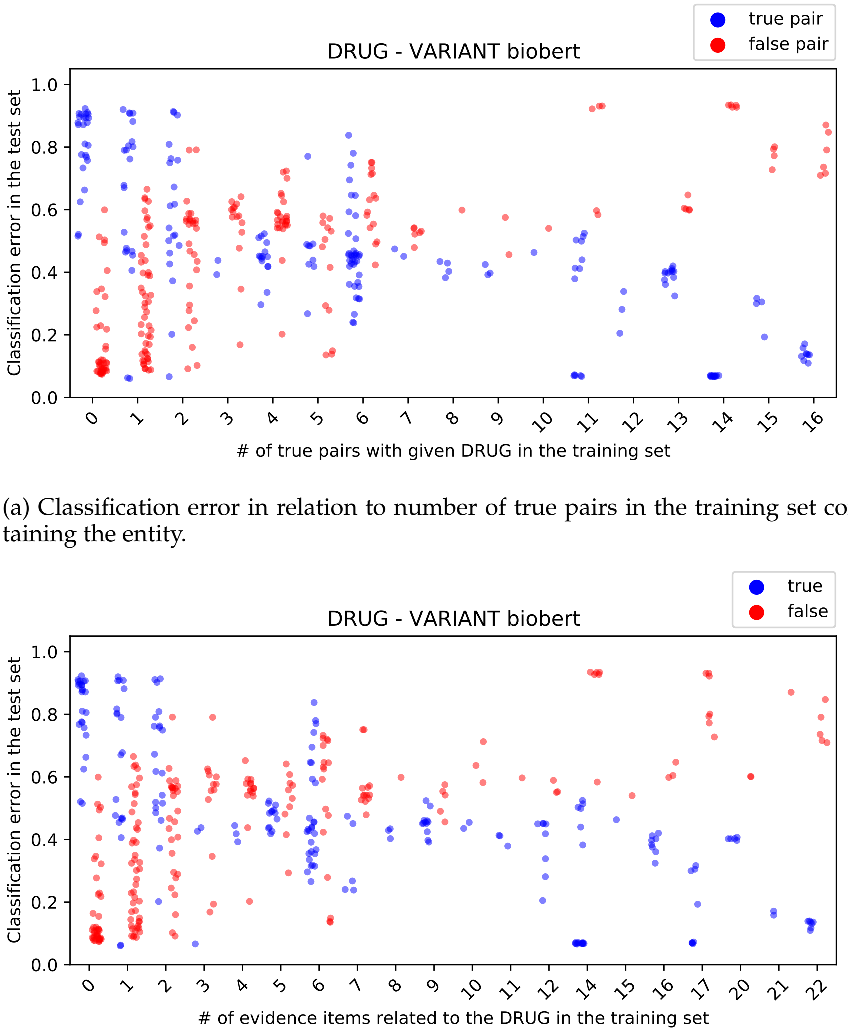
3.1.3 The Impact of Imbalance on the Model's Error. As we observed significant differences between performance on the imbalanced and balanced test sets, we investigated further the specifics of this phenomenon, namely, classification error for individual pairs. One or more evidence items can represent each pair (i.e., each pair can be found in one or more scientific papers). Similar to entities distribution, there is an imbalance in the number of evidence items related to pairs. For example, **73.3%** of variant-drug pairs are supported only by one, **12.5%** by > 2, and **1.1%** by $\geq$10 evidence items. Details for all 3 types of pairs are shown in the Table 3.

Classification error on the balanced test set varied according to the frequency of true pairs in the dataset—for drugs that occurred frequently in the training set (Figure 3a) or in the knowledge base (Figure 3b), true drug-variant pairs were typically classified correctly, whereas false drug-variant pairs were typically misclassified.

The analysis of error quantifies the impact of the imbalance in the dataset on the performance (RQ4). It shows that if an entity occurs in many true pairs in the training set, an unseen pair containing the entity from the test set is likely to be classified as true, regardless of biological meaning. Fine-tuned transformers are highly influenced by learned statistical regularities. For instance, pairs with drugs that occur in 15 true pairs in the training set obtain error <0.1 for true pairs and error >0.7 for false pairs (Figure 3a) as to all of them the model assigns a high probability of being true. This applies to the drug (significant Spearman correlation, $p < 0.001$), gene ($p < 0.001$), and variant entities ($p < 0.05$). All correlations are summarized in Supplementary Table A.1.

Similar correlation is observed regarding the error and the number of evidence items in the KB. The more evidence items related to an entity, the higher chance of a pair (containing this entity) being classified as true. For instance, if a pair contains a drug that is supported by only one evidence item, the pair is more likely to be labeled as false (Figure 3b).

This can be a major concern in applications in cancer precision medicine. There is little value of being accurate for well-known relations and facts. The true potential is for the less-obvious queries, which the experts are less familiar with. However, as shown above, biomedical transformers suffer from reduced performance for underrepresented cases in the dataset (RQ4).



(a) Classification error in relation to number of true pairs in the training set containing the entity.

(b) Classification error in relation to number of evidence items (i.e., scientific papers) describing the entity.

Figure 3
Evaluation of the impact of the dataset imbalance on model's performance: The more true pairs in the training set containing a DRUG entity (a), or the more evidence items related to a DRUG entity in the knowledge base (b), the higher change for a pair (containing the DRUG entity) of being classified as true.

3.2 Can Transformers Recognize Clinical Significance of a Relation? - Task 2

3.2.1 Distribution of Entities in Quadruples. A total of 2,989 quadruples were included in this analysis, 897 in the test set. As a result of balancing the test set, 207 quadruples are left for further investigation of the output vectors. It comprised 147 unique variants, 67 genes, 43 diseases, and 89 drugs (see Table 4).

Similar to the observed distribution of entity pairs, the distribution of entities among the quadruples was also non-uniform, with a Pareto distribution: The most

Table 4
Statistics about the datasets used in Task 2: Number of unique quadruples and entities.

Dataset	**Unique (n)**				
	Quadruples	**Variant**	**Gene**	**Disease**	**Drug**
Total	2,989	1,015	302	215	733
Training set	2,092	803	258	186	579
Test set	897	432	165	135	339
Balanced test set	207	147	67	43	89

common variant entity was *MUTATION*, the most common gene entity was *EGFR*, the most common disease was *Lung Non-small Cell Carcinoma*, and the most common drug was *Erlotinib* (see Supplementary Figure A.2).

In most cases (64%), the clinical significance of quadruples in the dataset was *Sensitivity/Response*. The imbalance between *Sensitivity/Response* and *Resistance* was most evident for the most common variants (*MUTATION, OVEREXPRESSION, AMPLIFICATION, EXPRESSION, V600E, LOSS, FUSION, LOSS-OF-FUNCTION and UNDEREXPRESSION*), where approximately 80% of quadruples related to drug sensitivity.

3.2.2 Performance. We evaluated the performance of the models in predicting the clinical significance of quadruples using AUC. In all cases, performance of the transformer models was superior to that of the KNN (non-pre-trained) baseline. Similar to the results for classification of entity pairs, performance was superior for the imbalanced dataset compared with the balanced dataset. Nevertheless, both BioBERT and BioMegatron achieved high accuracy (AUC >0.8) on the balanced dataset (Table 5). No significant difference between BioBERT and BioMegatron was observed (RQ3). Compared to the performance in Task 1, we observe a smaller drop in AUCs between the imbalanced and balanced test set, while the difference between transformers and KNN is significantly higher. This suggests that in the more complex Task 2, fine-tuned BioBERT and BioMegatron exploit some of the biological knowledge encoded within the architecture (RQ1). This accentuated difference between pre-trained and transformer-based baselines (when contrasted to the previous task) demonstrates that the benefit of the pre-training component of transformers can be better observed in the context of complex n-ary relations (RQ2).

Table 5
AUC in classification task 2 for imbalanced and balanced test set. CV = Cross Validation; sd = standard deviation.

AUC	Binary classification of quadruples			
	Imbalanced		Balanced	
	Test set	10fold CV (sd)	Test set	10fold CV (sd)
KNN (baseline)	0.878	.864 (.023)	0.753	.655 (.065)
BioBERT	0.898	.904 (.024)	0.806	.835 (.060)
BioMegatron	0.905	.910 (.022)	0.826	.833 (.037)

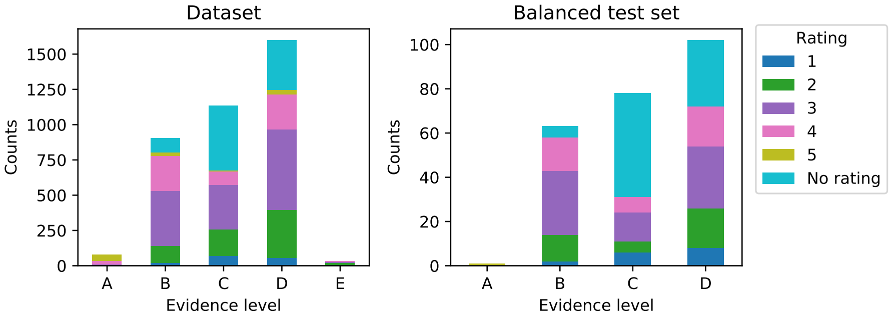


Figure 4
Number of evidence items in the datasets stratified by evidence level and evidence rating.

3.2.3 Model's Error vs. Strength of Biomedical Evidence. High confidence associations (*Evidence rating* = *5*) were rare—most quadruples in the balanced test set were either unrated or evidence level 3 (*Evidence is convincing, but not supported by a breadth of experiments*).

The most common type of evidence (denoted by the *Evidence level* attribute) described by quadruples in the dataset was *D - Preclinical evidence*; validated associations (*Evidence level* = *A*) were rare—only a single example remained in the test set after balancing. No inferential associations (*Evidence level* = *E*) remained in the balanced test set (Figure 4).

In the balanced test set, considering all levels of evidence, there was no correlation between level of evidence and model performance (p >0.05, Spearman correlation). Thus, we do observe that transformers are not better at classifying relations that are supported by strong evidence in the KB. Quite the opposite, AUCs for evidence level B were lower (.683 and .703) than for C and D (BioBert: .900 and .812; BioMegatron: .939 and .816, see Supplementary Table A.3). Considering pre-clinical evidence only (*Evidence level D*), the KNN model had significantly higher error compared with BioBERT (Mann-Whitney U test: $p = 0.014$) and BioMegatron ($p = 0.007$). This finding was supported by AUC and Brier scores (Supplementary Table A.3).

3.2.4 Misclassified Well-known Relations. A total of 16 well-known relations, defined as Evidence level A (*Validated association*) or B (*Clinical evidence*) and Evidence rating 5 (*Strong, well supported evidence from a lab or journal with respected academic standing*) or 4 (*Strong, well supported evidence*) were identified in the balanced test set (Table 6).

Despite the higher confidence assigned to these quadruples, the models did not perform better against these relations compared with the overall balanced test set—AUC for these quadruples was 0.75, 0.78, and 0.75 for BioBERT, BioMegatron, and KNN, respectively. For example, high classification error rates (≥ .6) were observed for transformer models for the following quadruples:

- *EXPRESSION - HSPA5 - Colorectal Cancer - Fluorouracil*
- *EXPRESSION - PDCD4 - Lung Cancer - Paclitaxel*
- *V600E - BRAF - Colorectal Cancer - Cetuximab and Encorafenib and Binimetinib* (BioMegatron only)

Table 6
List of 16 well-known relations and corresponding classification error. R stands for *Resistance* and S/R is for *Sensitivity/Response*.

