Transformers and the representation of biomedical background knowledge

TRANSFORMERS AND THE REPRESENTATION OF BIOMEDICAL BACKGROUND KNOWLEDGE

Oskar Wysocki*[1,2], Zili Zhou[1,2], Paul O'Regan[2], Deborah Ferreira[1],
Magdalena Wysocka[2], Dónal Landers[2], and André Freitas[1,2,3]

[1]Department of Computer Science, The University of Manchester
[2]digital Experimental Cancer Medicine Team, Cancer Biomarker Centre,
CRUK Manchester Institute, University of Manchester
[3]Idiap Research Institute

ABSTRACT

BioBERT and BioMegatron are Transformers models 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 analyse how the models behave with regard to biases and imbalances in the dataset.

1 Introduction

Transformers are deep-learning based models which are able to capture linguistic patterns at scale. By using unsupervised learning tasks which can be defined over large-scale textual corpora, these models are able to capture both linguistic and domain knowledge, which can be later specialised for specific inference tasks. The representation produced by the model is a high-dimensional linguistic space which 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 challenge their application in areas such as biomedicine.

In this work we elucidate the internal properties of transformers in the context of a well-defined cancer precision medicine inference task as expressed in the biomedical literature with an emphasis on the entities of gene, gene variant, drug and disease. 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 gefitinib"*

Additionally, this work provides a critical exploration of the internal representation properties of these models, using probing and clustering methods to compare and contrast the semantic behavior of different models addressing fundamental questions for the application of these models in biomedical inference such as:

[*] Corresponding author: oskar.wysocki@manchester.ac.uk

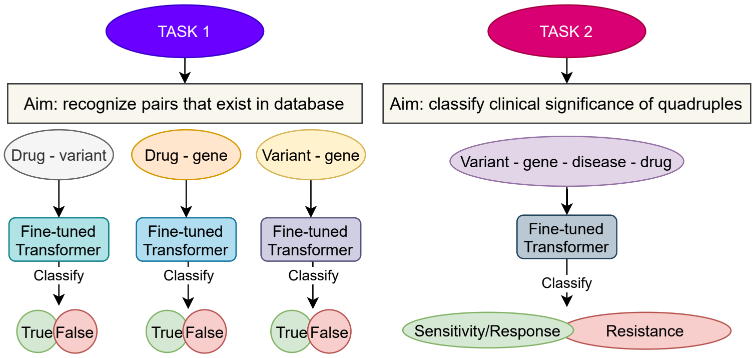


Figure 1: An overview of classification task 1 and 2. Each Transformer block represents a separate model which was fine-tuned separately for each classification. Two Transformers were used: BioBERT and BioMegatron.

1. Do these models encode fundamental semantic knowledge from the domain, at entity level (e.g. gene, gene variant, disease drug) and at a relational level?
2. How do these models cope with encoding complex n-ary relations?
3. Are there significant differences on how different models encode domain knowledge?
4. How so these models cope with distributional biases (e.g. are facts more frequently expressed in the literature, elicited in the models)?

In this analysis, we used state-of-the-art transformers: BioBERT [16] and BioMegatron [27]. Both models are pre-trained over large biomedical literature text corpora (PubMed[2]). These models have demonstrated the ability to address complex domain-specific linguistic tasks, such as answering biomedical questions [27]. Yet, the internal representation properties of these models are not fully understood and require exploration in order to use such models safely and reliably, exploiting the full potential on natural language processing methods, i.a. in cancer research.

We contribute the following:

- First, we compared the performance in specific classification tasks (Fig.1) between biomedical fine-tuned transformers (BioBERT and BioMegatron) and a naive simple classifier (KNN). The aim of the tasks were: i) identifying the existence of a given pair of entities in the dataset and ii) identifying the clinical significance of a sentence containing 4 entities: variant, gene, disease and drug. We hypothesised, that the transformer models would show superior performance as a result of knowledge embedded in the model, and not only from the fine-tuning fit to the training set. We also checked whether the model's error depends on the quality of the biological evidence, its occurrence in the literature, or its recognition among domain experts.
- Second, 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 whether the model captures the fundamental biomedical/semantic categories to support interpretation. We quantified how much semantic structure is lost in fine-tuning, and how biologically meaningful is the remaining.
- 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, e.g. groups with genes from the same pathways. Additionally, using homogeneity of clusters, we quantified the associations between the representations and the entity type and target labels.

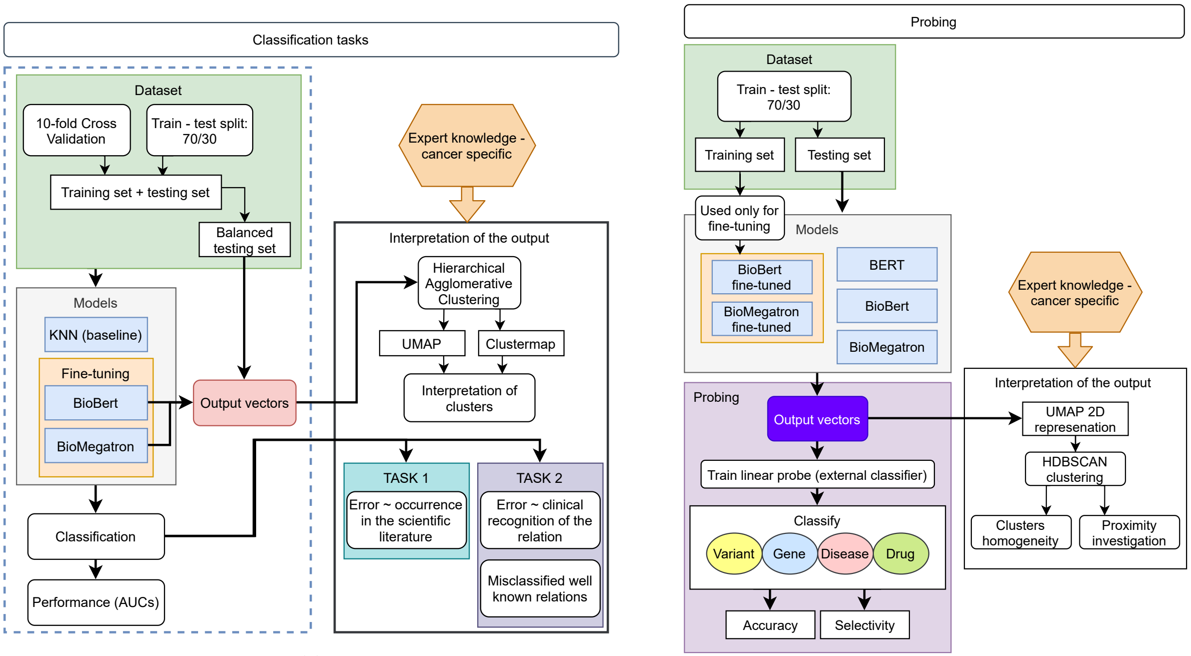
The workflow of the analysis is summarized in Fig.2.

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 characterisation of their tumour, has the potential to improve patient outcomes. For example, activating mutations in the epidermal growth

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


Classification tasks
Dataset
10-fold Cross Validation
Train - test split: 70/30
Training set + testing set
Balanced testing set
Models
KNN (baseline)
Fine-tuning
BioBert
BioMegatron
Output vectors
Classification
Performance (AUCs)
Expert knowledge - cancer specific
Interpretation of the output
Hierarchical Agglomerative Clustering
UMAP
Clustermap
Interpretation of clusters
TASK 1
Error ~ occurrence in the scientific literature
TASK 2
Error ~ clinical recognition of the relation
Misclassified well known relations
Probing
Dataset
Train - test split: 70/30
Training set
Testing set
Used only for fine-tuning
Models
BioBert fine-tuned
BioMegatron fine-tuned
BERT
BioBert
BioMegatron
Probing
Output vectors
Train linear probe (external classifier)
Classify
Variant
Gene
Disease
Drug
Accuracy
Selectivity
Expert knowledge - cancer specific
Interpretation of the output
UMAP 2D represenation
HDBSCAN clustering
Clusters homogeneity
Proximity investigation

(a)

(b)

Figure 2: The workflow of the performed analysis.

