Evolution Is All You Need: Phylogenetic Augmentation for Contrastive Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Self-supervised representation learning of biological sequence embeddings alleviates computational resource constraints on downstream tasks while circumventing expensive experimental label acquisition. However, existing methods mostly borrow directly from large language models designed for NLP, rather than with bioinformatics philosophies in mind. Recently, contrastive mutual information maximization methods have achieved state-of-the-art representations for ImageNet. In this perspective piece, we discuss how viewing evolution as natural sequence augmentation and maximizing information across phylogenetic "noisy channels" is a biologically and theoretically desirable objective for pretraining encoders. We first provide a review of current contrastive learning literature, then provide an illustrative example where we show that contrastive learning using evolutionary augmentation can be used as a representation learning objective which maximizes the mutual information between biological sequences and their conserved function, and finally outline rationale for this approach.

1 Introduction

Self-supervised learning representation learning of biological sequences aims to capture meaningful properties for downstream analyses, while pretraining only on labels derived from the data itself. Embeddings alleviate computational constraints, and yield new biological insights from analyses in a rich latent space; to do so in a self-supervised manner further circumvents the expensive and time-consuming need to gather experimental labels. Though recent works have successfully demonstrated the ability to capture properties such as fluorescence, pairwise contact, phylogenetics, structure, and subcellular localization, these works mostly use methods designed for natural language processing (NLP) [56, 12, 46, 47, 3, 25, 18, 5, 21, 40, 19]. This leaves open the question of how best to design self-supervised methods which align with biological principles.

Recently, contrastive methods for learning representations achieve state-of-the-art results on ImageNet [43, 26, 51, 24, 14]. Two "views" v_1 and v_2 of an input are defined (e.g. two image augmentation strategies), and the contrastive objective is to distinguish one pair of "correctly paired" views from $N - 1$ "incorrectly paired" dissimilar views. This incentivizes the encoder to learn meaningful properties of the input, while disregarding nuisance factors. Theoretically, it can be shown that such an objective maximizes the lower-bound on the mutual information, $I(v_1, v_2)$ [44].

In this piece, we first provide a review of current contrastive learning literature for obtaining representations in non-biological modalities. Then, we propose that molecular evolution is a good choice of augmentation to provide "views" for contrastive learning in computational biology, from both the theoretical and biological perspectives. Finally, we illustrate how evolutionary augmentation can be used to optimize a deep neural network encoder to preserve the information in biological sequences that pertains to their function.

Submitted to Machine Learning for Computational Biology (MLCB) 2020. Do not distribute.

2 Contrastive Learning for Mutual Information Maximization

2.1 Contrastive Learning and Mutual Information Estimation

The InfoMax optimization principle [36] aims to find a mapping g such that the Shannon mutual information between the input and output is maximized, i.e. $\max_{g \in \mathcal{G}} I(X; g(X))$. Recent years revive this principle as a representation learning objective to train deep encoders as g, and yield empirically desirable representations in the modalities of imaging [43, 27, 8, 51, 26, 37, 24, 14, 52, 55], text [48, 43, 32], and audio [37, 43].

Most follow a variation of this optimization objective: given input x, and transformations t_1 and t_2, define $v_1 = t_1(x)$ and $v_2 = t_2(x)$ as two different "views" of x. These "transformations" can be parameterless augmentations [14], or another neural network summarizing global information [27, 43]. Further, define encoder(s) and latent representations $z_1 = g_1(v_1)$ and $z_2 = g_2(v_1)$. The encoder mappings may be constrained by $\mathcal{G}_1$ and $\mathcal{G}_2$ (e.g. architecturally). In some works, g_1 and g_2 may share some [27] or all [14] parameters. The goal is to find encoder mappings which maximize the mutual information between the outputs:

$$\max_{g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2} I'(g_1(v_1); g_2(v_2)) \tag{1}$$

