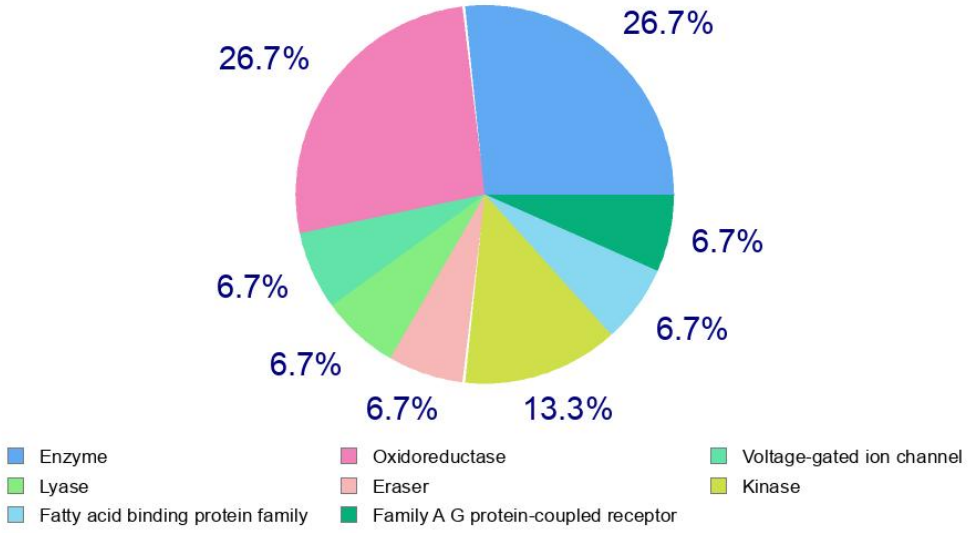

Figure S1. Distribution of Protein/Enzyme Categories Associated with NNK (4-(methyl nitrosamino)-1-(3-pyridyl)-1-butanone).

This pie chart illustrates the relative proportion of major molecular categories implicated in the metabolism, signaling, and biomolecular interactions of the carcinogen NNK within a biological context.

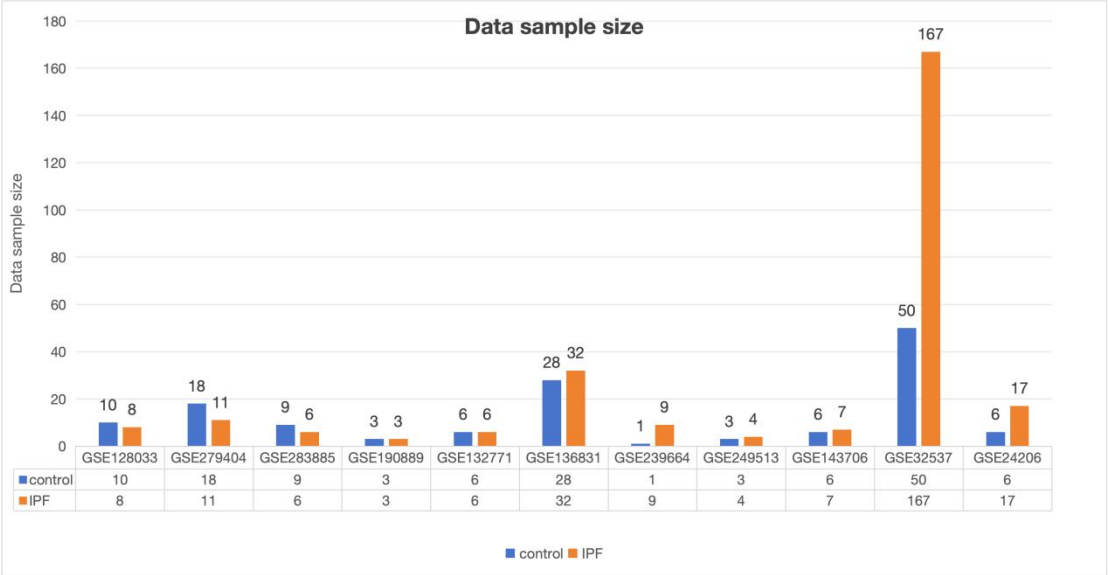

Figure S2. Sample sizes of idiopathic pulmonary fibrosis (IPF) and control cohorts across 11 public transcriptomic datasets.

A bar chart shows the number of control (blue) and IPF (orange) samples in each dataset. The x-axis lists the GEO accession numbers; the y-axis indicates the sample count. The collective samples from all datasets form the basis for subsequent integrated analyses.

