Deep Learning Identifies Morphological Determinants of Sex Differences in the Pre-Adolescent Brain

Ehsan Adeli\textsuperscript{a,1}, Qingyu Zhao\textsuperscript{a,1}, Natalie M. Zahr\textsuperscript{a,b}, Aimee Goldstone\textsuperscript{b}, Adolf Pfefferbaum\textsuperscript{a,b}, Edith V. Sullivan\textsuperscript{a}, Kilian M. Pohl\textsuperscript{a,b,*}

\textsuperscript{a}Department of Psychiatry & Behavioral Sciences, Stanford University, Stanford, CA 94305
\textsuperscript{b}Center for Biomedical Sciences, SRI International, Menlo Park, CA 94025

Abstract

The application of data-driven deep learning to identify sex differences in developing brain structures of pre-adolescents has heretofore not been accomplished. Here, the approach identifies sex differences by analyzing the minimally processed MRIs of the first 8,144 participants (age 9 and 10 years) recruited by the Adolescent Brain Cognitive Development (ABCD) study. The identified pattern accounted for confounding factors (i.e., head size, age, puberty development, socioeconomic status) and comprised cerebellar (corpus medullare, lobules III, IV/V, and VI) and subcortical (pallidum, amygdala, hippocampus, parahippocampus, insula, putamen) structures. While these have been individually linked to expressing sex differences, a novel discovery was that their grouping accurately predicted the sex in individual pre-adolescents. Another novelty was relating differences specific to the cerebellum to pubertal development. Finally, we found that reducing the pattern to a single score not only accurately predicted sex but also correlated with cognitive behavior linked to working memory. The predictive power of this score and the constellation of identified brain structures provide evidence for sex differences in pre-adolescent neurodevelopment and may augment understanding of sex-specific vulnerability or resilience to psychiatric disorders and presage sex-linked learning disabilities.

\textsuperscript{*}Corresponding author

Email address: kilian.pohl@stanford.edu (Kilian M. Pohl)

\textsuperscript{1}These authors contributed equally to this work.

Preprint submitted to NeuroImage August 20, 2020
1. Introduction

The concept of sex differences is based on biology and genetics. Since the 1930s (e.g., [1]), identifying sex differences in the Central Nervous System (CNS) has been explored in animal models [2, 3, 4, 5, 6] and histology of postmortem human brain samples [7, 8, 9]. More recently, in vivo neuroimaging [10, 11, 12, 3, 13, 14, 15] and computer-based learning tools [16, 17, 18, 19, 20] have been implemented in the search for a CNS basis of sexual differentiation. Beyond sex-linked risks for disease [21, 22, 23, 24, 25, 7], this search is motivated by adolescence being a period of particular vulnerability to the emergence of sex-linked neuropsychiatric disorders such as schizophrenia [7, 26] and autism [27, 28, 29, 25, 30], which have a higher prevalence in boys than girls, and depression, which girls by age 15 develop twice as likely as boys [18, 31].

In vivo structural magnetic resonance imaging (MRI) studies characterize brain development as following heterogeneous growth trajectories [32, 33] during which sex-specific behaviors emerge [34]. While physical signs of sex differences are present at birth [35], brain structural and functional differences between the sexes continue to develop over childhood through late adolescence [36, 37, 38, 39, 40]. For example, both cortical and subcortical gray matter volumes exhibit inverted U-shaped trajectories reflecting growth followed by synaptic pruning, with boys showing a slightly larger rate of change throughout childhood and adolescence than girls [41]. With respect to white matter, the volume increases with age in both sexes, but boys generally show a more rapid increase during adolescence [41]. These sex specific changes in brain structure during adolescence [42] are accompanied with asexual developments, such as structural volume [43, 44, 45, 46, 47, 48], cortical thickness [48], cortical surface area [44, 48], individual’s behavior [49], and testosterone effects [42].

Many of the differences in brain development between the sexes are actually
linked to head size [50, 51]. As boys on average have larger brains than girls, identifying sex differences in the brain beyond head size is challenging and might explain the inconsistent findings in the literature. For example, whether sex differences are present within the corpus callosum has been a matter of debate [52, 23, 53, 54, 55]. Beyond properly accounting for head size [53, 56, 39, 51], discrepancies in findings may be due to small sample sizes [57, 11, 58], wide age distributions (sometimes across several decades so age-specific sex differences are obscured) [52, 59], or a priori assumptions that reduce the rich information encoded in MRIs to a few brain measurements (e.g., volumes of a limited number of brain regions of interest (ROIs)) [52, 20]. The study presented herein accounts for these issues by building on recent advancements in the field of deep learning [60, 61, 19] to identify patterns not driven by study confounders, which are extraneous variables (such as age) that may induce undesired class differences if not properly controlled.

Specifically, we present a deep learning framework (see Figure 1) predicting sex from the minimally processed T1-weighted (T1w) MRIs [62] of 8,144 pre-adolescents (ages 9 and 10 years) of the ABCD study (http://abcdstudy.org). The variance in the prediction scores is related to the cognition test scores of the National Institutes of Health (NIH) Toolbox® [63]. Finally, we qualitatively assess the average saliency map [64] across all MRIs, which encodes the contribution of each voxel of the MRI in predicting sex while removing the effects driven by the confounders, i.e., age and pubertal and socioeconomic status.

