# Extended Abstract

**Motivation** Cancer target discovery requires choosing a gene whose perturbation kills cancer cells. This process entails integrating several evidence sources which, in practice, is expensive; computing expression, copy-number (CNA), damaging mutations, and hotspot mutations costs time, compute, and experimental burden. Prior dependency-map work builds static multi-omic rankers (Tsherniak et al., 2017; Pacini et al., 2024) that treat all omics as simultaneously available, which is unrealistic in a budget-constrained pipeline. Current feature acquisition learns generic feature-query policies (Janisch et al., 2018; Bernardino et al., 2022) but not target-specific acquisition. Some evidence is redundant or lineage-dependent and exhaustive multi-omic querying may not be worth its cost.

**Method** We formulate cancer target selection as a finite-horizon, cost-aware sequential decision process. We train a Double DQN that adapts acquisition to partial observations and stops when extra queries are not worth their cost, learning which gene-modality pairs to query before committing to a target. At each step, the model can either QUERY(`gene`, `modality`), incurring a cost, or SELECT(`gene`), terminating the episode. Each episode contains one DepMap cancer cell line and $16$ shuffled candidate genes: $2$ dependent positives (CRISPR gene effect $\leq -0.5$) and $14$ non-dependent negatives. Query actions reveal Ridge-predicted dependency evidence scores for expression, CNA, damaging mutation, or hotspot mutation, rather than raw omics values. The primary reward is total reward $=$ $-$selected dependency $-$ query cost, so the agent is rewarded for selecting strongly dependent genes while avoiding unnecessary evidence queries.

**Implementation** We implemented a Double DQN with experience replay, a target network, $\epsilon$-greedy exploration, and action masking. The Structured DQN encodes each candidate gene with a shared per-gene encoder, then uses separate QUERY and SELECT heads to score the $16 \times 4$ query actions and $16$ select actions. This parameter sharing is important because the model learns reusable gene-modality decision rules instead of treating all 80 actions independently. Policies train for 20k episodes, choose checkpoints by validation reward, and evaluate on 500 held-out cell-line episodes. Context variants add OncoTree lineage at different points: during acquisition, through modality fusion, through lineage-specific Ridge query scores, or only at final SELECT.

**Results** The Structured DQN outperformed baselines with reward $1.035$ using $12.59$ queries per episode, compared with $16$ queries for full CNA or expression policies. The best Context DQN uses lineage only at SELECT, improving reward to $1.043$ while using $12.37$ queries. Both Structured and Context DQN reach Hit@$3 = 1.0$, so the context gain is not from fixing missed targets; it comes from slightly better dependency scores and slightly lower query cost. The ablations also reveal two strong structured policies with different acquisition behavior: an earlier Structured 1-step run was CNA-heavy, while the tuned Structured DQN became expression-heavy, suggesting CNA and expression may contain overlapping useful dependency signal.

**Discussion** We learned that architecture and action representation matter before context helps. Moving from a flat MLP to a per-gene Structured DQN made the Q-function easier to learn because the same query/select logic can transfer across candidate genes. The context sweep then shows that cancer lineage is not automatically useful at every stage: acquisition-time context and fusion often changed modality usage toward CNA or mutation evidence, but lowered reward. In contrast, SELECT-only context preserved the reliable Structured DQN acquisition protocol and used lineage only to interpret collected evidence at the final target choice. Most cancers remain expression-driven, but some lineages rely more on mutation evidence: thyroid, lymphoid, liver, and skin show higher damaging-mutation query rates, while hotspot mutation is rare and most visible in liver.

**Conclusion** SEACTS shows that cancer target selection can be modeled as sequential, cost-aware evidence acquisition rather than static multi-omic ranking. Structured action sharing enables effective learning in the large gene-modality action space. The agent also learns a triage-like policy: query enough evidence to identify strong candidates, then stop before extra evidence becomes too costly. Lineage helps most after evidence is collected, improving final target selection without disrupting a strong query policy. Future work should calibrate query costs to realistic assay or computational burden, test larger candidate sets, add richer biological context features, and replace simple Ridge evidence scores with stronger modality-specific predictors.

