Title: A growth chart of brain function from infancy to adolescence based on electroencephalography

Authors: Kartik K. Iyer[1, 2]*; James A. Roberts[1]; Michaela Waak[2,3]; Simon J. Vogrin[4]; Ajay Kevat[2,3]; Jasneek Chawla[2,3]; Leena M. Haataja[5]; Leena Lauronen[5] ; Sampsa Vanhatalo[5] ; Nathan J. Stevenson[1]*

Affiliations:

[1] Brain Modeling Group, QIMR Berghofer Medical Research Institute; Brisbane, Australia.

[2] Faculty of Medicine, The University of Queensland; Brisbane, Australia.

[3] Queensland Children's Hospital; Brisbane, Australia.

[4] St Vincent's Hospital; Melbourne, Australia.

[5] BABA Center, Pediatric Research Center, Departments of Physiology and Clinical Neurophysiology, Children's Hospital, Helsinki University Hospital and University of Helsinki; Helsinki, Finland.

*Corresponding authors. Email: Kartik.Iyer@qimrberghofer.edu.au ; Nathan.Stevenson@qimrberghofer.edu.au

One Sentence Summary: Brain age from electroencephalography provides an accurate functional measure of cortical maturation from infancy to adolescence

Abstract:

In children, objective, quantitative tools that determine functional neurodevelopment are scarce and rarely scalable for clinical use. Direct recordings of cortical activity using routinely acquired electroencephalography (EEG) offer physiologically reliable measures of brain function. Here, we develop a novel measure of functional brain age (FBA) using a residual neural network based interpretation of the pediatric EEG. We show that the FBA from a 10 to 15 minute segment of 18-channel EEG during light sleep (stages 1 and 2) in typically developing children and adolescents was strongly associated with chronological age (R^2 = 0.96, 95%CI: 0.94 - 0.96, n = 1062, age range: 1 month to 18 years). The mean absolute error (MAE) between FBA and age was 0.6 years (n = 1062), with an MAE of 2.1 years following validation on an independent set of EEG recordings (n = 723). The FBA detected group level maturational delays in a small cohort of children with abnormal neurodevelopment (p = 0.00053, n = 40). Our work offers a practical, scalable and powerful automated tool for tracking maturation of brain function throughout childhood with an accuracy comparable to that of widely used physical growth charts.

INTRODUCTION

Deriving brain age in relation to a person's chronological age has become a burgeoning focus for scientists, clinicians, and education providers (*1*). The physiological maturation of the brain is fundamentally shaped by an interaction between an individual's genetic traits and acquired/environmental effects (*2, 3*). The chronological age of an individual is a particularly important benchmark in the assessment of pediatric populations, given that measures of brain age, whether functional or structural, are always judged as being 'correct, or delayed for age'.

The best known example of this practice is the use of anatomical growth charts, such as height, weight and head circumference, which provide an important, widely-used tool for early childhood assessment (*4*). In recent times, there has been a surge of alternative physiological and phenotypic measures aimed at capturing developmental trajectories across the lifespan (*5-9*). Biological age surrogates have been advanced by the development of measures across system-wide ageing processes occurring in cardiovascular, metabolic, and brain systems, demonstrating that tracking biological predictors of ageing may provide an important proxy for quantifying health status (*9, 10*).

Assessment of brain age has gained the most attention in recent studies of biological age predictors. Brain age models are predominantly built upon morphometric measurements of brain structure such as cortical thickness, grey matter, white matter, and intracranial volumes (*2, 3, 11-13*) derived from large bio-banking collections of magnetic resonance imaging (MRI) scans. Across these studies, the key biomarker assessed is referred to as 'brain age delta', "brain age gap" or 'brain-predicted age difference' ('brain-PAD'), and defines the difference between predicted brain age and chronological age as a biomarker (*14*). Prior work has also focused on adult populations to show that individual brain age trajectories are associated with

neurodegenerative illness and other conditions associated with compromised neurological health (*15, 16*) and can be used to monitor the response to intervention (*13, 17, 18*) .

The applicability of measures of functional brain maturation is arguably more promising in pediatrics than later parts of the lifespan, due to rapid functional and structural changes occurring in the brain during childhood and adolescence (*19-22*) and given that over 10% of children worldwide (*23*) are affected by neurodevelopmental problems. Anatomical measures of MRI modalities of brain age have emerged in detecting deviations in brain structure (*24-26*), but brain structure is not perfectly correlated with function; the pre-dominant concern for clinicians. The desire to assess function has promoted the use of behavioral assessments such as neuropsychological testing and gross motor tests. While these tests are well-aligned with primary clinical tools that assess core developmental functions, they often succumb to subjective interpretations that affect inter-rater reliability and diverge due to the evolving nature of the toolsets that practitioners employ (*23, 27*). Objective tools are necessary to overcome potentially confounding effects (e.g. anatomical variations unrelated to development) and account for effects of plasticity in the developing brain that accurately reflects the true neurodevelopment stage at the time of the test (*28*). Furthermore, recent studies have also revealed that altered gene expression and genetic mutations, such as those present in trisomy 21, autism spectrum disorder or attention-deficit/hyperactivity disorder, effect brain function at the neuronal level (*29-31*), suggesting that more direct measurement of brain function through techniques such as the electroencephalogram (EEG) provide an opportunity for neurodevelopmental assessment. The EEG's high temporal resolution enables the capture of subtle changes in brain function across dynamic brain states such as evoked responses, resting states and sleep providing valuable insights into typical and atypical maturation during neurodevelopment (*32-35*). Developing

neurodevelopmental biomarkers that are associated with these cellular-level changes to brain functions could thus provide key actionable indicators that guide early intervention and personalization of clinical care to ultimately impact long term outcomes.

Our study aimed to design a 'functional brain age growth chart' (FBA) for children and adolescents based on the cortical activity measured by the electroencephalogram (EEG). The EEG is well-suited for scalable and non-invasive measures of brain function in children, with wide availability in clinical settings. We used EEG activity during sleep to harmonize vigilance states across participants at varying stages of childhood. We collected EEG from three cohorts of children: two large cohorts of typical neurological development to develop and validate the FBA and a small cohort of atypical neurodevelopment to test the diagnostic potential of the FBA. Our modeling approach sought to develop a normative brain age measure with modern machine learning (ML) methods, and rigorous statistical analysis (*36, 37*), based on (i) a deep learning, graph neural network architecture with the raw EEG signals as input and, (ii) a combination of EEG summary metrics linked to spontaneous cortical activity features combined via multivariable regression models. These approaches yielded a 'functional brain age growth chart' of typical development.

RESULTS

The framework for predicting functional brain age (FBA) from the sleep EEG is presented in **Figure 1**. A FBA was developed using a supervised residual neural network architecture (Res-NN). A 'bag of features', Gaussian process regression (GPR) predictor of age was also assessed for further benchmarking of age estimation methods. Our input for the FBA model consisted of 60 second epochs of EEG, where an FBA per recording was calculated as an average across

multiple epochs extracted from a 10 or 15 minute segment. We developed the FBA on a primary training dataset (D1) which comprised 15 minutes of 18-channel EEGs from 1062 subjects recorded at the Helsinki University Children's Hospital, Finland. We then validated our trained model on a dataset (D2) which comprised of 2-channel EEGs recorded from 723 subjects at Queensland Children's Hospital, Brisbane, Australia.

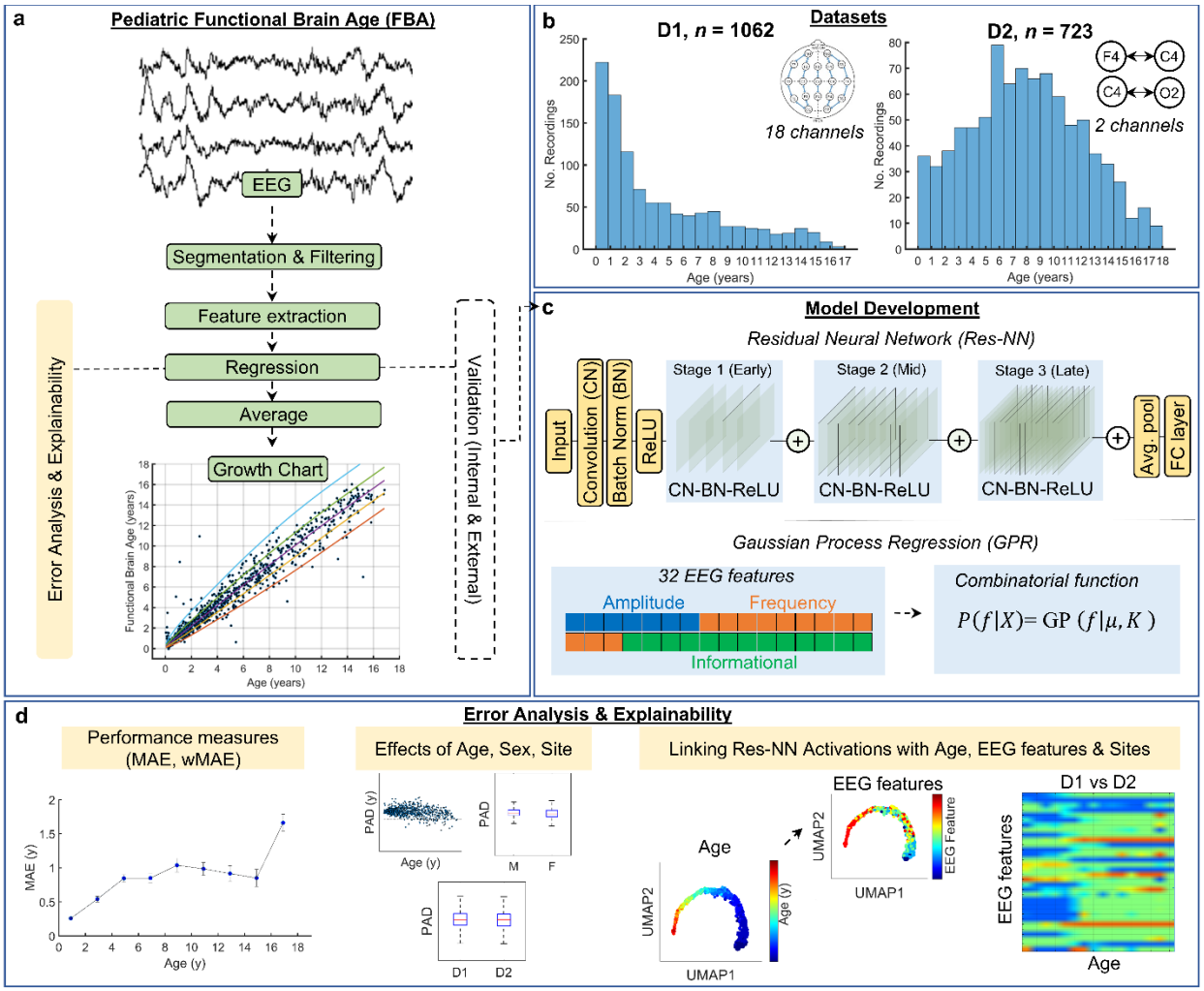


Figure 1 *– Study design.* **a.** EEG acquisition from routine sleep studies from 2 sites (D1 – Helsinki and D2 – Brisbane). The development of a functional brain age (FBA) growth chart within a supervised learning framework. Single, 60 s EEG epochs per recording were used to form a preliminary estimate of FBA that was then averaged across all available epochs within a 10-15 minute segment of EEG. Model evaluation (FBA) involved cross-validation procedures from the primary developmental dataset (D1) with external validation on an independently collected dataset (D2). Growth charts for D1, D2 and a combined D1/D2 dataset were computed. **b.** Distributions for

training (D1) and validation (D2) datasets across age and the EEG channel montages of acquisition. Data consisted of 5 minute epochs of N1 sleep followed by 10 minute epochs of N2 sleep from 18 channels for D1, and 10 minute epochs of N2 sleep from 2 channels for D2. **c.** A trained Residual Neural Network (Res-NN) was our primary method of feature extraction for FBA estimation. Features were combined to form an estimate of FBA using linear regression for Res-NN outputs, with performance benchmarked against a Gaussian process regression model on *a priori* engineered EEG summary measures (*38*); (see Methods for further details). **d**. Performance of the FBA was assessed via measures including the mean and weighted absolute error (MAE, wMAE) and predicted age differences (PAD = FBA minus actual age). Effects of age, sex, and recording site were examined using statistical tests. The behavior of the final trained Res-NN was explained using links between network activation, age, and engineered EEG features across sites.

EEGs from D1 contained recordings that captured the transition between N1 and N2 sleep and were used to optimize the FBA. Each 15 minute segment consisted of 5 minutes of N1 followed by 10 minutes of N2. Recordings in D2 contained 10 minutes of EEG from 2-channels during the N2 sleep state. The performance accuracies of all models tested were evaluated via mean absolute error (MAE), an age-adjusted MAE or weighted (wMAE; the average MAE across discrete age bins), and root mean square error (RMSE).

Functional brain age.

The FBA generated by the Res-NN was the most accurate predictor of chronological age (**Figure 2a-b**; MAE = 0.57 years, weighted MAE = 0.88 years, RMSE = 0.83 years; 10 fold cross-validation on D1, *n = 1062*), providing relatively uniform age predictions across the age range, with a median deviation of 13.9% from chronological age (see **Figure 3a**). Our comparative benchmarking of the Res-NN against a GPR approach and physical growth measures indicated that the neural network architecture exceeded (i) the age estimation accuracy of a GPR model

(Cohen's d = 0.14, p = 1.2 x 10^{-3}, n = *1062*, df = 1060) which had a MAE of 0.78 years (**Supplementary Figure 1c, d**; wMAE = 1.1 years, RMSE = 1.08 years); (ii) the highest-performing individual EEG feature predictor (5th percentile of EEG amplitude; MAE = 1.3 years, wMAE = 1.6 years, RMSE = 1.7 years; **Supplementary Table 1**); iii) a prediction based on head circumference (estimated MAE = 1.72 years, wMAE = 2.54 years – see Methods for details); and iv) a prediction based on the mean age (MAE = 3.52 years, wMAE = 4.47 years; n.b. this was performed to determine an upper bound for MAE based on the age distribution of the cohort).

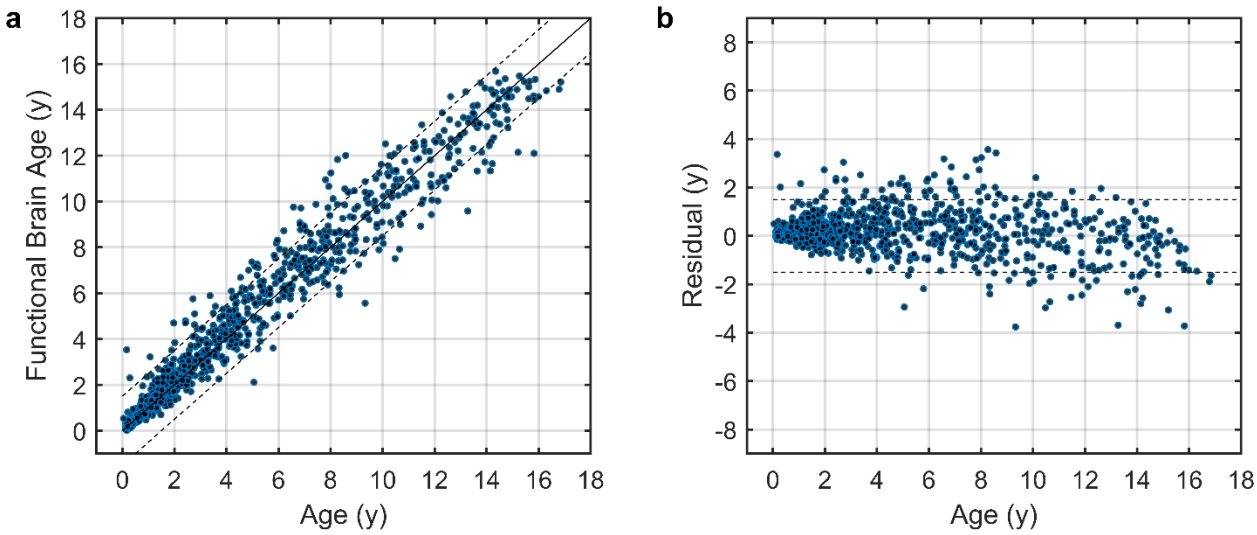


Figure 2 *– Functional brain age estimated in children and performance in training dataset (D1).* **a.** Res-NN model (R^2=0.96, MAE=0.57 years). The dashed line indicates an error bound of ±1.5 years. Individual EEG recordings (blue filled circles) are plotted across ages. **b.** The residual error (predicted age difference; PAD) represents the difference between the FBA measure and chronological age of an individual, in years.

<u>FBA model interpretability.</u>

To ascertain the presence of any bias in the data, we first examined the effects of age and sex on the FBA. The residual error between FBA and chronological age (referred to here as the predicted age difference or PAD) indicated a small bias in FBA estimates as age increased (β = -

0.031, p = 4 x 10^{-9}, n = 1062, df = 1060). There were no significant differences in PAD between males and females (p = 0.1, t-statistic = 1.7, unpaired t-test, n = 1062, df = 1060). Interactions between age and sex did not confound the relationship between PAD and age but a significant interaction with age alone was observed; PAD increased with age (PAD ~ sex: β = 0.06, p = 0.37; PAD ~ age*sex: β = 0.011, p = 0.32, PAD ~ age^2: β =-0.012, p = 3.2x10^{-17}, n = 1062, df = 1060; **Figure 3a**).

