HEX model architecture and training strategy

The HEX model was designed to predict the expression level for 40 protein biomarkers simultaneously, given an input H&E image patch (typically 224 × 224 pixels in this study). An overview of the model architecture is visualized in Fig. 1a and Extended Data Fig. 1. The backbone of HEX is built on a pretrained pathology foundation model, such as MUSK, to extract visual features from H&E patches. A two-stage regression head follows: a linear layer reducing the visual embedding to 256, followed by ReLU and dropout, then another linear layer projecting to 128 dimensions with ReLU and dropout and finally a linear output layer producing 40 biomarker predictions.

To improve the predictive robustness and handle the inherent challenges in multiplex imaging data, we integrated two key techniques during fine-tuning: FDS[61] and an ALF[62]. Spatial proteomics data such as CODEX exhibit substantial target imbalance: some biomarkers are ubiquitously expressed, whereas others appear infrequently or in sparse regions. To mitigate this, we adopted FDS, a post-hoc feature calibration technique that reduces the negative impact of data imbalance by explicitly smoothing features across similar target values. During training, features from the penultimate layer are first collected and stored for each target bin of biomarker *j* and then updated via the exponential moving average[63]. The calibrated features $\bar{h}$ are obtained by smoothing these bin-level features using a Gaussian kernel $g\left(\bullet\right)$:

$$\bar{h} = \bar{C}_b^{\frac{1}{2}} C_b^{-\frac{1}{2}} \left(h - \mu_b\right) + \bar{\mu}_b$$

where

$$\mu_b = \frac{1}{N_b - 1} \sum_{i \in b} h_i, C_b = \frac{1}{N_b - 1} \sum_{i \in b} (h_i - \mu_b)(h_i - \mu_b)^T$$

$$\bar{\mu}_b = \sum_{m \in B} g\left(y_b, y_m\right) \mu_m, \bar{C}_b = \sum_{m \in B} g\left(y_b, y_m\right) C_m$$

