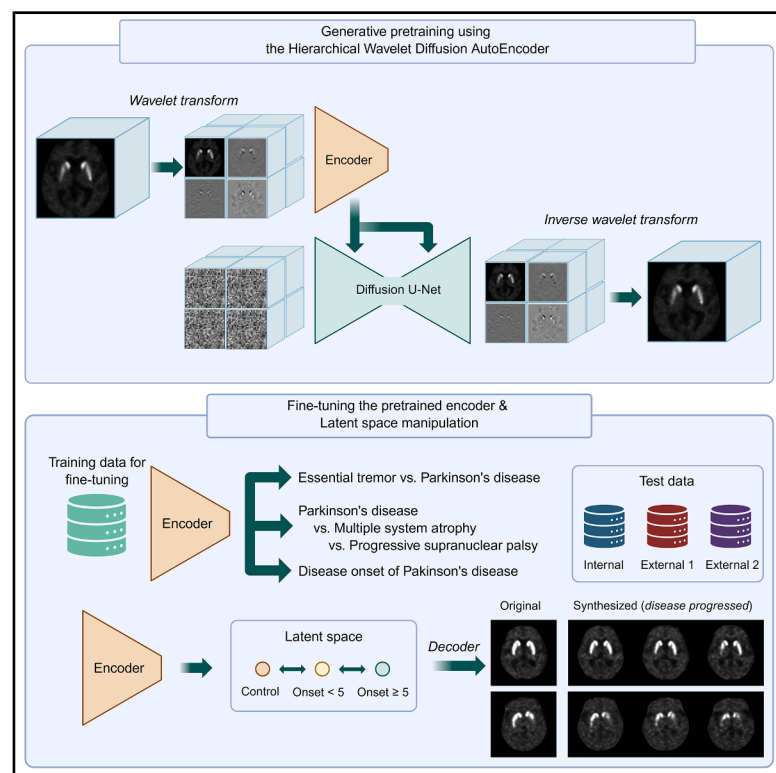


# Enhancing 3D dopamine transporter imaging as a biomarker for Parkinson's disease via self-supervised learning with diffusion models

## Graphical abstract



## Authors

Jongjun Won, Grace Yoojin Lee, Sungyang Jo, ..., Jieun Yum, Sun Ju Chung, Namkug Kim

## Correspondence

sjchung@amc.seoul.kr (S.J.C.), namkugkim@gmail.com (N.K.)

## In brief

Won et al. apply the hierarchical diffusion autoencoder to wavelet-transformed  $^{18}\text{F}$ -FP-CIT PET images to develop a self-supervised pretrained model for various clinical tasks. Their approach enhances the model's performance in differential diagnosis and disease progression estimation of Parkinson's disease, highlighting the potential of  $^{18}\text{F}$ -FP-CIT PET as a valuable disease biomarker.

## Highlights

- Evaluation of  $^{18}\text{F}$ -FP-CIT PET as a potential biomarker for Parkinson's disease
- Self-supervised learning using a diffusion model (HWD AE) on  $^{18}\text{F}$ -FP-CIT PET scans
- Assessment through differential diagnosis and disease progression estimation tasks
- External validation using multi-scanner, multi-center datasets



## Article

# Enhancing 3D dopamine transporter imaging as a biomarker for Parkinson's disease via self-supervised learning with diffusion models

Jongjun Won,<sup>1,8</sup> Grace Yoojin Lee,<sup>2,8</sup> Sungyang Jo,<sup>3</sup> Jihyun Lee,<sup>3</sup> Sangjin Lee,<sup>3</sup> Jae Seung Kim,<sup>4</sup> Changhwan Sung,<sup>4</sup> Jungsu S. Oh,<sup>4</sup> Kyum-Yil Kwon,<sup>5</sup> Soo Bin Park,<sup>6</sup> Joonsang Lee,<sup>1</sup> Jieun Yum,<sup>1</sup> Sun Ju Chung,<sup>3,\*</sup> and Namkug Kim<sup>1,7,9,\*</sup>

<sup>1</sup>Department of Convergence Medicine, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul 05505, Republic of Korea

<sup>2</sup>Department of Medical Science, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul 05505, Republic of Korea

<sup>3</sup>Department of Neurology, Asan Medical Center, University of Ulsan College of Medicine, Seoul 05505, Republic of Korea

<sup>4</sup>Department of Nuclear Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul 05505, Republic of Korea

<sup>5</sup>Department of Neurology, Soonchunhyang University Seoul Hospital, Soonchunhyang University School of Medicine, Seoul 04401, Republic of Korea

<sup>6</sup>Department of Nuclear Medicine, Soonchunhyang University Seoul Hospital, Soonchunhyang University School of Medicine, Seoul 04401, Republic of Korea

<sup>7</sup>Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul 05505, Republic of Korea

<sup>8</sup>These authors contributed equally

<sup>9</sup>Lead contact

\*Correspondence: [sjchung@amc.seoul.kr](mailto:sjchung@amc.seoul.kr) (S.J.C.), [namkugkim@gmail.com](mailto:namkugkim@gmail.com) (N.K.)

<https://doi.org/10.1016/j.xcrm.2025.102207>

## SUMMARY

Accurate diagnosis and precise estimation of disease progression states are crucial for developing effective treatment plans for patients with parkinsonism. Although various deep learning-based computer-aided diagnostic models have demonstrated benefits, they have been relatively underexplored in parkinsonism owing to limited data and lack of external validation. We introduce the hierarchical wavelet diffusion autoencoder (HWDAE), a generative self-supervised model trained with 1,934 dopamine transporter positron emission tomography (DAT PET) images. HWDAE learns relevant disease traits during generative training, prior to supervision with human labels, as evidenced by its ability to synthesize realistic images representing different disease states of Parkinson's disease. The pretrained HWDAE is subsequently adapted for two differential diagnostic tasks and one disease progression estimation task, tested on images from two medical centers. Our training approach introduces a paradigm for deep learning research utilizing PET and expands the potential of DAT PET as a biomarker for Parkinson's disease.

## INTRODUCTION

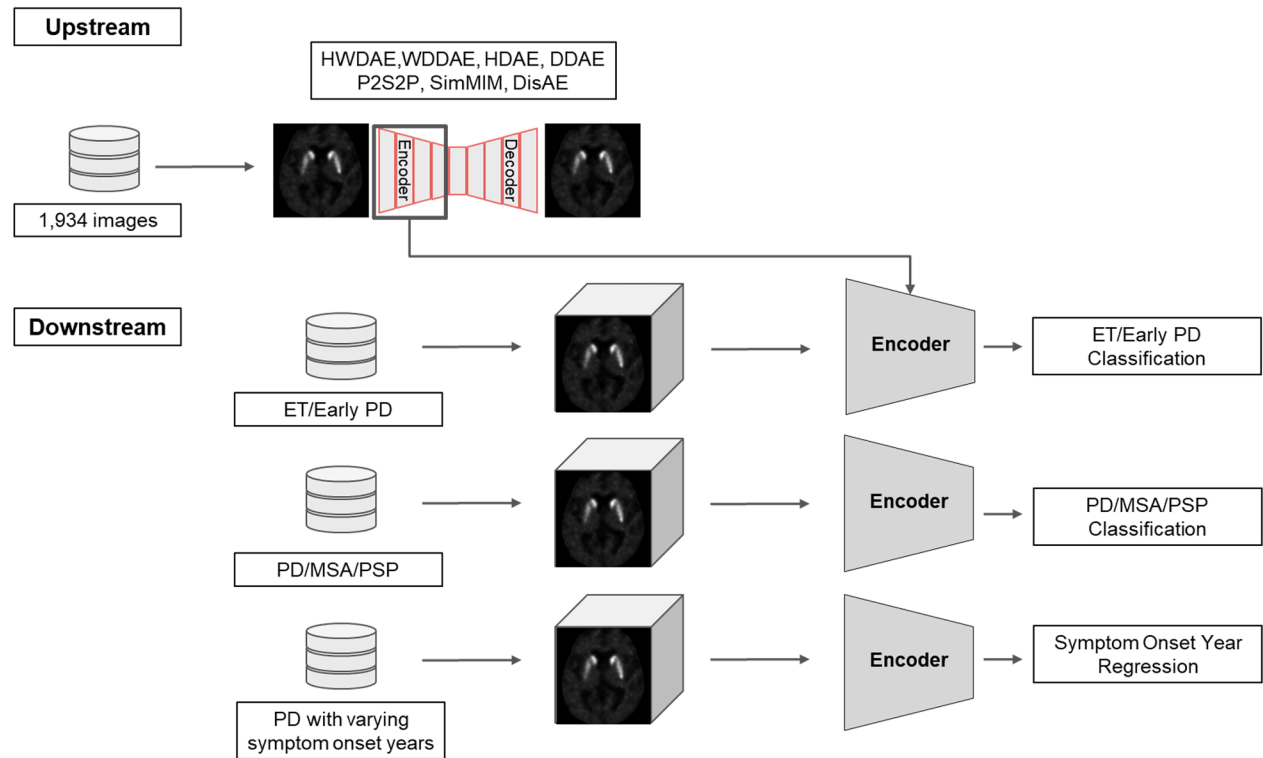
Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder after Alzheimer's disease. Although its incidence rate is increasing globally with the rise in life expectancy,<sup>1</sup> a validated diagnostic or disease progression biomarker for PD is lacking. This has caused clinicians to primarily rely on clinical findings, including patient history, neurological examination, medication response, and, when available, structural and functional imaging, such as presynaptic dopaminergic imaging using single-photon emission computed tomography (SPECT) or positron emission tomography (PET). Dopamine transporter (DAT) SPECT has been used to differentiate patients with parkinsonism for over a decade following its approval by the US Food and Drug Administration<sup>2</sup> and the European Medicines Agency.<sup>3</sup> DAT SPECT has demonstrated greater than 90% sensitivity in differ-

entiating between PD, multiple system atrophy (MSA), and progressive supranuclear palsy (PSP) from non-presynaptic dopaminergic parkinsonism and has led to diagnosis and treatment changes in 31% and 54% of patients, respectively.<sup>4,5</sup> In addition, PET imaging using DAT radiotracers discriminated patients with PD from controls or those with essential tremor (ET) with performance comparable to that of DAT SPECT.<sup>6–8</sup> Additionally, several studies have longitudinally observed the changes in the brain structures or radioligand uptake patterns in patients with PD.<sup>9</sup> Various imaging modalities, such as T1-weighted structural magnetic resonance imaging (MRI), iron-sensitive MRI, and dopaminergic PET/SPECT, have been identified as potential disease progression biomarkers, although their roles appear to vary by disease stage.<sup>9</sup>

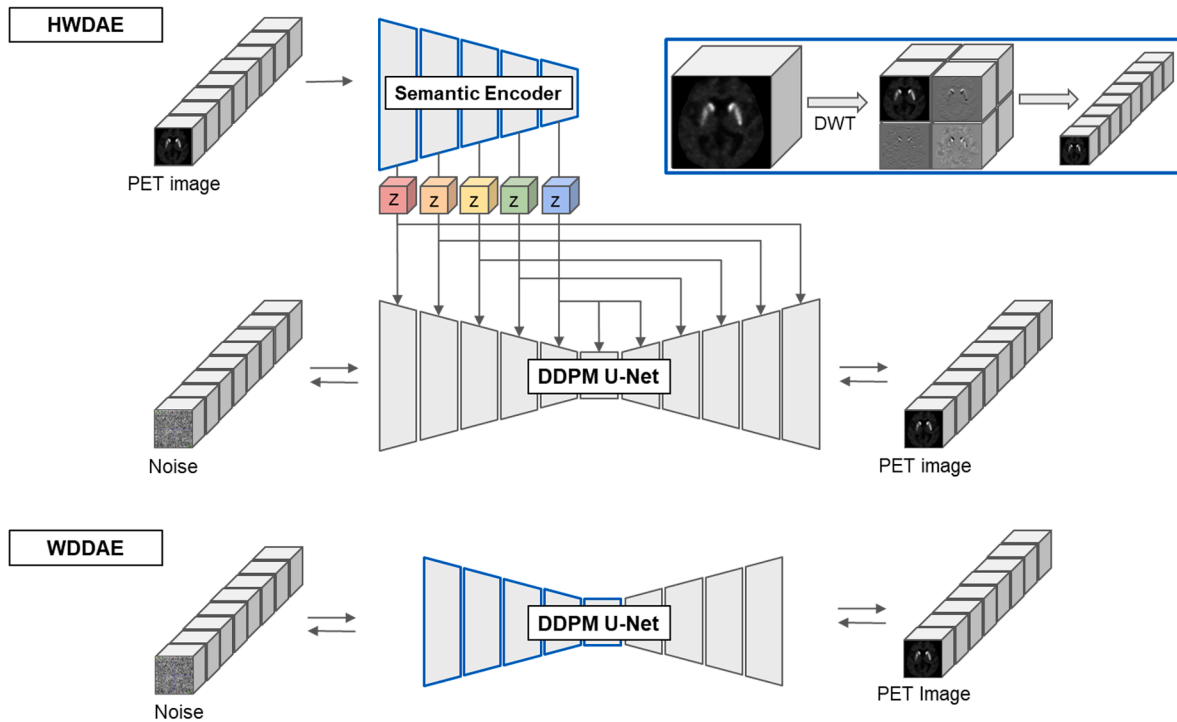
Previously, most research on DAT SPECT or PET images involved calculating the DAT availability for the regions of interest



A



B



(legend on next page)

(ROIs), such as the caudate, putamen, and ventral striatum. This required precise spatial registration of individual PET images to structural brain region atlases defined on computed tomography, MRI, or PET templates; alternatively, physicians manually drew the boundaries of these ROIs directly on the PET images.<sup>10–16</sup> In addition, calculating the DAT availability reduced the information in the PET images to a few scalar values, potentially resulting in the loss of important details, such as the gradual uptake gradient from the anterior to posterior putamen. However, with the advent of deep learning (DL) techniques, more in-depth and varied levels of the image features in big datasets can be extracted, allowing for the exploration of information that may have been previously overlooked. For instance, DL has been applied to presynaptic dopamine PET imaging for binary classification of PD and healthy controls, achieving an area under the curve (AUC) of 0.96 and a sensitivity of 91.5%.<sup>17</sup> Surprisingly, without additional clinical information, machine learning (ML) and DL models were able to discriminate between patients with PD and atypical parkinsonism with an AUC of at least 0.90.<sup>17,18</sup>

Although DAT PET images have high resolutions and are effective in discriminating patients with parkinsonism, their high costs, the relatively small number of medical centers with available PET scanners, and the scarcity of nuclear medicine specialists with expertise in interpreting DAT PET images hinder their practical use. Accordingly, building DL models, which usually require extensive training data, is particularly challenging. As a promising solution to this dilemma, self-supervised learning (SSL) has been introduced for pretext upstream tasks. In SSL, DL models learn semantic image features through the data rather than from handcrafted ground-truth labels. Various SSL strategies have demonstrated performance improvements across diverse medical imaging modalities, such as chest radiography, fundus images, MRI, and pathology slides.<sup>19–22</sup> Among various SSL strategies, generative models (GMs) learn latent representations and data distributions simultaneously, enabling latent space manipulation and new image generation. In particular, among the numerous GMs, diffusion models have dominated the image generation field, reaching state-of-the-art performance in image synthesis,<sup>23–25</sup> text-to-image synthesis,<sup>26–28</sup> and image editing.<sup>29</sup> However, it remains unclear whether the ability to generate realistic images also translates to capturing image semantics relevant for discrimination tasks such as differential diagnosis or disease progression prediction. Building on prior studies that proposed diffusion models as foundation models for classification tasks, from natural images<sup>30</sup> to chest radiographs,<sup>31</sup> or for constructing semantically meaningful latent spaces using brain MRI,<sup>32</sup> we adopted diffusion models to assess their ability to represent brain PET images. Unlike masked autoencoders and other SSL methods, diffusion models also support latent vector manipulation, offering capabilities for PET image editing and enhanced model explainability.