We next examined the behavior of the trained Res-NN using a dimensionality reduction method (UMAP (*39*)) (**Figure 1c**; see also **Supplementary Figure 2** for general architectures). We observed that early layers of network activations showed distinct age related clustering, a pattern that becomes increasingly resolved with network depth (**Figure 3b**). We also observed that network activations were related to a range of EEG amplitude, frequency and entropy-based features. We found that EEG features, used in our GPR model were highly correlated with activations present across early, middle and late stages of the NN architecture (median R^2 = 0.48, IQR = 0.31, across all 32 features), suggesting that the Res-NN architecture captures several fundamental time and frequency domain characteristics of the EEG signal that are measureable with independent summary measures within its training phases (see **Supplementary Figure 3** and **Supplementary Figure 4**). These relationships suggest that EEG features can be viewed as data surrogates that track the behavior of neural network activations.

In addition to these tests, we also examined whether EEG electrode location on the scalp and the timing of an epoch during an EEG segment (i.e., with respect to transitions between N1 and N2 sleep) influenced the FBA. FBA accuracy was higher for posterior channels (**Figure 3c**), e.g. central, occipital (average MAE = 0.63 years), whereas anterior channels (e.g. frontal) had lower

accuracies (average MAE = 0.79 years). Furthermore, these spatial differences contributed to prediction accuracy ($p = 3 \times 10^{-24}$, Kruskal-Wallis test). Applying the predictor to data with swapped anterior and posterior channel positions resulted in a reduced performance accuracy (MAE = 1.10 years) whereas data with left hemisphere channels swapped with right hemisphere channels did not alter overall performance accuracy (MAE = 0.55 years). Temporally, the accuracy of the FBA was highest during the transition between N1 and N2 sleep states (**Figure 3d**; MAE = 0.61 years). Taken together, averaging FBA estimates across time and space improved overall accuracy.

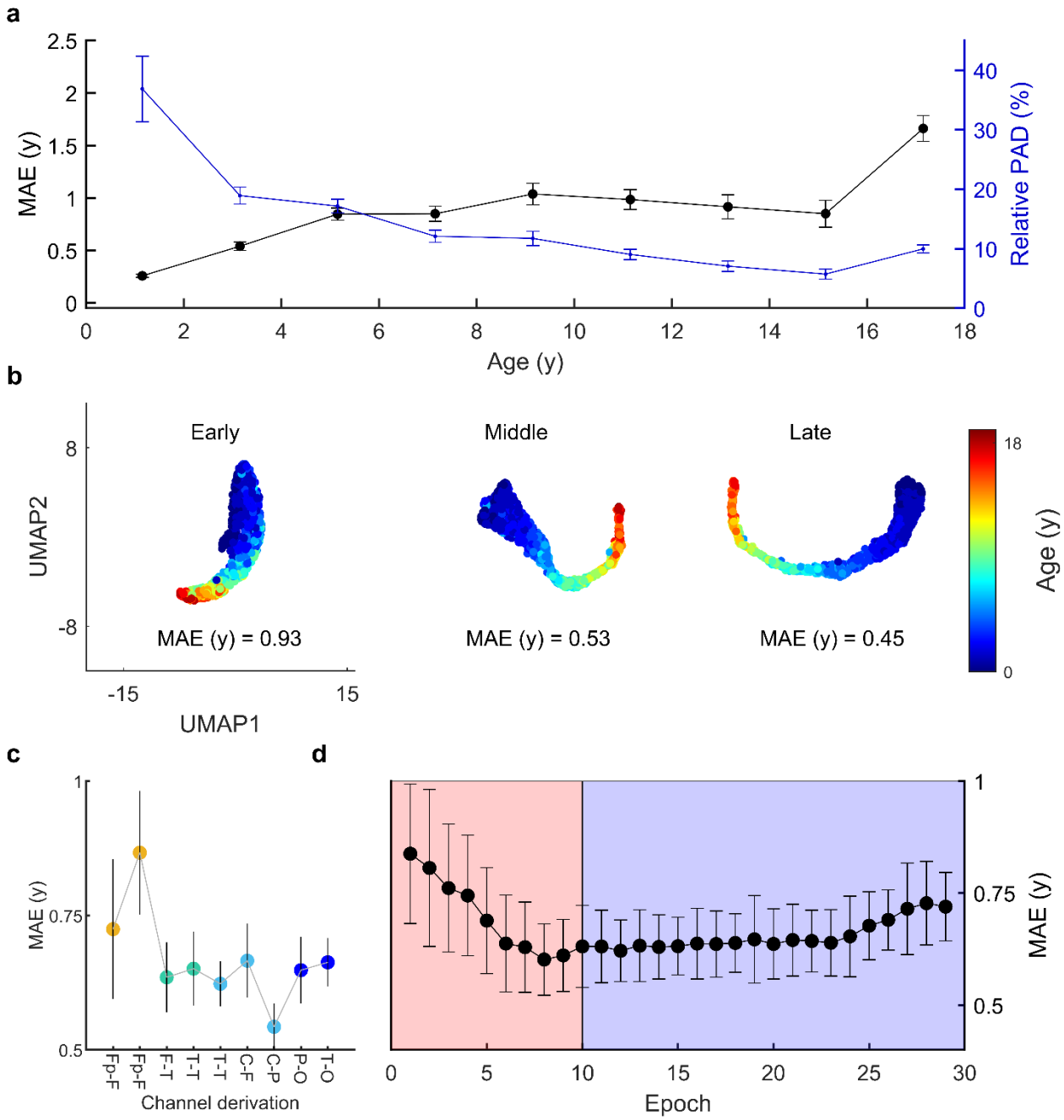


Figure 3 *– Interpreting the EEG-based FBA.* **a.** The mean absolute error (MAE) of the Res-NN (black circles, left axis) across 2 yearly age bins and the relative accuracy of PAD (in blue, right axis) across 2 yearly age bins. For both plots, the mean and standard error of the mean (SEM) are plotted to reflect the sample distribution within age bins. **b.** A lower-dimensional representation of the Res-NN network generated by UMAP on 18-channel D1 data. The network is composed of 63 layers represented by early (layer 6), middle (layer 29) and late stages (layer 61). Here, the Res-NN of the EEG clusters into younger age groups (blue) and older age groups (red) throughout the

training phase, MAE was calculated using UMAP values as predictors (no cross-validation). **c.** The performance accuracy of the Res-NN model (MAE, years) following individual cross-validation per EEG channels. Colors are ordered by right/left channel derivations, where frontal channels (Fp and F in yellow) collectively had higher MAEs than temporal (T in teal), central (C in light blue) and occipital channels (in dark blue) which had comparatively lower MAEs. The average MAE across channel regions is shown with the standard deviation shown as error bars. **d.** The temporal change in MAE of epoch sequences obtained sequentially from N1 (light pink) and N2 (light purple) indicated that the lowest MAEs were observed during a transition between sleep N1 and N2 stages. The MAE is shown with error bars indicating the standard deviation.

Validation.

We then validated the FBA using a secondary dataset (D2) composed of 723 subjects recorded at the Queensland Children's Hospital. To homogenize datasets across the different recording configurations in D1 and D2, we re-trained D1 on a 2-channel bipolar montage (F4-C4, C4-O2) and selected only periods of N2 sleep. In applying this retrained model to the validation set (trained D1 2-channel, tested D2 2-channel), the MAE of the FBA was significantly higher (**Supplementary Figure 5a;** MAE = 2.15 years; wMAE = 2.28 years) when compared to 2-channel models trained and tested on D1 and D2 individually (**Table 1**). Notably though, the validation performance still outperformed the estimated accuracy of simulated head circumference based models (wMAE = 2.54 years; see *Growth Charts* in Methods).

The decrease in FBA performance was attributed to three key site differences: age distribution, number of EEG channels, and site-specific differences of the EEG across age. In comparing MAE to wMAE, distributions of age contribute approximately 0.2 years to the increase in MAE (**Table 1**). Similarly, a reduction in the number of EEG channels from 18 to 2 resulted in an increase of 0.2 years in the wMAE (**Table 1**). Despite the effectiveness of the Res-NN in

capturing fundamental characteristics of the EEG signal (**Supplementary Figure 5b**), several age-specific differences in individual summary EEG features were observed across sites (**Supplementary Figure 5c** and **d.** Additionally, we found noteworthy distinctions in spectral estimates of the EEG recording noise floor between sites, whereby EEG recordings in D1 exhibited a higher frequency noise floor compared to D2 (70 to 80 Hz band power; $p = 3 \times 10^{-9}$, unpaired t-test). Training a FBA model constructed from a combined dataset D1 and D2 considerably improved overall prediction accuracy with a MAE of 1.04 years for D2 (n = 723) suggesting site specific differences were incorporated into the model.

Functional growth charts.

We next combined both datasets (D1 + D2) to generate a FBA 'growth chart', wherein generalized additive models were applied to construct age-appropriate centiles (*4, 40*). The resultant FBA had an MAE of 1.04 years with a wMAE of 1.32 years (**Figure 4a**; see also **Table 1**). Further, FBA growth charts based on an age-stratification of infants (0 to 2 years; MAE = 0.55 years) and children (2 to 18 years; MAE = 1.23 years) indicated a high degree of accuracy relative to their respective age group (**Supplementary Figure 6**).

The practical utility of an FBA is that it enables stratification of children, by quantifying brain functions associated with a child's diagnostic status and underlying neurodevelopmental issues. To demonstrate this, we compared typically developing children from D2 with an additional small cohort of children from the same site whom were diagnosed with Trisomy 21 (n = 40; 29/40 were recorded at less than 7 years of age). The FBAs (trained on D2 only and tested on children diagnosed with Trisomy 21 children) relative to age, despite having normal sleep studies, were found to be significantly lower than typically developing children (**Figure 4b**; p =

0.00053, t-statistic = 3.5, unpaired t-test, n = 763, df = 761). Similar comparisons using the FBA on (i) trained on D1 and tested on D2 + Trisomy 21; and, (ii) trained on D1 + D2 and tested on D2 + Trisomy 21 also indicated significant differences (i): p = 0.012 (t-statistic = 2.5, unpaired t-test, df = 761) and (ii): p = 0.03 (t-statistic = 2.2, unpaired t-test, df = 761). The observations suggest that at the group level, deviant neurodevelopmental trajectories in Trisomy 21 may translate to delayed maturation of their cortical function.

Finally, we benchmarked our FBA model against conventional growth chart trajectories of head circumference and height in children (*41*, *42*). Here, the maximal variation of age for the FBA (difference between 97th and 3rd centiles) falls between head circumference and height for age for a simulated cohort with similar age and sex demographics to D1 and D2 combined (**Figure 4c**). Our FBA growth chart thus exhibits comparable age variability to that of widely-used physical growth charts. Code for converting EEG into FBA and centiles is available at https://github.com/brain-modelling-group/functional-brain-age.

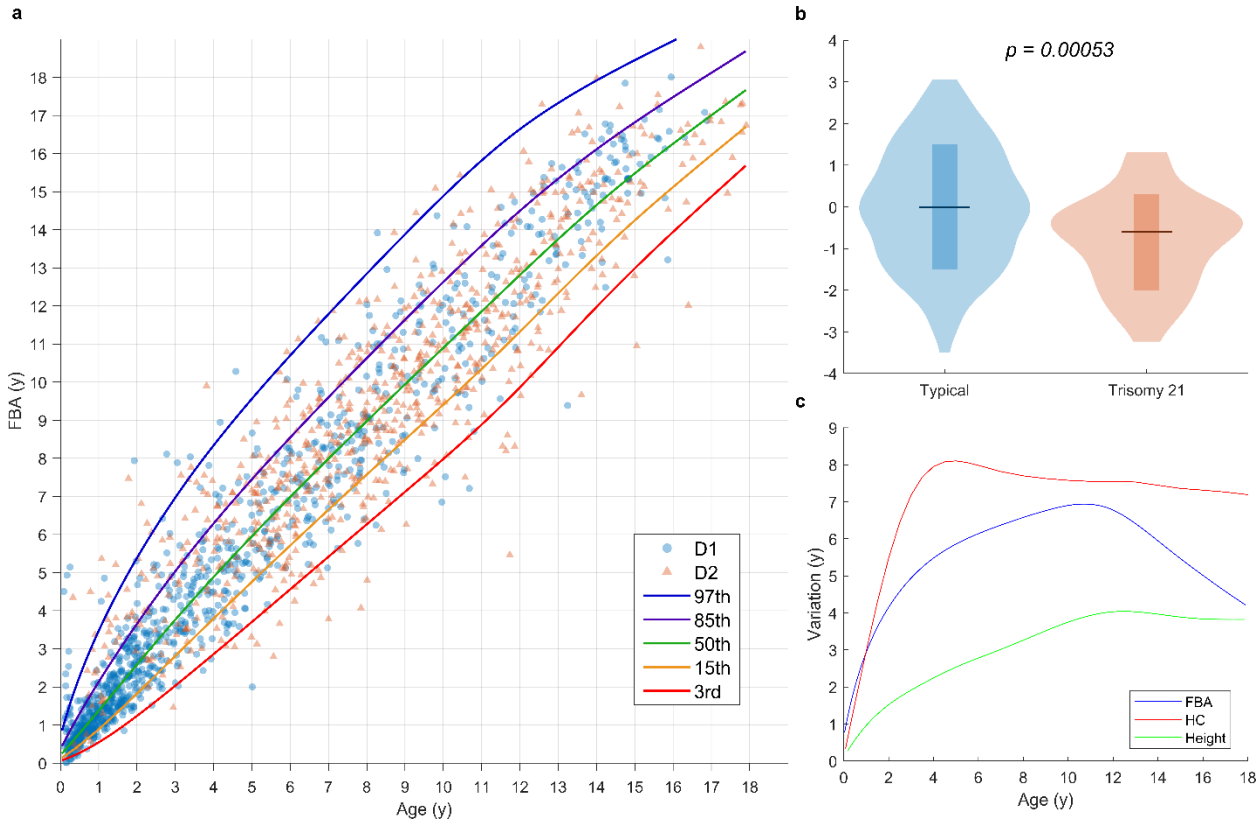


Figure 4 – The functional brain age (FBA) growth chart derived from cross-validation of a combination of D1 and D2 datasets. **a.** The chart is based on the FBA derived from the Res-NN model. Translucent colored triangles represent individual EEG recordings from D1 (blue) and D2 (red) datasets. The 3rd (red), 15th (yellow), 50th (green), 85th (purple) and 97th (blue) centiles are indicated. **b.** Differences in relative PAD between subjects with typically developing neurodevelopment and Trisomy 21. Significance values ($p < 0.05$) were determined by conducting an unpaired t-test between groups, where all data was checked for normality. Typically developing groups (n = 723; blue) and Trisomy groups (n = 40; pink) from D2 are plotted as violin plots, with the median (black line) and interquartile ranges (rectangles) shown, **c.** Variation in pediatric predictors of age. The variation (in years) represents the difference between the 97th and 3rd centiles across age for the FBA (in blue), head circumference (HC, in red) and height (in green). HC and Height predictors were simulated with the same age distribution as the combined D1 and D2 dataset.

Table 1 *– Overall performance of FBA across datasets.* The performances of the FBA, using a Res-NN model, across all training, test, and combined datasets. MAE, wMAE and RMSE are shown. CI is confidence interval, [a]only N2 from D1 used, [b]10-fold cross-validation.

Train	**# channels, # recordings, # epochs**	**Test**	**# channels, # recordings, # epochs**	**MAE (in years; 95% CI)**	**wMAE (in years; 95% CI)**	**RMSE (in years)**
D1	19, 1062, 30798	D1[b]	19, 1062, 31001	0.57 (0.53 - 0.61)	0.88 (0.73 – 1.03)	0.83
D1	2, 1062, 20311	D1[b]	2, 1062, 20311	0.65 (0.60 – 0.69)	1.10 (0.88 – 1.29)	0.88
D1[a]	2, 1062, 20311	D2	2, 723, 13737	2.15 (2.03 – 2.28)	2.28 (1.87 – 2.70)	2.06
D2	2, 723, 13737	D2[b]	2, 723, 13737	1.46 (1.38 – 1.55)	1.73 (1.45 – 2.04)	1.49
D1+D2	2, 1792, 34390	D1+D2[b]	2, 1792, 34048	1.04 (0.99 – 1.09)	1.32 (1.14 – 1.51)	1.35

DISCUSSION

In infants, children and adolescents, it is well known that EEG activity exhibits clear, consistent, and rapid changes with age (*33, 43-45*). We formalize this knowledge using a targeted deployment of deep learning algorithms to form a prediction of EEG age (FBA) underpinned by key human operator expertise and decision-making in specific stages of the process. We achieved an accurate FBA prediction when applying deep neural networks directly to the EEG signal, relying on a summary of only brief epochs (60 seconds) within a 10 - 15 minute segment of routinely recorded sleep EEG during N1/N2. The proposed FBA demonstrated state-of-the-art age prediction accuracy, was validated in an independent cohort and detected group level maturational delays in a small cohort of young children with a defined neurodevelopmental disorder. Tracking maturational delays in young children is of particular interest to clinicians, with the FBA demonstrating its potential for providing an objective, early correlate for neurological development. Our study presents a conceptual advance on prior EEG knowledge by showing that an FBA prediction with modern machine learning and artificial intelligence algorithms facilitates a highly accurate prediction of developmental age, providing similar accuracies with widely used anatomical growth charts (*4, 41, 42*).