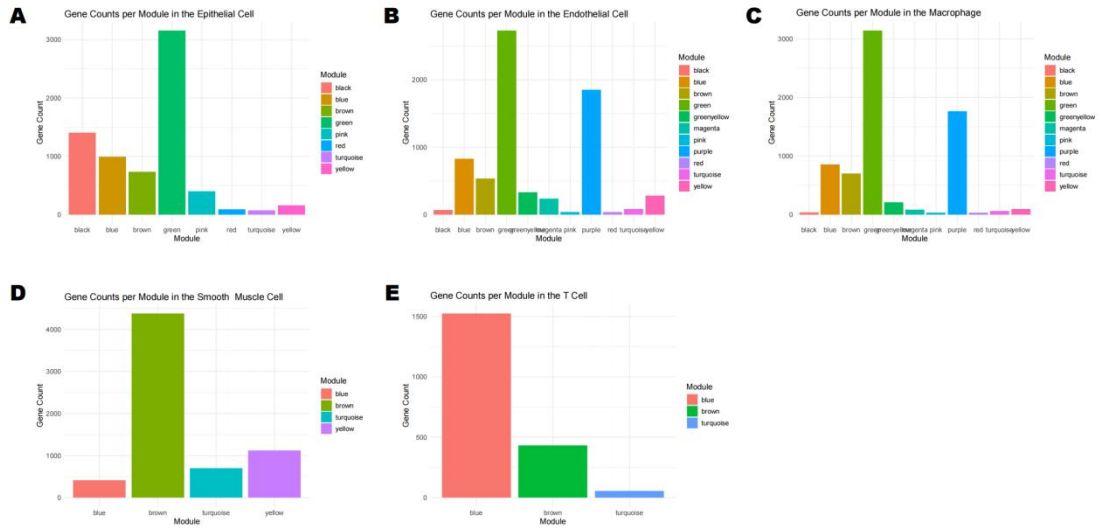

Figure S3. Distribution of WGCNA module gene counts across cell types in the training set.

(A-E) Module size in five lung cell types. Horizontal bar plots show the number of genes assigned to each Weighted Gene Co-expression Network Analysis (WGCNA) module with in five major cell types from the training cohort: (A) epithelial cells, (B) endothelial cells, (C) macrophages, (D) smooth muscle cells, and (E) T cells.

Visualization. In each plot, the y-axis lists the WGCNA modules, which are automatically assigned distinct color labels (e.g., black, blue, brown, green, red). The x-axis indicates the gene count (i.e., the number of genes clustered into each module). The length of each colored bar corresponds to the size (gene richness) of that module within the specified cell type. Interpretation. This comparative visualization reveals the cell-type-specific architecture of gene co-expression networks. It highlights which modules are large and potentially dominant in certain cell types (e.g., the green module in structural cells, the red module in T cells), thereby identifying cellular contexts where specific co-expression programs are most active. This information is foundational for subsequent module-trait correlation analyses.

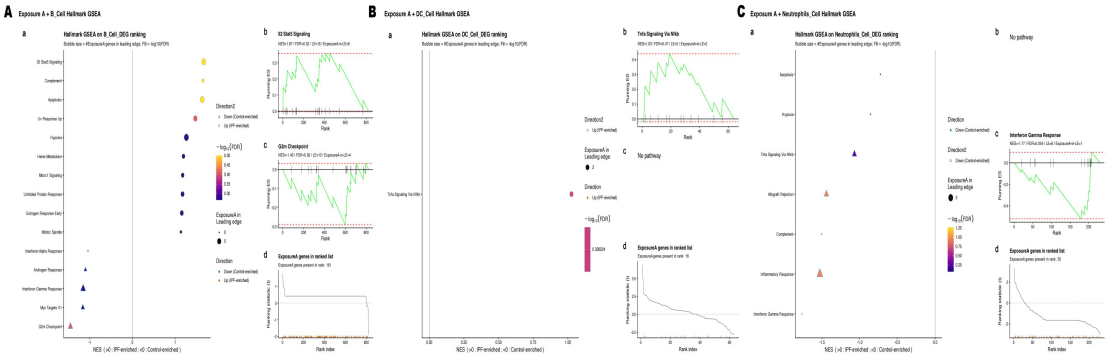

Figure S4. Hallmark pathway enrichment analysis of exposure-associated genes in immune cell types from the training set.

(A-C) Gene Set Enrichment Analysis (GSEA) in test-set immune cells. Hallmark pathway enrichment was assessed in B cells (A), dendritic cells (DCs) (B), and neutrophils (C) from the training set. Each panel contains four components:

(a) Overview of enriched pathways. Bubble plot summarizing GSEA results. Each bubble represents a Hallmark gene set, positioned horizontally by its Normalized Enrichment Score (NES) and vertically by the gene set. Bubble size indicates the number of exposure-associated ("ExposureA") genes in the leading edge subset. Bubble color represents the statistical significance of enrichment (-log$_{10}$ [FDR]). Symbol shape (△/○) denotes the direction of enrichment (up/down in IPF).

(b) Enrichment profile for the top upregulated pathway. Running enrichment score plot for the most significantly positively enriched Hallmark pathway (e.g., "Complement" in B cells), showing the distribution of exposure-associated genes within the ranked list of all genes. In Neutrophils where no pathways met the significance threshold for positive enrichment.

(c) Enrichment profile for the top downregulated pathway. Running enrichment score plot for the most significantly negatively enriched Hallmark pathway. In DC cell where no pathways met the significance threshold for negative enrichment.

