Generative Adversarial Networks in Precision Oncology

Leandro von Werra
ETH Zurich
Switzerland

Marcel Schöngens
ETH Zurich
Switzerland

Ece D. Gamsiz Uzun
Brown University
USA

Carsten Eickhoff
Brown University
USA

ABSTRACT

Precision medicine strives to deliver improved care based on genetic patient information. Towards this end, it is crucial to find effective data representations on which to perform matching and inference operations. We develop and evaluate a generative adversarial neural network (GAN) approach to representation learning with the goal of patient-centric literature retrieval and treatment recommendation in precision oncology. Several large-scale corpora including the COSMIC Cancer Gene Census, COSMIC Mutation Data, Genomic Data Commons (GDC) and 26M MEDLINE abstracts are used to train GANs for synthesizing genetic mutation patterns that likely correspond to patient properties such as their demographics or cancer type. The introduction of GANs into the literature retrieval and treatment recommendation process results in significant improvements in performance by increasing the recall of a range of methods at stable precision. Finally, we propose a method to discover novel gene-gene interaction hypotheses to guide future research.

KEYWORDS

Deep Learning; GAN; Precision Medicine; Oncology

ACM Reference Format:
Leandro von Werra, Marcel Schöngens, Ece D. Gamsiz Uzun, and Carsten Eickhoff. 2019. Generative Adversarial Networks in Precision Oncology. In *The 2019 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '19), October 2–5, 2019, Santa Clara, CA, USA*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3341981.3344238

1 INTRODUCTION

Precision medicine is an active field of study that aims to use genetic information in finding effective personalized treatments for patients. The term "precision oncology" is used to describe diverse strategies in cancer medicine ranging from the use of targeted therapies in general to employing data from next-generation sequencing to select therapy for a person independent of cancer type. Due to the popularity of the novel paradigm, the volume of annually published scholarly precision oncology articles has been growing rapidly in recent years. While this considerable amount of scientific research holds a rich and ever increasing well of knowledge, its sheer scale makes it intractable for manual inspection and mandates the development of dedicated automatic retrieval and reasoning facilities [13, 17]. In this context, existing methods [12, 16] rely on exact term matching of gene names in patient records and literature. While offering high precision, these approaches are known to miss many potentially relevant matches that are described using different, synonymous or related wording.

This paper proposes to address this issue by using a generative adversarial network (GAN) [7] in the precision oncology domain. The strength of the neural network approach lies in its ability to generalize from term identity (*e.g.*, individual gene names) to term semantics (*e.g.*, classes of genes with comparable functional properties) [9, 18]. Based on patient demographics as well as information regarding the type of tumor and its genetic composition, we investigate three use cases of key clinical relevance: (1) We rank scholarly articles as well as clinical trials with respect to their relevance for the patient. (2) We propose the most promising means of treatment for the patient's condition. (3) We discover new gene-gene interactions that hold significant potential for prospective investigation but that are not currently discussed in the literature.

Our experiments are based on all 26 million Medline abstracts, the COSMIC [4] and GDC [10] databases for model training as well as 30 synthetic oncology patient descriptions from MD Anderson Cancer Center for performance evaluation [15]. The results suggest a significant improvement in retrieval and treatment suggestion performance over state-of-the-art term-based models while, at the same time, enabling altogether new use cases such as the discovery of previously unseen gene-gene interactions.

The remainder of this paper is structured as follows: After listing the used data sources in Section 2, we begin by discussing the formal background of the GAN framework as well as our proposed literature retrieval, treatment prediction and interaction discovery schemes. Subsequently, in Section 4, we assess their relative performance on a set of held-out test patient cases. Finally, we conclude by discussing the observed merit of the method as well as its implications for future inquiry.