In this study, we aimed to build upstream SSL models using the DAT PET scans of patients with parkinsonism and evaluate the models using three downstream tasks: (1) binary classification of ET vs. early PD (motor symptom duration < 1 year) (EP task), (2) multi-class classification of PD vs. MSA vs. PSP (PMP task), and (3) regression of the onset year of motor symptoms of PD (SOY task) (Figure 1A). For qualitative evaluation of the upstream models, we visualized their latent spaces and manipulated them to synthesize new PET images. For the upstream models, we suggested diffusion models with two modifications—a hierarchical semantic encoder<sup>33</sup> and wavelet transformation of the input images—and compared them with other types of GMs. Notably, we expanded some of the GM architectures from two dimensions (2D) to three dimensions (3D), which are yet to be sufficiently explored in this field. This expansion was essential, as the DAT uptake patterns of target brain structures, such as the striatum, are inherently in 3D, and subregional reduced uptake patterns are not easily captured in individual cross-sectional slices. Previous studies presenting high AUC values when classifying patients with parkinsonism often lacked the use of external validation datasets.<sup>18,34,35</sup> We evaluated our models on the data from three different PET scanners at two medical centers to verify the model's generalizability. Through this study, we demonstrated that SSL using diffusion models can learn semantic image representations in the absence of labels, which signifies its potential as a general-purpose DL model in PD-related tasks. Lastly, we reconfirmed the significance of DAT PET as a potential disease biomarker of PD.

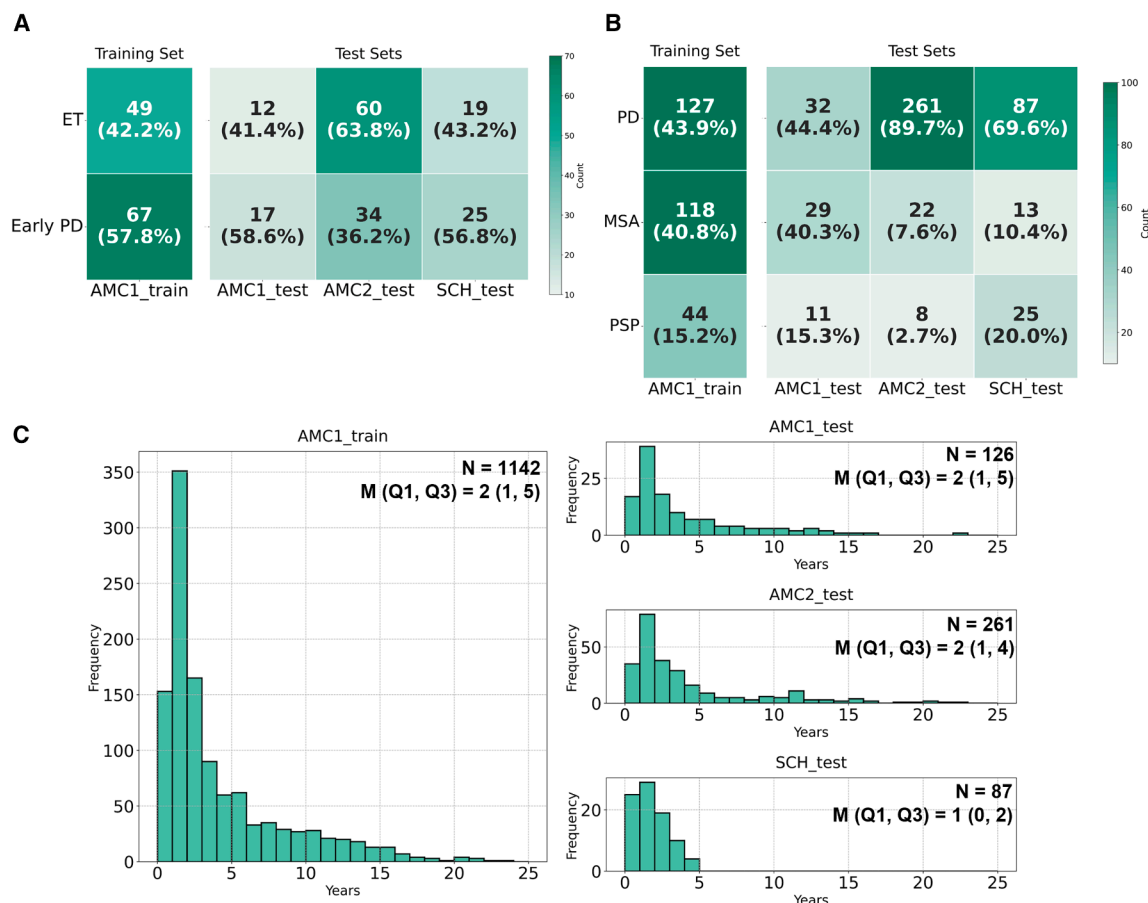
## RESULTS

For the upstream task, we propose a modified version of the diffusion models, called hierarchical wavelet diffusion autoencoder (HWDAAE), which adopts a semantic encoder outside the U-Net architecture as in the hierarchical diffusion autoencoder (HDAE)<sup>33</sup> and discrete wavelet transform (DWT)<sup>36</sup> preprocessing of input images. HWDAAE was compared with the denoising diffusion autoencoder (DDAE),<sup>30</sup> which uses the bottleneck layer of U-Net as the latent space instead of a separate semantic encoder. The same wavelet transform preprocessing was applied to the DDAE, which we denoted as the wavelet DDAE (WDDAE) (Figure 1B). In addition to the diffusion models (HWDAAE, HDAE, DDAE, and WDDAE), we also evaluated the following upstream models: simple masked image modeling (SimMIM),<sup>37</sup> disruptive autoencoder (DisAE),<sup>38</sup> and pixel2style2pixel (P2S2P).<sup>39</sup> We set the anchor model as 3D ResNet34. For the downstream tasks, we used three different methods to train each upstream model: linear probing (LP), scratch (SC), and fine-tuning (FT). The models were trained for the downstream tasks using 5-fold cross-validation (5CV). Thus, the mean performance of the five models evaluated on the same hold-out test set was reported. Paired t tests were conducted to compare different upstream models, using the five performance values obtained from each model architecture.

### Figure 1. Overall workflow and model architectures

(A) Workflow from upstream to downstream tasks.

(B) Simplified architectures of the HWDAAE and WDDAE, with input images preprocessed using discrete wavelet transform (DWT). The encoders of the HWDAAE and WDDAE used in the downstream tasks are highlighted as blue lines. The HDAE and DDAE have similar architectures to those of the HWDAAE and WDDAE, respectively, but lack DWT preprocessing and include an additional layer in the encoder and decoder of the U-Net architecture.



**Figure 2. Data composition of downstream tasks**

(A) Data composition of the EP task depicted using heatmap.

(B) Data composition of the PMP task.

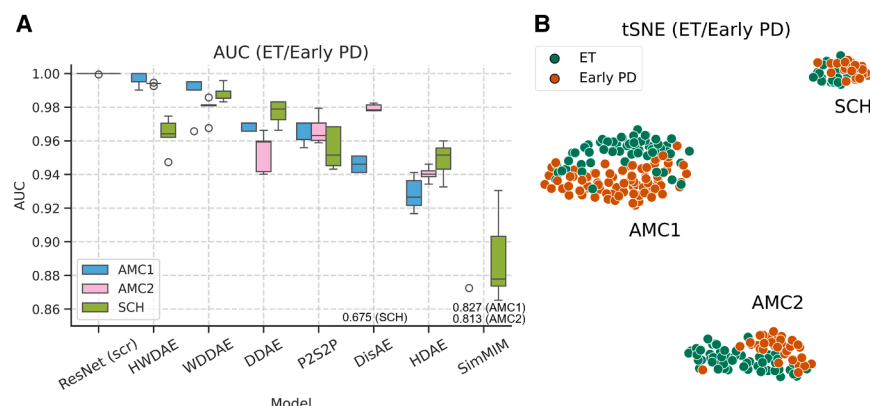
(C) Data composition of the SOY task. The total number, median, quartile 1, and quartile 3 values are provided for each training/test set.

See also [Figure S2](#) and [Table S1](#).

## Dataset

Of the 3,102 3D N-(3-fluoropropyl)-2 $\beta$ -carboxymethoxy-3 $\beta$ -(4-iodophenyl) nortropane ( $^{18}$ F-FP-CIT) PET images from 2,939 patients at the Asan Medical Center (AMC), 2,672 images from 2,600 patients were selected after applying the exclusion criteria and removing the images that failed the visual quality check during preprocessing ([Figure S1](#)). All of these images were acquired using a PET scanner, either Bio40 or Vision. We used the images from Bio40 (AMC1) for upstream model training, downstream model training, and downstream internal testing, whereas the images from Vision (AMC2) were used solely for downstream external testing. An additional dataset from Soonchunhyang University Seoul Hospital (SCH) using a Bio128 scanner was included for further external testing. The test sets for the downstream tasks comprised unseen images and patients from both the upstream and downstream training stages. A total of 1,934 images from AMC1 were used for upstream model training. The dataset configuration and sizes are described in [Figures 2](#) and [S2](#). For the EP and PMP tasks, the AMC1 training and test sets were adjusted to balance the number of images from pa-

tients with PD with those from other classes, while the AMC2 and SCH test sets remained unchanged. Reflecting real-world clinical distributions, images from patients with PSP were less frequent than those from patients with PD and MSA in the AMC2 test set. In comparison, the SCH test set, being registry-based, had a different patient composition from the other test sets. Overall, while patients with ET and PD had similar mean ages, patients with MSA were generally younger and patients with PSP older than patients with PD at the time of image acquisition across all training and test sets. For the SOY task, as most patients underwent PET scans during their initial evaluation, the data were right skewed toward longer onset years. The median time interval between motor symptom onset and image acquisition was 2 years for both the AMC1 training and test sets, as well as the AMC2 test set. Additional details on patient sex and age at image acquisition for both upstream and downstream models are provided in [Table S1](#). Lastly, for the SOY task, we created an additional test set of 30 patients with PD who underwent two PET scans using the same Bio40 scanner at different time intervals.



**Figure 3. Boxplot and t-SNE of the EP task**

(A) AUC of each model across test sets. For each model and test set, five AUC values from 5-fold cross-validation are depicted in a boxplot. Models are listed in the decreasing order of the AUC on the test set AMC1.

(B) t-SNE drawn using the latent space of the HWDAAE. Each circle corresponds to each PET image. Images from both training and test sets are included.

See also [Figures S3 and S7](#) and [Tables S2, S12, and S14](#).

### Classification of ET and early PD (EP task)

The model performance of the EP task is presented in [Figures 3A](#) and [S3](#) and [Table S2](#). Images labeled as early PD were defined as those from patients with a motor symptom duration of less than 1 year. For the EP task, we tested only with LP, as the HWDAAE achieved a mean AUC (mAUC) close to one in the AMC1 test set, making SC or FT unnecessary. On the internal test set AMC1 and external test set AMC2, HWDAAE achieved the highest mAUCs of 0.997 (95% confidence interval [CI]: 0.994, 1) and 0.994 (0.993, 0.994), respectively, among all upstream models. However, on the external test set SCH, WDDAE showed the highest mAUC of 0.988 (0.984, 0.992). In comparison, the supervised model ResNet34 reached a perfect mAUC of 1 on all test sets, probably owing to the relative simplicity of the EP task. On AMC1, HWDAAE and WDDAE had comparable mAUCs to that of ResNet34 (paired  $t$  test  $p = 0.21$ , 0.08), but ResNet34 significantly outperformed all upstream models on the external test sets ( $p < 0.003$  for all). The t-distributed stochastic neighbor embedding (t-SNE) of the latent space of the HWDAAE semantic encoder, which was frozen for this task, is shown in [Figure 3B](#). Although the semantic encoder was trained with images from AMC1 only, it effectively distinguishes images from AMC2 and SCH. In addition, the images are automatically clustered based on the ET and early PD labels. The occlusion sensitivity map<sup>40</sup> of the HWDAAE with LP revealed that masking regions corresponding to the putamen resulted in a higher probability of predicting PD and a lower probability of predicting ET ([Figure S7](#)).

