scientific reports

Check for updates

OPEN

Introducing and applying Newtonian blurring: an augmented dataset of 126,000 human connectomes at `braingraph.org`

László Keresztes[1,3], Evelin Szögi[1,3], Bálint Varga[1,3] & Vince Grolmusz[1,2✉]

Gaussian blurring is a well-established method for image data augmentation: it may generate a large set of images from a small set of pictures for training and testing purposes for Artificial Intelligence (AI) applications. When we apply AI for non-imagelike biological data, hardly any related method exists. Here we introduce the "Newtonian blurring" in human braingraph (or connectome) augmentation: Started from a dataset of 1053 subjects from the public release of the Human Connectome Project, we first repeat a probabilistic weighted braingraph construction algorithm 10 times for describing the connections of distinct cerebral areas, then for every possible set of 7 of these graphs, delete the lower and upper extremes, and average the remaining 7 – 2 = 5 edge-weights for the data of each subject. This way we augment the 1053 graph-set to 120 × 1053 = 126,360 graphs. In augmentation techniques, it is an important requirement that no artificial additions should be introduced into the dataset. Gaussian blurring and also this Newtonian blurring satisfy this goal. The resulting dataset of 126,360 graphs, each in 5 resolutions (i.e., 631,800 graphs in total), is freely available at the site https://braingraph.org/cms/download-pit-group-connectomes/. Augmenting with Newtonian blurring may also be applicable in other non-image-related fields, where probabilistic processing and data averaging are implemented.

Gaussian blur operations[1–3] are standard tools in image processing for different purposes. In Artificial Intelligence (AI), it is widely used for image-data augmentation, i.e., increasing the size of the training and testing image databases by blurring some parts of the original images and adding the resulting pictures to the dataset[4–6]. The augmented dataset is applicable for pattern recognition tasks with AI tools, and—since no new objects were artificially introduced, just some parts of the images blurred—the integrity and the validity of the image-set will not be hurt. Additionally, in many applications in pattern recognition, the positive or negative identification of patterns is desirable even when some noise (i.e., blurring) is present in the images.

Augmenting can also be applied successfully for the prevention of overfitting in the training of AI applications: introducing small perturbations in the data would help the AI tool to learn the "important" features of the input[7,8].

In non-image datasets, and especially biological datasets, similar augmenting techniques are generally missing. In the present contribution, we describe an augmenting method, which is called "Newtonian blurring", paying tribute to Isaac Newton's Binomial Theorem and binomial coefficients[9]. The Newtonian blurring is introduced and first applied here for augmenting a set of human braingraphs.

Human braingraphs. Human braingraphs (or connectomes) describe the cerebral connections between the anatomically identified areas of the human brain. The human brain, in a simplified setting, consists of gray matter and white matter. Gray matter is—again, with simplifications—formed from the bodies (or somas) of the neurons, and the white matter from the axonal fibers of the neurons: these fibers connect the somas of the neurons. The white matter also contains myelin covers of these fibers, providing electrical insulation of the axonal

nature portfolio

fibers. The gray matter is situated on the outer surface (the cortex) of the brain and also in some sub-cortical regions; the white matter can be found under the cortex, in the internal parts of the brain.

It is a very interesting question to map the neuronal level connections of the brain: in this mapping, a graph is defined, where the nodes correspond to neurons, and an edge connects two neurons if the axon of one of them is connected to the other's dendrite. Unfortunately, no such graph can be measured for the human brain, which has 80 billion neurons: we simply do not have methods for discovering all the connections of such a great number of vertices. The nematode worm *Caenorhabditis elegans* is the only developed organism, for which such a graph is described[10], but it has only 203 neurons. Very recently, it is announced that the neuronal level connectome of the fruit fly *Drosophila melanogaster*[11] is within our reach in one or two years, but today only the connections between some 25,000 neurons are available from the total of 100,000 neurons of the fruit fly brain[12].