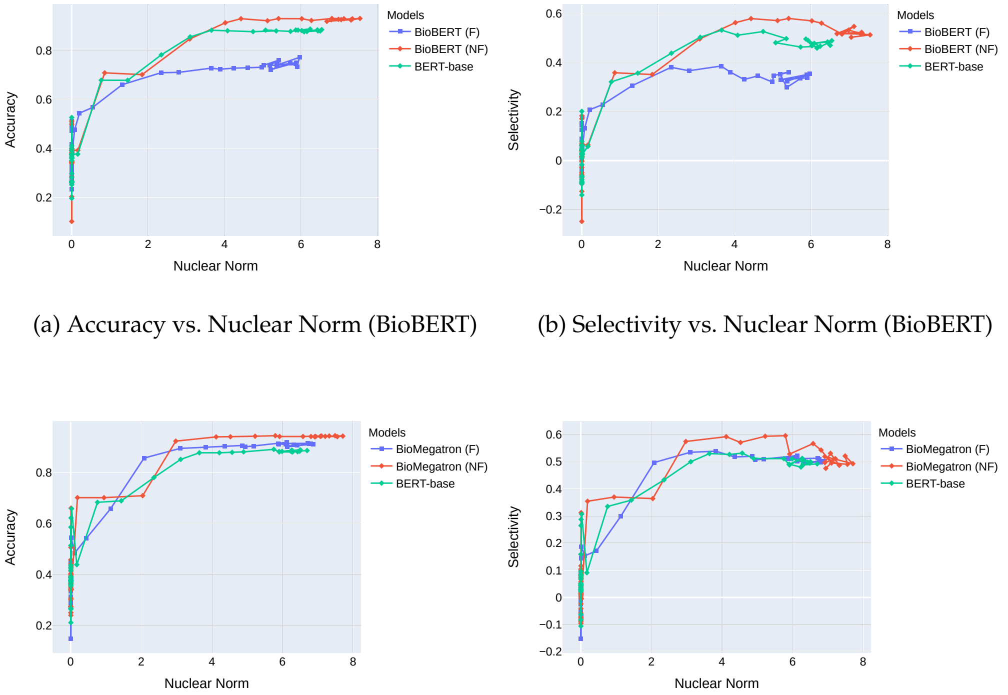
Variant	Gene	Diseases	Drugs	Clinical significance	BioBERT error	BioMegatron error	KNN error	Evidence level	Rating
EXON 2 MUTATION	KRAS	Pancreatic Cancer	Erlotinib and Gemcitabine	R	0.895	0.270	0.2	B	4
EXPRESSION	EGFR	Colorectal Cancer	Cetuximab	S/R	0.280	0.296	0.4	B	4
EXPRESSION	FOXP3	Breast Cancer	Epirubicin	S/R	0.153	0.776	0.6	B	4
EXPRESSION	HSPA5	Colorectal Cancer	Fluorouracil	S/R	0.845	0.608	0.4	B	4
EXPRESSION	PDCD4	Lung Cancer	Paclitaxel	S/R	0.954	0.939	0.4	B	4
EXPRESSION	AREG	Colorectal Cancer	Panitumumab	S/R	0.434	0.120	0.4	B	4
EXPRESSION	EREG	Colorectal Cancer	Panitumumab	S/R	0.345	0.202	0.6	B	4
ITD	FLT3	Acute Myeloid Leukemia	Sorafenib	S/R	0.418	0.355	0.6	B	4
K751Q	ERCC2	Osteosarcoma	Cisplatin	R	0.285	0.827	0.2	B	4
LOSS-OF-FUNCTION	VHL	Renal Cell Carcinoma	Anti-VEGF Monoclonal Antibody	R	0.074	0.360	0.8	B	4
MUTATION	KRAS	Colorectal Cancer	Cetuximab and Chemotherapy	R	0.067	0.021	0	B	4
MUTATION	SMO	Basal Cell Carcinoma	Vismodegib	R	0.062	0.039	0	B	4
OVEREXPRESSION	IGF2	Pancreatic Adenocarcinoma	Gemcitabine and Ganitumab	S/R	0.068	0.100	0.6	B	4
OVEREXPRESSION	ERBB3	Breast Cancer	Patritumab Deruxtecan	S/R	0.006	0.028	0.2	B	4
PML-RARA A216V	PML	Acute Promyelocytic Leukemia	Arsenic Trioxide	R	0.161	0.015	0.4	B	4
V600E	BRAF	Colorectal Cancer	Cetuximab and Encorafenib and Binimetinib	S/R	0.264	0.761	0.4	A	5

From a cancer precision medicine perspective, these significant misclassifications elicit the safety limitations of these models when considering clinical applications. In previous paragraphs we show that high error is expected for underrepresented relations, while here we demonstrate that transformers can fail even for well-known, strong evidence relations (RQ1).

3.3 Does the Fine-tuning Corrupt the Representation of Pre-trained Models?

3.3.1 Recognizing Entity Types from Representations of Pairs. Figure 5 presents the probing results for Task 1, with the left column containing the Accuracy results and the right column containing the Selectivity results. Selectivity was greater than zero for a control task containing random labels. For BioBERT, both accuracy and selectivity were higher for the non-fine-tuned models compared with the fine-tuned model. In fact, performance of the BERT (base) model was greater than that of the fine-tuned model for this task. This suggests that BioBERT loses some of the accuracy of background knowledge as a result of fine-tuning. This finding aligns with other works (Durrani, Sajjad, and Dalvi 2021; Merchant et al. 2020; Rajaee and Pilehvar 2021). For BioMegatron, performance of the fine-tuned model was slightly worse than the non-fine-tuned one, suggesting a similar behavior for BioMegatron, but in lower magnitude (RQ3).

3.3.2 Recognizing Entity Types from Representations of Quadruples. Figure 6 presents the probing results for Task 2, following the same task design as Task 1. Similar to Task 1, selectivity was greater than zero for a control task containing random labels, and BERT-base and BioBERT both had higher accuracy compared with fine-tuned BioBERT. For this task, we can observe minimal differences between the performance of the fine-tuned and non-fine-tuned versions of BioMegatron, which outperform BERT and BioBERT models. For probes with a lower value for their nuclear norm (i.e., less complex probes),



(a) Accuracy vs. Nuclear Norm (BioBERT) (b) Selectivity vs. Nuclear Norm (BioBERT)

(c) Accuracy vs. Nuclear Norm (BioMegatron) (d) Selectivity vs. Nuclear Norm (BioMegatron)

Figure 5
Probing results for models fine-tuned (F) on Task 1, together with the original (non fine-tuned) models (NF).

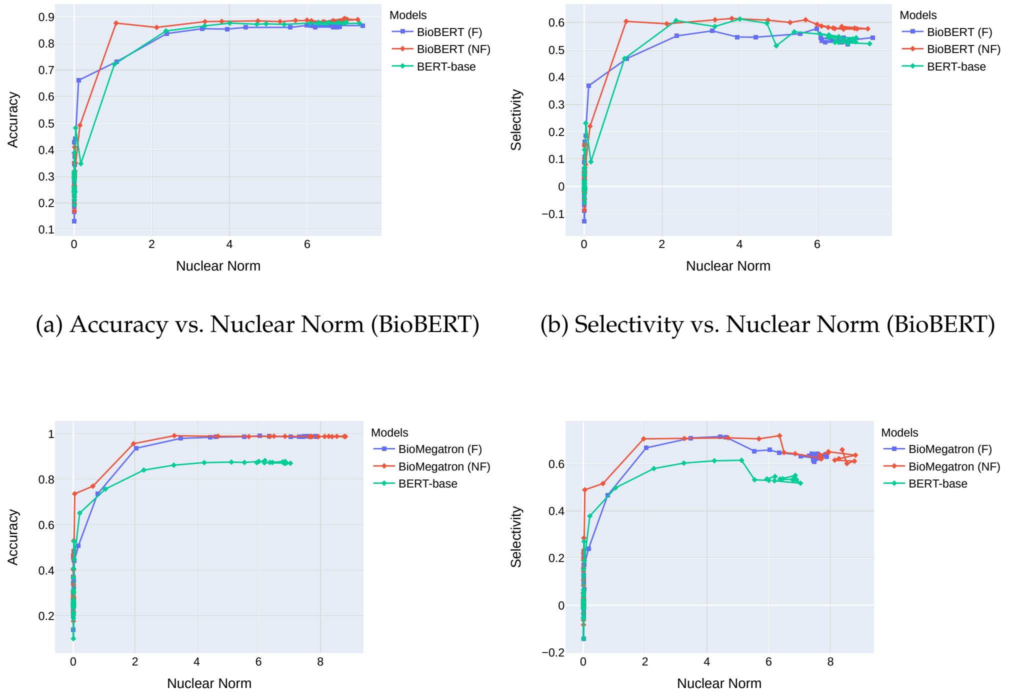
the performance of the original model is slightly better. However, the difference is nonexistent for more complex probes.

Probing results suggest that when fine-tuned for encoding complex *n*-ary relations (in Task 2), BioMegatron preserves more semantic information about entity type in the top layer than BioBERT (RQ3), as the difference in selectivity between fine-tuned (F) and non-fine-tuned (NF) versions is smaller (Figure 6). Both BioBERT and BioMegatron achieve acceptable selectivity (both F and NF), suggesting that they do encode semantic domain knowledge at entity level (RQ1).

3.4 How Much Biological Knowledge do Transformers Embed?

3.4.1 Biologically Relevant Clusters in Representations of Pairs. Based on clustering of BioBERT representations of variant-gene pairs in the balanced test set, and visual inspection of the clustermap and dendrogram, a cut point was applied that resulted in 5 clusters (Figure 7).

The dendrogram shows that cluster 5 (brown) contained 11 gene-variant pairs and remained separated from the other pairs until late in the merging process. The gene-variant pairs in this cluster involved only the PIK3CA and ERBB3 genes, and these genes did not occur in any other clusters. BioBERT classified all these pairs as true, with probability >0.60, although 4 of 11 pairs were false (Supplementary Table A.4).



(a) Accuracy vs. Nuclear Norm (BioBERT)

(b) Selectivity vs. Nuclear Norm (BioBERT)

(c) Accuracy vs. Nuclear Norm (BioMegatron)

(d) Selectivity vs. Nuclear Norm (BioMegatron)

Figure 6
Probing results for models fine-tuned on Task 2, following the same experiment design as Task 1, with fine-tuned (F) and non-fine-tuned (NF) models.

Interestingly, these genes participate in the same signaling pathways, including PI3K/AKT/mTOR.

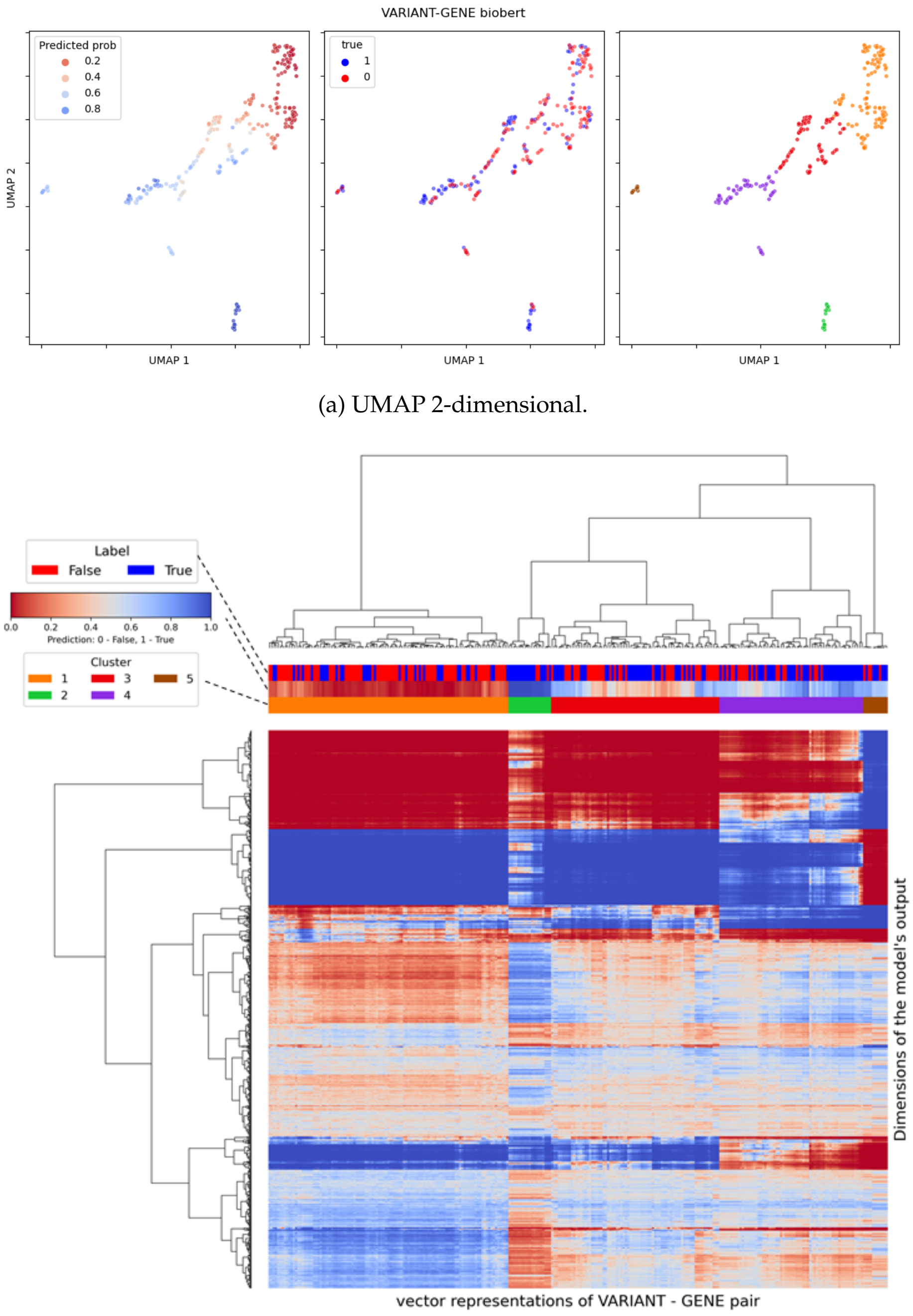
Cluster 2 (green) contained 19 gene-variant pairs; 14 of 19 variants in this cluster represented gene fusions, denoted by the notation *gene name - gene name*. All pairs were assigned as true, with probability >0.96, although 3 of 19 pairs were false (Supplementary Table A.5).