2. Materials and Methods

2.1. ABCD Participants and Study Design

The model was evaluated on data collected by the ABCD study (http://abcdstudy.org). Demographic information (Table 1), cognitive test scores from the NIH toolbox (Table 2), and T1-weighted (T1w) MR images [62] from 8,670 participants were distributed by the ABCD-Neurocognitive Prediction Chal-
The ABCD-NP-Challenge 2019 [65] via the National Database for Autism Research (NDAR) portal (Release 2.0), of which 8,144 subjects contained the data needed for this analysis. Socioeconomic status (SES) was estimated by identifying the maximum level of education across parents/guardians as done elsewhere [66]. Pubertal status was determined by self-assessment with the Pubertal Development Scale (PDS) [67, 68], a validated measure of pubertal stage that shows modest concordance with a physical exam and that correlates with basal gonadal hormone levels. An average PDS was calculated for each participant by adding up scales on five self-reports obtained from parents’ responses to a questionnaire, where each scale ranged from 1 to 4. Based on this computation, PDS categorized ABCD youth as either 1) pre-pubertal, 2) early-pubertal, 3) mid-pubertal, 4) late-pubertal 5) post-pubertal. Participants of multiple ethnicities were categorized according to their minority ethnicity (e.g., a report of Asian and Caucasian was classified as Asian) [39]. Body Mass Index (BMI) was calculated based on published methods [69]. Observed Sex for all the participants was defined as the sex at birth.

Recruitment for the ABCD study closely represented the general U.S. population of 9 and 10 year-old children with respect to key demographic variables including sex, ethnicity, household income, parental education, and parental marital status [70]. Parents provided informed consent and were fluent in either English or Spanish; children had to be fluent in English and provide assent for participation. Exclusionary criteria included poor English-language proficiency; the presence of severe sensory, intellectual, medical, or neurological issues that would affect the validity of data or ability to comply with the protocol; and contraindications to MRI (see [71] for complete description of details regarding recruitment and inclusion/exclusion criteria).

2.2. MRI Data Acquisition and Processing

Details on T1w-MRI acquisition are provided by https://abcdstudy.org/images/Protocol_Imaging_Sequences.pdf. Processing of T1w-MRI were subjected to the ABCD minimal-processing pipeline [62] followed by noise removal.
and field-inhomogeneity correction via N4ITK (Version 2.1.0) [73]. Brain masks were determined via majority voting [74] over the segmentations generated by applying the following tools to both bias and non-bias corrected T1w-images: FSL BET (Version 5.0.6) [75], AFNI 3dSkullStrip (Version AFNI_2011_12_21_1014) [76], FreeSurfer mri-gcut (Version 5.3.0) [77], and Robust Brain Extraction (ROBEX) (Version 1.2) [78]. The resulting brain mask was used to refine correction for image-inhomogeneity and skull stripping. MRIs were then affinely registered to the SRI24 template [79], down-sampled to 2mm isotropic voxel size, and re-scaled to 64x64x64 volumes. The affine registrations ensured that all MRIs of the ABCD study had similar head size as measured by supratentorium volume (svol) (see also Table 1 for the resulting insignificant difference in head size between boys and girls).

Figure 1 outlines the deep learning framework used to predict sex from minimally processed MRI data. The framework was composed of a Predictor/Extractor and a Classifier [60, 80]. The Predictor/Extractor identified a set of Predictor variables \( P = \{ P^1, P^2, ..., P^M \} \) from MR images based on a deep convolutional network [61]. The Classifier was a set of fully connected layers reducing \( P \) into a continuous Prediction Score \( S \), which was the probability \( \pi \) computed by the classifier of an MRI being associated with a girl (i.e., \( \pi(\text{girl}) = S \)) or a boy (i.e., \( \pi(\text{boy}) = 1 - S \)). Appendix B provides a more in-depth description of the deep learning architecture.

The prediction accuracy of the model was determined in two steps. Assuming that sex affects the brain bilateral [81, 82, 83, 84, 85] and to simplify the interpretation of the findings, the left hemisphere was first flipped to create a 2nd “right” hemisphere. Then, 5-fold cross-validation [86] was performed by splitting the data based on subjects. At each iteration of the cross-validation, the four folds of the data used for training were first augmented to ensure that the learning was based on a balanced and sufficient number of boys and girls [87], i.e., 5000 for each group. Data augmentation consisted of applying random rigid transformations (within one voxel shifting and 1° rotation along the three axes) to the minimally processed (and flipped) MRIs. On this augmented data set,
the entire deep model, which included the predictor extractor and the classifier, was trained from scratch in an end-to-end manner [60]. Next, the prediction of the individual’s sex was recorded on the fifth fold (which was not augmented) by computing the average prediction score \( (S) \) across both hemispheres. The training and testing processes were repeated until the prediction score was reported for each subject. The average accuracy of the method on all folds was then computed by first binarizing \( S \) of each participant to 1 (girl) or 0 (boy) and then comparing the predictions to their observed sex via commonly used metrics: balanced classification accuracy [88] (a.k.a. accuracy), true positive rate, false positive rate, and the area under the receiving operating characteristic curve.

To put the prediction accuracy in perspective and compare with widely used machine learning methods, the cross-validation was repeated with respect to logistic regression [89, 90], support vector machines [91] and random forest [92] applied to the volumes of 116 brain ROIs defined according to the SRI24 atlas [79]. Measuring the volumes of ROIs consisted of non-rigidly registering the SRI24 atlas to each brain-size corrected MRI via ANTS (Version: 2.1.0) [93] and overlaying parcellations with the tissue segmentations from Atropos [94]. The experiment was repeated using the 906 regional scores generated by Freesurfer based on the Destrieux atlas [95], which were provided by the ABCD Study Release 2.0 (http://abcdstudy.org). These regional scores consisted of cortical thickness, sulcal depth, surface area, and volume of cortical ROIs and the average T1 intensities within the white and gray matter.

