**1. Habitat-Level Spatial Feature Framework for AI Agent Training**

The spatial organization of cellular states in NSCLC is not stochastic but instead reflects a hierarchically encoded ecological logic, wherein topological coherence across diverse cellular populations defines a conserved spatial grammar of tumor biology. To resolve these mesoscale communities, we implemented a sliding-window algorithm that aggregates lineage-annotated single cells within a fixed 40-pixel radius, thereby delineating spatially coherent cellular neighborhoods (CNs). This unsupervised framework revealed ten canonical CNs that recapitulate archetypal tumor ecologies with distinct immunologic or stromal imprints, spanning proliferative tumor hubs, fibrotic stromal scaffolds, and immune-engaged or immune-excluded compartments. The resulting taxonomy encompassed: a tumorigenic core (CN01), macrophage-enriched niche (CN02), B cell cluster (CN03), fibrotic stromal domain (CN04), plasma cell zone (CN05), neutrophil-dominant habitat (CN06), tumor–stroma interface (CN07), T lymphocyte-enriched compartment (CN08), pan-immune activation focus (CN09), and vascular niche (CN10).

We next profiled the distribution of CNs across tumor cores and aligned their spatial architectures with clinicopathologic variables. In squamous cell carcinoma (SCC), CN06 (neutrophil dominant) was preferentially enriched in high-grade tumors, whereas CN08, CN09, and CN10, reflective of lymphoid and vascular activation, were more prominent in non-invasive and early-stage lesions. In adenocarcinoma (ADC), CN06 was likewise elevated in poorly differentiated, large-diameter tumors; in contrast, CN07 and CN08 were enriched in smaller lesions, with CN08 and CN09 also displaying selective enrichment in high-grade tumors of limited size. Although within-subtype prognostic associations were modest and only ADC CN08 demonstrated a positive correlation with overall survival, more consistent patterns emerged through integrative modeling of the entire NSCLC cohort. Univariate Cox analyses identified CN01, CN04, and CN06 as tumor–myeloid consortia conferring adverse prognosis, while CN03, CN05, CN07, and CN08, enriched in B cells, plasma cells, and vascular immune circuitry, delineated immuno-supportive niches consistently predictive of favorable outcomes. Importantly, these associations were independently validated in a separate TMA cohort, underscoring the reproducibility and generalizability of CN-based spatial signatures. Collectively, these findings establish CN-defined ecosystems as biologically coherent and prognostically informative spatial units that transcend histologic subtype and refine clinical risk stratification in lung cancer.

The discordant association of CANVAS ecotypes with immunotherapeutic signatures across cohorts underscores the limitations of fixed spatial clustering in resolving the complex immunoarchitecture of treatment response, highlighting the need for integrative and multiscale spatial biomarkers in immuno-oncology. To overcome this limitation, we developed CANVAS (Cellular Architecture and Neighborhood Informed Virtual AI Driven Spatial Profiling), a spatial recomposition framework designed to reassemble fragmented contextual niches into integrative and functionally coherent spatial modules. Unlike conventional approaches that rely primarily on habitat abundance or categorical classification, CANVAS focuses on emergent spatial features that capture ecological diversity, mesoscale spatial organization, intermodular connectivity, topographic distance metrics, and entropy based measures of spatial transition independent of distance.

To translate this spatial modeling framework into actionable analytic output, we applied CANVAS to the Stanford neoadjuvant immunotherapy discovery cohort (n = 149) and extracted a structured panel of 262 spatially resolved features. These features were hierarchically organized into six functional domains, each reflecting a distinct spatial principle that integrates higher order intercellular and extracellular ecological structuring: (1) Composition, quantifying the relative abundance of each habitat across individual tumor sections (n = 10); (2) Diversity, capturing the ecological complexity of habitat mixtures (n = 6); (3) Spatial dispersion, characterizing intra-habitat organization using multiscale statistical descriptors (n = 90); (4) Interaction, inferring inter-habitat coupling through pairwise proximity modeling (n = 100); (5) Distance, measuring absolute spatial segregation between habitat pairs (n = 55); and (6) Transition, an entropy based, distance independent metric that reflects the degree of spatial intermixing (n = 1).

