ema-tool: a Python Library for the Comparative Analysis of Embeddings from Biomedical Foundation Models

Pia Francesca Rissom[1,2]✉, Paulo Yanez Sarmiento[1,2], Jordan Safer[2], Connor W. Coley[3,4], Bernhard Y. Renard[1,5], Henrike O. Heyne[1,5,6,7,*]✉, and Sumaiya Iqbal[2,7,8,9,*]✉

[1]Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, Potsdam, Germany
[2]Broad Institute of MIT and Harvard, Center for the Development of Therapeutics, Cambridge, USA
[3]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, USA
[4]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, USA
[5]Hasso Plattner Institute, Mount Sinai School of Medicine, New York, USA
[6]Finnish Institute for Molecular Medicine (FIMM), University of Helsinki, Helsinki, Finland
[7]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, USA
[8]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, USA
[9]Cancer Data Sciences, Dana-Farber/Harvard Cancer Center, Boston, USA
[*]Co-supervision

The increasing use of foundation models in biomedical applications raises opportunities and challenges to analyze the information captured in the high-dimensional embedding spaces of different models. Existing tools offer limited capabilities for comparing information represented in the embedding spaces of different models. We introduce *ema-tool*, a Python library designed to analyze and compare embeddings from different models for a set of samples, focusing on the representation of groups known to share similarities. *ema-tool* examines pairwise distances to uncover local and global patterns and tracks the representations and relationships of these groups across different embedding spaces. We demonstrate the use of *ema-tool* through two examples. In the first example, we analyze the representation of ion channel proteins across versions of the ESM protein language models. In the second example, we analyze the representation of genetic variants within the *HCN1* gene across these models. The source code is available at https://github.com/broadinstitute/ema.

embeddings | protein language models | interpretability

Correspondence: *francesca.rissom@hpi.de, henrike.heyne@hpi.de, sumaiya@broadinstitute.org*

Introduction

The emergence of foundation models has revolutionized the field of natural language processing (NLP) (1–4), and foundation models are gaining popularity in biomedical applications (5–10). Trained on extensive datasets in a self-supervised manner, these models learn to represent the data points and their relationships in a numerical vector space. By leveraging these learned embedding spaces, the foundation models achieve competitive performance on several downstream tasks (11–13). With an increasing number of models, the comparison of what has been learned becomes more important, for example, pre- and post-fine-tuning, models with different numbers of parameters, or models trained on different data modalities. Whilst information learned by the different models is commonly evaluated by comparing its performance on downstream tasks (6, 13, 14), there has been little emphasis on directly comparing learned information, such as embedding spaces, across biomedical foundation models (6).

In contrast, the analysis of embedding spaces is more established in fields like NLP, leading to the development of tools dedicated to simplifying the comparative analysis of embedding spaces. The published tools for the comparison of embedding spaces primarily examine nearest-neighbor overlap using Euclidean or Cosine distances. embComp (15) and recomp (16) aggregate changes in nearest neighbors, offering a global overview of similarity between two embedding spaces. Other tools allow for the visualization of the nearest neighbors of a query data point in both embedding spaces using dimensionality reduction techniques such as Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), or Uniform Manifold Approximation and Projection (UMAP) (17, 18). Further, the Emblaze tool (19) visualizes how a query data point's nearest neighbors shift in a different embedding space, showing where these new neighbors are situated in the current embedding space. Emblaze also highlights data points with substantial changes in their neighbors and neighborhoods that exhibit relatively large changes between embedding spaces.

Whilst the presented tools (15–19) can also be applied to embeddings of biomedical data, we argue that the comparison of embedding spaces from biomedical foundation models can further be deepened by augmenting the analysis with prior knowledge of natural groupings of the embedded data points. The clustering analysis of natural groupings is already prominent in single embedding space evaluation (6, 20); our goal is to enhance accessibility and integrate this information into a comparative analysis.

To bridge this gap and further guide our understanding of the information captured by foundation models, we introduce *ema-tool*, a Python library, designed for a straightforward initial comparison of diverse embedding spaces in biomedical data. By enabling users to include various user-defined metadata on the natural grouping of data, users can not only compare global statistics of multiple embedding spaces but also understand the differences in the clustering of natural group-

ings across different embedding spaces. This approach allows for insights into the proximity of individual data points as well as the relationships and clustering of entire groups, e.g., genes or proteins, across embedding spaces. We provide two example applications of *ema-tool* for analyzing the representation of gene families and gene variants across multiple large Protein Language Models (PLMs).

