Proteomic study for the prediction of µCT imaging with iodine

Valentin Wesp[1,*], Lisa-Marie Barf[2], and Heiko Stark[2]

[1]Department of Bioinformatics, Friedrich-Schiller-University Jena, Germany
[2]Institute of Zoology and Evolutionary Research with Phyletic Museum, Friedrich-Schiller-University Jena, Germany
[*]valentin.wesp@uni-jena.de

ABSTRACT

Iodine-based staining techniques are commonly used in histological imaging and micro-computed tomography (µCT) due to the affinity of iodine to bind to specific molecules. However, the basis for tissue-specific contrast has not yet been sufficiently explored. In this study, we analyse the human proteome and try to identify proteins with high potential for iodine binding, especially those rich in aromatic heterocyclic amino acids. Using bioinformatic methods, we assess the occurrence of aromatic/non-aromatic heterocyclic, homocyclic and the remaining 15 other amino acids in over 20,000 proteins, including isoforms. The proteins were categorised into families using curated (HGNC) and automated datasets (UniProt Similarity Index, InterPro). In particular, structural proteins such as titin, nebulin, obscurin, mucin, filaggrin, and hornerin show a high absolute number of aromatic heterocyclic amino acids. These results may explain the high µCT contrast in muscle, skin and mucosal tissues. Our results provide a general baseline for iodine-based tissue imaging and are a potential stepping stone for future research, e.g. for cross-species applications and structural and functional effects of iodination.

Introduction

Iodine has been used in histology for many years to stain tissue. The most common and cheapest form is Lugol's solution, which is a dissolved mixture of iodine and potassium iodide. It was first described by Jean Guillaume Auguste Lugol in the early 1800s[1]. In solution, iodine (I2) and potassium iodide (KI) form triiodides ($I3^+$) and potassium ions (K^-)[2]. This causes the normal, non-cornified squamous epithelium of the normal oesophagus, for example, to discolour differently from pathologically altered tissue[1,3]. This is explained by the affinity of glycogen for iodine-containing substances[4,5]. However, lipid staining is also known[6,7]. The staining mechanism for high-contrast 3d imaging is not yet fully understood[5].

However, it is known that pure iodine is an inert halogen that usually requires catalysis to form a bond[8]. These reactions include oxidations of disulfide bridges (Cys), hydrophilic and ion-ion interactions (amino & carbonyl groups). Another possibility of staining can be explained by intercalation processes (detection of starch)[9,10]. Electron-rich aromatic heterocycles (pyrrole, furan, or thiophene), on the other hand, can react very easily directly with iodine[11]. Such aromatic heterocyclic compounds are a widespread class of compounds in nature with important functions in biological processes. They include the nucleic bases of RNA and DNA, cytochromes, cobalamins, some natural colourants and odorants, as well as the amino acids histidine and tryptophan[12,13]. Although Proline is also a heterocycle, it does not follow Hückel's rule and is therefore not aromatic. As a result, it does not react spontaneously with pure iodine.

Two reviews by Metscher[5,14] show the strength of iodine staining. He compares several contrast agents (Gallocyanin-chromalum, I2E, I2M, IKI, PTA, Osmium tetroxide) for the staining of soft tissue. Iodine provides results comparable to those of the other contrast agents. However, it is preferable in terms of usability and toxicity[14]. It is also shown that iodine primarily stains yolk, skin, eyes, nerve, and muscle tissue, which are protein-rich tissues.

There is also a special analysis method, called the iodine value determination, that measures the degree of unsaturated compounds in fats[15]. This utilises the fact that mixed halogens (iodine chloride) specifically add to olefinic bonds and can be used for quality control. However, it should be noted that the iodine must not be pure; otherwise, it will not attach to the alkenes. This method also explains why some authors use iodine to stain lipids.

So why is iodine used to stain soft tissue for µCT examinations? This is because soft tissue has a very similar density and offers little contrast. Iodine helps here because of its atomic radius. The radius of a covalently bonded iodine atom is around 133 picometres, which falls within the wavelength range of X-rays: 10 to 10,000 picometres[16]. X-rays are therefore easily absorbed by the iodine atoms and are suitable for visualising iodine-dense regions.

From the above-mentioned facts, aromatic heterocycles and olefinic double bonds can be selectively stained and spatially visualised through imaging techniques. This raises the question of which metabolites or proteins are being labelled. In this study, we were particularly interested in identifying proteins that are rich in aromatic heterocycles (histidine and tryptophan)

and, therefore, expected to exhibit intense staining.