This objective is shown [53] to lower-bound the true InfoMax objective. Perhaps the most widely adapted estimator is the InfoNCE estimator [43] which provides an unnormalized lower bound on the mutual information by optimizing the objective [43]:

$$\mathcal{L}_{NCE} := -\mathbb{E}_{v_1, v_2^-, v_2^+}\Big[\log \frac{\exp(f(g_1(v_1), g_2(v_2^+)))}{\exp(f(g_1(v_1), g_2(v_2^+))) + \sum_{j=1}^{N-1} \exp(f(g_1(v_1), g_2(v_{2_j}^-)))}\Big], \tag{2}$$

where $(v_1, v_2^+) \sim p(v_1, v_2)$ is a "real" pair of views drawn from their empirical joint distribution, and we draw negative samples $v_2^- \sim p(v_2)$ from the marginal distribution to form $N - 1$ "fake" pairs. N denotes the total number of pairs (and, in practice, often refers to the batch size). In Arora et al. [6], losses in this general form are termed "contrastive learning". Note that this is essentially a cross-entropy to distinguish one positive pair from $N - 1$ negative pairs, where f is a "critic" classifier (reminiscent of adversarial learning), and should learn to return high values for the "real" pair. As is common in deep learning, the expectation is calculated over multiple batches.

For a more detailed discussion of the connection between the InfoNCE loss, the InfoMax objective for representation learning, and other mutual information estimators, see Appendix A.

2.2 Choice of "Views" in Contrastive Learning Literature

Existing works select "views" of the input in different ways. These include using different time steps of an audio or video sequence [43, 49] or using different patches of the same image [43, 26, 27, 8]. Recently, contrastive learning between local and sequentially-global embeddings is used to establish representations for proteins [38]. Augmentations are an oft-used strategy for constructing different views [28, 24, 14], sometimes applied in conjunction with image patching [26, 8].

In this work, we argue that using evolution as a sequence augmentation strategy is a biologically and theoretically desirable choice to construct views. Previous work have explored evolutionary conservation as a means of sequence augmentation during training, such as augmenting a HMM using simulated evolution [34], or generating from a PSSM [7]. Other methods include using generative adversarial networks (GANs) for -omics data augmentation [17, 41] or injecting noise by replacing amino acids from an uniform distribution [31]. For genomic sequences, augmentations can be formed using reverse complements and extending (or cropping) genome flanks [13, 33].

3 Evolution as Sequence Augmentation

Here, we outline how phylogenetic augmentation fits into the contrastive learning framework, using SimCLR [14] as an example contrastive learning method. As outlined in Figure 1, homologous sequences can be considered as "evolutionary augmented views" of a common ancestor, x. Sequences v_1 and v_2 are encoded by an encoder $g(\cdot)$ to obtain embeddings z_1 and z_2. The pair of embeddings augmented from the same ancestor – that is, embeddings of homologous sequences – will be the positive pair $(v_1, v_2^+) \sim p(v_1, v_2)$. To sample negative samples $\{v_{2_j}^-\}_{j=1}^{N-1}$ from $p(v_2)$, we can draw negatives from all non-homologous sequences.

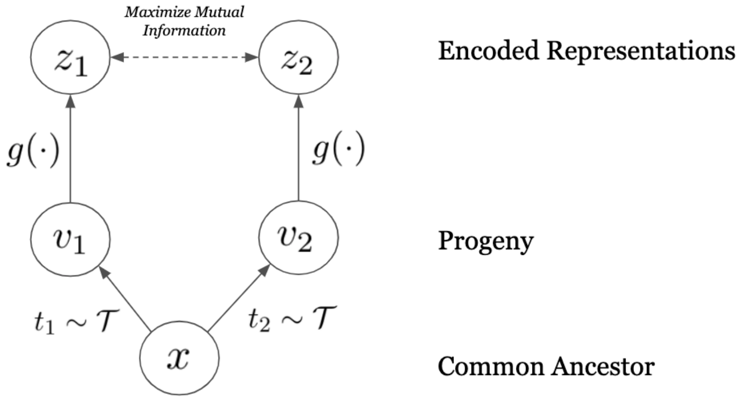

Figure 1: SimCLR [14] can be re-casted as a phylogenetic tree where augmentations are evolution. In the original Chen et al. [14] paper, x is an input image, and two image augmentation methods, t_1 and t_2, are sampled from a set of image augmentation methods $\mathcal{T}$, to produce image augmentation v_1 and v_2, which are then passed into a trainable encoder $g(\cdot)$ (i.e. g_1 and g_2 share parameters entirely). In conceptualizing evolution as an augmentation strategy, x can be viewed as a common ancestor, while $\mathcal{T}$ are possible evolutionary trajectories, characterized by different evolutionary distance, mutation and genetic drift, and t_1, t_2 are two example trajectories that lead to v_1 and v_2, sampled from a set of homologs. Note that notations are adapted from the original SimCLR paper for consistency with the current work.

