# Seeing Unseen: Discover Novel Biomedical Concepts via Geometry-Constrained Probabilistic Modeling (Supplementary Material)

Jianan Fan[1], Dongnan Liu[1], Hang Chang[2], Heng Huang[3], Mei Chen[4], and Weidong Cai[1]

[1]University of Sydney [2]Lawrence Berkeley National Laboratory [3]University of Maryland at College Park [4]Microsoft
jfan6480@uni.sydney.edu.au dongnan.liu@sydney.edu.au hchang@lbl.gov
henghuanghh@gmail.com Mei.Chen@microsoft.com tom.cai@sydney.edu.au

## 1. Details for Theoretical Analysis

### 1.1. Monotonicity between Distributional Concentration and Semantic Ambiguity

Here, we first present the detailed proof for the following proposition:

**Proposition 1.** *Let $\zeta_{\boldsymbol{x}}$ be the continuous entropy of the posterior vMF distribution parametrized by $\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}} \in \mathbb{S}^{d-1}$ and $\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}} \in \mathbb{R}^{d}_{>0}$. We have $\zeta_{\boldsymbol{x}}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})$ behave as a monotonically decreasing function in the interval $(0, +\infty)$.*

*Proof.* For any vMF distribution characterized as $q(\boldsymbol{z}|\boldsymbol{x}) \sim \mathbf{vMF}(\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}, \tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})$, its continuous entropy can be derived as:

$$\begin{aligned}\mathcal{H}_q(\boldsymbol{z}|\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}},\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}) &= \mathbb{E}_q[-\log(C_d(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})\exp(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}^T\boldsymbol{z}))] \\ &= -\int_{\mathbb{S}^{d-1}}[\log C_d(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}) + \tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}^T\boldsymbol{z}]q(\boldsymbol{z})d\boldsymbol{z} \\ &= -(\frac{d}{2}-1)\cdot\log\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}} + \log\mathcal{I}_{d/2-1}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}) \\ &\quad - \frac{\mathcal{I}_{d/2}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})}{\mathcal{I}_{d/2-1}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})}\cdot\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}} + (d/2)\log 2\pi.\end{aligned} \tag{1}$$

The derivation is based upon the vMF properties that $\|\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}\| = 1$ and $\mathbb{E}_q(\boldsymbol{z}) = A_d(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})\cdot\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}$, where $A_d(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}) = \frac{\mathcal{I}_{d/2}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})}{\mathcal{I}_{d/2-1}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})}$. It is noted that $\mathcal{H}_q(\boldsymbol{z})$ is a univariate function of $\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}$, regardless of $\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}$. We thereby model the derivative of the continuous entropy with respect to the distributional concentration parameter $\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}$:

$$\begin{aligned}&\frac{\partial\mathcal{H}_q(\boldsymbol{z}|\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}},\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})}{\partial\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}} = \nabla_{\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}}[-(\frac{d}{2}-1)\cdot\log\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}} + \\ &\log\mathcal{I}_{d/2-1}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}) + (d/2)\log 2\pi] - \nabla_{\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}}[\frac{\mathcal{I}_{d/2}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})}{\mathcal{I}_{d/2-1}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})}\cdot\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}] \\ &= \frac{\mathcal{I}_{d/2}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})}{\mathcal{I}_{d/2-1}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})} - \nabla_{\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}}[\frac{\mathcal{I}_{d/2}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})}{\mathcal{I}_{d/2-1}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})}]\cdot\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}} - \frac{\mathcal{I}_{d/2}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})}{\mathcal{I}_{d/2-1}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})} \\ &\propto \tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}\cdot[\frac{\mathcal{I}_{d/2}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})\cdot(\mathcal{I}_{d/2-2}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}) + \mathcal{I}_{d/2}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}))}{\mathcal{I}^2_{d/2-1}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})} - \frac{\mathcal{I}_{d/2+1}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})}{\mathcal{I}_{d/2-1}(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}})} \\ &\quad - 1],\end{aligned} \tag{2}$$

which is always negative for $\forall\, d \in (1, +\infty)$ and $\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}} \in (0, +\infty)$ given the properties of modified Bessel functions [9]. □

The proposition provides a theoretical foundation for our method, which aims to model the latent bias incurred by inconsistent imaging protocols across cohorts with the distributional concentration parameter. In this sense, the task-irrelevant interference attributes can be well characterized and decoupled from the informative semantic context. The above derivations also hold true for the KL divergence regularization term in Eq. (7) of the main text, in which the mathematical formulation is analogous to the continuous entropy. The monotonicity indicates that the statistical parameter $\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}$ can behave as an *instance-adaptive scaling factor* to dynamically rectify the weight of each instance according to its semantic ambiguity. The formulation delivers a heuristic solution to compensate for the latent bias issued from non-i.i.d. data aggregated from numerous cohorts.

