A Decision Support Framework for the Interpretation of Genetic Alterations in Molecular Tumor Boards

Dissertation
for the award of the degree

"Doctor rerum naturalium" (Dr. rer. nat.)
of the Georg-August-Universität Göttingen

within the doctoral program

International Max Planck Research School for Genome Science (IMPRS-GS)

of the Georg-August University School of Science (GAUSS)

submitted by
Nadine Sina Kurz

Göttingen, 2026

Thesis Advisory Committee

Dr. Jürgen Dönitz
Department of Medical Bioinformatics
University Medical Center Göttingen

Prof. Dr. Elisabeth Heßmann
Department of Gastroenterology and Intestinal Oncology
University Medical Center Göttingen

Dr. Johannes Söding
Quantitative and Computational Biology
Max Planck Institute for Multidisciplinary Sciences

Members of the Examination Board

1st Referee: Dr. Jürgen Dönitz
Department of Medical Bioinformatics
University Medical Center Göttingen

2nd Referee: Prof. Dr. Elisabeth Heßmann
Department of Gastroenterology and Intestinal Oncology
University Medical Center Göttingen

Further members of the Examination Board

Dr. Johannes Söding
Quantitative and Computational Biology
Max Planck Institute for Multidisciplinary Sciences

Prof. Dr. Anna Lena Illert
Department of Hematology and Medical Oncology
University Medical Center Göttingen

Prof. Dr. Ulrich Sax
Department of Medical Informatics
University Medical Center Göttingen

Prof. Dr. Anne-Christin Hauschild
Department of Predictive Deep Learning for Medicine and Healthcare
Justus-Liebig University Gießen

Date of oral examination: April 27, 2026

Abstract

Precision medicine promises novel effective cancer therapies through selecting therapeutics based on the individual molecular profiles of tumors. However, realizing this potential is critically bottlenecked by the interpretation of somatic variants, where distinguishing clinically actionable driver variants from benign polymorphisms and variants of uncertain significance demands rigorous, multidisciplinary integration of genomic, functional, and clinical evidence, a complex and time-intensive process for molecular tumor boards.

This thesis introduces a suite of integrated computational frameworks and tools to support the processing of molecular data, and somatic variant interpretation in particular, in precision medicine and biomedical research. To address the initial challenge of identifying tumor driver mutations in whole-genome and whole-exome sequencing data, this work presents *AdaGenes*, a framework for high-throughput variant annotation and filtering that is capable of annotating and filtering large-scale variant data with specific features. Leveraging tumor-specific genomic features, *AdaGenes* efficiently reduces the vast number of somatic variants to a prioritized shortlist of high-confidence candidate drivers.

The subsequent, more complex task, detailed biological and clinical interpretation of prioritized variants, requires assessing pathogenicity, protein-level functional consequences, and therapeutic implications. For this, two complementary frameworks were developed within this thesis to support the interpretation of genetic variants: The clinical variant interpretation framework *MTB-Report*, which features a unique variant classification, and the comprehensive clinical decision support framework *Onkopus*, which comprises a multitude of annotation modules to analyze a multitude of biological consequences as well as targeted therapies for genetic variants. Critically, *Onkopus* incorporates protein structural modeling and biochemical feature analysis to predict variant-induced disruptions in protein stability, interactions, and catalytic activity.

Beyond point mutations, structural variants like copy-number alterations are significant drivers of cancer. This thesis establishes *PanCNV-Explorer*, a pan-cancer database of somatic copy number alterations (CNAs) across cancer entities and normal tissue. *PanCNV-Explorer* enables systematic analysis of cancer-type-specific CNA frequencies and identifies dosage-sensitive genes whose copy-number changes confer selective advantages.

To expand the analysis of molecular profiles beyond interpreting genetic variants, such as integrating multi-layered omics analyses and analyzing drug screen data, *AstraKit* was developed, which provides a set of modular, reproducible workflows for a multitude of processing and analysis tasks in biomedical research and precision medicine.

Genomic data harmonization across reference genomes and notations remains a persistent hurdle. This thesis introduces *SeqCAT*, a robust toolkit for precise coordinate conversion and cross-referencing of genes/variants between genome assemblies and notations at genomic, transcript and protein level. Finally, for structural variants involving gene fusions, a major oncogenic mechanism, this work develops *FusionPath*, a deep learning method that predicts fusion pathogenicity by integrating protein language models with domain architecture features. *FusionPath* significantly outperforms existing tools in accuracy and biological interpretability.

Collectively, these frameworks bridge the critical challenge in the precision medicine to translate molecular profiles into clinical action. By standardizing and automating complex analytical steps, this thesis delivers scalable solutions that enhance the accuracy, efficiency, and clinical utility of molecular tumor boards, providing a significant benefit in advancing personalized cancer therapy.

Acknowledgements

First and foremost, I want to express my deepest gratitude to my supervisor Dr. Jürgen Dönitz, without whose commitment and support this thesis would not have been possible. Jürgen was a most effective supervisor with unwavering commitment, ideas, advice and steadfast encouragement that were indispensable for the completion of this work. His collaborative approach fostered an environment of intellectual rigor and teamwork, enabling the success of our large-scale projects. I am very grateful for the opportunity to engage in cancer research, and the prospect of contributing to finding effective treatments for one of the most lethal diseases has consistently motivated me throughout my entire doctorate.

I sincerely thank my thesis advisory committee, Prof. Elisabeth Heßmann and Dr. Johannes Söding, for their invaluable expertise, constructive feedback, and thoughtful counsel. Their perspectives were instrumental in refining my research, navigating challenges, and maintaining steady progress. I especially admire Elisabeth's exceptional clinical acumen and Johannes' encyclopedic knowledge across all aspects of computational biology.

My appreciation extends to the members of the molecular tumor board of the University Medical Center Göttingen for welcoming us to their conferences, offering profound insights into precision medicine in practice. Special thanks are due to Dr. Alexander König, Dr. Kirsten Reuter-Jessen, Dr. Li Beißbarth, Dr. Julia Beck and Dr. Nelia Nause for their regular collaboration, rigorous testing of our software, and indispensable feedback that significantly enhanced our applications.

I would like to thank Prof. Dr. Tim Beißbarth for his illuminating discussions and expert advice, Daniela Großmann for her efficient management of my administrative affairs, and Torsten Schöps for his support in maintaining our servers and virtual machines.

I would like to acknowledge my co-authors Kevin Kornrumpf, Klara Drofenik, Tim Tucholski, Julia Perera-Bel, Martin Stoves and Charlotte Höltermann, for their pivotal contributions to our bioinformatic frameworks. I also want to thank my lab colleagues Jingyu Yang, Yacoub Nsangou, and Can Nguyen Duc. Our discussions within our Journal Club have greatly fueled my scientific enthusiasm, and I will fondly remember the barbecue evenings in the rock garden in front of our institute.

I extend my thanks to Dr. Manuel Nietert for our thoughtful discussions concerning

how to continue with a PhD and the prospects and difficulties involved in planning one's own academic life. And I want to thank my former office colleagues Robin Kosch, Thomas Sterr, Laurenz Engel and Stefan Schrod for many lighthearted conversations.

My gratitude also goes to my students who contributed to my projects, Anna-Rosa Krüger and Irem Berna Güven. I also thank all members of the Department of Medical Bioinformatics at the University Medical Center Göttingen for their collegial support.

I would like to sincerely thank Dr. Henriette Irmer and Frauke Bergmann, coordinators of the IMPRS-GS, for all their support throughout my whole PhD. And I want to thank my friends from the IMPRS-GS, especially Nicole Seifert, Youngjun Park, Nisha Hemandhar Kumar and Christoph Elfmann, who made my PhD time a great experience.

And I would like to thank all students of the IMPRS-GS for their trust in electing me as their student representative in 2024. I did my best to represent the students' interests and support them in their PhD studies. I also want to thank Clemens Mauksch and Maaike Bierenbroodspot, for their dedicated help in organizing our IMPRS-GS Retreat 2024, who made our conference an unforgettable experience.

And finally, I am incredibly grateful to my family, especially my parents, Susann and Peter, and my brother Jens, who strongly encouraged me to do a PhD in Göttingen and always supported me throughout my entire doctoral journey.

This achievement is as much yours as it is mine.

Publication list

- Kurz, N.S. , Kornrumpf, K., Tucholski, T., Drofenik, K., König, A., Beißbarth, T., & Dönitz, J. (2025). Onkopus: precise interpretation and prioritization of sequence variants for biomedical research and precision medicine. Nucleic Acids Research, gkaf376. doi: 10.1093/nar/gkaf376
- Kurz, N.S. , Perera Bel, J., Höltermann, C., Tucholski, T., Yang, J., Beißbarth, T., & Dönitz, J. (2022). Identifying actionable variants in cancer - The dual web and batch processing tool MTB-report. Stud Health Technol Inform. 2022 Aug 17; 296: 73-80. doi: 10.3233/shti220806
- Kurz, N.S. , Drofenik, K., Kornrumpf, K., Reuter-Jessen, K., & Dönitz, J. AdaGenes: A streaming processor for high-throughput annotation and filtering of sequence variant data. bioRxiv, 2025-06. doi: 10.1101/2025.06.17.659929
- Kurz, N.S. , Kornrumpf, K., Krüger, A.R., & Dönitz, J. PanCNV-Explorer: Deciphering copy number alterations across human cancers. bioRxiv, 2026-01. doi: 10.64898/2026.01.26.701884
- Kurz, N.S. , Kornrumpf, K. & Dönitz, J. (2025) AstraKit: Customizable, reproducible workflows for biomedical research and precision medicine. bioRxiv, 2026-01. doi: 10.64898/2026.01.25.701622
- Kurz, N.S. , Güven, I.B., Beißbarth, T. & Dönitz, J. FusionPath: Gene fusion pathogenicity prediction using protein structural data and contextual protein embeddings. bioRxiv, 2026-01. doi: 10.64898/2026.01.24.701525
- Kornrumpf, K., Kurz, N.S. , Drofenik, K., Krauß, L., Schneider, C., Koch, R., Beißbarth, T. & Dönitz, J. SeqCAT: Sequence Conversion and Analysis Toolbox. Nucleic acids research, gkae422. doi: 10.1093/nar/gkae422

In addition to these publications, the author also contributed to other related works during her PhD research, which are not covered in this thesis:

