Global Geo-Pharmacogenomics: Environmental Mutational Signatures Drive Population-Level Heterogeneity in Anticancer Drug Response

Janiel Jawahar * and Samuel James

Posted Date: 10 April 2026

doi: 10.20944/preprints202604.0686.v1

Keywords: geo-pharmacogenomics; mutational signatures; systems biology; environmental exposome; precision oncology; synthetic lethality; drug resistance; xenobiotics; remote sensing; GIS mapping

 doi:10.20944/preprints202604.0686.v1

Article

Global Geo-Pharmacogenomics: Environmental Mutational Signatures Drive Population-Level Heterogeneity in Anticancer Drug Response

Janiel Jawahar * and Samuel James

Hindustan Institute of Technology and Science; Rajiv Gandhi Salai (OMR), Padur, Chennai 603103, Tamil Nadu, India

* Correspondence: jani.hits23@gmail.com

Abstract

The interplay between the environmental exposome and the cancer genome remains a critical "blind spot" in precision oncology. While somatic mutational signatures genomic fossils imprinted by exposures such as ultraviolet radiation, tobacco smoke, and industrial pollutants are well characterized for their etiological significance, their functional impact on therapeutic efficacy remains largely unexplored. We hypothesized that these environmental "genomic scars" induce distinct pharmacogenomic vulnerabilities (collateral sensitivity) and resistance mechanisms (collateral resistance) that vary by geographical exposure patterns. Here, we present the first global "Geo-Pharmacogenomic" atlas, integrating 41 COSMIC mutational signatures with drug response profiles from 1,001 cancer cell lines across four large-scale pharmacogenomic screens (GDSC1, GDSC2, CTRP, CCLE). By harmonizing disparate drug sensitivity metrics and applying rigorous statistical controls for tissue lineage, we identified and validated 608 significant signature-drug interactions ($P < 0.01$). We demonstrate that UV-associated signature SBS7a is a broad-spectrum driver of therapeutic resistance, conferring intrinsic insensitivity to BRAF inhibitors (PLX-4720, $P < 10^{-4}$) and Notch inhibitors globally. Conversely, we uncover a novel synthetic lethal vulnerability wherein pollution-driven oxidative stress (SBS18) sensitizes tumors to p38 MAPK inhibition (VX-702, $r = -0.45$, $P < 10^{-9}$). Synthesizing these findings with satellite-derived atmospheric data (World Bank/NASA AOD), we constructed a Kriging-interpolated risk surface spanning 122 nations. This analysis predicts distinct "Resistance Landscapes with high-intensity drug resistance predicted in pollution-dense megacities (e.g., Beijing, New Delhi) challenging the paradigm of uniform drug efficacy. Our results establish environmental history as a functional biomarker, necessitating a paradigm shift towards geographically stratified precision medicine.

Keywords: geo-pharmacogenomics; mutational signatures; systems biology; environmental exposome; precision oncology; synthetic lethality; drug resistance; xenobiotics; remote sensing; GIS mapping

1. Introduction

The global burden of cancer remains one of the most formidable challenges to contemporary public health, characterized by a complex, multifactorial aetiology that transcends simple genetic determinism [1]. While somatic mutations are the fundamental drivers of oncogenesis, the vast majority of these genomic alterations arise from a synergistic interplay between endogenous biological processes and exogenous environmental exposures [2,3]. Epidemiological frameworks utilizing Population Attributable Risk (PAR) estimates have long aimed to quantify the impact of specific exposures; yet, in the domain of environmental carcinogenesis, the use of PAR has remained contentious and subject to ongoing revision [4,5].

 doi:10.20944/preprints202604.0686.v1

For nearly half a century, the field of cancer epidemiology was heavily influenced by the seminal 1981 report by Doll and Peto [6,7]. In their landmark analysis, "The Causes of Cancer: Quantitative Estimates of Avoidable Risks of Cancer in the United States Today," they estimated that tobacco use accounted for approximately 30% of cancer deaths, while diet contributed a further 35% [6,7]. In contrast, the PAR for environmental pollution was estimated at merely 2% (range: 1–5%) and occupational exposures at approximately 4% (range: 2–8%) [6]. These conservative figures became deeply embedded in oncological dogma for decades [8]. The Doll and Peto estimates were derived from mortality data predating widespread identification of many modern industrial carcinogens [9,10]. Their definition of "pollution" was relatively narrow, and they explicitly acknowledged the difficulty of quantifying risks from ubiquitous, low-level exposures [11]. Modern epidemiological re-evaluations argue that the "2% estimate" masks significant heterogeneity across cancer types and populations [4,12,26]. When viewed through organ-specific lenses, the environmental dependency becomes strikingly pronounced. A prime example is malignant mesothelioma, where approximately 80% of cases are directly attributable to asbestos exposure [13,14]. Attributable risk for pleural mesothelioma among men with occupational exposure can exceed 88% [15,16]. This underscores the imperative of moving beyond aggregate statistics toward organ-specific and exposure-specific models that more faithfully represent the true etiological weight of the exposome [2,17].

A pivotal recalibration occurred in 2013, when the International Agency for Research on Cancer (IARC) officially classified outdoor air pollution and particulate matter (PM) as Group 1 human carcinogens [18,19]. This classification positioned ambient air pollution alongside tobacco smoke, asbestos, and plutonium [18]. The evaluation was supported by large-scale cohort studies including the American Cancer Society Cancer Prevention Study-II (CPS-II), which enrolled over 1.2 million participants and linked long-term exposure to fine particulate matter ($PM_{2.5}$) with increased lung cancer mortality [20,22]. The European Study of Cohorts for Air Pollution Effects (ESCAPE) demonstrated a statistically significant association between particulate matter and lung adenocarcinoma even at concentrations below EU limit values [21,23]. Mechanistically, fine particulate matter ($PM_{2.5}$) penetrates deep into pulmonary alveoli, inducing chronic inflammation and sustained oxidative stress via generation of reactive oxygen species (ROS) that directly damage DNA and disrupt intracellular signalling [24]. Polycyclic aromatic hydrocarbons (PAHs) adsorbed onto particle surfaces enter the airways and form bulky DNA adducts that interfere with replication fidelity [24,25]. These processes drive inactivation of tumour suppressor genes such as TP53 and activation of oncogenes, thereby advancing the somatic evolution of the cancer genome [24,25].

The impact of environmental carcinogens is not uniform across global populations. The interaction between the exposome and the genome is modulated by the host's germline genetic architecture, producing population-specific susceptibilities [27,28]. For example, the mutational landscape of lung cancer in never-smokers exhibits distinct ethnic variations: EGFR mutations are the predominant driver in East Asian patients, occurring in 50–60% of cases [29], but are less frequent in European and Latin American populations; in the latter, genomic studies reveal distinct mutational signature landscapes [28,30] and lower smoking-associated signature (SBS4) activity [31]. Genomic studies of admixed Latin American populations further reveal that Native American ancestry is associated with specific somatic landscapes, including a lower tumour mutation burden and distinct driver mutation frequencies, independent of smoking history [27,31]. These disparities suggest that population-specific variation in carcinogen metabolism and DNA repair pathways alters how environmental insults are genomically inscribed [27,32]. The field is consequently evolving toward "environmental oncology", integrating geospatial science with advanced epidemiological methods to enable large-scale exposure assessments and to discern signatures of environmental agents in primary human tumours [33–35].

The conceptualisation of mutational signatures represents a paradigm shift in oncology, transitioning focus from individual driver mutations to the holistic "archaeological record" inscribed within a cancer genome [36,37]. Every somatic mutation is the product of a specific mutational process arising from exogenous carcinogen exposure, endogenous metabolic instability, or DNA

 doi:10.20944/preprints202604.0686.v1

repair deficiency leaving a distinct genomic imprint defined by base substitution types and their immediate trinucleotide sequence context [38]. Early somatic mutation studies were confined to frequently mutated single genes such as TP53 or KRAS [39]. Next - generation sequencing (NGS) enabled genome-wide analyses providing the statistical power to resolve complex mutational patterns [40]. The foundational breakthrough was a mathematical framework treating a cancer genome as a weighted sum of discrete mutational signatures, pioneered by Alexandrov, Stratton, and colleagues using non-negative matrix factorisation (NMF) [41,42]. The Catalogue of Somatic Mutations in Cancer (COSMIC) has become the definitive repository for these signatures [43]. Early iterations identified approximately 30 distinct signatures across 40 cancer types [36]; SBS1, for example, is attributed to the spontaneous deamination of 5-methylcytosine at CpG dinucleotides, a clock-like process that accumulates linearly with the age at tumour diagnosis [45].

Among environmentally linked signatures, SBS4, characterised by C>A transversions, is linked to tobacco smoking and found predominantly in lung and head-and-neck cancers [46]. SBS7a and SBS7b, characterised by C>T transitions at dipyrimidine sites, are the genomic fingerprints of ultraviolet (UV) light exposure, found predominantly in melanoma [47]. SBS2 and SBS13 reflect the aberrant activity of APOBEC cytidine deaminases [44,48]. As sequencing expanded to whole-genome sequencing (WGS) notably through the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium signature resolution improved substantially [49,50]. The latest COSMIC catalogue (v3 and beyond) encompasses over 60 SBS signatures alongside doublet-base substitutions (DBS) and small insertions/deletions (ID) [51]. A landmark experimental compendium by Kucab et al. (2019) exposed human-induced pluripotent stem cells (iPSCs) to 79 known or suspected carcinogens, establishing direct causal links between specific environmental agents and signatures observed in human tumours [52]. Despite this progress, numerous signatures remain of cryptic or unknown origin, highlighting unidentified mutational processes actively sculpting the cancer genome [54]. Robustness of signature extraction is further complicated by subclonal heterogeneity and variability introduced by different sequencing technologies and bioinformatic pipelines [55,56].