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 [10].

Routine molecular characterisation of patients' tumours has become feasible because of improved turnaround times and reduced costs of molecular diagnostics [24]. 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 effective interventions, in order to increase survival and reduce the likelihood of adverse drug reactions [3].

Even considering only licensed treatments, the number of alternative treatments available may be very large. For example, in the USA, there are over 70 drugs approved by the FDA for the treatment of NSCLC [4]. 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 has the potential to scale-up the interpretation of this evidence space which could be integrated into decision support tools.

2.2 Database

CIViC[5] (Clinical Interpretation of Variants in Cancer) is a community-edited knowledge base of associations between genetic variations (or other alterations), drugs and outcomes in cancer [12]. The goal of CIViC is to support the implementation of personalized medicine in cancer. Data are freely available and licensed under a Creative Commons

[3] https://www.england.nhs.uk/genomics/nhs-genomic-med-service/
[4] https://www.cancer.gov/about-cancer/treatment/drugs/lung
[5] https://civicdb.org/home

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 [3] and My Cancer Genome [22], it is one of the most commonly used knowledge bases [2] for this purpose.

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 (e.g. 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 dump options and is integrated into various recent annotation tools and it follows an ontology driven conceptual model. It allows users to transparently generate current and accurate variant interpretations because it receives monthly updates. As of 5 January 2022, the database holds 8,441 interpretations of clinical relevance for 2,969 variants among 460 genes associated with 322 diseases and 479 drugs. Its accessibility and tabular format of the data allows for easy integration into ML pipelines, both as input data and domain knowledge incorporated in the model.

2.3 Data preprocessing

The process of downloading the data via API for the purpose of this study is detailed in Supp Methods 5.

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 (Fig.1) 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, can the model correctly classify T790M as a variant of the EGFR gene, but not the KRAS gene?

Three types of pairs 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 generated by randomly selecting entities from CIViC, and excluding those that already exist (i.e., negative sampling). The dataset includes equal # of false and true pairs. Of note, a pair can occur in multiple evidence items, i.e. 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 (Fig.1) 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 database contains a variant, gene, disease, drugs and CS, so a quadruple can be extracted directly from the database, and there are no false quadruples. Only unique quadruples were used to create the dataset. In the case of 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 preprocessing 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), i.e. 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 (Supp Methods 5). 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 paper, we used a naive classification model (Nearest Neighbors Classification model [9]) as a baseline. Briefly, each entity was represented as a sparse, one-hot encoded vector such that, e.g. 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, whilst 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 for use as 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 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]".

For more details refer to Supp Methods 5.

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 a probing framework, Probe-Ably [6], 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 [8, 14, 23]. Probing is often performed as a *post hoc* analysis, taking a pre-trained or fine-tuned model and analysing the obtained embeddings. For example, previous probing studies [25] 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, i.e., 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?* 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. We split the representations into training, validation and test set, using a 20/40/40 rate. Each model is trained for 5 epochs, with the validation set being used to selected the best performing model (in terms of accuracy). In this case, we want a small training set to avoid any overfitting.
3. After obtaining all representations for each model and respective entity types, we train a total of 50 linear probe 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 for 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 [23], measuring the complexity of a linear model $\hat{y} = W\mathbf{x} + \mathbf{b}$ 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. In order to obtain 50 different models, we randomly initialize the dropout and λ parameter. 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

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

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, we know 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 randomised labels, indicates that the probe architectures used are not expressive enough to ”memorise” 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-parameterised probes (i.e, probes that contain several learnable parameters) could have explained them.

2.6 Clustering

In addition to the evaluation of models' performance, we investigated these output vectors to identify potential relationships between entity pairs and/or quadruples. For clustering the output in Task 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 [20, 19], with parameter min cluster size = 120, and the rest kept their default values.

We applied Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [21] to compare patterns observable after dimensionality reduction into 2 dimensions with clusters obtained via HAC. UMAP parameters: default (n components= 2, n neighbors= 15)

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 is equal to the ratio of the count of most common label in the cluster and the total count in the cluster, e.g. 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 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: Pairs and unique entities the dataset used in Task 1.

	n (both true and false)	n in train set	n in test set	n in balanced test set (% of test set)	Unique genes (n)	Unique variants (n)	Unique drugs (n)
drug - variant	3676	2272	1104	418 (38%)	-	897	242
drug - gene	2480	1736	744	396 (53%)	302	-	432
variant - gene	1876	1312	564	276 (49%)	125	910	-

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 (SuppFig.S.1a). A similar pattern was observed for drug-variant pairs (Supp Fig.S.1b). In contrast, for variant-gene pairs, the majority of pairs involving the most common variant entities were false (Supp Fig.S.1c).

3.1.2 Performance

We evaluated 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. Performance of the transformer models was superior to the naive model in all cases, except for drug-gene classification against the imbalanced dataset.

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)	0.59	.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)

3.1.3 The impact of imbalance on 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, i.e. 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. Similarly to entities distribution, there is an imbalance in the # of evidence items related to pairs. For example, 73% of variant-drug pairs are supported only by one, 2% by $>= 2$, and 1.4% by$>=$10 evidence items. Details for all 3 types of pairs are shown in the Table 3.

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

Pair	Number of evidence items				
	1	>1	>=2	>=10	>=20
gene-drug ($n = 1240$)	792 (64.1%)	442 (32.9%)	127 (12.7%)	82 (6.9%)	12 (0.97%)
variant-gene ($n = 938$)	296 (63.2%)	342 (36.2%)	117 (12.2%)	49 (2.2%)	2 (0.2%)
variant-drug ($n = 1838$)	1347 (73.3%)	491 (26.7%)	91 (2%)	22 (1.4%)	1 (0.02%)

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 (Fig.3a) or in the knowledge base (Fig.3b), true drug-variant pairs were typically classified correctly, whilst false drug-variant pairs were typically misclassified .

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 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 Supp Table S.1.

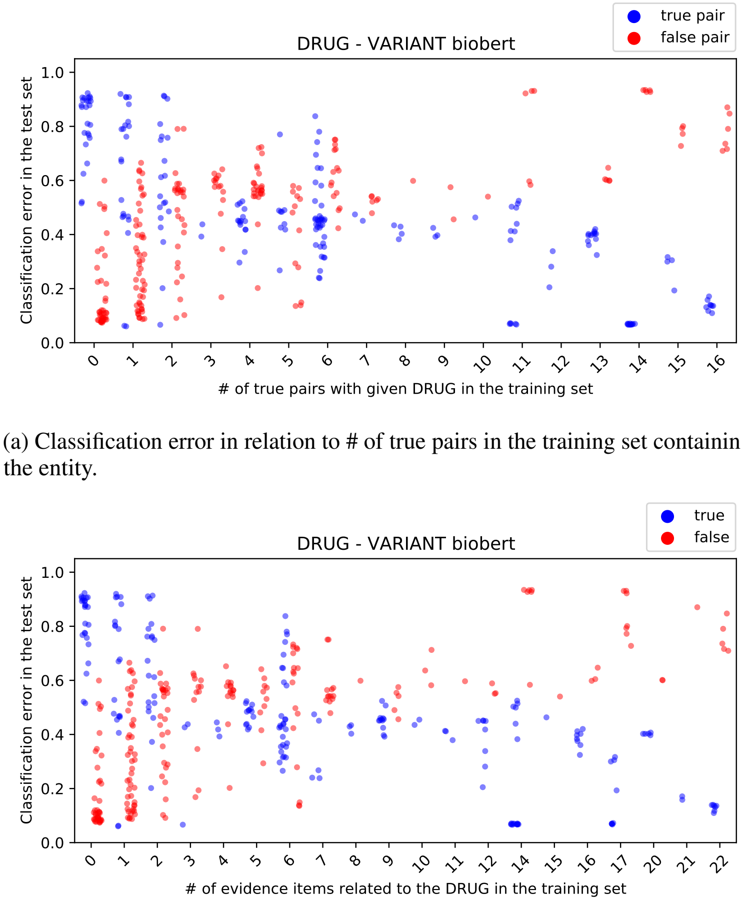
3.2 Can transformers recognize clinical significance of a relation? - Task 2

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 Supp Table S.2).