The key idea is that properties of the ancestral sequence that were important for its biological function will be preserved in both descendants (i.e. views). By training the encoder to project these to nearby locations in the latent space, we ensure that proximity in the latent space corresponds to similar biological functions without explicit labels during pretraining, analogous to how SimCLR learns semantic content without image labels. We see that contrastive learning frameworks such as SimCLR can be directly adapted to capture phylogenetic principles.

4 Why Evolution as Biological Sequence Augmentation?

4.1 Invariant Representations Across Evolutionary "Noisy-Channels" Mirrors Comparative Genomics

Biological sequences are vehicles for information transmission. As such, information theoretic principles are directly applicable to biological sequence analyses, and therefore, this may be more a more powerful approach than methods based on the analogy with natural language [3, 47, 45, 19].

The analogy between molecular evolution and noisy-channel coding is well-rooted in prior work [20, 39, 54, 35]: DNA dictates information transmission across generations, which must be transferred through a noisy "mutation and drift channel". Further, as noted in Kimura [30], as the genotype-to-phenotype manifestation is information transfer, and genomic information is passed down by heredity, we may view functional phenotypes as "decoded" information that was transmitted from a common ancestor via molecular evolution. Drawing from these writings, we argue that using maximizing mutual information across homologs is a good proxy for structure and function [1], which are the central aims for biological sequence embeddings [45].

Even without relying on the mutual information estimation interpretation of the InfoNCE loss, the contrastive learning objective directly encourage representational invariance to shared features across views [14]. Therefore, in using phylogenetic relationships to create views, learned representations directly capture the philosophy of *evolutionary conservation* in comparative genomics: functional elements will be preserved in comparisons of related sequences, while non-functional sequences will decay. Hence, functional elements in biological sequences can be identified through sequence comparisons [23]. This is perhaps the most successfully employed presumption in bioinformatics [16, 4].

We therefore argue that InfoMax-based deep learning on evolutionary augmentation has two attractive features from the biological perspective: (1) Molecular evolution and the genotype-to-phenotype relationship has a clear analogy to information transmission; and (2) contrastive learning in this setting encourages agreement between important features across evolutionary views (homologous sequences), which directly mirrors comparative genomics.

4.2 Evolutionary Augmentation is a Theoretically Desirable View

Tian et al. [51] proposes the "InfoMin" principle for selecting optimal views. The authors theoretically and empirically demonstrate that good views should have *minimize* their shared MI while keeping *task-relevant* information intact for downstream uses. More formally, for a downstream classification task C to predict label $y \in \mathcal{Y}$ from x, the optimal representation $z^* = g_1(x)$ is the is the minimal sufficient statistic for task C, such the representation is as useful as access to x while disregarding all nuisance in x [52, 50]. Then, the optimal views of task C is $(v_1^*, v_2^*) = \min_{v_1;v_2} I(v_1; v_2)$, subject to $I(v_1; y) = I(v_2; y) = I(x; y)$. Given (v_1^*, v_2^*), the learned representations (z_1^*, z_2^*) are optimal for task C.[1]