- Kosch, R., Limm, K., Staiger, A.M., Kurz, N.S. , Seifert, N., Ohlh, B., ... & Altenbuchinger, M. (2025). Integration of high-throughput proteomic data and complementary omics layers with PriOmics. Genome Research, gr-279487. doi: 10.1101/gr.279487.124

ˆ Seifert, N., Reinke, S., Kurz, N.S. , Demmer, J., Kornrumpf, K., Beibarth, T., ... & Altenbuchinger, M. (2025). TCRanalyzer: A user-friendly tool for comprehensive analysis of T-cell diversity, dynamics and potential antigen targets. bioRxiv, 2025-05. doi: 10.1101/2025.05.09.652820

ˆ Steinbrenner, I., Kotsis, F., Kosch, R., Meiselbach, H., Barthlein, B., Stockmann, H., ..., Kurz, N.S. , ... & Schultheiss, U. T. (2024). Interactive exploration of adverse events and multimorbidity in CKD. Nephrology Dialysis Transplantation, 39(12), 2016-2024. doi: 10.1093/ndt/gfae092

Table of Contents

1 Introduction 1
- 1.1 Precision medicine and targeted cancer therapy . . . 2
  - 1.1.1 Molecular tumor boards . . . 3
  - 1.1.2 Genetic alterations in cancer . . . 5
- 1.2 Challenges of genetic variant interpretation . . . 7
  - 1.2.1 Genomic, transcript and protein associations . . . 8
  - 1.2.2 Computational variant pathogenicity prediction . . . 9
  - 1.2.3 Biochemical and protein structural characteristics of variants . 10
  - 1.2.4 Clinical implications . . . 13
- 1.3 Structural variant analysis . . . 15
- 1.4 Functional validation for variant interpretation and therapeutic vulnerability pro ling . . . 17
- 1.5 Computational solutions for genetic variant interpretation . . . 18
  - 1.5.1 Clinical decision support systems . . . 19
  - 1.5.2 Automated work ows . . . 21
  - 1.5.3 Deep learning architectures for variant pathogenicity prediction 21
- 1.6 Thesis Outline . . . 23
  - 1.6.1 Research aims . . . 23
  - 1.6.2 Organization of this thesis . . . 25

2 High-throughput processing of sequence variant data 27

3 Clinical interpretation of genetic variants 38

4 A computational framework for comprehensive variant interpretation 47

5 A pan-cancer database of copy number alterations in the human genome 57

6 Reproducible, customizable work ows for biomedical research 67

7 Sequence conversion and analysis 79

8 Gene fusion pathogenicity prediction 85

9 Discussion 94
9.1 E ciency gains through high-throughput variant annotation and l-tering 95
9.2 Advancing cancer variant interpretation 97
9.2.1 CDSS as a universal resource for MTBs 97
9.2.2 Uni ed clinical-biological variant interpretation 100
9.2.3 Insights from molecular and protein structural analysis 102
9.2.4 Sequence analysis and conversion 104
9.3 Pan-cancer analysis of human copy number variation 105
9.4 Work ow automation for biomedical research 107
9.5 Gene fusion pathogenicity prediction 109

10 Conclusion 112

Bibliography 114

A List of modules 137

B List of API endpoints (Onkopus) 140

C List of web front ends 142

D Declaration: Use of AI tools 143

List of Figures

1.1 The Molecular Tumor Board work ow . The work ow of a molecular tumor board (MTB) involves collecting a tissue sample from a patient, performing variant calling through sequencing to identify genetic variants, and interpreting these variants for clinical signi cance. The MTB conference convenes a multidisciplinary team to evaluate the ndings and develop a personalized treatment recommendation based on the patient's molecular pro le. Based on Strantz et al. (2025). 3

1.2 Thesis structure This thesis develops an integrated suite of computational frameworks and tools to advance precision oncology and biomedical research, with a focus on cancer variant interpretation: A framework for annotation and ltering of WES/WGS data (AdaGenes), a framework for clinical variant interpretation (MTB-Report), a framework for comprehensive biological and clinical interpretation of genetic alterations (Onkopus). A tool for the harmonization and analysis of genetic data (SeqCAT) and a novel method for gene fusion pathogenicity prediction (FusionPath) support biomedical research and the variant interpretation tools. To further support the analysis of structural variants, a novel copy number variant database and analysis framework (PanCNV-Explorer) supports preclinical and foundational research in biomarker discovery and pan-cancer databases. In addition, platform-independent functionalities for a multitude of tasks in precision medicine, including multi-omics and drug screen analyses, are provided by a suite of automated work ows (AstraKit). 24

List of Tables

A.1 Overview of modules for variant interpretation of sequence variants . 138
A.2 Overview of variant interpretation modules for structural variants . . 139
A.3 Overview of gene annotation modules 139
A.4 Overview of infrastructure modules 139

B.1 Overview of API endpoints 141

Abbreviations

ACMG American College of Medical Genetics and Genomics 14
AI Arti cial Intelligence 21
AMP Association for Molecular Pathology 14
API Application Programming Interface 20
ASA Accessible Surface Area 12
AUC Area Under the Curve 22
ASCO American Society of Clinical Oncology 14
BAM Binary Alignment Map 20
BED Browser Extensible Data 20
BLOSUM BLOcks SUbstitution Matrix 11
bps base pairs 7
CADD Combined Annotation Dependent Depletion 9
CAP College of American Pathologists 14
cBioPortal cBio Cancer Genomics Portal 19
CDSS Clinical Decision Support System 18
CGC Cancer Genomics Consortium 14
CIViC Clinical Interpretation of Variants in Cancer 13
CNV Copy Number Variation 7
CRC Colorectal Cancer 2
CRISPR Clustered Regularly Interspaced Short Palindromic Repeats . . . 17
DMS Deep Mutational Scanning 100
DNA Deoxyribonucleic Acid 4
EHR Electronic Health Record 19
ESCAT ESMO Scale for Clinical Actionability of molecular Targets 14
ESM Evolutionary Scale Modeling 9

ESMO European Society for Medical Oncology 14
ExAC Exome Aggregation Consortium 9
FAIR Findable, Accessible, Interoperable and Reusable 109
FDA Food and Drug Administration 104
GEO Gene Expression Omnibus 17
GFF General Feature Format 20
GISTIC Genomic Identi cation of Signi cant Targets in Cancer 16
GoF Gain-of-Function 6
GO Gene Ontology 12
gnomAD Genome Aggregation Database 9
GR Growth Rate 17
GRAVY Grand Average of Hydropathy 11
GUI Graphical User Interface 20
HGNC HUGO Gene Nomenclature Committee 104
HGVS Human Genome Variation Society 8
HI Haploinsu ciency 16
InDel Insertions and Deletions 6
HL7/FHIR Fast Healthcare Interoperability Resources 99
KNIME Konstanz Information Miner 21
LLM Large Language Model 22
LOEUF Loss-of-Function Observed/Expected Upper bound Fraction 10
LoF Loss-of-Function 6
MAF Mutation Annotation Format 20
MAVE Multiplexed Assays of Variant E ect 100
MaxASA Maximum possible solvent Accessible Surface Area 12
MIRACUM Medical Informatics in Research and Care in University Medicine 19

MPRA Massively Parallel Reporter Assay 100
MSI Microsatellite Instability 7
MTB Molecular Tumor Board 4
NCCN National Comprehensive Cancer Network 14
NGS Next-Generation Sequencing 2
NSCLC Non-Small-Cell Lung Carcinoma 1
PDAC Pancreatic Ductal Adenocarcinoma 2
PD-L1 Programmed Death-Ligand 1 7
PFS Progression-Free Survival 4
pLDDT predicted Local Distance Di erence Test 13
pLI probability of Loss-of-Function intolerance 10
pLM protein Language Models 9
PolyPhen2 Polymorphism Phenotyping v2 9
REVEL Rare Exome Variant Ensemble Learner 9
RNA Ribonucleic Acid 4
RNA-seq RNA sequencing 2
ROC Receiver Operating Characteristic 22
RSA Relative accessible Surface Area 11
SAM Sequence Alignment Map 20
SeqCAT Sequence Converter and Analysis Toolbox 104
SHAP SHapley Additive exPlanations 23
SIFT Sorting Intolerant From Tolerant 9
SNV Single Nucleotide Variant 6
T2T-CHM13 Telomere-to-Telomere assembly 20
TCGA The Cancer Genome Atlas 15
TMB Tumor Mutational Burden 7

TS Triplosensitivity 16
TTD Therapeutic Target Database 15
VEP Variant E ect Predictor 20
VCF Variant Call Format 20
VICC Variant Interpretation for Cancer Consortium 14
VUS Variants of Unknown Signi cance 5
VQSR Variant Quality Score Recalibration 4
WGS Whole Genome Sequencing 2
WES Whole Exome Sequencing 2

1 Introduction

Cancer represents one of the most signi cant global health challenges of the modern era. With nearly 10 million annual fatalities worldwide [1] and incidence rates projected to rise by approximately 50% over the next two decades [1], the disease underscores an urgent need for innovative therapeutic strategies. This profound clinical challenge stems from cancer's inherent biological complexity and evolutionary dynamism: Cancer encompasses more than 100 distinct histological types, each further fragmented into molecularly de ned subtypes driven by unique compositions of somatic mutations, epigenetic alterations, and dysregulated signaling pathways [2, 3]. Critically, these alterations converge on core cancer hallmark capabilities, such as sustained proliferative signaling, evasion of growth suppressors, and resistance to cell death, yet are achieved through diverse molecular mechanisms even within the same organ site or histological subtype [2].