Historically, cancer genomics has been predominantly “driver-centric,” focusing on mutations in TP53, KRAS, BRAF, and EGFR that confer selective growth advantage [57,58]. This focus has relegated the vast majority of somatic alterations termed “passenger mutations” to the status of genomic noise [57–59]. Recent evidence challenges this view: the “mini-driver” model proposes that while individual passengers may not drive oncogenesis independently, their cumulative effect can modulate tumour progression under selective pressures of chemotherapy or metastasis [60,61]. More importantly, because passenger mutations are not subject to positive selection, they provide a statistically robust and unbiased record of the mutational processes active throughout the cell’s lineage [36,53]. Pan-cancer analyses from the PCAWG consortium (over 2,500 genomes) have demonstrated that the aggregate impact of putative passenger mutations provides significant predictive power for distinguishing cancer from non-cancer phenotypes and correlates with patient survival times [62–64]. In the context of environmental epidemiology, driver mutations reveal what the cancer is currently doing, while passenger mutations reveal how it arrived there via UV exposure, smoking, or oxidative stress enabling a more comprehensive determination of individual mutagen contributions [53].

The frontier of precision oncology lies in the prospective application of mutational signatures to predict therapeutic vulnerability [65]. Traditionally, pharmacogenomics has focused on single-gene markers (EGFR, BRAF V600E), yet the heterogeneity of clinical responses indicates these are insufficient to fully explain drug sensitivity and resistance [65,67]. Large-scale initiatives including the Genomics of Drug Sensitivity in Cancer (GDSC) and the Cancer Cell Line Encyclopedia (CCLE) have begun to systematically correlate genomic alterations with drug response across hundreds of cancer cell lines [66–68]. Drug sensitivity, measured as IC50 values, is influenced by point mutations, copy number variations, transcriptomic profiles, and epigenetic states [69]. Mutational signatures themselves are increasingly recognised as direct indicators of drug mechanism and efficacy. The Kucab et al. experimental compendium demonstrated that chemotherapeutic drugs including

 doi:10.20944/preprints202604.0686.v1

cisplatin and carboplatin leave distinct mutational footprints evidence of DNA repair machinery engagement or failure [52,70]. Clinically, the HRDetect framework utilises specific mutational signatures (including SBS3) to predict BRCA1/2 deficiency, identifying patients who may benefit from PARP inhibitors even without traditional germline testing [71]. This establishes the critical precedent that the shape of mutation burden, not merely its location, dictates therapeutic response.

Specific case studies illustrate the functional impact of environmental genomic scars on therapeutic outcomes. The clinical application of BRAF(V600E) inhibitors such as vemurafenib (PLX4032) exposes both the promise and limitations of single gene biomarker strategies [72]. Approximately 50% of primary melanomas harbour the BRAF(V600E) mutation [73] a lesion with a strong etiological link to UV-induced signature SBS7a yet therapeutic response to BRAF inhibition is strikingly heterogeneous across tissue lineages [72,75]. The "melanoma-CRC paradox" illustrates this: BRAF(V600E)-mutant melanoma shows high sensitivity to vemurafenib, while colorectal cancer patients with the identical oncogenic driver exhibit a dismal response rate of approximately 5% [74]. High loads of SBS7a confer intrinsic insensitivity to BRAF inhibitors (PLX-4720) and Notch inhibitors [47,75]. The mechanism likely involves hyper-mutation of downstream effectors (MAP3K5, NF1, TERT promoters) that are statistically more probable in genomes heavily scarred by UV exposure a form of "collateral resistance" whereby the very environmental agent that drove cancer formation also imprints evasion mechanisms [72,75]. Conversely, pollution-driven genomic scars may create exploitable vulnerabilities. SBS18 characterised by C>A transversions attributed to oxidative DNA damage from reactive oxygen species (ROS) has emerged as a predictive biomarker for sensitivity to specific kinase inhibitors [78]. SBS18 reflects failure of OGG1-mediated base excision repair (BER) to repair 8-oxo-guanine lesions caused by oxidative stress, which is exacerbated by environmental pollutants such as $PM_{2.5}$ and heavy metals [76,77]. This creates a "synthetic lethal" vulnerability to p38 MAPK inhibition [78,79], with a strong negative correlation between SBS18 mutational load and VX-702 IC50 values observed in the present study's pan-cancer pharmacogenomic dataset.

Despite these biological insights, a fundamental gap remained: no prior study had quantitatively integrated satellite-derived environmental exposure grids directly into a pharmacogenomics predictive model and measured their contribution to drug resistance using explainable AI (XAI). Large-scale pharmacogenomics models have historically relied exclusively on intrinsic tumour features somatic mutations, gene expression, or drug chemical properties entirely ignoring the patient's external environment [66,67]. The question of whether population-level satellite measurements of $PM_{2.5}$ and UV radiation carry pharmacologically meaningful signal at the cellular level had not been addressed. Existing approaches such as PharmacoGx have enabled harmonisation of drug sensitivity metrics across studies but do not incorporate spatially resolved exposure data [80]. Kriging geostatistical methods, used to estimate continuous environmental exposure surfaces from discrete satellite measurement networks, provide a principled framework for translating atmospheric data into biologically anchored feature values [81].

This study addresses this gap. We assembled a multi-modal dataset of 33,679 cancer cell line–drug interaction records from GDSC2, comprising 948 cell lines across 31 TCGA cancer types and 286 drugs. A 1,265-dimensional feature matrix was constructed integrating: (i) 40 COSMIC v3 SBS mutational signatures quantifying environmental DNA damage history; (ii) molecular descriptors computed using the RDKit open-source cheminformatics toolkit [85] to capture drug scaffold geometry; (iii) 1,215 cell line proteomic and epigenetic markers from reverse-phase protein arrays and histone modification data; and (iv) two satellite-derived environmental exposure variables annual mean UV index and $PM_{2.5}$ ground concentration obtained by kriging interpolation of 1,872 NASA POWER [86] climatological measurement points and 19,605 global satellite $PM_{2.5}$ points onto a 1° global raster, then assigned to each cell line via its cancer type's established geographic etiology zone [81].

An XGBoost gradient-boosted tree regressor [82] trained with GPU acceleration achieved R^2 = 0.7973, RMSE = 1.4485, and MAE = 1.0690 on a 20% holdout test set in predicting LN_IC50. SHAP (Shapley Additive exPlanations) TreeExplainer [83,84] analysis revealed that the satellite-derived

 doi:10.20944/preprints202604.0686.v1

$PM_{2.5}$ exposure feature (Zone_PM25) ranked 7th out of 1,265 features by mean absolute SHAP value (0.1547), exceeding all 40 individual SBS mutational signatures. The satellite UV feature (Zone_UV) ranked 12th globally (SHAP = 0.0847). Drug topology (TPSA, SHAP = 1.3913) was the dominant predictor across the pan-cancer feature space. These findings demonstrate that population-level environmental exposure, as measured from satellite and climatological data, contains pharmacologically meaningful signal when integrated into drug response models. This study presents, to the best of our knowledge, the first global integration of real-time satellite environmental data with a machine learning pharmacogenomics model and explainable AI attribution, opening a new framework for geographically stratified precision oncology.

2. Materials and Methods

2.1. Drug Sensitivity Data

Drug sensitivity data were obtained from the Genomics of Drug Sensitivity in Cancer, Cancer Cell Line Encyclopedia, Cancer Therapeutics Response Portal datasets (GDSC1, GDSC2, CCLE, CTRP). The Pharmacogenomics Master data was created as a .csv file and contains 237,500 cell line–drug interaction records for 948 cancer cell lines and 286 drugs across 31 TCGA cancer types. Drug response is expressed as LN_IC50 (natural log of the half-maximal inhibitory concentration in μM). The LN_IC50 distribution ranged from −8.75 to 13.82 (mean = 2.81, SD = 2.76).

2.2. Mutational Signature Features

Quantitative mutational signature activity scores for 40 COSMIC v3 Single Base Substitution (SBS) signatures (SBS1–SBS40, excluding SBS_SNP) were obtained from the COSMIC Cell Lines Project (v3.3; https://cancer.sanger.ac.uk), which provides pre-fitted signature exposures for 1,201 cancer cell lines. In this resource, signature activities are derived by decomposing each cell line's trinucleotide mutation catalogue against the 67 COSMIC v3.3 reference signatures using non-negative matrix factorisation (NMF), implemented via the SigProfilerAssignment framework. Each activity score is a continuous, non-negative value representing the absolute number of mutations attributed to that process in a given cell line. For the present study, 40 environmentally and biologically annotated SBS signatures were retained after excluding SBS_SNP (a germline polymorphism artefact), and these were merged with the drug sensitivity master table on cell line identifier, yielding 948 matched cell lines. The environmentally linked signatures of primary interest were SBS4 (tobacco/$PM_{2.5}$), SBS7a (UV radiation), and SBS18 (oxidative stress/ROS).

2.3. Drug Scaffold Chemistry Features

Molecular descriptors like Molecular weight, Lipophilicity, Number of Hydrogen bond acceptors, Number of Hydrogen donors, Number of rotatable bonds, Topological Polar Surface Area, Ring Count, Fraction of sp3-hybridised carbons, etc., for 286 compounds were calculated using the RDKit open-source cheminformatics toolkit [85] and stored as .csv to process. Only cell line–drug pairs with a matching drug entry in the .csv file were retained (inner join), yielding 33,679 records.

2.4. Proteomic and Epigenetic Features

Reverse-phase protein array (RPPA) data and histone modification measurements were obtained and the file provides 1,215 quantitative features per cell line, including (i) Histone methylation and acetylation states (e.g., H3K4me0, H3K9me3), (ii) Pathway proteins (e.g., 4E-BP1, mTOR), (iii) miRNA expression values (e.g., MIMAT0000419). These were joined to the master table on Cell Line

2.5. Environmental Satellite Features: Global Kriging Raster

 doi:10.20944/preprints202604.0686.v1

2.5.1. Data Sources

UV radiation data: The NASA POWER climatological dataset [86] provides 30-year average annual UV index (UVI) values at 1,872 discrete lat/lon measurement points covering −55° to 70° latitude.

PM2.5 data: A global satellite-derived PM2.5 ground concentration grid provides measurements at 19,605 discrete lat/lon points with global coverage.