(d) Ranked list of exposure-associated genes. Plot showing the position of exposure-associated genes (vertical bars) within the globally ranked list of all genes based on their association with the cellular phenotype.

Interpretation: Positive NES values (right side of the bubble plot) indicate pathway upregulation, while negative NES values (left side) indicate downregulation in the disease-associated or exposure-relevant condition within the training set. The analysis identifies which Hallmark biological processes are coordinately enriched or suppressed with exposure-associated genes in specific immune cell populations, providing exploratory insights into potential cell-type-specific mechanisms.

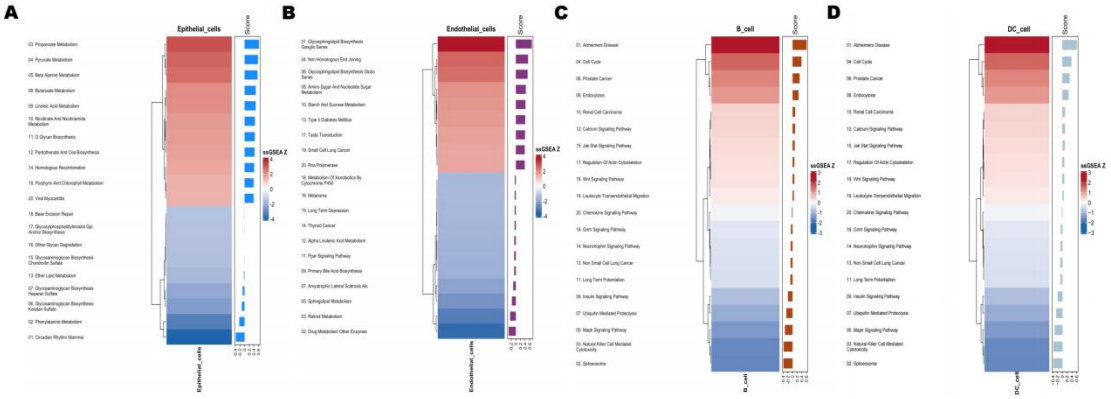

Figure S5. Single-sample enrichment of biological processes and exposure-associated signatures across lung cell types in the training set.

(A-D) Cell-type-specific enrichment profiles. This figure presents the single-sample Gene Set Enrichment Analysis (ssGSEA) scores for a curated set of biological processes and computational signatures across four major lung cell types within the training cohort. Each panel corresponds to a specific cell type: Eplithelial_cells (A), Endothelial_cells (B), B_cell(C), and DC_cells (D).

Visualization.For each panel, the rows (y-axis) list the functional terms or gene signatures. The enrichment level for each term is visualized by a horizontal bar, which is divided and colored to represent its aggregated activity score. The color gradient within each bar segment follows a scale from blue (low enrichment, z-score $\leq$ -2.5) to red (high enrichment, z-score $\geq$ 2.5).

Interpretation.This "heatmap-style" bar plot enables direct comparison of pathway activation across the four cell types. Patterns of high (red) or low (blue) enrichment identify which biological processes or exposure signatures are most active within each specific cellular compartment in the training data, informing downstream analyses.

