Distinguishing Proteins From Arbitrary Amino Acid Sequences

Stephen S.-T. Yau\(^1\)#, Wei-Guang Mao\(^1\)#, Max Benson\(^2\) & Rong Lucy He\(^3\)

\(^1\)Department of Mathematical Sciences, Tsinghua University, Beijing, 100084, China, \(^2\)Department of Computer Science, Seattle Pacific University, Seattle, WA 98119, USA, \(^3\)Department of Biological Sciences, Chicago State University, Chicago, IL 60628, USA.

What kinds of amino acid sequences could possibly be protein sequences? From all existing databases that we can find, known proteins are only a small fraction of all possible combinations of amino acids. Beginning with Sanger’s first detailed determination of a protein sequence in 1952, previous studies have focused on describing the structure of existing protein sequences in order to construct the protein universe. No one, however, has developed a criteria for determining whether an arbitrary amino acid sequence can be a protein. Here we show that when the collection of arbitrary amino acid sequences is viewed in an appropriate geometric context, the protein sequences cluster together. This leads to a new computational test, described here, that has proved to be remarkably accurate at determining whether an arbitrary amino acid sequence can be a protein. Even more, if the results of this test indicate that the sequence can be a protein, and it is indeed a protein sequence, then its identity as a protein sequence is uniquely defined. We anticipate our computational test will be useful for those who are attempting to complete the job of discovering all proteins, or constructing the protein universe.

Structurally, proteins are sequences of amino acids\(^1\). Not all sequences of amino acids correspond to proteins, however. Previous studies\(^2\)–\(^4\) have focused on describing the structure of existing protein sequences in order to construct the protein universe. Recent studies have considered the expansion of the protein universe and its relationship to the Big Bang\(^5\)–\(^7\). So far, however, no study has developed a criteria for distinguishing a protein that is part of the protein universe from an amino acid sequence that is not a protein.

In this paper we will describe a 99.69% accurate computational test for determining if an arbitrary amino acid sequence can be a protein. Our computational test makes it possible to quickly determine whether a certain amino acid sequence can be a protein.

Our test is based on a geometric representation of the protein universe that we have constructed using the natural vector representation\(^8\). According to this approach, each protein sequence is represented by a unique point in coordinate space.

Our testing has shown that the points corresponding to proteins cluster together and it has revealed a fundamental limit on the distribution of each amino acid within proteins that was not evident previously.

Results

The distribution of a specific amino acid “\(k\)” within an amino acid sequence can be described by three quantities:

- \(n_k\) the number of occurrences of the amino acid “\(k\)” within the sequence.
- \(\mu_k\) the mean distance of the amino acid “\(k\)” from the first position.
- \(D_k^2\) the second normalized central moment of the distribution of amino acid “\(k\)”.

Here is a more precise definition of the second and third quantities:

\[\mu_k: \text{Let } s[k][i] \text{ be the distance from the first position of the sequence to the location of the } i^{th} \text{ occurrence of the amino acid } k \text{ and let } T_k = \sum_{i=1}^{n_k} s[k][i] \text{ be the total of the distances of all occurrences of } k \text{ from the start position. Then } \mu_k \text{ is simply } T_k/n_k\]

\(D_k^2\): The second normalized central moment.
$D^2_2$: Let n denote the length of the sequence, then $D^2_2$ is defined by the following formula.

$$D^2_2 = \sum_{i=1}^{n} \frac{(s[k][i] - \mu_k)^2}{n \cdot n_k} \tag{1}$$

Using the standard abbreviations (A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V) for the 20 amino acids, each amino acid sequence can be represented by a point (called its natural vector) in 60-dimensional space with coordinates:

$$<n_A, n_R, \ldots, n_V, \mu_A, \mu_R, \ldots, \mu_V, D_A^1, D_R^1, \ldots, D_V^1> \tag{2}$$

In general, the natural vector representation is not one to one. Many different amino acid sequences will correspond to the same point in 60-dimensional space. But that is not the case with actual protein sequences. The first key result of this paper is that after collecting all the known, reviewed complete protein sequences available in the UniprotKB database and computing their natural vectors, we verified that the natural vector representation is in fact one to one when it comes to known protein sequences.