Results

Data

In our analysis, 322 protein sequences were not curated and removed from the dataset, with a median length of 93.5 amino acids compared to 415 amino acids of the 20,328 curated sequences. We categorise proteins (labelled with gene names and relative amount of aromatic heterocycles) with a high absolute and relative proportion of cycles. This shows that there are significant differences between aromatic heterocycles (histidine, tryptophan), non-aromatic heterocycles (proline) and homocycles (phenylalanine, tyrosine) compared to the remaining 15 other amino acids (Fig. 1).

Among proteins with the highest aromatic heterocyclic amino acid content, Titin (*TTN*) stands out with 944 occurrences in its 34,350-residue sequence, followed by Mucin-3B (*MUC3B*) with 533 in 13,477 residues and Filaggrin (*FILA*) with 434 in 4,061 residues (Table 1). Naturally, larger proteins tend to have higher absolute counts simply because of their size, although their proportion of heterocyclic amino acids relative to total length remains low. For example, only 2.7% of titin's residues are heterocyclic, which is below the dataset-wide average of 3.9%.

In contrast, among proteins with a relative heterocyclic amino acid content exceeding 10.0%, Histidine-rich carboxyl terminus protein 1 (*HRCT1*) has the highest proportion, with 23.4% of its 115 residues being aromatic heterocyclic (Table 2). It is followed by Proline, histidine, and glycine-rich protein 1 (*PHGR1*) and Placenta-specific protein 4 (*PLAC4*), with 18.3% and 16.0% in sequences of 82 and 150 residues. Since even a few occurrences significantly impact small proteins, they are more likely to exhibit high relative proportions despite their low absolute counts. For example, the Proline, histidine, and glycine-rich protein 1 contains only 15 aromatic heterocyclic amino acids, also below the dataset-wide average of 21.5 residues.

Families

Data was extracted from three sources to identify protein families enriched in aromatic heterocyclic amino acids: the HGNC database, the UniProt Similarity Index, and the InterPro database. The HGNC database identified the fewest clusters (1,448 families) compared to InterPro and UniProt, which classified 6,941 and 5,101 families, respectively.

The HGNC dataset shows the least fragmentation compared to the UniProt and InterPro datasets when looking at the number of proteins in each assigned family (Fig. 2). In UniProt and InterPro, families contain an average of 4.05 and 2.98 proteins, respectively. In addition, over 50% of families in both databases consist of only one protein (2,879 or 56.4% in UniProt and 4,191 or 60.4% in InterPro). In contrast, the protein families in the HGNC database are significantly more prominent, with an average of 14.26 proteins per family and only 10.2% (148) of families containing only one protein.

Gene Ontology analysis

Among the GO terms associated with the three specified keywords, the 'cellular_components' category stands out as particularly relevant to our analysis. Based on absolute values, 78 proteins exhibit a Z-score of at least three. The highest score is observed for Titin (*TTN*, $Z = 6.19$), while Apolipoprotein B-100 (*APOB*) marks the lower threshold ($Z = 3.0$). The four most frequently occurring terms are 'membrane' (46), 'cytoplasm' (43), 'plasma membrane' (39), and 'nucleus' (26) (3).

When considering relative values, 45 proteins meet the same Z-score threshold. Histidine-rich carboxyl terminus protein 1 (*HRCT1*) shows the highest score ($Z = 5.05$), whereas Acyl-CoA 6-desaturase (*FADS2*) is at the lower end ($Z = 3.01$). Naturally, this also leads to a reduced number of GO terms. In this case, only 49 unique terms were extracted, with the four most frequently occurring being 'membrane' (13), 'extracellular matrix' (8), 'endoplasmatic reticulum' (8), and 'nucleus' (8).

For the GO term analysis of the keywords 'biological_process' and 'molecular_function' see the Data availability section.

Discussion

Proteins with numerous aromatic heterocycles

Of all proteins, titin contains the largest number of aromatic heterocycles. This is unsurprising as it is the largest known human protein with a total amino acid length of 30,000. In addition, it is highly expressed in the muscles, as it is a structural protein responsible for their basic tension[17]. Other important structural proteins in the skeletal musculature are myosin (*MYH2*: 2.3%; *MYH4*: 2.5%; *MYH1*: 2.4%), actin (*ACTS*: 3.4%), nebulin (*NEBU*: 4.6%), and obscurin (*OBSCN*: 3.7%)´[18,19]. Nevertheless, a relative comparison of titin with these structural proteins shows that they are rich in aromatic heterocycles (2.7%). On the other hand, special isoforms of myosin (*MYH6*: 2.3%; *MYH7*: 2.6%) and actin (*ACTC1*: 3.4%) are found in the heart muscle (Fig. 4). The high proportion of aromatic heterocycles in the structural proteins also explains why the muscles, and in particular the heart muscles, are well stained by iodine[14,20–24].

