Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PrimateX: a unified multi-agentic framework for standardized MRI analysis across primate species

Zheng Wang

zheng.wang@pku.edu.cn

Peking University https://orcid.org/0000-0001-7138-8581

Xiaohui Yao
Harbin Engineering University

Yichun Huang
Peking University

Kailong Cui
Harbin Engineering University

Yunfeng Wang
Peking University

Junlong Huang
Hainan University

Hong Liang
Harbin Engineering University

Haoran Zhai
Harbin Engineering University

Chengtao Cai
Harbin Engineering University

Shiqi Yu
Peking University

Yiheng Hu
Peking University

Jiahao Huang
Zhejiang University of Technology

Hongyi Li
The People's Hospital of Liaoning Province

Ronghui Ju
The People's Hospital of Liaoning Province, Shenyang

Chenyuan Bian
The Affiliated Hospital of Qingdao University

Jing Liu

Peking University First Hospital
Yuanjing Feng
Zhejiang University of Technology
Sze Chai Kwok
Duke Kunshan University
William Hopkins
The University of Texas MD Anderson Cancer Center
Chet Sherwood
George Washington University
Junichi Hata
Tokyo Metropolitan
Daisuke Yoshimaru
RIKEN Center for Brain Science
Hideyuki Okano
Keio University Regenerative Medicine Research Center
Ken Nakae
Fukui University https://orcid.org/0000-0003-4183-1532
Stefan Everling
University of Western Ontario
David Leopold
National Institutes of Health
Ching-Yu Cheng
National University of Singapore https://orcid.org/0000-0003-0655-885X
Huazhu Fu
Technology and Research (A*STAR) https://orcid.org/0000-0002-9702-5524
Alejandro Frangi
University of Manchester https://orcid.org/0000-0002-2675-528X
Ravi Menon
Western University
Shan Cong
Harbin Engineering University

Article

Keywords: Primate neuroimaging, Multi-agent pipelines, Structural MRI standardization, Cross-species comparison

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11234334/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Yes there is potential Competing Interest. Z.W., S.C., X.H.Y., K.L.C., Y.C.H., and Y.F.W. are listed as co-inventors on filed patents covering multi-agent orchestrated standardized pipelines for cross-species primate brain MRI processing. All other authors declare no relevant financial conflicts of interest associated with this research.

PrimateX: a unified multi-agentic framework for standardized MRI analysis across primate species

Xiaohui Yao[1,2,3,4,#], Yichun Huang[5,#], Kailong Cui[1,2,3,#], Yunfeng Wang[5], Junlong Huang[6], Hong Liang[1], Haoran Zhai[1,2,3], Chengtao Cai[1], Shiqi Yu[5], Yiheng Hu[5], Jiahao Huang[7], Hongyi Li[8,9], Ronghui Ju[8,9], Chenyuan Bian[10], Jing Liu[11], Yuanjing Feng[7], Sze Chai Kwok[12,13], William D. Hopkins[14], Chet C. Sherwood[15], Junichi Hata[16,17,18], Daisuke Yoshimaru[19,20], Hideyuki Okano[18], Ken Nakae[21], Stefan Everling[22,23], David A. Leopold[24,25], Ching-Yu Cheng[4], Huazhu Fu[26], Alejandro F. Frangi[27,28,,29,30,31,32], Ravi S. Menon[33,34], Shan Cong[1,2,3,35*], Zheng Wang[5,6*]

[1]College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin, Heilongjiang, China

[2]Qingdao Innovation and Development Center, Harbin Engineering University, Qingdao, Shandong, China

[3]Sanya Nanhai Innovation and Development Base, Harbin Engineering University, Sanya, Hainan, China

[4]Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[5]School of Psychological and Cognitive Sciences; Key Laboratory of Machine Perception (Ministry of Education); Beijing Key Laboratory of Behavior and Mental Health; IDG/McGovern Institute for Brain Research; Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China

[6]School of Biomedical Engineering, Hainan University, Sanya, Hainan, China

[7]Institute of Advanced Technology, College of Information Engineering, Zhejiang University of Technology, Hangzhou, Zhejiang, China

[8]Department of Radiology, The People's Hospital of Liaoning Province, Shenyang, Liaoning, China

[9]Liaoning Provincial Key Laboratory of Neurointerventional Therapy and Biomaterials Research and Development, Shenyang, Liaoning, China

[10]Shandong Provincial Key Laboratory of Digital Medicine and Computer-Assisted Surgery, The Affiliated Hospital of Qingdao University, Qingdao, Shandong, China

[11]Department of Radiology, Peking University First Hospital, Beijing, China

[12]Division of Natural and Applied Sciences, Duke Kunshan University, Kunshan, Jiangsu, China

[13]Key Laboratory of Brain Functional Genomics (Ministry of Education), East China Normal University, Shanghai, China

[14]Michale E. Keeling Center for Comparative Medicine and Research, The University of Texas MD Anderson

Cancer Center, Bastrop, TX, USA
[15]Department of Anthropology and Center for the Advanced Study of Human Paleobiology, The George Washington University, Washington, DC, USA
[16]Laboratory for Marmoset Models of Brain Diseases, RIKEN Center for Brain Science, Wako, Saitama, Japan
[17]Graduate School of Human Health Sciences, Tokyo Metropolitan University, Tokyo, Japan
[18]Keio University Regenerative Medicine Research Center, Kanagawa, Japan
[19]Laboratory for Marmoset Neural Architecture, RIKEN Center for Brain Science, Wako, Saitama, Japan
[20]Division of Regenerative Medicine, The Jikei University School of Medicine, Tokyo, Japan
[21]Graduate School of Engineering, University of Fukui, Fukui, Japan
[22]Centre for Functional and Metabolic Mapping, Robarts Research Institute, University of Western Ontario, London, Ontario, Canada
[23]Department of Physiology and Pharmacology, University of Western Ontario, London, Ontario, Canada
[24]Section on Cognitive Neurophysiology and Imaging, Systems Neurodevelopment Laboratory, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA
[25]National Institute for Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA
[26]Institute of Advanced Intelligence and Computing (IAIC), Agency for Science, Technology and Research (A*STAR), Singapore
[27]Centre for Computational Imaging and Modelling in Medicine (CIMIM), The Christabel Pankhurst Institute, The University of Manchester, Manchester, UK
[28]Department of Computer Science, School of Engineering, University of Manchester, Manchester, UK
[29]Division of Informatics, Imaging and Data Sciences, University of Manchester, Manchester, UK
[30]NIHR Manchester Biomedical Research Centre, Manchester Academic Health Sciences Centre, University of Manchester, Manchester, UK
[31]Department of Cardiovascular Sciences, KU Leuven, Leuven, Belgium
[32]Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium
[33]Department of Medical Imaging, Schulich School of Medicine and Dentistry, Western University, London, ON, Canada
[34]Department of Medical Biophysics, Schulich School of Medicine and Dentistry, Western University, London, ON, Canada
[35]Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), Singapore

[#]These authors contributed equally. Correspondence: Shan Cong (shan.cong@hrbeu.edu.cn), Zheng Wang (zheng.wang@pku.edu.cn).

Abstract = 150 words

Main text = 3822 words

Number of Display Items = 7 Figures

Extended Data = 4 Figures + 3 Tables

Supplemental Material = 4 Method Notes + 8 Figures + 4 Tables

Running title: Multi-agentic MRI analysis across primates

Keywords: Primate neuroimaging, Multi-agent pipelines, Structural MRI standardization, Cross-species comparison

Abstract

Standardized preprocessing has transformed human neuroimaging into a reproducible, population-scale enterprise, yet equivalent computational frameworks for non-human primates (NHPs) remain absent, forcing continued reliance on fragmented, labor-intensive workflows. Here we introduce PrimateX, a multi-agentic computational framework in which LLM-guided orchestration dynamically routes structural MRI scans through species-tailored, topology-preserving workflows, extracting cortical surfaces directly from raw volumes with full provenance tracking across marmosets, macaques, chimpanzees, and humans. Validated on 1,221 NHP and 1,113 human scans, PrimateX outperformed existing NHP pipelines in segmentation and surface reconstruction while matching dedicated human tools such as FSL in their native domain. Through MRI-style data augmentation, the framework achieved robust zero-shot generalization to 8 unseen primate species across 31 cohorts from 19 global sites (n = 157). Critically, the routing architecture and closed-loop quality-control mechanisms, demonstrated through diverse downstream neuroimaging analyses and cross-species research, establish PrimateX as a widely-applicable computational paradigm. PrimateX is available at https://primatex.pku.edu.cn.

Introduction

The maturation of standardized computational pipelines - exemplified by FreeSurfer and the Human Connectome Project (HCP) workflows[1-7] - has transformed human neuroimaging into a reproducible, population-scale enterprise, enabling multi-site studies[8-13] and catalysing the emerging generation of neuroimaging foundation models[14, 15]. This computational consensus, however, has not extended beyond the human domain[16-18]. Non-human primate (NHP) neuroimaging, despite its privileged phylogenetic and neuroanatomic position for understanding human cognition and brain disorders[19-23], remains dependent on fragmented, protocol-specific workflows that lack the algorithmic generality needed to accommodate orders-of-magnitude variation in brain volume (marmoset to human), tissue contrast, cortical folding, and white-matter architecture across species[13, 24, 25].

The challenge is fundamentally computational. While initiatives like PRIME-DE[26, 27] and the NHP Consortium for Brain Research[28, 29] have aggregated diverse species datasets to propel deep learning solutions[30, 31], structural MRI preprocessing entails a chain of interdependent operations in which early-stage errors - in bias correction, skull stripping, or tissue segmentation - propagate relentlessly into downstream registration and surface reconstruction, systematically biasing statistical outcomes. Tools trained on human data routinely fail on NHP images: skull stripping misclassifies periorbital tissues or excludes ventral cortex[30-32]; white-matter segmentation errors create topological discontinuities that cascade into surface reconstruction failures[33],[34]. Prevailing segmentation models exhibit negligible cross-species generalizability[27, 31], and while biomedical foundation models (including MedSAM/2,

BiomedParse, and MedSegX) have advanced human-domain performance[35-38], their human-dominated training corpora and voxel-centric loss functions neglect the topological constraints essential for cross-species robustness. Recent self-supervised pretraining strategies achieve site-invariance for human brain MRI[15, 39], but have not been extended across species boundaries. Consequently, the field relies on labor-intensive manual review and ad-hoc reprocessing[30, 31, 40, 41] - a mode of operation that is computationally unsustainable and fundamentally incompatible with scalable cross-species harmonization.

To resolve these limitations, we developed PrimateX, an agent-mediated computational framework that reformulates expert-guided structural MRI analysis as an adaptive, self-correcting workflow. Three methodological advances distinguish the system. First, a state-machine-based multi-agent architecture couples LLM-guided workflow orchestration with a natural-language interface, enabling transparent pipeline execution, real-time quality inspection, and closed-loop reprocessing without scripting overhead. Second, connectivity-aware segmentation and topology-preserving surface reconstruction strategies enforce white-matter continuity and cortical-surface stability across species[34, 42] - constraints that conventional voxel-centric models do not capture. Third, the framework provides a unified computational substrate for standardized processing across phylogenetically diverse primates, from marmosets to humans, thereby enabling scalable comparative analyses. We validated PrimateX on 2,334 scans primarily aggregated from four public consortia (common marmoset, n = 446; cynomolgus macaque, n = 162; rhesus macaque, n = 583; chimpanzee, n = 30; and human, n = 1,113). The framework achieves state-of-the-art segmentation and reconstruction

fidelity at operational throughput. Critically, it exhibits zero-shot generalization across 31 NHP cohorts (157 individuals, 19 sites worldwide), including 8 species unseen during development (bonobo, gorilla, orangutan, gibbon, baboon, sooty mangabey, capuchin monkey, squirrel monkey), while maintaining performance parity with established human-specific pipelines such as FSL[1] in their native domain. By delivering adaptive neuroanatomical processing through a transparent, reusable computational architecture, PrimateX establishes a standardized foundation for primate neuroimaging and removes a central obstacle to large-scale comparative and translational neuroscience.

Results

The PrimateX Framework

As illustrated in Fig. 1, PrimateX consists of five key components: 1) a user-proxy agent that handles semantic task representation and prompt parsing; 2) a planning agent responsible for generating reproducible processing sequences; 3) an execution agent that manages task decomposition, module routing, and state-based workflow control, 4) an analysis agent for downstream analytics; and 5) a tool provider that interfaces with external application programming interfaces (APIs) and computational tools. The system begins with an input image and a user-defined request, constructing a structured task specification by integrating imaging attributes with analytical intent. This specification is converted into a fixed execution plan, which chains preprocessing, skull stripping, white-matter segmentation, cortical surface reconstruction, and downstream analysis into an end-to-end workflow. During execution, PrimateX automatically infers species identity and task requirements, directs data to the

appropriate analytical modules, and dynamically orchestrates computational tools under runtime monitoring. This architecture ensures a unified processing framework across five primate species while maintaining standardized outputs. Beyond neuroimaging derivatives, PrimateX generates standardized reports and execution logs documenting intermediate decisions, workflow progression, and final results - facilitating quality control, traceability, and reproducibility (Supplementary Note 1).

Species-aware agentic execution

To demonstrate PrimateX's agentic execution, we tracked a cynomolgus macaque T1-weighted brain scan through the full processing pipeline (Fig. 2). An LLM-based proxy agent first decomposed the natural-language instruction "perform brain reconstruction on this scan" into a standardized task definition - clarifying analytical goals, operational steps, and logical dependencies. Rather than assuming complete input information, the agent verified the integrity of core execution parameters. When essential information was missing (species, skull status, output format), the system queried the user for supplementary details or invoked embedded capabilities. Relevant parameters were automatically inferred via pre-trained species–task routing modules and directory inspection utilities. The task definition and inferred attributes were then passed to a planning agent, which constructed a sequential executable scheme (Fig. 2a-b). Users could interactively review and modify the plan, filling vacant parameters and adjusting output settings. The input scan was identified as skull-intact cynomolgus macaque, activating a tailored optimized processing pipeline (Fig. 2b).

The execution engine operated as a state machine, each discrete state corresponding to an

individual processing step and yielding archived intermediate outputs. Preprocessing produced a normalized, reoriented image; skull stripping and white-matter segmentation then generated a brain-extracted T1-weighted volume and species-specific brain and white-matter masks. These outputs fed into an NHP-optimized pipeline that reconstructed cortical surfaces in both native and standardized spaces using species-matched atlases. Automated quality checks assessed connectivity and surface integrity, followed by surface-based analysis on the derived cortical representations. Throughout the entire operational procedure, the system continuously monitored running status, intermediate documents, module outputs, and log records. Users could conduct real-time inquiries to check progress, view log files, retrieve generated data, troubleshoot abnormal steps, and obtain intermediate QC summaries. All state transitions, module decisions, intermediate results, and final outputs were logged in real time and compiled into a standardized participant-level report. This case demonstrates PrimateX's ability to transform vague natural-language commands into species-adaptive brain reconstruction workflows with interactive user engagement, multi-stage intelligent execution, full operational transparency, and reproducible analytical results.

Cross-species skull stripping performance

We assessed PrimateX skull stripping performance across four NHP species and human subjects using withheld test subsets sourced from seven independent datasets (Extended Data Table 1). Following quality control, 444 of the 446 common marmoset scans were retained for subsequent analyses; two scans were excluded because of corrupted T1-weighted images. For NHP datasets with voxel-level annotations (Extended Data Table 2), model performance was

compared against conventional and deep learning-based approaches via overlap and boundary-related metrics, with supplementary boundary quantification summarized in Supplementary Table 2. As voxel-wise ground truth was inaccessible for human scans, evaluation relied on volumetric and structural consistency against FSL outputs, incorporating correlation strength, inter-subject stability, systematic deviation and spatial similarity assessments.

