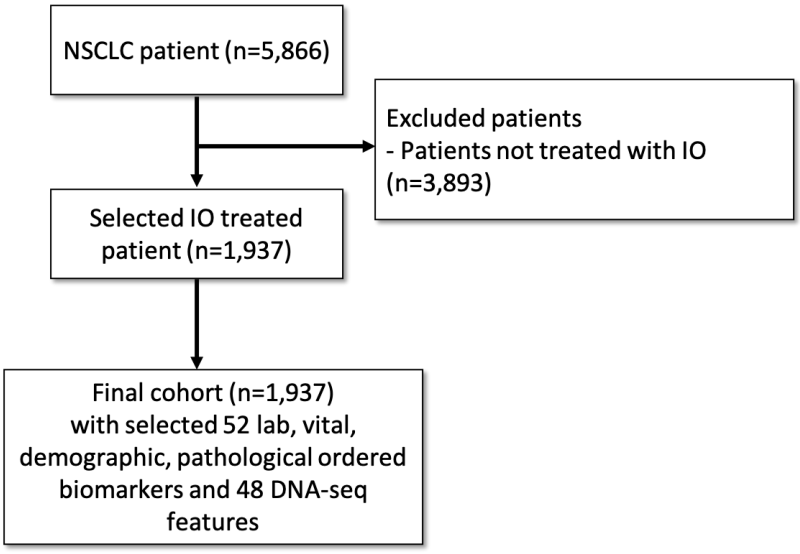
DeePaN: A deep patient graph convolutional network integrating clinico-genomic evidence to stratify lung cancers benefiting from immunotherapy