The connectome (or the braingraph) of humans can be described today in a coarse resolution: the vertices correspond to around 1000 anatomically identified areas (called ROIs, Regions of Interest) of the gray matter. We write here "areas" instead of "volumes", since, in the cortex, the gray matter is relatively thin, typically 2–4 mm thick. The edges of the graph connect those pairs of vertices, between which a magnetic resonance imaging (MRI) workflow discovers streamlines[13,14]. More exactly, diffusion MRI is capable of identifying the distribution of directional freedom of movement of water molecules in the brain, notably in and around neurons. This distribution is isotropic in the large body (soma) of the neurons, but it is directionally constrained (anisotropic) in and around the long, thin axons that are aligned with each other. Therefore, streamlines can be discovered in the white matter of the brain, and one can construct braingraphs (or connectomes) as follows: two vertices, corresponding to two ROIs, are connected by an edge if at least one streamline is discovered between them. The number of fibers detected is assigned as weights (fiber number) to the edges. For more details, we refer to Refs.[13,15–17].

As a result of a large, NIH-funded research project, the Human Connectome Project (HCP)[18,19], high-quality diffusion MRI data were published from hundreds of healthy, young adults. Applying this resource, together with an integrated toolset[20] for computing braingraphs from the MRI data, our research group successfully introduced numerous mathematical and graph-theoretical techniques into the analysis of the human braingraphs[15,16,21–32].

We have also published the braingraphs, computed by us at the site https://braingraph.org. First, we had made available several graph-sets, based on the HCP 500 Subjects Release[15,16,23], later we have published 1064 human connectomes, each in five resolutions[17], which were based on the 1200 Subjects Release of the Human Connectome Project. For each set and each resolution, the edges of the graphs are weighted by the fiber numbers detected between their endpoints.

The 1064 braingraphs, described in Ref.[17], were computed by using probabilistic algorithms, and typically, any two runs on the same input yielded slightly different outputs. For increasing the robustness, and the reproducibility, we have applied an averaging and extreme-value deleting strategy, as follows.

Basic averaging strategy.

- For all subjects, the tractography step of the processing, which determined the streamlines, connecting the ROIs of the brain, was computed 10 times.
- For each subject and each resolution (i.e., 83, 129, 234, 463, and 1015 nodes), the braingraph of the subject was computed, and ten interim weights were assigned for each edge. The ten interim weights corresponded to the number of fibers detected in the 10 tractography runs, respectively.
- Those edges, which appeared with 0 fibers in at least one of the 10 tractography runs, were deleted.
- For the remaining edges, the maximum and minimum edge weights were deleted, and the remaining eight weights were averaged (by simple arithmetic mean). This value was assigned to the edge as its (final, non-interim) weight.

In Ref.[17] the particular choice of 10 repetitions is analyzed and explained in detail.

In the present contribution, we modify the averaging strategy above and define the Newtonian blurring for braingraphs.

We remark that the averaging process called basic averaging strategy[17] clearly increased the robustness of the results and decreased the variance of the graph weights due to the probabilistic processing. In the introduction of the Newtonian blurring, we also focus on robustness, and we do not intend to add any artifacts to the data: we just slightly perturb the error-correcting, averaging steps of the real, unmodified data.

We also note that even in the augmented graphs, we have used averaging error correction for increasing robustness.

Discussion and results

Newtonian blurring.
Here, we describe the new Newtonian blurring method as a modification of the basic averaging strategy for braingraphs.

- For all subjects, the tractography step of the processing, which determined the streamlines, connecting the ROIs of the brain, is computed 10 times.
- For each subject and each resolution, the braingraph of the subject is computed, and ten interim weights were assigned for each edge. In other words, ten weighted graphs are generated for each subject and each resolution.
- For each subject and each resolution, 7 graphs from the 10 weighted graphs are chosen in every possible way (i.e., $\binom{10}{7} = 120$ ways).

- Those edges, which appeared with 0 fibers in at least one of the 7 chosen graphs, are deleted from all the 7 graphs.
- For each edge, the maximum and minimum edge weights out of the 7 are deleted, and the remaining five weights are averaged (by simple arithmetic mean). This value is assigned to the edge as its (final, non-interim) weight. In other words, from the 7 weighted graphs we prepared one graph in this workflow.