NHP skull stripping. PrimateX attained the highest Dice similarity coefficient and the lowest 95% Hausdorff (HD95) distance across all 4 NHP species, surpassing all competing algorithms (Fig. 3a, Supplementary Table 2). Taking rhesus macaque scans (n = 8) as an example, PrimateX yielded a Dice score of 0.994 ± 0.004 and HD95 of 0.500 ± 0.000 mm. It outperformed FSL (Dice: 0.518 ± 0.041, $p<0.001$; HD95: 37.940 ± 3.793 mm, $p = 0.014$) and three deep learning models, namely nBEST[31], DeepBet[32] and MedSAM2[36]. Corresponding statistical results were recorded as nBEST (Dice: 0.969 ± 0.002, $p < 0.001$; HD95: 1.044 ± 0.061 mm, $p = 0.012$), DeepBet (Dice: 0.984 ± 0.001, $p < 0.001$; HD95: 0.655 ± 0.096 mm, $p = 0.020$), and MedSAM2 (Dice: 0.493 ± 0.028, $p < 0.001$; HD95: 44.406 ± 8.572 mm, $p = 0.014$). Statistical comparisons adopted paired t-tests for normally distributed data and Wilcoxon signed-rank tests for non-normal distributions. PrimateX outperformed the second-best approach by 0.004 to 0.010 across species. More pronounced gaps emerged in HD95 measurement, where PrimateX lowered the distance metric by 0.075 to 0.236 mm relative to the runner-up method. Visual examination identified consistent segmentation defects in comparative methods across primate species (Fig. 4). Near orbital anterior and ventral areas, nBEST and DeepBet tended to erroneously incorporate periorbital soft tissue or partially

eliminate ventral cortical tissues. Noticeable boundary mismatches also appeared on lateral cortical surfaces, particularly evident in cynomolgus macaque samples. Competitor models frequently failed to fully capture inferior brainstem and cerebellar tissues, producing fragmented marginal segmentation boundaries. In contrast, PrimateX generated anatomically accurate and continuous segmentation contours throughout these vulnerable regions. Computational efficiency remained comparable across all tested algorithms. PrimateX achieved processing speeds of 0.618 ± 0.059 s/MB for common marmoset scans and 0.732 ± 0.047 s/MB for cynomolgus macaque scans, versus 0.778 ± 0.031 s/MB recorded by DeepBet (Fig. 3a). Qualitative segmentation results of unannotated scans from the seen-domain datasets are displayed in Supplementary Figures 2-5, with corresponding volumetric evaluations against manual annotations shown in Supplementary Figure 6.

Human skull stripping. PrimateX yielded brain volume measurements strongly correlated with FSL results across Human Connectome Project test participants. The Pearson correlation coefficient reached 0.999 (95% CI: 0.998-0.999, $p < 0.001$, $n = 223$). Excellent consistency was also validated at the individual subject level, with ICC(2,1) = 0.998 (95% CI: 0.992-0.999) and ICC(3,1) = 0.999 (95% CI: 0.998-0.999, both $p < 0.001$). Bland-Altman analysis indicated a mean volumetric discrepancy of 5.243 mL (95% CI: 4.400-6.086 mL). The 95% limits of agreement spanned −7.274 mL to 17.760 mL, with corresponding confidence intervals of −8.726 to −5.822 mL for the lower bound and 16.308 to 19.212 mL for the upper bound. The structural similarity index between the two methods stood at 0.991 ± 0.002 (95% bootstrap CI: 0.990-0.991; Extended Data Fig. 3a). Visual inspection confirmed highly

consistent overall brain outlines and cortical coverage between PrimateX and FSL segmentations (Fig. 4). Minor local discrepancies emerged around the inferior frontal cortex and ventral brain margin. PrimateX delivered smoother, uninterrupted tissue contours, whereas FSL occasionally displayed jagged boundaries and incomplete brain tissue delineation. In terms of processing speed, PrimateX achieved an average skull-stripping runtime of 0.896 ± 0.063 s/MB for HCP scans, while FSL ran at 0.062 ± 0.003 s/MB (Fig. 3).

Cross-species white-matter segmentation performance

We further evaluated PrimateX on cross-species white matter (WM) segmentation, adopting the identical benchmarking framework and assessment criteria applied to skull stripping analysis (Extended Data Tables 1-2, Supplementary Table 2).

NHP white-matter segmentation. PrimateX achieved the highest mean Dice coefficient and the lowest mean HD95 value across all 4 NHP species (Fig. 3b; Supplementary Table 2). Taking cynomolgus macaque data as an example, PrimateX reached a Dice of 0.932 ± 0.012 and an HD95 of 0.500 ± 0.000 mm, outperforming FSL (Dice, 0.842 ± 0.029, $p < 0.001$; HD95, 1.831 ± 0.403 mm, $p = 0.004$), DeepBet (Dice, 0.871 ± 0.010, $p < 0.001$; HD95, 0.818 ± 0.117 mm, $p = 0.004$), nBEST (Dice, 0.809 ± 0.013, $p < 0.001$; HD95, 1.656 ± 0.365 mm, $p = 0.004$), and MedSAM2 (Dice, 0.681 ± 0.033, $p < 0.001$; HD95, 4.442 ± 0.850 mm, $p = 0.004$) (Fig. 3b). Across all species, PrimateX consistently surpassed other comparative methods, with Dice improvements ranging from 0.015 to 0.080 and HD95 reductions falling between 0.031 to 2.897 mm. Visual assessment revealed that competing methods struggled to accurately segment geometrically intricate white matter structures, including slender, curved

and highly convoluted tissue architectures (Fig. 4). Segmentation outputs frequently presented fragmented contours in compact white matter regions, accompanied by blurred and anatomically unreasonable boundaries caused by tissue leakage into adjacent deep gray matter. These methods also failed to fully delineate superficial white matter along the cortical mantle, exhibiting evident segmentation gaps at cortical surfaces. By contrast, PrimateX yielded continuous, spatially coherent segmentation masks with well-defined boundaries and stable morphological consistency across diverse primate species. Qualitative segmentation results of unannotated scans from the seen-domain datasets are displayed in Supplementary Figures 2-5, with corresponding volumetric evaluations against manual annotations shown in Supplementary Figure 6.

Human white-matter segmentation. Volume measurements derived from PrimateX showed strong consistency with FSL outputs for human white matter segmentation. The Pearson correlation coefficient was 0.999 (95% CI, 0.998-0.999, $p < 0.001$, $n = 223$). Excellent inter-subject reproducibility was confirmed, with ICC(2,1) = 0.998 (95% CI, 0.970-0.999) and ICC(3,1) = 0.999 (95% CI, 0.998-0.999; both $p < 0.001$). Bland-Altman analysis reported a mean difference of 2.892 ml (95% CI, 2.571-3.213), with 95% limits of agreement ranging from -1.874 to 7.658 ml. The lower and upper boundary confidence intervals were -2.427 to -1.321 mL and 7.106 to 8.211 mL, respectively. A high structural similarity index of 0.965 ± 0.003 was obtained (95% bootstrap CI: 0.964-0.966; Extended Data Fig. 3b). Visual comparison verified highly matched major white matter architectures and overall spatial distribution between the two approaches (Fig. 4). Local discrepancies existed in superficial

cortical white matter and borders adjoining deep gray matter. PrimateX generated more intact and uninterrupted segmentation contours, whereas FSL produced minor breaks and irregular margins. In terms of computational efficiency, PrimateX completed segmentation at an average speed of 1.960 ± 0.277 s/MB, substantially faster than FSL's 35.357 ± 2.509 s/MB (Fig. 3). While PrimateX outputs closely correlate with FSL references, neither method represents absolute ground truth, and both automated pipelines carry inherent systematic biases and segmentation errors.

Generalization performance across independent acquisition sites

To validate the generalization capability of PrimateX toward unseen scanning conditions, the pretrained segmentation model was directly deployed on 23 species-site cohorts sourced from 19 distinct scanning sites within the PRIME-DE consortium[26] and National Chimpanzee Brain Resource[28] , covering cynomolgus macaques, rhesus macaques, and chimpanzees (Methods, Extended Data Table 3). No retraining or fine-tuning was conducted prior to inference. The unseen-domain datasets were acquired using three scanner manufacturers, over 10 distinct scanner models, and magnetic field strengths ranging from 1.5 T to 7 T.

Across diverse scanning sites, PrimateX yielded anatomically rational segmentation boundaries, intact superficial and deep white matter architectures, and stable morphological features within each primate species. Representative cross-site segmentation cases are presented in Fig.5, with additional results from cynomolgus and rhesus macaque scanning sites illustrated in Extended Data Fig.2. Cross-site robustness was further verified via qualitative visual evaluation.

Zero-shot generalization to unseen NHP species

We further examined the zero-shot generalization capacity of PrimateX across eight NHP species excluded from training data, namely bonobo, gorilla, orangutan, gibbon, baboon, sooty mangabey, capuchin monkey and squirrel monkey. Relevant datasets were retrieved from the National Chimpanzee Brain Resource (Extended Data Table 3). These species cover broad phylogenetic lineages including hominoids, Old World monkeys and New World monkeys, posing a rigorous test for cross-species adaptability amid prominent anatomical divergence. The model was deployed directly for inference without species-specific fine-tuning or annotated supervision, and segmentation outcomes were validated via visual qualitative assessment. PrimateX generated anatomically reasonable and structurally unified segmentation masks for all tested species, reliably retaining key neuroanatomical boundaries (Supplementary Fig 8). The results demonstrate strong zero-shot transfer performance beyond the original training data scope.

Cross-species cortical surface reconstruction

We evaluated PrimateX for NHP cortical surface reconstruction against NHPPipeline, an NHP adaptation of the openly available HCP structural processing pipeline[3], and a hybrid PrimateX skull-stripping with FreeSurfer white matter segmentation (Methods, Fig. 6, Supplementary Table 3). All analyses were limited to datasets with paired T2-weighted scans (n = 470), a mandatory input for NHPPipeline. Reconstruction quality was quantified using topology metrics derived from the Euler characteristic, including Euler number, topological defect count and hole count.

PrimateX outperformed other two competing pipelines across all measures (Fig. 6). Taking macaque data (n = 19) as an example (Fig. 6b), it produced the highest Euler number (−10.316 ± 5.493), exceeding NHPPipeline (−41.421 ± 13.379) and the NHP-adapted FreeSurfer workflow (−114.158 ± 19.686). PrimateX also yielded the lowest defect count (12.316 ± 5.493), compared with 43.421 ± 13.379 for NHPPipeline and 116.158 ± 19.686 for the adapted FreeSurfer pipeline. Likewise, its hole count (6.158 ± 2.746) was markedly lower than those of NHPPipeline (21.711 ± 6.690) and FreeSurfer (58.079 ± 9.843). In macaques, one-way ANOVA confirmed significant pipeline effects for all three metrics ($P < 0.001$ for all). Post-hoc Tukey's HSD tests further showed significant differences between PrimateX and each reference pipeline across all three metrics ($P < 0.001$ for all). For common marmosets, PrimateX successfully reconstructed all 444 evaluated samples (100.0%). In comparison, NHPPipeline successfully reconstructed 442/444 samples (99.55%), and the hybrid PrimateX skull-stripping with NHP-adapted FreeSurfer workflow successfully reconstructed 443/444 samples (99.77%). Visual inspection confirmed PrimateX's superiority, particularly in artifact-prone regions (e.g., frontal pole and superior temporal sulcus), where it preserved native gyral and sulcal morphology better than the alternatives (Fig. 6). PrimateX also supports human cortical reconstruction, standardizing outputs for cross-species analysis. This unified framework facilitated the macaque-human comparison in Fig. 7, where surfaces from both species were processed and aligned to a common reference.

Downstream surface-based analyses

Cortical surfaces reconstructed via PrimateX are fully compatible with a broad range of

downstream surface-based analyses (Fig. 7). When aligned to a standard reference space, these surfaces support standard vertex-wise morphometric quantification, such as cortical thickness, sulcal depth and curvature. Group-level maps derived from these metrics recapitulate well-characterized cortical organizational patterns (Fig. 7a). The outputs also enable geometric analyses of cortical architecture, including geodesic and Euclidean distance matrices, as well as distance-to-region maps anchored to the primary visual cortex (Fig. 7b).

At the regional scale, PrimateX-generated surfaces facilitate structural covariance analysis using cortical thickness, alongside morphometric inverse divergence (MIND) network analysis based on multivariate surface features. Corresponding node degree distributions can be projected onto the cortical surface for visualization (Fig. 7c). The reconstructed surfaces further support geometric eigenmode decomposition: low-order modes capture global cortical morphology, while high-order modes resolve fine-grained anatomical details (Fig. 7d). Leveraging standardized surface representations, we also performed cross-species projection of cortical areal expansion between macaque and human brains within a unified surface framework (Fig. 7e). Collectively, these results confirm that PrimateX yields high-quality cortical surfaces applicable to morphometric, geometric, network and cross-species comparative analyses.

Robust neuroimaging workflow execution under diverse prompts

We evaluated agent planning and execution across ten neuroimaging workflow steps, five simulated user profiles and 39 prompt configurations spanning nine condition categories (Extended Data Fig. 4). To prevent upstream execution failures from precluding the evaluation

of downstream steps, we adopted a stepwise protocol in which each agent received validated outputs from the preceding step before planning and executing the next operation. This enabled consistent evaluation of individual workflow steps without the confounding effects of earlier failures.

PrimateX consistently preserved the required steps across the four analysis endpoints, with inclusion rates of 98.3% for preprocessing (n = 780), 98.5% for white-matter segmentation (n = 585), 100% for cortical surface reconstruction (n = 390) and 94.4% for surface-based analysis (n = 195). General-purpose coding agents (Codex, Claude and Qwen) achieved comparable rates for preprocessing (98.2–100%) and cortical surface reconstruction (99.2–100%) but performed substantially worse in white-matter segmentation (35.9–39.5%) and surface-based analysis (51.3–68.2%) (Extended Data Fig. 4b). Step-level analysis revealed frequent omissions of domain-specific procedures, particularly ACPC alignment (100% across all three agents) and native-resolution rescaling (53–60%). FreeSurfer-to-GIFTI conversion was also frequently omitted by Codex (88%) and Qwen (84%), although rarely by Claude (2%) (Extended Data Fig. 4c). These omissions concern procedures needed for anatomical alignment, restoration of native physical scale and surface-format interoperability. In comparison, PrimateX recorded 74 omissions across 4,875 applicable workflow-step instances (1.5%), involving six of the ten steps, while the remaining four showed no omissions across the evaluated prompt conditions (Extended Data Fig. 4d).

Discussion

Reproducible primate neuroimaging hinges on standardized structural MRI processing, where

skull stripping and white-matter segmentation remain persistent bottlenecks. While human studies benefit from robust tools like fMRIPrep and BIDS-Apps[4, 5, 25], comparable NHP solutions must contend with interspecies neuroanatomical variation and protocol heterogeneity across multi-site initiatives like PRIME-DE. PrimateX addresses these challenges through integrated technical advances. First, its architecture combines white-matter topological constraints with MRI-style augmentation to enable robust cross-species segmentation across field strengths (1.5-7T) and scanner platforms, eliminating the need for site-specific tuning. This dual approach outperforms conventional tools like NHPPipeline and generalist models such as MedSAM2, reducing downstream topological defects by 1.6-9.4× and yielding analysis-ready cortical surfaces. Second, intelligent workflow adaptation mitigates protocol heterogeneity. An LLM-powered pipeline automatically adjusts to species-specific neuroanatomy, while MRI-style augmentation separates scanner artifacts from true anatomical features, enabling reliable zero-shot generalization. This ensures standardized processing across diverse acquisitions without compromising dataset-specific characteristics. Third, PrimateX implements comprehensive provenance tracking beyond human-focused systems like NeuroAgent[43], logging all planner decisions through PROV-DM-compliant records[44]. This combination of robust processing with complete traceability supports collaborative analysis at scales previously exclusive to human neuroimaging.

Accessible via web interface (https://primatex.pku.edu.cn), PrimateX provides both computational specialists and non-expert users with a powerful tool for neuroimaging analysis while maintaining full reproducibility through containerized, version-pinned execution. Future

expansions by adding specialized agents/skills could incorporate diverse MRI modalities – including task-based and resting-state functional MRI, diffusion MRI, and MR spectroscopic imaging, facilitating holistic structural, functional, and metabolic analyses. More fundamentally, it creates a blueprint for building reproducible, AI-powered pipelines that maintain transparency while managing increasingly complex multi-modal, multi-site, and multi-species datasets.

PrimateX serves as a transformative platform that uniquely integrates bioinformatics, evolutionary neuroscience, AI methodology, and clinical applications. Its standardized primate neuroimaging datasets enables integrative analyses with genomic and transcriptomic data[45, 46], while automated artifact removal and containerized deployment ensure reproducibility and facilitate high-throughput data mining[9, 47]. The platform transforms evolutionary neuroscience through three key capabilities: precise homology-based cortical comparisons across species[24], quantitative tracking of neuroanatomical divergence relative to speciation events, and integration with evolutionary genomics to identify brain evolution drivers. Aligned with contemporary neuroimaging initiatives including white matter micro- and macrostructure brain charts[13], PrimateX delivers topologically-accurate white-matter segmentation and stable cortical surface reconstructions, supporting comparative lifespan studies of brain organization[11-13]. Crucially, PrimateX provides cross-species neuroimaging resources essential for developing foundation models (e.g., BrainIAC for structural MRI[15], NeuroSTORM for fMRI[48]) and neuroscience-inspired AI. These standardized datasets accelerate preclinical translation by validating novel MRI diagnostics, surgical planning and interventional strategies

– effectively connecting primate research with human clinical practice[49].