We define protein space to be the set of all points in 60-dimensional space corresponding to protein sequences, and amino acid space to be the set of points that correspond to amino acid sequences with lengths ranging between the minimum and maximum lengths of the protein sequences.

In order to visualize how the points of protein space are distributed within amino acid space, we plotted them along two of the coordinate axes corresponding to the amino acids Alanine (A) and (B) show the two dimensional projection onto the (n_A, \mu_A) coordinate plane, while (C) and (D) show its projection onto the (n_A, D_A^1) coordinate plane.

In the (n_A, \mu_A) coordinate plane, the projection of amino acid space is bounded by the following three lines

$$\mu_{A, min} : \mu_A = \frac{n_A - 1}{2} \tag{3}$$

$$\mu_{A, max} : \mu_A = \frac{2n - 1 - n_A}{2} \tag{4}$$

$$n_A = 0. \tag{5}$$

In the (n_A, D_A^1) coordinate plane, the projection of the amino acid space is bounded by the following three curves

$$D_{A, min}^2 : D_A^2 = \frac{n_A^2 - 1}{12n} \tag{6}$$

$$D_{A, max}^2 : D_A^2 = \begin{cases} \frac{n - n_A}{4} \frac{n_A^2 - 1}{12n} - \frac{1}{12n} & \text{if } n_A \text{ is even} \\ \frac{(n_A - 1)(n_A + 1)(n - n_A)}{4n_A^2} + \frac{n_A^2}{12n} - \frac{1}{12n} & \text{if } n_A \text{ is odd} \end{cases} \tag{7}$$

$$n_A = 0 \tag{8}$$

In these equations, n_A stands for the maximum length of proteins in the dataset. For the collection of known, reviewed complete protein sequences found in the UniprotKB data as of March 6, 2013, we found that n = 35213.

From Figure 1, we can clearly see that the points of protein space are clustered, rather than being broadly distributed. This makes us believe that as new protein sequences are described, their points will lie approximately within the convex hull of the points corresponding to known protein sequences.

Since there are no efficient algorithms to compute the convex hull in such a high dimensional space, we looked at three dimensional projections of the problem. For each amino acid “k”, we refer to the convex hull of points the form <n_k, \mu_k, D_k^2> corresponding to protein sequences as the “k”-protein area. The “k”-amino acid area is the set of points corresponding to amino acid sequences with lengths in the range of actual protein sequences.

We performed several tests to check whether the boundaries of these protein areas will remain stable over time as more proteins are described. We computed the protein areas using the protein sequences listed in an earlier snapshot of the UniprotKB database. Then we tested to see whether more recently added protein sequences would lie within those convex hulls. We found that nearly all of the new proteins sequences lie within the convex hulls, and the few that did not were found to lie very close to the convex hull boundaries.

Thus, a second key result of this paper is that we have found strong evidence for the validity of the following computational test:

To check whether an arbitrary amino acid sequence can be a protein sequence, we start by computing its natural vector. Next we search a pre-computed database containing natural vectors that correspond to known protein sequences. If the natural vector is found within that database, we check whether the amino acid sequence is identical to the protein sequence we found. If they are identical, we have our answer. If they are not, the first key result of the paper allows us to conclude that the amino acid sequence is not a protein.

If the natural vector of the amino acid sequence is not found in the database, we proceed to check whether each of the 20 points <n_k, \mu_k, D_k^2>, k = A, R, N, \ldots, V in 3-dimensional space lie within their corresponding “k”-protein area convex hull. If all these checks succeed, we conclude that this amino acid sequence could be a known or unknown protein. If on the other hand, not all points lie within the corresponding convex hulls, by this paper’s second key result we would not expect the amino acid sequence to be a protein sequence.

**Methods**

We took three snapshots of the UniprotKB database to test our hypotheses: Uniprot 2013_03 (March 6, 2013), Uniprot 2014_03 (March 19, 2014), and Uniprot 2014_06 (June 11, 2014). In each case, only the reviewed, complete proteins were selected.

