DeNoFo: a file format and toolkit for standardised, comparable de novo gene annotation

Elias Dohmen,[1,*] **Margaux Aubel**,[1] **Lars A. Eicholt**,[1] **Paul Roginski**,[2] **Victor Luria**,[3, 4, 5] **Amir Karger**[6] **and Anna Grandchamp**[7]

[1]Institute for Evolution and Biodiversity, University of Münster, Hüfferstrasse 1, 48151, Münster, Germany, [2]Institute for Integrative Biology of the Cell (I2BC), Université Paris-Saclay, CEA, CNRS, 1 Av. de la Terrasse Bâtiment 21, 91198, Gif-sur-Yvette, France, [3]Department of Neuroscience, Yale School of Medicine, New Haven, 06510, CT, USA, [4]Department of Systems Biology, Harvard Medical School, Boston, 02115, MA, USA, [5]Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, 02115, MA, USA, [6]IT-Research Computing, Harvard Medical School, Boston, 02115, MA, USA and [7]Aix Marseille University, INSERM, TAGC, UMR_S1090,Marseille, France

*Corresponding author: e.dohmen@uni-muenster.de

Abstract

Motivation *De novo* genes emerge from previously non-coding regions of the genome, challenging the traditional view that new genes primarily arise through duplication and adaptation of existing ones. Characterised by their rapid evolution and their novel structural properties or functional roles, *de novo* genes represent a young area of research. Therefore, the field currently lacks established standards and methodologies, leading to inconsistent terminology and challenges in comparing and reproducing results.
Results This work presents a standardised annotation format to document the methodology of *de novo* gene datasets in a reproducible way. We developed DeNoFo, a toolkit to provide easy access to this format that simplifies annotation of datasets and facilitates comparison across studies. Unifying the different protocols and methods in one standardised format, while providing integration into established file formats, such as fasta or gff, ensures comparability of studies and advances new insights in this rapidly evolving field.
Availability and Implementation DeNoFo is available through the official Python Package Index (PyPI) and at `https://github.com/EDohmen/denofo`. All tools have a graphical user interface and a command line interface. The toolkit is implemented in Python3, available for all major platforms and installable with pip and uv.

Key words: de novo genes, proto-genes, annotation format, standardisation and comparability

Introduction

De novo genes are defined as genes that emerge from previously non-genic sequences within a genome. Since the early 2000s, significant progress has been made in understanding the mechanisms underlying their emergence (Long et al., 2003; Begun et al., 2007; Cai et al., 2008; Knowles and McLysaght, 2009; Tautz and Domazet-Lošo, 2011; Carvunis et al., 2012; Wissler et al., 2013; Schmitz et al., 2018; Van Oss and Carvunis, 2019; Heames et al., 2020; Zheng and Zhao, 2022; Parikh et al., 2022; Vakirlis et al., 2022; Zhao et al., 2024; Rich and Carvunis, 2023; Iyengar and Bornberg-Bauer, 2023; Aldrovandi et al., 2024; Peng and Zhao, 2024). The prevalence of *de novo* genes across all domains of life is now well-documented, and their biological roles within respective species remain an active area of research (Li et al., 2010; Gubala et al., 2017; Baalsrud et al., 2018; Zhang et al., 2019; Rivard et al., 2021; Lange et al., 2021; Broeils et al., 2023; Vakirlis and Kupczok, 2024).

A major challenge in the study of *de novo* genes is the accurate discrimination of recently originated *de novo* genes from those that have emerged via alternative mechanisms (Schmitz et al., 2018). This challenge stems from the scarcity of automated pipelines capable of automatically detecting and annotating *de novo* genes, which have only recently been developed (Arendsee et al., 2019; Casola et al., 2022; Roginski et al., 2024). Consequently, the majority of *de novo* genes have been identified through the implementation of non-standardized pipelines that integrate diverse software and custom scripts. Moreover, the characterisation criteria of *de novo* emergence, which varies across studies (Van Oss and Carvunis, 2019), is a crucial factor in the identification of *de novo* genes. Past definitions have varied, incorporating different evolutionary stages and criteria (Keeling et al., 2019; Weisman, 2022), thresholds for how much of a gene must have arisen *de novo* to qualify as such (McLysaght and Hurst, 2016), criteria for characterising the absence of homology (Casola, 2018; Vakirlis et al., 2020; Weisman et al., 2022), and different modes of emergence (Pereira et al., 2024). The ambiguities in

defining and identifying *de novo* gene emergence have prompted the usage of diverse methodologies. While exploring various methods is essential for a relatively young field of research, the lack of consensus on definitions and the use of varying methodologies, have hindered the effective comparison of results across studies (Schmitz et al., 2018; Aubel et al., 2023; Grandchamp et al., 2025).