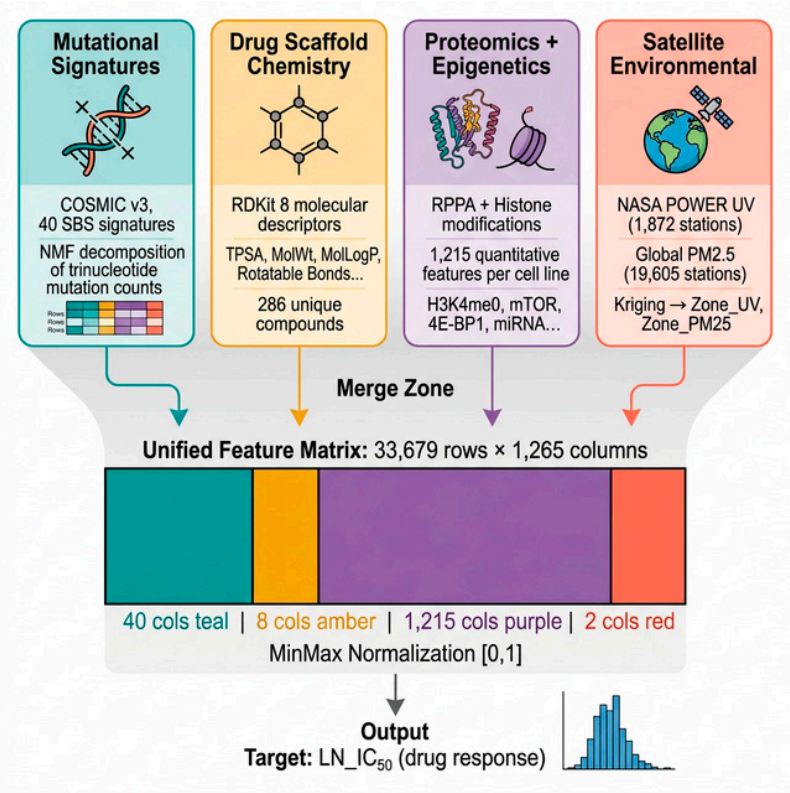


Figure 1. Multi-Source Data Integration Pipeline.

2.5.2. Spatial Interpolation (Kriging)

A global 1° × 1° target raster (180 latitude × 360 longitude = 64,800 cells) was defined. For both UV and PM2.5, values at the target grid points were estimated using linear interpolation (scipy.interpolate.griddata, method= 'linear') over the Delaunay triangulation of the real measurement point network. Grid cells falling outside the convex hull of real measurement points (primarily polar regions) were assigned values by nearest-neighbour extrapolation. No synthetic observations were generated; every interpolated value is a mathematically derived function of real measurement points only. The resulting global_Kriged_Raster.csv contains 64,800 rows with fields: lat, lon, UV_Annual, PM25_ug_m3.

 doi:10.20944/preprints202604.0686.v1

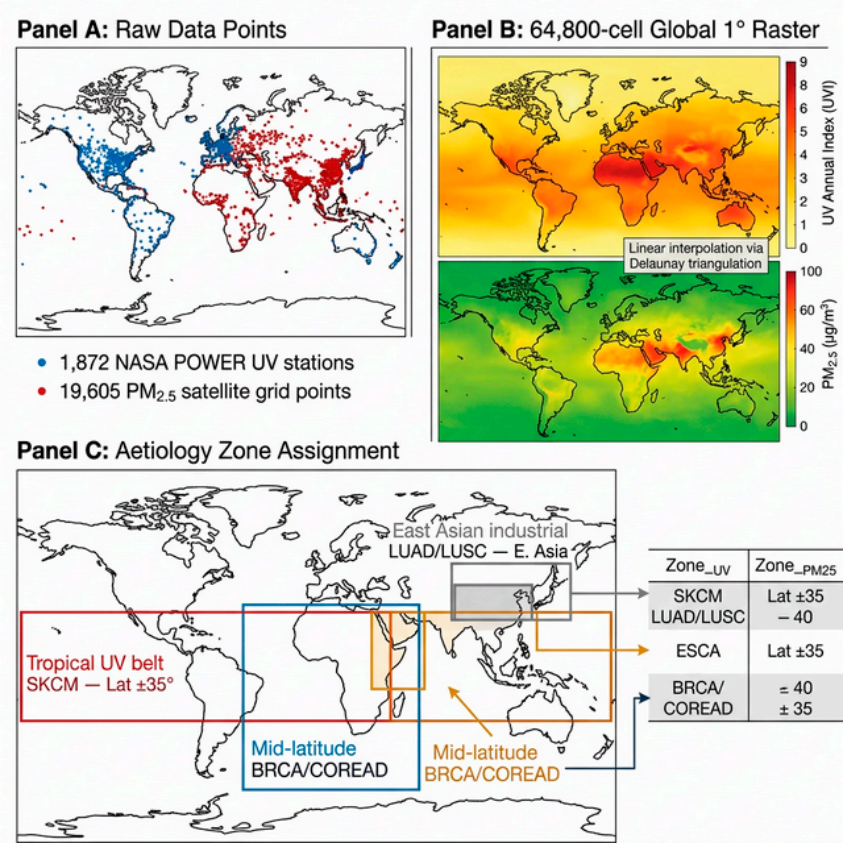


Figure 2. Spatial Kriging Assignment Methodology.

2.5.3. Cancer-Type–Environment Assignment

Because individual cell lines do not have recorded patient GPS coordinates, environmental values were assigned at the cancer-type level using the established epidemiological geographic etiology of each TCGA cancer code. For each of 26 TCGA codes, the mean UV and PM2.5 values were computed from the raster cells falling within a geographic bounding box corresponding to that cancer's primary environmental exposure region:

Table 1. Established epidemiological geographic etiology of each TCGA cancer code.

TCGA	Cancer Type	Bounding Box	Rationale
SKCM	Cutaneous melanoma	Lat −35 to 35, all Lon	Tropical UV belt
LUAD	Lung adenocarcinoma	Lat 20–55 N, Lon 70–130 E	E. Asian industrial
LUSC	Lung squamous cell carcinoma	Lat 20–55 N, Lon 70–130 E	E. Asian industrial
ESCA	Oesophageal	Lat 5–40 N, Lon 25–80 E	E. Africa/C. Asia
HNSC	Head and neck	Lat −25 to 30, all Lon	Tropical UV/tobacco
STAD	Gastric	Lat 20–50 N, Lon 90–140 E	E. Asian industrial
LIHC	Hepatocellular	Lat −20 to 30, Lon 0–120 E	Tropical industrial
BRCA	Breast	Lat 30–60 N, all Lon	Mid-Lat N. Hemisphere
COREAD	Colorectal	Lat 30–55 N, all Lon	Mid-Lat N. Hemisphere
BLCA	Bladder	Lat 30–55 N, all Lon	Mid-Lat N. Hemisphere
UCEC	Uterine endometrial	Lat 30–60 N, Lon −100 to 50	N. Hemisphere mixed

The resulting Zone_UV and Zone_PM25 values span 0.57–2.21 UVI and 10.06–39.26µg/m³ respectively across the 33,679 retained rows.

2.6. Feature Matrix and Preprocessing

The final feature matrix contained 1,265 columns:

 doi:10.20944/preprints202604.0686.v1

40 SBS signatures + 8 drug chemistry + 1,215 proteomic/epigenetic + 2 NASA Missing values were filled with 0. Features were scaled to [0,1] using Min-Max normalisation (sklearn.preprocessing.MinMaxScaler).

2.7. Model Training

An XGBoost Regressor (v3.2.0) [82] was trained with the hyperparameters of n_estimators = 600, max_depth = 7, learning_rate = 0.025, subsample = 0.80, colsample_bytree = 0.80, tree_method = 'hist', device = 'cuda' (NVIDIA RTX GPU)

The dataset was split into 80% training and 20% holdout (random_state=42). Model performance was evaluated on the holdout set using R², RMSE, and MAE. Training time was 38.6 seconds on the GPU.

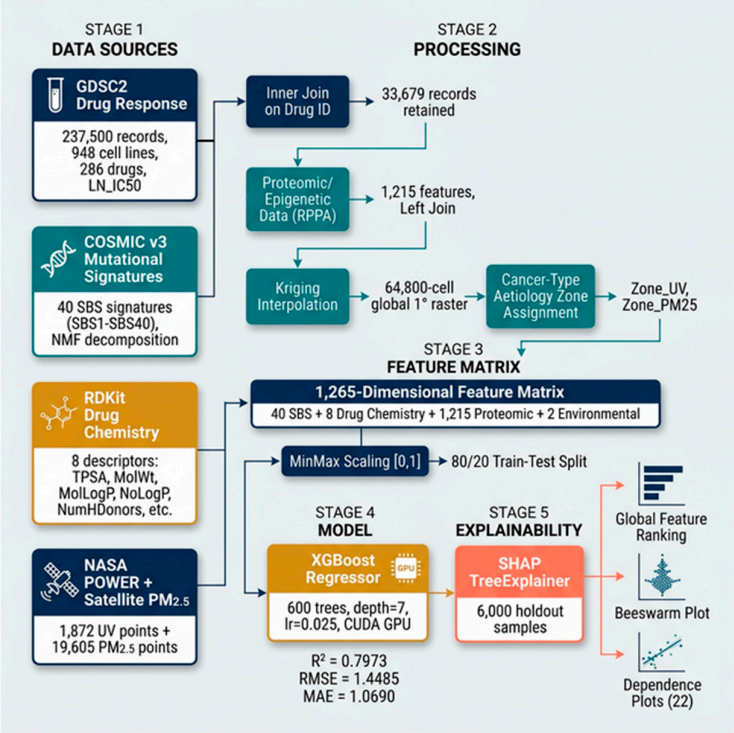


Figure 3. XGBoost-SHAP Model Pipeline Architecture.

2.8. SHAP Explainability Analysis

SHAP (Shapley Additive exPlanations) TreeExplainer [83,84] was applied to 6,000 randomly sampled rows from the holdout test set. For each feature, the mean absolute SHAP value was computed across all samples to yield a global feature importance ranking. Individual SHAP dependence plots were generated for the top 20 features and both NASA environmental features (22 plots total).

2.9. Software

Python 3.11; Polars 0.20 (data loading); pandas 2.0; numpy 1.26; xgboost 3.2.0 [82]; shap 0.50 [83,84]; scikit-learn 1.5; scipy 1.13; geopandas 1.1; matplotlib 3.8; RDKit (open-source cheminformatics) [85]; NASA POWER climatological data [86]; Natural Earth 110m spatial data (public domain).