2 DATASETS

TREC Precision Medicine. In collaboration with precision oncologists at MD Anderson Cancer Center, the TREC 2017 Precision Medicine track [15] made 30 synthetic oncology patients as well as their demographic and genetic information available in the context of a patient-centric retrieval benchmarking effort. For these 30 patients, a total of 22,642 Medline abstracts and 13,441 clinical trial descriptions have been manually annotated in terms of relevance

ICTIR '19, October 2–5, 2019, Santa Clara, CA, USA

ACM ISBN 978-1-4503-6881-0/19/10…$15.00
https://doi.org/10.1145/3341981.3344238

and suitability for clinical use. These patients form the centerpiece of our investigation for which the methods developed in the following sections aim at recommending specifically personalized literature. The full database of reference patients is available from (http://www.trec-cds.org/).

COSMIC Cancer Gene Census. In order to study gene similarities, we need to first identify a relevant range of genes to be included in our system. The Catalogue Of Somatic Mutations In Cancer (COSMIC) curates the Cancer Gene Census (CGC), a list of genes related to cancer genesis [4, 6]. The CGC lists official names of cancer genes as well as a list of synonyms which we use to extract gene occurrences from raw text. In total, 614 cancer-related genes are listed, leading to 188,191 possible gene pairings.

MEDLINE Abstracts. To further reduce this theoretical maximum to the number of gene pairs, we parse all scientific abstracts in the Medline database for cancer gene occurrences via the gene2pubmed list [14]. We count gene co-occurrences in this list and empirically define a minimal occurrence cut-off ($n_{cutoff} = 10$) resulting in 9,500 unique pairs of genes that are retained for model training.

COSMIC Mutation Data. As a second source of gene co-occurrence data, we use the COSMIC mutation data, a dataset with DNA screens of tumor cells [3]. Similarly to the Medline articles, we count co-occurrences in patient DNA screens and set an occurrence cutoff ($n_{cutoff} = 50$) for gene pairs, yielding another 13,400 training pairs.

GDC Database. Finally, in order to associate genetic variation to effective treatments, we access the Genomic Data Commons (GDC) database curated by the National Cancer Institute [10]. This database contains information on over 30,000 cancer patients. Through several APIs, we access information on age, gender, mutations and treatments, including information on administered drugs, which yields data on approximately 6,000 cancer patients with information across all categories.

3 METHODS

A GAN consists of two neural networks, a generator G and a discriminator D, which are connected via an adversary game. Given a dataset, the generator's goal is to mimic real data, while the discriminator aims at differentiating real from synthetic data. First, we train a GAN on gene pairs which are likely to co-occur (either within a scholarly article or within the same patient). While the generator learns to produce plausible gene pairs, the discriminator learns to differentiate observed from synthetic gene pairs, which we then use to compare gene mutations in a patient record to those found in the literature. Alternatively, we use a conditional GAN for treatment prediction. In this configuration, the data is split into conditions (patient information) and targets (treatments) and the generator is tasked to output a suitable target for a given condition, whereas the discriminator assesses the compatibility of the condition with the target. Figure 1 schematically illustrates these processing pipelines.

3.1 Pairwise Gene Compatibility

This section presents a GAN method to discriminating possible gene pairs for estimating patient-document compatibility. We modify the existing architecture of the medGAN model [2] to learn gene mutation similarities from co-occurrences.

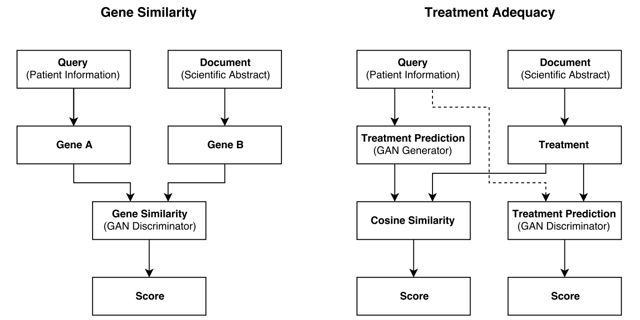
Gene Similarity
Query (Patient Information)
Document (Scientific Abstract)
Gene A
Gene B
Gene Similarity (GAN Discriminator)
Score
Treatment Adequacy
Query (Patient Information)
Document (Scientific Abstract)
Treatment Prediction (GAN Generator)
Treatment
Cosine Similarity
Treatment Prediction (GAN Discriminator)
Score
Score

Figure 1: Scoring schemes of patient-document pairs for gene similarity and treatment adequacy.