There are two implications in adapting the InfoMin principle to biology which render evolutionary augmentations desirable. Firstly, sampling evolutionary trajectories $t_1, t_2 \sim \mathcal{T}$ to create $v_1 = t_1(x)$ and $v_2 = t_2(x)$ provide a simple way to reduce $I(v_1, v_2)$ by selecting paired views (v_1, v_2^+) with a greater phylogenetic distance between them. Secondly, note that in order to choose views based on the InfoMin principle, access to labels $y \in \mathcal{Y}$ and knowledge of task C is needed. In fact, *supervised* contrastive learning [29] empirically yields improved results by explicitly sampling negatives from a different downstream class. If given labels for a downstream biological task of interest (e.g. remote homology), one can explicitly negative sample from dissimilar classes (e.g. different folds); however, owing to the difficult label-acquisition process and open-ended nature of biological questions, access to $\mathcal{Y}$ – or even task C – may not be always possible. Further, it is often desirable for biological sequence embeddings to be "universal representations" [3] and applicable for a variety of downstream tasks [45]. As noted in Section 4.2, evolutionary conservation is a good proxy for many tasks of interest (e.g. structure and function).

Thus, we see that in transferring theoretical results for optimal view selection to the biological setting, evolution as augmentation is desirable, as: (1) It is easy to control shared mutual information between views; and (2) evolutionary conservation is a good semantic proxy for downstream labels, and implicitly performs *supervised* contrastive learning while still circumventing expensive experimental label gathering. Hence, it may be best considered a general strategy for weakly-supervised contrastive learning.

5 Conclusion

Current methods for self-supervised representation learning in biology are mostly adapted from NLP methods. Contrastive learning achieves state-of-the-art results in the image modality, and has a desirable theoretical property of being a lower-bound estimator of mutual information. We demonstrate how evolution can be used as a sequence augmentation strategy for contrastive learning, and provide justifications for doing so from biological and theoretical perspectives. More generally, data augmentation is a critical preprocessing step in many image analysis applications of deep learning, but is it less clear how to augment data for biological sequence analysis. As research in applications of deep learning in biology expand, we hope the view of evolution as augmentation will guide the ideation of deep learning methods in computational biology.

[1] The optimal property of representations (z_1^*, z_2^*) assumes access to an encoder which serve as a minimal sufficient statistic of the input [50]. More formally, a "sufficient encoder" $g_{\text{sufficient}}$ require that $g_{\text{sufficient}}(v_1)$ has kept all information about v_2 in v_1, and a "minimal sufficient encoder" $g_1 \in \mathcal{G}_{\text{sufficient}}$ discards all irrelevant "nuisance" information such that $I(g_1(v_1); v_1) \leq I(g_{\text{sufficient}}(v_1); v_1), \forall g_{\text{sufficient}} \in \mathcal{G}_{\text{sufficient}}$.

References

[1] Christoph Adami, Charles Ofria, and Travis C Collier. Evolution of biological complexity. *Proceedings of the National Academy of Sciences*, 97(9):4463–4468, 2000.

[2] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

[3] Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-only deep representation learning. *bioRxiv*, page 589333, 2019.

[4] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[5] Jose Juan Almagro Armenteros, Alexander Rosenberg Johansen, Ole Winther, and Henrik Nielsen. Language modelling for biological sequences–curated datasets and baselines. 2019.

[6] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

[7] Ehsaneddin Asgari, Nina Poerner, Alice McHardy, and Mohammad Mofrad. Deepprime2sec: Deep learning for protein secondary structure prediction from the primary sequences. *bioRxiv*, page 705426, 2019.

[8] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15509–15519, 2019.

[9] David Barber and Felix V Agakov. The im algorithm: a variational approach to information maximization. In *Advances in neural information processing systems*, page None, 2003.

[10] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

[11] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[12] Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure. *arXiv preprint arXiv:1902.08661*, 2019.

[13] Zhen Cao and Shihua Zhang. Simple tricks of convolutional neural network architectures improve dna–protein binding prediction. *Bioinformatics*, 35(11):1837–1843, 2019.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[15] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36 (2):183–212, 1983.

[16] Sean R. Eddy. Profile hidden markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763, 1998.

[17] Majid Ghorbani Eftekhar. Prediction of protein subcellular localization using deep learning and data augmentation. *bioRxiv*, 2020.

[18] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, and Burkhard Rost. End-to-end multitask learning, from protein language to protein features without alignments. *bioRxiv*, page 864405, 2019.

[19] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Debsindhu Bhowmik, et al. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225*, 2020.

[20] Lila L Gatlin et al. *Information theory and the living system.* Columbia University Press, 1972.