Another structural protein family is the collagens, which are rich in proline but not in aromatic heterocycles[25–27]. Exceptions are collagen types XX (FACITs), XVIII (Endostatin precursor), XXVI (Other), VI (network-forming) and XIV (FACITs)

(*COKA1*: 3.5%; *COIA1*: 3.2%; *COQA1*: 2.7%; *CO6A5*: 2.7%; *COEA1*: 2.6%). Almost all type VI collagens (network-forming), which occur in the extracellular matrix, show a high number of aromatic heterocycles (*CO6A1* - *CO6A6*: 1.6% - 2.7%). In contrast, the most frequently occurring collagen types I (fibrillar), IV (network-forming), VIII (network-forming), IX (FACITs) and X (network-forming) contain only a few aromatic heterocycles (*CO1A1*: 1.0%; *CO1A2*: 1.5%; *CO4A5*: 1.1%; *CO8A1*: 1.6%; *COAA1*: 1.8%; *CO8A2*: 1.4%; *CO9A2*: 1.2%), with the exception of *CO4A4* (2.5%).

However, structural proteins are not the only ones with high aromatic heterocyclic content. Keratin-associated proteins (0.0%-6.8%), fillagrin (*FILA*: 10.7% and *FILA2*: 10.6%), and hornerin (*HORN*: 10.0%)[28]. They thus represent a group of iodine-sensitive proteins in the skin. This could explain the high contrast of this tissue by iodine[5].

Directly downstream in the gastrointestinal tract, the mucins (*MUC1*: 4.7%; *MUC2*: 2.1%; *MUC3A*: 2.5%; *MUC3B*: 4.0%; *MUC4*: 4.0%; *MUC5A*: 2.8%; *MUC5B*: 3.5%; *MUC6*: 4.7%; *MUC7*: 3.2%) with their high number of aromatic heterocycles represent another important group. Mucins are found as a structuring component of mucus on mucous membranes, and thus also in the gastrointestinal tract. We know from Fennerty (1999) that the oesophagus can be easily stained. This suggests that mucins play a significant role in this process. Furthermore, increased mucin production is observed in many types of cancer[29,30]. *PHGR1* (18.3%), another gene expressed in the gastrointestinal tract, tops the relative proportions of aromatic heterocycles.

The best-known reaction with iodine is found in the thyroid gland[31]. However, this is an enzymatic reaction and is based on the production of iodine-containing thyroid hormones, such as triiodothyronine (T3) and thyroxine or tetraiodothyronine (T4) and the peptide hormone calcitonin. The bound amounts of iodine are negligible (1.17 ng/dL), as these hormones only occur in very low concentrations compared to the contrast with the aromatic heterocycles.

Conclusions

In conclusion, iodine binds to proteins to varying extents, enabling the targeted staining and classification of specific tissues. Importantly, it is primarily the overall protein content within the tissue that determines staining intensity, rather than the aromatic heterocycle content of individual proteins. Nonetheless, there are notable exceptions — for instance, in skeletal and cardiac muscle tissue. Despite their similar structural organisation, the isoforms of their structural proteins differ in aromatic heterocycle composition.

Outlook

The method presented here is particularly well-suited for application to other organisms, offering a targeted approach for staining specific tissues. Future studies could also explore how iodination affects protein stability, similar to what was performed earlier for various types of non-enzymatic modification of proteins, such as bromination and chlorination[32]. Susceptibility scores can be determined for particular amino acids and then combined into overall scores for an entire protein. Notably, many top protein hits are localised in the skin and digestive tract, tissues frequently exposed to high microbial loads. Given that iodine, particularly in the form of Lugol's solution, is known for its antimicrobial properties[33], this observation provides a compelling basis for further investigation.

Methods

For our study, we retrieved the human proteome (containing 20,650 proteins) in FASTA format from the UniProt database (accessed 14 February 2025) using a custom Python script (v3.12). We chose a human dataset as a proof of concept since it is highly curated and relevant. We first removed all non-curated sequences and then analysed the remaining ones for their aromatic and non-aromatic heterocyclic, homocyclic, and remaining canonical amino acids content, all in absolute and relative terms. Absolute values potentially indicate the total amount of iodine that can bind to a protein. In contrast, relative values highlight proteins with a high proportion of heterocyclic and homocyclic amino acids, independent of sequence length.