### 1.2. Towards Bounded Open Space Risk with Uniform Proxies

In our proposed method, we specifically devise open space structuring with uniform proxies defined a priori to organically shape the geometric layout of embedding manifold and regulate the open space risk associated with novel classes. We hereby prove its effectiveness towards a tighter error bound for novel class positioning and discrimination.

**Lemma 2.** *Let $\hat{\boldsymbol{\Upsilon}} = \{\hat{\boldsymbol{v}}_1, ..., \hat{\boldsymbol{v}}_n\}$ denote a set of proxies uniformly distributed on a unit hyperspherical manifold $\mathbb{S}^{d-1}$. Assume $\eta = \frac{d}{n} \in (0, 1)$ as $n \to \infty$, then we have for the smallest singular value $\lambda_{min}$ of $\hat{\boldsymbol{\Upsilon}}$ that:*

$$\lim_{n\to\infty}\lambda_{min}(\hat{\boldsymbol{\Upsilon}}) \geq (1-\sqrt{\eta})\cdot(\min_i\frac{\sqrt{d}}{\|\hat{\boldsymbol{r}}_{\boldsymbol{i}}\|}), \tag{3}$$

*where $\hat{\boldsymbol{r}}_{\boldsymbol{i}}$ are i.i.d. $N(0, 1)$ random variables.*

The proof can be derived from [16]. Then, we consider the instance adjacency estimation procedure, which identifies

the instance pairs holding strong semantic correlations according to the following distributional overlap metric:

$$\mathcal{O}_{\boldsymbol{x}}^{\hat{\boldsymbol{\Upsilon}}} = \{C_d(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}}) \exp(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}} \tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}^T \hat{\boldsymbol{\upsilon}}_i), \forall \hat{\boldsymbol{\upsilon}}_i \in \hat{\boldsymbol{\Upsilon}}\}. \quad (4)$$

The error rate of distributional overlap measurement and subsequent instance pairing determines the efficacy of Eq. (11) in the main text and is therefore in consonance with the open space risk. Specifically, we represent the estimated error for distributional overlap score with least squares error as: $\Pi(\mathcal{O}_{\boldsymbol{x}}^{\hat{\boldsymbol{\Upsilon}}}) = \frac{1}{2N_{\mathcal{U}}} \sum_{i=1}^{N_{\mathcal{U}}} (\hat{y}_i - \mathcal{O}_{\boldsymbol{x}}^{\hat{\boldsymbol{\Upsilon}}})^2$, where $\hat{y}_i$ is the expected estimation. Afterwards, we can obtain its partial derivative with respect to the proxy placement as:

$$\frac{\partial \Pi(\mathcal{O}_{\boldsymbol{x}}^{\hat{\boldsymbol{\Upsilon}}})}{\partial \hat{\boldsymbol{\Upsilon}}} = \frac{1}{N_{\mathcal{U}}} \sum_{i=1}^{N_{\mathcal{U}}} (\mathcal{O}_{\boldsymbol{x}}^{\hat{\boldsymbol{\Upsilon}}} - \hat{y}_i) \cdot \tilde{\boldsymbol{C}} \exp(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}} \tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}^T \hat{\boldsymbol{\Upsilon}}), \quad (5)$$

where $\tilde{\boldsymbol{C}}$ denotes the normalization constant. By integrating $\tilde{\boldsymbol{C}} \exp(\tilde{\boldsymbol{\kappa}}_{\boldsymbol{x}} \tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}^T \hat{\boldsymbol{\Upsilon}})$ in their matrix form as $\mathbf{M}$, we can then reformulate the equation to characterize the error of distributional overlap measurement as:

