Research Square

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

PhenoGenX: A Dual-Engine, Data-Driven Platform for HIV-1 Drug Resistance Interpretation Integrating Ensemble Machine Learning and Rule-Based Algorithms

Yimam Getaneh
yimamgetaneh@gmail.com
Ethiopian Public Health Institute, Addis Ababa
Belete Woldesemayat
Ethiopian Public Health Institute, Addis Ababa
Kidist Zealiyas
Ethiopian Public Health Institute, Addis Ababa
Ghion Mengistu
The World Health Organization
Minilik Demissie
Ethiopian Public Health Institute, Addis Ababa
Zelalem Messele
Cliniton Health Access Initiative
Gemechu Leta
Ethiopian Public Health Institute, Addis Ababa
Yenew Kebede
Africa Centers for Disease Control and Prevention
Getachew Tolera
Ethiopian Public Health Institute, Addis Ababa
Lingjie Liao
Chinese Center For Disease Control and Prevention
Yiming Shao
Chinese Center For Disease Control and Prevention

Article

Keywords: HIV-1 drug resistance; Machine Learning, Genotype-Phenotype prediction, Ensemble Learning, Clinical Decision Support

Posted Date: March 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9056343/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background: HIV-1 drug resistance (HIVDR) interpretation relies on expert rule-based algorithms that translate mutation–drug relationships into clinical categories but do not directly model phenotypic susceptibility and may have limited sensitivity to complex mutational patterns. We developed PhenoGenX (PGX), a dual-engine platform combining a phenotype-trained machine learning (ML) model with an extended rule-based system to integrate data-driven inference with expert knowledge for resistance interpretation in LMICs.

Methods: From 45,039 HIV-1 clinical isolates, we curated 42,587 genotype–phenotype pairs with phenotypic fold-change (FC) measurements across 22 antiretroviral drugs. PGX integrates two independent engines: an ensemble ML model trained on mutation-level features and a rule-based interpreter derived from curated mutation knowledge bases. Model selection was guided by a Composite Resistance Performance Score (CRPS) incorporating predictive fit, error magnitude, rank correlation, categorical accuracy, and cross-validation stability. Ensemble predictions were calibrated to the PhenoSense assay scale and mapped to clinical resistance categories using safety-oriented cutoffs prioritizing minimization of very major errors. The ML engine was evaluated using an independent phenotypic dataset of 11,769 clinical isolates. The rule-based engine was benchmarked against Stanford HIVDB using 1,945 HIV-1 pol sequences (23,329 drug–sequence pairs) for NRTIs, NNRTIs, and PIs, with an additional 2,539 integrase sequences for INSTI validation.

Findings: Ensemble ML models showed consistent predictive performance across drugs (R^2 range 0.50–0.95). Calibration improved agreement with measured phenotypes (mean log-scale correlation r=0.78), and optimized cutoffs achieved high diagnostic accuracy with low very major error rates. Most drugs achieved AUC values ≥0.80. The rule-based engine demonstrated high concordance with Stanford HIVDB (overall agreement 85.6%, weighted κ=0.72), with exact agreement exceeding 92% for integrase inhibitors.

Interpretation: By integrating phenotype-calibrated ensemble ML with an extended rule-based interpreter, PhenoGenX provides a standardized framework for HIVDR interpretation that preserves biological plausibility and concordance with expert systems while maintaining a safety-weighted error profile. This approach may support HIV drug resistance surveillance and treatment decision-making where interpretation relies primarily on genotypic data in the next-generation sequencing era.

1. Background

HIV drug resistance (HIVDR) remains a major global threat to the long-term success of antiretroviral therapy (ART). Recent surveillance data from the World Health Organization highlight a growing public health challenge: pretreatment resistance to non-nucleoside reverse-transcriptase inhibitors (NNRTIs) now exceed 10% in many low- and middle-income countries, with rates surpassing 15% in parts of East and Southern Africa [1]. As countries transition to dolutegravir-based first-line regimens, emerging signals of integrase strand transfer inhibitor (INSTI) resistance underscore the need for vigilant, scalable resistance monitoring [2]. Compounding this, acquired resistance continues to accumulate in treatment-experienced populations, limiting effective second- and third-line options and increasing the risk of virological failure and onward transmission of resistant strains [3]. As ART coverage expands and treatment programs mature, the widespread use of nucleos(t)ide reverse-transcriptase inhibitors (NRTIs), NNRTIs, protease inhibitors (PIs), and INSTIs has intensified selective pressure, driving the evolution of complex resistance patterns across genetically diverse global HIV-1 populations [4], [5]. This diversity challenges existing interpretation frameworks, which often struggle to account for novel or regionally specific mutation combinations, highlighting the urgent need for more generalizable, evidence-based resistance interpretation tools.

Current HIVDR interpretation relies primarily on expert-derived, rule-based systems such as the Stanford HIVdb (https://hivdb.stanford.edu) and ANRS (https://anrs.fr/en/scientific-research/diseases-and-pathogens/hiv). These approaches encode decades of virological and clinical evidence linking specific mutations to clinical decision support. Despite their central role in clinical care, rule-based methods face inherent constraints. They assume largely additive mutation effects, struggle to represent nonlinear or higher-order interactions, and often fail to resolve rare, emerging, or uncharacterized mutation patterns. Importantly, many genotype patterns-particularly those involving mixtures of accessory, compensatory, or polymorphic mutations-do not map cleanly to existing rule sets, resulting in incomplete or ambiguous interpretations [5], [6]. These limitations apply broadly across global datasets where novel or complex mutation patterns increasingly appear.

At the same time, advances in machine learning (ML) and large-scale genotype-phenotype resources have created new opportunities to model viral resistance more continuously and holistically. Recent ML approaches have demonstrated the ability to infer fold-change susceptibility directly from mutational profiles, capture nonlinear interactions between mutations, and identify unrecognized pathways of resistance. However, most ML systems remain proof-of-concept rather than deployable tools: they are siloed, difficult to reproduce, insufficiently validated across diverse populations, and rarely integrated with established clinical interpretation frameworks [7]–[9]. As a result, clinicians and program managers lack unified platforms that combine the strengths of rule-based interpretability with the predictive resolution of ML. Given these limitations, there is a clear and urgent need for an HIVDR interpretation system that moves beyond the traditional dichotomy of rule-based versus machine-learning approaches. Such a system must retain the clinical interpretability that makes rule-based algorithms indispensable in guiding treatment decisions, while simultaneously taking advantage of large-scale genotype–phenotype datasets capable of capturing nonlinear and combinatorial patterns of resistance. It must provide explicit uncertainty estimates, particularly in the presence of rare, emerging, or previously uncharacterized mutations where biological evidence is limited. Equally important, it should deliver coherent predictions, ensuring that the resulting interpretations are reliable for both frontline patient management and large-scale public health surveillance programs in LMICs.

PhenoGenX (https://pgx.icvanalytics.org/) was developed to address this gap, which integrates enhanced rule-based interpretation with a multi-model ML ensemble trained on 42,587 HIV-1 clinical isolates from diverse populations and subtypes. The platform incorporates robust sequence quality control, alignment, mutation extraction, and a phenotype-informed ML prediction engine capable of generating continuous fold-change estimates, probabilistic resistance levels, and mutation-level contribution explanations. Its design reflects the realities of global HIV treatment programs: high heterogeneity, evolving mutation landscapes, and the need for reproducible, scalable, and transparent analytic systems. By unifying expert rules with modern data-driven modeling,

PhenoGenX aims to overcome longstanding constraints of both frameworks and provide a next-generation, integrated solution for HIV drug resistance interpretation across diverse global settings.

2. Methods

2.1. Data Curation and Sources

We assemble a large, curated dataset of HIV-1 genotype–phenotype pairs to train and validate PhenoGenX. Data were sourced from two main repositories: the Stanford HIVdb (https://hivdb.stanford.edu/pages/genotype-phenotype.html) with explicit genotype and phenotypic measurements, and GenBank (https://www.ncbi.nlm.nih.gov/genbank/) with HIV-1 genotype data. After applying quality filters from 45,039 isolates with mutations, requiring complete *pol* gene sequences, and quantitative fold-change (FC) values from standardized assays such as PhenoSense, we retained 42,587 clinical isolates (**Supplementary Figure-1**).

