Learning to Ground Medical Text in a 3D Human Atlas

Dusan Grujicic*, Gorjan Radevski*, Tinne Tuytelaars, Matthew B. Blaschko
Department of Electrical Engineering (ESAT-PSI)
KU Leuven
firstname.lastname@esat.kuleuven.be

Abstract

In this paper, we develop a method for grounding medical text into a physically meaningful and interpretable space corresponding to a human atlas. We build on text embedding architectures such as BERT and introduce a loss function that allows us to reason about the semantic and spatial relatedness of medical texts by learning a projection of the embedding into a 3D space representing the human body. We quantitatively and qualitatively demonstrate that our proposed method learns a context sensitive and spatially aware mapping, in both the inter-organ and intra-organ sense, using a large scale medical text dataset from the "Large-scale online biomedical semantic indexing" track of the 2020 BioASQ challenge. We extend our approach to a self-supervised setting, and find it to be competitive with a classification based method, and a fully supervised variant of approach.

1 Introduction

The quantity of available medical literature increases daily (Wang et al.; Tsatsaronis et al., 2015), however, it is often provided in a non-systematized, free form. The development of BERT (Devlin et al., 2018), and the increased popularity of transfer learning in natural language processing (NLP), prompted notable works that aim to leverage publicly available medical and scientific articles to develop domain specific pre-trained language models (Lee et al., 2019; Alsentzer et al., 2019; Beltagy et al., 2019; Jin et al., 2019). High quality sentence representations that capture the semantics and structure of the text can be obtained by training models to solve the Natural Language Inference (NLI) task on open-domain datasets (Bowman et al., 2015; Williams et al., 2017) and predict whether two pieces of text entail, contradict or are neutral to each other (Conneau et al., 2017). The aforementioned BERT-based models can serve as the encoder backbone for such approaches (Reimers and Gurevych, 2019), and the setup can be trivially extended to enable searching through and retrieving relevant documents from large datasets. Despite proving useful in a variety of settings, these works suffer from the following limitations:

(i) The documents are embedded in a non-interpretable space. (ii) There is no clear visually intuitive indication of how similar two retrieved documents are, i.e., black box retrieval. (iii) Visualizing the embeddings requires dimensionality reduction techniques (Hotelling, 1933; Maaten and Hinton, 2008).

By contrast, we propose a method that embeds medical text into a universal, small dimensional space corresponding to the human body that is easy to navigate and interpret (Figure 1). The propensity of functionally similar organs towards being physically close represents an inductive bias that can be leveraged for computing compact, 3D text representations that are competitive with standard higher dimensional text embeddings. Additionally, our approach allows to search through and retrieve documents grounded within the physical space of the human body. To that end, our contributions are:

Figure 1: Given the text with implicit reference to lungs: "Divided into two lobes, an upper and a lower lobe, by the oblique fissure, which extends from the costal to the mediastinal surface" (Drake et al., 2009), our model learns the grounding indicated by the star.

*Equal Contribution

(i) We propose the task of the grounding medical text in the physical space of the human body, where anatomically related substructures tend to be close to one another. (ii) We develop a loss function that allows us to reason about the semantic relatedness of medical texts. (iii) We develop a concrete use-case for medical text retrieval where we outperform several competitive baselines.

We perform extensive evaluation to measure the performance of our method in two scenarios, namely, grounding in the human atlas (relevant for visualization and navigation), and medical text-to-text retrieval (directly assessing the performance of our model in an information retrieval setting). Furthermore, we set up an experimental setting explicitly tailored to measure the spatial reasoning ability of our model within an organ, a setting never directly imposed during training. We empirically demonstrate that our method is highly successful in all aforementioned experimental settings, effectively addressing the previously stated limitations. The codebase and the trained models are released at: `www.github.com/gorjanradevski/text2atlas`

2 Related work

(Medical) text embeddings. Before the development of BERT, a common approach to embedding text was leveraging a pre-trained recurrent neural network (RNN) language model (LM) (Peters et al., 2018; Kiros et al., 2015). An extension of such LM for the biomedical domain is BioELMO (Jin et al., 2019). Despite being successful in a transfer learning setting, the usefulness of the generated embeddings for medical text navigation and retrieval is arguably limited. Furthermore, BERT-based medical language representation models such as BIOBERT (Lee et al., 2019) and CLINICALBERT (Alsentzer et al., 2019), despite outperforming RNN based LMs on a variety of downstream tasks, also make embedded text navigation impractical. We, on the other hand, directly focus on learning embeddings that are rich with visual information, i.e., are by default represented in a physically meaningful space of the human body.

(Medical) text grounding. There has been a variety of approaches (Krishnamurthy and Kollar, 2013; Kong et al., 2014; Rohrbach et al., 2016; Hu et al., 2016; Wang et al., 2018) and datasets, such as ReferIt (Kazemzadeh et al., 2014) and RefCOCO (Yu et al., 2016), focusing on the visual grounding of natural language in the general domain. However, the application of text grounding in the medical domain has been limited, and to the best of our knowledge, there are no works that ground medical text in the human body. The main differences between these works and ours are: (i) we perform a grounding which is universal, and not specific to a single environment (e.g. the image), (ii) their models are trained with expensive bounding box annotations for the desired grounding location, (iii) the methods rely on explicit annotations of every concept referred in the text, i.e., these models can not reason about the particular referred region unless explicitly trained to do so. Furthermore, our method is designed to reduce the labeling costs, as it relies on high level annotations of the referred organs in a paragraph, which, in a self-supervised setting, can be inferred from the text itself.

(Medical) Document retrieval. Retrieving a set of relevant documents given a query requires that both the query and the documents are embedded in a joint latent space. A straight-forward approach to obtaining a single text representation is to use the `[CLS]` token representation concatenated with the mean-pooled and max-pooled representations of the remaining tokens from a pre-trained BERT model. However, it is shown that this often leads to a worse representation than averaging GloVe embeddings (Pennington et al., 2014; Reimers and Gurevych, 2019). Recently, such embeddings are obtained using a pre-trained BERT subsequently fine-tuned as a Siamese model (Reimers and Gurevych, 2019) on the NLI task using general domain datasets (Bowman et al., 2015; Williams et al., 2017). The proxy-task is proven to be effective as models trained this way generate embeddings in which documents that share similar semantics map nearby. Despite this useful feature, without inspecting the documents' content, it is not immediately obvious why a set of documents is clustered together in the latent space, and why they are considered to be semantically similar. We address this issue by embedding documents in the physical space of the human body, where the document similarity is expressed in terms of physical proximities in 3D. This leads to an intuitive interpretation of why a set of documents are considered to be similar.