The two most prevalent cancer types worldwide, breast and lung cancer, exemplify profound genetic heterogeneity: Histologically identical diagnoses often conceal molecularly distinct diseases with divergent clinical behaviors and therapeutic responses. For instance, breast cancer comprises Luminal A, Luminal B, HER2-enriched, and Basal-like subtypes (de ned by transcriptomic pro ling), each characterized by distinct genomic alteration patterns, therapeutic vulnerabilities, and clinical trajectories [4, 5, 1]. Similarly, Non-Small-Cell Lung Carcinoma (NSCLC) is routinely strati ed by actionable genomic alterations, including *EGFR* mutations and *ALK* fusions, rendering broad "lung cancer" classi cation clinically inadequate for treatment selection [6]. Consequently, despite advances in early detection, survival rates remain dismal for many advanced malignancies, falling below 6.6-10.9% in pancreatic ductal adenocarcinoma for 3-year survival [7] and 2.2% in glioblastoma for 5-year survival [8]. Conventional cancer therapies, including surgery, chemotherapy, and radiation, have long operated under the premise that cancers of the same histological origin respond uniformly to treatment. While e ective for some patients, these modalities embody a "one-size- ts-all" approach that targets rapidly dividing cells indiscriminately, ignoring the molecular heterogeneity de ning individual tumors and their microenvironments. These therapies have two critical limitations: First, severe on-target/o -tumor toxicity arises from collateral damage to healthy proliferating tissues. Second, and more insidiously, therapeutic e cacy is

undermined by cancer's inherent diversity. Tumor heterogeneity fuels therapeutic resistance through multiple, overlapping mechanisms, including clonal selection, adaptive DNA repair, and microenvironmental shielding, rendering standard therapies ine ective for molecularly distinct subpopulations [9]. Critically, empirical dosing based on population averages overlooks individual genomic signatures, leaving many patients with suboptimal outcomes in molecularly distinct subpopulations. Driven by these limitations and accelerated by advances in genomic pro ling, cancer therapy is shifting toward precision oncology. This paradigm explicitly incorporates the molecular complexity of cancer and leverages individual patient tumor pro ling to optimize treatment e cacy, minimize toxicity, and overcome resistance through biomarker-driven therapies and dynamic treatment adaptation [10].

1.1 Precision medicine and targeted cancer therapy

The limitations of histology-driven oncology have catalyzed the emergence of precision medicine, a strategy that deciphers each tumor's unique genetic molecular pro- le to match therapies to its speci c drivers. Precision medicine integrates patient-speci c multi-omics data, including genomic, transcriptomics, and proteomics, to identify cancer driver mutations, stratify patients into subgroups with shared molecular pro les, and match drugs to patients [11]. This approach is grounded in two key principles: (1) treatment e cacy is substantially enhanced by incorporating the tumor's genomic context, and (2) patients with molecularly similar tumors, including shared driver alterations, will exhibit comparable responses to targeted therapies. For instance, RAF-directed therapy is highly e ective in certain cancer types, including melanoma [12], but demonstrates minimal clinical bene t in BRAF V600E-mutant colon cancer [13, 14], highlighting how tissue context in u-ences therapeutic response. Among the rst validated targeted cancer therapies were the development of trastuzumab for HER2-positive breast cancer [15], and BCR-ABL tyrosine kinase inhibitors for chronic myeloid leukemia [16, 17]. These successes established the paradigm of targeting oncogenic drivers de ned by molecular biomarkers. While early genome sequencing (2000-2010) relied on locus-speci c techniques like Sanger sequencing and FISH, Next-Generation Sequencing (NGS) revolutionized genomics by enabling high-throughput, unbiased pro ling via Whole Exome Sequencing (WES), Whole Genome Sequencing (WGS), and RNA sequencing (RNA-seq). Among the signi cant successes of precision medicine is the development of targeted therapies for Kirsten rat sarcoma viral oncogene homologue (KRAS), an oncogene and one of the main driver genes for Pancreatic Ductal Adenocarcinoma (PDAC), Colorectal Cancer (CRC) and NSCLC, that was considered a di cult therapeutic target and even labeled "undruggable" for decades [18]. The development of therapeutics that directly target KRAS, in particular KRAS G12C, such as sotorasib and adagrasib, has yielded promising results in clinical trials [18].

Figure 1.1: The Molecular Tumor Board work ow . The work ow of a molecular tumor board (MTB) involves collecting a tissue sample from a patient, performing variant calling through sequencing to identify genetic variants, and interpreting these variants for clinical signi cance. The MTB conference convenes a multidisciplinary team to evaluate the ndings and develop a personalized treatment recommendation based on the patient's molecular pro le. Based on Strantz et al. (2025).

This breakthrough exempli es how structural biology and innovative drug design can overcome perceived therapeutic barriers. It has be shown that more than one third of patients could be treated more e ectively with targeted therapies [13]. Between 2017 and 2022, the proportion of cancer patients with tumors harboring biomarkers predictive of response to standard targeted therapies nearly tripled (from 8.9% to 31.6%), while tumors carrying nonactionable drivers decreased by almost half (44.2% to 22.8%), demonstrating signi cant expansion in precision oncology [19].

Despite remarkable progress, critical challenges persist. A crucial challenge in precision medicine is to select suitable therapy options based on a patient's molecular pro le, with the objective of ensuring the most e ective therapy [20]. Selecting optimal therapies based on molecular pro les requires distinguishing pathogenic driver mutations from benign polymorphisms or passenger alterations, a task complicated by tumor heterogeneity, clonal evolution, and the functional ambiguity of many variants. Molecular tumor boards, multidisciplinary teams integrating genomic, clinical, and functional data, address this by evaluating variant pathogenicity and matching patients to targeted therapies or clinical trials.

1.1.1 Molecular tumor boards

Translating molecular pro les of cancer patients into actionable clinical decisions requires precise variant analysis and interpretation within the context of targeted

therapy response. To address this complexity, a Molecular Tumor Board (MTB) leverages multidisciplinary integration of genomic data and clinical expertise to guide personalized cancer treatment [21]. Typically convened weekly or biweekly, MTBs comprise core specialists, including medical oncologists, molecular pathologists, clinical geneticists, bioinformaticians, and pharmacologists, among others [22], who collectively evaluate genomic findings within the context of a patient's clinical history, tumor type, prior therapies, and available therapeutic options. Core activities include identifying pathogenic driver variants, reviewing evidence-based literature, and prioritizing targeted therapies with the highest likelihood of clinical efficacy [22, 23]. Recommendations emerge from structured, patient-specific discussions during dedicated MTB conferences, embodying precision medicine's individualized ethos [23]. Crucially, while each patient is evaluated individually, the MTB functions strictly as an advisory body. The treating physician retains ultimate decision-making authority regarding treatment implementation. Patients with advanced cancer who received therapy matched to their molecular alterations via a multidisciplinary MTB demonstrate improved outcomes, with 30-40% achieving significantly longer Progression-Free Survival (PFS), overall survival, and higher clinical benefit rates compared to lower-matched or unmatched therapies [9, 24, 25, 26, 27, 28]. While significant survival benefits have been demonstrated for patients with actionable driver mutation (particularly BRAF V600E), only limited efficacy could be shown for tumors lacking such alterations [29]. Additionally, the cancer type is of significant importance when evaluating therapy options. While substantial benefits have been achieved in colon cancer with patients with BRAF V600E-mutated tumors, only BRAF/MEK inhibition has been shown to be effective in treating glioblastoma with patients harboring the same variant [30].

The MTB workflow begins with the collection of a cancer tissue sample from the patient Figure 1.1, which undergoes comprehensive genomic sequencing typically NGS of tumor Deoxyribonucleic Acid (DNA) and Ribonucleic Acid (RNA) to decode somatic alterations. Raw sequencing data is then processed using bioinformatic pipelines that align reads to a reference genome and apply variant-calling algorithms. Stringent quality filters, e.g. using depth, allele balance, and Variant Quality Score Recalibration (VQSR), eliminate technical artifacts, ensuring clinical-grade variant calls [31]. Subsequently, curated variants are annotated and interpreted. This process assesses pathogenicity, oncogenicity, and biomarker relevance to identify driver mutations and match actionable alterations to targeted therapies or clinical trials. For this purpose, a multitude of different data sources and tools has to be evaluated, including functional predictions, clinical databases, and literature evidence. The curated genomic findings and therapeutic hypotheses are presented during the multidisciplinary MTB conference, where the board members collaboratively evaluate clinical relevance, drug availability, and patient-specific factors. Finally, a consensus-based, prioritized therapeutic strategy is generated, emphasizing evidence-supported targeted agents or trials tailored to the patient's molecular profile. This expert consensus translates into a treatment recommendation, pri-

oritizing evidence-based targeted interventions or clinical trial options. Critically, the MTB serves as an advisory board, with the nal decision made by the treating physician [9].

However, a core challenge for MTBs is the required workload associated with evaluating the molecular pro les of patients and identifying potential personalized therapy, while in-depth review of molecular data can be time-consuming and could limit the scalability of MTBs [32]. The manual e ort for the MTB preparation is high [33], especially for looking up information distributed across many data sources, while the workload is expected to even increase in the coming years [32]. This process directly contributes to extended turnaround times, which can be approximately three weeks from DNA acquisition to molecular analysis completion for the majority of patients [32]. Overcoming this bottleneck to support and automate the interpretation of genetic variants is therefore an important requirement for speeding up and improving the e ciency of a MTB, for making targeted therapies available to a large number of cancer patients.

1.1.2 Genetic alterations in cancer

A typical human genome varies from the reference human genome at approximately 4.1 million to 5.0 million sites, including 149-182 sites with protein truncating variants [34]. The number of identi ed cancer driver genes varies between studies, ranging from 140 [35] to 299 [36]; critically, not all variants within these genes are pathogenic drivers. Mutational processes exhibit signi cant variability among tumors and cancer types. For instance, tumors with exogenous mutagen exposure such as lung adenocarcinoma and cutaneous melanoma have the highest mutation burden, while pediatric brain tumors and leukemias typically have the lowest number of mutations [37]. Unlike germline variants, somatic variants accumulate over time, enabling tumors to progress from benign to malignant lesions [35]. While the number of driver mutations varies by cancer type, most tumors harbor only 2 to 6 driver alterations [38, 35], highlighting that oncogenesis requires relatively few critical events. A central challenge is therefore to understand how driver mutations cooperatively cause tumorigenesis [39]. Even when somatic variants occur in a driver gene, their biological consequences can vary signi cantly. Variants that promote tumor growth and are implicated in cancer development are designated as driver mutations, while the remaining variants are referred to as passenger mutations [40, 41, 42]. Critically, Variants of Unknown Signi cance (VUS) represent alterations where current evidence is insu cient to classify them as pathogenic (driver) or benign (passenger), posing challenges for clinical interpretation.

