Bioinformatics, 2025, 0–0
DOI/10.5281/zenodo.XXXXXXX
Advance Access Publication Date: 28 July 2025
Bioinformatics

Report

HMM-Based Detection of Kunitz Domains from Structure-Derived Alignments

Kianoush Keshani[1,*]

[1]Department of Pharmacy and Biotechnology, Alma Mater Studiorum – Università di Bologna, Via F. Selmi 3 - 4th Floor, Room 183.

[*]To whom correspondence should be addressed.

Abstract

Motivation:
Kunitz domains are structurally conserved protease inhibitors with low sequence identity, complicating their identification by conventional alignment methods. This study aims to build a sensitive and reproducible Hidden Markov Model (HMM) for Kunitz domain detection based on structure-informed multiple sequence alignment.

Results:
We developed a computational pipeline integrating PDBeFold structural alignment, TM-align validation, and HMMER-based profile construction. Evaluation against curated datasets showed high sensitivity and specificity, with an optimal E-value threshold of 1e-6 yielding 0.989 precision, 1.0 recall, and 0.994 F1-score. Four false negatives were identified, likely due to atypical domain positioning or multidomain architectures, suggesting potential refinements. The workflow outperformed sequence-based models in remote homology detection and is fully reproducible.

Availability:
All scripts, evaluation metrics, and environment files are available at https://github.com/kianinsilico/HMM_Kunitz_domain .

Contact:
kianoush.keshani@studio.unibo.it

Supplementary information:
Supplementary tables and figures are available online.

1. Introduction

The Kunitz domain is a small protein motif predominantly associated with serine protease inhibition and is implicated in diverse physiological and pathological processes, including coagulation cascades, inflammation, and cancer progression. These domains are relatively small (50–60 amino acids) and characterized by a disulfide-rich α+β fold stabilized by three highly conserved disulfide bonds (Ranasinghe and McManus, 2013). Most Kunitz-containing sequences belong to the MEROPS inhibitor family I2, clan IB, where they primarily inhibit S1 family proteases and are largely restricted to metazoans, with few known exceptions (Rawlings et al., 2004).

Several Kunitz-type protease inhibitors have been extensively characterized. Aprotinin (bovine pancreatic trypsin inhibitor, BPTI), isolated independently by Kraut et al. (1930) and Kunitz and Northrop (1936), is a classic example, comprising 58 amino acids and approximately 6.5 kDa in mass (Ascenzi et al., 2003). Other biologically significant examples include the Kunitz domain in the Alzheimer's amyloid precursor protein (APP), which acts as protease nexin-II when secreted (Oltersdorf et al., 1989), and tissue factor pathway inhibitor (TFPI), which features three tandem Kunitz domains that modulate blood coagulation by inhibiting factor VIIa and factor Xa (Bajaj et al., 2001).

Despite their structural and functional conservation, Kunitz domains often display low sequence identity, making them difficult to detect using traditional sequence alignment methods. To address this, profile Hidden Markov Models (HMMs) have proven effective in modeling conserved domains across evolutionarily distant sequences. These probabilistic models capture evolutionary patterns by encoding position-specific residue conservation and indel probabilities derived from multiple sequence alignments (Eddy, 2011). Compared to traditional sequence profiles, profile HMMs offer superior performance in detecting remote homologs and generating high-quality alignments (Mishra, 2020).

The objective of this study was to construct a robust, structure-informed HMM profile capable of accurately identifying Kunitz-type domains in protein sequences. By integrating structural alignment, clustering, model construction, and performance evaluation within a reproducible computational workflow, this work aims to contribute a reliable and generalizable tool for protein domain annotation. All scripts, software specifications, and data handling protocols are available at https://github.com/kianinsilico/HMM_Kunitz_domain , enabling reproducibility and community-based validation.

2. Methods

2.1 Computational Environment

To ensure reproducibility and consistent execution across platforms, a dedicated software environment was created using the conda package manager. Conda has gained wide adoption for managing cross-language software dependencies and ensuring platform-independent environments (Grüning et al., 2018). All analyses were performed within a custom bioinformatics environment, which included the following tools:

- HMMER (v3.3.2) – Used for HMM profile construction and sequence scanning via hmmbuild and hmmsearch. HMMER employs profile HMMs to model the evolution of sequence families with high sensitivity (Eddy, 2011).