[21] Vladimir Gligorijevic, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Kyunghyun Cho, Tommi Vatanen, Daniel Berenberg, Bryn C Taylor, Ian M Fisk, Ramnik J Xavier, et al. Structure-based function prediction using graph convolutional networks. *bioRxiv*, page 786236, 2019.

[22] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[23] Ross C Hardison. Comparative genomics. *PLoS Biol*, 1(2):e58, 2003.

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[25] Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. Modeling the language of life-deep learning protein sequences. *bioRxiv*, page 614313, 2019.

[26] Olivier J Hénaff, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

[27] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[28] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1558–1567. JMLR. org, 2017.

[29] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

[30] Motoo Kimura. Natural selection as the process of accumulating genetic information in adaptive evolution. *Genetics Research*, 2(1):127–140, 1961.

[31] Satoshi Koide, Keisuke Kawano, and Takuro Kutsuna. Neural edit operations for biological sequences. In *Advances in Neural Information Processing Systems*, pages 4960–4970, 2018.

[32] Lingpeng Kong, Cyprien de Masson d'Autume, Wang Ling, Lei Yu, Zihang Dai, and Dani Yogatama. A mutual information maximization perspective of language representation learning. *arXiv preprint arXiv:1910.08350*, 2019.

[33] Wolfgang Kopp, Remo Monti, Annalaura Tamburrini, Uwe Ohler, and Altuna Akalin. Deep learning for genomics using janggu. *Nature communications*, 11(1):1–7, 2020.

[34] Anoop Kumar and Lenore Cowen. Augmented training of hidden markov models to recognize remote homologs via simulated evolution. *Bioinformatics*, 25(13):1602–1608, 2009.

[35] Ercan E Kuruoglu and Peter F Arndt. The information capacity of the genetic code: Is the natural code optimal? *Journal of Theoretical Biology*, 419:227–237, 2017.

[36] Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

[37] Sindy Löwe, Peter O'Connor, and Bastiaan Veeling. Putting an end to end-to-end: Gradient-isolated learning of representations. In *Advances in Neural Information Processing Systems*, pages 3033–3045, 2019.

[38] Amy X Lu, Haoran Zhang, Marzyeh Ghassemi, and Alan Moses. Self-supervised contrastive learning of protein representations by mutual information maximization. *bioRxiv*, 2020.

[39] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[40] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.

[41] Mohamed Marouf, Pierre Machart, Vikas Bansal, Christoph Kilian, Daniel S Magruder, Christian F Krebs, and Stefan Bonn. Realistic in silico generation and augmentation of single-cell rna-seq data using generative adversarial networks. *Nature communications*, 11(1):1–12, 2020.

[42] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[43] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[44] Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019.

[45] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. In *Advances in Neural Information Processing Systems*, pages 9686–9698, 2019.

[46] Adam J Riesselman, Jung-Eun Shin, Aaron W Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C Kruse, and Debora S Marks. Accelerating protein design using autoregressive generative models. *bioRxiv*, page 757252, 2019.

[47] Alexander Rives, Siddharth Goyal, Joshua Meier, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*, page 622803, 2019.

[48] Morgane Rivière, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. Unsupervised pretraining transfers well across languages. *arXiv preprint arXiv:2002.02848*, 2020.

[49] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1134–1141. IEEE, 2018.

[50] Stefano Soatto and Alessandro Chiuso. Visual representations: Defining properties and deep approximations. *arXiv preprint arXiv:1411.7676*, 2014.

[51] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[52] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020.

[53] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

[54] Susana Vinga. Information theory applications for biological sequence analysis. *Briefings in bioinformatics*, 15(3):376–389, 2014.

[55] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *arXiv preprint arXiv:2005.10242*, 2020.

[56] Kevin K Yang, Zachary Wu, Claire N Bedbrook, and Frances H Arnold. Learned protein embeddings for machine learning. *Bioinformatics*, 34(15):2642–2648, 2018.