In addition to the comparison to other methods, a sex-agnostic test correlated the prediction score \( S \) of the individuals with the test scores of the age-corrected NIH toolbox (see Table 4 left; significant \( p \)-value < 0.05 according to Pearson’s \( R \)). Identifying variance in the prediction score partially induced by an NIH toolbox score (Figure 2, Table 4 right) was done via the partial mediation model [96]. Partial mediation required that 1) observed sex significantly correlated with the NIH toolbox score; 2) the NIH toolbox score significantly correlated with the prediction score when accounting for observed sex as an
additional covariate; and 3) the correlation between observed sex and the pre-
diction score was significantly reduced (p-value inferred from a permutation
test of 10000 permutations) when accounting for the NIH toolbox score as an
additional covariate.

Finally, we performed bootstrapping (5 runs) to determine the effects of
PDS (the most significant confounder of this study according to Table 1) on
the sex predictions of our approach. Each of the 5 runs was defined by 5-fold
cross-validation consisting of a unique random split of the data into 5-folds. The
correctly classified subjects in all 5 runs were assigned to one group, and the ones
that were incorrectly classified in all 5 runs were assigned to a second group.
For each sex separately, differences in the PDS between the two groups were
defined by the p-value of the χ² test [97]. Across the two groups, the prediction
accuracy (for both boys and girls) was determined for cohorts confined to the
same PDS. We then reported on the cohorts with a sufficient number of samples
for each sex, which were the cohorts for PDS 1, 2, and 3.

All methods were implemented using Python 3.7.0 and its libraries including
SciPy 1.1.0, NumPy 1.15.1, Scikit-Learn 0.19.2, pygrowup 0.8.2 toolbox [98],
Tensorflow 1.7.0 [99], and Keras 2.2.2 [100]. The codes of our deep learning
implementation are publicly available at https://github.com/QingyuZhao/
Confounder-Aware-CNN-Visualization and the tests at https://github.com/
eadeli/ABCD_SexDiff.

2.3. Identifying Confounder-Free Patterns and ROIs Relevant to Sex

To derive a single pattern informative for identifying sex differences, we re-
trained our proposed approach on the entire dataset. For each participant, the
discriminative power of each voxel to predict sex was recorded using a saliency
map [64]. The initial salience map was computed by applying the minimally
processed and flipped MRI to the trained prediction model and then perform-
ing back-propagation [101]. Note, saliency computation did not require data
augmentation nor estimating prediction accuracy.

Next, the map was further corrected for the effects of potential confounders
on the decision process of the model. Confounders were demographic factors significantly different between sexes according to Table 1, i.e., age ($z^{age}$), PDS ($z^{pds}$), and SES ($z^{ses}$). To determine if a confounder significantly influenced the decision process of $S$, a general linear model (GLM) [102] was fit across all samples with respect to each predictor variable $P_j$ of $P$:

$$P_j = \beta_0 + \beta_1 S + \beta_2 z^{pds} + \beta_3 z^{age} + \beta_4 z^{ses}.$$ (1)

If the predictor variable $P_j$ of the GLM significantly correlated ($p \leq 0.05$) with one of the demographic variables, the predictor was considered confounded and omitted from computing the saliency maps. The lenient $p$-value threshold of 0.05 was not corrected for multiple comparison as we wanted our analysis to be sensitive towards identifying confounded predictors so that the resulting pattern accurately represented sex differences. The pattern encoding the relevance of each voxel in predicting sex was defined by the average across the confounder-free saliency maps of all participants. Conversely, a pattern encoding the effect of a specific confounder was created by computing the saliency maps based on confounded predictors.

To relate the identified voxels to previously defined brain ROIs (using SRI24 atlas [79]), we computed the average saliency value of each ROI from the confounder-free saliency map of each participant. For each ROI, follow-up t-tests evaluated whether the average saliency value within that region was significantly different between groups ($p$-value $< 0.05$ with Bonferroni multiple comparison correction [103]).

3. Results

The accuracy of the prediction score in correctly assigning MRIs to either sex was 89.6%, which was significantly better than chance ($p < 0.001$ according to a Fisher exact test [104]). The prediction accuracy was stable across 5 runs of 5-fold cross-validation based on random splitting of folds (89.6% $\pm$ 0.13%) but was slightly lower (87.3%) on a subset of 2464 boys and 2464 girls matched
on head size (matched according to [105]). Furthermore, the True Positive Rate (TPR) of the deep learning model was 87.4% and True Negative Rate (TNR) was 91.5% (girls=1, boys=0). Compared with the correctly classified pre-adolescents, misclassified boys had significantly higher PDS while misclassified girls had significantly lower PDS ($p$-value $< 10^{-6}$ according to $\chi^2$ test). The prediction confined to individuals with the same PDS was 88.9% for participants with PDS = 1, 89.5% for PDS = 2, and 90.1% for PDS = 3.

The prediction of our approach was significantly more accurate (Delong test [106], $p$-value<0.001) than the results reported by Logistic Regression, Support Vector Machine, and Random Forest applied to the 116 ROI volume measures or the 906 Destrieux parcellation measures (see Table 3). To gain a better understanding of this improvement, we recomputed the accuracy of our model across 5 runs of 5-fold cross-validation with respect to the number of predictors. The average accuracy remained relatively high (86.5%) even when extracting only 128 predictors from each MRI (see Fig. 3(a)). Furthermore, similarly high accuracy was achieved by the other approaches when trained on the predictors extracted by our deep model (Fig. 3(b)).

A visual confirmation of the significant prediction accuracy of our model were the two distinct distributions shown in Figure 4(a), which plotted the Prediction Score ($S$) of each participant as a function of their observed sex. Furthermore, projecting the high dimensional Predictors ($P$) learned from one training run into 2D via the t-distributed Stochastic Neighbor Embedding (tSNE) [107] also resulted in a cluster for boys and a separate one for girls (Figure 4(b)).