We emphasize the difference in the basic averaging strategy and the Newtonian blurring: in the former the ten tractography runs for a fixed subject and resolution result one graph, in the latter the ten tractography runs result 120 graphs.

We note that for a given subject and a given resolution, if a graph-edge "e" is present in all the 10 tractography runs with a non-0 fiber number, then "e" will be present in all the 120 augmented graphs as well.

If "e" is present with a 0 fiber number in exactly one tractography run out of the 10, then edge "e" will be deleted from the augmented graphs where the 0 weight is chosen, i.e., from $\binom{9}{6}$ graphs, and will be present in the remaining $\binom{9}{7}$ graphs (by the Pascal-triangle identity $\binom{9}{6} + \binom{9}{7} = \binom{10}{7}$).

This way, we augment the starting braingraph-set to a 120-times larger set. The augmented braingraphs do not contain any artificial components: they even contain strict error correction by deleting the extremal edge weights and averaging the remaining five weights.

We point out that Gaussian blurring in image processing introduces "artificial noise" in several parts of an image by—essentially—an intricate pixel-averaging strategy. We have two reasons for calling the new augmenting method "blurring": The first is paying tribute to the Gaussian blurring method. The second is as follows: for each edge, instead of deleting the two extremes and averaging the remaining 8 weights, we choose all the possible 7 from the 10 weights, delete the extremes, and average the remaining 5 edge weights: that is, we use a data-cleaning method (i.e., averaging) even in the augmentation step, but not so strong as it were possible without the augmenting step. Without the augmenting step, 8 weights are averaged; with the augmenting step, only 5 weights are averaged. This weaker averaging strategy possibly corrects fewer inconsistencies in the measured data and, consequently, it could be seen as a "blurring" relative to the case when 8 weights are averaged.

On the deviation of the augmented graphs. The following is a natural question: How diverse are the augmented graphs compared to the starting ones? Clearly, one needs certain diversity in the augmented set for the applications in artificial intelligence since simply repeating the starting graph several times is useless. However, as in the Gaussian blurring, where the blurred images are very similar to the original ones, here we can also find a similar situation: the Newtonian-blurred graphs are also similar to the starting one.

For a distance measure, we apply Jaccard-distance for the edges $E(G)$ of graphs G:

$$J(G_1, G_2) = \frac{|(E(G_1) \cup E(G_2)) - (E(G_1) \cap E(G_2))|}{|E(G_1) \cup E(G_2)|},$$

where $|A|$ denotes the number of elements of set A. The Jaccard distance of any two graphs is a non-negative number between 0 and 1; it describes the fraction of edges in which the graphs differ.

Figure 1 depicts the distribution of the Jaccard-distances of all the pairs formed from the 120 augmented graphs of subjects 101915 and 654350, the two most-similar subjects as measured by Jaccard distance of their (basic averaging strategy) graphs.

We have chosen the most similar graphs in order to show that the augmenting technique not only changes the weights of the edges, but also the edge-sets even in the augmented versions of the two most similar graphs. The same finding from more different graphs would be less striking.

From the 120 + 120 = 240 graphs, we can form $\binom{240}{2} = 28,680$ pairs. The pairs are partitioned into three classes in Fig. 1:

- *Red class* both members of the pair belong to subject 101915.
- *Blue class* both members of the pair belong to subject 654350.
- *Green class* one member belong to subject 101915, the other to subject 654350.

On the x-axis, the Jaccard distance is given; on the y-axis, the count of the pairs of graphs with the given Jaccard-distance (it is a histogram).

In the figure, one can observe that the red, blue, and green classes form Gaussian distributions. Additionally, the expectation of the fraction of the different edges within the 120 graphs of the same origins are a little larger than 0.02, while the expectation of the Jaccard distance in the green class is more than 0.14, i.e., at least 14% of the edges differ.

Therefore, typically, more than 2% of the edges differ in the graphs of the same subject and in more than 14% in graphs of different subjects.

