CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	A machine learning approach to 3D protein structure sonification
Authors	Ronan, Isabel;Mi, Yanlin;Yallapragada, Venkata Vamsi Bharadwaj;Ó Nuanáin, Cárthach;Tabirca, Sabin
Publication date	2024
Original Citation	Ronan, I., Mi, Y., Yallapragada, V. V. B., Ó Nuanáin, C. and Tabirca, S. (2024) 'A machine learning approach to 3D protein structure sonification', International Journal of Music Science, Technology and Art, 6(1), pp. 9-18. https://doi.org/10.48293/IJMSTA-106
Type of publication	Article (peer-reviewed)
Link to publisher's version	10.48293/IJMSTA-106
Rights	© 2024, the Authors. Licensed to IJMSTA. This is an open access article distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unlimited use, distribution and reproduction in any medium so long as the original work is properly cited. - https://creativecommons.org/licenses/by/4.0/
Download date	2026-11-01 21:40:49
Item downloaded from	https://hdl.handle.net/10468/17084

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

International Journal of Music Science, Technology and Art
(IJMSTA) 6 (1): 9-18, Music Academy "Studio Musica", 2024 ISSN: 2612-2146 (Online)

A Machine Learning Approach to 3D Protein Structure Sonification

Isabel Ronan[1]✉, Yanlin Mi[1], Venkata Vamsi Bharadwaj Yallapragada[2], Cárthach Ó Nuanáin[2], Sabin Tabirca[1]

[1]University College Cork, firstinitial.lastname@cs.ucc.ie
[2]Munster Technological University
✉ Corresponding author: i.ronan@cs.ucc.ie

ARTICLE INFO

Received: March 23, 2024
Accepted: May 6, 2024
Published: May 28, 2024
DOI: https://doi.org/10.48293/IJMSTA-106

Keywords:
machine learning
knn
protein
sonification
music

ABSTRACT

Proteins are intricate structures that can be analysed by biologists and presented to the public using visualisations. However, with an increase in the amount of readily available protein-related information, new forms of data representation are needed. Sonification offers multiple advantages when conveying large amounts of complex data to interested audiences. Previous attempts have been made to sonify protein data; these techniques mainly focus on using amino acid sequences and secondary structures. This paper proposes a novel protein sonification algorithm involving atomic coordinates, B-factors, and occupancies to investigate new ways of displaying 3D protein structure data. This study culminates in creating a cultural showcase involving some of nature's most significant molecular structures. Results of both musical analysis and the showcase indicate that protein sonification has the potential to act as a helpful outreach and engagement tool for biologists while also helping experts in the field glean new insights from complex data.

Copyright © 2024 Author *et al.*, licensed to IJMSTA. This is an open access article distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/3.0/), which permits unlimited use, distribution and reproduction in any medium so long as the original work is properly cited.

1 Introduction

Sonification is the transformation of data into non-speech audio to facilitate communication and interpretation [1]. Sonification is naturally interdisciplinary; it can merge the arts with a plethora of different sciences. Recently, sonification has grown in popularity. With the advent of big data, sonification is used to make large volumes of information comprehensible [2]. Additionally, sonification can make complex, non-textual data accessible for those who are visually impaired [3].

Human beings are inherently able to respond to sound [3]. Due to the subjective nature of the communication medium, we can find patterns in the audio environments presented to us. This inherent ability to glean new insights from otherwise unremarkable data makes sound an ideal communication medium for multi-dimensional, complex data. Sonification can also be used as an outreach tool for education and public engagement (EPE) to bring complex protein structures to a larger audience.

This paper proposes a novel sonification method for protein data involving supervised machine learning. The algorithm proposed in this paper is one of the first to determine musical pitch based on 3D protein structures in relation to their centres of mass. Additionally, B-factors and occupancies of each atom within a protein are used. B-factors measure atom oscillation amplitudes around their equilibrium [4]. Occupancies monitor atom presence at its mean position [5]. As B-factors and occupancies are related, our proposed algorithm aims to link both factors to determine note duration [4]. Furthermore, this algorithm is used to generate pieces for an artistic showcase to demonstrate the enjoyable and educational nature of protein data exploration.