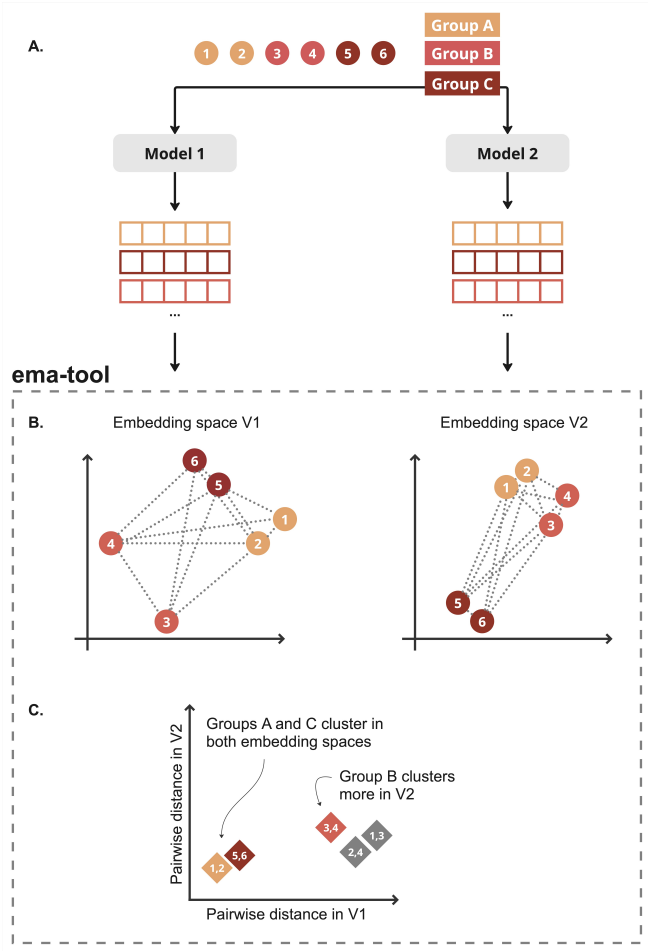

Figure 1: Overview of the *ema-tool* workflow. **A** Six samples are embedded into two high-dimensional spaces, V_1 and V_2. The two embedding spaces are passed to the *ema-tool* library along with metadata, grouping the samples into groups A, B, and C. **B** In addition to visualization options of the clustering of groups of samples in each embedding space, *ema-tool* calculates the pairwise distances between all sample pairs within each embedding space. **C** Visualization tools in *ema-tool* enable users to explore changes in pairwise distances between the two embedding spaces, helping to identify common patterns and groups that show differences in their relative position in each space.

ema-tool: Methods and Features

Figure 1 provides an overview of the *ema-tool* workflow. Given a set of n samples $(s_1, ..., s_n)$ and at least two embedding spaces $V_i \subset \mathbb{R}^{d_i}$, i = 1, 2 (i.e., two functions that map the sample s to an embedding space: $E_i : \mathcal{S} \to V_i : s \mapsto v^{(i)}$), *ema-tool* provides overview statistics and visualizations of the distribution of samples in each embedding space. Further, *ema-tool* computes the pairwise distances between the embeddings of all samples. The embedding spaces can then be compared based on the differences in the pairwise distances, both on an individual data point level and by comparing user-defined groups of sample points.

Input Data. *ema-tool* takes the following input objects:

- *Sample information*: A pandas DataFrame where the first column contains unique identifiers for each sample. Additional columns can include user-defined categorical and continuous metadata about the samples such as group labels.

- *Embedding spaces*: NumPy arrays with rows containing the embeddings of each sample in the respective embedding space. In our application examples, each row contains the embedding of a specific protein sequence. Multiple embedding spaces can be loaded into the library object for comparison. At least two embedding spaces are required for the full analysis. The dimensions of the embedding spaces of interest can be of different dimensions.