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 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 Supp Fig.S.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,*



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

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

Figure 3

OVEREXPRESSION, AMPLIFICATION, EXPRESSION, V600E, LOSS, FUSION, LOSS-OF-FUNCTION and UNDER-EXPRESSION), where approximately 80% of quadruples related to drug sensitivity.

3.2.2 Performance

We evaluated the performance of the models in predicting clinical significance of quadruples using AUC. In all cases, performance of the transformer models was superior to that of the naive model. 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 greater than 0.8) on the balanced dataset (Table 4).

Table 4: AUC in classification task 2.

AUC	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)

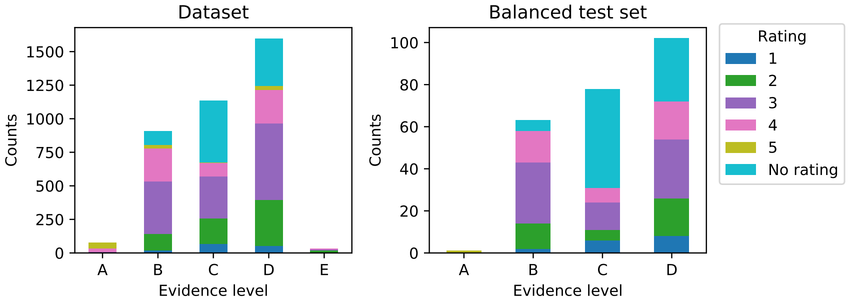


Figure 4: Number of evidence items in the datasets stratified by evidence level and evidence rating.

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 (Fig.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). Considering preclinical evidence only (*Evidence level D*), the KNN model had significantly higher error (MWU test) compared with BioBERT (p=0.014) and BioMegatron (p=0.007). This finding was supported by AUC and Brier scores (Supp Table S.4).

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 5).

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 ($\geq .6$) 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)

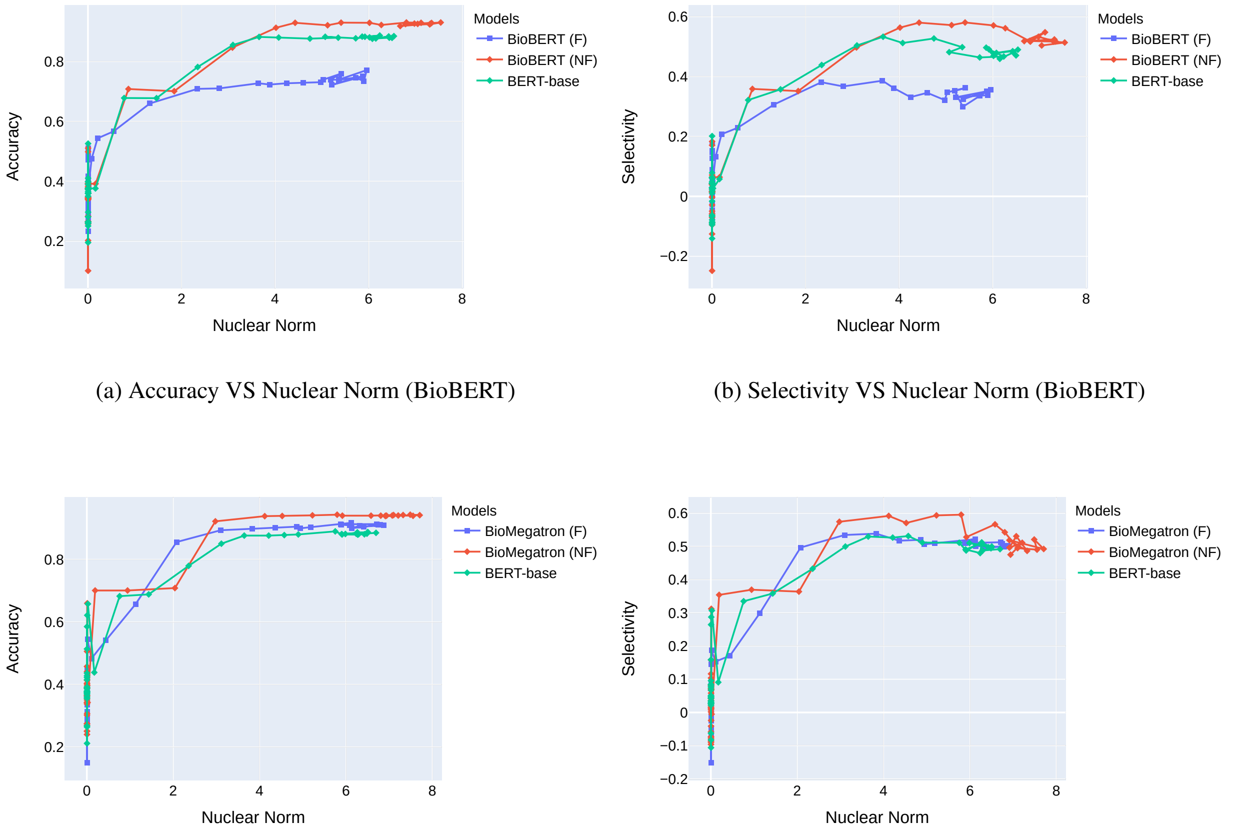
3.3 Do the fine-tuned models lose the generalizability of representations?

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. For BioMegatron, performance of the fine-tuned model was no better than that for either of the non-fine-tuned models.

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 see very little difference between the performance of the fine-tuned and non fine-tuned versions of BioMegatron, which outperform BERT and BioBERT models. For



(a) Accuracy VS Nuclear Norm (BioBERT)

(b) Selectivity VS Nuclear Norm (BioBERT)

(c) Accuracy VS Nuclear Norm (BioMegatron)