The main contributions of this paper are the definition of a novel sonification algorithm, the musical analysis of resultant sounds to discern structural features of proteins and the use of protein sonification in a real-life outreach application to assess its

effectiveness. The rest of this paper is divided into sections regarding related work, methodology, algorithmic results, and conclusions.

2 Related Work

Sonification has many applications and can be used in various fields. Sonification has been explored in depth within the context of sports, select forms of scientific communication, and healthcare [6], [7], [8]. The Sonification Handbook [9] gives a comprehensive overview of such auditory processes; this book defines sonification as the transformation of data into acoustic signals to facilitate communication or interpretation. Additionally, the aforementioned book outlines various types of sonification and musification, such as parameter mapping [9]. With regards to parameter mapping, there often exists a compromise between accuracy and aesthetics [9]; this compromise is of particular concern when dealing with complex data. Dubus and Bresin's systematic literature review on physical quantity sonification [10] revealed that the most common motivations for data sonification include data exploration, aesthetics, accessibility, motion perception, monitoring, complements to visualisation, and psychoacoustic study. Low-level synthesis, sample-based synthesis, musical sounds, and MIDI are the most common musical products of sonification processes.

Many pre-existing studies regarding protein sonification focus heavily on using amino acids to generate music. Dunn and Clark [11] proposed a sonification method involving analysis of the primary amino acid sequence and the folding patterns of proteins. They used changes in instrumentation and pitch to mark region turns within a protein. They arranged protein amino acids in order of water solubility and subsequently created solubility scales to map each amino acid to a pitch. Their project resulted in an artistic CD to provide empirical proof of the aesthetic patterning of proteins. Their protein music contains rich counter melodies and tune offsets, which create deeply intricate sonic motifs.

Yu, Qin, Martin-Martinez, and Buehler [12] demonstrated the use of the natural vibrational frequencies of amino acids in creating playable musical instruments. These amino acids were subsequently mapped to piano keyboard notes to create musical sequences. They used Python scripts to translate any sequence into a musical score. Additionally, the secondary structure of the protein geometry was computed using DSSP; this secondary structural information is used to determine note duration and volume. The hierarchical nature of a protein is subsequently captured by mapping these features to music; this process is furthered by training a recurrent neural network (RNN) on this protein music. They highlight the need for further work to generate new proteins. They also report on the outreach potential found in such music; a broader range of individuals could understand proteins and their importance by using sound-based representations.

Martin, Meagher, and Barker [13] developed five different sonification algorithms for protein sequences. Their algorithms used amino acid, entropy, and hydrophobicity features when mapping protein data to musical notes. They also used a reduced alphabet of four letters to decrease the complexity of the amino acid sequences. They subsequently tested their algorithms for effectiveness with a range of bioinformatics participants. The results of their usability tests indicated that such sonification can convey particular protein features; these representations lead to heightened interest in proteins.

Gena and Strom [14] created musical pieces from DNA files by converting the list of 64 codons into musical events according to their physical properties. The dissociation constant of each amino acid was used to define pitch. The hydrogen bonding in each codon determined the velocity of each note. Dissociation constants, along with the atomic weights of amino acids, are used to determine note duration. Their work describes one of the most intricate mappings of protein properties to music.

Meytin [15] proposed three sonification methods involving amino acid molecular vibrations and manual pitch assignments. They found that it is possible to sonify proteins consistently for future scientific applications. They also highlighted artistic implications for protein structure sonification; the highly patterned nature of the results can be applied to musical composition to create novel forms of scientific entertainment.

Bouchara and Montès [16] have proposed an immersive sonification method involving protein surfaces. Their method involves decimating protein point clouds and using immersive spatialisation to place users at the centre of the protein for analysis. Their work consists of the use of interactive features to filter data. Subsequent results find consistent point position and density-related audio patterns.