Figure 5(a) visualizes the initial saliency map with voxel values above 0.1 before correcting for confounders. The highlighted area significantly contributed to predicting sex, which partly consisted of the temporal lobes, subcortical regions, cerebellum, and corresponding white matter. Figure 5(b) shows the area of sexual differentiation according to the confounder-free saliency map (i.e., with age, PDS, and SES removed), which is more spatially concentrated than the initial saliency map (Figure 5(a)). According to the confounder-free saliency values, the 10 ROIs most relevant for predicting sex were insula, pallidum, para
hippocampus, and putamen (larger in boys than girls); hippocampus, corpus medullare, and cerebellum VI (larger in girls than boys) (Figure 6). Although deep learning identified insula, amygdala, and cerebellar lobules III and IV/V as significant predictors of sex, their volume differences by sex were not forthcoming. The cerebellum was also the region mostly confounded by PDS (Figure 5(c)), the most significant confounder in the model.

Table 4 lists the correlation and mediation effect of NIH toolbox scores with respect to the prediction score $S$. Significant correlations ($p$-value < 0.05) between $S$ and NIH toolbox scores were confined to the List Sorting Working Memory Test, Pattern Comparison Processing Speed, Picture Sequence Memory Test, and Picture Vocabulary Test. Further, a partial mediation model examined whether the NIH toolbox scores could partially explain the variance in $S$ in addition to the observed sex (Fig. 2). Only the List Sorting Working Memory Test score met the 3 significance conditions of the mediation model ($p$-value < 0.05): 1) observed sex significantly correlated with the NIH toolbox score; 2) the NIH toolbox score significantly correlated with $S$ when accounting for observed sex as an additional covariate; and 3) the correlation between observed sex and $S$ was significantly reduced when accounting for the NIH toolbox score as an additional covariate.

4. Discussion

The deep learning model presented herein not only successfully predicted the sex of 8144 pre-adolescents from (head-size normalized) T1w MRI but also was more accurate than several other commonly used machine learning approaches, e.g., logistic regression, support vector machine, and random forest. While these machine learning approaches relied on a priori defined regional measurements (as is commonly used for neuroscience studies [43, 108, 105, 109, 59]), the improved accuracy of the deep learning model was mostly due to its ability to simultaneously extract predictors directly from the MRIs and perform classification (see Figure 3). A novel discovery of that search for discriminative infor-
Information was that sex could be accurately predicted in individual pre-adolescents through a pattern composed of subcortical and cerebellar regions. Also unknown for pre-adolescence was that the cerebellum was most strongly affected by PDS, the most significant confounder of the study. Finally, reducing the pattern to a single score revealed that its variance was not only explained by sex but also by cognitive behavior linked to working memory.

Critical for interpreting the pattern was the notion that sex differences on brain structure are bilateral [82, 83, 84, 85]. We modeled that by ‘flipping’ the left hemisphere and then training the algorithm on two ‘right’ hemispheres for each subject. When omitting flipping, the prediction accuracy was 89.1% when just trained on the left hemisphere, 88.5% when only trained on the right hemisphere, and 90.1% when trained on both hemispheres (omitting flipping). These accuracy scores were insignificantly different ($p > 0.1$; DeLong’s test) from those of the ‘flipped’ approach confirming the bilateral nature of sex differences.

Another critical aspect in analyzing the pattern was computing a saliency map that displayed brain areas exhibiting sex differences while accounting for confounders; something that had not been attempted by prior data-driven analyses [16, 17, 50, 19, 20]. Removing confounding effects after training a machine learning model is potentially a more conservative approach compared with removing effects through preprocessing (e.g., matching), i.e., before the training. Unlike removing confounding effects after training, preprocessing generally cannot completely remove those effects so that learning approaches can still leverage the remaining confounding effects to ‘improve’ predictions [88]. Of the three confounders considered, PDS was the most significant one, which was generally larger in girls than in boys within the pre-adolescent age range (Table 1). While misclassified boys had significantly higher PDS and misclassified girls had significantly lower PDS than correctly classified individuals of the same sex, the prediction accuracy of our deep learning model was not affected by PDS as the overall accuracy of 89.6% remained stable when confining the evaluation to individuals with the same PDS. The region most confounded by PDS was the cerebellum (Figure 5(c)) suggesting that pubertal status may be specifically
associated with cerebellum development at this young age. This hypothesis is difficult to test on the baseline data of ABCD as the majority (∼73%) of individuals were categorized as pre- or early pubescent. However, as the ABCD cohort ages, the variability in PDS will be considerably greater, and as such, will allow us to explore in more detail the potential interaction between sex and puberty in terms of cerebellar development.

In addition to the relationship to PDS, structures of the cerebellum were critical to predicting sex in individual, which is inline with a number of adult studies [108, 10, 110, 111]. However, sex differences in cerebellar volume became generally negligible once studies corrected for intracranial volume (e.g., [112, 113, 114]). More specifically, the corpus medullare of the cerebellum in this study was significantly larger in girls than boys. By contrast, the longitudinal study by [111] did not detect significant sex differences in the corpus medullare but reported that total cerebellar volume was larger in boys than girls, and that this total volume peaked at age 15.6 years in boys and at age 11.8 years in girls. The discrepancy in age range of the participants between that study (spanning pre-adolescents to young adults) and our analysis (ages 9 and 10 years) might reflect variance in cerebellar developmental trajectories during critical developmental years. Indeed, a recent review of the literature on language and brain development concluded that sex differences were most often found in studies limited to tight age ranges [52]. Sex differences in regional brain volumes may be apparent in some but negligible in other developmental stages, likely due to different rates of brain maturation between girls and boys [115].

