Contrastive RNA Representation Learning Through Maximizing Mutual Information Between Splice Isoforms

Philip Fradkin[1,2,*] **Ruian Shi**[1,2,*], **Keren Isaev**[3], **Caitlin Harrigan** [1,2]
Quaid Morris[4], **Bo Wang**[1,2,5], **Brendan Frey**[1,6], **Leo J. Lee** [1,6]

[1] Vector Institute
[2] Department of Computer Science, University of Toronto
[3] Columbia University, New York Genome Center
[4] Memorial Sloan Kettering Cancer Center
[5] Peter Munk Cardiac Center, UHN
[5] Department of Electrical & Computer Engineering, University of Toronto

Abstract

In the face of rapidly accumulating genomic data, our understanding of the RNA regulatory code remains incomplete. Recent self-supervised methods in other domains have demonstrated the ability to learn rules underlying the data-generating process, such as sentence structure in language. Inspired by this, we extend contrastive learning techniques to genomic data by utilizing functional similarities between sequences generated through alternative splicing and gene duplication. We introduce IsoCLR, a model trained on a novel dataset with a contrastive objective enabling the learning of generalized RNA isoform representations. We validate representation utility on downstream tasks such as RNA half-life and mean ribosome load prediction. Our pre-training strategy yields competitive results using linear probing across 6 tasks, along with up to a two-fold increase in Pearson correlation in low-data conditions. Importantly, our exploration of the learned latent space reveals that our contrastive objective yields semantically meaningful representations, underscoring its potential as a valuable initialization technique for RNA property prediction.

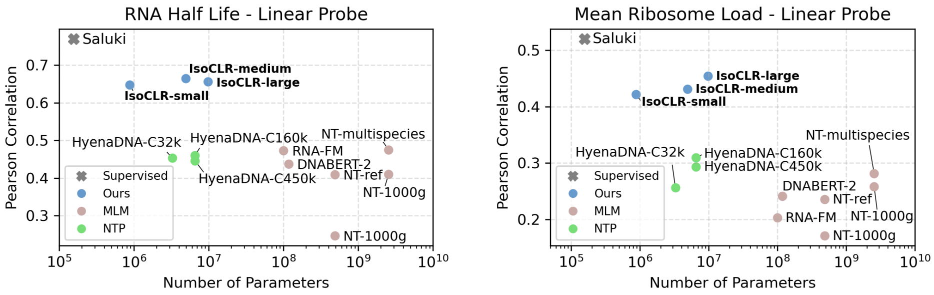
1 Introduction

Self-supervised learning (SSL) techniques have recently enabled the generation of effective representations that can be fine-tuned on related downstream tasks. This has reduced reliance on labeled data and demonstrated impressive generalization capabilities to a diversity of tasks (Tomasev et al. (2022); Radford et al. (2021)). SSL can be formulated through a data reconstruction objective, where a model is required to reconstruct a portion of the input data. Typical formulations have included next token prediction (NTP) and masked language modeling (MLM) (Devlin et al. (2018); Radford & Narasimhan (2018); Vaswani et al. (2017)). Recent self-supervised methods, including those by Ji et al. (2021), Chen et al. (2022), and Nguyen et al. (2023), have applied the self-supervised learning (SSL) paradigm to genomic data. However, the unique properties inherent to genomic data pose challenges for implementing reconstruction-based SSL objectives or supervised learning approaches.

Genomic sequences in the natural world are constrained by evolutionary viability, resulting in low natural diversity[1] and high mutual information across genomes from the same species (Taliun et al. (2021)). Latest estimates propose that up to five percent of the human genome is under constraint and

[*] Co-first Authors; correspondence to phil.fradkin@mail.utoronto.ca ljlee@psi.toronto.edu

[1] In the coding region (2% of human DNA), an average individual carries 27 ± 13 unique SNPs (Gudmundsson et al. (2021)).



(a)

(b)

Figure 1: Pearson correlation of linear regressions trained on representations by different self-supervised methods. RNA half-life and mean ribosome load are important cellular properties for regulating protein abundance. IsoCLR's contrastive objective outperforms existing SSL approaches.

can be considered high information content (Chen et al. (2024); Lindblad-Toh et al. (2011)). The remaining 95% of the genetic sequence lacks evidence of negative selection, meaning mutations may have little to no impact on organism fitness (Chen et al. (2024)). Without a strong biological inductive bias, existing reconstruction-based SSL models often reconstruct non-informative tokens, which can result in suboptimal representations. Due to the high-mutual information between samples, it is also difficult to scale the effective size of the training dataset to circumvent this issue. We find that recent applications of SSL methods to genomics (Dalla-Torre et al. (2023); Ji et al. (2021); Nguyen et al. (2023)) learn latent representations that are not well linearly separated (Figure 1). The gap between baseline SSL methods and supervised approaches remains large, while no clear trend exists between model size and performance.

In this work, we develop IsoCLR, a contrastive technique applied to genomic data with the purpose of learning effective RNA representations. Contrastive learning, a type of SSL, utilizes data augmentations to alter samples in a semantically meaningful way to learn effective representations (Koch (2015); Chen et al. (2020)). (Poole et al. (2019); Tschannen et al. (2019)). IsoCLR utilizes stronger biologically motivated inductive biases, making it less reliant on limited sequence diversity and capable of learning representations without extensive training on experimental data (Figure 2). To generate RNA augmentations, we rely on naturally occurring cellular and evolutionary processes: alternative splicing, and gene homology. Byproducts of these processes often generate RNAs with different sequences and similar functions (Pertea et al. (2018)). We identify paired RNA sequences generated by these processes and use them as augmentations for learning RNA embeddings. We investigate the effectiveness of learned representations by evaluating IsoCLR on six tasks including RNA half-life (HL) and mean ribosome load prediction (MRL) (Agarwal & Kelley (2022); Sugimoto & Ratcliffe (2022)). We find that IsoCLR outperforms other self-supervised methods and matches or exceeds supervised performance when fine-tuned. Our main contributions are:

- We create a novel RNA pre-training dataset by proposing augmentations for genomic sequences produced through homology, and alternative splicing processes.
- We propose IsoCLR, a novel method that employs a contrastive learning objective to learn robust RNA isoform representations across species.
- We conduct extensive evaluations of IsoCLR on tasks such as RNA half-life and mean ribosome load prediction to demonstrate improvements, particularly in the low data regime.

2 Methods

Contrastive Learning Dataset: Our proposed dataset for the contrastive learning objective is composed of annotated RNA transcriptomes (Frankish et al. (2021); O'Leary et al. (2016)). Using this information, we generate a six-track mature RNA representation, consisting of four one-hot encoded tracks encoding genomic sequence, a track indicating the 5' location of splice sites, and a track indicating the first nucleotide of every codon. The addition of splice site and coding sequence locations

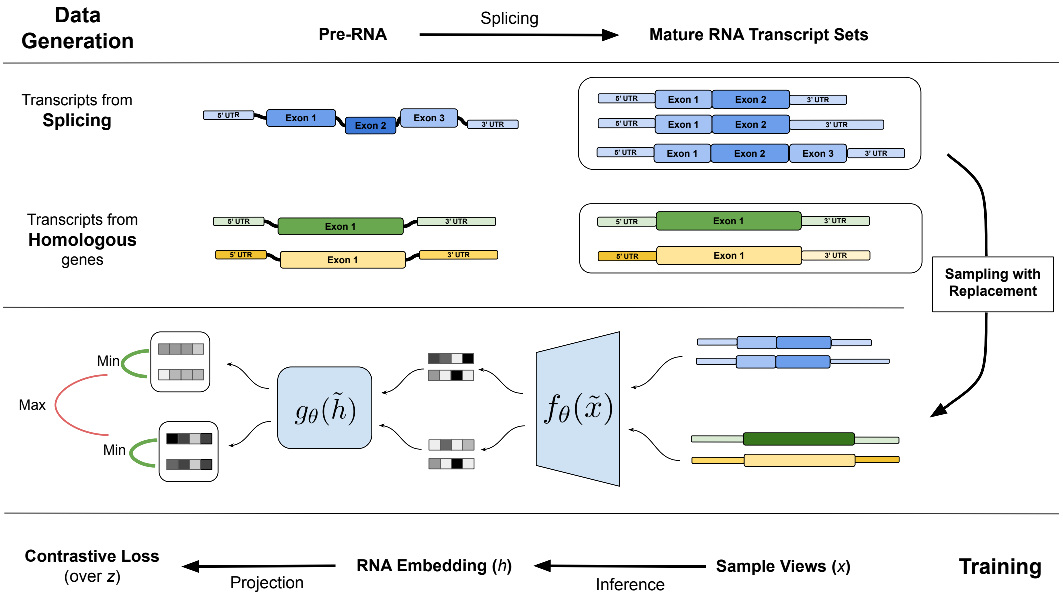


Figure 2: Description of the data generation and training processes for IsoCLR. The **upper** half of the figure demonstrates hypothetical examples for creating mature RNA sets from which positive data pairs are sampled. The first example demonstrates that a positive mature RNA set can be constructed from splicing. The second example demonstrates RNA set construction from gene homology. The **lower** half of the figure demonstrates the training process utilizing the generated mature RNA sets. First, RNAs are sampled with replacement from the sets and an RNA embedding is generated using a dilated convolutional residual encoder f. Then the representations are passed through a projector g, the normalized output of which is used to compute the decoupled contrastive loss.

has been shown to be beneficial for downstream genomic tasks (Agarwal & Kelley (2022)). Depending on the species analyzed and the transcriptome annotation resource used, between 25% and 50% of genes contain multiple isoforms which we then sample to use as augmentations (Table 3). Additionally, we used homology as a source of RNA isoform invariances. Homologous genes, which share structural similarities and encode similar functions, include paralogous and orthologous relationships, the former of which we use for training (Lesk (2020)). To annotate these relationships, we used the Homologene database (Sayers et al. (2023)).