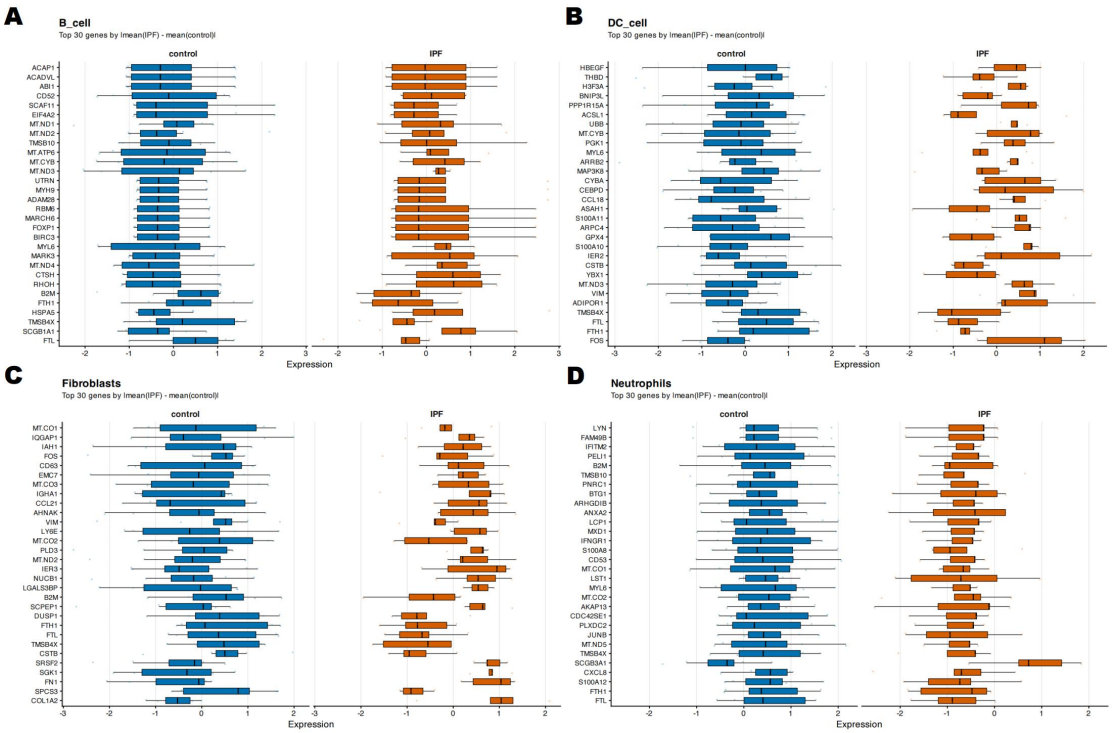

Figure S6. Cell-type-specific expression patterns of the top 30 differentially expressed genes in the training set.

(A-D) Expression of signature genes across immune and structural cell types. The figure displays the expression levels of the 30 genes with the largest absolute difference in mean expression between IPF and control groups (|mean(IPF) - mean(control)|) in the training set. Analysis is shown for four key lung cell types: B cells (A), dendritic cells (B), fibroblasts (C), and neutrophils (D).

Visualization. For each cell type, a horizontal box plot is shown for each of the top 30 genes (listed on the y-axis). Gene expression values (z-scored or normalized) are plotted on the x-axis. Each gene's expression distribution is compared between control samples (blue) and idiopathic pulmonary fibrosis (IPF) samples (orange) from the training cohort.

Interpretation. The box plots show the median (center line), interquartile range (box), and data range (whiskers) for each group. This visualization identifies cell-type-specific dysregulation patterns, highlighting genes that are consistently upregulated or downregulated in IPF within distinct cellular compartments of the lung.

Figure S7. Gene Ontology (GO) enrichment analysis for the top 30 cell-type-specific genes in the training set.

(A-D) Cell-type-resolved functional enrichment. The figure displays the GO enrichment results for the 30 genes most strongly associated with the exposure signature within each of four cell types from the training cohort: B cells (A), dendritic cells (B), fibroblasts (C), and neutrophils (D). Analysis and Visualization. For each cell type, enrichment is analyzed for the three standard GO domains: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF). The results for each domain are presented in two complementary plots:

Left (Bar Plot): Shows the number of genes (Count) from the cell-type-specific top 30 list that are annotated to each of the significantly enriched GO terms.

Right (Dot Plot): Illustrates the enrichment strength and statistical significance of the same terms. The x-axis represents the gene ratio (proportion of query genes mapped to the term), and the y-axis represents the $-\log_{10}$ (adjusted p-value). Each dot corresponds to a GO term, colored for distinction.

Interpretation. This analysis, performed on the training set, highlights the distinct biological processes, cellular components, and molecular activities most strongly linked to the exposure response within each immune and structural cell compartment. The results provide cell-type-resolved functional insights that informed downstream model development and hypothesis generation. The consistent presentation across all four cell types facilitates direct comparison of enrichment patterns.

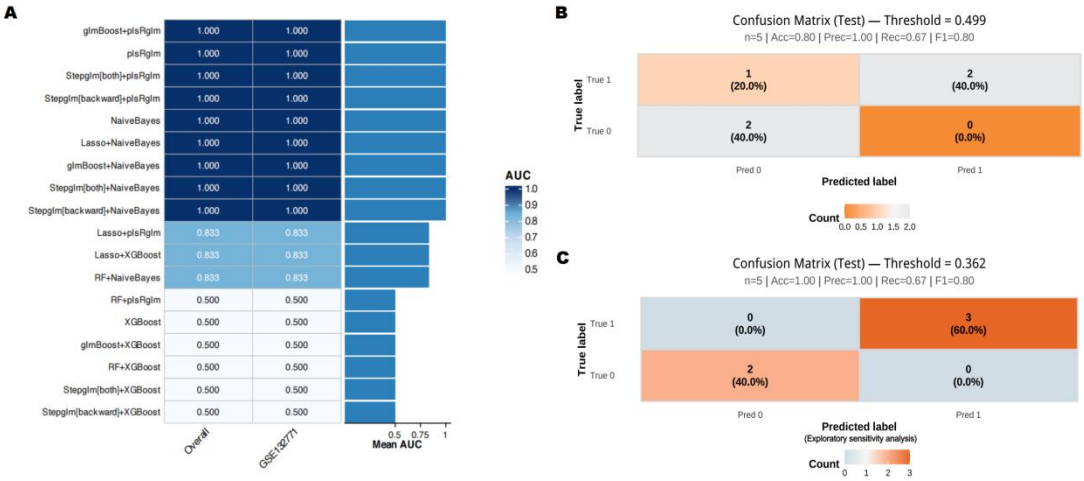

Figure S8. Model performance evaluation on the independent test set: AUC ranking and confusion matrices at two decision thresholds.

(A) Model ranking by average AUC. Bar plot showing the mean Area Under the ROC Curve (AUC) achieved by different candidate models (or model combinations) on the independent test set. Models are ranked by their AUC values, which provide an aggregate measure of discriminative ability, with values closer to 1.0 indicating better performance.