Following the clustering of BioMegatron representations on variant-gene pairs in the balanced test set, a cut point was applied that resulted in 6 clusters (Figure 8).

BioMegatron cluster 1 contained 16 of the 19 gene-variant pairs found in BioBERT cluster 2 (Supplementary Table A.5) as observed for BioBERT, BioMegatron determined all these pairs to be true with high confidence (probability >0.96).

Clustering analysis reveals an evident dataset artefact, that is, gene fusions as *gene name - gene name*, which is reflected in the representation. Both models encoded these fusions in a significantly different way compared with other pairs.

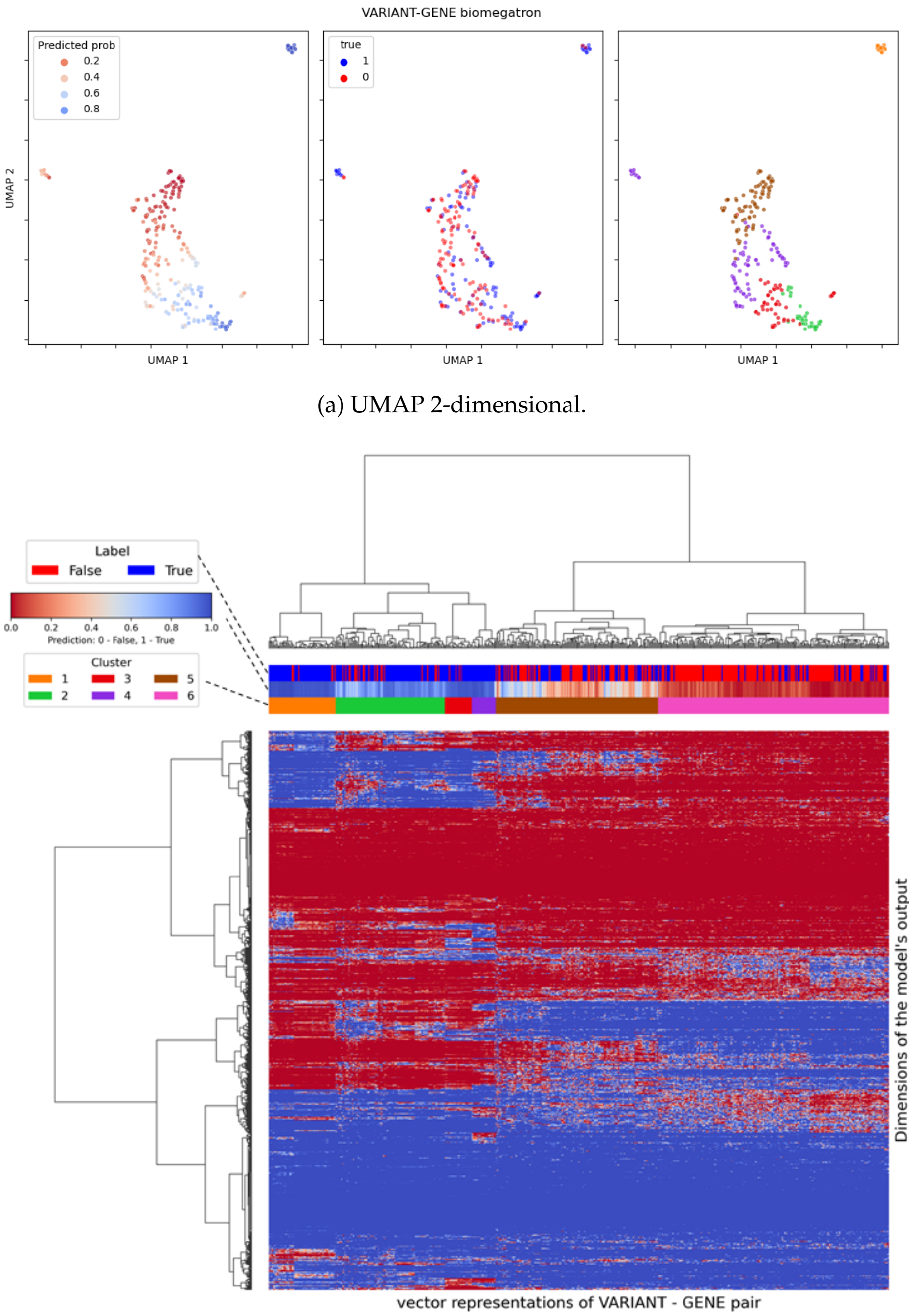
3.4.2 Biologically Relevant Clusters in Representations of Clinical Relations. Following clustering of BioMegatron representations of quadruples, a cut-off point was applied that resulted in 6 clusters (Figure 9).



(a) UMAP 2-dimensional.

(b) Clustermap based on Hierarchical Agglomerative Clustering.

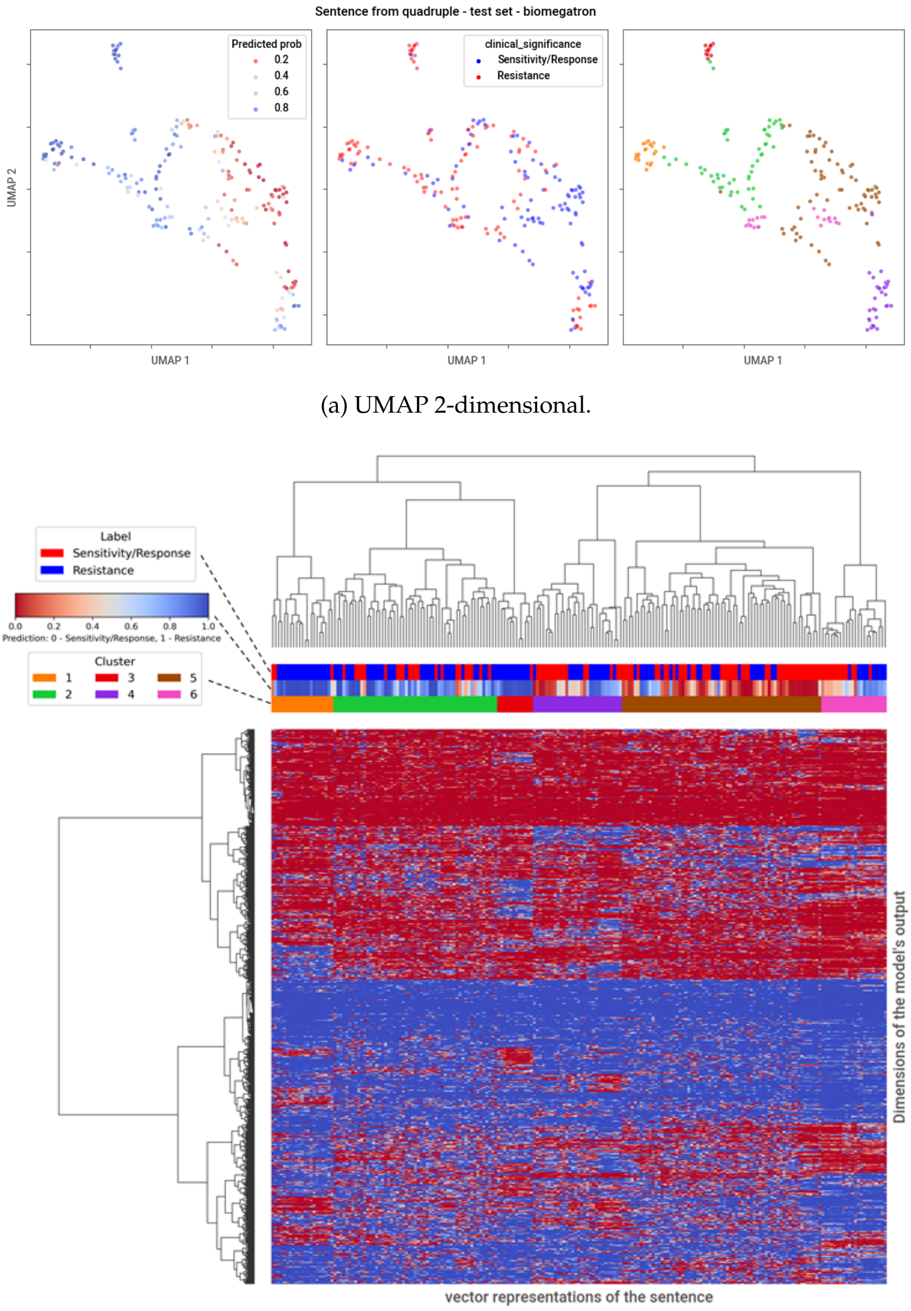
Figure 7
Representations of BioBERT output for variant-gene pairs in the balanced test set.



(a) UMAP 2-dimensional.

(b) Clustermap based on Hierarchical Agglomerative Clustering.

Figure 8
Representations of BioMegatron output for variant-gene pairs in the balanced test set.



(a) UMAP 2-dimensional.

(b) Clustermap based on Hierarchical Agglomerative Clustering.

Figure 9
Representations of BioMegatron output for quadruples in the balanced test set.

- Cluster 1 included 21 quadruples, all of which related to colorectal cancer. Most quadruples involved either BRAF, EGFR, or KRAS genes.
- Cluster 3 included 11 quadruples, all of which related to the drug vemurafenib. Most (9/11) related to melanoma, and 10 of 11 were associated with resistance.
- Cluster 4 included 30 quadruples, all of which related to the KIT gene, gastrointestinal stromal tumor, and either sunatinib or imatinib drugs; KIT was not associated with any other clusters.
- Cluster 6 included 22 quadruples, all of which related to the ABL gene and fusions with the BCR gene (denoted by *Variant BRCA-ABL*)

Similarly, 6 clusters were defined based on the BioBERT representations (Figure 10). Quadruples in BioBERT clusters were less homogeneous compared with those for the BioMegatron clusters. The two small clusters 5 and 6 are described in Supplementary Table A.6. Cluster 5 included 10 quadruples, involving 7 different genes, 7 diseases, and 6 drugs; cluster 6 included 11 quadruples, with 4 genes, 5 diseases, and 11 drugs; no clear pattern was evident in either cluster.

Clustering analysis reveals that representations encoded by fine-tuned BioMegatron form biologically meaningful clusters, in terms of gene-variant-disease-drug (RQ2). For BioBERT, the patterns are less apparent and may require deeper, more granular investigation (RQ3).

3.4.3 Entity Types Clusters in Fine-tuned Models. In this section, we investigated the clustering of the latent vectors. These vectors were also used for the probing task. Each vector represents one entity contextualized inside sentences from the test set (from both Task 1 and Task 2; more in Supplementary Methods).