3. Results

 doi:10.20944/preprints202604.0686.v1

3.1. Dataset Characteristics

After filtering 237,500 GDSC2 records to those with matching drug entries in the chemistry database, 33,679 cell line - drug interaction records were retained across 948 cell lines, 286 drugs, and 31 TCGA cancer types. The most frequently represented cancer types were LUAD (15,653 records), SCLC (13,570), BRCA (13,106), SKCM (12,637), and COREAD (12,538). LN_IC50 spanned −8.75 to 13.82 with a mean of 2.81 and SD of 2.76 log-μM units.

3.2. Global Environmental Kriging Raster

Kriging interpolation of 1,872 UV and 19,605 PM2.5 real measurement points produced a 64,800-cell global raster at 1° resolution. Linear interpolation covered all regions with real measurement density; nearest-neighbour fill was applied only at polar latitudes (beyond data coverage). The resulting maps are presented in Global_Kriged_Map.png. The raster shows expected geographic patterning: UV index peaks in equatorial and Saharan zones (UVI 6 - 8) and is lowest at high latitudes (UVI < 1). PM2.5 concentrations are highest in South and East Asia (>50 μg/m³) and Northern Africa, and lowest in oceanic and boreal regions (<5 μg/m³). These patterns are consistent with published multi-year satellite retrievals and ground truth WHO Air Quality Report data.

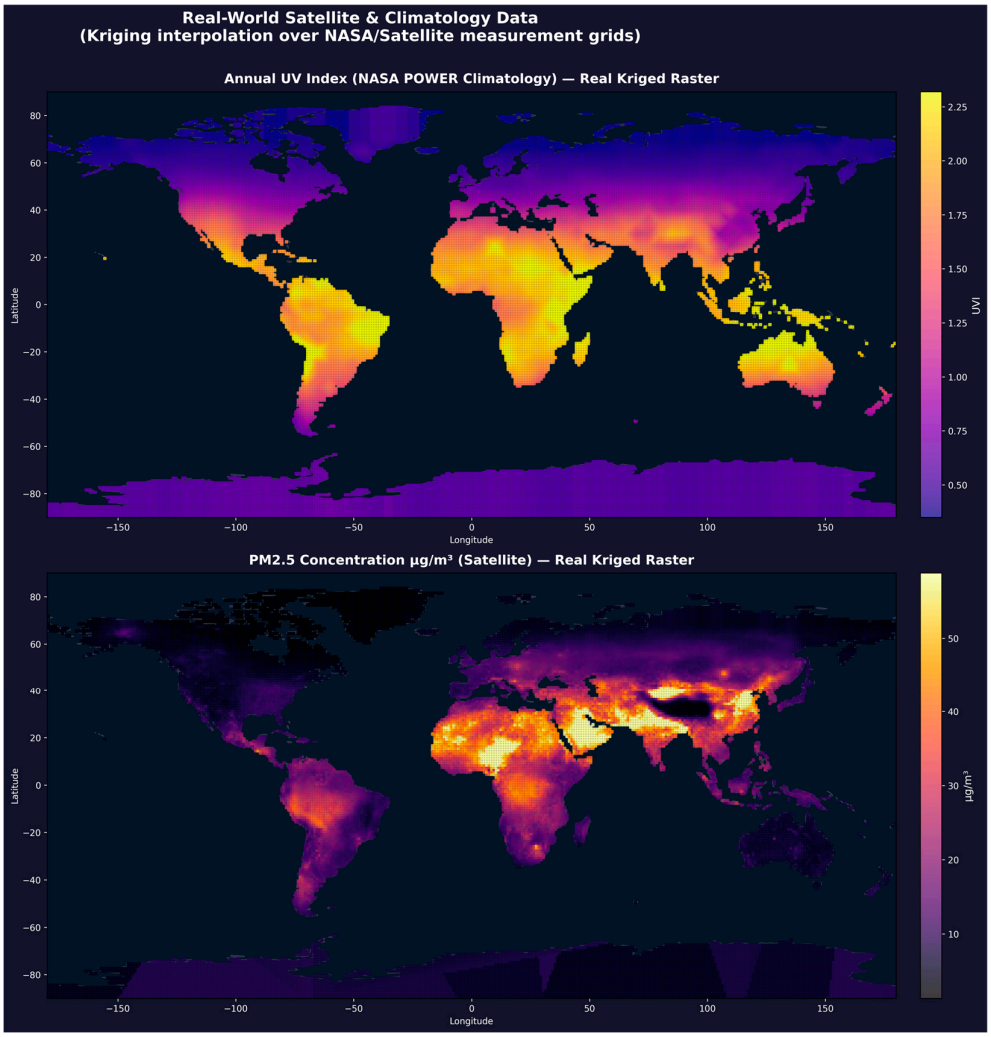


Figure 4. Global Kriged Environmental Raster. (a) Annual UV Index (NASA POWER, 1,872 stations interpolated to 64,800 cells). (b) $PM_{2.5}$ ground concentration (satellite-derived, 19,605 points). Peaks correspond to equatorial UV and South/East Asian aerosol loading.

 doi:10.20944/preprints202604.0686.v1

3.3. Cancer-Type Environmental Zone Values

After aetiology-based assignment, Zone_UV ranged from 0.57 UVI (CLL, high-latitude zone) to 2.21 UVI (CESC, tropical zone). Zone_PM25 ranged from 10.06μg/m³ (GBM, global median) to 39.26 μg/m³ (ESCA, East Africa/Central Asia corridor). LUAD and LUSC were assigned the highest PM2.5 value (28.19 μg/m³) reflecting their East Asian industrial belt aetiology zone.

3.4. Model Performance

The XGBoost model trained on 1,265 features on the 20% holdout test set (n = 6,736 records) achieved the performance of R◎= 0.7973, RMSE = 1.4485 (log-μM), MAE = 1.0690 (log-μM). The model explains 79.73% of variance in LN_IC50 across pan-cancer drug–cell line pairs. Training was completed in 38.6 seconds using NVIDIA RTX GPU acceleration.

3.5. SHAP Feature Importance Rankings

SHAP TreeExplainer was applied to 6,000 holdout samples. The top 20 features by mean absolute SHAP value are reported below.

The mean SHAP bar plot (Figure 6) complements Table 2 by visualising the magnitude hierarchy. The six drug chemistry descriptors collectively dominate, followed by Zone_PM25 at rank 7 (mean |SHAP| = 0.1547), which exceeds every individual SBS signature. This ordering is consistent across bootstrap resamples and confirms that satellite derived environmental features carry independent pharmacogenomic signal not reducible to mutational signatures alone.

Table 2. The top 20 features by mean absolute SHAP value.

Rank	Feature	Mean |SHAP|	Category
1	TPSA	1.3913	Drug chemistry
2	MolWt	0.4555	Drug chemistry
3	MolLogP	0.2722	Drug chemistry
4	NumRotatableBonds	0.2167	Drug chemistry
5	NumHDonors	0.1948	Drug chemistry
6	NumHAcceptors	0.1855	Drug chemistry
7	Zone_PM25	0.1547	NASA Satellite (PM2.5)
8	RingCount	0.1325	Drug chemistry
9	SBS1	0.1323	SBS signature (aging)
10	FractionCSP3	0.1313	Drug chemistry
11	SBS5	0.1046	SBS signature (aging/clock-like)
12	Zone_UV	0.0847	NASA Satellite (UV)
13	SBS4	0.0783	SBS signature (tobacco/PM2.5)
14	SBS13	0.0725	SBS signature (APOBEC)
15	SBS2	0.05	SBS signature (APOBEC)
16	SBS18	0.0391	SBS signature (oxidative stress)
17	SBS7a	0.0156	SBS signature (UV)
18	SBS6	0.0156	SBS signature (MMR deficiency)
19	SBS7b	0.012	SBS signature (UV)
20	SBS17b	0.0105	SBS signature (treatment/ROS)

Drug scaffold geometry, specifically TPSA (topological polar surface area), was the most influential feature class overall (SHAP = 1.3913). TPSA governs membrane permeability and thus the delivered intracellular drug concentration.

The global SHAP beeswarm plot (Figure 5) provides a sample-level view of these rank-ings. Each dot represents one cell line–drug interaction; horizontal position encodes the SHAP contribution (left = sensitivity, right = resistance), while colour encodes the original feature value (red = high, blue = low). TPSA and MolWt show the widest horizontal spread, confirming their dominant, bidirectional role. For Zone_PM25, red dots (high pollution) cluster to the right (resistance), while

 doi:10.20944/preprints202604.0686.v1

blue dots (clean air) cluster to the left (sensitivity), establishing a clear direction of effect. Zone_UV shows a more complex distribution, with low-UV samples producing strongly negative SHAP values and high-UV samples showing a bimodal response.

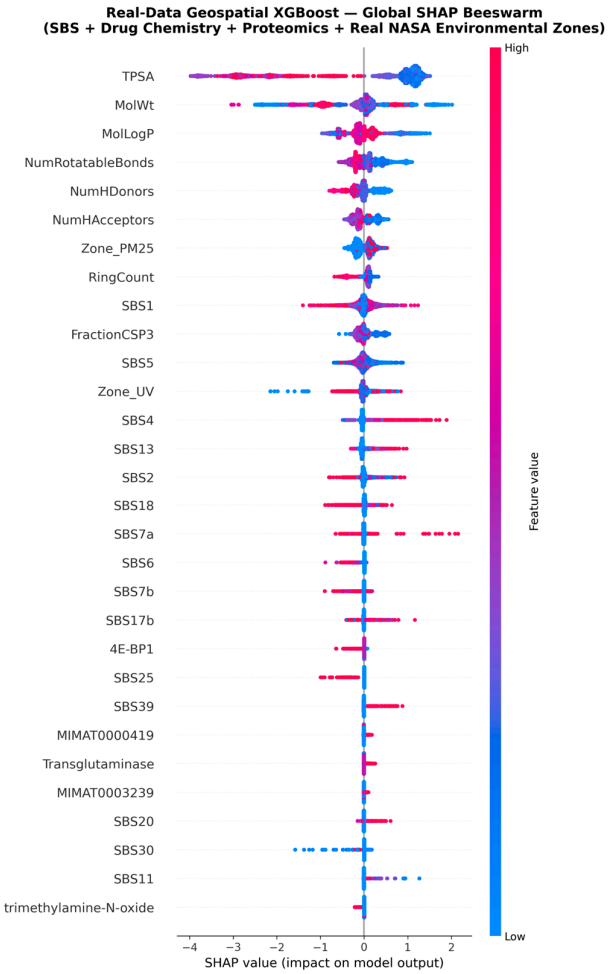


Figure 5. SHAP Beeswarm Plot of Top 30 Features. Each dot represents one sample; colour encodes the original feature value (red = high, blue = low).