Analyses Within an Embedding Space. To gain deeper insights into the properties of each embedding space, *ema-tool* offers several analytical tools. Within the same embedding space V_i, *ema-tool* lets the user analyze the distribution of the embedding values to gain an initial understanding of the scale and variance of these values in each embedding space. The clustering of samples in relation to the metadata can be analyzed by comparing the correlation of user-defined groups with clusters identified by unsupervised clustering approaches. Furthermore, dimensionality reduction methods, including PCA, UMAP, and t-SNE, can be applied to visually inspect the variance and clustering within the embedding space. At each stage, the analysis can be stratified by the user-defined groups of samples.

Measuring Pairwise Distances. For each pair of samples s_j and s_k, the distance of their respective embeddings $dist\left(v_j^{(i)}, v_k^{(i)}\right)$ is measured in each embedding space V_i. The use of conventional distance metrics to measure meaningful distances in high dimensional spaces has been debated (21, 22). Thus, *ema-tool* offers a range of different distance metrics for a comprehensive analysis of the embedding space, including Manhattan, Euclidean, and Cosine distances as well as normalized and scaled versions of these. The range of distance metrics provides different viewpoints on the proximity of data points in the embedding space. In a range of visualizations, users can explore the distances of groups of samples to each other.

Analyses Between Embedding Spaces. To understand the similarities and differences in the relationship of samples across embedding spaces $dist\left(v_j^{(1)}, v_k^{(1)}\right)$ vs. $dist\left(v_j^{(2)}, v_k^{(2)}\right)$, *ema-tool* provides visualization tools to compare sample pairs between two embedding spaces. The introduced distance metrics can be used. The provided metadata can be leveraged to identify

differences between the relationships of groups of samples among the embedding spaces.

Application Examples

Embedding of Protein Families Across Protein Language Models. In this example, we use *ema-tool* to analyze the embeddings of 102 wild-type protein sequences across the PLMs ESM-1b (6), ESM-1v (11) and ESM-2 models (12). For each protein, we retrieve one embedding of the dimension 1 x 1280 for each of the three ESM models. The analyzed proteins belong to eight different ion channel families, which are used as metadata to group the proteins. The families are voltage-gated potassium channels (Kv), calcium-activated potassium channels (KCa), inwardly rectifying potassium channels (Kir), two-pore potassium channels (K2P), cyclic nucleotide-gated channels (CNG), hyperpolarization and cyclic-nucleotide gated channels (HCN), transient receptor potential channels (TRP), and transient receptor potential related channels (TRPML). We find that pairwise distances between proteins are correlated between ESM-1b and ESM-2 (Spearman's $\rho = 0.41$, $p < 0.0001$) and even more pronounced between ESM-1b and ESM-1v (Spearman's $\rho = 0.73$, $p < 0.0001$). When visually comparing the pairwise distances between protein families, we observe that all three models seem to embed proteins related to the Kir family (15 proteins) closer to each other than to proteins from other families (see Figure 2A I). The embeddings of other protein families are more heterogeneous between the models. For example, by visual inspection proteins of the CNG and HCN families seem more separated in the ESM-1b space compared to ESM-2 (see Figure 2A II). Our experiments demonstrate the varying representation of protein families for ion channel proteins in the distances of embeddings across three PLMs.

Embedding of Genetic Variants Across Protein Language Models. In a second example, we use *ema-tool* to analyze the embeddings of the ESM-1b and ESM-2 model for 880 genetic missense variants (i.e., single amino acid changes in the encoded protein sequence) in the *HCN1* ion channel gene. These variants are collected from two sources. 861 variants are collected from the Genome Aggregation Database (gnomAD) (23), the largest publicly available database for genetic variants that are observed in the general population and are therefore expected to not increase the risk for severe diseases with large effect sizes. Additionally, 19 variants are obtained from the ClinVar database (24), a large public community data source for variants contributed from clinical diagnostic settings. The selected ClinVar variants are expected to be disease-causing with high confidence. Embeddings are obtained by incorporating the specific amino acid alteration into the wild-type protein sequence corresponding to the *HCN1* gene and retrieving an embedding for this modified sequence. Each embedding is of dimension 1 x 1280.

By comparing the embedding representations of all 880 *HCN1* missense variants to a set of protein features collected from UniProt (25) and the Genomics 2 Proteins Portal (26), we find that embeddings of the subset of variants that are located in disordered regions of the protein (397 variants) are clustered in the embedding space of ESM-1b, but not that of ESM-2, when visually inspecting the first two principle components of each embedding space (see Figure 2 B). Interestingly, 100% of the 397 variants located in the disordered regions are putatively benign population variants from the gnomAD database. This association between gnomAD missense variant positions in proteins and disordered regions in proteins is in agreement with previous studies (27–29).