Of the predictive regions within the subcortex, the hippocampus was larger in girls than boys after correcting for head size (see Fig. 6). The hippocampus has often been associated with sex-specific differences in memory and learning in adolescence [116, 117]. This observation comports with the finding that girls participating in the ABCD study had significantly higher scores on the NIH Toolbox Picture Sequence Memory Test, which is a validated measure of episodic memory [118]. The finding that girls had relatively larger hippocampi
than boys is also supported by MRI studies of young adults [11, 119, 114] that linked sex differences in hippocampal volume to hormonal responsivity [120, 121] and memory performance [81, 122, 123]. Other studies noted relations between hippocampal volumes and clinical characteristics of psychiatric disorders [22, 119, 124], where sleep disturbances are more severe [? ], depressive episodes are more frequent and longer, and higher frequency of migraines occurs in depressed female compared to depressed male patients [125].

Other regions relevant for predicting sex included putamen, pallidum, and amygdala. These regions are frequently noted with reference to sex differences in brain maturation. An early imaging study of children aged 4-18 years suggested that while the caudate is relatively larger in girls, the pallidum is larger in boys [126]. A more recent study based on data from the Pediatric Imaging, Neurocognition, and Genetics (PING) study with 1,234 participants (ages 3 to 21 years) [127] showed that volumes of putamen and pallidum had greater variance in boys than girls: these differences may contribute to the variability in cognition and general intelligence in developing boys [128, 129]. Likewise, the amygdala has been linked to sex differences in animal and human studies across the lifespan [130, 109]. A surface-based modeling approach showed that men had a larger mean radius of amygdala subregions than women [59]. Further, sex differences in amygdala volume may contribute to the expression of selective psychotic disorders occurring more commonly in men than women [130] and depressive disorders, which are more common in women [18, 31].

Like the amygdala, the insula was important for predicting sex but its volume was insignificantly different between the two cohorts. Functional studies have frequently shown the significant role of these two regions in working memory performance [131]. Interestingly in our study, sex prediction by the deep learning model was mediated by the List Sorting Working Memory test score, which was higher for boys than girls (see Table 4). These results suggest that the deep learning approach of directly analyzing intensity values at a voxel level is potentially more powerful in extracting morphological characteristics linked to cognitive differences between the sexes than traditional approaches that focus
In addition to the mediation analysis, the predictive score was significantly correlated to most of the cognitive scores by the NIH Toolbox. These early and pervasive sex differences in neurocognitive measures echoed those identified on the 10,000 youth of the Philadelphia Neurodevelopmental Cohort (PNC) [132], in which girls performed better than boys on tasks assessing verbal memory and social cognition, whereas boys excelled on spatial processing and motor speed [132, 133]. Similar results were reported with the National Consortium on Alcohol and Neurodevelopment in Adolescence (NCANDA) data, whose cognitive test battery included those of the PNC study [66]. Further consistency in sex differences on performance is forthcoming between our results and those published by the PING study, which, like the ABCD study, used the NIH Toolbox Battery. The PING study found that girls performed better than boys on tests assessing cognitive flexibility, problem solving, and episodic memory, whereas boys performed better on a list sorting task, assessing working memory for sorting and sequencing information [134]. Taken together, the consistency of sex differences in the development of component processes of selective cognitive skills transcended cohort differences and specific testing materials, which provide evidence for generalization of these identified sex differences.

Limitation. Our analysis did not detect significant sex differences in the cortex possibly because the MRIs were affinely aligned to a template, thereby minimizing headsize differences. While a common practice in end-to-end training [135, 60], affine registration might poorly align the cortical gyri and sulci given their high inter-subject variability [136]. Non-rigid registration achieves better voxel-wise correspondence across MRIs enabling learning algorithms to focus on fine-grained regional cues [137, 138]. Now identifying cues differentiating between groups highly depends on the ‘stiffness’ of the deformation field [139, 140], which can substantially modify the shape and appearance of brain structures. As affine and non-rigid registration can both negatively affect analysis, their effect on our deep learning findings needs to be further investigated.
5. Conclusion

The voxel-level analysis on the large number (N=8,144) of pre-adolescents (age 9 and 10) confirmed and extended the common finding of smaller neuroimaging studies that cerebellum and subcortical structures (including hippocampus, amygdala, pallidum, and putamen) differed in size between boys and girls. Not known before, however, was that the constellation of those brain structures accurately predicted the sex of individual pre-adolescents. The predictive power of the pattern provides evidence for sex differences in pre-adolescent, pubertal development, which may show even greater differentiation as the cohort ages. Tracking these disparities is a normative process that could augment understanding of sex-specific vulnerability or resilience to psychiatric disorders and presage sex-linked learning disabilities.

6. Acknowledgments

Funding for this study was received from the U.S. National Institutes Health (NIH) grants AA026762, DA041123, AA021697, and AA010723.

None of the authors has conflicts of interest with the reported data or their interpretation.

References


[88] S. H. Park, Y. Zhang, D. Kwon, Q. Zhao, N. M. Zahr, A. Pfefferbaum, E. V. Sullivan, K. M. Pohl, Alcohol use effects on adolescent brain develop-


[96] R. M. Baron, D. A. Kenny, The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical con-

[97] K. Pearson, On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50 (302) (1900) 157–175.