### Classification of PD, MSA, and PSP (PMP task)

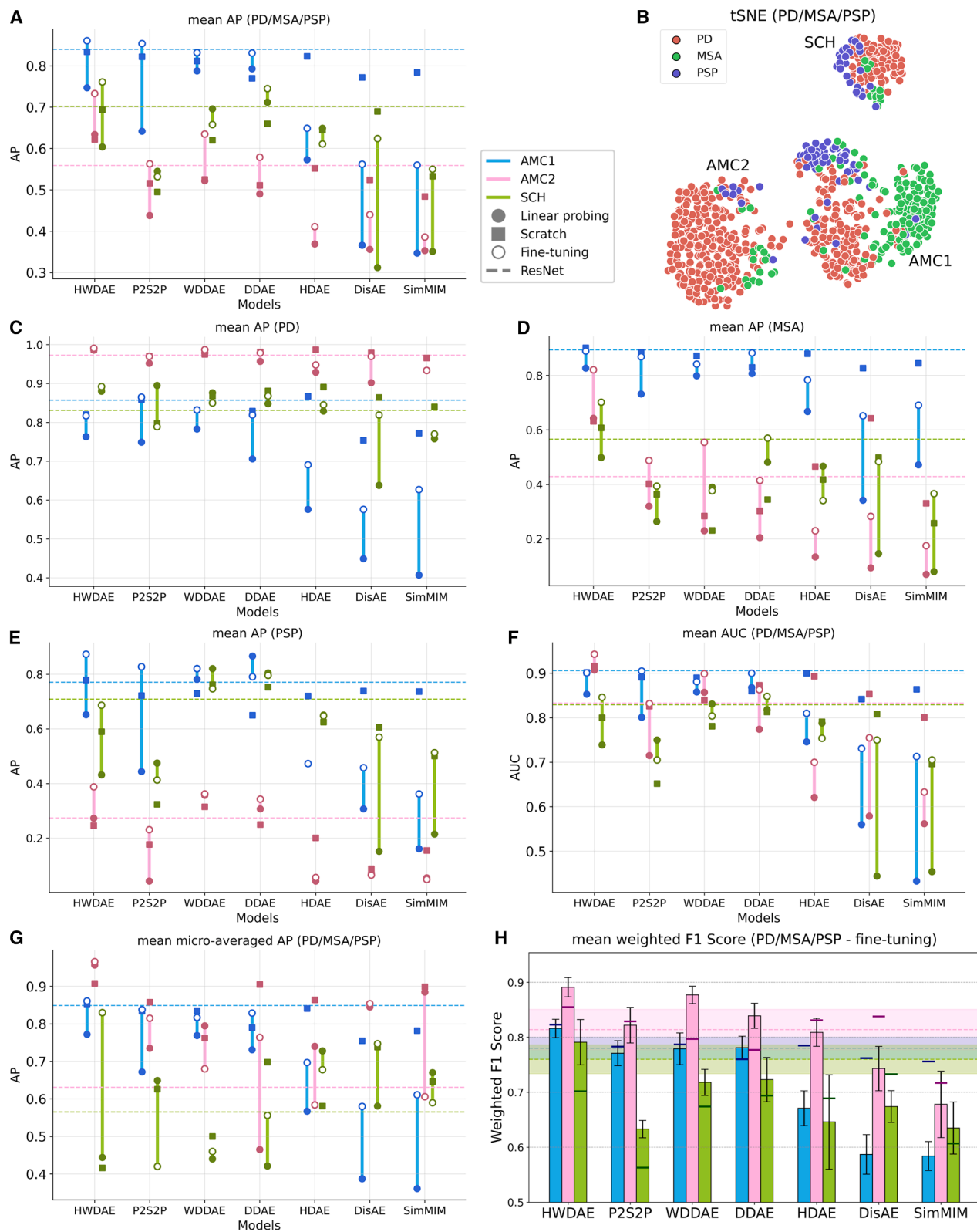
The model performance of the PMP task is presented in [Figures 4](#) and [S4](#) and [Tables S3–S8](#). For the PMP task, we tested all three methods (LP, SC, and FT) to determine whether FT outperformed LP and SC. We selected the macro-averaged area under the precision-recall curve (AP) as the primary evaluation metric, owing to the class imbalance of the test sets, and additionally calculated the micro-averaged AP, macro/micro-averaged AUC, and weighted F1 score. For the HWDAAE, we used the third block of its semantic encoder out of five for LP, SC, and FT. The use of all five blocks led to overfitting on the internal test set AMC1; hence, we chose the third block that best distinguished the PD, MSA, and PSP classes based on visual

inspection of the t-SNE plots of each block following FT ([Figure S5](#)).

The HWDAAE with FT achieved the highest mean macro-averaged AP across all three test sets: 0.861 (95% CI: 0.854, 0.868) for AMC1, 0.733 (0.705, 0.762) for AMC2, and 0.761 (0.742, 0.78) for SCH, compared with all upstream models and ResNet34 (HWDAAE vs. ResNet34 paired  $t$  test  $p = 0.06$ , 0.008, and 0.17, respectively). Although P2S2P with FT reached the second-highest mean macro-averaged AP on the internal test set AMC1, the diffusion models WDDAE and DDAE ranked second or third on the two external test sets AMC2 and SCH. While some upstream models reached mean macro-averaged AP values comparable to those of the HWDAAE with FT, most models exhibited performance variability across different test sets, unlike the HWDAAE with FT.

In general, for most upstream models, FT outperformed SC, and SC outperformed LP on the test sets. However, some models, such as P2S2P, WDDAE, DDAE, and HDAE, showed an interesting trend where LP achieved a higher mean macro-averaged AP than FT or SC, particularly on the SCH test set. Additionally, in models like HDAE, DisAE, and SimMIM, FT did not improve the mean macro-averaged AP compared with SC, whereas in other models, it consistently enhanced performance across all test sets. Consistent with its superior performance in terms of mean macro-averaged AP, the HWDAAE with FT generally achieved the highest mean macro/micro-averaged AUC, mean micro-averaged AP, mean weighted F1 score, and mean balanced accuracy across the three test sets.

The t-SNE of the latent space from the third block of HWDAAE after FT is shown in [Figure 4B](#). The three datasets form distinct clusters, with images from patients with PD, MSA, and PSP grouped within each cluster. The occlusion sensitivity maps and Guided GradCAMS<sup>40</sup> of the HWDAAE with FT for PD, MSA-parkinsonian (MSA-P) subtype, MSA-cerebellar (MSA-C) subtype, and PSP are shown in [Figures S8–S10](#). Although the PMP task was not designed to differentiate between MSA subtypes, we generated separate maps for the MSA-P and MSA-C subtypes to observe potential differences between them. In the occlusion sensitivity maps, masking regions corresponding to the putamen, ventral striatum, and caudate increased the probability of predicting PD, MSA-P (or MSA-C), and PSP, respectively. However, the Guided GradCAM results



(legend on next page)

were less informative; although activation was focused on the striatum, no consistent subregional patterns emerged across classes.

### Regression of motor symptom onset prediction (SOY task)

The model performance of the SOY task is presented in Figure 5 and Table S9. Here, the downstream task was to predict the symptom onset year, which was defined as the time interval in years between the initiation of motor symptoms and PET image acquisition. We used the mean coefficient of determination ( $R^2$ )<sup>41</sup> and concordance correlation coefficient (CCC)<sup>42</sup> as the primary and secondary evaluation metrics, respectively. Additionally, we calculated the Pearson correlation coefficient (R) using scatter-plots of the predicted vs. true symptom onset years. The predicted onset years were averaged over the results from five folds. The SCH dataset was excluded for this task, as all its symptom onset years fell within the 0–4 range, lacking images from advanced disease stages.

On the internal test set AMC1, the WDDAE with FT had the highest mean  $R^2$  of 0.519 (95% CI: 0.5, 0.537) across all upstream models and ResNet34 (WDDAE vs. ResNet34 paired t test  $p = 0.129$ ). On the external test set AMC2, the HWDAAE with FT significantly outperformed all models with a mean  $R^2$  of 0.588 (0.579, 0.596) (HWDAAE vs. ResNet34  $p = 0.042$ ). A similar trend was observed for the mean CCC, with 0.736 (0.727, 0.745) for the WDDAE with FT on AMC1 (WDDAE vs. ResNet34  $p = 0.092$ ) and 0.733 (0.727, 0.738) for the HWDAAE with FT on AMC2 (HWDAAE vs. ResNet34  $p = 0.07$ ). Upon visual inspection of the AMC1 test set images with residual years (predicted onset year – true onset year) outside one standard deviation, we observed that the extent of putaminal uptake degradation deviated from that of the other images with similar onset years (Figure S6).

As in the PMP task, the mean  $R^2$  and CCC values increased for most upstream models as the training method was changed from LP to SC to FT. Notably, regarding LP, the HWDAAE showed exceptional performance in both AMC1 and AMC2, whereas the other upstream models had mean  $R^2$  values near zero or even negative. All upstream models and ResNet34 had higher mean  $R^2$  and CCC values than the multivariate linear regression model using specific/nonspecific binding ratios (SNBRs) of the striatal subregions (ventral striatum, anterior/posterior caudate, and anterior/posterior putamen), which resulted in mean  $R^2$  values of 0.34 (0.336, 0.344) for AMC1 and 0.311 (0.308, 0.314) for AMC2 and mean CCC values of 0.490 (0.482, 0.5) for AMC1 and 0.458 (0.451, 0.464) for AMC2.

Scatter and Bland-Altman plots were generated for the models with the highest  $R^2$  values for AMC1 (Figures 5C and 5E) and AMC2 (Figures 5D and 5F). For both AMC1 and AMC2, most of the estimation errors fell within the 95% limits of agreement for the images from patients with onset years of <5 years, whereas several images from those with onset years of  $\geq 5$  years fell outside the 95% limit. For patients with onset years of <5 years, the model showed a downward trend of the Bland-Altman plots, predicting onset years closer to the middle values. Moreover, the model tended to underestimate the onset year for those with onset years of >10 years.

Figures 6A and 6B display the t-SNE of the latent space from the WDDAE after FT. Each circle, representing an image, showed a general trend of shifting toward the right as the onset year increased, indicating that the WDDAE successfully embedded the progression of PD from the early to advanced stages. A similar trend was found in Figure 6D, where the baseline and follow-up images of 30 patients with PD with repeated scans are marked with arrows per patient.

The Guided GradCAMs of the HWDAAE with FT for patients with onset years of <5 and >10 years revealed that the model primarily focused on the putamen in cases with onset years of <5 years while displaying more varied activation across striatal subregions and the raphe nuclei in cases with onset years of >10 years (Figure S11).

Additionally, we analyzed the relationship between the acquisition time interval between the baseline and follow-up images and their model-predicted time interval. Because the model struggled to accurately predict the onset years for images from patients with advanced PD (Figure 6E), the discrepancy between the true and predicted time intervals increases with an increase in the true time interval (Figure 6C).

### Qualitative evaluation of upstream model latent spaces

Figure 7 shows the t-SNE plots for downstream tasks of the upstream HWDAAE model before any FT. In other words, the latent space of the HWDAAE was used immediately after the upstream SSL process. The HWDAAE was selected over other upstream models because it most frequently demonstrated the best performance across the three downstream tasks and test sets.

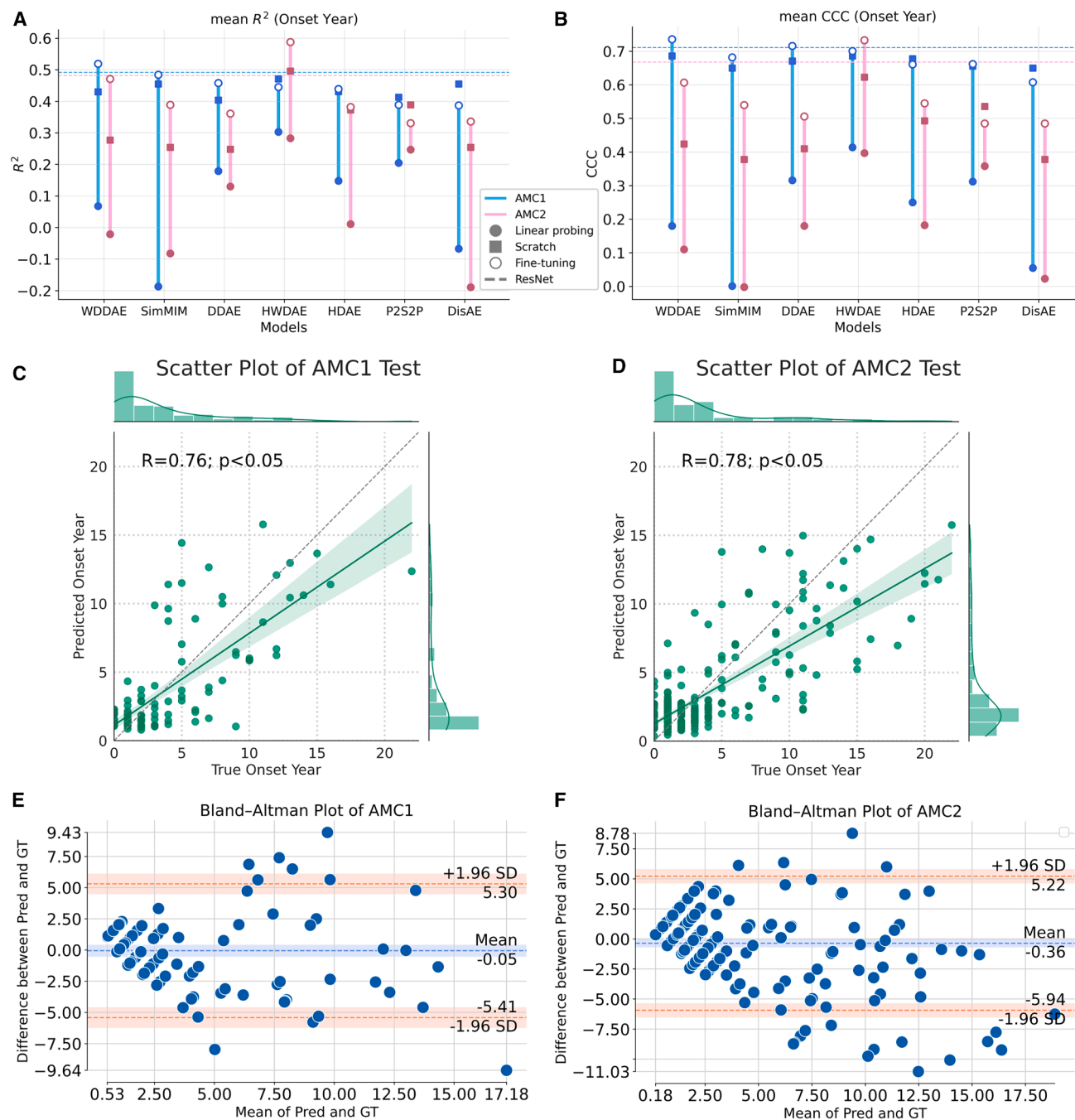
Latent vectors from images with normal PET findings were clustered together and clearly distinguished from those of patients with PD (Figures 7A and 7D), whereas no distinct clustering patterns were observed among those of patients with PD, MSA, and PSP (Figure 7B), most likely owing to the insufficient number of images from the patients with MSA and PSP in

### Figure 4. Bar plots and t-SNE of the PMP task

Models are listed in the decreasing order of AP with fine-tuning on the test set AMC1.