# SEACTS: Sequential Evidence Acquisition for Cancer Target Selection

**Ria Garg**
Department of Computer Science
Stanford University
riagarg@stanford.edu

**Nathan Zhou**
Department of Computer Science
Stanford University
natezhou@stanford.edu

## Abstract

Cancer target discovery requires integrating heterogeneous biological evidence while deciding which genes are worth perturbing. Most dependency-map methods treat expression, copy-number, and mutation data as simultaneously available, but real target-discovery pipelines face costs for acquiring, computing, or interpreting each evidence source. We introduce SEACTS, a sequential evidence-acquisition framework for cancer target selection. Each episode presents one DepMap cancer cell line and a small candidate gene set; an agent may query modality-specific evidence for gene-modality pairs before selecting a final target. Query actions incur costs, while selection is rewarded according to hidden CRISPR dependency. We train Double DQN policies with candidate-structured Q-networks. On held-out cell-line episodes, the tuned Structured DQN reaches mean total reward $1.035$ with $12.6$ queries per episode, beating fixed full-modality query baselines while using fewer queries. A SELECT-only cancer-context variant further improves reward to $1.043$ with $12.4$ queries by using OncoTree lineage only at final target scoring. Ablations show that candidate-structured action representation is critical, while context is most useful for interpreting acquired evidence rather than broadly reshaping acquisition.

## 1 Introduction

Modern cancer target discovery remains a major bottleneck in oncology, with clinical drug development failure rates often exceeding 90% (Sun et al., 2022). This process relies on integrating diverse biological evidence sources like gene expression, mutations, and pathway context to identify genes whose perturbation impairs tumor viability. In practice, these modalities are highly heterogeneous, uneven in quality, and often redundant, yet most existing methods are designed under the assumption that all relevant data are available at decision time and should be used simultaneously. This obscures a key decision-making problem: which evidence is actually necessary for a given cancer context, and when does additional information cease to be worth its cost?

We focus on cancer dependency prediction using DepMap data and approach target selection as a sequential decision-making problem under a budget constraint. Given a cancer cell line and a set of candidate genes, an agent must decide which biological evidence to acquire, in what order, and when to stop before selecting a gene to target. We model this process using deep reinforcement learning, where the agent is rewarded for selecting genes with strong dependency, using CRISPR dependency as a proxy for intervention effectiveness, while minimizing the cost of evidence acquisition. In our simulated environment, query costs serve as proxies for experimental, computational, or assay burden in real target-discovery pipelines. Our objective is to determine whether context-conditioned, adaptive evidence acquisition can approach the performance of static multi-omics models while using fewer modalities, and to understand how optimal evidence strategies vary across cancer types.

## 2 Related Work

Cancer dependency prediction has been extensively studied using large-scale data such as DepMap, which integrates CRISPR gene knockout screens with genomic, transcriptomic, and copy-number data across hundreds of cancer cell lines (Broad Institute, 2026; Tsherniak et al., 2017). These resources have enabled predictive models of gene essentiality and therapeutic vulnerabilities, with recent work further developing clinically informed dependency maps for target prioritization (Pacini et al., 2024) and extending these ideas to translational settings by learning predictors that generalize from cell lines to patient tumors (Shi et al., 2024). However, these approaches rely on static multi-omics integration, treating all modalities as simultaneously available rather than modeling how evidence should be acquired under constraints.

A closely related line of work is active feature acquisition (AFA), which formulates prediction as a sequential decision problem in which an agent selects features to observe while trading off predictive accuracy and acquisition cost. Reinforcement learning (RL) has been widely applied in this setting, including early deep RL approaches for cost-aware feature selection (Janisch et al., 2018) and more recent work on active modality selection in medical diagnosis (Bernardino et al., 2022). Recent advances have explored structured acquisition strategies and information-theoretic objectives, highlighting limitations of both RL-based policies and greedy approaches (Huang et al., 2026). While these methods capture sequential decision-making, they typically treat inputs as homogeneous features and focus on classification tasks. In contrast, biological evidence sources are heterogeneous and semantically distinct, and the downstream objective is often target ranking or intervention selection rather than simple prediction.