Contrastive Learning Objective: During our contrastive training phase, we pool together sequences of splicing isoforms from homologous genes and treat them as views of the same object. Given a batch of N sequences (e.g. RNA isoforms) $x_1, ...x_N$ let x_i^1, x_i^2 be two splicing isoforms from a set of homologous genes. We pass these augmented views through a dilated convolutional encoder f resulting in the outputs h_i^1 and h_i^2. These representations are then fed into a multi-layer perceptron projection head, g the output of which is used to calculate normalized projections z_i as shown in Figure 2.

Normalized projections z_i are used to compute the contrastive loss, utilizing samples from the rest of the batch as negative examples. We use decoupled contrastive learning (DCL) as it has been shown to require smaller batch sizes, is less sensitive to hyperparameters such as learning rate, and the positive loss term can be weighted by sample difficulty (Yeh et al. (2021)). DCL iterates on the normalized temperature-scaled cross-entropy loss by splitting the contrastive objective into two terms: a similarity loss (positive) and a dissimilarity loss (negative) (Sohn (2016)). More formally, the positive and negative losses are calculated:

$$\mathcal{L}_{DCL,i}(\theta) = \log \sum_{z_k \in \mathcal{Z}, l \in 1,2}^{N} \mathbb{1}_{k \neq i} \exp(\langle z_i^1 \cdot z_k^l \rangle / \tau) - w_i \langle z_i^1, z_i^2 \rangle / \tau. \quad (1)$$

Table 1: Linear probing results for self-supervised methods. The embeddings were computed for each method and then linear regression was computed analytically using the corresponding labels for each task. Bolded numbers indicate the best performing model. IsoCLR-S; small, IsoCLR-M; medium and IsoCLR-L; large models. MSE; Mean square error.

Model Name	RNA HL Human MSE ↓	RNA HL Human R ↑	RNA HL Mouse MSE ↓	RNA HL Mouse R ↑	MRL MSE ↓	MRL R ↑	GO Mol Func ROC AUC ↑	Protein loc ROC AUC ↑	mRFP Expr R ↑
IsoCLR-S (ours)	0.59	0.65	**0.58**	**0.66**	0.76	0.42	0.86	**0.85**	0.65
IsoCLR-M (ours)	**0.56**	**0.66**	**0.58**	**0.66**	0.75	0.43	0.84	0.83	**0.71**
IsoCLR-L (ours)	0.57	**0.66**	**0.58**	**0.66**	**0.73**	**0.45**	**0.87**	0.84	0.66
DNA-BERT2	0.84	0.44	0.83	0.38	0.84	0.24	0.72	0.77	0.41
NT-500m-1000g	1.03	0.25	0.98	0.30	0.97	0.17	0.67	0.70	0.33
NT-500m-human-ref	0.90	0.41	0.85	0.40	0.91	0.24	0.72	0.73	0.52
NT-2.5b-1000g	0.89	0.41	0.85	0.37	0.92	0.26	0.73	0.70	0.51
NT-2.5b-multi-species	0.89	0.48	0.88	0.44	1.00	0.28	0.78	0.73	0.44
Hyena-32K-seqlen	0.83	0.45	0.77	0.44	0.84	0.26	0.75	0.79	0.43
Hyena-160K-seqlen	0.81	0.46	0.79	0.46	0.81	0.29	0.75	0.79	0.62
Hyena-450K-seqlen	0.80	0.45	0.80	0.46	0.82	0.29	0.75	0.78	0.55
RNA-FM	0.78	0.47	0.83	0.44	0.89	0.20	0.78	0.81	0.55

In the above z^1 and z^2 correspond to two views of the same object, z_k are views from other objects, τ is the temperature parameter set to 0.1, and $\mathbb{1}_{k \neq i}$ is an indicator function that evaluates to 1 when $k \neq i$. The above loss is computed for all the samples in the batch for both the sampled views $l \in 1, 2$. Due to the non-uniform number of views per set of sample, we use the term w_i for sample evidence weighting. Additional implementation details and datasets can be found in the appendix B.

3 Experimental Results

We demonstrate that contrastive pre-training across homologous genes and splicing isoforms improves downstream prediction across six tasks, including RNA HL and MRL prediction. We evaluate the effectiveness of the learned representation with three strategies: linear probing, full model fine-tuning, and latent space evaluations. In addition, we highlight the effectiveness of pre-trained representations in low-data settings, which can be found in the appendix along with additional findings C.

3.1 IsoCLR embeddings are predictive of diverse phenotypes

To evaluate the effectiveness of our pre-trained representations, we followed the conventional evaluation strategy of linear probing. The learned latent embedding is effective if $\exists$ $\mathbf{w}$ s.t. $\mathbf{w}^T\mathbf{X} + b = \hat{y}$, where $\mathbf{X}$ is a matrix of embeddings and $\hat{y}$ approximates y. To evaluate the above, we freeze the weights of the dilated convolutional encoder f and train a linear layer to predict labels for regression and classification tasks. Further experimental details are described in Appendix D.1. We demonstrate that IsoCLR outperforms other evaluated self-supervised methods on a diverse set of tasks by a substantial margin in Figure 1 and Table 1.

We observe mixed results with regard to scaling the number of model parameters in terms of linear probing. We see a clear improvement trend in MRL and GO class prediction, but we do not observe the same trend for other datasets. Similarly, we observe that for other self-supervised models, the number of parameters does not consistently improve performance. The clearest improvement trend we observe is in the Nucleotide Transformer work, where increasing the diversity of the training set by scaling the number of species improves performance. Similarly in our work, we aggregate highly informative sequences across 10 species. This demonstrates a path to further improve model effectiveness in genomic property prediction.

3.2 Fine tuning IsoCLR yields effective predictors

To assess whether the IsoCLR pre-training objective provides utility beyond an effective representation, we evaluate its performance by fully fine-tuning it and comparing it to a supervised model with matched architecture. We also evaluate its performance against a published method for the RNA half-life prediction, Saluki (Agarwal & Kelley (2022)). We find that the fully fine-tuned IsoCLR model matches the performance of Saluki on the RNA half-life task (Table 2). Furthermore,

Table 2: MSE and Pearson correlations (R) of full model fine-tuning on RNA half-life (HL), mean ribosome load (MRL), protein localization and mRFP expression tasks. Best models are shown in bold. Confidence intervals were computed using standard deviation over three random seeds. Additional experimental details are described in appendix D.2

Model Name	RNA HL Human MSE	RNA HL Human R	RNA HL Mouse MSE	RNA HL Mouse R	MRL MSE	MRL R	Protein Loc. ROC AUC	mRFP Expr. R
IsoCLR-S	**0.44** ± **1e-2**	**0.76** ± **8e-3**	**0.53** ± **3e-2**	**0.70** ± **1e-2**	**0.70** ± **3e-2**	**0.50** ± **2e-2**	0.84 ± 5e-3	**0.85** ± 1e-2
IsoCLR-M	0.48 ± 1e-2	0.74 ± 7e-3	0.56 ± 2e-2	0.69 ± 2e-2	0.76 ± 3e-2	0.49 ± 3e-2	**0.85** ± 5e-3	0.82 ± 1e-2
HyenaDNA-Tiny	0.79 ± 2e-2	0.47 ± 9e-3	0.79 ± 6e-2	0.48 ± 2e-2	0.91 ± 2e-2	0.05 ± 2e-2	0.82 ± 2e-4	0.13 ± 1e-1
HyenaDNA-Small	0.78 ± 1e-2	0.46 ± 5e-3	0.78 ± 3e-2	0.48 ± 8e-3	0.91 ± 2e-2	0.04 ± 4e-2	0.82 ± 1e-5	0.24 ± 2e-1
Supervised-S	0.50 ± 1e-2	0.71 ± 8e-3	0.59 ± 5e-2	0.66 ± 3e-3	0.69 ± 6e-2	0.50 ± 5e-2	0.84 ± 1e-2	0.79 ± 2e-2
Supervised-M	0.53 ± 3e-2	0.63 ± 3e-2	0.64 ± 8e-2	0.69 ± 2e-2	0.82 ± 6e-2	0.43 ± 6e-2	0.84 ± 3e-3	0.15 ± 2e-1
Saluki	**0.44** ± **1e-2**	**0.76** ± **1e-2**	**0.55** ± **5-e2**	**0.70** ± **3e-2**	**0.67** ± **4e-2**	**0.52** ± **2e-2**	0.80 ± 2e-3	0.38 ± 2e-2

we retrain the Saluki architecture for other tasks and identify that IsoCLR significantly outperforms those models for protein localization and mRFP expression prediction tasks. In addition, IsoCLR outperforms fine-tuned HyenaDNA models across an assortment of tasks. Other baseline SSL methods such as DNA-BERT2 and RNA-FM have limited input context windows, and cannot be easily applied to these tasks. For certain tasks, we observe that scaling the models results in performance degradation, but note that IsoCLR still significantly outperforms baselines with similar parameter counts.

4 Discussion

In this work, we demonstrate that by minimizing the distance between mature RNAs generated through gene duplication and alternative splicing, we are able to generate representations useful for RNA property prediction tasks. The pre-training is especially helpful in low data regimes when there are 200 or fewer data points with labels 4. These situations arise in molecular biology applications, especially in therapeutic domains where manufacturing and experiments can be expensive. We demonstrate that self-supervised pre-training is an approach for addressing data efficiency challenges present in genomics, and scaling to additional species can be an effective dataset expansion strategy.