3 Data collection

3.1 Human body atlas

We leverage the **S**egmented **I**nner **O**rgans (SIO) (Pommert et al., 2001) (see Appendix, Figure A.1), though the approach is readily extended to other models of the human anatomy. We refer to this 3D model as the *atlas*. We base the 3D atlas on the segmentation labels of the tissues in the human body provided in SIO, which come in the form of image slices that form a 3D voxel model of the male torso when stacked on top of one another. The stacked images from the torso represent a volume of $573 \times 330 \times 774$ voxels, with 1-millimeter resolution along each axis. The value of each voxel represents the segmentation label of its corresponding organ or tissue. An organ can be represented as the set of indices of voxels in the aforementioned volume which contain the value corresponding to the organ's segmentation label.[1]

3.2 Dataset

The dataset used in this work is built upon the training set of the Task 8a: "Large-scale online biomedical semantic indexing" of the 2020 BioASQ challenge (Tsatsaronis et al., 2015). Originally, it consists of 14,913,939 samples, where each sample pertains to one medical article, and contains the abstract text and the Medical Subject Headings (MeSH) (Lipscomb, 2000) vocabulary terms of the organs. We consider the grounding of article abstracts to the locations in the atlas that correspond to the article MeSH terms. Therefore, we use the articles that contain one or more MeSH terms that match the names or the alias terms of the organs in the atlas glossary. To accommodate the maximal sequence length of $\text{BERT}_{\text{BASE}}$, we keep the articles whose abstracts have fewer than 512 WordPiece (Wu et al., 2016) tokens. For each organ in the atlas glossary, we take 500 articles that mention it individually, and another 500 articles that mention it in addition to another organ(s). Subsequent removal of duplicates resulted in the final dataset of 25,552 abstracts annotated with organ MeSH terms, of which 70% are used for training, 15% for validation and 15% for testing.[2]

[1] Details about the creation of the 3D human atlas can be found in the Appendix Section A.

[2] Additional details about the dataset found in the Appendix Section B.

4 Proposed task and methods

4.1 Text-to-atlas grounding objective

Our goal is to ground medical texts into the 3D space of the human body. To achieve this, we project the representations of text referring to one or more atlas organs into the 3D volume in the atlas that corresponds to the mentioned organs. The volume of each organ is characterized by a set of voxels in the atlas, which capture its position, size and shape. The voxels of one organ can, in turn, be represented by a point cloud in 3D space, where each point represents the coordinate indices of one voxel[3]. The most straightforward way to associate texts with predefined regions of a physical space, is to have a model trained to simply minimize the cross-entropy between the predicted probability distribution over the set of all organs (e.g., each indexing their predefined location), and the target vector with 1's at the indices corresponding to the target organs and 0's elsewhere. This approach is expected to yield a high accuracy of selecting the right organ, however, has a critical downside of not providing any meaningful within organ reasoning, i.e., during inference, it grounds all articles pertaining to a single organ to either a random location within the organ, or a single predefined one. On the other hand, framing the task as minimization of the mean squared error between the predicted 3D location and an average of the ground truth organ positions would result in a grounding to some midway location, potentially belonging to some other, unrelated organ. To overcome both of these issues, and retain as much of the predictive power as possible, we frame the task as predicting a set of 3D coordinates within the human body, while forcing the prediction to snap to the most nearby ground truth organ. Namely, we design a loss function – Soft Organ Distance loss, henceforth abbreviated as SOD (Section 4.3), which gives the model freedom to choose the most relevant organ in case there are multiple organ annotations for a particular sample.

4.2 Model

We use BERT (Devlin et al., 2018) as our model backbone due to its applicability in a wide range of domains. As per Devlin et al. (2018), we tokenize the input text using WordPiece (Wu et al., 2016), and take the representation of the `[CLS]` token as the sequence representation. Finally, to obtain the

[3] For brevity, we will also use the term voxel or organ point for a vector of indices of the actual voxel in the 3D volume.

3D atlas grounding for a piece of medical text, we project BERT's output with a linear layer, mapping from BERT's hidden space to the 3D space:

$$\hat{y} = \text{Linear}(\text{BERT}(x)), \quad (1)$$

where x is a vector of tokens representing the medical text and $\hat{y}$ is the 3D grounding in the human body. During training, we normalize $\hat{y}$ by applying $\tanh$, which is subsequently rescaled to the dimensions of the atlas during inference.

4.3 Soft Organ Distance loss

The proposed loss function – SOD, allows us to sacrifice the least amount of predictive power and in turn, achieve within organ contextual reasoning, i.e., not only grounding the medical article to the right organ but also to the appropriate location within the organ without any explicit annotations at that level of granularity. Furthermore, a medical text may simultaneously refer to a single or multiple organs in the human body. In the former setting, we would like to have an approach based on mean squared error minimization, while in the latter, we would like to relax the target and pull the model's prediction to the location of the closest ground truth organ. Finally, the organs themselves are distributed in nature, and their volumes are characterized by a set of points in 3D space, rather than just one.

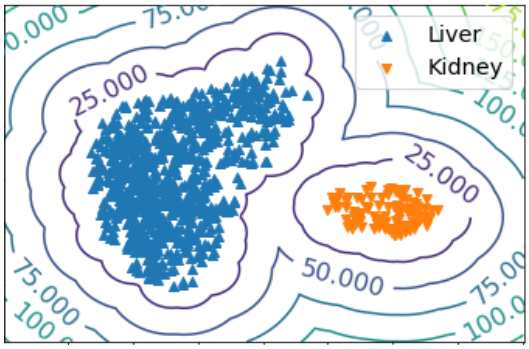


Figure 2: Loss isocurves around "liver" and "kidney" point clouds projected into 2D with PCA (Hotelling, 1933).

In Figure 2, we observe our desired scenario when there are two ground truth organs – "liver" and "kidney". As the grounding approaches the "liver", we observe that the loss contribution from the "kidney" organ voxels diminishes, and vice versa. This effect extends to the loss contributions of individual voxel points. Namely, as the grounding approaches a particular region in the organ, the loss contribution from the other voxel points diminishes – thus allowing the model to ground the input text within the most appropriate organ substructure.

In order to account for the distributed nature of the organs and take a step towards the desired within organ semantic reasoning, for each sample during training, we randomly sample a set of N points from the point cloud of each of its organs. Then, we calculate (1) the Euclidean distances between the prediction and each sampled organ point, and (2) the soft-min[4] across these distances as weights for the contributions of individual points. The loss contribution $\mathcal{L}_p$ of an organ point y is the product of its distance from the predicted point $\hat{y}$ and its corresponding weight produced by the soft-min:

$$\mathcal{L}_p = \|\hat{y} - y\|_2 \frac{\exp(-\|\hat{y} - y\|_2/\gamma_p)}{\sum_{i=1}^{N} \exp(-\|\hat{y} - y_i\|_2/\gamma_p)}, \quad (2)$$

where N is the number of points sampled from the organ point cloud and γ_p is a temperature term. We calculate the loss for one organ $\mathcal{L}_o$ as the sum of contributions of its points: $\mathcal{L}_o = \sum_{i=1}^{N} \mathcal{L}_p^i$.