As a consequence, datasets comprising *de novo* genes from the same species can differ considerably, and in some cases, may not even be overlapping, depending on the methodologies and definitions employed to determine the *de novo* status of a given gene. Several studies investigating *de novo* gene emergence in the same species show large discrepancies. For example, in Roginski et al. (2024), the authors detected 89 *de novo* genes in humans, while Vakirlis et al. (2022) identified 155 de novo Open Reading Frames (ORFs), and Dowling et al. (2020) identified 2,749 human-specific *de novo* ORFs. The observed discrepancies can be attributed to methodological differences. Specifically, Roginski et al. (2024) detected *de novo* genes through the analysis of annotated genomes, Vakirlis et al. (2022) utilized ORFs identified via ribosome profiling as candidates, and Dowling et al. (2020) sought candidates within transcriptomic data.

Similarly, Heames et al. (2020) detected 66 *de novo* genes in *Drosophila melanogaster*, while Roginski et al. (2024) detected 92 and Peng and Zhao (2024) 555 *de novo* genes. In the same species Zheng and Zhao (2022) identified 993 *de novo* ORFs and a fifth study detected an average of 1,548 *de novo* ORFs (Grandchamp et al., 2023). These discrepancies primarily stem from variations in methodology and underlying data sources.

These methodological discrepancies and different definitions are discussed in more detail in Grandchamp et al. (2025), along with the categorisation and standardisation of these methods for the annotation format presented here. The characterisation and understanding of *de novo* gene emergence is a young research field, and for future advancements, the exploration of different methodologies remains crucial. However, the wide variety of methodologies complicates the comparability of datasets and hinders progress.

Consequently, there is an essential need for a standardised and comparable approach to ensure reproducibility and facilitate the comparison of results across studies without restrictions on the applied methods and definitions. To address this need, we present DeNoFo, a toolkit designed to automate and streamline the annotation process for *de novo* gene detection and validation as well as rapid and straightforward comparison of methodology across studies.

Tool Description

The standardised annotation format

The present diversity in definitions, terms, protocols, tools and pipelines mentioned above to investigate *de novo* gene evolution impedes comparability of datasets and studies.

While it is important to enable a certain flexibility regarding the methodology to detect and describe *de novo* genes, other factors such as the meaning of terms or comparable description of approaches should be standardised to enable reproducibility and comparability. We propose the following annotation format to categorise and standardise the methods used for *de novo* gene detection while still allowing for the required flexibility to include a great variety of approaches. With this, information on used protocols and tools from both bioinformatic and wet lab approaches can be integrated into widely-accepted file formats in genomics.

The here described format focuses on the methodology used for the detection of *de novo* genes in different studies, in contrast to providing information about individual genes. By omitting gene-centric properties and only covering methodological aspects, all genes of a study can be covered by a single description of their methodology and studies become comparable between each other.

Based on the standardised and categorised methods for bioinformatic detection of *de novo* genes we compiled (Grandchamp et al., 2025), the format is structured into six main sections (see Fig. 1): input data, evolutionary information, homology filter, non-coding homologs, lab verification and hyperlinks(URLs/DOIs).

Our annotation format with the here described information can be stored either in a separate file and shared as such or can be included in widely-accepted bioinformatic file formats. The default DeNoFo file format is JSON-based and human-readable to a high degree (*.dngf file extension). To allow for integration into widely-accepted file types and formats, such as fasta or gff, we developed additionally a short string encoding of the DeNoFo format, which can be added to the fasta header or the additional info column in gff files. This short string encoding of the DeNoFo format is not human-readable due to its compressed information content. For easy use and transfer of information into this proposed unified format as well as extraction of information from the mentioned file types annotated with this format, we developed a toolkit to facilitate the usage.

