AstraKit: Customizable, reproducible workflows for biomedical research and precision medicine

Nadine S. Kurz[1,2,*], Kevin Kornrumpf[1], Martin Stoves[1], and Jürgen Dönitz[1,2,3]

[1]Department of Medical Bioinformatics, University Medical Center Göttingen, 37075 Göttingen, Niedersachsen, Germany
[2]Comprehensive Cancer Center Göttingen (G-CCC), 37075 Göttingen, Germany
[3]Campus-Institut Data Science (CIDAS), Section Medical Data Science (MeDaS), 37075 Göttingen, Niedersachsen, Germany

Abstract

Motivation: The success of precision medicine and biomedical research depends on the availability of efficient software solutions for processing and interpreting genetic variants, interpreting multi-omics data, and integrating drug screen analyses. However, fragmented bioinformatics tools compel researchers and clinicians to resort to error-prone manual pipelines. **Results:** We present AstraKit, a unified KNIME workflow suite enabling end-to-end precision medicine analytics. AstraKit introduces three transformative innovations: 1) Dynamic variant interpretation with customizable annotation and filtering for disease-specific genomic contexts; 2) Multi-layered omics analyses integrating genomic, transcriptomic, and epigenetic data; and 3) Translational drug matching that correlates in vitro drug screens with clinical outcomes. Validated across oncology cohorts, AstraKit demonstrates concordance between experimental drug sensitivity and clinical outcomes, resolving discordances to uncover resistance mechanisms. By unifying variant analysis, multi-omics, and drug response modeling on a single customizable platform, AstraKit eliminates siloed workflows, accelerating biomarker validation and enabling clinicians to directly link molecular profiles to therapeutic decisions. As all AstraKit workflows are open-source and platform-independent, we provide a versatile comprehensive software suite for a multitude of tasks in bioinformatics and precision medicine.

Availability and implementation: The KNIME workflows are available at KNIME Hub https://hub.knime.com/bioinf_goe/spaces/Public/AstraKit~lfVsGBY2HnPYc1h1/. The source code is available at https://gitlab.gwdg.de/MedBioinf/mtb/astrakit.

Keywords: Variant interpretation, Multi-omics integration, Drug screen analysis, Precision medicine, Automated workflows

1 Introduction

Precision medicine demands robust computational frameworks to transform heterogeneous molecular data into clinically actionable insights. While next-generation sequencing (NGS) and high-throughput drug screening have exponentially expanded our capacity to generate genomic, transcriptomic, epigenomic, and functional drug response data, the lack of integrated, end-to-end analytical pipelines remains a critical barrier to reproducible and clinically translatable research. Software packages typically operate in silos: variant calling (GATK [1]), genomic operations (samtools [2], bedtools [3]), annotation (ANNOVAR [4], VEP [5], Onkopus [6]), and filtering (SNPSift [7], AdaGenes [8]) address isolated steps in the analytical continuum. Consequently, researchers must manually chain tools across incompatible data formats (VCF [9], MAF, BED), perform auxiliary tasks like genome assembly coordinate conversion (LiftOver), and implement custom filtering criteria—compromising traceability, reproducibility, and accessibility for non-programmers. This fragmentation impedes two critical translational needs:

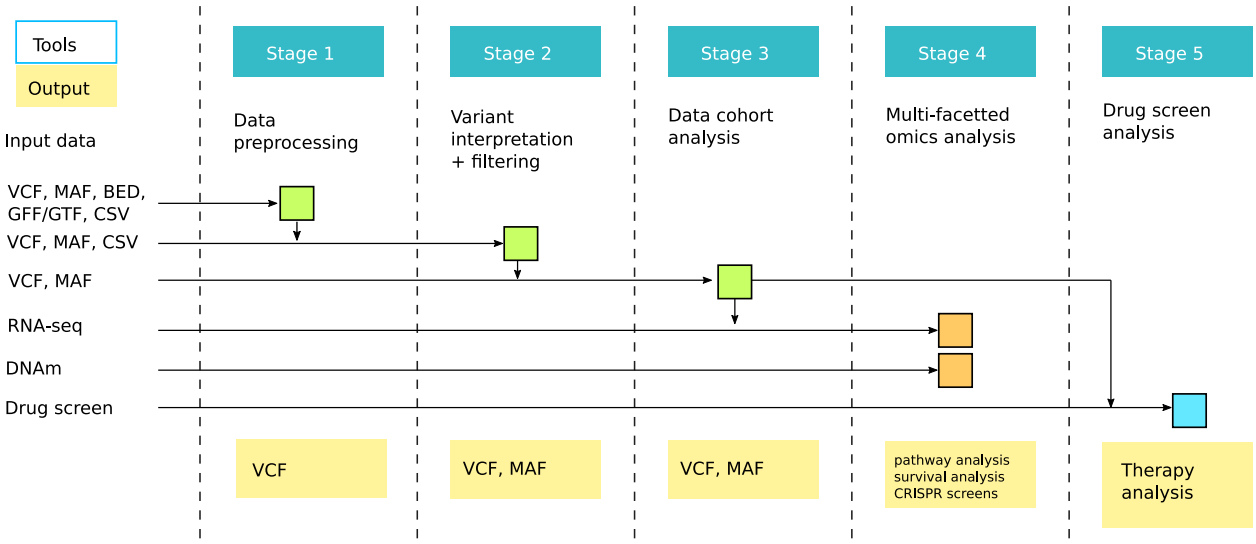


Figure 1: Overview of the AstraKit scope. AstraKit is able to preprocess mutational data in various formats, including VCF, MAD, BED, GFF/GTF, and CSV, for file format conversions and genome assembly coordinate conversion (LiftOver). Variant interpretation workflows support modular annotation and filtering using customizable rules. Cohort analysis leverages tools like *maftools* for mutation pattern detection, while multi-omics analyses facilitates joint analysis of genomic, transcriptomic, and epigenomic data. Additionally, drug sensitivity validation is performed through integrated drug screen analysis workflows.

