Cell Systems

CellPress
OPEN ACCESS

Perspective

On knowing a gene: A distributional hypothesis of gene function

Jason J. Kwon,[1,2,8] Joshua Pan,[1,2,7,8] Guadalupe Gonzalez,[3] William C. Hahn,[1,2,*] and Marinka Zitnik[2,4,5,6,*]
[1]Dana-Farber Cancer Institute and Harvard Medical School, Department of Medical Oncology, Boston, MA 02215, USA
[2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[3]Department of Computing, Faculty of Engineering, Imperial College, London SW7 2AZ, UK
[4]Harvard Medical School, Department of Biomedical Informatics, Boston, MA 02115, USA
[5]Harvard Data Science Initiative, Harvard University, Cambridge, MA 02138, USA
[6]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Allston, MA 02134, USA
[7]Present address: DeepMind, London, UK
[8]These authors contributed equally
*Correspondence: william_hahn@dfci.harvard.edu (W.C.H.), marinka@hms.harvard.edu (M.Z.)
https://doi.org/10.1016/j.cels.2024.04.008

SUMMARY

As words can have multiple meanings that depend on sentence context, genes can have various functions that depend on the surrounding biological system. This pleiotropic nature of gene function is limited by ontologies, which annotate gene functions without considering biological contexts. We contend that the gene function problem in genetics may be informed by recent technological leaps in natural language processing, in which representations of word semantics can be automatically learned from diverse language contexts. In contrast to efforts to model semantics as "is-a" relationships in the 1990s, modern distributional semantics represents words as vectors in a learned semantic space and fuels current advances in transformer-based models such as large language models and generative pre-trained transformers. A similar shift in thinking of gene functions as distributions over cellular contexts may enable a similar breakthrough in data-driven learning from large biological datasets to inform gene function.

INTRODUCTION

An important goal of molecular biology is to uncover the function of genes. "Functionalizing" a gene sequence involves identifying a biological state in which the gene or protein is expressed and experimentally assessing its impact on that state.[1] These discoveries are typically represented as new edges in biological knowledge graphs, such that a gene's function is described in relationship to known complexes or biological pathways ("gene A is a member of pathway B"). However, this approach to understanding genes presents several areas that could benefit from improvement. First, gene sequence discovery far outpaces the rate of deciphering gene function, leaving many genes uncharacterized.[2] The development of computational methods can bridge the gap from discovery to understanding more efficiently. Second, new connections are often biased toward well-understood and experimentally tractable complexes and pathways,[3] a systematic bias known as the "streetlight effect."[4] Third, and most critically, gene function depends on the state of the cell, meaning that a gene can be involved in different biological functions depending on the cell in which it is located.[4,5]

Cell states are described by their molecular and biochemical characteristics that can change over time as a cell responds to various stimuli. Although the methods and process of defining cell state are still an active area of discussion,[6–8] it is well understood that biological "context" is important for gene function. For example, the function of a protein enzyme is incomplete without understanding in which context or cell state the enzyme is operational, such as the presence of a cofactor, cellular microenvironments, and organism identity. Furthermore, the context of a gene's function can be extended to additional emergent levels of biology, ranging from cells to tissues to organs comprising multiple cellular types. In our framework, context encompasses a broad spectrum of environmental, temporal, and spatial influences that modulate gene function. Although intrinsic features of a gene, such as promoter regions or gene regulatory elements, are inherent, the context in which they operate can differ markedly. It is this variability in contextual settings that we seek to elucidate. By doing so, we aim to unravel the mechanisms through which a single gene can assume varied roles or exhibit differential activities "across" diverse tissues or states. Focusing on well-studied biological contexts, such as reliable cell lines or model organism strains, can lead to gaps in which we are uncertain how well our annotations will generalize across the panoply of cellular states.

Considering these challenges, what is the best way for molecular biologists to study and represent gene functions? In searching for an answer to this question, we found parallels to the study of word semantics in the 1990s, summarized in George Miller's seminal essay "To Know A Word."[9] In this perspective, taking direct inspiration from Miller, we summarize his critique of relational models of semantics and describe how transitioning to Miller's distributional model of word semantics unlocked powerful inductive insights exploited by large language models (LLMs)

Please cite this article in press as: Kwon et al., On knowing a gene: A distributional hypothesis of gene function, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.04.008

Box 1. Glossary

Semantics: The study of linguistics and logic that deals with the meaning of language and words.
Sentence context: The set of words surrounding a word or phrase within a sentence.
Embedding: A compact latent signature given as a vector representation of an entity from high-dimensional data into low-dimensional space that preserves entity relationships.
Distributional semantics: The assumption that one can "know a word by the company it keeps," whereby words that occur in similar contexts have similar meanings.
WordNet: A lexical database of defining words based on the hierarchical structure of terms and relational semantics, pioneered by George Miller in 1985.
Biological contexts and cell states: The molecular and biochemical configuration of a cell that uniquely defines a cell's phenotype.
Gene Ontology: A database that annotates the functions of genes based on the hierarchical structure of biology and gene relationships. Notably, "molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions and do not specify where, when, or in what context the action takes place."
Pleiotropy: A gene that affects multiple unrelated phenotypic traits.

today.[10] We then nominate structural correspondences between linguistics and genetics that suggest a distributional hypothesis of gene function. We conclude with broad recommendations for better understanding how these parallels between genetics and linguistics might inform gene function studies in the future. Excellent reviews cover the current state of protein language models and their methodological capabilities.[11,12] We will not review these models; instead, this perspective is intended for a broad multi-disciplinary audience to help biologists better understand the underlying principles that drive the success of inductive computational models. Our focus extends to deciphering the core principles of transformer technology—such as attention mechanisms and large-scale pretraining—that contribute significantly to their success in complex pattern recognition tasks and how these principles can be analogously applied to the interpretation and prediction of gene functions in biological systems. Finally, by shedding light on distributional representations and their potential applications in biology, we propose an approach to characterize gene function from a broader and more holistic perspective, along with suggestions for practical steps necessary to implement the approach.