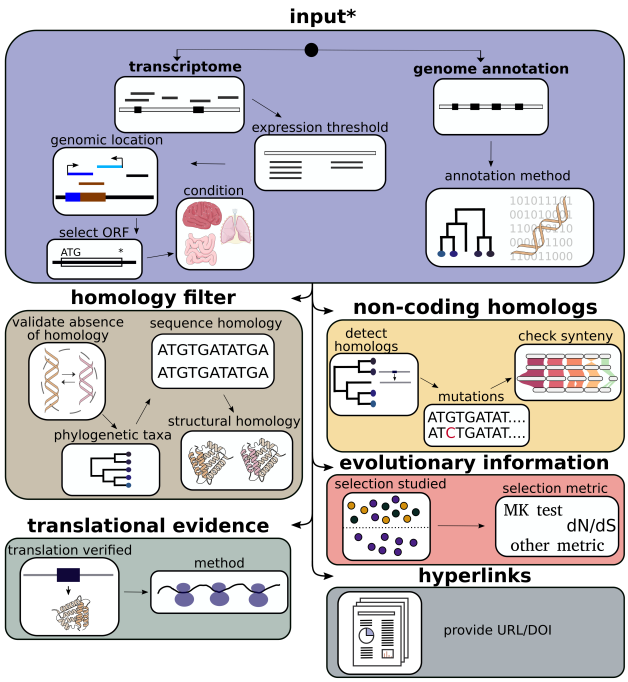

Fig. 1: Structure of the de novo gene annotation format with six main sections and related subsections. The star indicates non-optional sections. This figure has been designed using resources from https://bioicons.com/.

Toolkit Functionality

The tools of the DeNoFo toolkit assist in the conversion of workflows into the required format, the automated annotation of *de novo* genes with the format, and the extraction of information in an easily accessible and interpretable manner. Additionally, a tool is provided for the comparison of two such *de novo* gene set annotations with regard to their similarities and differences in methodological approaches. All tools presented here are available with a graphical user interface (GUI) and a command-line interface (CLI).

The **DeNoFo-Questionnaire** represents the core component of the toolkit, serving as a guide that directs users through the required sections of the format via interactive queries. These queries offer either pre-populated options or the option to enter custom answers. To ensure user convenience, the user is permitted to move between questions and modify previous answers as required. The resulting DeNoFo annotation is saved to a user-specified file in a JSON-like *de novo* gene file format (dngf, that is also the file extension).

The **DeNoFo-Converter** is a tool designed for the purpose of converting the DeNoFo format from one file type to another. To illustrate this function, one may consider the process of annotating sequences in a fasta or gff file with an annotation file in the dngf format that has already been saved. To enhance user convenience, the tool offers various options for annotating or extracting only annotations from a selection of sequences.

The **DeNoFo-Comparator** is a tool designed for the purpose of comparing two studies or datasets that have been annotated in the DeNoFo format. This tool facilitates the loading of two different DeNoFo annotations and the generation of a report that either highlights the similarities or differences between the two studies or datasets. The report is presented in an easily interpretable, human-readable formatting.

The help function and tooltips of the GUI and CLI versions of each tool, as well as the DeNoFo user manual, contain further explanations and examples with regard to the options and functionality of the toolkit and its constituent tools.

Implementation

The DeNoFo toolkit has been developed in modern Python (version 3.10 or higher) and is released under the GPL v3 license. All GUI versions are implemented using PyQt6 (Limited, 2024). The DeNoFo format is implemented as a set of pydantic models (Colvin et al., 2025) which contain the definitions and validations for the DeNoFo format.

The pydantic models are independent of the tools with specific user interfaces. Decoupling the models from the tools allows anyone to interact with the format programmatically without using our tools, while ensuring a high level of robustness of the format. For example, any *de novo* gene detection software could implement its own routine to automatically generate a valid annotation format file, so that a user does not need to manually generate one using the DeNoFo-Questionnaire.