First, robust cohort-level variant analysis, essential for identifying recurrent mutations and clinical associations, often requires ad hoc integration of specialized tools (e.g., *maftools* [10]) with upstream processing, hindering scalable biomarker discovery. The joint analysis of various omics data, including transcriptomics and epigenetics, places further demands on data integration. Second, while clinical interpretation tools (e.g., MTB-Report [11], Pandrugs [12]) suggest targeted therapies from mutational profiles, and drug screen platforms (e.g., Breeze [13]) enable in vitro data exploration, no existing framework systematically links clinical genomic classifications to functional drug response validation.

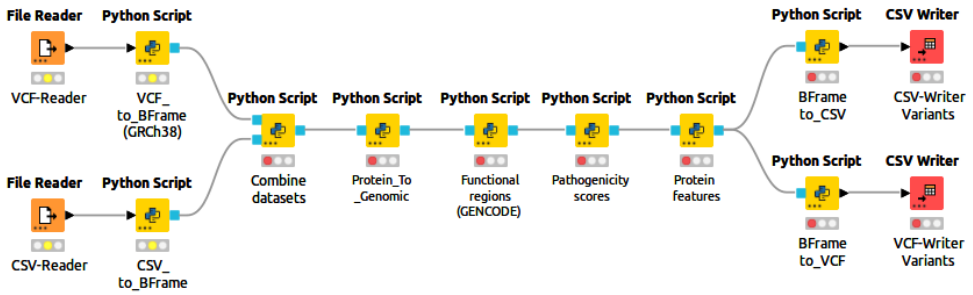
Workflow systems like Snakemake [14] and Nextflow [15] improve reproducibility but require coding expertise, while cloud-based platforms like Galaxy [16] may pose challenges for sensitive clinical data due to privacy regulations. Although the Jupyter environment [17] enhances codes reproducibility, it lacks native workflow orchestration and standardized sharing mechanisms for complex pipelines. KNIME [18] provides a local, GUI-driven environment for workflow assembly. Prior workflow toolkits for precision medicine focused narrowly on variant calling/annotation [19, 20, 21, 22] or reference genome processing [23], without unifying variant interpretation, multi-omics integration, or drug response validation. While bioinformatic workflow managers offer reproducibility, scalability, and portability [24, 25], the potential for dynamically combining heterogeneous omics data with clinical evidence and functional validation—critical for precision oncology—remains largely unrealized.

Here we present AstraKit, a comprehensive suite of modular, scalable, and reproducible workflows supporting a multitude of processing steps in bioinformatics and precision medicine. Unlike existing tools, AstraKit provides integrated workflows spanning from molecular profile processing and analysis to clinical interpretation and experimental validation. It supports seamless, format-agnostic data preprocessing, rule-based variant annotation/filtering, and cohort-level analysis. Critically, AstraKit bridges three previously disconnected domains: Multi-omics integration for joint analysis of mutations, gene expression, and epigenetic data—enabling pathway enrichment, survival modeling, direct linkage of clinical variant interpretations to drug sensitivity profiling, where patient-derived genomic biomarkers are systematically validated against high-throughput drug screen data (e.g., IC_{50}, GR metrics) to identify actionable therapeutic targets. By unifying computational genomics with functional validation in a single framework, AstraKit accelerates the translation of molecular insights into clinically testable hypotheses.

A

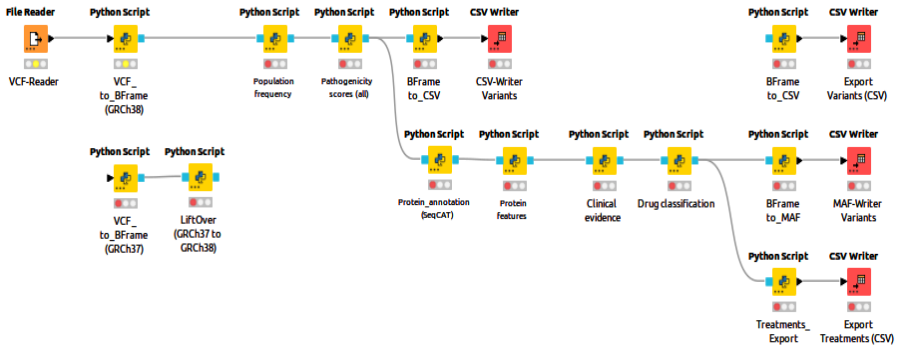
Gene	AA Exchange
TP53	E2K
TP53	E2Q
TP53	E2A

Gene	AA Exchange
KRAS	G12C
BRAF	V600E



B

CHROM	POS	ID	REF	ALT	....
1	46278251	.	G	A	.
1	158627717	.	G	C	.
2	61499712	.	C	T	.



C

Genomic location	Gene symbol	AA exchange	transcript	Protein ID	Uniprot ID	type	1000Genomes_AF	AlphaMissense_Score	CIViC_exact
chr12:25245351C>A	KRAS	G12C	NM_004985.5:c.34G>T	NP_004976.2	P01116	protein_coding	.	0.998	34

D

Associated biomarker	Match type	Disease	Drug class	Drug	Evidence level	Response	Source	Citation ID
KRAS G12C	Exact match	Lung Non-small cell carcinoma	KRAS Inhibitor	Sotorasib	1	Sensitive	CIViC	34096690

Figure 2: Customizable variant interpretation workflows in AstraKit. (**A**) A KNIME variant annotation workflow start with reading in one or multiple data files. Users can then dynamically align variant processing nodes to annotate the variants with selected features. The annotated variants can be written out again in different data formats. (**B**) For clinical interpretation, AstraKit provides the functionality to add clinical evidence and drug classification annotation, and the option to export therapeutic options for a patient. (**C**) AstraKit can export the functional annotations of variants, as well as (**D**) the retrieved evidence-based biomarkers for selected targeted therapies.