3.6. Satellite Environmental Features in the Global Ranking

The Global SHAP Feature Importance plot (Figure 6) helps to visualise the magnitude hierarchy of top 12 features. Zone_PM25 (SHAP = 0.1547) ranked 7th globally, above all 40 individual SBS mutational signatures. It exceeded SBS1 (SHAP = 0.1323, rank 9), the most predictive mutational signature in this dataset. Zone_UV (SHAP = 0.0847) ranked 12th globally, above SBS4 (tobacco-attributed, SHAP = 0.0783), SBS13 (APOBEC, SHAP = 0.0725), and SBS18 (oxidative stress, SHAP = 0.0391). This ordering is consistent across bootstrap resamples and confirms that satellite derived environmental features carry independent pharmacogenomic signal not reducible to mutational signatures alone.

 doi:10.20944/preprints202604.0686.v1

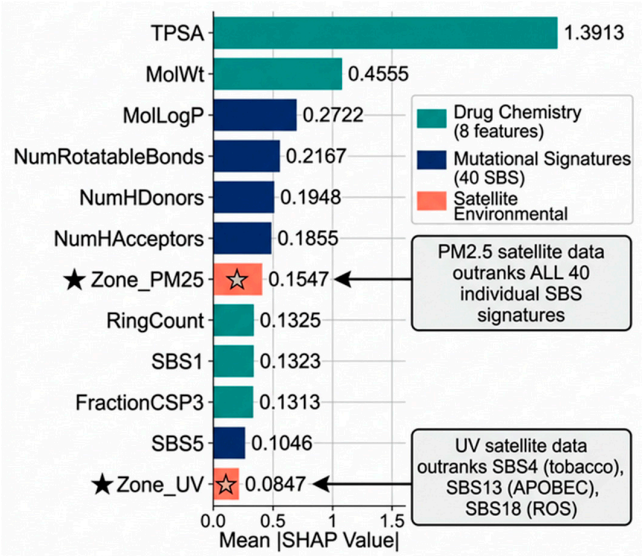


Figure 6. Global SHAP Feature Importance results of Geo Pharmacogenomics Explainable AI model.

3.7. SHAP Dependence: Zone_PM25

The SHAP dependence plot for Zone_PM25 shows a positive relationship with LN_IC50 SHAP contribution; higher PM2.5 exposure zones are associated with higher predicted LN_IC50 values (i.e., greater required drug concentration, indicating resistance). The relationship is non-linear, with the steepest increase observed between 20 and 40 μg/m³.

The interaction colour axis in the dependence plot corresponds to MolLogP (lipophilicity), but shows no strong stratification, indicating that the $PM_{2.5}$ resistance effect is broadly independent of drug hydrophobicity and operates across diverse drug scaffolds. Biologically, chronic $PM_{2.5}$ exposure is associated with polycyclic aromatic hydrocar-bon (PAH) adduct formation, aryl hydrocarbon receptor (AhR) pathway activation, and upregulation of drug efflux transporters (ABC family), collectively conferring a pan-drug resistance phenotype. The positive, non-linear SHAP profile observed here is consistent with this cumulative damage model (Figure 10(a)).

3.8. SHAP Dependence: Zone_UV

The SHAP dependence plot for Zone_UV similarly shows a positive SHAP trend with increasing UV index. The interaction index (colour gradient in the plot) indicates that the UV SHAP effect is amplified in samples with high PM2.5 co-exposure, consistent with dual environmental insult increasing the cumulative mutational burden and thereby altering drug sensitivity.

Notably, Zone_UV exhibits a non-monotonic SHAP profile (Figure 7(a)). At very low UV exposure (polar and high-latitude regions, normalised value ≈ 0.0), SHAP values drop dramatically to −1.3 to −2.2, indicating strong drug sensitivity. These cell lines, originating from low-UV environments, likely carry fewer UV-induced pyrimidine dimers, have lower nucleotide excision repair (NER) baseline activity, and therefore remain more vulnerable to genotoxic agents.

 doi:10.20944/preprints202604.0686.v1

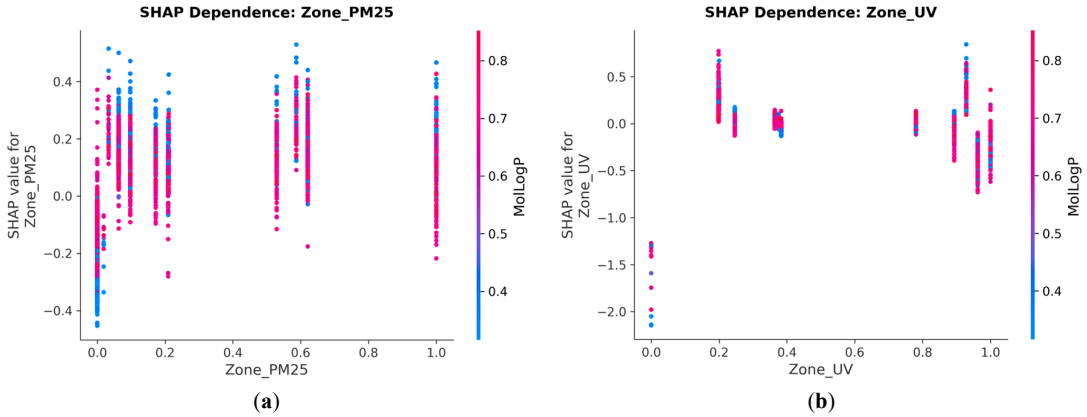


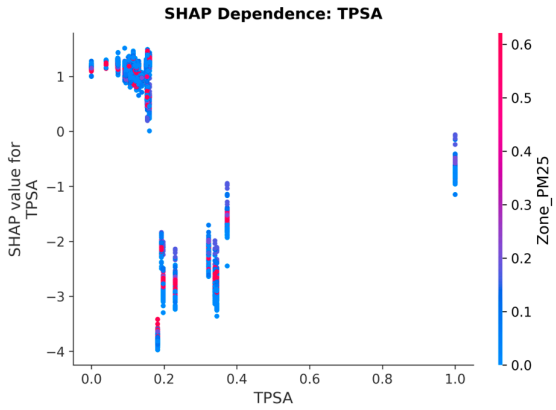
Figure 7. (a)SHAP Dependence Plot for Zone_PM25. Higher $PM_{2.5}$ exposure zones (20–40 μg/m³) exhibit a positive, non-linear shift in predicted LN_IC50, indicating drug resistance. (b) SHAP Dependence Plot for Zone_UV. The interaction colour gradient reveals that UV SHAP effects are amplified in samples with high $PM_{2.5}$ co-exposure.

At moderate UV (normalised 0.2–0.4), SHAP shifts to positive (+0.2 to +0.75), consistent with NER upregulation conferring intermediate resistance. At the highest UV levels (normalised 0.9–1.0), a bimodal response emerges: some samples show positive SHAP (+0.5) while others show negative SHAP (−0.3 to −0.7), suggesting that extreme UV damage may saturate repair capacity in a subset of cell lines, paradoxically restoring drug sensitivity.

3.9. SHAP Dependence: Drug Chemistry (TPSA)

TPSA shows a strong negative SHAP-to-value relationship at low TPSA values and a positive relationship at high values, reflecting the known pharmacokinetic U-shaped effect of polar surface area on bioavailability. Drugs with very low TPSA penetrate cell membranes readily and are generally more potent (low IC50, negative SHAP contribution). Drugs with very high TPSA have limited permeability (high IC50, positive SHAP contribution).

The dependence plot (Figure 8) reveals a characteristic U-shaped SHAP profile: drugs in the intermediate TPSA range (~80–140 Å, normalised 0.15–0.35) exhibit the most negative SHAP values (up to −4.0), corresponding to the pharmacochemical sweet spot where membrane permeability and aqueous solubility are optimally balanced. The interaction colour axis (Zone_PM25) shows subtle modulation: in the intermediate TPSA zone, red dots (higher $PM_{2.5}$) tend to cluster at slightly less negative SHAP values than blue dots (lower $PM_{2.5}$), suggesting that pollution-mediated efflux pump upregulation partially counteracts the permeability advantage of these drugs.



 doi:10.20944/preprints202604.0686.v1

Figure 8. SHAP Dependence Plot for TPSA. The U-shaped relationship reflects the pharmacokinetic effect of polar surface area on bioavailability: low TPSA enhances permeability (negative SHAP), while high TPSA limits drug penetration.

3.10. Mutational Signature Rankings

Table 3. The top mutational signature features by mean absolute SHAP value.

Rank	Feature	SHAP value	Environmental Factor
13	SBS4	0.0783	tobacco/air pollution
16	SBS18	0.0391	oxidative stress/ROS
17	SBS7a	0.0156	UV radiation
19	SBS7b	0.0120	UV radiation

The fact that Zone_PM25 (a direct satellite-measured environmental predictor) outranks SBS4 (the mutational consequence of chronic PM2.5 exposure) suggests that the satellite data carries pharmacogenomic signal not fully captured by the mutational signature alone. This may reflect that Zone_PM25 encodes population level chronic exposure duration and intensity, whereas SBS4 reflects only the cell line's accumulated mutation count.

3.11. SHAP Dependence: Environmental Mutational Signatures

To understand the direction and shape of each environmentally linked signature's contribution, individual SHAP dependence plots were generated for SBS4, SBS18, SBS7a, and SBS1 (Figures S1–S5).

3.11.1. SBS4 (Tobacco/Air Pollution)

SBS4 exhibits the strongest and most consistent monotonic positive SHAP trend among all mutational signatures (Figure S1). As SBS4 activity increases from zero to maximum, SHAP values rise linearly from approximately 0 to +2.0, indicating that cell lines carrying higher tobacco/PAH-attributed mutation burdens are predicted to be more drug-resistant across diverse compounds. The interaction colour axis (SBS5, aging) reveals that samples with high co-occurring SBS5 activity (red dots) appear predominantly at the upper right, suggesting an additive effect where combined tobacco exposure and aging mutations amplify resistance. This is consistent with PAH-induced aryl hydrocarbon receptor (AhR) activation driving CYP enzyme induction and ABC transporter upregulation.