For easy and fast setup the DeNoFo toolkit supports installation via pip or uv and has been tested on all major platforms (Linux, Windows, MacOS).

Conclusion

The DeNoFo format and toolkit presented here provide the community with an easily accessible, unified and comparable standard for methodology annotation for *de novo* genes. The overarching objective of this initiative is to enhance the comparability, interpretability and reproducibility of data sets and findings within this new and heterogeneous field of research.

The format has been designed to be as flexible and open as possible to accommodate the diverse methodological approaches in this field, while ensuring a unified annotation and reporting with a focus on best practice. To this end, the format has been designed to allow for changes and adaptation as this field of research moves forward. In order to facilitate this, the format has been implemented in such a way that additions and changes can be made with as few compatibility issues as possible. This approach enables the incorporation of new databases or software for future studies in the format, without requiring researchers to update their existing annotations.

The community is encouraged to contribute to the format by adding new approaches, databases, software and more, as well as requesting or contributing features and functionality that would improve the accessibility, usability or comparability of the format, toolkit and datasets.

In conclusion, the unified format and toolkit for *de novo* gene annotation, developed with and for the community, will hopefully pave the way for faster progress and deeper understanding in this field of research, while being continuously adapted and improved in conjunction with innovative methods and novel tools to advance the field of *de novo* gene evolution.

Competing interests

No competing interest is declared.

Author contributions statement

AG was responsible for the conceptualisation of the format with contributions from all authors. ED was responsible for the software development and implementation of the toolkit, with feedback and testing from all authors, particularly AG, MA and LE. AG and ED jointly handled project administration. MA and AG were responsible for the visualisation of the format as a figure for the paper and designed the layout and colour scheme of the toolkit. VL and AK contributed to developing the evidence criteria and the questionnaire for evaluating de novo gene status. The original draft of the manuscript was written by ED and AG, with subsequent editing and review of the first version by MA and LE. The final manuscript was edited and reviewed by all authors. Erich Bornberg-Bauer provided general administrative support and acquisition of the financial support for the project leading to this publication.

Acknowledgments

We thank Erich Bornberg-Bauer (EBB), Nikolaos Vakirlis, Li Zhao, Anne Lopes, Bharat Ravi Iyengar and Andreas Lange for their useful feedback during the planning phase of the project. ED was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 503348080. This grant with the additional grant number BO2544-22-1 was awarded to EBB. MA

received funding from the Volkswagen foundation with grant code 98183 awarded to EBB. AG was supported by the Deutsche Forschungsgemeinschaft priority program "Genomic Basis of Evolutionary Innovations" (SPP 2349) BO 2544/20-1 to EBB, and by the (HFSP) Postdoctoral Fellowship (Grant No. 2023-981550) awarded to EBB, Anne-Ruxandra Carvunis and Christine Brun. LAE has been supported by EMBO Scientific Exchange Grant 10944. VL was supported by NIH grant R01NS095654 (to Nenad Sestan).

For Figure 1 the following icons were taken from `https://bioicons.com/`: brain-1 icon (colour changed), healthy-lung icon and intestine icon by Servier `https://smart.servier.com/` licensed under CC-BY 3.0 simple_DNA_backbone icon (colour changed) by Marnie-Maddock `https://github.com/MarnieMaddock` licensed under CC-BY 4.0 Riparian_plots icon by Chenxin-Li `https://github.com/cxli233` licensed under CC-BY 4.0 Protein_monochrome icon (colour changed) by DBCLS `https://togotv.dbcls.jp/en/pics.html` licensed under CC-BY 4.0

References

Aldrovandi, S., Fajardo Castro, J., Ullrich, K., Karger, A., Luria, V., and Tautz, D. Expression of random sequences and de novo evolved genes from the mouse in human cells reveals functional diversity and specificity. *Genome Biology and Evolution*, 16 (12):evae175, 2024.

Arendsee, Z., Li, J., Singh, U., Bhandary, P., Seetharam, A., and Wurtele, E. S. Fagin: synteny-based phylostratigraphy and finer classification of young genes. *BMC bioinformatics*, 20(1):1–14, 2019.