For the training and validation of the ML engine, we utilized a dataset where each isolate consisted of a patient-specific amino-acid mutation profile (protease, reverse transcriptase, and integrase relative to HXB2) paired with its corresponding phenotypic susceptibility, measured as the fold-change (FC) in IC_{50} via the PhenoSense assay. The dataset initially spanned 27 antiretroviral drugs across nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs), protease inhibitors (PIs), and integrase strand transfer inhibitors (INSTIs). To ensure adequate statistical support for mutation effect estimation, we excluded drugs represented by fewer than 200 isolates (that would give sufficient number of mutations for the permutation analysis of each drug), resulting in a final set of 22 drugs that form the therapeutic backbone of global HIV treatment programs, including abacavir (ABC), lamivudine (3TC), efavirenz (EFV), dolutegravir (DTG), and lopinavir (LPV).

2.2. Rule-Based System Development

We developed an extended rule-based interpreter that integrates mutations from three authoritative sources: Stanford HIVDB (www.stanford.edu), IAS-USA 2025 review [10], and WHO Surveillance Drug Resistance Mutation (SDRM) lists[11]. Resistance outputs were expressed on a normalized 0-100 scale and assigned to four clinical categories: Susceptible (S), Low-Level Resistance (LLR), Intermediate Resistance (IR), and High-Level Resistance (HLR). This framework was implemented to ensure consistent and clinically interpretable resistance calls across drugs and mutation profiles (**Supplementary Table 1**).

2.3. Validation Strategy

For the ML engine, we used an independent phenotypic dataset of 11,769 clinical isolates, balanced across drug classes. Drugs with fewer than 200 validation samples were excluded, leaving 20 drugs for final evaluation. For the rule-based engine, we retrieved 2,472 HIV-1 *pol* sequences from the Los Alamos HIV Database (https://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html). After quality filtering, 1,945 sequences (23,329 drug–sequence pairs) were retained for benchmarking against Stanford HIVDB across NRTIs, NNRTIs, and PIs; an additional 2,539 INSTI sequences were analyzed separately.

2.4. Sequence Processing Pipeline

Raw sequences, submitted in FASTA format, are processed through a multi-stage workflow designed to accommodate the variability of clinical data. Initial preprocessing removes extraneous characters and performs basic quality assessment, flagging, but not excluding sequences with high ambiguity or short length. Alignment is performed hierarchically: beginning with MAFFT optimized for HIV-1, escalating to MUSCLE and Clustal Omega if needed, and finally invoking a codon-aware fallback algorithm to reconstruct viable reading frames for challenging sequences. This cascading approach ensures successful alignment even with noisy or truncated data. Following alignment, sequences are translated and compared positionally to HXB2 to extract mutations in standard notation. Quality control annotations are recorded throughout, such as alignment method used, ambiguity levels, or codon inconsistencies, but do not halt analysis, allowing users to weigh reliability in a clinical context.

2.5. Machine Learning Model Training

For each drug, we trained an ensemble of four complementary model families using mutation-level features. Elastic Net regression provided an interpretable linear baseline, Random Forest captured nonlinear and epistatic interactions, and gradient boosted trees (XGBoost/LightGBM) modeled fine-grained, complex relationships. These models were combined into a stacked ensemble, with predictions weighted by a Composite Resistance Performance Score (CRPS).

The CRPS was calculated within a 10-fold cross-validation framework and integrated multiple performance dimensions, including predictive fit (R^2), error magnitude (MAE), ranking accuracy (Spearman correlation), clinical categorical accuracy, and stability across folds. CRPS thresholds were empirically defined based on the observed distribution of model performance, enabling separation of poorly performing and reliable predictors. Accordingly, drugs with CRPS values < 60 showed inadequate or unstable performance and were therefore excluded from the final ML engine. To limit overfitting driven by sparse observations, mutations detected in fewer than three isolates were excluded from automated model fitting and instead assigned effects through manual overrides informed by published phenotypic evidence.

2.6. Safety and Explainability Features

To ensure clinical safety and biological plausibility, multiple safeguards were embedded within the ML framework. For mutation patterns falling outside the effective training distribution, predictions are conservatively regularized toward neutral fold-change estimates to limit unsupported extrapolation. For a predefined set of well-established, high-impact drug-resistance mutations that were underrepresented in the training data, calibrated resistance overrides are applied to prevent clinically unsafe underestimation of resistance. In addition, the contribution of weakly supported mutations is capped to avoid extreme or

unstable predictions. For every prediction, a mutation-level contribution table is generated, reporting each variant's estimated effect size, prevalence within the training dataset, and whether any regularization, capping, or override was applied. This transparency enables clinicians and researchers to interpret not only the predicted resistance level but also the underlying biological and data-driven rationale.

2.7. Calibration to Phenotypic Scale

Raw ML predictions were generated on an algorithm-derived scale that does not directly correspond to biologically measured fold-change. We therefore implemented a drug-specific calibration pipeline to map predictions onto the PhenoSense assay scale. Following log-transformation and outlier filtering, predictions were aligned using Theil–Sen regression and non-parametric percentile matching. Calibration markedly improved agreement with PhenoSense measurements, shifting predictions toward the identity line across drugs and constraining values to a biologically plausible range (0.1–1000 FC). Post-calibration, the mean log-scale correlation increased to r = 0.78, with consistent improvement observed across all evaluated drugs (Fig. 1).

2.8. Clinical Cutoff Derivation

Following calibration, clinical cutoffs were derived to translate continuous FC values into discrete resistance categories (S, LLR, IR, and HLR). We first established phenotypic "ground truth" using the PhenoSense data itself, defining categories based on empirical percentiles: S ≤ 60th, LLR = 60–75th, IR = 75–90th, HLR > 90th. We then optimized PGX-specific cutoffs via a grid search that balanced multiple objectives: maximizing accuracy and weighted Kappa while minimizing very major errors (VMEs)-the most dangerous error where high-level resistance is missed. The optimization was guided by a composite objective function that heavily penalized VMEs, reflecting a "safety-first" design philosophy. Final cutoffs were reviewed for biological plausibility and consistency within drug classes.

2.9. Rule-Based Benchmarking

The rule-based engine was benchmarked against the Stanford HIVdb using 1,945 sequences (23,329 drug–sequence pairs) spanning NRTIs, NNRTIs, and PIs. Mutations were derived from aligned sequences and interpreted using the PGX rule-based scoring framework, which applies curated mutation knowledge bases and evidence-informed penalty weights to generate drug-specific resistance scores and categorical resistance calls. Agreement between PGX and Stanford HIVdb interpretations was evaluated using exact category agreement, ± 1 category agreement, and clinical error classifications (VME, ME, and minor error). Discordant cases were examined to assess the impact of extended mutation coverage, particularly for drugs with complex accessory mutation patterns. For integrase strand transfer inhibitors (INSTIs), benchmarking was performed using a curated dataset of 2,539 HIV-1 integrase sequences with corresponding reference resistance interpretations. Mutations were identified from aligned sequences and interpreted using the same PGX rule-based framework. Drug-level agreement was evaluated for BIC, CAB, DTG, EVG, and RAL using exact category agreement, agreement within ± 1 resistance category, and major discrepancy rates. Because resistance prevalence was low and class distributions were highly imbalanced, ordinal concordance was quantified using mean absolute error (MAE), and reliability was assessed using prevalence-adjusted bias-adjusted kappa (PABAK) and Matthews correlation coefficient (MCC), which are more appropriate under severe class imbalance.