The development of PrimateX establishes a strong computational framework for continued innovation, with several promising areas identified for future research progress. Our results demonstrate that training-time MRI-style transfer augmentation achieves exceptional cross-site generalization performance across all 31 evaluated cohorts. While the current approach demonstrates significant advantages, its application may benefit from additional extensions to accommodate future datasets with acquisition protocols or scanner characteristics beyond our current training distribution. Importantly, we identify inference-time harmonization strategies[50], as a promising pathway to enhance framework adaptability by enabling dynamic adjustment to novel imaging domains. A critical next step is developing a next-generation primate neuroimaging foundation model trained on high-throughput, phylogenetically diverse, multi-modal datasets with unified neuroanatomical representations. This would move the field beyond species-specific analyses toward a comprehensive, generalizable framework for comparative primate brain research, thereby enabling systematic investigations of evolutionary neurobiology and cognitive architecture.

In brief, PrimateX represents a paradigm shift, integrating connectivity-aware segmentation, topology-preserving reconstruction, and LLM-guided workflow orchestration to transition primate neuroimaging from fragmented processing to standardized, large-scale science. By enabling robust cross-species generalization while ensuring reproducibility - directly supporting multi-institutional consortia like PRIME-DE, it unlocks new possibilities in evolutionary and translational neuroscience.

References

1. Smith, S.M., M. Jenkinson, Woolrich Mark W., and et al. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* **23**, S208-S219 (2004).
2. Fischl, B. FreeSurfer. *NeuroImage* **62**, 774-781 (2012).
3. Glasser, M.F., S.N. Sotiropoulos, Wilson J. Anthony, and et al. The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* **80**, 105-124 (2013).
4. Gorgolewski, K.J., et al. BIDS apps: improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLOS Computational Biology* **13**, e1005209 (2017).
5. Esteban, O., C.J. Markiewicz, Blair Ross W., and et al. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods* **16**, 111-116 (2019).
6. Ad-Dab'bagh, Y., et al. The CIVET image-processing environment: a fully automated comprehensive pipeline for anatomical neuroimaging research. *Proceedings of the 12th annual meeting of the organization for human brain mapping* **2266**, (2006).
7. Gorgolewski, K., et al. Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics* **Volume 5 - 2011**, (2011).
8. Glasser, M.F., et al. The Human Connectome Project's neuroimaging approach. *Nat Neurosci* **19**, 1175-87 (2016).
9. Elliott, L.T., et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562**, 210-216 (2018).
10. Marek, S., et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654-660 (2022).
11. Bethlehem, R.A.I., et al. Brain charts for the human lifespan. *Nature* **604**, 525-533 (2022).
12. Zhuo, Z., et al. Charting brain morphology in international healthy and neurological populations. *Nat Neurosci* **29**, 420-434 (2026).
13. Kim, M.E., et al. White matter micro- and macrostructure brain charts for the human lifespan. *Nature* (2026).
14. Lyu, Y., et al. Learning neuroimaging models from health system-scale data. *Nature Biomedical Engineering* 1-13 (2026).
15. Tak, D., et al. A generalizable foundation model for analysis of human brain MRI. *Nature Neuroscience* 1-12 (2026).
16. Autio, J.A., et al. Minimal specifications for non-human primate MRI: challenges in standardizing and harmonizing data collection. *NeuroImage* **236**, 118082 (2021).
17. Hayashi, T., et al. The nonhuman primate neuroimaging and neuroanatomy project. *Neuroimage* **229**, 117726 (2021).
18. Milham, M., et al. Toward next-generation primate neuroscience: A collaboration-based strategic plan for integrative neuroimaging. *Neuron* **110**, 16-20 (2022).

19. Capitanio, J.P. and M.E. Emborg. Contributions of non-human primates to neuroscience research. *Lancet* **371**, 1126-35 (2008).
20. Miller, C.T., et al. Marmosets: A Neuroscientific Model of Human Social Behavior. *Neuron* **90**, 219-33 (2016).
21. Bernardi, S. and C.D. Salzman. The contribution of nonhuman primate research to the understanding of emotion and cognition and its clinical relevance. *Proc Natl Acad Sci U S A* **116**, 26305-26312 (2019).
22. Scott, J.T. and J.A. Bourne. Modelling behaviors relevant to brain disorders in the nonhuman primate: Are we there yet? *Prog Neurobiol* **208**, 102183 (2022).
23. Stone, R. Ambitious collaboration will map primate brains—including ours—in unprecedented detail. *Science* **390**, (2025).
24. Donahue, C.J., M.F. Glasser, T.M. Preuss, J.K. Rilling, and D.C. Van Essen. Quantitative assessment of prefrontal cortex in humans relative to nonhuman primates. *Proceedings of the National Academy of Sciences* **115**, E5183-E5192 (2018).
25. Eichner, C., et al. Detailed mapping of the complex fiber structure and white matter pathways of the chimpanzee brain. *Nat Methods* **21**, 1122-1130 (2024).
26. Milham, M.P., et al. An open resource for non-human primate imaging. *Neuron* **100**, 61-74 (2018).
27. Milham, M., et al. Accelerating the evolution of nonhuman primate neuroimaging. *Neuron* **105**, 600-603 (2020).
28. Hof, P.R., D. Falk, Erwin Joseph M., and et al. Chimpanzee brain morphometry utilizing standardized MRI preprocessing and macroanatomical annotations. *eLife* **9**, e60136 (2020).
29. Okano, H., S. Muraoka, Kawano Hiroshi, and et al. Brain/MINDS: brain-mapping by integrated neurotechnologies for disease studies. *Philosophical Transactions of the Royal Society B* **370**, 20140310 (2015).
30. Wang, X., et al. U-net model for brain extraction: Trained on humans for transfer to non-human primates. *Neuroimage* **235**, 118001 (2021).
31. Zhong, T., et al. nBEST: deep-learning-based non-human primates brain extraction and segmentation toolbox across ages, sites and species. *NeuroImage* **295**, 120652 (2024).
32. Fisch, L., et al. deepbet: fast brain extraction of T1-weighted MRI using convolutional neural networks. *Computers in Biology and Medicine* **179**, 108845 (2024).
33. Rahmani, M., et al. Evolution of white matter hyperintensity segmentation methods and implementation over the past two decades; an incomplete shift towards deep learning. *Brain Imaging and Behavior* **18**, 1310-1322 (2024).
34. Shit, S., J.C. Paetzold, Ezhov Ivan, and et al. clDice: a novel topology-preserving loss function for tubular structure segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 16006-16015 (2021).
35. Ma, J., et al. Segment anything in medical images. *Nature Communications* **15**, 654 (2024).
36. Ma, J., et al. MedSAM2: segment anything in 3D medical images and videos. *arXiv* (2025).
37. Zhao Theodore and et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature Methods* (2024).

38. Zhang Siteng and et al. A generalist foundation model and database for open-world medical image segmentation. *Nature Biomedical Engineering* (2025).
39. Wang, L., et al. A lifespan-generalizable skull-stripping model for magnetic resonance images that leverages prior knowledge from brain atlases. *Nat Biomed Eng* **9**, 700-715 (2025).
40. Froudist-Walsh, S., M. Browning, Xia Zoe, and et al. CIVET-Macaque: an automated pipeline for MRI-based cortical surface generation and cortical thickness in macaques. *NeuroImage* **225**, 117519 (2021).
41. García-Saldivar, P., H. Merchant, and L. Concha. PREEMACS: pipeline for preprocessing and extraction of the macaque brain surface. *NeuroImage* **225**, 117517 (2021).
42. Kirchhoff, J., M. Ritter, Ritter Katrin, and et al. Skeleton recall loss for connectivity conserving and resource efficient segmentation of thin tubular structures. *European Conference on Computer Vision (ECCV)* (2024).
43. Liu Z. and et al. NeuroAgent: LLM agents for multimodal neuroimaging analysis and research. *arXiv preprint* (2026).
44. Souza, R., A. Gueroudji, DeWitt Stephen, and et al. PROV-AGENT: unified provenance for tracking AI agent interactions in agentic workflows. *IEEE eScience* (2025).
45. Bo, T., et al. Imaging Genomics Across Primate Species: Advancing Understanding of Neurological and Psychiatric Disorders. *Neurosci Bull* (2026).
46. Zhuo, L., et al. MAPbrain: a multi-omics atlas of the primate brain. *Nucleic Acids Research* **53**, D1055-D1065 (2025).
47. Bo, T., et al. Brain-wide and cell-specific transcriptomic insights into MRI-derived cortical morphology in macaque monkeys. *Nature Communications* **14**, 1499 (2023).
48. Dong, J., Y. Sun, Yu X., and et al. Towards a general-purpose foundation model for functional MRI analysis. *Nature Biomedical Engineering* (2026).
49. Zhan, Y., et al. Diagnostic Classification for Human Autism and Obsessive-Compulsive Disorder Based on Machine Learning From a Primate Genetic Model. *Am J Psychiatry* **178**, 65-76 (2021).
50. Karani, N., E. Erdil, K. Chaitanya, and E. Konukoglu. Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis* **68**, 101907 (2021).

Methods

Cohorts

Internal cohorts. A total of 2,334 T1-weighted structural MRI scans were aggregated from five primate species for model development and internal evaluation, including common marmoset (n = 446; Brain/MINDS NA216[29]), cynomolgus macaque (n = 162; 29 scans from

Institute of Neuroscience and 133 scans from Kunming Institute of Zoology, Chinese Academy of Sciences[51]), rhesus macaque (n = 583; University of Wisconsin-Madison), chimpanzee (n = 30; National Chimpanzee Brain Resource[28]), and human (n = 1,113; HCP S1200[3, 52]), with each scan treated as an independent input sample for model development and internal evaluation. During initial quality control, two common marmoset scans (T1WI_410 and T1WI_412) were excluded because their T1-weighted image files were corrupted, leaving 444 scans available for subsequent analyses. All animal imaging procedures were approved by the institutional ethical review boards of respective acquisition sites and performed in accordance with local experimental protocols (Extended Data Table 1). All macaque, chimpanzee, and human datasets were acquired at 3.0 T, while common marmoset scans were obtained at 9.4 T. The voxel resolution varied across cohorts, ranging from an in-plane resolution of 0.27 mm (common marmoset) to an isotropic resolution of 0.9 mm (chimpanzee). Anisotropic acquisition protocols were applied for common marmoset ($0.27 \times 0.27 \times 0.54$ mm^3) and rhesus macaque ($0.273 \times 0.5 \times 0.273$ mm^3) imaging.

Dataset partitioning. Species-specific training and held-out test datasets were predefined prior to model training (Extended Data Table 2). All test scans were strictly excluded from model training, data augmentation, and hyperparameter selection procedures to ensure unbiased performance evaluation. The number of held-out scans used for evaluation differed across the four analytical tasks. Species classification was evaluated on 1,394 held-out scans (994 NHP and 400 human scans), while task classification was evaluated on 1,426 held-out scans (713 scans with skull and 713 skull-stripped scans). Skull-stripping performance was quantitatively

evaluated on 254 held-out scans, comprising 31 annotated NHP scans and 223 human scans. White-matter segmentation was quantitatively evaluated on 250 held-out scans, comprising 27 annotated NHP scans and 223 human scans. All held-out test scans were strictly excluded from model training, data augmentation, and hyperparameter selection to ensure independent evaluation of model performance.

External cohorts (PRIME-DE and NCBR). For cross-site generalization evaluation, the primary dataset consisted of 23 species-site cohorts from rhesus macaques, cynomolgus macaques, and chimpanzees, collected across 19 independent acquisition sites from the PRIME-DE consortium[26] and the National Chimpanzee Brain Resource (NCBR)[28]. To assess cross-species generalization, zero-shot evaluation was further performed on 8 additional species-site cohorts from the NCBR, including Bonobo, Gorilla, Orangutan, Gibbon, Baboon, Sooty mangabey, Capuchin monkey, and Squirrel monkey. These datasets covered major scanner vendors (such as Siemens, Philips, and Bruker) and magnetic field strengths ranging from 1.5 T to 7 T (Extended Data Table 3). Eligible cohorts were selected based on three criteria: (1) availability of complete T1-weighted structural MRI data; (2) availability of species information for cohort characterization; (3) complete public metadata regarding scanner vendor, model, and field strength for cohort characterization (Supplementary Figure 1). Notably, no imaging data, manual annotations, or style transfer references from these external PRIME-DE cohorts were incorporated at any stage of model development. These fully unseen-domain datasets therefore enable rigorous assessment of the model's generalization ability toward novel, out-of-distribution imaging conditions.

Agent-assisted orchestration of structural MRI analysis

PrimateX utilizes four specialized LLM agents - a user-proxy agent, a planner agent, an execution agent and an analysis agent - to automate structural MRI processing through a unified Agent Skills registry and a dedicated tool-provider module (Supplementary Table 4). Each Skill encapsulated a discrete computational operation, including versioned input-output schemas, standardized procedural rules, and required computational resources. By invoking Skills rather than interacting directly with underlying software modules, the system minimized LLM context overhead and enforced strict operational boundaries, reducing unintended side effects. The canonical workflow integrated dozens of predefined Skills (see Supplementary Note 1 for details), along with complete agent prompts and standardized JSON schemas for inter-agent communication.

(1) User-Proxy Agent: processes unstructured natural-language requests, resolves input ambiguities, and prompts users for missing basic information. Additionally, it enables interactive querying, allowing users to retrieve processing summaries and derived outputs via plain-language instructions without requiring programming expertise. (2) Planner Agent: dynamically constructs execution plans by querying the Skill registry. It incorporates user-defined objectives, inferred species-specific parameters (e.g., working resolution, registration templates, and tissue priors), and cross-stage data dependencies to select and parameterize Skills. (3) Execution Agent: operates as a deterministic state machine, sequentially invoking scheduled Skills while monitoring real-time outputs and logs. If execution errors occur, it parses unstructured module outputs to diagnose issues and trigger failover protocols. (4)

Analysis Agent: performs downstream surface-based morphometric analyses using reconstructed cortical data. It further generates structured reports that detail processing trajectories, methodological parameters, and final results in natural language. All agents interact through a shared state memory that persistently stores intermediate outputs, inferred metadata, execution histories and dependency graphs, thereby enabling cross-agent coordination, context-aware decision making and recovery from interrupted or failed processing stages.

PrimateX leveraged a modular multi-agent architecture that decouples semantic parsing, task planning, execution monitoring, and result synthesis for systematic pipeline orchestration. During runtime, the agent iteratively invoked skills, interprets module outputs, and dynamically schedules subsequent steps - either until task completion or upon hitting predefined execution limits. Crucially, a state-machine controller enforced valid transitions across preprocessing, segmentation, reconstruction, and quality control, ensuring strict adherence to version-controlled workflows. All Skill invocations used version-controlled interfaces, and each state transition was logged with timestamps, inputs/outputs, operational states, and runtime metrics, ensuring full provenance tracking for every processing run. This end-to-end auditability ensured reproducibility without sacrificing usability, making cross-species MRI analyses both intuitive (via natural language) and rigorously reliable (via deterministic execution).

Construction of the agent workflow benchmark

We constructed a benchmark spanning four neuroimaging analysis tasks and ten workflow

steps: task routing, preprocessing, skull stripping, ACPC alignment, white-matter segmentation, cortical surface reconstruction, native-resolution rescaling, FreeSurfer-to-GIFTI conversion, surface-based analysis and report generation. Five simulated user profiles represented different levels of expertise: non-specialist, neuroscience user, MRI-tool user, primate MRI practitioner and pipeline expert. Each task–profile combination was tested under 39 prompt configurations, comprising an unperturbed baseline and eight perturbation types (typographical errors, missing information, field reordering, paraphrasing, colloquial language, verbosity, distractors and mixed-language prompts) at up to five levels (L1–L5). This yielded 780 test cases per agent (Extended Data Fig. 4a).

We compared PrimateX with three general-purpose coding agents: Codex (gpt-5.6-sol, Codex CLI v0.154.0-alpha.6.2), Claude Code (claude-opus-5-5, v2.1.310) and Qwen Code (qwen3.8-max, v0.26.0). All agents received the same benchmark requests. PrimateX used its integrated neuroimaging workflow knowledge, while the comparator agents were provided with task requirements and documentation for the corresponding neuroimaging tools. To prevent upstream failures from affecting downstream evaluations, we adopted a stepwise execution protocol in which agents received validated outputs from the preceding workflow step before proceeding to the next. Agents were required to plan and execute the corresponding operations using the available tools. Each test case was initiated in a fresh conversation. All evaluation results were manually verified.