ON KNOWING A WORD: FROM RELATIONAL TO DISTRIBUTIONAL SEMANTICS

> Knowing a word involves knowing its meaning, and therefore, in my view, knowing a word involves knowing its contexts of use. So my present concern is how to characterize that contextual knowledge—George Miller (On Knowing a Word, 1999)

The term *semantics*, coined by Michel Breal in 1883, is the study of the meaning of words in natural language (Box 1). Different representations of meaning have predominated over time. In the 1980s and 1990s, relational semantics specified that meaning could be broken down into sparse semantic relationships (Figure 1A). For example, two similar words, such as "happiness" and "elation," can be mapped to a general term describing "the state of extreme happiness."[13] Such semantic relationships can be represented as a graph with words as nodes and relationships as links; these words "link" to the same meaning and are designated synonyms. To systematically capture these semantic relations at a large scale, George Miller led a group of cognitive scientists to form the seminal WordNet project in 1985.[14] WordNet collated four lexical databases, one each for nouns, verbs, adjectives, and adverbs, each of which contains sets of synonyms as building blocks associated with one another by semantic relations (Figure 1A). Once complete, WordNet exhibited advantages over machine-readable dictionaries for computational linguistics. For example, synonym sets could be used in information retrieval to expand a user's query and thus retrieve relevant items that might otherwise have been missed.

However, throughout the project, Miller found a critical flaw in relational semantics—its failure to discriminate between alternative meanings of polysemous words (a word with multiple meanings). Polysemic words are easily represented as nodes linking to multiple parental meanings (Figure 1B). Still, such a representation only catalogs the list of possible meanings but fails to discriminate between the meanings a word can have within specific sentences. In his essay "On Knowing a Word," Miller opined that the theory of relational semantics was incorrect and favored a competing approach of distributional semantics to develop and study theories and methods of word meaning.

Distributional semantics was once summarized as the following: "you shall know a word by the company it keeps."[15] Distributional semantics posits that a word's meanings can be empirically derived from sentence contexts in which the word is found,[10] "a cognitive representation of the set of contexts in which a given word form can be used to express a given word meaning." In this framework, a polysemous word would have different contextual representations for each of its various definitions, for example, "apple" computer vs. apple fruit (Figure 1B). Miller advocated for linguistics to embrace learning the meaning of words from large language datasets instead of compiling semantic ontologies such as WordNet.

THE MODERN RISE OF DISTRIBUTIONAL SEMANTICS

Miller's observation was profoundly ahead of its time. Although distributional semantics is conceptually simple (Box 1), exploiting and implementing this idea was technologically challenging during Miller's era. Distributional semantics was realized decades later with advancements such as computational power through Moore's law, the accrual of sizable digital text

Please cite this article in press as: Kwon et al., On knowing a gene: A distributional hypothesis of gene function, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.04.008

CellPress
OPEN ACCESS

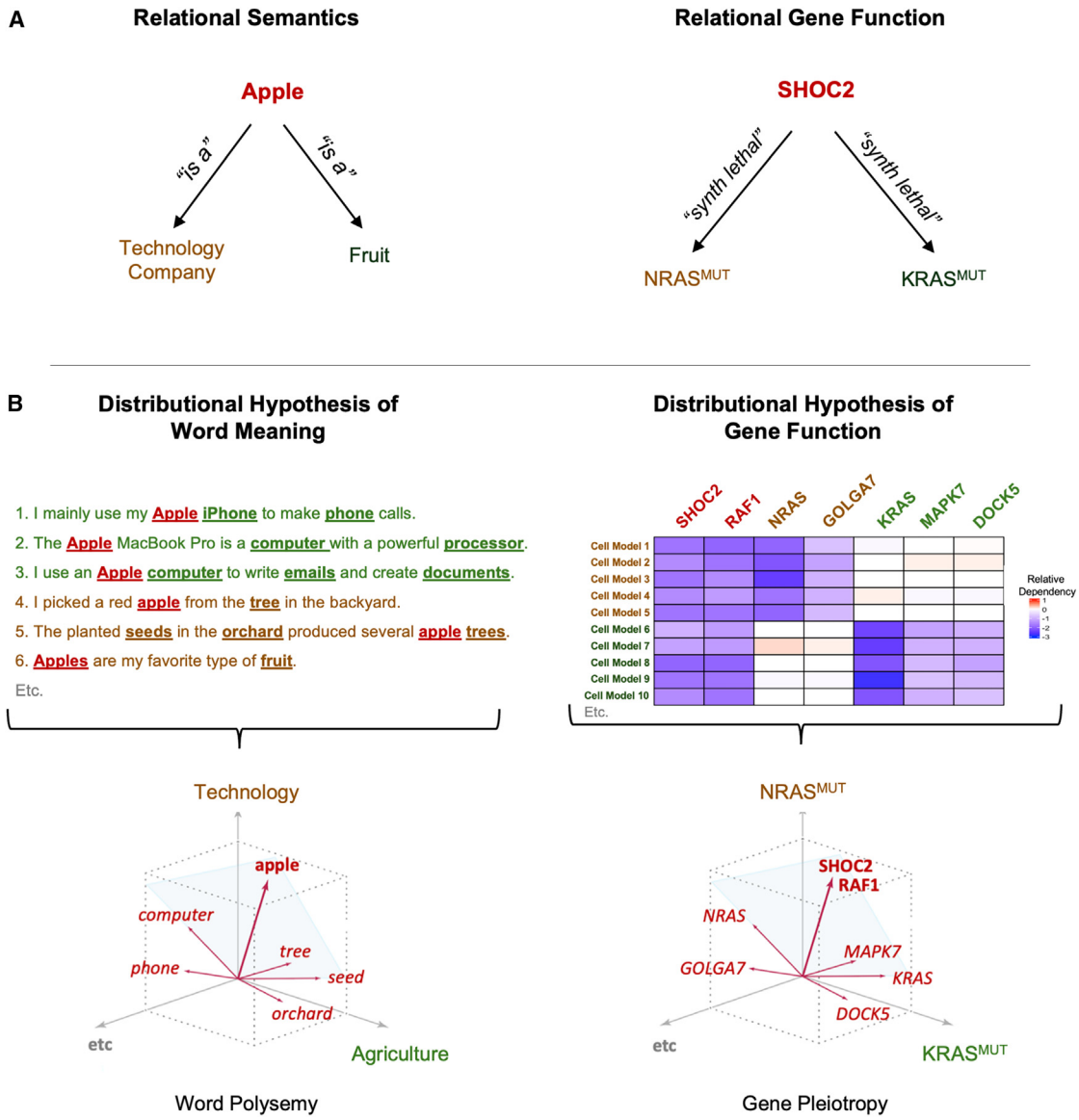

Figure 1. Exploring the parallels between the distributional hypothesis of word semantics and gene function

(A) Relational semantics is an approach to studying the meaning of words by analyzing the relationships between objects, individuals, and propositions. The definition of a word is not viewed as a separate entity but rather as a relationship between different elements of the language. Accordingly, this one-to-one mapping between words and meanings cannot disambiguate words with multiple senses (e.g., polysemy). Similarly, conventional approaches to studying gene function are often relationally defined, making it challenging to disambiguate context-based gene functions.