An application example with support vector machines. Here we concisely describe an application of our augmented dataset for training and testing linear Support Vector Machines (SVM)[33], on a simple classification task of braingraphs.

A typical use of data augmentation is preventing the overfitting in the training of AI tools. Here we demonstrate an application of our augmented braingraph set in this aspect.

In the SVM classification, we have k elements from the n-dimensional real space, denoted by $x_1, x_2, \ldots, x_k$, and k labels $y_1, y_2, \ldots, y_k$, where for all i $y_i = 1$ or $y_i = -1$, corresponding to a binary classification. The goal is
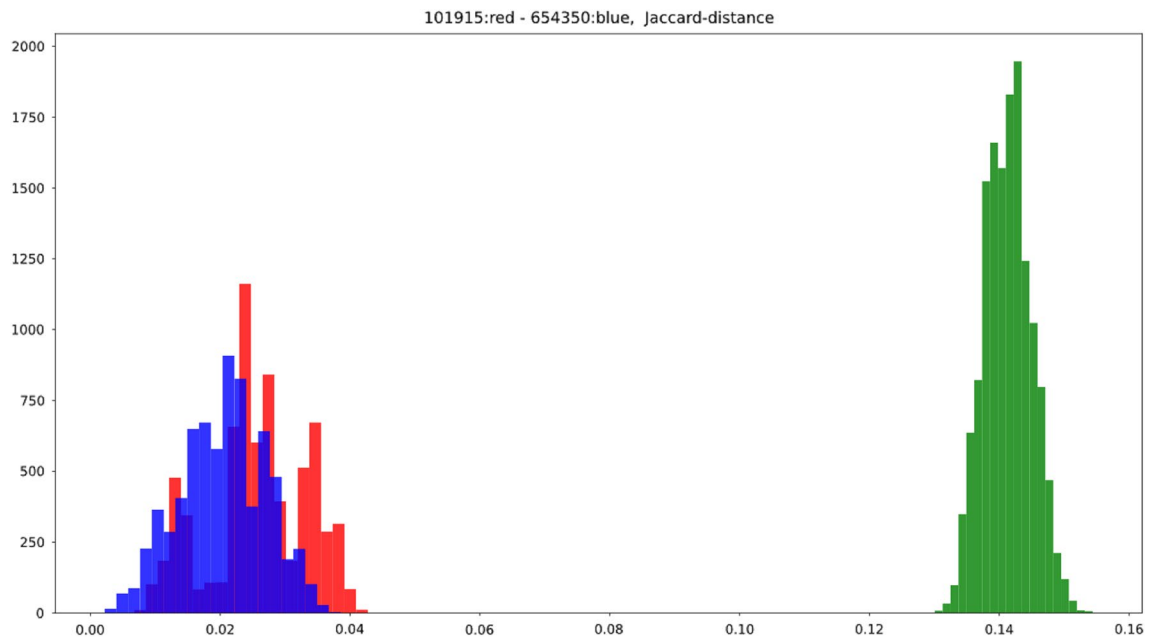
Figure 1. The distribution of the Jaccard-distances of all the pairs formed from the 120 augmented graphs of the two closest (measured in Jaccard distance) graphs from the 1053 ones: the graphs No. 101915 and 654350. From the 120 + 120 = 240 augmented graphs, we can form 28,680 pairs. The pairs are partitioned into three classes: Red class: both members of the pair belong to subject 101915; Blue class: both members of the pair belong to subject 654350; Green class: one member belongs to subject 101915, the other to subject 654350. On the x-axis, the Jaccard distance is given; on the y-axis, the count of the pairs of graphs with the given Jaccard distance is shown (it is a histogram).

finding the n-dimensional hyperspace, which "best" separates the x_i points with $y_i = 1$ from the x_j points, where $y_j = -1$. The hyperspace is characterized by the equation $wx + b = 0$, and ideally, the separation is performed by the sign of $wx + b$, that is, for all i, $\text{sgn}(wx_i + b) = y_i$. From a short computation, the optimal hyperspace is characterized by w and b, satisfying

$$\min_{w,z,b} \left(\frac{1}{2} w \cdot w + C \sum_{i=1}^{k} z_i \right),$$

(1)

where $z_i \geq 0$ and $y_i(wx_i - b) \geq 1 - z_i$.

