Prediction of KRAS mutation status from H&E foundation model embeddings in non-small cell lung cancer

Marc Robbins* ROBBINSMARC@VERILY.COM
Verily Life Sciences, San Bruno USA
Jessica Loo* JESSICALOO@VERILY.COM
Verily Life Sciences, San Bruno USA
Saurabh Vyawahare SAURABHV@VERILY.COM
Verily Life Sciences, San Bruno USA
Yang Wang WAYANG@VERILY.COM
Verily Life Sciences, San Bruno USA
Carson Mcneil CMCNEIL@VERILY.COM
Verily Life Sciences, San Bruno USA
Dave Steiner DAVESTEINER@GOOGLE.COM
Google Research, Mountain View USA
Sudha Rao SUDHARAOK@GMAIL.COM
Seton Medical Center, Daly City USA
Pok Fai Wong
Ehud Rivlin EHUD@VERILY.COM
Verily Life Sciences, San Bruno USA
Shamira Weaver SHAMIRASW@VERILY.COM
Verily Life Sciences, San Bruno USA
Roman Goldenberg RGOLDENBERG@VERILY.COM
Verily Life Sciences, San Bruno USA

Editor:

Abstract

We predicted KRAS mutation status on non-small cell lung cancer (NSCLC) H&E images from foundation model embeddings. We evaluated a variety of attention-based multiple instance learning (MIL) models and aggregation strategies for a tilewise linear classifier. MIL with self-attention performed the best (AUC=0.822) followed by the minimum over tiles classified with the linear model (AUC=0.810). Self-attention was necessary for MIL to surpass tilewise linear classification when a wide range of aggregation techniques was considered.

Keywords: KRAS, NSCLC, multiple instance learning, foundation model, adenocarcinoma, squamous cell carcinoma, histopathology, mutation

1 Introduction

Genomic testing has proven vital for informing prognosis and treatment strategy for a wide range of cancers Planchard et al. (2018). However, these tests are expensive, have high

*. These authors contributed equally to this work

turnaround time, and exhaust scarce tissue Rusch et al. (2018). Since genetic alterations correlate with phenotypic signatures in tumor cells and their microenvironments, recent work has sought to predict genetic signatures directly from routinely stained H&E slides Coudray et al. (2018); Chen et al. (2020); Kather et al. (2020). In addition to aiding in clinical application, relationships between histology and genomics could also help elucidate novel biological mechanisms and biomarkers.

Lung cancer is the most common cancer globally Bray et al. (2024). Approximately 85% of lung cancers are classified as non-small cell lung cancer (NSCLC), of which lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) are the two predominant subtypes Ganti et al. (2021). Identification of genetic mutations in NSCLC - in particular KRAS, EGFR, and ALK - are vital for treatment decisions and enrollment in clinical trials Planchard et al. (2018). Kristen rat sarcoma (KRAS) is the most commonly mutated isoform of rat sarcoma (RAS), the most frequently mutated oncogene in human cancer Reck et al. (2021). While KRAS normally acts as a switch for activating downstream pathways, point mutations can cause dysregulation due to a constitutively active GTP-bound state Reck et al. (2021).

Recent work has predicted these and other common LUAD genetic mutations from H&E images with varying degrees of success Kather et al. (2020); Coudray et al. (2018). For example, in one study, an AUC of 0.733 was achieved for KRAS classification in LUAD Coudray et al. (2018).

Recent therapies have proven effective for treating NSCLC with p.G12C KRAS mutation Skoulidis et al. (2021); Jänne et al. (2022). While KRAS mutations are much more common in LUAD (prevalence 30%) than LUSC (prevalence 1-7%) Skoulidis and Heymach (2019); Acker et al. (2021), a single model identifying KRAS mutations in NSCLC H&E images could help the entire eligible patient population.