To identify proteins with an over-representation of these amino acid groups compared to the rest, we calculated the Z-score, which measures how many standard deviations a value deviates from the mean[34]. This means that a Z-score of one is one standard deviation from the mean, two are two standard deviations, three are three standard deviations (typically defined as outliers), and so on. Since a Z-score assumes a normal distribution, the data was first normalised using the Yeo-Johnson method[35]. Proteins were ranked based on their absolute and relative values and their Z-scores to emphasise those with particularly high proportions of aromatic heterocycles, non-aromatic heterocycles, homocycles, and all other amino acids.

Gene ontology (GO) terms were retrieved from QuickGO for all proteins with an aromatic heterocyclic Z-score of at least three based on the keywords 'biological_process', 'cellular_component' and 'molecular_function'[36]. Word clouds were generated based on their frequencies for the visualisation of the GO terms.

Lastly, proteins were grouped into families and ranked based on the averages of their absolute and relative values to provide a more comprehensive overview. This classification was conducted using three sources: the HUGO Gene Nomenclature Committee (HGNC), the Similarity Index from UniProt, and the InterPro database[37–39]. InterPro integrates protein signatures from multiple databases to categorise proteins based on predicted domains. Similarly, UniProt employs the Similarity Index,

which reflects sequence similarity between proteins. Consequently, the datasets from InterPro and UniProt, which are automatically annotated, exhibit more significant fragmentation than HGNC, which is highly curated.

In addition, all available isoforms of the human proteome (84,854 proteins) were also analysed using the mentioned workflow (see Data availability).

References

1. Fennerty, M. B. Tissue staining (chromoscopy) of the gastrointestinal tract. *Can. J. Gastroenterol.* **13**, 423–429, DOI: 10.1155/1999/763162 (1999).
2. Degenhardt, K., Wright, A. C., Horng, D., Padmanabhan, A. & Epstein, J. A. Rapid 3D phenotyping of cardiovascular development in mouse embryos by micro-CT with iodine staining. *Circ. Cardiovasc. Imaging* **3**, 314–322, DOI: 10.1161/CIRCIMAGING.109.918482 (2010).
3. Canto, M. I. Staining in gastrointestinal endoscopy: The basics. *Endoscopy* **31**, 479–486, DOI: 10.1055/s-1999-8041 (1999).
4. Bock, W. J. & Shear, C. R. A staining method for gross dissection of vertebrate muscles. *Anat. Anzeiger* **130**, 222–227 (1972).
5. Metscher, B. D. Micro CT for comparative morphology: Simple staining methods allow high-contrast 3D imaging of diverse non-mineralized animal tissues. *BMC Physiol.* **9**, DOI: 10.1186/1472-6793-9-11 (2009).
6. Gignac, P. M. & Kley, N. J. Iodine-enhanced micro-CT imaging: Methodological refinements for the study of the soft-tissue anatomy of post-embryonic vertebrates. *J. Exp. Zool. Part B: Mol. Dev. Evol.* **322**, 166–176, DOI: 10.1002/jez.b.22561 (2014).
7. Gignac, P. M. *et al.* Diffusible iodine-based contrast-enhanced computed tomography (diceCT): An emerging tool for rapid, high-resolution, 3-D imaging of metazoan soft tissues. *J. Anat.* **228**, 889–909, DOI: 10.1111/joa.12449 (2016).
8. Ajvazi, N. & Stavber, S. Electrophilic Iodination of Organic Compounds Using Elemental Iodine or Iodides: Recent Advances 2008–2021: Part I. *Compounds* **2**, 3–24, DOI: 10.3390/compounds2010002 (2022).
9. Pesek, S., Lehene, M., Brânzanic, A. M. V. & Silaghi-Dumitrescu, R. On the Origin of the Blue Color in The Iodine/Iodide/Starch Supramolecular Complex. *Mol. (Basel, Switzerland)* **27**, DOI: 10.3390/molecules27248974 (2022).
10. Pesek, S. & Silaghi-Dumitrescu, R. The Iodine/Iodide/Starch Supramolecular Complex. *Mol. (Basel, Switzerland)* **29**, DOI: 10.3390/molecules29030641 (2024).
11. Stavber, S., Jereb, M. & Zupan, M. Electrophilic iodination of organic compounds using elemental iodine or iodides. *Synthesis* 1487–1513, DOI: 10.1055/s-2008-1067037 (2008).
12. Vollhardt, K., Peter, C. & Schore Neil, E. *Organic Chemistry* (Freeman & Company, 2007), 5th editio edn.
13. Schuster, S. & Malycheva, T. Enumeration of saturated and unsaturated substituted n-heterocycles (2025). 2309.02343.
14. Metscher, B. D. MicroCT for developmental biology: A versatile tool for high-contrast 3D imaging at histological resolutions. *Dev. Dyn.* **238**, 632–640, DOI: 10.1002/dvdy.21857 (2009).
15. Thomas, A. Fats and Fatty Oils. In *Ullmann's Encyclopedia of Industrial Chemistry*, DOI: 10.1002/14356007.a10_173 (Wiley, 2000).
16. Dean, J. A. & Lange, N. A. *Lange's handbook of chemistry* (McGraw-Hill, 1999).
17. Freundt, J. K. & Linke, W. A. Titin as a force-generating muscle protein under regulatory control. *J. Appl. Physiol.* **126**, 1474–1482, DOI: 10.1152/japplphysiol.00865.2018 (2019).
18. Holmes, K. C. Myosin Structure. *Myosins* 35–54, DOI: 10.1007/978-1-4020-6519-4_2 (2007).
19. Dominguez, R. & Holmes, K. C. Actin Structure and Function. *Annu. Rev. Biophys.* **40**, 169–186, DOI: 10.1146/annurev-biophys-042910-155359 (2011).
20. Jeffery, N. S., Stephenson, R. S., Gallagher, J. A., Jarvis, J. C. & Cox, P. G. Micro-computed tomography with iodine staining resolves the arrangement of muscle fibres. *J. Biomech.* **44**, 189–192, DOI: 10.1016/j.jbiomech.2010.08.027 (2011).
21. Kupczik, K. *et al.* Reconstruction of muscle fascicle architecture from iodine-enhanced microCT images: A combined texture mapping and streamline approach. *J. Theor. Biol.* **382**, 34–43, DOI: 10.1016/j.jtbi.2015.06.034 (2015).
22. Dickinson, E., Stark, H. & Kupczik, K. Non-Destructive Determination of Muscle Architectural Variables Through the Use of DiceCT. *Anat. Rec.* **301**, 363–377, DOI: 10.1002/ar.23716 (2018).