Here, in (1), $C > 0$, the amount of penalty for wrongly classifying a point. If in (1) we simply repeat all x_i's 120 times, then we need to use $C/120$ instead of the original C if we want to retain the value of the minimum. In the augmented case, when all data points are exchanged to 120 new, augmented data points, we also need to use $C/120$ for comparisons with the original value of (1).

Figure 2 shows the accuracy of applying our augmenting method for linear SVM training and testing for the prediction of the sex of the subjects in our original and augmented braingraph set as the function of $\log C$.

More exactly, in the non-augmented application, we have trained on 737 braingraphs (397 females and 340 males) and tested on 316 graphs (170 women and 146 men); we have applied the 83-node resolution, 0–1 weighted graphs in the example (we need to add that in our related work[34] the fiber-number weighted graphs were used in the SVM classification).

In the augmented application, the original 737 training braingraphs were augmented 120-fold to 120×737 graphs, and the test set is identical to the non-augmented case, it contains 316 graphs.

We remark that a similar advantage of the augmented data set can be demonstrated by logistic regression classification.

Conclusions

We have introduced a new augmenting method for non-image data, called Newtonian blurring. With the new method, we have prepared 631,800 augmented braingraphs for artificial intelligence applications (training and testing) and made the set publicly available at the https://braingraph.org/cms/download-pit-group-connectomes/. One important feature of the Newtonian blurring that it would not introduce any artifacts, any artificial perturbation into the data, and, still, it is capable of considerable augmenting for AI applications. Another important feature of the new method that it can be used for avoiding overfitting and improving classification accuracy, as we have shown in a Support Vector Machine application example.
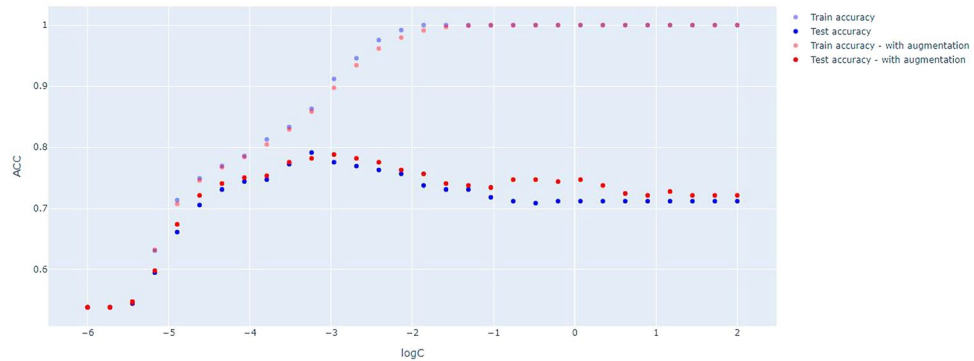
Figure 2. The comparison of train- and test-accuracies with and without augmentation, as the function of the logarithms of parameter C. The figure shows the better test accuracy in almost all domain of log C with augmentation.

Data availability

The data source of this work is published at the Human Connectome Project's website at http://www.human connectome.org[18]. The parcellation data, containing the anatomically labeled ROIs, is listed in the CMTK nypipe GitHub repository https://github.com/LTS5/cmp_nipype/blob/master/cmtklib/data/parcellation/lausanne2008/ ParcellationLausanne2008.xls. The braingraphs, computed by us, can be accessed at the http://braingraph.org/ cms/download-pit-group-connectomes/ site. For each resolution, a compressed archive is available, containing 126,360 graphs. Technical remark: uncompressing an archive containing that number of files takes several minutes, even on fast systems.