The discriminator and generator neural networks play a min-max game where the discriminator's goal is to assign scores of 1 to real data $\mathbf{x}_{real}$ and 0 to synthetic data $\mathbf{x}_{fake}$. Meanwhile, the generator aims at deceiving the discriminator. This game is translated to loss functions used for generator and discriminator network training as follows:

$$loss_G = -\log(D(\mathbf{x}_{fake})) \text{ and}$$
$$loss_D = -\log(D(\mathbf{x}_{real})) - \log(1 - D(\mathbf{x}_{fake})).$$

Our goal is to measure the similarity of two gene mutations with the discriminator's valuation of the pair. The training data, therefore, encodes mutation pairs as two-hot vectors with dimensions corresponding to the number of considered genes:

$$\mathbf{x}(\text{gene}_i, \text{gene}_j) = (0, \ldots, 0, \underset{i}{1}, 0, \ldots, 0, \underset{j}{1}, 0, \ldots, 0) \in \mathbb{R}^{n_{genes}}.$$

To avoid mode collapse [7], the generator is configured to output an entire batch of samples at each training step. The discriminator is exposed to one sample along with the batch average at a time forcing the generator to establish a data distribution similar to real data, which prevents mode collapse. However, now the discriminator critically depends on the disclosure of batch averages. For the task of evaluating gene pairs from queries and documents, there is no sensible definition of such batch averages. As a consequence, in addition to batch-training, we introduce an online training scheme. In this mode, the discriminator receives only a single sample from the batch with the batch average set to zero along with a binary flag indicating the training mode. We train the GAN with both modes in parallel by adding the losses, such that:

$$loss = loss_{online} + loss_{batch}.$$

Empirically, we find that a discriminator/generator training ratio of 1/5 to yield reliable results. With this setup, the discriminator trains once, while the generator is trained five times. In order to improve stability, we apply shortcut-connections to the generator and L2-regularization losses as well as dropout to the autoencoder and discriminator. Finally, we introduce a metric to rank patient-document pairs. For this purpose, we denote patients by p and documents by d and look for mentions of genes g in both. We introduce four aggregation scores:

$$\begin{aligned}
s_{mean,id} &= \text{mean}(s) \quad D(\mathbf{x}(g_i, g_i)) = 1,\\
s_{mean} &= \text{mean}(s) \quad D(\mathbf{x}(g_i, g_i)) = 0,\\
s_{min} &= \min(s) \quad\;\; D(\mathbf{x}(g_i, g_i)) = 0 \text{ and}\\
s_{max} &= \max(s) \quad\;\; D(\mathbf{x}(g_i, g_i)) = 0, \text{ where}\\
s &= \{D(\mathbf{x}(g_p, g_d)) \mid \forall g_p \in p, \forall g_d \in d\}
\end{aligned}$$

is the set of pairwise similarities between genes in the query and genes in the document. We vary the value for self-similarity ($g_p = g_d$) to incorporate exact and semantic term matching separately. We summarize these scores in an aggregate vector $\mathbf{s}_{gene}$ for later use