We further leveraged *ema-tool* to compute the pairwise Euclidean distances between embeddings of variants within disordered regions of the protein and between variants inside and outside the disordered regions of the protein. In the embedding space of the ESM-1b model, the distances between embeddings of variants within disordered regions, $n_1 = 78{,}606$, are notably lower compared to the distances between embeddings of variants inside and outside the disordered regions, $n_2 = 191{,}751$ ($\text{median}_1 = 0.056$, $\text{median}_2 = 0.067$, respectively, $p < 0.0001$ in Mann-Whitney U test, one-sided). In contrast, we did not observe a statistically significant difference between ESM-2 embeddings of variants within the disordered regions and those of variants inside and outside of the disordered regions ($\text{median}_1 = 0.045$, $\text{median}_2 = 0.044$, respectively, $p = 1.0$ in Mann-Whitney U test, one-sided). By clustering the variants in the disordered regions of the *HCN1* gene, the ESM-1b model captures more information in the distances between embeddings about the disordered region feature, compared to the ESM-2 model. This result indicates that protein embeddings from the ESM-1b model may be more suitable as input for variant pathogenicity prediction for the *HCN1* gene compared to those from ESM-2.

Conclusion

In summary, we introduce *ema-tool*, a Python library for comparing embeddings in different latent spaces of foundation models. With an emphasis on incorporating available metadata, *ema-tool* aims to foster the exploration and comparison of embeddings to gain insights into the learned representations and ultimately guide the downstream use of the models in biomedical applications. While the current implementation of *ema-tool* supports the analysis of small sets of samples, further efforts are presently being made to streamline distance computations for larger sets of samples. Our future directions also include the development of additional quantitative analysis methods to deepen the insights provided by embedding space comparisons.

ACKNOWLEDGEMENTS

We would like to express our gratitude to Katharina Baum, Jakub Bartoszewicz, Melania Nowicka, Marta Lemanczyk, Eugenia Alleva, and Hilary Finucane for their valuable discussions and insightful input regarding this project. The project was supported by the Designing for Sustainability research program, a joint initiative of the Morningside Academy for Design at MIT and the Hasso Plattner Institute, generously funded by the Hasso Plattner Foundation.