A InfoMax Principle and Mutual Information Estimation for Representation Learning

A.1 InfoMax for Representation Learning

Using InfoMax for representation learning extends as far back as ICA [11]. As described in Equation 1, in recent years, works typically maximize mutual information of two *encoded* "views" of an input (e.g. different patches of an image, or augmentations). By the data processing inequality, Tschannen et al. [53] show that:

$$I(g_1(v_1); g_2(v_2)) \leq I(x; g_1(v_1), g_2(v_2)), \tag{3}$$

such that maximizing Equation 1 is equivalent to maximizing a lower bound on the true InfoMax objective. This ability to minimize the mutual information in the latent embedding space rather than directly between the input and the encoded output (as per the original InfoMax formulation) has two advantages [53]: (1) MI is notoriously difficult to estimate in high dimensions, and this allows for MI estimation in a lower-dimensional space; (2) creative choices of $\mathcal{G}$ can be used, which accommodates specific modeling needs and data intricacies.

A.2 InfoNCE Estimator

InfoNCE is one of many mutual information estimators, and following the rationale in Section A.1, the original Oord et al. [43] paper does this estimation in the embedding space. For the InfoNCE loss (Equation 2) which estimates $I(z_1; z_2) = I(g_1(v_1); g_2(v_2))$ in Equation 3, the optimal critic function is $f^*(z_1, z_2) = \frac{p(z_2|z_1)}{p(z_1)}$ [43]. Inserting this in the InfoNCE loss function (Equation 1) and rearranging, we have the bound [43, 44]:

$$I(z_1, z_2) \geq \log(N) - \mathcal{L}^*_{\text{NCE}} \tag{4}$$

where N is the number of samples. From Equation 4, note that the bound is tight when: (1) We use more samples for N which increases the $\log(N)$ term; and (2) we have a better f which results in a lower $\mathcal{L}_{\text{NCE}}$. Empirically, most works corroborate the former theoretical observation regarding N (exceptions being Arora et al. [6], Lu et al. [38]), while the latter observation regarding f does not usually hold, as will be further discussed in Section A.3.

The contrastive nature of the InfoNCE loss stems from its direct adaptation of the noise-contrastive estimation (NCE) method [22]. Noise-contrastive estimation was originally proposed for the problem of estimating parameters for unnormalized statistical models in high dimensions, by reducing the problem to simply estimating logistic regression parameters to distinguish between observed data and noise. In InfoNCE, the distinction is made between "similarity scores", as scored by critic $f(z_1, z_2)$, for one positive pair and $N - 1$ negative pairs of encoded views.

A.3 Other Mutual Information Estimators

The InfoNCE estimator is one of many approaches which builds on advancements in variational methods to create differentiable and tractable sample-based mutual information estimators in high dimensions [15, 9, 42, 2, 10, 43, 27]. Many of these estimations involve a "critic" classifier, f. In practice, f might be a bilinear model $z_1^T W z_2$ [43, 26, 51], separate models $\phi(z_1)^T \phi(z_2)$ [8], modelling concatenated data $\phi([z_1, z_2])$ [27], or a simple dot-product $z_1^T z_2$ [14, 38]. There may be a different f for each view [43], or a global f [27]. f is often trained jointly with g_1 and g_2.

The aim of f is often to approximate the unknown densities $p(B)$ and $p(B|A)$, or density ratios $\frac{p(A|B)}{p(B)} = \frac{p(B|A)}{p(A)}$ [44]. If $I(A, B)$ is high, then f should intuitively be able to easily assign high probabilities to those samples drawn from $p(A, B)$ [53]. The InfoNCE estimator reduces variance as compared to other estimators, by depending on multiple samples, but trades off bias to do so [44].

Importantly, it should be noted that whether the empirical success of the InfoNCE loss should be attributable to mutual information estimation has been questioned [44, 53], instead attributing success to geometric properties in the latent space [55]. For example, a higher-capacity f should increase tightness of the bound, as noted in Section A.2, yet hinders performance [53]. The development of MI-estimators useful for neural network training – and demystifying their empirical success – remains an active area of research. For the purposes of ideas in this work, we note that SimCLR-like contrastive losses itself intuitively maximizes agreement between views in the representation space without relying on the mutual information framing of the loss [14], and hence the connection to comparative genomics still hold.