Previous self-supervised works for genomic sequence property prediction have focused on reconstruction objectives like masked language modeling or next token prediction (Ji et al. (2021); Dalla-Torre et al. (2023)). As previously discussed, most genomic positions are under little to no negative selection, and are not as informative for model training. Thus, predicting the corresponding tokens introduces little new information to the model. In this work, we instead choose to utilize a stronger inductive bias, minimizing the distance between functionally similar sequences. By relying on a more structured objective, we are able to outperform models that are multiple orders of magnitude larger in terms of parameter count. A possible limitation of our approach is that by minimizing the representational distance between related sequences, we remove important signals for predicting certain properties. Are there property prediction tasks for which our inductive bias is actually detrimental compared to a randomly initialized model? For RNA half-life, Spies et al. (2013) demonstrated that in more than 85% of genes, isoform choice has no statistically discernible effect.

An important question to address is why we expect that minimizing distances between RNA isoforms would be useful for predicting seemingly unrelated phenotypes like RNA half-life or codon optimality in mRFP prediction. One hypothesis is that alternative splicing and gene duplication preserve core functional RNA segments. Through the contrastive pre-training procedure, we identify these shared regions between diverse sequences. Indeed, a recent work proposes that contrastive methods are effective due to block separating latent variables shared between views (von Kügelgen et al. (2022)). By utilizing decoupled contrastive learning, diverse sequences are pushed apart, thus uniformly distributing samples in the latent space which helps with downstream tasks (Yeh et al. (2021); Wang & Liu (2021)). Through encoding these invariances, we find that IsoCLR is able to learn complex RNA properties such as cellular component localization and RNA half-life.

5 Conclusions

In this work, we propose a novel, self-supervised contrastive objective for learning mature RNA isoform representations. We show that this approach is an effective strategy to address major challenges for cellular property prediction: data efficiency, and model generalizability. We demonstrate that IsoCLR representations are effective in the low data setting, paving the path to true few-shot learning for RNA property prediction. Finally, fine-tuning IsoCLR matches the performance of supervised models, beating out other self-supervised methods.

References

V. Agarwal and D. R. Kelley. The genetic and biochemical determinants of mRNA degradation rates in mammals. *Genome Biol*, 23(1):245, Nov 2022.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. doi: 10.1038/75556.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A Cookbook of Self-Supervised Learning. *arXiv e-prints*, April 2023. doi: 10.48550/arXiv.2304.12210.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. *arXiv e-prints*, May 2021. doi: 10.48550/arXiv.2105.04906.

Ido Ben-Shaul, Ravid Shwartz-Ziv, Tomer Galanti, Shai Dekel, and Yann LeCun. Reverse Engineering Self-Supervised Learning. *arXiv e-prints*, May 2023. doi: 10.48550/arXiv.2305.15614.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, Irwin King, and Yu Li. Interpretable RNA Foundation Model from Unannotated Data for Highly Accurate RNA Structure and Function Predictions. *arXiv e-prints*, art. arXiv:2204.00300, April 2022. doi: 10.48550/arXiv.2204.00300.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *arXiv e-prints*, art. arXiv:1606.00915, June 2016. doi: 10.48550/arXiv.1606.00915.

Siwei Chen, Laurent C. Francioli, Julia K. Goodrich, Ryan L. Collins, Masahiro Kanai, Qingbo Wang, Jessica Alföldi, Nicholas A. Watts, Christopher Vittal, Laura D. Gauthier, Timothy Poterba, Michael W. Wilson, Yekaterina Tarasova, William Phu, Riley Grant, Mary T. Yohannes, Zan Koenig, Yossi Farjoun, Eric Banks, Stacey Donnelly, Stacey Gabriel, Namrata Gupta, Steven Ferriera, Charlotte Tolonen, Sam Novod, Louis Bergelson, David Roazen, Valentin Ruano-Rubio, Miguel Covarrubias, Christopher Llanwarne, Nikelle Petrillo, Gordon Wade, Thibault Jeandet, Ruchi Munshi, Kathleen Tibbetts, Maria Abreu, Carlos A. Aguilar Salinas, Tariq Ahmad, Christine M. Albert, Diego Ardissino, Irina M. Armean, Elizabeth G. Atkinson, Gil Atzmon, John Barnard, Samantha M. Baxter, Laurent Beaugerie, Emelia J. Benjamin, David Benjamin, Michael Boehnke, Lori L. Bonnycastle, Erwin P. Bottinger, Donald W. Bowden, Matthew J. Bown, Harrison Brand, Steven Brant, Ted Brookings, Sam Bryant, Sarah E. Calvo, Hannia Campos, John C. Chambers, Juliana C. Chan, Katherine R. Chao, Sinéad Chapman, Daniel I. Chasman, Rex Chisholm, Judy Cho, Rajiv Chowdhury, Mina K. Chung, Wendy K. Chung, Kristian Cibulskis, Bruce Cohen, Kristen M. Connolly, Adolfo Correa, Beryl B. Cummings, Dana Dabelea, John Danesh, Dawood Darbar, Phil Darnowsky, Joshua Denny, Ravindranath Duggirala, Josée Dupuis, Patrick T. Ellinor, Roberto Elosua, James Emery, Eleina England, Jeanette Erdmann, Tõnu Esko, Emily Evangelista, Diane Fatkin, Jose Florez, Andre Franke, Jack Fu, Martti Färkkilä, Kiran Garimella, Jeff Gentry, Gad Getz, David C. Glahn, Benjamin Glaser, Stephen J. Glatt, David Goldstein, Clicerio Gonzalez, Leif Groop, Sanna Gudmundsson, Andrea Haessly, Christopher

Haiman, Ira Hall, Craig L. Hanis, Matthew Harms, Mikko Hiltunen, Matti M. Holi, Christina M. Hultman, Chaim Jalas, Mikko Kallela, Diane Kaplan, Jaakko Kaprio, Sekar Kathiresan, Eimear E. Kenny, Bong-Jo Kim, Young Jin Kim, Daniel King, George Kirov, Jaspal Kooner, Seppo Koskinen, Harlan M. Krumholz, Subra Kugathasan, Soo Heon Kwak, Markku Laakso, Nicole Lake, Trevyn Langsford, Kristen M. Laricchia, Terho Lehtimäki, Monkol Lek, Emily Lipscomb, Ruth J. F. Loos, Wenhan Lu, Steven A. Lubitz, Teresa Tusie Luna, Ronald C. W. Ma, Gregory M. Marcus, Jaume Marrugat, Kari M. Mattila, Steven McCarroll, Mark I. McCarthy, Jacob L. McCauley, Dermot McGovern, Ruth McPherson, James B. Meigs, Olle Melander, Andres Metspalu, Deborah Meyers, Eric V. Minikel, Braxton D. Mitchell, Vamsi K. Mootha, Aliya Naheed, Saman Nazarian, Peter M. Nilsson, Michael C. O'Donovan, Yukinori Okada, Dost Ongur, Lorena Orozco, Michael J. Owen, Colin Palmer, Nicholette D. Palmer, Aarno Palotie, Kyong Soo Park, Carlos Pato, Ann E. Pulver, Dan Rader, Nazneen Rahman, Alex Reiner, Anne M. Remes, Dan Rhodes, Stephen Rich, John D. Rioux, Samuli Ripatti, Dan M. Roden, Jerome I. Rotter, Nareh Sahakian, Danish Saleheen, Veikko Salomaa, Andrea Saltzman, Nilesh J. Samani, Kaitlin E. Samocha, Alba Sanchis-Juan, Jeremiah Scharf, Molly Schleicher, Heribert Schunkert, Sebastian Schönherr, Eleanor G. Seaby, Svati H. Shah, Megan Shand, Ted Sharpe, Moore B. Shoemaker, Tai Shyong, Edwin K. Silverman, Moriel Singer-Berk, Pamela Sklar, Jonathan T. Smith, J. Gustav Smith, Hilkka Soininen, Harry Sokol, Rachel G. Son, Jose Soto, Tim Spector, Christine Stevens, Nathan O. Stitziel, Patrick F. Sullivan, Jaana Suvisaari, E. Shyong Tai, Kent D. Taylor, Yik Ying Teo, Ming Tsuang, Tiinamaija Tuomi, Dan Turner, Teresa Tusie-Luna, Erkki Vartiainen, Marquis Vawter, Lily Wang, Arcturus Wang, James S. Ware, Hugh Watkins, Rinse K. Weersma, Ben Weisburd, Maija Wessman, Nicola Whiffin, James G. Wilson, Ramnik J. Xavier, Anne O'Donnell-Luria, Matthew Solomonson, Cotton Seed, Alicia R. Martin, Michael E. Talkowski, Heidi L. Rehm, Mark J. Daly, Grace Tiao, Benjamin M. Neale, Daniel G. MacArthur, and Konrad J. Karczewski. Author correction: A genomic mutational constraint map using variation in 76, 156 human genomes. *Nature*, January 2024. ISSN 1476-4687.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv e-prints*, art. arXiv:2002.05709, February 2020. doi: 10.48550/arXiv.2002.05709.

The Gene Ontology Consortium, Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil dos Santos, Steven Marygold, Victor Strelets, Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C C Saverimuttu, Renzhi Su, Kate E Thurlow, Ruth C Lovering, Colin Logie, Snezhana Oliferenko, Judith Blake, Karen Christie, Lori Corbani, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A Aleksander, J Michael Cherry, Stacia R Engel, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Marek S Skrzypek, Shuai Weng, Edith D Wong, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena

Zarowiecki, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 03 2023. ISSN 1943-2631. doi: 10.1093/genetics/iyad031.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023. doi: 10.1101/2023.01.11.523679.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, art. arXiv:1810.04805, October 2018. doi: 10.48550/arXiv.1810.04805.