We show that the application of deep learning directly to the EEG signal improves the accuracy of the FBA over traditional 'bag of features' approaches. Improved accuracy via these methods suggest that deep neural networks are better able to cope with the non-monotonic relationships between key EEG characteristics such as EEG amplitude, sample entropy, fractal dimension and relative spectral power and age that pose challenges for feature based age predictions (**Supplementary Figure 4**). Our FBA estimates had MAEs comparable to the highest performing MRI-based (*2, 3, 12, 13, 16, 46*) and EEG-based studies (*47-51*) with reported MAEs

in the literature ranging from 1.0 to 4.6 years compared to our best wMAE of 0.88 years. The accuracy of the FBA may be directly attributable to the use of residual neural network architectures over conventional multivariable age regression approaches that are typically used in brain age studies (e.g. GPR). We show that deep learning networks capture well-established EEG characteristics (such as amplitude, frequency, bursting behavior, and entropy) and that the representation of these latent patterns improve FBA estimates through higher-dimensional abstractions of the EEG signal. It is also important to recognize that the fundamental physical and physiological differences in measures of brain function (EEG) versus structure (MRI): the EEG captures neuronal level signatures that reflect rapidly maturing cortical function (*52*), which correlate well with multiple domains of neurocognitive performance and associated neurocognitive conditions (*53, 54*). The underlying gross brain structure measured by MRI is at best only sensitive to larger structural alterations such as malformations of cortical development (*55*).

The present FBA measures were developed solely using large EEG datasets that are routinely collected and widely available in hospitals worldwide. Although the performance error of the FBA increases with age, the relative accuracy of the FBA is comparatively uniform across the age range. Here, additional variations to FBA accuracy were linked to the spatial and temporal organization of the EEG. The effect of spatial organization for instance was primarily a frontal-occipital gradient which is a well-established phenomenon in the maturing EEG within this age group (*56, 57*). Temporally, the accuracy of the FBA was maximum at the onset of N2 sleep characterized by the presence of sleep spindles which are key cortical signatures that emerge in the first few months of life and remain present in the EEG through adulthood (*58*). Waking EEG could also be used in older cooperative children, but collecting several minutes of good quality

EEG signal from alert infants and toddlers is difficult. Thus, the only manual selection done prior to our computational analyses was the identification of the first sleep spindle as a sign of N2 sleep, which was necessary to harmonize vigilance state across a cohort with a wide age range. The N2 sleep state offers a well standardized vigilance state that can be considered much more homogeneous across individuals than the wake or other sleep states (*35*). A brief period of N2 sleep is also often recorded in routine EEG studies as it is rich in EEG signatures, relatively homogeneous across individuals, known to be sensitive for observing pathological phenomena, (such as epileptiform events (*59, 60*)), and is also minimally contaminated by the common artefacts due to subjects' movements. Our additional sleep stage-specific tests (**Supplementary Table 4**) show that deeper sleep stages such as N3 and REM are potentially valuable for further disentangling the relationship between finer grained cortical maturation processes related to growth and homeostatic regulation of sleep during childhood (*61*). Our findings support N2 sleep as an ideal target for the measurement of cortical activity, suggesting that the FBA can be further optimized, providing potential avenues for finer-grained studies of brain maturation.

It is likely that a multitude of age-related changes in the synaptic circuitry underpinning the EEG signal contribute to the high performance of our FBA measure. The development of the EEG signal across childhood is well-established and includes, for instance, notable increases in the posterior dominant frequency during wakefulness and sleep, as well as a decrease in the posterior amplitudes during early childhood (*62, 63*). These developmental changes coincide with maturing brain cytoarchitecture, including changes in neuronal cell density, dendritic growth, myelination, and synaptic remodeling (*52*). However, despite concerted efforts to quantify visually observable changes to the pediatric EEG (*43, 45, 64, 65*), establishing normative values to directly assess brain maturity have largely been challenging. In addition, biophysical

constraints on the underlying scalp EEG in childhood also change due to anatomical changes in non-neuronal structures implicated in EEG propagation, such as skull ossification, cortical thickness, and head circumference (*41*, *52*). Our comparison to head circumference (**Figure 5c**) shows, expectedly, that head size was associated with age, but provides higher variation in age prediction in comparison to the EEG-based FBA.

The present study has some potential limitations. Firstly, the performance of external validation was markedly lower than the overall performance of cross-validated results in each site independently. The drop-off in accuracy due to several factors, namely: site specific differences and differences across age, a lower electrode density, and inherent differences in acquisition of the EEG recordings. We show that site differences in EEG features effect network activations and, therefore, underlying age prediction. There were also differences in the neurodevelopmental profiling at each site. Cohort selection of D2 was based on normal studies of children with no overt neurological illnesses nor pathological EEG activity identified in their overnight PSG, and were not subject to the four-year neurodevelopmental follow up performed in the D1. Nevertheless, despite the trade-off in performance accuracy, externally validated results still outperformed head circumference growth chart measures simulated over the same age range. The combination of datasets largely corrects for these site differences, which suggests that functional growth charts such as the 'normative' FBA proposed, are likely to benefit from assimilating more diversity in datasets across multiple sites (*66*).

Across multiple sites, all children included in this study were identified as typically developing and within clinical neurological norms (except for the trisomy 21 group). However, it is essential to note that these children are not representative of the larger, healthy pediatric population but rather a subgroup of children clinically referred from the primary care level to a tertiary care

center for diagnostic assessment. Estimates of FBA in such groups are clinically interesting; however, it is essential to benchmark the FBA in healthy neurotypical cohorts to ensure further clarity and confidence in applying FBA to a broader pediatric population. Conducting these characterizations across wake and sleep states to establish the dynamic range within which EEG-based brain age estimates can be effectively utilized and adequately interpreted. The FBA measures in this study are also built upon well-resourced diagnostic centers placed within high-income populations, but have equal applicability to providing rapid objective tools (e.g. EEG wearables (*53*)) in lower middle income countries where there is a significant lack of diagnostic resources (*67, 68*). Further studies are necessary to calibrate the FBA-derived growth charts for populations with potentially different traits, including other aspects of population demographics such as race and ethnicity. Continued development of tools such as the FBA in heterogeneous clinical landscapes are likely to improve overall performance, resulting in better delineation of diverse normative population profiles and atypical neurodevelopment.

Secondly, we did not find any significantly confounding effect of sex on our brain age estimates, and have, therefore, generated a single FBA growth chart for both sexes. This does not discount the fact that sex related differences in cortical activity exist across age, rather, that sex specific effects in the EEG were accounted for and adjusted out by the model (*44*). Future applications of the FBA could be used to study genuine sex-related differences in cortical maturation (*44*) by separating data according to biological sex at the training stage. The FBA also appears to compensate for, or is indifferent to, growth spurts, hormonal and pubertal changes in both sexes, and changes to brain structure that occur at differing time points between the sexes, such as increased rates of cortical thinning in males in adolescence (*69, 70*). This ability of trained models to inherently adjust for potential confounders is a key aspect of artificial intelligence (AI)

methods in medicine and means that what we know about EEG and age should be reevaluated singularly in the context of the AI outputs, requiring detailed new studies across a range of exquisitely designed cohorts. Growth charts can also be calibrated for diagnosis, prognosis, and stratification; here the optimal tradeoff between cohort heterogeneity, EEG acquisition, training data size and MAE is not entirely resolved, with evidence from MRI-based studies suggesting that the clinical utility is not necessarily inversely proportional to MAE (*12*, *66*).

The FBA provides an accurate and objective maturational measure of cortical function that is linked to a child's age. Future iterations of the FBA are needed to rigorously validate its performance using independent clinical cohorts. While recognizing FBA's clinical potential, a series of targeted evaluations of the FBA within clinical populations are necessary to determine its efficacy prior to endorsing its widespread use. It is also essential to evaluate the links and dependencies between FBA and the host of complementary modalities of neurodevelopmental assessments, such as standardized neurological and behavioral assessments methods. Rather than replacing or challenging the existing techniques, the EEG-derived FBA is perhaps best seen as a valuable complement to support current neurodevelopmental assessment, offering a tool towards personalization that both benefits the patient and healthcare practitioner alike. Further, the intricacies of AI necessitate a new set of studies that are informed by, but unbiased to previous understanding of EEG maturation in the developing brain. These extra studies are required not only determine the clinical utility of the algorithm but to also enable other researchers and institutions to identify appropriate safeguards that can be built to ensure decision safety and efficacy. We, therefore, publicly release the FBA estimation algorithm to enable continued validation and development via federated learning processes that utilize additional data collected by independent researchers (*71*). Our approach ensures the preservation of earlier feature

extraction layers while retraining later layers using smaller data samples designed to address clinical concerns related to demographics, diseases, or health outcomes. These methods-based outcomes thus generate an 'online' resource that facilitates the continual refinement of targeted algorithms for tracking childhood brain function and neurodevelopment.

MATERIALS AND METHODS

Study Design

Demographics. Our functional brain age chart aimed to cover the pediatric age range following the neonatal period (4 weeks) to adulthood (18 years). Our primary training dataset comprised of electroencephalography (EEG) recordings that were collected as routine sleep EEGs and judged to be typically developing by visual review. These children were later followed up for at least four years after the EEG to confirm typical neurodevelopment, which was established as the absence of major neurological, developmental, or psychiatric diagnoses. The external validation dataset comprised of EEG recordings obtained as a part of routine polysomnography (PSG; overnight sleep study) in hospital-based sleep laboratories. Normal outcomes for the external validation dataset were defined as the absence of sleep disorders based on clinical assessment of the PSG, and absence of neurological diagnoses. Outcomes were defined by clinical experts responsible for the EEG/PSG studies. Our functional brain age chart, hence, only included studies that were defined as having: (i) EEG activity within normal ranges of background activity for age; and (ii) children identified as being typically developing based on neurodevelopmental and/or physical outcomes for their age. Children with any of the following outcomes were

excluded: (i) any seizure-related (epileptiform) or aberrant paroxysmal discharges present in the EEG; (ii) underlying genetic or neurodevelopmental abnormalities disclosed prior to the recording; and (iii) incomplete data due to technical difficulties (e.g. PSG was incompletely recorded as the child did not tolerate the PSG monitoring leads). Site-specific characteristics for each of the EEG datasets are described further below.

Datasets. Our primary training dataset (referred hereto as *D1*) was collected from a convenience sample of EEGs collected at the Helsinki University Children's Hospital in Helsinki, Finland. The original cohort included over 6800 recordings, out of which 3200 were reported to be normal. Following the four-year follow-up period, children who presented with any diagnoses of neurological diseases (including epilepsy), psychiatric disorders, brain-acting medications, presence of tumors/cancer, congenital and/or perinatal disorders and malformations were excluded (2131 recordings). A total of 1062 children with typical neurodevelopment were available for analysis over the 18 year period. For D1, the Institutional Research Review Board at Helsinki and Uusimaa Hospital district approved the study (HUS/244/2021) including waiver of consent due to the retrospective collection of data acquired as part of standard of care. These datasets were acquired and reviewed between 2011 and 2016. The ages of children for D1 ranged between 4 weeks and 17 years of age (median=2.6 years, IQR=6.1 years), with 549 males (min=1.5 months, max=15.8 years; median=2.9 years) and 520 females (min=1 month, max=16.8 years; median=2.5 years). EEG was recorded using 10-20 electrode positions with an electrode on the vertex as reference (either Fz or Cz) using a Nicolet One EEG (Natus Medical Inc. Middleton, WI, USA). All EEG recordings were sampled at 250 Hz and the referential montage was saved in a pseudonymized EDF file format. A total of 19 channels (Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, O2) were collected as part of the

EEG. These recordings were clinically reviewed to define normality of the record, where the most common montage was a standard longitudinal bipolar montage (double banana). Then, the EEG record was re-reviewed by EEG technicians trained for the purpose (and approved by L.L), and the first occurrence of sleep spindle or K-complex was annotated as the onset of an N2 sleep state (see the guidelines of the American Academy of Sleep Medicine, AASM (*72*)). A 15 minute segment was then extracted and saved in EDF format, consisting of 5 minutes before the sleep spindle and regarded as N1 sleep, as well as 10 minutes after the sleep spindle regarded as N2 sleep. It should be noted that these 15 minute segments may include brief arousals or short awake periods that arise during sleep state transitions. As the recordings were done in the routine clinical EEG setting, the only preparatory measure given to patients' families was a request to wake up the child several hours (2-4 hours) earlier than their normal wake-up time to ensure sufficient sleep pressure for falling asleep in the laboratory; however, no external measures of possible individual variations in the sleep pressure could be obtained. This period around N1/N2 transition was selected for FBA training because it allows the best comparability with respect to behavioral states across the whole age range of interest, and because it is often the only state that contains sufficient amount of artefact-free EEG signal in the younger children.

Our secondary dataset, *D2*, was used for external validation of the FBA algorithm, and included 723 children with normal outcomes following a PSG. Ethics approval for data in D2 was granted by Children's Health Queensland (LNR/2021/QCHQ/73595) and the QIMR Berghofer Medical Research Institute (No. P3736, P3727). PSG for these children were used to diagnose the presence of a pediatric sleep disorder (e.g. central sleep apnea, nocturnal hypoventilation, obstructive sleep apnea). For this dataset, we retrospectively screened a convenience sample of 3800 sleep studies from the Queensland Children's Hospital (Respiratory and Sleep Medicine

Department), reviewed between 2014 and 2021. Of those, 3376 were exclusively diagnostic PSG studies, with the remaining group comprising of treatment studies (e.g. oxygen supplementation, ventilator support). From 3800 studies, 1206 individual studies were identified as having a 'normal sleep study' or 'no evidence of sleep disordered breathing'. Out of these 1206 subjects, 465 children were identified as having neurological sequelae (e.g. ASD, Duchenne Muscular dystrophy, spina bifida, spinal muscular atrophy, Trisomy 21, seizure activity, central sleep apnea events higher than 5 events per hour (*73*)). Upon visual inspection of data, we excluded a further 18 subjects for having any direct compromise to EEG activity during sleep segments captured (i.e. EEG channel leads that lost connection). Following exclusion of these subjects, the final cohort included 723 children who were identified as having a normal sleep study outcome for age, without any underlying neurological and/or neurodevelopmental conditions.

The age range for D2 was between 2 months and 18 years of age (median age of 8.1 years, IQR=5.6 years), with 291 females (min= 4 months; max 17.8 years; median=7.9 years) and 432 males (min= 3 months, max=17.9 years, median=8.3 years). As per D1, all sleep stages were seen and scored by a clinician according to the AASM guidelines. PSG was acquired via the EMBLA N7000 (Natus Neuro, Middleton, WI, USA). EEG was recorded using 10-20 electrode positions with reference electrodes attached to the mastoid (A1/M1 and A2/M2). EEG recordings were sampled at 200 Hz or 500 Hz. For D2, a total of 3 channels were recorded overnight (F4, C4 and O2) as part of the PSG. Based on epoch and sleep stage availability, the first 10 minute period of N2 classified sleep was selected from EEG recordings across all ages based on study hypnograms.

The age of subjects in both D1 and D2 was resolved in months.

Comparison of datasets. While the overall age range, sample size, and the essential technical descriptors of the EEG signals are comparable between the datasets, there are also some apparent differences. First, the data from D1 consisted of a full routine EEG setup with 18 bipolar channel derivations, while D2 only included three electrodes with two bipolar derivations (F4-C4, C4-O2). Second, the normality of the EEG signal could be more thoroughly characterized in D1 due to a higher channel count and a longer neurodevelopmental follow-up time, whereas normality in D2 was based on normal-appearing signals, with additional physiological monitoring used to assess sleep-disordered breathing and associated clinical outcomes, as well as the absence of significant neurological illness by the time of PSG study.

Data processing. All EEG data were zero-phase filtered in both forward and reverse directions with an infinite impulse response, bandpass, 4th order Butterworth filter with a lower cutoff of 0.5 Hz and an upper cutoff of 30 Hz (Gaussian process regression) or 15 Hz (residual neural network regression). EEG data were resampled to 64 Hz and 32 Hz as inputs to the feature extraction components of the Gaussian process regression and residual network regression, respectively. Residual network regression approaches employed EEG data inputs at a lower sampling rate to maximize computational efficiency, with several qualitative tests performed to ensure that important aspects of sleep EEG were retained (e.g. delta and alpha rhythms in sleep). Our feature-based methods were examined at a higher sampling frequency (e.g. GPR; **Supplementary Table 1**) and ensured that higher frequency components of the EEG were also captured (e.g. beta activity). For D1, a bipolar montage was computed from the monopolar/referential EEG data resulting in 18 EEG derivations/channels: Fp2-F4, F4-C4, C4-P4, P4-O2, Fp1-F3, F3-C3, C3-P3, P3-O1, Fp2-F8, F8-T4, T4-T6, T6-O2, Fp1-F7, F7-T3, T3-T5, T5-O1, Fz-Cz, and Cz-Pz.