Beyond mutation type, the biological consequences of protein-coding variants depend critically on the functional class of the a ected gene. During carcinogenesis, oncogenes and tumor suppressor genes regulate neoplastic cell growth through opposing mechanisms [43, 44]. Genetic variants are commonly classi ed by their func-

tional consequences, with a fundamental distinction between Loss-of-Function (LoF) and Gain-of-Function (GoF) variants. LoF variants reduce or abolish the normal activity of a gene product (e.g., through premature stop codons, frameshifts, or splice-site disruptions), often resulting in haploinsu ciency in diploid organisms or nonfunctional proteins. In contrast, GoF variants confer novel or enhanced biochemical activity to the gene product, which can dysregulate cellular pathways and drive disease.

Oncogenes drive malignancy when activated by gain-of-function variants, ampli cations, or fusions that induce constitutive signaling or novel functions. Oncogenes such as*KRAS*, *BRAF*, and *PIK3CA* are frequently activated by GoF variants that promote hyperactivity or dysregulated proliferation [35]. Conversely, tumor suppressor genes inhibit cell proliferation and maintain genomic stability. Their inactivation promotes carcinogenesis, caused by loss-of-function variants or deletions [45]. Common inactivating mechanisms include premature stop codons, frameshift mutations, or LoF variants that abrogate protein function. The tumor suppressor gene *TP53*, the most frequently mutated gene throughout all cancer types [46], orchestrates cell division, DNA repair, and apoptosis to suppress malignancy. Notably, aggressive cancers often exhibit concurrent overexpression of oncogenic proteins and inactivation of multiple tumor suppressors [47]. Driver alterations can further be classi ed by scale as sequence variants (e.g., single-nucleotide variants) or structural variants (e.g., copy-number alterations, translocations), with distinct mechanistic and clinical implications.

Sequence variants

Sequence variants represent alterations at the nucleotide level, encompassing Single Nucleotide Variant (SNV)s, insertions or deletions of individual bases or small DNA segments. Among protein-coding variants in cancer, SNVs predominate, accounting for approximately 90% of such alterations [35, 40], where a single nucleotide is substituted by another. These substitutions exhibit a spectrum of functional consequences depending on genomic context and the speci c nucleotide change. When occurring in coding regions, SNVs may be synonymous (silent) or non-synonymous (missense) [48]. Missense SNVs alter a single amino acid, potentially impairing protein function (e.g., destabilizing enzymatic activity). Stop-gain (nonsense) or stop-loss variants, which create or eliminate a stop codon, respectively, leading to truncated or elongated proteins. Additionally, SNVs at splice-site boundaries disrupt canonical splicing, causing exon skipping, intron retention, or cryptic splice-site usage, which frequently yields nonfunctional protein isoforms. Conversely, synonymous SNVs preserve the amino acid sequence but can still exert pathogenic e ects by altering mRNA stability or splicing e ciency. Beyond SNVs, Insertions and Deletions (InDel)s involve the insertion or deletion of nucleotides. Frameshift InDels, occurring when the change is not a multiple of three nucleotides, shift the translational reading frame, typically generating premature stop codons and truncated

proteins. Frameshifts are significantly rarer than in-frame events [49], paralleling their disruptive roles in tumor suppressors.

Structural variants

Structural variants, defined as genomic alterations exceeding 50 base pairs (bps) [50, 51, 52] in length, drive oncogenesis through large-scale genomic reorganization, including Copy Number Variation (CNV), inversions, translocations, and complex genomic rearrangements. CNVs are a major class of structural variants involving DNA segment duplications or deletions, hold substantial clinical significance in cancer. For instance, *ERBB2* amplification in breast cancer is a validated biomarker enabling targeted anti-ERBB2 therapies (e.g. trastuzumab) [15, 53], while deletion of the tumor suppressor *PTEN* is significantly associated with aggressive phenotypes in prostate adenocarcinomas [54, 55]. Inversions disrupt regulatory landscapes of gene structure without altering the copy number and are often classified as benign [56]. Translocations reposition chromosomal segments, creating fusion genes like BCR-ABL1 in chronic myeloid leukemia, which encode chimeric proteins with constitutive kinase activity. Gene fusions, formed by chromosomal rearrangements or abnormal splicing, generate chimeric proteins that may have oncogenic characteristics and are possible drug targets [57], including *BCR-ABL1* in chronic myeloid leukemia [58].

Beyond genetic alterations, other predictive biomarkers are now critical for guiding therapy selection, particularly for immunotherapy. These include Tumor Mutational Burden (TMB), defined as the total number of somatic mutations per megabase of DNA, which correlates with neoantigen load and response to immune checkpoint inhibitors [59]. Microsatellite Instability (MSI) status, a genomic signature indicating defective DNA mismatch repair that predicts robust response to immunotherapy across multiple cancer types [60]. Another biomarker is the use of Programmed Death-Ligand 1 (PD-L1) expression as a biomarker to identify patients most likely to benefit from immune checkpoint inhibitors, which block the PD-1/PD-L1 pathway to unleash anti-tumor immune responses [61].

1.2 Challenges of genetic variant interpretation

The precise interpretation of genetic variants presents multifaceted challenges that require a comprehensive evaluation of the biological consequences of a variant and its clinical significance as a potential biomarker for targeted therapies. This process requires a comprehensive assessment of the characteristics of a variant, including the deciphering of genomic, transcriptomic and protein associations, evaluating pathogenicity through computational prediction, computing biochemical and protein features, and cancer-specific driver evidence [62, 63]. Genetic variant interpretation systematically addresses these complexities, detailing the interplay of genomic, tran-

scriptomic, and protein-level characteristics; advances in computational pathogenicity prediction; and the translation of variant evidence into clinical actionability, enabling evidence-based selection of personalized therapies in oncology.

1.2.1 Genomic, transcript and protein associations

Accurate variant interpretation requires precise mapping of genomic alterations to functional biological units, a process fundamentally dependent on the reference genome assembly (e.g., GRCh38) and standardized nomenclature. This process begins with standardized variant representation using the Human Genome Variation Society (HGVS) nomenclature [64], the international standard for unambiguous description of DNA, RNA, and protein-level changes. HGVS syntax resolves critical ambiguities, such as distinguishing genomic (g.) from coding DNA (c.) and protein (p.) coordinates, enabling consistent cross-database queries and interoperability between clinical and research resources. However, inconsistent implementation across tools and reference sequences can propagate errors in automated pipelines, particularly for complex InDels or structural variants.

A core challenge is transcript selection: SNVs and short InDels need to be assigned to speci c transcripts to assess functional impact, as alternative splicing causes a single genomic variant to yield divergent consequences across isoforms. Since SNV callers like GATK HaplotypeCaller [65] or FreeBayes [66] output variants solely as chromosomal coordinates relative to a reference genome, assigning them to a speci c transcript is inherently ambiguous. Due to alternative splicing, a single genomic position may correspond to multiple transcripts with divergent consequences (e.g., missense in one isoform, synonymous in another). To map a genomic position to a protein position, one option is to use the gene's MANE Select transcript [67], a curated representative isoform for each human gene, as the reference for mapping a genomic position to a protein. In addition to selecting the transcript, translating a genomic position of a protein-coding gene into the position of an amino acid within the protein is a challenging task, as not only must the codons be converted, but the intron regions must also be excised, and only a few software tools are capable of performing this task [68]. For structural variants and larger indels, functional annotation is critical for clinical interpretation. Identifying a ected elements, including genes, transcripts, promoters, coding sequences (CDSs), or untranslated regions (UTRs) using resources like GENCODE [69] determines clinical severity. For instance, a deletion in a tumor suppressor gene may disrupt regulation, while one spanning exons could cause LoF. Crucially, variants a ecting non-coding regulatory elements require specialized annotation to capture disease mechanisms beyond protein-coding changes.

1.2.2 Computational variant pathogenicity prediction

Identifying driver mutations remains a particular challenge in the analysis of molecular cancer pro les. As the complex biological factors behind variant pathogenicity are still lacking and many cell-speci c factors are not yet known [10], computational methods and curated databases are an important data source for assessing the impact of variants on carcinogenesis. Population frequency databases serve as foundational lters for benign variants. Resources like dbSNP [70], Exome Aggregation Consortium (ExAC) [49], the 1000 Genomes Project Consortium [34], and Genome Aggregation Database (gnomAD) [71] catalog population allele frequencies. Studies demonstrate that high population allele frequency strongly indicates benignity for common variants [72, 73], though rare pathogenic variants may still exhibit low frequencies in speci c populations. The ClinVar database [74] provides manually curated clinical interpretations of variant pathogenicity and often serves as the ground truth dataset for evaluating machine learning-based methods for missense variant pathogenicity prediction.

Computational variant pathogenicity prediction methods leverage machine learning to assess the functional impact of genetic variants, particularly missense substitutions. AlphaMissense [75], built on the architecture of AlphaFold [76], leverages evolutionary conservation and structural context to infer amino acid substitutions. Another approach is the hypothesis that variants conserved between by humans and primates are more likely to be benign. Based on this assumption, the methods PrimateAI [77] and PrimateAI-3D [78] incorporate non-human primate sequencing data for pathogenicity prediction, with the latter using voxelized 3D protein representations to detect structural perturbations. Evolutionary Scale Modeling (ESM)-1b leverages protein Language Models (pLM) to capture inherent knowledge derived from the mutated protein sequences to assess variant pathogenicity, while SIGMA [79] explicitly integrates 3D structural features to enhance missense variant assessment. Rare Exome Variant Ensemble Learner (REVEL) [80] integrates scores from 13 individual tools, using machine learning to weigh features like evolutionary constraint and biochemical impact, enhancing accuracy for rare variants. These tools collectively address the nuanced biophysical consequences of missense changes, re ning their clinical interpretation in MTB work ows. Foundational tools like Sorting Intolerant From Tolerant (SIFT) [81] and Polymorphism Phenotyping v2 (PolyPhen2) [82] established early frameworks for functional prediction, and MutationTaster [83] incorporates disease association and splice e ects. Three-dimensional spatial relationships have been examined by HotSpot3D [84] to interpret potential functions of variants. In order to make the predictions of these numerous methods, as well as population allele frequency data, available to the public, dbNSFP [85] has become a valuable resource that provides precomputed scores of missense variants. Predicting the pathogenicity of InDels presents unique challenges due to frameshifts, splice disruptions, or premature truncations. Combined Annotation Dependent Depletion (CADD) [86] addresses this by integrating ge-

nomic features, including sequence conservation, epigenetic marks, and functional annotations|into a uni ed deleteriousness score for both SNVs and InDels.