References

 1. Mair, B. A., Wilson, D. C. & Reti, Z. Deblurring the discrete Gaussian blur. In *Proc. Workshop Mathematical Methods in Biomedical Image Analysis*, 273–277 (1996).
 2. Cho, C. & Don, H. Blur identification and image restoration using a multilayer neural network. In *Proc. 1991 IEEE International Joint Conference Neural Networks*, Vol. 3, 2558–2563 (1991).
 3. Erdem, A. T. & Tekalp, A. M. Blur identification using bispectrum. *Proc. Signal Process. Int. Conf. Acoust. Speech* **4**, 1961–1964 (1990).
 4. Talukdar, J., Biswas, A. & Gupta, S. Data augmentation on synthetic images for transfer learning using deep cnns. In *Proc. 5th International Conference Signal Processing and Integrated Networks (SPIN)*, 215–219 (2018).
 5. Andriyanov, N. & Andriyanov, D. Pattern recognition on radar images using augmentation. In *Proc. Radioelectronics and Information Technology (USBEREIT) 2020 Ural Symposium Biomedical Engineering*, 0289–0291 (2020).
 6. Lands, A., Kottarathil, A. J., Biju, A., Jacob, E. M. & Thomas, S. Implementation of deep learning based algorithms for diabetic retinopathy classification from fundus images. In *Proc. 4th International Conference Trends in Electronics and Informatics (ICOEI) (48184)*, 1028–1032 (2020).
 7. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *J. Big Data* **6**, 1–48. https://doi.org/10.1186/s40537-019-0197-0 (2019).
 8. Wong, S. C., Gatt, A., Stamatescu, V. & McDonnell, M. D. Understanding data augmentation for classification: When to warp? In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 1–6. https://doi.org/10.1109/DICTA.2016.7797091 (2016).
 9. Newton, I. *Philosophiae Naturalis Principia Mathematica* 1st edn. (Adee, 1687).
10. White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. The structure of the nervous system of the nematode *Caenorhabditis elegans*: The mind of a worm. *Philos. Trans. R. Soc. Lond* **314**, 1–340 (1986).
11. Zheng, Z. *et al.* A complete electron microscopy volume of the brain of adult drosophila melanogaster. *Cell* **174**, 730–743. https://doi.org/10.1016/j.cell.2018.06.019 (2018).
12. Scheffer, L. *et al.* A connectome and analysis of the adult drosophila central brain. *eLife* **9**, e57443. https://doi.org/10.7554/eLife.57443 (2020).
13. Hagmann, P. *et al.* Mapping the structural core of human cerebral cortex. *PLoS Biol.* **6**(7), e159. https://doi.org/10.1371/journal.pbio.0060159 (2008).
14. Cammoun, L. *et al.* Mapping the human connectome at multiple scales with diffusion spectrum MRI. *J. Neurosci. Methods* **203**(2), 386–397. https://doi.org/10.1016/j.jneumeth.2011.09.031 (2012).
15. Kerepesi, C., Szalkai, B., Varga, B. & Grolmusz, V. The braingraph.org database of high resolution structural connectomes and the brain graph tools. *Cogn. Neurodyn.* **11**(5), 483–486 (2017).
16. Szalkai, B., Kerepesi, C., Varga, B. & Grolmusz, V. High-resolution directed human connectomes and the consensus connectome dynamics. *PLoS ONE* **14**(4), e0215473 (2019).
17. Varga, B. & Grolmusz, V. The braingraph.org database with more than 1000 robust human structural connectomes in five resolutions. *Cogn. Neurodyn.* **15**, 915. https://doi.org/10.1007/s11571-021-09670-5 (2021).
18. McNab, J. A. *et al.* The human connectome project and beyond: Initial applications of 300 mT/m gradients. *Neuroimage* **80**, 234–245. https://doi.org/10.1016/j.neuroimage.2013.05.074 (2013).
19. Sotiropoulos, S. N. *et al.* Advances in diffusion MRI acquisition and processing in the human connectome project. *Neuroimage* **80**, 125–143. https://doi.org/10.1016/j.neuroimage.2013.05.057 (2013).
20. Daducci, A. *et al.* The connectome mapper: An open-source processing pipeline to map connectomes with MRI. *PLoS ONE* **7**(12), e48121. https://doi.org/10.1371/journal.pone.0048121 (2012).