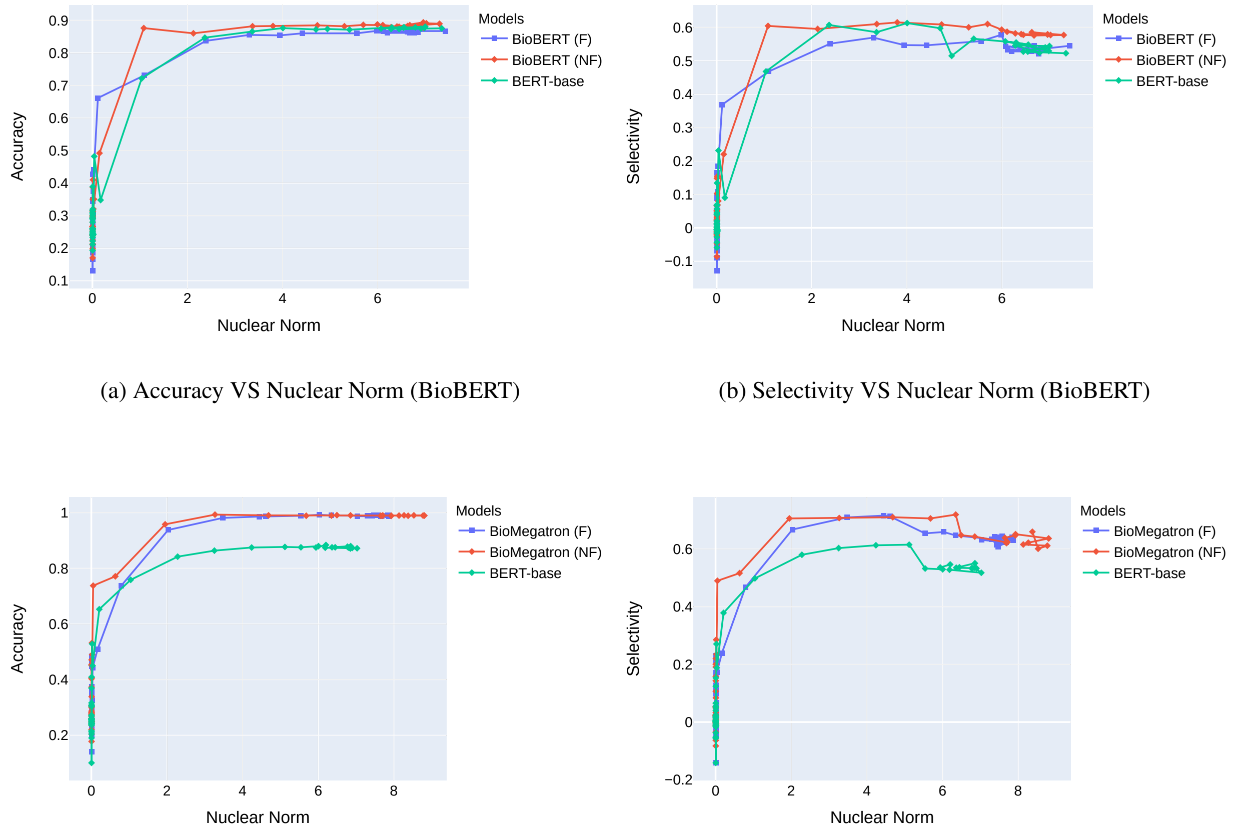
(d) Selectivity VS Nuclear Norm (BioMegatron)

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

Table 5: 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

probes with a lower value for nuclear norm (i.e., less complex probes), the performance of the original model is slightly better. However, the difference is non-existent for more 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 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

3.4 How much biological knowledge do transformers embed?

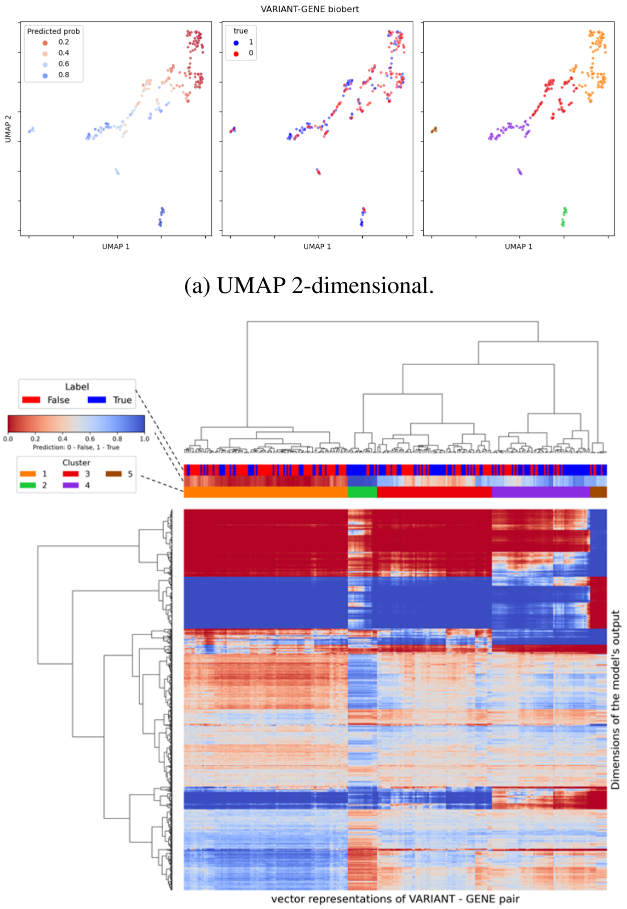
3.4.1 Biologically relevant clusters in representations of pairs

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 (Fig.7b).

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 4/11 pairs were false (Supp Table S.5). Interestingly, these genes participate in the same signalling pathways, including PI3K/AKT/mTOR.

Cluster #2 (green) contained 19 gene-variant pairs; 14/19 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 3/19 pairs were false S.6).

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 (Fig.8).



(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.

BioMegatron cluster #1 contained 16 of the 19 gene-variant pairs found in BioBERT cluster #2 (Supp Table S.6) As observed for BioBERT, the BioMegatron called all these pairs as true with high confidence (probability >0.96).

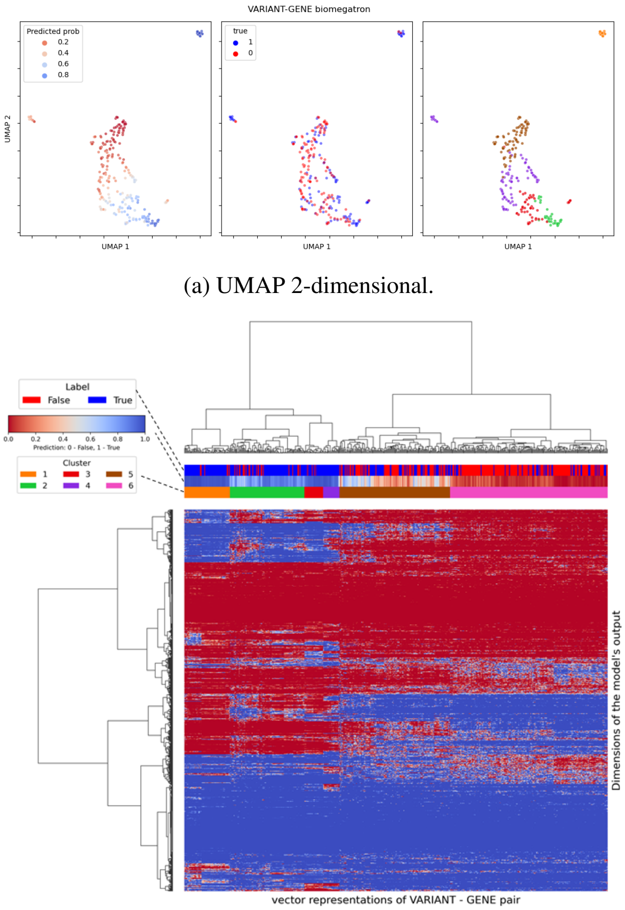
3.4.2 Biologically relevant clusters in representations of clinical relations

Following clustering of BioMegatron representations of quadruples, a cut point was applied that resulted in 6 clusters (Fig.9).

- 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 10/11 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 for the BioBERT representations (Fig.10).

Quadruples in BioBERT clusters were less homogeneous compared with those for the BioMegatron clusters. Two small clusters 5 and 6 are described in Supp Table S.7. 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.



(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.

3.4.3 Representations proximity defined by entity type

In this section, we investigated the input vectors to the probing task. Each vector represents one entity contextualised inside sentences from the test set (From both Task 1 and Task 2, more in Supp Methods 5).