2 Materials and Methods

2.1 Datasets and analytical workflows

To test and validate the workflows for data preprocessing and variant interpretation, we used publicly available datasets, including somatic variants in VCF format [26, 27]. For the variant data cohort analysis, we downloaded somatic variant data, gene expression, and DNA methylation files from 32 cancer types in The Cancer Genome Atlas (TCGA) using the TCGAbiolinks package [28]. TP53 tumor variants were downloaded from the TP53 database [29]. Drug sensitivity data was retrieved from cancer cell line data from DepMap [30] and GDSC [31] (release 8.5), which includes IC_{50} and AUC values from drug sensitivity assays. The AstraKit framework was constructed using the KNIME analytics platform (v5.8.0) using Python (v3.11) and R (v4.5.2) extensions.

2.2 Variant annotation and interpretation

Data parsing and genome assembly conversion were implemented using *AdaGenes* [8], which supports VCF, MAF, FASTA, CSV, and plain text formats, as well as LiftOver between GRCh37/GRCh38/T2T-CHM13 assemblies. For variant annotation and interpretation, we developed custom KNIME nodes leveraging the *Onkopus* Python package [6]. Annotated variants were serialized as pickled objects for inter-node transfer. Genetic variants at transcript/protein levels were mapped to genomic coordinates using `SeqCAT` [32]. To standardize cancer type nomenclature across heterogeneous sources, we retrieved targeted therapy evidence from CIViC [33], MetaKB [34], and OncoKB [35]. Cancer entities were mapped to the OncoTree ontology [36] via a two-step process: 1) Exact string matching against OncoTree's canonical cancer type names, and 2) fuzzy matching for unmatched entries, selecting the highest-scoring OncoTree term. The token sort ratio normalizes Levenshtein distance by string length by first tokenizing, sorting, and rejoining words to handle word-order variations:

$$R = \left(1 - \frac{D(S1_{\text{sorted}}, S2_{\text{sorted}})}{\max(\text{len}(S1_{\text{sorted}}), \text{len}(S2_{\text{sorted}}))}\right) \times 100 \tag{1}$$

where $S1sorted$ and $S2sorted$ are token-sorted versions of the input strings. Drug names were standardized to ChEMBL IDs [37] by resolving synonyms via the ChEMBL Onkopus API, ensuring consistent aggregation of drug response data.

2.3 Cohort and multi-omics analyses

Cohort-specific analyses and visualizations were implemented using maftools, scikit-learn [38], and seaborn [39]. Protein structural analyses leveraged the Onkopus Application Programming Interface (API) to derive 3D features. Biochemical properties were computed using *BioPython* [40], and $\Delta\Delta G$ (change in folding free energy) was predicted with *FoldX* (v5.0) [41] via the `BuildModel` command:

$$\Delta\Delta G = \Delta G_{mut} - \Delta G_{wt}$$

2.4 Drug screen analysis

To analyze drug sensitivity in relation to clinical outcomes, IC_{50} values were first converted to molar concentration (M) based on user-specified input units (e.g. nM, μM). IC_{50} values were then transformed to pIC_{50} values:

$$pIC_{50} = -log_{10}(IC_{50})$$

This logarithmic transformation linearizes the concentration-response relationship and stabilizes variance, with higher pIC_{50} values indicating greater drug sensitivity (i.e., lower molar IC_{50}). For sample-wise normalization (to account for inter-sample variability in assay performance), we implemented a robust percentile-based approach within each sample (grouped by sample ID). Normalization was performed directly on the pIC_{50} scale to preserve the linearized relationship:

$$\text{pIC}_{50}\text{-norm} = \frac{\text{pIC}_{50} - P_5}{P_{95} - P_5}$$

where P_5 and P_{95} represent the 5th and 95th percentiles of pIC_{50} values within the sample-specific drug response distribution. This scaling maps the central 90% of values to [0, 1], with values below P_5 assigned 0 (maximal resistance) and those above P_{10} and those above P_{95} assigned 1 (maximal sensitivity). For binary classification, normalized values were categorized using a user-defined threshold, defaulting to 0.5:

$$\text{drug sensitivity}(\text{IC}_{50}) = \begin{cases} 1 & \text{if pIC}_{50}\text{-norm} > 0.5 \\ 0 & \text{if pIC}_{50}\text{-norm} \leq 0.5 \end{cases}$$

This threshold defines sensitivity as responses in the upper half of the normalized scale (i.e., above the midpoint between P_5 and P_{95}), not the raw median. Note that 0.5 corresponds to the scaled midpoint of the central 90% distribution, which may differ from the true median if the distribution is skewed. The predicted in vitro sensitivity was then compared against clinical response data. Clinical outcomes r_i were defined per standard criteria, where:

$$\text{Clinical Response} = \begin{cases} 1 & \text{if } r_i = \text{sensitive} \\ 0 & \text{if } r_i = \text{resistant} \end{cases}$$

To quantify agreement between predicted and clinical responses, we evaluated binary classification performance using receiver operating characteristic (ROC-AUC).