Aubel, M., Eicholt, L., and Bornberg-Bauer, E. Assessing structure and disorder prediction tools for de novo emerged proteins in the age of machine learning. *F1000Research*, 12, 2023.

Baalsrud, H. T., Tørresen, O. K., Solbakken, M. H., Salzburger, W., Hanel, R., Jakobsen, K. S., and Jentoft, S. De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. *Molecular biology and evolution*, 35(3):593–606, 2018.

Begun, D. J., Lindfors, H. A., Kern, A. D., and Jones, C. D. Evidence for de novo evolution of testis-expressed genes in the drosophila yakuba/drosophila erecta clade. *Genetics*, 176(2): 1131–1137, 2007.

Broeils, L. A., Ruiz-Orera, J., Snel, B., Hubner, N., and van Heesch, S. Evolution and implications of de novo genes in humans. *Nature ecology & evolution*, 7(6):804–815, 2023.

Cai, J., Zhao, R., Jiang, H., and Wang, W. De novo origination of a new protein-coding gene in saccharomyces cerevisiae. *Genetics*, 179(1):487–496, 2008.

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., et al. Proto-genes and de novo gene birth. *Nature*, 487(7407):370–374, 2012.

Casola, C. From de novo to "de nono": the majority of novel protein-coding genes identified with phylostratigraphy are old genes or recent duplicates. *Genome Biology and Evolution*, 10 (11):2906–2918, 2018.

Casola, C., Owoyemi, A., Pepper, A. E., and Ioerger, T. R. Accurate identification of de novo genes in plant genomes using machine learning algorithms. *bioRxiv*, pages 2022–11, 2022.

Colvin, S., Jolibois, E., Ramezani, H., Garcia Badaracco, A., Dorsey, T., Montague, D., Matveenko, S., Trylesinski, M., Runkle, S., Hewitt, D., Hall, A., and Plot, V. Pydantic, January 2025. URL `https://github.com/pydantic/pydantic`.

Dowling, D., Schmitz, J. F., and Bornberg-Bauer, E. Stochastic gain and loss of novel transcribed open reading frames in the human lineage. *Genome biology and evolution*, 12(11):2183–2195, 2020.

Grandchamp, A., Kühl, L., Lebherz, M., Brüggemann, K., Parsch, J., and Bornberg-Bauer, E. Population genomics reveals mechanisms and dynamics of de novo expressed open reading frame emergence in drosophila melanogaster. *Genome Research*, 33(6):872–890, 2023.

Grandchamp, A., Aubel, M., Eicholt, L. A., Roginski, P., Luria, V., Karger, A., and Dohmen, E. De novo gene emergence: Summary, classification, and challenges of current methods. *EcoEvoRxiv*, 2025. doi: 10.32942/X2DP88. URL `https://ecoevorxiv.org/repository/view/8918/`.

Gubala, A. M., Schmitz, J. F., Kearns, M. J., Vinh, T. T., Bornberg-Bauer, E., Wolfner, M. F., and Findlay, G. D. The goddard and saturn genes are essential for drosophila male fertility and may have arisen de novo. *Molecular biology and evolution*, 34(5):1066–1082, 2017.

Heames, B., Schmitz, J., and Bornberg-Bauer, E. A continuum of evolving de novo genes drives protein-coding novelty in drosophila. *Journal of molecular evolution*, 88(4):382–398, 2020.

Iyengar, B. R. and Bornberg-Bauer, E. Neutral models of de novo gene emergence suggest that gene evolution has a preferred trajectory. *Molecular Biology and Evolution*, 40(4):msad079, 2023.

Keeling, D. M., Garza, P., Nartey, C. M., and Carvunis, A.-R. The meanings of'function'in biology and the problematic case of de novo gene emergence. *Elife*, 8:e47014, 2019.

Knowles, D. G. and McLysaght, A. Recent de novo origin of human protein-coding genes. *Genome research*, 19(10):1752–1759, 2009.

Lange, A., Patel, P. H., Heames, B., Damry, A. M., Saenger, T., Jackson, C. J., Findlay, G. D., and Bornberg-Bauer, E. Structural and functional characterization of a putative de novo gene in drosophila. *Nature communications*, 12(1):1667, 2021.