- MMseqs2 (release 13-45111) – Used for clustering protein sequences based on sequence identity. MMseqs2 supports scalable, accurate clustering and demonstrates near-linear runtime with high sensitivity, outperforming traditional tools like CD-HIT (Steinegger and Söding, 2017).

- TM-align – Applied for pairwise structural alignment and TM-score evaluation. TM-align offers superior accuracy and speed compared to other structural alignment tools such as CE and DALI (Zhang and Skolnick, 2005).

- UCSF ChimeraX (v1.6) – Used for structural visualization, domain superposition, and manual inspection. ChimeraX offers advanced rendering capabilities and interactive tools for 3D molecular analysis (Goddard et al., 2018; Pettersen et al., 2021).

- Python (v3.10) with biopython, pandas, numpy, and matplotlib – Used for data parsing, result processing, and evaluation script development.

The complete environment specification file is available in the project repository (https://github.com/kianinsilico/HMM_Kunitz_domain).

2.2 Data Acquisition and Clustering

Protein structures annotated with Kunitz domains were retrieved from the Protein Data Bank (PDB) through domain- and keyword-based queries. To reduce sequence redundancy while preserving structural diversity, the sequences were clustered using MMseqs2 at a 90% identity threshold. MMseqs2 was selected over CD-HIT due to its improved scalability, ability to handle short domain-like sequences efficiently, and more flexible clustering parameters. It also achieves BLAST-like sensitivity at a fraction of the runtime (Steinegger and Söding, 2017).

2.3 Structural Alignment and Curation

Cluster representatives were aligned using PDBeFold, a structure-based alignment tool employing the Secondary Structure Matching (SSM) algorithm to identify residues in equivalent spatial positions (Krissinel and Henrick, 2004). Structure-based alignment was preferred over sequence-based alternatives to better capture conserved 3D motifs characteristic of the Kunitz fold.

To ensure structural homogeneity, TM-align was used to calculate pairwise TM-scores across the aligned structures. TM-scores above 0.5 were used as a threshold to confirm that all included sequences shared the same fold, consistent with standard structural classification practices (Zhang and Skolnick, 2005). Proteins with poor TM-scores or incompatible topologies were excluded.

Further manual inspection and refinement were carried out in UCSF ChimeraX, enabling visual verification of structural alignment, assessment of disulfide bonding, and detection of potential anomalies. ChimeraX also facilitated the annotation of functional residues and visualization of conserved structural elements (Pettersen et al., 2021).

2.4 Profile HMM Construction

A final curated multiple structure-based alignment served as input to hmmbuild from the HMMER suite. This tool constructs probabilistic models from aligned sequences, capturing both residue conservation and indel distributions (Eddy, 2011). The resulting profile HMM is tailored to reflect the structural and functional constraints of the Kunitz domain and is intended to improve remote homology detection beyond sequence-derived models.

2.5 Validation Datasets and Performance Assessment

The constructed HMM profile was evaluated against four distinct datasets designed to assess both sensitivity and specificity:

1) Human proteins containing annotated Kunitz domains.
2) Human proteins without Kunitz annotations.
3) Non-human Kunitz domain-containing proteins.
4) A representative background set from UniProt/Swiss-Prot.

All datasets were scanned using hmmsearch, and output hits were analyzed using a custom Python script. Evaluation metrics included:

- Precision: proportion of correctly predicted positives over all predicted positives.

- Recall: proportion of correctly predicted positives over all actual positives.

- F1-score: harmonic mean of precision and recall, offering a balanced performance indicator (Powers, 2011; Zheng et al., 2020).

- Accuracy: overall correct predictions over all instances.

These metrics allowed robust performance benchmarking across datasets and informed subsequent refinement.

