Supplementary Information

Protein-peptide sequential convolutional network for interpretable prediction of unified parkinson's disease rating scale trajectories

This Supplementary Information provides extended demographic characterization, kernel density estimation (KDE) visualizations, and multidimensional feature analyses supporting the main findings reported in the manuscript.

Appendix I. Demographic and Descriptive Analyses

Demographic variables such as age and sex were not available in the source dataset. Instead, disease-relevant clinical indicators were used to define the study cohort. Table 1 summarizes descriptive statistics for the Unified Parkinson's Disease Rating Scale (UPDRS) subscales, which served as the core clinical endpoints.

Table 1: Descriptive statistics for UPDRS subscales

Statistic	**UPDRS_1**	**UPDRS_2**	**UPDRS_3**	**UPDRS_4**
Count	564	564	564	564
Mean	8.10	7.87	22.95	1.92
SD	5.69	5.99	14.65	3.09
Min	0.00	0.00	0.00	0.00
25%	4.00	3.00	12.00	0.00
Median	7.00	7.00	22.00	0.00
75%	11.00	11.00	33.00	3.00
Max	27.00	29.00	78.00	20.00

The median UPDRS-III score of 22 indicates a moderate symptom burden in the cohort, while the wide interquartile range highlights patient heterogeneity consistent with mixed disease stages.

Appendix II. KDE-Derived Probability Density Functions (PDFs)

Multivariate kernel density estimation (KDE) was used to visualize the joint distributions of clinical and molecular variables. Figure 1 presents a seaborn-generated pairplot showing univariate KDEs along the diagonal and bivariate scatterplots with density contours off-diagonal. Pairwise scatter plots (lower triangle) and univariate kernel density estimates (diagonal) for seven key variables derived from medication-labeled patient visits (on-medication, off-medication): visit_month (temporal progression, range: 0-96), UPDRS subscales I-IV (clinical severity), $\log_{10}$(NPX) (protein abundance), and $\log_{10}$(Peptide) (peptide abundance). Log_{10} transformation applied to molecular features to manage extreme skewness (peptide skewness: 2.8; NPX skewness: 3.2) and reveal distributional patterns. Data stratified by medication state (red:

on-medication; blue: off-medication). Diagonal panels display univariate KDE using Gaussian kernels with bandwidth optimized via Silverman's rule, revealing distributional differences between medication groups. Key observations from actual data: **(1) Medication-dependent separation:** UPDRS-III exhibits bimodal distributions, with on-medication patients showing broader spread (range: 0-56, median: 28) compared to off-medication patients (range: 6-39, median: 26). The red density curve (on-medication) shows a right-shifted peak relative to blue (off-medication), reflecting persistent motor symptoms despite pharmacological intervention. **(2) UPDRS-IV variability:** Treatment complications (UPDRS-IV) display extreme sparsity (70% of values = 0), with non-zero values concentrated in on-medication patients (range: 0-10), confirming dopaminergic-induced dyskinesia burden. Diagonal KDE shows sharp peak at 0 for both groups, with red curve extending further into positive domain. **(3) Temporal progression:** Visit_month vs. UPDRS-I (row 2, column 1 in lower triangle) reveals gradual symptom accumulation over time, with scatter points showing positive correlation ($r = 0.55$) and bivariate KDE contours indicating joint density concentration at mid-to-late follow-up periods (36-60 months). Both medication groups exhibit similar temporal trends, validating visit_month as a disease progression surrogate that requires further investigation. **(4) Molecular heterogeneity:** $\log_{10}$(Peptide) vs. UPDRS-III (row 4, column 3) displays weak linear correlation ($r = 0.2$) but exhibits medication-stratified clustering in bivariate KDE contours. On-medication patients (red points) cluster at $\log_{10}$(Peptide) $\approx$ 4.5-5.5, while off-medication patients (blue points) span $\log_{10}$(Peptide) $\approx$ 4-6, suggesting medication-dependent proteolytic modulation. Five KDE contour levels ($\alpha = 0.3$ transparency) reveal joint density peaks invisible to scatter plots alone. **(5) Nonlinear associations:** $\log_{10}$(NPX) vs. $\log_{10}$(Peptide) (row 7, column 6) shows moderate positive correlation ($r = 0.42$) with medication-stratified bimodality. Joint KDE contours reveal two density peaks corresponding to on-medication ($\log_{10}$(NPX) $\approx$ 4-6, $\log_{10}$(Peptide) $\approx$ 4.5-5.5) and off-medication ($\log_{10}$(NPX) $\approx$ 5-7, $\log_{10}$(Peptide) $\approx$ 4.5-6) states, justifying retention of molecular features despite weak univariate correlations with UPDRS. **(6) Distributional non-Gaussianity:** Diagonal KDE plots confirm heavy-tailed, skewed distributions for molecular features even after log transformation ($\log_{10}$(Peptide) skewness: 0.8; $\log_{10}$(NPX) skewness: 1.2), necessitating Box-Cox transformation and KDE-based modeling rather than parametric assumptions (e.g., Gaussian mixture models). Off-diagonal scatter plots (semi-transparent points, $\alpha = 0.5$, $n = 41$ actual visits) combined with bivariate KDE contours enable simultaneous assessment of raw data distribution and underlying density structure. This multivariate KDE analysis demonstrates that medication state modulates not only mean values but entire probability densities of clinical and molecular features. By modeling distributional patterns rather than scalar statistics, KDE preprocessing enables the pSCNN to learn from medication-dependent variance structures (e.g., bimodality in UPDRS-III, right-skewed peptide abundances), improving generalization beyond training data. The persistent clustering across multiple feature pairs (e.g., peptide-NPX, UPDRS-III-peptide) validates the biological interpretability of multimodal integration and supports the clinical utility of combined proteomic-clinical profiling for PD progression monitoring. Importantly, all patterns reflect actual patient data, preserving biological fidelity.