We followed the same procedure for generating all three datasets. We added the keyword “Complete proteome[KW-01811]” to select only complete sequences, eliminating sequences with missing amino acids. We also added the keyword “Reviewed” to restrict to only reviewed proteins. Sequence redundancy was not reduced. After the download, the datasets were normalized by removing protein sequences containing Selenocysteine (U) and Pyrrolysine (O), as well as protein sequences containing placeholders (B, Z, J, X).

Only a small number of protein sequences were eliminated by this normalization process: The number of protein sequences containing Selenocysteine, Pyrrolysine or one of the four placeholders is 971 in the Uniprot 2013_03 dataset, 1010 in the Uniprot 2014_03 dataset, and 1044 in the Uniprot 2014_06 dataset. See Table 1 for details of the counts of sequences in the three snapshots of UniprotKB:

We used the Uniprot 2013_03 and Uniprot 2014_03 datasets to verify the first key result (the natural vector mapping produces a one-to-one correspondence between proteins and 60-dimensional natural vectors). As can be seen from Table 2, the number of distinct protein sequences in each of these datasets was equal to the number of distinct natural vectors.

We also performed several tests to check whether our computational test was effective for determining whether a given amino acid sequence could be a protein or not.

First, we used the 391704 distinct protein sequences contained in both Uniprot 2013_03 and Uniprot 2014_03 to construct twenty 3-dimensional convex hulls using the Qhull algorithm. Then we checked to see how many of the 3810 sequences contained solely in Uniprot 2014_03 failed to lie outside of the twenty convex hulls. We found that only 14 sequences (only 0.37% of the 3810 sequences) lie outside one of the convex hulls. We also used all 392455 sequences contained in Uniprot 2013_03 to construct convex hulls and the results were the same. None of these 14 sequences were far from the boundaries of the convex hulls. In Table 3, we list the 14 sequences and...
Figure 1 | Alanine convex hull computed from the Uniprot 2013_03 dataset. Blue points in each of these four subfigures stand for vectors corresponding to proteins. (A) shows the picture in $(n_A, \mu_A)$ coordinate plane and (C) shows the picture $(n_A, D_2^A)$ coordinate plane. (B) is the enlarged view of the protein area in (A). The black lines stand for the boundaries of the convex hull for protein area. (D) is the enlarged view of the protein area in (C). The black lines stand for the boundaries of the convex hull for the protein area.
Table 1 | Detailed counts of sequences in the three snapshots of UniprotKB

| Sequence Count Description | 2013_03 | 2014_04 | 2014_06 |
|----------------------------|---------|---------|---------|
| Number of Distinct Sequences | 392455  | 395514  | 379748  |
| Number of Sequences in Uniprot | 471284  | 478131  | 471284  |
| Number of Sequences in Uniprot 2014_04 & Uniprot 2014_06 | 399254  | 399254  | 399254  |
| Number of Distinct Sequences in both | 17652  | 17652  | 17652  |
| Number of Distinct Sequences in Uniprot 2013_03 & Uniprot 2014_06 | 751  | 751  | 751  |

Discussion

Although there were a few proteins in each test which lay outside at least one of the convex hulls, our results are quite promising. The results of our testing showed that only 18 sequences (0.31% of the 5820 sequences) lie outside one of the convex hulls. As in the previous test, none of these 18 sequences were far from the boundaries of the convex hulls. See the supplementary information for tabulation of the distance of each sequence from the convex hulls. Table 4 lists the 18 sequences and their distance from their respective convex hulls.

As a final test, we tested to see whether the Top 711, HOP212, and GLUT113 protein sequences lie inside the twenty convex hulls. All three proteins lie inside the twenty convex hulls as expected.

Table 2 | Corresponding between sequence counts and natural vectors

| Dataset | Number of Sequences Before Normalization | Number of Sequences | Number of Distinct Sequences | Number of Natural Vectors | Number of Distinct Natural Vectors |
|---------|------------------------------------------|---------------------|----------------------------|----------------------------|-----------------------------------|
| Uniprot 2013_03 | 472284 | 471313 | 392455 | 471313 | 392455 |
| Uniprot 2014_03 | 475547 | 474537 | 395514 | 474537 | 395514 |
amino acid \( k \). The following lemmas show that any distribution which doesn’t have amino acid \( k \) in both the first and last positions will have a second normalized moment which is not maximal.