Due to the large size of histopathology slides, images must first be partitioned into thousands of tiles from which information can later be recombined Chen et al. (2022). One approach assigns slidewise labels to all tiles to train a model that classifies them independently. An aggregation strategy is then required to obtain a slidewise score from tilewise scores Bilal et al. (2023). In another approach known as multiple instance learning (MIL), labels are applied to an entire bag of tiles simultaneously. An attention mechanism learns to identify and weigh the relative importance of tiles, and additional mechanisms such as self-attention can also learn to capture relationships between tiles Ilse et al. (2018); Rymarczyk et al. (2021). MIL models learn to output a single slidewise score from each bag, so an independent aggregation strategy is not required during inference.

Many previous classifiers process H&E tiles directly Coudray et al. (2018); Kather et al. (2020); Chen et al. (2020). Recently, embeddings derived from self-supervised foundation models have proven an effective starting point for a wide range of downstream pathology tasks, such as tumor grading and region of interest (ROI) retrieval Vorontsov et al. (2023); Chen et al. (2024); Lai et al. (2023). Such techniques involve pretraining on a large corpus of unlabeled data and enable rapid finetuning to a range of downstream tasks on small amounts of labeled data. The low-dimensionality of such embeddings enables direct comparison between simple techniques such as linear regression and more sophisticated attention-based MIL approaches.

	Train	Validation	Test
Mutation Status			
KRAS+	65	47	27
KRAS-	402	187	214
Cancer Subtype			
LUAD	226	119	118
LUSC	241	115	123

Table 1: Dataset composition by partition

	LUAD	LUSC
Mutation Status		
KRAS+	131	8
KRAS-	332	471

Table 2: Dataset composition by cancer subtype

2 Materials and Methods

Our dataset consisted of 942 NSCLC cases from The Cancer Genome Atlas (TCGA, https://www.cancer.gov/tcga) with 1,038 corresponding H&E stained tissue slide images and KRAS mutation status, whereby KRAS+ indicates any KRAS mutation and KRAS- indicates wild-type. The cases were randomly split into train, validation, and test partitions, with compositions described in Tables 1 and 2. From each case, 2,048 tiles of size 224x224 were randomly sampled at 20X magnification from all tissue regions. If a case contained multiple slides, tiles were sampled across all slides.

Each tile was converted to a 384 dimensional embedding using a foundation model. The foundation model was trained on 6,249 H&E whole slide images (WSIs) from a wide range of TCGA studies through the masked siamese networks self-supervised approach, and has been previously published Lai et al. (2023). The train, validation, and test partitions of our dataset follow the dataset partitions used for the development of the foundation model to avoid data leakage.

We trained a variety of models to predict KRAS mutation status from these embeddings. We trained a tilewise logistic regression model with three aggregation strategies:

1. Tilewise mean
2. Percentage of tiles above dataset quantile, q_d
3. Tilewise score at slidewise quantile, q_s

The tilewise mean and percentage of positively classified tiles have previously been used for prediction of mutation status in LUAD Coudray et al. (2018). We extend this approach by exploring a range of values for q_d, as well as q_s; we trade off targeting a small subset of especially high and low scoring tiles hypothesized to contain the most signal, with a larger subset of more representative tiles. Values for q_s were evaluated at 10% increments,

including the 0th and 100th quantiles for evaluation of the minimum and maximum tile scores, respectively. Values for q_d were obtained by considering values at 10% increments of all tiles in the validation set.

We also trained three variations of an attention-based MIL model Ilse et al. (2018):

1. Baseline Ilse et al. (2018)
2. Self-attention Rymarczyk et al. (2021)
3. Additive MIL Javed et al. (2022)

The baseline attention-based MIL model consists of a single gated attention mechanism followed by a fully-connected output Ilse et al. (2018). The self-attention MIL introduces a self-attention layer before the gated attention mechanism to capture dependencies between tiles Rymarczyk et al. (2021). Additive MIL applies the final fully-connected layer to tilewise features before aggregation, rather than after Javed et al. (2022), such that the additive effect of each tile on the final slidewise score can be observed directly.