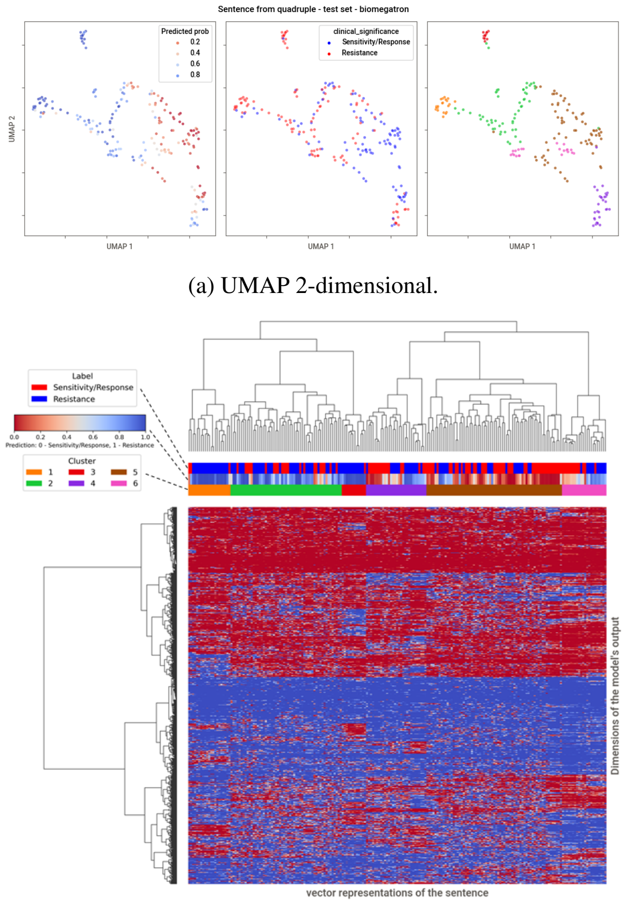
Results from HDBSCAN evaluation of UMAP representations are summarized in Table 6 and Fig.11c,12e.

For Task 1, the non-fine-tuned transformer models clustered entities according to their type - 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 Fig12).

However, for fine-tuned models, the majority of representations get closer together in 2D UMAP. In fine-tuned BioBERT, drugs stick to variants and some genes. As a result, a large cluster (#5) with mixed entity types emerges. The fine-tuned BioMegatron is even more disorganized compared to its non fine-tuned counterpart. Clustering shows one massive 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 labels proportion are 0.50 and 0.65, respectively.



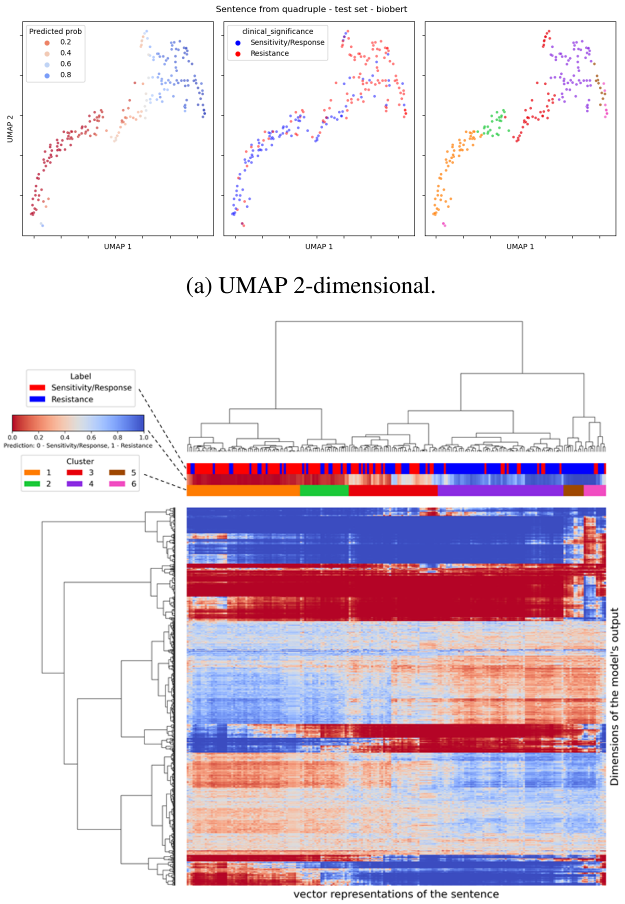
(a) UMAP 2-dimensional.

(b) Clustermap based on Hierarchical Agglomerative Clustering.

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

Table 6: 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	



(a) UMAP 2-dimensional.

(b) Clustermap based on Hierarchical Agglomerative Clustering.

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

4 Discussion

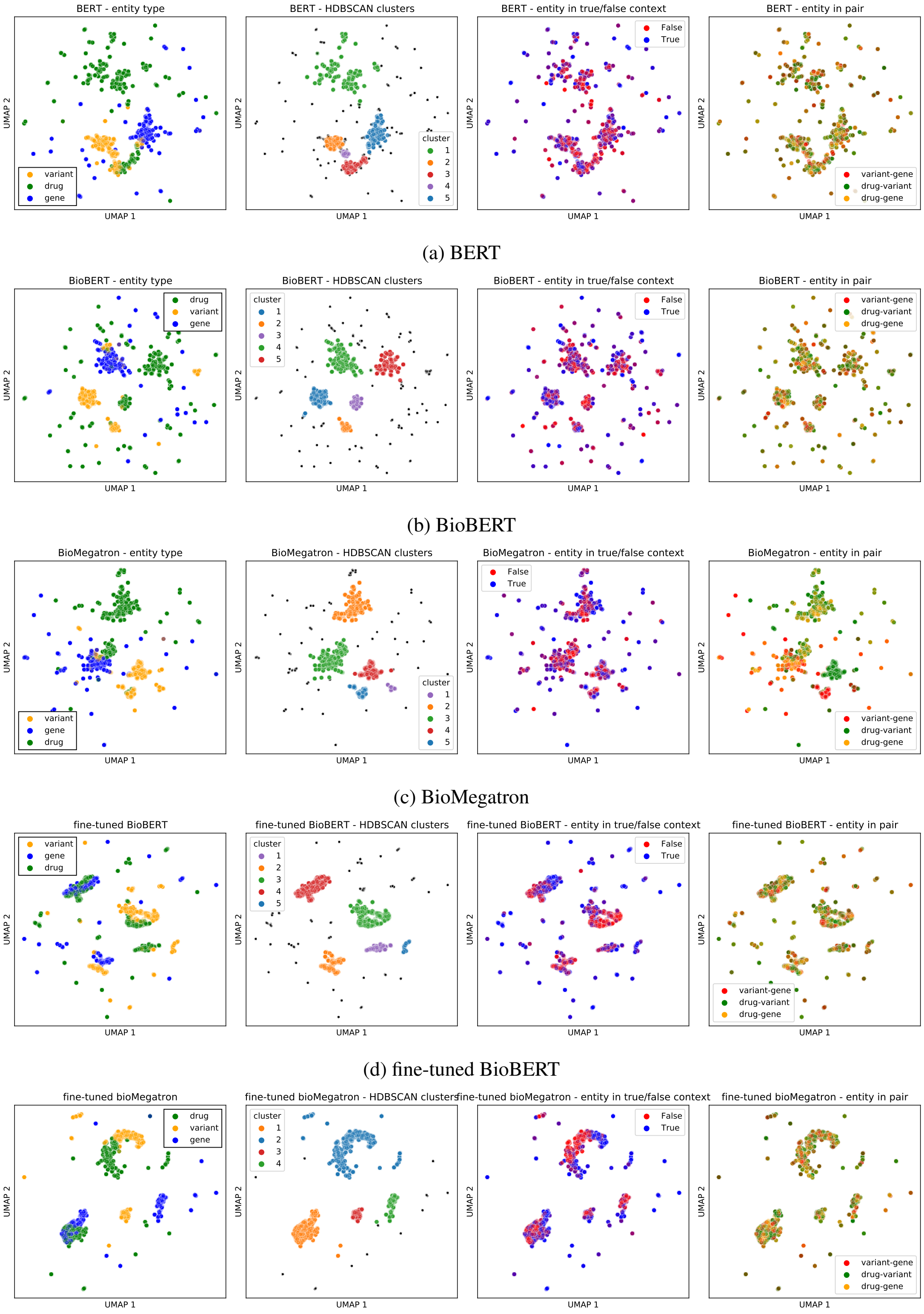
4.1 Summary of main findings

In this study we performed a detailed semantic analysis on the embedding 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 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 superior performance compared with a naive classifier. Results for both tasks support this hypothesis. For Task 1, both BioBERT and BioMegatron were superior to 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 exploratory data analysis focusing on the imbalance in the dataset, as it has a major impact on the performance. Furthermore, we argue that the imbalance is an inherent property of any biological real-world textual corpus, and should be always considered in model's evaluation. 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 individual entities. Similarly, the error is associated with