Lemma 2. Let \( 1 \leq x_1 < x_2 < \cdots < x_n \leq n \) be a distribution of the amino acid \( k \). Let \( 1 \leq x_1 < x_2 < \cdots < x_n \leq n \) be another distribution with the properties that \( x_i = x_1 \) for \( 1 \leq i \leq n_k - 1 \) and \( x_n = x_n + 1 \). Then

\[
D_k^2(x) < D_k^2(x')
\]

Proof: \[
\mu' = \frac{(x_1' - 1) + \cdots + (x_m' - 1)}{n_k} = \frac{(x_1 - 1) + \cdots + (x_m - 1) + 1}{n_k} = \mu + \frac{1}{n_k}
\]

Lemma 3. let \( 1 < x_1 < x_2 < \cdots < x_n \leq n \) be a distribution of the amino acid \( k \). Let \( 1 < x_1' < x_2' < \cdots < x_n' \leq n \) be another distribution with the properties that \( x_1' = x_1 - 1 \) and \( x_i' = x_i \) for \( 2 \leq i \leq n_k \). Then

\[
D_k^2(x) < D_k^2(x')
\]

Proof:

\[
\mu' = \frac{(x_1' - 1) + (x_2' - 1) + \cdots + (x_n' - 1)}{n_k} = \frac{(x_1 - 1) + (x_2 - 1) + \cdots + (x_n - 1) - 1}{n_k} = \mu - \frac{1}{n_k}
\]
Table 4 | The 18 protein sequence outliers from Uniprot 2014_06 and their distances from the convex hulls

| Sequence No. | Length | Access ID | Convex hull(s) the sequences fall outside | Distance from point to convex hull |
|--------------|--------|-----------|------------------------------------------|-----------------------------------|
| 1            | 20     | P82867    | Aspartic acid (D)                        | 0.0045                            |
| 2            | 19     | P68214    | Aspartic acid (D)                        | 0.0309                            |
| 3            | 15     | P80612    | Alanine (A)                              | 0.3850                            |
| 4            | 267    | P14918    | Arginine (R)                             | 0.0000042949                      |
| 5            | 150    | P27787    | Phenylalanine (F)                        | 0.00025511                        |
| 6            | 372    | Q5AKU5    | Histidine (H)                            | 0.00013742                        |
| 7            | 105    | Q2RB28    | Leucine (L)                              | 0.000053612                       |
| 8            | 105    | B9G8M3    | Leucine (L)                              | 0.000053612                       |
| 9            | 94     | Q5GB23    | Aspartic acid (D)                        | 0.0022                            |
| 10           | 391    | P46525    | Serine (S)                               | 0.1044                            |
| 11           | 838    | P08489    | Methionine (M)                           | 0.0840                            |
| 12           | 848    | P10388    | Glutamic acid (E)                        | 0.1269                            |
| 13           | 240    | P04702    | Valine (V)                               | 3.0627                            |
| 14           | 240    | P06677    | Glycine (G)                              | 0.000036147                       |
| 15           | 240    | P04703    | Glycine (G)                              | 0.000036147                       |
| 16           | 240    | P06677    | Glycine (G)                              | 0.000036147                       |
| 17           | 267    | P04698    | Glycine (G)                              | 0.000070559                       |
| 18           | 187    | B6U769    | Proline (P)                              | 0.0037                            |

\[
D_k^n(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - 1)^2 \tag{15}
\]

\[
D_k^n(x') = \frac{1}{n} \sum_{i=1}^{n} (x'_i - 1)^2
= D_k^n(x) \left( \frac{(n-1)^2}{n^2} + \frac{n_k - 1}{n_k} + 2 n_k - 1 \right) \tag{16}
\]

\[
D_k^n(x) \left[ \frac{n_k - 1}{n_k} + 2 \frac{n_k}{n_k} \sum_{i=2}^{n} (x_i - x_1) \right] > D_k^n(x)
\]

We can now determine the precise distribution of the amino acid \( k \) where the maximum value of the second normalized moment is assumed.