$$\pi(\mathcal{O}_{\boldsymbol{x}}^{\hat{\boldsymbol{\Upsilon}}}) = \frac{1}{N_{\mathcal{U}}} \sum_{i=1}^{N_{\mathcal{U}}} (\mathcal{O}_{\boldsymbol{x}}^{\hat{\boldsymbol{\Upsilon}}} - \hat{y}_i) = \frac{1}{\mathbf{M}} \cdot \frac{\partial \Pi(\mathcal{O}_{\boldsymbol{x}}^{\hat{\boldsymbol{\Upsilon}}})}{\partial \hat{\boldsymbol{\Upsilon}}}, \quad (6)$$

$$\|\pi(\mathcal{O}_{\boldsymbol{x}}^{\hat{\boldsymbol{\Upsilon}}})\| \leq \frac{1}{\lambda_{min}(\mathbf{M})} \cdot \|\frac{\partial \Pi(\mathcal{O}_{\boldsymbol{x}}^{\hat{\boldsymbol{\Upsilon}}})}{\partial \hat{\boldsymbol{\Upsilon}}}\|. \quad (7)$$

Combined with Lemma 2, the error associated with instance pairing and open space structuring can be well-bounded when $\hat{\boldsymbol{\Upsilon}}$ holds spatial uniformity over the hyperspherical manifold.

# 2. Experimental Setup Details

## 2.1. Datasets

### 2.1.1 Pneumonia Infectious Organisms

With the rapid emergence of COVID-19, there has been a significant surge in interest and efforts to facilitate fundamental research in the field of pneumonia with the advancement of radiology. Besides vanilla discrimination between normal and infected patients, exactly identifying the responsible infectious microorganisms, such as viruses or bacteria, could bring clinical benefits for patient management and symptom explanation [5]. In this regard, we consider the discovery scenario aimed at recognizing and grouping pneumonia incurred by unseen microorganisms according to the anatomical and pathological information present in X-rays.

Specifically, we adopt MDTD [11] and COV-iDC [4] as the base and unlabeled sets of data, respectively. MDTD contains 5,863 chest X-ray images collected from retrospective cohorts of pediatric patients in a Chinese hospital. The radiographs are firstly categorized by whether the patients are infected with pneumonia, and then the positive ones are annotated with their infectious organisms (viral pneumonia). COV-iDC accumulates the public data available from the Internet and formalizes a database comprising two new classes of pneumonia infectious microorganisms, bacterium and fungus. It also suggests a fine-grained categorization schema in which pneumonia patients are further differentiated with their specific subclasses, leading to in total eight classes, including Normal, COVID-19, SARS, MERS-CoV, Streptococcus, Klebsiella, Legionella, and Pneumocystis. Notably, data distribution shifts intrinsically exist across the two sets of data due to different imaging device and archiving procedure across hospitals and countries. Additionally, in COV-iDC, the number of patients infected by the two novel classes of pneumonia is much fewer than the patients corresponding to base classes, which conforms to our argument. The quantitative comparisons of caseload between base and novel sets of classes are shown in Table 1.

Table 1. Statistics of the datasets for novel biomedical concept discovery. The total numbers of instances belonging to base and novel sets of classes are separately presented.

| | **Pneumonia** | **Cell Nuclei** | **Skin Lesion** | **Retinopathy** |
|---|---|---|---|---|
| Base | 3,220 | 244,831 | 2,368 | 14,547 |
| Novel | 80 | 4,508 | 232 | 488 |

### 2.1.2 Cell Nuclei

In the biological system of multicellular organism, cells would become specialized to perform different functions according to the regulation of genetic information contained in nuclei. The rich information exhibited by the structural architecture and morphological characteristics of nuclei conveys essential clues for tissue microenvironment profiling and disease analysis [1]. We therefore propose to evaluate the effectiveness of our method towards discovering novel classes of nuclei autonomously, to verify its potential to accelerate fundamental biomedical research.

In particular, we adopt PanNuke [7] and Lizard [8] as the base and unlabeled datasets, respectively. PanNuke is composed of 189,744 annotated nuclei obtained from digitalized pathology specimens across various organs, where the nuclei can be generally categorized into three types: epithelial, inflammatory, and connective. Lizard comprises a total of 495,179 nuclei acquired from colon tissue. Compared with PanNuke, it introduces two novel classes of nuclei: neutrophil and eosinophil. From the fine-grained perspective, Lizard is principally encompassing nine subclasses of cells: neoplastic, non-neoplastic epithelial, lymphocyte, plasma, neutrophil, eosinophil, fibroblasts, muscle, and endothelial cells. We use all images from the breast tissue in PanNuke

and the DPath set of data in Lizard to ensure evident data distribution discrepancy.