We calculate the loss for each individual target organ in the way described above. Then, we compute the soft-min over the set of such loss terms as contribution weights for each organ. The total loss is the sum of soft-min-weighted losses over each organ:

$$\mathcal{L}_t = \sum_{i=1}^{M} \mathcal{L}_o^i \frac{\exp(-\mathcal{L}_o^i/\gamma_o)}{\sum_{j=1}^{M} \exp(-\mathcal{L}_o^j/\gamma_o)}, \quad (3)$$

where M is the total number of target organs, $\mathcal{L}_o^i$ is the organ loss for the i-th organ, and γ_o is a temperature term.

5 Experimental setup

We use $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2018) as the backbone of the trained models. We use AdamW (Loshchilov and Hutter, 2017) with a learning rate of 10^{-5} as per Devlin et al. (2018), weight decay of 10^{-2} and clip the gradients when the global norm exceeds 2.0. We perform early stopping by saving the model with the best performance on the validation set. We only tune the hyperparameters related to the SOD loss function, and we keep everything else fixed as per the standard practice

[4] Soft-max on the inputs reversed in sign, used to emphasize smaller quantities - in this case, shortest distances.

(Devlin et al., 2018). Our implementation utilizes PyTorch (Paszke et al., 2019) and the HuggingFace Transformers library (Wolf et al., 2019).

6 Evaluation

We quantitatively evaluate our trained models in two scenarios: (1) Grounding to the human atlas – measuring to what extent our trained model can ground medical articles to the correct location. (2) Medical information retrieval – to directly assess the quality of the document embeddings, i.e., evaluate to what extent medical articles characterized by a certain set of MeSH terms are grouped together in the physical space of the human body.

6.1 Grounding to the human atlas

To evaluate the quality of the grounding, we measure each of the models performance on three evaluation metrics (more details in Appendix C):

(1) Rate at which the texts are grounded within, or sufficiently close[5], to the volume of the correct organ, or the hit rate, which we denote as Inside Organ Ratio – **IOR**, expressed as percentage. (2) Distance to the nearest voxel of the nearest correct organ, denoted as Nearest Voxel Distance – **NVD**, expressed in centimeters. (3) Distance to the nearest voxel of the nearest correct organ, calculated only on samples for which the projection is outside the organ volume, denoted as Nearest Voxel Distance Outside – **NVD-O**, expressed in centimeters.

We compute the aforementioned metrics in four distinct inference scenarios specifically tailored to measure the grounding ability of our models. In the following experiments we show that our approach has an advantage over multiple baselines and demonstrate its ability to reason within the substructures of the organ and generalize to out-of-atlas organs, in addition to its other desirable properties that we demonstrate qualitatively.

6.1.1 General setting

We generate a 3D grounding for each of the articles in the test set and measure our model's performance against the following baselines:

(i) **Random** – We predict a randomly sampled point within a randomly chosen organ for each sample. (ii) **Center** – We use the center of the 3D atlas as the prediction. (iii) **Frequency (Freq.)** – We measure the frequency of the organ terms in the training set, and always predict the point within the most frequent organ. (iv) **MSE** – We frame the task as regression, and minimize the mean squared error (MSE) between the prediction and the average of a set of randomly sampled points from all the target organs. (v) **CLS** – We frame the task as classification and train a model to predict an organ index. The model is trained to minimize the cross-entropy between the output probability distribution and the target vector with 1's at the positions corresponding to the indices of organs present in the text and 0's elsewhere. During evaluation, the prediction is considered to be correct when it corresponds to any one of the target MeSH terms. When measuring NVD and NVD-O, we randomly sample a voxel point from the predicted organ as a 3D grounding.[6]

Method	IOR	NVD	NVD-O
Random	8.9 ± 0.5	17.9 ± 0.3	19.0 ± 0.3
Center	6.7 ± 0.4	13.3 ± 0.2	13.3 ± 0.2
Freq.	10.9 ± 0.5	13.9 ± 0.2	15.4 ± 0.2
MSE	$9{,}9 \pm 0.5$	6.8 ± 0.1	7.0 ± 0.1
CLS	$\mathbf{90.8 \pm 0.5}$	0.9 ± 0.1	8.2 ± 0.5
SOD	89.4 ± 0.5	$\mathbf{0.8 \pm 0.1}$	$\mathbf{2.5 \pm 0.1}$

Table 1: Mean IOR, NVD and NVD-O measured on the test set. The error bars represent the standard error.

In Table 1 we observe that SOD outperforms all baselines, and achieves nearly the same IOR as CLS. Furthermore, SOD significantly outperforms CLS according to the NVD and NVD-O metrics, which give a strong indication of the overall grounding performance, as per the one-sided Wilcoxon signed-rank test (Wilcoxon, 1945) ($p \approx 0$). We conclude that despite framing the task as soft regression, we sacrificed the least amount of predictive power (as per IOR), and exploited the atlas's inductive bias to achieve successful grounding (as per NVD and NVD-O).

6.1.2 Within organ reasoning

We perform a simulation to demonstrate that the grounding can infer anatomical substructures not present at the granularity of labeling in a specific atlas. Therefore, we perform experiments in which we merge the voxels of two different organs – effectively treating them as a single organ, and keep only instances from the training set that contain

[5] As some organs are hollow (small intestine, colon, etc.), we record a "hit", when the grounding is less than 1cm away from the most nearby voxel.

[6] We do not use the organ voxels centroids as prediction, as they can be outside of the organ volume for non-convex organs and yield a non-zero distance even when the correct organ is predicted, unfairly penalizing the classification baseline.

these "super-organs".[7] Then, we train a new model on each of these subsets and subsequently generate 3D groundings for each of the test set samples that only contain the individual occurrences of the two merged organs. The merged organ pairs are: (i) the "lung" and the "stomach", functionally different organs that belong to different groups, respiratory and digestive, respectively; and (ii) the "duodenum"[8] and the "small intestine", organs which are functionally related and frequently jointly referred to as "small intestine" in the literature.

Then, we train three different models for each merger: (1) **SMP** – We train a classification baseline on each of the subsets. During inference, we substitute the organ index with a randomly sampled voxel point within the predicted organ (2) **SOD w/** – We train a model using SOD with (w/) individual organs from the filtered training set. (3) **SOD w/o** – We train a model with the two organs merged into one "super-organ", effectively training without (w/o) the per-organ annotations.