Although amino acid sequences are often used to sonify proteins, their geometrical structures have yet to be explored in depth. While surface-based mapping techniques have been developed, further work is needed to analyse these structures thoroughly. Using machine learning to enable efficient parameter mapping has yet to be thoroughly investigated. Additionally, while many

studies discuss the outreach capabilities of such sonification, there is a lack of work surrounding the use of public engagement to assess general enthusiasm for such projects

3 Methodology

The parameter-mapping algorithm developed for this project involves transforming atom coordinate points within a protein structure into meaningful measurements using the protein's centre of mass. A protein's centre of mass is an artificial point for detecting structural features, shapes, and associations [17]. This paper outlines one of the first sonification methods, which uses centre of mass to discern meaning from atomic coordinates within a 3D protein structure. Ten different biologically significant protein structures are chosen for testing and presentation purposes.

Music generation is carried out using Python scripts. Post-processing is subsequently carried out using Logic Pro. Protein PDB files downloaded from the RCSB Protein Data Bank were used in the music generation process [18]. Biopython, pretty midi, mir eval, scikit-learn and Matplotlib libraries were used to organise protein material, visualise algorithmic strategies, analyse resultant sounds and generate appropriate musical outputs for further aesthetic processing.

The protein-to-music conversion process is outlined as follows:

- Step 1: Extracting Protein Data
- Step 2: Labelling Training Data
- Step 3: Training KNN Model
- Step 4: Predicting Notes and Creating MIDI
- Step 5: Analysing Music and Selecting MIDI
- Step 6: Post-Processing.

3.1 Extracting Protein Data

The algorithm developed involves extracting the centre of mass, atomic coordinates, B-factors, and occupancy measures from a protein PDB file. Note durations are calculated based on the B-factors and occupancies of the protein points used. Specifically, B-factors are normalised with respect to their occupancies. Table 1 outlines the quantisation thresholds for B-factor occupancies.

TABLE I. B-FACTOR VALUES.

B-Factor Threshold	Mapped Amount
b < 25	0.25
25 <= b =< 50	0.5
50 < b < 75	0.75
b >= 75	1.0

Once the centre of mass and all of the coordinates of the protein are calculated, a standard 3-dimensional distance formula is used to discern the distances between all points. The distance measurements calculated are used for both pitch predictions and velocity note parameter mappings. This distance formula for d is outlined below where (x2, y2, z2) is an atomic coordinate and (x1, y1, z1) is the centre of mass of the protein.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

3.2 Labelling Training Data

The protein point distances are then divided into eight different distance bands representing the eight notes of a C major scale. The eight musical bands are evenly arranged around the centre of mass; the furthest point from the centre of mass is used to calculate the distribution. The points at each of the eight distances from the centre of mass are mapped to their corresponding notes. This labelled data is then used to train a KNN machine-learning model to assign notes to the rest of the points in the protein. The KNN-supervised machine-learning technique is used due to its simplicity. Each protein analysed and converted into music will have its own KNN model trained specifically for note-mapping purposes. Figure 1 shows the labelled protein

data and the data for KNN categorisation.

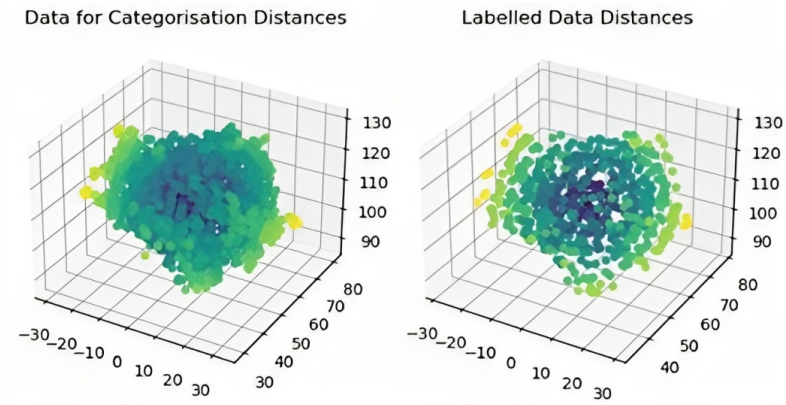


Figure 1. Unclassified Versus Labelled Data.

3.3 Training KNN Model

The labelled data is divided into training and testing sets. Subsequently, the KNN model is trained; two neighbours were found to be the optimal number for this algorithm after multiple training and testing cycles. The average accuracy of the KNN models trained for each protein was 99.3%. Results for an exemplary selection of proteins mapped are outlined in Table 2.