Further, for D1, each 15 minute segment of EEG was divided into 60 s epochs for analysis. Epochs were extracted with a 30 s overlap (29 epochs per recording). We assumed that 60 s was sufficiently long to capture important EEG signal characteristics and sufficiently short to reduce any effects of non-stationarity in the EEG while generating a sufficiently large and diverse set for model training.

Due to the limited availability of channels for D2, we limited our analysis in this dataset to a simplified bipolar derivation of the EEG; i.e., F4–C4, C4–O2. Each 10 minute EEG recording in D2 was segmented into 60 s epochs with a 30 s overlap (19 epochs per recording) and used for training and testing as described in *Cross-validation, optimization and training.*

At the end of these data curation steps, we then developed a residual neural network regression (Res-NN) for age prediction (*74*). We also used Gaussian process regression (GPR) model as a benchmark (GPR) (*75*).

Residual neural network regression. EEG epochs were first resampled to 32 Hz (resampling included anti-aliasing filtering).

We added variability to the residual neural network by changing the temporal filter width (FW), filter channel depth (FD) and filter number (FN) within the convolutional layers as well as increasing the network depth (ND). We used the file *generate_networks_v2.m* to generate networks with different configurations and architectures (see the GitHub page for more details). The general structure of the residual network is shown in **Supplementary Figure 2**.

Several parameters specific to the definition of these neural network architectures were selected during each training iteration. In general, parameters defined the filter size (temporal width and channel depth), filter number and network depth. Training options such as solver type and mini-

batch size were selected based on preliminary analysis (see Supplementary Material for more details; **Supplementary Figures 7 to 14**), and an alternate architecture based on inception layers was also evaluated (*76*).

GPR. This approach combined several summary measures of the EEG features to form a prediction of age. We chose GPR due to its consistent performance across various imaging modalities in estimating brain age following feature extraction due to its capability to model the underlying latent distribution (via Gaussian process prior distribution) and quantifying any associated uncertainty to provide probabilistic predictions from data (*2, 3, 25, 26, 37*). All EEG epochs were resampled to 64 Hz (resampling included anti-aliasing filtering). We used a total of 32 features from the EEG. Of these, 7 features were general measures of EEG amplitude activity, 12 features were frequency dependent representations present within the EEG, and the remaining 13 features were further computations of information content drawn from amplitude, frequency, and entropy based transformations of the EEG signal. A descriptor of the EEG features used is listed in **Supplementary Table 1** and detailed via our GitHub codes for their implementation in MATLAB, see *Code Availability* section).

For both D1 and D2, the set of EEG features was estimated for each available channel and then averaged across channels (with a median operation used). For training, each EEG recording was, therefore, summarized by an M-by-32 feature matrix (where M was 29 epochs per recording for D1 and 19 epochs per recording for D2, and there were 32 features per epoch). A combinatorial function (shown in **Figure 1**) was used to determine the final brain age estimate. This was achieved by first assuming a Gaussian prior (GP) where $P(f\,|X) = \mathrm{GP}(f|\mu, K)$ denotes the dependence of f (i.e. FBA) on X (i.e. age) across all observed data points and μ and K represent the mean function and kernel parameters, respectively.

Performance assessment, cross-validation, optimization, and independent validation.

<u>*Performance assessment:*</u> We utilized two commonly used measures to compare the predicted age to actual age and, therefore, define the accuracy of prediction: mean absolute error and root mean square error. The root mean square error (RMSE, Eq.1) between predicted age per recording and actual age was defined as

$$RMSE = \sqrt{\frac{1}{N}\sum_{\forall n}(a_n - b_n)^2} \qquad \text{(Eq.1)}$$

where a_n is the age of the nth EEG recording in years, b_n is the median predicted age across all extracted EEG epochs per recording, and N is the number of EEG recordings (1062 for D1 and 723 for D2). The mean absolute error (MAE, Eq.2) was defined as:

$$MAE = \frac{1}{N}\sum_{\forall n}|a_n - b_n| \qquad \text{(Eq.2)}$$

We supported these performance measures with weighted MAE. The wMAE is a variation of the MAE that normalizes across the distribution of the data with respect to age; i.e., an approximation of the MAE for a cohort with uniformly distributed age. We defined it as the average MAE across age binned averages of the MAE (bin width = 2 years). We also computed the relative error with respect to age, by calculating the percentage change between predicted age (FBA) and chronological age.

<u>*Cross-validation, optimization and training*</u>:

We used D1 as our primary training set for developing a functional brain age. To evaluate the accuracy of prediction, we used 10-fold cross-validation. This entailed a train-test split where 90% of the cohort of EEG recordings were used for training and 10% for testing. This process was repeated 10 times so that all EEG recordings were excluded from training for at least one iteration (fold).

For residual neural network regression, due to computational limits, initial experiments of randomly selected combinations of network architectures were used to select several model parameters based on maximum regression accuracy (minimum RMSE) on 10-fold cross-validation outputs. These parameters included training solver and mini-batch size. It was assumed that the selection of these parameters was less prone to overfitting than internal network weights from variability in architecture. Architecture parameters were selected using an internal 10-fold cross-validation (9:1 training data split). We used a small fixed grid approach and selected the parameter combination (filter width, filter depth, filter number, network depth) that minimized the RMSE. A single optimal architecture was defined by averaging internal validation RMSE across all 10 training iterations (a point of data leakage assumed to result in negligible overfitting). Remaining training hyperparameters were not optimized: initial learning rate was 0.0001 for training for initial selection and nested architecture selection, training drop factor was 0.1 every 8 epochs, maximum number of training epochs was 50, 10% of recordings within the training set were used for internal validation, and no changes in RMSE on the validation set within 6 training iterations (epochs) stopped training. The squared gradient decay factor was 0.99 for the ADAM solver. Data extraction and network implementation and training was performed in MATLAB (The MathWorks Inc, Natick, Massachusetts, USA: Deep Learning Toolbox;

R2020a or R2021a with GTX 1080 or 3070 graphics cards, for dataset D1 and datasets D1 and D1+D2, respectively).

Additional parameters of the Res-NN model were also benchmarked to gauge performance limits, including epoch length (30 s, 60 s, 120 s and 300 s in length; **Supplementary Figure 7**), the type of solver of min-batch size (**Supplementary Figure 8**), the effect of uniform versus non-uniform age distribution of ages during training (**Supplementary Note 1** and **Supplementary Figure 9**) and various network parameters and architectures tested (**Supplementary Figure 10 to Supplementary Figure 14**). Based on these extensive tests, we use did not use data balancing, but ensured the distribution of age in each training fold was similar with filter width = 9 samples, filter depth = 6 channels, filter number = 16, network depth = 3, with a mini-batch size of 128 samples for our training dataset (D1; 18 channel EEG). For training and testing on 2-channel EEG derivations, further optimization of network parameters yielded a filter depth = 2 channels, network depth = 2, filter number = 2, with the remaining parameters consistent with our training set. These additional robustness tests confirmed that the bipolar montage with 60 s epochs offer optimal FBA estimates.

To explore how the EEG signal was represented by the neural network, we extracted activations of network layers at several stages of the deep learning network when trained and tested on all available data (for this exploratory analysis no cross-validation was used). Here, the optimal network was composed of 63 layers, where an early layer output was taken at layer 6, the middle layer output was taken at layer 29 and the late layer output was extracted at layer 61. A Uniform Manifold Approximation and Project method (UMAP (*39*)) was used to reduce the high-dimensional network activation space into a lower dimensional 2D space (UMAP1, UMAP2), wherein the output space was coded according to age. The use of UMAP here enables network

activations to be condensed into low-dimensional manifolds (UMAP1, UMAP2), while retaining the structural similarity of higher dimensional activations. Visualizations via UMAP allowed us to qualitatively assess how the activations of a network layer cluster with respect to age (**Figure 3b**). To quantify the association between UMAP1 and UMAP2 dimensions and age, we used GPR with UMAP1 and UMAP2 as inputs to predict age (via 10-fold cross-validation). The accuracy of age prediction for each network layer (**Figure 3b**) was then derived. We used the same approach to track how UMAP representations of network activations were linked to individual features of the EEG (**Supplementary Figure 4 and Supplementary Figure 5b**).

For GPR, hyperparameters were selected using Bayesian optimization within a nested cross-validation and included kernel function (e.g. rational quadratic) and sigma values. Shapley values (*77*) were also computed to quantify each feature in terms of its contribution to the overall prediction of FBA (**Supplementary Figure 5c**), extending on the notion of linear model predictions.

Statistical analyses

We first evaluated the prediction accuracy of FBA between residual neural network and Gaussian process regressions. The key variable used in this analysis was the residual error — also referred to as the predicted age difference, PAD (the difference in years between the FBA and chronological age). Comparisons were made using the non-parametric Mann-Whitney rank sum test.

We then analyzed potential confounding effects of age and sex on PAD using multiple linear regression. Here, our multiple regression models for D1 and D2 first tested the effect of sex and age (Eq.3).

$$PAD \sim \beta_0 + \beta_1 age + \beta_2 sex + \beta_1 age * sex + \beta_1 age^2 \qquad \text{(Eq.3)}$$

We also assessed the performance of the FBA across spatial and temporal combinations of the EEG. This included examining changes in prediction accuracy for limited 2-channel FBAs based on bilateral channels of the bipolar montage. A FBA was trained and tested for each bilateral channel and then examined with respect to its designation across the scalp to see if particular regions of the brain were associated with higher accuracy. As a further adversarial test, we also tested variations of channel laterality and channel location in 18 channel D1 data, to gauge changes to performance accuracy with respect to the spatial location of EEG channels. Potential differences in channel laterality were assessed by swapping left hemisphere EEG channels swapped with right hemisphere EEG channels prior to training and testing. The same procedure was used to assess differences in anteroposterior directions, where anteriorly positioned (e.g. frontal) electrodes were swapped with posteriorly positioned electrodes (i.e. parietal, occipital) prior to training and testing. We also examined the temporal variation in FBA across all available epochs, i.e., from the start to the end of the recording period to observe optimal times to evaluate FBAs during N1 to N2 transitions.

We also performed additional sleep stage-specific tests to ascertain the performance accuracy of FBAs across wake, N2, N3 and REM sleep. Using Res-NN and GPR models, we estimated FBAs for all available EEG epochs for subjects in D2 (based on availability of sleep states). Age prediction within N2 sleep states were found to be the most accurate within our data (**Supplementary Table 3**).

Validation: We re-trained models using all available data in D1 and performed an out-of-sample validation on EEG data from D2. As D2 contained a limited number of recording electrodes, we used the same electrodes when training on D1 in a bipolar configuration (F4-C4 and C4-O2). Furthermore, only periods of N2 of D1 (10 minute segments) were included in the training dataset. We compared the mean absolute error between predicted and chronological age from the 10-fold cross-validation results from D1 to D2 to determine if the accuracy of the FBA trained on D1 was preserved when applied to D2. Mann-Whitney rank sum tests were used to compare development versus validation errors. All validation analyses performed for the Res-NN model were repeated for the GPR model (**Supplementary Figure 15**). **Supplementary Table 4** summarizes model performances across test folds for all datasets, respectively.

To test site differences in EEG recordings (**Supplementary Figure 5c**), we compared the distribution of EEG feature values using Kolmogorov-Smirnov tests corrected for multiple comparisons (Bonferroni's method). We used a small random sampling of EEG features within several age defined bins (bin width was 1 year) to minimize the effect of age on feature distribution. Here, differences between EEG features per site and across age were denoted by the number of times the null hypothesis (EEG features between sites were from the same distribution) was rejected across 1000 samplings of each bin (n = 30 samples without replacement per feature per bin). The comparisons were also encoded to delineate if the distributional difference in EEG feature resulted in values greater than expected (older appearing) or less than expected (younger appearing) when comparing D1 to D2. At an acquisition level, we also estimated the total (summed) band power of the raw EEG signal, between 70 and 80 Hz, using Welch's power density spectral estimate, to determine whether

significant differences in the high frequency noise floor were present across sites (unpaired t-test).

Growth Charts: We generated growth charts for the FBA: a chart based on the full 10-20 International System (D1) and a limited 2-channel EEG growth chart for sleep recordings based on the AASM criteria, in a combined D1 and D2 dataset (2-channel EEG). These charts are available in our GitHub repository (see *Data Availability* section). Centiles were estimated using generalized additive models with a protocol similar to the WHO (*4*) child growth charts for height and weight for age. Here, we used the GAMLSS package (*78*) in RStudio (version 1.4.1717). These centiles were optimized by a Box-Cox-power exponential distribution with a cubic spline smoothing function, with distribution GAMLSS parameters *sigma* set as a cubic spline fit over age (df = 3) with *nu* = 1 and *tau* = 1. Centiles were estimated at the 3rd, 15th, 50th, 85th and 97th centiles as per WHO guidelines.

We also compared the performance of our FBA chart with pediatric head circumference and height measures. Using respective reference centiles across the pediatric ranges (*41, 42*), we generated 1000 simulated training and testing cohorts from similar distributions of age, sex and cohort size to the combined D1 and D2 dataset. Using 10-fold cross-validation, we then trained and tested a GPR model with head circumference or height as input, and evaluated the results using MAE and wMAE.

To evaluate the FBA as a tool to determine group-wide differences amongst clinical cohorts, we included an additional cohort of children from D2 who were diagnosed with Trisomy 21 (n = 40). This group of children were chosen as they are known to be predisposed to neurodevelopmental delays. To test this ‘proof-of-concept’, we examined differences between

groups using unpaired t-tests, where variables were checked to ensure normality (*79*) (confirmed via a Lilliefors test (*80*)). The null hypothesis tested was that children with Trisomy 21 will have a PAD that is not different from children with typical neurodevelopment.

All statistical tests employed for analyses were two-sided and the level of significance was 0.05.

References

1. J. Jylhävä, N. L. Pedersen, S. Hägg, Biological age predictors. *EBioMedicine* **21**, 29-36 (2017).
2. J. H. Cole, R. E. Marioni, S. E. Harris, I. J. Deary, Brain age and other bodily 'ages': implications for neuropsychiatry. *Molecular psychiatry* **24**, 266-281 (2019).
3. J. H. Cole *et al.*, Brain age predicts mortality. *Molecular psychiatry* **23**, 1385-1392 (2018).
4. W. H. Organization, W. H. O. N. f. Health, *WHO child growth standards: growth velocity based on weight, length and head circumference: methods and development*. (World Health Organization, 2009).
5. T. Kaufmann *et al.*, Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nature neuroscience* **22**, 1617-1623 (2019).
6. D. W. Belsky *et al.*, Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm. *Elife* **9**, e54870 (2020).
7. D. Karasik, S. Demissie, L. A. Cupples, D. P. Kiel, Disentangling the genetic determinants of human aging: biological age as an alternative to the use of survival measures. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* **60**, 574-587 (2005).
8. J. Van Dongen *et al.*, Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nature communications* **7**, 1-13 (2016).
9. J. Rutledge, H. Oh, T. Wyss-Coray, Measuring biological age using omics data. *Nature Reviews Genetics*, 1-13 (2022).
10. Y. E. Tian *et al.*, Biological aging of human body and brain systems. *medRxiv*, (2022).
11. R. A. Bethlehem *et al.*, Brain charts for the human lifespan. *Nature* **604**, 525-533 (2022).
12. V. M. Bashyam *et al.*, MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain* **143**, 2312-2324 (2020).
13. K. Franke, C. Gaser, Ten years of BrainAGE as a neuroimaging biomarker of brain aging: What insights have we gained? *Frontiers in neurology*, 789 (2019).
14. S. M. Smith, D. Vidaurre, F. Alfaro-Almagro, T. E. Nichols, K. L. Miller, Estimation of brain age delta from brain imaging. *Neuroimage* **200**, 528-539 (2019).
15. S. Shahab *et al.*, Brain structure, cognition, and brain age in schizophrenia, bipolar disorder, and healthy controls. *Neuropsychopharmacology* **44**, 898-906 (2019).
16. J. Lee *et al.*, Deep learning-based brain age prediction in normal aging and dementia. *Nature Aging* **2**, 412-424 (2022).
17. D. Beck *et al.*, Cardiometabolic risk factors associated with brain age and accelerate brain ageing. *Human brain mapping* **43**, 700-720 (2022).
18. A.-M. G. De Lange *et al.*, Multimodal brain-age prediction and cardiovascular risk: The Whitehall II MRI sub-study. *NeuroImage* **222**, 117292 (2020).
19. Y. Chen, T. Z. Baram, Toward understanding how early-life stress reprograms cognitive and emotional brain networks. *Neuropsychopharmacology* **41**, 197-206 (2016).
20. M. d. Onis *et al.*, Development of a WHO growth reference for school-aged children and adolescents. *Bulletin of the World health Organization* **85**, 660-667 (2007).
21. A. B. Archibald, J. A. Graber, J. Brooks-Gunn, Pubertal processes and physiological growth in adolescence. *Blackwell handbook of adolescence*, 24-47 (2003).
22. R. M. Malina, C. Bouchard, O. Bar-Or, *Growth, maturation, and physical activity*. (Human kinetics, 2004).
23. D. E. Astle, J. Holmes, R. Kievit, S. E. Gathercole, Annual Research Review: The transdiagnostic revolution in neurodevelopmental disorders. *Journal of Child Psychology and Psychiatry* **63**, 397-417 (2022).
24. B. Tunç *et al.*, Deviation from normative brain development is associated with symptom severity in autism spectrum disorder. *Molecular autism* **10**, 1-14 (2019).
25. G. Ball, C. E. Kelly, R. Beare, M. L. Seal, Individual variation underlying brain age estimates in typical development. *Neuroimage* **235**, 118036 (2021).
26. C. Kelly *et al.*, Investigating brain structural maturation in children and adolescents born very preterm using the brain age framework. *NeuroImage* **247**, 118828 (2022).
27. P. J. Anderson, A. Burnett, Assessing developmental delay in early childhood—concerns with the Bayley-III scales. *The Clinical Neuropsychologist* **31**, 371-381 (2017).