With the functionally different "lung" and "stomach" merged together, in Table 2 we observe that SOD w/o significantly outperforms SMP, which can predict the coarse label corresponding to the super-organ, but is unable to reason about the organ's subregions. We also observe that SOD w/o performance is relatively close to SOD w/, which is trained with the separated organs. In Figure 3 we observe the grounding of 136 articles related to the "lung" and the "stomach" generated with SOD w/o. A notably harder problem is the "small intestine"-"duodenum" merger, which involves functionally related organs. We again observe that SOD w/o significantly outperforms SMP in both the micro-averaged performance and the grounding within the "duodenum". SMP achieves higher IOR on the articles that belong to the "small intestine," which is a result of the roughly 3 times larger number of "small intestine" voxels compared to the "duodenum" making the SMP performance skewed. We further examine the approximate locations of anatomical structures that co-occur most frequently with a given organ. For the organs co-occurring with the "lung", the frequency-weighed arithmetic mean of their centroids lies roughly 13.8 centimeters above that of the organs that co-occur with the "stomach". Similarly, such mean location of organs co-occurring with the "duodenum" lies 6.5 centimeters to the upper-left of the one of the "small intestine".

We conclude that despite the lack of explicit within organ annotations, SOD w/o learns to spatially reason about substructures within the organ based on the target organs' co-occurrences. In particular, the model learns to disambiguate between the organ regions because terms associated with different sub-regions tend to co-occur with different organs, typically the ones to which they are closer to (See Appendix Section F). This is an important observation from the following aspects: (i) Medical articles would get mapped to the appropriate organ regions they refer to, even though never explicitly annotated as such during training. (ii) Given an atlas with increased granularity, our method would, to a degree, accommodate for the newly added sub-regions without the need for re-training.

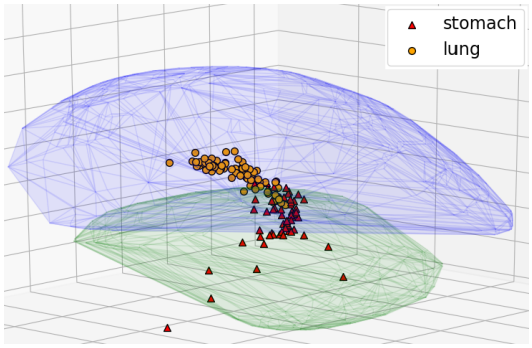


Figure 3: Groundings of articles referring exclusively to either the "lung" or the "stomach", obtained from a model trained with the two organs fused into one.

6.1.3 Generalization to unseen organs

To verify that our approach captures the locations of organs which are absent in the atlas segmentation labels, we evaluate the generalization ability of our method to organs unseen during training. For every organ, we remove its annotation from the training set, train a separate model, perform inference on the test set samples referring to it, and finally report the metrics averaged over all held-out organs.[9] In addition to NVD, we measure the rate at which the prediction is within the convex hull enveloping the organs of the same functional group as the held out organ, denoted as Inside Group Ratio – **IGR**. We compare SOD's performance against Random, Center and CLS, defined in Section 1.

[7] It may occur individually or co-occur with other organs.

[8] The duodenum is the first section of the small intestine in most higher vertebrates, including mammals.

[9] Samples referring solely to the held-out organ are removed, and the ones referring to it in addition to some other organs retain only the annotations of the other organs.

Method	IOR	NVD	IOR	NVD	IOR	NVD
	Lung		Stomach		Total	
SOD w/	100.0 ± 0.0	0.0 ± 0.0	95.4 ± 2.6	0.2 ± 0.1	97.8 ± 1.3	0.1 ± 0.0
SMP	46.5 ± 6.0	2.5 ± 0.3	44.6 ± 6.2	4.2 ± 0.6	45.6 ± 4.3	3.3 ± 0.3
SOD w/o	$\mathbf{94.4 \pm 2.8}$	$\mathbf{0.2 \pm 0.1}$	$\mathbf{81.5 \pm 4.8}$	$\mathbf{0.5 \pm 0.1}$	$\mathbf{88.2 \pm 2.8}$	$\mathbf{0.3 \pm 0.1}$
	Small intestine		Duodenum		Total	
SOD w/	97.4 ± 1.8	0.1 ± 0.0	90.9 ± 3.6	0.2 ± 0.1	94.4 ± 1.9	0.2 ± 0.0
SMP	$\mathbf{71.1 \pm 5.2}$	$\mathbf{1.0 \pm 0.2}$	33.3 ± 5.8	5.1 ± 0.6	53.5 ± 4.2	2.9 ± 0.3
SOD w/o	50.0 ± 5.8	1.1 ± 0.1	$\mathbf{93.9 \pm 3.0}$	$\mathbf{0.2 \pm 0.0}$	$\mathbf{70.4 \pm 3.8}$	$\mathbf{0.7 \pm 0.1}$

Table 2: Within organ reasoning evaluated on test set subsets obtained according choice of organs merged.

Method	IGR	NVD
Random	34.5 ± 4.0	21.1 ± 1.5
Center	37.0 ± 9.5	15.8 ± 2.0
CLS	72.1 ± 5.8 (84.21)	8.3 ± 1.0 (6.9)
SOD	$\mathbf{76.5 \pm 5.3}$ **(86.15)**	$\mathbf{7.5 \pm 1.0}$ **(5.9)**

Table 3: Results on test set samples referring to organs held out during training. Median values are in parentheses.

Method	IOR	NVD	NVD-O
Occ	68.7 ± 0.7	3.2 ± 0.1	9.7 ± 0.3
CLS	74.1 ± 0.7	2.5 ± 0.1	8.4 ± 0.3
CLS + M	80.4 ± 0.6	1.7 ± 0.1	7.3 ± 0.3
SOD	79.7 ± 0.7	1.6 ± 0.1	5.1 ± 0.2
SOD + M	$\mathbf{83.2 \pm 0.6}$	$\mathbf{1.2 \pm 0.1}$	$\mathbf{3.9 \pm 0.2}$

Table 4: Results on the full test set when the models are trained in a self-supervised fashion.

In Table 3, we observe that SOD significantly outperforms Random and Center. We also confirm a significant advantage of SOD over CLS by performing a Wilcoxon signed-rank test (IGR: $p = 0.0063$; NVD: $p = 0.0014$). Therefore, we conclude that besides grounding texts regarding organs present in the atlas, SOD reasons about unannotated structures, i.e., it learns to leverage the shared context between the held out organ and the functionally similar organs nearby. Consequently, we conclude that SOD learns to relate the articles' context with the spatial domain of the human body, and exploits this knowledge to improve generalization in a *zero-shot* setting. This suggests that our approach is robust to the granularity of the atlas used in training.