2.10. Implementation, System Architecture, and Availability

PhenoGenX is implemented in Python 3.13 using standard scientific libraries (e.g., pandas, numpy, and scikit-learn) and deployed as a modular system for standardized HIV-1 sequence processing and drug resistance interpretation (**Supplementary Fig. 2**). The platform accepts HIV-1 sequences in FASTA format or pre-derived mutation tables in CSV format and applies a unified processing pipeline for quality control, alignment, and mutation calling prior to resistance interpretation. For FASTA inputs, sequences undergo quality assessment and alignment against a reference sequence, followed by mutation extraction. For mutation table inputs, variants are parsed directly and routed to the interpretation layer without re-alignment. Resistance interpretation is performed in parallel by two independent components: a rule-based engine that applies curated mutation knowledge bases and a ML engine that generates probabilistic resistance estimates from trained predictive models. Outputs from both engines are subsequently harmonized into a common ategorical framework to ensure consistent clinical interpretation across drug classes. Results are delivered through standardized resistance reports, mutation-level contribution summaries, and an interactive web interface supporting batch analyses (**Supplementary Fig. 3**). The system is deployed in a modular, containerized environment to ensure reproducibility and facilitate maintenance and updates. PhenoGenX is accessible at https://pgx.icvanalytics.org/. This architecture separates sequence processing, interpretation logic, and presentation layers, enabling transparent evaluation and extensibility of both rule-based and machine-learning components.

3. Results

3.1. Dataset Characteristics and Composition

The PGX ML engine was developed using a globally representative dataset of 45,039 HIV-1 clinical isolates with paired genotypic and phenotypic data, of which 42,587 were retained after quality filtering. This curated collection spanned all major antiretroviral drug classes, providing a robust foundation for modeling genotype–phenotype relationships. The training dataset comprised 15,248 NRTI, 6,548 NNRTI, 17,881 PI, and 5,362 INSTI isolates. Phenotypic FC distributions reflected known class-specific resistance dynamics. NRTIs exhibited the broadest dynamic range (median FC=1.8, IQR 1.0–5.5), consistent with their complex, multi-mutation resistance pathways. NNRTIs showed moderately elevated resistance profiles (median FC=2.1, IQR 0.7–45.0), aligning with their lower genetic barrier to resistance. PIs demonstrated slightly higher median FC values (median = 2.5, IQR 0.9–26.0) and included isolates with extreme resistance (FC >100) in multi-mutation backgrounds. INSTIs displayed the lowest baseline FC distribution (median=1.7, IQR 1.0–11.0), in line with their high genetic barrier, while still capturing high-level resistance associated with established INSTI resistance pathways.

For independent phenotypic validation, we utilized a separate panel of 11,769 drug–isolate pairs derived from the PhenoSense assay, providing balanced representation across drug classes and enabling direct comparison of resistance landscapes between training and validation cohorts. Median FC values in

the validation dataset closely mirrored those observed in training: 2.3 for NRTIs, 2.5 for NNRTIs, 9.4 for PIs, and 3.1 for INSTIs, confirming the comparability of resistance distributions across cohorts. Extreme resistance (FC >100) was observed across all major drug classes in both datasets. In the training cohort, 1,798 NRTI, 1,282 NNRTI, 1,837 PI, and 539 INSTI isolates exceeded this threshold, with a smaller subset exhibiting very high resistance (FC >1000). The validation cohort showed a similar pattern, supporting the representativeness of the training data and the suitability of the validation panel for independent performance assessment. Together, these results demonstrate that the PGX training dataset provides a class-balanced and phenotypically diverse representation of HIV-1 resistance, and that the validation cohort accurately reflects the resistance landscape captured during model development (**Supplementary Figure 1**).

For rule-based benchmarking, a dataset of 2,472 HIV-1 pol sequences covering nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs), and protease inhibitors (PIs) was processed. After quality filtering and harmonization with Stanford HIVDB outputs, 1,945 sequences were retained for comparative analysis. For INSTIs, benchmarking was conducted separately using a curated dataset of 2,539 HIV-1 integrase sequences with corresponding reference resistance interpretations. This validation set included sequences from diverse geographic and treatment backgrounds, with resistance profiles ranging from wild-type to highly multidrug-resistant viruses. Notably, 42% of sequences harbored at least one major NRTI resistance mutation (e.g., M184V/I, thymidine analog mutations), 38% contained major NNRTI mutations (K103N, Y181C, G190A), and 29% featured PI-associated mutations, offering a challenging and clinically realistic test set for algorithm concordance assessment.

3.2. Machine Learning Model Development and Performance

We trained and evaluated four ML models-Random Forest, XGBoost, LightGBM, and Elastic Net-across 22 antiretroviral drugs, using mutation-encoded feature matrices and continuous phenotypic fold-change as the prediction target. For each drug, we employed a stacked ensemble approach that averaged predictions across all four algorithms, maximizing generalization and stability. Ensemble R^2 values ranged from **0.53 to 0.91**, reflecting the variable predictability of genotype–phenotype relationships across drug classes. The highest R^2 values were observed for TPV (0.91), 3TC (0.84), DTG (0.78), NVP (0.78), EFV (0.79), FPV (0.81), and RTV (0.81). Model calibration, assessed via CRPS, ranged from **64.17 to 86.21**, indicating strong probabilistic reliability across most prediction tasks. Drugs with well-defined phenotypic boundaries- such as 3TC, EFV, D4T, and many PIs-achieved the highest CRPS values (**Supplementary Tables 2 and 3)**.

3.3. Calibration to Phenotypic Reference and Cutoff Optimization

A critical step in translating raw ML predictions into clinically interpretable outputs was calibrating them to the PhenoSense assay scale. Pre-calibration evaluation revealed a fundamental scale misalignment, with raw model predictions reaching orders of magnitude (e.g., >1000 FC) that were biologically implausible. Through a robust, drug-specific calibration pipeline combining Theil-Sen regression and percentile matching, we successfully mapped predictions onto the PhenoSense scale, constraining values to a biologically plausible range (0.1–1000 FC). Post-calibration, the mean log-scale correlation between PGX and PhenoSense values improved to **r=0.78**, with 75% of drugs achieving a correlation >0.70 (**Figure 1)**.

We then established optimized clinical cutoffs to categorize continuous FC values into four resistance levels: S, LLR, IR, and HLR. Using PhenoSense data as the phenotypic "ground truth," we defined initial percentile-based thresholds (S ≤60^{th}, LLR 60–75^{th}, IR 75–90^{th}, HLR >90^{th}) and refined them via a multi-objective grid search that prioritized and minimized VMEs. This optimization yielded drug-specific thresholds that balanced accuracy with clinical safety. For example, the susceptible cutoff for 3TC was rationally lowered from 115.3 FC to **49.7 FC**, appropriately capturing the high-level resistance conferred by M184V, while RAL's cutoff increased from 2.4 to **2.9 FC** to enhance sensitivity without compromising specificity. INSTIs exhibited characteristically low thresholds (1.1–3.3 FC), aligning with their high genetic barrier and clinical pharmacology. The full set of optimized cutoffs is provided in (**Supplementary Table 4)**.

3.4. Phenotypic Validation and Diagnostic Performance

Receiver operating characteristic (ROC) analysis showed strong discrimination across antiretroviral drugs, with most drugs achieving AUC values above 0.80 and a substantial subset exceeding 0.90. The highest discriminatory performance was observed for several integrase strand transfer inhibitors and protease inhibitors, including RAL, DTG, EVG, FPV, IDV, and LPV, which clustered at the upper end of the AUC distribution (**Figure 2A**).

When visualized jointly by ROC AUC and VME rate, 19 of 20 drugs (95%) achieved high discriminative performance (AUC ≥0.80), and 14 drugs (70%) met the stringent safety criterion (VME <1.5%), placing them in the optimal performance–safety region. An additional five drugs (25%) showed acceptable safety with VME <3.0%, consistent with established performance thresholds used in antimicrobial susceptibility evaluation[12]. Overall, PGX exhibited a strongly safety-weighted error profile, with most misclassifications occurring near phenotypic decision boundaries rather than as false-susceptible calls. This distribution reflects a conservative classification behavior that prioritizes minimizing missed resistance while maintaining high diagnostic accuracy across drug classes (**Figure 2C**).