28. B. Kolb, A. Harker, R. Gibb, Principles of plasticity in the developing brain. *Developmental Medicine & Child Neurology* **59**, 1218-1223 (2017).
29. S. W. MacDonald, L. Nyberg, L. Bäckman, Intra-individual variability in behavior: links to brain structure, neurotransmission and neuronal activity. *Trends in neurosciences* **29**, 474-480 (2006).
30. S. K. Yadav *et al.*, Genetic variations influence brain changes in patients with attention-deficit hyperactivity disorder. *Translational Psychiatry* **11**, 349 (2021).
31. J. O. Edgin, Cognition in Down syndrome: a developmental cognitive neuroscience perspective. *Wiley Interdisciplinary Reviews: Cognitive Science* **4**, 307-317 (2013).
32. C. A. Nelson, J. P. McCleery, Use of event-related potentials in the study of typical and atypical development. *Journal of the American Academy of Child & Adolescent Psychiatry* **47**, 1252-1261 (2008).
33. A. J. Anderson, S. Perone, Developmental change in the resting state electroencephalogram: Insights into cognition and the brain. *Brain and cognition* **126**, 40-52 (2018).
34. M. Angriman, B. Caravale, L. Novelli, R. Ferri, O. Bruni, Sleep in children with neurodevelopmental disabilities. *Neuropediatrics* **46**, 199-210 (2015).
35. C. J. Chu, J. Leahy, J. Pathmanathan, M. Kramer, S. S. Cash, The maturation of cortical sleep rhythms and networks over early development. *Clinical Neurophysiology* **125**, 1360-1370 (2014).
36. A. F. Marquand *et al.*, Conceptualizing mental disorders as deviations from normative functioning. *Molecular psychiatry* **24**, 1415-1424 (2019).
37. S. Rutherford *et al.*, The normative modeling framework for computational psychiatry. *Nature Protocols*, 1-24 (2022).
38. N. J. Stevenson *et al.*, Automated cot□side tracking of functional brain age in preterm infants. *Annals of clinical and translational neurology* **7**, 891-902 (2020).
39. L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, (2018).
40. M. De Onis, A. W. Onyango, WHO child growth standards. *The Lancet* **371**, 204 (2008).
41. A. F. Roche, D. Mukherjee, S. Guo, W. M. Moore, Head circumference reference data: birth to 18 years. *Pediatrics* **79**, 706-712 (1987).
42. R. J. Kuczmarski, *2000 CDC Growth Charts for the United States: methods and development*. (Department of Health and Human Services, Centers for Disease Control and …, 2002).
43. M. Eisermann, A. Kaminska, M.-L. Moutard, C. Soufflet, P. Plouin, Normal EEG in childhood: from neonates to adolescents. *Neurophysiologie Clinique/Clinical Neurophysiology* **43**, 35-65 (2013).
44. A. R. Clarke, R. J. Barry, R. McCarthy, M. Selikowitz, Age and sex effects in the EEG: development of the normal child. *Clinical neurophysiology* **112**, 806-814 (2001).
45. A. Kaminska, M. Eisermann, P. Plouin, Child EEG (and maturation). *Handbook of clinical neurology* **160**, 125-142 (2019).
46. M. J. Lund *et al.*, Brain age prediction using fMRI network coupling in youths and associations with psychiatric symptoms. *NeuroImage: Clinical* **33**, 102921 (2022).
47. M. M. Vandenbosch, D. van't Ent, D. I. Boomsma, A. P. Anokhin, D. J. Smit, EEG□based age□prediction models as stable and heritable indicators of brain maturational level in children and adolescents. *Human brain mapping* **40**, 1919-1926 (2019).
48. S. I. Dimitriadis, C. I. Salis, Mining time-resolved functional brain graphs to an EEG-based chronnectomic brain aged index (CBAI). *Frontiers in human neuroscience* **11**, 423 (2017).
49. O. Al Zoubi *et al.*, Predicting age from brain EEG signals—A machine learning approach. *Frontiers in aging neuroscience* **10**, 184 (2018).
50. D. A. Engemann *et al.*, A reusable benchmark of brain-age prediction from M/EEG resting-state signals. *NeuroImage*, 119521 (2022).
51. H. Sun *et al.*, Brain age from the electroencephalogram of sleep. *Neurobiology of aging* **74**, 112-120 (2019).
52. G. Z. Tau, B. S. Peterson, Normal development of brain circuits. *Neuropsychopharmacology* **35**, 147-168 (2010).
53. A. Lau-Zhu, M. P. Lau, G. McLoughlin, Mobile EEG in research on neurodevelopmental disorders: Opportunities and challenges. *Developmental cognitive neuroscience* **36**, 100635 (2019).
54. N. Hayashi-Kurahashi *et al.*, EEG for predicting early neurodevelopment in preterm infants: an observational cohort study. *Pediatrics* **130**, e891-e897 (2012).
55. T. J. Silk, A. G. Wood, Lessons about neurodevelopment from anatomical magnetic resonance imaging. *Journal of Developmental & Behavioral Pediatrics* **32**, 158-168 (2011).

56. L. Novelli *et al.*, Mapping changes in cortical activity during sleep in the first 4 years of life. *Journal of Sleep Research* **25**, 381-389 (2016).
57. V. Jaramillo *et al.*, Characterization of overnight slow-wave slope changes across development in an age-, amplitude-, and region-dependent manner. *Sleep* **43**, zsaa038 (2020).
58. M. Grigg-Damberger *et al.*, The visual scoring of sleep and arousal in infants and children. *Journal of Clinical Sleep Medicine* **3**, 201-240 (2007).
59. J. L. Moore, D. Z. Carvalho, E. K. St Louis, C. Bazil, Sleep and epilepsy: a focused review of pathophysiology, clinical syndromes, co-morbidities, and therapy. *Neurotherapeutics* **18**, 170-180 (2021).
60. J. Carpay *et al.*, The diagnostic yield of a second EEG after partial sleep deprivation: a prospective study in children with newly diagnosed seizures. *Epilepsia* **38**, 595-599 (1997).
61. A. Buchmann *et al.*, EEG sleep slow-wave activity as a mirror of cortical maturation. *Cerebral Cortex* **21**, 607-615 (2011).
62. J. W. Britton *et al.*, Electroencephalography (EEG): an introductory text and atlas of normal and abnormal findings in adults, children, and infants. (2016).
63. L. Cornelissen, S.-E. Kim, P. L. Purdon, E. N. Brown, C. B. Berde, Age-dependent electroencephalogram (EEG) patterns during sevoflurane general anesthesia in infants. *elife* **4**, e06513 (2015).
64. P. Laoprasert, *Atlas of Pediatric EEG*. (McGraw-Hill, 2011), vol. 1.
65. J. M. Lee *et al.*, A prospective study of age-dependent changes in propofol-induced electroencephalogram oscillations in children. *Anesthesiology* **127**, 293-306 (2017).
66. T. Hahn *et al.*, From 'loose fitting'to high-performance, uncertainty-aware brain-age modelling. *Brain* **144**, e31-e31 (2021).
67. J. Zhang *et al.*, Moving towards vertically integrated artificial intelligence development. *NPJ digital medicine* **5**, 1-9 (2022).
68. J. Zhang *et al.*, An interactive dashboard to track themes, development maturity, and global equity in clinical artificial intelligence research. *The Lancet Digital Health* **4**, e212-e213 (2022).
69. P. C. M. Koolschijn, E. A. Crone, Sex differences and structural brain maturation from childhood to early adulthood. *Developmental cognitive neuroscience* **5**, 106-118 (2013).
70. R. K. Lenroot, J. N. Giedd, Sex differences in the adolescent brain. *Brain and cognition* **72**, 46-55 (2010).
71. A. Zhang, L. Xing, J. Zou, J. C. Wu, Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, 1-16 (2022).
72. R. B. Berry *et al.*, The AASM manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine* **176**, 2012 (2012).
73. S. Ghirardo, A. Amaddeo, L. Griffon, S. Khirani, B. Fauroux, Central apnea and periodic breathing in children with underlying conditions. *Journal of Sleep Research* **30**, e13388 (2021).
74. K. He, X. Zhang, S. Ren, J. Sun, in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2016), pp. 770-778.
75. C. E. Rasmussen, C. Williams, Gaussian processes for machine learning. *MIT Press. Cambridge, MA* **32**, 68 (2006).
76. C. Szegedy *et al.*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2015), pp. 1-9.
77. E. Winter, The shapley value. *Handbook of game theory with economic applications* **3**, 2025-2054 (2002).
78. M. D. Stasinopoulos, R. A. Rigby, G. Z. Heller, V. Voudouris, F. De Bastiani, *Flexible regression and smoothing: using GAMLSS in R*. (CRC Press, 2017).
79. J. B. Burbidge, L. Magee, A. L. Robb, Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association* **83**, 123-127 (1988).
80. H. W. Lilliefors, On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association* **62**, 399-402 (1967).

Acknowledgements

We would like to acknowledge and thank Dr Paula Sanz-Leon for technical assistance with the deployment of functional brain age software provided on our GitHub repository. For Dataset 1, we thank Mr Jarmo Mäki and other EEG technicians within the HUS diagnostic center in annotating EEGs, as well as Mr Eero Ahtola for assisting with the preprocessing of EEG data. For Dataset 2, we thank all technicians, nurses and doctors involved with data collection, overnight annotations and clinical assessment of polysomnography recordings conducted at the Respiratory and Sleep Medicine Department located within the Queensland Children's hospital.

Funding:

National Health and Medical Research Council (NHMRC) Grant no. 2002135

The HUS Children's Hospital/HUS diagnostic center research funds

Finnish Academy (335788, 332017),

Finnish Pediatric Foundation (Lastentautiensäätiö)

Sigrid Juselius Foundation,

QIMR Berghofer Medical Research Institute, Early Career Seed Funding grant.

Author contributions:

Conceptualization: KKI, JAR, MW, SV, NJS

Methodology: KKI, MW, LL, SV, NJS

Investigation: KKI, SV, NJS

Visualization: KKI, NJS

Funding acquisition: MW, SV, NJS

Project administration: KKI, LL, SV, NJS

Supervision: LL, SV, NJS

Writing – original draft: KKI, JAR, MW, LL, SV, NJS

Writing – review & editing: KKI, JAR, MW, JC, SJV, AK, LMH, LL, SV, NJS

Competing interests: None

Data and materials availability:

Trained Res-NN and GPR models for age prediction are available on our GitHub repository, accessible at: https://github.com/brain-modelling-group/functional-brain-age. Data used to train and validate the FBA will be made available to researchers on reasonable request to S.V., N.J.S. K.I.: data sharing is subject to a material transfer agreement, approved by the legal departments of the requesting researcher and by all legal departments of the institutions that provided data and ethical clearances for the study. EEG feature matrices (z-scored) and subjects ages (quantized by 3 monthly intervals for ages 1 month to 4 years, and 1 year intervals for ages 5 and beyond) for each EEG recording in our training dataset (D1, 18 channel and 2 channel versions) and external validation set (D2, 2 channel version) are also available in our GitHub repository. These files have been pseudonymized.

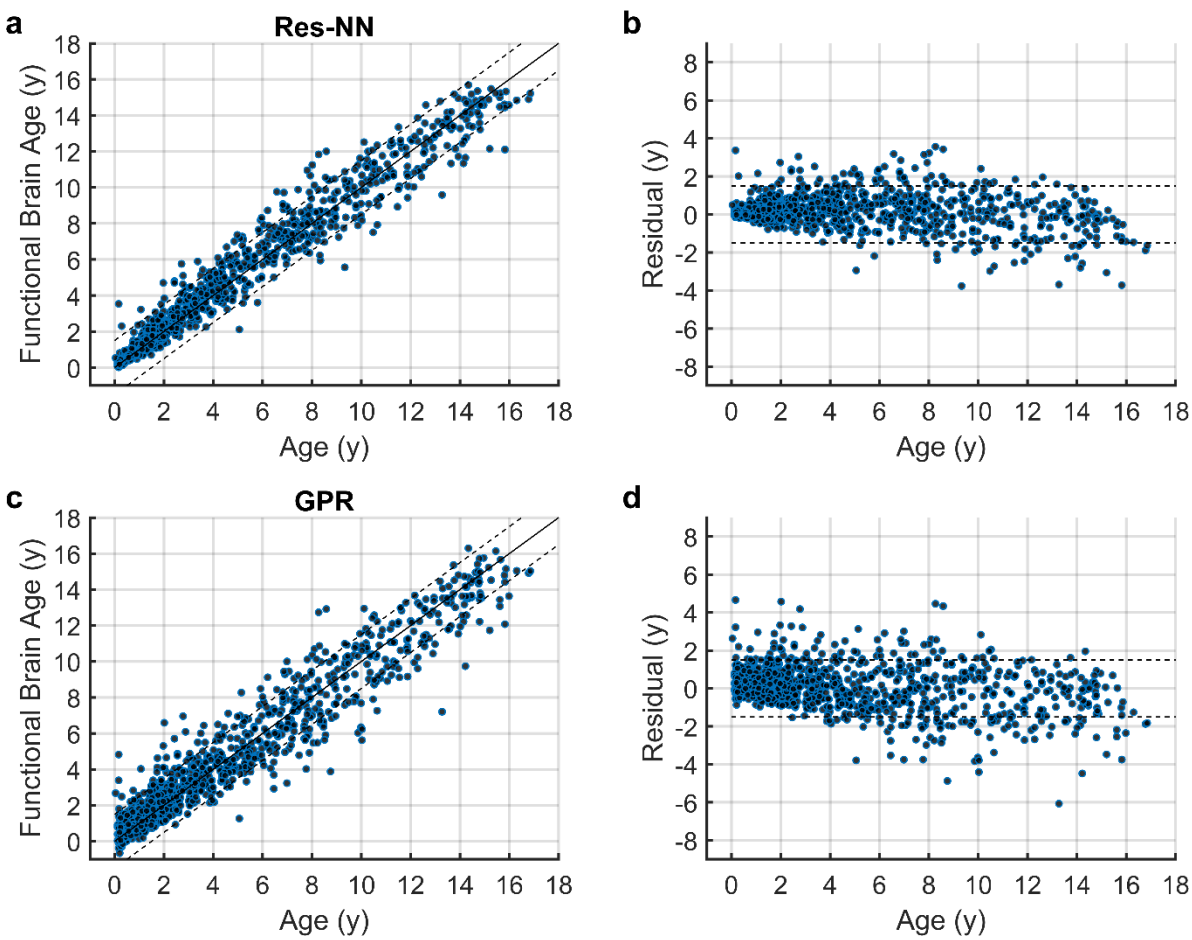
MATLAB and Python code for the estimation of functional brain age by Res-NN based methods on 18- and 2-channel EEG are available on our GitHub repository.