2.6 Reproducibility and Workflow Availability

All steps in the computational pipeline—from data acquisition and preprocessing to model construction and validation—were implemented using modular shell and Python scripts. The approach adheres to the five pillars of computational reproducibility: literate programming, version control, environment standardization, persistent data access, and clear documentation (Ziemann et al., 2023). Full documentation, scripts, and environment files are hosted in the project's public repository at https://github.com/kianinsilico/HMM_Kunitz_domain , enabling transparent reproduction and extension of this work.

3. Results and Discussion

3.1 Structural Alignment and Consistency Analysis

The structural multiple sequence alignment (MSA) generated via PDBeFold successfully preserved core features of the Kunitz domain, including the conserved disulfide bonding pattern and the compact α+β fold. Visual inspection using ChimeraX confirmed accurate spatial superposition of the selected structures. Structural consistency across aligned domains was quantified using three key metrics: pairwise TM-scores, RMSD values, and sequence identity.

A pairwise TM-score matrix revealed a consistent fold across the dataset, with most scores exceeding 0.5—indicating strong topological similarity (Zhang and Skolnick, 2005). The heatmap visualizations provide an overview of these relationships:

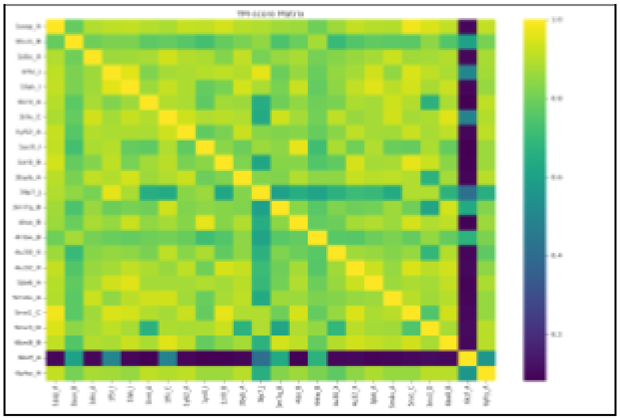

Figure 1 | *Pairwise TM-scores between Kunitz domain representatives, highlighting structural similarity among the dataset.*

In parallel, RMSD values supported these findings, with most aligned domains showing RMSD values below 3 Å:

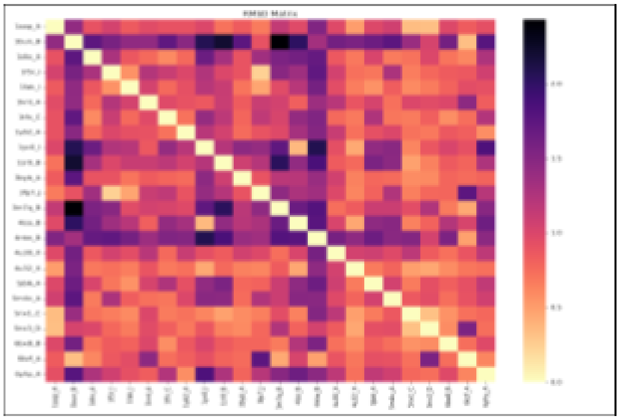

Figure 2 | Heatmap of pairwise RMSD values, indicating low structural deviation across aligned Kunitz domains.

PDBeFold analysis revealed, sequence identity, although generally low (as expected for remote homologs), was sufficient to cluster functionally similar variants, and further justifies the choice of structure-based alignment over sequence-based methods:

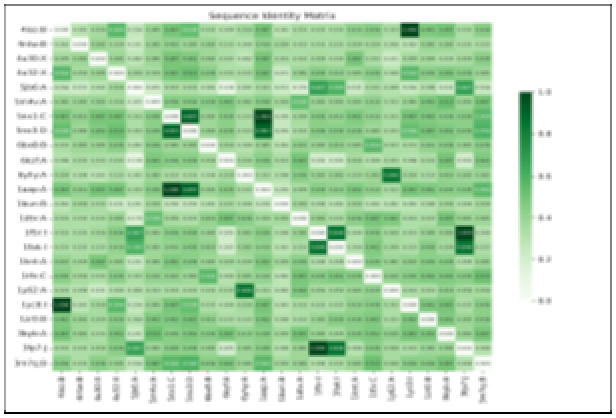

Figure 3a | Matrix of pairwise sequence identity among selected Kunitz domain sequences.