TABLE II. KNN ACCURACIES.

Protein	KNN Accuracy
1N8Z	99.19%
3BIK	99.13%
3J6R	99.37%
6PV7	99.51%

3.4 Predicting Notes and Creating MIDI

Once all protein points have been assigned a note, the protein is traversed in 20 arbitrary directions in a scan-like fashion. These 20 random scans are used to organise the notes as they are written to MIDI files. Protein distance values related to each assigned pitch are used as velocity values for the MIDI notes. Additionally, the B-factor-occupancy values calculated during the earlier stages of the algorithmic process are used to assign durations to notes. Figure 2 provides a conceptual representation of how this process was undertaken.

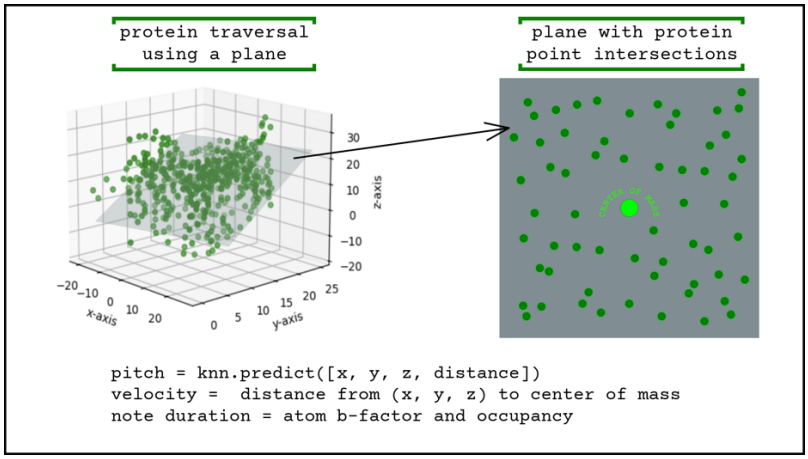


Figure 2. Protein Point Plane Traversal.

3.5 Analysing Music and Selecting MIDI

Each of the 20 MIDI files is musically analysed based on duration, tempo, rhythm, and pitch stability. Duration is assessed using the pretty-midi library's get end time() function. Piece length is categorised as 'appropriate' when it is of one to four minutes in duration. The piece's tempo is calculated using prettymidi's estimate tempo() function. This function returns a beats-per-minute amount; the higher the score, the quicker the piece and the more exciting the music. Additionally, pretty-midi's get pitch class histogram() function calculates the strongest pitch within a piece. The strongest pitch number is then added to the total score with respect to its strength. The higher the pitch score, the more stable and aesthetically pleasing the pitch centre is in the piece. Rhythmic consistency is calculated using the mir eval library's information gain() function.
The resultant scores for each musical feature are combined to create an overall score. This overall score determines which MIDI file is the most 'musical' and should subsequently be subject to post-processing. The MIDI file with the highest overall score is used to represent the protein.

3.6 Post-Processing

The MIDI files chosen during the musical analysis phase are inputted into Logic Pro for further processing. Post-processing ensures that aesthetically appealing soundscapes are generated from the protein MIDI data. The pitch and note duration values outputted during the MIDI generation phase remain untouched during post-processing; however, each MIDI file is truncated to facilitate optimal performance length. All post-processing decisions are made to facilitate the optimal mix of aesthetic appeal and accuracy. Instruments and sound effects are all chosen with the relevant protein in mind. For example, as shown in Figure 3, breath sound effects are added to the music of the Human Oxyhaemoglobin protein 1HH0 [19] to facilitate an oxygen-focused soundscape. Additionally, reverb, compression, and EQ plugins were added to create ambient atmospheres.

4 Results and Discussion

Both musical analysis and public performance assessments support the feasibility of the proposed sonification algorithm. Musical analysis revealed insightful patterns within the proteins. A public performance undertaken as part of this research showed the use of biological music within an outreach context. Results provide justification for future work, both within scientific and artistic contexts.