3 Results

3.1 The AstraKit workflow suite

AstraKit delivers a comprehensive suite of modular, automated workflows for a multitude of tasks in precision medicine and biomedical research, including genomic data processing, multi-omics integration, and

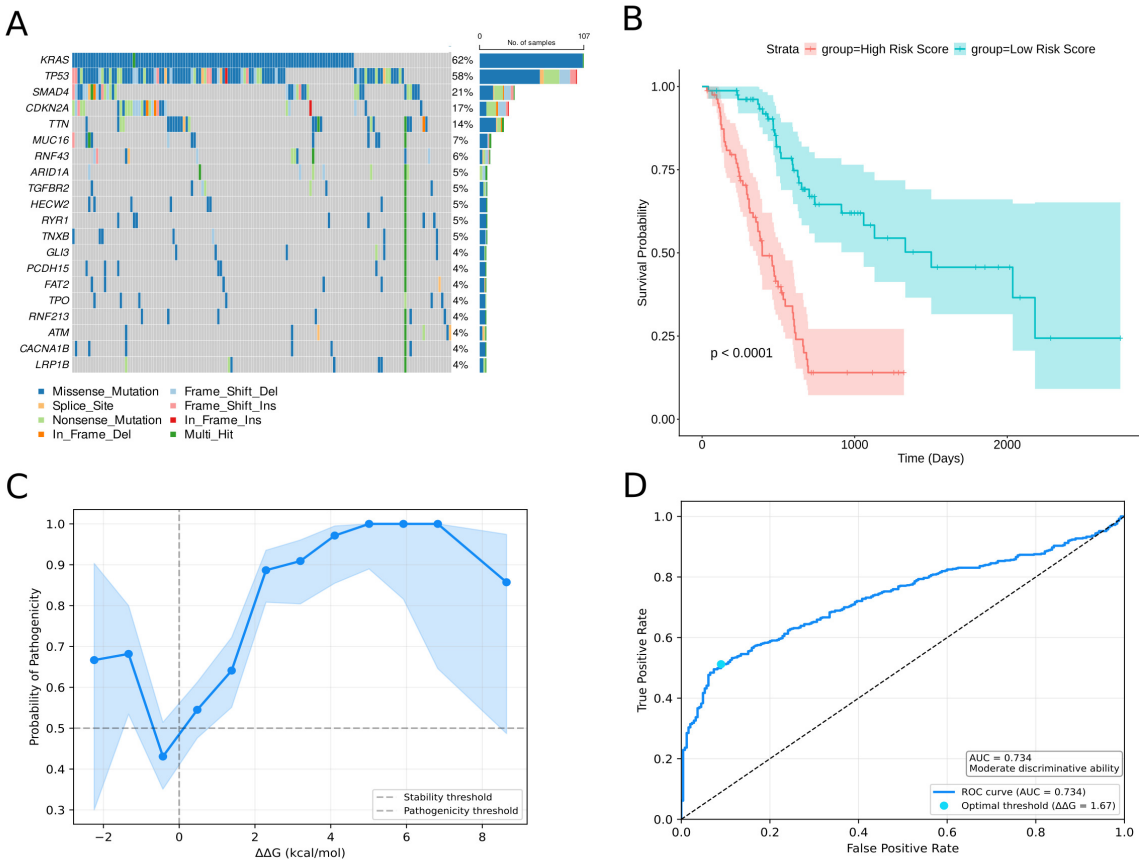


Figure 3: AstraKit variant cohort analysis overview. (**A**) Automated oncoprint analysis of the 20 most recurrently altered genes in the TCGA PAAD (pancreatic adenocarcinoma) cohort, generated directly using AstraKit's variant cohort analysis workflows. (**B**) Kaplan-Meier curve for overall survival in TCGA-PAAD stratified by risk score from Lasso-regularized Cox model (48 MSigDB Hallmark pathways, top 50 mutated genes with median split). (**C**) Correlation analysis between annotated variant features, including computational predictions of protein stability changes (e.g., using FoldX) and AlphaMissense-predicted pathogenicity scores for somatic TP53 variants. (**D**) Receiver operating characteristic (ROC) between pathogenicity prediction and protein stability change, indicating a modest correlation (AUC = 0.73).

drug screen analysis (Figure 1). These pipelines support end-to-end genetic data analysis—from raw sequencing data through variant calling, molecular data preprocessing, annotation, filtering, and clinical interpretation. They further facilitate robust cohort-level analyses and adaptive integration of heterogeneous omics data. Critically, AstraKit introduces innovative tools for preclinical model assessment and clinical trial evaluation by integrating molecular data with experimental drug screen results. Implemented in KNIME, AstraKit provides a graphical user interface that allows researchers to customize, modify, and execute workflows without coding expertise. Built-in workflow saving ensures reproducibility and seamless sharing of research results.

3.2 Customizable mutational profile processing

AstraKit's variant interpretation framework empowers users to construct annotation and filtering pipelines through a unique modular architecture. Each processing step (e.g., LiftOver, annotation, filtering) is encapsulated in standalone, interoperable nodes. This modularity is enabled by a standardized branched JSON data structure that serves as the universal input/output format across all nodes. By enforcing consistent data handling, nodes can be freely combined without manual data reformatting to generate reproducible, context-specific workflows. The suite provides extensive adaptable nodes for annotation (e.g., pathogenicity predictors, pro-

tein features, clinical evidence), filtering, LiftOver, and protein-to-genome conversion (resolving coordinate system mismatches between protein and genomic variants). Workflows dynamically begin with format-specific readers (VCF, MAF, CSV) and conclude with export nodes, supporting flexible pipeline design. For instance, a workflow can conduct a functional annotation of variant cohorts (Figure 2A), where protein-level variants are mapped to genomic coordinates, merged, and annotated with functional context—including transcript isoforms, protein surface accessibility, alpha-carbon distances, and secondary structures—to reveal mutation-driven mechanistic insights into disease pathogenesis. For clinical interpretation, a workflow (Figure 2B) may begin by annotating with population allele frequencies (gnomAD), then filtering common variants before layering pathogenicity predictions (e.g., SIFT, PolyPhen), protein structural data, and ClinVar classifications. Benign/non-coding variants are iteratively excluded, enabling efficient prioritization of clinically actionable alterations. Final outputs include annotated VCFs, column-wise variant reports (Figure 2C), and evidence-based therapy recommendations ranked by clinical actionability (Figure 2D), facilitating seamless transition from analysis to clinical decision support.