23. Nyakatura, J., Baumgarten, R., Baum, D., Stark, H. & Youlatos, D. Muscle internal structure revealed by contrast-enhanced μCT and fibre recognition: The hindlimb extensors of an arboreal and a fossorial squirrel. *Mammalian Biol.* **99**, DOI: 10.1016/j.mambio.2019.10.007 (2019).

24. Kolmann, M. A. *et al.* DiceCT for fishes: recommendations for pairing iodine contrast agents with μCT to visualize soft tissues in fishes. *J. Fish Biol.* **102**, 893–903, DOI: 10.1111/jfb.15320 (2023).

25. Mayne, R. & Brewton, R. G. New members of the collagen superfamily. *Curr. Opin. Cell Biol.* **5**, 883–890, DOI: 10.1016/0955-0674(93)90039-S (1993).

26. Ricard-Blum, S. The Collagen Family. *Cold Spring Harb. Perspectives Biol.* **3**, 1–19, DOI: 10.1101/cshperspect.a004978 (2011).

27. Wesp, V., Scholz, L., Ziermann-Canabarro, J. M., Schuster, S. & Stark, H. Constructing networks for comparison of collagen types. *J. Integr. Bioinforma.* **21**, 20240020, DOI: 10.1515/jib-2024-0020 (2024).

28. Henry, J. *et al.* Hornerin is a component of the epidermal cornified cell envelopes. *The FASEB J.* **25**, 1567–1576 (2011).

29. Singh, R. & Bandyopadhyay, D. MUC1: a target molecule for cancer therapy. *Cancer biology & therapy* **6**, 481–6, DOI: 10.4161/cbt.6.4.4201 (2007).

30. Singh, A. P. *et al.* Clinical potential of mucins in diagnosis, prognosis, and therapy of ovarian cancer. *The Lancet Oncol.* **9**, 1076–1085, DOI: 10.1016/S1470-2045(08)70277-8 (2008).

31. Colledge, N. R., Walker, B. R., Ralston, S. & Davidson, L. S. P. *Davidson's principles and practice of medicine* (Elsevier Health Sciences, 2014).

32. Fichtner, M., Schuster, S. & Stark, H. Determination of scoring functions for protein damage susceptibility. *BioSystems* **187**, 104035, DOI: 10.1016/j.biosystems.2019.104035 (2020).

33. Cooper, R. A. Iodine revisited. *Int. Wound J.* **4**, 124–137, DOI: 10.1111/j.1742-481X.2007.00314.x (2007).

34. Huck, S. W., Cross, T. L. & Clark, S. B. Overcoming Misconceptions About Z-Scores. *Teach. Stat.* **8**, 38–40, DOI: 10.1111/j.1467-9639.1986.tb00624.x (1986).