6.1.4 Self-supervised extension

We additionally evaluate our method in a self-supervised setting. Specifically, we ground medical abstracts in the atlas using only self-supervision in the form of occurrences of organ related terms. For that, we aggregate a list of all organ names corresponding to the MeSH terms, together with their UMLS synonyms (Bodenreider, 2004). During training, instead of providing the ground truth MeSH term annotation as target organs, we provide the target organs that correspond to the elements of the aggregated list of organ terms that appear in the abstract. We then train two different variants of our method: (1) **SOD** – A model trained with our regular SOD loss function. (2) **SOD + M** – During training, we stochastically substitute the occurrences of organ names or their synonyms in the text with a `[MASK]`[10] token with 50% probability.

We evaluate the performance of our method against the following baselines: (i) **Occ** – A naive model that predicts one of the organ names that appear in the text at random. When there is no explicit organ occurrence, it predicts the center of the atlas. (ii) **CLS** – A classification baseline, trained to predict one of the organ name occurrences from the text. (iii) **CLS + M** – A classification baseline boosted with the "masking" extension. Finally, we perform inference on the annotated test set and measure the IOR, NVD and NVD-O.

In Table 4 we observe that our method outperforms all baselines when trained both without (SOD), and with the masking extension (SOD + M). Since masking the organ names and their synonyms puts additional emphasis on their surrounding context, it allows the model to generalize better to the semantically annotated test set, yielding a considerable improvement for all metrics. It is noteworthy that in spite of training the model using organ names + synonym occurrences that appear

[10]The `[MASK]` token is included in BERT's vocabulary.

within the medical articles as ground truth targets, we obtain performance competitive to the fully-supervised training, included in Table 1. This data efficiency feature of our method is especially important since obtaining annotated data for medically relevant NLP tasks requires the time and effort of medical experts.

6.2 Medical information retrieval

We formulate a text-to-text retrieval setting where each test set article serves as a query and the remaining articles as the database from which we retrieve the relevant ones. We measure the retrieval quality using the standard Recall@K metric, i.e., the fraction of queries for which the correct article is retrieved among the top K articles. A retrieved article is considered correct when it has an identical set of MeSH term annotations as the query article. We fix K to 1, 5 or 10. We evaluate the performance of our method against the following supervised (w/) and pre-trained (w/o) baselines:

(i) **3D-Sms** (w/) – We train a Siamese BERT to group articles by optimizing the triplet loss, enforcing the embedding of articles with matching sets of MeSH annotations to nearby locations, and the non-matching ones to distant locations in the embedding space. We set the embedding space dimension to 3, and use the Euclidean distance measure and online triplet mining to obtain the positives and negatives for each sample during training (Hermans et al., 2017). (ii) **Large-Sms** (w/) – We follow the same procedure as 3D-Sms, however, we extend the embedding space dimension to 768[11]. (iii) **BaseBert** (w/o) – We use a general domain pre-trained BERT and concatenate the mean-pooled, max-pooled and `[CLS]` representations into a 2304 dimensional vector for each of the test articles. (iv) **BioBert** (w/o) – We use BERT pre-trained on PubMed abstracts and follow the same procedure as with BASEBERT. (v) **SciBert-NLI** – We use the mean-pooled embeddings from SCIBERT, fine-tuned for the NLI task on the datasets of Bowman et al. (2015); Williams et al. (2017).

In all baselines, we perform retrieval by taking the top K elements from the list of articles ranked by the Euclidean distance between their representation vectors and that of the query. The distance is computed in the representation space for the models trained on the retrieval task and the pre-trained sentence representation models, while for the SOD models we consider the physical distance in the 3D atlas.

Method	Dims.	R@1	R@5	R@10
Large-Sms (w/)	768	**42.9 ± 0.8**	**68.7 ± 0.7**	**75.4 ± 0.7**
3D-Sms (w/)	3	34.6 ± 0.8	61.4 ± 0.8	69.7 ± 0.7
SOD (w/)	3	37.4 ± 0.8	64.3 ± 0.8	71.3 ± 0.7
BaseBert (w/o)	2304	9.8 ± 0.5	26.1 ± 0.7	37.5 ± 0.8
BioBert (w/o)	2304	13.6 ± 0.6	35.2 ± 0.8	48.5 ± 0.8
SciBert-NLI (w/o)	768	16.9 ± 0.6	41.8 ± 0.8	55.4 ± 0.8
SOD+M (w/o)	3	**26.7 ± 0.7**	**56.8 ± 0.8**	**65.9 ± 0.8**

Table 5: Medical text retrieval. **Top:** Methods that use MeSH term supervision (w/), and **Bottom:** Methods that do not use MeSH annotation (w/o).

In Table 5 (upper), we compare our method with supervised (w/) Siamese models trained to group documents based on their MeSH term annotations. An interesting observation is that despite being trained to choose between the target organs when there are multiple, SOD outperforms 3D-Sms, which is explicitly trained to group articles in 3D based on the whole set of MeSH annotations, without being limited to organizing the article embeddings in the rather constrained 3D human atlas. It is worth noting that SOD falls slightly short compared to Large-Sms, most likely because Large-Sms is trained to embed text in 768 dimensions, thus having a higher representational power.

In Table 5 (lower), we evaluate the retrieval performance of our self-supervised method SOD+M (trained on occurrences of atlas glossary terms and their synonyms, see Section 6.1.4) against the pre-trained BERT baselines, in a setting that does not rely on ground truth MeSH term annotations. We observe that SOD+M significantly outperforms all of them, including SciBert-NLI[12].

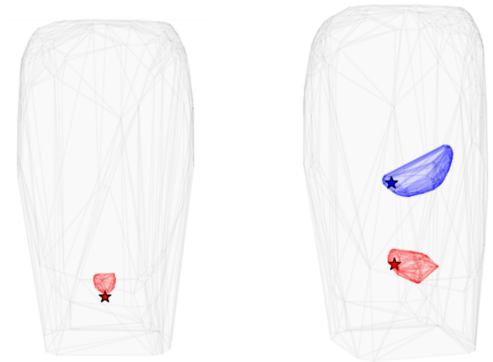
We observe a performance gap between SOD+M (w/o) and SOD (w/), as well as the methods that are explicitly trained to optimize the retrieval performance (3D-Sms, Large-Sms). However, in a (realistic) scenario of having large quantities of unannotated medical texts that require systematization, such fully-supervised approaches would not be feasible.

6.3 Qualitative evaluations and use-cases

We further demonstrate several desirable properties of our approach in a qualitative fashion. Although training was performed using a single male atlas, in Figure 4a, we observe the grounding of a paragraph

[11] The dimensionality of the BERT embedding.

[12] We report SciBert-NLI as it outperformed BioBert-NLI.