URL https://pypi.org/project/pygrowup/


[141] R. C. Gershon, M. V. Wagster, H. C. Hendrie, N. A. Fox, K. F. Cook,


Appendix A. Descriptions of the NIH Toolbox® Cognitive Tests

The NIH Toolbox® cognition measures were developed as part of the NIH Blueprint for Neuroscience Research (http://www.nihtoolbox.org). The tests assess episodic memory, executive function, attention, working memory, processing speed, and language abilities, enabling generation of composite scores [141, 142]. Use of a common tool for cognitive assessment valid for ages spanning the ABCD cohort’s current and future range allows for longitudinal tracking of the developmental trajectories of this cohort in addition to harmonization and comparison of cognitive performance with numerous other studies. The tasks were selected based on a consensus building process and developed and validated using assessment methods that included item response theory (IRT) and computerized adaptive testing (CAT) where appropriate and feasible. Each
Toolbox® task produces a number of scores, some of which are adjusted for age, sex, and ethnicity. All tasks provide raw scores, uncorrected standard scores, and age-corrected standard scores based on a normative sample of 2917 children and adolescents [143]. This study used age-corrected measures to compare the two cohorts of boys and girls, as there was a significant difference between our two cohorts. These tests are comprehensively described elsewhere [63] and briefly below.

1. **Language/Vocabulary Comprehension:** The Toolbox Picture Vocabulary Task® (TPVT) is a variant of the Peabody Picture Vocabulary Test (PPTV) [144, 145, 146].

2. **Language/Reading Decoding:** The Toolbox Oral Reading Recognition Task® (TORRT) asks individuals to pronounce single letters or words presented in the middle of the iPad screen [144, 145] and measures exposure to language materials and cognitive skills involved in reading.

3. **Processing Speed:** The Toolbox Pattern Comparison Processing Speed Test® (TPCPST) [147, 148, 149] was modeled on the Pattern Comparison Task developed by Salthouse [150] and is a measure of rapid visual processing.

4. **Working Memory:** The Toolbox List Sorting Working Memory Test® (TLSWMT) is a variant of the letter-number sequencing test [151] that uses pictures rather than words or letters [152, 153].

5. **Episodic Memory:** The Toolbox Picture Sequence Memory Test® (TPSMT) was modeled after memory tests asking children to imitate a sequence of actions with props developed by Bauer and colleagues. [154, 155].

6. **Executive Function/Attention/Inhibition:** The Toolbox Flanker Task® (TFT), a variant of the Eriksen Flanker task [156], was adapted from the Attention Network Task [157, 158] and assessed response inhibition.

7. **Executive Function/Cognitive Flexibility:** The Toolbox Dimensional Change Card Sort Task® (TDCCS) was based on the work of Zelazo and colleagues [159] and measures problem solving and cognitive flexibility.
Appendix B. Deep Learning Model Architecture & Hyperparameters

Input to the deep learning model was the 3D MRI of one hemisphere of size 64 × 64 × 32. The predictor extraction network contained 4 stacks of 3 × 3 × 3 convolutional layers, ReLu activation, batch normalization, and 2 × 2 × 2 max-pooling layers. The size of feature channel for the 4 convolution layers was (16, 32, 64, 128). Then the resulting 4096 features were fed into a set of fully connected layers (Multi-Layer Perceptron) classifier composed of three Fully Connected (FC) layers of dimension (64, 32, 1). \texttt{tanh} activation was used for the first two FC layers, and \texttt{sigmoid} activation was used for the last FC layer resulting in the final prediction score $S \in [0, 1]$. An L2 regularization of weight 0.1 was applied to the FC layers (see Fig. 7).

Appendix C. Receiver Operating Characteristic Curve

As included in the main paper, our deep learning framework led to an accuracy of nearly 90% for predicting the sex of individuals based their structural MRI data. The receiver operating characteristic (ROC) curve of this classification model is depicted in Appendix Figure 8, which shows an area under the curve (AUC) of 0.96.
Figure 1: Overview of the proposed analysis. The convolutional neural network (CNN) automatically extracts predictors ($P$) from the minimally processed MRI. Based on $P$, the classifier computes a prediction score ($S$) that assigns the MRI to either sex. This deep learning analysis operates directly on voxel-level data omitting any hypothesis or assumption related to brain regions or tissue measurements (like regional volumes). Statistical analysis relates obtained results to NIH Toolbox cognitive test scores, creates confounder-free visualization of the patterns predicting sex (a.k.a. saliency map), and examines volume scores of those regions that contribute significantly to the prediction according to the saliency map.
Mediating (indirect) Effect

Sex \rightarrow \text{Direct Effect} \rightarrow \text{Prediction Score (S)}

\text{Cognitive Test Score} \leftarrow \text{Mediating (indirect) Effect}

Figure 2: Mediation analysis to observe how much of the variance in the prediction score was explained by the observed sex and how much was influenced by the NIH toolbox score.
(a) Accuracy of our deep learning classification as a function of the number of extracted predictors, $P$. 

(b) Accuracy of different models using 116 regional volumetric measures (left) and predictors extracted by the deep learning (middle & right).