Sensitivity and specificity at the optimized classification cutoffs showed that most antiretroviral drugs achieved simultaneously high resistance detection and susceptible classification performance. The majority of drugs clustered in a high-sensitivity, high-specificity region, indicating that the selected thresholds preserved both dimensions of diagnostic accuracy. Drugs classified as having very good performance achieved sensitivity and specificity above 0.8, whereas only a small subset exhibited lower balanced performance. Overall, the selected operating points reflected a favorable trade-off between false-susceptible and false-resistant classifications (**Figure 2B**). Across drugs, sensitivity increased as fold-change thresholds became more permissive, with a corresponding decline in specificity, generating characteristic drug-specific inflection points in the sensitivity–specificity trade-off curves. The optimized cutoffs consistently occurred near these inflection regions, indicating that threshold selection was guided by intrinsic performance structure rather than

arbitrary parameter choices. For most drugs, the final cutoffs achieved balanced sensitivity and specificity at biologically plausible fold-change values, supporting the stability of the calibrated classification framework (**Figure 3**).

Across antiretroviral agents, PGX demonstrated high sensitivity for identifying susceptible isolates, typically ranging from approximately 70% to over 95%, together with consistently high negative predictive values. This indicates that isolates classified as susceptible were unlikely to exhibit phenotypic resistance. For resistance detection, specificity remained conservative across drugs, generally exceeding 85% for most agents, thereby limiting false-susceptible classifications. Performance varied by drug, reflecting differences in resistance prevalence and the degree of phenotypic separation; however, overall error patterns were dominated by misclassifications near decision boundaries rather than unsafe false-susceptible calls. A subset of drugs showed particularly favorable safety profiles, with no observed very major errors, consistent with a safety-oriented classification strategy (**Table 1)**.

3.5. Rule-Based Engine Benchmarking Against Stanford HIVdb

The PGX rule-based engine demonstrated **85.6% overall concordance** with Stanford HIVdb, with a mean accuracy of 81.1% and a weighted Kappa of 0.72. Protease inhibitors showed the highest agreement: DRV achieved **92.7% accuracy** (93.2% concordance), ATV 89.2%, and LPV 88.4% (93.4% concordance). Among NRTIs and NNRTIs, strong performance was observed for 3TC/FTC (85.3% accuracy, Kappa=0.71), NVP (82.6%, Kappa=0.69), EFV (79.0%, Kappa=0.66), and ETR (77.4%, Kappa=0.60) (**Table 2**). **Supplementary Table 5 summarizes chi-square tests demonstrating statistically significant agreement between PGX and Stanford HIVdb interpretations across all evaluated antiretroviral drugs, with observed agreement ranging from 0.66 to 0.93 (all p <0.001).** Clinical error analysis revealed a **very major error rate of 1.20%,** with major and minor errors at 3.05% and 10.25%, respectively. Discrepancies primarily reflected PGX's more conservative interpretation of accessory and surveillance-relevant mutations, leading to a modest upward shift from Susceptible to Low-Level or Intermediate Resistance categories compared to Stanford. This pattern is visualized in **Table 2,** which was further disaggregated to compare resistance-level distributions between Stanford HIVdb and PGX rule-based engine, **Supplementary Figure 4**.

Across all INSTIs, PGX demonstrated high concordance with Stanford HIVdb despite extreme class imbalance, with HIVdb-defined resistance ranging from 2.6% for BIC and DTG to approximately 7.8% for CAB, EVG, and RAL. Exact agreement between PGX and HIVDB ranged from 92.1% to 97.5%, while agreement within ±1 resistance category exceeded 97.6% for all drugs. Major discrepancies were uncommon, occurring in less than 2.5% of sequences across INSTIs. Directional analysis showed that PGX was slightly more conservative than HIVDB, with a modest tendency toward higher resistance calls. Ordinal disagreement was minimal, with mean absolute error values ≤0.112 across all drugs, indicating very small average category differences. Prevalence-adjusted agreement remained high (PABAK ≥0.842), confirming strong reliability after accounting for resistance rarity. Imbalance-aware binary performance metrics showed consistent discrimination, with balanced accuracy ranging from 0.547 to 0.615 and MCC values from 0.22 to 0.39, with higher values observed for EVG and RAL, where resistance prevalence was greater (**Supplementary Table 6)**.

3.6. Score Calibration and Discordance Network Analysis

PhenoGenX scores showed clear, monotonic separation across Stanford resistance categories for all six representative drugs (3TC, EFV, AZT, ATV, DRV, and LPV). Susceptible isolates clustered at low PGX scores, while intermediate and high-level resistant isolates exhibited progressively higher score distributions. Concordance with Stanford HIVDB was highest for protease inhibitors, particularly DRV (92.7%), ATV (89.2%), and LPV (88.4%), and remained strong for 3TC (85.3%) and EFV (79.0%), with lower agreement observed for AZT (66.3%). These distributions demonstrate that PGX scores are well calibrated to established resistance categories while providing finer quantitative resolution within each class (**Supplementary Figure 5**). Discordance network analysis (**Supplementary Figure 6**) revealed that disagreements between PGX and Stanford were biologically interpretable and drug-specific, rather than systematic. The highest individual discordance rates were observed for AZT (33.7%), TDF (28.8%), and ETR (22.6%), reflecting the influence of extended mutation penalties from IAS-USA and WHO sources. Strongest co-discordance occurred between drugs sharing resistance pathways: ABC↔TDF (18.5%), ETR↔RPV (17.4%), and EFV↔NVP (17.3%). Protease inhibitors formed a distinct low-discordance cluster, underscoring their more deterministic mutational patterns.

4. Discussion

The development and validation of PGX represent a significant advance in the interpretation of HIV-1 drug resistance by integrating two separate paradigms: expert rule-based algorithms and phenotype-trained ML. By harmonizing these approaches into a single standardized platform, PGX provides a more holistic, accurate, and clinically safe tool for resistance interpretation that is particularly suited for genomic surveillance and treatment needs. Our results demonstrate that PGX successfully balances the interpretability and clinical trust of rule-based systems with the predictive precision and nonlinear modeling capability of modern ML, all while maintaining a stringent focus on minimizing dangerous under-calls of resistance.

Current HIVDR interpretation relies heavily on expert-curated rule-based systems such as Stanford HIVdb (https://hivdb.stanford.edu/), ANRS (https://anrs.fr/en/scientific-research/diseases-and-pathogens/hiv/), which encode decades of clinical and virological evidence into discrete, interpretable[13], [14]. While indispensable, these systems inherently struggle with complex, non-additive mutation interactions and emerging patterns not captured by existing rule sets [15], [16]. Conversely, ML models offer powerful capabilities for capturing nonlinear relationships and continuous phenotype prediction directly from genotype [17]–[19], but often operate as "black boxes" with limited clinical transparency and uncertain safety profiles [20], [21]. PGX addresses this dichotomy not by choosing one paradigm over the other, but by architecting a dual-engine system where each component mitigates the weaknesses of the other. The rule-based layer, extended with Stanford HIVdb (https://hivdb.stanford.edu/), IAS-USA [10] and WHO SDRM [11] mutations, ensures broad DRM coverage and maintains high concordance (85.6%) with the established Stanford HIVdb standard. Simultaneously, the ML ensemble, trained on over 42,587 phenotypic measurements, provides a quantitatively accurate, continuous estimate of fold-change resistance that captures interaction

effects and phenotypic gradients invisible to discrete rules. This integrative aligns with the growing consensus in computational medicine that the most effective clinical decision-support tools will hybridize data-driven learning with expert domain knowledge [22], [23].