MRI volume preprocessing

Non-human primate MRI volumes underwent standardized preprocessing prior to

segmentation using an nnU-Net-inspired self-configuring pipeline[53] with species-specific parameter settings. The volumes were reoriented to RAS space relative to the sphinx anatomical coordinate system, corrected for intensity non-uniformity using N4 bias-field correction, structurally denoised, spatially normalized to species-specific anatomical templates and resampled to a 0.5-mm isotropic working resolution. For human evaluation, we used T1-weighted images processed with the HCP minimal preprocessing pipeline. These images were provided at 0.7-mm isotropic resolution and were used directly as inputs to PrimateX segmentation.

Species-task router

Architecture and Training. The species-task router utilized a compact 3D convolutional neural network (CNN) architecture designed to simultaneously predict species identity (5-class classification) and skull-stripping status (2-class classification). The network consisted of three sequential 3D convolutional blocks employing 3×3×3 kernels with respective strides of 1, 2, and 2, each followed by batch normalization and ReLU activation. These convolutional layers fed into an adaptive average pooling layer before branching into two parallel linear classification heads. Prior to processing, input volumes underwent z-score normalization and resampling to an isotropic 64^3 resolution. Training leveraged the multi-species datasets detailed in Extended Data Table 2. The model was optimized over 500 epochs using the Adam optimizer with a learning rate of 10^{-4} and batch size of 64. The composite loss function combined two cross-entropy terms corresponding to the joint classification tasks. Comprehensive performance metrics are available in Supplementary Table 1.

Online Data Augmentation. During training, we implemented a stochastic augmentation pipeline applied to each minibatch that included both spatial and intensity transformations. Spatial augmentations comprised random rigid rotations within ±15° about each principal axis and random single-axis flips along the x, y, or z dimensions. Intensity augmentations incorporated additive intensity shifts, multiplicative gamma perturbations, and injected Gaussian noise. These transformation parameters were sampled independently for each volume after z-scoring and resizing. Notably, all training samples including human data received identical augmentation treatment without holdouts. This online augmentation strategy differs from the offline MRI-style transfer procedure used for segmentation model training (Extended Data Table 2).

Inference. During pipeline operation, the router performed a single efficient forward pass on the preprocessed 64^3 input volume, typically completing in under 0.5 seconds per scan. The resulting species and skull-stripping predictions immediately informed downstream processing: the species prediction guided selection of appropriate segmentation models, while the skull-stripping determination controlled whether the skull-removal stage should be executed. Both predictions and their associated confidence scores were systematically recorded in the scan's provenance metadata, enabling retrospective verification of routing decisions during quality control. This automated routing mechanism ensured proper species-specific parameter selection while contributing minimal computational overhead to the overall processing pipeline.

Segmentation network architecture

PrimateX employed a U-shaped 3D encoder-decoder architecture combining two

complementary approaches for volumetric brain segmentation: hierarchical gated convolutional networks (HGCN) for local feature extraction and MoE-Mamba blocks for efficient long-range dependency modeling (Supplementary Note 3). This hybrid design replaced conventional transformer modules, avoiding their prohibitive $O(n^2)$ computational complexity for high-resolution 3D volumes[54].

Encoder. Early stages utilized HGCN blocks with stacked gated convolutions to capture multi-scale spatial interactions. Deeper stages implemented MoE-Mamba blocks combining state-space models[55] with mixture-of-experts routing. The Mamba operation processed volumetric data as serialized sequences with selective scanning, maintaining O(n) complexity per expert evaluation

MoE-Mamba blocks. Each MoE-Mamba block contained $K=3$ expert sub-networks, inspired by BlackMamba[56]. A lightweight MLP router generated data-dependent mixture weights $\mathrm{w} = \mathrm{Gate(x)}$ from the input features x. Parallel expert outputs were combined via weighted summation: $\mathbf{y} = \sum_k w_k \cdot Expert_k(\mathbf{x})$. The design achieved dense expert mixing with linear computational scaling. This combination of MoE and Mamba was intentionally adapted for 3D brain segmentation.

Expert specialization. A de-duplication regularizer penalized pairwise cosine similarity between expert parameter matrices to prevent expert collapse. KL divergence regularization encouraged uniform gating distribution[57]. Unlike Switch Transformer[58] or GShard[59] approaches, explicit load balancing was avoided (full formulation in Supplementary Note 3).

Decoder and skip connections. The decoder structurally mirrored the encoder architecture,

incorporating upsampling blocks at each hierarchical scale. Residual-fused skip connections bridged paired encoder and decoder stages to retain fine-grained anatomical boundary details across feature propagation. Deep supervision losses[60] were applied at all decoder scales, with exponentially decaying weights assigned from coarse to fine hierarchical levels to prioritize fine-resolution structural learning.

White-matter segmentation enhancements. PrimateX incorporated a specialized loss function to preserve white-matter tract connectivity, combining two key components. First, a skeleton-informed Dice loss evaluated soft skeletons of predictions and ground truth, explicitly penalizing topological discontinuities in thin tracts[34, 42]. Second, an uncertainty-guided resampling scheme dynamically focused training on challenging regions by sample-level weighting 1×, 5×, or 10× based on predictive entropy in both target and boundary regions. The complete mathematical formulation, including skeleton computation and entropy weighting, is detailed in Supplementary Note 4.

Training protocol. The skull-stripping objective combined: (1) standard Dice loss, (2) octant-partitioned Tversky loss ($\alpha = 0.3$, $\beta = 0.4$) with adaptive weights based on local error statistics, (3) MoE gating regularization (KL divergence from uniform prior), (4) expert diversity regularization (pairwise cosine similarity), and (5) deep supervision across scales. For white-matter segmentation, we augmented this with the connectivity-aware skeleton loss (see White-matter segmentation enhancements). All models were trained using Adam (lr = 1×10^{-4}, weight decay = 3×10^{-7}) for 300 epochs. We maintained consistent hyperparameters across species despite varying dataset sizes (15-890 native samples per species). MRI-style transfer provided

~21× augmentation for non-human primates, with marmoset training volumes increasing to 315. Validation loss tracking prevented overfitting. Key tuned parameters included: connectivity loss weight (λ_{skel}), MoE auxiliary loss weight (L_{MoE}), diversity regularization (L_{div}), deep supervision decay rate, boundary uncertainty width (δ), and MRI-transfer coefficients (λ_c, λ_s). Final values are documented in Supplementary Notes 4. PrimateX routes each scan to one of ten species-and-task-specific networks using an automated classifier (five species × skull stripping/white-matter segmentation). This single-network approach maintains specialization while optimizing computational efficiency.

MRI-style transfer augmentation

We developed offline MRI-style transfer to harmonize NHP training data across sites while preserving anatomical fidelity. The model adapted the StyTr² architecture[61] to MRI, employing Transformer-based encoders for style and content separation with cross-attention decoding. Key features included 2.5D adjacent-slice input for slice continuity, content-aware positional encoding for voxel spacing invariance, and a VGG-19-based loss balancing structural and stylistic fidelity. Uniform human (HCP) data remained unmodified. Implementation details are in Supplementary Note 2.

Cortical surface reconstruction

Following skull stripping, brain volumes underwent anterior commissure-posterior commissure (ACPC) alignment using species-specific reference frames to standardize brain orientation. PrimateX then employed a modified FreeSurfer[2] workflow adapted for non-human primate anatomy. Key adaptations included representing all inputs with 1-mm headers for

compatibility, substituting FreeSurfer's default brain and white matter masks with deep-learning-derived counterparts, and performing species-adjusted template alignment. The pipeline maintained FreeSurfer's robust topological correction while replacing its primate-vulnerable components with learned anatomical priors, followed by native-resolution rescaling of all outputs.

For human data (HCP S1200), we used FastSurfer instead, with identical downstream processing. All reconstructed surfaces underwent HCP-style standardization, including conversion to GIFTI format, generation of multiple surface representations (white, pial, midthickness), and computation of morphometric features. Species-specific spherical registration aligned surfaces[62] to appropriate templates (e.g., marmoset to Brain/MINDS MBM v3, macaques to Yerkes19[63], chimpanzee to chimpyerkes29[64], human to fs_LR[3]), followed by resampling to standardized 32k meshes. This dual-output approach preserved native-resolution anatomical accuracy while enabling cross-species surface-based analysis through compatible derivatives.

Downstream surface-based analyses

All analyses were performed on macaque cortical reconstructions registered to fsLR-32k space, employing both vertex-wise and parcellated approaches (91-hemisphere Markov parcellation[65]). We conducted five primary analytical workflows:

Vertex-wise morphometrics. Cortical thickness, sulcal depth, and curvature values were resampled to template space and averaged across subjects to create group maps.

Cortical geometry characterization. We computed both geodesic (surface-constrained) and Euclidean (straight-line) distance measures between parcels, with the primary visual cortex (V1) serving as a reference seed region for distance mapping.

Morphometric network construction. Structural covariance networks were derived from parcel-level thickness correlations. Morphometric inverse divergence (MIND) networks[66] were computed to represent multivariate feature relationships across five surface-based measures (cortical thickness, cortical volume, surface area, mean curvature, and sulcal depth).

Geometric eigenmodes. For spectral decomposition, we generated an average cohort midthickness surface and computed eigenmodes by solving the Laplace-Beltrami operator[67]. This produced a hierarchical set of spatial basis functions, ranked from low- to high-frequency geometric patterns. We then evaluated surface reconstruction fidelity by progressively approximating the cortical geometry using subsets of these eigenmodes.

Cross-species cortical expansion. Cortical expansion patterns between humans and macaques were quantified by comparing surface area ratios of corresponding cortical facets across species. This surface-based approach first established inter-species topological correspondence, then calculated local expansion factors while preserving spatial relationships. The resulting expansion maps were projected bi-directionally between species templates, revealing evolutionary scaling patterns through intrinsic surface geometry without volumetric distortions.

Benchmarking and statistical analysis

Segmentation accuracy was assessed by comparing automated skull stripping and white-matter

segmentation against manual annotations performed by trained imaging experts on held-out test sets for each species. These annotations were subsequently reviewed by a senior neuroanatomy specialist. Performance metrics included the Dice similarity coefficient for volumetric overlap and the 95th-percentile Hausdorff distance (HD95) for boundary precision, with average symmetric surface distance (ASSD) and normalized surface Dice at a 0.5-mm tolerance (NSD@0.5 mm) included as additional boundary metrics (Supplementary Table 2). PrimateX was compared separately with each alternative method within each species. Dice was analysed using two-sided paired t-tests, whereas HD95, ASSD and NSD@0.5 mm were analysed using two-sided paired Wilcoxon signed-rank tests. No multiple-comparison correction was applied to these prespecified segmentation comparisons; all segmentation P values are unadjusted. Processing efficiency was standardized as seconds per megabyte (s/MB) to account for interspecies differences in brain size and complexity, with all cases processed sequentially.

For cortical surface reconstruction, PrimateX was compared against the Washington University NHPPipeline and an NHP-adapted FreeSurfer implementation. The NHPPipeline comparison was limited to datasets with available T2-weighted scans due to its protocol requirements. For this comparison, paired T2-weighted scans were available for common marmoset (n = 446), chimpanzee (n = 7) and macaque (n = 19) and were acquired with species-specific protocols: common marmoset, 9.4-T Bruker BioSpec RARE, 0.27 × 0.27 × 0.54 mm^3, TR/TE = 4,000/22 ms; chimpanzee, 3-T Siemens TrioTim 3D SPACE, 0.8-mm isotropic, TR/TE = 3,200/449 ms; macaque, 3-T Siemens TrioTim 2D TSE/3D SPACE, 0.5 × 0.5 × 1.0–

1.5 mm³ or 0.5-mm isotropic, TR/TE = 7,000/43 ms or 3,000/372 ms. These T2 scans were used only for NHPPipeline benchmarking and not for PrimateX training. Reconstruction quality was assessed through Euler characteristic-derived topological metrics (defect and hole counts). Statistical differences across workflows were evaluated using one-way ANOVA with Tukey's HSD post hoc testing, supplemented by Kruskal-Wallis with Dunn's test when normality assumptions were violated ($\alpha = 0.05$ threshold).

Cross-site and Cross-species generalization analyses

We evaluated PrimateX's generalizability using MRI scans from 23 species-site, unseen-domain cohorts of rhesus macaques, cynomolgus macaques, and chimpanzees, collected across 19 independent acquisition sites from the PRIME-DE and the NCBR (Extended Data Table 3), along with unannotated seen-domain datasets from Brain/MINDS, KIZ, National Chimpanzee Brain Resource, and UW-Madison (Extended Data Table 1). Since ground truth segmentations were unavailable for these datasets, performance was assessed through qualitative evaluation of skull-stripping/white-matter masks (Supplementary Fig 7). In addition, zero-shot cross-species generalization was assessed on 8 species-site cohorts from previously unseen NHP species (Extended Data Table 3).

Human Concordance Validation

For human validation, we compared PrimateX against FSL using the HCP dataset. Agreement measures included Pearson's correlation, ICC(2,1) (absolute agreement), ICC(3,1) (consistency), and Bland-Altman analysis for total brain and white matter volumes (Fig. 4c). We additionally computed voxel-wise structural similarity (SSIM) between PrimateX and FSL

outputs (Extended Data Fig. 3).

Methodological comparisons

We benchmarked PrimateX against DeepBet and MedSAM2, retraining both models on identical training data with matching optimization parameters (Adam optimizer, learning-rate schedules, batch sizes, weight decay and epoch budget) (see Methods for details). All models shared the same spatial and intensity augmentations during training, though only PrimateX benefited from our specialized MRI-style transfer augmentation. For classical reference, we included FSL applied with default settings and without NHP-specific adaptation. The DeepBet evaluation included both its pretrained public weights and our retrained version, with the latter showing consistently superior performance across all four NHP species and thus serving as our primary comparison. The nBEST was tested using its official inference-only Docker container (wxyabc/nbest:1.3; sha256:a1c2268493d1da4d25c81aefe0426c2610a1d49c49d604467192751ff3407e7b), which precluded retraining due to unavailable training code.

Statistical analysis

We compared PrimateX against alternative methods using paired t-tests when per-pair differences met normality assumptions (Shapiro-Wilk test, $\alpha=0.05$); otherwise, the paired Wilcoxon signed-rank test was applied. Primary segmentation metrics (Dice, HD95) are reported as species-specific means ± standard deviations, with 95% confidence intervals derived from stratified bootstrap resampling ($B = 10{,}000$ replicates). Boundary metrics and full statistical details are provided in Supplementary Table 2.

Hardware and runtime

Training and inference for deep learning were conducted on an NVIDIA RTX A6000 GPU (48GB VRAM) in a shared cluster environment. CPU-intensive tasks (registration, surface reconstruction, and file conversion) were performed on a workstation equipped with dual Intel Xeon Gold 6226R CPUs (32 physical cores total) and 256GB RAM. Runtime benchmarks - reported in seconds per megabyte (s/MB) of input data - reflect per-volume wall-clock timing, with cases processed sequentially to enable standardized cross-method comparisons. The containerized release supports deployment on any CUDA-enabled Linux system matching these dependencies, simulating a typical single-workstation setup.

Software and container release

PrimateX is implemented as a modular monorepo that unifies preprocessing, deep learning-based segmentation, cortical reconstruction, and workflow management. The system features both a command-line interface (primx) for task coordination and a web-based graphical interface (Node.js) that enables pipeline monitoring, report generation, and real-time visualization using NiiVue (https://github.com/niivue/niivue) for inspecting T1 images, segmentations, and surfaces. For simplified deployment, the complete software stack - including neuroimaging tools (FreeSurfer 7.2.0, ANTs 2.4.3, FSL 6.0.7.8, MRtrix3 3.0.2, Connectome Workbench 2.1.0, FastSurfer 7.2.0) and deep learning libraries (Python 3.10, PyTorch 2.0.1) - is packaged in a Singularity/Docker container built via NeuroDocker. Pretrained models and container images are distributed via cloud storage to minimize installation effort, while source code remains separately available.

Standardized outputs

PrimateX generates standardized neuroimaging outputs for marmosets, cynomolgus macaques, rhesus macaques, chimpanzees, and humans, comprising preprocessed T1-weighted images alongside derived structural measures. The pipeline generates brain masks, skull-stripped volumes, white matter segmentations, and reconstructed cortical surfaces at multiple anatomical boundaries (white, pial, midthickness, and inflated surfaces) with corresponding surface-based parcellations and quantitative geometric measurements.