Figure 3: Results of the deep learning model predicting sex with different numbers of predictors (a), and different classifiers (b).
Figure 4: Visualization of Predictors and the Prediction Score as determined by the deep learning model. (a) Prediction Score ($S$) of each participant as a function of their observed sex. These two figures show that our deep learning model can effectively reduce the MRIs to a vector of predictors ($P$) and then to a scalar value ($S$) that distinguishes girls from boys. (b) t-Distributed Stochastic Neighbor Embedding (tSNE) [107] projection of extracted Predictors ($P$) in 2D space. Each point indicates one adolescent; color represents sex. The axes show the relative location of each individual with respect to their neighbors in 2D. tSNE preserves the neighborhood of the high dimensional space.
Figure 5: Saliency maps defining predictive brain areas for distinguishing boys from girls in the ABCD study; (a) original and (b) corrected for confounding factors. In the developing brain of 9 and 10-year-olds, the factors distinguishing boys from girls mainly lie in the subcortical and cerebellar regions. (c) Regional brain pattern of sex differences confounded by PDS. Note, computing saliency maps requires scaling of the maps so that the resulting importance values are only meaningful within one saliency map but cannot be directly compared across maps.
<table>
<thead>
<tr>
<th>Regions of Interest (ROIs)</th>
<th>Mean Saliency</th>
<th>Female Volume Mean ± SD</th>
<th>Male Volume Mean ± SD</th>
<th>Cohen’s d</th>
<th>Group Differences</th>
<th>p-value</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pallidum</td>
<td>0.1605</td>
<td>80.00 ± 29.25</td>
<td>81.53 ± 29.12</td>
<td>0.053</td>
<td>0.02</td>
<td>F &lt; M</td>
<td></td>
</tr>
<tr>
<td>Cerebellum III</td>
<td>0.1500</td>
<td>803.53 ± 115.78</td>
<td>805.16 ± 120.03</td>
<td>0.014</td>
<td>NS</td>
<td>F = M</td>
<td></td>
</tr>
<tr>
<td>Amygdala</td>
<td>0.1460</td>
<td>1169.53 ± 94.08</td>
<td>1173.33 ± 94.87</td>
<td>0.040</td>
<td>NS</td>
<td>F = M</td>
<td></td>
</tr>
<tr>
<td>Hippocampus</td>
<td>0.1398</td>
<td>3408.35 ± 244.99</td>
<td>3322.98 ± 249.41</td>
<td>0.350</td>
<td>&lt; 10^{-6}</td>
<td>F &gt; M</td>
<td></td>
</tr>
<tr>
<td>Corpus Medullare</td>
<td>0.1387</td>
<td>7720.72 ± 660.95</td>
<td>7493.03 ± 663.39</td>
<td>0.343</td>
<td>&lt; 10^{-6}</td>
<td>F &gt; M</td>
<td></td>
</tr>
<tr>
<td>Para Hippocampal</td>
<td>0.1384</td>
<td>4700.50 ± 400.66</td>
<td>4726.10 ± 416.05</td>
<td>0.064</td>
<td>0.005</td>
<td>F &lt; M</td>
<td></td>
</tr>
<tr>
<td>Cerebellum VI</td>
<td>0.1384</td>
<td>9902.93 ± 1207.29</td>
<td>9478.25 ± 1151.66</td>
<td>0.360</td>
<td>&lt; 10^{-6}</td>
<td>F &gt; M</td>
<td></td>
</tr>
<tr>
<td>Corpus Medullare</td>
<td>0.1340</td>
<td>5440.99 ± 585.16</td>
<td>5441.95 ± 582.88</td>
<td>0.002</td>
<td>NS</td>
<td>F = M</td>
<td></td>
</tr>
<tr>
<td>Insula</td>
<td>0.1308</td>
<td>7322.80 ± 559.90</td>
<td>7305.74 ± 539.40</td>
<td>0.030</td>
<td>NS</td>
<td>F = M</td>
<td></td>
</tr>
<tr>
<td>Putamen</td>
<td>0.1298</td>
<td>2237.28 ± 502.88</td>
<td>2344.62 ± 502.94</td>
<td>0.213</td>
<td>&lt; 10^{-6}</td>
<td>F &lt; M</td>
<td></td>
</tr>
</tbody>
</table>

Figure 6: Top 10 regions relevant for distinguishing sex as determined by the deep learning framework. Some of these regions are smaller in girls (cerebellar lobules III and IV/V, amygdala; and insula, pallidum, para hippocampus, and putamen), while hippocampus, corpus medullare, and cerebellar lobule VI are smaller in boys. P-values of group differences of ROI volumes were calculated using two sample t-test. NS denotes not significant.
Figure 7: Architecture of our deep learning model.
Figure 8: Receiver Operating Characteristics (ROC) curve of the classifier differentiating boys and girls based on MR images. The blue curve shows the results of the model based on ABCD data.
Table 1: Demographic information (mean ± standard deviation).

<table>
<thead>
<tr>
<th>Measure</th>
<th>Female (F)</th>
<th>Male (M)</th>
<th>p-value&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Group Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Subjects</td>
<td>3,895</td>
<td>4,249</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Age (years)</td>
<td>9.92±0.62</td>
<td>9.95±0.62</td>
<td>0.04</td>
<td>F &lt; M</td>
</tr>
<tr>
<td>Head size&lt;sup&gt;b&lt;/sup&gt; (svol cm&lt;sup&gt;3&lt;/sup&gt;)</td>
<td>1341.5±15.3</td>
<td>1342.0±16.0</td>
<td>NS</td>
<td>F=M</td>
</tr>
<tr>
<td>Socioeconomic Status (SES)</td>
<td>18.0±66.8</td>
<td>18.5±67.8</td>
<td>NS</td>
<td>F=M</td>
</tr>
<tr>
<td>Pubertal Development Scale (PDS)</td>
<td>2.0±1.0</td>
<td>1.3±0.6</td>
<td>( \leq 10^{-6} )</td>
<td>F &gt; M</td>
</tr>
<tr>
<td>Ethnicity (%)&lt;sup&gt;c&lt;/sup&gt;:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian/African American/Caucasian/Other</td>
<td>269/631/2650/346</td>
<td>282/629/2975/363</td>
<td>NS</td>
<td>F = M</td>
</tr>
<tr>
<td>Body Mass Index (BMI) z-scores&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.23±5.6</td>
<td>0.15±14.1</td>
<td>NS</td>
<td>F = M</td>
</tr>
</tbody>
</table>

<sup>a</sup>measured by \( \chi^2 \)-test or t-test: NS ‘=’ not significantly different by \( p=0.05 \); ‘<’ or \( > \) significantly different at \( p \leq 0.05 \).