All MIL models were trained on bags of 2,048 patches. A dropout layer set attention weights to 0 with frequency 20%. Models were trained with the Adam optimizer Kingma and Ba (2014) (β=0.9, β_2=0.99, ϵ=1e-7) and an exponentially decaying learning rate with linear warmup (α=1e-6 with 5000 warmup steps, decay rate = 0.9, and 10,000 steps per decay). An L2 regularization was used (λ=1e-4). The best epoch was selected by performance on the validation set. Confidence intervals (95%) were obtained by bootstrapping over 1000 samplings of tiles from the test set.

3 Experiments and Results

	AUC (95% CI)
Linear Model Aggregation Techniques	
Tilewise mean	0.767 (0.672, 0.855)
Percentage of tiles above dataset median ($q_d = 0.5$)	0.753 (0.651, 0.836)
Minimum tile score ($q_s = 0$)	0.810 (0.727, 0.879)
MIL Model	
Baseline	0.753 (0.663, 0.837)
Additive MIL	0.776 (0.679, 0.849)
Self-attention	**0.822 (0.743, 0.884)**

Table 3: Results for a selection of aggregation techniques for the linear model and MIL models.

The self-attention MIL model performed best overall, predicting KRAS mutation status with AUC=0.822 (0.743, 0.884) (Table 3). The baseline MIL model achieved AUC=0.753 (0.663, 0.837) and the additive MIL model achieved AUC=0.776 (0.679, 0.849).

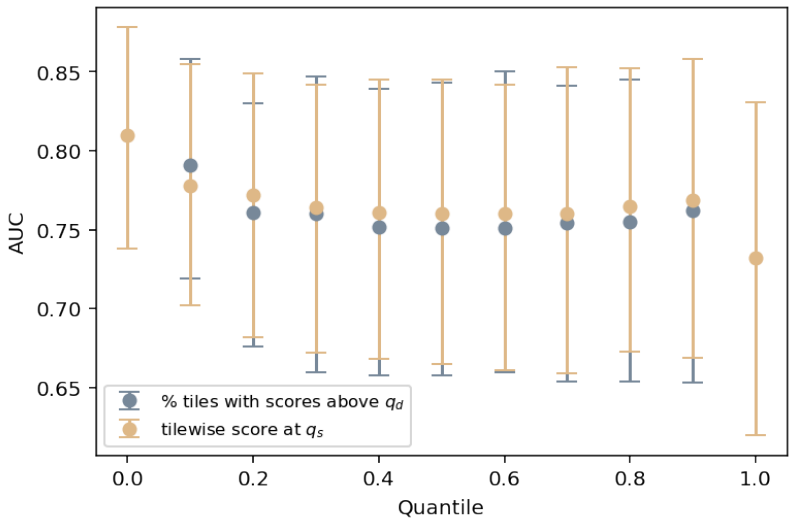

Figure 1: Comparison of aggregation strategies for obtaining slidewise from tilewise KRAS mutation status predictions

The tilewise linear model predicted KRAS mutation status best when aggregating by the minimum tilewise score, with AUC=0.810 (0.727, 0.879) (Figure 1). Aggregating by the tilewise mean achieved AUC=0.767 (0.727, 0.879). The worst predictive power was achieved via the maximum score, with AUC=0.732 (0.629,0.825). All other aggregation strategies resulted in AUCs between 0.750 (0.652, 0.842) and 0.789 (0.713, 0.853). In general, metrics sensitive to particularly high and low tile scores outperformed those that considered larger numbers of tiles.

We qualitatively examined the highest attention tiles for the self-attention MIL model and highest and lowest scored tiles for the tilewise linear model. Figure 2 shows some representative examples on KRAS+ and KRAS- slides. On both KRAS+ and KRAS- slides, the highest attention patches exhibited a mixture of tumor and benign tissue, including alveolar-like epithelium and macrophages.