$$\mathbf{s}_{gene} = (s_{mean,id}, s_{mean}, s_{min}, s_{max}).$$

3.2 Treatment Adequacy

In addition to gene expression similarity, we can further model the adequacy of treatments mentioned in the literature for the patient at hand. In this setting, we use the generator and discriminator separately resulting in two scores, following the scheme introduced earlier in this section. The conditional GAN losses are obtained by extending the discriminator's and generator's arguments with the condition vector $\mathbf{c}$, corresponding to the patient information vector. In this setup, we pre-train the generator like a standalone feed-forward network to predict treatments from patient information and the discriminator to differentiate between real and synthetic vectors. While we apply the main training loss to the discriminator, the generator loss during pre-training (with noise vector $\vec{z}$ set to 0) is given by the squared difference between output and label:

$$loss_{G,pre} = \|\mathbf{x} - G(\mathbf{0}, \mathbf{c})\|^2 .$$

In contrast to the gene similarity GAN, the conditions $\mathbf{c}$ and target vectors $\mathbf{x}$ are no longer restricted to two-hot vectors, but can contain real numbers for patient information (*e.g.*, age, weight, *etc.*) or n-hot entries in case of treatment categories (one patient can receive several treatments). Furthermore, the condition restricts the generator sufficiently, such that mode collapse poses no threat and batch mode training can be omitted.

We use the generator and discriminator separately to score patient and document pairs. First, we compare the generator output conditioned on the patient information in the query with the document treatment vector using cosine similarity to obtain a generator score. Second, we use the discriminator to evaluate the query condition and the document treatment vector for compatibility. The two scores can be written as:

$$\begin{aligned}
s_{gen} &= \cos(G(\mathbf{z}, \mathbf{c}_q), \mathbf{x}_d) \text{ and}\\
s_{disc} &= D(\mathbf{c}_q, \mathbf{x}_d).
\end{aligned}$$

As a final step, we combine both scores in a score vector $\mathbf{s}_{treat}$ for later score fusion:

$$\mathbf{s}_{treat} = (s_{gen}, s_{disc}).$$

3.3 Score Fusion

We describe the process of fusing all obtained scores into a single value used to rank documents with the help of manually annotated query document pairs. The objective is to maximize the commonly applied normalized Discounted Cumulative Gain (*nDCG*) metric [11]. To leverage exact term matching, we enrich the GAN scores with the score of a vector space model (VSM) based on TF-IDF scores. As a first step towards document scoring, we combine all score values into a single vector

$$\mathbf{s} = s_{vsm} \oplus \mathbf{s}_{gene} \oplus \mathbf{s}_{treat}.$$

Then we fuse all scores in a weighted sum with a weight vector $\mathbf{w}$ that maximizes the *nDCG* score

$$\begin{aligned}
s_{fused} &= \mathbf{s} \cdot \mathbf{w}, \text{ where}\\
\mathbf{w} &= \arg\max_{\mathbf{w}} \left[nDCG(s_{fused}) \right] .
\end{aligned}$$

4 RESULTS

The literature retrieval performance of the presented models is assessed on TREC 2017 precision medicine data. Under this task, 30 model oncology patients are to be associated with a ranking of relevant biomedical literature that can help in devising treatments, as well as with clinical trials for which the patient is eligible. Following the original benchmark's guidelines, we measure the performance of each run in terms of inferred nDCG [19], precision at 10 and 30 retrieved documents, R-precision for abstract retrieval and precision at 5, 10, 15 and 30 retrieved documents for clinical trial retrieval. In both settings, we also report recall at 10 retrieved documents. We compare the standard TF-IDF VSM to a variant enriched with GAN scores.

The results in Table 1 show that, in both settings and across all evaluation metrics, additional GAN scores considerably improve the absolute retrieval performance. Due to the low sample size of only 30 instances, statistical significance of performance differences (measured by a paired t-test at $p < 5\%$ and denoted by an asterisk) can only be attained for some of the metrics.

Table 1: Patient-centric retrieval results.