C. B. Do, M. S. Mahabhashyam, M. Brudno, and S. Batzoglou. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*, 15(2):330–340, Feb 2005.

A. Frankish, M. Diekhans, I. Jungreis, J. Lagarde, J. E. Loveland, J. M. Mudge, C. Sisu, J. C. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, C. Boix, S. Carbonell Sala, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. n, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, K. L. Howe, T. Hunt, O. G. Izuogu, R. Johnson, F. J. Martin, L. nez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, F. C. Riera, M. Ruffier, B. M. Schmitt, E. Stapleton, M. M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, M. Y. Wolf, J. Xu, Y. T. Yang, A. Yates, D. Zerbino, Y. Zhang, J. S. Choudhary, M. Gerstein, R. ó, T. J. P. Hubbard, M. Kellis, B. Paten, M. L. Tress, and P. Flicek. GENCODE 2021. *Nucleic Acids Res*, 49(D1):D916–D923, Jan 2021.

J. Frazer, P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, Y. Gal, and D. S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, Nov 2021.

Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann Lecun. On the duality between contrastive and non-contrastive self-supervised learning. *arXiv e-prints*, art. arXiv:2206.02574, June 2022. doi: 10.48550/arXiv.2206.02574.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised Learning. *arXiv e-prints*, June 2020. doi: 10.48550/arXiv.2006.07733.

Sanna Gudmundsson, Moriel Singer-Berk, Nicholas A. Watts, William Phu, Julia K. Goodrich, Matthew Solomonson, Heidi L. Rehm, Daniel G. MacArthur, and Anne O'Donnell-Luria and. Variant interpretation using population databases: Lessons from gnomAD. *Human Mutation*, 43 (8):1012–1030, December 2021. doi: 10.1002/humu.24309.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv e-prints*, art. arXiv:1512.03385, December 2015. doi: 10.48550/arXiv.1512.03385.

Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15): 2112–2120, Aug 2021.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. dek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, Aug 2021.

D. R. Kelley, Y. A. Reshef, M. Bileschi, D. Belanger, C. Y. McLean, and J. Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res*, 28(5):739–750, May 2018.

Gregory R. Koch. Siamese neural networks for one-shot image recognition. 2015.

N. K. Lee, Z. Tang, S. Toneyan, and P. K. Koo. EvoAug: improving generalization and interpretability of genomic deep neural networks with evolution-inspired data augmentations. *Genome Biol*, 24(1):105, May 2023.

Arthur M. Lesk. *Chapter 4 Alignments and phylogenetic trees*. Oxford University Press, 2020.

Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. Dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, Mar 2023.

Kerstin Lindblad-Toh, Manuel Garber, Or Zuk, Michael F. Lin, Brian J. Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, Evan Mauceli, Lucas D. Ward, Craig B. Lowe, Alisha K. Holloway, Michele Clamp, Sante Gnerre, Jessica Alföldi, Kathryn Beal, Jean Chang, Hiram Clawson, James Cuff, Federica Di Palma, Stephen Fitzgerald, Paul Flicek, Mitchell Guttman, Melissa J. Hubisz, David B. Jaffe, Irwin Jungreis, W. James Kent, Dennis Kostka, Marcia Lara, Andre L. Martins, Tim Massingham, Ida Moltke, Brian J. Raney, Matthew D. Rasmussen, Jim Robinson, Alexander Stark, Albert J. Vilella, Jiayu Wen, Xiaohui Xie, Michael C. Zody, Kim C. Worley, Christie L. Kovar, Donna M. Muzny, Richard A. Gibbs, Wesley C. Warren, Elaine R. Mardis, George M. Weinstock, Richard K. Wilson, Ewan Birney, Elliott H. Margulies, Javier Herrero, Eric D. Green, David Haussler, Adam Siepel, Nick Goldman, Katherine S. Pollard, Jakob S. Pedersen, Eric S. Lander, and Manolis Kellis. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482, October 2011. ISSN 1476-4687. doi: 10.1038/nature10530.

J. Linder, S. E. Koplik, A. Kundaje, and G. Seelig. Deciphering the impact of genetic variation on human polyadenylation using APARENT2. *Genome Biol*, 23(1):232, Nov 2022.

Amy X. Lu, Alex X. Lu, and Alan Moses. Evolution Is All You Need: Phylogenetic Augmentation for Contrastive Learning. *arXiv e-prints*, December 2020. doi: 10.48550/arXiv.2012.13475.

Amy X. Lu, Haoran Zhang, Marzyeh Ghassemi, and Alan Moses. Self-supervised contrastive learning of protein representations by mutual information maximization. *bioRxiv*, 2020. doi: 10.1101/2020.09.04.283929.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021. doi: 10.1101/2021.07.09.450648.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Chris Ré. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. *arXiv e-prints*, June 2023. doi: 10.48550/arXiv.2306.15794.

Thijs Nieuwkoop, Barbara R Terlouw, Katherine G Stevens, Richard A Scheltema, Dick de Ridder, John van der Oost, and Nico J Claassens. Revealing determinants of translation efficiency via whole-gene codon randomization and machine learning. *Nucleic Acids Res.*, 51(5):2363–2376, March 2023.

N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, 44(D1):D733–745, Jan 2016.

M. Pertea, A. Shumate, G. Pertea, A. Varabyou, F. P. Breitwieser, Y. C. Chang, A. K. Madugundu, A. Pandey, and S. L. Salzberg. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol*, 19(1):208, Nov 2018.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On Variational Bounds of Mutual Information. *arXiv e-prints*, May 2019. doi: 10.48550/arXiv.1905.06922.

Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *arXiv e-prints*, pp. arXiv:2103.00020, February 2021. doi: 10.48550/arXiv.2103.00020.

Jose Manuel Rodriguez, Fernando Pozo, Daniel Cerdán-Vélez, Tomás Di Domenico, Jesús Vázquez, and Michael L Tress. APPRIS: selecting functionally important isoforms. *Nucleic Acids Res.*, 50 (D1):D54–D59, January 2022.

E. W. Sayers, E. E. Bolton, J. R. Brister, K. Canese, J. Chan, D. C. Comeau, C. M. Farrell, M. Feldgarden, A. M. Fine, K. Funk, E. Hatcher, S. Kannan, C. Kelly, S. Kim, W. Klimke, M. J. Landrum, S. Lathrop, Z. Lu, T. L. Madden, A. Malheiro, A. Marchler-Bauer, T. D. Murphy, L. Phan, S. Pujar, S. H. Rangwala, V. A. Schneider, T. Tse, J. Wang, J. Ye, B. W. Trawick, K. D. Pruitt, and S. T. Sherry. Database resources of the National Center for Biotechnology Information in 2023. *Nucleic Acids Res*, 51(D1):D29–D38, Jan 2023.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

N. Spies, C. B. Burge, and D. P. Bartel. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res*, 23(12):2078–2090, Dec 2013.

Yoichiro Sugimoto and Peter J. Ratcliffe. Isoform-resolved mRNA profiling of ribosome load defines interplay of HIF and mTOR dysregulation in kidney cancer. *Nature Structural Molecular Biology*, 29(9):871–880, September 2022. doi: 10.1038/s41594-022-00819-2.

D. Taliun, D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech, R. Torres, S. A. G. Taliun, A. Corvelo, S. M. Gogarten, H. M. Kang, A. N. Pitsillides, J. LeFaive, S. B. Lee, X. Tian, B. L. Browning, S. Das, A. K. Emde, W. E. Clarke, D. P. Loesch, A. C. Shetty, T. W. Blackwell, A. V. Smith, Q. Wong, X. Liu, M. P. Conomos, D. M. Bobo, F. Aguet, C. Albert, A. Alonso, K. G. Ardlie, D. E. Arking, S. Aslibekyan, P. L. Auer, J. Barnard, R. G. Barr, L. Barwick, L. C. Becker, R. L. Beer, E. J. Benjamin, L. F. Bielak, J. Blangero, M. Boehnke, D. W. Bowden, J. A. Brody, E. G. Burchard, B. E. Cade, J. F. Casella, B. Chalazan, D. I. Chasman, Y. I. Chen, M. H. Cho, S. H. Choi, M. K. Chung, C. B. Clish, A. Correa, J. E. Curran, B. Custer, D. Darbar, M. Daya, M. de Andrade, D. L. DeMeo, S. K. Dutcher, P. T. Ellinor, L. S. Emery, C. Eng, D. Fatkin, T. Fingerlin, L. Forer, M. Fornage, N. Franceschini, C. Fuchsberger, S. M. Fullerton, S. Germer, M. T. Gladwin, D. J. Gottlieb, X. Guo, M. E. Hall, J. He, N. L. Heard-Costa, S. R. Heckbert, M. R. Irvin, J. M. Johnsen, A. D. Johnson, R. Kaplan, S. L. R. Kardia, T. Kelly, S. Kelly, E. E. Kenny, D. P. Kiel, R. Klemmer, B. A. Konkle, C. Kooperberg, A. ttgen, L. A. Lange, J. Lasky-Su, D. Levy, X. Lin, K. H. Lin, C. Liu, R. J. F. Loos, L. Garman, R. Gerszten, S. A. Lubitz, K. L. Lunetta, A. C. Y. Mak, A. Manichaikul, A. K. Manning, R. A. Mathias, D. D. McManus, S. T. McGarvey, J. B. Meigs, D. A. Meyers, J. L. Mikulla, M. A. Minear, B. D. Mitchell, S. Mohanty, M. E. Montasser, C. Montgomery, A. C. Morrison, J. M. Murabito, A. Natale, P. Natarajan, S. C. Nelson, K. E. North, J. R. O'Connell, N. D. Palmer, N. Pankratz, G. M. Peloso, P. A. Peyser, J. Pleiness, W. S. Post, B. M. Psaty, D. C. Rao, S. Redline, A. P. Reiner, D. Roden, J. I. Rotter, I. Ruczinski, C. Sarnowski, S. Schoenherr, D. A. Schwartz, J. S. Seo, S. Seshadri, V. A. Sheehan, W. H. Sheu, M. B. Shoemaker, N. L. Smith, J. A. Smith, N. Sotoodehnia, A. M. Stilp, W. Tang, K. D. Taylor, M. Telen, T. A. Thornton, R. P. Tracy, D. J. Van Den Berg, R. S. Vasan, K. A. Viaud-Martinez,