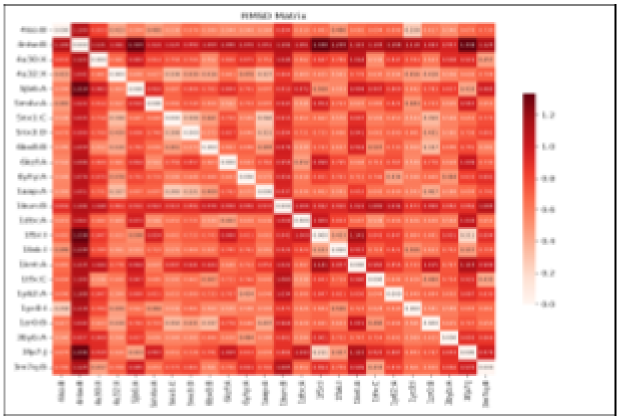

Figure 3b | Matrix of pairwise RMSD among selected Kunitz domain sequences.

All raw values and matrices are provided in supplementary files (Supplementary Table S1).

3.2 HMM Model Performance on Known Structures

To test whether the HMM profile captured the structural features used during training, it was applied to the original structural dataset. All structures were correctly recognized with extremely low E-values and high alignment scores (Supplementary Table S1). These results confirm that the model encapsulates the conserved core of the Kunitz fold while tolerating peripheral variation. Full model alignment scores and domain annotations are reported in Supplementary (Supplementary Table S2).

3.3 HMM Generalization: Independent Dataset Assessment

To evaluate the model's generalization capability, we constructed a positive validation set by combining annotated Kunitz domain sequences from both human and non-human sources. This ensured phylogenetic diversity and included various domain architectures, such as tandem Kunitz repeats and proteins with Kunitz domains in different structural contexts.

This positive set was tested against two negative/background datasets: one composed of human proteins lacking any Kunitz domain annotation, and another consisting of a broad, taxonomically diverse collection of curated protein sequences from UniProt/Swiss-Prot. These negative sets were used to evaluate the specificity and robustness of the model in realistic scenarios.

The profile HMM successfully identified the majority of known Kunitz sequences with high confidence. Notably, distant homologs such as *6BX8_B* and *4NTW_B* were correctly detected with E-values of 9.5e-26 and 3.5e-24, respectively. No high-confidence false positives were observed in the negative sets, indicating strong specificity and minimal overfitting.

These findings demonstrate the model’s ability to detect true Kunitz domains across evolutionary distances and varied protein contexts, supporting its application to large-scale proteome annotation tasks.

3.4 Threshold Optimization and Benchmarking

A comprehensive benchmark was conducted using four validation datasets: human Kunitz, human non-Kunitz, non-human Kunitz, and general UniProt/Swiss-Prot sequences. Performance metrics were calculated using both full-sequence and best-domain E-values at multiple thresholds.

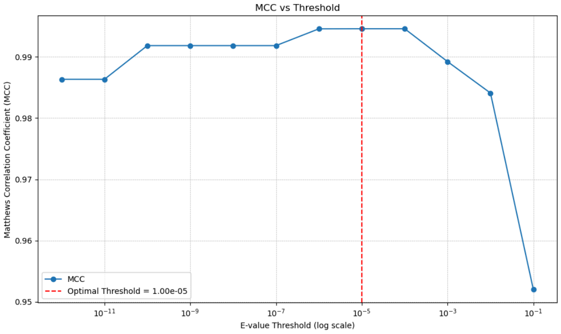

Figure 4a | MCC as a function of the E-value threshold.

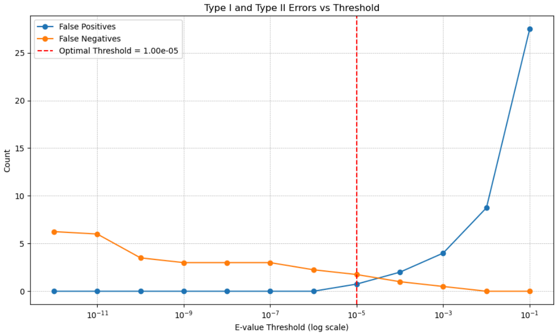

Figure 4b | Type I and Type II errors as a function of the E-value.