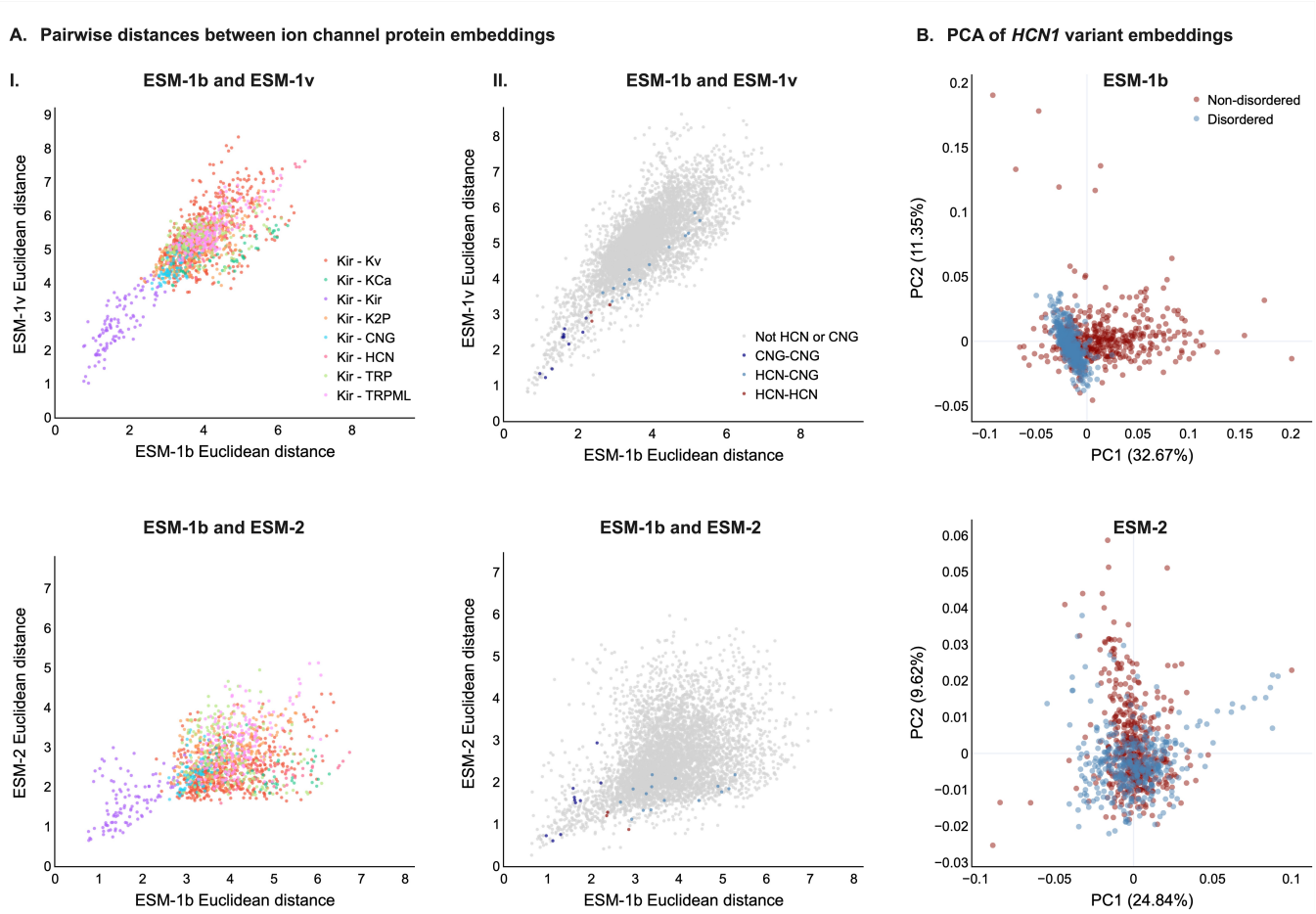

Figure 2: **A. Application example 1: Visualising ion channel families in different embedding spaces.** Each dot represents the Euclidean distance between the embeddings of two of the selected ion channel proteins, as indicated on the x and y axes. The color of the dots indicates the respective family of the ion channels. The upper panels show the distances of the embeddings in the ESM-1b and ESM-1v models. The lower panels show the distances between the embeddings of the ESM-1b and ESM-2 models. **I. Distances of proteins of the Kir family**. Embeddings of proteins of the Kir family are closer to each other than to proteins from other families. This trend is observed in all three embedding spaces. **II. Distances of proteins of the HCN and CNG family**. In the embeddings from the ESM-1b and ESM-1v models, distances between embeddings of proteins from the HCN family are closer to each other than they are to embeddings from proteins of the CNG family, and vice versa. This trend is not observed in the embeddings from the ESM-2 model, where proteins from the HCN family are embedded at similar distances to proteins from their family than to proteins from the CNG family. The two families are clustered more distinctly in the ESM-1b and ESM-1v models, compared to the ESM-2. **B. Application example 2: Scatter plot of the first two principal components of the *HCN1* variant embeddings**. Each dot shows the first (x-axis) and second (y-axis) principle components representing embeddings of an amino acid sequence with a missense mutation. The upper plot shows the first two principal components representing the ESM-1b embeddings, the lower plot the first two principal components representing the ESM-2 embeddings. The color of the dots indicates whether the mutation occurs within a disordered region of the protein ("Disordered", blue) or outside a disordered region ("Non-disordered", red). The percentage of variance explained by each principal component is shown in brackets on the respective axis label. Mutations in disordered regions cluster more closely in the first principal component of the ESM-1b embeddings compared to the ESM-2 embeddings.

Bibliography

1. Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56 (2):1–40, 2023.
2. Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
4. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.
5. Giorgio Valentini, Dario Malchiodi, Jessica Gliozzo, Marco Mesiti, Mauricio Soto-Gomez, Alberto Cabri, Justin Reese, Elena Casiraghi, and Peter N Robinson. The promises of large language models for protein design and modeling. *Frontiers in Bioinformatics*, 3, 2023.
6. Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.
7. Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.
8. Micaela E Consens, Cameron Dufault, Michael Wainberg, Duncan Forster, Mehran Karimzadeh, Hani Goodarzi, Fabian J Theis, Alan Moses, and Bo Wang. To transformers and beyond: Large language models for the genome. *arXiv preprint arXiv:2311.07621*, 2023.
9. Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.
10. Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36, 2024.
11. Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.
12. Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.
13. Eric Nguyen, Michael Poli, Matthew G. Durrant, Armin W. Thomas, Brian Kang, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aman Patel, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. Sequence modeling and design from molecular to genome scale with evo. *bioRxiv*, 2024. doi:

10.1101/2024.02.27.582234.

14. Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 02 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083.
15. Florian Heimerl, Christoph Kralj, Torsten Möller, and Michael Gleicher. embcomp: Visual interactive comparison of vector embeddings. *IEEE Transactions on Visualization and Computer Graphics*, 28(8):2953–2969, 2022. doi: 10.1109/TVCG.2020.3045918.
16. Dan Shiebler. repcomp. *https://github.com/dshiebler/RepresentationComparison*, 2018.
17. Angie Boggust, Brandon Carter, and Arvind Satyanarayan. Embedding Comparator: Visualizing Differences in Global Structure and Local Neighborhoods via Small Multiples. page 746–766, 2022. doi: 10.1145/3490099.3511122.
18. Agence Data Services PE Nantes. Embcompare. *https://github.com/OSS-Pole-Emploi/embcompare*, 2022.
19. Venkatesh Sivaraman, Yiwei Wu, and Adam Perer. Emblaze: Illuminating machine learning representations through interactive comparison of embedding spaces. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, page 418–432, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391443. doi: 10.1145/3490099.3511137.
20. Michelle M. Li, Yepeng Huang, Marissa Sumathipala, Man Qing Liang, Alberto Valdeolivas, Ashwin N. Ananthakrishnan, Katherine Liao, Daniel Marbach, and Marinka Zitnik. Contextualizing protein representations using deep learning on protein networks and single-cell data. *bioRxiv*, 2024. doi: 10.1101/2023.07.18.549602.
21. Evgeny M. Mirkes, Jeza Allohibi, and Alexander Gorban. Fractional norms and quasinorms do not help to overcome the curse of dimensionality. *Entropy*, 22(10), 2020. ISSN 1099-4300. doi: 10.3390/e22101105.
22. Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In Jan Van den Bussche and Victor Vianu, editors, *Database Theory — ICDT 2001*, pages 420–434, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-44503-6.
23. 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, Konrad J. Karczewski, and Genome Aggregation Database Consortium. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*, 625(7993):92–100, 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06045-0.
24. Melissa J. Landrum, Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42:D980 – D985, 2013.
25. The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 11 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac1052.
26. Seulki Kwon, Jordan Safer, Duyen T. Nguyen, David Hoksza, Patrick May, Jeremy A. Arbesfeld, Alan F. Rubin, Arthur J Campbell, Alex Burgin, and Sumaiya Iqbal. Genomics 2 proteins portal: A resource and discovery tool for linking genetic screening outputs to protein sequences and structures. *bioRxiv*, 2024. doi: 10.1101/2024.01.02.573913.
27. Sumaiya Iqbal, Eduardo Pérez-Palma, Jakob B. Jespersen, Patrick May, David Hoksza, Henrike O. Heyne, Shehab S. Ahmed, Zaara T. Rifat, M. Sohel Rahman, Kasper Lage, Aarno Palotie, Jeffrey R. Cottrell, Florence F. Wagner, Mark J. Daly, Arthur J. Campbell, and Dennis Lal. Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. *Proceedings of the National Academy of Sciences*, 117(45):28201–28211, 2020. doi: 10.1073/pnas.2002660117.
28. Shehab S. Ahmed, Zaara T. Rifat, Ruchi Lohia, Arthur J. Campbell, A. Keith Dunker, M. Sohel Rahman, and Sumaiya Iqbal. Characterization of intrinsically disordered regions in proteins informed by human genetic diversity. *PLOS Computational Biology*, 18(3):1–28, 03 2022. doi: 10.1371/journal.pcbi.1009911.
29. Sumaiya Iqbal, Tobias Brünger, Eduardo Pérez-Palma, Marie Macnee, Andreas Brunklaus, Mark J Daly, Arthur J Campbell, David Hoksza, Patrick May, and Dennis Lal. Delineation of functionally essential protein regions for 242 neurodevelopmental genes. *Brain*, 146(2):519–533, 10 2022. ISSN 0006-8950. doi: 10.1093/brain/awac381.