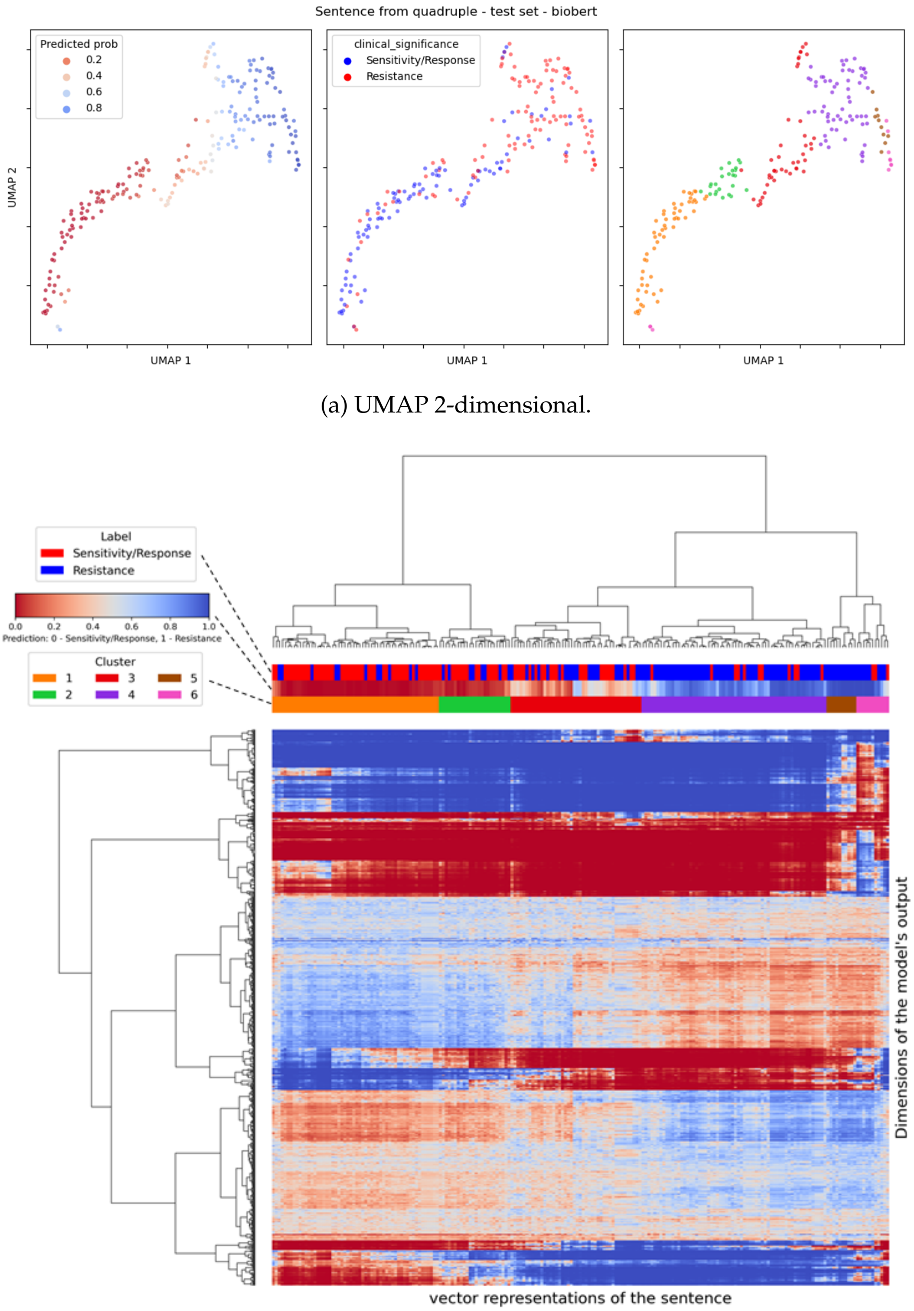
Results from HDBSCAN evaluation of UMAP representations are summarized in Table 7.

For Task 1, the non-fine-tuned transformer models clustered entities according to their type (Figure 11)—the average homogeneity of clusters was 0.940 for BioBERT, 0.911 for BioMegatron, and 0.883 for BERT. In contrast, clusters generated by the fine-tuned transformer models were less homogeneous (0.758 and 0.726 for BioMegatron and BioBERT, respectively)—this was observed across all types of entity-pairs.

For Task 2, clusters generated by the non-fine-tuned models were almost perfectly homogeneous (homogeneity >98.8%), except for cluster 5, consisting of both gene and variant entities (black dashed box in Figure 12).

However, for fine-tuned models, the majority of entities get projected closely under a 2D UMAP projection, similar to the findings in Rajaee and Pilehvar (2021) and Durrani, Sajjad, and Dalvi (2021). In fine-tuned BioBERT, drugs are projected to variants and some genes. As a result, a large cluster (5) with mixed entity types emerges. A similar type of clustering behavior is observed in the fine-tuned BioMegatron, showing one large cluster (2) containing portions of all types of entities.

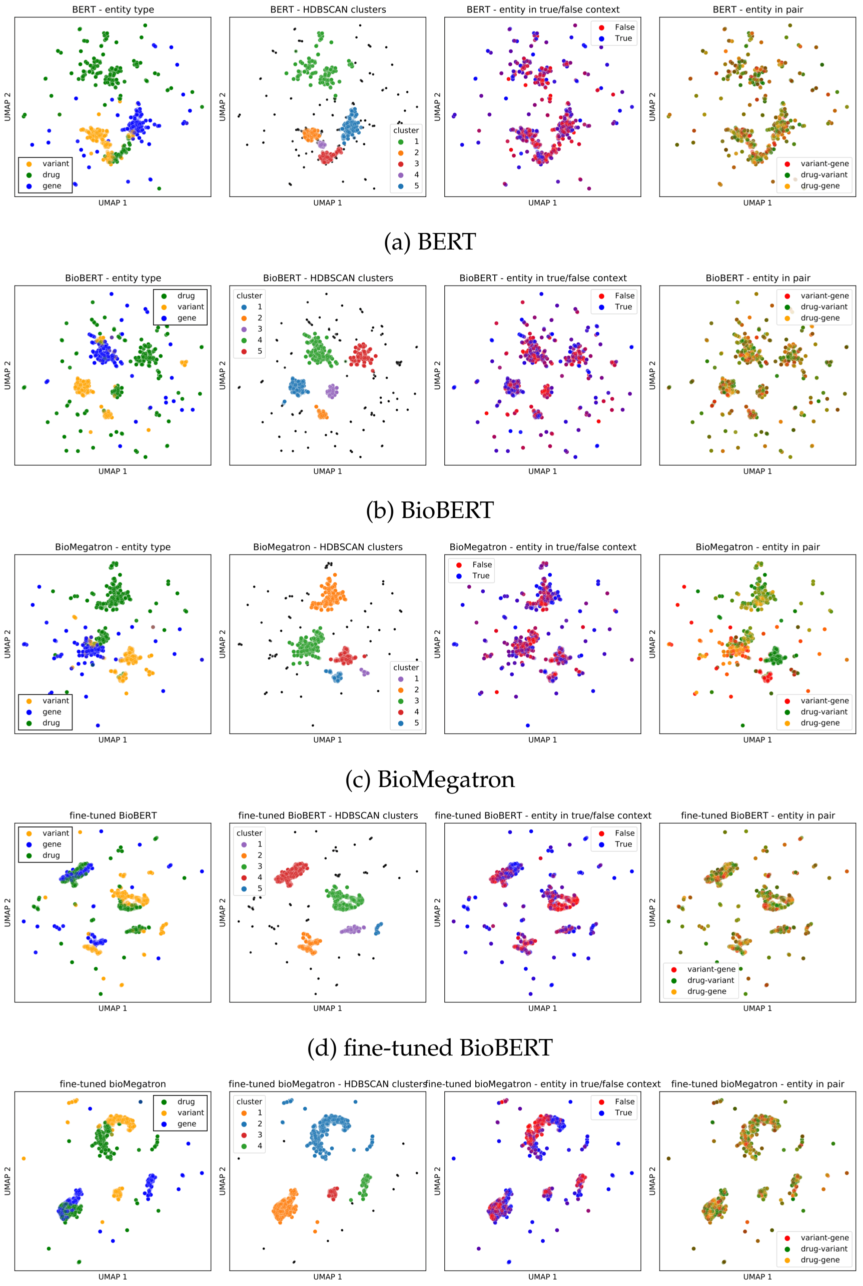
In all the 5 models, the representations do not group according to target labels in Task 1 nor Task 2. Homogeneity of clusters regarding true/false labels equals on average .570, and regarding "Sensitivity/Response"/"Resistance" .680. They are close to a random distribution of labels over clusters, because the labels proportions are 0.50 and 0.65, respectively.



(a) UMAP 2-dimensional.

(b) Clustermap based on Hierarchical Agglomerative Clustering.

Figure 10
Representations of BioBERT output for quadruples in the balanced test set.



(a) BERT

(b) BioBERT

(c) BioMegatron

(d) fine-tuned BioBERT

(e) fine-tuned BioMegatron

Figure 11
UMAP representation of entities from Task 1 used as input to Probing. In BERT, BioBERT, and BioMegatron, the clusters are homogeneous regarding the entity type (left plots). Fine-tuned models lose this property.

Table 7
Mean homogeneity in clusters.

Task	Model	Entity type (gene, variant, drug)	Target label pair type (True or False)	Pair type (d-g, g-v,d-v)
Task 1	**BERT**	0.883	0.553	0.471
	BioBERT	0.940	0.572	0.478
	BioMegatron	0.911	0.548	0.708
	FT BioBERT	0.726	0.638	0.488
	FT BioMegatron	0.758	0.538	0.474
		gene, variant, drug, disease	**Sensitivity/Response or Resistance**	
Task 2	**BERT**	.996; .599 in #5 (genes and variants)	0.695	
	BioBERT	.998; .793 in #5 (genes and variants)	0.679	
	BioMegatron	1.0; .773 in #5 (genes and variants)	0.656	
	FT BioBERT	.990; .514 in #5 (drugs, variants, genes)	0.680	
	FT BioMegatron	.380 in large cluster #2	0.691	

Clustering analysis and homogeneity evaluation confirm that both BioBERT and BioMegatron encode fundamental semantic knowledge at the entity level, in this case genes, variants, drugs, and diseases. However, a significant part of the latent semantics is changed during fine-tuning, which is particularly apparent for a more complex Task 2 (RQ1).

4. Discussion

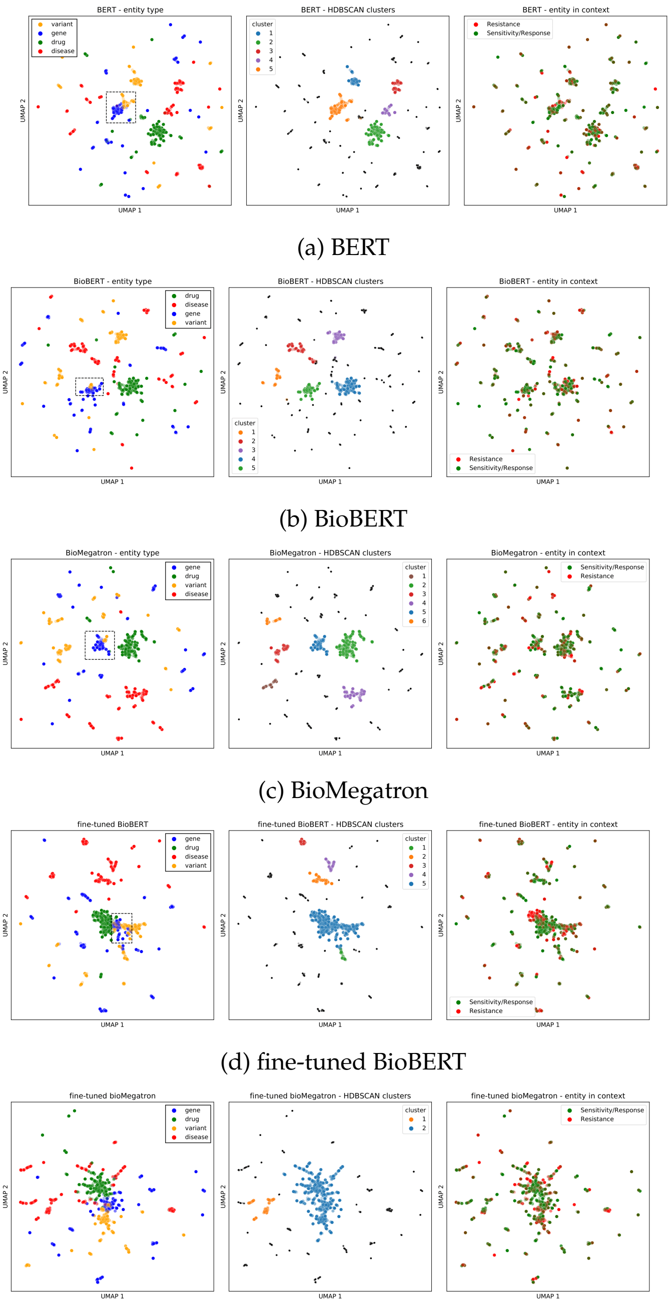
4.1 Summary of Main Findings

In this study we performed a detailed analysis of the embeddings of biological knowledge in transformer-based neuro-language models using a cancer genomics knowledge base.