(B) Distributional semantics is a technique utilized in machine learning to analyze and understand the meanings of words by representing words as vectors in a latent space, which is derived from the co-occurrence distribution of words and their usage. By applying sparse dictionary learning to these vectors, it is possible to capture the multiple meanings of a word (polysemy) and gain a more interpretable understanding of its semantics. Similarly, the effects of genes can be represented as vectors defined by their essentiality in different cell contexts. By studying the co-function of genes in various cell states, latent functions can be discovered, and pleiotropic genes can be modeled as mixtures of these latent biological functions. As more perturbational screens are conducted, it is hoped that the resulting data will serve as a corpus from which gene function can be automatically inferred in a way that is analogous to how machine learning is used to derive word semantics from language datasets.

repositories, and the development of new machine learning approaches.

Modern language models are explicitly trained to model linguistic context, as Miller envisioned.[9] Early models generated word embeddings by directly modeling the distribution of word occurrences in large datasets. In such embedding spaces, a polysemic word can be represented as pointing in multiple directions in semantic space (Figure 1B).[16,17] Furthermore, the implementation of simple, sparse coding has been shown to extract multiple meanings of polysemous words from its embedding.[18] Such embedded representations are used in every natural language processing application that relies on modeling the word meaning.[19]

Advances in transformer architectures using attention further leveraged context to represent words.[20–29] When using attention, a word embedding is mixed with the embeddings of nearby

Please cite this article in press as: Kwon et al., On knowing a gene: A distributional hypothesis of gene function, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.04.008

Table 1. The curious parallels between words and genes

Hierarchical organization	
Syntactic hierarchy	Cellular hierarchy
letters/characters	nucleic acids/amino acids
morpheme	domain
word	gene
sentence	complex
topic	functional module
lexicon/dictionary	genome
writer	cell
Dynamic changes/processes	
Word semantics	Gene function
closed-class/open-class words	*essential/selective genes*
word frequency of occurrence	*gene expression*
cloze task	*gene knockout*

words in a sentence, resulting in a refined embedding that incorporates the word's surroundings. Such contextual embeddings underlie modern language features, including autocomplete, Google search, and the family of generative pre-trained transformers (GPTs), such as ChatGPT.[26]

CONCEPTUAL CORRESPONDENCES BETWEEN THE MEANING OF WORDS AND FUNCTIONS OF GENES

Can the stunning advances in representing word semantics be informative of gene function? The answer depends on whether they share the same inductive biases. Inductive biases are assumptions that allow algorithms to make predictions for inputs they have not encountered during training and thus generalize beyond training datasets. Below, we explain why the similarities between genes and words merit further exploration. We focus on the functional characterization of protein-coding genes while noting that other functional genetic elements, such as gene regulatory regions and non-coding genes, may be understood through a similar distributional lens.

First, there is an apparent organizational hierarchy both in linguistics (morphology)[30] as well as biology (Gene Ontology [GO]),[31] where the combination of fundamental, lower-level elements gives rise to emergent, higher-level entities (Table 1). For example, in the case of semantics, combinations of morphemes such as affixes and roots arrange together to give rise to words with meaning, such as the affix "trans-" in words such as "transport" and "transition." Similarly, proteins are known to possess discrete structural units called domains that can arrange in ways to give rise to functionally distinct genes. For example, kinase domains are structurally conserved and confer the capacity to catalyze the transfer of a phosphate onto a substrate. In the same way that the prefix trans- can imbue the meaning of across to certain words, proteins with shared domains, such as a kinase domain, may provide similar functional capacity but depending on the combination of other domains within a gene can result in different substrate engagement and physiological impact.

Second, both words and genes are susceptible to analogous dynamic changes in fitness over time (Table 1). Further recognition of the "evolvable" characteristics of words is appreciated in evolutionary linguistics.[32] Linguists have developed taxonomical methods of tracing the origins of words through phylogenetic methods.[33] The changes in the meaning of words occur in both passive (semantic drift) and active manner (semantic narrowing, resignification), much like genes (e.g., genetic drift vs. gene flow). The methodological approaches etymologists use to track neologism in etymological dictionaries are similar to how evolutionary geneticists study *de novo* gene birth through evolutionary tracing to catalog allopatric speciation events.

Third, both genes and words are amenable to study through perturbations. For example, the cloze test is a standard method in psycholinguistics studies to test reading comprehension, whereby particular words are occluded, and the subject must fill in the missing word from a word bank.[34] In its modern form as masked auto-encoding,[20] this approach can infer word similarity if they can interchangeably appear in similar contexts. Similarly, genes can be deleted from a cell line, and other genes that rescue that gene's phenotype are considered to have compatible functions.[35] Furthermore, advances in forward genetic screening technologies[36–44] that permit genome-scale perturbational studies have lent insights into the distributions of shared gene function through phenotypic similarity.[44–49]

Fourth, there has been an astronomical increase in data availability, both linguistic and biological. Akin to large language datasets that power language models today, genome-wide omics approaches in biomedicine have enabled the development of biological datasets, where DNA, RNA, and protein biomolecules can be measured at scale across many different biological contexts and diverse samples. Examples of biorepositories for such datasets include the Genotype-Tissue Expression (GTEx) platform with tissue-specific gene expression information, The Cancer Genome Atlas (TCGA) for DNA sequences and mutations, the Human Protein Atlas for protein expression patterns distributed in time and space, and Dependency Map (DepMap) for genetic perturbational data. These datasets have also similarly exhibited sparsity and lower-dimensional structure.[50] RNAi and CRISPR perturbational screens also probe the phenotypic consequence of gene depletion across contexts.[51] Guilt-by-association studies have served as harbingers[52] of this idea, finding that co-variation of gene perturbations at scale across various biological contexts and different phenotypic endpoints can be informative of gene co-function.[44,45,48,53–56]