The optimal trade-off was observed at an E-value threshold of 1e-06, using full-sequence scores, yielding the following metrics:

- Precision: 0.989
- Recall: 1.0
- F1-score: 0.994
- Matthews Correlation Coefficient (MCC): 0.995
- Accuracy (Q2): >0.9999

These results confirm excellent sensitivity and specificity across diverse inputs. The performance plateaued between 1e-06 and 1e-09, indicating that domain detection remained stable within this confidence range.

A complete benchmark dataset, including thresholds and metrics, is available in supplementary files (Supplementary Table S3).

3.5 Analysis of False Negatives

While the trained HMM accurately identified the vast majority of Kunitz-containing proteins at the 1e-6 threshold, a total of four sequences were classified as false negatives. Manual inspection confirmed that all four sequences possess at least one annotated PF00014 (Kunitz_BPTI) domain according to UniProt. However, they were not detected by the profile HMM.

This discrepancy is likely attributable to atypical domain contexts—such as domains near the C-terminus, tandem domain architectures, or short protein lengths—that deviate from the structural patterns captured in the training alignment. These findings highlight a limitation of the model: reduced sensitivity to less canonical domain arrangements or multidomain proteins.

TABLE 1 | False negatives at *1e-06* E-value threshold and their domain annotation

UniProt ID|	Length	Domains	Domain Position(s)	Comments
1A0A1Q1NL17	101	1	32-88	Short sequence
O62247	202	1	138-184	Domain near C-terminal
Q8WPG5	134	2	17-69, 83-129	Tandem domains
D3GGZ8	195	1	120-190	Domain near C-terminal

3.6 Limitations and Future Directions

While the HMM profile demonstrated high accuracy in identifying Kunitz domains across diverse datasets, several limitations should be acknowledged. First, the training dataset was derived exclusively from structurally resolved sequences in the Protein Data Bank (PDB), which introduces potential bias toward well-studied and canonical Kunitz domain variants. As a result, the model may not fully capture rare, divergent, or non-canonical forms present in less-characterized organisms.

Second, while structural alignment via PDBeFold was crucial for capturing conserved spatial motifs, the current workflow does not include a direct comparison with models built from sequence-based MSAs (e.g., using MAFFT or Clustal Omega). Future work should address this gap by constructing a parallel HMM from a high-quality sequence-based alignment and systematically comparing the resulting models in terms of precision, recall, and generalizability. This would provide insights into the added value of structure-informed alignment and quantify the trade-offs between sensitivity and computational cost.

Finally, expanding the validation to include predicted structures (e.g., AlphaFoldDB entries) or unannotated proteomes from non-metazoan taxa could test the limits of the model's generalizability and uncover novel domain variants not yet described in current databases.

4. Conclusions

In this study, we developed and validated a structure-informed Hidden Markov Model (HMM) for the detection of Kunitz-type domains in protein sequences. By leveraging structure-based multiple sequence alignment via PDBeFold and incorporating rigorous quality control with TM-align and ChimeraX, we constructed a high-confidence model that generalizes effectively across taxonomic boundaries and domain arrangements.

Benchmarking against curated datasets demonstrated excellent classification performance, achieving high precision (0.989), perfect recall (1.0), and an F1-score of 0.994 at the optimal threshold. These metrics confirm that structure-based alignment improves the model's sensitivity to functionally conserved but sequence-divergent proteins. Nonetheless, a small number of false negatives revealed limitations in detecting domains with unusual positioning or domain multiplicity, suggesting room for refinement.

The workflow—modular, reproducible, and publicly documented—provides a solid framework for domain-centric modeling efforts and may serve as a reference for future studies involving other structurally conserved protein families. Future directions include benchmarking against sequence-derived alignments and testing the model on large-scale proteomes or predicted structures.