The performance profile of PGX was designed to prioritize clinical safety, a requirement for any tool influencing antiretroviral therapy decisions. The rule-based engine demonstrated a very major error rate of only 1.14% compared to Stanford HIVdb, well below the clinically concerning threshold of 3%, indicating a minimal risk of under-calling resistance[24], [25]. The discrepancies that did occur were predominantly overcalls (PGX calling a higher resistance level than Stanford) or minor category shifts, reflecting the more conservative and inclusive mutation penalty system of PGX that incorporates accessory and WHO surveillance mutations. This conservative bias is clinically preferable, as it reduces the chance of continuing a failing regimen. The ML engine, after calibration to the PhenoSense biological scale, achieved a similarly low VME rate of 1.14% against the phenotypic gold standard, with an overall categorical accuracy of 90.8% across a challenging four-class system. It is noteworthy that the majority of errors were overcalls or adjacent-category misclassifications, a pattern consistent with the inherent biological ambiguity at phenotype category boundaries and preferable to under-calls [26]–[28]. The high discriminative performance, evidenced by AUC values exceeding 0.90 for key drugs like Lopinavir, confirms that the learned genotype-phenotype relationships are robust and biologically coherent.

A central challenge in deploying machine learning for phenotypic prediction is the fundamental scale misalignment between raw model outputs and biologically anchored laboratory measurements. Our pre-calibration assessment revealed that unadjusted ensemble predictions could reach orders of magnitude that are pharmacologically meaningless. This is not a failure of the ML models, which successfully learned relative resistance rankings, but rather an expected consequence of models optimizing for relative error on an unconstrained numerical scale [29], [30]. The critical contribution of our calibration pipeline was to provide a rigorous, drug-specific mapping from this internal model space to the biologically constrained scale of the PhenoSense assay. By employing a hybrid of regression (Theil-Sen) and non-parametric percentile matching, we preserved the models' learned ordinal relationships while forcing the final predictions into a plausible fold-change range (0.1–1000). This step is not a mere post-processing fix but an essential translational bridge for any ML system aiming for clinical utility, ensuring that numerical outputs correspond directly to real-world drug susceptibility concepts used in treatment guidelines.

The mutation-level effect estimates derived from the PGX ensemble models recapitulate well-established HIV resistance biology, providing an internal validation of the learning process. For NRTIs, mutations in the TAM pathways and M184V were assigned the highest impact weights, consistent with their known role in high-level resistance [6]. For NNRTIs, signature mutations like K103N and Y181C were correctly identified as major drivers of resistance. Notably, the models also captured the critical role of mutation combinations; for example, the co-occurrence of K65R and M184V showed a distinct resistance profile different from either mutation alone, and PI resistance was consistently predicted to require multi-mutation clusters, reflecting the high genetic barrier of this drug class [29], [30]. This ability to model epistatic interactions is a key advantage of the tree-based ensemble over purely additive rule-based or linear models. The explainability layer of PGX, which reports each mutation's estimated contribution to the final fold-change, makes these learned relationships transparent, allowing experts to understand not just the prediction but the mutational rationale behind it-a crucial feature for building trust and enabling scientific discovery [31].

PGX was developed with the realities of global HIV treatment programs in mind, particularly in LMIC, where the burden of HIVDR is growing, and resources for phenotypic testing are scarce[32]. The platform's ability to process raw, often imperfect Sanger or NGS sequences through a fault-tolerant pipeline makes it suitable for high-throughput national surveillance. The integration of WHO SDRMs directly supports the standardized reporting required by global monitoring initiatives[11]. For clinical management, PGX offers a more nuanced view of resistance than binary susceptible/resistant calls. The four-level resistance categorization, informed by optimized, drug-specific clinical cutoffs, can help clinicians gauge the degree of resistance, potentially informing choices between partially active drugs in salvage therapy scenarios. The platform's conservative design, which errs on the side of calling resistance, provides a safety net in settings where genotypic testing may be performed infrequently, and treatment failure carries significant individual and public health consequences.

Despite its strengths, PGX has limitations that define important directions for future development. First, the ML models are trained on existing phenotypic datasets, which under-represent rare subtypes, circulating recombinant forms, and emerging mutation patterns; given the known heterogeneity of resistance mutations by subtype, systematic subtype-stratified validation and model calibration are needed and are planned as a key future extension. Although built-in safeguards (neutral assignment for unknown mutations and critical mutation overrides) mitigate this limitation, continuous retraining with newly generated and subtype-diverse phenotypic data will be essential as the viral landscape evolves. Second, while PGX was validated against established phenotypic and genotypic reference standards, prospective clinical validation linking PGX interpretations to patient-level virological outcomes remains the definitive test of clinical utility and is required before large-scale clinical implementation. Third, the current platform focuses on the pol gene; expansion to additional genomic regions, such as the envelope for CCR5/coreceptor tropism prediction, could enable a more comprehensive resistance and treatment optimization framework. Finally, deployment in LMICs will require careful consideration of infrastructure, data governance, and integration within existing clinical and laboratory workflows.

5. Conclusion

PhenoGenX provides a dual-engine, data-driven framework for HIV-1 drug resistance interpretation by integrating a phenotype-calibrated ensemble machine learning model with an extended rule-based system. Using representative genotype–phenotype datasets spanning all major antiretroviral drug classes, the ML engine generated biologically plausible fold-change predictions that, after calibration, showed strong agreement with phenotypic reference measurements and favorable discrimination across drugs. Optimization of clinical cutoffs produced a conservative classification strategy that maintained high diagnostic accuracy while prioritizing to minimize VME. The rule-based engine demonstrated high concordance with Stanford HIVDB across reverse transcriptase inhibitors, protease inhibitors, and integrase strand transfer inhibitors, including under conditions of marked class imbalance. Discrepancies

were primarily attributable to more conservative handling of accessory and surveillance-relevant mutations, rather than systematic misclassification. Together, these results show that the two engines contribute complementary strengths: data-driven modeling of quantitative resistance patterns and transparent, knowledge-based interpretation aligned with established clinical standards. By unifying these complementary approaches within a standardized analytical pipeline, PGX preserves established biological relationships while extending resistance interpretation to complex mutational patterns that are difficult to encode using rules alone. This dual-engine architecture supports scalable HIV drug resistance surveillance and clinically interpretable decision support in LMICs.

Declarations

Ethics Statement

No ethics approval was required for this study as it exclusively used publicly available, de-identified datasets from the Stanford HIVdb, GenBank, and LANL HIV database. All data were accessed and analyzed in accordance with the terms of use of the respective source repositories.

Funding Statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' Contribution

YG conceived and designed the study, developed the platform and computational pipeline, designed validation and reporting frameworks, performed data analysis, and wrote the manuscript. BW contributed to data curation and reviewed analytical outputs. KZ contributed to the study design and validation strategy. MD reviewed the data and analytical results. ZM supported data analysis and comment the draft manuscript. GL contributed to the study design. YK contributed to the study design. GM and GT contributed to data review and interpretation of results. LL generated the figures and reviewed the draft manuscript. YS contributed to study design, provided scientific oversight, and critically reviewed the manuscript.

Declaration of Interests

We declare no competing interests.

Data Availability

All genotype–phenotype data used in this study were obtained from publicly available sources, including the Stanford HIVdb, the LANL HIV Sequence Database, and GenBank, in accordance with their respective data use policies. The curated datasets generated for model training and validation are derived from these public resources and contain no directly identifiable participant information. De-identified mutation-level datasets, rule-based scoring tables, optimized clinical cutoffs, and analysis code will be made publicly available upon publication via the PhenoGenX platform website (https://pgx.icvanalytics.org) and archived in a persistent repository with a DOI. Individual-level clinical or participant data are not shared to protect patient confidentiality.