While missense variant pathogenicity prediction is well-supported by numerous computational frameworks, structural variants, particularly CNVs and gene fusions, pose distinct challenges due to their size, genomic context-dependence, and complex mechanistic impacts. For CNVs, pathogenicity prediction hinges on the role of the a ected genes and functional regions, in particular dosage sensitivity and functional impact [87]. Advances for CNV pathogenicity prediction address this gap through integrative approaches: CNVoyant [88] applies machine learning on features such as gene constraint (gnomAD pLI scores), regulatory element overlap, and cancer-speci c recurrence patterns to score CNVs. dbCNV [89] aggregates population frequencies, functional annotations, and disease associations into a searchable database, while X-CNV [90] employs an ensemble of CNV-speci c algorithms to predict oncogenic ampli cations or tumor-suppressor deletions. Complementary tools like ClassifyCNV (which adapts ACMG/ClinGen guidelines for CNV interpretation [91]), ISV (automating SV classi cation via regulatory impact scoring [92]), and SVFX (integrating genomic, epigenomic, and evolutionary features for general structural variant pathogenicity [93]) further re ne clinical assessment.

In addition, multiple methods have been developed to predict the pathogenicity of gene fusions, including ChimerDriver [94], DEEPrior [95] and Pegasus [96]. Fusion proteins pose a particular challenge for predicting pathogenicity, since the consequences of in-frame fusions can vary for each individual fusion, due to the potential di erences in breakpoints within the genes.

Beyond binary pathogenicity classi cation, emerging methods predict speci c functional consequences such as LoF or GoF e ects. LoGoFunc [97] and 3DHotspots [98] identify activating/deactivating variants, a database for LoF and GoF variants has been established with GoFCards [99]. Genes tolerant to high-con dence LoF variants exhibit reduced evolutionary conservation, including in promoter regions [100]. To quantify gene-level vulnerability to disruptive mutations, constraint metrics estimate intolerance to LoF variants. The probability of Loss-of-Function intolerance (pLI) score measures intolerance to protein-truncating variants [49]. The Loss-of-Function Observed/Expected Upper bound Fraction (LOEUF) score provides a continuous metric of LoF constraint, with lower values indicating stronger intolerance [71].

1.2.3 Biochemical and protein structural characteristics of variants

Despite signi cant advancements in experimental and computational assessment of genetic variants, the biological consequences of a substantial proportion of variants remain unknown. Another step toward illuminating the biological consequences of

protein-coding variants is therefore to analyze the biochemical and structural characteristics of variants and the mutated protein [101]. Missense variants have been shown to a ect protein stability and dynamics, as well as structural features such as hydrogen bonds and activity [48]. Additionally, signi cant di erences have been identi ed between the e ects on protein structure between LoF and GoF mutations [102].

Charge alterations are particularly consequential, as they disrupt electrostatic interactions essential for enzymatic activity, ligand binding, or allosteric regulation. Aromatic amino acids, such as phenylalanine (F), tyrosine (Y) and tryptophan (W), are crucial in protein structure and function due to their hydrophobic and stacking interactions. The aromaticity of a protein, de ned as the relative frequency of aromatic amino acids [103], is calculated as

$$A = \sum_{i=1}^{20} {}_i f_i \tag{1.1}$$

where f_i is the relative frequency of amino acid of kind i in the protein and ${}_i = 1$ for aromatic amino acids (F,Y,W). The Grand Average of Hydropathy (GRAVY) scale [104] quanti es the overall hydrophobicity or hydrophilicity of a protein sequence. Hydropathy indices, empirically derived from transfer-free energy measurements, assign numerical values to each amino acid (e.g., +4.5 for isoleucine, -4.5 for arginine), where positive values denote hydrophobicity and negative values indicate hydrophilicity. The GRAVY score is computed as the arithmetic mean of these indices across the entire sequence. For a protein sequence with n amino acids, the GRAVY score is de ned as:

$$\text{GRAVY} = \frac{1}{n} \sum_{i=1}^{n} H(a_i) \tag{1.2}$$

where ${}_i$ represents the ith amino acid in the sequence and $H({}_i)$ is its standardized hydropathy index (e.g. Kyte-Doolittle values). Biochemical compatibility of substitutions is assessed using the BLOcks SUbstitution Matrix (BLOSUM) matrix [105], which scores log-odds ratios of observed vs. expected substitutions in conserved protein blocks.

Changes in the spatial protein structure and folding caused by genetic variants can disrupt critical function sites, including active sites and binding sites, abolish catalytic activity, or causing complete loss of function. Variants inactivating tumor suppressors in cancer have been shown to primarily occur in the core of the protein, while the activation of oncogenes is often associated with variants at functional sites [106]. This nding indicates that the position of a variant within the folded protein structure and its functional elements has a substantial in uence on the e ect of the protein. A key metric in the analysis of protein structure is the surface accessibility at the mutated site of the protein. The Relative accessible Surface Area (RSA)

[107] of a position quanti es residue exposure to solvent or binding partners and is calculated as the ratio of its Accessible Surface Area (ASA) in the folded structure to the Maximum possible solvent Accessible Surface Area (MaxASA) observed for that residue in an extended tripeptide conformation:

$$RSA = \frac{ASA}{MaxASA}$$

Another metrics are the interatomic distances relative to the mutated site. The alpha carbon atom distances (C) are calculated as the spatial distance of each amino acid's alpha carbon atom to the variant's alpha carbon atom:

$$d_{C_i} = \sqrt{(x_i - x_m)^2 + (y_i - y_m)^2 + (z_i - z_m)^2}$$

Domain annotation identi es variants within functional units (e.g., kinase domains), which may disrupt autoinhibition or substrate recognition. Secondary structure context is equally critical: Variants in -helices and -sheets potentially destabilize entire folds, such as mutations in rigid helices or turns.

Beyond global topology, variants disrupt critical atomic interactions that maintain structural delity. Substitutions of di erentially charged amino acids may generate salt bridge rearrangements between a positively charged residue (LYS (Lysine), ARG (arginine)) and a negatively charged residue (ASP (Aspartate), GLU (Glutamate)) if spatially close to each other, which may comprise electrostatic stabilization. To calculate potential new salt bridges in case of an amino acid substitution, the residue centroids are computed as

$$C_R = \frac{1}{n}\sum_{i=1}^{n} a_i$$

Then the Euclidean distance between the target and candidate residue is calculated as

$$d = ||C_{target} - C_{candidate}|| = \sqrt{\sum_{i=1}^{3} (C_{target;i} - C_{candidate;i})^2}$$

A potential salt bridge is typically identi ed if the Euclidean distance d is within the range of a prede ned distance cuto (e.g. 0.4 A). Potential disul de bonds, forming bonds between cysteine residues critical for protein stability, are identi- ed if the Euclidean distance between two residue's cysteine atoms is smaller than a certain radius. Hydrogen bond perturbations alter folding energy landscapes, potentially causing misfolding. Binding site prediction, e.g. via ScanNet [108], identi es variants disrupting molecular interactions. Functional context is enriched via Gene Ontology (GO) terms [109, 110], standardizing annotations of molecular functions (e.g. "ATP binding"), biological processes (e.g. "cell cycle regulation"), and cellular components. Another approach for predicting variant e ects is protein stability

analysis [111].

Another possible approach for predicting the e ects of protein-coding variants is to predict the e ects on protein stability [111]. Previous studies indicate that protein-destabilizing variants are enriched among disease-associated mutations [112], and about 60% of missense variants have been associated with the reduction of protein stability [113, 114]. Thermodynamic models predicting the energetic e ects of mutations have demonstrated strong correlation with experimental stability measurements [115]. Thermodynamic models predict stability changes via the change in folding free energy (G) [116, 117]:

$$\Delta G = G_{mut}^{folding} - G_{wt}^{folding} \tag{1.3}$$

where $G^{folding}$ is the Gibbs free energy change for the folded unfolded transition. Critically, $\Delta G > 0$ indicates destabilization, while $\Delta G < 0$ indicates protein stabilization. Computational prediction of G, e.g. using FoldX [118], thus provides a mechanistically grounded metric for variant pathogenicity assessment, complementing sequence-based predictors.

The structural analysis of protein structures relies on high-quality 3D models. With the release of AlphaFold DB [119], researchers now have ubiquitous access to high-con dence structural models even for proteins lacking experimental characterization. A potential disadvantage of utilizing AlphaFold-predicted protein structure is that, as with any computational prediction, there is a possibility of inaccuracy. Central to AlphaFold's reliability is the predicted Local Distance Di erence Test (pLDDT) score that quanti es per-residue con dence of the model. The pLDDT is a metric between 0 and 100, where higher values indicate greater con dence in the local structural accuracy around a residue. Scores ≥ 90 indicate high-accuracy backbone predictions, while values < 50 suggest intrinsically disordered regions or modeling uncertainty.

1.2.4 Clinical implications

The clinical translation of genomic variants relies on their validation as biomarkers, objective biological indicators that guide cancer diagnosis, prognosis, or therapeutic decisions. The clinical translation of genomic variants into actionable therapeutic strategies relies on curated knowledge bases that aggregate evidence on biomarker signi cance, drug associations, and molecular mechanisms. OncoKB [120] catalogs expert-curated therapeutic implications of somatic variants and provides its own tiered classi cation system of alterations. The Clinical Interpretation of Variants in Cancer (CIViC) database [121] employs a community-driven model to aggregate diagnostic, prognostic, and therapeutic annotations with detailed literature provenance. Emerging meta databases like MetaKB [122] harmonize multiple repositories to resolve con icts and unify variant interpretations.

In precision oncology, clinical guidelines for variant assessment can be broadly categorized into two distinct types: those that evaluate the pathogenicity or oncogenicity of variants (biological impact), and those that assess the clinical signi cance or actionability of variants for therapeutic decision-making (clinical relevance). The rst category comprises guidelines focused on variant classi cation based on biological evidence. The American College of Medical Genetics and Genomics (ACMG) guidelines [123] represent the gold standard for germline variant pathogenicity classi cation, establishing criteria to categorize variants as pathogenic, likely pathogenic, variant of uncertain signi cance, likely benign, or benign.