(a) Ovaries (b) Colon segments

Figure 4: **Left**: Grounding of the paragraph about the "ovaries" (Appendix Section D). The red structure is the "urinary bladder", which serves as a location reference. **Right:** Grounding of Wikipedia articles describing the "transverse colon" (upper) and "sigmoid colon" (lower), which were contained within the common label "colon" during training.

describing the "ovaries" (See Appendix Section D) to a reasonably close vicinity of their actual location. We additionally qualitatively evaluate the results of Section 6.1.2 by mapping Wikipedia articles referring to the "transverse colon" and the "sigmoid colon" to the 3D atlas. In Figure 4b, we observe that the articles are mapped to the actual locations of the colon segments, despite that the terms shared a common label ("colon") during training.

The low dimensional text embeddings in the 3D atlas space can be put to use in multiple real-world applications. Integrated with a speech recognition system, they could be used to provide real time localization of the steps taken during medical procedures based on the narrative operative reports. Additionally, the grounding to a 3D atlas can be used as a way to systematize large corpora of unannotated text while being able to observe the relationship between embedded texts in an intuitively meaningful setting. Another advantage of text retrieval in the physical 3D space is the ability to retrieve information by directly specifying an observable locations in the human atlas space, as opposed to using textual queries. To demonstrate this, we built a tool which accepts a query in the form of 3D coordinates and matches articles related to Covid-19 based on the proximity of their embeddings in 3D space (Grujicic et al., 2020). The tool for visual-based retrieval of Covid-19 related articles can be accessed at: `www.github.com/dusangrujicic/cord19-visualizer`

7 Discussion and conclusions

One limitation of our method is that it does not explicitly take into account spatial descriptions and other modifier expressions. Rather, it uses abstract level annotation to ground whole abstracts to the most semantically relevant regions, and uses the co-occurrences between terms (which also reflect their spatial relationships to a significant degree) to organize and distribute the grounding to within the same organ or to out-of-atlas organs. A natural extension of this work would be to move up from the entity level, and explicitly address the spatial language and descriptions of relationships between anatomical structures.

In this paper, we formulated a novel task of medical text grounding within an atlas of the human body. We proposed a loss function, Soft Organ Distance, which enables us to reason about inter-organ and intra-organ relatedness of medical text, without explicit annotations for the latter. In particular, we addressed the following limitations of prior work: (i) The text is embedded within a non-interpretable space – we embed, and systematically organize all articles in the 3D model of the human body, thus interpretability is intrinsic to our approach. (ii) There is no immediate, visually intuitive indication of the similarity between the retrieved articles – we perform retrieval directly in the 3D atlas, where the text embeddings and the relationships between them are visually comprehensible. Namely, while standard embedding and visualization techniques uncover hidden data clusters, the underlying similarity grouping the articles is not clear. On the other hand, our approach provides semantically and spatially meaningful grounding together with off-the-shelf successful retrieval, which we believe to be essential for many NLP applications involving medical information retrieval and visualization.

Acknowledgements

This work was supported by KU Leuven Internal Funds (MACCHINA) and the Flemish Government under the Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen programme.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Andreas Christ, Wolfgang Kainz, Eckhart G Hahn, Katharina Honegger, Marcel Zefferer, Esra Neufeld, Wolfgang Rascher, Rolf Janka, Werner Bautz, Ji Chen, et al. 2009. The virtual family—development of surface-based anatomical models of two adults and two children for dosimetric simulations. *Physics in Medicine & Biology*, 55(2):N23.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Richard Drake, A Wayne Vogl, and Adam WM Mitchell. 2009. *Gray's Anatomy for Students E-Book*. Elsevier Health Sciences.

Marie-Christine Gosselin, Esra Neufeld, Heidi Moser, Eveline Huber, Silvia Farcito, Livia Gerber, Maria Jedensjö, Isabel Hilber, Fabienne Di Gennaro, Bryn Lloyd, et al. 2014. Development of a new generation of high-resolution anatomical models for medical device evaluation: the virtual population 3.0. *Physics in Medicine & Biology*, 59(18):5287.

Dusan Grujicic, Gorjan Radevski, Tinne Tuytelaars, and Matthew B Blaschko. 2020. Self-supervised context-aware covid-19 document exploration through atlas grounding. In *ACL Workshop on Natural Language Processing for COVID-19*.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Karl Heinz Höhne, Bernhard Pflesser, Andreas Pommert, Martin Riemer, Rainer Schubert, Thomas Schiemann, Ulf Tiede, and Udo Schumacher. 2001. A realistic model of human structure from the visible human data. *Methods of information in medicine*, 40(02):83–89.

Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564.

Qiao Jin, Bhuwan Dhingra, William W Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. *arXiv preprint arXiv:1904.02181*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What are you talking about? text-to-image coreference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3558–3565.

Jayant Krishnamurthy and Thomas Kollar. 2013. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics*, 1:193–206.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Andreas Pommert, Karl Heinz Höhne, Bernhard Pflesser, Ernst Richter, Martin Riemer, Thomas Schiemann, Rainer Schubert, Udo Schumacher, and Ulf Tiede. 2001. Creating a high-resolution spatial/symbolic model of the inner organs based on the visible human. *Medical Image Analysis*, 5(3):221–228.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer.

Thomas Schiemann, Ulf Tiede, and Karl Heinz Höhne. 1997. Segmentation of the visible human for high-quality volume-based visualization. *Medical image analysis*, 1(4):263–270.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. 2018. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. Cord-19: The covid-19 open research dataset. *ArXiv*.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics*, 1(6):80–83.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer.

A Human atlas

There are multiple digital anatomical models available. The Virtual Population (Christ et al., 2009; Gosselin et al., 2014) of the *IT'IS Foundation*[13] contains anatomical models of 10 different persons obtained from MRI procedures. The **S**egmented **I**nner **O**rgans **(SIO)** from the *Voxel-Man* project (Höhne et al., 2001; Pommert et al., 2001; Schiemann et al., 1997) [14] is based on the *Visible Human Male* (U.S. National Library of Medicine [15]) and contains 202 labeled anatomical objects within the human torso. The model consists of 774 slices obtained by CT and MRI imaging, where each slice contains a cryosection image, a CT image and a segmentation label image where the grayscale level corresponds to a segmentation label of the tissue (Figure A.1).

The Segmented Inner Organs (SIO) contains a glossary of medical terms and their associated segmentation labels. A list of synonyms and closely related wordforms for each glossary term were retrieved. The ScispaCy UmlsEntityLinker (Neumann et al., 2019) was used for searching the UMLS Metathesaurus *(The Unified Medical Language System)* (Bodenreider, 2004) for all word forms of the SIO glossary [16]. The parameters of the UmlsEntityLinker were kept at default values.