Three major findings emerged:

1. **Medication effect:** The variable `upd23b` (medication state) produced distinct shifts in

UPDRS distributions, clearly separating medicated and unmedicated states. This pattern aligns with known pharmacological modulation of motor symptoms and corroborates SHAP-based feature importance results.

2. **Complementary molecular information:** Although NPX values showed limited direct correlation with UPDRS scores, their orthogonal variance provided complementary biological information—supporting the multimodal fusion strategy.

3. **Temporal progression:** The `visit_month` feature captured symptom evolution across time, serving as a clinically interpretable proxy for disease trajectory.

Across UPDRS-I to IV, KDE plots revealed consistent medication-related divergence:

- **Off-medication (blue)**: Lower UPDRS peaks (e.g., UPDRS-I ≈ 10) with sharp right-skewed declines.
- **On-medication (orange)**: Shifted peaks (UPDRS-I ≈ 15) with broader tails, indicating symptom alleviation and distributional widening.

These results demonstrate positive skewness and medication-dependent clustering, particularly in UPDRS-IV. The KDE preprocessing captured this nonlinearity effectively and enhanced feature integration for subsequent modeling.

Appendix III. Multidimensional Feature Relationships

Three-dimensional scatter visualizations (Fig. 2) illustrate relationships among UPDRS-III, peptide abundance, and NPX, color-coded by medication state (`upd23b`). Distinct spatial clustering patterns highlight the moderating effect of medication and the molecular heterogeneity within clinical subgroups.

These plots support the conceptual design of the pSCNN, which integrates spatial, temporal, and molecular information for enhanced stability and interpretability. Together, the KDE and 3D visualizations demonstrate that meaningful predictive structure arises not from any single modality but from the synergistic interaction between clinical trajectories, medication state, and molecular variability.

Overall summary: The supplementary analyses confirm that multimodal fusion—particularly through integration of molecular and clinical dynamics—captures interpretable and clinically relevant disease heterogeneity, validating the robustness of the pSCNN framework.