For somatic variants, the ClinGen Somatic Cancer Working Group, in collaboration with the Cancer Genomics Consortium (CGC) and Variant Interpretation for Cancer Consortium (VICC), developed a comprehensive standard operating procedure for somatic variant oncogenicity classi cation [124]. This evidence-based framework categorizes somatic variants into ve classi cations (oncogenic, likely oncogenic, VUS, likely benign, and benign) using a point system that evaluates evidence strength (Very Strong, Strong, Moderate, or Supporting) across multiple data types including functional studies, population frequency, computational predictions, and cancer-speci c databases.

The second category encompasses frameworks that translate biological evidence into clinical actionability. The Association for Molecular Pathology (AMP)/American Society of Clinical Oncology (ASCO)/College of American Pathologists (CAP) guidelines [125] classify variants into tiers based on clinical evidence strength: Tier I (variants with strong clinical signi cance, e.g., FDA-recognized biomarkers), Tier II (potential clinical signi cance), Tier III (unknown signi cance), and Tier IV (benign). Similarly, the European Society for Medical Oncology (ESMO)/ESMO Scale for Clinical Actionability of molecular Targets (ESCAT) scale [126] ranks biomarkers from Tier I (highest evidence, e.g., biomarkers with survival bene t in prospective trials) to Tier V (preclinical evidence only). These frameworks weigh factors like study design, cancer-type speci city, and clinical outcomes (e.g., overall survival vs. response rate). The NCT/DKFZ MASTER program [127] exempli es integrated variant assessment by decoupling biological and clinical evaluation. First, somatic variant classi cation follows the ClinGen/CGC/VICC SOP to determine oncogenicity. Then, for variants deemed oncogenic or likely oncogenic, clinical actionability is assessed using the NCT's proprietary evidence grading system. In addition, clinical evidence is supported by scores like the Genetic Priority Score (GPS) [128] that integrates human genetic features to prioritize drug target genes based on their likelihood of clinical success. While standard cancer therapy guidelines, including National Comprehensive Cancer Network (NCCN) and the ESMO Clinical Practice Guidelines, synthesize evidence to recommend speci c treatments for de ned patient groups or cancer types, precision oncology variant interpretation guidelines provide standardized frameworks speci cally for classifying the clinical signi cance of individual genomic variants identi ed through molecular testing.

In order to select e ective therapies, multiple other databases provide data for analyzing drugs and their mechanisms of action. The Therapeutic Target Database (TTD) [129] and Open Targets [130] map therapeutic targets to disease pathways and drug development stages. The DGIdb [131] aggregates drug-gene interaction evidence to identify druggable targets, while pharmacological resources (DrugBank [132], PubChem [133], ChEMBL [134]) provide chemical structures, and bioactivity data. PharmGKB [135] uniquely curates clinically actionable gene-drug relationships, including pharmacogenomic biomarkers, dosing guidelines, and toxicity risks based on metabolizer phenotypes.

1.3 Structural variant analysis

While SNVs and small InDels present well-documented interpretation hurdles, CNVs represent a particularly challenging class of genomic alteration due to their size, complexity, and potential to disrupt multiple genes or regulatory elements simultaneously. Beyond clinical variant interpretation, structural variant analysis is indispensable in preclinical cancer research for identifying tumor-speci c biomarkers and therapeutic targets. Among the potential consequences of CNVs is the alteration of gene expression critical for cancer hallmarks. Ampli cations of oncogenes can lead to overexpression (e.g ERBB2/HER2 in breast and colorectal cancer [136]), while deletions of tumor suppressors (e.g. CDKN2A in glioblastoma) disrupt cell-cycle control. Moreover, copy number events exhibit signi cant di erences across tumor types, with numerous genes showing context-dependent functionality as either tumor suppressor genes or oncogenes [137].

To navigate the complexities of genetic variant heterogeneity, curated repositories provide data sources of genetic variation, such as specialized databases cataloging somatic sequence and structural variants. Based on the classi cation of [138], databases for CNV interpretation can be divided in three main categories: First, databases containing genetic and phenotypic data on individuals. Second, databases containing information about population studies of healthy people. And third, there are datasets customly created by laboratories. While repositories like The Cancer Genome Atlas (TCGA) [139] and DepMap [140] provide raw calls of somatic CNVs, repositories for germline CNVs include dbVAR [52], DGV [141], and gnomAD [71]. These resources address the critical need for standardized variant representation, as the complex nature of CNVs makes them less amenable to automated analysis compared to other genomic variants [142]. By aggregating population frequency data and clinical annotations across diverse cohorts, CNV databases enable reliable distinction between pathogenic variants and benign polymorphisms, essential for both research interpretation and clinical decision-making.

Although CNVs were found to a ect primarily non-coding sequences, their pathogenicity often manifests through a ecting dosage-sensitive genes [87, 143]. This sensitivity

arises from two key phenomena:

- Haploinsu ciency (HI): Loss of one functional allele reduces gene dosage below a critical threshold, causing numerous genes to be unable to maintain normal gene function [144]. HI is a hallmark of tumor suppressor genes, e.g. TP53 and PTEN, where even partial loss disrupts tumor-suppressive activity. In these cases, CNV deletions have the same e ect as LoF variants [145]. However, not all genes exhibit HI, and approximately 100 genes have been identi ed that can be completely deleted (nullizygosity) without observable phenotypic consequences [51].
- Triplosensitivity (TS): Gene duplication elevates dosage beyond physiological limits, potentially causing oncogene overexpression.

ClinGen's HI/TS scores [146] provide standardized metrics for gene-level dosage sensitivity, where high scores indicate strong evidence for pathogenicity upon deletion (HI) or duplication (TS).

A notable challenge when analyzing CNVs or structural variants is that vary across samples, necessitating aggregation of events a ecting the same genomic region for cohort-level analysis. CNV analysis therefore begins with segmentation algorithms that partition genomic pro les into regions of constant copy number. The segmentation mean represents the average $\log_2$-transformed copy number ratio of a genomic segment identi ed by segmentation algorithms, quantifying the relative copy number state (e.g. deletion, neutral, ampli cation) compared to a reference sample (typically diploid) [147]:

$$S = \frac{1}{n} \sum_{i=1}^{n} \log_2 \left(\frac{R_i}{R_{ref;i}} \right)$$

Distinguishing driver from passenger CNVs requires robust statistical approaches. Among the conceptual frameworks for identifying signi cant regions of copy number alteration, the Genomic Identi cation of Signi cant Targets in Cancer (GISTIC) algorithm [148, 149] addresses this by identifying genomic regions with signi cantly recurrent copy number alterations, after correcting for background mutation rates and chromosomal arm-level events. Unlike simple frequency-based methods, GISTIC weights alterations by amplitude and spatial clustering to prioritize focal driver events over broad passenger changes, a key advancement in the eld.

One method to identify regions that are more or less a ected by copy number variation is to calculate the CNV frequency per segment, de ned as the number of CNVs overlapping with a genomic region, divided by the total number of samples. In addition to frequency, the CNV burden refers to the total genomic footprint of CNVs in an individual, typically measured as the cumulative length (in megabases) or number of events. This metric correlates with genomic instability and has prognostic value in cancers like ovarian carcinoma.

1.4 Functional validation for variant interpretation and therapeutic vulnerability pro ling

While accurate variant interpretation forms the foundation of precision oncology, translating genomic ndings into e ective therapeutic strategies requires functional validation that bridges molecular observations with clinical outcomes. Beyond the identi cation of genetic variants, complementary molecular layers, particularly transcriptomic pro ling and functional genomic screens, provide critical context for assessing phenotypic impact and therapeutic relevance. Two areas in particular o er high potential here: The analysis of multi-omics data and the integration of experimental clinical data, such as drug screening data, with molecular pro le data.

Gene expression analysis reveals the functional consequences of genomic alterations by quantifying active molecular pathways that directly in uence disease phenotypes and therapeutic responses. Public repositories like the Gene Expression Omnibus (GEO) [150] and TCGA provide vast compendia of expression data across diverse conditions, enabling robust comparative analyses and biomarker discovery. Critically, the Genotype-Tissue Expression (GTEx) [151] project serves as an essential reference for baseline tissue-speci c expression patterns in normal physiology. This layer of functional genomics bridges the gap between static genetic variants and their phenotypic manifestations, re ning patient strati cation and target identi cation in precision oncology.

Complementing transcriptomic approaches, Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-based functional genomic screens represent powerful tools for systematically identifying genes essential for speci c cellular phenotypes. These high-throughput approaches enable unbiased discovery of therapeutic targets by modeling gene essentiality through explicit mechanistic representations of cell population dynamics following gene perturbation [152, 153]. The integration of such functional genomic data with variant pro les allows researchers to distinguish passenger from driver alterations and prioritize targets with demonstrated biological signi cance.

While genomic and transcriptomic analyses identify potential therapeutic targets, translating these discoveries into e ective treatments requires functional validation through drug response pro ling. High-throughput drug screens provide the critical link between molecular aberrations and therapeutic vulnerability. The quantitative assessment of drug e cacy in cell-based assays is most commonly performed by tting dose{response to extract metrics such as the Half-maximal inhibitory concentration (IC_{50}), the concentration yielding 50% endpoint inhibition, and related Growth Rate (GR) metrics. The IC_{50} value is calculated as the concentration C obtained by tting a sigmoidal dose-response curve to inhibition values $I(C)$ relative

to an initial density N_0:

$$I(C_i) = \frac{N_{ctrl}(t) - N(C_i;t)}{N_{ctrl}(t) - N_0}: \tag{1.4}$$

Due to the wide dynamic range of IC_{50} values, potency is reported as

$$pIC_{50} = -\log_{10} IC_{50}\,[M] \tag{1.5}$$

which linearizes comparisons and improves statistical robustness. Critically, IC_{50} reflects endpoint signal reduction but ignores intrinsic proliferation kinetics. For instance, slow-growing cell lines may exhibit artificially low IC_{50} values even with weak cytotoxicity, as their baseline proliferation minimally offsets drug-induced death. To decouple drug effects from proliferation artifacts, the GR normalizes responses to exponential growth dynamics.

Reporting both IC_{50} and GR-based metrics (e.g., GR_{50}, GR_{max}) provides a comprehensive efficacy profile: IC_{50} reflects endpoint potency in a specific assay context, while GR metrics reveal intrinsic drug action independent of proliferation artifacts. This dual-parameter approach is essential for cross-cell-line comparisons and high-throughput profiling, where uncorrected IC_{50} values may misrepresent true drug potency.