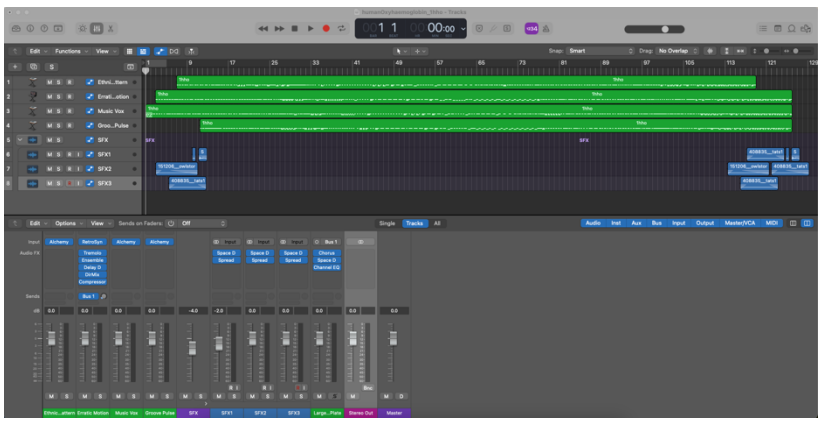


Figure 3. Human Oxyhaemoglobin in Logic Pro.

4.1 Musical Analysis

The results of this algorithmic process were empirically analysed by the authors of this paper to discern meaning from the music. Observations highlighted the underlying natures of the proteins themselves. For example, the tobacco mosaic virus [19] generated a strong pulsing rhythm; one can coincidentally observe the mottled mosaic-like appearance of infected plants in the music.

All protein music was generated using MIDI notes related to the C major scale. The musical notes used, along with their corresponding MIDI numbers, are outlined in Table 3.

TABLE IV. MIDI MUSICAL NOTES.

Musical Note	C_4	D_4	E_4	F_4	G_4	A_4	B_4	C_5
MIDI Number	60	62	64	65	67	69	71	72

Each protein used within this study generated music with a different key centre. The key centres discerned from the post-processed protein tracks are outlined in Table 4. From these results, one can see that C minor is the most popular key centre. Figure 4 shows the C minor pitch histogram of the Immunoglobulin protein molecule.

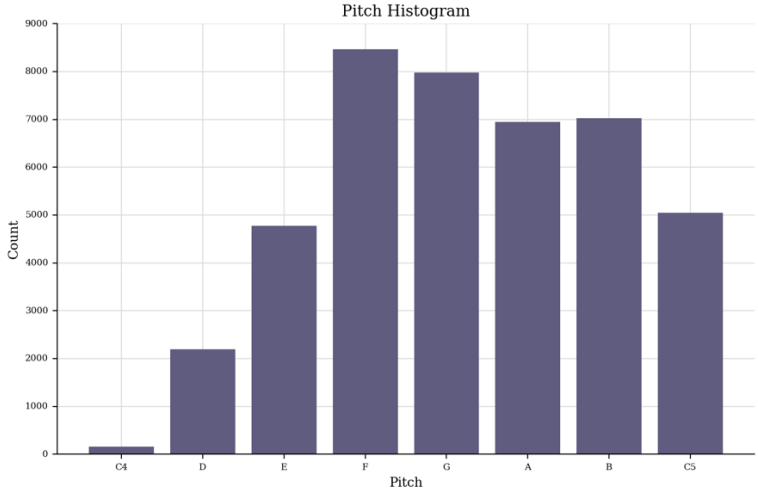


Figure 4. Pitch Histogram of Immunoglobulin.

Additionally, many human-related proteins tend to focus on the same key; for example, the music generated from human alpha thrombin and genome proteins results in the same pitch centre. Key signatures involving accidentals, such as Pig Insulin's F# minor, are attributed to post-processing effects and not to failures on the part of the algorithm itself.

The structure of all MIDI files generated revealed the symmetrical nature of many proteins. As each piece progresses, the different musical lines get increasingly more complex. After intense sonic complexity, the sounds become more sparse and aesthetically spacious. Additionally, there are pitch descent and ascent-related patterns present in each file; these patterns vary according to the file chosen during the music analysis section of the proposed algorithm.