The most positively scored tiles on both KRAS+ and KRAS- slides frequently exhibited glandular and mucinous patterns of LUAD as well as tumor-stromal interfaces (TSIs) and tumor-infiltrating lymphocytes (TILs). The most negatively scoring tiles on KRAS- slides (true negatives) consistently depicted flattened, poorly-differentiated tumor cells and intercellular bridges characteristic of LUSC. The most negatively scoring tiles on KRAS+ slides (false negatives) were more varied.

We further examined the spatial distribution of attention scores and tilewise predictions for characteristic KRAS+ and KRAS- slides in Figure 3. Non-tumor regions had the lowest attention scores and the most variation in tilewise prediction. Tumor regions had fairly uniform patterns of attention scores and tilewise predictions, though with some slides experiencing significantly more variation. In Figure 3, for example, the KRAS+ tumor received a broad range of attention scores and prediction values, while the KRAS- tumor received consistently high attention scores and low prediction values.

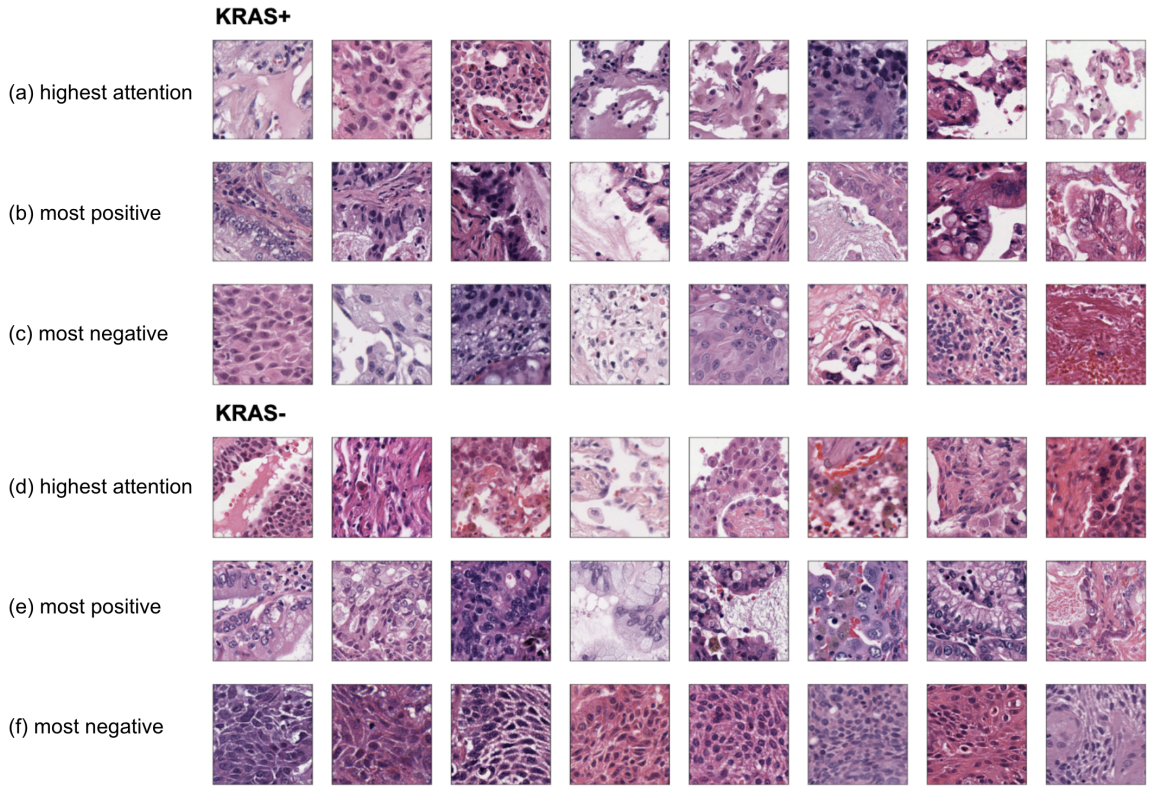

Figure 2: Representative examples of the highest attention tiles of the self-attention MIL model and most confidently predicted KRAS+ and KRAS- tiles of the linear model.