Within the discovery cohort, five habitats (H01, H04, H06, H07, and H08) emerged as dominant spatial units that collectively defined the architectural scaffold of the NSCLC tumor immune microenvironment. Spatial co-localization analysis revealed recurrent juxtaposition of H03 with H08 and H02 with H10, signifying the formation of immune–vascular convergence zones. In contrast, H01 was spatially segregated from both H04 and H08, establishing an exclusionary topology that reflects immune-privileged niches within hyperproliferative tumor regions. Clinically, tumors enriched for H01, H06, and H10 demonstrated inferior progression-free survival, whereas the presence of H02, H07, and H08 delineated immunologically engaged ecosystems predictive of durable benefit from immune checkpoint blockade. To further evaluate habitat-level spatial alignments associated with therapeutic response, tumors were stratified by clinical outcome and the Scaled Jaccard Index was applied across habitat pairs. Responders displayed increased spatial contiguity between H08 or H02 and tumor-associated habitats such as H01, H04, and H06, suggesting preferential topologic coupling between immune-permissive and tumor-intrinsic niches that facilitates immune infiltration and enhances therapeutic efficacy.

In parallel, we evaluated the prognostic relevance of spatial architecture using habitat-level spatial metrics. Elevated Ripley’s K and L values in H06 and H10 indicated pronounced aggregation and were associated with shorter progression-free survival, implicating tightly clustered neutrophilic and vascular niches in the emergence of immunosuppressive configurations that restrict effector cell infiltration and blunt therapeutic efficacy. In contrast, higher F function in H01 and increased kernel density in H03 and H08 correlated with prolonged progression-free survival, suggesting that spatially permissive or hypocellular tumor regions facilitate immune cell trafficking and amplify antitumor immune engagement. To prioritize predictive features, we applied a robust rank aggregation strategy and visualized representative constellations. One pattern, defined by high abundance of H03 and H08, elevated Shannon diversity, and increased kernel density in H08, was observed in a responder sample characterized by dense TIL infiltration and well-formed tertiary lymphoid structures. Conversely, a non-responder sample exhibited high frequencies of H01 and H04, increased Fisher’s alpha, and elevated F function in H01, consistent with an immune-desert and fibrotic phenotype. These two cases also occupied opposite ends of the spatial transition entropy distribution, each within the top quartile of either high or low values, underscoring the contribution of transition complexity to immunotherapy responsiveness. At the network level, graph-based spatial connectivity analysis identified H01, H04, H06, H07, and H08 as central interaction hubs that anchor the structural framework of the NSCLC tumor immune microenvironment. Collectively, these multiscale spatial features delineate an immunologic landscape in which CANVAS resolves architectural signatures that predict immunotherapy outcomes with high fidelity.

A focused subset of spatial features has previously been shown to capture essential clinico-pathologic determinants. Building on this rationale, we evaluated whether a compact panel of descriptors derived directly from routine H&E images could serve as a reliable predictor of immunotherapy outcome. Using the internal cohort, randomly partitioned into a 75 percent training set and a 25 percent validation set, we applied two independent feature selection strategies following redundancy reduction. Thirteen features consistently overlapped across both methods, converging into a stable and parsimonious predictive signature. A multivariate Cox model trained on this signature demonstrated progressively increasing prognostic accuracy across extended time horizons, with mean AUROCs over 100 iterations exceeding 0.70 at 180 days, 365 days, and 730 days. These findings highlight the capacity of a minimal, interpretable spatial signature to anticipate durable therapeutic benefit with clinical precision.

The 13-feature immunotherapy model demonstrates coherent biological interpretability, highlighting the predictive prominence of immune-centric spatial patterns within the tumor immune microenvironment over tumor-intrinsic compartments. Protective features (β < 0) predominantly reflect immune-activated architectures. The spatial juxtaposition of H08, enriched in helper and cytotoxic T lymphocytes, with H03, a B cell–dominant niche, indicates coordinated T–B cell interactions and suggests the presence of tertiary lymphoid structures, a hallmark of adaptive immune engagement and a robust indicator of immunotherapy responsiveness. The adjacency of H01 and H03 may further potentiate antigen presentation, fostering T cell priming that is essential for therapeutic efficacy. Supporting this interpretation, low J-function values in H03 reflect a spatially regular B cell distribution, consistent with organized immune zones linked to favorable outcomes. Elevated habitat richness, reflecting diverse yet structured ecological architectures, similarly encodes organized immunologic compartmentalization conducive to effective antitumor responses rather than disordered heterogeneity.