(B, C) Confusion matrices for the best-performing model at two decision thresholds. Detailed classification performance of the top model (selected from panel A) is shown at a threshold of 0.499 (B) and 0.362 (C).

Axes: The x-axis represents the model's predicted labels; the y-axis represents the true labels.

Quadrants: Each matrix is divided into four quadrants showing counts (and/or proportions) of: True Negatives (TN), False Positives (FP), False Negatives (FN), and True Positives (TP).

Interpretation of threshold selection: The comparison illustrates the trade-off between sensitivity and specificity. The higher threshold (B, 0.499) favors specificity (minimizing FP, useful when false alarms are costly), while the lower threshold (C, 0.362) favors sensitivity (minimizing FN, useful for high-recall screening). This analysis guides the selection of an optimal operating point based on clinical or research priorities.

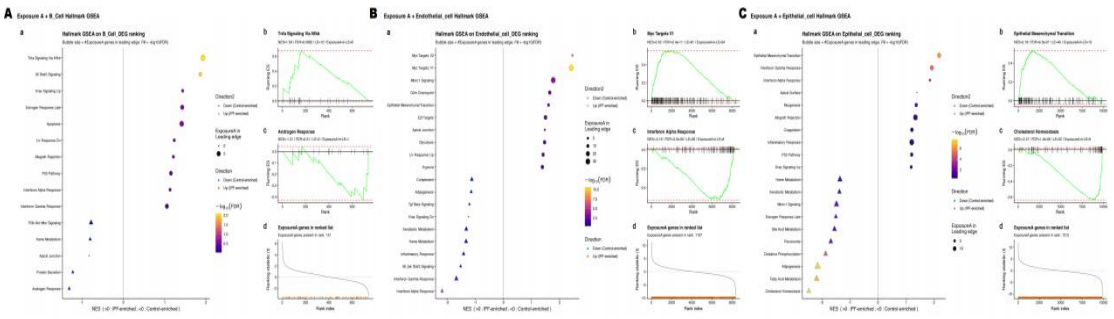

Figure S9. Cell-type-specific Hallmark pathway enrichment of exposure-associated genes in the test set.

(A-C) Gene Set Enrichment Analysis (GSEA) in B cells (A), endothelial cells (B), and epithelial cells (C). For each cell type, the analysis is presented in four integrated panels:

(a) Overview of enriched Hallmark pathways. Bubble plot summarizing GSEA results. Each bubble represents a pathway, positioned horizontally by its Normalized Enrichment Score (NES) and colored by its statistical significance ($-\log_{10}$ [FDR]). Bubble size indicates the number of exposure-associated ("ExposureA") genes in the core enrichment subset. Symbol shape (△/○) denotes the direction of enrichment (up/down in IPF).

(b, c) Detailed enrichment profiles. Running enrichment score plots for two selected, significant pathways from panel (a), highlighting the position of exposure-associated genes (vertical black bars) within the ranked gene list. Key statistics (NES, FDR, core gene counts) are displayed.

(d) Ranked list of all genes. Gray line traces the ranking statistic; orange vertical bars mark the positions of exposure-associated genes, providing context for the GSEA.

Interpretation: The analysis reveals distinct, cell-type-specific pathway regulation associated with the exposure. In B cells, the TNFα/NFκB pathway is significantly activated. In endothelial cells, MYC targets and interferon response pathways are strongly enriched. In epithelial cells, a coordinated activation of Epithelial-Mesenchymal Transition and suppression of Cholesterol Homeostasis is observed. These patterns suggest that the exposure factor may drive IPF progression through complementary, cell-type-specific mechanisms.

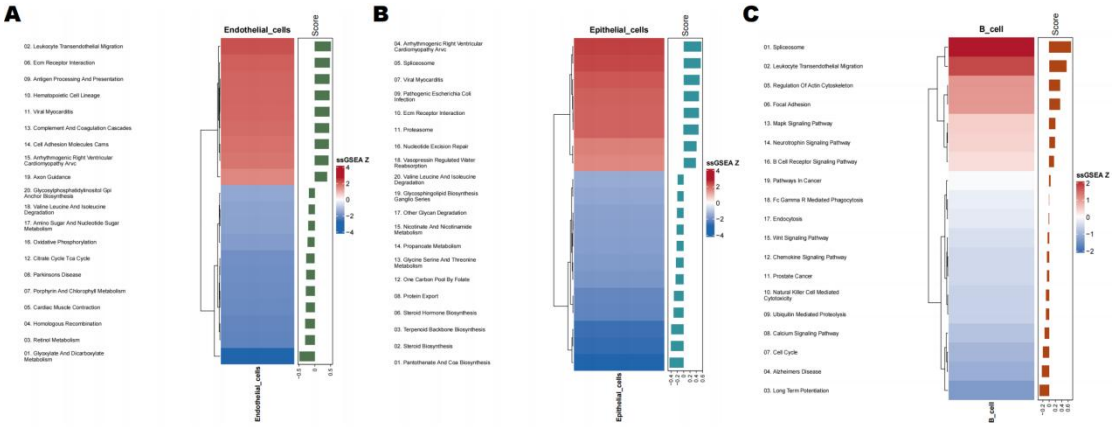

Figure S10. Sample-level enrichment patterns of hallmark pathways in test-set cell types.