4 Discussion and Conclusion

We trained a variety of models to predict KRAS mutation status from NSCLC H&E embeddings, further validating the use of foundation model embeddings for downstream tasks.

MIL techniques did not automatically outperform simpler tilewise classification. Previous work has found tilewise classification to exceed more sophisticated MIL pipelines, but with feature extractors trained on different datasets Ghaffari Laleh et al. (2022). A wide range of aggregation techniques was needed to more accurately identify the advantages of MIL training, as was an additional self-attention mechanism. Baseline attention-based MIL architectures may have difficulty learning aggregations that focus on very few tiles, such as minimum or maximum score.

Surprisingly, the highest attention patches did not exhibit features clearly associated with KRAS or cancer subtype. Analysis of minimum-scoring tiles, however, revealed defining characteristics of LUSC, in which KRAS mutations are rare. Analysis of positively predicted patches revealed characteristics specific to LUAD, including glandular and mucinous patterns, as well as TILs and TSIs. Mucinous patterns and TILs have been associated with KRAS mutation in LUAD Kadota et al. (2014); Rekhtman et al. (2013).

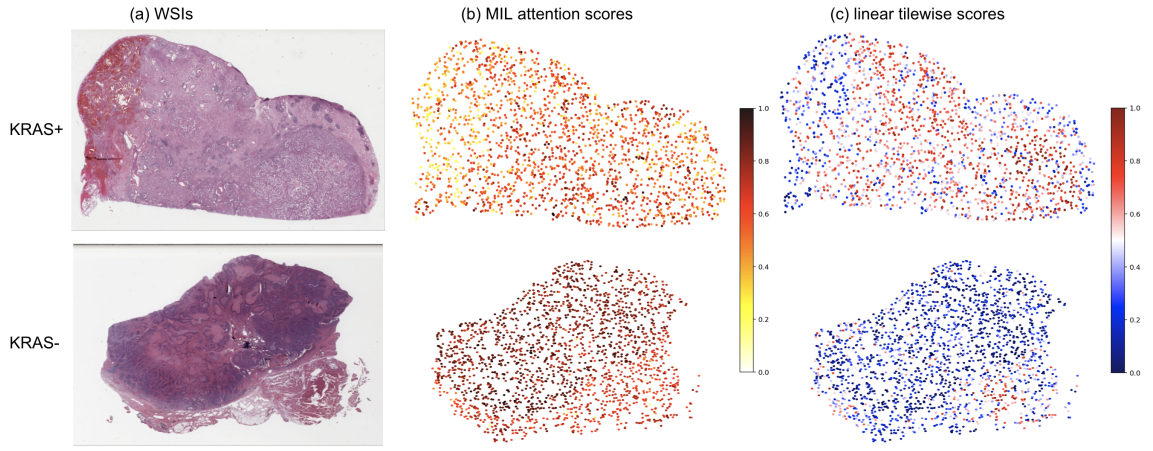

Figure 3: Spatial distribution of self-attention MIL attention scores and tilewise linear model predictions on characteristic KRAS+ and KRAS- slides. Attention scores have been normalized per slide.

Spatial analysis of the models revealed relatively homogenous prediction patterns within individual tumors, though with significant variation between cases. More work is needed to understand tumor characteristics that result in more or less spatial consistency.

Future work would aim to reproduce these findings on an external test set. Analysis and training of models on LUAD and LUSC subgroups separately would help differentiate signals of KRAS mutation status from cancer subtype, though more KRAS mutant LUSC cases are needed. Reproducing the analysis across a wide range of mutations would differentiate general findings from those specific to KRAS mutation.

Acknowledgments and Disclosure of Funding

This work was supported by Verily Life Sciences LLC. All authors at Verily Life Sciences LLC have equity interests.