Theorem 4. Let \( 0 < n_k \leq n \) be fixed positive integers. Let \( 1 \leq x_1 < x_2 < \cdots < x_{n_k} \leq n \) be a distribution of the amino acid \( k \). If \( D_k^n(x) \) attains the maximum value \( D_{k,\text{max}}^n \) among all possible distributions, then \( x_1, \ldots, x_{n_k} \) is of the following form:

Case 1 \( n_k \) even integer

\[
x_1 = x_2 = \cdots = x_{n_k/2} = x_{n_k/2+1} = n + 1 - \frac{n_k}{2} \\text{ and } x_{n_k+1} = n + 2 - \frac{n_k}{2} = \cdots = x_n = n
\]

Case 2 \( n_k \) odd integer

\[
x_1 = x_2 = \cdots = x_{n_k/2} = x_{n_k/2+1} = n + 1 - \frac{n_k}{2} \\text{ and } x_{n_k+1} = n + 2 - \frac{n_k}{2} = \cdots = x_n = n
\]

Proof: We shall only prove the case when \( n_k \) is an even integer since the proof for \( n_k \) an odd integer is the same. We use induction on \( n_k \). Let \( 1 \leq x_1 < x_2 < \cdots < x_{n_k} \leq n \) be any distribution of the amino acid \( k \). In view of Lemma 2 and Lemma 3, we have

\[
D_{k,\text{max}}^n = \frac{n_k^n - n_k^m (n^m - n_k)}{n_k n} = \frac{1}{4} + \frac{n_k^2 - 1}{12 n} - \frac{1}{4} n_k
\]
(2) If \( n_k \) is odd, then in view of Theorem 4
\[
\mu_k = \frac{1}{n_k} \left[ \frac{n_k-1}{2} \sum_{i=1}^{n_k-1} (i-1) + \frac{n_k}{2} \sum_{i=1}^{n_k-1} (n-i - \frac{n_k-1}{2}) + 1 \right]
\]
\[
= \frac{1}{n_k} \left[ \frac{n_k}{2} \sum_{i=1}^{n_k-1} i + \frac{n_k+1}{2} + \frac{n_k-1}{2} \left( n - \frac{n_k-1}{2} \right) - n_k \right]
\]
\[
= \frac{1}{n_k} \left[ \frac{n_k}{2} \sum_{i=1}^{n_k-1} i + \frac{n_k+1}{2} + \frac{n_k-1}{2} \left( 2n - n_k - 1 \right) - n_k \right]
\]
\[
= \frac{1}{n_k} \left[ n_k + \frac{n n_k}{2} - \frac{n}{n_k} \right]
\]
\[
n_k^{-1} D_{2, \text{max}}^{n_k} = \left( \frac{n_k-1}{2} \right)^2 + \sum_{i=1}^{n_k-1} \left( n-i - \frac{n_k-1}{2} \right)^2
\]
\[
= \left( \frac{n_k-1}{2} \right)^2 + \sum_{i=1}^{n_k-1} \left( n-i - \frac{n_k-1}{2} \right)^2
\]
\[
= \left( \frac{n_k-1}{2} \right)^2 + \sum_{i=1}^{n_k-1} \left( n-i - \frac{n_k-1}{2} \right)^2
\]
\[
= \left( n-k \right)^2 \left( \frac{n}{n_k} \right)^2
\]

Therefore,
\[
D_{2, \text{max}}^{n_k} = \frac{1}{4} (c-n_k) + \frac{n_k^2-1}{12c}
\]

Corollary 7. Let \( c \) be the maximum length of the sequences in the dataset. Let \( n_k \), the number of occurrences of the amino acid \( k \), be a fixed odd integer. Then \( D_{2, \text{max}}^{n_k} \) in the data set is given by
\[
D_{2, \text{max}}^{n_k} = \frac{(n_k-1)(n_k+1)(c-n_k)}{4n_k^2} + \frac{n_k^2-1}{12c}
\]

Proof: Let \( f(n) = D_{2, \text{max}}^{n_k} \) for \( n_k \leq n \leq c \). In this case, the minimum value of \( f(n) \) is assuming \( n = \frac{n_k}{\sqrt{c}} \). When \( n \) is larger than \( \frac{n_k}{\sqrt{c}} \), \( f(n) \) increases as \( n \) increases. Since \( \sqrt{\frac{n_k^2-1}{3}} < n_k \leq n \), \( f(n) \) gets its maximum value at point \( c \).