References

1. World Health Organization, *HIV drug resistance, Brief report*. Geneva, 2024. [Online]. Available: https://www.who.int/publications/i/item/9789240086319
2. S. Rhee *et al.*, "A systematic review of the genetic mechanisms of dolutegravir resistance," *J Antimicrob Chemother*, no. July, pp. 3135–3149, 2019.
3. R. L. Hamers *et al.*, "Effect of pretreatment HIV-1 drug resistance on immunological, virological, and drug-resistance outcomes of first-line antiretroviral treatment in sub-Saharan Africa: A multicentre cohort study," *Lancet Infect. Dis.*, vol. 12, no. 4, pp. 307–317, 2012, doi: 10.1016/S1473-3099(11)70255-9.
4. D. M. Tebit and E. J. Arts, "Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease," *Lancet Infect. Dis.*, vol. 11, no. 1, pp. 45–56, 2011, doi: 10.1016/S1473-3099(10)70186-9.
5. S. Wagner, M. Kurz, and T. Klimkait, "Algorithm evolution for drug resistance prediction: Comparison of systems for HIV-1 genotyping," *Antivir. Ther.*, vol. 20, no. 6, pp. 661–665, 2015, doi: 10.3851/IMP2947.
6. S. B. et al. Dana S Clutter, Michael R Jordan, "HIV-1 Drug Resistance and Resistance Testing," *Infect Genet Evol.*, vol. 46, no. 3, pp. 292–307, 2019.
7. D. Poojitha, T. Darak, U. Samaddar, C. S. Vasavi, B. Karthikeyan, and D. B. Korlepara, "Enhancing HIV Drug Resistance Prediction Using Bidirectional LSTM Neural Networks," *Procedia Comput. Sci.*, vol. 258, pp. 2888–2898, 2025, doi: 10.1016/j.procs.2025.04.549.
8. M. C. Steiner, K. M. Gibson, and K. A. Crandall, "Drug resistance prediction using deep learning techniques on HIV-1 sequence data," *Viruses*, vol. 12, no. 5, pp. 1–24, 2020, doi: 10.3390/v12050560.
9. O. Tarasova, N. Biziukova, D. Filimonov, and V. Poroikov, "A computational approach for the prediction of HIV resistance based on amino acid and nucleotide descriptors," *Molecules*, vol. 23, no. 11, 2018, doi: 10.3390/molecules23112751.
10. A. M. Wensing *et al.*, "2022 Update of the Drug Resistance Mutations in HIV-1," *Top. Antivir. Med.*, vol. 30, no. 4, pp. 560–574, 2022.
11. World Health Organization, *HIV Drug Resistance Report 2021*, no. November. 2021. [Online]. Available: https://www.who.int/publications/i/item/9789240038608
12. U.S. Department of Health and Human Services Food and Drug Administration Center for Devices and Radiological Health, "Guidance for Industry and FDA Class II Special Controls Guidance Document: Antimicrobial Susceptibility Test (AST) Systems Preface Public Comment : Additional Copies :," pp. 1–42, 2009.

13. A. D. Revell *et al.*, “Modelling response to HIV therapy without a genotype: An argument for viral load monitoring in resource-limited settings,” *J. Antimicrob. Chemother.*, vol. 65, no. 4, pp. 605–607, 2010, doi: 10.1093/jac/dkq032.
14. B. Larder *et al.*, “The development of artificial neural networks to predict virological response to combination HIV therapy,” *Antivir. Ther.*, vol. 12, no. 1, pp. 15–24, 2007, doi: 10.1177/135965350701200112.
15. W. Heneine, “When do minority drug-resistant HIV-1 variants-have a major clinical impact?,” *J. Infect. Dis.*, vol. 201, no. 5, pp. 647–649, 2010, doi: 10.1086/650545.
16. R. Paredes *et al.*, “Pre-existing minority drug-resistant HIV-1 variants, adherence, and risk of antiretroviral treatment failure,” *J. Infect. Dis.*, vol. 201, no. 5, pp. 662–671, 2010, doi: 10.1086/650543.
17. N. Beerenwinkel *et al.*, “Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes,” *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3850–3855, 2003, doi: 10.1093/nar/gkg575.
18. A. R. Z. et al. Soo-Yon Rhee, W. Jeffrey Fessel, “HIV-1 Protease and Reverse-Transcriptase Mutations,” *J Infect Dis.*, vol. 192, no. 3, pp. 456–465, 2008, doi: 10.1086/431601.HIV-1.
19. S. Y. Rhee *et al.*, “HIV-1 protease and reverse-transcriptase mutations: Correlations with antiretroviral therapy in subtype B isolates and implications for drug-resistance surveillance,” *J. Infect. Dis.*, vol. 192, no. 3, pp. 456–465, 2005, doi: 10.1086/431601.
20. S. Joshi *et al.*, “AI as an intervention: improving clinical outcomes relies on a causal approach to AI development and validation,” *J. Am. Med. Informatics Assoc.*, vol. 32, no. 3, pp. 589–594, 2025, doi: 10.1093/jamia/ocae301.
21. C. Pou-Prom, J. Murray, S. Kuzulugil, M. Mamdani, and A. A. Verma, “From compute to care: Lessons learned from deploying an early warning system into clinical practice,” *Front. Digit. Heal.*, vol. 4, no. September, pp. 1–11, 2022, doi: 10.3389/fdgth.2022.932123.
22. R. Goel, “Artificial intelligence in medicine: its working, potentials and challenges,” *Int. J. Adv. Med.*, vol. 10, no. 1, p. 108, 2022, doi: 10.18203/2349-3933.ijam20223412.
23. G. Rong, A. Mendez, E. Bou Assi, B. Zhao, and M. Sawan, “Artificial Intelligence in Healthcare: Review and Prediction Case Studies,” *Engineering*, vol. 6, no. 3, pp. 291–301, 2020, doi: 10.1016/j.eng.2019.08.015.
24. R. Winand *et al.*, “Assessing transmissibility of HIV-1 drug resistance mutations from treated and from drug-naive individuals,” *Aids*, vol. 29, no. 15, pp. 2045–2052, 2015, doi: 10.1097/QAD.0000000000000811.
25. C. Chu, D. Armenia, C. Walworth, M. M. Santoro, and R. W. Shafer, “Genotypic Resistance Testing of HIV-1 DNA in Peripheral Blood Mononuclear Cells,” *Clin. Microbiol. Rev.*, vol. 35, no. 4, 2022, doi: 10.1128/cmr.00052-22.
26. H. F. Günthard and A. U. Scherrer, “HIV-1 Subtype C, Tenofovir, and the Relationship with Treatment Failure and Drug Resistance,” *J. Infect. Dis.*, vol. 214, no. 9, pp. 1289–1291, 2016, doi: 10.1093/infdis/jiw214.
27. R. Kantor *et al.*, “Pretreatment HIV Drug Resistance and HIV-1 Subtype C Are Independently Associated with Virologic Failure: Results from the Multinational PEARLS (ACTG A5175) Clinical Trial,” *Clin. Infect. Dis.*, vol. 60, no. 10, pp. 1541–1549, 2015, doi: 10.1093/cid/civ102.
28. V. Kouamou and A. M. Mcgregor, “https://www.scientificarchives.com/journal/journal-of-aids-and-hiv-treatment High Levels of Pre-Treatment HIV Drug Resistance in Zimbabwe: Is this a Threat to HIV/AIDS Control? Dedication Conflict of Interest,” *J AIDS HIV Treat*, vol. 3, no. 3, pp. 42–45, 2021, [Online]. Available: https://www.scientificarchives.com/journal/journal-of-aids-and-hiv-treatment
29. A. Cozzi-Lepri *et al.*, “Low-frequency drug-resistant HIV-1 and risk of virological failure to first-line NNRTI-based ART: A multicohort European case-control study using centralized ultrasensitive 454 pyrosequencing,” *J. Antimicrob. Chemother.*, vol. 70, no. 3, pp. 930–940, 2015, doi: 10.1093/jac/dku426.
30. M. Noguera-Julian *et al.*, “Contribution of APOBEC3G/F activity to the development of low-abundance drug-resistant human immunodeficiency virus type 1 variants,” *Clin. Microbiol. Infect.*, vol. 22, no. 2, pp. 191–200, 2016, doi: 10.1016/j.cmi.2015.10.004.
31. C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.
32. N. Parkin, P. R. Harrigan, S. Inzaule, and S. Bertagnolio, “Need assessment for HIV drug resistance testing and landscape of current and future technologies in low- and middle-income countries,” *PLOS Glob. Public Heal.*, vol. 3, no. 10 October, pp. 1–19, 2023, doi: 10.1371/journal.pgph.0001948.