References

Fabian Acker, Jan Stratmann, Lukas Aspacher, Ngoc Thien Thu Nguyen, Sebastian Wagner, Hubert Serve, Peter J Wild, and Martin Sebastian. KRAS mutations in squamous cell carcinomas of the lung. *Front. Oncol.*, 11:788084, December 2021.

Mohsin Bilal, Robert Jewsbury, Ruoyu Wang, Hammam M AlGhamdi, Amina Asif, Mark Eastwood, and Nasir Rajpoot. An aggregation of aggregation methods in computational pathology. *Med. Image Anal.*, 88:102885, August 2023.

Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.*, 74(3):229–263, April 2024.

Mingyu Chen, Bin Zhang, Win Topatana, Jiasheng Cao, Hepan Zhu, Sarun Juengpanich, Qijiang Mao, Hong Yu, and Xiujun Cai. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. *NPJ Precis Oncol*, 4:14, June 2020.

Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical Self-Supervised learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16134. IEEE, June 2022.

Richard J Chen, Tong Ding, Ming Y Lu, Drew F K Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L Weishaupt, Judy J Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nat. Med.*, 30(3):850–862, March 2024.

Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.*, 24(10):1559–1567, October 2018.

Apar Kishor Ganti, Alyssa B Klein, Ion Cotarla, Brian Seal, and Engels Chou. Update of incidence, prevalence, survival, and initial treatment in patients with Non-Small cell lung cancer in the US. *JAMA Oncol*, 7(12):1824–1832, December 2021.

Narmin Ghaffari Laleh, Hannah Sophie Muti, Chiara Maria Lavinia Loeffler, Amelie Echle, Oliver Lester Saldanha, Faisal Mahmood, Ming Y Lu, Christian Trautwein, Rupert Langer, Bastian Dislich, Roman D Buelow, Heike Irmgard Grabsch, Hermann Brenner, Jenny Chang-Claude, Elizabeth Alwers, Titus J Brinker, Firas Khader, Daniel Truhn, Nadine T Gaisa, Peter Boor, Michael Hoffmeister, Volkmar Schulz, and Jakob Nikolas Kather. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. *Med. Image Anal.*, 79:102474, July 2022.

Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. February 2018.

Pasi A Jänne, Gregory J Riely, Shirish M Gadgeel, Rebecca S Heist, Sai-Hong I Ou, Jose M Pacheco, Melissa L Johnson, Joshua K Sabari, Konstantinos Leventakos, Edwin Yau, Lyudmila Bazhenova, Marcelo V Negrao, Nathan A Pennell, Jun Zhang, Kenna Anderes, Hirak Der-Torossian, Thian Kheoh, Karen Velastegui, Xiaohong Yan, James G Christensen, Richard C Chao, and Alexander I Spira. Adagrasib in Non-Small-Cell lung cancer harboring a KRASG12C mutation. *N. Engl. J. Med.*, 387(2):120–131, July 2022.

Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive MIL: Intrinsically interpretable multiple instance learning for pathology. June 2022.

Kyuichi Kadota, Yi-Chen Yeh, Sandra P D'Angelo, Andre L Moreira, Deborah Kuk, Camelia S Sima, Gregory J Riely, Maria E Arcila, Mark G Kris, Valerie W Rusch, Prasad S Adusumilli, and William D Travis. Associations between mutations and histologic patterns of mucin in lung adenocarcinoma: invasive mucinous pattern and extracellular mucin are associated with KRAS mutation. *Am. J. Surg. Pathol.*, 38(8):1118–1127, August 2014.

Jakob Nikolas Kather, Lara R Heij, Heike I Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M Niehues, Kai A J Sommer, Peter Bankhead, Loes F S Kooreman, Jefree J Schulte, Nicole A Cipriani, Roman D Buelow, Peter Boor, Nadi-Na Ortiz-Brüchle, Andrew M Hanby, Valerie Speirs, Sara Kochanny, Akash Patnaik, Andrew Srisuwananukorn, Hermann Brenner, Michael Hoffmeister, Piet A van den Brandt, Dirk Jäger, Christian Trautwein, Alexander T Pearson, and Tom Luedde. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Cancer*, 1 (8):789–799, August 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014.