First, we compared the performance between biomedical fine-tuned transformers (BioBERT and BioMegatron) and a naive simple classifier (KNN) for two specific classification tasks. Specifically, these tasks aimed to determine whether each transformer model captures biological knowledge about: pairwise associations between genes, variants, drugs, and diseases (Task 1), and the clinical significance of relationships between gene variants, drugs, and diseases (Task 2).

The hypothesis under test was that transformers would show better performance compared with a naive classifier, eliciting the role of the pre-trained component of the model (RQ4). Results for both tasks support this hypothesis. For Task 1, both BioBERT and BioMegatron outperformed the naive classifier for distinguishing true versus false associations between pairs of biological entities. Similarly, for Task 2, both transformer models outperformed the naive classifier for predicting the clinical significance of quadruples of entities. For Task 2, the transformer models achieved an acceptable performance (AUC > 0.8), although performance in Task 1 was lower (AUC approx. 0.6).

We highlighted the need for addressing the role of dataset imbalance within the assessment of embeddings (RQ4). Specifically, in our analysis, we found significant differences between AUCs for the imbalanced and balanced test sets. Furthermore, we found significant correlations between the classification error and imbalance for



(a) BERT

(b) BioBERT

(c) BioMegatron

(d) fine-tuned BioBERT

(e) fine-tuned BioMegatron

Figure 12
UMAP representation of entities from Task 2: (left) entity types; (center) clusters from HDBSCAN; (right) target label in classification task. Dashed box corresponds to entities from quadruples, in which variant entity contains the gene entity name. Representations from non fine-tuned models form more distinctive clusters, more homogeneous in terms of entity type.

individual entities. Similarly, the error is associated with a co-occurrence bias (within the corpus based on the biomedical literature): That is, in Task 1: a true pair that occurs in the literature multiple times is more likely to be classified as true, compared to pairs that occur less frequently.

Second, we used probing methods to inspect the consistency of the representation for each type of biological entity, and we compared pre-trained versus fine-tuned models (RQ1, RQ2). More specifically, we determined the performance of each model in classifying the type (gene, variant, drug, or disease) of entities based on their representation in the model via accuracy and selectivity. We quantified how much semantic structure is lost in fine-tuning, and how biologically meaningful is the remaining. For BioBERT, both accuracy and selectivity were lower for the fine-tuned models compared with the base models, including BERT-base, which is not specific for the medical/biological domain. For BioMegatron, there was only a slight difference in performance between the fine-tuned and non-fine-tuned models. Probing experiments demonstrated that fine-tuned BioMegatron better preserves the pre-trained knowledge when compared with fine-tuned BioBERT (RQ3).

Finally, we provide a qualitative and quantitative analysis of clustering patterns of the embeddings, using UMAP, HDBCAN, and HAC. We show that entities of the same type cluster together, and that this is more pronounced for the non-fine-tuned models compared with the fine-tuned models (RQ1, RQ2). A cluster analysis revealed biological meaning. For instance, we found a cluster with the vast majority of sentences related to resistant response to vemurafenib in melanoma treatment. Another example: a cluster specific to KIT gene, gastrointestinal stromal tumor (GIST), sunatinib, and imatinib. According to domain-expert knowledge, imatinib, a KIT inhibitor, is a standard first-line treatment for metastatic GIST, whereas sunatinib is the second option.

4.2 Strengths and Limitations

Strengths:

- We have used the CIViC database as the basis of our analysis. We consider this to be a high-quality dataset, because: (i) it entails a set of relationships curated by domain experts; (ii) most relationships include a confidence score; (iii) it has been developed for a closely related use case, namely, to support clinicians in the evaluation of the clinical significance of variants.
- We use state-of-the-art, bidirectional transformer models trained on a biomedical text corpus (PubMed abstracts) containing over 29M articles and 4.5B words.
- Patterns in representations are investigated using 2 methods (UMAP and HAC), instead of relying on a single method. Clusters are thoroughly described and quantified using homogeneity metrics.
- We include input from domain experts in data preparation, evaluation, and interpretation of results. It allows for: (i) the correct filtering of evidence; (ii) assessment of the relevance of investigated biomedical relations; and (iii) granular analysis of clusters in search for biological meaning.

Limitations:

- The distribution of entities among the dataset has the potential to lead to overfitting. For example, if the EGFR gene is over-represented among true gene-drug pairs compared with other genes, a model could classify gene-drug pairs solely on whether the gene = EGFR and performs better than expected. Indeed, the distributions of entities in our dataset were highly right-skewed (Pareto distribution). This issue refers to the well-known imbalance problem, which leads to an incorrect performance evaluation. Although we applied a balancing procedure, it is infeasible to create a perfectly balanced dataset.
- In CIViC, drug interaction types can be either *combination*, *sequential*, or *substitutes*. In the generation of evidence sentences, we did not account for that variation, which for sentences with multiple drugs may slightly alter the representation of clinical significance in the model.
- In CIViC, there are evidence items that claim contradicting clinical significance for the same relation. We excluded them from our dataset; however, their future investigation would be of relevance.

4.3 Related Work

4.3.1 Supporting Somatic Variant Interpretation in Cancer. There is a critical need to evaluate the large amount of relevant variant data generated by tumor next generation sequencing analyses, which predominantly have unknown significance and complicate the interpretation of the variants (Good et al. 2014). One of the ways to streamline and standardize cancer curation data in electronic medical records is to use the Web resources from the CIViC curatorial platform (Danos et al. 2018)—an open source and open access CIViC database, built on community input with peer-reviewed interpretations, already proven to be useful for this purpose (Barnell et al. 2019). The authors used the database to develop the Open-sourced CIViC Annotation Pipeline (OpenCAP), providing methods for capturing variants and subsequently providing tools for variant annotation. It supports scientists and clinicians who use precision oncology to guide patient treatment. In addition, Danos et al. (2019) described improvements at CIViC that include common data models and standard operating procedures for variant curation. These are to support a consistent and accurate interpretation of cancer variants.

Clinical interpretation of genomic cancer variants requires highly efficient interoperability tools. Evidence and clinical significance of the CIViC database was used in a novel genome variation annotation, analysis, and interpretation platform, the TGex (the Translational Genomics expert) (Dahary et al. 2019). By providing access to a comprehensive KB of genomic annotations, the TGex tool simplifies and speeds up the interpretation of variants in clinical genetics processes. Furthermore, Wagner et al. (2020) provided CIViCpy, an open-source software for extracting and inspection of records from the CIViC database. The delivery of CIViCpy enables the creation of downstream applications and the integration of CIViC into clinical annotation pipelines.

4.3.2 Text-mining Approaches using CIViC. The development of guidelines (Li et al. 2017) for the interpretation of somatic variants, which include complexity of multiple dimensions of clinical relevance, allow for a better standardization of the assessment of cancer variants in the oncological community. In addition, they can enhance the rapidly growing use of genetic testing in cancer, the results of which are critical to accurate prognosis and treatment guidance. Based on the guidelines, He et al. (2019) demonstrated computational approaches to take pre-annotated files and to apply criteria for the assessment of the clinical impact of somatic variants. In turn, Lever et al. (2018) proposed a text-mining approach to extract the data on thousands of clinically relevant biomarkers from the literature; and, using a supervised learning approach, they constructed a publicly accessible KB called CIViCmine. They extracted key parts of the evidence item, including: cancer type, gene, drug (where applicable), and the specific evidence type. The CIViCmine contains over 87K biomarkers associated with 8k genes, 337 drugs, and 572 cancer types, representing more than 25k abstracts and almost 40k full-text publications. This approach allowed counting the number of mentions of specific evidence items—cancer type, gene, drug (where applicable), and the specific evidence type—in PubMed abstracts and PubMed Central Open Access full-text articles and comparing them with the CIViC knowledge base. A similar approach was previously proposed by Singhal, Simmons, and Lu (2016), who proposed a method to automate the extraction of disease-gene-variant triples from all abstracts in PubMed related to a set of ten important diseases.

Ševa, Wackerbauer, and Leser (2018) developed an NLP pipeline for identifying the most informative key sentences in oncology abstracts by assessing the clinical relevance of sentences implicitly based on their similarity to the clinical evidence summaries in the CIViC database. They used two semi-supervised methods: transductive learning from positive and unlabeled data and self-training by using abstracts summarized in relevant sentences as unlabeled examples. Wang and Poon (2018) developed deep probabilistic logic as a general framework for indirect supervision, by combining probabilistic logic with deep learning. They used existing KBs with hand-curated drug-gene-mutation facts: the Gene Drug Knowledge Database (GDKD) (Dienstmann et al. 2015) and CIViC, which together contained 231 drug-gene-mutation triples, with 76 drugs, 35 genes, and 123 mutations. Recently, Jia, Wong, and Poon (2019) proposed a novel multiscale neural architecture for document-level *n*-ary relation extraction, which combines representations learned over various text spans throughout the document and across the subrelation hierarchy. For distant supervision, they used CIViC, GDKD (Dienstmann et al. 2015), and OncoKB (Chakravarty et al. 2017) KBs.

This section summarized the usage of the CIViC database in the development of NLP pipelines as well as approaches to using NLP with cancer-related literature. However, we did not find any study using cancer genomic databases (such as CIViC) to investigate the semantic characterization of biomedically trained neural language models.

4.4 Model Bias Caused by the Unbalanced Training Set

Our findings regarding the bias in the models caused by the unbalanced dataset align with findings in the previous works. McCoy, Pavlick, and Linzen (2019) show that NLI models rely on adopted heuristics from statistical regularities in training sets, which are valid for frequent cases, but invalid for less-frequent ones. This results in low performance in HANS (Heuristic Analysis for NLI Systems), which is attributed to invalid heuristics rather than deeper understanding of language. Gehman et al. (2020) recommend a careful examination of the dataset due to possible toxic, biased, or otherwise degenerate behavior of language models. Similarly, in Nadeem, Bethke, and Reddy (2021), a strong stereotypical bias was reported in pre-trained BERT, GPT2, ROBERTA, and XLNET. Distribution in the dataset affects the performance (Zhong, Friedman, and Chen 2021), leading to overestimation of model's inference and deeper understanding of language (Gururangan et al. 2018; Min et al. 2020). In our study, we confirmed the importance of integrating a balancing strategy for embedding studies.

4.5 Evaluation of Semantic Knowledge in Transformer-based Models

Fine-tuning distorts the original distribution within pre-trained models: Higher layers are more adjusted to the specific task and lower layers retain their representation (Durrani, Sajjad, and Dalvi 2021; Merchant et al. 2020). Although fine-tuning affects top layers, it is interpreted to be a conservative process and there is no catastrophic forgetting of information in the entire model (Merchant et al. 2020). However, it has been reported that fine-tuned models can fail to leverage syntactic knowledge (McCoy, Pavlick, and Linzen 2019; Min et al. 2020) and rely on pattern matching or annotation artifacts (Gururangan et al. 2018; Jia and Liang 2017). It is expected that fine-tuned representations will differ significantly from the pre-trained ones (Rajaee and Pilehvar 2021) and architectures will deliver different representations of background and linguistic knowledge (Durrani, Sajjad, and Dalvi 2021).

Probing proved to be an effective method to investigate what information is encoded in the model and how it influences the output (Adi et al. 2017; Belinkov 2021; Hupkes, Veldhoen, and Zuidema 2018). In recent work, probing was used to verify the model's understanding of scale and magnitude (Zhang et al. 2020) or whether a model can reflect an underlying foundational ontology (Jullien, Valentino, and Freitas 2022). In Jin et al. (2019), probing was used to determine what additional information is carried intrinsically by BioELMo and BioBERT.