Supplementary Information

The following figures and tables are recommended for inclusion as supplementary material:

- Supplementary Table S1: Raw structural consistency data (consistency_check.txt)

- Supplementary Table S2: Full HMMER results on training set (final_assessment_set1.txt)
- Supplementary Table S3: Benchmarking thresholds and performance metrics (final_assessment_set2.txt)
- Supplementary Figure S1: MSA preview (pdb_kunitz_PDBeFold_alignment_cleaned.fasta.png)
- Supplementary Figure S2: Raw MSA before cleaning (pdb_kunitz_PDBeFold_alignment.fasta.png)
- Supplementary Figure S3: Structural superposition (kunitz_superposition.png)
- Supplementary Figure S4: Q-score matrix (Q-score_matrix.png)

Each supplementary element is referenced at the appropriate point in the text.

Acknowledgements

I would like to express my sincere gratitude to Professor Emidio Capriotti for his guidance, constructive feedback, and continuous support throughout the development of this project. I am also thankful to the University of Bologna and the Department of Pharmacy and Biotechnology for providing the academic environment and resources necessary to complete this work.

Funding

This work was supported in part by the ER.GO Regional Authority for the Right to Higher Education in Emilia-Romagna, through an international student scholarship. Their financial assistance is gratefully acknowledged.

Conflict of Interest: none declared.

References

Ascenzi, P., Bocedi, A., Bolognesi, M., Spallarossa, A., Coletta, M., De Cristofaro, R., & Menegatti, E. (2003). The bovine basic pancreatic trypsin inhibitor (Kunitz inhibitor): a milestone protein. Current Protein & Peptide Science, 4(3), 231-251.

Bajaj, M. S., Birktoft, J. J., Steer, S. A., & Bajaj, S. P. (2001). Structure and biology of tissue factor pathway inhibitor. Thrombosis and Haemostasis, 86(4), 959-972.

Eddy, S. R. (2011). Accelerated profile HMM searches. PLoS Computational Biology, 7(10), e1002195.

Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H., & Ferrin, T. E. (2018). UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein Science, 27(1), 14-25.

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., ... & Köster, J. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. Nature Methods, 15(7), 475-476.

Krissinel, E., & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallographica Section D, 60(12), 2256-2268.

Kunitz, M., & Northrop, J. H. (1936). Isolation from beef pancreas of crystalline trypsinogen, trypsin, a trypsin inhibitor, and an inhibitor-trypsin compound. The Journal of General Physiology, 19(6), 991-1007.

Kraut, H., Frey, E. K., & Werle, E. (1930). Der Nachweis eines Kreislauf-Hormons in der Pankreasdrüse. IV. Mitteilung über dieses Kreislauf-Hormon. Hoppe-Seyler's Zeitschrift für Physiologische Chemie, 189(1-3), 97-106.

Mishra, M. (2020). Evolutionary aspects of the structural convergence and functional diversification of Kunitz-domain inhibitors. Journal of Molecular Evolution, 88(7), 537-548.

Oltersdorf, T., Fritz, L. C., Schenk, D. B., Lieberburg, I., Johnson-Wood, K. L., Beattie, E. C., ... & Sinha, S. (1989). The secreted form of the Alzheimer's amyloid precursor protein with the Kunitz domain is protease nexin-II. Nature, 341(6238), 144-147.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., ... & Ferrin, T. E. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Science, 30(1), 70-82.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37-63.

Ranasinghe, S., & McManus, D. P. (2013). Structure and function of invertebrate Kunitz serine protease inhibitors. Developmental & Comparative Immunology, 39(3), 219-227.

Rawlings, N. D., Tolle, D. P., & Barrett, A. J. (2004). Evolutionary families of peptidase inhibitors. Biochemical Journal, 378(3), 705-716.

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology, 35(11), 1026-1028.

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research, 33(7), 2302-2309.

Zheng, W., Zhou, X., Wuyun, Q., Pearce, R., Li, Y., & Zhang, Y. (2020). FUpred: detecting protein domains through deep-learning-based contact map prediction. Bioinformatics, 36(12), 3749-3757.

Ziemann, M., Poulain, P., & Bora, A. (2023). The five pillars of computational reproducibility: bioinformatics and beyond. Briefings in Bioinformatics, 24(6), bbad375.