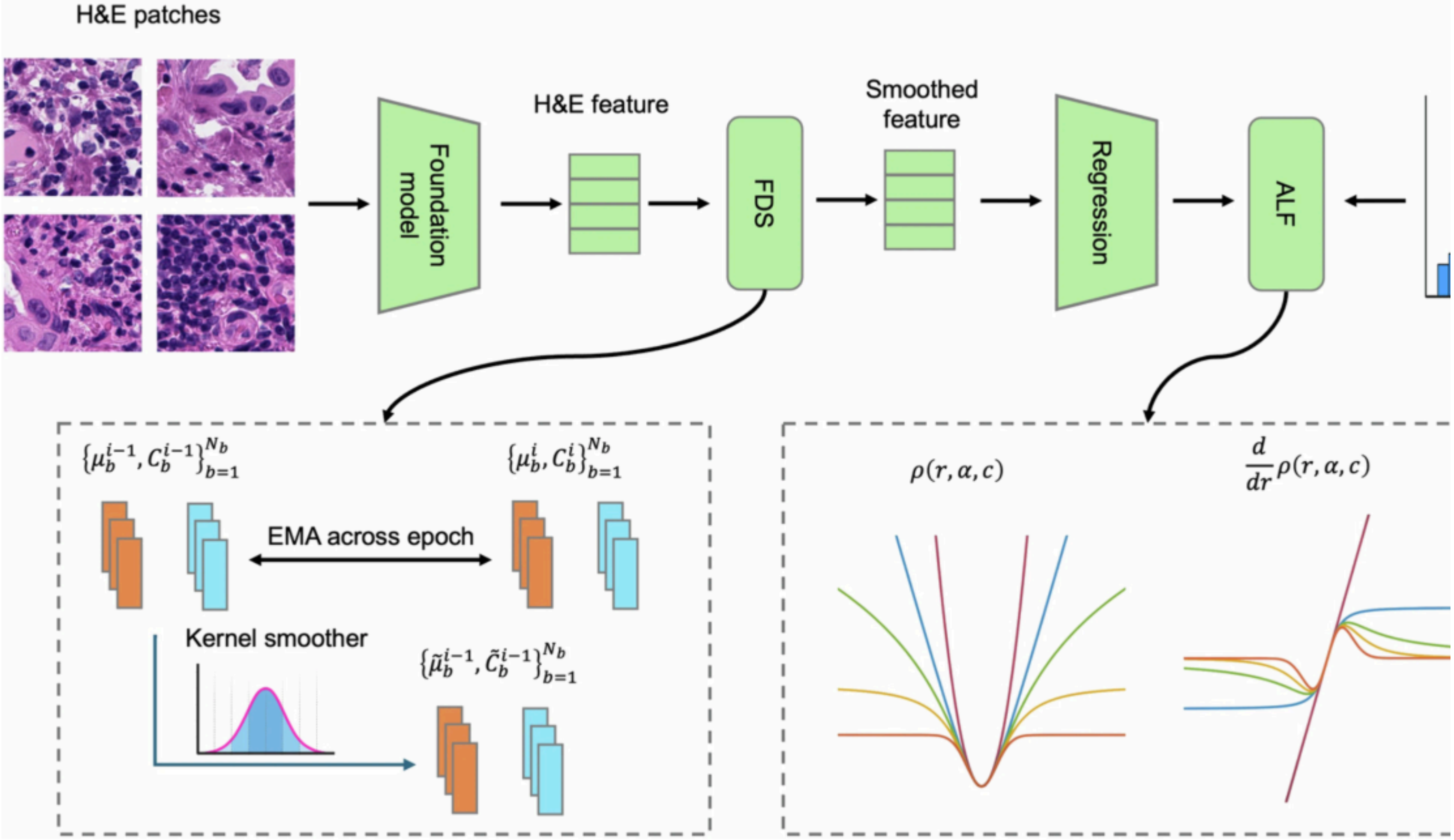

H&E patches
Foundation model
H&E feature
FDS
Smoothed feature
Regression
ALF
$\{\mu_b^{i-1}, C_b^{i-1}\}_{b=1}^{N_b}$
$\{\mu_b^{i}, C_b^{i}\}_{b=1}^{N_b}$
EMA across epoch
Kernel smoother
$\{\tilde{\mu}_b^{i-1}, \tilde{C}_b^{i-1}\}_{b=1}^{N_b}$
$\rho(r, \alpha, c)$
$\frac{d}{dr}\rho(r, \alpha, c)$

Figure S1. HEX is built on a pre-trained pathology foundation model (for example, **MUSK**) to extrac **features from H&E** image patches. A three-layer regression head maps visual embeddings to 40 bior predictions via intermediate 256- and 128-dimensional representations with ReLU activations and dro regularization. To enhance predictive robustness, the model incorporates **Feature Distribution Smoo (FDS)** and an **Adaptive Loss Function (ALF)** during training to address the challenges of multiplexe imaging data. EMA, exponential moving average.

We finally evaluated the impact of using different foundation model backbones and training strategies models initialized from the **MUSK**[22] backbone achieved the highest overall accuracy, whereas model the **CONCH**[24] backbone showed faster inference at the expense of lower accuracy (Supplementary Figs. 1 and 2). To dissect the role of training strategies, we conducted ablation experiments by remov feature distribution smoothing (**FDS**) or adaptive loss function (**ALF**) from the original HEX model. cases, model performance decreased notably across all metrics, indicating that these components are r for achieving robust and generalizable predictions (Supplementary Figs. 3 and 4).