For each processed subject, an interactive HTML report comprehensively documents processing outcomes, including subject identifiers, acquisition metadata, image resolution parameters, and quality-controlled visualization of preprocessing outputs. These reports further incorporate skull-stripping validations, white matter segmentation results, cortical surface reconstructions, and region-specific geodesic/Euclidean distance heatmaps accompanied by statistical summaries to support downstream morphometric analyses across species. Currently, reports for other primate species are limited to preprocessing validation, skull-stripping results, and white matter segmentation outputs without surface reconstruction capabilities due to the absence of species-specific neuroanatomical templates.

Data availability

Common marmoset MRI scans were obtained from the Brain/MINDS Marmoset MRI NA216 public repositories (https://dataportal.brainminds.jp/marmoset-mri-na216#datadownload). For cynomolgus macaques, we used in-house training datasets provided by Zheng Wang's laboratory and independent validation datasets from the PRIME-DE consortium

(http://fcon_1000.projects.nitrc.org/indi/indiPRIME.html). The in-house cynomolgus macaque MRI data are not publicly available but may be obtained from the corresponding authors upon reasonable request. Rhesus macaque data were similarly sourced from PRIME-DE. Chimpanzee neuroimaging data were accessed through the National Chimpanzee Brain Resource (https://www.chimpanzeebrain.org/) following standard data access procedures. Human brain MRI scans came from the extensively characterized HCP Young Adult S1200 dataset (https://www.humanconnectome.org/study/hcp-young-adult).

Code availability

PrimateX source code and the corresponding documentation are hosted at https://github.com/YCHuang0610/PrimateX. The repository is currently private during peer review. Editors and reviewers may request access from the corresponding author. The repository will be made publicly available under the AGPLv3 License upon publication. An interactive online platform of PrimateX is available at https://primatex.pku.edu.cn.

Acknowledgements

This study was supported by Brain Science and Brain-like Intelligence Technology - National Science and Technology Major Project (No. 2021ZD0204002), the National Natural Science Foundation of China (No. U25C2016, 82151303, 62673148, 62102115, 62103116), the Shandong Provincial Natural Science Foundation (ZR2024QF081), the Hainan Provincial Natural Science Foundation of China (526QN0672, 626MS0266), and grants from Peking-Tsinghua Center for Life Sciences. We particularly acknowledge the vital contribution from open datasets used in this study: the PRIME-DE, the UNC-Wisconsin Rhesus Macaque MRI

Database, the Brain/MINDS Marmoset MRI database and the National Chimpanzee Brain Resource.

Author contributions

Z.W., S.C., and X.H.Y. conceptualized and led the project. X.H.Y. and K.L.C. designed the MoE-Mamba segmentation network and implemented the segmentation training framework, harmonization training, and data augmentation strategies. Y.C.H. developed the LLM-agent architecture, AgentSkills, data processing pipelines, and implemented the PrimateX website with the help of Y.F.W. X.H.Y., Y.C.H., and K.L.C. organized and preprocessed the data, conducted the analyses and drafted the manuscript. K.N. analyzed and validated segmentation and reconstruction results of common marmosets. Z.W. and S.C. edited the manuscript, acquired the funding support, and supervised the overall project. All other authors were involved in one or more of the following tasks: data acquisition and collection, organizing datasets, analyzing/validating results, interpreting findings, editing the manuscript or reviewing the manuscript.

Competing interests

Z.W., S.C., X.H.Y., K.L.C., Y.C.H., and Y.F.W. are listed as co-inventors on filed patents covering multi-agent orchestrated standardized pipelines for cross-species primate brain MRI processing. All other authors declare no relevant financial conflicts of interest associated with this research.

References

51. Lv, Q., et al. Normative analysis of individual brain differences based on a population MRI-based atlas of cynomolgus macaques. *Cerebral Cortex* **31**, 341-355 (2021).
52. Van Essen, D.C., et al. The WU-Minn human connectome project: an overview. *Neuroimage* **80**, 62-79 (2013).
53. Isensee, F., P.F. Jaeger, S.A.A. Kohl, J. Petersen, and K.H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **18**, 203-211 (2021).
54. Cong, S., et al. Comprehensive review of Transformer-based models in neuroscience, neurology, and psychiatry. *Brain-X* **2**, e57 (2024).
55. Gu, A. and T. Dao. Mamba: linear-time sequence modeling with selective state spaces. *arXiv* (2023).
56. Anthony, Q., Y. Tokpanov, P. Glorioso, and B. Millidge. BlackMamba: mixture of experts for state-space models. *International Conference on Learning Representations (ICLR)* (2024).
57. Shazeer, N., et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. *International Conference on Learning Representations (ICLR)* (2017).
58. Fedus, W., B. Zoph, and N. Shazeer. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* **23**, 1-39 (2022).
59. Lepikhin, D., et al. GShard: scaling giant models with conditional computation and automatic sharding. *International Conference on Learning Representations (ICLR)* (2021).
60. Lee, C.-Y., S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)* **38**, 562-570 (2015).
61. Deng, Y., et al. StyTr^2: image style transfer with transformers. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 11326-11336 (2022).
62. Robinson, E.C., et al. Multimodal surface matching with higher-order smoothness constraints. *NeuroImage* **167**, 453-465 (2018).
63. Donahue, C.J., et al. Using diffusion tractography to predict cortical connection strength and distance: a quantitative comparison with tracers in the monkey. *Journal of Neuroscience* **36**, 6758-6770 (2016).
64. Van Essen, D.C., et al. The brain analysis library of spatial maps and atlases (BALSA) database. *Neuroimage* **144**, 270-274 (2017).
65. Markov, N.T., et al. A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cerebral Cortex* **24**, 17-36 (2014).
66. Sebenius, I., et al. Robust estimation of cortical similarity networks from brain MRI. *Nature Neuroscience* **26**, 1461-1471 (2023).
67. Pang, J.C., et al. Geometric constraints on human brain function. *Nature* **618**, 566-574 (2023).

Figure Legends

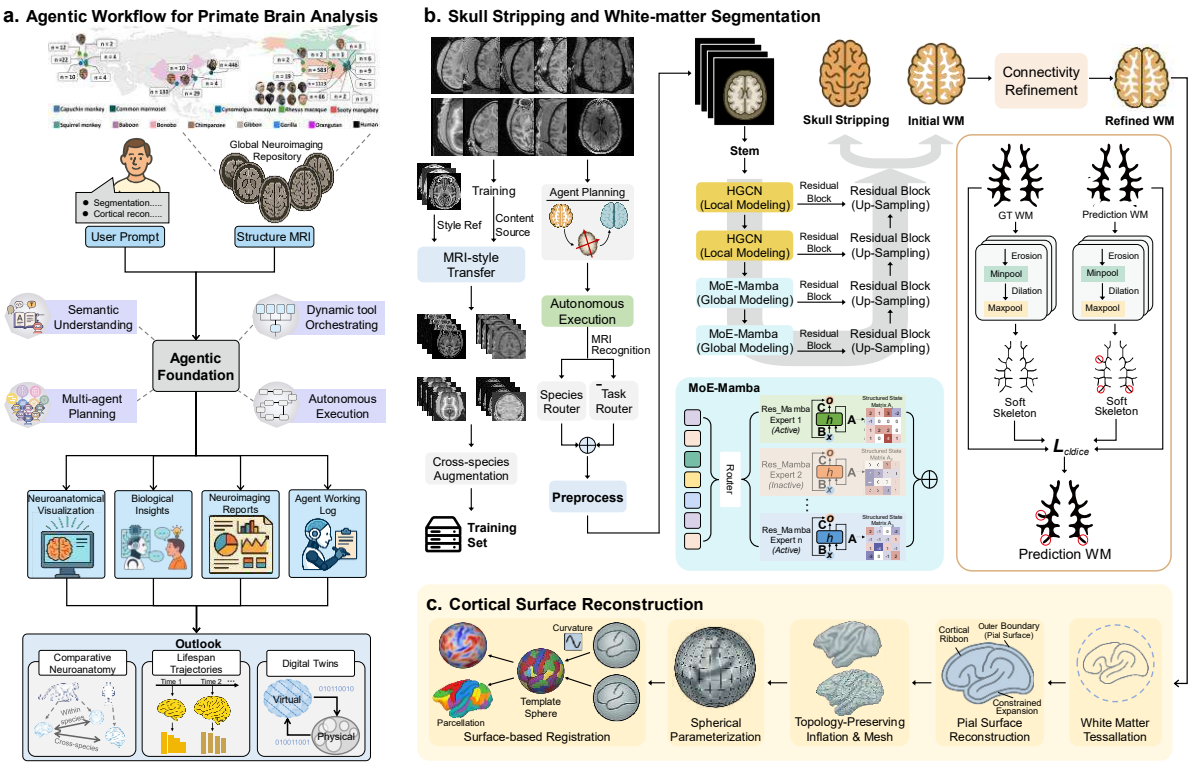


Figure 1 | PrimateX: agentic conversational AI for standardized primate neuroimaging analysis. a, Agentic neuroimaging workflow. PrimateX integrates user inputs with multi-species structural MRI data from global neuroimaging repositories through a unified pipeline coordinating (i) semantic parsing of instructions, (ii) multi-agent planning, (iii) dynamic tool orchestration, and (iv) autonomous execution. Processed outputs include species-specific neuroanatomical visualization, quantifiable biological summaries, neuroimaging reports, and provenance-tracked reports. From processed structural outputs, PrimateX facilitates comprehensive downstream analyses spanning cross-species neuroanatomical comparisons, longitudinal population morphometry, foundation models/digital twin simulation. **b,** Species-specific trained models. Input volumes undergo MRI-style transfer augmentation and agent-guided preprocessing before routing through species-specific pathways. Segmentation

combines hierarchical gated convolutional networks (HGCNs; local feature extraction) with mixture-of-experts Mamba blocks (MoE-Mamba; global context modeling), followed by white-matter connectivity-aware refinement using soft-skeleton representations. **c,** Topology-preserving surface reconstruction. White-matter masks are tessellated into initial surfaces, followed by pial surface generation, inflation with spherical topology constraints, and registration to a species-appropriate atlas for parcellation.

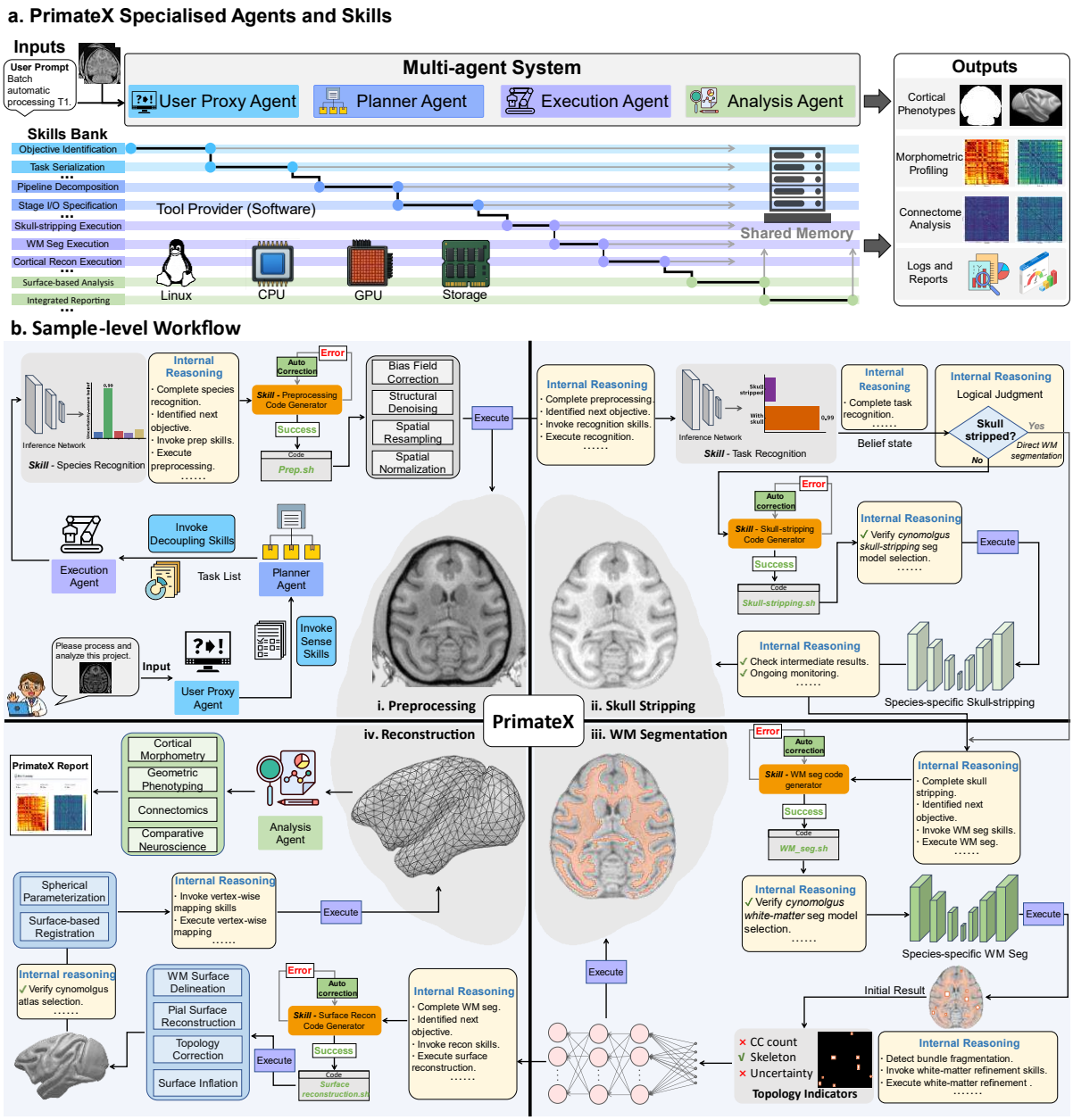


Figure 2 | Agentic analysis pipeline of PrimateX. a, Multi-agent/skill architecture. User requests and structural MRI data are processed through four specialized AI agents: User Proxy Agent (mediates natural language inputs), Planner Agent (decomposes tasks into executable steps), Execution Agent (orchestrates tool/API calls) and Analysis Agent (derives phenotypes and generates reports). Agents interact via a shared memory module and a structured skills bank (Methods), enabling automated computation of cortical morphometry, connectomics, and

species-specific analytics. **b,** End-to-end workflow of a cynomolgus macaque MRI scan. (i) Input routing: Species-recognition skills direct data to the appropriate pipeline. (ii) Preprocessing: Auto-generated tasks (bias correction, denoising, resampling and spatial normalization) are executed. (iii) Segmentation: Species-specific models perform skull-stripping and white-matter segmentation. Topological refinement is implemented with connected-component analysis and soft-skeleton representations. (iv) Surface reconstruction: Using the refined white-matter segmentations, the pipeline executes: white-matter boundary delineation, pial surface reconstruction, followed by topology correction, surface inflation, spherical parameterization, and surface-based registration. (v) Output generation: Cortical thickness maps, geometric phenotypes, and connectomes are computed, followed by automated report synthesis.