(a) BERT

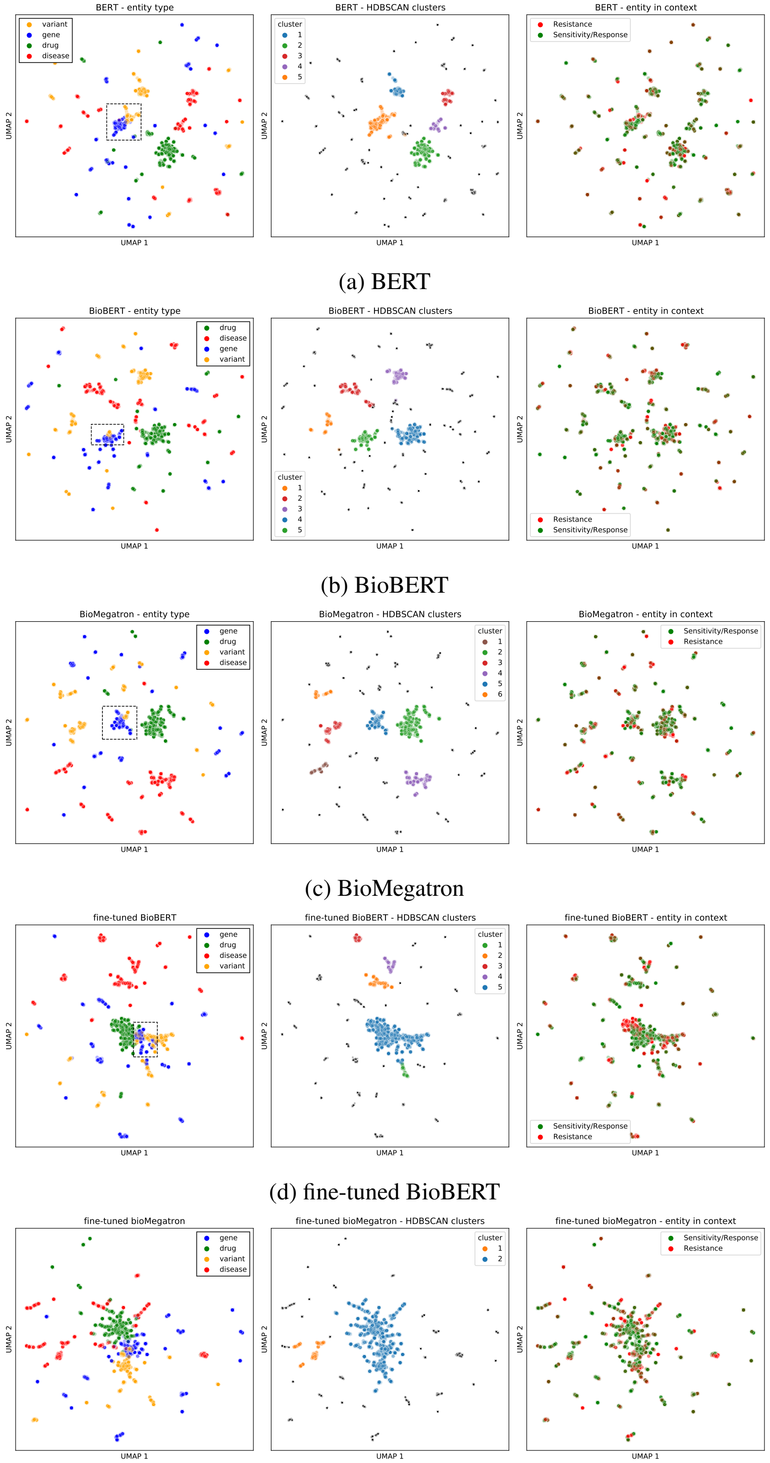
(b) BioBERT

(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.



(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.

the occurrence in the corpus (biomedical literature): i.e. in Task 1: a true pair which 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 representation for each type of biological entity, and we compared pre-trained versus fine-tuned models. 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 no difference in performance between the fine-tuned and non-fine-tuned models. Fine-tuned BioMegatron performed better than fine-tuned BioBERT in probing, which suggest that it preserves the semantic structure better after the fine-tuning.

Finally, we provide a qualitative and quantitative analysis of clustering patterns of the embeddings, using UMAP, HDBCAN and hierarchical agglomoerative clustering. 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. Representations in classification Tasks 1 and 2 reveal the biological meaning, especially for BioMegatron. For instance, we found a cluster with 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 the Imatinib, a KIT inhibitor, is a the 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: first, it entails a set of relationships curated by domain experts; second, most relationships include a confidence score; third, it has been developed for a closely-related use case, namely to support clinicians in evaluation of clinical significance of variants.
- We employ 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 only on one. Clusters are thoroughly described and quantified using homogeneity.
- 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; 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 the whether gene = EGFR and perform 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 in Machine Learning, which leads to an incorrect performance evaluation. Although we applied a balancing procedure, it is infeasible to create 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 it 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 (NGS) analyses, which predominantly have unknown significance and complicates the interpretation of the variants [11]. One of the ways to streamline and standardise cancer curation data in electronic medical records is to use the web resources from the CIViC curatorial platform [6]. An open source and open access CIViC database, built on community input with peer-reviewed interpretations, already proved to be useful for this purpose [1]. 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 patients' treatment. In addition, Danos et al. [5] 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 TGex (the Translational Genomics expert) [4]. By providing access to a comprehensive knowledge base of genomic annotations, TGex tool simplifies and speeds up the interpretation of variants in clinical genetics processes. Furthermore, Wagner et al. [29] 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 which used CIViC database

The development of guidelines [18] 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. [13] 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. [17] 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 knowledge base 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 et al. [28] who proposed a method to automate the extraction of disease-gene-variant triplets from all abstracts in PubMed related to a set of ten important diseases.

Seva et al. [26] developed a machine learning 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 machine learning approaches: transductive learning from positive and unlabelled data (PU Learning) and Self-Training by using abstracts summarised in relevant sentences as unlabelled examples. Wang and Poon [30] developed deep probabilistic logic (DPL) as a general framework for indirect supervision, by combining probabilistic logic with deep learning. They used existing knowledge bases with hand-curated drug-gene-mutation facts: the Gene Drug Knowledge Database (GDKD) [7] and CIViC, which together contained 231 drug-gene-mutation triples, with 76 drugs, 35 genes and 123 mutations. Recently, Jia et al. [15] 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 [7], and OncoKB [3] knowledge bases.

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

5 Conclusions

In this work we performed a detailed analysis of fundamental knowledge representation properties of transformers, demonstrating that they are biased towards more frequent statements. We recommend to account 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 in 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 [8]. We found that the fine-tuned models loose 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 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 some eminent clinical quadruples, the models misclassified the clinical significance (whether sensitive or resistant to treatment), highlighting the limitations of contemporary neural language models.

References

[1] Erica K. Barnell et al. "Open-Sourced CIViC Annotation Pipeline to Identify and Annotate Clinically Relevant Variants Using Single-Molecule Molecular Inversion Probes". In: *JCO Clinical Cancer Informatics* 3 (2019). PMID: 31618044, pp. 1–12. DOI: 10.1200/CCI.19.00077. eprint: https://doi.org/10.1200/CCI.19.00077. URL: https://doi.org/10.1200/CCI.19.00077.

[2] Florian Borchert et al. "Knowledge bases and software support for variant interpretation in precision oncology". In: *Briefings in Bioinformatics* (May 2021). bbab134. ISSN: 1477-4054. DOI: 10.1093/bib/bbab134. eprint: https://academic.oup.com/bib/advance-article-pdf/doi/10.1093/bib/bbab134/38657863/bbab134.pdf. URL: https://doi.org/10.1093/bib/bbab134.