Another relevant direction is mixture-of-experts (MoE) and gating models, which learn to route inputs to specialized predictors (Shazeer et al., 2017). While these approaches capture modality specialization, they generally make single-step routing decisions and do not model sequential querying, stopping, or cost-aware decision-making. More broadly, Deep Q-Networks (DQNs) provide a standard framework for learning value functions over discrete actions with delayed rewards (Mnih et al., 2015), making them a natural fit for a finite-horizon evidence-acquisition environment. Double DQN further reduces overestimation bias by decoupling action selection from target evaluation in the Bellman target (van Hasselt et al., 2016). However, this framework has not been applied to settings where biological evidence acquisition and target selection are jointly optimized.

Our work bridges these areas by reformulating cancer target selection as a sequential, cost-aware decision problem. Unlike prior AFA methods, we explicitly model target selection as an action and evaluate performance using dependency-based outcomes. Unlike static multi-omics models, we allow the policy to adapt its evidence acquisition strategy based on cancer context, enabling analysis of how different biological settings influence optimal decision-making. To our knowledge, prior work has not explored jointly optimizing sequential evidence acquisition and target selection in cancer dependency maps under cancer-context conditioning.

## 3 Method

### 3.1 Sequential Target-Selection Environment

We formulate cancer target selection as a finite-horizon Markov decision process. An episode is defined by a cancer cell line $c$ and a shuffled candidate gene set $G$. Let $d(c, g)$ denote the CRISPR dependency score for gene $g$ in cell line $c$, with more negative values indicating stronger dependency. These true dependency scores are hidden during the episode and used only for terminal reward and evaluation.

At each timestep, the agent either queries evidence, $\textsc{Query}(g, m)$, for gene $g$ and modality $m$, or terminates with $\textsc{Select}(g)$. Repeated queries are masked, and select actions can be masked until a minimum evidence budget is reached. Thus the policy must learn what evidence to acquire, when to stop, and which target to select.

The state records the partially observed evidence table. For each candidate gene and modality, it contains one observed-value slot and one binary query-mask slot. Unqueried values are encoded as zero, so the mask distinguishes missing observations from true zero-valued scores. Each candidate also receives a normalized slot-index feature that identifies its fixed action indices, but not biological

identity because candidate genes are shuffled. With $M$ modalities, the encoded state has $|G|(2M + 1) + 1$ features; in our main setting, $|G| = 16$ and $M = 4$, giving 145 state features.

Raw omics values are heterogeneous, so the environment converts each modality into comparable dependency evidence scores. For each modality and gene, we fit a one-feature Ridge regression on training cell lines,

$$\hat{d}_m(c, g) = \beta_{0,g,m} + \beta_{1,g,m} x_m(c, g), \qquad e_m(c, g) = -\hat{d}_m(c, g), \tag{1}$$

where $x_m(c, g)$ is the raw modality value and $e_m(c, g)$ is the returned evidence score, with larger values indicating stronger predicted dependency. These per-gene models are fit only on training cell lines. Missing scores, caused by too few valid examples or no modality variation, are returned as zero with the query mask set to one. Lineage-specific variants repeat the fit within OncoTree lineages when enough samples are available, otherwise falling back to global scores.

Query actions incur a modality-specific cost. A query reveals deterministic evidence from the dataset and gives immediate reward

$$R_{\text{Query}}(g, m) = -\text{cost}(m). \tag{2}$$

A select action ends the episode and receives reward

$$R_{\text{Select}}(g) = -d(c, g), \tag{3}$$

where more negative CRISPR dependency therefore corresponds to higher reward. The total episode return is

$$R_{\text{total}} = -d(c, g_{\text{selected}}) - \sum_{t \in \mathcal{Q}} \text{cost}(m_t), \tag{4}$$

where $\mathcal{Q}$ is the set of query timesteps. This objective directly trades off target quality against evidence cost.

### 3.2 Double DQN

We train Double DQN policies with replay, a target network, action masking, and $\epsilon$-greedy exploration. The online network selects the next action for the Bellman target while the target network evaluates it. Transitions store the encoded state, action, reward, next state, valid-action mask, done flag, and optional cancer-context index.