3.3 Cohort and multi-faceted omics analysis

To establish the analytical robustness of AstraKit for somatic mutation profiling, workflows for analyzing variants and mutated genes, mutual exclusivity, and affected pathways enable effective analysis of large patient cohorts (Figure 3A). These analyses identify significantly altered genes, co-occurring mutations, and dysregulated biological pathways. Beyond core genomic analysis, AstraKit integrates multi-omics data to enable functional validation of variant impact and therapeutic target prioritization. For instance, users can stratify cohorts by mutation status (e.g., *KRAS* amplification) to compare transcriptomic or proteomic profiles between mutant and wild-type groups. Critically, AstraKit's survival modeling workflow executes Lasso-penalized Cox regression using user-specified inputs (e.g., 48 MSigDB Hallmark pathway scores and binary mutation status of the top 50 altered genes), generating risk scores that automatically stratify cohorts for Kaplan-Meier survival analysis (Figure 3B). This enables identification of mutation-pathway signatures prognostic of clinical outcomes while controlling overfitting through cross-validated regularization.

Critically, AstraKit links variant annotation with downstream feature analysis. For instance, protein stability changes (predicted by FoldX) can be systematically correlated with pathogenicity scores, like AlphaMissense. When analyzing TP53 variants, we observed elevated pathogenicity probabilities for both stabilizing (negative $\Delta\Delta G$) and destabilizing (positive $\Delta\Delta G$) mutations, suggesting stability-independent pathogenic mechanisms (Figure 3C). A receiver operating characteristic (ROC) analysis using $\Delta\Delta G$ to predict pathogenicity yielded an area under the curve (AUC) of 0.73, indicating a fair correlation between protein stability perturbation and pathogenicity (Figure 3D).

3.4 Drug screen validation with clinical outcome

Complementary to molecular profile analysis, AstraKit provides streamlined workflows to analyze targeted therapies for patient cohorts, identifying the most sensitive/resistant drug classes (Figure 4A), drugs (Figure 4B), biomarkers (Figure 4C), and tumor types (Figure 4D). A key innovation is the correlation of preclinical drug screen results with clinical evidence using matched molecular profiles. The associated workflow loads genetic profiles (VCF) and drug screen data (e.g., IC_{50} or GR_{max}), with user-adjustable parameters for normalization, cancer type filtering, and clinical evidence sources. To validate this approach, we analyzed GDSC drug screens alongside the matching DepMap mutational profiles of the employed cell lines. We generated cell-line-wise VCF files and mapped their mutational profiles to therapeutic options using Onkopus using exact match searches. This integration reveals drugs with high experimental efficacy (e.g., elevated pIC_{50}) but limited clinical exploration (fewer clinical records) (Figure 4E). Conversely, comparing pIC_{50} values with clinical outcomes per cell line highlights models with strong translational potential (Figure 4F). Interactive sunburst visualizations (Figure 5A) enable multi-scale exploration of therapeutic options by drug class, biomarker, and tumor type. For TCGA PAAD patients, ROC analysis of cell line-derived predictions versus clinical outcomes showed strong concordance for specific cell lines (COSMIC ID 753595: AUC 0.89; 907285: AUC 0.88) but significant variability across other lines, underscoring the need for context-specific model selection (Figure 5B).