S. Vrieze, D. E. Weeks, B. S. Weir, S. T. Weiss, L. C. Weng, C. J. Willer, Y. Zhang, X. Zhao, D. K. Arnett, A. E. Ashley-Koch, K. C. Barnes, E. Boerwinkle, S. Gabriel, R. Gibbs, K. M. Rice, S. S. Rich, E. K. Silverman, P. Qasba, W. Gan, G. J. Papanicolaou, D. A. Nickerson, S. R. Browning, M. C. Zody, S. llner, J. G. Wilson, L. A. Cupples, C. C. Laurie, C. E. Jaquish, R. D. Hernandez, T. D. O'Connor, G. R. Abecasis, N. Abe, L. Almasy, S. Ament, P. Anderson, P. Anugu, D. Applebaum-Bowden, T. Assimes, D. Avramopoulos, E. Barron-Casella, T. Beaty, G. Beck, D. Becker, A. Beitelshees, T. Benos, M. Bezerra, J. Bis, R. Bowler, U. Broeckel, J. Broome, K. Bunting, C. Bustamante, E. Buth, J. Cardwell, V. Carey, C. Carty, R. Casaburi, P. Castaldi, M. Chaffin, C. Chang, Y. C. Chang, S. Chavan, B. J. Chen, W. M. Chen, L. M. Chuang, R. H. Chung, S. Comhair, E. Cornell, C. Crandall, J. Crapo, J. Curtis, C. Damcott, S. David, C. Davis, L. L. Fuentes, M. DeBaun, R. Deka, S. Devine, Q. Duan, R. Duggirala, J. P. Durda, C. Eaton, L. Ekunwe, A. El Boueiz, S. Erzurum, C. Farber, M. Flickinger, M. Fornage, C. Frazar, M. Fu, L. Fulton, S. Gao, Y. Gao, M. Gass, B. Gelb, X. P. Geng, M. Geraci, A. Ghosh, C. Gignoux, D. Glahn, D. W. Gong, H. Goring, S. Graw, D. Grine, C. C. Gu, Y. Guan, N. Gupta, J. Haessler, N. L. Hawley, B. Heavner, D. Herrington, C. Hersh, B. Hidalgo, J. Hixson, B. Hobbs, J. Hokanson, E. Hong, K. Hoth, C. A. Hsiung, Y. J. Hung, H. Huston, C. M. Hwu, R. Jackson, D. Jain, M. A. Jhun, C. Johnson, R. Johnston, K. Jones, S. Kathiresan, A. Khan, W. Kim, G. Kinney, H. Kramer, C. Lange, E. Lange, L. Lange, C. Laurie, M. LeBoff, J. Lee, S. S. Lee, W. J. Lee, D. Levine, J. Lewis, X. Li, Y. Li, H. Lin, H. Lin, K. H. Lin, S. Liu, Y. Liu, Y. Liu, J. Luo, M. Mahaney, B. Make, J. Manson, L. Margolin, L. Martin, S. Mathai, S. May, P. McArdle, M. L. McDonald, S. McFarland, D. McGoldrick, C. McHugh, H. Mei, L. Mestroni, N. Min, R. L. Minster, M. Moll, A. Moscati, S. Musani, S. Mwasongwe, J. C. Mychaleckyj, G. Nadkarni, R. Naik, T. Naseri, S. Nekhai, B. Neltner, H. Ochs-Balcom, D. Paik, J. Pankow, A. Parsa, J. M. Peralta, M. Perez, J. Perry, U. Peters, L. S. Phillips, T. Pollin, J. P. Becker, M. P. Boorgula, M. Preuss, D. Qiao, Z. Qin, N. Rafaels, L. Raffield, L. Rasmussen-Torvik, A. Ratan, R. Reed, E. Regan, M. S. Reupena, C. Roselli, P. Russell, S. Ruuska, K. Ryan, E. C. Sabino, D. Saleheen, S. Salimi, S. Salzberg, K. Sandow, V. G. Sankaran, C. Scheller, E. Schmidt, K. Schwander, F. Sciurba, C. Seidman, J. Seidman, S. L. Sherman, A. Shetty, W. H. Sheu, B. Silver, J. Smith, T. Smith, S. Smoller, B. Snively, M. Snyder, T. Sofer, G. Storm, E. Streeten, Y. J. Sung, J. Sylvia, A. Szpiro, C. Sztalryd, H. Tang, M. Taub, M. Taylor, S. Taylor, M. Threlkeld, L. Tinker, D. Tirschwell, S. Tishkoff, H. Tiwari, C. Tong, M. Tsai, D. Vaidya, P. VandeHaar, T. Walker, R. Wallace, A. Walts, F. F. Wang, H. Wang, K. Watson, J. Wessel, K. Williams, L. K. Williams, C. Wilson, J. Wu, H. Xu, L. Yanek, I. Yang, R. Yang, N. Zaghloul, M. Zekavat, S. X. Zhao, W. Zhao, D. Zhi, X. Zhou, and X. Zhu. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, 590(7845):290–299, Feb 2021.

Peter J Thul, Lovisa Åkesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, Anna Asplund, Lars Björk, Lisa M Breckels, Anna Bäckström, Frida Danielsson, Linn Fagerberg, Jenny Fall, Laurent Gatto, Christian Gnann, Sophia Hober, Martin Hjelmare, Fredric Johansson, Sunjae Lee, Cecilia Lindskog, Jan Mulder, Claire M Mulvey, Peter Nilsson, Per Oksvold, Johan Rockberg, Rutger Schutten, Jochen M Schwenk, Åsa Sivertsson, Evelina Sjöstedt, Marie Skogs, Charlotte Stadler, Devin P Sullivan, Hanna Tegel, Casper Winsnes, Cheng Zhang, Martin Zwahlen, Adil Mardinoglu, Fredrik Pontén, Kalle von Feilitzen, Kathryn S Lilley, Mathias Uhlén, and Emma Lundberg. A subcellular map of the human proteome. *Science*, 356 (6340), May 2017.

Nenad Tomasev, Ioana Bica, Brian McWilliams, Lars Buesing, Razvan Pascanu, Charles Blundell, and Jovana Mitrovic. Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet? *arXiv e-prints*, January 2022. doi: 10.48550/arXiv.2201.05119.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On Mutual Information Maximization for Representation Learning. *arXiv e-prints*, July 2019. doi: 10.48550/arXiv.1907.13625.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv e-prints*, July 2018. doi: 10.48550/arXiv.1807.03748.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Silvia Vidali, Raffaele Gerlini, Kyle Thompson, Jill E Urquhart, Jana Meisterknecht, Juan Antonio Aguilar-Pimentel, Oana V Amarie, Lore Becker, Catherine Breen, Julia Calzada-Wack, Nirav F Chhabra, Yi-Li Cho, Patricia da Silva-Buttkus, René G Feichtinger, Kristine Gampe, Lillian Garrett, Kai P Hoefig, Sabine M Hölter, Elisabeth Jameson, Tanja Klein-Rodewald, Stefanie Leuchtenberger, Susan Marschall, Philipp Mayer-Kuckuk, Gregor Miller, Manuela A Oestereicher, Kristina Pfannes, Birgit Rathkolb, Jan Rozman, Charlotte Sanders, Nadine Spielmann, Claudia Stoeger, Marten Szibor, Irina Treise, John H Walter, Wolfgang Wurst, Johannes A Mayr, Helmut Fuchs, Ulrich Gärtner, Ilka Wittig, Robert W Taylor, William G Newman, Holger Prokisch, Valerie Gailus-Durner, and Martin Hrabě de Angelis. Characterising a homozygous two-exon deletion in UQCRH: comparing human and mouse phenotypes. *EMBO Mol. Med.*, 13(12):e14397, December 2021.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style, 2022.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss, 2021.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled Contrastive Learning. *arXiv e-prints*, art. arXiv:2110.06848, October 2021. doi: 10.48550/arXiv.2110.06848.

T. Yu, H. Cui, J. C. Li, Y. Luo, G. Jiang, and H. Zhao. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, Mar 2023.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. *arXiv e-prints*, art. arXiv:1905.04899, May 2019. doi: 10.48550/arXiv.1905.04899.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. *arXiv e-prints*, art. arXiv:1710.09412, October 2017. doi: 10.48550/arXiv.1710.09412.