Recent work on applying language models to biomedical tasks are: MarkerGenie - identifies bioentity relations from texts and tables of publications in PubMed and PubMed Central (Gu et al. 2022); ScispaCy model—relevant for drug discovery, aims to cover disease-gene interactions significant from pharmacological perspective (Qumsiyeh and Jayousi 2021); and DisKnE—aims to evaluate pre-trained language models about the disease knowledge (Alghanmi, Espinosa Anke, and Schockaert 2021). In Vig et al. (2021), transformers are used for better understanding working mechanisms in proteins. Biomedical transformers has demonstrated to be highly effective in biomedical NLI tasks (Jin et al. 2019), but safety and validation of their usage is still an under-explored area. A promising direction of future research is to integrate structured knowledge into the models (Colon-Hernandez et al. 2021; Yuan et al. 2021).

5. Conclusions

In this work we performed a detailed analysis of fundamental knowledge representation properties of transformers, demonstrating that they are biased toward more frequent statements. We recommend accounting for this bias in biomedical applications. In terms of the semantic structure of the model, BioMegatron shows more salient biomedical knowledge embedding than BioBERT, as the representations cluster into more interpretable groups and the model better retains the semantic structure after fine-tuning.

We also investigated the representation of entities both in base and fine-tuned models via probing (Ferreira et al. 2021). We found that the fine-tuned models lose the general structure acquired at the pre-training phase and degrade the models with regard to cross-task transferability.

We found biologically relevant clusters, such as genes and variants that are present in the same biological pathways. Considering the vectors used in probing, we found that the distances are associated with entity type (gene, variant, drug, disease). However, the fine-tuning renders the representations internally more inconsistent, which was quantified by the evaluation of clusters' homogeneity. We investigated whether the models can capture the quality of evidence and found that they did not perform significantly better for well-known relations. Even for eminent clinical quadruples /statements, the models misclassified the clinical significance (whether sensitive or resistant to treatment), highlighting the limitations of contemporary neural language models.

Appendix

Supplementary Methods

Downloading the Data. The data was downloaded via CIViC API using the following queries:

- 'https://civicdb.org/api/variants/XYZ where XYZ is a 'variant id'

Variant id can be found in the list of all available variants:

- 'https://civicdb.org/api/variants?count=2634'

Balancing the Test Set. We excluded the imbalanced pairs/quadruples from the *test set* in order to create a *balanced test set* according to the following procedure.

First, we give two definitions of imbalanced entity, followed by the definitions of imbalanced pair and imbalanced quadruple. We define 2 types of imbalanced entity, true-imbalanced entity and false-imbalanced entity. An entity is considered as true-imbalanced entity if it meets the following criteria:

Over 70% of training pairs/quadruples containing this entity are true.

In reverse, the criteria for false-imbalanced entity is:

Less than 30% of training pairs/quadruples containing this entity are true.

Based on the definition of true-imbalanced entity and false-imbalanced entity, we can define imbalanced pair as:

Either one element of the pair is true-imbalanced entity and the other element is not false-imbalanced entity, or one element of the pair is false-imbalanced entity and the other element is not true-imbalanced entity.

Similar to the imbalanced pair definition, the imbalanced quadruple can be defined as the following:

Either one element of the quadruple is true-imbalanced entity and no other element is false-imbalanced entity, or one element of the quadruple is false-imbalanced entity and no other element is true-imbalanced entity.

Note, for quadruples *true|false* should be replaced with *sensitivity/response | resistance*.

The key intuition of the balancing is to remove the bias that some pairs/quadruples containing specific entities are almost always true (or false). Removing the bias allows us to compare the test results more fairly.

Note, we apply the balancing only to the pairs that are in the test set due to the following reasons. First, the training set after balancing would be too small. This is a common drawback when trying to balance the dataset without oversampling, and remains an open challenge for real world datasets. Second, in a Machine Learning pipeline the test set should be isolated at the very beginning, before any exploratory data analysis or feature engineering. As the balancing aims for better performance evaluation, we must consider ratios in the test set, but this information should not leak to any activity done on the training set. However, we do exclude pairs (from the test set) also looking at the occurrence in the training set, as we want to mitigate the possible impact of overfitting during training. Balancing the test set left us with 38% to 53% of pairs in the balanced test set.

Transformers. Because both BioBERT and BioMegatron models allow 512 tokens in the input sequences, which is far longer than the input sequences we defined, we do not consider the sentence truncation in this work.

Transformers models have multiple layers, with BioBERT having 12 layers and BioMegatron having 24 layers. One output is generated for each layer, but the output of the last layer is generally used as the output of transformers models since we want to fully use the neural network connection architecture through multiple layers. Multiple vectors are contained in the transformers model's last layer output, where each vector represents one input token in input sequence, respectively. A total of 512 vectors are contained in last layer outputs of both BioBERT and BioMegatron because they both allow the same vector size in the input sequences. Because we do a sentence-level classification task in this work and the first token of each input sequence is "[CLS]", we use the vector output of "[CLS]" token (first token) in the sequence as pooled output vector of transformers models. Although there are two major output vector pooling methods, either obtaining the first token vector or averaging the vector of all tokens, we choose to use the former, since it is used in most sentence level transformers' pre-training tasks such as sentence classification and next sentence prediction. The BioBERT model uses a 768-dimension output vector, while BioMegatron uses 1,024 dimensions.

$$V_r = f_\theta^{TRF}(seq)[0] \tag{1}$$

As shown in Equation (1), f_θ^{TRF} is last-layer output function of the transformers model, *seq* is input sequence of the tranformers model. We use first token's output vector,$f_\theta^{TRF}(seq)[0]$, as pooled output of the sequence, V_r.

For training purposes, we stack a classification layer on top of transformers models. For the Task 1, we need to classify the true and false pairs. We stack a fully connected N-to-1 linear layer and use sigmoid activation to constrain the output value from 0 to 1. Binary cross entropy loss function is used for true/false classification.

For Task 2, we need to classify the multiple clinical significance categories for each input sentence. There are 2 clinical significance categories, "Sensitivity/Response" and "Resistance" while more categories could be added in a future dataset. We use N-to-2 linear layer and softmax activation to get one probability score for each category; then cross entropy loss function is used for model parameter optimization.

Clustering the Probing Input. In total, 4,500 and 3,572 vectors were obtained from the pairs and quadruples test set, respectively (see Task 1 and Task 2). Vectors for pairs were aggregated from 3 fine-tuned models trained for each pair type. Each vector consists of 768 for BERT and BioBERT, and 1,024 dimensions for BioMegatron. We used UMAP for dimensionality reduction and HDBSCAN clustering algorithm to identify patterns in an unsupervised manner.

Supplementary Tables

Table A.1
Spearman correlations between the classification error and the number of pairs in the training set where an entity occurs. For example, for BioBERT, there is a significant negative correlation between the number of drug-gene pairs in the training set where a drug entity occurs and the classification error.

Pair type	Model	Entity	True/false pair vs. error	Spearman correlation	p-val	Significance
DRUG - VARIANT	BioBERT	DRUG	True	−0.75	0.0000	***
			False	0.73	0.0000	***
		VARIANT	True	0.23	0.0010	*
			False	0.06	0.3825	ns
	BioMegatron	DRUG	True	−0.69	0.0000	***
			False	0.68	0.0000	***
		VARIANT	True	0.15	0.0382	*
			False	0.05	0.4591	ns
DRUG - GENE	BioBERT	DRUG	True	−0.42	0.0000	***
			False	0.27	0.0000	***
		GENE	True	−0.55	0.0000	***
			False	0.41	0.0000	***
	BioMegatron	DRUG	True	−0.51	0.0000	***
			False	0.31	0.0000	***
		GENE	True	−0.48	0.0000	***
			False	0.45	0.0000	***
VARIANT - GENE	BioBERT	VARIANT	True	−0.30	0.0004	***
			False	0.05	0.5646	ns
		GENE	True	−0.47	0.0000	***
			False	0.61	0.0000	***
	BioMegatron	VARIANT	True	−0.29	0.0007	***
			False	0.07	0.4023	ns
		GENE	True	−0.47	0.0000	***
			False	0.63	0.0000	***

Table A.2
Examples of variant entities whose representations appear in the same cluster (5) as gene representations for all 3 base models (BERT, BioBERT, and BioMegatron) according to UMAP transformation. Variant representations stem from sentences where the variant entity contains a gene name.

Variant entry	Sentence constructed from quadruple
IGH-CRLF2	IGH-CRLF2 of CRLF2 identified in B-lymphoblastic leukemia/lymphoma, BCR-ABL1–like is associated with ruxolitinib
ZNF198-FGFR1	ZNF198-FGFR1 of FGFR1 identified in myeloproliferative neoplasm is associated with midostaurin
SQSTM1-NTRK1	SQSTM1-NTRK1 of NTRK1 identified in lung non-small cell carcinoma is associated with entrectinib
CD74-ROS1 G2032R	CD74-ROS1 G2032R of ROS1 identified in lung adenocarcinoma is associated with DS-6501b
BRD4-NUTM1	BRD4-NUTM1 of BRD4 identified in NUT midline carcinoma is associated with JQ1
KIAA1549-BRAF	KIAA1549-BRAF of BRAF identified in childhood pilocytic astrocytoma is associated with trametinib
TPM3-NTRK1	TPM3-NTRK1 of NTRK1 identified in spindle cell sarcoma is associated with larotrectinib
KIAA1549-BRAF	KIAA1549-BRAF of BRAF identified in childhood pilocytic astrocytoma is associated with vemurafenib and sorafenib
CD74-NRG1	CD74-NRG1 of NRG1 identified in mucinous adenocarcinoma is associated with afatinib
EWSR1-ATF1	EWSR1-ATF1 of EWSR1 identified in clear cell sarcoma is associated with crizotinib

Table A.3
AUCs and Brier scores for the balanced test set stratified by evidence level. KNN performs significantly worse for evidence level D compared with the transformers (**bold**).

Evidence level	AUC			Brier score loss		
	B	C	D	B	C	D
BioBERT	0.683	0.900	0.812	0.254	0.148	0.202
BioMegatron	0.703	0.939	0.816	0.274	0.103	0.178
KNN	0.682	0.910	**0.705**	0.231	0.122	**0.228**

Table A.4
Pairs in cluster 5 in BioBERT representations containing only PIK3CA and ERBB3 genes.

Cluster #5 (brown)	Variant	Gene	TRUE	Predicted probability
1	R93W	PIK3CA	1	0.678
2	H1047R	PIK3CA	1	0.664
3	D350G	PIK3CA	1	0.666
4	G1049R	PIK3CA	1	0.624
5	H1047L	PIK3CA	1	0.673
6	R103G	ERBB3	1	0.773
7	E545G	PIK3CA	1	0.657
8	E281K	ERBB3	0	0.756
9	C475V	ERBB3	0	0.780
10	F386L	PIK3CA	0	0.680
11	D816E	ERBB3	0	0.776

Table A.5
Cluster 2 for BioBERT and cluster 1 for BioMegatron, where the variant entities contain gene names.