For a collapsed transition $(s_t, a_t, R_t^{(n)}, s_{t+n})$, where $R_t^{(n)} = \sum_{i=0}^{n-1} \gamma^i r_{t+i}$, the Double DQN target is

$$y_t = R_t^{(n)} + (1 - \mathbf{1}_{t+n}^{\text{done}}) \gamma^n Q_{\bar{\theta}} \left( s_{t+n}, \arg \max_{a' \in \mathcal{A}(s_{t+n})} Q_\theta(s_{t+n}, a') \right). \tag{5}$$

The indicator removes bootstrapping at terminal states. $\mathcal{A}(s)$ excludes queried pairs and, when applicable, select actions before the minimum-query threshold. We minimize Smooth L1 Bellman error over replay minibatches and periodically copy $\theta$ into $\bar{\theta}$.

### 3.3 Structured DQN

The flat action space contains repeated decisions: the same modality can be queried for many genes, and each gene can be selected. A plain MLP maps the flattened state directly to action values without explicit sharing, whereas the Structured DQN applies a shared candidate encoder. For gene $g$, let $x_g$ denote observed evidence, $q_g$ the query mask, and $p_g$ the normalized slot feature:

$$\phi_g = f_\theta(x_g, q_g, p_g), \qquad \bar{\phi} = \frac{1}{|G|} \sum_{g \in G} \phi_g. \tag{6}$$

The pooled vector $\bar{\phi}$ summarizes the candidate set; separate heads score query and select actions:

$$Q_q(g, m) = h_q^m(\phi_g, \bar{\phi}), \qquad Q_s(g) = h_s(\phi_g, \bar{\phi}). \tag{7}$$

This produces one Q-value for every action while sharing parameters across repeated decision types. We also evaluate dueling and multi-step-return variants, but the main modeling contribution is this candidate-aware action representation.

### 3.4 Context DQN

To test whether cancer type can influence acquisition, we build context DQN variants using OncoTree lineage, which represents the type of cancer the cell line is from. We evaluate context throughout the structured network, lineage-specific evidence scores, context-fusion heads, and SELECT-only context. The SELECT-only variant preserves the Structured DQN query pathway and adds lineage embedding $c$ only when scoring final target selection,

$$Q_s(g) = h_s(\phi_g, \bar{\phi}, c). \tag{8}$$

These variants test whether lineage can change acquisition or mainly final interpretation.

## 4 Experimental Setup

### 4.1 Data and Environment

As depicted in Table 1, we use DepMap CRISPR gene-effect data with expression, copy-number alteration (CNA), damaging mutation, and hotspot mutation features. Each episode corresponds to one cancer cell line and contains 16 shuffled candidate genes: 2 dependent positives with CRISPR gene effect $\leq -0.5$ and 14 non-dependent negatives. CRISPR gene effect ranges from $-6.26$ to $5.77$. True dependency values are hidden until final selection.

| Name | Description | Query cost |
|---|---|---|
| Expression | Gene expression measurements (TPM log) | $0.02$ |
| CNA | Copy-number alteration values ($\log_2$) | $0.02$ |
| Damaging mutation | Binary indicator of damaging mutations | $0.02$ |
| Hotspot mutation | Binary indicator of known hotspot mutations | $0.02$ |
| Metadata | Cell line identifiers and lineage/context features | not queryable |
| CRISPR gene effect | Ground-truth dependency labels | hidden label |

Table 1: DepMap features.

In the main environment, each query costs $0.02$ and returns the Ridge-predicted dependency evidence score in Eq. (1) rather than a raw omics value. With 16 candidates and 4 queryable modalities, the action space contains $16 \times 4$ query actions and 16 select actions, for 80 total actions with invalid actions masked. The agent must make at least 8 queries before selecting and has a maximum horizon of 32 steps.

### 4.2 Baselines and Training

We compare against three groups of baselines. First, random selection selects uniformly without querying, and an oracle upper bound selects the candidate with the best hidden CRISPR dependency score. Second, direct raw-modality rankers score candidates without sequential interaction. Third, environment baselines use the same QUERY/SELECT interface and query costs as DQN: fixed full-modality policies query expression or CNA for all 16 candidates before selecting, budgeted policies query a smaller fixed number of candidates such as CNA budget 12, and query-all observes every candidate-modality pair. Separately, we also compare Structured DQN ablations and cancer-context DQN variants to test action representation and lineage conditioning.