N. Zhou, Y. Jiang, T. R. Bergquist, A. J. Lee, B. Z. Kacsoh, A. W. Crocker, K. A. Lewis, G. Georghiou, H. N. Nguyen, M. N. Hamid, L. Davis, T. Dogan, V. Atalay, A. S. Rifaioglu, A. ran, R. Cetin Atalay, C. Zhang, R. L. Hurto, P. L. Freddolino, Y. Zhang, P. Bhat, F. Supek, J. M. ndez, B. Gemovic, V. R. Perovic, R. S. ć, N. Sumonja, N. Veljkovic, E. Asgari, M. R. K. Mofrad, G. Profiti, C. Savojardo, P. L. Martelli, R. Casadio, F. Boecker, H. Schoof, I. Kahanda, N. Thurlby, A. C. McHardy, A. Renaux, R. Saidi, J. Gough, A. A. Freitas, M. Antczak, F. Fabris, M. N. Wass, J. Hou, J. Cheng, Z. Wang, A. E. Romero, A. Paccanaro, H. Yang, T. Goldberg, C. Zhao, L. Holm, P. nen, A. J. Medlar, E. Zosa, I. Borukhov, I. Novikov, A. Wilkins, O. Lichtarge, P. H. Chi, W. C. Tseng, M. Linial, P. W. Rose, C. Dessimoz, V. Vidulin, S. Dzeroski, I. Sillitoe, S. Das, J. G. Lees, D. T. Jones, C. Wan, D. Cozzetto, R. Fa, M. Torres, A. Warwick Vesztrocy, J. M. Rodriguez, M. L. Tress, M. Frasca, M. Notaro, G. Grossi, A. Petrini, M. Re, G. Valentini, M. Mesiti, D. B. Roche, J. Reeb, D. W. Ritchie, S. Aridhi, S. Z. Alborzi, M. D. Devignes, D. C. E. Koo, R. Bonneau, V. ć, M. Barot, H. Fang, S. Toppo, E. Lavezzo, M. Falda, M. Berselli, S. C. E. Tosatto, M. Carraro, D. Piovesan, H. Ur Rehman, Q. Mao, S. Zhang, S. Vucetic, G. S. Black, D. Jo, E. Suh, J. B. Dayton, D. J. Larsen, A. R. Omdahl, L. J. McGuffin, D. A. Brackenridge, P. C. Babbitt, J. M. Yunes, P. Fontana, F. Zhang, S. Zhu, R. You, Z. Zhang, S. Dai, S. Yao, W. Tian, R. Cao, C. Chandler, M. Amezola, D. Johnson, J. M. Chang, W. H. Liao, Y. W. Liu, S. Pascarelli, Y. Frank, R. Hoehndorf, M. Kulmanov, I. Boudellioua, G. Politano, S. Di Carlo, A. Benso, K. Hakala, F. Ginter, F. Mehryary, S. Kaewphan, J. rne, H. Moen, M. E. E. Tolvanen, T. Salakoski, D. Kihara, A. Jain, T. muc, A. Altenhoff, A. Ben-Hur, B. Rost, S. E. Brenner, C. A. Orengo, C. J. Jeffery, G. Bosco, D. A. Hogan, M. J. Martin, C. O'Donovan, S. D. Mooney, C. S. Greene, P. Radivojac, and I. Friedberg. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol*, 20(1):244, Nov 2019.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. *arXiv e-prints*, June 2023. doi: 10.48550/arXiv.2306.15006.

A RELATED WORKS

This work builds on top of foundational efforts spread across three main areas: contrastive representation learning, self-supervised applications in cellular property prediction, and methods for enriching genetic sequence input beyond one hot encoded representation. **Contrastive methods**: We build the IsoCLR approach for RNA sequences utilizing a rich body of work exploring contrastive learning for computer vision (Balestriero et al. (2023)). A fundamental deep metric learning approach is SimCLR in which the authors propose minimizing the representation distance between two views from the same sample while maximizing the distance between views from different samples (Chen et al. (2020)). This approach does not require labeled data and is based on the availability of domain-specific augmentations. Methods like BYOL and VicReg followed and were able to reformulate the contrastive approach by removing the need for in-batch negative samples (Grill et al. (2020); Bardes et al. (2021)). They propose solutions to the trivial solution collapse problem through a variance regularization loss term and architectural design choices. Recent work aims to unify these methods under the contrastive formulation by making a distinction between *sample* and *dimension* contrastive methods (Garrido et al. (2022)). **Self-supervised learning for cellular properties:** Due to the common sequence-based representation between genomics and language, self-supervised learning techniques have long been explored in genomic sequence property predictions. DNABert utilized the BERT problem formulation to learn an encoding for 500 nucleotide long sequences and demonstrated the value for splice site predictions and other tasks (Ji et al. (2021); Devlin et al. (2018); Zhou et al. (2023)). Nucleotide Transformer (NT), another masked language modeling method, demonstrated the utility of doing data collection from multiple species (Dalla-Torre et al. (2023)). RNA-FM was trained to predict non-coding RNA properties with masked language modeling using 23 million non-coding sequences (Chen et al. (2022)). Recently, HyenaDNA has demonstrated that applying long convolutions replacing the attention operation, can lead to effective DNA property prediction while scaling the input sequence length to a million tokens (Nguyen et al. (2023)). In the distinct protein representation learning space, there is a variety of protein language models utilizing auto-regressive and masked language modeling losses to predict protein properties like structure, variant effects, and functional properties (Meier et al. (2021); Lin et al. (2023)). Contrastive learning has also been used in more specialized domains such as enzyme property prediction while utilizing known shared enzyme properties as views of similar sequences (Yu et al. (2023)). Contrastive methods have also been used to learn a more general representation of protein function by maximizing the mutual information between global and local sequence representations (Lu et al. (2020)). We build on these works by exploiting domain-specific RNA augmentation to build general representations that are architecture-agnostic. **Beyond one hot encoded genomes:** Another important area for advancing cellular property prediction is iterating beyond the reference genome for representing genomic sequences. One such strategy is to integrate random biologically plausible augmentations during training (Lee et al. (2023)). By using domain-specific knowledge of the types of augmentations introduced during evolutionary processes, the authors demonstrate they can improve the performance of supervised models for predicting DNA properties. Using multiple sequence alignments is another way to use homology information, common in the protein modeling space (Do et al. (2005); Frazer et al. (2021); Jumper et al. (2021)). In another perspective, authors have argued that evolutionary homologs are a viable path for generating augmentations (Lu et al. (2020)).

B EXTENDED METHODS

Contrastive learning has been shown to be a bound on mutual information between two random variables X and Y corresponding to $I(X;Y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right]$. We utilize a variation of the classical InfoNCE loss, $\mathbb{E}\left[\log \frac{\exp(f(x_i,y_i))}{\Sigma \exp(f(x_i,y_j))}\right]$, where a model f is tasked with classifying the correct y_i which was jointly drawn with x_i (van den Oord et al. (2018)). Herein, the observations x_i, y_i correspond to splice isoforms or duplicate gene sequences which are interpreted as views of the same object while f is a neural network that we optimize to minimize the loss.

In the vision domain, contrastive learning strategies have had significant success by identifying augmentations that do not have a strong semantic effect, such as cropping, rotation, or Gaussian blur (Yun et al. (2019); Zhang et al. (2017); Chen et al. (2020)). In this work, we use RNA splicing isoforms and homologous genes as sources of functional invariance. By sampling RNA isoform sequences produced by alternative splicing, we identify sequence variation that is likey to maintain core functional properties. In addition, we use homology to pool RNA transcripts from evolutionarily related genes and generate sequence diversity (Pertea et al. (2018)). By minimizing the distance between functionally similar sequences, the model can learn regulatory regions critical for RNA property and function prediction. We pre-train a dilated convolutional residual model which has been demonstrated to be successful in applications for cellular property prediction by generalizing to long variable length sequences (Kelley et al. (2018); Linder et al. (2022); Chen et al. (2016); He et al. (2015)).

We use decoupled contrastive learning (DCL) as it has been shown to require smaller batch sizes, is less sensitive to hyperparameters such as learning rate, and the positive loss term can be weighted by sample difficulty (Yeh et al. (2021)). DCL iterates on the normalized temperature-scaled cross-entropy loss by splitting the contrastive objective into two terms: a similarity loss (positive) and a dissimilarity loss (negative) (Sohn (2016)). More formally, the positive and negative losses are calculated:

$$\mathcal{L}_{DCL,i}(\theta) = \log \sum_{z_k \in \mathcal{Z}, l \in 1,2}^{N} \mathbb{1}_{k \neq i} \exp(\langle z_i^1 \cdot z_k^l \rangle / \tau) - w_i \langle z_i^1, z_i^2 \rangle / \tau. \tag{2}$$

Sets of homologous genes with more transcripts are more informative than those with a single transcript thus, we weight the loss non-uniformly.

Unlike computer vision, the function q which can be used to generate views of the same object, is unknown. Thus, in genomics we have to use naturally observed views which may vary in number per gene and so the set of homologous genes will have a different number of splicing isoforms. Many non-protein coding genes will have only a single splicing isoform, resulting in the sampled two views being identical. To make the positive objective of identifying the augmented isoform non-trivial, we use dropout in our model and, randomly mask 15% of the transcript sequence Ji et al. (2021). This enforces the positive loss term for samples with a single RNA sequence to be non-zero. However, samples with multiple sequences generated through splicing and homology processes are more informative to the model. To reflect this imbalance between samples in our positive loss term, we introduce a sample evidence weighting term w_i to increase the importance of samples with a higher number of splicing isoforms:

$$w_i = log(t_i + c) \frac{T}{\Sigma_{k=1}^{N} log(t_k + c)}, \tag{3}$$

Where t is the number of transcripts per gene set, T corresponds to the total count of transcripts in the dataset, and c is a constant. The above objective increases the importance of samples with multiple RNA views while maintaining the overall norm of the total loss at the start of training. The weighting is applied only to the positive loss since the negative loss responsible for maximizing the distance between different samples is not affected by the number of transcripts per sample.

B.1 Downstream Evaluation Tasks

RNA half-life (RNA HL) is an important cellular property to measure due to its implications for protein expression regulation. Recently, it has been shown that the choice of method for measuring RNA half-life can have an outsize impact with no clear ground truth (Agarwal & Kelley (2022)). To

Table 3: Descriptive statistics for the contrastive learning dataset. As we utilize more species for dataset construction the number of sequences grows. Trans.;Transcripts.