Tables

Table-1: Validation of PGX in reference to PhenoSense

Drug	N	TP_S	TN_S	FP_S	FN_S	Sens_S	Spec_S	PPV_S	NPV_S	TP_R	TN_R	FP_R	FN_R	Sens_R	Spec_R	PPV_R
3TC	655	121	447	79	8	0.94	0.85	0.61	0.98	342	285	5	23	0.94	0.98	0.99
ABC	462	150	232	22	58	0.72	0.91	0.87	0.8	25	230	3	204	0.11	0.99	0.89
ATV	290	64	168	31	27	0.7	0.84	0.67	0.86	72	142	30	46	0.61	0.83	0.71
AZT	607	174	353	15	65	0.73	0.96	0.92	0.84	202	265	130	10	0.95	0.67	0.61
BIC	540	339	95	29	77	0.81	0.77	0.92	0.55	26	449	4	61	0.3	0.99	0.87
D4T	611	113	366	115	17	0.87	0.76	0.5	0.96	73	430	21	87	0.46	0.95	0.78
DDI	609	335	143	47	84	0.8	0.75	0.88	0.63	16	452	3	138	0.1	0.99	0.84
DTG	889	569	120	29	171	0.77	0.81	0.95	0.41	41	773	28	47	0.47	0.97	0.59
EFV	691	234	246	199	12	0.95	0.55	0.54	0.95	209	389	14	79	0.73	0.97	0.94
EVG	1532	31	606	894	1	0.97	0.4	0.03	1.00	348	990	57	137	0.72	0.95	0.86
FPV	797	119	461	206	11	0.92	0.69	0.37	0.98	237	430	39	91	0.72	0.92	0.86
IDV	801	319	428	35	19	0.94	0.92	0.9	0.96	166	338	8	289	0.36	0.98	0.95
LPV	501	143	322	21	15	0.91	0.94	0.87	0.96	204	211	37	49	0.81	0.85	0.85
NFV	836	266	509	28	33	0.89	0.95	0.9	0.94	321	398	65	52	0.86	0.86	0.83
NVP	706	373	273	53	7	0.98	0.84	0.88	0.98	242	397	16	51	0.83	0.96	0.94
RAL	1636	18	909	709	0	1	0.56	0.02	1.00	499	941	161	35	0.93	0.85	0.76
RTV	802	196	475	122	9	0.96	0.8	0.62	0.98	320	396	58	28	0.92	0.87	0.85
SQV	824	399	319	41	65	0.86	0.89	0.91	0.83	127	537	37	123	0.51	0.94	0.77
TDF	296	47	132	111	6	0.89	0.54	0.3	0.96	17	167	10	102	0.14	0.94	0.63
TPV	148	47	35	5	61	0.44	0.88	0.9	0.36	16	73	35	24	0.4	0.68	0.31

Table 2: PGX rule-based engine performance compared to the Stanford HIVDB

Drug	Total_Comparisons	Accuracy	VME_Rate	ME_Rate	Minor_Error_Rate	Overall_Error_Rate	Concordance_Rate
3TC	1,820	0.85	0.03	0	0.1	0.13	0.87
FTC	1,820	0.85	0.03	0	0.1	0.13	0.87
ABC	1,820	0.78	0	0.04	0.13	0.17	0.83
AZT	1,499	0.66	0.01	0.02	0.12	0.14	0.86
TDF	1,654	0.71	0.01	0.04	0.17	0.22	0.78
DOR	1,820	0.77	0.01	0.06	0.13	0.19	0.81
EFV	1,820	0.79	0.03	0.06	0.1	0.18	0.82
ETR	1,820	0.77	0	0.02	0.16	0.19	0.81
NVP	1,820	0.83	0.03	0.07	0.02	0.13	0.87
RPV	1,820	0.81	0.01	0.05	0.09	0.15	0.85
ATV	1,874	0.89	0	0.02	0.08	0.1	0.9
DRV	1,868	0.93	0	0	0.07	0.07	0.93
LPV	1,874	0.88	0	0.01	0.05	0.07	0.93

Note: *Total_Comparisons* indicates the total number of paired predictions and reference interpretations evaluated per drug. *Accuracy* represents the proportion of correct predictions among all comparisons. *VME* (very major error) denotes false-susceptible predictions; *ME* (major error) denotes false-resistant predictions; *Minor_Error* denotes misclassification between intermediate and adjacent resistance categories. *Overall_Error_Rate* is the sum of very major, major, and minor error rates. *Concordance_Rate* represents the proportion of predictions concordant with the reference interpretation.

Figures

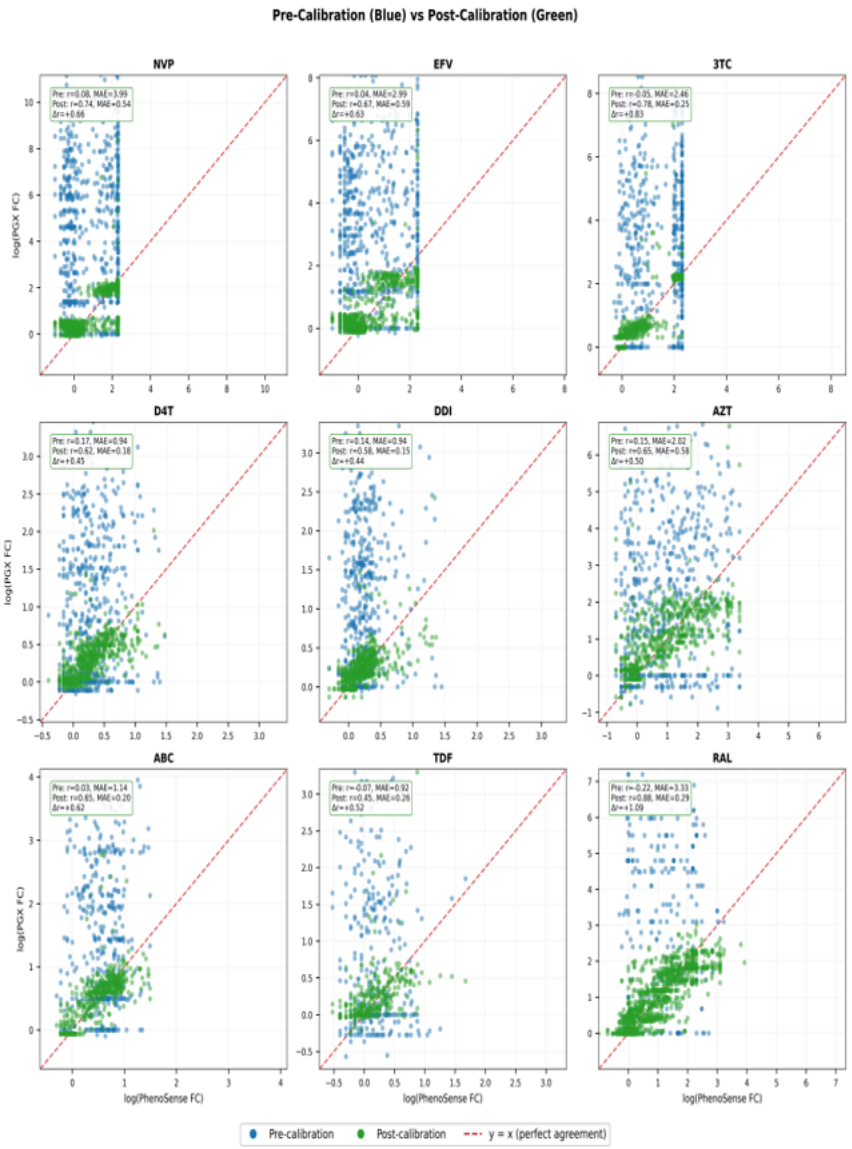


Figure 1

Drug-specific calibration of PGX predictions to the PhenoSense phenotypic scale.

