Supplementary Materials and Methods

1. Machine-learning-based candidate gene prioritization. LASSO regression, support vector machine–recursive feature elimination (SVM-RFE), and random forest were used to prioritize candidate genes associated with acrylonitrile exposure and lung cancer. Feature selection and model evaluation procedures are described in the main manuscript. The corresponding LASSO cross-validation and coefficient profiles, SVM-RFE performance curves, random forest feature-importance rankings, and overlap of prioritized genes are presented in Figure S1A–F.

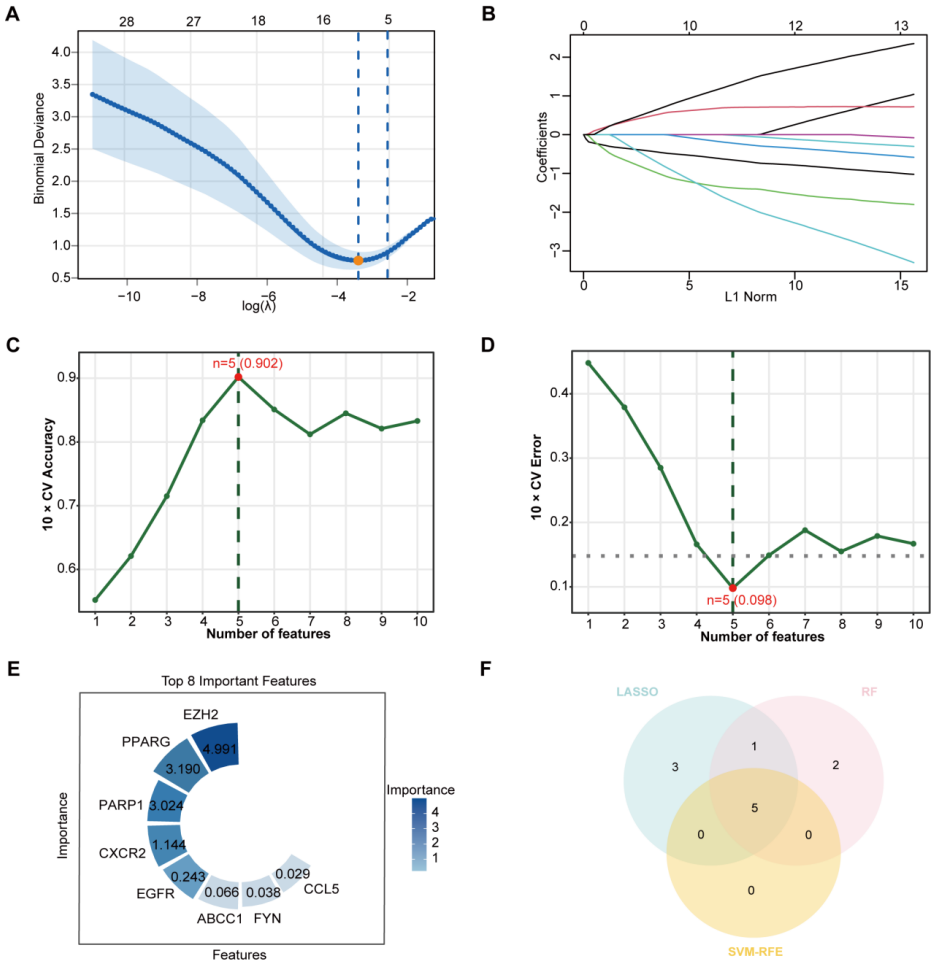


Figure S1. Exploratory machine-learning prioritization of candidate genes associated with acrylonitrile-related lung cancer. *(A) LASSO cross-validation curve showing binomial deviance versus log(λ). (B) LASSO coefficient profiles of candidate genes. (C) SVM-RFE curve showing cross-validation accuracy across different feature numbers. (D) SVM-RFE curve showing cross-validation error across different feature numbers. (E) Random forest circular bar plot showing feature-importance rankings. (F) Venn diagram showing the overlap among genes prioritized by LASSO, random forest, and SVM-RFE analyses. Feature selection was performed to prioritize candidate genes; the results do not establish independent predictive validity or acrylonitrile-specific biological effects.*

2. Molecular docking and molecular dynamics analysis of the AN–EGFR complex. Molecular docking and 100-ns molecular dynamics simulations of the predicted acrylonitrile (AN)–EGFR

complex were performed using the procedures described in the main manuscript. The predicted docking configuration, root mean square deviation, root mean square fluctuation, radius of gyration, solvent-accessible surface area, Gibbs free energy landscapes, and hydrogen-bond profiles are presented in Figure S2A–H. These computational results do not establish direct biochemical binding or functional regulation of EGFR by AN.