#	Variant	Gene	True/false	cluster # in BioBERT HAC	cluster # in BioMegatron HAC
1	D1930V	ATM	1	2	other
2	M2327I	ATM	0	2	other
3	R777FS	ATM	1	2	other
4	ZKSCAN1-BRAF	BRAF	1	2	1
2	IGH-CRLF2	CRLF2	1	2	1
6	DEK-AFF2	DEK	1	2	1
7	EWSR1-ATF1	EWSR1	1	2	1
8	FGFR2-BICC1	FGFR2	1	2	1
9	ATP1B1-NRG1	NRG1	1	2	1
10	CD74-NRG1	NRG1	1	2	1
11	NRG1	NRG1	1	2	1
12	ETV6-NTRK2	NTRK1	0	2	1
13	LMNA-NTRK1	NTRK1	1	2	1
14	SQSTM1-NTRK1	NTRK1	1	2	1
12	ETV6-NTRK2	NTRK2	1	2	1
16	NTRK1-TRIM63	NTRK2	0	2	1
17	RCSD1-ABL1	RCSD1	1	2	1
18	TFG-ROS1	ROS1	1	2	1
19	UGT1A1*60	UGT1A1	1	2	1

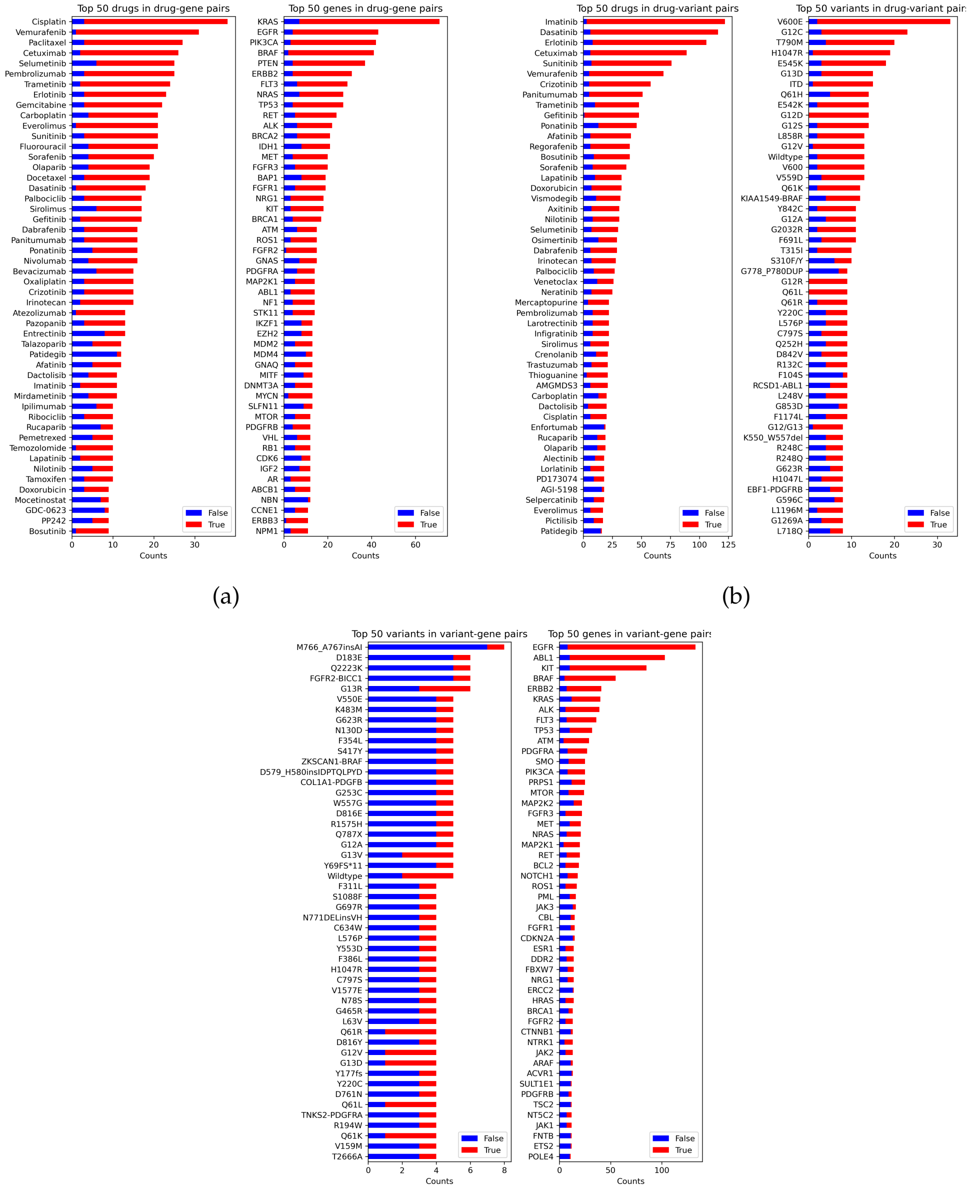
Table A.6
BioBERT quadruples from clusters #5 and #6. No obvious patterns. R stands for Resistance and S/R is for Sensitivity/Response.

E17K	AKT3	Melanoma	Vemurafenib	R	5
ALK FUSION G1202R	ALK	Cancer	Alectinib	R	5
D835H	FLT3	Acute Myeloid Leukemia	Sorafenib	R	5
G12D	KRAS	Colorectal Cancer	Panitumumab	R	5
G12R	KRAS	Colorectal Cancer	Panitumumab	R	5
K117N	KRAS	Clear Cell Sarcoma	Vemurafenib	R	5
OVEREXPRESSION	PIK3CA	Melanoma	Vemurafenib	R	5
LOSS	PTEN	Melanoma	Vemurafenib	R	5
M237I	TP53	Glioblastoma	AMGMDS3	R	5
L3 DOMAIN MUTATION	TP53	Breast Cancer	Tamoxifen	R	5
T790M	EGFR	Lung Non-small Cell Carcinoma	Cetuximab and Panitumumab and Brigatinib	S/R	6
Y842C	FLT3	Acute Myeloid Leukemia	Lestaurtinib	S/R	6
ITD D839G	FLT3	Acute Myeloid Leukemia	Pexidartinib	R	6
ITD I687F	FLT3	Acute Myeloid Leukemia	Sorafenib	R	6
D839N	FLT3	Acute Myeloid Leukemia	Pexidartinib	R	6
ITD Y842C	FLT3	Acute Myeloid Leukemia	Sorafenib and Selinexor	R	6
G12D	KRAS	Melanoma	Vemurafenib	R	6
G12S	KRAS	Lung Non-small Cell Carcinoma	Erlotinib	R	6
G12V	KRAS	Colon Cancer	Regorafenib	S/R	6
G12V	KRAS	Lung Cancer	Gefitinib	R	6
E545G	PIK3CA	Melanoma	Vemurafenib	R	6

Table A.7
Homogeneity in clusters obtained from 2-dimensional UMAP representation using HDBSCAN algorithm.

# cluster	BERT	BioBERT	BioMegatron
1	99.7% variant	99.6% variant	100% disease
2	100% drug	100% disease	100% drug
3	100% disease	99.7% variant	100% variant
4	98.8% disease	99.7% drug	100% disease
5	59.9% gene, 40.1% variant	79.3% gene, 20.7% variant	77.3% gene, 20.0% variant
6			100% variant

Supplementary Figures

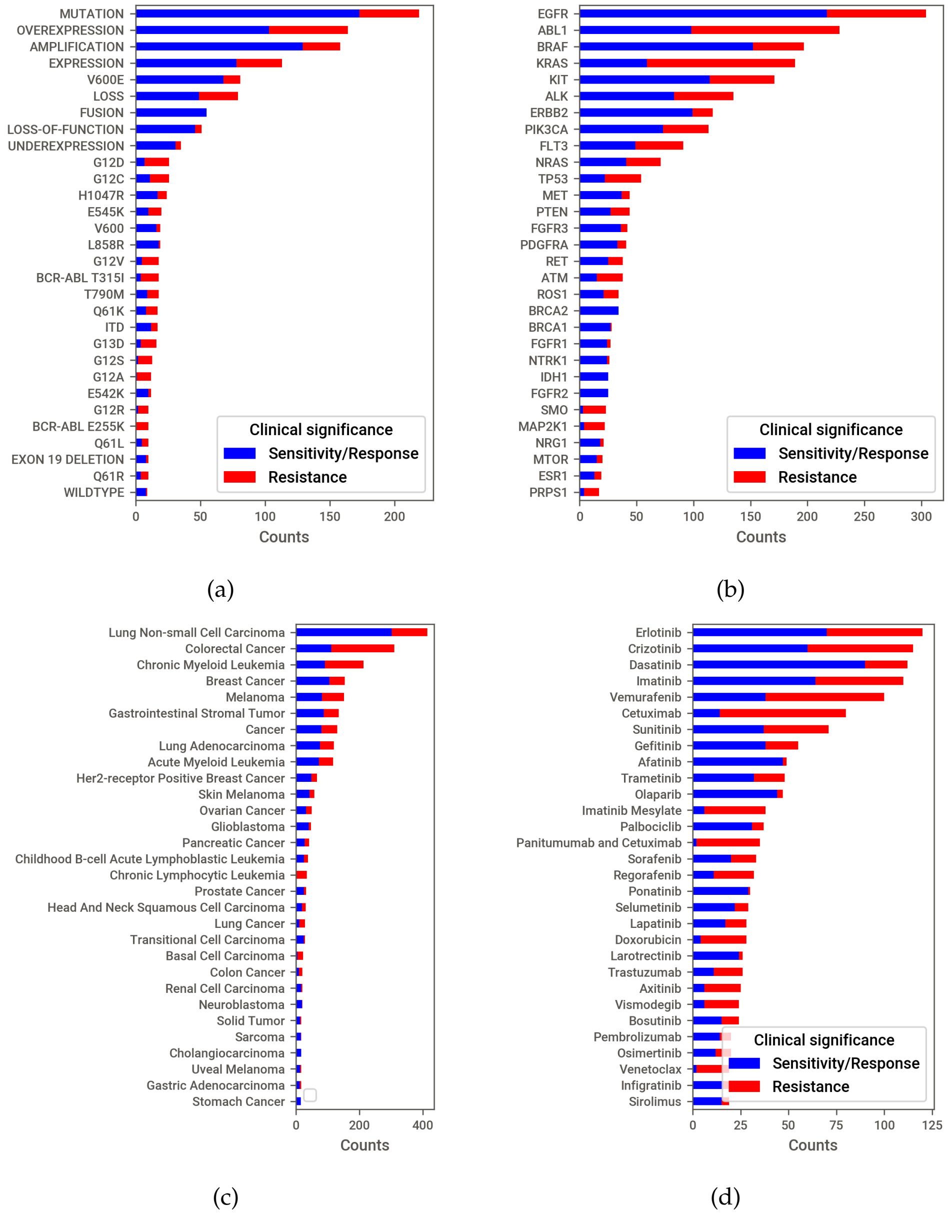


(a)

(b)

(c)

Figure A.1
Top 50 pairs in the dataset from Task 1. Most frequent entities occur mostly in true pairs, except for variants in variant-gene pairs.



(a)

(b)

(c)

(d)

Figure A.2
Top 30 entities of each type in the dataset from Task 2: (a) variants, (b) genes, (c) diseases, and (d) drugs.

Acknowledgments

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 965397. This project has also be supported by funding from the digital Experimental Cancer Medicine Team, Cancer Biomarker Centre, Cancer Research UK Manchester Institute (P126273).

References

Adi, Yossi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. (arXiv:1608.04207).

Alghanmi, Israa, Luis Espinosa Anke, and Steven Schockaert. 2021. Probing pre-trained language models for disease knowledge. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3023–3033. https://doi.org/10.18653/v1/2021.findings-acl.266

Barnell, Erica K., Adam Waalkes, Matt C. Mosior, Kelsi Penewit, Kelsy C. Cotto, Arpad M. Danos, Lana M. Sheta, Katie M. Campbell, Kilannin Krysiak, Damian Rieke, Nicholas C. Spies, Zachary L. Skidmore, Colin C. Pritchard, Todd A. Fehniger, Ravindra Uppaluri, Ramaswamy Govindan, Malachi Griffith, Stephen J. Salipante, and Obi L. Griffith. 2019. Open-sourced civic annotation pipeline to identify and annotate clinically relevant variants using single-molecule molecular inversion probes. *JCO Clinical Cancer Informatics*, (3):1–12. https://doi.org/10.1200/CCI.19.00077, PubMed: 31618044

Belinkov, Yonatan. 2021. Probing classifiers: Promises, shortcomings, and advances. (arXiv:2102.12452). https://doi.org/10.1162/coli_a_00422

Borchert, Florian, Andreas Mock, Aurelie Tomczak, Jonas Hügel, Samer Alkarkoukly, Alexander Knurr, Anna-Lena Volckmar, Albrecht Stenzinger, Peter Schirmacher, Jürgen Debus, Dirk Jäger, Thomas Longerich, Stefan Fröhling, Roland Eils, Nina Bougatf, Ulrich Sax, and Matthieu-P Schapranow. 2021. Knowledge bases and software support for variant interpretation in precision oncology. *Briefings in Bioinformatics*, 22(6):Bbab134. https://doi.org/10.1093/bib/bbab134, PubMed: 33971666

Chakravarty, Debyani, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E. Rudolph, Rona Yaeger, Tara Soumerai, Moriah H. Nissan, Matthew T. Chang, Sarat Chandarlapaty, Tiffany A. Traina, Paul K. Paik, Alan L. Ho, Feras M. Hantash, Andrew Grupe, Shrujal S. Baxi, Margaret K. Callahan, Alexandra Snyder, Ping Chi, Daniel C. Danila, Mrinal Gounder, James J. Harding, Matthew D. Hellmann, Gopa Iyer, Yelena Y. Janjigian, Thomas Kaley, Douglas A. Levine, Maeve Lowery, Antonio Omuro, Michael A. Postow, Dana Rathkopf, Alexander N. Shoushtari, Neerav Shukla, Martin H. Voss, Ederlinda Paraiso, Ahmet Zehir, Michael F. Berger, Barry S. Taylor, Leonard B. Saltz, Gregory J. Riely, Marc Ladanyi, David M. Hyman, José Baselga, Paul Sabbatini, David B. Solit, and Nikolaus Schultz. 2017. OncoKB: A precision oncology knowledge base. *JCO Precision Oncology*, (1):1–16. https://doi.org/10.1200/PO.17.00011, PubMed: 28890946

Colon-Hernandez, Pedro, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal. 2021. Combining pre-trained language models and structured knowledge. (arXiv:2101.12294).