In contrast, risk features (β > 0) trace spatial correlates of immune suppression. An increased frequency of H06 corresponds to the accumulation of neutrophils expressing HIF1A and ARG1, consistent with N2-like tumor-associated neutrophils implicated in immune evasion. Interactions between H04, a fibroblast-enriched region, and H09, comprising diverse immune cells, suggest that within fibrotic and regulatory T cell–rich immunosuppressive microenvironments, NK cells frequently undergo functional exhaustion, synergistically impairing antitumor immunity and reducing the efficacy of checkpoint blockade. Frequent H05–H03 contacts, involving plasma cells and B cells, may represent ineffective or exhausted humoral responses characterized by expansion of non-neutralizing antibody activity. Similarly, close proximity between H02, populated by macrophages and dendritic cells, and H09 may promote M2 polarization and anti-inflammatory cytokine signaling such as IL-10 and VEGF-A, thereby suppressing antigen presentation. Increased pair correlation within H02 indicates dense macrophage clustering, potentially marking immunosuppressive TAM hubs that restrict effector cell access. Collectively, these spatial motifs delineate the immune-architectural logic through which the tumor microenvironment resists immunotherapy and underscore the mechanistic basis for the predictive power of the 13-feature model.

Crucially, the CANVAS-derived risk score exhibited robust discriminatory capacity in stratifying patients according to immunotherapy responsiveness, revealing markedly divergent progression-free survival trajectories between responders and non-responders irrespective of PD-L1 expression status. Surpassing the predictive utility of PD-L1, CANVASX further delineated immunologically responsive subsets within EGFR- and KRAS-mutant tumors, molecular subtypes that have historically been refractory to immune checkpoint blockade. Notably, its prognostic fidelity remained stable across therapeutic modalities, including both monotherapy checkpoint inhibition and combination chemoimmunotherapy regimens. In multivariate Cox proportional hazards analysis, CANVAS emerged as the only independent predictor of progression-free survival aside from smoking history, underscoring its clinical relevance as a spatially encoded biomarker of therapeutic benefit.

**2. Habitat Identification and Nomenclature**

We defined ten canonical habitats that capture the ecological heterogeneity of the NSCLC tumor immune microenvironment. These include H01, the tumorigenic core, dominated by proliferating cancer cells; H02, a macrophage-enriched niche; H03, a B cell–enriched habitat; H04, a fibrotic activity hub characterized by cancer-associated fibroblasts; H05, a plasma cell–enriched zone; H06, a neutrophil-prominent compartment; H07, the tumor–stroma interface; H08, a T lymphocyte–enriched region; H09, a pan-immune activation zone with diverse immune infiltration; and H10, a vascular niche associated with endothelial and perivascular elements. Together, these habitats provide a structured taxonomy that links cellular composition with spatial architecture, forming the foundation for ecological modeling of tumor–immune dynamics.

**3. Categorization and interpretation of six spatial feature domains used in NSCLC immunotherapy prognostic modeling**

|  |  |  |  |
| --- | --- | --- | --- |
| Feature type | Definition & spatial role | Biological interpretation (in TIME context) | Calculation notes |
| Composition | Quantifies the frequency or proportion of each habitat per image | Reflects abundance of specific microenvironmental states; for example, high H01 indicates tumor core dominance | Proportion of habitat labels across all patches per sample |
| Diversity | Measures ecological complexity of habitat composition (e.g., Shannon, Simpson, Richness) | High diversity (e.g., high Shannon with enrichment of H02, H03, H08) indicates broader immune repertoire and better surveillance | vegan::diversity, specnumber, fisher.alpha; Shannon = entropy of habitats |
| Spatial dispersion | Describes intra-habitat distribution (Ripley’s K/L, Clark–Evans, F/G/J, Kernel, Quadrat) | Aggregation (high K/L) indicates immune exclusion barriers; uniformity (high Clark–Evans) indicates permissive environments | Point pattern analysis: Ripley’s K/L, Clark–Evans, Quadrat test, Kernel density |
| Interaction | Captures inter-habitat neighborhood relationships (e.g., GNN edge weights, SCIMAP Z-scores) | High interaction scores (e.g., H01–H03 or H01–H08) may indicate immune clustering or surveillance-associated niches | SCIMAP spatial\_interaction, Z-score, or GNN-derived edge weights |
| Distance | Pairwise nearest-neighbor or exclusion distances between habitats | Shorter distances suggest active immune infiltration; longer distances reflect exclusion or spatial segregation | Patch-level pairwise distances, averaged then ranked |
| Transition | Spatial Transition Entropy between habitats across image patches | High entropy reflects complex and intermixed microenvironments; low entropy indicates structured or compartmentalized tissue | Entropy of patch-wise habitat transition matrix (non-distance-based entropy) |