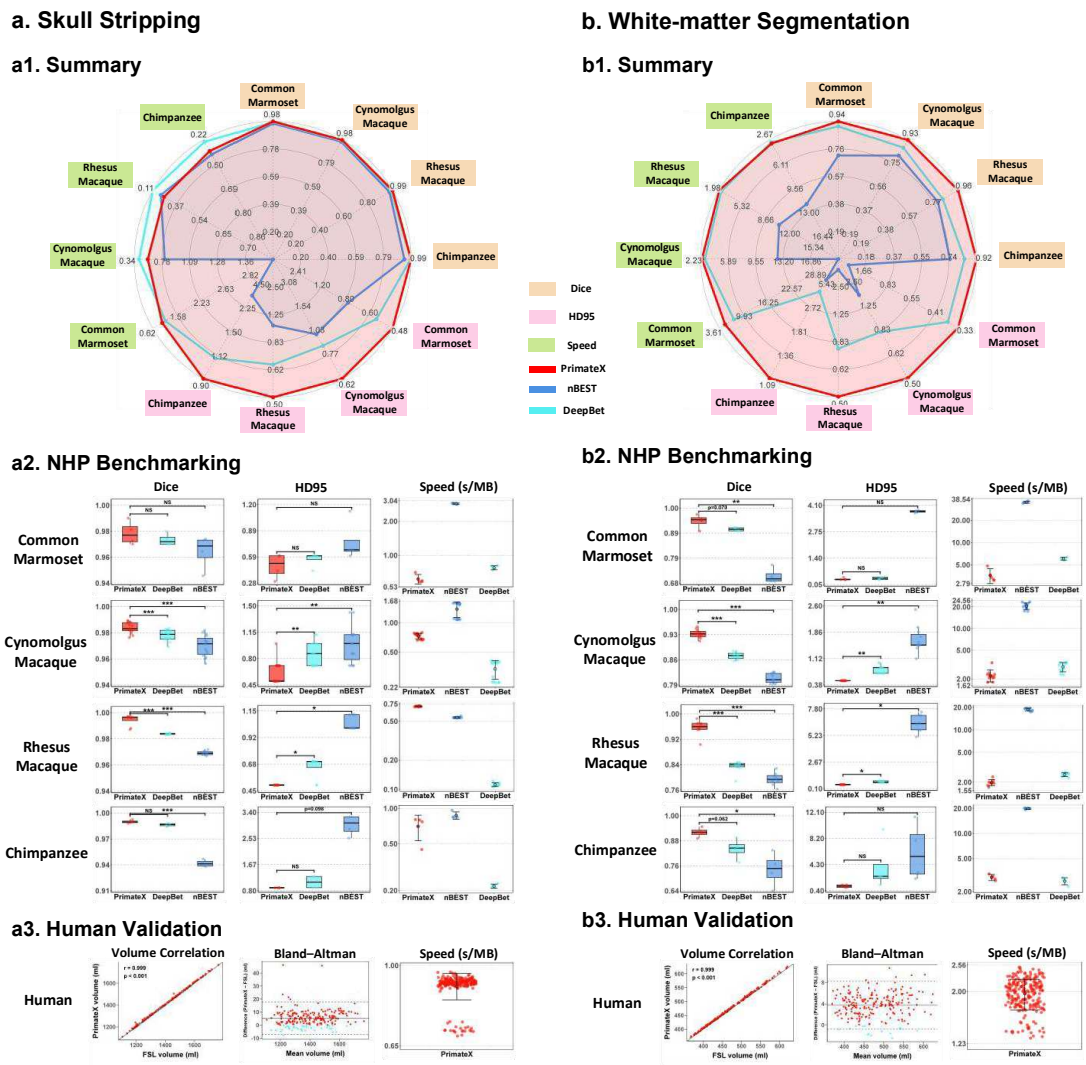


Figure. 3 | Cross-species performance of PrimateX in skull stripping and white-matter segmentation. a, Skull stripping evaluation. a1, Multi-species summary: Radar plots compare PrimateX, nBEST, and DeepBet across common marmoset (n = 4), cynomolgus macaque (n = 15), rhesus macaque (n = 8), and chimpanzee (n = 4) using Dice score (overlap), HD95 (boundary error), and processing speed (s/MB). a2, NHP benchmarking: Box plots reveal per-species distributions of the same metrics. PrimateX achieves superior Dice scores and lower HD95 than baselines ($P < 0.05$), with comparable efficiency. a3, Human validation: Volumetric correlation and Bland-Altman analysis demonstrate strong agreement between PrimateX and

FSL in human data (right: speed comparison). **b,** White-matter segmentation benchmarking. b1, Multi-species summary: Radar plots evaluate performance (Dice, HD95, speed) for marmoset (n = 4), cynomolgus macaque (n = 11), rhesus macaque (n = 8), and chimpanzee (n = 4). b2, NHP benchmarking: Box plots highlight PrimateX's higher accuracy (↑Dice) and lower boundary error (↓HD95) versus alternatives, especially in species with complex cortical morphology (e.g., chimpanzee). Competing methods exhibit greater variability in these cases. b3, Human validation: Agreement with FSL (volumetric correlation/Bland-Altman) and speed metrics confirm human-compatible performance. Statistics: paired two-sided t-tests (Dice) and Wilcoxon signed-rank tests (HD95). P values: *< 0.05, ** < 0.005, *** < 0.0005; NS, not significant.

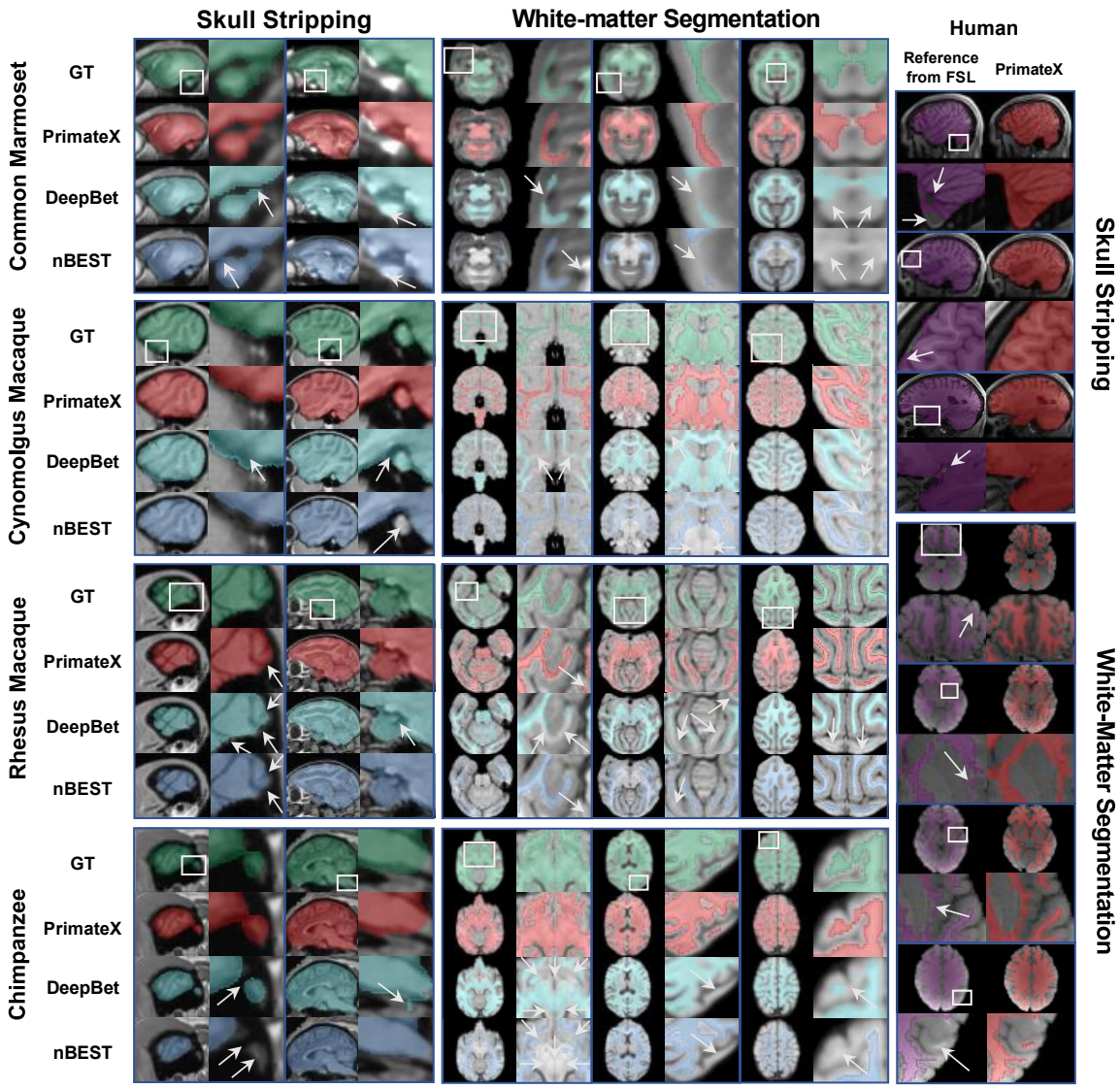


Figure. 4 | Qualitative superiority of PrimateX in cross-species skull stripping and white-matter segmentation. Representative results for four NHP species - common marmoset, cynomolgus macaque, rhesus macaque and chimpanzee – illustrate (left) skull stripping and (middle) white-matter segmentation performance, alongside (right) human validation. In skull stripping, PrimateX minimizes boundary leakage/under-segmentation versus baselines. Brain masks adhere more precisely to reference annotations (arrows: key differences in insets). In white-matter segmentation, PrimateX yields anatomically coherent delineation, with fewer discontinuities in complex regions (e.g., cortical folds), and outperforms baselines in boundary

precision, especially for species with rich cortical morphology. In human validation, PrimateX maintains close agreement with FSL reference while preserving fine-scale edge definition (see insets).

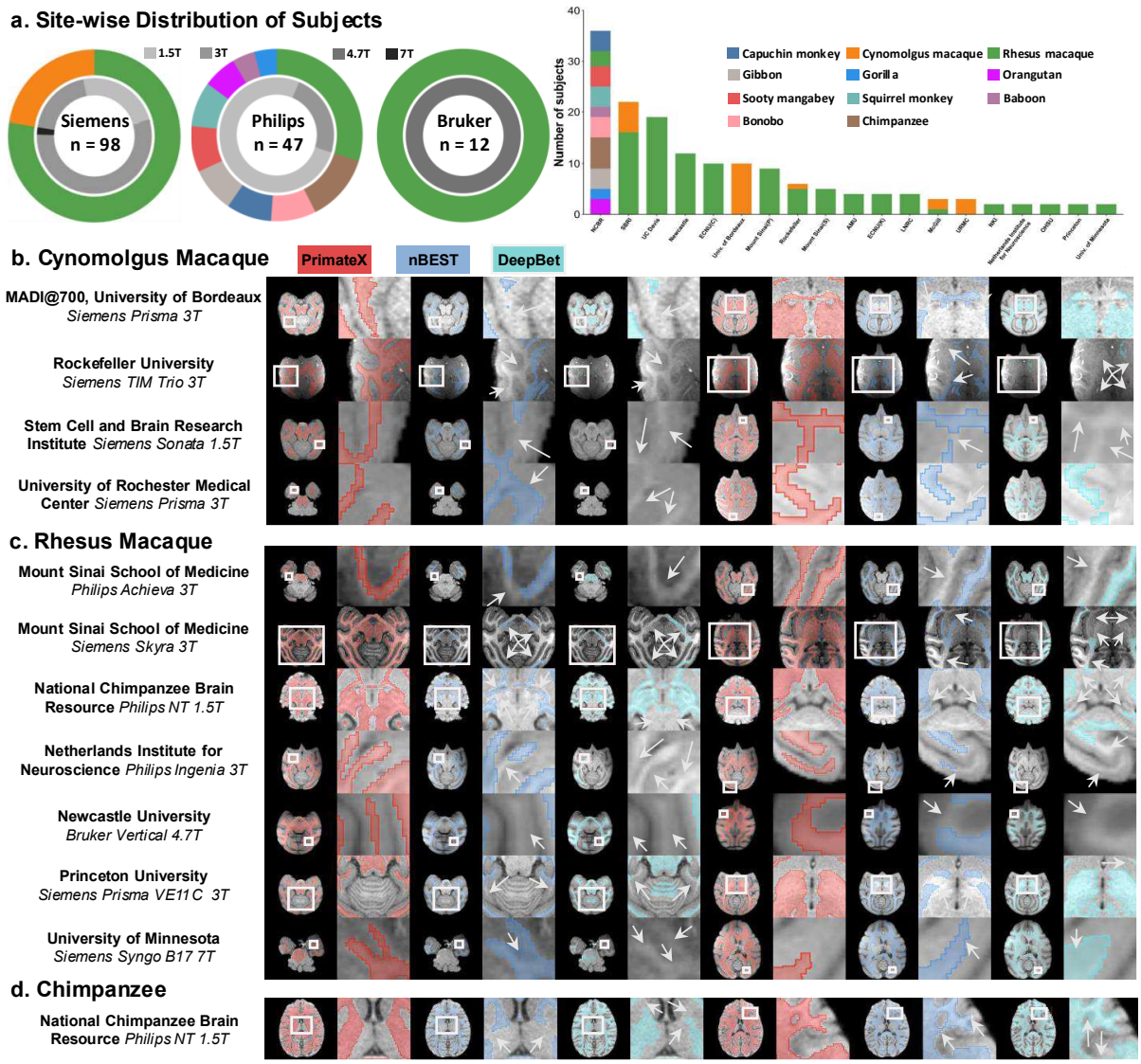
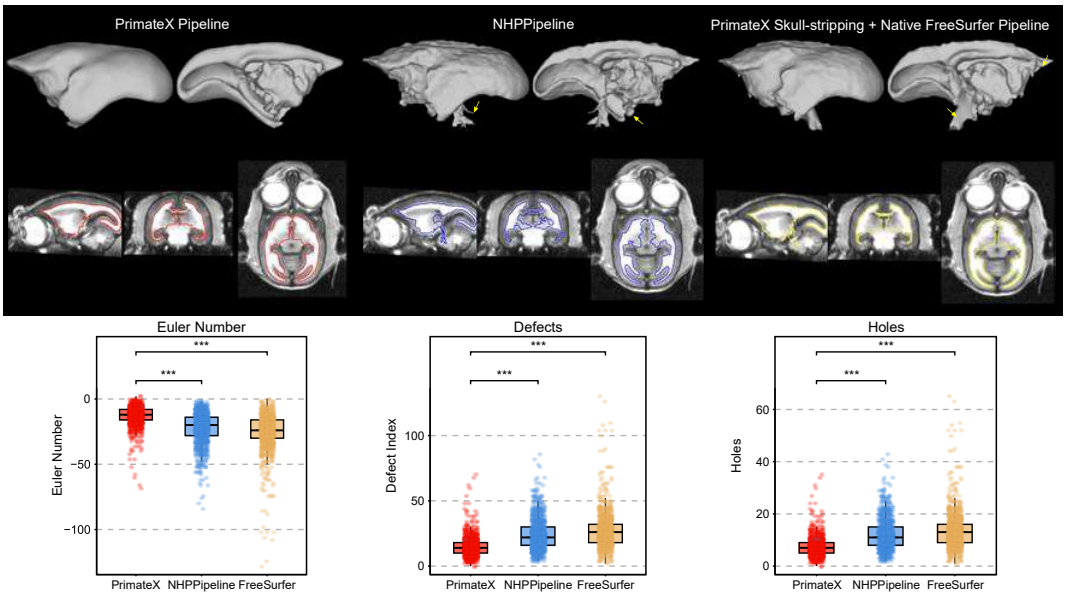
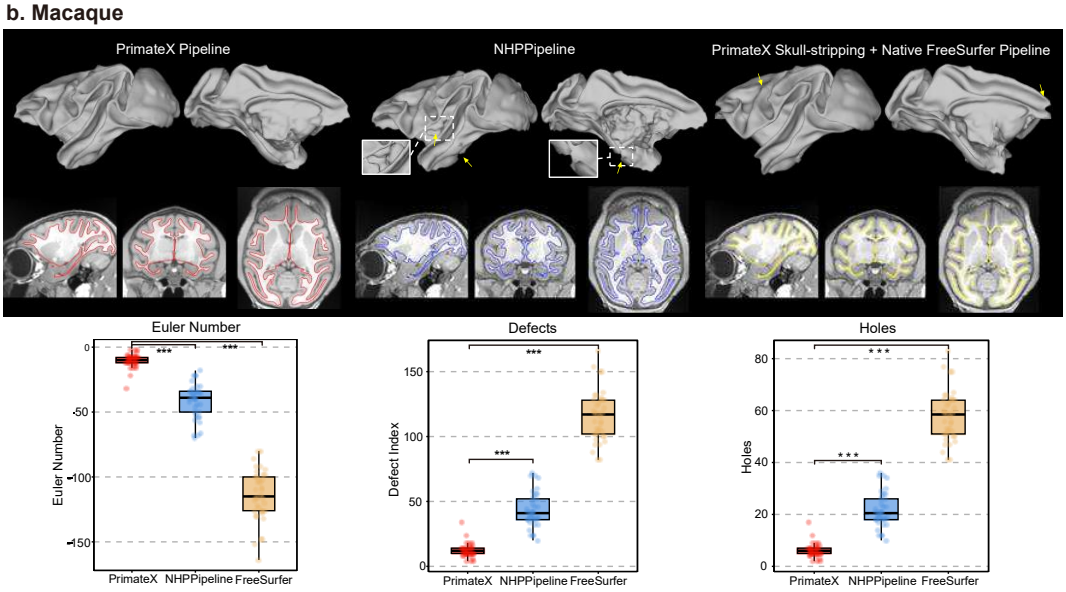


Figure. 5 | Cross-site robustness of PrimateX in white-matter segmentation across acquisition platforms. a, Dataset characteristics: distribution of scans from multiple sites shows 11 NHP species imaged across different scanner vendors and field strengths (1.5T-7T). **(b-d),** comparative performance of PrimateX versus nBEST and DeepBet in **b,** cynomolgus macaque, **c,** rhesus macaque and **d,** chimpanzee. PrimateX maintains anatomically precise white-matter segmentation despite drastic variations in scanner manufacturers, scanner model and magnetic field strengths.