Dahary, Dvir, Yaron Golan, Yaron Mazor, Ofer Zelig, Ruth Barshir, Michal Twik, Tsippi Iny Stein, G. Rosner, R. Kariv, F. Chen, Q. Zhang, Yiping Shen, M. Safran, D. Lancet, and Simon Fishilevich. 2019. Genome analysis and knowledge-driven variant interpretation with TGex. *BMC Medical Genomics*, 12:Article 200 (17 pp). https://doi.org/10.1186/s12920-019-0647-8, PubMed: 31888639

Danos, Arpad M., K. Krysiak, Erica K. Barnell, Adam C. Coffman, J. McMichael, Susanna Kiwala, N. Spies, Lana Sheta, Shahil Pema, Lynzey Kujan, Kaitlin A. Clark, Amber Z. Wollam, Shruti Rao, D. Ritter, Dmitriy Sonkin, G. Raca, Wan-Hsin Lin, C. Grisdale, Raymond H. Kim, Alex H. Wagner, S. Madhavan, M. Griffith, and O. Griffith. 2019. Standard operating procedure for curation and clinical interpretation of variants in cancer. *Genome Medicine*, 11:Article 76 (12 pp). https://doi.org/10.1186/s13073-019-0687-x, PubMed: 31779674

Danos, Arpad M., Deborah I. Ritter, Alex H. Wagner, Kilannin Krysiak, Dmitriy Sonkin, Christine Micheel, Matthew McCoy, Shruti Rao, Gordana Raca, Simina M. Boca, Angshumoy Roy, Erica K. Barnell, Joshua F. McMichael, Susanna Kiwala, Adam C. Coffman, Lynzey Kujan, Shashikant Kulkarni, Malachi Griffith, Subha Madhavan, Obi L. Griffith, and Clinical Genome Resource Somatic Working Group and Clinical Interpretation of Variants in Cancer team members. 2018. Adapting crowdsourced clinical cancer curation in CIViC to the ClinGen minimum variant level data community-driven standards. *Human Mutation*, 39(11):1721–1732. https://doi.org/10.1002/humu.23651, PubMed: 30311370

Dienstmann, Rodrigo, In Sock Jang, Brian Bot, Stephen Friend, and Justin Guinney. 2015. Database of genomic biomarkers for cancer drugs and clinical targetability in solid tumors. *Cancer Discovery*, 5(2):118–123. https://doi.org/10.1158/2159-8290.CD-14-1118, PubMed: 25656898

Durrani, Nadir, Hassan Sajjad, and Fahim Dalvi. 2021. How transfer learning impacts linguistic knowledge in deep NLP models? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4947–4957. https://doi.org/10.18653/v1/2021.findings-acl.438

Ferreira, Deborah, Julia Rozanova, Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2021. Does my representation capture X? Probe-ably. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 194–201. https://doi.org/10.18653/v1/2021.acl-demo.23

Fix, E. and J. L. Hodges. 1989. Discriminatory analysis - Nonparametric discrimination: Consistency properties. *International Statistical Review*, 57:238. https://doi.org/10.2307/1403797

Gehman, Samuel, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

Good, Benjamin M., Benjamin J. Ainscough, Josh F. McMichael, Andrew I. Su, and Obi L. Griffith. 2014. Organizing knowledge to enable personalization of medicine in cancer. *Genome Biology*, 15(8):438. https://doi.org/10.1186/s13059-014-0438-7, PubMed: 25222080

Griffith, Malachi, Nicholas C. Spies, Kilannin Krysiak, Joshua F. McMichael, Adam C. Coffman, Arpad M. Danos, Benjamin J. Ainscough, Cody A. Ramirez, Damian T. Rieke, Lynzey Kujan, Erica K. Barnell, Alex H. Wagner, Zachary L. Skidmore, Amber Wollam, Connor J. Liu, Martin R. Jones, Rachel L. Bilski, Robert Lesurf, Yan Yang Feng, Nakul M. Shah, Melika Bonakdar, Lee Trani, Matthew Matlock, Avinash Ramu, Katie M. Campbell, Gregory C. Spies, Aaron P. Graubert, Karthik Gangavarapu, James M. Eldred, David E. Larson, Jason R.Walker, Benjamin M. Good, Chunlei Wu, Andrew I. Su, Rodrigo Dienstmann, Adam A. Margolin, David Tamborero, Nuria Lopez-Bigas, Steven J. M. Jones, Ron Bose, David H. Spencer, Lukas D. Wartman, Richard K. Wilson, Elaine R. Mardis, and Obi L. Griffith. 2017. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics*, 49(2):170–174. https://doi.org/10.1038/ng.3774, PubMed: 28138153

Gu, Wenhao, Xiao Yang, Minhao Yang, Kun Han, Wenying Pan, and Zexuan Zhu. 2022. MarkerGenie: An NLP-enabled text-mining system for biomedical entity relation extraction. *Bioinformatics Advances*, 2(1):vbac035. https://doi.org/10.1093/bioadv/vbac035

Gururangan, Suchin, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112. https://doi.org/10.18653/v1/N18-2017

He, Max M., Quan Li, Muqing Yan, Hui Cao, Yue Hu, Karen Y. He, Kajia Cao, Marilyn M. Li, and Kai Wang. 2019. Variant Interpretation for Cancer (VIC): A computational tool for assessing clinical impacts of somatic variants. *Genome Medicine*, 11(1):53. https://doi.org/10.1186/s13073-019-0664-4, PubMed: 31443733

Hewitt, John and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Hupkes, Dieuwke, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and ‘diagnostic classifiers’ reveal how recurrent and recursive neural networks process hierarchical structure. (arXiv:1711.10203). https://doi.org/10.24963/ijcai.2018/796

Jia, Robin and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031. https://doi.org/10.18653/v1/D17-1215

Jia, Robin, Cliff Wong, and Hoifung Poon. 2019. Document-level *N*-ary relation extraction with multiscale representation learning. *CoRR*, abs/1904.02347. https://doi.org/10.18653/v1/N19-1370

Jin, Qiao, Bhuwan Dhingra, William W. Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. (arXiv:1904.02181). https://doi.org/10.18653/v1/W19-2011

Jullien, Mael, Marco Valentino, and Andre Freitas. 2022. Do transformers encode a foundational ontology? Probing abstract classes in natural language. (arXiv:2201.10262).

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. https://doi.org/10.1093/bioinformatics/btz682, PubMed: 31501885

Lever, Jake, Martin R. Jones, Arpad M. Danos, Kilannin Krysiak, Melika Bonakdar, Jasleen Grewal, Luka Culibrk, Obi L. Griffith, Malachi Griffith, and Steven J. M. Jones. 2018. Text-mining clinically relevant cancer biomarkers for curation into the CIViC database. *bioRxiv Genome Medicine*, 11(1):Article 78 (16 pp). https://doi.org/10.1186/s13073-019-0686-y, PubMed: 31796060

Li, Marilyn M., Michael Datto, Eric J. Duncavage, Shashikant Kulkarni, Neal I. Lindeman, Somak Roy, Apostolia M. Tsimberidou, Cindy L. Vnencak-Jones, Daynna J. Wolff, Anas Younes, and Marina N. Nikiforova. 2017. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: A joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *The Journal of Molecular Diagnostics*, 19(1):4–23. https://doi.org/10.1016/j.jmoldx.2016.10.002, PubMed: 27993330

McCoy, Tom, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448. https://doi.org/10.18653/v1/P19-1334

McInnes, Leland and John Healy. 2017. Accelerated hierarchical density based clustering. In *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*, pages 33–42. https://doi.org/10.1109/ICDMW.2017.12

McInnes, Leland, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205. https://doi.org/10.21105/joss.00205

McInnes, Leland, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. UMAP: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861. https://doi.org/10.21105/joss.00861

Merchant, Amil, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44. https://doi.org/10.18653/v1/2020.blackboxnlp-1.4

Min, Junghyun, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352. https://doi.org/10.18653/v1/2020.acl-main.212

Nadeem, Moin, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416

Pimentel, Tiago, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622.

https://doi.org/10.18653/v1/2020.acl-main.420

Qumsiyeh, Emma and Rashid Jayousi. 2021. Biomedical information extraction pipeline to identify disease-gene interactions from PubMed breast cancer literature. In *2021 International Conference on Promising Electronic Technologies (ICPET)*, pages 1–6. https://doi.org/10.1109/ICPET53277.2021.00007

Rajaee, Sara and Mohammad Taher Pilehvar. 2021. How does fine-tuning affect the geometry of embedding space: A case study on isotropy. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3042–3049. https://doi.org/10.18653/v1/2021.findings-emnlp.261

Rieke, Damian T., Mario Lamping, Marissa Schuh, Christophe Le Tourneau, Neus Basté, Mark E. Burkard, Klaus H. Metzeler, Serge Leyvraz, and Ulrich Keilholz. 2018. Comparison of treatment recommendations by molecular tumor boards worldwide. *JCO Precision Oncology*, (2):1–14. https://doi.org/10.1200/PO.18.00098

Rives, Alexander, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118 (12 pp). https://doi.org/10.1073/pnas.2016239118, PubMed: 33876751

Ševa, Jurica, Martin Wackerbauer, and Ulf Leser. 2018. Identifying key sentences for precision oncology using semi-supervised learning. In *Proceedings of the BioNLP 2018 Workshop*, pages 35–46. https://doi.org/10.18653/v1/W18-2305

Shin, Hoo Chang, Yang Zhang, Evelina Bakhturina, Raul Puri, M. Patwary, M. Shoeybi, and Raghav Mani. 2020. Bio-Megatron: Larger biomedical domain language model. In *EMNLP*, pages 4700–4706. https://doi.org/10.18653/v1/2020.emnlp-main.379

Singhal, Ayush, Michael Simmons, and Zhiyong Lu. 2016. Text mining genotype-phenotype relationships from biomedical literature for database curation and precision medicine. *PLoS Computational Biology*, 12(11):e1005017. https://doi.org/10.1371/journal.pcbi.1005017, PubMed: 27902695

Vig, Jesse, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. BERTology meets biology: Interpreting attention in protein language models. (arXiv:2006.15222). https://doi.org/10.1101/2020.06.26.174417

Wagner, Alex H., Susanna Kiwala, Adam C. Coffman, Joshua F. McMichael, Kelsy C. Cotto, Thomas B. Mooney, Erica K. Barnell, Kilannin Krysiak, Arpad M. Danos, Jason Walker, Obi L. Griffith, and Malachi Griffith. 2020. Civicpy: A Python software development and analysis toolkit for the CIViC knowledgebase. *JCO Clinical Cancer Informatics*, (4):245–253. https://doi.org/10.1200/CCI.19.00127, PubMed: 32191543

Wang, Benyou, Qianqian Xie, Jiahuan Pei, Prayag Tiwari, Zhao Li, and Jie Fu. 2021. Pre-trained language models in biomedical domain: A systematic survey. (arXiv:2110.05006).

Wang, Hai and Hoifung Poon. 2018. Deep probabilistic logic: A unifying framework for indirect supervision. *CoRR*, abs/1808.08485. https://doi.org/10.18653/v1/D18-1215

Yuan, Zheng, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. 2021. Improving biomedical pretrained language models with knowledge. (arXiv:2104.10344). https://doi.org/10.18653/v1/2021.bionlp-1.20

Zhang, Xikun, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do language embeddings capture scales? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4889–4896. https://doi.org/10.18653/v1/2020.findings-emnlp.439

Zhong, Zexuan, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033. https://doi.org/10.18653/v1/2021.naacl-main.398