35. Yeo, I.-K. A new family of power transformations to improve normality or symmetry. *Biometrika* **87**, 954–959, DOI: 10.1093/biomet/87.4.954 (2000).

36. Binns, D. *et al.* QuickGO: A web-based tool for Gene Ontology searching. *Bioinformatics* **25**, 3045–3046, DOI: 10.1093/bioinformatics/btp536 (2009).

37. Seal, R. L. *et al.* Genenames.org: the HGNC resources in 2023. *Nucleic acids research* **51**, D1003–D1009, DOI: 10.1093/nar/gkac888 (2023).

38. Bateman, A. *et al.* UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617, DOI: 10.1093/nar/gkae1010 (2025).

39. Blum, M. *et al.* InterPro: the protein sequence classification resource in 2025. *Nucleic acids research* **53**, D444–D456, DOI: 10.1093/nar/gkae1082 (2025).

40. Shiels, H. A. & White, E. The Frank–Starling mechanism in vertebrate cardiac myocytes. *J. Exp. Biol.* **211**, 2005–2013, DOI: 10.1242/jeb.003145 (2008).

41. Burbaum, L. *et al.* Molecular-scale visualization of sarcomere contraction within native cardiomyocytes. *Nat. Commun.* **12**, 4086, DOI: 10.1038/s41467-021-24049-0 (2021).

42. Yamada, Y., Namba, K. & Fujii, T. Cardiac muscle thin filament structures reveal calcium regulatory mechanism. *Nat. Commun.* **11**, 153, DOI: 10.1038/s41467-019-14008-1 (2020).

43. Dutta, D., Nguyen, V., Campbell, K. S., Padrón, R. & Craig, R. Cryo-EM structure of the human cardiac myosin filament. *Nature* **623**, 853–862, DOI: 10.1038/s41586-023-06691-4 (2023).

Data availability

The data supporting this study's findings are available at DOI: 10.17632/nd4ch2vhty

Acknowledgements

We would like to thank all members of the Department of Bioinformatics (especially Prof. Schuster) and the Institute of Zoology and Evolutionary Research. H.S. was supported by DFG FI 410/24-1 and 463/20-1.

Author contributions statement

H.S., V.W. and L.M.B. designed the study. V.W. performed the statistics. V.W. and H.S. drafted the manuscript and all figures. All authors contributed to the interpretation of the results and revised the manuscript.

Competing interests

The authors declare that they have no competing interests.