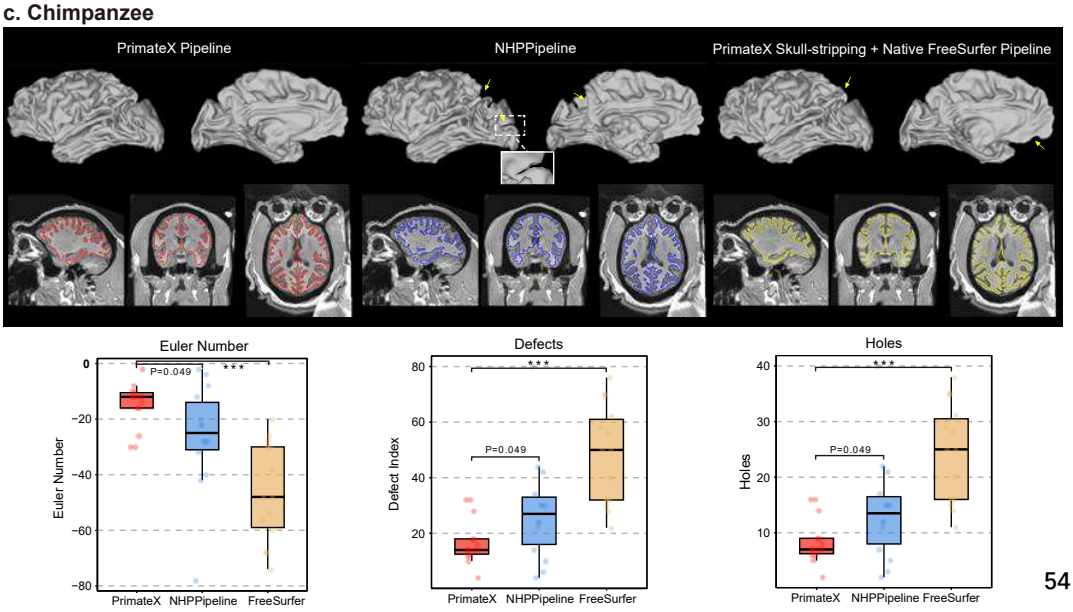
a. Common Marmoset



b. Macaque



c. Chimpanzee



1

Figure. 6 | PrimateX enhances cortical surface topological accuracy across species. Cortical surface reconstructions and their topological integrity were evaluated across three preprocessing workflows: (1) full PrimateX processing, including deep-learning-based brain extraction and white-matter segmentation; (2) the standard NHPPipeline (https://github.com/Washington-University/NHPPipelines); and (3) hybrid PrimateX skull-stripping with FreeSurfer white-matter segmentation. Representative cortical surface reconstructions and quantitative topological assessments are shown for (a) common marmosets, (b) macaques and (c) chimpanzees. For marmosets, valid reconstructions were obtained for 444, 442 and 443 subjects using PrimateX, NHPPipeline and the hybrid FreeSurfer workflow, respectively. The macaque and chimpanzee analyses included 19 and 7 subjects, respectively. Topological metrics were evaluated separately for each hemisphere. PrimateX demonstrated superior performance across species, as indicated by higher Euler numbers and fewer surface defects and holes than the alternative workflows. Topological quality was assessed using Euler-derived metrics, with higher Euler numbers and lower defect and hole counts indicating improved surface topology. Statistical comparisons used one-way ANOVA followed by Tukey's HSD post hoc tests. P values are denoted as * < 0.05, ** < 0.005, *** < 0.0005; NS, not significant.

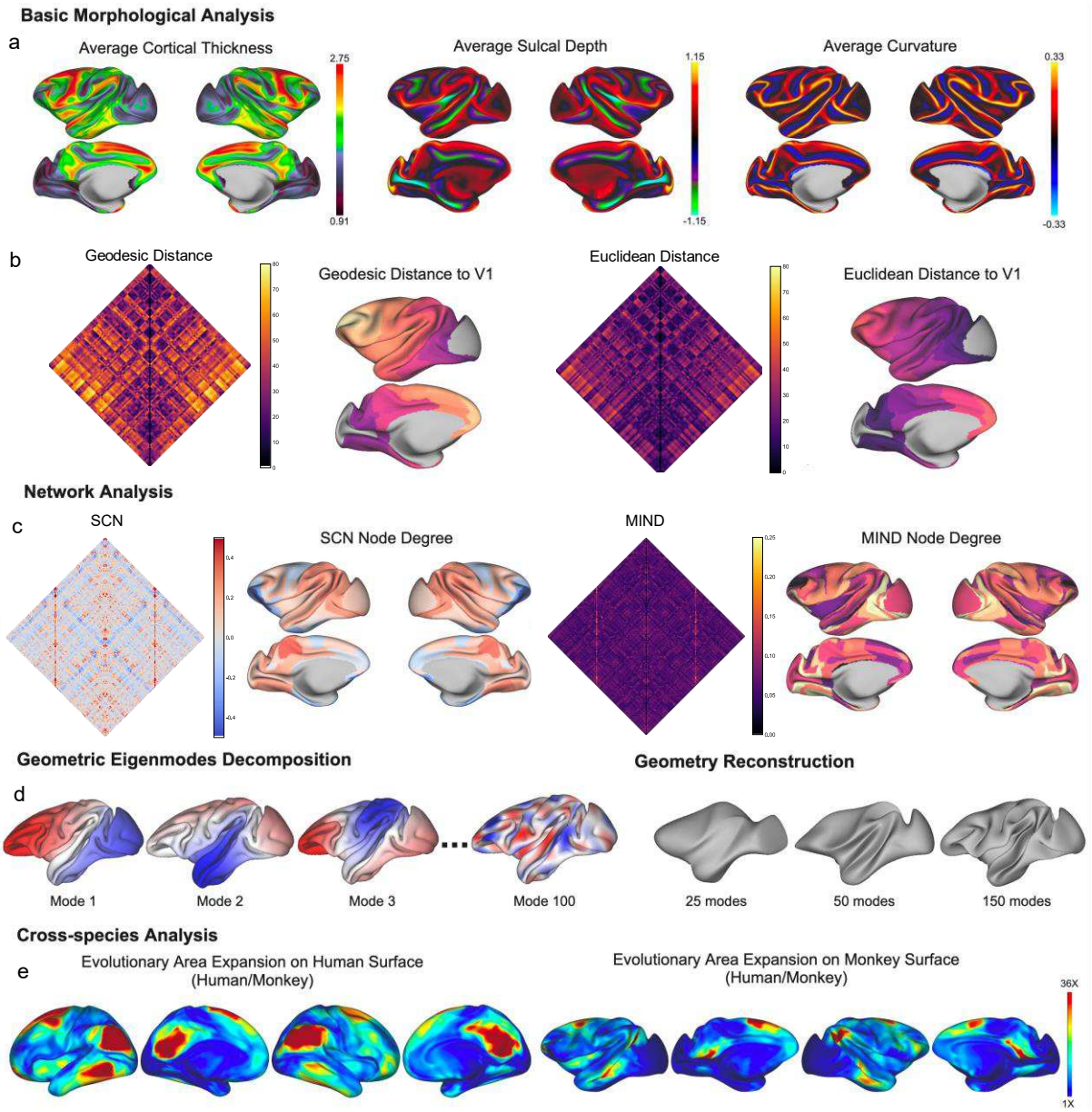
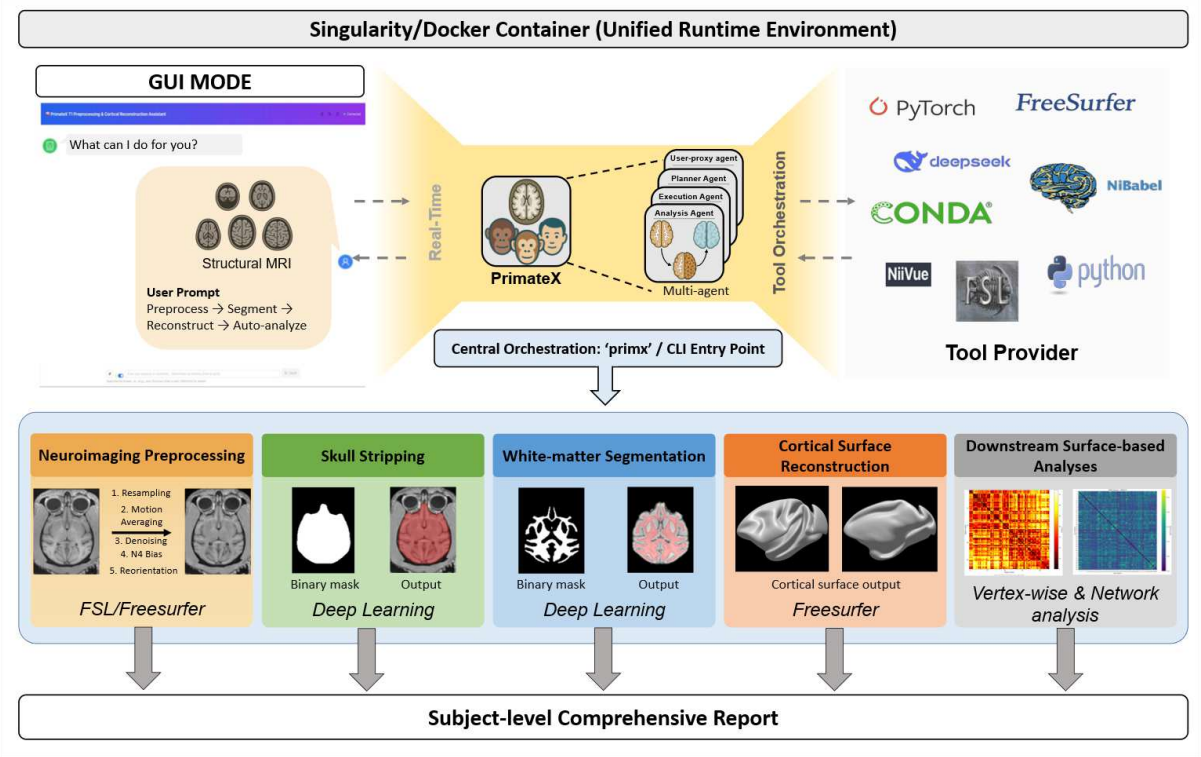


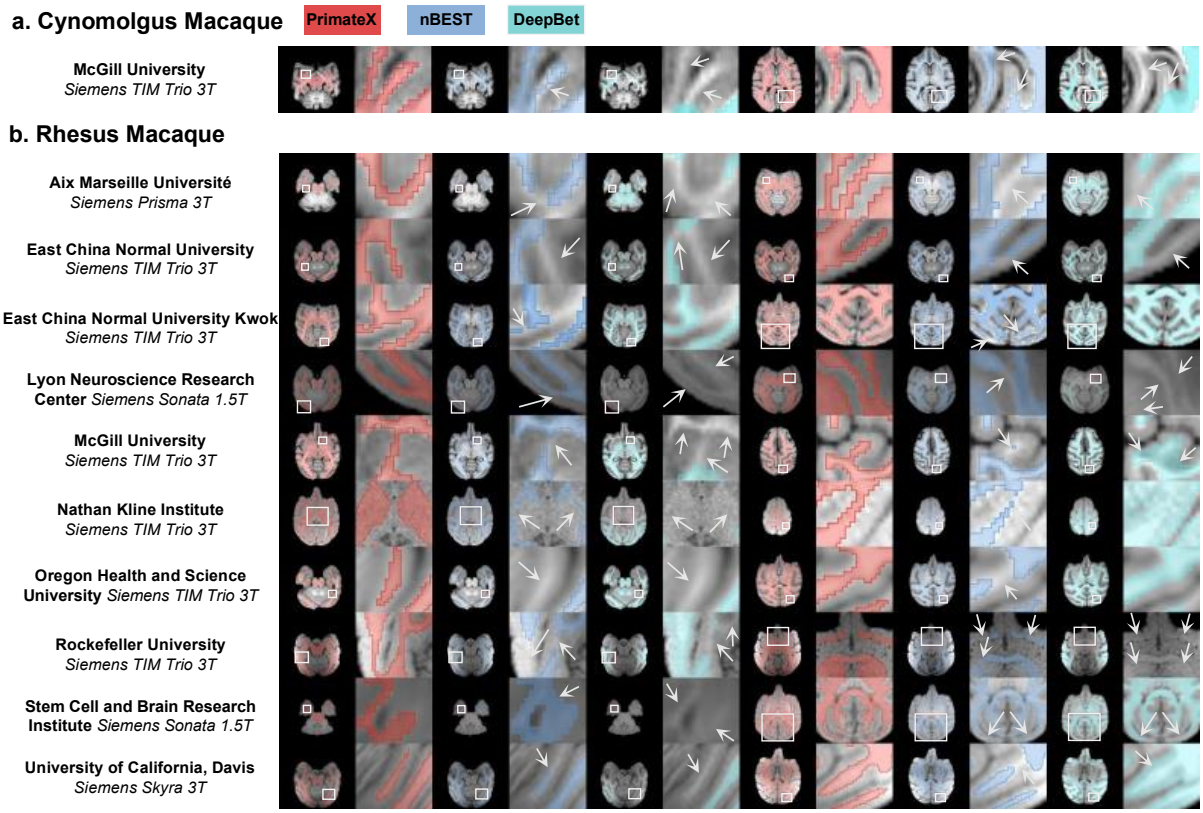
Figure. 7 | Multifaceted applications of PrimateX-enabled cortical surface-based analyses. **a,** Fundamental surface morphometry showing local cortical thickness, sulcal depth, and curvature patterns. **b,** Spatial mapping of geodesic and Euclidean distances from primary visual cortex (V1), visualized through distance matrices and cortical surface projections. **c,** Network-level analyses featuring structural covariance and morphometric inverse divergence networks, with corresponding nodal degree distributions mapped onto the surface. **d,** Spectral decomposition demonstrating the Laplace-Beltrami operator's eigenmodes (modes 1, 2, 3, and

100) and progressive surface reconstruction fidelity using 25, 50, and 150 modes. **e,** Evolutionary neuroanatomical comparison, with areal expansion patterns visualized through human-macaque surface registrations.