- (A) Mean macro-averaged AP.
- (B) t-SNE drawn using the latent space of the third block of the HWDAAE after fine-tuning.
- (C) Mean AP of PD versus others.
- (D) Mean AP of MSA versus others.
- (E) Mean AP of PSP versus others.
- (F) Mean AUC.
- (G) Mean micro-averaged AP.
- (H) Mean weighted F1 score of upstream models with fine-tuning. Thick lines in each bar represent the results of the model trained from scratch. Error bars and transparent ranges around dashed lines denote the 95% confidence intervals.

See also Figures S4, S5, S8–S10, and S15 and Tables S3–S8, S10–S11, S13, and S15.



**Figure 5. Bar, scatter, and Bland-Altman plots of the SOY task**

Models are listed in the decreasing order of  $R^2$  with fine-tuning on the test set AMC1.

(A) Bar plot of mean  $R^2$ .

(B) Bar plot of mean CCC.

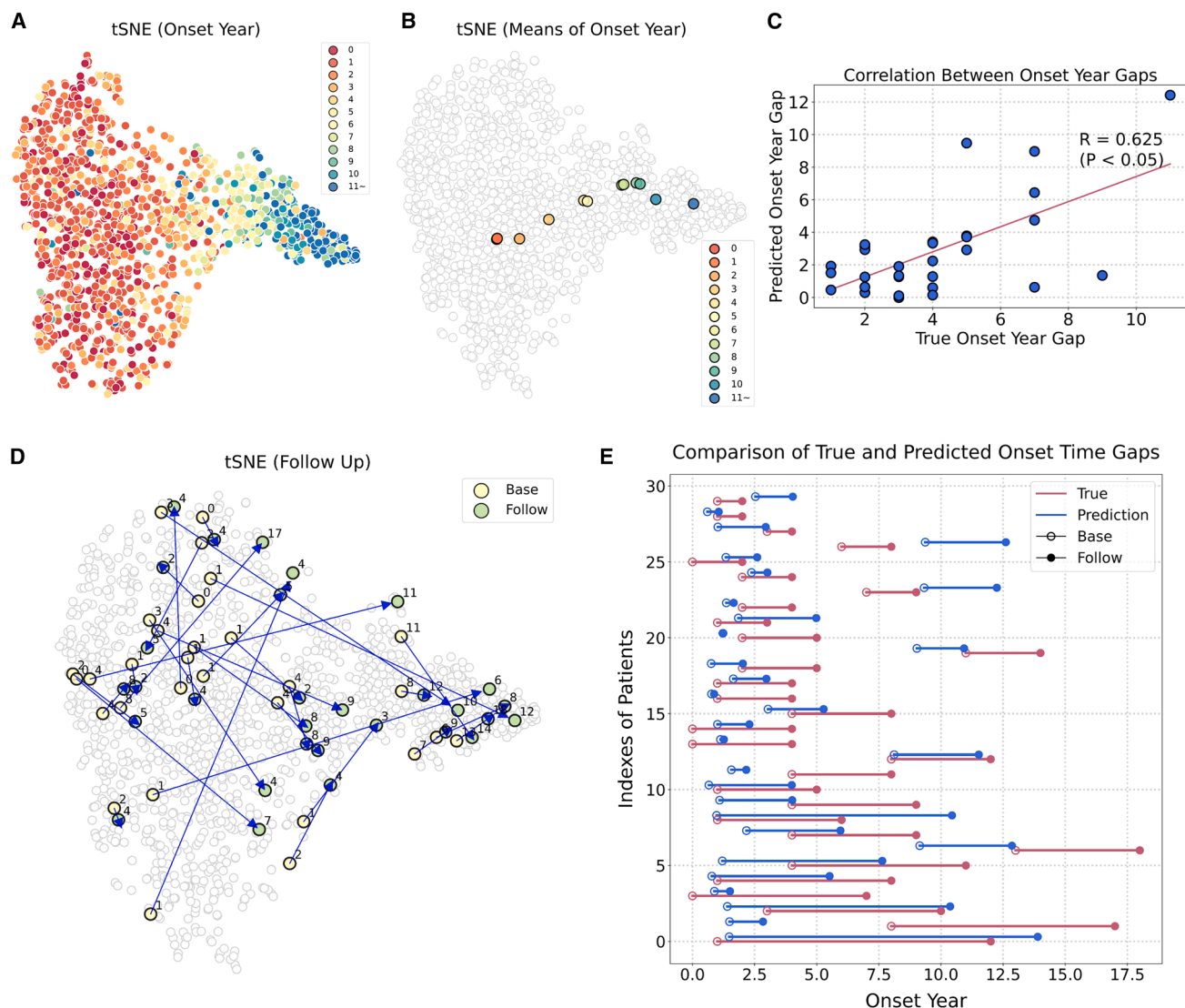
The scatter and Bland-Altman plots were generated using the upstream model with the highest mean  $R^2$  in each test set.

(C) Scatterplot of predicted vs. true onset years with a regression line and  $p$  value for AMC1 using the WDDAE upstream model.

(D) Scatterplot for AMC2 using the HWDAAE upstream model.

(E) Bland-Altman plot for AMC1 using the WDDAE upstream model.

(F) Bland-Altman plot for AMC2 using the HWDAAE upstream model. The y axis was calculated by subtracting the ground truth from the model-predicted values. See also [Figures S6 and S11](#) and [Tables S9 and S16](#).



**Figure 6. t-SNE, scatter, and line plots of the SOY task comparing baseline and follow-ups**

(A and B) were plotted with the training and test sets of AMC1. The onset years were categorized into 12 classes.

(C–E) Compare the baseline and follow-up images of the 30 patients with PD.

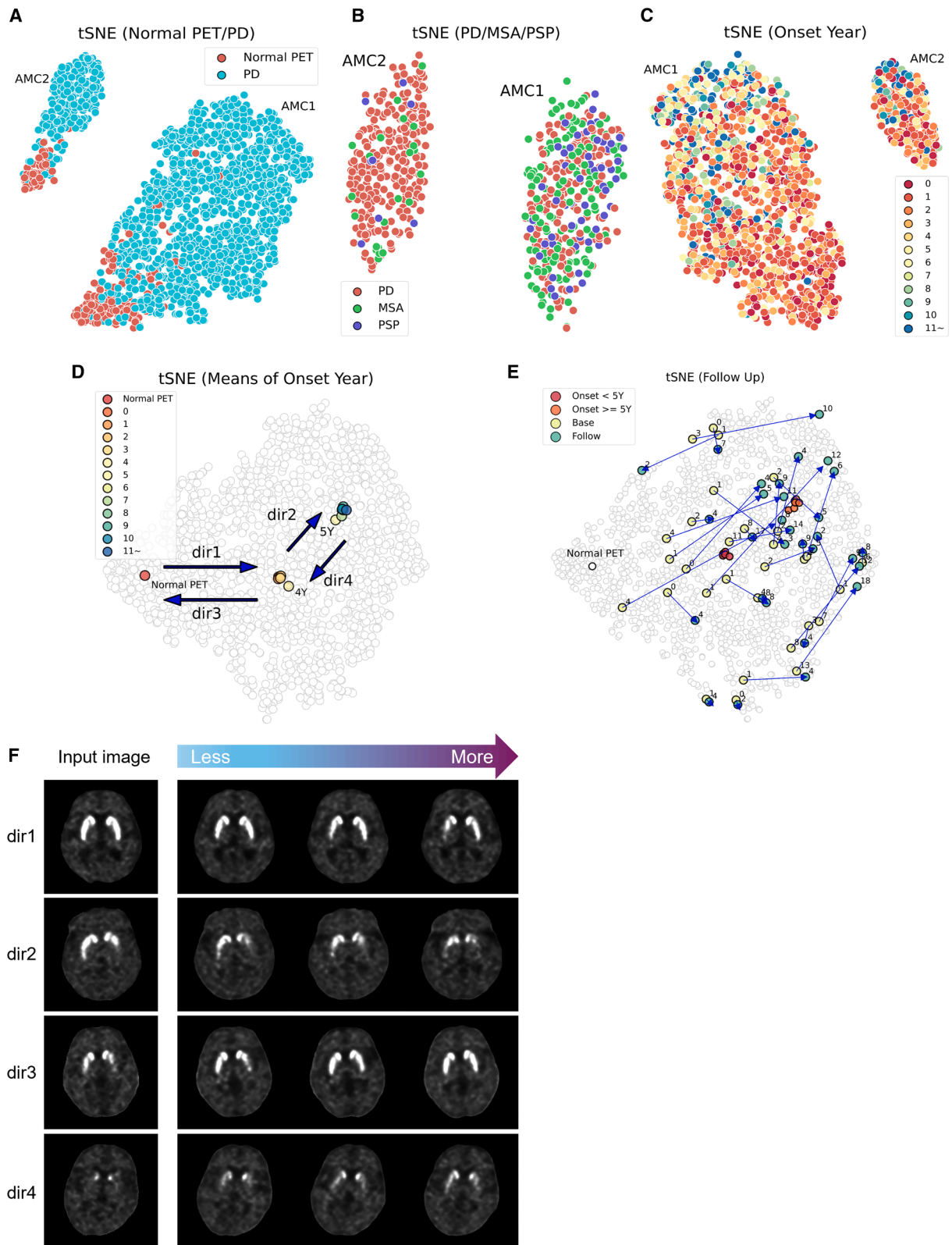
All images were generated using the WDDAE after fine-tuning. (A) t-SNE of the latent space. (B) Mean of each class (0–11~) added to (A). (C) Scatterplot of predicted vs. true onset year gap between the baseline and follow-up images with a regression line and  $p$  value. (D) Latent of baseline and follow-up images added to (A). The arrows connect the baseline and follow-up latents of the same patients, and the numbers next to the circles denote the true onset years. (E) Line plot of each patient's true and predicted onset years with the baseline and follow-up images. Each patient is represented by a pair of red and blue lines.

the upstream training set. Although the trend was not as pronounced as in the t-SNE of the WDDAE after FT (Figure 6A), the t-SNE showed a mild shifting trend from the bottom right to the top left as the onset years increased (Figure 7C). Most arrows connecting the baseline to the follow-up latent vectors of the 30 patients with PD were directed toward the top right, consistent with the results in Figure 7D. Lastly, for all three tasks, images acquired with different scanners formed separate clusters (Figures 7A–7C).

For a more in-depth evaluation of the latent space, we manipulated the images along the four direction vectors representing disease progression. The mean latent vectors of the images

with normal PET findings, patients with PD with onset years of <5 years, and those with onset years of  $\geq 5$  years were used to define the four direction vectors (Figure 7D). The manipulated images showed reasonable changes in the striatal uptake values along each direction vector (Figure 7F).

Examples of randomly and unconditionally generated images from the WDDAE, DDAE, and generator component (StyleGAN2) of P2S2P, as well as the reconstructed images from the HWDAAE, HDAE, P2S2P, SimMIM, and DisAE, are presented in Figures S12–S15. Upon visual inspection, the reconstructed images of the HWDAAE most closely resembled the striatal uptake patterns of the original images.



(legend on next page)

## DISCUSSION

In this study, we proposed a 3D generative SSL model, the HWDAAE, and evaluated its potential as a general-purpose DL model for analyzing  $^{18}\text{F}$ -FP-CIT PET images obtained from patients with parkinsonism. We assessed and compared the performance of the HWDAAE with those of other SSL models for various downstream tasks, as well as the visualization and manipulation of its latent space. Although the HWDAAE with LP or FT did not always outperform the reference model (3D ResNet34 with SC), its consistently strong performance across all three downstream tasks on both the internal and external test sets demonstrates that our generative pretraining method effectively captured the image representations of PD.

For  $^{18}\text{F}$ -FP-CIT PET to serve as a valid biomarker for PD, it should be a quantifiable, objective, reproducible, and accurate indicator of diagnosis, disease progression, and treatment response.<sup>43</sup> Since DL-based image feature extraction supports objectivity and quantifiability, we focused on improving the accuracy and reproducibility of  $^{18}\text{F}$ -FP-CIT PET across three prediction tasks using SSL. In the case of the EP task, although nuclear medicine physicians have reported high reliability in visually differentiating PD from ET or other non-neurodegenerative parkinsonism using DAT imaging,<sup>44</sup> sensitivity remains lower than the near-perfect specificity,<sup>45</sup> and some expert disagreements persist.<sup>46</sup> Furthermore, distinguishing PD from atypical parkinsonism remains challenging even for experienced neurologists.<sup>47</sup> In this study, along with diffusion models such as the WDDAE and DDAE, the HWDAAE attained high AUC and AP values in both the EP and PMP tasks. In the EP task, most models demonstrated outstanding performance in distinguishing ET from early PD. We also conducted an additional experiment for binary classification of PET images, comparing the images from patients with ET, drug-induced parkinsonism (DIP), or vascular parkinsonism (VP) who had normal PET findings with those from patients with early PD. As expected, most of the upstream models and ResNet34 consistently exhibited exceptional performances (Table S12). In comparison, both the mean AUC and AP values were generally lower in the PMP task. In our DL models evaluated on the AMC1 test set, a general trend of confusing the images from patients with PD and MSA was observed, with most misclassifications occurring in MSA-P subtype patients rather than MSA-C subtype patients (Figure S16; Table S13). Furthermore, in the AMC1 and AMC2 test sets, all images from MSA-P patients showed striatal uptake signal loss, whereas some from MSA-C patients had normal PET findings. Our results align with those of previous studies, suggesting

possible differences in the uptake pattern between MSA subtypes,<sup>48–50</sup> and indicate that separating the PET images by the MSA subtype may be necessary for a more accurate evaluation of DL models. Although some of the previous DL models using DAT PET images have achieved similar AUC results, they often lacked external validation,<sup>18,35</sup> raising concerns about their generalizability. We addressed this issue by using the data from multiple scanners and medical centers to conclude that self-supervised diffusion models can improve the model performance on unseen datasets. To rigorously verify reproducibility,  $^{18}\text{F}$ -FP-CIT PET images from the same patients obtained using different scanners at different locations are needed to evaluate the HWDAAE's classification performance. Due to the retrospective nature of our study, such datasets were unavailable. However, the HWDAAE's relatively consistent performance on external test sets suggests its potential reliability in same-patient settings.