**4. Biological interpretation of spatial dispersion features (Cox model coefficients β) in relation to prognosis**

|  |  |  |
| --- | --- | --- |
| Feature type | β > 0 (Adverse implication) | β < 0 (Favorable implication) |
| Ripley’s K / L | Greater aggregation, indicative of immune-exclusion barriers | More dispersed distribution, permissive for immune cell infiltration |
| F function | Increased sparsity (larger empty-to-cell distances) associated with better outcome | Tight clustering may mark suppressive niches |
| G function / Kernel density | Higher density favorable in immune-enriched regions | High density within tumor cores may predict poor prognosis |
| Clark–Evans index | Aggregated distribution unfavorable | More uniform distribution linked to favorable prognosis |
| J function | High values reflect aggregation-driven suppression (unfavorable) | Low values indicate regular and sparse patterns (favorable) |

**5. End-to-end analytic roadmap for habitat-informed prognostic modeling in NSCLC**

In this study we proceed through seven coordinated steps to convert patch-level CN predictions from H&E into clinically meaningful predictors of immunotherapy outcome: (1) Habitat composition profiling quantifies per-image CN frequencies and CN co-occurrence, followed by univariate Cox screening of each CN’s association with PFS to nominate composition features for modeling. (2) Ecological diversity analysis computes Richness, Shannon, Simpson, Inverse Simpson, Pielou evenness, and Fisher’s alpha at the image level; diversity indices with protective or adverse signals in Cox screening are retained as global context features. (3) Intra-habitat spatial dispersion assessment derives CN-specific point-pattern metrics (Ripley K and L, pair correlation g, G and F functions, J function, Clark–Evans index, quadrat χ², kernel density) and relates them to PFS using partial correlation; representative H&E tiles with overlaid CN labels are displayed to illustrate clustered versus permissive architectures. (4) Graph neural network connectivity mapping builds a cell-level graph, extracting edge counts and GCN-based scores to quantify network centrality and interaction intensity of each CN. (5) Inter-habitat interaction modeling applies SCIMAP spatial\_interaction at the same radius to obtain Z-scored CN–CN proximity networks, followed by univariate Cox evaluation to nominate interaction edges as candidate features. (6) Mesoscale distance quantification measures nearest-neighbor and exclusion distances between CN pairs independent of predefined radii, with Cox screening to capture prognostically relevant spatial segregation or contiguity. (7) Feature selection and prognostic modeling performs redundancy reduction, bootstrap LASSO-Cox and Random Forest–based importance ranking, then trains the final multivariate survival model in the training set and independent validation, accompanied by a schematic that summarizes the selected features and their biological interpretation.

**6. Habitat-Level Spatial Feature Analysis in NSCLC Immunotherapy Cohorts**

1. Frequency and Composition Features (CN level)

* Approach: For each H&E image, the proportion of patches annotated as CN01–CN10 was computed. Univariate Cox regression was used to evaluate associations with PFS.
* Key Findings:
* CN01 (tumorigenic core): Higher frequency associated with poor PFS (HR = 1.99, p = 0.0058).
* CN02 (macrophage-enriched): Higher frequency associated with favorable PFS (HR = 0.61, p = 0.0355).
* CN06 (neutrophil-prominent): Higher frequency associated with poor PFS (HR = 1.77, p = 0.04).
* CN07 and CN08: Frequencies associated with favorable PFS.
* Other CNs showed weaker or no significant associations.

Interpretation: CN composition captures tumor–immune balance; tumor-dominant or neutrophil-rich states predict poor outcome, whereas macrophage or T-cell enriched states reflect protective niches.

2. Diversity Metrics (image level)

* Metrics computed: Richness, Shannon, Simpson, Inverse Simpson, Pielou evenness, Fisher’s alpha.
* Key Findings:
* Shannon index: Significant protective factor (HR = 0.675, p = 0.046).
* Simpson / Inverse Simpson: Borderline protective (p ≈ 0.06, HR ≈ 0.70).
* Fisher’s alpha: Borderline adverse (HR = 1.47, p = 0.07), suggesting rare CN dominance may predict worse outcome.
* Richness and Pielou: Not significant but trended towards favorable associations.

Interpretation: Greater ecological diversity generally confers favorable outcomes, consistent with broader immune repertoire, while disproportionate enrichment of rare CNs (high Fisher alpha) may signal dysregulated immunity.

3. Intra-Habitat Spatial Dispersion Features (CN level)

* Metrics computed per CN: Ripley’s K/L, pair correlation g(r), G/F/J functions, Clark–Evans index, Quadrat χ², kernel density.
* Key Findings:
* Ripley’s K/L: Aggregation of CN05, CN06, CN10 linked to poor PFS.
* F function: Higher values in CN01 and CN05 associated with improved PFS.
* Clark–Evans index: Higher values in CN06 associated with improved PFS.
* Kernel density: Higher values in CN03, CN07, CN08 correlated with better survival.