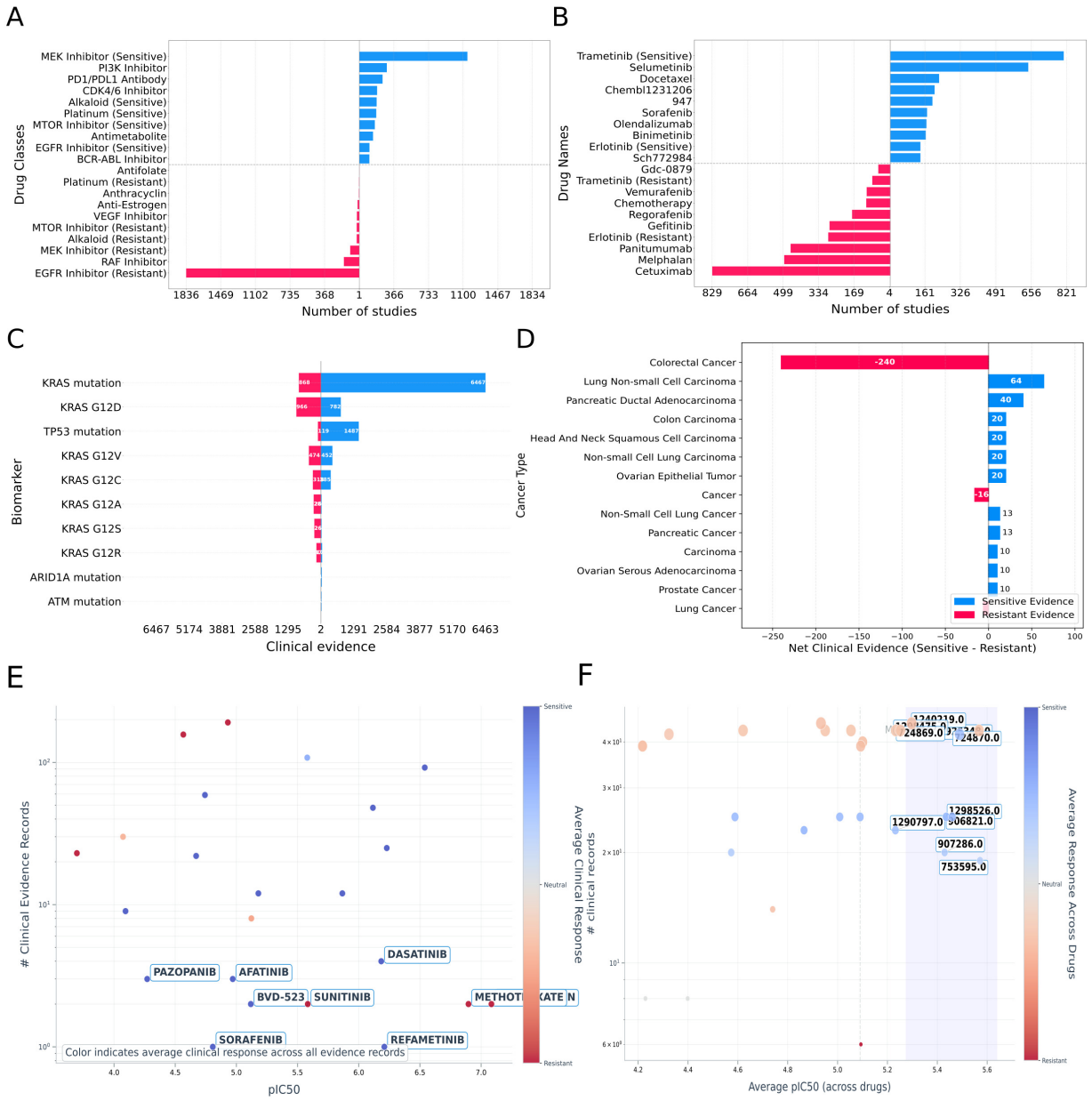


Figure 4: Clinical evidence analysis for patient cohorts and drug screen analysis. (**A**) Top 10 sensitive and top 10 resistant drug classifications for pancreatic cancer patients. Clinical studies were retrieved using Onkopus exact matches, which retrieves only evidence for the exact biomarker found in patient's mutational profile. (**B**) Top 10 sensitive and resistant therapeutics, and (**C**) number of studies per biomarker with sensitive (blue) and resistant (red) outcome. (**D**) Top sensitive and resistant clinical outcome per cancer type, based on PAAD patient biomarkers. (**E**) Comparison of clinical evidence records and drug screen results (pIC50) per therapeutic option. Drugs with low clinical evidence, but high pIC50 values are highlighted, indicating potentially underexplored, sensitive therapeutics. (**F**) Comparison of clinical evidence and drug screen results based on cell lines.

4 Discussion

We presented AstraKit, a comprehensive suite of modular, KNIME-based workflows designed for diverse applications in bioinformatics and precision medicine, including clinical-grade genetic variant interpretation, multi-layered omics analyses, and translational drug-response modeling. Unlike monolithic pipeline frameworks, AstraKit emphasizes visual configurability and end-to-end automation while main-

taining compatibility with clinical data governance requirements. As AstraKit leverages the KNIME analytics platform, all workflows operate through graphical, drag-and-drop nodes that eliminate the need for coding expertise. While command-line tools like Snakemake and Nextflow excel in scalable high-performance computing environments, and Galaxy offers cloud-based accessibility, KNIME provides a distinctive balance of visual workflow design, local execution capability, and integration with statistical/ML libraries, making it particularly suitable for processing sensitive clinical data under regulatory frameworks. AstraKit addresses critical gaps in existing tools by replacing error-prone manual tool chaining with fully automated genomic processing pipelines, encompassing variant interpretation, filtering, and coordinate conversion (LiftOver). They further enable scalable cohort analysis for mutational signature profiling, mutual exclusivity testing, and stratified omics investigations. Finally, they integrate clinical evidence with preclinical drug-response data through systematic correlation of patient genomic biomarkers with in vitro drug sensitivity metrics, thereby identifying candidate therapies even when clinical evidence is limited.

A key innovation is AstraKit's modular architecture for variant processing. Prior frameworks primarily automate variant calling, but offer limited flexibility in downstream annotation and filtering. In contrast, AstraKit decomposes the entire variant interpretation workflow into discrete, interchangeable nodes, allowing users to dynamically configure annotation sources (e.g., ClinVar, GTEx, CIViC), filtering criteria, and LiftOver steps without scripting. For cohort-scale analyses, AstraKit identifies mutational patterns, pathway enrichments, and mutual exclusivity across patient groups. Critically, while most dependencies are auto-installed via KNIME, license-restricted tools (e.g., FoldX for protein stability prediction) require pre-installation by the user, a limitation noted in our documentation.

In drug-response modeling, AstraKit bridges two complementary approaches: First, clinical evidence-based interpretation is enabled by annotating variants to match therapies with existing guidelines (e.g., OncoKB, CIViC). Second, AstraKit provides a preclinical data-driven discovery by correlating genomic alterations with high-throughput drug screen results to uncover novel biomarker-drug associations. This requires rigorous normalization of dose-response curves and statistically robust thresholding (e.g., using FDR correction) to discretize continuous drug sensitivity metrics values into clinically meaningful "sensitive" vs. "resistant" categories, a non-trivial challenge due to cell-line-specific confounders.

Validation of this approach revealed significant variability in drug response correlations across cell lines, suggesting that lineage-specific biomarkers (e.g., tissue-of-origin effects) critically influence sensitivity. This aligns with recent work by Schmidt et al. (2025) [42], who applied this methodology using MTB-Report for retrieving clinical evidence. The comparison to experimental drug sensitivity revealed T-cell malignancies and identified DDR pathway alterations as biomarkers for WEE1 inhibitor sensitivity, demonstrating how preclinical-clinical integration can reveal actionable targets absent from current databases. In summary, AstraKit delivers a unified, clinically adaptable platform covering the full precision oncology workflow, from raw sequencing data to therapeutic recommendations, while ensuring local execution for data privacy and regulatory compliance. By unifying fragmented tools into auditable, visual workflows, AstraKit accelerates biomarker validation and the translation of genomic insights into personalized treatment strategies, particularly benefiting clinicians and experimental biologists.