Another essential aspect of a biomarker is its ability to stratify patients by disease progression and treatment responses for possible clinical trials. The SOY task, trained to predict disease duration using only  $^{18}\text{F}$ -FP-CIT PET images, illustrates how DAT uptake changes from early to advanced disease stages. The mean  $R^2$  and CCC values of this task were generally moderate, and the differences between the model-predicted onset years and true onset years increased when the true onset years exceeded approximately 5 years. Although our training set for the SOY task exhibited a notable data imbalance with a right-skewed distribution (Figure 2C), potentially leading to an insufficient number of training images for patients in the advanced stages of PD, it has been reported that dopaminergic function imaging becomes less sensitive in these later stages.<sup>9,51,52</sup> We also showed that although  $^{18}\text{F}$ -FP-CIT PET images may have limitations as a disease progression biomarker covering all stages of PD, DL applied to  $^{18}\text{F}$ -FP-CIT PET can extract the image features more closely associated with disease duration when compared with striatal SNBRs. Additionally, the increasing variability in the differences between image-predicted and chronological disease duration among patients with advanced PD suggests that the rate of DAT availability decline may vary among patients, reflecting the heterogeneity of PD. While most clinical trials for disease-modifying treatments in neurodegenerative diseases tend to be concentrated on early-stage patients, identifying patients with advanced PD with relatively preserved DAT availability, as quantified by image-predicted disease duration, may help assess varying treatment responses among subjects. Thus, imaging features of  $^{18}\text{F}$ -FP-CIT PET may contribute to the stratification of patients with PD and potentially assist in

### Figure 7. t-SNE and latent space manipulation using HWDAAE without fine-tuning

(A–E) Plotted using (A–C) both AMC1 training/test set and AMC2 test set and (D and E) using AMC1 training/test set only. (A) t-SNE of images from patients with PD vs. those diagnosed with either vascular parkinsonism, drug-induced parkinsonism, or ET, all of whom had normal PET findings. (B) t-SNE of images from patients with PD vs. MSA vs. PSP. (C) t-SNE of images from patients with PD grouped by symptom onset years. (D) t-SNE of images from the patients with normal PET findings as in (A) and patients with PD grouped by symptom onset years. The mean latent vector of each group and four direction vectors. (E) Latent of baseline and follow-up images of the 30 patients added to (D). The arrows connect the baseline and follow-up latents of the same patients, and the numbers next to circles denote the true onset years.

(F) Example input and manipulated images. The input images were from patients with normal PET findings and with symptom onset years of <5 years, <5 years, and  $\geq 5$  years, corresponding to dir1 through dir4 from (D), respectively. The manipulated images are arranged in the order of increasing multiplication factors of each direction vector.

See also Figures S12–S15.

patient selection and treatment response interpretation in therapeutic interventions.

Rather than simply using the latent space of upstream SSL models for specific tasks through FT, we took a step further by dissecting the latent space by hierarchical blocks of the HWDAAE's encoder. Although the HWDAAE's encoder appeared to successfully reconstruct the input images and outperformed other upstream models on the test set AMC1, its results fluctuated on AMC2 and SCH. We hypothesized that the model was focusing on both high- and low-level features, possibly including features that were less relevant to the downstream tasks. By selecting the third block for FT, we observed a more stable performance across different test sets and hyperparameter tuning ablation studies (Tables S10 and S11). Thus, we suggest that utilizing specific blocks of the latent space in upstream SSL models could be a potential solution to mitigate overfitting. However, further studies are needed, as we could not identify specific blocks responsible for other downstream tasks.

After confirming that the HWDAAE with LP, where latent vectors were frozen, showed moderate performance in the SOY task and that t-SNE without FT displayed a gradient from low to high onset years—indicating the HWDAAE's ability to learn the semantics of  $^{18}\text{F}$ -FP-CIT PET imaging without human supervision—we generated PET images corresponding to different disease durations through latent space manipulation. The latent space of the HWDAAE synthesized plausible PET images that allowed us to move backward or forward in time specific to each patient. The synthesized images enabled individualized visualization of the nigrostriatal degeneration process over time, which can be valuable for educational purposes for both students and patients. Looking forward, manipulating the latent space based on other clinical features related to PD progression, such as Hoehn and Yahr scales and Unified Parkinson's Disease Rating Scale scores, could enhance the model's interpretability and elucidate the relationship between  $^{18}\text{F}$ -FP-CIT PET imaging and disease progression. This approach is particularly relevant, as disease progression or stage cannot solely be determined by onset years but also by symptom severity, given the variability in clinical progression among patients with PD with similar onset years.

We initially hypothesized that a single SSL upstream model would outperform other SSL models in all downstream tasks. However, we observed that differential diagnosis and disease progression tasks may rely on different image features from distinct SSL models, as noted in previous studies on Alzheimer's disease classification and brain age regression using MRI.<sup>53,54</sup> In our study, SimMIM showed the poorest performance in the EP and PMP tasks but achieved relatively high performance in the SOY task. The HWDAAE and WDDAAE with LP demonstrated similar performances in the EP and the PMP tasks, although the HWDAAE with LP outperformed the WDDAAE by a large margin in the SOY task. Furthermore, no single SSL model outperformed the reference model ResNet34 across all downstream tasks and test sets. Although developing a general-purpose DL model for various tasks and datasets remains challenging, the HWDAAE consistently emerged as the top SSL model in most of our tasks and datasets. In addition to its strong performance, the HWDAAE offers advantages in the model training process. In

the case of P2S2P, we had to adopt the cascade method for training its generator owing to its training instability. For the DDAE, an additional tuning process for the optimal noise level was needed, as its performance could vary with the time steps. Lastly, the performance of ResNet34 fluctuated considerably depending on the hyperparameter settings, such as the base learning rate and batch size (Tables S10 and S11).

Another technique that we applied to diffusion models was DWT preprocessing. We constructed our diffusion models using 3D convolutional neural networks to eliminate the need to select 2D image slices that encompass the ROIs, thereby mitigating experimenter bias and the labor-intensive process associated with large datasets. To accommodate the high-resolution input images and computational demands of 3D models, we used DWT preprocessing, which has been utilized in several studies to address data shortages<sup>55</sup> or to manage high-dimensional feature maps.<sup>56</sup> With DWT preprocessing, although models converged more slowly compared to those without DWT preprocessing, we could increase the initial number of channels in both the HDAE (from 8) and the DDAE (from 16) to 128. In addition, we observed consistent performance gains when DWT was added to the HDAE, forming the HWDAAE.

In conclusion, we developed a GM for SSL of  $^{18}\text{F}$ -FP-CIT PET images of patients with parkinsonism and validated its image representation ability across multiple center settings. On diverse downstream tasks encompassing both classification and regression, the HWDAAE showed strong performance in addition to an impressive ability to capture semantic image features without supervision. Notably, while the model showed satisfactory performance in classifying PD from ET, its ability to differentiate PD from atypical parkinsonism, such as MSA and PSP, highlights its potential clinical utility. Given the clinical challenge of distinguishing PD from other degenerative parkinsonism due to overlapping motor and nonmotor symptoms, the HWDAAE could be a valuable diagnostic aid in clinical practice. Through this study, we once again demonstrated that GMs are capable of learning both low- and high-level features, thus broadening their applicability in both image generation and discrimination. SSL-based general-purpose DL models of PET images offer a solution for adapting DL techniques to real clinical environments with limited GPU resources while also pushing the boundaries of  $^{18}\text{F}$ -FP-CIT PET imaging as a potential biomarker for PD.

### Limitations of the study

Our study has several limitations. First, the diagnoses of patients were based on clinical diagnostic criteria rather than pathological confirmation. Second, we did not have PET images from normal controls, as the data were retrospectively collected in a real-world setting. Third, in the SOY task, the motor symptom onset years were estimated based on patient interviews, which may not be perfectly accurate. Fourth,  $^{18}\text{F}$ -FP-CIT PET images from other atypical parkinsonism, such as dementia with Lewy bodies or corticobasal degeneration, were not included. Fifth, detailed demographics and clinical characteristics beyond sex and age at PET image acquisition were not reported, as data collection was limited to variables essential for downstream tasks.

We also encountered some limitations regarding model training. First, despite our efforts to create balanced training

and test sets, the images from patients with PD in the early stages of the disease were predominant when compared with those from patients with MSA, PSP, and advanced PD. This data imbalance was also present in the dataset used for upstream SSL model training, which may have limited the models' ability to learn latent representations of the minor classes. This could explain why FT was not always superior to SC and why our HWDAAE with FT or LP did not consistently outperform ResNet34. Second, the images used in the training process of the downstream tasks were also included in the upstream training, which may have led to overfitting on the internal test set AMC1. We faced a data shortage in training the upstream models, as DL models typically require large datasets, and  $^{18}\text{F}$ -FP-CIT PET is not a commonly available imaging modality. However, we plan to expand our dataset and address this limitation in the future. Third, SSL models are affected by the backbone architecture. Although we followed the encoder architecture used in previous studies by truncating the U-Net of the DDAE in the middle, the optimal encoder architecture for our downstream tasks requires further experimentation. Fourth, SSL models, including the HWDAAE, rely on large model sizes and require long training hours for optimization. However, once optimized, they can be shared among researchers as general-purpose DL models. Fifth, hyperparameters for the upstream models were not strictly standardized across architectures. For example, the maximum number of epochs was chosen to ensure model stability in generating or reconstructing images, resulting in variations across models. To address this, we conducted an experiment where all upstream models were uniformly trained for 500 epochs. While some performance fluctuations occurred across tasks and test sets, no model showed a dramatic change in ranking or consistent performance gain compared to previous results (Tables S14–S16). Lastly, the performance gains achieved using the HWDAAE as a pretrained model should be further validated on other brain imaging modalities, such as structural MRI sequences.

## RESOURCE AVAILABILITY

### Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Namkug Kim ([namkugkim@gmail.com](mailto:namkugkim@gmail.com)).

### Materials availability

This study did not generate new unique reagents.

### Data and code availability

The original image data utilized in this study are not publicly accessible owing to patient privacy concerns. However, data reported in this paper will be shared by the lead contact upon request with institutional review board approval. Synthesized images using the WDDAAE, model weights, and all original codes for preprocessing and model implementation have been deposited at [https://github.com/Krying/PD\\_SSL\\_ZOO](https://github.com/Krying/PD_SSL_ZOO) and are publicly available at <https://doi.org/10.5281/zenodo.15377683> as of the date of publication. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

## ACKNOWLEDGMENTS

This research was supported by the National Research Foundation of Korea (NRF) (RS-2023-00262527). This research was supported by a grant from

the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (RS-2023-00265820 and HR20C0026). This research was also supported by a grant from the MD-PhD/Medical Scientist Training Program through the KHIDI, funded by the Ministry of Health & Welfare, Republic of Korea.

## AUTHOR CONTRIBUTIONS

J.W. developed the DL models and performed the experiments. G.Y.L. collected the data, performed experiments, and contributed to the study design. J.W. and G.Y.L. wrote the manuscript. S.J., Jihyun L.e.e, and S.L. assisted with data collection and study design. J.S.K., C.S., and J.S.O. assisted with image preprocessing and interpretation. K.-Y.K. and S.B.P. collected the external test set from SCH. Joonsang Lee and J.Y. assisted with subgroup analysis. S.J.C. and N.K. jointly directed the study and reviewed the manuscript. S.J.C. contributed to the study design and oversaw data collection and curation from the perspective of a movement specialist, while N.K. supervised the model development from an engineering perspective, utilizing expertise in medical image analysis. All authors reviewed the manuscript.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

## DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

During the preparation of this work, the authors used ChatGPT-4 (OpenAI, <https://chat.openai.com/>) in order to improve grammar and clarity. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS**
  - Asan Medical Center (AMC) dataset
  - Soonchunhyang University Seoul Hospital (SCH) dataset
  - Ethics
- **METHOD DETAILS**
  - $^{18}\text{F}$ -FP-CIT PET image acquisition
  - T1 MRI image acquisition
  - $^{18}\text{F}$ -FP-CIT PET and T1 MRI image preprocessing
  - DWT
  - Diffusion model architecture and implementation
  - Other model architectures and implementation
  - Hyperparameter selection: Maximum number of epochs for upstream models
  - Adaptation to downstream tasks
  - Latent space manipulation
  - Computational hardware and software
- **QUANTIFICATION AND STATISTICAL ANALYSIS**
  - Evaluation metrics
  - Data fold splitting and statistical analysis

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.xcrm.2025.102207>.