Finally, genetics may have as much to gain from shifting from relational to distributional representations as semantics did throughout the 2010s. Miller's challenge of modeling polysemic words in the 1990s was later solved with distributional semantic models such as word2vec. Correspondingly, we and others have found that applying similar methods also recovers orthogonal gene functions when applied to genome-scale fitness screens.[52,57,58]

A DISTRIBUTIONAL HYPOTHESIS FOR GENE FUNCTION

Given these correspondences between word semantics and gene function, can the history of the former inform the future of the latter? If "you shall know a word by the company it keeps," perhaps we shall "know a gene by the company it keeps" as well. We join others in the call for a distributional hypothesis of gene function,[59–64] inspired by how distributional semantics

Please cite this article in press as: Kwon et al., On knowing a gene: A distributional hypothesis of gene function, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.04.008

has revolutionized our understanding of word meaning in natural language.

Concretely, we propose shifting away from mapping genes into fixed ontologies of function and toward learning distributional representations of gene function directly from biological data. From the perspective of a scientist using such a database, rather than each gene recovering a list of GO terms, each gene would be mapped to a vector of probabilities (summing to one) of size N, where N is the number of latent variables that capture the informative axes of variation present in the training data, possibly aligning to novel or known pathways. Additionally, the user can access a matrix of size $N \times M$, where M is the number of enumerated cell states that describe the organism of interest. This matrix captures the relationships between the learned variables and the independent biological contexts associated with the organism of interest. Below, we address critical questions about this proposal.

What information can be learned from systems that analyze distributional gene function? Going back to natural language, distributional semantics is grounded in the idea that a lower-dimensional space of "abstracted" "topics" shapes the high-dimensional lexicon (Figure 1B). We believe, analogously, that biological information may capture abstracted biological processes within the genome. Specifically, biological systems evolve and augment biological processes in response to selective pressures. This approach may learn and uncover this lower-dimensional manifold of pathways and processes. The N latent variables in a successfully trained system would model this lower-dimensional manifold.

What information can be captured by latent variables in distributional gene representations? In natural language, latent variables can be interpreted as "topics," into which similar words are mapped based on their co-occurrence across sentence contexts. Biologically, latent variables could capture independent biological pathways that correspond to distinct contexts. For example, circadian genes, like CLOCK and BMAL1, coordinate the temporal expression of other genes over a daily cycle.[65–67] As such, these genes represent the diurnal cycle that evolved in response to the rotation of the Earth.[68] From a dataset whose samples contain temporal diversity, the relationship between CLOCK and BMAL1 could be inferred directly from the bottom up and captured by a particular latent variable (out of N).

What datasets are needed to train learning systems of distributional gene function? The ideal datatype we envision has a structured format with at least three conceptual types of information, including (1) identifiers for perturbed genes, (2) annotated cell contexts, and (3) phenotypic readout(s). Examples of such datasets have been disseminated in recent years[69,70] and when appropriately dimensionally scaled, they furnish the requisite data for the inference of latent variables about gene function, as further discussed below.

Why is incorporating cell states important? Gene functions are distributed over natural states, and the same gene may have different roles in the hierarchical levels of emergent biological organizations (cells → tissues → organs → organisms) or at different stages of development. In this way, the functions of genes we infer from large-scale functional genomic datasets will be contingent on what biological context the model system was used to generate this "corpus" of gene activity. In particular, the set of biological contexts chosen for study defines the distribution of functions that will be observed, highlighting context as a critical choice in experimental design. Recent studies provide proof of concept for these ideas.[57,58] For example, in yeast, adaptive mutations respond to environmental shifts; a small number of phenotypes can predict fitness in native conditions, but additional phenotypes become significant in different environments.[57,58] Additionally, we have recently shown that graph-regularized sparse dictionary learning can geometrically recover the relationships of genes and biological processes that define gene function in a latent space where genes are represented as vectors.[58] Our approach, applied to extensive genomic fitness screening in human cellular models, has underscored the ability to discern distinct, contextually derived genetic functions that are generalizable to a spectrum of cell type contexts,[58] as well as contexts of environmental stressors.[71] Recent transformer models produce context-specific approaches to understanding biological datasets.[72–74] The deployment of similar machine learning methods on non-perturbational datasets, such as transcriptional data, can result in the classification of orthogonal cellular subtypes.[75] Notably, the utilization of comparable machine learning techniques on non-perturbational datasets, such as transcriptional data, primarily leads to the classification of distinct cellular subtypes, highlighting a different aspect of cellular context vs. function.

How will distributional learning systems compare to current approaches to studying gene function? In the pre-genomic era, gene functions were primarily studied through experimental and observational methods. Classical genetics, or forward genetics, seeks to identify the gene associated with an observed phenotype through natural variations across generations or induce mutations in organisms. This method was instrumental in identifying genes linked to specific traits or diseases. In fact, Gregor Mendel's work on pea plants laid the foundation for understanding inheritance patterns, which indirectly offered insights into gene function. Following the advent of the human genome project and gene annotations, gene function prediction has been guided by sequence-based methods, leveraging the inherent information embedded within the primary structure. Starting with a gene of interest, researchers relied on genetic mutations, biochemical assays, and the characterization of phenotypic changes to infer gene functions, often focusing on individual genes or small sets of genes in specific contexts without the comprehensive, high-throughput approaches enabled by genomic technologies today. The approach to predicting gene function based on primary sequences encompasses a range of computational methods that analyze sequence homology, structural motifs, and evolutionary relationships.[76–79] These methods leverage the principle that sequence similarity implies functional similarity, allowing researchers to rapidly annotate genes in newly sequenced genomes or identify functional domains within proteins. Despite the incredible enablement this approach provides, sequence-based methods have limitations, particularly in cases where gene functions are influenced by higher-ordered protein structures or are less conserved. The emergence of structure-based prediction techniques from sequence and protein structures offers detailed, residue-level annotations and has shown superiority over existing methods in accurately determining protein functions.[80–83] However, the reliance on

Please cite this article in press as: Kwon et al., On knowing a gene: A distributional hypothesis of gene function, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.04.008

experimentally determined structures, which are less abundant than sequence data, poses a significant limitation. Recent developments, such as ESMFold, AlphaFold, and OpenFold, highlight the complementary nature of sequence- and structure-based approaches.[81,84,85] Sequence-based methods excel in predicting evolutionary-constrained functions, yet they fall short in identifying diverse functional outcomes stemming from different protein folds, a strength of structure-based methods. Integrating multimodal learning approaches may be essential in capturing the full complexity of gene function. Recent representative studies have elucidated the significance of combining multiple data types.[80,86–90] This multimodal approach addresses the limitations of singular data type analyses and highlights how additional datatypes offer a greater comprehensive understanding of gene function across various biological contexts.