### 2.1.3 Skin Cancer Lesions

Here, we consider the discovery of another biomedical concept, i.e., lesions, which corresponds to an abnormal region suffering injury or disease. We evaluate our method towards novel lesion discovery on a large-scale miscellaneous skin lesion benchmark [3] consisting of more than 10k dermatoscopic images. We select two sets of data which are collected from different body locations (lower extremity and face) and exhibit significant visual difference and evident distribution shifts. The face dataset is adopted to formulate the base class set with three typical lesion types: benign keratinocytic lesions, melanocytic nevi, and melanomas. Then, the lower extremity dataset is considered as the unseen class set, which incorporates four novel categories of skin lesions: vascular lesions, dermatofibromas, basal cell carcinomas, and actinic keratoses intraepithelial carcinomas. Those novel classes to be discovered are intrinsically rare and much fewer in case amount than the common ones, with up to $100\times$ discrepancy in overall quantity.

### 2.1.4 Diabetic Retinopathy Severity

All the above tasks can be technically recognized as a classification problem where different classes independently exist and do not retain mutual correspondence. In this experiment, we conduct evaluations on a different technical scenario, ordinal regression, for which the variables to be predicted are only informative in the context of the relative ordering between different values.

Specifically, we consider the discovery of undefined levels of diabetic retinopathy severity. The severity of diabetic retinopathy can be exhibited by the pathological tissue changes in the retina and behaves as a fundamental indicator for the potential risks of vision loss and blindness. In this vein, we adopt two retina image datasets DDR [13] and APTOS [10] acquired with fundus photography which capture diabetic retinopathy under a variety of severity levels. We utilize DDR as the base set, with three standard levels of severity: no symptom, mild, and moderate. APTOS is then adopted as the unlabeled set comprising two uncommon severity degrees: severe and proliferative, which are regarded as the novel classes.

## References


[1] Harris Busch. *The Cell Nucleus V3*, volume 3. Elsevier, 2012.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[3] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[4] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. Covid-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*, 2020.

[5] Carole Dennie, Cameron Hague, Robert S Lim, Daria Manos, Brett F Memauri, Elsie T Nguyen, and Jana Taylor. Canadian society of thoracic radiology/canadian association of radiologists consensus statement regarding chest imaging in suspected and confirmed covid-19. *Canadian Association of Radiologists Journal*, 71(4):470–481, 2020.

[6] Zhen Fang, Jie Lu, Feng Liu, Junyu Xuan, and Guangquan Zhang. Open set domain adaptation: Theoretical bound and algorithm. *IEEE transactions on neural networks and learning systems*, 32(10):4309–4322, 2020.

[7] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *Digital Pathology: 15th European Congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15*, pages 11–19. Springer, 2019.

[8] Simon Graham, Mostafa Jahanifar, Ayesha Azam, Mohammed Nimir, Yee-Wah Tsang, Katherine Dodd, Emily Hero, Harvir Sahota, Atisha Tank, Ksenija Benes, et al. Lizard: A large-scale dataset for colonic nuclear instance segmentation and classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 684–693, 2021.

[9] CM Joshi and SK1094928 Bissu. Some inequalities of bessel and modified bessel functions. *Journal of the Australian Mathematical Society*, 50(2):333–342, 1991.

[10] Sohier Dane Karthik, Maggie. Aptos 2019 blindness detection. *Kaggle*, 2019.

[11] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

[12] Seung-Jean Kim, Alessandro Magnani, and Stephen Boyd. Robust fisher discriminant analysis. *Advances in neural information processing systems*, 18, 2005.

[13] Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501:511–522, 2019.

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[15] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. *Advances in neural information processing systems*, 33:16282–16292, 2020.

[16] Jack W Silverstein. The smallest eigenvalue of a large dimensional wishart matrix. *The Annals of Probability*, pages 1364–1368, 1985.

[17] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7492–7501, 2022.

[18] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.