ct **visual**
marker
opout
othing
ed

s. HEX
ls based on

ing either
In both
necessary

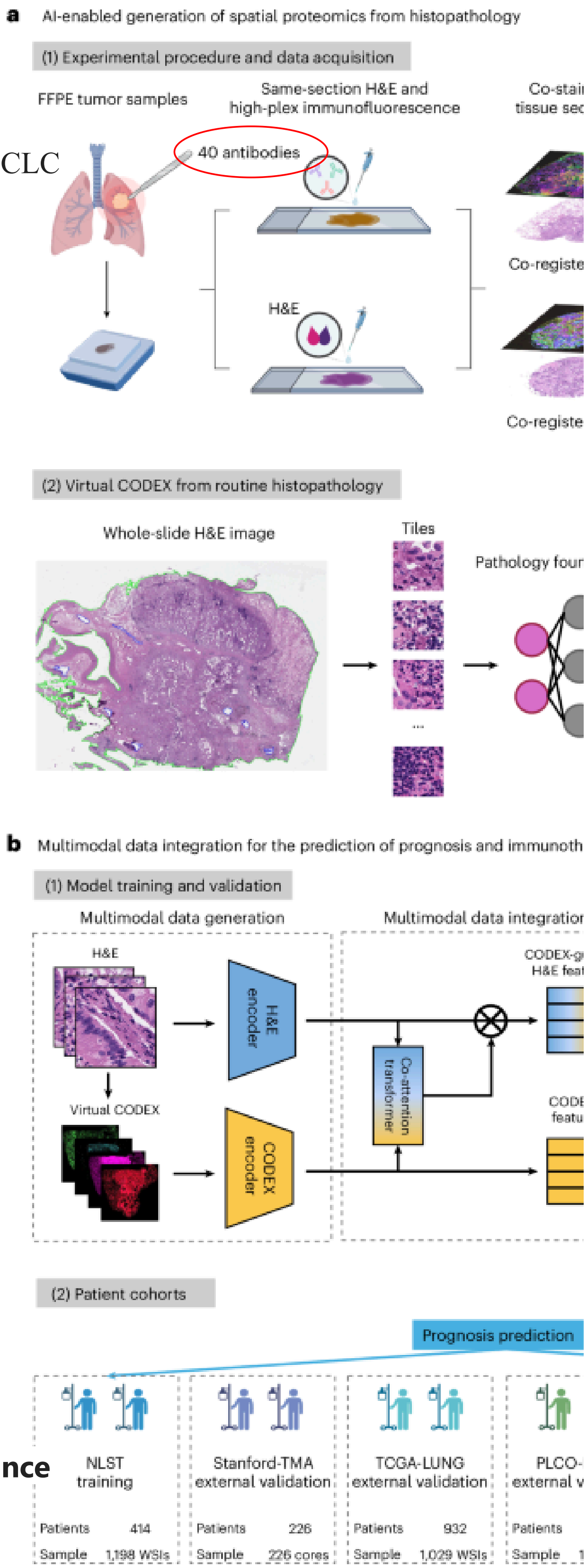
ten patients with NSCLC

The whole-slide images (WSIs) from the two experiments were co-registered and cropped into smaller image tiles measuring ~50 µm.

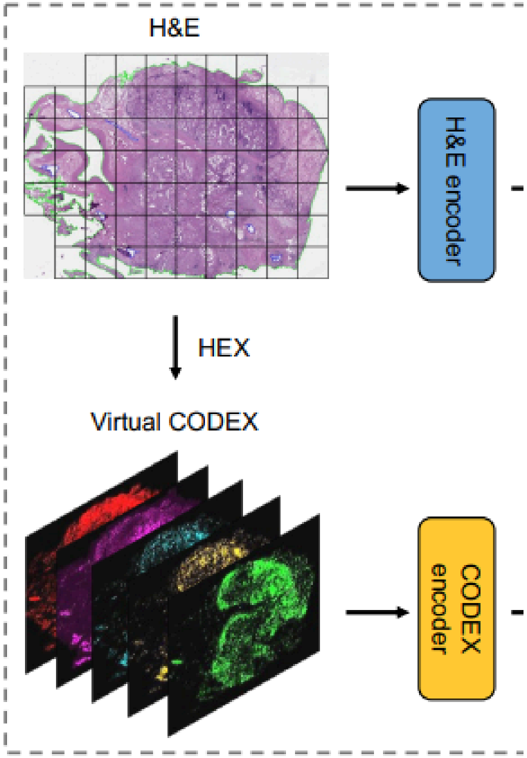
↓

755,000 image tiles with 40 protein biomarkers and matched H&E histopathology

HEX was trained by leveraging state-of-the-art pathology foundation models[22,23,24,25] to predict the expression of 40 protein biomarkers simultaneously based on H&E images