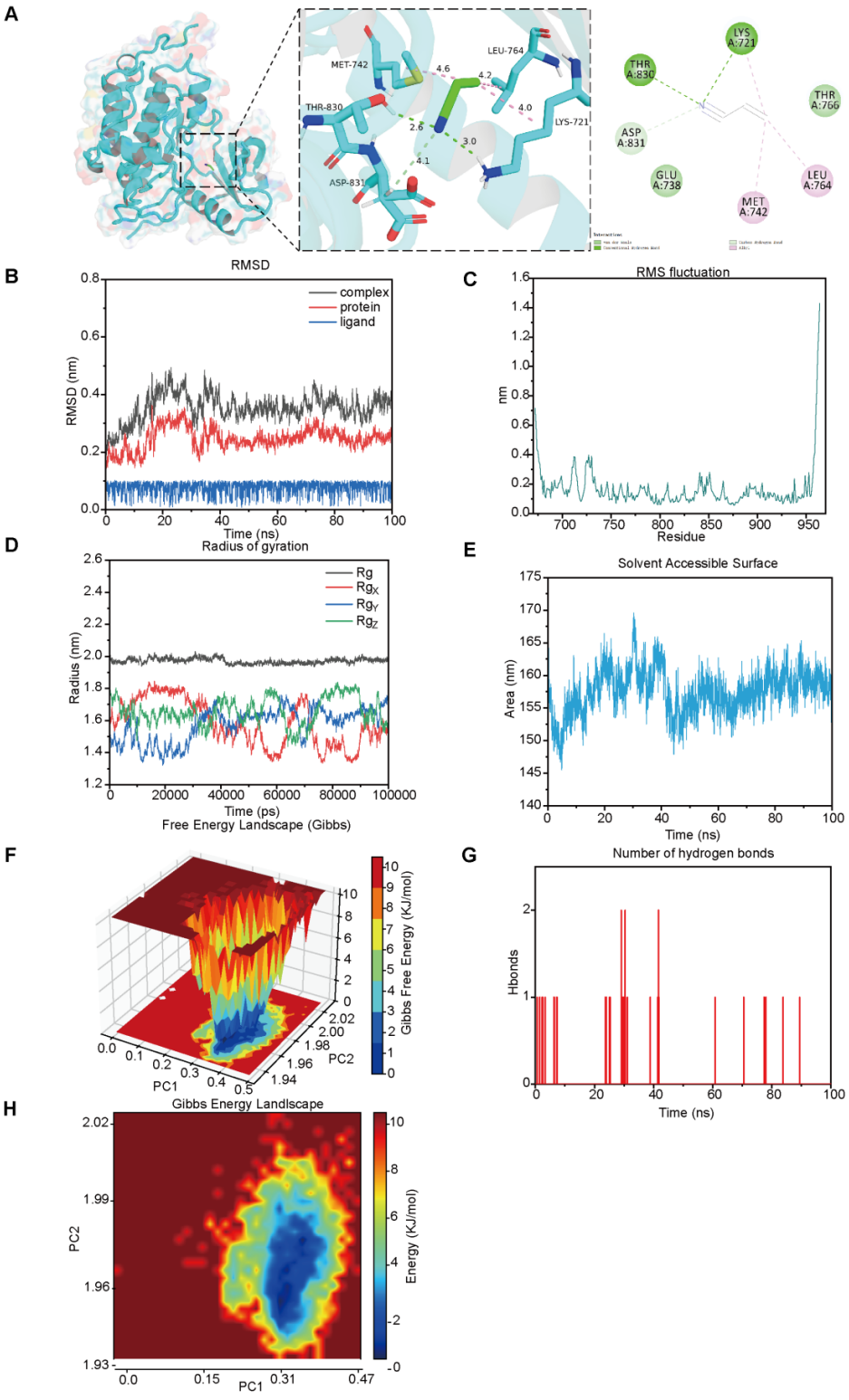


Figure S2. Molecular docking and molecular dynamics analyses of the AN–EGFR complex. *(A) Docking mode of AN with EGFR. (B) RMSD profiles of the complex, protein, and ligand during the 100 ns simulation. (C) RMSF profile of EGFR residues. (D) Radius of gyration (Rg) of the complex. (E) Solvent accessible surface area (SASA) of the complex. (F) Three-dimensional Gibbs free energy landscape of the AN–EGFR complex. (G) Number of hydrogen bonds formed between AN and EGFR. (H) Two-dimensional Gibbs free energy landscape of the AN–EGFR complex. These computational analyses describe predicted conformational behavior of the AN–EGFR complex and do not establish direct biochemical binding, covalent modification, or functional regulation of EGFR by acrylonitrile.*