Most pieces assessed as part of this project follow a similar density and pitch trajectory. One can infer from the musical representations of structures analysed that most proteins are densest at their centre of mass. One can also infer that proteins have fewer atoms around the edges of their structure. Figure 5 exhibits the symmetrical MIDI file shape of 1FQY as seen in Logic Pro. This figure clearly shows that the protein structure grows in density as the music moves closer to the centre of mass of the molecule.

In summary, the protein sounds generated exhibit clear musical patterns. The significance of this work lies within the use of a novel structure-focused sonification algorithm; this algorithm has clearly exhibited the ability to extract patterns from protein data. From these patterns, one can recognise 3D structural features of proteins. Future work could focus on establishing links between structures based on their musical features. Additionally, the parameter mapping techniques explored in this study could be further investigated within the context of "de novo" protein design; if proteins can be made into music, music can be made into proteins. Further investigation could lead to significant advances within the field of molecular biology. Both within educational and innovative contexts, sonification could help scientists glean new insights into pre-existing protein data and generate new proteins using music.

TABLE V.PROTEIN KEY CENTRES.

Protein Name	Protein Code	Key Centre
Immunoglobulin	1igt	C minor
Human Oxyhaemoglobin	1hho	C minor
DnaB Helicase	4esv	C minor
Tobacco Mosaic Virus	2tmv	E minor
Serotonin	4iar	E minor
Human Alpha Thrombin	1ppb	A minor
Genome	4un3	A minor
Zika Virus	5ire	G minor
Pig Insulin	4ins	F# minor
Water Channel	1fqy	C major

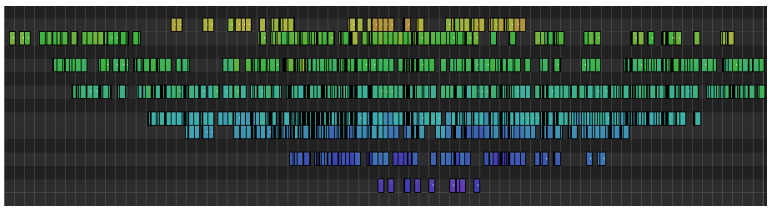


Figure 5. 1FQY MIDI in Logic Pro.

4.2 Public Performance

An artistic showcase was prepared to present the results of the developed algorithm to interested members of the public. This outreach event assessed the significance of algorithmically-generated soundscapes within a public engagement context. The event took place during Culture Night 2023, an annual, nationwide event developed by The Arts Council that celebrates cultural institutions and their activities[1]. The "Molecular Melodies" exhibition comprised three components:

- A quadraphonic audio-visual installation of a selection of the musical pieces accompanied by 3D visuals
- Physical 3D printed models of the protein structures along with descriptive materials
- Interactive virtual reality experiences with 3D models

The protein sounds presented at the showcase were chosen based on their importance in nature. Each protein was a well-known molecule from biological history or from key functional processes within human beings and animals. Each piece was roughly three to four minutes in duration. All generated music was made available on SoundCloud, so interested members of the general public could access the pieces [20]. The event was open to all members of the public; interested observers were free to view 3D models of the proteins in virtual reality to accompany the sonic soundscapes. Figure 6(a) shows the layout of the installation space for the event, and Figure 6(b) shows some interested audience members experiencing protein in VR to accompany the music.



(a) (b)

Figure 6. 3D Soundscape Event

[1] https://culturenight.ie/about/

The audience response to the showcase was very positive. People of all ages attended the event. Members of the public engaged actively with those involved in the project and were eager to learn more about the same. Those involved in the project presentation were present to answer questions and assist members of the public in understanding the showcase. Informal feedback was gleaned from audience members through the use of a small comment book placed at the entrance of the room. 46 responses were collected from the comment book. Some of the most detailed responses highlighted future applications of the showcase.

"I'm amazed by how science and music can be brought together so skillfully. One has to be probably very passionate about both. I'd love to have this introduced to schools as well to make biochemistry more relatable and appealing."

Additionally, audience members focused on the comprehensibility of the showcase; they outlined the positive impact the generated music had on understanding the protein visualisations.