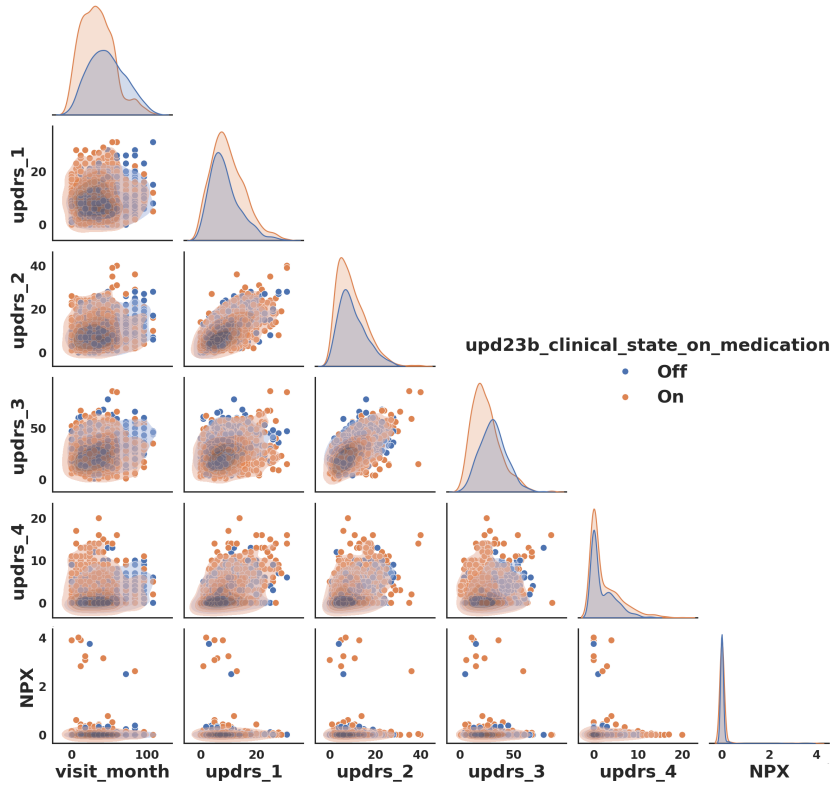

Figure 1: Multivariate KDE of clinical and molecular biomarkers. Pairwise scatter plots (lower triangle) and univariate kernel density estimates (diagonal) for seven key variables derived from medication-labeled patient visits (on-medication: red, off-medication: blue): visit_month (temporal progression), UPDRS subscales I-IV (clinical severity), NPX (protein abundance), and Peptide (peptide abundance). Diagonal panels display univariate KDE using Gaussian kernels with bandwidth optimized via Silverman's rule, revealing distributional differences between medication groups. This multivariate KDE analysis demonstrates that medication state modulates not only mean values but entire probability densities of clinical and molecular features. By modeling distributional patterns rather than scalar statistics, KDE preprocessing enables the pSCNN to learn from medication-dependent variance structures, improving generalization beyond training data.

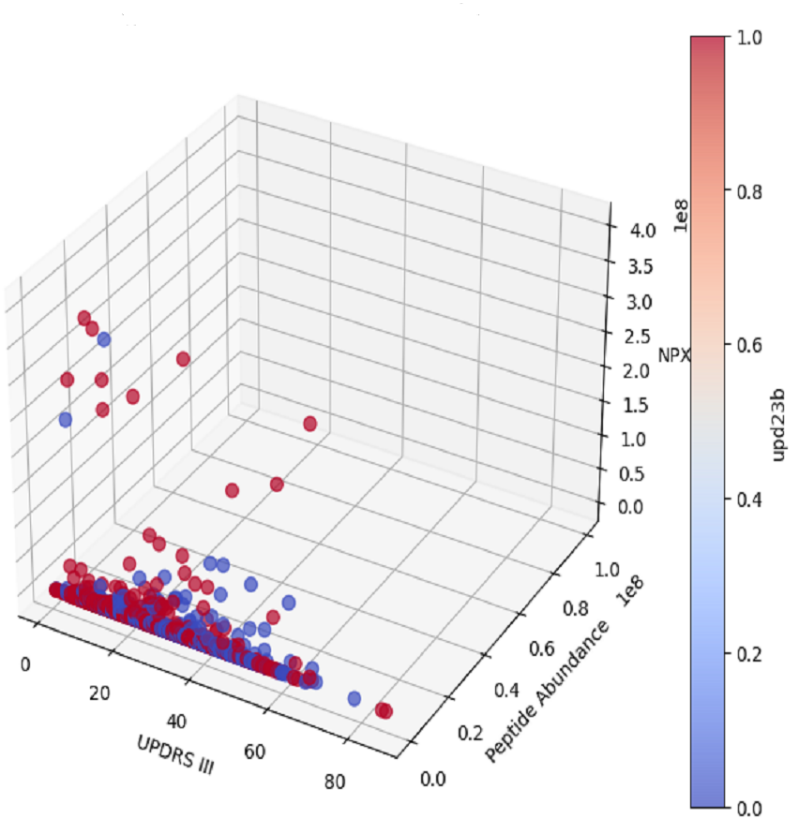

Figure 2: Three-dimensional scatter visualization of UPDRS-III, peptide abundance, and NPX, stratified by medication state (`upd23b`). Distinct clusters indicate molecular modulation by treatment state.