Supplementary Material

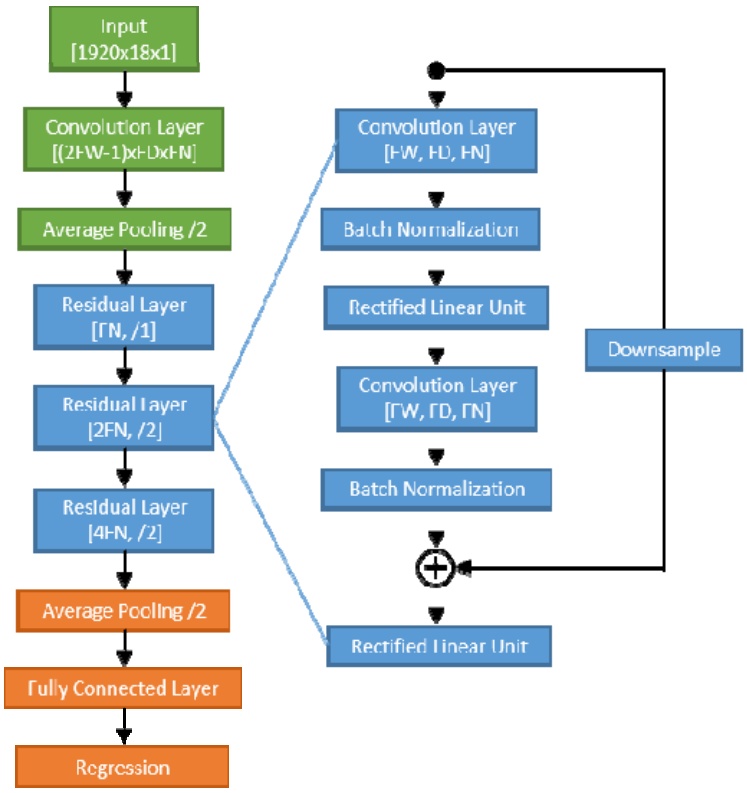
Supplementary Table 1 – Performance of individual EEG features in predicting age following Gaussian process regression (GPR) and 10-fold cross-validation of training data (D1). Results are based on 60 second EEG epochs acquired in the bipolar montage. Amplitude metrics (ID1 to ID7) denote the 5th, 50th (median), 95th centiles based on the absolute value of the Hilbert transform of the EEG (amplitude envelope), along with four central moments of the signal: mean, standard deviation, skewness and kurtosis. Frequency metrics (ID8 to ID19) summarize frequency domain metrics based on the power spectrum density of the EEG signal, and included features such as peak frequency, signal amplitude at this peak, along with mean frequency, bandwidth and relative spectral power in common EEG oscillatory bands. Informational metrics (ID20 to ID32) summarized various non-linear and statistical properties of the EEG. SNLEO stands for the smoothed non-linear energy operator.

		GPR			Shapley	
Feature ID	**EEG Features**	**R^2**	**RMSE**	**MAE**	**Value**	**Rank**
1	5th percentile Amplitude Envelope	0.802	1.698	1.312	0.524	1
2	50th percentile amplitude envelope	0.802	1.700	1.341	0.246	21
3	95th percentile amplitude envelope	0.755	1.832	1.532	0.069	14
4	Mean amplitude envelope	0.773	1.784	1.460	0.187	22
5	Standard deviation amplitude envelope	0.742	1.865	1.577	0.051	26
6	Skewness amplitude envelope	0.183	1.664	3.021	0.113	7
7	Kurtosis amplitude envelope	0.307	2.002	2.714	0.311	31
8	Peak Frequency (PSD)	0.084	1.146	3.298	0.110	17
9	Power in peak frequency	0.145	1.537	3.236	0.044	2
10	Signal amplitude of peak frequency component	0.781	1.762	1.500	0.185	13
11	Mean frequency	0.555	2.111	2.200	0.023	19
12	Bandwidth	0.581	2.101	2.079	0.159	4
13	Relative spectral power (delta 1, 0.5-2Hz)	0.194	1.699	3.153	0.231	10
14	Relative spectral power (delta 2, 2-4Hz)	0.401	2.091	2.322	0.401	15
15	Relative spectral power (theta, 4-8Hz)	0.253	1.842	2.890	0.181	20
16	Relative spectral power (alpha, 8-12Hz)	0.590	2.100	1.978	0.036	12
17	Relative spectral power (beta, 12-30Hz)	0.620	2.073	1.951	0.262	30
18	Total power in all bands	0.763	1.809	1.488	0.075	25
19	Spectral slope (decay in power spectrum)	0.572	2.109	2.093	0.204	6
20	Sample Entropy	0.547	2.119	2.233	0.172	32
21	Fractal Dimension (Higuchi)	0.573	2.101	2.175	0.523	8
22	Spectral entropy	0.568	2.110	2.169	0.372	18
23	Spectral difference	0.786	1.735	1.438	0.031	24
24	Hjorth 1 (Activity)	0.768	1.795	1.478	0.073	3

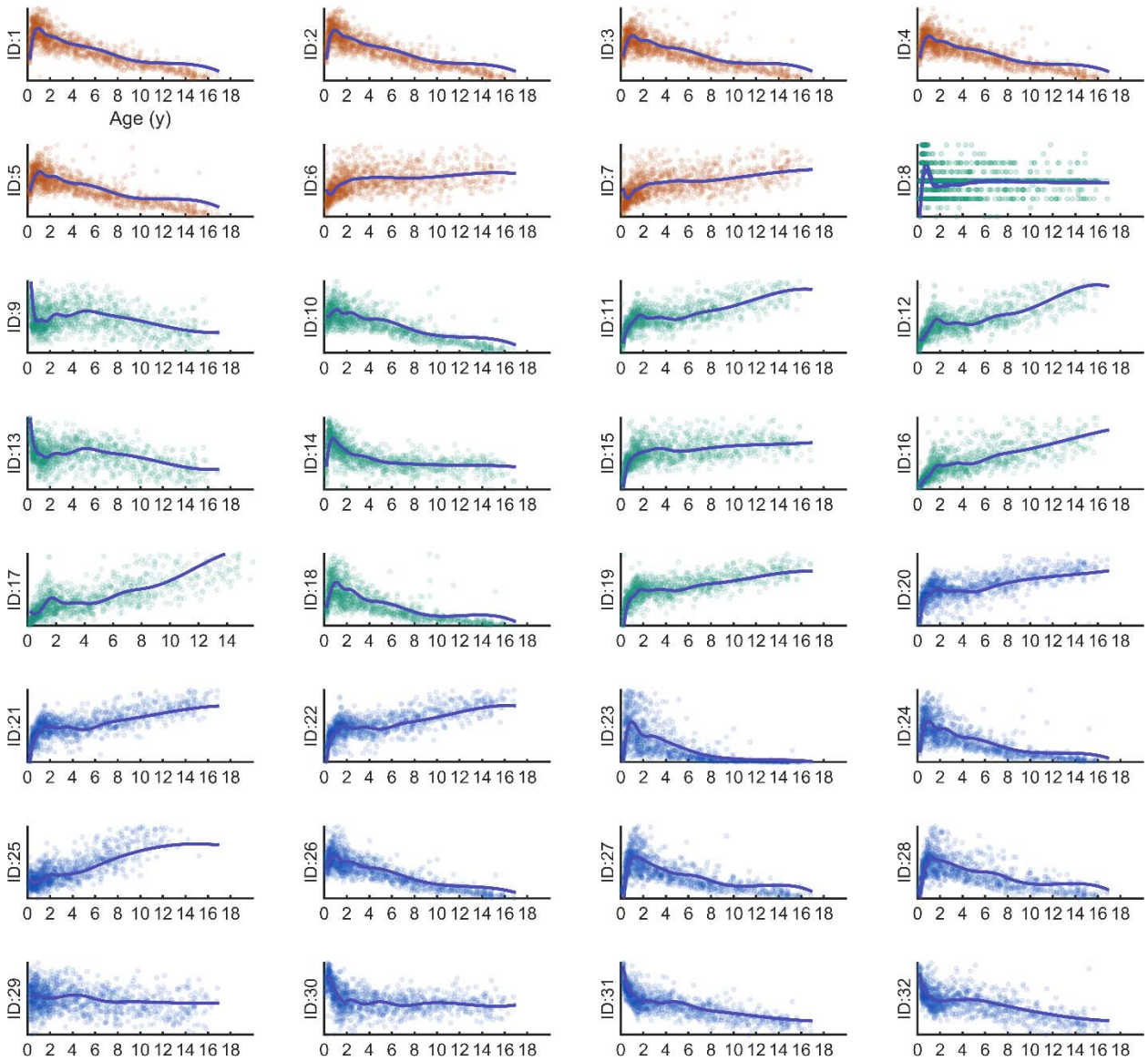
25	Hjorth 2 (Mobility)	0.803	1.693	1.443	0.115	27
26	Hjorth 3 (Complexity)	0.827	1.613	1.324	0.361	5
27	SNLEO mean	0.506	2.150	2.063	0.068	9
28	SNLEO standard deviation	0.408	2.090	2.402	0.007	16
29	Burst shape skewness (all bursts)	0.073	1.124	3.328	0.024	23
30	Burst shape kurtosis (all bursts)	0.207	1.712	2.912	0.118	29
31	Burst duration (mean)	0.677	1.989	1.801	0.305	11
32	Burst duration (standard deviation)	0.536	2.124	2.299	0.112	28



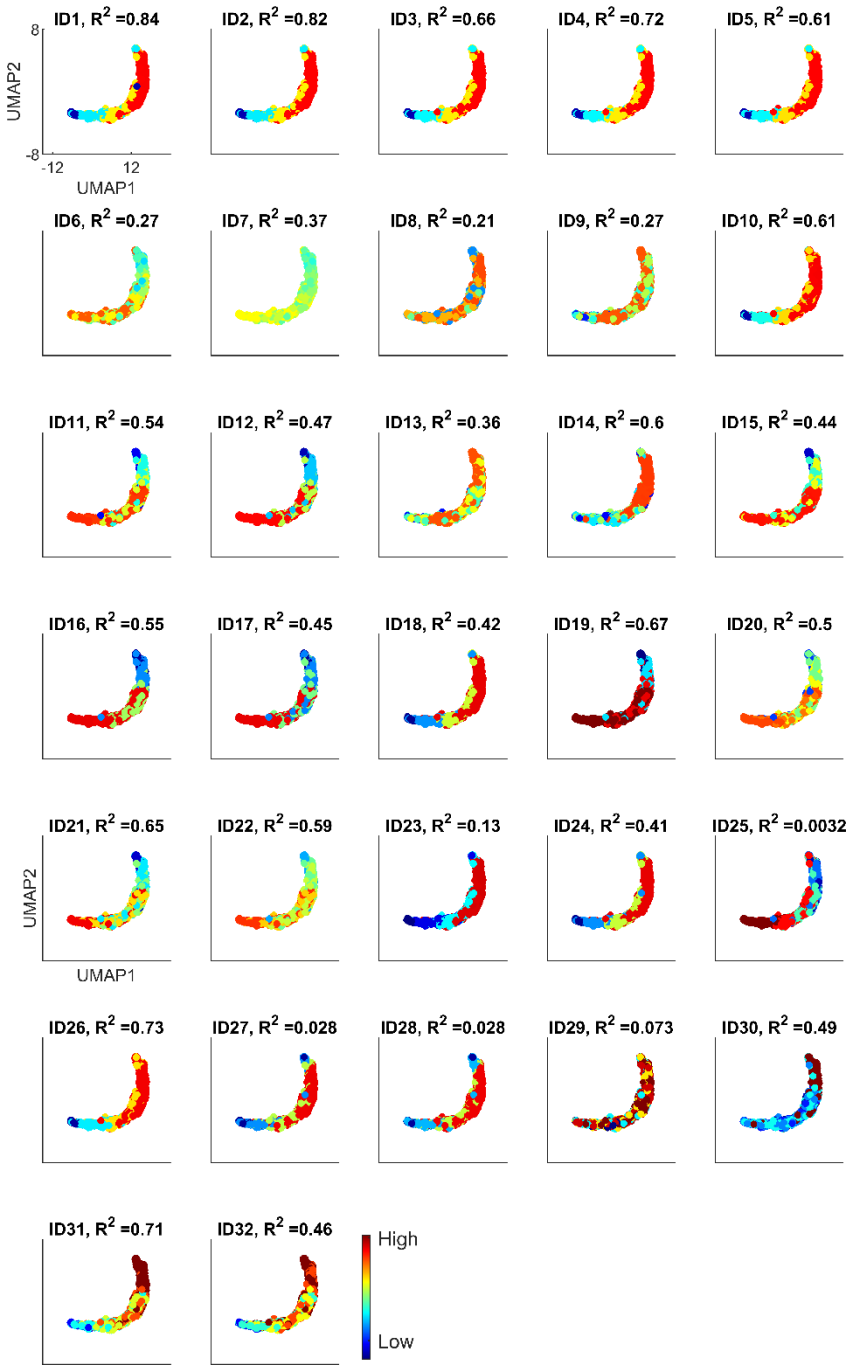
Supplementary Figure 1 – Comparative performance of FBA estimation from a GPR model with the Res-NN **a.** The Res-NN model (R^2=0.96, MAE=0.57 years) with error bounds of ±1.5 years and **b.** The residual error (PAD), in years. **c.** GPR results based on a multivariable model of EEG features (R^2=0.93, MAE=0.79 years). **d.** Residual error (PAD), in years, for the GPR model. Individual EEG recordings (blue filled circles) are plotted across ages.



Supplementary Figure 2 *- The general architecture of the residual neural network used in this work.* For inception networks, we used elements of the inceptionv3 network in MATLAB with a final regression layer. The scale of this pre-trained network was too large to accurately train on the dataset used in our work, so we tested architectural aspects of the network rather than the entire network. Similarly to the residual neural network we added variability by changing the temporal filter width (FW), filter channel depth (FD) and filter number (FN) within the convolutional layers. We used the file generate_networks_v2.m to generate networks with different configurations and architectures (see the GitHub page for more details).

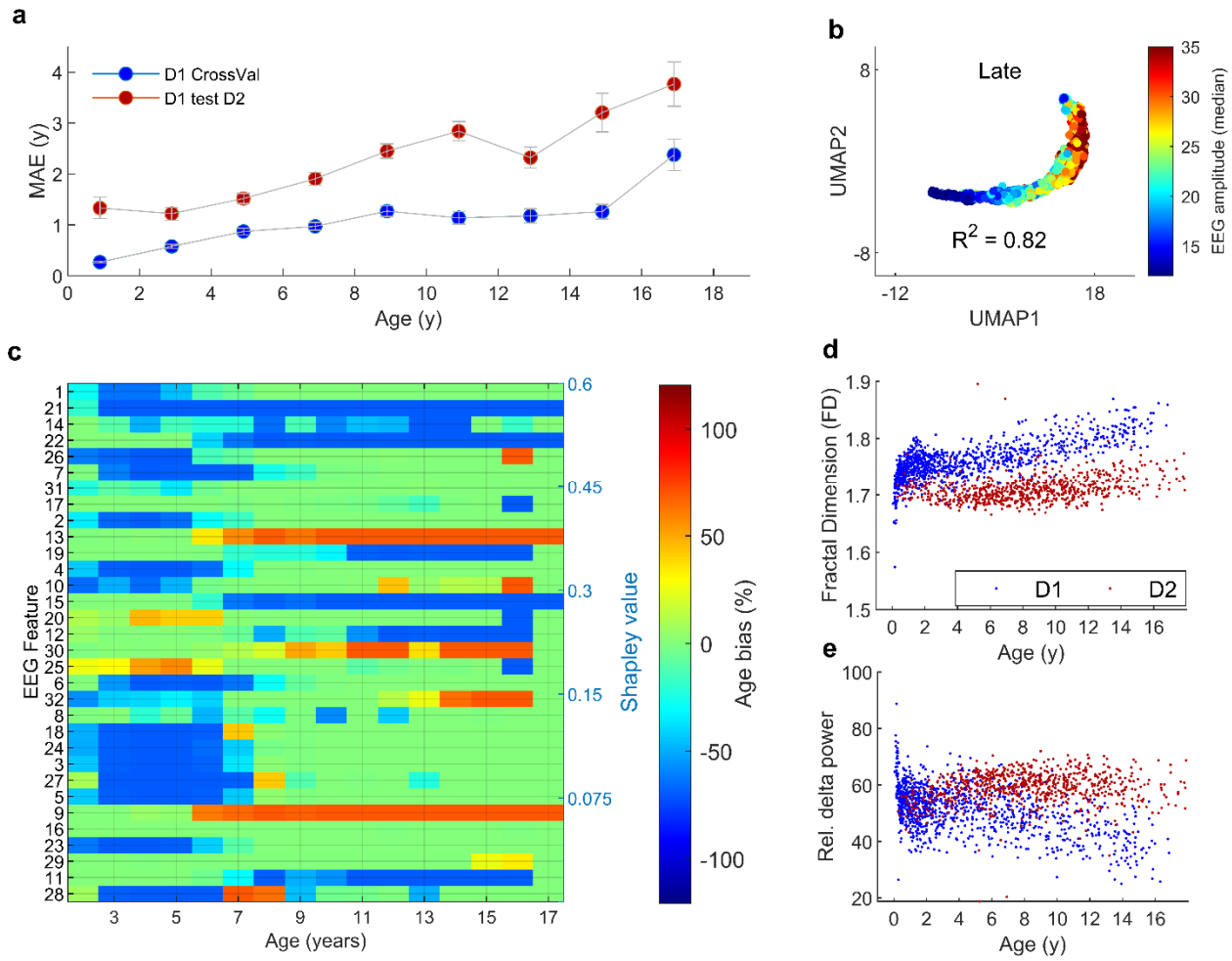


Supplementary Figure 3 *– Individual EEG features versus age (all subjects from training dataset D1).* As per Supplementary Table 1, each EEG feature and their respective ID are based on 60 epochs from the bipolar montage. Feature types are indicated by colors, where amplitude features (reddish hue), frequency domain features (green hue) and informational metrics (blue hue). A spline is fitted across the median values derived across 2 yearly age bins.