#Species	#Genes	#Trans.	Mean #Trans.	%Genes with ≥ 2 Trans.
10	228,800	926,628	4.0	29%
2	65,600	286,390	4.36	41%
1	42,800	222,492	5.19	51%

address this problem, Agarwal and Kelley (2022) utilized the first principal component of over 40 different RNA half-life experiments. The dataset consists of 10,432 human and 11,008 mouse RNA sequences with corresponding measurements. The low data availability and high inter-experiment variation underscore the importance of data efficiency, and generalizability in computational models to be developed for this task.

Mean ribosome load (MRL) is a measure of the translational efficiency of a given mRNA molecule. It measures the number of ribosomes translating a single mRNA molecule at a point in time. Accurate MRL measurement is crucial as it offers insights into the efficiency of protein translation, a key process in cellular function. The dataset in question, derived from the HP5 workflow, captures this metric across 12,459 mRNA isoforms from 7,815 genes (Sugimoto & Ratcliffe (2022)). This dataset was derived from a single experiment, so we can expect a higher amount of noise associated than the RNA half-life dataset.

Protein localization Protein function is often linked to its subcellular location, which can be determined using cells that are immunofluorescently stained. We downloaded a dataset of 10,409 genes, whose protein localization was determined by the Human Protein Atlas (Thul et al. (2017)). We included the 12 most common locations including Nucleoplasm, Cytosol, Vesicles, Mitochondria, Plasma Membrane, Golgi apparatus and others. We utilized one transcript per gene (defined to be the canonical isoform by Rodriguez et al. (2022)) to obtain IsoCLR embeddings.

mRFP Expression We utilized 1,459 RNA sequences based on mRFP (monomeric Red Fluorescent Protein) with induced synonymous codon randomization. The data is from experiments conducted in Escherichia coli (E. coli) where protein production levels for various gene variants were quantified (Nieuwkoop et al. (2023)). We obtained IsoCLR embeddings for each sequence (all 678 bases long).

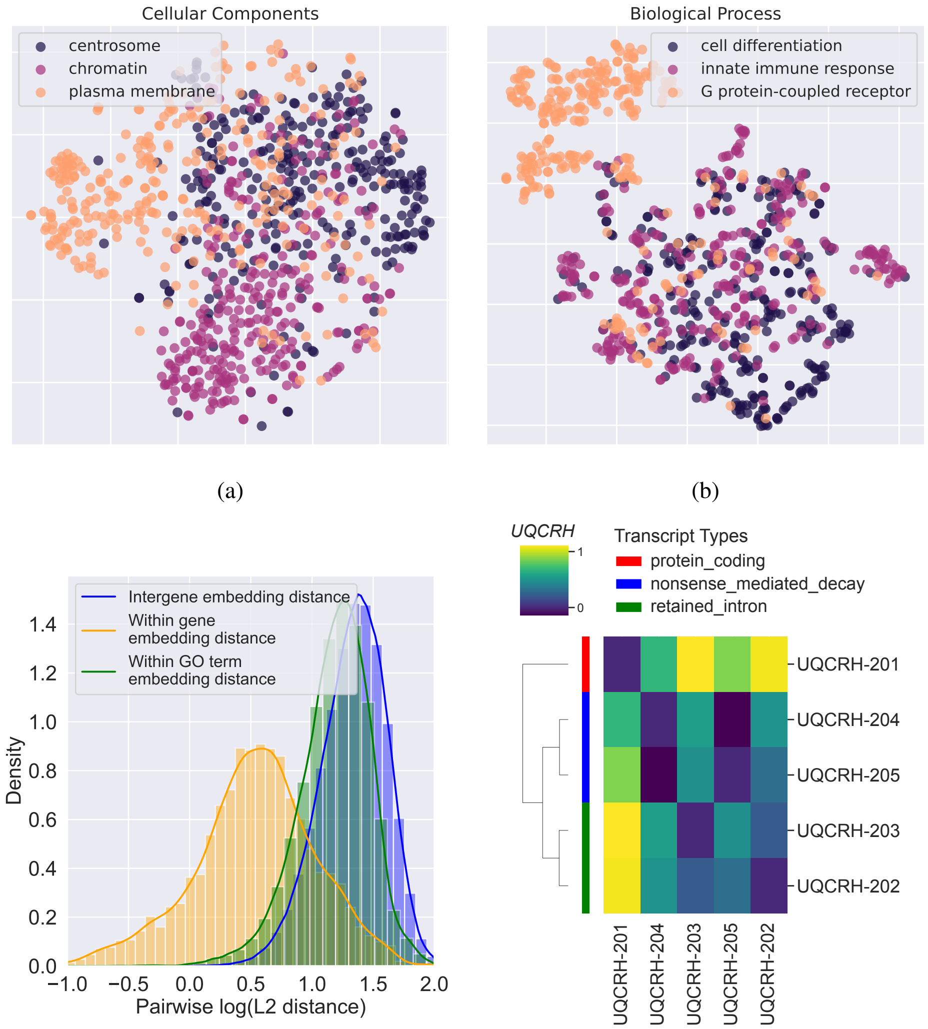
Gene ontology (GO) terms are a hierarchical classification system used for assigning function to genes and their products (Consortium et al. (2023); Ashburner et al. (2000); Zhou et al. (2019)). In this work, we utilize GO classes to visualize model latent embeddings and classification. GO term hierarchical systems allow for fine-grained annotation of function, with broader terms at the top of the hierarchy and increased specificity closer to the bottom. To annotate genes with gene ontology terms, we subset GO classes three levels from the root labeling all available genes.

C Extended Results

C.1 Latent Space Analysis

We evaluate whether IsoCLR's pre-trained representations capture fundamental biological information. First, we examine IsoCLR's ability to capture gene-ontology terms associated with cellular components, and biological processes (Figure 3a, b). We generate the representation with the encoder f and reduce the dimensionality of the embedding with t-sne (van der Maaten & Hinton (2008)). To quantitatively verify the latent structure, we perform linear probing over ten GO classes and find they are linearly separable in IsoCLR's latent space. (Table **??**, Appendix D.3).

In addition, we examine the embedding distances learned by IsoCLR across three different settings (Figure 3c). We measure the distances of splice isoforms within genes, across genes - by taking the principal isoform per gene, and within GO classes by measuring the distances between principal isoforms of genes from the same GO term. Consistent with our training objective, we find that 'within gene' distances are significantly smaller compared to inter-gene distances. In addition when sampling genes with the same GO class we observe statistically significant difference in distances compared to randomly sampled genes (p=2.2e-16, two sided t-test). In the literature, there are well annotated examples where transcripts belonging to the same gene have drastically different function. We find evidence of that reflected in IsoCLR representations by identifying that more than 4% of 'within gene' transcript pairs have a distance greater than that of two randomly chosen genes, indicating that the IsoCLR training objective preserves within gene sequence diversity. 'Within gene' diversity could potentially help delineate differential isoform protein functions, a very active area of research. We examine more closely one such gene corresponding to *UQCRH*, a gene whose protein product is localized in the mitochondrion and is involved in the electron transport chain (Vidali et al. (2021)). Out of its five annotated splice isoforms in our dataset, two are subjected to nonsense mediated decay and two possess a retained intron while *UQRCH-201* is the principal protein-coding



(a)

(b)

(c)

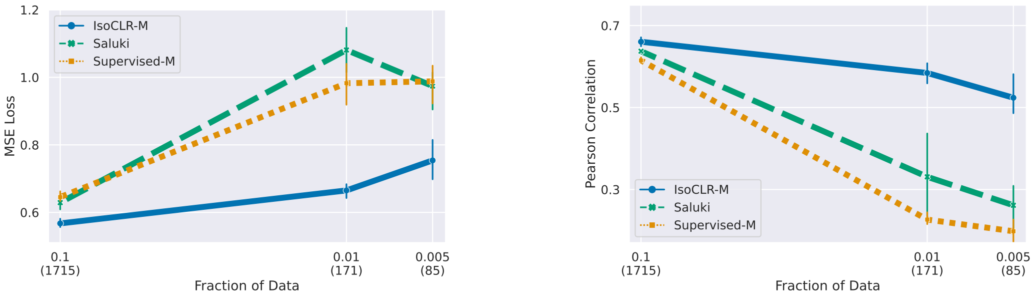
(d)

Figure 3: **(a, b)** Visualization of the learned latent representations with stochastic neighbor embedding. Each dot is an RNA transcript from a unique gene colored by the correspondingly annotated gene ontology. **(c)** Distributions of log(L2) distances are shown between IsoCLR embeddings in three settings. Intergene distances represent measurements between random genes (one transcript per gene), while within-gene embeddings show distances among transcripts within the same gene, sampled from 500 genes. Within GO term distances are calculated for genes grouped by the same gene set. **(d)** An example of a within gene embedding distances heatmap showing pairwise similarity across transcripts. Values indicate log(L2) distances.

isoform (Figure 3d). IsoCLR learns to linearly separate the protein coding isoform from the four others without any supervised labels. We also observe that NMD and intron retention isoforms cluster together suggesting further compartmentalization of embeddings driven by underlying splicing outcomes.

C.2 Ablations: splicing augmentations are key

Finally, we investigate the IsoCLR augmentations that contribute towards effective performance. We find that sampling splicing isoforms from the same gene is the primary driver of performance on downstream tasks(Table 4). Homology, and masking a small percentage of the input sequence provide small additional gains. Homologous gene mapping can be interpreted as removing wrong



(a)

(b)

Figure 4: Data sub-sampling analysis demonstrating IsoCLR's strong performance in the low data regime. Confidence intervals were computed using standard deviation over three random seeds.

negatives from our training set, where duplicated genes will have highly similar sequences. Masking the input space can also be thought of as a regularization method which has recently been shown as vital for learning effective representations with contrastive learning (Ben-Shaul et al. (2023)).