The integration of these diverse data types|genomic variants, transcriptomic profiles, functional genomic screens, and drug response data|presents significant computational challenges. Aggregating information across these heterogeneous data sources manually requires considerable time and expertise, creating bottlenecks in clinical decision-making. The fragmentation of bioinformatics tools across isolated analytical steps further complicates reproducible analysis, as researchers must manually chain tools across incompatible data formats and implement custom filtering criteria.

1.5 Computational solutions for genetic variant interpretation

The emergence of numerous new databases and computational methods that elucidate the pathogenicity and functional impact of genetic variants opens significant opportunities for MTBs and biomedical research to interpret patient molecular profiles and select effective personalized therapies. However, aggregating and synthesizing this diverse, multi-source information manually, and evaluating individual cases with biological and clinical context, requires time-intensive efforts and highly trained personnel. To address this bottleneck, robust computational foundations have been developed. In precision oncology and broader clinical genomics, Clinical Decision Support System (CDSS)s bridge the gap between biomarker discovery and clinical

application by integrating evidence from literature, databases, and guidelines to prioritize actionable variants. Concurrently, in biomedical research, scalable work ow frameworks automate complex, multi-step analyses|enabling reproducible processing of genomic, transcriptomic, and proteomic data. Additionally, machine-learning based methods enhance biological knowledge. Together, these tools accelerate the translation of genomic insights into actionable clinical strategies while reducing human error.

1.5.1 Clinical decision support systems

CDSS have evolved to improve healthcare delivery by supporting various clinical tasks, including diagnostics, patient decisions, clinical management, and administrative functions [154, 155]. For MTBs, CDSS serve as critical tools that automate the aggregation of evidence from various sources, including genomic databases, biomedical literature, and clinical guidelines, streamline the interpretation of genomic and multi-omics data, and contribute to improve therapy recommendations and streamline MTB case preparation [33].

Web-based interfaces have become pivotal in modern CDSS, o ering accessible, scalable platforms that bridge genomic research with clinical practice. The cBio Cancer Genomics Portal (cBioPortal) [156] provides cohort-level analysis of genetic pro les with survival correlations and supports MTB work ows through extension like MIRACUM-Pipe [157], PM4Onco [158] and the documentation of therapy recommendations [159]. The Medical Informatics in Research and Care in University Medicine (MIRACUM) consortium [160] leverages web-based infrastructure to integrate heterogeneous Electronic Health Record (EHR) data across institutions, facilitating real-time CDSS for precision oncology in routine care. Similarly, MTB Portal [161] facilitates secure clinical data exchange and therapy matching. For variant interpretation, SNPnexus [162] specializes in functional annotation of both coding and non-coding variants, while CancerVar [163] integrates clinical evidence with an own deep-learning-based variant pathogenicity prediction method. MTP-Pilot [164] speci cally focuses on prioritizing and annotating clinically actionable biomarkers. In clinical practice, commercial platforms like QIAGEN Clinical Insight (QCI), SOPHIA DDM, and VarSeq leverage proprietary annotation pipelines as rule-based engines to process NGS data intro structured reports that comply with regulatory standards.

A critical implementation challenge for CDSSs is ensuring clinical validity and utility. Tools must undergo rigorous analytical and clinical validation, and their recommendations should align with consensus guidelines. A crucial question for CDSS implementation involving the interpretation of genetic variants is variant annotation methodology. Given the heterogeneity of the biological and clinical aspects of variants, the number of di erent databases, and scale, enriching genetic data with contextual information represents a major hurdle. CDSS backbones may thus

utilize existing annotation pipelines or proprietary solutions. In research settings, widely adopted pipelines include ANNOVAR [165], SnpE /SnpSift [166] and the Ensembl Variant E ect Predictor (VEP) [167], which annotate functional variant consequences. To lter variants, command line-based tools like SnpSift [168] and VCFtools [169] allow the manual generation of ltering pipelines. This process is facilitated by Graphical User Interface (GUI)-based tools that allow interactive l-tering [170, 171] and exploration [172, 173] of variant data.

A key limitation of these annotation pipelines is that they usually have to be installed as a complete package, which often restricts modular customization of annotation sources or algorithms. One method for simplifying the annotation process is to uti-lize web services that o er annotation functionality through Application Program-ming Interface (API) endpoints. This architectural concept has been implemented by VarSome [174], providing an API endpoint for the annotation of variants. Plat-forms focusing primarily on the biological interpretation of genetic variants include wANNOVAR [175], REEV [176], VarCards2 [177], FASMIC [178] and OpenCravat [179].

Another challenge is genomic data format heterogeneity, which demands distinct software tools and introduces interoperability barriers. Raw sequencing data (FASTQ) requires alignment tools to generate Sequence Alignment Map (SAM) les, con-verted to compressed Binary Alignment Map (BAM) les. Variant calls stored in Variant Call Format (VCF) format [169] provide a structured and exible format (CHR, POS, ID, REF, ALT, QUAL, FILTER, and INFO), but annotation scala-bility is limited as all metadata must reside in the INFO column. The Mutation Annotation Format (MAF) consolidates annotations into separate columns for sim-plicity, yet introduces new challenges due to non-standardized column de nitions and inconsistent eld usage across studies. Structural variants are commonly stored in Browser Extensible Data (BED) format, which speci es genomic intervals but lacks standardized schema for variant attributes, complicating automated process-ing. For functional context, General Feature Format (GFF) les describe genomic features (e.g., genes, regulatory elements), though inconsistent feature hierarchies across databases hinder integration.

A fundamental issue in genomic analysis arises when datasets are aligned to dif-ferent reference genome assemblies. A reference genome serves as a standardized template representing a species' DNA sequence, enabling variant detection by align-ing sequenced DNA fragments to this template and identifying mismatches. Key human references include GRCh37, its updated successor GRCh38 [180] with im-proved accuracy and centromere representation, and the Telomere-to-Telomere as-sembly (T2T-CHM13) [181], which resolves gaps and complex regions absent in prior versions. Genomic variants are de ned by their chromosomal coordinates within a speci c assembly. When comparing or integrating datasets based on di erent references, direct analysis is impossible because of coordinate systems divergence. LiftOver solves this problem by computationally converting genomic positions from

a source assembly (e.g. GRCh37) to their equivalent locations in a target assembly (e.g. GRCh38), enabling cross-assembly data integration.

1.5.2 Automated work ows

While standalone bioinformatics tools and web applications excel at isolated tasks, such as sequence alignment or variant calling, most have project-speci c features and functionality [11]. Critically, these tools face limitations in addressing the systemic challenge of orchestrating multi-step, interdependent analyses required in modern biomedical research, where reproducibility, scalability, and integration of heterogeneous data are paramount. Considering the processing of molecular clinical data for a MTB, the whole processing pipeline including raw sequencing data must undergo quality control, alignment, variant calling, functional annotation, and clinical interpretation, a cascade of interdependent steps where manual execution introduces error, ine ciency, and irreproducibility. Such pipelines exemplify the limitations of ad hoc tool combinations, particularly when scaling to population-level cohorts or integrating multi-omics data (e.g., genomics, proteomics, and imaging).

Bioinformatic work ow management systems address these challenges by automating, standardizing, and scaling complex analyses across genomics, proteomics, and other life science domains. They enforce reproducibility through version-controlled pipelines, manage computational dependencies, and optimize resource utilization for large-scale data. Key frameworks include Snakemake [182] and Next ow [183], which use declarative, domain-speci c languages to de ne portable, scalable pipelines with native support for parallelization, fault tolerance, and containerization (e.g., Docker [184], Singularity [185]). For executable documentation and interactive analysis, the Jupyter ecosystem [186] enables browser-based notebooks that integrate code, visualizations, and narrative text, facilitating transparent and shareable research. GUI-driven, cloud-based platforms like Galaxy [187] provide accessible, web-based environments for visual work ow construction via drag-and-drop interfaces, lowering barriers for non-programmers. Meanwhile, Konstanz Information Miner (KNIME) [188] o ers a local, open-source platform with modular, visual work ow design that extends into diverse domains through nodes for data integration, machine learning, and visualization, though its bioinformatics adoption is often augmented with specialized extensions.

1.5.3 Deep learning architectures for variant pathogenicity prediction

The application of Arti cial Intelligence (AI) models is increasingly pivotal in precision medicine, enabling tasks such as preclinical diagnostics, disease activity forecasting, and clinical decision support [189, 190] while o ering signi cant potential for improving MTB work ows [191].

This is acutely evident in pathogenicity prediction for missense variants, where AlphaMissense and REVEL, among others, have demonstrated how deep learning architectures transcend conventional feature engineering by learning latent biological constraints directly from sequence and evolutionary data. These advances reveal a critical paradigm: Rather than relying on hand-curated rules or linear combinations of isolated features, the most impactful models now leverage models that implicitly encode the functional grammar of proteins.

This is particularly transformative for predicting the pathogenicity of missense variants, a critical challenge given the high number of variants of uncertain signi cance, hindering clinical interpretation. Here, AlphaMissense and PrimateAI-3D exemplify how deep learning architectures enhance conventional approaches by jointly learning latent biological constraints from sequence, evolutionary, and structural data. Critically, these advances revealed that the most e ective models leverage architectures that implicitly encode multiple sequence alignments and protein functional grammar.

Central to this shift are protein language models, which acquire remarkable capacity to model probable biophysical and structural consequences of amino acid substitutions through self-supervised pretraining on vast evolutionary sequence corpora. Unlike earlier statistical or machine learning methods, pLMs e ectively capture statistical patterns re ecting long-range epistatic dependencies and context-sensitive mutational e ects that are di cult to model with traditional approaches. Key models like ESM-2 [192] and ProtTrans [193] process amino acid sequences analogously to how text-based Large Language Model (LLM)s handle language, treating sequences as "sentences" and amino acids as tokens, learning contextual representations by predicting masked residues during pretraining. This yields rich protein embeddings that encode evolutionary, structural, and functional information, which can be extracted and ne-tuned for diverse tasks|including pathogenicity prediction, protein structure/function inference, and design.