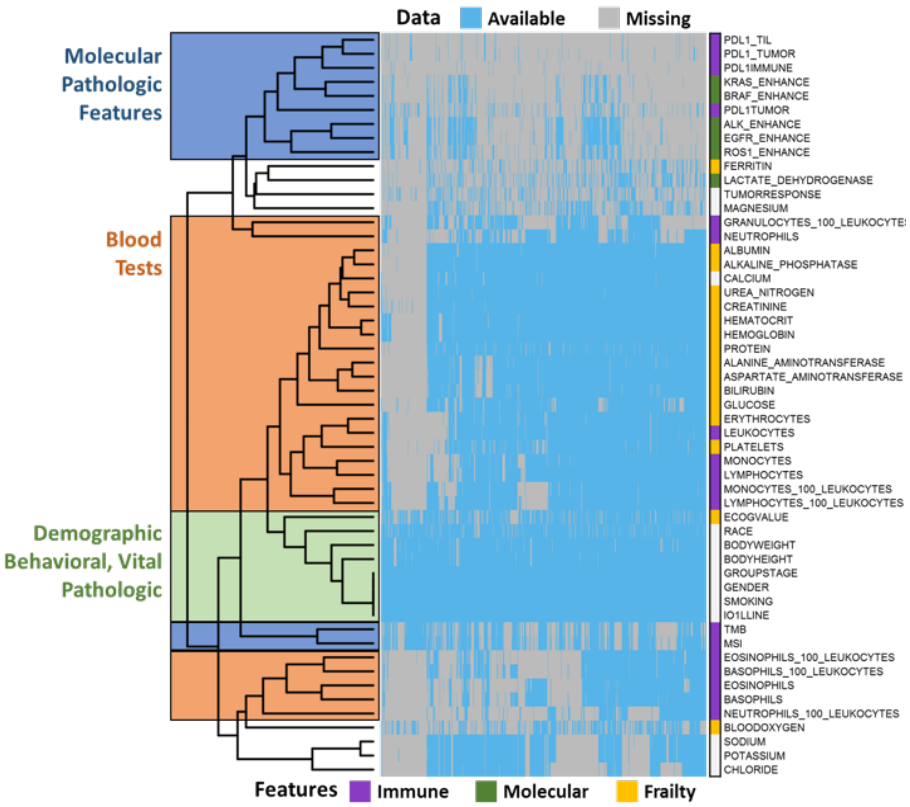
I. Supplemental Figures

Cohort characteristics

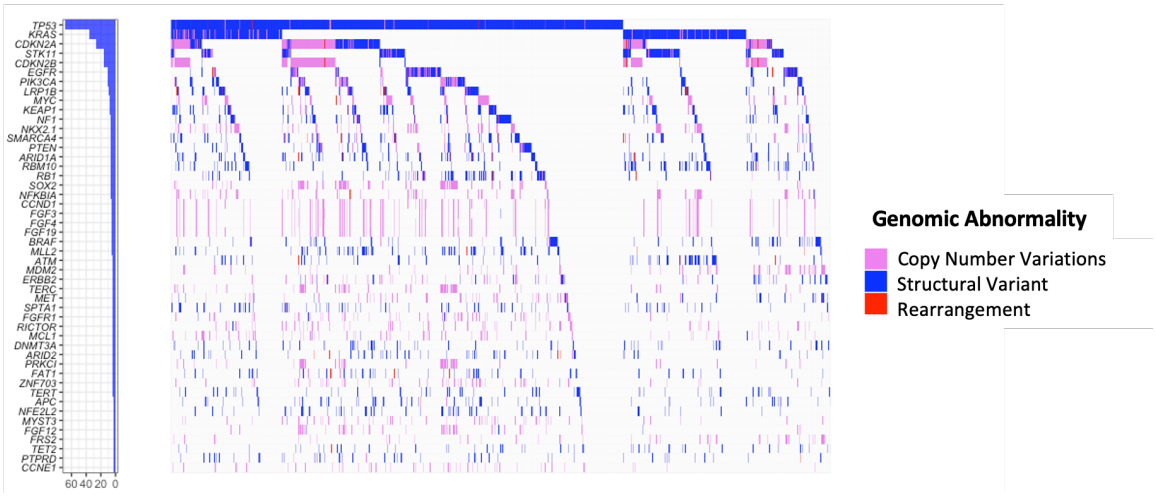
The cohort selection process, baseline demographic and pathologic characteristics is shown in Supplemental Figure 1: **(A)** shows how the cohort was identified with inclusion and exclusion criteria. **(B)** shows the clinical features, such as hemoglobin, erythrocytes, hematocrit, etc. They can be categorized as molecular pathological features, blood test, and demographic behavioral and vital pathologic features. **(C)** shows the genomic features visualized by the waterfall plot of gene mutation profiling. Each row represents a gene and each column represents a patient. The mutation type can be SV (structural variant), CN (copy number variations), and RE (rearrangement). The potentially actionable somatic mutations found in this study are consistent with prior studies[1-3].



(A)



(B)



(C)

Supplemental Figure 1. NSCLC IO treated patient cohort and visualization of their clinical and genomic features A) Cohort Identification: an illustration of how patient cohort was identified using inclusion and exclusion criteria in this study. **B)** visualization

of clinical and genomic features in the study cohort. Features are categorized into molecular pathology features, blood test features, etc. Gray color indicates missingness in the feature. Note that “Tumor response” is not included as an input feature. **C)** Waterfall plot of DNA alterations in the study cohort. The genes are sorted based on frequency.

II. Supplemental Tables

Table 1. Baseline Demographic and Pathologic Characteristics			
Characteristics	**All**	**IO Beneficial Subgroup**	**IO Non-beneficial Subgroups**
number of patients	**1,937**	**400**	**897**
Age (year)			
Median, MAD	67.0, 10.4	67.0, 10.4	67.0, 10.4
Range	26.0-85.0	26.0-85.0	28.0-85.0
Sex: no., %			
Male	984 (50.8)	148 (37.0)	522 (58.2)
Female	953 (49.2)	252 (63.0)	375 (41.8)
Race			
African American	144 (7.4)	31 (7.8)	80 (8.9)
White	1,428 (73.7)	289 (72.3)	648 (72.2)
Asian	46 (2.4)	10 (2.5)	19 (2.1)
Other Race	143 (7.4)	41 (10.3)	68 (7.6)
Histology			
Non-squamous cell carcinoma	1,433 (73.9)	329 (82.2)	601 (67.0)
Squamous cell carcinoma	419 (21.6)	62 (15.5)	259 (28.9)
NSCLC histology NOS	75 (3.8)	9 (2.3)	37 (4.1)
Stage: no., %			
Stage I	164 (8.5)	45 (11.3)	69 (7.7)
Stage II	122 (6.3)	30 (7.5)	55 (6.1)
Stage III	372 (19.2)	95 (23.8)	176 (19.6)
Stage IV	1,241 (64.1)	225 (56.3)	571 (63.7)
ECOG Score: no., %			
0	375 (19.4)	104 (26.0)	129 (14.4)
1	856 (44.2)	176 (44.0)	430 (47.9)
2	273 (14.1)	27 (6.8)	149 (16.6)
3	50 (2.6)	6 (1.5)	27 (3.0)
4	2 (0.1)	0 (0.0)	2 (0.2)
Smoking Status: no., %			
History of smoking	1,657 (85.5)	342 (85.5)	775 (86.4)
No history of smoking	276 (14.2)	57 (14.3)	120 (13.4)
Previous Treatment: no., %			
No	718 (37.1)	170 (42.5)	228 (25.4)
Yes	1,219 (62.9)	230 (57.5)	669 (74.6)
Eastern Cooperative Oncology Group (ECOG)			
MAD: Median Absolute Deviation.			
(·) represents percentage of patients			