(A-C) **Single-sample Gene Set Enrichment Analysis (ssGSEA) in endothelial cells (A), epithelial cells (B), and B cells (C).** For each cell type, the left panel lists significantly enriched Hallmark gene sets. The right panel is a corresponding heatmap where each row represents a gene set and each column represents an individual sample from the test set.

Visualization and Data. Gene sets are ordered by their enrichment significance or pattern. Within the heatmap, the color of each cell represents the **Normalized Enrichment Score (NES)**-derived activity (z-scored) for that gene set in a specific sample, as indicated by the adjacent color key. The gradient from **blue to red** indicates low to high pathway activity, reflecting relative up- or down-regulation of the gene set in that sample.

Interpretation. This visualization reveals the cell-type-specific activation landscape of hallmark biological processes across all test samples. Consistent red bands across samples for a given gene set indicate a pathway that is broadly activated in that cell type within the cohort, highlighting key, shared biological states (e.g., inflammatory response, metabolic processes) within each cellular compartment.

Figure S11. Cell-type-specific Gene Ontology (GO) enrichment of the top 30 genes in the test set.

(A-C) GO enrichment in B cells (A), endothelial cells (B), and epithelial cells (C). The figure presents the GO enrichment results for the 30 most significant genes associated with each cell type in the independent test set. For each cell type, analysis is shown for the three standard GO domains: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF).

Visualization. For each cell type and GO domain, two complementary plots are presented:

Left (Bar Plot): Displays the number of genes (Count) from the cell-type-specific top 30 list that are annotated to all significantly enriched GO terms within that domain.

Right (Dot Plot): Shows the enrichment strength and statistical significance for the top enriched terms. The x-axis represents the Gene Ratio (proportion of query genes in the term), and the y-axis lists the specific GO terms. Dot size corresponds to the number of enriched genes, and dot color represents the statistical significance ($-\log_{10}$ [adjusted p-value]).

Interpretation. The analysis reveals highly distinct functional profiles for the top genes of each cell type, consistent with their specialized biology. B cell genes are enriched for immune and signaling functions, endothelial cell genes for vascular and adhesive functions, and epithelial cell genes for structural and developmental processes. This confirms that the most salient genes captured by the analysis within each compartment are directly related to their core cellular identities and functions in the test cohort.