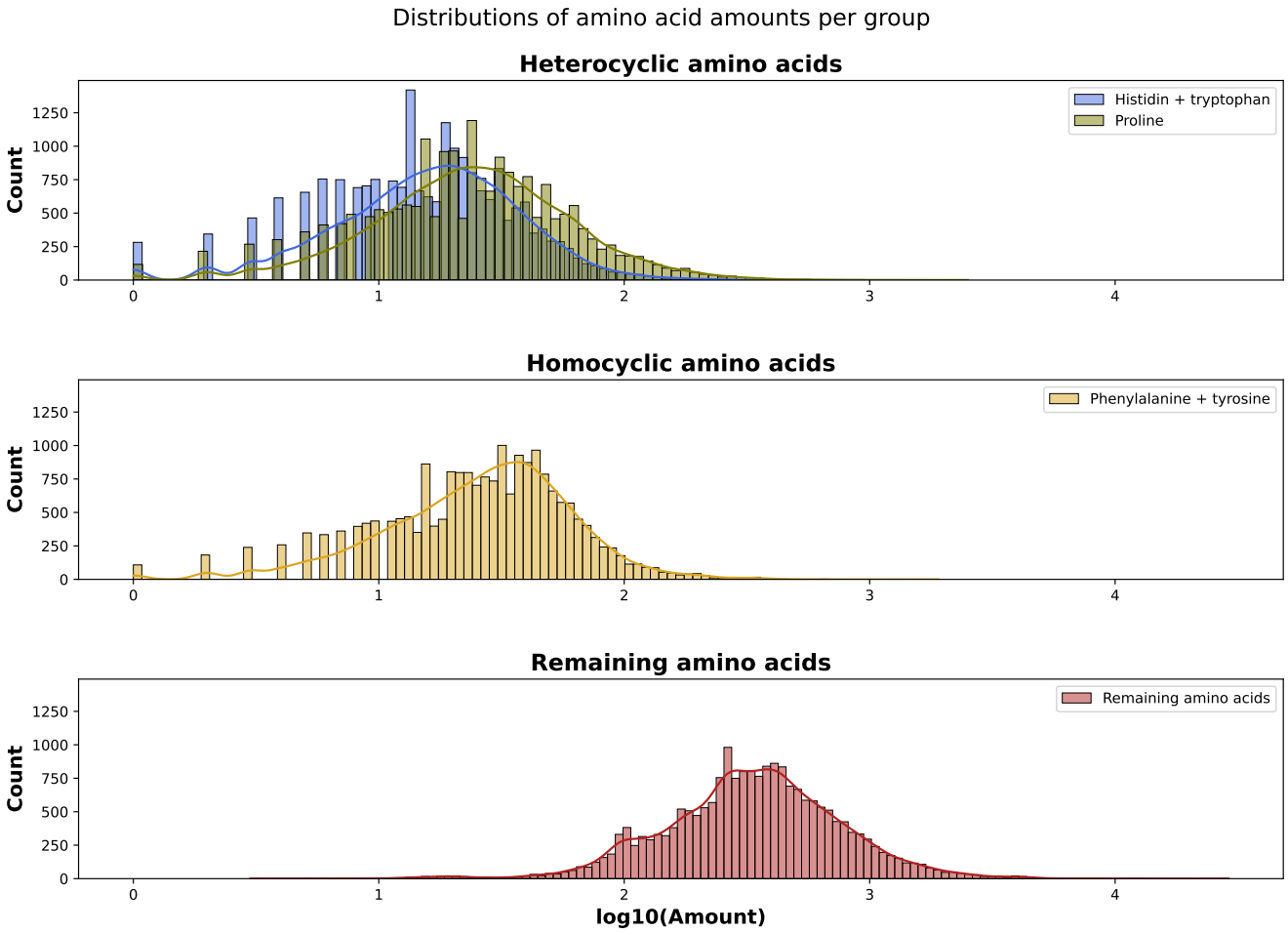

Figure 1. Distribution of proteins, measured by the number of aromatic and non-aromatic heterocycles, homocycles and non-aromatic amino acids.

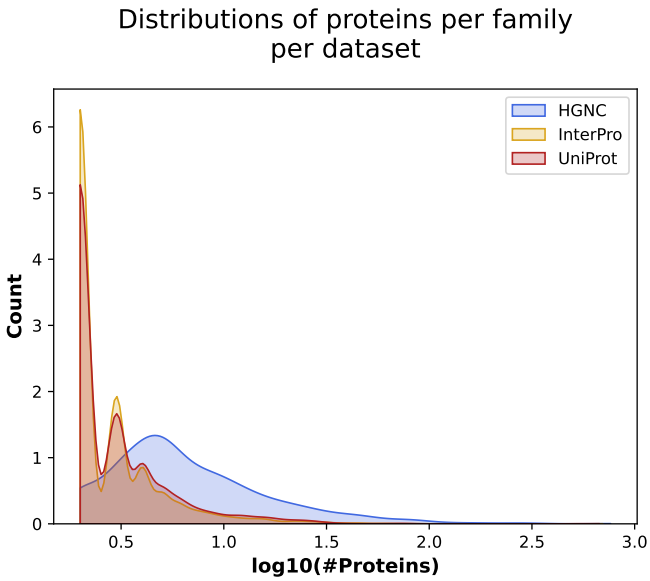

Figure 2. Distribution of the number of proteins per protein family in the HGNC, UniProt and InterPro databases.

Gene name	Full name	#Aromatic heterocycles	Z-score
TITIN	Titin	944	6.19
MUC3B	Mucin-3B	533	5.12
FILA	Filaggrin	434	4.75
MUC16	Mucin-16	408	4.64
NEBU	Nebulin	389	4.55
SYNE1	Nesprin-1	369	4.46
SSPO	SCO-spondin	314	4.19
OBSCN	Obscurin	298	4.10
HORN	Hornerin	284	4.02
DNHD1	Dynein heavy chain domain-containing protein 1	267	3.92

Table 1. The top ten human proteins with the highest number of aromatic heterocycles, along with their corresponding Z-scores.