Supplementary Figure 4 *– Relating individual EEG features to the late network layer of the Res-NN model.* For each individual feature we used GPR to assess the strength of correlation between individual EEG features with network activation layers (here the late layer, i.e., layer 61 from 2-channel EEG data from D1 is shown) to derive a predicted feature value based on UMAP1 and UMAP2 dimensions derived from this training phase. The strength of correlation (R^2 value) indicates how individual EEG features and predicted individual EEG features are linked by

UMAP representations of the Res-NN model, following 10-fold cross-validation. The color bar indicates the relative range of the EEG feature pertaining to its minimum and maximum values (e.g. for feature ID4, low mean amplitude versus high mean amplitude).

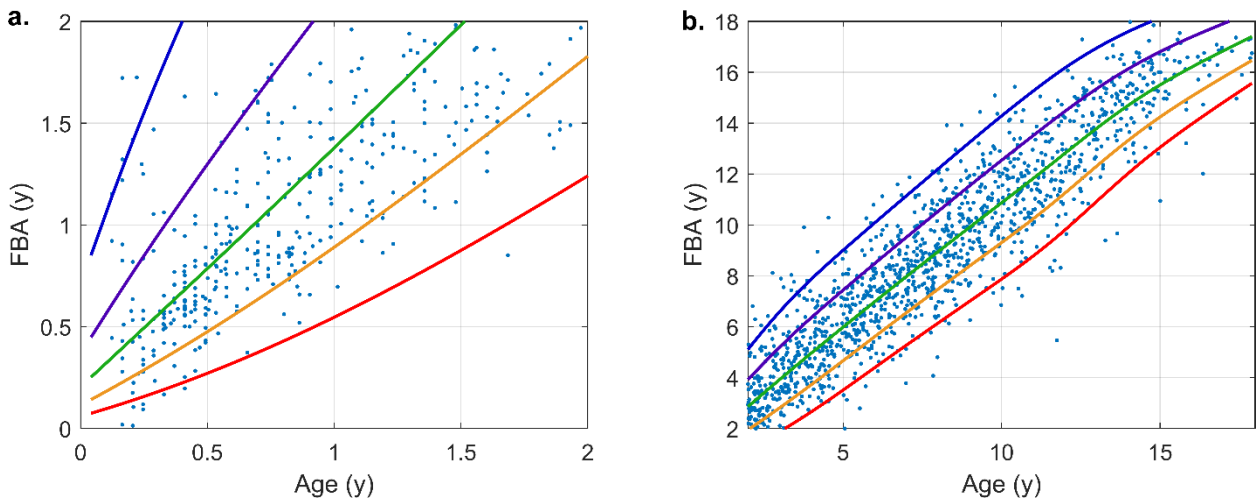


Supplementary Figure 5 *- Validation of the FBA across sites.* **a.** Differences in MAE between D1 and D2: cross-validation (train and test D1 CrossVal, indicated by blue circles) versus external validation on D2 (train D1 test D2, red circles). The MAE across age is higher when testing D1's model on D2. The MAE is shown with error bars indicating the standard deviation. **b.** Mapping EEG features onto the late layer network UMAP (layer 61) derived from 2-channel D1 data; here we show the EEG median amplitude as an exemplar. The R^2 value indicates the strength of correlation between UMAP values and feature value, based on a GPR prediction. **c.** Differences between EEG features and site across age. EEG features (IDs 1 to 32, see also *Supplementary Table 1* for feature names) were ordered by Shapley values (highest to lowest) to indicate the relative contribution of the feature to the overall

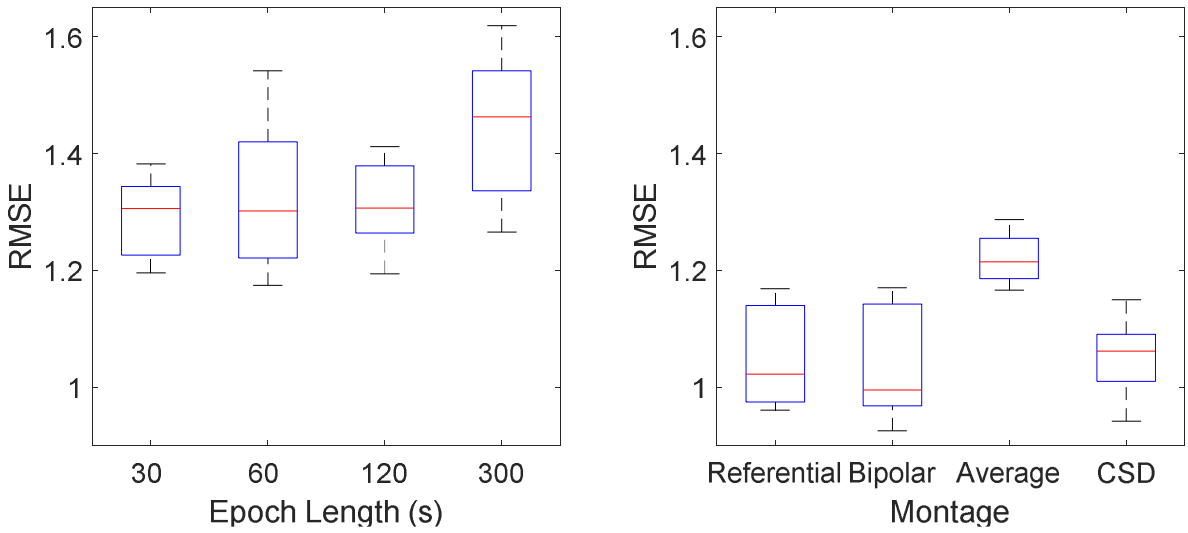
model. Red and blue colors indicate significant age biases following multiple comparisons correction (Bonferroni's method). A positive age bias percentage (red) indicates EEG feature values being 'older' in D2 compared to D1 and a negative age bias percentage (blue) indicates EEG feature values being 'younger' in D2 compared to D1. Approximately 37% of EEG features/age bins combinations differed significantly across age and sites with 191 out of 512 hypothesis tests meeting significance at $p<0.05$ following correction for multiple comparisons. See methods for further details. **d.** Fractal dimension and relative delta power (0.5 to 2 Hz) are exemplar EEG features that show site related differences (D1=blue; D2=maroon).

Supplementary Table 2 – *Overall performance of FBA across datasets.* The performances of the FBA, using a GPR model, across all training, test, and combined datasets. MAE, wMAE and RMSE are shown. CI is confidence interval, [a]only N2 from D1 used, [b]10-fold cross-validation.

Train	**# channels, # recordings, # epochs**	**Test**	**# channels, # recordings, # epochs**	**MAE (in years; 95CI)**	**wMAE (in years; 95CI)**	**RMSE (in years)**
D1	19, 1062, 30798	D1[b]	19, 1062, 30798	0.79 (0.74 – 0.84)	1.10 (0.91 – 1.29)	1.09
D1	2, 1062, 20178	D1[b]	2, 1062, 20178	0.93 (0.88 – 0.99)	1.25 (1.02 – 1.51)	1.24
D1[a]	2, 1062, 20178	D2	2, 723, 13737	1.96 (1.83 – 2.06)	2.55 (2.17 – 2.85)	1.35
D2	2, 723, 13737	D2[b]	2, 723, 13737	1.25 (1.17 – 1.31)	1.34 (1.11 – 1.57)	1.41
D1+D2	2, 1792, 33915	D1+D2[b]	2, 1792, 33915	1.08 (1.04 – 1.12)	1.30 (1.14 – 1.47)	1.36

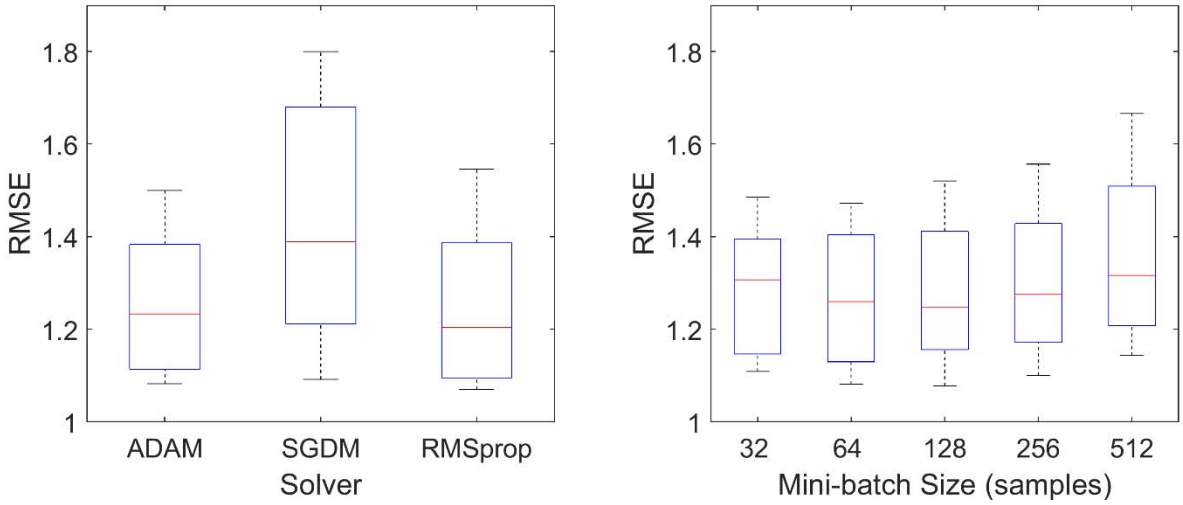


Supplementary Figure 6 *– FBA growth charts stratified into infant and childhood age groups.* **a.** FBA charts for infants (0 to 2 years) and **b.** children and adolescents (2 to 18 years) were generated to offer a magnified view of the FBA with respect to chronological age. In both age-stratified versions of the chart, over 95% of the data (blue dotted points) are captured between the 3rd and 97th centiles estimated. The 3rd (red), 15th (yellow), 50th (green), 85th (purple) and 97th (blue) centiles are indicated.



Supplementary Figure 7 **-** The effect of epoch length and montage on cross-validated RMSE on a random selection of 7 network architectures. An epoch length of 60s and the bipolar montage were selected *a priori*. Different EEG epoch durations (30 s, 60 s, 120 s, 300 s) and EEG montages (referential, bipolar, average, current source density)

were tested. The root mean square error between predicted age (FBA) and age across all testing data from a 10-fold cross-validation was used to determine the optimal selection, with a minimum RMSE indicating the optimal results.

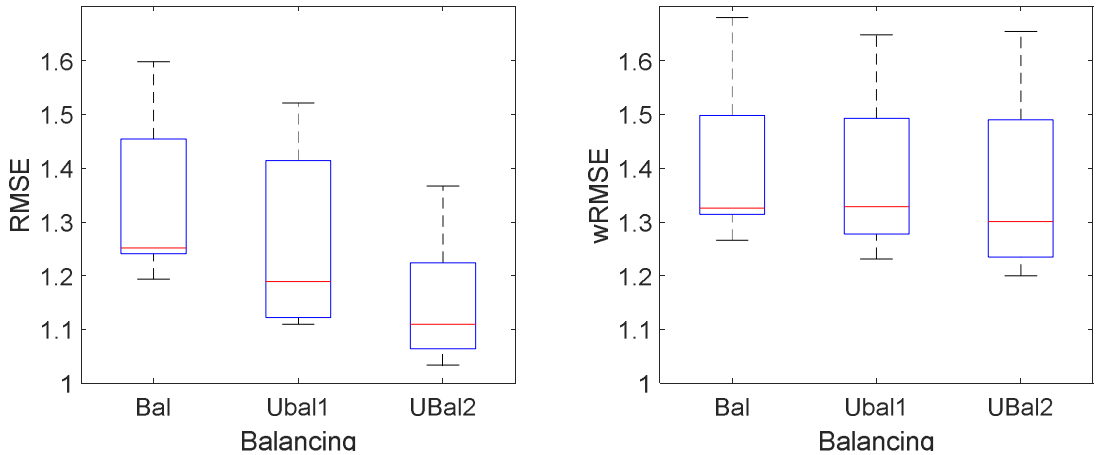


Supplementary Figure 8 **-** The effect of solver and mini-batch size on cross-validated RMSE on a random selection of network architectures, epoch length, and montage (n = 8 per group for solver and n = 7 per group for mini-batch size). For training the neural networks, two hyper-parameters of training were selected based on an initial run on dataset D1 (Solver Type and Mini-batch size). A random selection of network architectures were made and the network was trained and tested within a 10-fold cross-validation with the RMSE on the accumulated left-out test data used to select the hyper-parameters. The ADAM solver and a mini-batch size of 128 samples were selected.

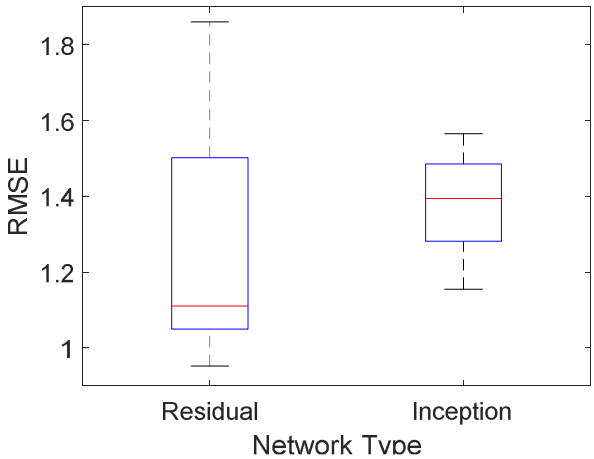
Supplementary Note 1 **- The effect of data balancing on NN results**

Balancing the dataset is an important component of constructing viable training datasets. In this case, the number of recording per age bin should be equal for the best general regression results; increased diversity of data (typically by adding more data) also provides an improved fit. Furthermore, the training data distribution should best represent the distribution of the larger population. We show the effect of data balancing by training 10 randomly selected network architectures with training datasets containing different levels of heterogeneity with respect to age (***Supplementary Figure 9***). To ensure the number of training points is equal we extract more epochs per recording as the dataset becomes more balanced by changing the overlap of epoch extraction. We select three levels of balance: balanced (uniform distribution of subjects across age) extracts at most 40 recordings from each age group (ages are grouped at

yearly intervals), Unbalanced1 (more samples at early ages) extracts at most 80 recordings from each age group, and Unbalanced2 (most samples at early ages) extracts at most 160 recordings from each age group. The RMSE and weighted RMSE (an age adjusted RMSE) are calculated on all EEG recordings that were left out during the 10-fold cross-validation (an unbalanced cohort).



Supplementary Figure 9 *- The general effects of data balancing on cross-validated RMSE for 18-channel data (D1 datasets).* The left plot is the RMSE and the right plot is the wRMSE which is an RMSE averaged across age bins (yearly) rather than per EEG recording. In both cases, the increased diversity in recordings associated with unbalanced training dataset offers lower errors, although the reduction in error is smaller when considering wRMSE.

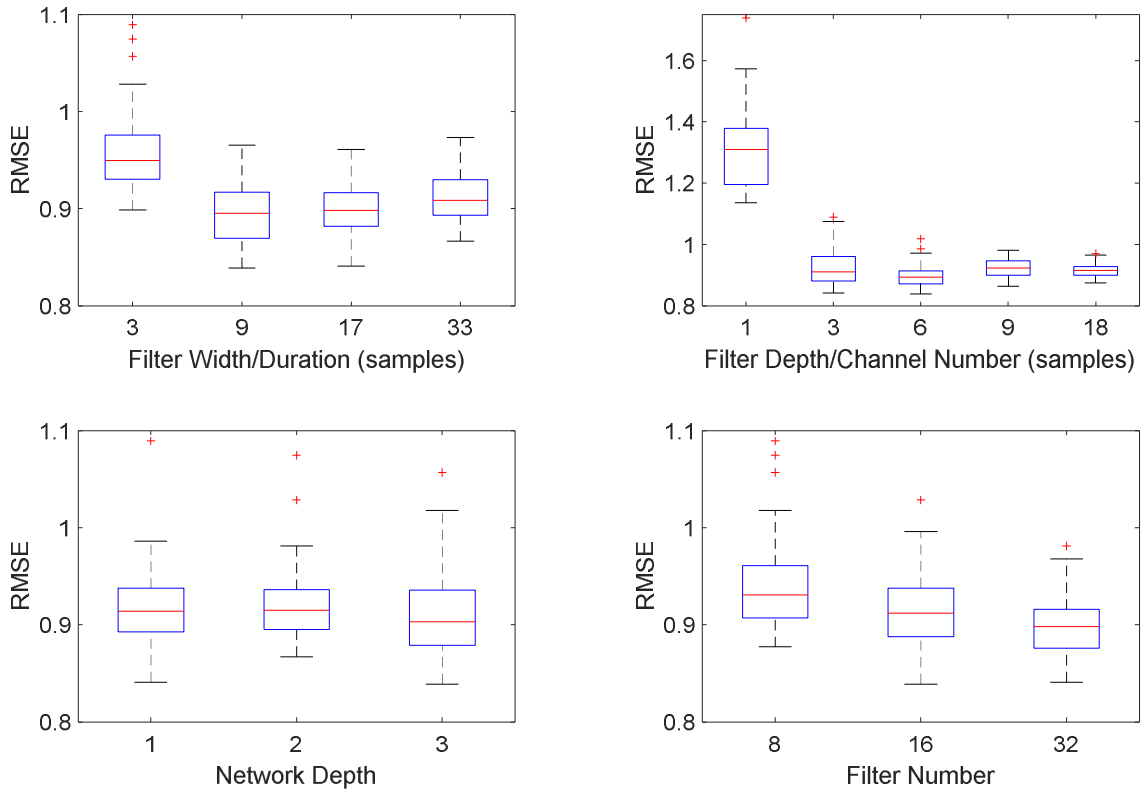


Supplementary Figure 10 *- The general effects of data network architecture type (Residual vs Inception) on cross-validated RMSE for 18-channel data (D1 datasets).* Once general EEG parameters and network training hyper-parameters were initialized, we tested two general architectures based on residual neural networks and inception based neural networks. We randomly selected 10 networks based on both architectures and computed RMSE. Based

on these parameter evaluations, residual networks were selected as the most suitable candidate architecture for estimating age.