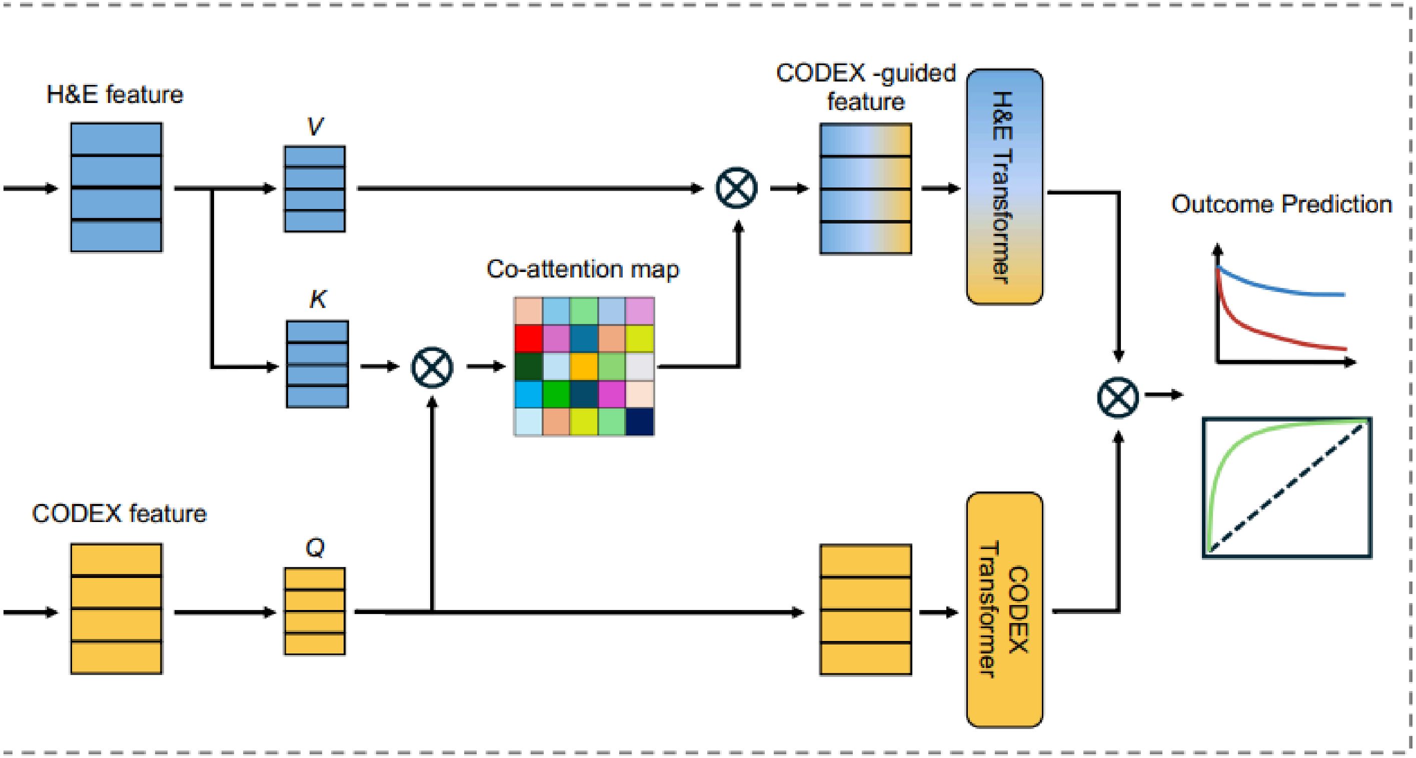
Cross-validation performance



Supplementary Fig. 5 | O
used to extract features fror
to CODEX images to ger
attention layers learn cross
modality-specific multiple-ir
features for outcome predic

Rsult1 : HEX improves prognosis

With the ability to generate spatial pr
clinically relevant outcomes by addir
CODEX maps offer complementary
data types, we developed **multimoda**
virtual CODEX data at an early st
interactions and spatial relationships,



verview of the MICA framework. First, a pathology foundation model is n H&E images, forming histology feature bags. Second, DINOv2 is applied ıerate corresponding CODEX feature bags. Third, CODEX-guided co- ;-modal interactions between histology and CODEX features. Finally, two ıstance learning Transformers with global average pooling aggregate the :tion.

s prediction in early-stage lung cancer

roteomics from standard H&E images, HEX enables biologically interpretable prediction of ıg a new layer of molecular insight. Although H&E provides detailed tissue morphology, virtual information about spatially resolved protein expression. To integrate these distinct yet synergistic **al integration via co-attention (MICA), a deep learning framework that fuses H&E and age** (Supplementary Fig. 5 and Methods). This approach explicitly models cross-modal , enhancing its ability to identify clinically relevant features predictive of patient outcomes.