Supplemental Table 1: Baseline demographic and pathologic characteristics for the overall IO cohort, IO beneficial subgroup, and IO non-beneficial subgroup.

Feature name	FDR	IO Beneficial Subgroup	IO Non-beneficial Subgroup
HEMOGLOBIN	1.60E-178	Normal	Low
HEMATOCRIT	2.49E-168	Normal	Low
ERYTHROCYTES	5.12E-130	Normal	Low
ALBUMIN	1.66E-23	Normal	Low
FERRITIN	5.39E-17	Normal	High
LYMPHOCYTES	2.92E-14	Normal	Low
GENDER	2.67E-10	Female	Male
NEUTROPHILS	2.52E-08	Normal	High
BASOPHILS	4.01E-06	Abnormal	Normal
PROTEIN	6.31E-06	Normal	Abnormal
MONOCYTES	3.70E-05	Normal	High
LYMPHOCYTES %	6.99E-04	Normal	Low
PLATELETS	1.16E-03	Normal	High
GRANULOCYTES %	1.35E-03	Normal	High
LEUKOCYTES	5.54E-03	Normal	High
EOSINOPHILS	1.29E-02	Normal	Abnormal
NKX2.1	2.04E-02	Altered	Wild-type
KRAS	3.55E-02	Altered	Wild-type
CALCIUM	3.61E-02	Normal	Abnormal
GLUCOSE	4.74E-02	Normal	High
TMB	4.90E-02	High	Low

Supplemental Table 2: Enriched clinical and genomics characteristics differentiating the IO beneficial vs non-beneficial subgroups. Enriched features with significant difference between the IO beneficial and non-beneficial subgroups were identified by chi-square test with the multiple-hypothesis adjustment of FDR less than 0.05.

III Supplemental Methods

Clinico-genomic feature encoding and defining linked patients based on feature similarity in the deep patient graph convolutional network modeling

In the deep patient graph convolutional network modeling, patients are represented as nodes in the graph with associated clinic-genomic features, and patients with similar clinico-genomic features are linked by edges. The node features are encoded by categorical feature vectors X. In particular, the genomic features are binary encoding, i.e. if a patient carries one or more known or likely genetic alternations in a gene, the corresponding gene feature is 1; otherwise, 0. For numerical features, we used the high- and low-bound measurement annotations provided by EHRs to bin the numerical features into categorical features. For example, a patient has the hemoglobin measurement as 8.3 grams per deciliter, the low- and high-bound references for hemoglobin is 14 and 18 grams per deciliter, respectively. Since it falls between two bounds, it is categorized as the "normal" class. The two nodes are connected if the node feature vectors are similar. Here we employed cosine similarity to define similarity [4] and used the cosine similarity of 0.5 as an empirical cutoff. If cosine similarity is less than 0.5, then there is not a link between two nodes; otherwise, connected.