<sup>b</sup>head size was measured after being affinely registered to the SRI24 template.

<sup>c</sup>individuals who self-identified as Hispanic were included in the Caucasian group: 493 girls and 574 boys.

<sup>d</sup>z-scores of the BMI (instead of percentile) are calculated by the *pygrowup* toolbox [98] for each individual to enable group comparison using t-test.
Table 2: Scores of cognitive tests (mean ± standard deviation).

<table>
<thead>
<tr>
<th>Test</th>
<th>Cognitive Process</th>
<th>Female (F) N=3895</th>
<th>Male (M) N = 4249</th>
<th>Cohen's $d$-value $^*$</th>
<th>Group Difference $^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIH Toolbox: Flanker®</td>
<td>cognitive control; attention</td>
<td>96.29±13.37</td>
<td>97.09±14.39</td>
<td>0.058</td>
<td>NS F = M</td>
</tr>
<tr>
<td>NIH Toolbox: List Sorting Working Memory Test®</td>
<td>working memory; categorization; information processing</td>
<td>101.68±14.08</td>
<td>102.64±14.68</td>
<td>0.067</td>
<td>0.038 F &lt; M</td>
</tr>
<tr>
<td>NIH Toolbox: Dimensional Change Card Sort®</td>
<td>flexible thinking; concept formation; set shifting</td>
<td>98.89±15.07</td>
<td>97.44±15.54</td>
<td>0.095</td>
<td>0.3429e-2 F &gt; M</td>
</tr>
<tr>
<td>NIH Toolbox: Oral Reading Recognition Test®</td>
<td>reading ability; language; academic achievement</td>
<td>104.45±19.53</td>
<td>103.65±18.52</td>
<td>0.042</td>
<td>NS F = M</td>
</tr>
<tr>
<td>NIH Toolbox: Pattern Comparison Processing Speed®</td>
<td>processing speed; information processing</td>
<td>96.70±20.92</td>
<td>93.72±22.18</td>
<td>0.140</td>
<td>0.1875e-4 F &gt; M</td>
</tr>
<tr>
<td>NIH Toolbox: Picture Sequence Memory Test®</td>
<td>visuospatial sequencing &amp; memory</td>
<td>103.47±16.47</td>
<td>100.62±15.81</td>
<td>0.180</td>
<td>&lt; $10^{-6}$ F &gt; M</td>
</tr>
<tr>
<td>NIH Toolbox: Picture Vocabulary Test®</td>
<td>language; verbal intellect</td>
<td>108.53±16.98</td>
<td>109.35±17.05</td>
<td>0.056</td>
<td>NS F = M</td>
</tr>
</tbody>
</table>

$^*$measured by $\chi^2$-test or t-test: NS ‘=’ not significant; ‘<’ or ‘>’ significant at $p \leq 0.05$. 

50
Table 3: Accuracy (Acc), true positive rate (TPR), true negative rate (TNR), area under the ROC curve (AUC) of different methods for predicting sex from MRIs.

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc</th>
<th>TPR</th>
<th>TNR</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours (end-to-end deep learning)</td>
<td>89.6%</td>
<td>87.4%</td>
<td>91.5%</td>
<td>0.96</td>
</tr>
<tr>
<td>116 SRI24 volume Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>74.2%</td>
<td>74.3%</td>
<td>74.0%</td>
<td>0.80</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>74.2%</td>
<td>73.0%</td>
<td>75.5%</td>
<td>0.81</td>
</tr>
<tr>
<td>Random Forest</td>
<td>70.9%</td>
<td>66.7%</td>
<td>74.5%</td>
<td>0.75</td>
</tr>
<tr>
<td>906 Destrieux Parcellation Measures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>80.0%</td>
<td>80.8%</td>
<td>79.2%</td>
<td>0.88</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>79.1%</td>
<td>78.1%</td>
<td>79.9%</td>
<td>0.84</td>
</tr>
<tr>
<td>Random Forest</td>
<td>74.2%</td>
<td>72.2%</td>
<td>76.0%</td>
<td>0.79</td>
</tr>
</tbody>
</table>
Table 4: *p*-values of the correlation and mediation analysis with respect to the NIH Toolbox Scores. Correlation analysis was examined by Pearson’s *R*. Mediation analysis examined the indirect effect of NIH Toolbox scores on sex prediction; Significant mediation effect (*p* <0.05 for all 3 conditions of the partial mediation model) is marked by bold typeface. NS denotes not significant.

<table>
<thead>
<tr>
<th>Test</th>
<th>Correlation</th>
<th>Mediation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prediction</td>
<td>Observed Sex</td>
</tr>
<tr>
<td></td>
<td>Score (Condition 1)</td>
<td>Score (Condition 2)</td>
</tr>
<tr>
<td>NIH Toolbox Flanker®</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>NIH Toolbox List Sorting Working Memory Test®</td>
<td>0.001</td>
<td><strong>0.03817</strong></td>
</tr>
<tr>
<td>NIH Toolbox Dimensional Change Card Sort®</td>
<td>NS</td>
<td>0.00342</td>
</tr>
<tr>
<td>NIH Toolbox Oral Reading Recognition Test®</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>NIH Toolbox Pattern Comparison Processing Speed®</td>
<td>0.0025</td>
<td>0.00002</td>
</tr>
<tr>
<td>NIH Toolbox Picture Sequence Memory Test®</td>
<td>0.00001</td>
<td>&lt; 10^-6</td>
</tr>
<tr>
<td>NIH Toolbox Picture Vocabulary Test®</td>
<td>0.0309</td>
<td>NS</td>
</tr>
</tbody>
</table>