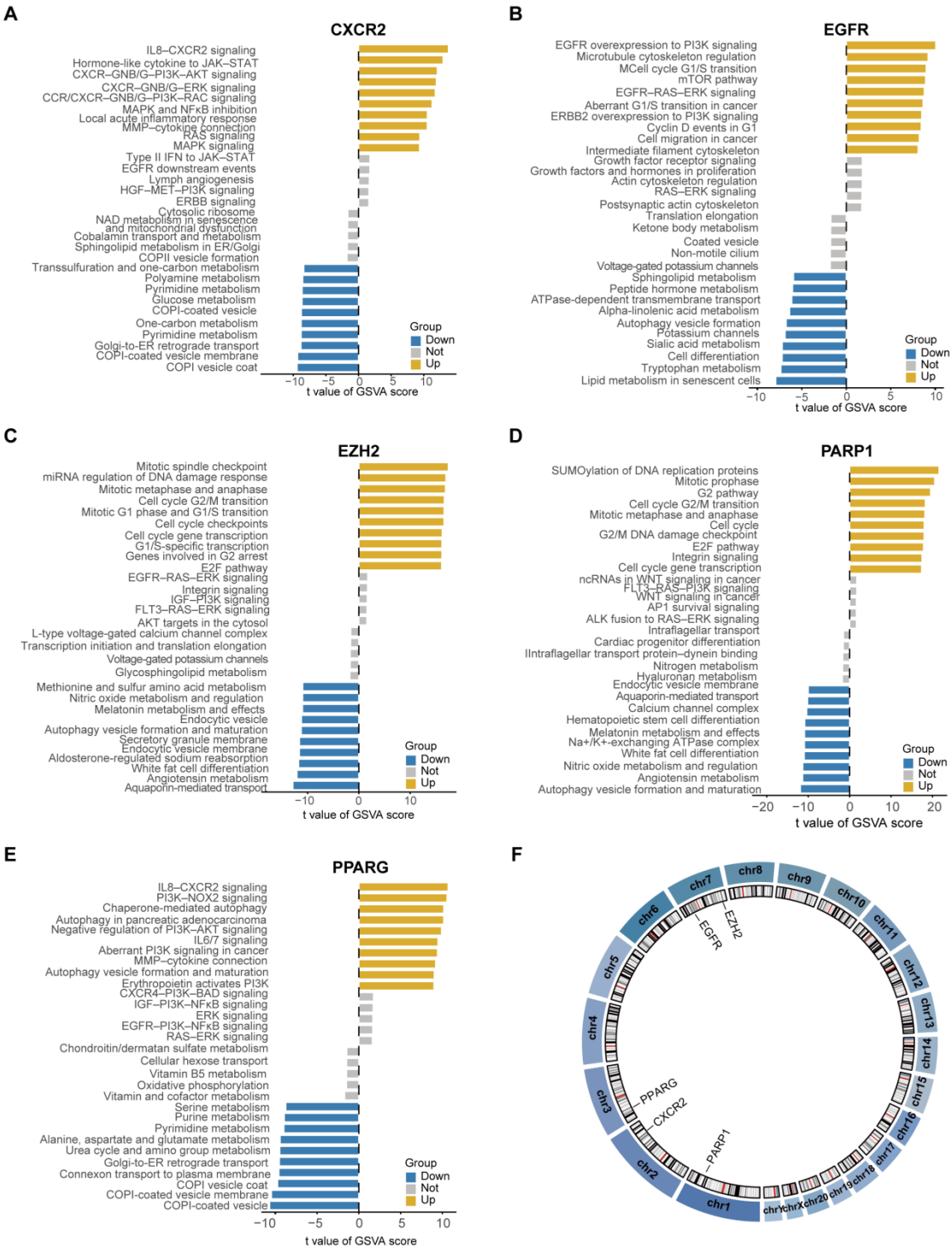


Figure S3. GSVA enrichment analysis and chromosomal distribution of candidate genes associated with acrylonitrile-related lung cancer. *(A–E) GSVA bar plots illustrating pathway enrichment patterns associated with CXCR2, EGFR, EZH2, PARP1, and PPARG, respectively. Yellow and blue bars indicate pathways with positive and negative enrichment scores, respectively, whereas gray bars indicate pathways without a clear directional enrichment pattern. These analyses are exploratory and do not establish pathway activation or suppression by acrylonitrile exposure. (F) Circos plot showing the chromosomal locations of the five candidate genes.*

3. Gene set variation analysis (GSVA). Gene set variation analysis was performed using the GSVA R package to evaluate pathway-associated transcriptional signatures of the five prioritized genes in lung cancer. Gene sets were obtained from the Molecular Signatures Database (MSigDB). Pathway scores were calculated using the single-sample gene set enrichment analysis (ssGSEA) method. Spearman correlation analysis was performed between the expression levels of the five prioritized genes and ssGSEA-derived pathway scores to identify gene-associated transcriptional signatures (Figure S3A–E). The chromosomal locations of the five prioritized genes were visualized using a Circos plot (Figure S3F).

4. Multi-omics Analysis

4.1 Transcriptomics. For transcriptomic analysis, A549 cells from the control and AN-treated groups were analyzed with three biological replicates per group. Total RNA was extracted from cell lysates using the RNAeasy™ Animal RNA Isolation Kit. RNA integrity and library quality were assessed using the Agilent 2100 Bioanalyzer with the RNA Nano 6000 Kit. Sequencing libraries were constructed and subjected to 150-bp paired-end sequencing on the Illumina NovaSeq platform. Clean reads were aligned to the reference genome using HISAT2 v2.0.5, and gene-level read counts were quantified using featureCounts v1.5.0-p3. Differentially expressed genes (DEGs) were identified using DESeq2 v1.20.0 with an adjusted P-value < 0.05 and $|\log2FoldChange| \geq 1.0$. Functional enrichment analyses, including Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses, were performed using the clusterProfiler R package.