μ_b and C_b are the mean and covariance of the features with each bin $b \in B$, and N_b is the total number of samples in the bth bin. FDS was applied directly across all biomarkers to jointly regularize and smooth the feature distributions. This explicit calibration in feature space reduces bias in under-represented targets and stabilizes training in imbalanced regression settings.

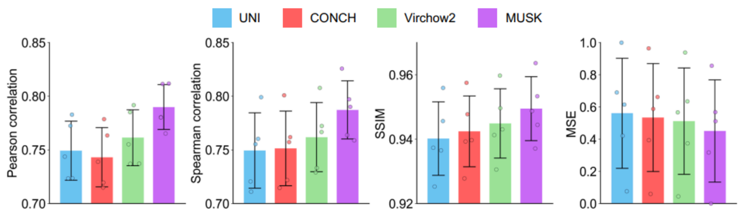
To further mitigate the impact of image noise and outliers commonly observed in CODEX data, we incorporated the ALF into the training objective. The ALF generalizes robust loss functions by introducing the learnable shape α and scale c parameters that modulate the tail behavior of the error distribution. This design allows the model to dynamically interpolate between different loss regimens based on data characteristics. When interpreted as the negative log-likelihood of a univariate probability distribution, ALF enables robustness to be automatically adapted during training, improving generalization to noisy or heterogeneous regions in histopathology. The ALF in this study is defined as

$$\mathcal{L}\left(r,\alpha,c\right)=\rho\left(r,\alpha,c\right)+\log Z\left(\alpha\right)$$

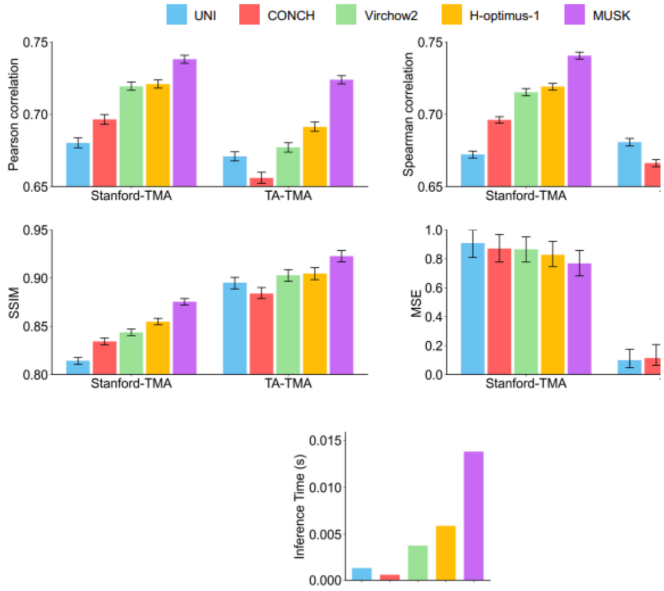
where

$$\rho\left(r,\alpha,c\right)=\frac{|\alpha-2|}{\alpha}\left(\left(\frac{\left(r/c\right)^{2}}{\alpha-2}+1\right)^{\alpha/2}-1\right)$$

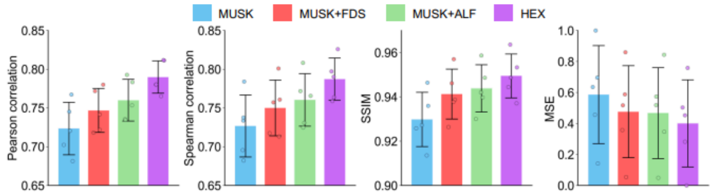
and $Z\left(\alpha\right)$ is the partition function, α is the shape parameter, c is the scale parameter and r is the residual error.