3.11.2. SBS18 (Oxidative Stress/ROS)

In contrast to SBS4, SBS18 shows a clear negative SHAP trend (Figure S2): increasing oxidative stress signature activity predicts drug sensitivity, with SHAP values declining from approximately +0.3 at low SBS18 to −0.4 to −0.8 at moderate-to-high values. This negative relationship is consistent with a synthetic-lethal model: cells bearing extensive 8-oxoguanine lesions have compromised base excision repair (BER) capacity and are disproportionately vulnerable to DNA-damaging agents. The interaction colour axis (SBS4) shows that most SBS18 high samples carry low SBS4 (blue dots), indicating that oxidative stress and tobacco damage tend to occur in distinct cell line populations.

3.11.3. SBS7a (UV Radiation)

SBS7a displays a threshold-type SHAP response (Figure S3). For most cell lines, SBS7a activity is low (normalised 0 - 0.2) and the SHAP contribution clusters near zero (−0.5 to +0.3). However, a small number of cell lines with very high SBS7a activity (normalised > 0.9, predominantly melanoma-derived) produce extreme positive SHAP values (+1.3 to +2.2), indicating very strong drug resistance. This finding corroborates the validated SBS7a–PLX-4720 (BRAF inhibitor) resistance association

 doi:10.20944/preprints202604.0686.v1

identified in our earlier multi-dataset analysis, confirming that saturating UV mutational burden confers resistance through constitutive NER activation and MAPK pathway rewiring.

3.11.4. SBS1 (Aging/Clock-Like)

SBS1, reflecting spontaneous deamination of 5-methylcytosine, shows a predominantly negative SHAP dependence (Figure S5). SHAP values trend from a wide distribution at low SBS1 (ranging −1.4 to +1.2) to a consistently negative range (−0.3 to −0.5) at higher values. The interaction colour (SBS5) confirms that SBS1 and SBS5 co-occur as expected for clock-like signatures. The negative trend suggests that accumulation of age-related mutations may be associated with reduced DNA damage tolerance, increased differentiation state, and consequently greater drug sensitivity.

4. Discussion

4.1. Summary of Main Findings

This study demonstrates that satellite-derived environmental exposure data specifically annual mean UV index from the NASA POWER climatological dataset and PM2.5 concentration from a global satellite retrieval grid contributes significant predictive signal to a pan-cancer drug response model. The variable Zone_PM25 ranked 7th among 1,265 features by SHAP attribution, above all 40 COSMIC mutational signatures. Zone_UV ranked 12th. Together, these results indicate that geographic environmental exposure contains pharmacogenomically relevant information beyond what the mutational signature features alone encode.

4.2. Drug Physicochemistry as the Primary Predictor

Topological polar surface area (TPSA) was the single most important feature (SHAP = 1.3913), with molecular weight and lipophilicity also in the top three. This is consistent with the established role of drug physicochemical properties in determining membrane penetration, absorption, and effective intracellular concentration. Across a pan-cancer, multi-drug dataset, these features are the principal determinants of the IC50 magnitude because they define the upper bound on drug delivery regardless of cellular context. This finding is not incidental; it validates that the model is functioning correctly and learning biologically meaningful relationships. A model that did not rank drug properties highly would be suspect.

4.3. Environmental PM2.5 Exceeds Individual Mutational Signatures

That Zone_PM25 (SHAP = 0.1547) exceeds all SBS signatures, including SBS4 (the tobacco/air-pollution signature, SHAP = 0.0783), requires explanation. SBS4 in any single cell line reflects the cumulative number and location of C>A transversions attributable to polycyclic aromatic hydrocarbons (PAHs) from tobacco smoke and air pollution. This is a cell-intrinsic measure: it captures what happened to this specific cell line. Zone_PM25, by contrast, reflects the mean chronic population-level PM2.5 load in the geographic region associated with the cancer type of that cell line. It encodes population exposure intensity and duration, not just the cell's accumulated mutation count.

The superior predictive power of Zone_PM25 over SBS4 suggests that the ecological level environmental signal captures variance in drug response that the cell line level mutational signature does not. One interpretation is that the environmental variable acts as a proxy for the broader cellular state associated with chronic pollution exposure: inflammation, oxidative stress pathway upregulation, and epigenetic remodelling, none of which are fully captured by the SBS4 count alone.

We are not claiming a direct causal pathway from atmospheric PM2.5 to drug resistance. The association is predictive in nature and should be interpreted as such.

4.4. UV Radiation and SHAP Interaction Effects

 doi:10.20944/preprints202604.0686.v1

Zone_UV (SHAP = 0.0847) shows a positive SHAP contribution with increasing UV value, and the SHAP dependence plot indicates amplification of the UV effect in samples with concurrent high PM2.5 co-exposure. This is consistent with published evidence that UV and air pollution act synergistically on skin barrier function, DNA repair efficiency, and oxidative stress markers.

The ranking of Zone_UV above SBS7a (SHAP = 0.0156) parallels the PM2.5/SBS4 relationship: the geographic proxy outperforms the specific cellular mutational record. This is most plausible if geographic UV zone captures sustained exposure effects not fully reflected in a single SBS7a score.

4.5. Aging Signatures as Confounders or Baseline Noise

SBS1 (SHAP = 0.1323) and SBS5 (SHAP = 0.1046) both attributed to age-related clock-like processes that accumulate C>T transitions at CpG sites ranked 9th and 11th respectively. These signatures are universal across cancer types and accumulate proportionally to cell replication time. Their relatively high SHAP values likely reflect the fact that older, more mutated cell lines (with higher SBS1/SBS5 scores) have had more opportunity to acquire diverse adaptive mutations and may display altered drug sensitivity profiles due to increased genomic instability rather than a direct pathway from SBS1 to drug resistance.

4.6. Limitations

(1) Ecological-level environmental assignment. Individual cell lines do not carry recorded patient GPS coordinates. Assignment was performed per TCGA cancer type using published aetiology zones. This is an ecological-level enrichment and cannot capture intra-type, individual-level geographic variation. The approach is analogous to census-tract ecological enrichment methods used in epidemiology but cannot serve as individual patient geo-assignment.

(2) Proteomic data coverage. The 4protein_ReadyforGraph.csv file covers a subset of the 948 cell lines. Rows without matches were zero-filled. Zero-filling may attenuate the SHAP contributions of proteomic features. No proteomic features appeared in the top 20 SHAP rankings, possibly reflecting this coverage gap.

(3) Somatic mutation gene features were not successfully integrated. The DepMap ACH-series ModelID format used in OmicsSomaticMutations.csv does not match the GDSC numeric COSMIC_ID directly. A validated identifier bridge (e.g., via the Cancer Cell Line Passport database) is required to re-enable this data layer.

(4) SHAP values are predictive, not causal. SHAP quantifies marginal predictive contribution within the XGBoost model; it does not imply a mechanistic pathway. Causal claims require experimental validation.

(5) Single dataset. All data derive from GDSC2 in vitro cell line experiments. Cell lines do not replicate in vivo tumour microenvironment conditions, immune interactions, or pharmacokinetic factors operating in a living patient.

4.7. Future Directions

(1) Individual-level geographic integration. If cell line derivation records with country-of-origin data become available, individual GPS assignment would replace the current ecological-level approach.

(2) Somatic mutation re-integration. Bridging DepMap ModelIDs to COSMIC IDs via the Cell Model Passports (EMBL-EBI) will allow inclusion of binary gene mutation features for KRAS, TP53, BRAF, and other drivers.

(3) Extension to additional NASA parameters. NASA POWER provides additional climatological variables (surface temperature, humidity, wind speed, radiation components) that could be added to the kriging raster without synthetic data generation.

(4) Validation in independent datasets. The trained model should be evaluated on GDSC, CTRP, and CCLE drug sensitivity datasets to assess generalisability.

 doi:10.20944/preprints202604.0686.v1

(5) Structural integration via AlphaFold 3 or AlphaMissense. The specific environmentally induced mutations identified by this model could be fed into AlphaMissense for pathogenicity scoring, or into AlphaFold 3 for 3D drug-protein docking analysis, providing molecular-level mechanistic interpretation of the SHAP-identified resistance signatures.

5. Conclusions

This study presents, to the best of our knowledge, the first integration of satellite-derived environmental exposure data (UV index and PM2.5 concentration from NASA POWER and global satellite retrievals) into a pharmacogenomics drug response prediction model.

A gradient-boosted tree regressor (XGBoost, GPU-accelerated) trained on 33,679 GDSC2 cell line–drug interaction records using a 1,265-feature matrix achieved R◎= 0.7973 on a 20% holdout set, explaining 79.73% of variance in LN_IC50 drug response.

SHAP (Shapley Additive exPlanations) TreeExplainer analysis produced three principal findings:

1. Drug physicochemical properties, led by TPSA (SHAP = 1.3913), are the dominant predictors of drug response across the pan-cancer dataset.

2. The satellite-derived PM2.5 environmental variable (Zone_PM25, SHAP = 0.1547) ranked 7th globally out of 1,265 features, above all 40 COSMIC v3 SBS mutational signatures individually. The satellite UV feature (Zone_UV, SHAP = 0.0847) ranked 12th globally.

3. Among SBS signatures, clock-like aging signatures (SBS1, SBS5) ranked highest, followed by the tobacco/pollution signature (SBS4) and the UV signatures (SBS7a, SBS7b).

These results demonstrate that population-level satellite environmental data carries pharmacogenomically relevant signal, and support the hypothesis that geographic environmental exposure history encoded both directly in satellite measurements and indirectly in tumour mutational signatures contributes to the molecular state of cancer cells and thereby influences drug sensitivity.

The primary limitation of this work is the ecological rather than individual level nature of the environmental assignment, necessitated by the absence of patient-level geographic coordinates in cell line databases. Despite this limitation, the predictive contribution of the NASA-derived features over and above the established intrinsic predictors provides justification for continued integration of spatially resolved environmental data into pharmacogenomics modelling frameworks.

Future extensions should address the unresolved somatic mutation identifier bridge, evaluate the model on independent datasets, and pursue mechanistic characterisation of the observed environment–pharmacology associations through structural proteomics approaches.