"I don't have a science background and this made it much easier for me to visualise proteins."

In summary, the showcase was very successful, emphasising the future educational and entertainment-related advantages of such outreach activities. The use of such an event is significant within the field of protein sonification as, to the best of our knowledge, no prior related work has implemented an artistic showcase utilising experimental sonic results. This paper suggests that through artistic events, proteins can be made more appealing, entertaining, and understandable to the general public. Future work could focus on conducting more rigorous usability tests within a similar installation-based context; these usability tests could further establish the viability of such events

5 Conclusion

Protein-generated music and biologically-influenced soundscapes can be used in a variety of ways. Outreach activities can be envisaged involving these structures, and educational programs can be developed to help increase the comprehensibility of complex biological forms. This paper proposes a novel music-generation algorithm using supervised machine learning. Atomic coordinates within protein structures are mapped to musical notes using distance calculations; these distance calculations are also used to determine velocities. B-factors and occupancies are also used to determine note durations.

This algorithm has proven useful both in the discernment of new patterns within protein data and in the development of new outreach activities. Outreach activities can make protein structures more accessible while simultaneously allowing artistic insights to be gleaned from complex biological data. In-depth analysis of such music can also be used as a tool for molecular biologists. Visualisations of proteins can be augmented by the use of sound, helping those who study such structures to discover new features of well-known forms.

Protein data has been turned into music; the next steps in this work involve refining the algorithm with more detailed parameter mapping; this refinement will lead to decorative and elaborate musical arrangements along with deeper data insights. Additionally, future work will include inverting the proposed algorithm to turn music into protein data; this backward biological mapping could be used to discover new proteins through the power of sound.

Visualisation is one of our most common forms of data representation. However, visual depictions of proteins can be insufficient when dealing with intricate structures. Sonification offers scientists the ability to examine complex data using a sense other than sight. These auditory representations can not only augment the visuals we are already so familiar with, but can lead to new and intuitive insights into proteins. The more we sense, the more we understand. If there is meaning in music and power in proteins, protein music holds the ability to be as meaningful as it is powerful; such sounds can help us sense new potentials for the building blocks of life and new futures for both science and art.

Notes

The website dedicated to housing this project's materials can be found via the following link: https://cs1.ucc.ie/~imr1/ampc/.

Acknowledgements

This project is funded by the MTU Arts Office Create Le Chéile Arts Project Award.
This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant number 18/CRT/6222. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[1] Kramer, G., Walker, B., Bonebright, T., Cook, P., Flowers, J., Miner, N., Neuhoff, J. (2010). Sonification Report: Status of the Field and Research Agenda. Faculty Publications, Department of Psychology, March 2010. URL https://digitalcommons.unl.edu/psychfacpub/ 444.

[2] Rimland, J., Ballora, M., Shumaker, W. (2013). Beyond visualization of big data: A multi-stage data exploration approach using visualization, sonification, and storification, volume 8758. May 2013. https://doi.org/10.1117/12.2016019. Journal Abbreviation: Proceedings of SPIE - The International Society for Optical Engineering Publication Title: Proceedings of SPIE - The International Society for Optical Engineering.

[3] Barrass, S., Kramer, G. (1999). Using sonification. Multimedia Systems, 7(1):23–31, January 1999. ISSN 1432-1882. https://doi.org/10.1007/s005300050108. URL https://doi.org/10. 1007/s005300050108.

[4] Carugo, O. (2018). How large B-factors can be in protein crystal structures. BMC Bioinformatics, 19(1):61, February 2018. ISSN 1471-2105. https://doi.org/10.1186/s12859-018-2083-8. URL https://doi.org/10. 1186/s12859-018-2083-8.

[5] Carugo, O. (1999). Correlation between occupancy and B factor of water molecules in protein crystal structures. Protein Engineering, Design and Selection, 12(12):1021–1024, December 1999. ISSN 1741-0126. https://doi.org/10.1093/protein/12.12.1021. URL https://doi.org/10. 1093/protein/12.12.1021.