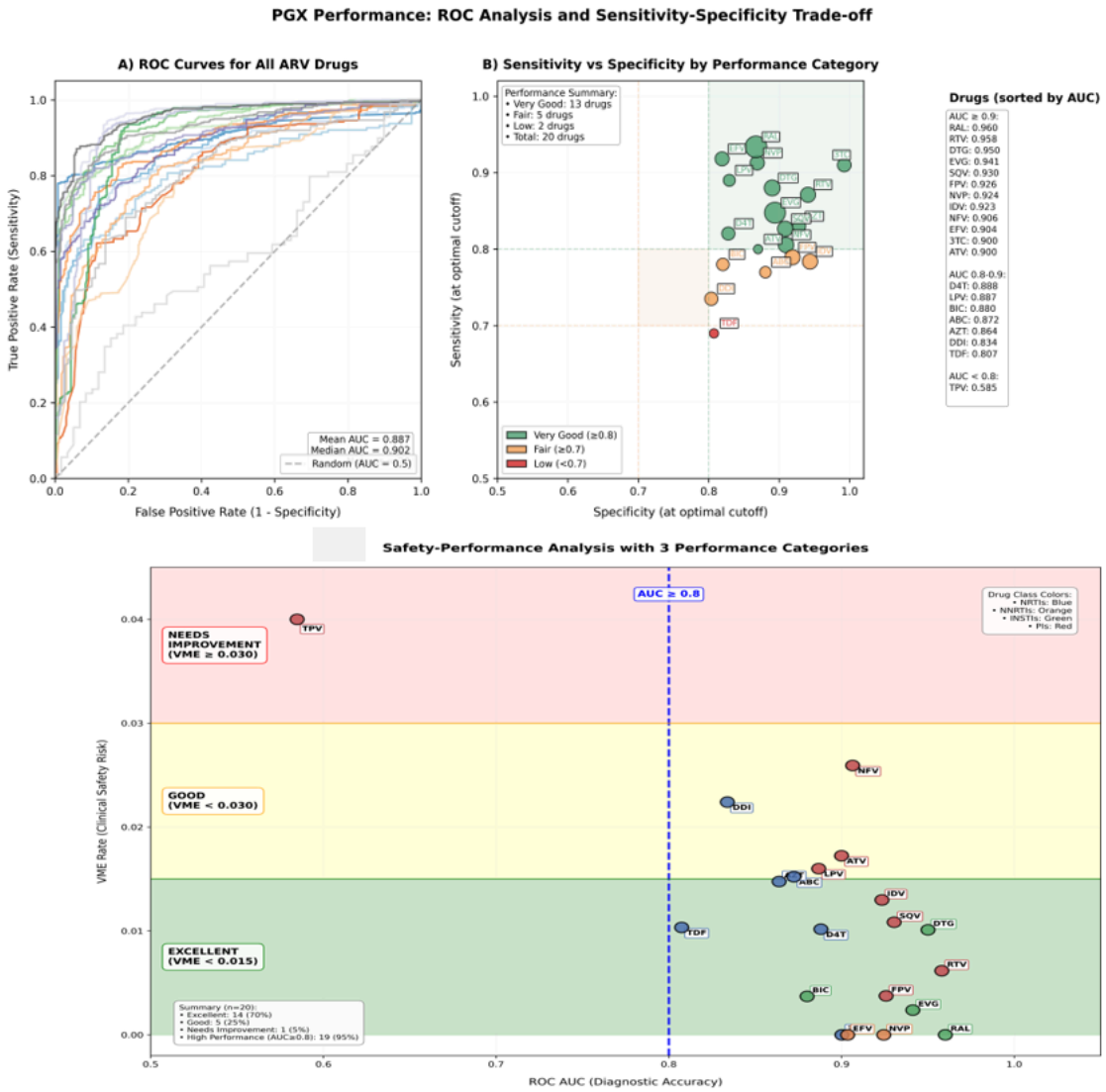


Figure 2

ROC-based performance of the PGX machine-learning model across 20 antiretroviral drugs.
ROC curves and sensitivity–specificity distributions demonstrate strong and consistent discrimination across drug classes, with most agents achieving AUC ≥0.80 at optimized thresholds.

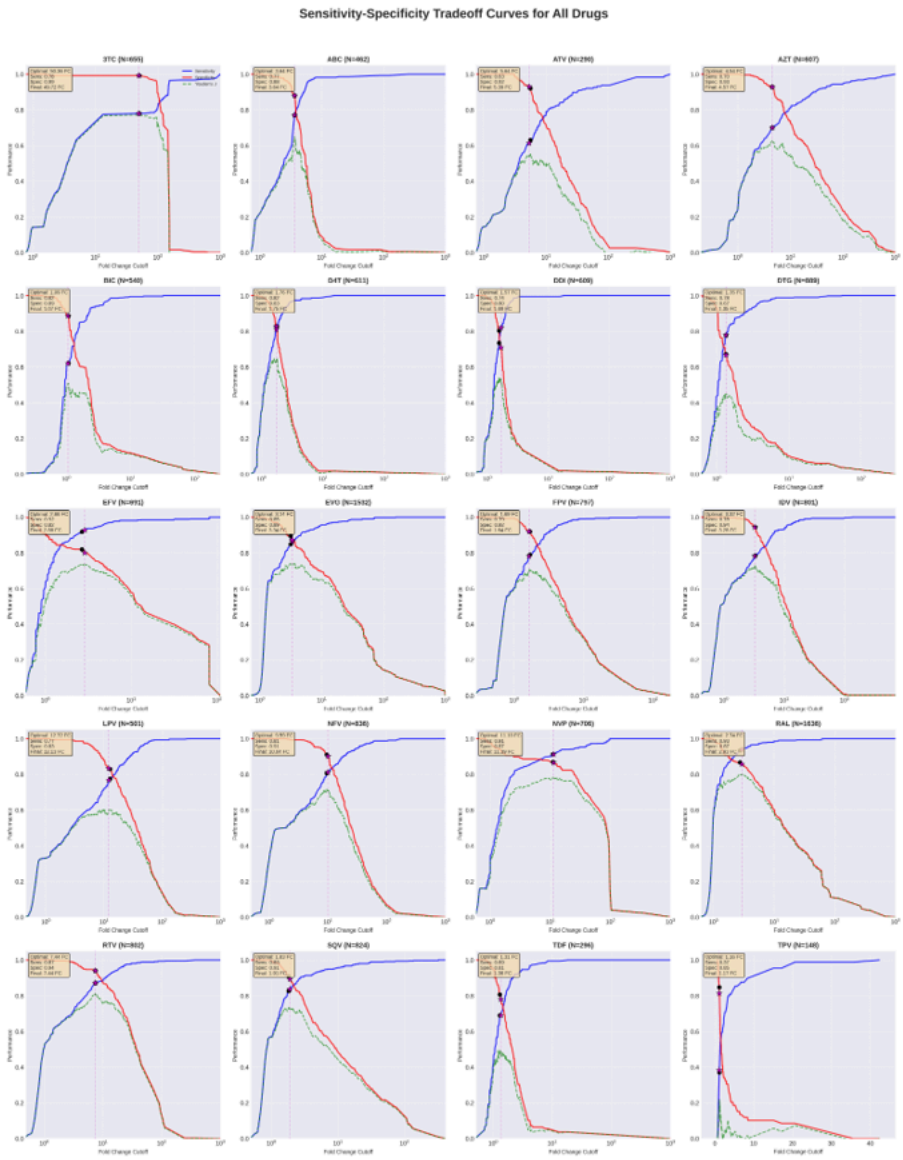


Figure 3

Sensitivity-specificity trade-off curves across 20 antiretroviral drugs.

For each drug, sensitivity (blue) and specificity (red) were computed across calibrated fold-change thresholds relative to PhenoSense. Vertical dashed lines indicate optimized cutoffs selected to balance sensitivity and specificity, with historical susceptible cutoffs shown for comparison. The curves demonstrate drug-specific inflection points and support the use of calibrated, drug-specific thresholds for resistance classification across NRTIs, NNRTIs, PIs, and INSTIs.

Supplementary Files

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

- PGXsupplementariesNPJV2.0.docx