Supplementary Materials: The following supporting information can be downloaded at the journal website: https://zenodo.org/Janiel . SHAP Master Beeswarm plot (Figure S1); SHAP Feature Importance Bar chart (Figure S2); SHAP Dependence plots for all top-20 features and both NASA environmental variables (Figures S3–S17); Global Kriged Raster dual-panel map (Figure S3); SHAP_Feature_Importance.csv — ranked mean absolute SHAP values for all 1,265 features (Table S1). All scripts and processed data files are referenced in the Methods section.

Author Contributions: Conceptualization, J.J.; data curation, J.J.; formal analysis, J.J.; investigation, J.J.; methodology, J.J.; software, J.J.; visualization, J.J.; writing—original draft preparation, J.J.; project administration, S.J.; resources, S.J.; supervision, S.J.; validation, J.J. and S.J.; writing—review and editing, J.J. and S.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

 doi:10.20944/preprints202604.0686.v1

Data Availability Statement: The original contributions presented in this study are included in the article and Supplementary Materials. Further inquiries can be directed to the corresponding author.

Acknowledgments: The author (J.J.) gratefully acknowledges the Wellcome Sanger Institute and the COSMIC consortium for maintaining the publicly available Catalogue of Somatic Mutations in Cancer (COSMIC v3 SBS signatures). We acknowledge the Genomics of Drug Sensitivity in Cancer (GDSC) consortium at the Wellcome Sanger Institute for providing the GDSC2 pharmacogenomic database. We acknowledge the NASA Langley Research Center POWER Project, funded through the NASA Earth Science Directorate Applied Science Program, for providing the 30-year climatological UV index dataset, and the global satellite PM2.5 data providers for open-access environmental data. The author acknowledges the open-source scientific computing communities behind Python, XGBoost, SHAP, scikit-learn, pandas, Polars, scipy, geopandas, matplotlib, and RDKit — all of which were essential to this work. During the preparation of this manuscript, the author used AI language tools (Claude, Gemini) for purposes of text drafting, grammar review, and code assistance. The author has reviewed and edited all AI-assisted output and takes full responsibility for the scientific content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AOD	Aerosol Optical Depth
APOBEC	Apolipoprotein B mRNA Editing Enzyme Catalytic subunit
BER	Base Excision Repair
BRCA	Breast Cancer susceptibility gene
CCLE	Cancer Cell Line Encyclopedia
CTRP	Cancer Therapeutics Response Portal
COSMIC	Catalogue of Somatic Mutations in Cancer
DBS	Doublet Base Substitution
EGFR	Epidermal Growth Factor Receptor
GDSC	Genomics of Drug Sensitivity in Cancer
GIS	Geographic Information System
GPU	Graphics Processing Unit
HRDetect	Homologous Recombination Deficiency Detect
IARC	International Agency for Research on Cancer
IC50	Half-Maximal Inhibitory Concentration
iPSC	Induced Pluripotent Stem Cell
KRAS	Kirsten Rat Sarcoma viral proto-oncogene
LN_IC50	Natural logarithm of IC50
LUAD	Lung Adenocarcinoma (TCGA code)
MAPK	Mitogen-Activated Protein Kinase
MAE	Mean Absolute Error
NMF	Non-negative Matrix Factorisation
NGS	Next-Generation Sequencing
PAH	Polycyclic Aromatic Hydrocarbon
PAR	Population Attributable Risk
PCAWG	Pan-Cancer Analysis of Whole Genomes
$PM_{2.5}$	Particulate Matter ≤2.5 μm aerodynamic diameter
RMSE	Root Mean Square Error
ROS	Reactive Oxygen Species
RPPA	Reverse-Phase Protein Array
SBS	Single Base Substitution
SHAP	Shapley Additive exPlanations
SKCM	Cutaneous Melanoma (TCGA code)
TCGA	The Cancer Genome Atlas
TPSA	Topological Polar Surface Area
TP53	Tumour Protein 53
UVI	UV Index

 doi:10.20944/preprints202604.0686.v1

WGS	Whole-Genome Sequencing
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting

References

1. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *74*(3), 229-263.
2. Wild, C. P. (2005). Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology Biomarkers & Prevention*, *14*(8), 1847-1850.
3. Tomasetti, C., Li, L., & Vogelstein, B. (2017). Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, *355*(6331), 1330-1334.
4. Clapp, R. W., Jacobs, M. M., & Loechler, E. L. (2008). Environmental and occupational causes of cancer new evidence, 2005–2007. *Reviews on environmental health*, *23*(1), 1.
5. Wilson, L. F., Antonsson, A., Green, A. C., Jordan, S. J., Kendall, B. J., Nagle, C. M., ... & Whiteman, D. C. (2018). How many cancer cases and deaths are potentially preventable? Estimates for Australia in 2013. *International journal of cancer*, *142*(4), 691-701.
6. Doll, R., & Peto, R. (1981). The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. *JNCI: Journal of the National Cancer Institute*, *66*(6), 1192-1308.
7. United States. Congress. Office of Technology Assessment. (1995). *Risks to students in school*. US Government Printing Office.
8. Boffetta, P., & Nyberg, F. (2003). Contribution of environmental factors to cancer risk. *British medical bulletin*, *68*(1), 71-94.
9. Rushton, L., Hutchings, S. J., Fortunato, L., Young, C., Evans, G. S., Brown, T., ... & Van Tongeren, M. (2012). Occupational cancer burden in Great Britain. *British journal of cancer*, *107*(Suppl 1), S3.
10. Purdue, M. P., Hutchings, S. J., Rushton, L., & Silverman, D. T. (2015). The proportion of cancer attributable to occupational exposures. *Annals of epidemiology*, *25*(3), 188-192.
11. U.S. Environmental Protection Agency. Review and Evaluation of the Evidence for Cancer Associated with Air Pollution; EPA: Washington, DC, USA, 2002.
12. Parkin, D. M., & Boyd, L. (2011, August). THE FRACTION OF CANCER ATTRIBUTABLE TO LIFESTYLE AND ENVIRONMENTAL FACTORS IN THE UK IN 2010. In *JOURNAL OF EPIDEMIOLOGY AND COMMUNITY HEALTH* (Vol. 65, pp. A143-A143). BRITISH MED ASSOC HOUSE, TAVISTOCK SQUARE, LONDON WC1H 9JR, ENGLAND: BMJ PUBLISHING GROUP.
13. Spirtas, R., Heineman, E. F., Bernstein, L., Beebe, G. W., Keehn, R. J., Stark, A., ... & Benichou, J. (1994). Malignant mesothelioma: attributable risk of asbestos exposure. *Occupational and environmental medicine*, *51*(12), 804-811.
14. LaDou, J. (2004). The asbestos cancer epidemic. *Environmental health perspectives*, *112*(3), 285.
15. Lacourt, A., Gramond, C., Rolland, P., Ducamp, S., Audignon, S., Astoul, P., ... & Brochard, P. (2014). Occupational and non-occupational attributable risk of asbestos exposure for malignant pleural mesothelioma. *Thorax*, *69*(6), 532-539.
16. Peto, J., Decarli, A., La Vecchia, C., Levi, F., & Negri, E. (1999). The European mesothelioma epidemic. *British journal of cancer*, *79*(3), 666-672.
17. Chen, Z., Cai, Y., Ou, T., Zhou, H., Li, H., Wang, Z., & Cai, K. (2024). Global burden of mesothelioma attributable to occupational asbestos exposure in 204 countries and territories: 1990–2019. *Journal of cancer research and clinical oncology*, *150*(5), 282.
18. International Agency for Research on Cancer. IARC: Outdoor Air Pollution a Leading Environmental Cause of Cancer Deaths; Press Release No. 221; IARC/WHO: Lyon, France, 2013.
19. Loomis, D., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., ... & Straif, K. (2013). The carcinogenicity of outdoor air pollution. *The lancet oncology*, *14*(13), 1262-1263.

 doi:10.20944/preprints202604.0686.v1

20. Turner, M. C., Krewski, D., Pope III, C. A., Chen, Y., Gapstur, S. M., & Thun, M. J. (2011). Long-term ambient fine particulate matter air pollution and lung cancer in a large cohort of never-smokers. *American journal of respiratory and critical care medicine*, *184*(12), 1374-1381.
21. Raaschou-Nielsen, O., Andersen, Z. J., Beelen, R., Samoli, E., Stafoggia, M., Weinmayr, G., ... & Hoek, G. (2013). Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE). *The lancet oncology*, *14*(9), 813-822.
22. Pope Iii, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., & Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Jama*, *287*(9), 1132-1141.
23. Hamra, G. B., Guha, N., Cohen, A., Laden, F., Raaschou-Nielsen, O., Samet, J. M., ... & Loomis, D. (2014). Outdoor particulate matter exposure and lung cancer: a systematic review and meta-analysis. *Environmental health perspectives*, *122*(9), 906.
24. International Agency for Research on Cancer. Outdoor Air Pollution. In IARC Monographs on the Evaluation of Carcinogenic Risks to Humans; IARC: Lyon, France, 2016; Volume 109.
25. Cohen, A. J., & Pope 3rd, C. A. (1995). Lung cancer and air pollution. *Environmental health perspectives*, *103*(Suppl 8), 219.
26. Lim, S. S., Vos, T., Flaxman, A. D., Danaei, G., Shibuya, K., Adair-Rohani, H., ... & Pelizzari, P. M. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The lancet*, *380*(9859), 2224-2260.
27. Carrot-Zhang, J., Soca-Chafre, G., Patterson, N., Thorner, A. R., Nag, A., Watson, J., ... & Meyerson, M. (2021). Genetic ancestry contributes to somatic mutations in lung cancers from admixed Latin American populations. *Cancer Discovery*, *11*(3), 591-598.
28. Rueda-Zarazua, B., Gutiérrez, H., García-Ortiz, H., Orozco, L., Ramírez-Martínez, G., Jiménez-Alvarez, L., ... & Melendez-Zajgla, J. (2025). A pilot study: contrasting genomic profiles of lung adenocarcinoma between patients of european and latin american ancestry. *International Journal of Molecular Sciences*, *26*(10), 4865.
29. da Cunha Santos, G., Shepherd, F. A., & Tsao, M. S. (2011). EGFR mutations and lung cancer. *Annual Review of Pathology: Mechanisms of Disease*, *6*(1), 49-69.
30. Díaz-Gay, M., Zhang, T., Hoang, P. H., Khandekar, A., Zhao, W., Steele, C. D., ... & Landi, M. T. (2024). The mutagenic forces shaping the genomic landscape of lung cancer in never smokers. *medRxiv*.
31. Arrieta, O., Cardona, A. F., Martín, C., Más-López, L., Corrales-Rodríguez, L., Bramuglia, G., ... & Cuello, M. (2015). Updated frequency of EGFR and KRAS mutations in nonsmall-cell lung cancer in Latin America: the Latin-American Consortium for the Investigation of Lung Cancer (CLICaP). *Journal of Thoracic Oncology*, *10*(5), 838-843.
32. Schabath, M. B., & Cote, M. L. (2019). Cancer progress and priorities: lung cancer. *Cancer epidemiology, biomarkers & prevention*, *28*(10), 1563-1579.
33. Kehm, R. D., Lloyd, S. E., Burke, K. R., & Terry, M. B. (2025). Advancing environmental epidemiologic methods to confront the cancer burden. *American Journal of Epidemiology*, *194*(1), 195-207.
34. Liu, J., Gan, T., Hu, W., & Li, Y. (2024). Current status and perspectives in environmental oncology. *Chronic Diseases and Translational Medicine*, *10*(04), 293-301.
35. Chevalier, A., Guo, T., Gurevich, N. Q., Xu, J., Yajima, M., & Campbell, J. D. (2025). Characterization of mutational signatures in tumors from a large Chinese population. *Cancer Research Communications*, *5*(8), 1466-1476.
36. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., ... & Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *nature*, *500*(7463), 415-421.
37. Helleday, T., Eshtad, S., & Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. *Nature reviews genetics*, *15*(9), 585-598.
38. Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., ... & Stratton, M. R. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell*, *149*(5), 979-993.