And finally, how do we get there? A new initiative would be required to complement the current GO paradigm, operating in multiple stages. In the first stage, we propose curating a large and consistently processed corpus of biological datasets covering the underlying relationships between genes and their possible functions across contexts. In the second stage, we could recommend self-supervised objectives, such as language modeling and contrastive learning, coupled with benchmarks for assessing model performance on known biological relationships. In the third stage, the focus would shift to pretraining a model on this dataset on a larger scale. By adapting the model across a broad array of functional tasks, the objective would be to encapsulate the diverse aspects of interrelationships present within the data, mirroring the approach taken with words in sentences as per the distributional hypothesis. In the fourth stage, we envision a close interdisciplinary collaboration with experimentalists to generate additional functional measurements to test generalization and assess whether novel insights were obtained from the pretraining stage. This setup would propel the theory toward practical application and pave the way for pioneering models in gene function prediction.

We envision that this systematic investigation of gene function would be enacted practically by leveraging genetic perturbation and small molecule modulation, as these methods are readily deployable and poised to generate expansive datasets, ideal for tokenization to produce unified multimodal data representations that can be leveraged using transformer and other emerging neural architectures. This endeavor will harness the power of self-supervised learning for the initial exploration of the dataset for gene function, followed by deploying a scalable modeling framework capable of predicting further perturbations and their outcomes. This analytical approach would be paired with a "self-driving" lab to iteratively direct and refine biological experimentation toward a more comprehensive and functionally insightful understanding of gene function. These steps will be instrumental in forging a revolutionary tool for predicting gene function in the future.

CONCLUSION

In the face of complexity, biology has relied on consilient metaphors to evoke principles from other disciplines. The most impactful analogies have ranged from modularity in engineering design,[91] landscapes,[92] cell circuits,[93] switches,[94] and social graphs.[94] Here, we have summarized historical parallels and structural correspondences, suggesting that genetics and natural language may have much in common. Genetics follows conceptually corresponding structural rules as the language to encode information about an agent's environment. Darwin noted the "curious parallels" between biological and linguistic evolution in the descent of man.[95] Just as words have meanings that depend on their context, genes have functions that rely on the cellular context in which they are expressed. By shifting our conceptualization from relational to distributional representations of gene function, we may benefit from the inductive biases powering successful self-supervised models of natural language.

However, the distinction between protein complexes and pathways is not always apparent in biological systems. Genes may interact dynamically and are not necessarily ordered the way words are. Genetic elements, such as promoters, introns, exons, etc., also lack an equivalent in the distributional semantics of natural language. This highlights the need for thoughtfully developing bespoke machine learning models for biology. Although the analogy between semantics and genetics is informative, it is essential to recognize the distinct differences between words and genes, warranting caution against overinterpreting these similarities.

To do so, structured databases of perturbations are required. These should span diverse cell contexts, be captured by different biological assays, and be harmonized with the burgeoning single-cell atlases of cell types across organisms. Machine learning and artificial intelligence can play a significant role in this data curation process, ensuring that models can maximally exploit biological datasets. If appropriately constructed, these biological corpora may enable reasoning about how individual genes contribute to biological complexity. Conversely, developing bespoke computational models for biological data will help unlock new insights into the fundamental principles governing the language of life.

ACKNOWLEDGMENTS

J.J.K. was supported by NIH/NCI K99CA270290. W.C.H. was supported by NIH/NCI U01CA176058. M.Z. gratefully acknowledges the support of NIH R01-HD108794, NSF CAREER 2339524, US DoD FA8702-15-D-0001, and awards from the Harvard Data Science Initiative, Amazon Faculty Research, the Google Research Scholar Program, AstraZeneca Research, the Roche Alliance with Distinguished Scientists, the Sanofi iDEA-iTECH Award, Pfizer Research, the Chan Zuckerberg Initiative, the John and Virginia Kaneb Fellowship Award at Harvard Medical School, the Aligning Science Across Parkinson's (ASAP) Initiative, the Biswas Computational Biology Initiative in partnership with the Milken Institute, and the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

AUTHOR CONTRIBUTIONS

J.J.K., J.P., G.G., W.C.H., and M.Z. conceptualized and wrote the manuscript.

DECLARATION OF INTERESTS

W.C.H. is a consultant for Thermo Fisher, Solasta Ventures, MPM Capital, KSQ Therapeutics, Tyra Biosciences, Jubilant Therapeutics, RAPPTA Therapeutics, Function Oncology, Frontier Medicines, Riva Therapeutics, Serinus

Biosciences, Kestrl Biosciences, and Calyx. J.J.K. is a consultant for A2A Pharmaceuticals and Longitude Capital.

REFERENCES

1. Keeling, D.M., Garza, P., Nartey, C.M., and Carvunis, A.-R. (2019). The meanings of "function" in biology and the problematic case of de novo gene emergence. eLife *8*, e47014. https://doi.org/10.7554/eLife.47014.

2. Ellens, K.W., Christian, N., Singh, C., Satagopam, V.P., May, P., and Linster, C.L. (2017). Confronting the catalytic dark matter encoded by sequenced genomes. Nucleic Acids Res. *45*, 11495–11514. https://doi.org/10.1093/nar/gkx937.

3. Stoeger, T., Gerlach, M., Morimoto, R.I., and Nunes Amaral, L.A. (2018). Large-scale investigation of the reasons why potentially important genes are ignored. PLoS Biol. *16*, e2006643. https://doi.org/10.1371/journal.pbio.2006643.

4. Thul, P.J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L.M., et al. (2017). A subcellular map of the human proteome. Science *356*, eaal3321. https://doi.org/10.1126/science.aal3321.

5. Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419. https://doi.org/10.1126/science.1260419.

6. Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Res. *25*, 1491–1498. https://doi.org/10.1101/gr.190595.115.

7. Clevers, H. (2017). What is your conceptual definition of "'cell type'" in the context of a mature organism? What is an adult cell type, really? Cell Syst. *4*, 255–259. https://doi.org/10.1016/j.cels.2017.03.006.

8. Morris, S.A. (2019). The evolving concept of cell identity in the single cell era. Development *146*, dev169748. https://doi.org/10.1242/dev.169748.

9. Miller, G.A. (1999). On knowing a word. Annu. Rev. Psychol. *50*, 1–19. https://doi.org/10.1146/annurev.psych.50.1.1.

10. Miller, G.A., and Charles, W.G. (1991). Contextual correlates of semantic similarity. Lang. Cogn. Process *6*, 1–28. https://doi.org/10.1080/01690969108406936.

11. Bepler, T., and Berger, B. (2021). Learning the protein language: Evolution, structure, and function. Cell Syst. *12*, 654–669.e3. https://doi.org/10.1016/j.cels.2021.05.017.

12. Ferruz, N., and Höcker, B. (2022). Controllable protein design with language models. Nat. Mach. Intell. *4*, 521–532. https://doi.org/10.1038/s42256-022-00499-z.

13. Fellbaum, C. (1998). WordNet 1. 6: An Electronic Lexical Database (Bradford Books).

14. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K.J. (1990). Introduction to WordNet: An on-line lexical database. Int. J. Lexicography *3*, 235–244. https://doi.org/10.1093/ijl/3.4.235.

15. Firth, J.R. (1952). A synopsis of linguistic theory 1930-1955. In Studies in Linguistic Analysis, J.R. Firth and F.R. Palmer, eds. (Longman), pp. 1–32.

16. Yun, Z., Chen, Y., Olshausen, B., and LeCun, Y. (2021). Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. In Proceedings of the Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures (Association for Computational Linguistics), pp. 1–10. https://doi.org/10.18653/v1/2021.deelio-1.1.

17. Zhang, J., Chen, Y., Cheung, B., and Olshausen, B.A. (2019). Word embedding visualization via dictionary learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.1910.03833.

18. Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. (2016). Linear algebraic structure of word senses, with applications to polysemy. Preprint at arXiv. https://doi.org/10.48550/arXiv.1601.03764.

19. Jurafsk, D. and Martin, J.H. Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition (Prentice Hall).

20. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional Transformers for language understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.04805.

21. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog *1*, 9.

22. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q.V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1906.08237.

23. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. Preprint at arXiv. https://doi.org/10.48550/arXiv.1909.11942.

24. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. Preprint at arXiv. https://doi.org/10.48550/arXiv.1907.11692.

25. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P.J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. Preprint at arXiv. https://doi.org/10.48550/arXiv.1910.10683.

26. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language Models are Few-Shot Learners. Preprint at arXiv. https://doi.org/10.48550/arXiv.2005.14165.

27. Clark, K., Luong, M.-T., Le, Q.V., and Manning, C.D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. Preprint at arXiv. https://doi.org/10.48550/arXiv.2003.10555.

28. He, P., Liu, X., Gao, J., and Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with disentangled attention. Preprint at arXiv. https://doi.org/10.48550/arXiv.2006.03654.

29. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al. (2022). PaLM: Scaling language modeling with Pathways. Preprint at arXiv. https://doi.org/10.48550/arXiv.2204.02311.

30. Mark Aronoff, A.K.F. (2007). What is Morphology? For. Mod. Lang. Stud. *43*, 93.

31. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29. https://doi.org/10.1038/75556.

32. Dunn, M. (2014). Evolutionary Linguistics by April McMahon and Robert McMahon. American Anthropologist *116*, 690–691. https://doi.org/10.1111/aman.12136_17.

33. Platnick, N.I., and Cameron, H.D. (1977). Cladistic methods in textual, linguistic, and phylogenetic analysis. Systematic Biology *26*, 380–385. https://doi.org/10.1093/sysbio/26.4.380.

34. Taylor, W.L. (1953). "cloze procedure": A new tool for measuring readability. Journalism Q. *30*, 415–433. https://doi.org/10.1177/107769905303000401.

35. Yook, K. (2005). Complementation. WormBook, 1–17.

36. Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature *462*, 108–112. https://doi.org/10.1038/nature08460.

37. Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelson, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., and Zhang, F. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. Science *343*, 84–87. https://doi.org/10.1126/science.1247005.

38. Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. Science *343*, 80–84. https://doi.org/10.1126/science.1246981.

39. Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., et al. (2014). Genome-scale CRISPR-mediated control of gene repression and activation. Cell *159*, 647–661. https://doi.org/10.1016/j.cell.2014.09.029.

40. Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert, L.A., Horlbeck, M.A., Hein, M.Y., et al. (2016). A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. Cell *167*, 1867–1882.e21. https://doi.org/10.1016/j.cell.2016.11.048.

41. Jaitin, D.A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T.M., Tanay, A., van Oudenaarden, A., and Amit, I. (2016). Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. Cell *167*, 1883–1896.e15. https://doi.org/10.1016/j.cell.2016.11.039.

42. Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell *167*, 1853–1866.e17. https://doi.org/10.1016/j.cell.2016.11.038.

43. Feldman, D., Singh, A., Schmid-Burgk, J.L., Carlson, R.J., Mezger, A., Garrity, A.J., Zhang, F., and Blainey, P.C. (2019). Optical pooled screens in human cells. Cell *179*, 787–799.e17. https://doi.org/10.1016/j.cell.2019.09.016.

44. Replogle, J.M., Saunders, R.A., Pogson, A.N., Hussmann, J.A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E.J., Adelman, K., Lithwick-Yanai, G., et al. (2022). Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Cell *185*, 2559–2575.e28. https://doi.org/10.1016/j.cell.2022.05.013.

45. Pan, J., Meyers, R.M., Michel, B.C., Mashtalir, N., Sizemore, A.E., Wells, J.N., Cassel, S.H., Vazquez, F., Weir, B.A., Hahn, W.C., et al. (2018). Interrogation of Mammalian Protein Complex Structure, Function, and Membership Using Genome-Scale Fitness Screens. Cell Syst. *6*, 555–568.e7. https://doi.org/10.1016/j.cels.2018.04.011.