Data availability

The KNIME workflows are publicly available at KNIME HUB at https://hub.knime.com/bioinf_goe/spaces/Public/~Zd6-FraRjnV7MMpu/. The source code of the KNIME workflows and the data preprocessing scripts are available at https://gitlab.gwdg.de/MedBioinf/mtb/astrakit. The source code of all Onkopus modules is available at https://gitlab.gwdg.de/MedBioinf/mtb/onkopus, including instructions on how to install the Onkopus modules locally. All modules have been generated from freely available data sources.

Funding

This work was supported by the Gemeinsamer Bundesauschuss (01NVF20006), the Volkswagen Foundation (11-76251-12-1/19), the Deutsche Krebshilfe (70114018), the Deutsche Forschungsgemeinschaft (KFO5002) and the Bundesministerium für Bildung und Forschung (BMBF) (01KD2437, 01KD2401B, 01KD2208A, 01KD2414A).

Competing interests

No competing interest is declared.

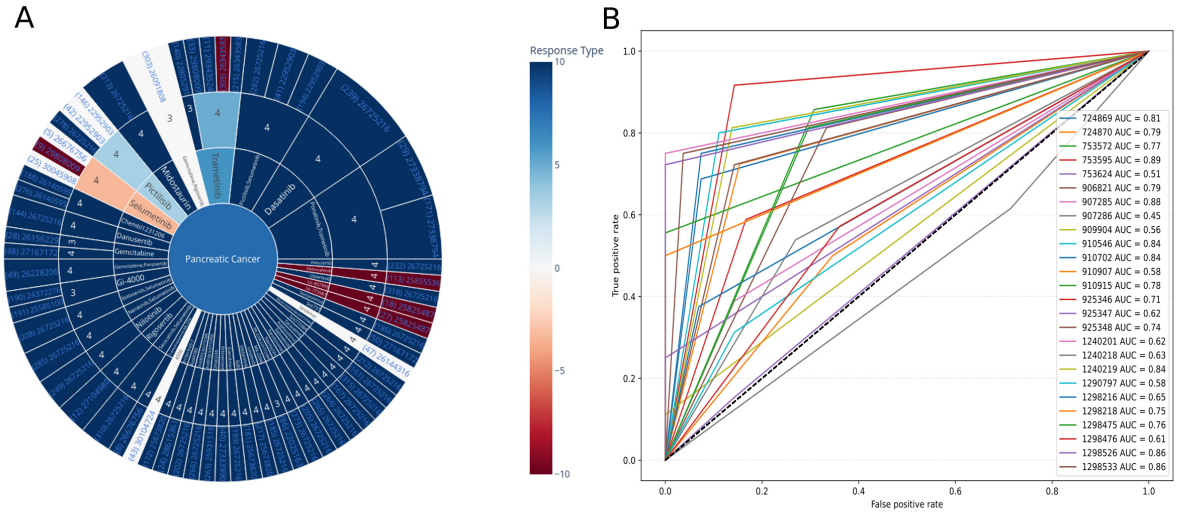


Figure 5: Drug screen analysis: Comparison between experimental drug sensitivity and clinical outcome. (**A**) Clinical studies for KRAS:p.G12D, filtered to clinical evidence for pancreatic cancer, visualized in the Onkopus sunburst visualization. Blue results indicate sensitive, red resistant outcome. (**B**) Receiver operating characteristic of clinical outcome of TCGA PAAD patient biomarkers predicting experimental drug sensitivity (IC_{50}) per cell line. The correspondence between drug screen and clinical studies' results varies significantly across cell lines, ranging from an area under the ROC curve (AUC) from 0.45 up to 0.89.

Author contributions

J.D. and N.S.K. developed the conceptualization and methodology. N.S.K., K.K. and M.S. implemented the KNIME workflows. N.S.K. wrote the manuscript. All authors approved the manuscript.

Acknowledgments

The authors thank the Molecular Tumor Board of the University Medical Center Göttingen (UMG), namely Alexander König, Li Beißbarth and Kirsten Reuter-Jessen, for providing constant feedback and discussions on clinical needs and use cases. The authors also thank the International Max Planck Research School for Genome Science (IMPRS-GS).

References

[1] Babadi, M., Fu, J., Lee, S., Smirnov, A., Gauthier, L., Walker, M., Benjamin, D., Zhao, X., Karczewski, K., Wong, I. & Others GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. *Nature Genetics*. **55**, 1589-1597 (2023)

[2] Danecek P, Bonfield J, Liddle J et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

[3] Quinlan A, Hall I. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

[4] Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*. **38**, e164-e164 (2010)

[5] McLaren, W., Gil, L., Hunt, S., Riat, H., Ritchie, G., Thormann, A., Flicek, P. & Cunningham, F. The ensembl variant effect predictor. *Genome Biology*. **17**, 122 (2016)

[6] Kurz NS, Kornrumpf K, Tucholski T et al. Onkopus: precise interpretation and prioritization of sequence variants for biomedical research and precision medicine. *Nucleic Acids Res* pp. gkaf376 (2025).

[7] Cingolani, P., Patel, V., Coon, M., Nguyen, T., Land, S., Ruden, D. & Lu, X. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers In Genetics*. **3** pp. 35 (2012)

[8] Kurz NS, Drofenik K, Kornrumpf K et al. AdaGenes: A streaming processor for high-throughput annotation and filtering of sequence variant data. *BioRxiv* pp. 2025-06 (2025).

[9] Danecek, P., Auton, A., Abecasis, G., Albers, C., Banks, E., DePristo, M., Handsaker, R., Lunter, G., Marth, G., Sherry, S. & Others The variant call format and VCFtools. *Bioinformatics.* **27**, 2156-2158 (2011)

[10] Mayakonda A, Lin D, Assenov Y et al. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res* **28**, 1747 (2018).