Interpretation:

* Aggregation (high K/L) reflects immune-exclusion barriers.
* Dispersed or uniform distributions (F function, Clark–Evans) indicate more permissive immune environments.
* Dense T/B-cell clusters (kernel density) capture TLS-like or immune-hot regions predictive of responsiveness.

4. GNN-Derived Connectivity (400 µm radius)

* Approach: Graph neural networks were built at a fixed 800-pixel radius, extracting edge weights and GCN centrality scores.
* Purpose: To quantify which CNs act as spatial hubs with strong interconnectivity.
* Outcome: CN01, CN04, CN06, CN07, CN08 consistently emerged as central nodes anchoring the TIME.

5. Inter-Habitat Interaction Features

* Approach: SCIMAP spatial\_interaction with permutation-based Z-score normalization (400 µm radius) across all CN pairs.
* Use: Identified enriched or depleted CN–CN proximities relative to null expectation.
* Application: Univariate Cox analysis linked certain proximities (e.g., H03–H08, H02–H10) to differential survival.

6. Inter-Habitat Distance Metrics

* Approach: Euclidean nearest-neighbor distances between CN pairs computed independent of radius.
* Interpretation:
* Shorter distances → immune infiltration into tumor habitats (favorable).
* Longer distances → immune exclusion and compartmentalization (unfavorable).

7. Integrated Clustering of Features

* Correlation analysis identified 5 clusters:
* Cluster 1 (immune-active): freq\_CN02/03/08, kernel\_density\_CN03/08 → favorable PFS.
* Cluster 2 (Fisher-aberrant / atypical immune): freq\_CN04/06/09, Fisher\_alpha → unstable, suppressive patterns.
* Cluster 3 (non-aggregated immune-supportive): F\_function\_CN01/05, Clark–Evans\_CN06 → favorable.
* Cluster 4 (immune-complexity/diversity): freq\_CN07, kernel\_density\_CN07, Richness, Shannon, Simpson, Pielou → favorable.
* Cluster 5 (tumor-suppressive barrier): Ripley’s K/L of CN05/06/10, freq\_CN05 → poor outcome.

8. Spatial Transition Entropy (STE)

* Definition: Non-distance-based entropy metric quantifying complexity of CN label transitions across k-nearest neighbors (k = 6).
* Interpretation:
* High STE = highly intermixed or blurred niche boundaries, reflecting ecological complexity and immune–tumor interdigitation.
* Low STE = sharply compartmentalized niches, consistent with immune exclusion.
* Relevance: Provides a complementary measure of structural mixing beyond distance metrics.
* References: Shannon entropy–based transition approaches in ecology and spatial statistics.

Overall Summary:

* Composition (CN frequency), diversity, and dispersion metrics capture complementary ecological principles of tumor architecture.
* GNN, interaction, distance, and STE extend analysis to higher-order connectivity and intermixing.
* Combined, these multi-scale spatial features provide a robust framework for machine-learning–based prognostic modeling in NSCLC immunotherapy.

**7. Hierarchical levels of spatial ecology terminology and their relevance to feature modeling**

|  |  |  |  |
| --- | --- | --- | --- |
| Term level | Description | Suggested cellular layer | Explanation |
| Cell | Single-cell level (e.g., H&E nuclear segmentation, cell type annotation) | Intra-cellular | Captures intrinsic cell features such as maker expression, nuclear morphology, or transcriptional states |
| Intercellular interaction | Direct relationships between two neighboring cells (e.g., adjacency, L–R analysis) | Inter-cellular | Describes spatial interactions between individual cells, including immunosuppressive axes such as PD1–PDL1 |
| CN with cell state | Aggregation of cells into neighborhoods (CNs) incorporating lineage and activation/exhaustion status | Inter-cellular or transition-zone | Reflects construction of micro-niches that span multiple cell states and functional phenotypes |
| Habitat | Mesoscale ecological unit composed of multiple cells | Inter-/extra-cellular | Defines intra-habitat cell interactions and their interfaces with surrounding tissue compartments |
| Ecotype | Sample-level ecological composition integrating multiple habitats | Extra-cellular | Represents macro-environmental signatures that infer system-level immune states |
| Feature recomposition | Reconstructed spatial descriptors based on multi-scale principles (e.g., composition, diversity, STE) | All layers | Crosses intra-, inter-, and extra-cellular levels; captures integrated features of tumor–immune ecology |
| Model prediction | Application of spatial features for prognostic or immunotherapy-response modeling | All layers | Incorporates spatial descriptors from all levels as input for predictive modeling |