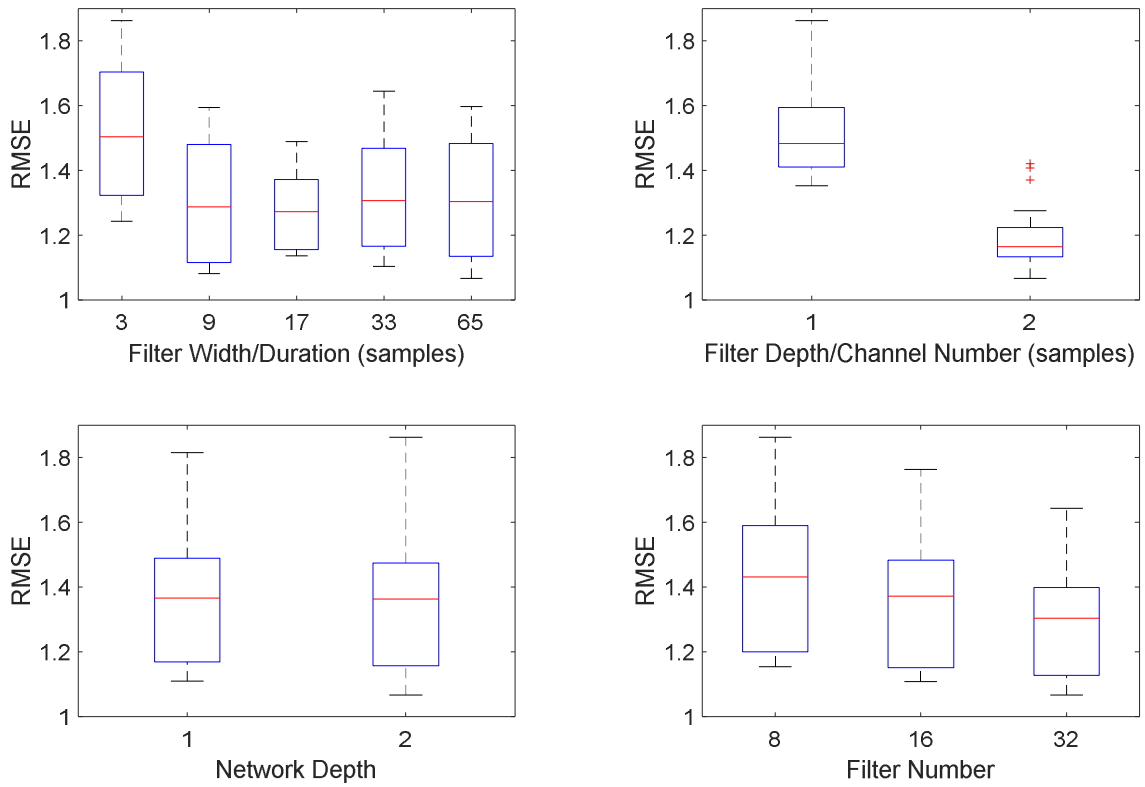
Supplementary Note 2 – Testing various of network parameters (18 channel FBA for D1)

We also tested parameters of filter width, filter depth, network depth and filter number within the residual network. The optimal combination was selected using the internal validation data, averaged across all folds. The latter was performed to ensure only 1 network architecture was used across fold. A fixed grid optimization search across was performed across an array of network parameters. For the 18-channel FBA based on dataset D1, this resulted in 180 different network architectures (***Supplementary Figure 8***). The general trends are shown in Figure 4 with an optimal combination of network parameters selected: a filter width of 9 samples (n = 45 per group), filter depth of 6 channels (n = 36 per group), network depth of 3 (n = 60 per group), and a filter number of 16 (n = 60 per group): a total of 867,601 learnable parameters.

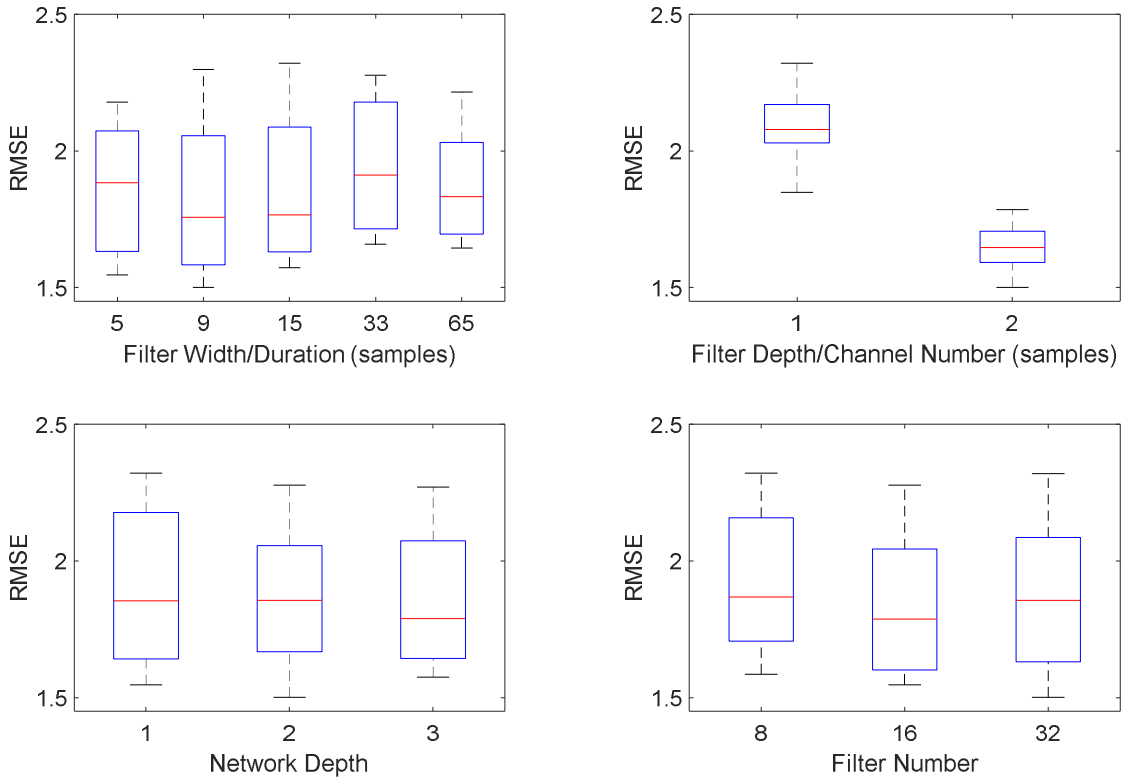


Supplementary Figure 11 **-** The general effects of residual network parameters on cross-validated RMSE for 18-channel data (D1 dataset). Internal single fold cross-validation averaged across all 10 training folds selected. Note

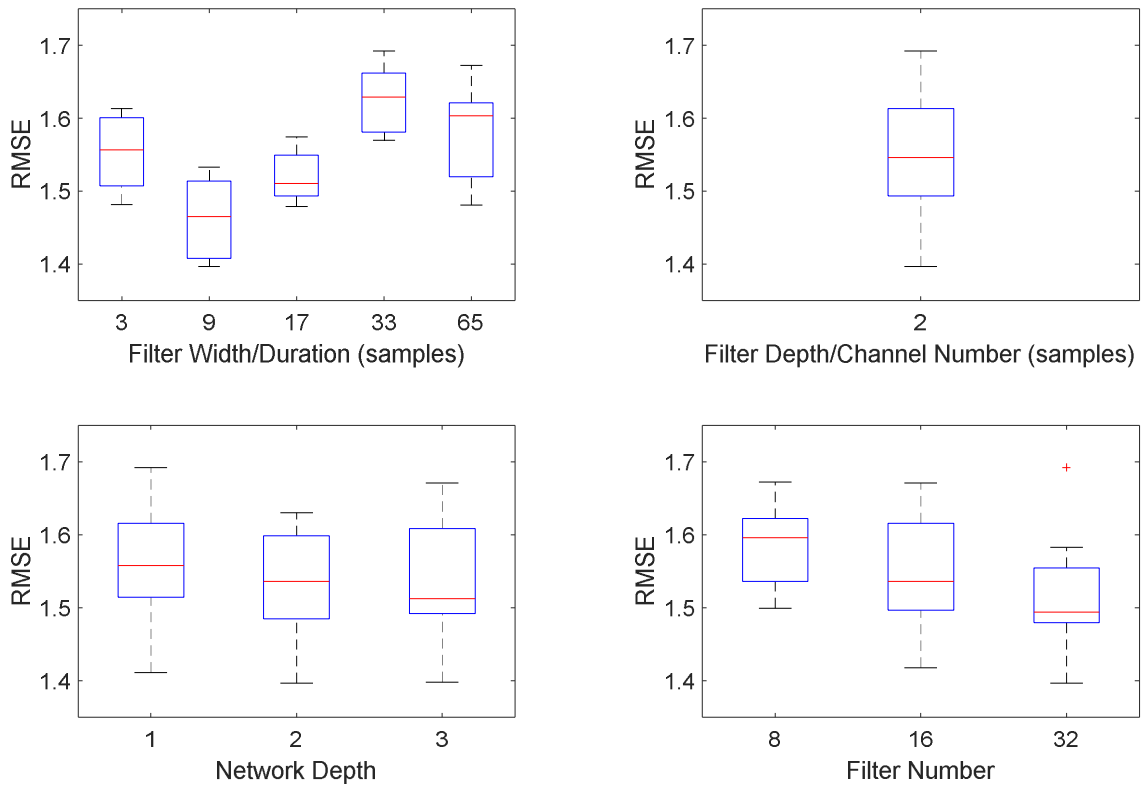
that, as filter depth of 1 channel had a considerably higher RMSE than other filter depths, RMSE points associated with a 1 channel network were removed before generating the remaining boxplots for visualization purposes.



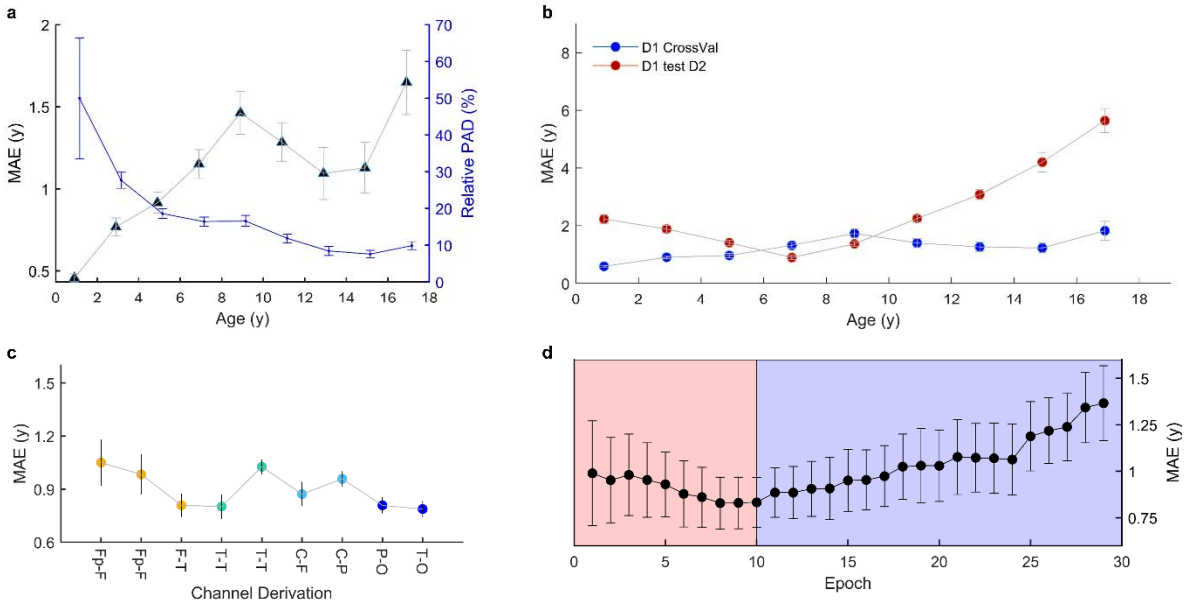
Supplementary Figure 12 *- The general effects of residual network parameters on cross-validated RMSE for a 2-channel FBA based on dataset D1.* This resulted in 60 different network architectures. The optimal combination of network parameters selected were a filter width of 9 samples (n = 18 per group), filter depth of 2 channels (n = 30 per group), network depth of 2 (n = 30 per group), and a filter number of 2 (n = 20 per group): a total of 906,657 learnable parameters.



Supplementary Figure 13 *- The general effects of residual network parameters on cross-validated RMSE for 2-channel data (D2 dataset).* This resulted in 90 different network architectures. The optimal parameters were a filter width of 9 samples (n = 18 per group), filter depth of 2 samples (n = 45), network depth of 2 (n = 30 per group), and a filter number of 32 (n = 30 per group): a total of 906,657 learnable parameters.



Supplementary Figure 14 *- The general effects of network parameters on cross-validated RMSE for 2-channel data (D1 and D2 combined).* This resulted in 45 different network architectures. The RMSE of an internal single fold cross-validation, average across all 10 training folds selected a filter width of 9 samples (*n* = 9 per group), a network depth of 2 (*n* = 45 per group), network depth of 2 (*n* = 15 per group), and a filter number of 32 (*n* = 15 per group): a total of 906,657 learnable parameters. Note, we did not vary the filter depth/channel number as previous experiments on D1 and D2 alone showed that 2 channels was optimal.



Supplementary Figure 15 *- Summary plot of GPR model performance*. **a.** The mean absolute error (MAE) of the GPR (dark blue triangles, left axis) across 2 yearly age bins and the relative accuracy of PAD (in blue, right axis) across 2 yearly age bins. For both plots, the mean and standard error of the mean (SEM) are plotted to reflect the sample distribution within age bins. **b.** Differences in MAE of D1 following cross-validation (D1 CrossVal) versus external validation on D2 (D1 test on D2) for GPR. Here differences in sites are more noticeable for younger and older age groups. **c.** Change in MAE across EEG channel locations, with average MAE shown and error bars indicated by standard deviation. **d.** Temporal transitions in the EEG across N1 (light pink) and N2 (light purple). Similar to the Res-NN, the lowest MAE is observed during a transition between sleep N1 and N2 stages. Error bars denote standard deviation for each epoch.

Supplementary Table 3 *– Performance of FBA across sleep stages in the validation dataset (D2)*. The performances of the FBA following 10-fold cross-validation, using either a Res-NN or GPR model, in the D2 dataset. MAE, wMAE and RMSE are shown. N.B. N1 was excluded due to inconsistent availability across ages in D2.

		Res-NN		**GPR**	
Sleep stage	**# channels, # recordings, # epochs**	**MAE (in years)**	**wMAE (in years)**	**MAE**	**wMAE**
N3	2, 723, 13737	1.46	1.76	1.75	2.00
REM	2, 714, 13339	1.55	1.98	1.68	1.89
WAKE	2, 564, 8810	1.89	2.53	2.06	2.61
N2	2, 723, 13737	1.46	1.73	1.25	1.34

Supplementary Table 4 *– Performance summary across test folds for Res-NN and GPR models*. Within each of the 10 test folds, the mean absolute error (MAE) is summarized for both model approaches.

	Res-NN	**GPR**
Folds	**MAE**	**MAE**
	Test D1	*Test D1*
1	0.63	0.72
2	0.52	0.73
3	0.61	0.73
4	0.52	0.78
5	0.61	0.75
6	0.55	0.76
7	0.50	0.98
8	0.58	0.80
9	0.62	0.86
10	0.53	0.80
Fold-wise Average	0.57	0.79
	Test D2	*Test D2*
1	1.69	1.20
2	1.43	1.25
3	1.63	1.25
4	1.40	1.15
5	1.43	1.34
6	1.63	1.11
7	1.39	1.21
8	1.35	1.34
9	1.32	1.20
10	1.51	1.33
Fold-wise Average	1.46	1.24
	Test D1 + D2	*Test D1 + D2*
1	1.15	1.07
2	0.97	1.05
3	1.03	0.91
4	0.95	1.14
5	0.96	1.21
6	1.06	1.16
7	0.89	1.11
8	1.36	1.04
9	1.00	1.13
10	1.11	1.03
Fold-wise Average	1.05	1.09