Robust evaluation is essential for clinical translation. For classi er performance, Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) are among the most widely used metrics. The ROC curve visualizes the trade-o between true positive rate (sensitivity) and false positive rate (1-speci city) across classi cation thresholds, while the AUC quanti es the model's overall discriminative ability, with higher values (closer to 1.0) indicating better performance at ranking positive instances higher than negatives [194]. The AUC [195] is computed as

$$\text{AUC} = \frac{1}{n_+ \, n} \sum_{i=1}^{n_+} \sum_{j=1}^{n} 1\left(p_i > p_j\right)$$

where n_+ and n denote the number of positive and negative instances, p_i is the prediction score for the i-th positive instance, and p_j for the j-th negative instance.

Despite strong predictive performance, clinical adoption requires not only accuracy but also interpretability to build trust and align with biological mechanisms. SHapley Additive exPlanations (SHAP) [196] values provide a robust, game theory-based framework for interpreting individual predictions by quantifying each input feature's contribution relative to a background dataset. Complementarily, attention mechanisms o er dynamic insight into feature weighting during prediction. Additive attention [197] computes feature importance via a learnable alignment function:

While performance metrics con rm a model's discriminative power in pathogenicity prediction, clinical adoption demands interpretable decision logic that aligns with biological mechanisms. To bridge this gap, attention mechanisms provide critical visibility into how models weigh genomic features during prediction. Additive attention [197] computes feature importance through a learnable alignment function:

$$_j = \text{softmax}\quad v^{>} \tanh\left(Wh_j + b\right) \tag{1.6}$$

where h_j represents the embedding of the j-th input feature (e.g. an amino acid position). This mechanism dynamically highlights biologically relevant functional units by assigning higher weights to features that collectively in uence pathogenicity predictions. Scaled dot-product attention [198] generalized this concept for parallel sequence processing, but additive attention remains optimal for structured genomic feature interpretation due to its explicit feature-scoring capability.

1.6 Thesis Outline

1.6.1 Research aims

The primary research aim of this thesis was to conceptualize and implement a new CDSS to support cancer variant interpretation in MTBs. This should address key challenges in the MTB work ow, including the need for rapid annotation and l-tering of WGS data to pre-identify potential biomarkers. The main focus should then be on a comprehensive framework that provides detailed biological and clinical interpretation of the potential biomarkers identi ed in the rst step and o ers direct treatment recommendations for targeted therapy for a patient. Furthermore, given that structural variant interpretation, particularly of CNVs, poses signi cant challenges in precision medicine, a comprehensive pan-cancer database should be established alongside dedicated analysis capabilities.

The second research objective was to develop computational solutions that should extend the functionalities for genetic variant interpretation, and address the critical need for reproducible, scalable outcomes in biomedical research. In addition, they should integrate additional omics and experimental data, such as transcriptomics and experimental drug screen data, enabling expansion and validation of variant interpretation results, as well as providing preclinical research analysis tools.

Figure 1.2: Thesis structure This thesis develops an integrated suite of computational frameworks and tools to advance precision oncology and biomedical research, with a focus on cancer variant interpretation: A framework for annotation and ltering of WES/WGS data (AdaGenes), a framework for clinical variant interpretation (MTB-Report), a framework for comprehensive biological and clinical interpretation of genetic alterations (Onkopus). A tool for the harmonization and analysis of genetic data (SeqCAT) and a novel method for gene fusion pathogenicity prediction (FusionPath) support biomedical research and the variant interpretation tools. To further support the analysis of structural variants, a novel copy number variant database and analysis framework (PanCNV-Explorer) supports preclinical and foundational research in biomarker discovery and pan-cancer databases. In addition, platform-independent functionalities for a multitude of tasks in precision medicine, including multi-omics and drug screen analyses, are provided by a suite of automated work ows (AstraKit).

Finally, the third research objective focused on developing novel computational methods and tools to enhance the frameworks established in the rst two aims, thereby advancing broader biomedical research capabilities.

The following solutions were developed based on these objectives:

1. Development of new computational solutions for the interpretation of somatic cancer variants, including
    a) Robust annotation and ltration of WES/WGS sequencing data
    b) The comprehensive biological and clinical interpretation of SNVs and small InDels
    c) Analysis and interpretation of structural variants, in particular CNVs, accompanied by an annotated novel database of healthy and tumor CNVs
2. Development of automated work ows for diverse biomedical research tasks, including genetic variant interpretation, multi-omics integration, and drug screen analyses
3. Development of new computational methods and tools to support precision medicine, including the
    a) Development of a tool for sequence conversion and analysis of variants at the genome, transcript and protein level
    b) Development of a new computational method for the pathogenicity prediction of fusion proteins

Collectively, these aims address a critical gap in clinical genomics: the lack of adaptable, transparent tools that empower MTB teams to interpret complex genomic data without compromising data sovereignty or work ow exibility. By prioritizing clinician-centric design and open interoperability, this work provides a foundation for evidence-based, patient-speci c decision support that evolves alongside precision oncology standards.

1.6.2 Organization of this thesis

This thesis advances precision oncology through seven integrated computational contributions, each developed to address speci c gaps in cancer variant interpretation and presented in dedicated chapters below. The thesis comprises three computational frameworks for providing cancer variant interpretation support for precision medicine and preclinical research, including the comprehensive CDSS Onkopus (Chapters 2-5). To extend this functionality for reproducibility and customization, a set of work ows was developed (Chapter 6). As methods and tools to support biomedical research, a toolkit for versatile processing of genetic data and a method for predicting the pathogenicity of gene fusions were developed (Chapters 7-8).

1. Chapter 2: High-throughput variant annotation and filtering — Introduces AdaGenes, a streaming processor for annotation and filtering WES/WGS data.
2. Chapter 3: Clinical variant interpretation — Presents MTB-Report, a framework implementing ACMG/AMP-inspired rules for somatic variant classification.
3. Chapter 4: Comprehensive biological and clinical variant interpretation — Details Onkopus, a modular CDSS for unified analysis of SNVs, InDels, and structural variants with protein-level impact assessment.
4. Chapter 5: Pan-cancer CNV database and interpretation — Introduces PanCNV-Explorer, an annotated resource of germline and cancerous CNV profiles.
5. Chapter 6: Reproducible, customizable workflows for biomedical research — Describes AstraKit, a suite of adaptable workflows integrating a multitude of different tasks in precision medicine, including genomic, transcriptomic, and drug sensitivity data.
6. Chapter 7: Sequence conversion and analysis — Proposes SeqCAT, a toolkit for processing and analyzing genetic data across genomic, transcriptomic, and protein notation levels.
7. Chapter 8: Gene fusion pathogenicity prediction — Develops FusionPath, a deep learning method for prioritizing oncogenic fusion events.

Chapter 9 synthesizes these contributions, discusses clinical implementation challenges, and outlines future directions for precision oncology decision support. Collectively, these contributions significantly advance the interpretation and analysis of molecular cancer data, providing both innovative methodologies and tangible clinical benefits for molecular tumor boards.

2 High-throughput processing of sequence variant data

Reference

Kurz, N. S., Drofenik, K., Kornrumpf, K., Reuter-Jessen, K., & Donitz, J. (2025). AdaGenes: A streaming processor for high-throughput annotation and filtering of sequence variant data. bioRxiv. doi: 10.1101/2025.06.17.659929

Original Contribution

Nadine Kurz and Jurgen Donitz conceptualized the architecture. Nadine Kurz implemented the database (INF08) and the web front end (INF07). Nadine Kurz and Klara Drofenik implemented the AdaGenes Server (INF09). Nadine Kurz set up and maintained the publicly available instance of AdaGenes. Kirsten Reuter-Jessen provided an evaluation of medical use cases.
Nadine Kurz wrote the manuscript. Nadine Kurz created the figures. All authors proofread the manuscript.

3 Clinical interpretation of genetic variants

Reference

Kurz, N. S. , Perera-Bel, J. , Holtermann, C., Tucholski, T., Yang, J., Beissbarth, T., & Donitz, J. (2022). Identifying Actionable Variants in Cancer-The Dual Web and Batch Processing Tool MTB-Report. Studies in health technology and informatics, 296, 73-80. doi: 10.3233/shti220806

Equal contribution

This is a pre-copyedited, author-produced PDF of an article accepted for publication in Studies in Health Technology and Informatics following peer review.

Original Contribution

Nadine Kurz and Jurgen-Donitz conceptualized the methodology. Nadine Kurz, Charlotte Holtermann, Julia Perera-Bel, Tim Tucholski and Jingyu Yang implemented the application (INF01, see App. A). Julia Perera-Bel and Tim Bei barth developed the variant classi cation methodology [199] utilized by MTB-Report. Nadine Kurz, Jurgen Donitz and Charlotte Holtermann wrote the manuscript. Nadine Kurz created the gures. All authors proofread the manuscript. Nadine Kurz revised the manuscript.

4 A computational framework for comprehensive variant interpretation

Reference

Kurz, N. S., Kornrumpf, K., Tucholski, T., Drofenik, K., König, A., Beißbarth, T., & Dönitz, J. (2025). Onkopus: precise interpretation and prioritization of sequence variants for biomedical research and precision medicine. Nucleic Acids Research, gkaf376. doi: 10.1093/nar/gkaf376

This is a pre-copyedited, author-produced PDF of an article accepted for publication in Nucleic Acids Research following peer review.

Original Contribution

Nadine Kurz, Jürgen Dönitz and Tim Beißbarth conceptualized the architecture. Nadine Kurz implemented the Onkopus modules (SEQ01-SEQ02, SEQ04-SEQ10, SEQ12-SEQ23, CNV01-CNV06, INF02-INF04), the Onkopus Aggregator (INF05), G01, and the Onkopus server I (INF06]), see App. A. Klara Drofenik and Nadine Kurz implemented the protein annotation module (SEQ11). Nadine Kurz implemented the client software package for accessing the Onkopus modules. Nadine Kurz and Tim Tucholski implemented the web front end (INF07). Nadine Kurz and Kevin Kornrumpf implemented the KNIME workflows. Nadine Kurz set up and maintained the publicly available instances all Onkopus modules, the Onkopus server and the Onkopus web front end. Klara Drofenik set up testing environments for the public web instance. Alexander König provided medical use cases and evaluated the manuscript.
Nadine Kurz wrote the manuscript. Nadine Kurz created the figures. All authors proofread the manuscript. Nadine Kurz revised the manuscript.