Received: January 10, 2025

Revised: April 6, 2025

Accepted: June 2, 2025

Published: June 27, 2025

## REFERENCES

- Ben-Shlomo, Y., Darweesh, S., Llibre-Guerra, J., Marras, C., San Luciano, M., and Tanner, C. (2024). The epidemiology of Parkinson's disease. *Lancet* 403, 283–292. [https://doi.org/10.1016/S0140-6736\(23\)01419-8](https://doi.org/10.1016/S0140-6736(23)01419-8).
- Food and Drug Administration. FDA DaT Scan Approval Letter. (2011). [https://www.accessdata.fda.gov/drugsatfda\\_docs/label/2011/022454Orig1s000Lbl.pdf](https://www.accessdata.fda.gov/drugsatfda_docs/label/2011/022454Orig1s000Lbl.pdf).
- Summary of the European Public Assessment Report for DaTSCAN. (2011). [ema.europa.eu/en/medicines/human/EPAR/datscan#product-information-section](https://ema.europa.eu/en/medicines/human/EPAR/datscan#product-information-section).
- Hastings, A., Cullinane, P., Wrigley, S., Revesz, T., Morris, H.R., Dickson, J.C., Jaunmuktane, Z., Warner, T.T., and De Pablo-Fernández, E. (2024). Neuropathologic Validation and Diagnostic Accuracy of Presynaptic Dopaminergic Imaging in the Diagnosis of Parkinsonism. *Neurology* 102, e209453. <https://doi.org/10.1212/WNL.000000000000209453>.
- Bega, D., Kuo, P.H., Chalkidou, A., Grzeda, M.T., Macmillan, T., Brand, C., Sheikh, Z.H., and Antonini, A. (2021). Clinical utility of DaTscan in patients with suspected Parkinsonian syndrome: a systematic review and meta-analysis. *NPJ Parkinsons Dis.* 7, 43. <https://doi.org/10.1038/s41531-021-00185-8>.
- Eshuis, S.A., Jager, P.L., Maguire, R.P., Jonkman, S., Dierckx, R.A., and Leenders, K.L. (2009). Direct comparison of FP-CIT SPECT and F-DOPA PET in patients with Parkinson's disease and healthy controls. *Eur. J. Nucl. Med. Mol. Imaging* 36, 454–462. <https://doi.org/10.1007/s00259-008-0989-5>.
- Lee, I., Kim, J.S., Park, J.Y., Byun, B.H., Park, S.Y., Choi, J.H., Moon, H., Kim, J.Y., Lee, K.C., Chi, D.Y., et al. (2018). Head-to-head comparison of (18) F-FP-CIT and (123) I-FP-CIT for dopamine transporter imaging in patients with Parkinson's disease: A preliminary study. *Synapse (N. Y.)* 72, e22032. <https://doi.org/10.1002/syn.22032>.
- Huang, W.-S., Chiang, Y.-H., Lin, J.-C., Chou, Y.-H., Cheng, C.-Y., and Liu, R.-S. (2003). Crossover study of 99mTc-TRODAT-1 SPECT and 18F-FDOPA PET in Parkinson's disease patients. *J. Nucl. Med.* 44, 999–1005.
- Mitchell, T., Lehericy, S., Chiu, S.Y., Strafella, A.P., Stoessl, A.J., and Vailancourt, D.E. (2021). Emerging Neuroimaging Biomarkers Across Disease Stage in Parkinson Disease: A Review. *JAMA Neurol.* 78, 1262–1272. <https://doi.org/10.1001/jamaneurol.2021.1312>.
- Chung, S.J., Kim, Y.J., Kim, Y.J., Lee, H.S., Jeong, S.H., Hong, J.M., Sohn, Y.H., Yun, M., Jeong, Y., and Lee, P.H. (2022). Association Between White Matter Networks and the Pattern of Striatal Dopamine Depletion in Patients With Parkinson Disease. *Neurology* 99, e2672–e2682. <https://doi.org/10.1212/WNL.000000000000201269>.
- Chung, S.J., Yoo, H.S., Lee, H.S., Oh, J.S., Kim, J.S., Sohn, Y.H., and Lee, P.H. (2018). The Pattern of Striatal Dopamine Depletion as a Prognostic Marker in De Novo Parkinson Disease. *Clin. Nucl. Med.* 43, 787–792. <https://doi.org/10.1097/RLU.0000000000002251>.
- Labrador-Espinosa, M.A., Grothe, M.J., Macías-García, D., Jesús, S., Adames-Gómez, A., Muñoz-Delgado, L., Fernández-Rodríguez, P., Martín-Rodríguez, J.F., Huertas, I., García-Solis, D., and Mir, P. (2021). Levodopa-Induced Dyskinesia in Parkinson Disease Specifically Associates With Dopaminergic Depletion in Sensorimotor-Related Functional Subregions of the Striatum. *Clin. Nucl. Med.* 46, e296–e306. <https://doi.org/10.1097/RLU.00000000000003609>.
- Niemi, K.J., Sunikka, J., Soltanian-Zadeh, H., Davoodi-Bojd, E., Rahmim, A., Kaasinen, V., and Joutsma, J. (2024). Rest Tremor in Parkinson's Disease Is Associated with Ipsilateral Striatal Dopamine Transporter Binding. *Mov. Disord.* 39, 2014–2025. <https://doi.org/10.1002/mds.29997>.
- García-Gómez, F.J., García-Solis, D., Luis-Simón, F.J., Marín-Oyaga, V.A., Carrillo, F., Mir, P., and Vázquez-Albertino, R.J. (2013). Elaboration of the SPM template for the standardization of SPECT images with 123I-ioflupane. *Rev. Española Med. Nucl. Imagen Mol.* 32, 350–356.
- Gómez, F.J.G., Huertas, I., Ramírez, J.A.L., and Solís, D.G. (2018). Elaboración de una plantilla de SPM para la normalización de imágenes de PET con 18F-DOPA. *Imagen Diagnóstica* 9, 23–25.
- Alae Eddine, E.B., Scheiber, C., Grenier, T., Janier, M., and Flaus, A. (2024). CT-guided spatial normalization of nuclear hybrid imaging adapted to enlarged ventricles: Impact on striatal uptake quantification. *Neuroimage* 294, 120631. <https://doi.org/10.1016/j.neuroimage.2024.120631>.
- Wang, J., Xue, L., Jiang, J., Liu, F., Wu, P., Lu, J., Zhang, H., Bao, W., Xu, Q., Ju, Z., et al. (2024). Diagnostic performance of artificial intelligence-assisted PET imaging for Parkinson's disease: a systematic review and meta-analysis. *NPJ Digit. Med.* 7, 17. <https://doi.org/10.1038/s41746-024-01012-z>.
- Zhao, Y., Wu, P., Wu, J., Brendel, M., Lu, J., Ge, J., Tang, C., Hong, J., Xu, Q., Liu, F., et al. (2022). Decoding the dopamine transporter imaging for the differential diagnosis of parkinsonism using deep learning. *Eur. J. Nucl. Med. Mol. Imaging* 49, 2798–2811. <https://doi.org/10.1007/s00259-022-05804-x>.
- Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al. (2023). A foundation model for generalizable disease detection from retinal images. *Nature* 622, 156–163. <https://doi.org/10.1038/s41586-023-06555-x>.
- Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al. (2024). A whole-slide foundation model for digital pathology from real-world data. *Nature* 630, 181–188. <https://doi.org/10.1038/s41586-024-07441-w>.
- Tiu, E., Talus, E., Patel, P., Langlotz, C.P., Ng, A.Y., and Rajpurkar, P. (2022). Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng.* 6, 1399–1406. <https://doi.org/10.1038/s41551-022-00936-9>.
- Zhao, Q., Liu, Z., Adeli, E., and Pohl, K.M. (2021). Longitudinal self-supervised learning. *Med. Image Anal.* 71, 102051. <https://doi.org/10.1016/j.media.2021.102051>.
- Dhariwal, P., and Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* 34, 8780–8794.
- Peebles, W., and Xie, S. (2023). Scalable diffusion models with transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.
- Karras, T., Aittala, M., Aila, T., and Laine, S. (2022). Elucidating the design space of diffusion-based generative models. *Adv. Neural Inf. Process. Syst.* 35, 26565–26577.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2204.06125>.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. p. 10684–10695.
- Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghaseмпour, K., Gontijo Lopes, R., Karagol Ayan, B., and Salimans, T. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Adv. Neural Inf. Process. Syst.* 35, 36479–36494.
- Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. (2022). Prompt-to-prompt image editing with cross attention control. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2208.01626>.
- Xiang, W., Yang, H., Huang, D., and Wang, Y. (2023). Denoising diffusion autoencoders are unified self-supervised learners. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15802–15812.
- Han, T., Žigutytė, L., Huck, L., Huppertz, M.S., Siepmann, R., Gandelsman, Y., Blüthgen, C., Khader, F., Kuhl, C., Nebelung, S., et al. (2024). Reconstruction of patient-specific confounders in AI-based radiologic

- image interpretation using generative pretraining. *Cell Rep. Med.* 5, 101713. <https://doi.org/10.1016/j.xcrm.2024.101713>.
32. Ijshakin, A., Abdulaal, A., Hadjivasilou, A., Martin, S., and Cole, J. (2023). Interpretable Alzheimer's Disease Classification Via a Contrastive Diffusion Autoencoder. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2306.03022>.
33. Lu, Z., Wu, C., Chen, X., Wang, Y., Bai, L., Qiao, Y., and Liu, X. (2024). Hierarchical diffusion autoencoders and disentangled image manipulation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5374–5383.
34. Sun, J., Cong, C., Li, X., Zhou, W., Xia, R., Liu, H., Wang, Y., Xu, Z., and Chen, X. (2024). Identification of Parkinson's disease and multiple system atrophy using multimodal PET/MRI radiomics. *Eur. Radiol.* 34, 662–672. <https://doi.org/10.1007/s00330-023-10003-9>.
35. Piccardo, A., Cappuccio, R., Bottoni, G., Cecchin, D., Mazzella, L., Cirone, A., Righi, S., Ugolini, M., Bianchi, P., Bertolaccini, P., et al. (2021). The role of the deep convolutional neural network as an aid to interpreting brain [(18)F]DOPA PET/CT in the diagnosis of Parkinson's disease. *Eur. Radiol.* 31, 7003–7011. <https://doi.org/10.1007/s00330-021-07779-z>.
36. Mallat, S.G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 11, 674–693.
37. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. (2022). Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9653–9663.
38. Valanarasu, J.M.J., Tang, Y., Yang, D., Xu, Z., Zhao, C., Li, W., Patel, V.M., Landman, B., Xu, D., and He, Y. (2023). Disruptive Autoencoders: Leveraging Low-level features for 3D Medical Image Pre-training. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2307.16896>.
39. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., and Cohen-Or, D. (2021). Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2287–2296.
40. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* 128, 336–359. <https://doi.org/10.1007/s11263-019-01228-7>.
41. Chicco, D., Warrens, M.J., and Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* 7, e623. <https://doi.org/10.7717/peerj-cs.623>.
42. Lawrence, I., and Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45, 255–268.
43. Strimbu, K., and Tavel, J.A. (2010). What are biomarkers? *Curr. Opin. HIV AIDS* 5, 463–466. <https://doi.org/10.1097/COH.0b013e32833ed177>.
44. Suwijn, S.R., Verschuur, C.V.M., Slim, M.A., Booi, J., and de Bie, R.M.A. (2019). Reliability of visual assessment by non-expert nuclear medicine physicians and appropriateness of indications of [(123)I]FP-CIT SPECT imaging by neurologists in patients with early drug-naïve Parkinson's disease. *EJNMMI Res.* 9, 63. <https://doi.org/10.1186/s13550-019-0537-2>.
45. Vlaar, A.M.M., van Kroonenburgh, M.J.P.G., Kessels, A.G.H., and Weber, W.E.J. (2007). Meta-analysis of the literature on diagnostic accuracy of SPECT in parkinsonian syndromes. *BMC Neurol.* 7, 27. <https://doi.org/10.1186/1471-2377-7-27>.
46. Choi, H., Ha, S., Im, H.J., Paek, S.H., and Lee, D.S. (2017). Refining diagnosis of Parkinson's disease with deep learning-based interpretation of dopamine transporter imaging. *Neuroimage. Clin.* 16, 586–594. <https://doi.org/10.1016/j.nicl.2017.09.010>.
47. Hughes, A.J., Daniel, S.E., Ben-Shlomo, Y., and Lees, A.J. (2002). The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service. *Brain* 125, 861–870.
48. Wallert, E.D., van de Giessen, E., Knol, R.J.J., Beudel, M., de Bie, R.M.A., and Booi, J. (2022). Imaging Dopaminergic Neurotransmission in Neurodegenerative Disorders. *J. Nucl. Med.* 63, 27S–32S. <https://doi.org/10.2967/jnumed.121.263197>.
49. Bu, L.L., Liu, F.T., Jiang, C.F., Guo, S.S., Yu, H., Zuo, C.T., Wu, P., and Wang, J. (2018). Patterns of dopamine transporter imaging in subtypes of multiple system atrophy. *Acta Neurol. Scand.* 138, 170–176. <https://doi.org/10.1111/ane.12932>.
50. Yang, Y., Li, X., Lu, J., Ge, J., Chen, M., Yao, R., Tian, M., Wang, J., Liu, F., and Zuo, C. (2025). Recent progress in the applications of presynaptic dopaminergic positron emission tomography imaging in parkinsonism. *Neural Regen. Res.* 20, 93–106. <https://doi.org/10.4103/1673-5374.391180>.
51. Bruck, A., Aalto, S., Rauhala, E., Bergman, J., Marttila, R., and Rinne, J.O. (2009). A follow-up study on 6-[18F]fluoro-L-dopa uptake in early Parkinson's disease shows nonlinear progression in the putamen. *Mov. Disord.* 24, 1009–1015. <https://doi.org/10.1002/mds.22484>.
52. Nandhagopal, R., Kuramoto, L., Schulzer, M., Mak, E., Cragg, J., McKenzie, J., McCormick, S., Ruth, T.J., Sossi, V., de la Fuente-Fernandez, R., and Stoessl, A.J. (2011). Longitudinal evolution of compensatory changes in striatal dopamine processing in Parkinson's disease. *Brain* 134, 3290–3298. <https://doi.org/10.1093/brain/awr233>.
53. Ouyang, J., Zhao, Q., Adeli, E., Zaharchuk, G., and Pohl, K.M. (2022). Self-supervised learning of neighborhood embedding for longitudinal MRI. *Med. Image Anal.* 82, 102571. <https://doi.org/10.1016/j.media.2022.102571>.
54. Barbano, C.A., Brunello, M., Dufumier, B., and Grangetto, M. (2024). Anatomical Foundation Models for Brain MRIs. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2408.07079>.
55. Zhu, J., Ma, H., Chen, J., and Yuan, J. (2022). Few-shot image generation with diffusion models. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2211.03264>.
56. Hoogeboom, E., Heek, J., and Salimans, T. (2023). simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pp. 13213–13232.
57. Gibb, W.R., and Lees, A.J. (1988). The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* 51, 745–752.
58. Gilman, S., Wenning, G.K., Low, P.a., Brooks, D.J., Mathias, C.J., Trojanowski, J.Q., Wood, N.W., Colosimo, C., Dürr, A., Fowler, C.J., et al. (2008). Second consensus statement on the diagnosis of multiple system atrophy. *Neurology* 71, 670–676.
59. Hoglinger, G.U., Respondek, G., Stamelou, M., Kurz, C., Josephs, K.A., Lang, A.E., Mollenhauer, B., Müller, U., Nilsson, C., Whitwell, J.L., et al. (2017). Clinical diagnosis of progressive supranuclear palsy: The movement disorder society criteria. *Mov. Disord.* 32, 853–864. <https://doi.org/10.1002/mds.26987>.
60. Elble, R.J. (2000). Diagnostic criteria for essential tremor and differential diagnosis. *Neurology* 54, S2–S6.
61. Erro, R., Schneider, S.A., Stamelou, M., Quinn, N.P., and Bhatia, K.P. (2016). What do patients with scans without evidence of dopaminergic deficit (SWEDD) have? New evidence and continuing controversies. *J. Neurol. Neurosurg. Psychiatry* 87, 319–323. <https://doi.org/10.1136/jnnp-2014-310256>.
62. Lee, S.J., Oh, S.J., Chi, D.Y., Kang, S.H., Kil, H.S., Kim, J.S., and Moon, D. H. (2007). One-step high-radiochemical-yield synthesis of [18F]FP-CIT using a protic solvent system. *Nucl. Med. Biol.* 34, 345–351. <https://doi.org/10.1016/j.nucmedbio.2007.02.007>.
63. Hoopes, A., Mora, J.S., Dalca, A.V., Fischl, B., and Hoffmann, M. (2022). SynthStrip: skull-stripping for any brain image. *Neuroimage* 260, 119474.
64. Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.P., Heiland, S., Wick, W., et al. (2019). Automated brain extraction of multisequence MRI using artificial