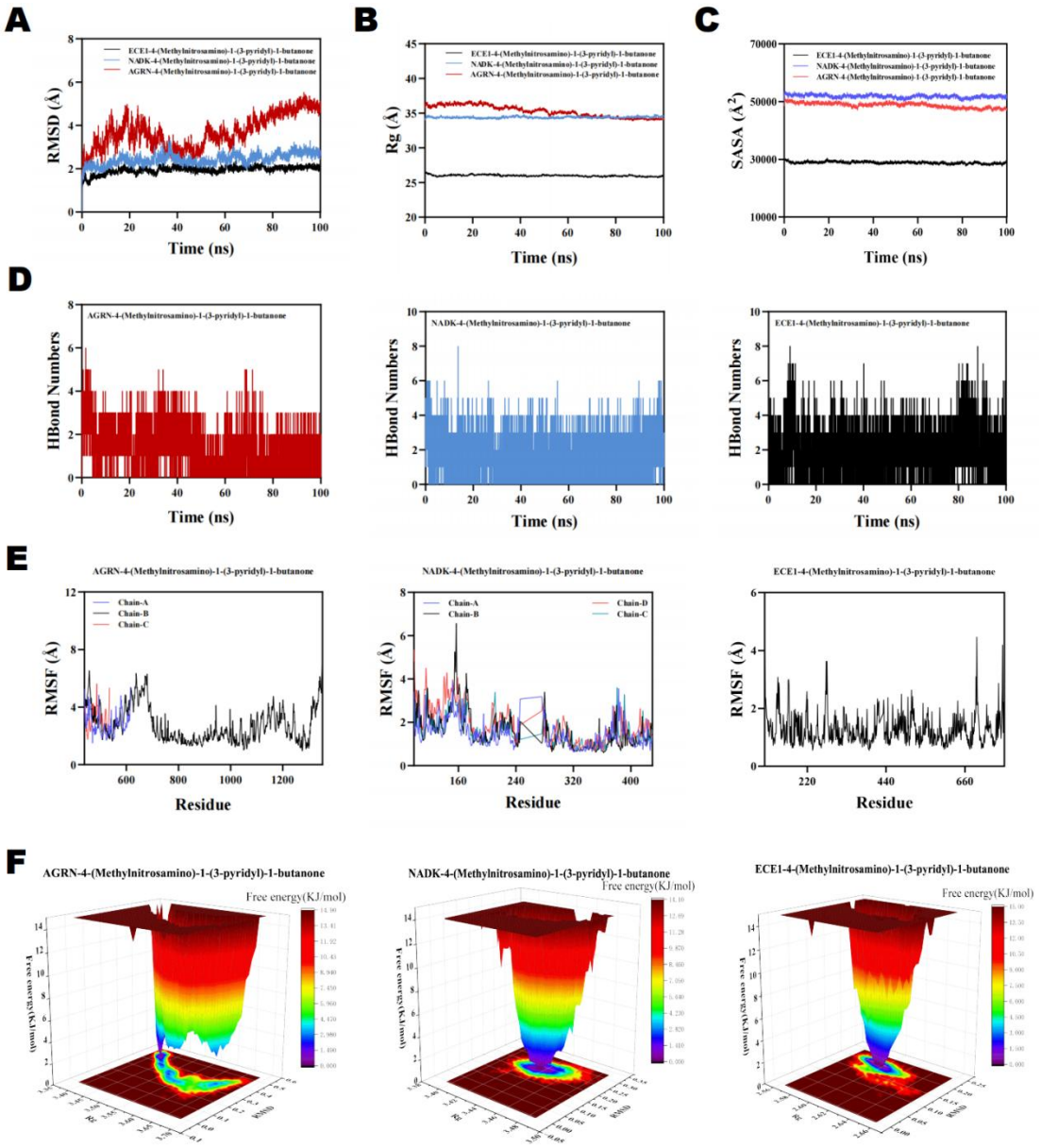

Figure S12. Molecular dynamics simulation analysis of protein-NNK complexes.

(A-F) Stability and interaction analysis of three candidate complexes. Molecular dynamics simulations (100 ns) were performed for the AGRN-NNK (magenta), NADK-NNK (dark green), and ECE1-NNK (orange) complexes. The analysis includes:

(A) Root Mean Square Deviation (RMSD). Time evolution of the backbone RMSD (Å), measuring overall conformational stability. Lower and stable values indicate a more stable complex.

(B) Radius of Gyration (Rg). Time evolution of Rg (Å), reflecting the global compactness of the protein.

(C) Solvent Accessible Surface Area (SASA). Time evolution of SASA (nm²), indicating changes in solvent exposure.

(D) Intermolecular Hydrogen Bonds. Number of hydrogen bonds formed between the protein and NNK over time, a key indicator of binding interaction stability.

(E) Root Mean Square Fluctuation (RMSF). Per-residue RMSF (Å) of the protein Cα atoms, highlighting local flexible regions.

(F) Free Energy Landscape (FEL). Two-dimensional FELs projected onto the first two principal components. The contour map shows the Gibbs free energy (G, in kJ/mol), with deep blue regions representing the most stable conformational states.

Interpretation: The integrated metrics provide a comparative assessment of complex stability. The AGRN-NNK complex shows lower final RMSD, lower SASA, and a deep, single minimum in its FEL, suggesting high conformational stability and a tightly packed structure. The ECE1-NNK complex maintains a higher number of intermolecular hydrogen bonds, indicating strong specific interactions. These simulations help elucidate the structural basis for potential differential binding stability among the candidates.