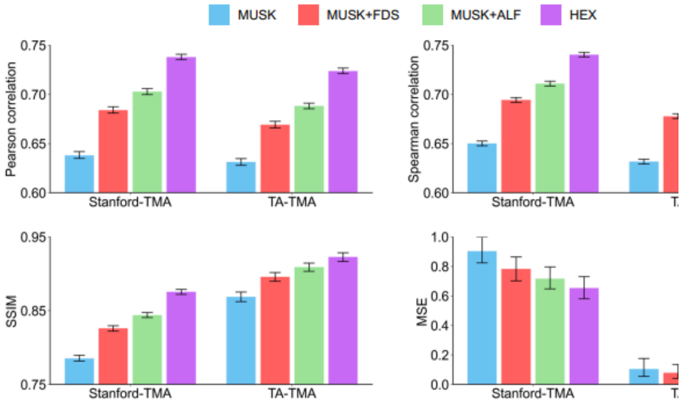
Supplementary Fig. 1 | Performance comparison of foundation model backbones for HEX using five-fold cross-validation on Stanford-WSI cohort. Among the tested backbones, HEX initialized with MUSK achieved the highest predictive accuracy. Bars represent the mean across five-fold cross-validation on the Stanford-WSI dataset (n = 10 WSIs); dots show individual folds and error bars indicate standard deviation.



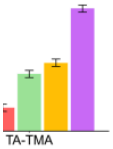
Supplementary Fig. 2 | Performance comparison of foundation model backbon
independent validation cohorts. MUSK-based HEX consistently outperformed oth
achieving the highest predictive accuracy across external datasets (Stanford-TMA, n =
TMA, n = 108 cores). Models based on CONCH, UNI, Virchow2, and H-optimus-1 shov
of performance and speed, reflecting different trade-offs between accuracy and
efficiency. Bars represent point estimates and error bars indicate 95% bootstrap (
resamples).

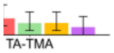


Supplementary Fig. 3 | Performance comparison when using different training strategies for HEX on Stanford-WSI cohort. Ablation experiments were conducted by removing either feature distribution smoothing or adaptive loss function from the full HEX model. In both cases, performance declined notably across all evaluation metrics, highlighting the necessity of these components for achieving robust and generalizable predictions on the training cohort. Bars represent the mean across five-fold cross-validation on the Stanford-WSI dataset (n = 10 WSIs); dots show individual folds and error bars indicate standard deviation.

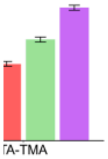


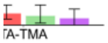
Supplementary Fig. 4 | Performance comparison when using different training s
HEX on independent validation cohorts. Removing either feature distribution
adaptive loss function significantly degraded performance across two independent coh
TMA, n = 264 cores; TA-TMA, n = 108 cores), reinforcing the critical role of these c
enabling generalization across distinct datasets. Bars represent point estimates a
indicate 95% bootstrap CIs (n = 1,000 resamples).





ıes for HEX on
her backbones,
: 264 cores; TA-
w varying levels
computational
CIs (n = 1,000