Li, D., Dong, Y., Jiang, Y., Jiang, H., Cai, J., and Wang, W. A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell research*, 20(4):408–420, 2010.

Limited, R. C. Pyqt6: Python bindings for the qt application framework, 2024. URL `https://www.riverbankcomputing.com/software/pyqt/`. Version 6.8.1.

Long, M., Betrán, E., Thornton, K., and Wang, W. The origin of new genes: glimpses from the young and old. *Nature Reviews Genetics*, 4(11):865–875, 2003.

McLysaght, A. and Hurst, L. D. Open questions in the study of de novo genes: what, how and why. *Nature Reviews Genetics*, 17(9):567–578, 2016.

Parikh, S. B., Houghton, C., Van Oss, S. B., Wacholder, A., and Carvunis, A.-R. Origins, evolution, and physiological implications of de novo genes in yeast. *Yeast*, 39(9):471–481, 2022.

Peng, J. and Zhao, L. The origin and structural evolution of de novo genes in drosophila. *Nature Communications*, 15(1):810, 2024.

Pereira, A. B., Marano, M., Bathala, R., Zaragoza, R. A., Neira, A., Samano, A., Owoyemi, A., and Casola, C. Orphan genes are not a distinct biological entity. *BioEssays*, page 2400146, 2024.

Rich, A. and Carvunis, A.-R. De novo gene increases brain size. *Nature Ecology & Evolution*, 7(2):180–181, 2023.

Rivard, E. L., Ludwig, A. G., Patel, P. H., Grandchamp, A., Arnold, S. E., Berger, A., Scott, E. M., Kelly, B. J., Mascha, G. C., Bornberg-Bauer, E., et al. A putative de novo evolved gene required for spermatid chromatin condensation in drosophila melanogaster. *PLoS genetics*, 17(9):e1009787, 2021.

Roginski, P., Grandchamp, A., Quignot, C., and Lopes, A. De novo emerged gene search in eukaryotes with dense. *Genome Biology and Evolution*, 16(8):evae159, 2024.

Schmitz, J. F., Ullrich, K. K., and Bornberg-Bauer, E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nature ecology & evolution*, 2(10): 1626–1632, 2018.

Tautz, D. and Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12(10):692–702, 2011.

Vakirlis, N. and Kupczok, A. Large-scale investigation of species-specific orphan genes in the human gut microbiome elucidates their evolutionary origins. *Genome Research*, 34(6):888–903, 2024.

Vakirlis, N., Carvunis, A.-R., and McLysaght, A. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *elife*, 9:e53500, 2020.

Vakirlis, N., Vance, Z., Duggan, K. M., and McLysaght, A. De novo birth of functional microproteins in the human lineage. *Cell reports*, 41(12), 2022.

Van Oss, S. B. and Carvunis, A.-R. De novo gene birth. *PLoS genetics*, 15(5):e1008160, 2019.

Weisman, C. M. The Origins and Functions of De Novo Genes: Against All Odds? *Journal of Molecular Evolution*, 90(3):244–257, August 2022. ISSN 1432-1432. doi: 10.1007/s00239-022-10055-3. URL `https://doi.org/10.1007/s00239-022-10055-3`.

Weisman, C. M., Murray, A. W., and Eddy, S. R. Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Current Biology*, 32(12):2632–2639, 2022.

Wissler, L., Gadau, J., Simola, D. F., Helmkampf, M., and Bornberg-Bauer, E. Mechanisms and dynamics of orphan gene emergence in insect genomes. *Genome biology and evolution*, 5 (2):439–455, 2013.

Zhang, L., Ren, Y., Yang, T., Li, G., Chen, J., Gschwend, A. R., Yu, Y., Hou, G., Zi, J., Zhou, R., et al. Rapid evolution of protein diversity by de novo origination in oryza. *Nature ecology & evolution*, 3(4):679–690, 2019.

Zhao, L., Svetec, N., and Begun, D. J. De novo genes. *Annual Review of Genetics*, 58, 2024.

Zheng, E. B. and Zhao, L. Protein evidence of unannotated orfs in drosophila reveals diversity in the evolution and properties of young proteins. *Elife*, 11:e78772, 2022.