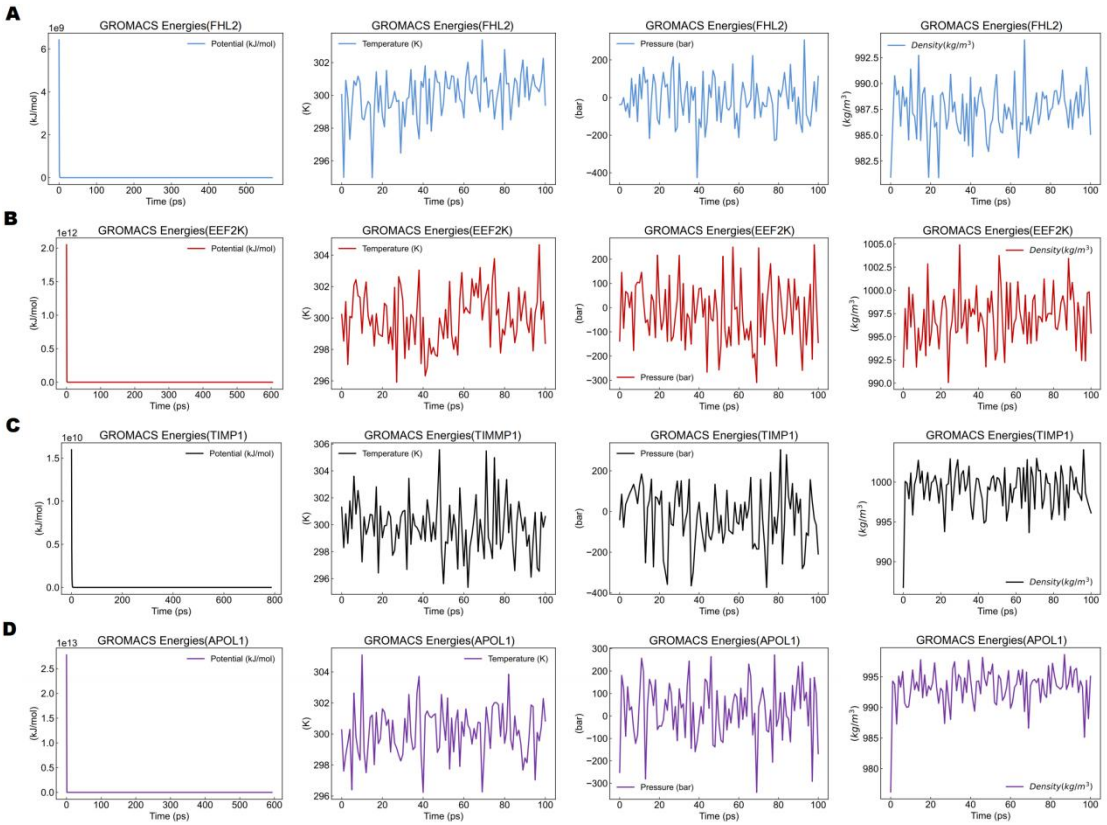

Figure S13. System equilibration monitoring during 100-ns molecular dynamics simulations (training set).

Overall Layout. The figure presents the time evolution of key thermodynamic parameters during the equilibration phase of molecular dynamics simulations for four protein-NNK: FHL2 (A), EEF2K(B), TIMP1 (C), and APOL1(D). The layout is a 4×4 matrix: each row tracks a specific parameter for all complexes, and each column corresponds to a single complex.

Parameters Monitored. From left to right, the following quantities are plotted against simulation time for each complex:

Left Panel: Potential Energy (kJ/mol) for all four complexes.

Middle-Left Panel: Temperature (K) for all four complexes.

Middle-Right Panel: Pressure (bar) for all four complexes.

Right Panel: Density (kg/m³) for all four complexes.

Interpretation. The convergence and stability of all four parameters for each complex confirm that the simulations reached thermodynamic equilibrium, validating the subsequent 100-ns production runs used for detailed analysis.

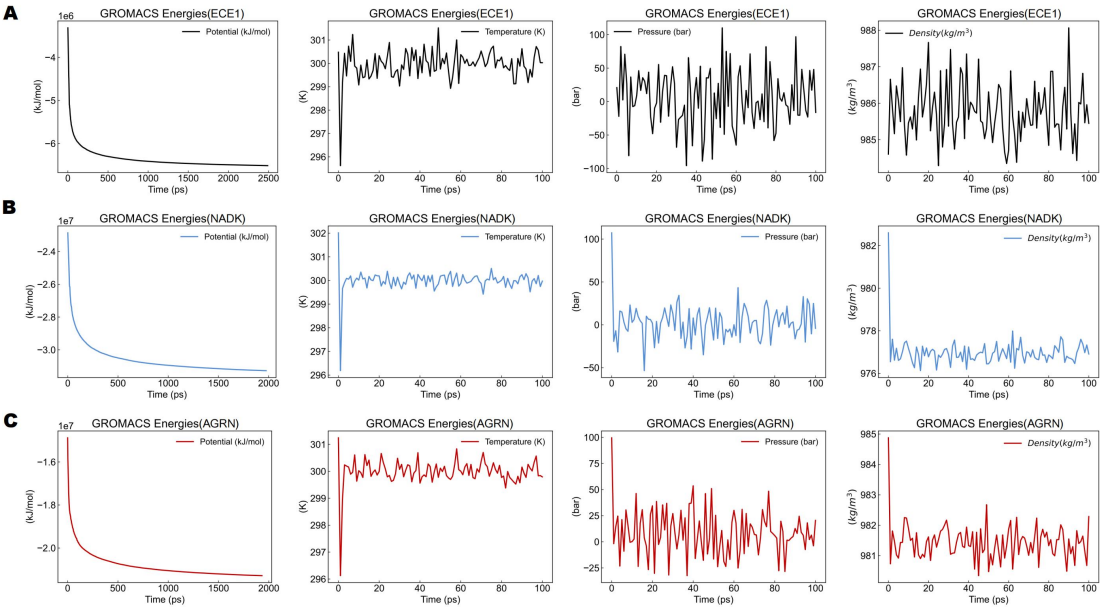

Figure S14. System equilibration monitoring during 100-ns molecular dynamics simulations (test set).

Overall Layout. The figure presents the time evolution of key thermodynamic parameters during the equilibration phase of molecular dynamics simulations for three protein-NNK: ECE1(A), NADK(B) and AGRN(C). The layout is a 2×2 matrix: each row tracks a specific parameter for all complexes, and each column corresponds to a single complex.

Parameters Monitored. From left to right, the following quantities are plotted against simulation time for each complex:

Left Panel: Potential Energy (kJ/mol) for all four complexes.

Middle-Left Panel: Temperature (K) for all four complexes.

Middle-Right Panel: Pressure (bar) for all four complexes.

Right Panel: Density (kg/m³) for all four complexes.

Interpretation. The convergence of all four parameters to stable plateaus indicates that the molecular dynamics simulation reached a robust state of thermodynamic equilibrium. This validates the stability of the simulation and the reliability of the subsequent production phase for structural analysis.