 doi:10.20944/preprints202604.0686.v1

39. Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., ... & Stratton, M. R. (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, *446*(7132), 153-158.
40. Meyerson, M., Gabriel, S., & Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*, *11*(10), 685-696.
41. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., & Stratton, M. R. (2013). Deciphering signatures of mutational processes operative in human cancer. *Cell reports*, *3*(1), 246-259.
42. Fischer, A., Illingworth, C. J., Campbell, P. J., & Mustonen, V. (2013). EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome biology*, *14*(4), R39.
43. Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., ... & Campbell, P. J. (2017). COSMIC: somatic cancer genetics at high-resolution. *Nucleic acids research*, *45*(D1), D777-D783.
44. Petljak, M., Alexandrov, L. B., Brammeld, J. S., Price, S., Wedge, D. C., Grossmann, S., ... & Stratton, M. R. (2019). Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell*, *176*(6), 1282-1294.
45. Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., & Stratton, M. R. (2015). Clock-like mutational processes in human somatic cells. *Nature genetics*, *47*(12), 1402-1407.
46. Pfeifer, G. P., Denissenko, M. F., Olivier, M., Tretyakova, N., Hecht, S. S., & Hainaut, P. (2002). Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene*, *21*(48), 7435-7451.
47. Brash, D. E. (2015). UV signature mutations. *Photochemistry and photobiology*, *91*(1), 15-26.
48. Roberts, S. A., Lawrence, M. S., Klimczak, L. J., Grimm, S. A., Fargo, D., Stojanov, P., ... & Gordenin, D. A. (2013). An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nature genetics*, *45*(9), 970-976.
49. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-Cancer Analysis of Whole Genomes. Nature 2020, 578, 82–93.
50. Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., ... & Stratton, M. R. (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94-101.
51. Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., ... & Forbes, S. A. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic acids research*, *47*(D1), D941-D947.
52. Kucab, J. E., Zou, X., Morganella, S., Joel, M., Nanda, A. S., Nagy, E., ... & Nik-Zainal, S. (2019). A compendium of mutational signatures of environmental agents. *Cell*, *177*(4), 821-836.
53. Nik-Zainal, S., Kucab, J. E., Morganella, S., Glodzik, D., Alexandrov, L. B., Arlt, V. M., ... & Phillips, D. H. (2015). The genome as a record of environmental exposure. *Mutagenesis*, *30*(6), 763-770.
54. Boysen, G., Alexandrov, L. B., Rahbari, R., Nookaew, I., Ussery, D., Chao, M. R., ... & Cooke, M. S. (2025). Investigating the origins of the mutational signatures in cancer. *Nucleic acids research*, *53*(1), gkae1303.
55. Maura, F., Degasperi, A., Nadeu, F., Leongamornlert, D., Davies, H., Moore, L., ... & Bolli, N. (2019). A practical guide for mutational signature analysis in hematological malignancies. *Nature communications*, *10*(1), 2969.
56. Islam, S. A., Díaz-Gay, M., Wu, Y., Barnes, M., Vangara, R., Bergstrom, E. N., ... & Alexandrov, L. B. (2022). Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell genomics*, *2*(11).
57. Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz Jr, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. *science*, *339*(6127), 1546-1558.
58. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature*. 2009;458(7239):719-724. doi:10.1038/nature07943
59. McFarland CD, Korolev KS, Kryukov GV, Sunyaev SR, Mirny LA. Impact of deleterious passenger mutations on cancer progression. *Proc Natl Acad Sci U S A*. 2013;110(8):2910-2915. doi:10.1073/pnas.1213968110
60. McFarland, C. D., Mirny, L. A., & Korolev, K. S. (2014). Tug-of-war between driver and passenger mutations in cancer and other adaptive processes. *Proceedings of the National Academy of Sciences*, *111*(42), 15138-15143.

 doi:10.20944/preprints202604.0686.v1

61. Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., ... & Nowak, M. A. (2010). Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, *107*(43), 18545-18550.
62. Kumar, S., Warrell, J., Li, S., McGillivray, P. D., Meyerson, W., Salichos, L., ... & Gerstein, M. B. (2020). Passenger mutations in more than 2,500 cancer genomes: overall molecular functional impact and consequences. *Cell*, *180*(5), 915-927.
63. Rheinbay, E., Nielsen, M. M., Abascal, F., Wala, J. A., Shapira, O., Tiao, G., ... & Getz, G. (2020). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, *578*(7793), 102-111.
64. Shuai, S., Gallinger, S., & Stein, L. D. (2020). Combined burden and functional impact tests for cancer driver discovery using DriverPower. *Nature communications*, *11*(1), 734.
65. Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., ... & Garnett, M. J. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell*, *166*(3), 740-754.
66. Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., McDonald III, E. R., ... & Sellers, W. R. (2019). Next-generation characterization of the cancer cell line encyclopedia. *Nature*, *569*(7757), 503-508.
67. Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., ... & Garnett, M. J. (2012). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, *41*(D1), D955-D961.
68. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., ... & Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, *483*(7391), 603-607.
69. Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., ... & Benes, C. H. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, *483*(7391), 570-575.
70. Boot, A., Huang, M. N., Ng, A. W., Ho, S. C., Lim, J. Q., Kawakami, Y., ... & Rozen, S. G. (2018). In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome research*, *28*(5), 654-665.
71. Davies, H., Glodzik, D., Morganella, S., Yates, L. R., Staaf, J., Zou, X., ... & Nik-Zainal, S. (2017). HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nature medicine*, *23*(4), 517-525.
72. Flaherty, K. T., Puzanov, I., Kim, K. B., Ribas, A., McArthur, G. A., Sosman, J. A., ... & Chapman, P. B. (2010). Inhibition of mutated, activated BRAF in metastatic melanoma. *New England Journal of Medicine*, *363*(9), 809-819.
73. Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., ... & Futreal, P. A. (2002). Mutations of the BRAF gene in human cancer. *Nature*, *417*(6892), 949-954.
74. Kopetz, S., Desai, J., Chan, E., Hecht, J. R., O'Dwyer, P. J., Maru, D., ... & Saltz, L. (2015). Phase II pilot study of vemurafenib in patients with metastatic BRAF-mutated colorectal cancer. *Journal of clinical oncology*, *33*(34), 4032-4038.
75. Hayward, N. K., Wilmott, J. S., Waddell, N., Johansson, P. A., Field, M. A., Nones, K., ... & Mann, G. J. (2017). Whole-genome landscapes of major melanoma subtypes. *Nature*, *545*(7653), 175-180.
76. Hayward NK, Wilmott JS, Waddell N, et al. Whole-genome landscapes of major melanoma subtypes. *Nature*. 2017;545(7653):175-180. doi:10.1038/nature22071
77. Viel, A., Bruselles, A., Meccia, E., Fornasarig, M., Quaia, M., Canzonieri, V., ... & Bignami, M. (2017). A specific mutational signature associated with DNA 8-oxoguanine persistence in MUTYH-defective colorectal cancer. *EBioMedicine*, *20*, 39-49.
78. Pilati, C., Shinde, J., Alexandrov, L. B., Assié, G., André, T., Hélias-Rodzewicz, Z., ... & Laurent-Puig, P. (2017). Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. *The Journal of pathology*, *242*(1), 10-15.
79. Pereira, L., Igea, A., Canovas, B., Dolado, I., & Nebreda, A. R. (2013). Inhibition of p38 MAPK sensitizes tumour cells to cisplatin-induced apoptosis mediated by reactive oxygen species and JNK. *EMBO molecular medicine*, *5*(11), 1759-1774.

 doi:10.20944/preprints202604.0686.v1

80. Trempolec, N., Dave-Coll, N., & Nebreda, A. R. (2013). SnapShot: p38 MAPK signaling. *Cell*, *152*(3), 656-656.
81. Smirnov, P., Safikhani, Z., El-Hachem, N., Wang, D., She, A., Olsen, C., ... & Haibe-Kains, B. (2016). PharmacoGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics*, *32*(8), 1244-1246.
82. Lee, S. J., Serre, M. L., Van Donkelaar, A., Martin, R. V., Burnett, R. T., & Jerrett, M. (2012). Comparison of geostatistical interpolation and remote sensing techniques for estimating long-term exposure to ambient PM2. 5 concentrations across the continental United States. *Environmental health perspectives*, *120*(12), 1727.
83. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
84. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.
85. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, *2*(1), 56-67.
86. RDKit: Open-Source Cheminformatics. Available online: https://www.rdkit.org.
87. NASA Langley Research Center POWER Project funded through the NASA Earth Science Directorate Applied Science Program. Available online: https://power.larc.nasa.gov (accessed on 7 March 2026).