- neural networks. *Hum. Brain Mapp.* 40, 4952–4964. <https://doi.org/10.1002/hbm.24750>.
65. Tziortzi, A.C., Searle, G.E., Tzimopoulou, S., Salinas, C., Beaver, J.D., Jenkinson, M., Laruelle, M., Rabiner, E.A., and Gunn, R.N. (2011). Imaging dopamine receptors in humans with [<sup>11</sup>C]-(+)-PHNO: dissection of D3 signal and anatomy. *Neuroimage* 54, 264–277. <https://doi.org/10.1016/j.neuroimage.2010.06.044>.
66. Tziortzi, A.C., Haber, S.N., Searle, G.E., Tsoumpas, C., Long, C.J., Shotbolt, P., Douaud, G., Jbabdi, S., Behrens, T.E.J., Rabiner, E.A., et al. (2014). Connectivity-based functional analysis of dopamine release in the striatum using diffusion-weighted MRI and positron emission tomography. *Cereb. Cortex* 24, 1165–1177. <https://doi.org/10.1093/cercor/bhs397>.
67. Rolls, E.T., Huang, C.C., Lin, C.P., Feng, J., and Joliot, M. (2020). Automated anatomical labelling atlas 3. *Neuroimage* 206, 116189. <https://doi.org/10.1016/j.neuroimage.2019.116189>.
68. Graps, A. (1995). An introduction to wavelets. *IEEE Comput. Sci. Eng.* 2, 50–61.
69. Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* 33, 6840–6851.
70. Hang, T., Gu, S., Li, C., Bao, J., Chen, D., Hu, H., Geng, X., and Guo, B. (2023). Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7441–7451.
71. Loshchilov, I. (2017). Decoupled weight decay regularization. Preprint at arXiv. <https://doi.org/10.48550/arXiv.1711.05101>.
72. Song, J., Meng, C., and Ermon, S. (2020). Denoising diffusion implicit models. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2010.02502>.
73. Loshchilov, I., and Hutter, F. (2016). SGDR: Stochastic gradient descent with warm restarts. Preprint at arXiv. <https://doi.org/10.48550/arXiv.1608.03983>.
74. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119.
75. Lim, J.H., and Ye, J.C. (2017). Geometric GAN. Preprint at arXiv. <https://doi.org/10.48550/arXiv.1705.02894>.
76. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., and Xu, D. (2021). Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pp. 272–284.
77. Pinaya, W.H., Graham, M.S., Kerfoot, E., Tudosi, P.D., Dafflon, J., Fernandez, V., Sanchez, P., Wolleb, J., da Costa, P.F., and Patel, A. (2023). Generative AI for medical imaging: extending the MONAI framework. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2307.15208>.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE                                 | SOURCE                                                                                                                                                      | IDENTIFIER                                                                                    |
|-----------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
| <b>Deposited data</b>                               |                                                                                                                                                             |                                                                                               |
| Source codes, model weights, and synthesized images | GitHub: <a href="https://github.com/Krying/PD_SSL_ZOO">https://github.com/Krying/PD_SSL_ZOO</a>                                                             | <a href="https://doi.org/10.5281/zenodo.15377683">https://doi.org/10.5281/zenodo.15377683</a> |
| <b>Software and algorithms</b>                      |                                                                                                                                                             |                                                                                               |
| Python (3.10.13)                                    | <a href="https://www.python.org">https://www.python.org</a>                                                                                                 | RRID: SCR_008394                                                                              |
| PyTorch (2.1.2)                                     | <a href="https://pytorch.org">https://pytorch.org</a>                                                                                                       | RRID: SCR_018536                                                                              |
| NVIDIA CUDA (12.1)                                  | <a href="https://developer.nvidia.com">https://developer.nvidia.com</a>                                                                                     | N/A                                                                                           |
| accelerate (0.26.1)                                 | <a href="https://github.com/huggingface/accelerate">https://github.com/huggingface/accelerate</a>                                                           | N/A                                                                                           |
| ema-pytorch (0.3.0)                                 | <a href="https://github.com/lucidrains/ema-pytorch">https://github.com/lucidrains/ema-pytorch</a>                                                           | N/A                                                                                           |
| xformers (0.0.23.post1)                             | <a href="https://facebookresearch.github.io/xformers/">https://facebookresearch.github.io/xformers/</a>                                                     | N/A                                                                                           |
| monai (1.3.0)                                       | <a href="https://monai.io/">https://monai.io/</a>                                                                                                           | N/A                                                                                           |
| numpy (1.26.2)                                      | <a href="https://numpy.org/">https://numpy.org/</a>                                                                                                         | RRID: SCR_008633                                                                              |
| SimpleITK (2.2.0)                                   | <a href="https://simpleitk.org/">https://simpleitk.org/</a>                                                                                                 | RRID: SCR_024693                                                                              |
| PyWavelets (1.5.0)                                  | <a href="https://github.com/PyWavelets/pywt">https://github.com/PyWavelets/pywt</a>                                                                         | RRID: SCR_024421                                                                              |
| Scipy (1.14.1)                                      | <a href="https://scipy.org/">https://scipy.org/</a>                                                                                                         | RRID: SCR_008058                                                                              |
| Matplotlib (3.9.2)                                  | <a href="https://matplotlib.org">https://matplotlib.org</a>                                                                                                 | RRID: SCR_008624                                                                              |
| Captum (0.7.0)                                      | <a href="https://captum.ai/">https://captum.ai/</a>                                                                                                         | N/A                                                                                           |
| Scikit-learn (1.4.0)                                | <a href="http://scikit-learn.org/">http://scikit-learn.org/</a>                                                                                             | RRID: SCR_002577                                                                              |
| SimMIM, DisAE                                       | <a href="https://github.com/Project-MONAI/research-contributions/tree/main/DAE">https://github.com/Project-MONAI/research-contributions/tree/main/DAE</a>   | N/A                                                                                           |
| HWDAE, HDAE, WDDAE, DDAE                            | <a href="https://github.com/Project-MONAI/GenerativeModels/tree/main/generative">https://github.com/Project-MONAI/GenerativeModels/tree/main/generative</a> | N/A                                                                                           |
| WDDAE, DDAE                                         | <a href="https://github.com/lucidrains/denoising-diffusion-pytorch">https://github.com/lucidrains/denoising-diffusion-pytorch</a>                           | N/A                                                                                           |
| StyleGAN2                                           | <a href="https://github.com/lucidrains/stylegan2-pytorch">https://github.com/lucidrains/stylegan2-pytorch</a>                                               | N/A                                                                                           |
| P2S2P                                               | <a href="https://github.com/eladrich/pixel2style2pixel">https://github.com/eladrich/pixel2style2pixel</a>                                                   | N/A                                                                                           |
| SynthStrip                                          | <a href="https://surfer.nmr.mgh.harvard.edu/pub/dist/freesurfer/dev/">https://surfer.nmr.mgh.harvard.edu/pub/dist/freesurfer/dev/</a>                       | N/A                                                                                           |
| HD-BET                                              | <a href="https://github.com/MIC-DKFZ/HD-BET">https://github.com/MIC-DKFZ/HD-BET</a>                                                                         | N/A                                                                                           |
| MATLAB version 2022b                                | Mathworks                                                                                                                                                   | RRID: SCR_001622                                                                              |
| SPM12                                               | <a href="https://www.fil.ion.ucl.ac.uk/spm/software/spm12/">https://www.fil.ion.ucl.ac.uk/spm/software/spm12/</a>                                           | RRID: SCR_007037                                                                              |
| <b>Other</b>                                        |                                                                                                                                                             |                                                                                               |
| GPU RTX A6000 48 GB                                 | Nvidia Corp., Santa Clara, California.                                                                                                                      | N/A                                                                                           |

### EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

#### Asan Medical Center (AMC) dataset

We retrospectively collected the data of patients who visited the movement disorder clinic at the AMC from January 2005 to March 2022. Patients diagnosed with PD, MSA, PSP, ET, and other types of secondary parkinsonism were included. The United Kingdom PD Society Brain Bank criteria,<sup>57</sup> probable MSA criteria from the second consensus statement on the diagnosis of MSA,<sup>58</sup> probable PSP criteria from the Movement Disorder Society criteria for PSP,<sup>59</sup> and diagnostic criteria for ET<sup>60</sup> were utilized. For the diagnosis of PD, only patients whose <sup>18</sup>F-FP-CIT PET scans showed striatal uptake signal loss, visually confirmed by nuclear medicine specialists, were included. Although some patients with intact dopamine transporter binding may ultimately be diagnosed with PD, others may have different movement disorders,<sup>61</sup> potentially introducing false class labels and contaminating the model. The clinical diagnoses of secondary parkinsonism were based on the clinical presentation and imaging findings evaluated by movement specialists. To perform qualitative evaluations of our models, we also included the data of patients diagnosed as VP or DIP. Among patients with

normal PET findings, those with bilateral deep cerebral white matter hyperintensities or basal ganglia lesions on MRI were classified as having VP, whereas those with parkinsonism following dopamine-blocking or -depleting medications were classified as having DIP.

Only patients who undertook both  $^{18}\text{F}$ -FP-CIT PET and brain MRI scans at the AMC and who were assessed by the two designated movement specialists (SJC and SJ) were included. Exclusions were made for patients with PET and MRI scans over 5 years apart, significant PET image artifacts, ischemic striatum lesions on the PET, scans acquired post-deep brain stimulation surgery, or PET scans with slice thickness greater than 1.5 mm. Additionally, the time interval between the onset of parkinsonism and PET scan acquisition was calculated based on medical records from routine clinical visits.

### Soonchunhyang University Seoul Hospital (SCH) dataset

For the external validation of our trained model, we collected the  $^{18}\text{F}$ -FP-CIT PET scans from patients at the SCH movement disorder clinic between December 2016 and March 2024. The same diagnostic criteria were applied by a movement specialist (K-YK) to identify patients with PD, MSA, PSP, and ET. The same exclusion criteria were used, except that because all PET scans had a slice thickness of 3.4 mm, no images were excluded based on the slice thickness.

### Ethics

This study was approved by the institutional review board at the Asan Medical Center (IRB 2022-0614) and Soonchunhyang University Seoul Hospital, Seoul, South Korea (IRB 2024-01-014). Written informed consent from participants was waived by the ethics committees because of the retrospective nature of the study.

## METHOD DETAILS

### $^{18}\text{F}$ -FP-CIT PET image acquisition

#### Asan Medical Center (AMC)

N-(3-fluoropropyl)-2 $\beta$ -carboxymethoxy-3 $\beta$ -(4-iodophenyl) nortropane ( $^{18}\text{F}$ -FP-CIT) was synthesized according to the method previously described in literature.<sup>62</sup>  $^{18}\text{F}$ -FP-CIT PET scans were acquired utilizing two scanners, the Biograph TruePoint 40 (Bio40) and Vision 600 (Vision) (Siemens, Knoxville, TN, USA), 180 min after the intravenous administration of 185 MBq  $^{18}\text{F}$ -FP-CIT. These scanners achieve an in-plane spatial resolution of 2.0 mm full width at half maximum at the center of the field of view. A low-dose brain computed tomography (CT) scan, typically at 120 kVp and 20 mAs with a slice thickness of 1.5 mm, was performed immediately prior to the PET imaging to assist in image fusion and attenuation correction. The PET scanning duration was ten minutes for Bio40 and seven minutes for Vision, both conducted in the three-dimensional mode. For image reconstruction, we employed the TrueX algorithm for Bio40 and also time-of-flight implementation for Vision by using all-pass filters across  $336 \times 336$  (Bio40) and  $440 \times 440$  (Vision) matrices.

#### Soonchunhyang University Seoul Hospital (SCH)

$^{18}\text{F}$ -FP-CIT PET scans were completed using a single scanner, Biograph 128 mCT (Bio128; Siemens, Erlangen, Germany) with an in-plane spatial resolution of 2.0 mm full width at half maximum from the center of the field of view. Prior to PET image acquisition, CT scans were obtained using the CARE Dose4D system at 120 kVp and 100 mAs with a slice thickness of 3.0 mm and an increment of 1.5 mm. Two hours after the intravenous injection of 185 MBq  $^{18}\text{F}$ -FP-CIT, images were acquired for 10 min. Attenuation-corrected PET images were reconstructed using the TrueX algorithm and time-of-flight implementation, with Gaussian filters across  $400 \times 400$  matrices. Before being uploaded to the hospital's Picture Archiving and Communication System (PACS), the images were processed to have a slice thickness of 3.4 mm, an inter-slice distance of 3.0 mm, and voxel intensities within the range of [0, 10-13].

### T1 MRI image acquisition

#### Asan Medical Center (AMC)

All brain MRI T1 images were acquired in the axial orientation with parameters presented as median [IQR] values, reflecting the parameter variations resulting from the retrospective design of the study: TR 450.0 ms [6.5, 450.0], TE 8.0 ms [3.0, 10.8], flip angle 69.0° [9.0, 70.0], x-, y-voxel spacing 0.5  $\times$  0.5 mm [0.4, 0.5], slice thickness 5.0 mm [2.0, 5.0], and spacing between slices 7.0 mm [2.0, 7.5]. A total of nine scanners with magnetic fields of 1.5 T and 3.0 T from three different vendors were used: Achieva 1.5T and 3.0T, Ingenia 3.0T (Philips), Magnetom Avanto 1.5T, Vision 1.5 T, Skyra 3.0T, Vida 3.0T (Siemens), Signa Architect 3.0T, and Signa CV/i 1.5T (General Electric). Because the T1 images were used solely for PET image registration to the Montreal Neurological Institute template and no analysis was performed on the T1 images themselves, we imposed no restrictions on the MRI acquisition settings.

#### Soonchunhyang University Seoul Hospital

All T1 images were obtained in the axial orientation with the following parameters expressed as median [IQR] values: TR 2020.0 ms [7.5, 2275.1], TE 3.5 ms [2.6, 21.3], flip angle 9.0° [8.0, 111.0], x-, y-voxel spacing 0.4  $\times$  0.4 mm [0.4, 0.4], slice thickness 2.0 mm [2.0, 4.0], and spacing between slices 5.0 mm [5.0, 6.4]. The scanners used for image acquisition are as follows: Ingenia CX 3T (Philips), Sonata 1.5T, Magnetom Skyra 3.0T, Vida 3.0T (Siemens), Discovery MR750w 3T, and Signa Pioneer 3T (General Electric).