SIO includes 202 anatomical objects with their distinct segmentation labels. Tissues such as skin, gray matter, white matter, and unclassified tissues were removed from the set of labeled terms, as they denote general medical concepts not characterized by specific compact locations in the human body. The vertebrae, bones, and muscles of the locomotor system were discarded as well. The blood vessels, being small, elongated and often not particular to any single region of the body, were also removed. Additionally, we remove the remaining small organs with fewer than 1000 associated voxels. The SIO includes the model of the human head as well, which we do not use.

In the case of categories for bilateral organs located symmetrically on both the left and the right side of the body, which are seldom mentioned explicitly in the texts, only the atlas voxels pertaining to the left organ were kept for every bilateral pair. Atlas labels that appear infrequently in medical literature, but are functionally related to other, more frequently occurring organs, or are colloquially referred to under a single, umbrella term, were merged. The aforementioned steps reduced the list of distinct anatomical objects of interest to 27: "ampulla", "bronchi", "caecum", "diaphragm", "gallbladder", "larynx", "liver", "myocardium", "pancreas", "pericardium", "prostate", "rectum", "seminal gland", "small intestine", "spleen", "testis", "thyroid gland", "urinary bladder", "stomach", "colon", "penis", "trachea", "ventricle", "atrium", "kidney", "lung", "duodenum".

B Dataset

Focusing on the articles pertaining to the human anatomy, we remove the samples that contain any of the following MeSH terms: "Animals", "Rats", "Mice", "Rats, Sprague-Dawley", "Rats, Wistar", "Mice, Inbred C57BL", "Rats, Inbred Strains", "Disease Models, Animal", "Dogs", "Rabbits", "Swine", "Mice, Inbred BALB C", "Guinea Pigs", "Mice, Knockout", "Cattle", "Animals, Newborn", "Mice, Transgenic", "Chickens", "Sheep", "Mice, Inbred Strains", "Rats, Inbred F344", which typically correspond to articles that describe clinical trials on animals. Subsequently, we discard the MeSH terms which are not a name or a synonym of an atlas organ.

C Metrics

For the IOR, the predicted 3D point lies inside the organ volume (*hit*) when its coordinates, rounded to the nearest integer to represent voxel indices, are within the set of indices of voxels that make up the corresponding organ. However, in the case of hollow organs, such as the intestines and the stomach, scoring a *hit* would require predicting a point that lies exactly within a (usually very thin) organ wall, as the region of the organ's lumen is typically not included. Therefore, we use a more relaxed criterion, and record a *hit* when the predicted 3D point to is inside or sufficiently close to the organ volume, which we consider to be the case when its coordinates are less than 1cm away from the nearest voxel of the target organ. In cases of multiple target organs, we measure a *hit* when the predicted coordinates lie within or sufficiently close to any one of the given organs.

When the projection is exactly inside the volume of the organ, the NVD is zero, and otherwise, it

[13] www.itis.swiss/
[14] www.voxel-man.com/
[15] www.nlm.nih.gov/research/visible/
[16] ScispaCy version 0.2.3 and en_core_sci_lg pipeline

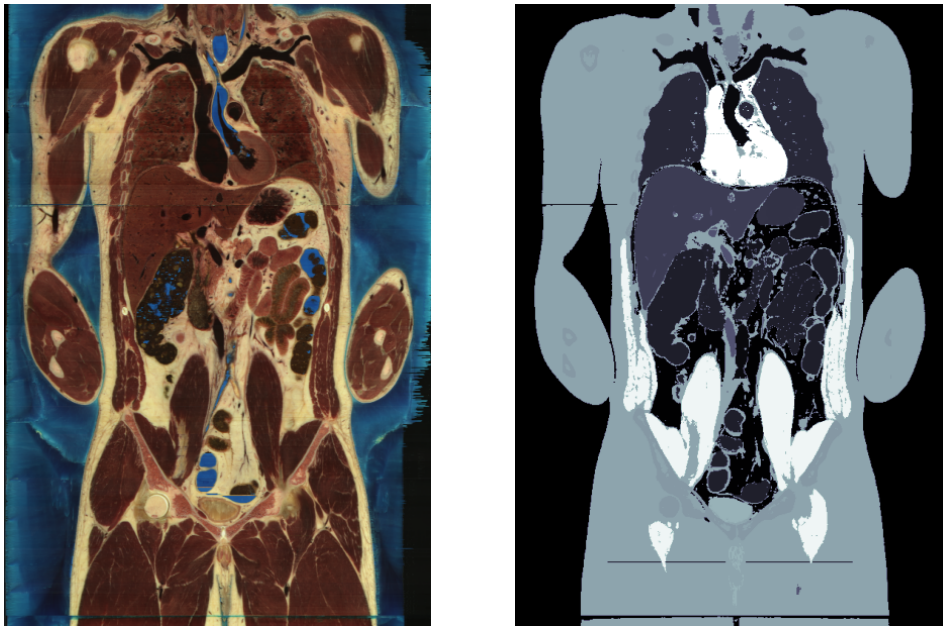


Figure A.1: Cross-sections of the RGB volume (left) and the grayscale volume representing segmentation labels (right) (Pommert et al., 2001).

is measured as the distance to the surface of the nearest organ in the text. The NVD-O metric complements the NVD metric, such that it gives insight into how far off the prediction is when it misses the correct organ.

D Qualitative Example of Ovaries

The Wikipedia paragraph describing the structure of the ovaries [17]: *"The ovaries are considered the female gonads. Each ovary is whitish in color and located alongside the lateral wall of the uterus in a region called the ovarian fossa. The ovarian fossa is the region that is bounded by the external iliac artery and in front of the ureter and the internal iliac artery. This area is about 4 cm x 3 cm x 2 cm in size. The ovaries are surrounded by a capsule, and have an outer cortex and an inner medulla. The capsule is of dense connective tissue and is known as the tunica albuginea. Usually, ovulation occurs in one of the two ovaries releasing an egg each menstrual cycle. The side of the ovary closest to the fallopian tube is connected to it by infundibulopelvic ligament, and the other side points downwards attached to the uterus via the ovarian ligament. Other structures and tissues of the ovaries include the hilum."*

The first sentence - "The ovaries are considered the female gonads", was removed, as it mentions the term *gonads*, which is a strong clue for the model, as the male gonad ("testis"), was present in the atlas.

[17] https://en.wikipedia.org/wiki/Ovary

E Varying temperatures and sampled voxels

Table E.1 and E.2 showcase the how varying the temperature terms γ_p and γ_o affect the results measured by IOR, NVD and NVD-O on the full test set of medical articles. In particular, table E.1 shows the results when the model is trained with 100 voxel points sampled for each organ during training, while table E.2 shows the results when the model is trained by sampling 1000 voxel points during training.