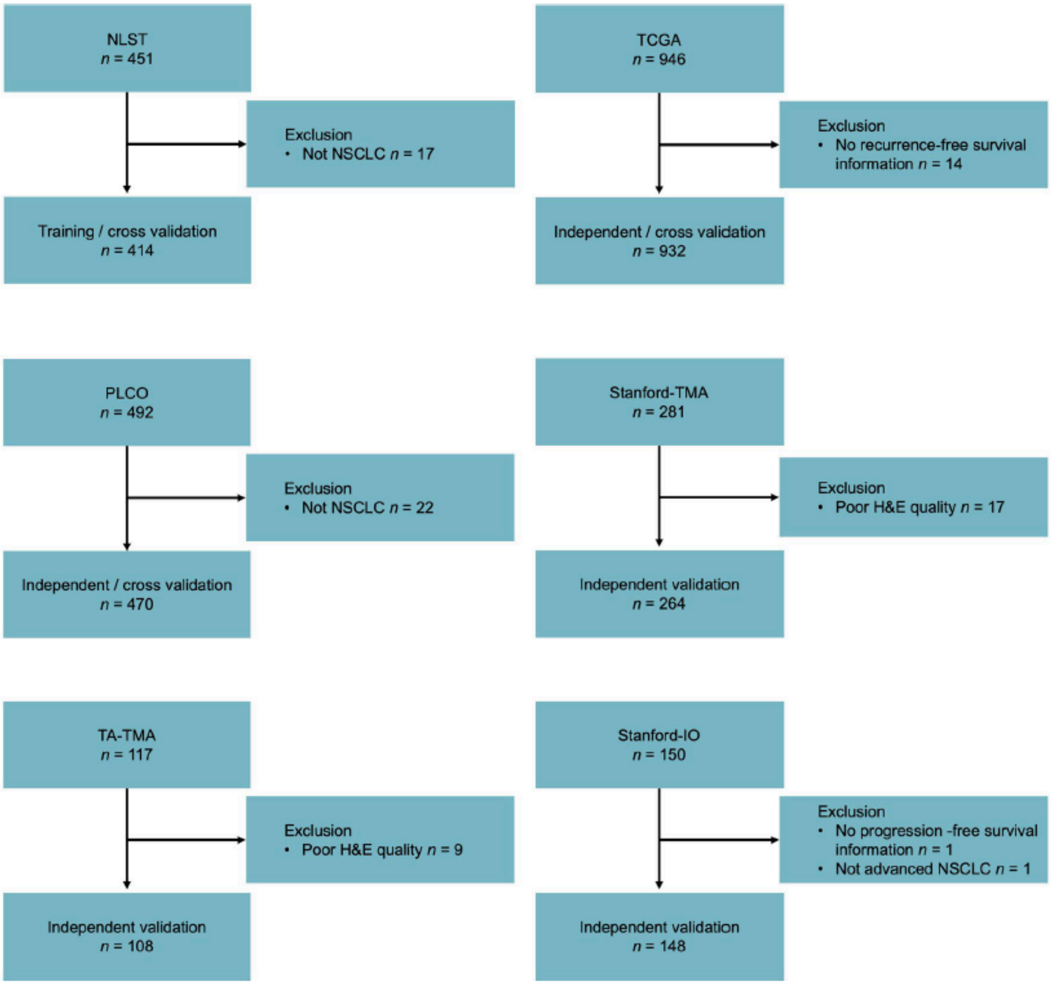


strategies for
smoothing or
orts (Stanford-
components in
and error bars

Independent validation perf

We assessed the model **accuracy** for protein expression on two independent datasets with CODEX and H&E-stained tissue sections for 372 tumor samples.

Two independent tissue microarray (TMA) cohorts—Stanford-TMA and tissue array TMA (TA-TMA)



NLST
n = 451

Exclusion
- Not NSCLC n = 17

Training / cross validation
n = 414

TCGA
n = 946

Exclusion
- No recurrence-free survival information n = 14

Independent / cross validation
n = 932

PLCO
n = 492

Exclusion
- Not NSCLC n = 22

Independent / cross validation
n = 470

Stanford-TMA
n = 281

Exclusion
- Poor H&E quality n = 17

Independent validation
n = 264

TA-TMA
n = 117

Exclusion
- Poor H&E quality n = 9

Independent validation
n = 108

Stanford-IO
n = 150

Exclusion
- No progression -free survival information n = 1
- Not advanced NSCLC n = 1

Independent validation
n = 148

Extended Data Fig. 2 | Sample inclusion and exclusion across six cohorts. Flow chart depicting patient/sample inclusion and exclusion for the six study coho NLST, TCGA, PLCO, Stanford-TMA, TissueArray TMA (TA-TMA), and the Stanford immuno-oncology (Stanford-IO) cohort.

formance

Generalizability, we externally validated HEX on a pan-cancer dataset containing 57-plex CODEX and matched H&E images across 206 tumor samples from 34 tissue types

To evaluate HEX's adaptability to new tissue types and expanded biomarker panels, we conducted both retraining and fine-tuning experiments on 140 colorectal cancer (CRC) cores from the Bern dataset, using all 57 protein markers—33 of which were not present in the original NSCLC panel. The same data split was used for both experiments: 84 cores from 21 patients for training and 56 cores from 14 patients for testing.

rts–