46. Norman, T.M., Horlbeck, M.A., Replogle, J.M., Ge, A.Y., Xu, A., Jost, M., Gilbert, L.A., and Weissman, J.S. (2019). Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. Science *365*, 786–793. https://doi.org/10.1126/science.aax4438.

47. Bayraktar, E.C., La, K., Karpman, K., Unlu, G., Ozerdem, C., Ritter, D.J., Alwaseem, H., Molina, H., Hoffmann, H.-H., Millner, A., et al. (2020). Metabolic coessentiality mapping identifies C12orf49 as a regulator of SREBP processing and cholesterol metabolism. Nat. Metab. *2*, 487–498. https://doi.org/10.1038/s42255-020-0206-9.

48. Wainberg, M., Kamber, R.A., Balsubramani, A., Meyers, R.M., Sinnott-Armstrong, N., Hornburg, D., Jiang, L., Chan, J., Jian, R., Gu, M., et al. (2021). A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. Nat. Genet. *53*, 638–649. https://doi.org/10.1038/s41588-021-00840-z.

49. Kim, E., Novak, L.C., Lin, C., Colic, M., Bertolet, L.L., Gheorghe, V., Bristow, C.A., and Hart, T. (2022). Dynamic rewiring of biological activity across genotype and lineage revealed by context-dependent functional interactions. Genome Biol. *23*, 140. https://doi.org/10.1186/s13059-022-02712-z.

50. Petti, S., Reddy, G., and Desai, M.M. (2022). Inferring sparse structure in genotype-phenotype maps. Preprint at bioRxiv. https://doi.org/10.1101/2022.09.27.509675.

51. Przybyla, L., and Gilbert, L.A. (2022). A new era in functional genomics screens. Nat. Rev. Genet. *23*, 89–103. https://doi.org/10.1038/s41576-021-00409-w.

52. Petti, S., Reddy, G., and Desai, M.M. (2023). Inferring sparse structure in genotype–phenotype maps. Genetics *225*, iyad127. https://doi.org/10.1093/genetics/iyad127.

53. Tong, A.H.Y., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., et al. (2004). Global mapping of the yeast genetic interaction network. Science *303*, 808–813. https://doi.org/10.1126/science.1091317.

54. Boone, C., Bussey, H., and Andrews, B.J. (2007). Exploring genetic interactions and networks with yeast. Nat. Rev. Genet. *8*, 437–449. https://doi.org/10.1038/nrg2085.

55. Wang, T., Yu, H., Hughes, N.W., Liu, B., Kendirli, A., Klein, K., Chen, W.W., Lander, E.S., and Sabatini, D.M. (2017). Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. Cell *168*, 890–903.e15. https://doi.org/10.1016/j.cell.2017.01.013.

56. Funk, L., Su, K.-C., Ly, J., Feldman, D., Singh, A., Moodie, B., Blainey, P.C., and Cheeseman, I.M. (2022). The phenotypic landscape of essential human genes. Cell *185*, 4634–4653.e22. https://doi.org/10.1016/j.cell.2022.10.017.

57. Kinsler, G., Geiler-Samerotte, K., and Petrov, D.A. (2020). Fitness variation across subtle environmental perturbations reveals local modularity and global pleiotropy of adaptation. eLife *9*, e61271. https://doi.org/10.7554/eLife.61271.

58. Pan, J., Kwon, J.J., Talamas, J.A., Borah, A.A., Vazquez, F., Boehm, J.S., Tsherniak, A., Zitnik, M., McFarland, J.M., and Hahn, W.C. (2022). Sparse dictionary learning recovers pleiotropy from human cell fitness screens. Cell Syst. *13*, 286–303.e10. https://doi.org/10.1016/j.cels.2021.12.005.

59. Wagner, G.P., and Zhang, J. (2011). The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. Nat. Rev. Genet. *12*, 204–213. https://doi.org/10.1038/nrg2949.

60. Fraser, A.G., and Marcotte, E.M. (2004). A probabilistic view of gene function. Nat. Genet. *36*, 559–564. https://doi.org/10.1038/ng1370.

61. Civelek, M., and Lusis, A.J. (2014). Systems genetics approaches to understand complex traits. Nat. Rev. Genet. *15*, 34–48. https://doi.org/10.1038/nrg3575.

62. Wang, S., Ma, J., Fong, S., Rensi, S., Han, J., Peng, J., Pratt, D., Altman, R.B., and Ideker, T. (2019). Deep functional synthesis: a machine learning approach to gene functional enrichment. Preprint at bioRxiv. https://doi.org/10.1101/824086.

63. Costanzo, M., Kuzmin, E., van Leeuwen, J., Mair, B., Moffat, J., Boone, C., and Andrews, B. (2019). Global genetic networks and the genotype-to-phenotype relationship. Cell *177*, 85–100. https://doi.org/10.1016/j.cell.2019.01.033.

64. Kustatscher, G., Collins, T., Gingras, A.-C., Guo, T., Hermjakob, H., Ideker, T., Lilley, K.S., Lundberg, E., Marcotte, E.M., Ralser, M., and Rappsilber, J. (2022). Understudied proteins: opportunities and challenges for functional proteomics. Nat. Methods *19*, 774–779. https://doi.org/10.1038/s41592-022-01454-x.

65. Vitaterna, M.H., King, D.P., Chang, A.M., Kornhauser, J.M., Lowrey, P.L., McDonald, J.D., Dove, W.F., Pinto, L.H., Turek, F.W., and Takahashi, J.S. (1994). Mutagenesis and mapping of a mouse gene, Clock, essential for circadian behavior. Science *264*, 719–725. https://doi.org/10.1126/science.8171325.

66. Bunger, M.K., Wilsbacher, L.D., Moran, S.M., Clendenin, C., Radcliffe, L.A., Hogenesch, J.B., Simon, M.C., Takahashi, J.S., and Bradfield, C.A. (2000). Mop3 is an essential component of the master circadian pacemaker in mammals. Cell *103*, 1009–1017. https://doi.org/10.1016/s0092-8674(00)00205-1.