Graph convolutional network

Graph convolutional network (GCN) [5] applies the convolution operation on a graph from the spectral domain. Given the adjacency matrix A and content matrix X of a graph, the spectral convolution function f used to calculate layer-wise transformation is defined as:

$$Z^{(l+1)} = f(Z^{(l)}, A)$$

Here, $Z^{(l)} \in R^{n \times d}$ (n nodes and d features) defines the input for layer l. The input layer contains the patient clinical and genomic feature matrix for our problem. The feature dimension of the input layer is 227, which was derived from the original 100 features. Our graph model has three hidden layers and the embedding dimension is the same as input layer 227. The MGAE embedding method reconstructs the feature matrix of a node without hidden layers.

GCN [5] applies Chebyshev polynomials [6] to approximate the convolution filter. The layer-wise propagation rule for GCN can then be defined as:

$$f(Z^{(l)}, A) = \sigma(DZ^{(l)}W^{(l)})$$

Here, D is the degree matrix for A. $W^{(l)}$ is the learnable weights for the l-th layer. $\sigma(\cdot)$ is an activation function such as ReLU [7].

Marginalized graph autoencoder (MGAE)

The MGAE [8] is a content and structure augmented autoencoder. MGAE reconstructs the input $X = \{x_1, \dots, x_n\} \in R^{n \times d}$ by using a single mapping function $f(\cdot)$, that minimizes the squared reconstruction loss:

$$\|X - f(X)\|^2$$

For graph convolution networks, the loss function becomes:

$$\|X - DXW\|^2 + \lambda\|W\|_F^2$$

Here, D is the degree matrix for A. W is the parameter matrix. $\|W\|_F^2$ is a Frobenius norm regularization term with coefficient λ being a tradeoff.

The marginalized graph autoencoder provides an effective way to integrate both content and structure information. To encourage the interplay between content and structure information, MGAE introduces some random noises into the content features

during training. The corruption process can be randomly removing some features or setting them to 0. Given the corrupted version of the original input X, the corrupted version of original input X is:

$$\tilde{X} = \{\tilde{X}_1, \dots, \tilde{X}_n\}$$

The objective function becomes the following where m is the number of corruption times:

$$\frac{1}{m}\sum_{i=1}^{m}\left\|X - D\tilde{X}_i W\right\|^2 + \lambda\|W\|_F^2$$

And the final graph embedded representation Z is defined as:

$$Z = \hat{A}XW$$

Patient subtype clustering with MGAE

We applied the spectral clustering algorithm [8] for patient subtyping. The symbol used and pseudo-code is defined as follows:

Given the patient graph network G with n nodes, each patient node is a d-dimension attribute vector. The patient attribute matrix $X \in R^{n\times d}$ of G, the total number of patient subtypes k, the corruption probability p, and the number of stacked autoencoder layers l. In our problem formulation, $l = 3$. $Z = X$ is the input to the first layer.

Step-1:

For each layer:

Construct a single layer denoise autoencoder with input data Z

Learn the output representation Z according to MGAE algorithm

$$Z = \hat{A}XW$$

Step-2:

Refine representation by apply a linear kernel function:

$$Z_0 \Leftarrow Z$$

$$Z_1 \Leftarrow Z_0 Z_0{}^{\mathrm{T}}$$

Make representation symmetric and nonnegative:

$$Z_2 \Leftarrow \frac{1}{2}(|Z_1| + |Z_1^T|)$$

Step-3:

Run spectral clustering on Z_2 (Running k-means on the top-'numberOfCluster' eigenvectors of the normalized Laplacian)

Reference:

1 Jordan, E. J. *et al.* Prospective comprehensive molecular characterization of lung adenocarcinomas for efficient patient matching to approved and emerging therapies. *Cancer discovery* **7**, 596-609 (2017).

2 Network, C. G. A. R. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519 (2012).

3 Network, C. G. A. R. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543 (2014).

4 Li, L. *et al.* Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine* **7**, 311ra174-311ra174 (2015).

5 Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

6 Hammond, D. K., Vandergheynst, P. & Gribonval, R. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis* **30**, 129-150 (2011).

7 Nair, V. & Hinton, G. E. in *Proceedings of the 27th international conference on machine learning (ICML-10).* 807-814.

8 Wang, C., Pan, S., Long, G., Zhu, X. & Jiang, J. in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* 889-898 (ACM).