## A Computing Infrastructure

Data processing and experiments are run in a high performance cluster using Linux and Slurm.

**Data processing** Data processing is conducted using 40$\times$ Intel(R) Xeon(R) Gold 6248 CPU @ 2.50GHz processors and 200Go of RAM totalling 15 hours.

**Transformer** Transformer experiments are conducted using A100 gpus, 8$\times$ EPYC 7543 Milan AMD processors and 64Go of RAM totalling 2000 hours.

**MLP/XGBoost** MLP and XGBoost experiments are conducted using V100 gpus, 10$\times$ Intel(R) Xeon(R) Gold 6248 CPU @ 2.50GHz processors and 40Go of RAM totalling 1500 hours.

## B Hyper-parameter Search Protocol

Across all three model families—**Transformer**, **Multi-Layer Perceptron (MLP)**, and **XGBoost**—we perform a three-stage grid search. At each stage, we sweep the hyper-parameters listed in Tables 1–2–3 over all Cartesian products, fixing all other settings to the best configuration from the previous stage. Validation uses mean absolute error (MAE) on the delay prediction targets, with a variance-aware criterion: if two candidates have similar MAE, the one with lower training variance is selected. For Transformers and MLPs, we apply early stopping on the validation MAE with patience equal to $0.25$ of the maximum epoch count (e.g., 20 epochs for an 80-epoch run).

After tuning, the best configuration is retrained on the union of train and validation data and evaluated on the test split using ten random seeds (0–9), with all randomness controlled via PyTorch Lightning's global seeding. In all tables, a dash (—) indicates that the field is not applicable to the corresponding method.

**Transformer.** The Transformer sweep supports *Regression*, *Behavioural Cloning (BC)* and *Drift-Corrected Imitation Learning (DCIL)*. All variants share the optimiser and architectural defaults listed at the top of Table 1. Phase 1 explores model dimension, number of layers and learning rate, Phase 2 fine-tunes dropout, batch size and learning rate, and Phase 3 (DCIL only) searches over trajectory length, $\alpha$ and $\beta$.

**MLP.** The MLP sweep supports *Regression*, *Behavioural Cloning (BC)* and *Drift-Corrected Imitation Learning (DCIL)*. All variants share the optimiser and architectural defaults listed at the top of Table 2. Phase 1 explores hidden dimensions sizes and learning rate, Phase 2 fine-tunes batch size and learning rate, and Phase 3 (DCIL only) searches over trajectory length, $\alpha$ and $\beta$.

**XGBoost.** The XGBoost sweep supports *Regression* and *Behavioural Cloning (BC)*. All variants share the optimiser and architectural defaults listed at the top of Table 3. Phase 1 explores gamma, max depth, min child weight, subsample and colsample by tree and Phase 2 fine-tunes learning rate, number of estimators, reg $\alpha$ and reg $\lambda$.

## C Final configurations

The final configurations are given in Tables 4-5-6.

## Ethics Statement

We use railway operational data released under a CC0 public-domain license. The dataset contains only non-personal operational information (train times, stations, delays) and no data about individual passengers or staff. Our use of the data therefore complies with the provider's license and does not raise additional privacy concerns.

To ensure reproducibility and support further research, we release all scripts needed to reproduce our experiments; the corresponding GitHub link is provided in the Links section at the beginning of the paper.

## Acknowledgements

This work received financial support from SNCF through the research chair "AI and optimization for mobility" with École Polytechnique. This work was granted access to the HPC resources of IDRIS under the allocation AD011015656R1 made by GENCI. Finally, we thank Clément Mantoux and Mathis Le Bail for helpful discussions and feedback on this work.

Table 1. Default hyper-parameters and grid-search ranges for Transformer-based methods.

| Phase | Hyper-parameter | Regression | BC | DCIL |
|---|---|---|---|---|
| *Defaults* | | | | |
| | Optimiser | AdamW (default $\alpha,\ \beta$), weight decay 0.01 | | |
| | Activation | ReLU | | |
| | $d_{\mathrm{ff}}$ | $4\,d_{\mathrm{model}}$ | | |
| | Loss | L2 | Cross-entropy | Cross-entropy |
| | Training epochs | 80 | 80 | 600 |
| | Batch size | 64 | 64 | 128 |
| | Heads $n_{\mathrm{head}}$ | 8 | | |
| | Dropout | 0.2 | | |
| | Replay buffer | — | — | 60,000 |
| | Synthetic samples/epoch | — | — | 20,000 |
| | Trajectory length | — | — | 10 |
| | $\alpha$ | — | — | 0.5 |
| | $\beta$ | — | — | 2 |
| *Phase 1* | | | | |
| | $d_{\mathrm{model}}$ | {128, 256, 512, 1024} | | |
| | Layers | {4, 6} | | |
| | Learning rate | {1e-4, 5e-5, 1e-5} | | |
| *Phase 2* | | | | |
| | Batch size | {64, 128, 256} | | |
| | Dropout | {0.05, 0.10, 0.20} | | |
| | Learning rate | {1e-4, 5e-5, 2e-5} | {3e-4, 1e-4, 5e-5} | same as BC |
| *Phase 3 (DCIL only)* | | | | |
| | Trajectory length | — | — | {5, 10, 15, 20} |
| | $\alpha$ | — | — | {0.5, 0.8} |
| | $\beta$ | — | — | {1, 2, 3, 4} |