[6] Van Rheden, V., Grah, T., Meschtscherjakov, A. (2020). Sonification approaches in sports in the past decade: a literature review. In Proceedings of the 15th International Audio Mostly Conference, pages 199–205, Graz Austria, September 2020. ACM. ISBN 978-1-4503-7563-4. https://doi.org/10.1145/3411109.3411126.

[7] Sawe, N., Chafe, C., Trevino, J. (2020). Using Data Sonification to Overcome Science Literacy, Numeracy, and Visualization Barriers in Science Communication. Frontiers in Communication, 5, 2020. ISSN 2297- 900X. URL https://www.frontiersin.org/articles/10.3389/fcomm. 2020.00046.

[8] Guerra, J., Smith, L., Vicinanza, D., Stubbs, B., Veronese, N., Williams, G. (2020). The use of sonification for physiotherapy in human movement tasks: A scoping review. Science & Sports, 35(3):119–129, June 2020. ISSN 0765-1597. https://doi.org/10.1016/j.scispo.2019.12.004.

[9] The Sonification Handbook | edited by Hermann, Hunt, Neuhoff, . URL https://sonification.de/handbook/. Sonification of 3D Protein Structures Using Supervised Machine Learning 15

[10] Dubus, G., Bresin, R. (2013). A Systematic Review of Mapping Strategies for the Sonification of Physical Quantities. PLOS ONE, 8(12):e82491, December 2013. ISSN 1932-6203. https://doi.org/10.1371/journal.pone.0082491. URL https://journals. plos.org/plosone/article?id=10.1371/journal.pone.0082491. Publisher: Public Library of Science.

[11] Dunn, J., Anne Clark, M. (1999). The Sonification of Proteins. *Leonardo.*

[12] Chi-Hua Yu, Zhao Qin, Francisco J. Martin-Martinez, Markus J. Buehler (2019). A Self-Consistent Sonification Method to Translate Amino Acid Sequences into Musical Compositions and Application in Protein Design Using Artificial Intelligence. ACS Nano, 13(7):7471–7482, July 2019. ISSN 1936-0851, 1936-086X. https://doi.org/10.1021/acsnano.9b02180. URL https://pubs.acs.org/doi/10.1021/acsnano.9b02180.

[13] Martin, E.J., Meagher, T.R., Barker, D. (2015). Using sound to understand protein sequence data: new sonification algorithms for protein sequences and multiple sequence alignments. BMC Bioinformatics, 22(1): 456, September 2021. ISSN 1471-2105. https://doi.org/10.1186/s12859- 021-04362-7. URL https://doi.org/10.1186/s12859-021-04362-7.

[14] Peter Gena. Musical Synthesis of DNA Sequences.

[15] Meytin, S. (2021). A Novel Method for Protein-Protein Interface Analysis Using Sonification. In 2021 IEEE MIT Undergraduate Research Technology Conference (URTC), pages 1–5, October 2021. https://doi.org/10.1109/URTC54388.2021.9701622. URL https: //ieeexplore.ieee.org/document/9701622.

[16] Bouchara, T., Montes, M. (2020). Immersive sonification of protein surface. In 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), pages 380– 383, Atlanta, GA, USA, March 2020. IEEE. ISBN 978-1-72816- 532-5. https://doi.org/10.1109/VRW50115.2020.00082

[17] Costantini, S., Paladino, A., Facchiano, A.M. (2008). CALCOM: A software for calculating the center of mass of proteins. Bioinformation, 2(7):271–272, February 2008. ISSN 0973-2063. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2374369/.

[18] RCSB PDB: Homepage, . URL https://www.rcsb.org/.

[19] RCSB Protein Data Bank. RCSB PDB - 2TMV: VISUALIZATION OF PROTEIN-NUCLEIC ACID INTERACTIONS IN A VIRUS. REFINED STRUCTURE OF INTACT TOBACCO MOSAIC VIRUS AT 2.9 ANGSTROMS RESOLUTION BY X-RAY FIBER DIFFRACTION. URL https://www.rcsb.org/structure/2tmv.

[20] Soundscapes From 3D Protein Structures, September 2023. URL https://soundcloud.com/user-455170818/sets/soundscapes-from-3d-protein-structures.