Abstracts	**infNDCG**	**R-prec**	**P@10**	**P@30**	**S@10**
VSM	0.3085	0.2130	0.3700	0.3100	0.0380
VSM+GAN	0.3430*	0.2281*	0.4000	0.3311	0.0484

Trials	**P@5**	**P@10**	**P@15**	**P@30**	**S@10**
VSM	0.3867	0.3333	0.3111	0.2600	0.1452
VSM+GAN	0.3933	0.3667	0.3667*	0.2544	0.1718*

Turning from literature retrieval to a core classification task, we assess the method's performance at associating cancer types and genomic variants with their appropriate treatments. From the GDC database, we retrieve two condition datasets (patient information with and without mutation information) and two treatment datasets (with and without administered medication). We examine for all combinations of datasets the ability of the model to predict the correct treatments. The results are presented in Table 2. Within each setting, statistical significance between a standard term-based single-class support vector classifier and an alternative that additionally has access to GAN output is measured by a paired t-test at $p < 5\%$ and denoted by an asterisk. While we observe reasonable performance at predicting general treatments such as radiation, chemotherapy or surgery, the task of also predicting administered drugs is difficult with the available training data. In this difficult setting, the introduction of GAN information increases recall by up to 90% while retaining undiminished precision. In the easier setting of general treatment prediction, this effect is negligible.

Table 2: Experimental results of treatment prediction on a held-out validation set. Term-based SVC compared to a model additionally enriched by GAN output.

Condition	Target	Model	Recall	Precision
Pat. Info.	Treat. +	SVC	0.122	0.985
	Med.	SVC+GAN	0.209*	0.972
Pat. Info. +	Treat. +	SVC	0.135	0.983
Genes	Med.	SVC+GAN	0.208*	0.969
Pat. Info.	Treat.	SVC	0.481	0.866
		SVC+GAN	0.474	0.852
Pat. Info. +	Treat.	SVC	0.505	0.856
Genes		SVC+GAN	0.508	0.839

Up to this point, we have been using the GAN primarily as a means of estimating gene similarity or treatment compatibility. We can, however, also use the discriminator to explore previously unseen, yet likely, co-occurrence patterns of gene mentions. Recall that we employed an inclusion cut-off to gene co-occurrence frequency. Inspecting the discriminator rating of gene pairs beyond the cut-off frequency in the COSMIC dataset, we observe that "false-positive" pairs are gathered closely to the cut-off as presented in Figure 2.

In other words, the unseen pairs suggested by the discriminator are plausible and, in fact, have occurred in authentic patients but were merely excluded from the training process due to their lower prevalence. In the following, we will briefly discuss three concrete examples of gene pairs that were never present in the method's training data but that have nevertheless correctly been discovered.

BRCA1/WRN. BRCA1 encodes a well-studied breast cancer susceptibility protein, the production of which has been shown to be affected by the presence of the Werner Syndrome encoded on WRN.

SETD2/EZH2. Both of these are histone-modifying genes that interact in the expression of lymphoma.

COL1A1/FBXW7. While at first glance the osteogenesis-related COL1A1 and FBXW7 (primarily associated with ovarian and breast cancer) are unconnected, recent research has found overexpression of FBXW7 in C2C12 cells to result in down-regulation of Col1A1 mRNA.

While these examples are of mostly anecdotal value, they highlight the potential of generative models in gene interaction hypothesis generation.

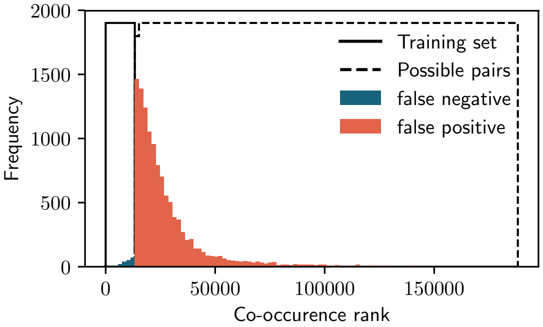

Figure 2: False positives classification of a discriminator trained on the COSMIC dataset.