F Organ Co-Ocurrences

We observe a significant degree of correlation between the membership of two organs in the same

Method	IOR	NVD	NVD-O
$\gamma_p = 0.1, \gamma_o = 0.1$	89.2 ± 0.5	0.9 ± 0.1	2.8 ± 0.1
$\gamma_p = 0.1, \gamma_o = 0.5$	88.8 ± 0.5	$\mathbf{0.8 \pm 0.1}$	2.7 ± 0.2
$\gamma_p = 0.1, \gamma_o = 1.0$	$\mathbf{89.4 \pm 0.5}$	$\mathbf{0.8 \pm 0.1}$	2.5 ± 0.2
$\gamma_p = 0.5, \gamma_o = 0.1$	82.5 ± 0.6	1.0 ± 0.1	2.3 ± 0.1
$\gamma_p = 0.5, \gamma_o = 0.5$	86.7 ± 0.5	0.9 ± 0.1	2.1 ± 0.1
$\gamma_p = 0.5, \gamma_o = 1.0$	85.0 ± 0.6	1.0 ± 0.1	2.3 ± 0.1
$\gamma_p = 1.0, \gamma_o = 0.1$	83.6 ± 0.6	1.0 ± 0.1	2.4 ± 0.1
$\gamma_p = 1.0, \gamma_o = 0.5$	82.2 ± 0.6	1.0 ± 0.1	2.3 ± 0.1
$\gamma_p = 1.0, \gamma_o = 1.0$	82.2 ± 0.6	1.0 ± 0.1	$\mathbf{1.9 \pm 0.1}$

Table E.1: Results on the full test set from models trained with varying inference while randomly sampling **100** voxles during training.

Method	IOR	NVD	NVD-O
$\gamma_p = 0.1, \gamma_o = 0.1$	89.1 ± 0.5	1.0 ± 0.1	3.6 ± 0.2
$\gamma_p = 0.1, \gamma_o = 0.5$	$\mathbf{89.3 \pm 0.5}$	$\mathbf{0.8 \pm 0.1}$	2.5 ± 0.2
$\gamma_p = 0.1, \gamma_o = 1.0$	89.2 ± 0.5	$\mathbf{0.8 \pm 0.1}$	2.4 ± 0.2
$\gamma_p = 0.5, \gamma_o = 0.1$	84.0 ± 0.6	1.0 ± 0.1	2.6 ± 0.2
$\gamma_p = 0.5, \gamma_o = 0.5$	86.8 ± 0.5	0.9 ± 0.1	$\mathbf{2.0 \pm 0.1}$
$\gamma_p = 0.5, \gamma_o = 1.0$	84.0 ± 0.6	0.9 ± 0.1	$\mathbf{2.0 \pm 0.1}$
$\gamma_p = 1.0, \gamma_o = 0.1$	84.2 ± 0.6	1.1 ± 0.1	2.8 ± 0.2
$\gamma_p = 1.0, \gamma_o = 0.5$	81.4 ± 0.6	1.0 ± 0.1	2.1 ± 0.1
$\gamma_p = 1.0, \gamma_o = 1.0$	83.1 ± 0.6	1.0 ± 0.1	2.1 ± 0.1

Table E.2: Results on the full test set from models trained with varying inference while randomly sampling **1000** voxels during training.

functional system [18] and their proximity (Pearson correlation coefficient of 0.31, increasing to 0.49 if the locomotor system, which is not localized to any particular area of the body is left out), as well as with their number of times they co-occur within the same sample (Pearson correlation coefficient of 0.37). Consequently, the organs that are close to one another tend to co-occur as well (Pearson correlation coefficient of 0.36). This suggests that the physical proximity of medical terms in our atlas reflects their functional and semantic relatedness to a significant degree, corroborated by the tendency of nearby organs to both co-occur in the dataset and the belong to a common category in a general categorization of anatomical terms.

In order to determine which organs co-occur most frequently, we sort the organs based on their positions along the vertical axis of the body, and calculate the number of co-occurrences for each organ pair. We then construct a matrix of the number of co-occurrences and normalize each row so that its values sum up to one. The matrix is shown in Figure F.1. The fact that the highest values tend to be near the main diagonal of the matrix confirms that the organs located at similar heights, and therefore nearby locations, tend to co-occur the most.

[18] respiratory, digestive, etc.

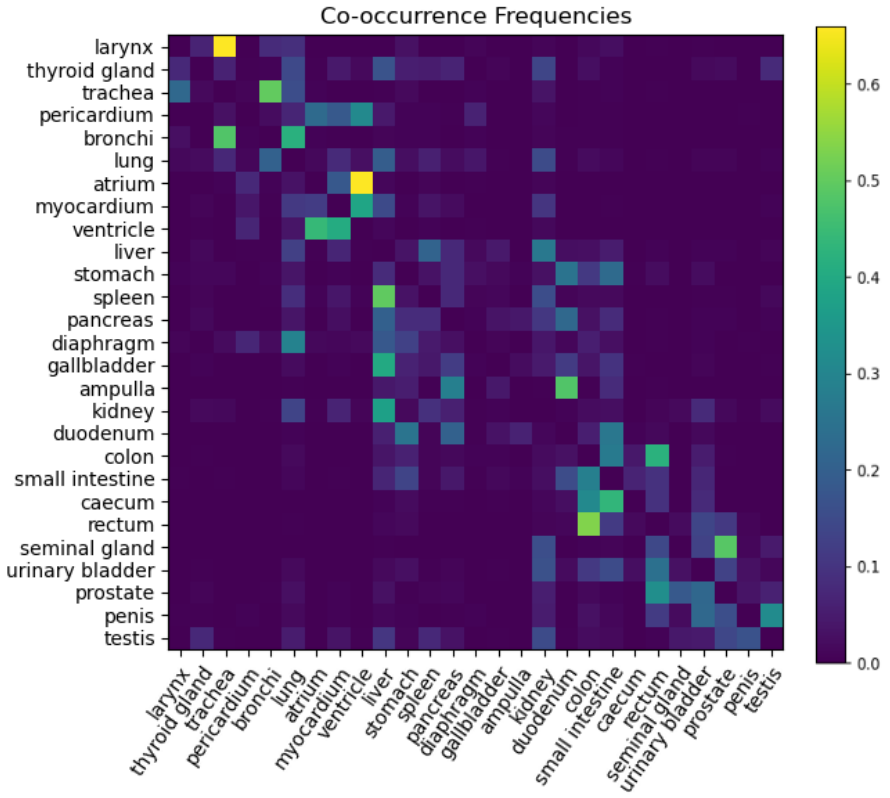


Figure F.1: Matrix of normalized co-occurrence frequencies between organ pairs.