Corollary 8. Let \( c \) be the maximum length of the sequences in the dataset. Let \( n_k \), the number of occurrences of the amino acid \( k \), be a fixed even integer. Then \( D_{2, \text{max}}^{n_k} \) in the data set is given by
\[
D_{2, \text{max}}^{n_k} = \frac{1}{4} (c-n_k) + \frac{n_k^2-1}{12c}
\]

Proof: Let \( f(n) = D_{2, \text{max}}^{n_k} \) for \( n_k \leq n \leq c \). We need to find the maximum value of \( f(n) \). It is easy to show that \( f(n) \) gets its minimum value at point \( n = \sqrt{\frac{n_k^2-1}{3}} \). When \( n \) is larger than \( \sqrt{\frac{n_k^2-1}{3}} \), \( f(n) \) increases as \( n \) increases. Since \( \sqrt{\frac{n_k^2-1}{3}} < n_k \leq n \), \( f(n) \) gets its maximum value at point \( c \).

1. Sanger, F. The arrangement of amino acids in proteins. Adv. Protein Chem. 7, 1–66 (1952).
2. Holm, L. & Sander, C. Mapping the protein universe. Science 273, 595–602 (1996).
3. Levitt, M. Nature of the protein universe. Proc. Natl. Acad. Science USA 106, 11079–11084 (2009).
4. Rackovsky, S. Sequence physical properties encode the global organization of protein structure space. Proc. Natl. Acad. Science USA 106, 14345–14348 (2009).
5. Koonin, E. V., Wolf, Y. I. & Karev, G. P. The structure of the protein universe and genome evolution. Nature 420, 218–223 (2002).
6. Dokholyan, N. V., Shakhnovich, B. & Shakhnovich, E. I. Expanding protein universe and its origin from the biological Big Bang. Proc. Natl. Acad. Science USA 99, 14132–14136 (2002).
7. Povolotskaya, I. S. & Kondrashov, F. A. Sequence space and the ongoing expansion of the protein universe. Nature 465, 922–926 (2010).
8. Yu, C. et al. Protein space: a natural method for realizing the nature of protein universe. J Theor Biol. 318, 197–204 (2013).
9. The Uniprot Consortium. Activities at the Universal Protein Resource (Uniprot) Nucleic Acids Res. 42, D191–D198 (2014).
10. Barber, C. B., Dobkin, D. P. & Huhdanpaa, H. T. The Quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software. 22, 469–483 (1996).
11. Kuhlman, B. et al. Design of a novel globular protein fold with atomic-level accuracy. Science. 302, 1364–1368 (2003).
12. Chan, Y.-L. et al. The third exon of the budding yeast meiotic recombination gene HOP2 is required for calcium-dependent and recombinase Dmc1-specific stimulation of homologous strand assimilation. J Biol Chem 289, 18076–86 (2014).
13. Deng, D. et al. Crystal structure of the human glucose transporter GLUT1. Nature 510, 121–125 (2014).
14. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. Nature 463, 457–463 (2010).
15. Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet 11, 345–55 (2010).

Acknowledgments
This research is supported by the USA Natural Science Foundation. (DMS-1120824 to S. S.-T. Yau), National Natural Sciences Foundation of China. (31271408 to S.S.-T. Yau), Tsinghua University start up fund (to S. S.-T. Yau) and Tsinghua University independent research project grant (to S. S.-T. Yau).

Author contributions
S.S.-T.Y. conceived the two criteria and designed the studies. W.G.M. carried out the data analysis including figures drawing. S.S.-T.Y., W.G.M. and R.L.H. provided the theory in Supplementary Materials. All authors participated in writing up the paper. The final version is done by M.B.

Additional information
Supplementary information accompanies this paper at http://www.nature.com/ scientificreports
Competing financial interests: The authors declare no competing financial interests.
How to cite this article: Yau, S.S.-T., Mao, W.-G., Benson, M. & He, R.L. Distinguishing Proteins From Arbitrary Amino Acid Sequences. Sci. Rep. 5, 7972; DOI:10.1038/srep07972 (2015).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/