4.2 Metabolomics. For metabolomic analysis, intracellular metabolites were extracted from control and AN-treated A549 cells, with three biological replicates per group. Cells were rapidly rinsed with PBS, quenched with precooled 80% methanol, and centrifuged to remove cellular debris. The supernatant containing intracellular metabolites was collected and diluted to 53% methanol before LC-MS analysis. Metabolic profiling was performed using a Q Exactive™ HF-X mass spectrometer in both positive and negative ion modes. Data were processed using Compound Discoverer 3.3 with blank-based background removal. Peak areas were normalized by dividing each raw peak area by the ratio of the sample's total metabolite signal to that of QC1. Metabolites with QC CV>30% were excluded. PLS-DA was performed using metaX to obtain VIP scores. Differential metabolites were selected using VIP>1, Student's t-test P<0.05, and fold change ≥2 or ≤0.5. P values were not adjusted for multiple

comparisons. Accordingly, the identified differential metabolites and associated pathway enrichment results were interpreted as exploratory rather than statistically confirmed discoveries.

4.3 Metabolomic-Transcriptomic Integrative Analysis. To investigate coordinated molecular and metabolic alterations following acrylonitrile (AN) exposure, differentially expressed genes (DEGs) and differentially abundant metabolites (DAMs) were integrated for multi-omics analysis. Top-ranked DEGs and DAMs were selected according to their P values. Pearson correlation analysis was performed using standardized transcriptomic and metabolomic data from corresponding samples to evaluate associations between gene expression and metabolite abundance. A nominal $P < 0.05$ threshold was used to identify candidate DEG–DAM associations for further investigation. These correlations were interpreted as exploratory associations rather than evidence of direct causal regulation. Candidate DEG–DAM pairs meeting the specified correlation threshold were used to construct an integrated metabolite–transcript association network. Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were performed for the transcriptomic and metabolomic datasets, and pathways shared by both datasets were considered candidate pathways potentially associated with AN exposure. Integrated multi-omics alterations were further mapped onto KEGG pathway diagrams, with genes and metabolites annotated according to their log2 fold-change values. All analyses were conducted using R version 4.4.3, with experimental and primary data-processing support provided by Novogene Co., Ltd.