[3] Debyani Chakravarty et al. "OncoKB: A Precision Oncology Knowledge Base". In: *JCO Precision Oncology* 1 (2017), pp. 1–16. DOI: 10.1200/PO.17.00011. eprint: https://doi.org/10.1200/PO.17.00011. URL: https://doi.org/10.1200/PO.17.00011.

[4] Dvir Dahary et al. "Genome analysis and knowledge-driven variant interpretation with TGex". In: *BMC Medical Genomics* 12 (2019).

[5] Arpad M Danos et al. "Standard operating procedure for curation and clinical interpretation of variants in cancer". In: *Genome Medicine* 11 (2019).

[6] Arpad M. Danos et al. "Adapting crowdsourced clinical cancer curation in CIViC to the ClinGen minimum variant level data community-driven standards". In: *Human Mutation* 39.11 (2018), pp. 1721–1732. DOI: https://doi.org/10.1002/humu.23651. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.23651. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.23651.

[7] Rodrigo Dienstmann et al. "Database of Genomic Biomarkers for Cancer Drugs and Clinical Targetability in Solid Tumors". In: *Cancer Discovery* 5.2 (2015), pp. 118–123. ISSN: 2159-8274. DOI: 10.1158/2159-8290.CD-14-1118. eprint: https://cancerdiscovery.aacrjournals.org/content/5/2/118.full.pdf. URL: https://cancerdiscovery.aacrjournals.org/content/5/2/118.

[8] Deborah Ferreira et al. "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*. Online: Association for Computational Linguistics, Aug. 2021, pp. 194–201. DOI: 10.18653/v1/2021.acl-demo.23. URL: https://aclanthology.org/2021.acl-demo.23.

[9] E. Fix and J. L. Hodges. "Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties". In: *International Statistical Review* 57 (1989), p. 238.

[10] Benjamin M Good et al. "Organizing Knowledge to Enable Personalization of Medicine in Cancer". In: *Genome Biology* 15.8 (Aug. 2014), p. 438. ISSN: 1474-760X. DOI: 10/gnp849. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0438-7 (visited on 01/25/2022).

[11] Benjamin M. Good et al. "Organizing knowledge to enable personalization of medicine in cancer". In: *Genome Biology* 15.8 (2014), p. 438. ISSN: 1474-760X. DOI: https://doi.org/10.1186/s13059-014-0438-7. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0438-7#citeas.

[12] Malachi Griffith et al. "CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer". English. In: *Nature Genetics* 49.2 (Jan. 2017), pp. 170–174. ISSN: 1061-4036. DOI: 10.1038/ng.3774.

[13] Max M. He et al. "Variant Interpretation for Cancer (VIC): a computational tool for assessing clinical impacts of somatic variants". In: *Genome Medicine* 11.1 (Aug. 2019), p. 53. ISSN: 1756-994X. DOI: 10/gf8dht. URL: https://doi.org/10.1186/s13073-019-0664-4.

[14] John Hewitt and Christopher D Manning. “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)*. 2019, pp. 4129–4138.

[15] Robin Jia, Cliff Wong, and Hoifung Poon. “Document-Level N-ary Relation Extraction with Multiscale Representation Learning”. In: *CoRR* abs/1904.02347 (2019). arXiv: `1904.02347`. URL: `http://arxiv.org/abs/1904.02347`.

[16] Jinhyuk Lee et al. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.

[17] Jake Lever et al. “Text-mining clinically relevant cancer biomarkers for curation into the CIViC database”. In: *bioRxiv* (2018). DOI: `10.1101/500686`. eprint: `https://www.biorxiv.org/content/early/2018/12/20/500686.full.pdf`. URL: `https://www.biorxiv.org/content/early/2018/12/20/500686`.

[18] Marilyn M. Li et al. “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”. In: *The Journal of Molecular Diagnostics* 19.1 (Jan. 1, 2017), pp. 4–23. ISSN: 1525-1578. DOI: `10/f9jwh7`. URL: `https://doi.org/10.1016/j.jmoldx.2016.10.002` (visited on 12/20/2021).

[19] Leland McInnes and John Healy. “Accelerated Hierarchical Density Based Clustering”. In: *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*. IEEE. 2017, pp. 33–42.

[20] Leland McInnes, John Healy, and Steve Astels. “hdbscan: Hierarchical density based clustering”. In: *The Journal of Open Source Software* 2.11 (2017), p. 205.

[21] Leland McInnes et al. “UMAP: Uniform Manifold Approximation and Projection”. In: *The Journal of Open Source Software* 3.29 (2018), p. 861.

[22] *My Cancer Genome*. `https://www.mycancergenome.org/`. Accessed: 2021-12-20.

[23] Tiago Pimentel et al. “Information-Theoretic Probing for Linguistic Structure”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 4609–4622.

[24] Damian T. Rieke et al. “Comparison of Treatment Recommendations by Molecular Tumor Boards Worldwide”. In: *JCO Precision Oncology* 2 (2018), pp. 1–14. DOI: `10.1200/PO.18.00098`. eprint: `https://doi.org/10.1200/PO.18.00098`. URL: `https://doi.org/10.1200/PO.18.00098`.

[25] Alexander Rives et al. “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences”. In: *Proceedings of the National Academy of Sciences* 118.15 (2021).

[26] Jurica Ševa, Martin Wackerbauer, and Ulf Leser. “Identifying Key Sentences for Precision Oncology Using Semi-Supervised Learning”. In: *Proceedings of the BioNLP 2018 workshop*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 35–46. DOI: `10.18653/v1/W18-2305`. URL: `https://aclanthology.org/W18-2305`.

[27] Hoo-Chang Shin et al. “Bio-Megatron: Larger Biomedical Domain Language Model”. In: *EMNLP*. 2020.

[28] Ayush Singhal, Michael Simmons, and Zhiyong Lu. “Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine”. In: *PLoS Computational Biology* 12.11 (Nov. 2016), e1005017. DOI: `10.1371/journal.pcbi.1005017`.

[29] Alex H. Wagner et al. “CIViCpy: A Python Software Development and Analysis Toolkit for the CIViC Knowledgebase”. In: *JCO Clinical Cancer Informatics* 4 (2020). PMID: 32191543, pp. 245–253. DOI: `10.1200/CCI.19.00127`. eprint: `https://doi.org/10.1200/CCI.19.00127`. URL: `https://doi.org/10.1200/CCI.19.00127`.

[30] Hai Wang and Hoifung Poon. “Deep Probabilistic Logic: A Unifying Framework for Indirect Supervision”. In: *CoRR* abs/1808.08485 (2018). arXiv: `1808.08485`. URL: `http://arxiv.org/abs/1808.08485`.

Supplementary

Supplementary Methods

Downloading the data

The data was downloaded via CIViC API using 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 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 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 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 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 ML 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 38-53% 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 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 former, since it is used in most sentence level transformers pre-training tasks such as sentence classification and next sentence prediction. BioBERT model uses 768-dimensions output vector, while BioMegatron uses 1024-dimensions.

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

As shown in Eq. 1, f_θ^{TRF} is last-layer output function of 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 further 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, BioBERT , and 1024 dimensions for BioMegatron. We used UMAP for dimensionality reduction and HDBSCAN clustering algorithm to identify patterns in an unsupervised manner.

Supplementary Tables

Table S.1: Spearman correlations between the classification error and the # pairs in the training set where an entity occurs. E.g. For BioBERT, there is a significant negative correlation between # of drug-gene pairs that a drug entity occurs in the training set 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 S.2: Unique entities in quadruples from Task 2.