5 CONCLUSION

This paper presents an adversarial learning approach to information retrieval in precision oncology[1]. Using data from four databases, we train two GAN architectures for the task of scoring patient-document pairs in terms of similarity of genetic expression as well as adequacy of discussed treatments.

Throughout our experiments with the proposed GAN architecture, we see consistent performance improvements in terms of recall. Both in the literature retrieval [1, 5, 8] as well as treatment prediction applications precision is retained at unchanged levels while recall significantly increases when introducing neural network output.

Clearly, the governing limitations of the present early-stage investigation are collection size as well as depth. Especially in case of the interaction discovery scenario, a prospective study is required to truly explore the accuracy and informativeness of novel interaction candidates. While this particular application is an early outlook, it holds considerable potential for identifying promising new directions for future inquiry, demonstrating a paradigm under which generative machine learning methods can help to chart the vast decision spaces faced by clinical researchers.

REFERENCES

[1] Prakrit Baruah, Riya Dulepet, Kyle Qian, and Carsten Eickhoff. Brown university at trec precision medicine 2018. In *TREC'18*. NIST.
[2] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Machine Learning for Healthcare Conference*, 2017.
[3] Simon A. Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G. Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, Raymund Stefancsik, Bhavana Harsha, Chai YinKok, Mingming Jia, Harry Jubb, Zbyslaw Sondka, Sam Thompson, Tisham De, and Peter J. Campbell. COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45, 2017.
[4] Simon A Forbes, David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, Sally Bamford, Charlotte Cole, Sari Ward, et al. Cosmic: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research*, 43, 2014.
[5] Negar Foroutan Eghlidi, Jannick Griner, Nicolas Mesot, Leandro von Werra, and Carsten Eickhoff. Eth zurich at trec precision medicine 2017. In *TREC'17*. NIST.
[6] P Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177, 2004.
[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS'14*.

[1] Code base available at https://github.com/lvwerra/geneGAN.

[8] Simon Greuter, Philip Junker, Lorenz Kuhn, Felix Mance, Virgile Mermet, Angela Rellstab, and Carsten Eickhoff. Eth zurich at trec 2016 clinical decision support. In *TREC'16*. NIST.

[9] Paulina Grnarova, Florian Schmidt, Stephanie L Hyland, and Carsten Eickhoff. Neural document embeddings for intensive care patient mortality prediction. In *NIPS'16 ML4Health Workshop*.

[10] Robert L Grossman, Allison P Heath, Vincent Ferretti, Harold E Varmus, Douglas R Lowy, Warren A Kibbe, and Louis M Staudt. Toward a shared vision for cancer genomic data. *NEJM*, 375(12):1109–1112, 2016.

[11] Kalervo Järvelin, Susan L Price, Lois ML Delcambre, and Marianne Lykke Nielsen. Discounted cumulated gain based evaluation of multiple-query ir sessions. In *ECIR'08*. Springer.

[12] Thorsten Joachims. Optimizing search engines using clickthrough data. In *KDD'02*. ACM.

[13] Lorenz Kuhn and Carsten Eickhoff. Implicit negative feedback in clinical information retrieval. In *SIGIR'16 Medical Information Retrieval Workshop*.

[14] Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 39, 2010.

[15] Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, W Hersh, Steven Bedrick, Alexander J Lazar, and Shubham Pant. Overview of the trec 2017 precision medicine track. *TREC'17*.

[16] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *Nist Special Publication Sp*, 109, 1995.

[17] Xing Wei and Carsten Eickhoff. Distant supervision in clinical information retrieval. In *AMIA'18*.

[18] Xing Wei and Carsten Eickhoff. Embedding electronic health records for clinical information retrieval. In *https://arxiv.org/abs/1811.05402*, 2018.

[19] Emine Yilmaz, Evangelos Kanoulas, and Javed A Aslam. A simple and efficient sampling method for estimating ap and ndcg. In *SIGIR'08*. ACM.