67. McNamara, P., Seo, S.B., Rudic, R.D., Sehgal, A., Chakravarti, D., and FitzGerald, G.A. (2001). Regulation of CLOCK and MOP4 by nuclear hormone receptors in the vasculature: a humoral mechanism to reset a peripheral clock. Cell *105*, 877–889. https://doi.org/10.1016/s0092-8674(01)00401-9.

68. Bhadra, U., Thakkar, N., Das, P., and Pal Bhadra, M. (2017). Evolution of circadian rhythms: from bacteria to human. Sleep Med. *35*, 49–61. https://doi.org/10.1016/j.sleep.2017.04.008.

69. Cui, A., Huang, T., Li, S., Ma, A., Pérez, J.L., Sander, C., Keskin, D.B., Wu, C.J., Fraenkel, E., and Hacohen, N. (2024). Dictionary of immune responses to cytokines at single-cell resolution. Nature *625*, 377–384. https://doi.org/10.1038/s41586-023-06816-9.

70. Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M., et al. (2017). Defining a cancer dependency map. Cell *170*, 564–576.e16. https://doi.org/10.1016/j.cell.2017.06.010.

71. Olivieri, M., Cho, T., Álvarez-Quilón, A., Li, K., Schellenberg, M.J., Zimmermann, M., Hustedt, N., Rossi, S.E., Adam, S., Melo, H., et al. (2020). A genetic map of the response to DNA damage in human cells. Cell *182*, 481–496.e21. https://doi.org/10.1016/j.cell.2020.05.040.

72. Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., Lu, H., and Yao, J. (2022). scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nat. Mach. Intell. *4*, 852–866. https://doi.org/10.1038/s42256-022-00534-z.

73. Theodoris, C.V., Xiao, L., Chopra, A., Chaffin, M.D., Al Sayed, Z.R., Hill, M.C., Mantineo, H., Brydon, E.M., Zeng, Z., Liu, X.S., and Ellinor, P.T. (2023). Transfer learning enables predictions in network biology. Nature *618*, 616–624. https://doi.org/10.1038/s41586-023-06139-9.

74. Li, M.M., Huang, Y., Sumathipala, M., Liang, M.Q., Valdeolivas, A., Ananthakrishnan, A.N., Liao, K., Marbach, D., and Zitnik, M. (2024). Contextualizing protein representations using deep learning on protein networks and single-cell data. Preprint at bioRxiv. https://doi.org/10.1101/2023.07.18.549602.

75. Khan, A., and Lee, B. (2023). DeepGene Transformer: Transformer for the gene expression-based classification of cancer subtypes. Expert Syst. Appl. *226*, 120047. https://doi.org/10.1016/j.eswa.2023.120047.

76. Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P. (1998). SMART, a simple modular architecture research tool: Identification of signaling domains. Proc. Natl. Acad. Sci. USA *95*, 5857–5864. https://doi.org/10.1073/pnas.95.11.5857.

77. Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L.L., et al. (2008). The Pfam protein families database. Nucleic Acids Res. *36*, D281–D288. https://doi.org/10.1093/nar/gkm960.

78. Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., et al. (2012). InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res. *40*, D306–D312. https://doi.org/10.1093/nar/gkr948.

79. Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R., and Finn, R.D. (2018). HMMER web server: 2018 update. Nucleic Acids Res. *46*, W200–W204. https://doi.org/10.1093/nar/gky448.

80. Gligorijević, V., Renfrew, P.D., Kosciolek, T., Leman, J.K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B.C., Fisk, I.M., Vlamakis, H., et al. (2021). Structure-based protein function prediction using graph convolutional networks. Nat. Commun. *12*, 3168. https://doi.org/10.1038/s41467-021-23303-9.

81. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

82. Kulmanov, M., and Hoehndorf, R. (2020). DeepGOPlus: improved protein function prediction from sequence. Bioinformatics *36*, 422–429. https://doi.org/10.1093/bioinformatics/btz595.

83. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science *373*, 871–876. https://doi.org/10.1126/science.abj8754.

84. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science *379*, 1123–1130. https://doi.org/10.1126/science.ade2574.

85. Ahdritz, G., Bouatta, N., Floristean, C., Kadyan, S., Xia, Q., Gerecke, W., O'Donnell, T.J., Berenberg, D., Fisk, I., Zanichelli, N., et al. (2022). OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. Preprint at bioRxiv. https://doi.org/10.1101/2022.11.20.517210.

86. Ektefaie, Y., Dasoulas, G., Noori, A., Farhat, M., and Zitnik, M. (2023). Multimodal learning with graphs. Nat. Mach. Intell. *5*, 340–350. https://doi.org/10.1038/s42256-023-00624-6.

87. Kabir, A., and Shehu, A. (2022). GOProFormer: A multi-modal transformer method for Gene Ontology protein function prediction. Biomolecules *12*, 1709. https://doi.org/10.3390/biom12111709.

88. Tang, X., Zhang, J., He, Y., Zhang, X., Lin, Z., Partarrieu, S., Hanna, E.B., Ren, Z., Shen, H., Yang, Y., et al. (2023). Explainable multi-task learning for multi-modality biological data analysis. Nat. Commun. *14*, 2546. https://doi.org/10.1038/s41467-023-37477-x.

89. Lei, Y., Li, S., Liu, Z., Wan, F., Tian, T., Li, S., Zhao, D., and Zeng, J. (2021). A deep-learning framework for multi-level peptide–protein interaction prediction. Nat. Commun. *12*, 5465. https://doi.org/10.1038/s41467-021-25772-4.

90. McDermott, M.B.A., Yap, B., Szolovits, P., and Zitnik, M. (2023). Structure-inducing pre-training. Nat. Mach. Intell. *5*, 612–621. https://doi.org/10.1038/s42256-023-00647-z.

91. Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. (1999). From molecular to modular cell biology. Nature *402*, C47–C52. https://doi.org/10.1038/35011540.

92. Huang, S. (2012). The molecular and mathematical basis of Waddington's epigenetic landscape: a framework for post-Darwinian biology? BioEssays *34*, 149–157. https://doi.org/10.1002/bies.201100031.

93. Chen, Z., and Elowitz, M.B. (2021). Programmable protein circuit design. Cell *184*, 2284–2301. https://doi.org/10.1016/j.cell.2021.03.007.

94. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). How Genetic Switches Work (Garland Science).

95. Darwin, C. (1888). The Descent of Man And Selection in Relation to Sex (Cambridge University Press).