### <sup>18</sup>F-FP-CIT PET and T1 MRI image preprocessing

We first converted the PET and MRI images from the DICOM to NIfTI format and removed the skulls using SynthStrip<sup>63</sup> for PET and HD-BET<sup>64</sup> for MRI images. After bias field correction, the T1 images were co-registered to the corresponding PET images. Since the time interval between T1 and PET acquisition ranged from 0 to 1,811 days, potentially affecting co-registration accuracy, we compared normalized mutual information values between 100 image pairs acquired on the same day and the 100 pairs with the longest intervals. Both groups had a mean of 1.34 and a standard deviation of 0.02, with an independent t test yielding a *p*-value of 0.43, indicating that co-registration performance did not significantly vary with time intervals. T1 images were then normalized to the MNI template using SPM12 software (Statistical Parametric Mapping, the Wellcome Trust Centre for Neuroimaging) on MATLAB 2022b. Next, we applied the inverse deformation map obtained from the previous normalization step to the following ROIs from the brain atlases defined in the same MNI space: ventral striatum, caudate, putamen, limbic, executive, and sensorimotor striatum from the Oxford-GSK-Imanova structural<sup>65</sup> and connectivity<sup>66</sup> striatal atlases, and calcarine cortex from the Automated Anatomical Labelling Atlas 3.<sup>67</sup> Accordingly, we obtained the ROI mask on the native space of the co-registered T1 and PET images. The mean uptake values of the ROIs were used to calculate the SNBR using the following equation: (mean uptake value of bilateral ROIs – mean uptake value of bilateral calcarine cortices)/(mean uptake value of bilateral calcarine cortices). For training and testing of the model, the PET images were intensity normalized by dividing by the mean uptake value of the bilateral calcarine cortices of each image. Subsequently, all images were normalized by clipping and dividing by the 99th percentile intensity value of all voxels across all images. Finally, we cropped the PET images to a uniform size of 192 × 192 × 96, centered on the nonzero region of each image.

### DWT

Discrete Harr Wavelet transform<sup>68</sup> was applied to the input images, producing eight channels of half-resolution wavelet coefficients (*lll*, *hll*, *lhl*, *llh*, *lhh*, *hlh*, *hhl*, *hhh*). After channel-wise concatenation, the input size changed from 1 × 192 × 192 × 96 (*channel* × *height* × *width* × *depth*) to 8 × 96 × 96 × 48. The model outputs, consisting of wavelet coefficients, were converted back into images using inverse DWT.

### Diffusion model architecture and implementation

Two model architectures (with and without a semantic encoder) and input types (images with and without DWT) resulted in four diffusion models: DDAE, HDAE, WDDAE, and HWDAAE (Figures 1B and S17).

DDAE<sup>30</sup> used the Denoising Diffusion Probabilistic Model (DDPM)<sup>69</sup> U-Net architecture, and the model performance was tested on 2D, small-sized (32 × 32 or 64 × 64), natural images. For the downstream tasks, the latent space from the U-Net encoder was processed with global average pooling and a linear classifier for the prediction. However, we modified the DDAE to accept 3D inputs. Additionally, DDPM was replaced by Simple Diffusion,<sup>56</sup> and the min-SNR technique<sup>70</sup> was added for more stable and faster convergence of input images of high resolution. For the implementation, we used the shifted cosine scheduler,<sup>56</sup> 2 batches per GPU, a learning rate (LR) of 2e-5, maximum iterations of 400,000 for DDAE and 650,000 for WDDAE, and the AdamW optimizer<sup>71</sup> (betas (0.9, 0.999), weight decay 0.05). Data augmentation techniques including rotation, flip, translation, Gaussian noise, and scaling were applied with a probability of 0.1. Lastly, the latent spaces of our DDAE and WDDAE model resulted in a 512-sized vector. The training took 7 days for both DDAE and WDDAE.

HDAE consists of a conditional Denoising Diffusion Implicit Model (DDIM)<sup>72</sup> and semantic encoder. The semantic encoder extracts semantic image features from the input images, which are then used as conditions for the DDIM U-Net. Importantly, these image features or latent vectors are fed into the corresponding DDIM encoder and decoder layers hierarchically, with low-level latent vectors to the low-level layers. Similar to DDAE, we changed the model architecture to accept 3D inputs and replaced DDIM with DDPM. Further, we used the linear beta noise scheduler<sup>69</sup> with 1,000 timesteps, 2 batched per GPU, cosine decay LR scheduler<sup>73</sup> (base 4e-5, end LR 2e-5), and a maximum of 300 epochs for HDAE and 500 epochs for HWDAAE, requiring 8 and 15 days, respectively. The resulting latent vectors, sized 512 × 6 for HDAE and 512 × 5 for HWDAAE, were concatenated for the downstream tasks.

### Other model architectures and implementation

We compared the diffusion models to two other types of self-supervised learning (SSL) models: Generative Adversarial Network (GAN) and Masked Autoencoder (MAE).

Among the GAN models, we selected StyleGAN2<sup>74</sup> and combined its generator with the Pixel2Style2Pixel (P2S2P)<sup>39</sup> encoder to extract the latent vectors of the input images (Figure S18). We modified the StyleGAN2 architecture to accept the 3D images as inputs and trained it using cascaded learning, generating images at progressively higher resolutions: 48 × 48 × 24, 96 × 96 × 48, and final full resolution of 192 × 192 × 96. The initial input size of the generator was 6 × 6 × 3. Throughout the cascaded learning process, we reduced the LR of the generator from 1e-5 to 5e-6 and finally to 1e-6, whereas the LR of the discriminator was reduced from 1.5e-5 to 7.5e-6 and then to 1.5e-6. We progressively adjusted the batch size per GPU from 20 to 5 and then to 2, and the number of epochs from 30 to 35 and finally to 40, with training durations of 2, 3, and 13 days, respectively. The Adam optimizer (betas (0.5, 0.9)) was used. The generator and discriminator were each composed of 6 blocks, and hinge loss<sup>75</sup> was used. After training StyleGAN2, we retained the generator and kept it frozen while training the P2S2P encoder. The encoder, which consisted of 24 squeeze-excitation blocks for feature extraction and 12 map2style blocks for latent vector generation, resulted in latent vectors (*N* + space) sized

$512 \times 12$ . L2 loss was used for training the encoder, with a LR of  $1e-6$  and batch size of two. The training lasted for 110 epochs over 4 days. We utilized the AdamW optimizer (betas (0.9, 0.999)) with a weight decay of  $5e-4$ .

For the MAE, we chose Simple Masked Image Modeling (SimMIM)<sup>37</sup> and Disruptive AutoEncoder (DisAE)<sup>38</sup> (Figures S19–S20). SimMIM learns image representations by reconstructing randomly masked areas of the input images. We used Swin-ViT<sup>76</sup> as the encoder with L1 loss, setting the mask patch size to  $32 \times 32 \times 32$  with a masking ratio of 0.6. SimMIM was trained with a batch size of 2, using a cosine decay LR scheduler with a base LR of  $5e-4$  and an end LR of  $5e-6$ . The training spanned 300 epochs and lasted 6 days. We used the AdamW optimizer (betas (0.9, 0.999)) with a weight decay of 0.05. DisAE shared the same encoder architecture and masking ratio as SimMIM, but masked random channels instead of patches. The model downsampled the input images by four scales and added Gaussian noise (variance 0.1), after which the channels were randomly masked. DisAE learned the latent representations by simultaneously reconstructing the masked channels, denoising, and performing super-resolution. L1 loss was used. DisAE was trained with a batch size of two, using a cosine decay LR scheduler with a base LR of  $4e-4$  and an end LR of  $1e-5$ . We trained the model for 300 epochs over the course of 7 days. The AdamW optimizer (betas (0.9, 0.999)) with a weight decay of 0.05 was used. The encoders of SimMIM and DisAE were used for the downstream tasks.

### Hyperparameter selection: Maximum number of epochs for upstream models

During upstream training, we validated the reconstructed or generated images at regular intervals to confirm convergence. For HDAE, DisAE, and SimMIM, we selected 300 epochs, as stable reconstruction was observed after 200 epochs. For P2S2P, we chose 110 epochs based on stable reconstruction after 65 epochs. Models without DWT, such as DDAE and HDAE, converged faster, so we scaled training epochs by a factor of 5/3 for DWT models. DDAE and HDAE achieved stable reconstruction at 400,000 iterations and 300 epochs, respectively. For DWT models, we extended training to 650,000 iterations for WDDAE and 500 epochs for HWDAAE.

### Adaptation to downstream tasks

We applied three training strategies—LP, SC, and FT—to the downstream tasks. All these methods share the same model architecture, wherein a trainable linear classifier was added to the last layer of the pretrained upstream models. We used 3D ResNet34 as the reference model, applying its best hyperparameter settings to adapt all upstream models.

In LP, the pretrained encoders of the upstream models were kept frozen, and a fully connected, trainable linear classifier was added. We used a base LR of  $1e-3$  with a polynomial LR scheduler, warm-up epoch of 10, and a maximum epoch of 50, 50, and 30 for the EP, PMP, and SOY tasks, respectively.

In FT, the pre-trained weights were updated through optimization. We used a cosine decay LR scheduler<sup>73</sup> with a base LR of  $2e-5$  for PMP and  $1e-4$  for SOY, with a maximum epoch of 40 and 30 and a warm-up epoch of 20 and 10, respectively. SC used randomly initialized weights instead of pretrained encoder weights. Except for SC of the PMP task, where we used a maximum epoch of 60, SC shared identical training settings with FT.

The following additional implementation details apply to all three strategies. We used the same optimizer and data augmentation settings as in the diffusion upstream models. Cross-entropy loss with a label smoothing of 0.1 was employed for the classification, whereas Huber loss was applied for regression. For the EP, PMP, and SOY tasks, we selected the best epochs based on the lowest cross-entropy loss, the highest micro-averaged AUC, and the lowest Huber loss in their respective validation sets. Owing to computational resource limitations, we set the batch size to three for SimMIM, DisAE, and P2S2P, and to six for the diffusion models. For the DDAE and WDDAE models, minimal noise was added to the input images with a timestep of 0.002.

For model explainability and visualization, we generated Guided GradCAM and occlusion sensitivity maps for selected AMC1 test set images. These included images for which the following models correctly predicted the class in classification tasks or achieved low mean absolute error in the regression task: HWDAAE with LP for the EP task, and HWDAAE with FT for the PMP and SOY tasks. The maps were min-max normalized to a range of 0 to 100 per image for visualization purposes. Occlusion sensitivity is shown for the EP task, both Guided GradCAM and occlusion sensitivity for the PMP task, and Guided GradCAM for the SOY task.

### Latent space manipulation

We first calculated the mean of 100 randomly selected latent vectors from each of the following groups: patients with normal PET findings and diagnosed with ET, VP, DIP, or PD with onset years  $<5$  y, and those with onset years  $\geq 5$  y (Figure 7D). The means were subtracted from each other and normalized to generate four direction vectors connecting the three groups. The entire process was repeated five times, and the results were averaged to obtain the final direction vectors. In Figure 7F, we added the direction vectors, multiplied by gradually increasing factors, to the latent vectors of input images.

### Computational hardware and software

All implementations were conducted using Python 3.10.13 with the PyTorch framework (version 2.1.2, CUDA 12.1). We used accelerate (v0.26.1), ema-pytorch (v0.3.0), xformers (v0.0.23.post1), and monai (v1.3.0) to train the upstream models. PET image processing was performed using numpy (v1.26.2) and SimpleITK (v2.2.0), DWT with PyWavelets (v1.5.0), statistical analysis with SciPy (v1.14.1), normalized mutual information calculation with Scikit-image (v0.24.0), plotting with Matplotlib (v3.9.2), Guided GradCAM with Captum (v0.7.0), and occlusion sensitivity with monai (v1.3.0). The evaluation metrics were calculated using Scikit-learn (v1.4.0).

We referred to publicly available source codes for model implementation. For DDAE, WDDAE, HDAE, and HWDAE, we used the MONAI Generative Models repository (<https://github.com/Project-MONAI/GenerativeModels/tree/main/generative>)<sup>77</sup> and a diffusion model implementation (<https://github.com/lucidrains/denoising-diffusion-pytorch>). SimMIM and DisAE were implemented based on the official DisAE repository (<https://github.com/Project-MONAI/research-contributions/tree/main/DAE>). For 3D style-GAN2 and P2S2P, we adapted existing 2D model repositories (<https://github.com/lucidrains/stylegan2-pytorch>, <https://github.com/eladrich/pixel2style2pixel>) to support 3D inputs. All experiments, including the upstream and downstream, were run on a single NVIDIA RTX A6000 48 GB GPU.

## QUANTIFICATION AND STATISTICAL ANALYSIS

### Evaluation metrics

In the EP task, the AUC was evaluated, and in the PMP task, the AP (macro- and micro-averaged), AUC, and weighted F1 were evaluated. Macro-average is the arithmetic mean of the class-wise metrics, whereas micro-average is the arithmetic mean calculated across the inputs of all classes to assess the contribution of each input to the overall metric. The weighted F1 score is the weighted mean of the F1 score of each class, which is calculated as the harmonic mean of precision and recall. The balanced accuracy is the arithmetic mean of sensitivity and specificity.

In the SOY task, we mainly reported  $R^2$  and CCC.  $R^2$  represents the proportion of variance in the dependent variable and measures the goodness of fit for a regression model.  $R^2$  is defined as

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}},$$

where  $SS_{res}$  denotes the sum of the squared residuals (differences between true and predicted values), and  $SS_{tot}$  represents the total sum of the squares (difference between true values and their mean). CCC measures the agreement between the true and predicted values as

$$CCC = \frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2},$$

where  $x$  is the true value,  $y$  the predicted value,  $\rho_{xy}$  the Pearson correlation between  $x$  and  $y$ ,  $\sigma_x^2$  and  $\sigma_y^2$  the variances, and  $\mu_x$  and  $\mu_y$  the means. The evaluation metrics for each task are reported in the Results subsections and visualized in Figure 3 (EP task), 4 (PMP task), and 5 and 6 (SOY task).

### Data fold splitting and statistical analysis

We trained all models using only the AMC1 training set and evaluated them on one internal (AMC1) and two external (AMC2 and SCH) test sets. The entire AMC1 dataset was randomly split into five folds, and one fold was selected as the internal test set. The remaining four folds were further divided into five folds again for 5CV. The models were ranked based on the mean performance of the five folds from 5CV and tested on the internal and external test sets. We compared the models using a paired t test on the five-fold results of each model, with the resulting  $p$ -values reported in the Results subsections. We applied an independent t test for the Pearson's correlation coefficients in Figure 5C, d, and 6c. All tests were conducted with a two-sided significance level of 0.05. The 95% CIs for the performance metrics were defined as the mean  $\pm$  (1.96  $\times$  standard error), where the standard error was calculated by dividing the standard deviation of the five folds by  $\sqrt{5}$ .