Dataset		**n unique**			
	n quadruples	**variant**	**gene**	**diasease**	**drug**
All	2989	1015	302	215	733
Training set	2092	803	258	186	579
Test set	897	432	165	135	339
Balanced set	207	147	67	43	89

Table S.3: 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 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 S.4: AUCs and Brier scores for the balanced test set stratified by evidence level. KNN performs significantly worse for evidence level D compared to 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 S.5: 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 S.6: 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

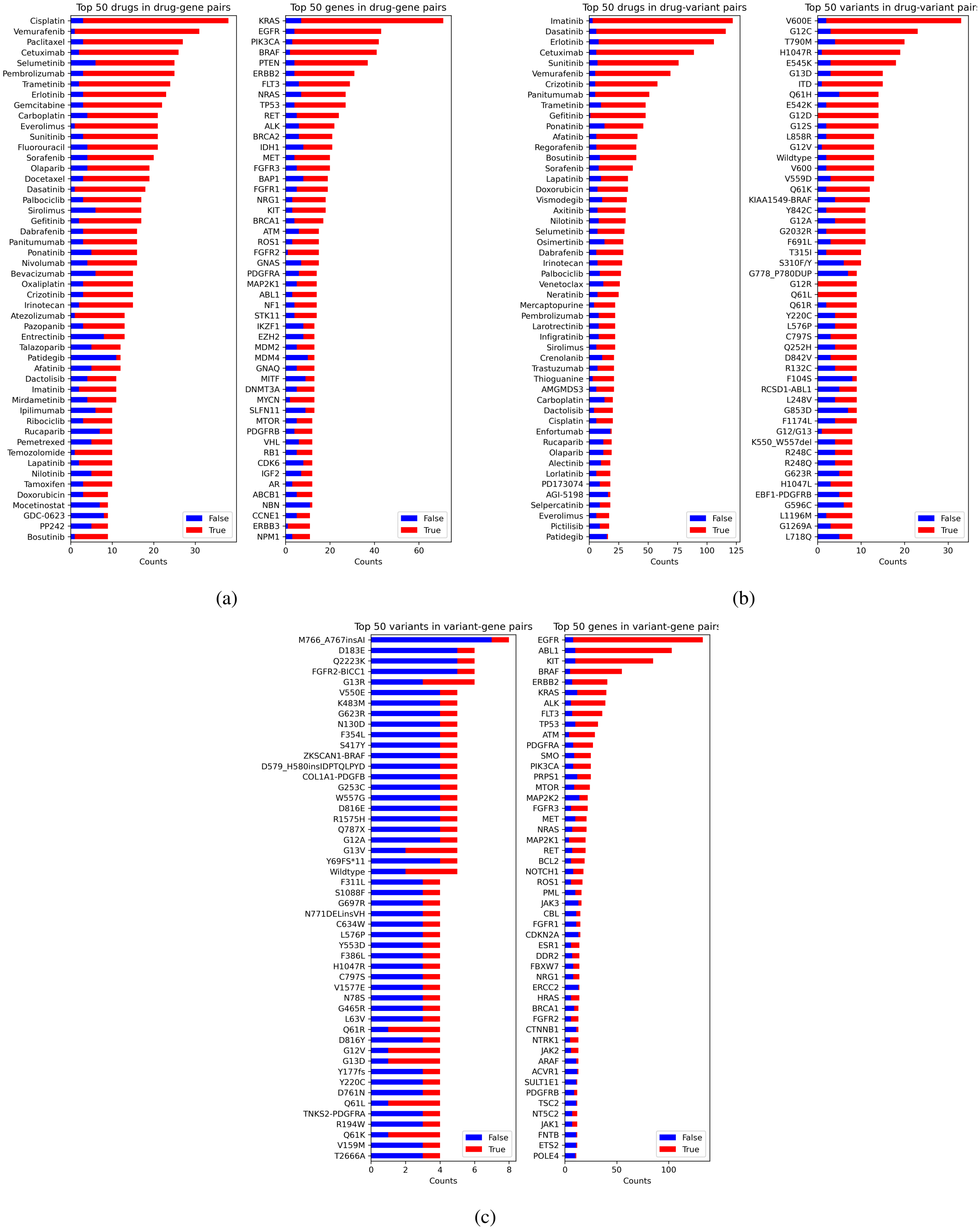
Table S.7: BioBERT quadruples from clusters #5 and #6. No obvious patterns. R stands for Resistance and S/R is for Sensitivity/Response.

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 S.8: 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

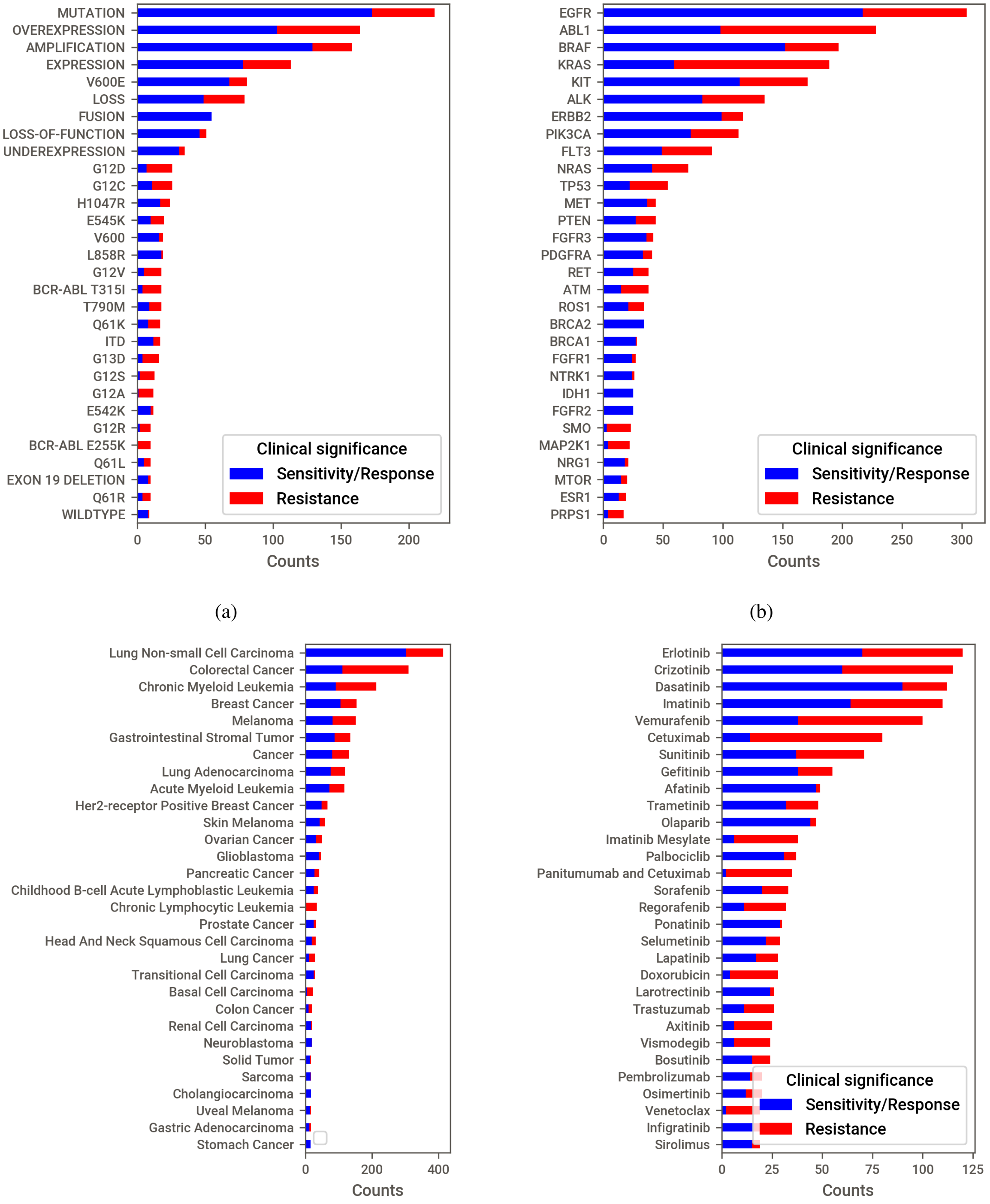


(a)

(b)

(c)

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