21. Szalkai, B., Kerepesi, C., Varga, B. & Grolmusz, V. The Budapest reference connectome server v2. 0. *Neurosci. Lett.* **595**, 60–62 (2015).
22. Szalkai, B., Varga, B. & Grolmusz, V. Graph theoretical analysis reveals: Women's brains are better connected than men's. *PLoS ONE* **10**(7), e0130045. https://doi.org/10.1371/journal.pone.0130045 (2015).
23. Kerepesi, C., Szalkai, B., Varga, B. & Grolmusz, V. How to direct the edges of the connectomes: Dynamics of the consensus connectomes and the development of the connections in the human brain. *PLoS ONE* **11**(6), e0158680 (2016).
24. Szalkai, B., Kerepesi, C., Varga, B. & Grolmusz, V. Parameterizable consensus connectomes from the Human Connectome Project: The Budapest reference connectome server v3.0. *Cogn. Neurodyn.* **11**(1), 113–116. https://doi.org/10.1007/s11571-016-9407-z (2017).
25. Szalkai, B., Varga, B. & Grolmusz, V. The graph of our mind. *Brain Sci.* **11**(3), 342. https://doi.org/10.3390/brainsci11030342 (2021).
26. Kerepesi, C., Szalkai, B., Varga, B. & Grolmusz, V. Comparative connectomics: Mapping the inter-individual variability of connections within the regions of the human brain. *Neurosci. Lett.* **662**(1), 17–21. https://doi.org/10.1016/j.neulet.2017.10.003 (2018).
27. Kerepesi, C., Varga, B., Szalkai, B. & Grolmusz, V. The dorsal striatum and the dynamics of the consensus connectomes in the frontal lobe of the human brain. *Neurosci. Lett.* **673**, 51–55. https://doi.org/10.1016/j.neulet.2018.02.052 (2018).
28. Szalkai, B., Varga, B. & Grolmusz, V. The robustness and the doubly-preferential attachment simulation of the consensus connectome dynamics of the human brain. *Sci. Rep.* **7**, 16118. https://doi.org/10.1038/s41598-017-16326-0 (2017).
29. Fellner, M., Varga, B. & Grolmusz, V. The frequent subgraphs of the connectome of the human brain. *Cogn. Neurodyn.* **13**(5), 453–460. https://doi.org/10.1007/s11571-019-09535-y (2019).
30. Fellner, M., Varga, B. & Grolmusz, V. The frequent network neighborhood mapping of the human hippocampus shows much more frequent neighbor sets in males than in females. *PLoS ONE* **15**(1), e0227910. https://doi.org/10.1371/journal.pone.0227910 (2020).
31. Fellner, M., Varga, B. & Grolmusz, V. The frequent complete subgraphs in the human connectome. *PLoS ONE* **15**(8), e0236883. https://doi.org/10.1371/journal.pone.0236883 (2020).
32. Fellner, M., Varga, B. & Grolmusz, V. Good neighbors, bad neighbors: The frequent network neighborhood mapping of the hippocampus enlightens several structural factors of the human intelligence on a 414-subject cohort. *Sci. Rep.* **10**, 11967. https://doi.org/10.1038/s41598-020-68914-2 (2020).
33. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995).
34. Keresztes, L., Szogi, E., Varga, B. & Grolmusz, V. Identifying super-feminine, super-masculine and sex-defining connections in the human braingraph. *Cogn. Neurodyn.* **15**(6), 949–959. https://doi.org/10.1007/s11571-021-09687-w (2021).

Acknowledgements

Author contributions

L.K. and E.S. invented the novel Newtonian blurring method for braingraphs, B.V. invented the basic averaging strategy, constructed the image processing system, computed the augmented braingraphs, analyzed the deviation of the braingraphs and prepared the figures, V.G. secured funding, analyzed data, and wrote the paper.

Funding

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to V.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.