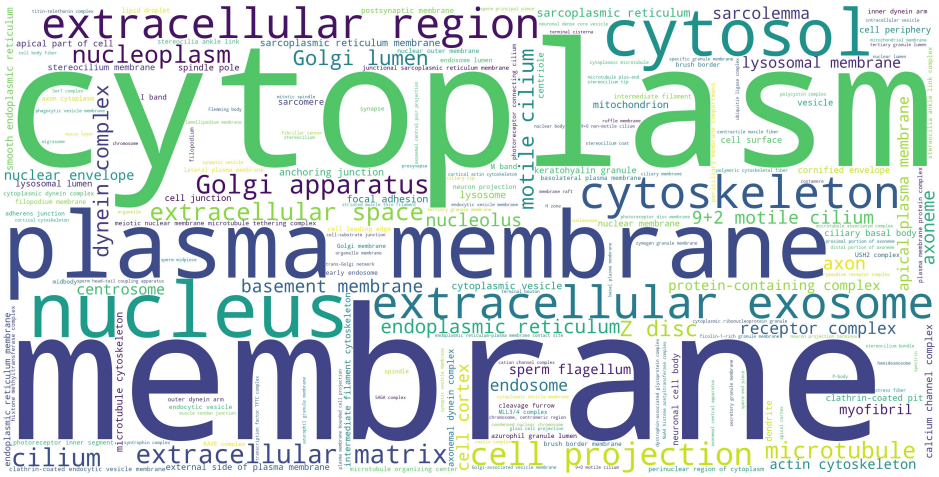

Figure 3. Word cloud illustrating 'cellular_component' Gene Ontology terms associated with the 78 proteins with an absolute aromatic heterocyclic Z-score of three or higher.

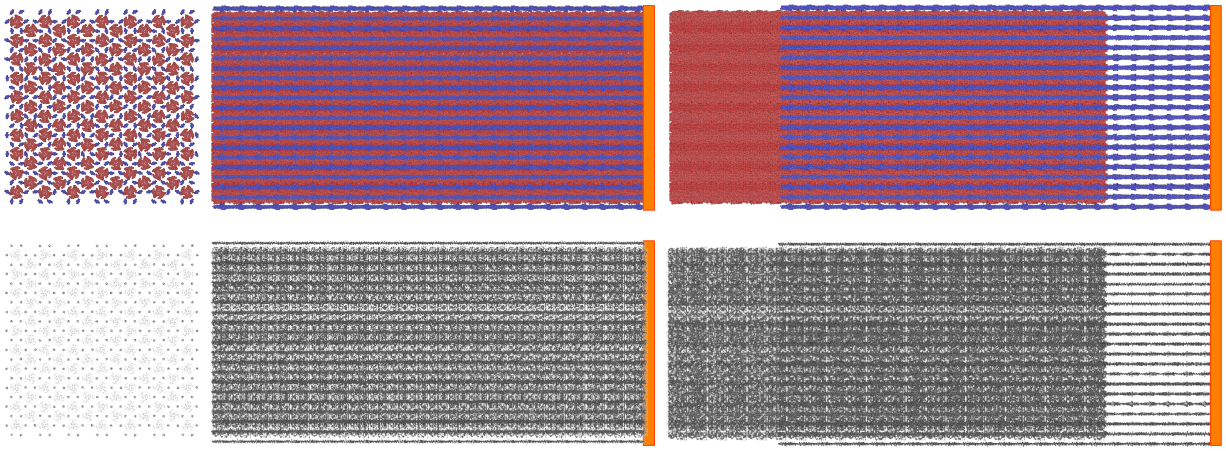

Figure 4. Half-sarcomere protein arrangements of the cardiac muscle (Thin filaments with 7056 protein complexes PDB:6KN7 with a total of 54,902,736 AAs, and thick filaments with 2728 protein complexes PDB:8G4L with a total of 140,322,864 AAs)[40–43]. From left to right: Cross-section at the level of the Z-line (orange), 100% overlap, and 75% overlap of the myosin (red) and actin filament (blue). Below, only the aromatic heterocycles are highlighted in black.

Gene name	Full name	%Aromatic heterocycles	Z-score
HRCT1	Histidine-rich carboxyl terminus protein 1	23.5	5.05
YD019	Putative uncharacterized protein FLJ46204	21.1	4.84
PHGR1	Proline, histidine and glycine-rich protein 1	18.3	4.53
PLAC4	Placenta-specific protein 4	16.0	4.20
SPGOS	Putative uncharacterized protein SPART-AS1	15.4	4.10
PRM2	Protamine-2	13.7	3.79
HIS3	Histatin-3	13.7	3.79
YV018	Putative uncharacterized protein DKFZp434K191	13.7	3.79
SNIT1	Uncharacterized protein encoded by SND1-IT1	13.6	3.77
S39A7	Zinctransporter SLC39A7	13.4	3.72

Table 2. The top ten human proteins with the relatively highest number of aromatic heterocycles, along with their corresponding Z-scores.