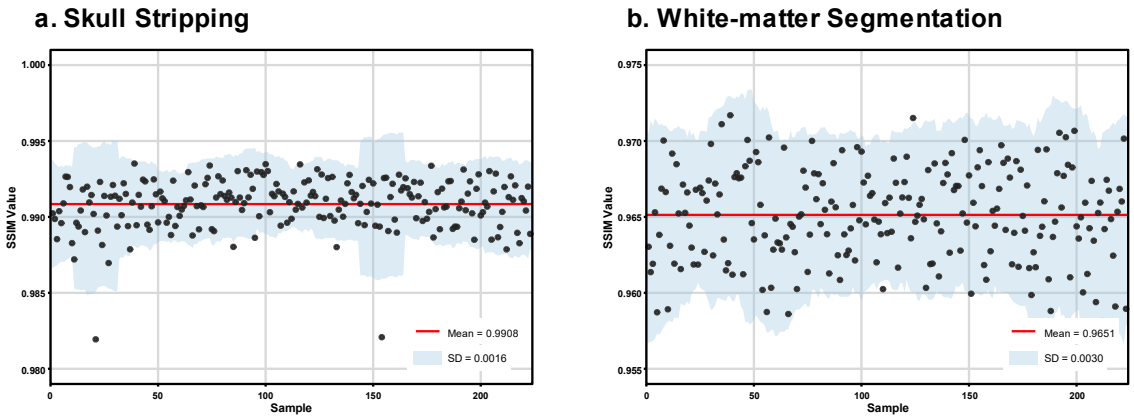
Extended Data



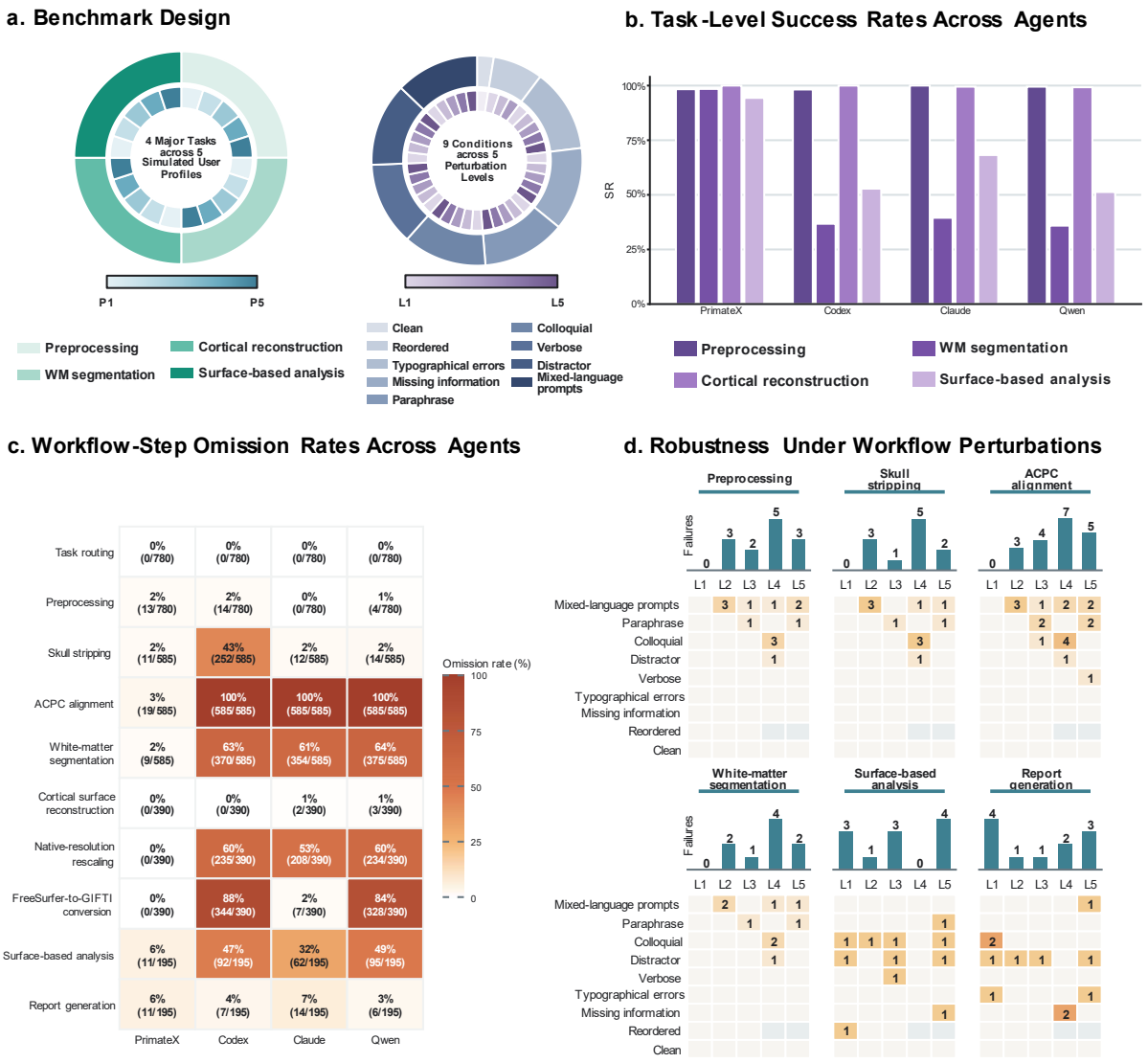
Extended Data Fig. 1 | Deployment architecture of PrimateX. The system comprises four core modules: a neuroimaging pipeline for preprocessing, registration, and surface reconstruction using established neuroimaging tools; a deep-learning segmentation module handling skull stripping and white-matter segmentation, including model inference and management; an integration layer for task coordination and automated execution; and a graphical user interface (GUI) for interactive visualization and user control. A central orchestration interface acts as the command-line entry point, coordinating all components. Processing occurs within a containerized runtime environment (Singularity or Docker), ensuring reproducibility via a unified software stack - including the operating system, neuroimaging tools, and deep-learning dependencies.



Extended Data Fig. 2 | Unseen-domain demonstration of white-matter segmentation in macaques Representative white-matter segmentation results from external sites acquired with different scanners and field strengths in **a,** cynomolgus and **b,** rhesus macaques. PrimateX is benchmarked against baseline methods, demonstrating robustness across acquisition protocols. These results further support the findings in Fig. 5.



Extended Data Fig. 3 | Concordance between PrimateX and reference FSL outputs in human neuroimaging data. Structural similarity index (SSIM) values between PrimateX-generated and reference FSL outputs for **a**, skull stripping and **b**, white-matter segmentation. Each black dot represents an individual sample, and red horizontal lines indicate the mean SSIM across all samples. Blue shaded regions represent the local mean ± 2 standard deviations (s.d.), calculated using a centred 21-sample moving window. The overall mean and s.d. are reported in each panel. SSIM values exceeded 0.95 across all samples, indicating close agreement between the two pipelines.



Extended Data Fig. 4 | Benchmark design and evaluation of the PrimateX agentic workflow. **a**, Benchmark design comprising four major neuroimaging tasks, five simulated user profiles (P1–P5) and 39 prompt configurations, including an unperturbed baseline and eight perturbation categories spanning up to five levels (L1–L5). Outer rings represent analysis tasks and perturbation categories, while inner segments represent user profiles and individual prompt configurations, respectively. **b**, Success rates (SR) of PrimateX and three general-purpose coding agents (Codex, Claude Code and Qwen Code) across preprocessing (n = 780), white-matter segmentation (n = 585), cortical reconstruction (n = 390) and surface-based analysis (n = 195). Sample sizes indicate the

number of applicable test cases per agent. **c**, Workflow-step omission rates across agents. Each cell shows the percentage of applicable cases in which a required step was omitted, with the corresponding omission count and total number of applicable cases in parentheses. **d**, Distribution of PrimateX workflow-step omissions across perturbation categories and levels. Bar charts show omission counts at each level, while heatmaps resolve these counts by perturbation category. Six workflow steps with observed omissions are shown, as the remaining four had no omissions across the evaluated cases. Unannotated cells indicate zero omissions for evaluated configurations, while grey cells denote combinations that were not tested.

Extended Data Table 1 | Data sources and acquisition.

Species	Sample size	Data source	Scanner / field strength	Imaging protocol (T1-weighted)	Anesthesia / preparation
Common marmoset	446	Brain/MINDS Marmoset MRI (NA216)	BioSpec 94/30 9.4 T	MPRAGE; 0.27×0.27×0.54 mm³; TR=6000 ms; TE=2 ms; TI=1600 ms; flip angle=12°	Isoflurane (2.0%) in oxygen/air mixture; supine positioning
Cynomolgus macaque	162	Institute of Neuroscience (n=29); Kunming Institute of Zoology, Chinese Academy of Sciences (n=133)	Siemens Tim Trio 3T; United Imaging UMR 790 3T	MPRAGE; 0.5×0.5×0.5 mm³; TR=2300 ms; TE=3 ms; TI=1000 ms; flip angle=9°	Ketamine (10 mg/kg) + atropine (0.05 mg/kg) induction; isoflurane (0.8-1.5%) maintenance (Institute of Neuroscience); pentobarbital sodium (10-16 mg/kg) maintenance (Kunming Institute); stereotaxic positioning
Rhesus macaque	583	UW-Madison Rhesus MRI dataset	GE Discovery MR750 3T; GE Signa EXCITE 3T	T1-weighted; ~0.273×0.5×0.273 mm³; TR=8.648-11.4 ms; TE=1.888-5.412 ms; TI=600 ms; flip angle=10°	Ketamine (15 mg/kg); medetomidine (0.03 mg/kg) or dexmedetomidine (0.015 mg/kg); reversed with atipamezole (0.15 mg/kg)
Chimpanzee	30	National Chimpanzee Brain Resource (NCBR)	Siemens Trio 3T	MPRAGE; 0.9×0.9×0.9 mm³; TR=2300 ms; TE=4.4 ms; TI=1100 ms; flip angle=8°	Ketamine (10 mg/kg) or telazol (3-5 mg/kg); propofol (40-60 mg/kg/h)
Human	1113	Human Connectome Project (HCP S1200)	Siemens Skyra 3T	MPRAGE; 0.7×0.7×0.7 mm³; TR=2400 ms; TE=2.14 ms; TI=1000 ms; flip angle=8°	Standard HCP protocol (healthy participants)

Extended Data Table 2 | Dataset composition and augmentation strategies. Training, augmented, and test set sizes for each species across all analytical tasks (species classification, task classification, skull stripping and white-matter segmentation). For classification tasks, data augmentation employed standard image transformations (including rotation and cropping). For segmentation tasks, augmentation used MRI-style transfer to preserve anatomical validity while expanding training diversity. For classification tasks, counts refer to input images, skull-intact and skull-stripped versions of the same scan were retained in the same data partition, and augmentation was applied only to the training set.

Task	Species	Training (real)	Training (augmented)	Test
Species classification	Common marmoset	596	4,768	292
	Cynomolgus macaque	200	1600	124
	Rhesus macaque	600	4,800	566
	Chimpanzee	28	224	12
	Human	600	4,800	400
Task classification	With skull	1,012	8,096	713
	Without skull	1,012	8,096	713
Skull stripping	Common marmoset	15	315	4
	Cynomolgus macaque	59	1,239	15
	Rhesus macaque	32	672	8
	Chimpanzee	16	336	4
	Human	890	0	223
WM segmentation	Common marmoset	15	315	4
	Cynomolgus macaque	44	924	11
	Rhesus macaque	32	672	8
	Chimpanzee	16	336	4
	Human	890	0	223

Extended Data Table 3 | External multi-site dataset composition.

Species	Sample size	Site	Scanner / field strength	Imaging protocol (T1-weighted)	Anaesthesia / preparation
Rhesus macaque	4	Aix-Marseille-Université	Siemens Prisma 3T	MPRAGE; 0.8 × 0.8 × 0.8 mm; TE 2.04 ms; TR 2900 ms; TI 1000 ms	Isoflurane; sphinx position; Kopf frame and ear bars
	10	East China Normal University - Chen	Siemens TIM Trio 3T	T1; 0.6 × 0.6 × 0.6 mm or 0.75 × 0.75 × 0.8 mm; TE 2.69-3.71 ms; TR 2200 ms; flip angle 7° or 9°	Atropine, ketamine and Zoletil; sphinx position
	4	East China Normal University - Kwok	Siemens TIM Trio 3T	T1; TE 77 ms; TR 3000 ms; flip angle 15°; voxel size not reported	Atropine and Zoletil; sphinx position
	4	Lyon Neuroscience Research Center	Siemens Sonata 1.5T / Siemens Prisma 3T	T1 parameters not reported	Not reported
	1	McGill University	Siemens TIM Trio 3T	MP2RAGE; 0.6 × 0.6 × 0.6 mm; TE 3.65 ms; TR 5000 ms; TI 700/2500 ms; flip angles 4°/5°	Isoflurane 0.6-1.2%; supine
	9	Mount Sinai School of Medicine - Philips	Philips Achieva 3T	T1; 0.5 × 0.5 × 0.5 mm; TE 6.93 ms; TR 1500 ms; TI 1100 ms; flip angle 8°	Ketamine/buprenorphine before intubation; isoflurane maintenance
	5	Mount Sinai School of Medicine - Siemens	Siemens Skyra 3T	MPRAGE; 0.5 × 0.5 × 0.5 mm; TE 2.11 ms; TR 4000 ms; TI 650/1820 ms; flip angle 8°; additional T1 also reported	Ketamine, meloxicam and buprenorphine; isoflurane 1-2%
	2	Nathan Kline Institute	Siemens TIM Trio 3T	T1; 0.5 × 0.5 × 0.5 mm; TE 3.87 ms; TR 2500 ms; TI 1200 ms; flip angle 8°	Atropine, dexdomitor and ketamine; 0.75% isoflurane
	3	National Chimpanzee Brain Resource (Comparative Sample of Anthropoid Primate 1.5T MRI Scans)	Philips NT 1.5T	T1w gradient-echo; pixel size 0.47-0.78 mm²; 0.6 mm slice interval; TE 8.5 ms; TR	Ketamine 10 mg/kg induction; propofol 10-20 mg/kg/hr IV infusion; prone (monkeys, knee coil);

				19.0 ms; FOV adjusted by brain size	supine (apes, head coil)
	2	Netherlands Institute for Neuroscience	Philips Ingenia 3T	3D-FFE; 0.6 × 0.6 × 0.6 mm; TE 6 ms; TR 13 ms; TI 900 ms; flip angle 8°	Medetomidine and ketamine; stereotactic device or sphinx position depending on condition
	12	Newcastle University	Bruker Vertical 4.7T	MDEFT; 0.5 × 0.5 × 2 mm; TE 6 ms; TR 750 ms; inversion delay 700 ms	Most animals scanned awake; two anaesthetized scans noted, details referred to publication
	2	Oregon Health and Science University	Siemens TIM Trio 3T	MPRAGE; 0.5 × 0.5 × 0.5 mm; TE 3.33 ms; TR 2600 ms; TI 900 ms; flip angle 8°	Ketamine 10 mg/kg; isoflurane 1-1.5%; dorsal recumbency
	2	Princeton University	Siemens Prisma VE11C 3T	T1w-MPRAGE; 0.5 × 0.5 × 0.5 mm; TE 2.32 ms; TR 2700 ms; TI 850 ms; flip angle 9°	Ketamine 10 mg/kg; isoflurane 2.5-3.0%; sphinx position
	5	Rockefeller University	Siemens TIM Trio 3T	MPRAGE; 0.5 × 0.5 × 0.5 mm; TE 2.95 ms; TR 2300 ms; TI 1100 ms; flip angle 8°	Ketamine 3 mg/kg and dexmedetomidine 0.02 mg/kg; isoflurane 0.5-0.6%; sphinx position
	16	Stem Cell and Brain Research Institute	Siemens Sonata 1.5T / Siemens Prisma 3T	T1 MPR 3D; 0.6 × 0.6 × 0.6 mm, 0.6 × 0.6 × 1 mm, or 0.5 × 0.5 × 0.5 mm; TE/TR vary by scanner	Zoletil + atropine, or ketamine + xylazine with 1% isoflurane
	19	University of California, Davis	Siemens Skyra 3T	T1; 0.3 × 0.3 × 0.3 mm; TE 3.65 ms; TR 2500 ms; TI 1100 ms; flip angle 7°	Ketamine, dexmedetomidine and buprenorphine; isoflurane 1-2%; sphinx position
	2	University of Minnesota	Siemens Syngo B17 7T	T1; 0.3 × 0.3 × 0.3 mm; TE 3.65 ms; TR 2500 ms; TI 1100 ms; flip angle 7°	Isoflurane 2.5%; sphinx position
Cynomolgus macaque	10	MADI@700, University of Bordeaux	Siemens Prisma 3T	3D-MPRAGE; 0.5 × 0.5 × 0.5 mm; TE 2.32 ms; TR 2000 ms; TI 1000 ms	Anatomical MRI: ketamine 10 mg/kg + xylazine 1 mg/kg; supine

	2	McGill University	Siemens TIM Trio 3T	MP2RAGE; 0.6 × 0.6 × 0.6 mm; TE 3.65 ms; TR 5000 ms; TI 700/2500 ms; flip angles 4°/5°	Isoflurane 0.6-1.2%; supine
	1	Rockefeller University	Siemens TIM Trio 3T	MPRAGE; 0.5 × 0.5 × 0.5 mm; TE 2.95 ms; TR 2300 ms; TI 1100 ms; flip angle 8°	Ketamine 3 mg/kg and dexmedetomidine 0.02 mg/kg; isoflurane 0.5-0.6%; sphinx position
	6	Stem Cell and Brain Research Institute	Siemens Sonata 1.5T / Siemens Prisma 3T	T1 MPR 3D; parameters vary by scanner as above	Zoletil + atropine, or ketamine + xylazine with 1% isoflurane
	3	University of Rochester Medical Center	Siemens Prisma 3T	T1w; 0.5 × 0.5 × 0.5 mm; TE 4.4 ms; TR 1880 ms; TI 1020 ms; flip angle 10°	Isoflurane 1-1.25%; sphinx position
Chimpanzee	6	National Chimpanzee Brain Resource (Comparative Sample of Anthropoid Primate 1.5T MRI Scans)	Philips NT 1.5T	T1w gradient-echo; pixel size 0.47-0.78 mm²; 0.6 mm slice interval; TE 8.5 ms; TR 19.0 ms; FOV adjusted by brain size	Ketamine 10 mg/kg induction; propofol 10-20 mg/kg/hr IV infusion; prone (monkeys, knee coil); supine (apes, head coil)
Capuchin monkey	4				
Sooty mangabey	4				
Squirrel monkey	4				
Baboon	2				
Bonobo	4				
Gibbon	4				
Gorilla	2				
Orangutan	3				

1

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PrimateXsupplementary.pdf
- PrimateXsupplementaryvideo.mp4
- NATCOMPUTSCI26407520261025PrimateXsupplementaryNCS.pdf