Jeremy Lai, Faruk Ahmed, Supriya Vijay, Tiam Jaroensri, Jessica Loo, Saurabh Vyawahare, Saloni Agarwal, Fayaz Jamil, Yossi Matias, Greg S Corrado, Dale R Webster, Jonathan Krause, Yun Liu, Po-Hsuan Cameron Chen, Ellery Wulczyn, and David F Steiner. Domain-specific optimization and diverse evaluation of self-supervised models for histopathology. October 2023.

D Planchard, S Popat, K Kerr, S Novello, E F Smit, C Faivre-Finn, T S Mok, M Reck, P E Van Schil, M D Hellmann, S Peters, and ESMO Guidelines Committee. Metastatic non-small cell lung cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.*, 29(Suppl 4):iv192–iv237, October 2018.

M Reck, D P Carbone, M Garassino, and F Barlesi. Targeting KRAS in non-small-cell lung cancer: recent progress and new approaches. *Ann. Oncol.*, 32(9):1101–1110, September 2021.

Natasha Rekhtman, Daphne C Ang, Gregory J Riely, Marc Ladanyi, and Andre L Moreira. KRAS mutations are associated with solid growth pattern and tumor-infiltrating leukocytes in lung adenocarcinoma. *Mod. Pathol.*, 26(10):1307–1319, October 2013.

Michael Rusch, Joy Nakitandwe, Sheila Shurtleff, Scott Newman, Zhaojie Zhang, Michael N Edmonson, Matthew Parker, Yuannian Jiao, Xiaotu Ma, Yanling Liu, Jiali Gu, Michael F Walsh, Jared Becksfort, Andrew Thrasher, Yongjin Li, James McMurry, Erin Hedlund, Aman Patel, John Easton, Donald Yergeau, Bhavin Vadodaria, Ruth G Tatevossian, Susana Raimondi, Dale Hedges, Xiang Chen, Kohei Hagiwara, Rose McGee, Giles W Robinson, Jeffery M Klco, Tanja A Gruber, David W Ellison, James R Downing, and Jinghui Zhang. Clinical cancer genomic profiling by three-platform sequencing of whole genome, whole exome and transcriptome. *Nat. Commun.*, 9(1):3962, September 2018.

Dawid Rymarczyk, Adriana Borowa, Jacek Tabor, and Bartosz Zieliński. Kernel Self-Attention for weakly-supervised image classification using deep multiple instance learning. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1720–1729. IEEE, January 2021.

Ferdinandos Skoulidis and John V Heymach. Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy. *Nat. Rev. Cancer*, 19(9):495–509, September 2019.

Ferdinandos Skoulidis, Bob T Li, Grace K Dy, Timothy J Price, Gerald S Falchook, Jürgen Wolf, Antoine Italiano, Martin Schuler, Hossein Borghaei, Fabrice Barlesi, Terufumi Kato, Alessandra Curioni-Fontecedro, Adrian Sacher, Alexander Spira, Suresh S Ramalingam, Toshiaki Takahashi, Benjamin Besse, Abraham Anderson, Agnes Ang, Qui Tran, Omar Mather, Haby Henary, Gataree Ngarmchamnanrith, Gregory Friberg, Vamsidhar Velcheti, and Ramaswamy Govindan. Sotorasib for lung cancers with KRAS p.G12C mutation. *N. Engl. J. Med.*, 384(25):2371–2381, June 2021.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D Kunz, Matthew C H Lee, Jan Bernhard, Ran A Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Juan Retamero, William A Moye, Razik Yousfi, Christopher Kanan, David Klimstra, Brandon Rothrock, and Thomas J Fuchs. Virchow: A Million-Slide digital pathology foundation model. September 2023.