Table 2. Default hyper-parameters and grid-search ranges for MLP-based methods.

| Phase | Hyper-parameter | Regression | BC | DCIL |
|---|---|---|---|---|
| *Defaults* | | | | |
| | Optimiser | AdamW (default $\alpha,\ \beta$), weight decay 0.001 | | |
| | Activation | ReLU | | |
| | Loss | L2 | Cross-entropy | Cross-entropy |
| | Training epochs | 100 | 160 | 1500 |
| | Batch size | 32 | | |
| | Dropout | 0.0 | | |
| | Replay buffer | — | — | 30,000 |
| | Synthetic samples/epoch | — | — | 10,000 |
| | Trajectory length | — | — | 10 |
| | $\alpha$ | — | — | 0.5 |
| | $\beta$ | — | — | 2 |
| *Phase 1* | | | | |
| | Hidden Dims | {(64, 128, 256, 128, 64) to (256, 512, 1024, 2048, 1024, 512, 256)} (8 configs) | | |
| | Learning rate | {1e-3, 5e-4, 1e-4} | | |
| *Phase 2* | | | | |
| | Batch size | {16 32 64 128 256} | | |
| | Learning rate | {3e-4, 1e-4, 5e-5, 3e-5} | {3e-3, 1e-3, 3e-4, 1e-4} | same as Regression |
| *Phase 3 (DCIL only)* | | | | |
| | Trajectory length | — | — | {5, 10, 15, 20} |
| | $\alpha$ | — | — | {0.5, 0.8} |
| | $\beta$ | — | — | {1, 2, 3, 4} |

Table 3. Default hyper-parameters and grid-search ranges for XGBoost-based methods.

| Phase | Hyper-parameter | Regression | BC |
|---|---|---|---|
| *Defaults* | | | |
| | Loss | L2 | Softprob |
| | # Estimators | 400 | |
| | Learning Rate | 0.1 | |
| | Reg $\alpha$ | 0 | |
| | Reg $\lambda$ | 1 | |
| *Phase 1* | | | |
| | $\gamma$ | $\{0, 1, 5\}$ | |
| | Max Depth | $\{4, 6, 9, 13\}$ | |
| | Min Child Weight | $\{1, 5\}$ | |
| | Subsample | $\{0.6, 0.8, 1.0\}$ | |
| | Colsample by Tree | $\{0.5\ 0.8\}$ | |
| *Phase 2* | | | |
| | Learning Rate | $\{0.03, 0.04, 0.06, 0.07, 0.09, 0.1\}$ | |
| | # Estimators | $\{200, 400, 800, 1000, 1600, 2000\}$ | |
| | Reg $\alpha$ | $\{0, 0.3, 1\}$ | |
| | Reg $\lambda$ | $\{0, 1, 5\}$ | |

Table 4. Best hyper-parameters for Transformer models.

| Hyper-parameter | Regression | BC | DCIL |
|---|---|---|---|
| $d_{\text{model}}$ | 512 | 512 | 512 |
| Layers | 4 | 6 | 4 |
| Learning rate | 5e-5 | 5e-5 | 1e-4 |
| Batch size | 64 | 128 | 64 |
| Dropout | 0.2 | 0.2 | 0.2 |
| Trajectory length | — | — | 10 |
| $\alpha$ | — | — | 0.8 |
| $\beta$ | — | — | 2 |

Table 5. Best hyper-parameters for MLP models.

| Hyper-parameter | Regression | BC | DCIL |
|---|---|---|---|
| Hidden dims | (512, 1024, 2048, 1024, 512) | (128, 256, 512, 1024, 512, 256, 128) | (256, 512, 1024, 512, 256) |
| Learning rate | 1e-4 | 1e-3 | 5e-5 |
| Batch size | 32 | 32 | 16 |
| Trajectory length | — | — | 5 |
| $\alpha$ | — | — | 0.5 |
| $\beta$ | — | — | 1 |

Table 6. Best hyper-parameters for XGBoost models.

| Hyper-parameter | Regression | BC |
|---|---|---|
| $\gamma$ | 0 | 1 |
| Max Depth | 13 | 13 |
| Min Child Weight | 5.0 | 5.0 |
| Subsample | 1.0 | 1.0 |
| Colsample by Tree | 0.8 | 0.8 |
| Learning Rate | 0.03 | 0.04 |
| # Estimators | 2000 | 1600 |
| Reg $\alpha$ | 1.0 | 0.3 |
| Reg $\lambda$ | 5.0 | 1.0 |