Table 4: Ablation analysis demonstrates that splicing is the key contributor to representation effectiveness. Linear probing results are generated by performing gradient descent to optimize a linear layer. 6t; Six track input corresponding to one hot encoded sequence, splicing and codon positions.

Augmentations	Half-life Human MSE	Half-life Mouse MSE	GO ROC AUC
Splice + Homology + Mask + 6t	0.57	0.62	0.85
Splice + Homology + Mask	0.73	0.74	0.84
Splice + Homology + 6t	0.58	0.70	0.84
Splice + Mask + 6t	0.58	0.65	0.85
Mask + 6t	0.88	0.86	0.77

D Data Efficiency

To simulate downstream tasks for which there is a lack of experimental data, we perform fine-tuning on RNA HL prediction where only a subset of the original training data set is available. We observe that supervised methods are ineffective in this regime, while IsoCLR maintains competitive performance at 10% and 1% of the data (Figure **??**). The performance differences are even more stark when using only 0.5% of the training data (Pearson R; IsoCLR = 0.50 versus Saluki = 0.26). These findings illustrate that IsoCLR advances towards the aim of few-shot learning for downstream tasks where data is too limited for traditional supervised learning approaches.

D.1 Linear Probe Experimental Details

In this section, we describe the experimental procedure to evaluate linear probing results.

We first performed a 70-15-15 data split on datasets. The data sequences are then embedded by the various self-supervised learning (SSL) models. For IsoCLR, we simply take the mean of the embeddings across the seqeunce dimension. For HyenaDNA, we take the mean and max of the embedding sequence dimension, as well as the last hidden state in the output sequence. Other SSL methods could not handle input sequences of more than 500 or 1000 nucleotides. Thus, when input sequences exceeded the allowable context window, each sequence was chunked to the maximum length allowed by a model. We then computed the mean of each chunk embedding across the sequence dimension, and then averaged the mean embedding of each chunk to obtain the final embedding.

After obtaining embedding vectors, we used the scikit-learn implementation of linear models to perform the linear probes of the embeddings. For the downstream regression tasks, we used either used linear regression or ridge regression with the regularization parameter selected by cross validation. The final linear model was selected using the validation split. The gene ontology and protein localization tasks are multi-label classification tasks. For this, we fit scikit-learn's LogisticRegression model to the labels using a MultiOutputClassifer, which essentially trains a separate linear classifier for each label class. We use the default logistic regression parameters, and set 5000 maximum iterations for the solver.

For the classification tasks we also calculate AU-PRC results but due to space constraints could not include them in the main text of the paper.

Model	GO AUROC	GO AUPRC	Protein Loc AUROC	Protein Loc AUPRC
IsoCLR-S (ours)	0.86	0.57	**0.85**	**0.41**
IsoCLR-M (ours)	0.84	0.53	0.83	0.38
IsoCLR-L (ours)	**0.87**	**0.59**	0.84	0.40
DNA-BERT2	0.72	0.30	0.77	0.26
NT-500m-1000g	0.67	0.25	0.70	0.20
NT-500m-human-ref	0.72	0.35	0.73	0.24
NT-2.5b-1000g	0.73	0.34	0.70	0.22
NT-2.5b-multi-species	0.78	0.42	0.73	0.26
Hyena-32K-seqlen	0.75	0.36	0.79	0.29
Hyena-160K-seqlen	0.75	0.33	0.79	0.29
Hyena-450K-seqlen	0.75	0.35	0.78	0.29
RNA-FM	0.78	0.40	0.81	0.32

Table 5: Area under the Precision Recall curve metrics computed for imbalanced datasets.

D.2 Fine-tuning Experimental Details

We fine-tune IsoCLR by first initializing most of the model with weights from pre-training, the penultimate two layers with random initialization, and the final layer with zero init. We don't apply any weight decay to weights that were initialized from pre-training while the final three layers have an l2 weight decay term of 1e-5. We fine-tune on downstream tasks using the Adam optimizer with a learning rate of 0.01. We apply exponential learning rate decay with a factor of 0.95. The models are trained with a single Nvidia T4 GPU in a mixed precision setting.

HyenaDNA models initialized with fine-tuning head. We perform a small learning rate hyperparameter grid search around the suggested hyperparameters of 6e-4. The suggested AdamW optimizer is used. Models were trained for a maximum of 100 epochs on Nvidia T4 GPUs with a batch size of 28 for the HyenaDNA-tiny and a batch size of 8 for HyenaDNA-small. Models were stopped early based on validation loss using an epoch patience of three. After selecting learning rate using the validation split, the runs were repeated using different random initializations to generate confidence intervals.

D.3 Additional comparisons for IsoCLR Representations

To test IsoCLR's latent space semantic interpretability, we take a single transcript from every single human gene and annotate it with gene ontology terms from the three main hierarchies: biological processes, cellular components, and molecular function. We subset the gene ontology terms to third from the root to identify a broad and yet high-level number of functions. From the subset, we select the three most common gene ontology terms to use for the visualization. We also compare the representations to two baselines: a stochastic neighborhood embedding with randomly initialized labels from IsoCLR representation, and a supervised model with a matched architecture trained to predict RNA half-life (van der Maaten & Hinton (2008), **??**). We observe that the supervised embedding also produces a distinct cluster for the biological process go hierarchy confirming the unique structure of g protein-coupled receptors. Upon visual inspection, however, the clusters from the supervised model are less separated compared to the IsoCLR representation. For example, we observe a distinct cluster for sequences associated with the centrosome function in the IsoCLR representation, whereas, for the supervised model, the samples are interspersed throughout the representation **??**.

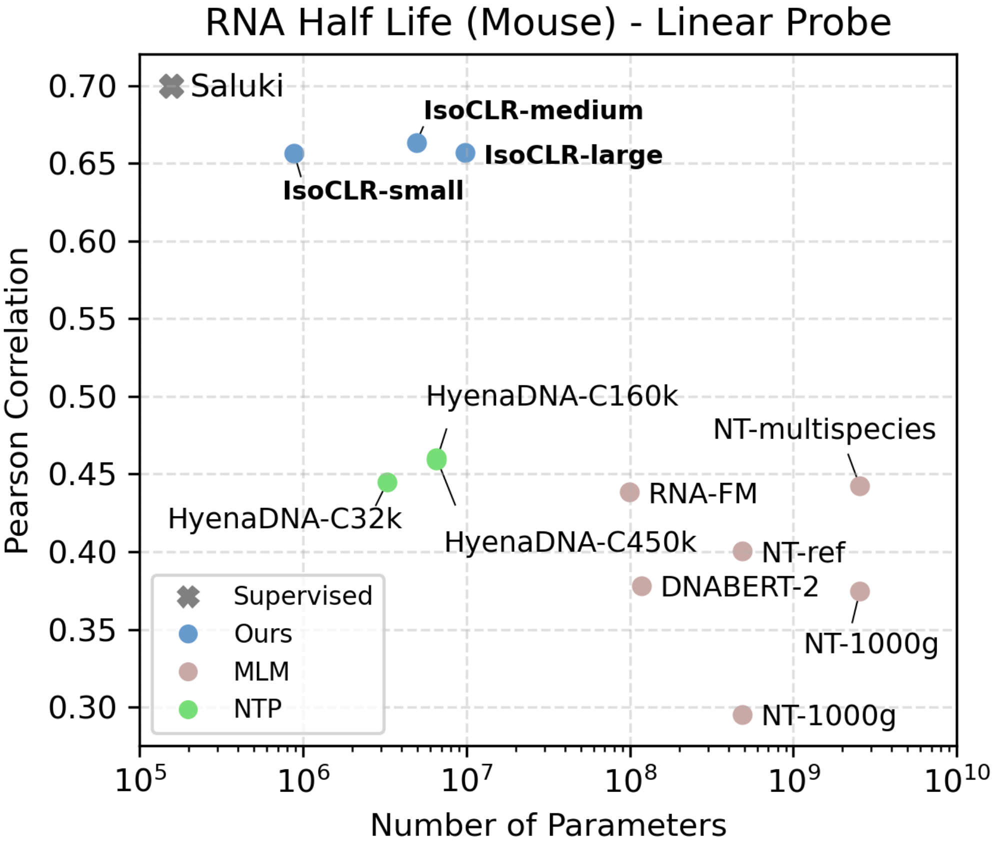


Figure 5: Comparing linear probing performance for self-supervised methods on RNA half-life mouse data.

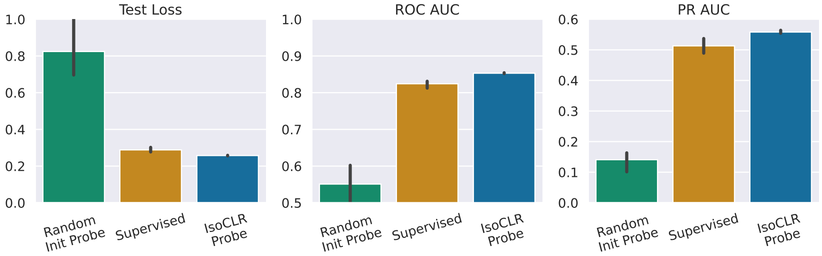


Figure 6: Gene ontology multi-label classification comparison with the supervised model. This is a ten-class multi-label classification task for the molecular function gene ontology category.

To quantitatively validate IsoCLR latent space with gene ontology, we perform linear probing over a 10 class multi-label classification task. Each class corresponds to a gene ontology term, and the samples are RNA sequences with corresponding GO labels. We find that performing linear probing on IsoCLR embeddings exceeds performance of supervised models trained with full fine-tuning (Figure 6.