[11] Kurz, N., Perera-Bel, J., Höltermann, C., Tucholski, T., Yang, J., Beissbarth, T. & Dönitz, J. Identifying Actionable Variants in Cancer-The Dual Web and Batch Processing Tool MTB-Report. *Studies In Health Technology And Informatics.* **296** pp. 73-80 (2022)

[12] Jiménez-Santos MJ, Nogueira-Rodríguez A, Piñeiro-Yáñez E et al. PanDrugs2: prioritizing cancer therapies using integrated individual multi-omics data. *Nucleic Acids Res* **51**, W411-W418 (2023).

[13] Potdar S, Ianevski F, Ianevski A et al. Breeze 2.0: an interactive web-tool for visual analysis and comparison of drug response data. *Nucleic Acids Res* **51**, W57-W61 (2023).

[14] Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520-2522 (2012).

[15] Di Tommaso P, Chatzou M, Floden E et al. Nextflow enables reproducible computational workflows. *Nat Biotechnology* **35**, 316-319 (2017).

[16] Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. & Others The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research.* **46**, W537-W544 (2018)

[17] Beg, M., Taka, J., Kluyver, T., Konovalov, A., Ragan-Kelley, M., Thiéry, N., Fangohr, H. Using Jupyter for reproducible scientific workflows. *Computing In Science & Engineering.* **23**, 36-46 (2021)

[18] Fillbrunn A, Dietz C, Pfeuffer J et al. KNIME for reproducible cross-domain analysis of life science data. *J Biotechnol* **261** pp. 149-156 (2017).

[19] Herrick N, Walsh S. ILIAD: a suite of automated Snakemake workflows for processing genomic data for downstream applications. *BMC Bioinformatics* **24**, 424 (2023).

[20] Yang J, Beißbarth T, Dönitz J. Onkopipe: A Snakemake Based DNA-Sequencing Pipeline for Clinical Variant Analysis in Precision Medicine. *Stud Health Technol Inf* **307** pp. 60-68 (2023).

[21] Ardin M, Cahais V, Castells X et al. MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. *BMC Bioinformatics* **17** pp. 1-10 (2016).

[22] Hiltemann S, Mei H, Hollander M et al. CGtag: complete genomics toolkit and annotation in a cloud-based Galaxy. *GigaScience* **3**, 2047-217X (2014).

[23] Larivière D, Abueg L, Brajuka N et al. Scalable, accessible and reproducible reference genome assembly and evaluation in Galaxy. *Nat Biotechnology* **42**, 367-370 (2024).

[24] Wratten L, Wilm A, Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nat Methods* **18**, 1161-1168 (2021).

[25] Fjukstad B, Dumeaux V, Hallett M, Bongo L. Reproducible data analysis pipelines for precision medicine. *2019 27th Euromicro International Conference On Parallel, Distributed And Network-Based Processing (PDP)* pp. 299-306 (2019).

[26] Pereira B, Chin S, Rueda O et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat Commun* **7** pp. 11479 (2016).

[27] Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Res* **48**, D941-D947 (2020).

[28] Colaprico A, Silva T, Olsen C et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research* **44**, e71-e71 (2016).

[29] Andrade, K., Lee, E., Tookmanian, E., Kesserwan, C., Manfredi, J., Hatton, J., Loukissas, J., Zavadil, J., Zhou, L., Olivier, M. & Others The TP53 database: transition from the International Agency for Research on Cancer to the US National Cancer Institute. *Cell Death & Differentiation.* **29**, 1071-1073 (2022)

[30] Tsherniak, A., Vazquez, F., Montgomery, P., Weir, B., Kryukov, G., Cowley, G., Gill, S.,

Harrington, W., Pantel, S., Krill-Burger, J. & Others Defining a cancer dependency map. *Cell.* **170**, 564-576 (2017)

[31] Yang W, Soares J, Greninger P et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* **41**, D955-D961 (2012).

[32] Kornrumpf K, Kurz NS, Drofenik K et al. SeqCAT: Sequence Conversion and Analysis Toolbox. *Nucleic Acids Res* **52**, W116-W120 (2024).

[33] Griffith M, Spies N, Krysiak K et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genetics.* **49**, 170-174 (2017).

[34] Wagner A, Walsh B, Mayfield G et al. A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. *Nat Genetics* **52**, 448-457 (2020).

[35] Chakravarty D, Gao J, Phillips S et al. OncoKB: a precision oncology knowledge base. *JCO Precis Oncol* **1** pp. 1-16 (2017).

[36] Kundra, R., Zhang, H., Sheridan, R., Sirintrapun, S., Wang, A., Ochoa, A., Wilson, M., Gross, B., Sun, Y., Madupuri, R. & Others OncoTree: a cancer classification system for precision oncology. *JCO Clinical Cancer Informatics.* **5** pp. 221-230 (2021)

[37] Gaulton A, Hersey A, Nowotka M et al. The ChEMBL database in 2017. *Nucleic Acids Res* **45**, D945-D954 (2017).

[38] Pedregosa F, Varoquaux G, Gramfort A et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res* **12** pp. 2825-2830 (2011).

[39] Waskom M. Seaborn: statistical data visualization. *J Open Source Software* **6**, 3021 (2021).

[40] Cock P, Antao T, Chang J et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422 (2009).

[41] Delgado J, Radusky L, Cianferoni D, Serrano L. FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics* **35**, 4168-4169 (2019).

[42] Schmidt, N., Kornrumpf, K., Ries, L., Markus, K., Ksionsko, N., Mazzeo, P., Lohrberg, M., Sitte, M., Salinas, G., Braun, T. & Others Discovery of molecularly-informed therapeutic strategies for mature T-cell leukemias and lymphomas. *Blood Advances.* **9**, 5327-5340 (2025)