All main DQN experiments use Double DQN with experience replay, a target network, $\epsilon$-greedy exploration, and validation-based checkpoint selection. Unless swept, models use discount $\gamma = 0.95$, Adam learning rate $10^{-4}$, batch size $64$, target-network updates every $500$ optimization steps, learning after $500$ replay transitions, and gradient clipping at norm $10$. Exploration is gradually decayed from $\epsilon = 1.0$ to $0.05$. We use 1-step targets in Eq. (5) ($n$=1); ablations include 3-step returns ($n$=3).

Reported long-run DQN and ablation models train for $20{,}000$ episodes with replay capacity $50{,}000$ and validation every $250$ episodes. Tuned Structured and Context DQN models use hidden dimension $256$; architecture ablations use $128$ unless stated otherwise. Final DQN evaluation averages over $500$ held-out cell-line episodes. Direct data and environment baselines use the same held-out split and episode generator, averaged over $1000$ episodes.

### 4.3 Metrics

The primary objective is total reward defined in Eq. (4), which combines the terminal select reward in Eq. (3) with query costs in Eq. (2). Because lower CRISPR gene-effect values indicate stronger dependency, this rewards more essential selected genes while penalizing unnecessary evidence acquisition. We also report selected dependency, query cost, query count, modality usage, Hit@3, NDCG@3, and MRR@3. Hit@3 checks whether the predicted top three contain any true top-three dependency gene; NDCG@3 uses relevance $\max(-d(c, g), 0)$; and MRR@3 is the reciprocal rank of the single most dependent gene if it appears in the predicted top three.

## 5 Results

### 5.1 Quantitative Evaluation

The main ablation result is that candidate-structured action representation matters more than multi-step returns or dueling heads. With 1-step returns, the MLP reaches total reward $0.845$, whereas the Structured DQN ablation reaches $1.023$, the largest gain among non-context ablations. Figure 1 shows the corresponding reward differences and modality-use patterns.

Multi-step returns and dueling heads do not consistently improve reward. The 3-step MLP drops to $0.799$, and the 3-step Structured DQN drops to $0.978$. Adding a dueling head raises the 3-step structured variant to $1.021$, but it does not exceed the strongest 1-step structured policies.

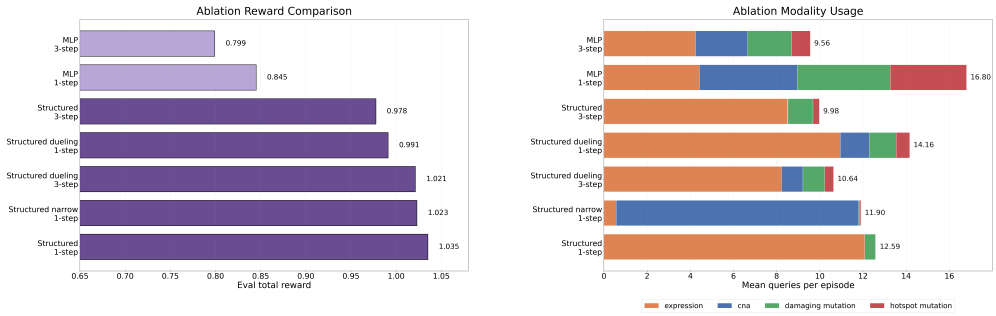


Figure 1: Structured DQN architecture ablations.

The original Structured 1-step ablation uses hidden dimension $128$, reaches reward $1.023$, and averages $11.23$ CNA queries out of $11.90$ total queries. The tuned Structured DQN uses hidden dimension $256$, reaches reward $1.035$, and averages $12.08$ expression queries out of $12.59$ total queries.

Table 2 compares the metrics from the tuned Structured DQN against the main baselines. The oracle upper bound obtains total reward $1.423$. Full CNA and full expression querying select slightly stronger dependencies than the learned DQN, but always spend $16$ queries. The Structured DQN uses fewer queries ($12.59$) and obtains higher cost-adjusted reward ($1.035$) than either fixed full-query policy. Relative to full-expression and full-CNA baselines, it saves roughly $3.4$–$3.6$ queries per episode. Direct raw-modality rankers are weaker than sequential supervised-score policies; expression is the strongest single raw modality, but remains below the full-query and DQN policies.

We compare cancer-context variants against the tuned Structured DQN. Figure 2 compares the variants which include context from start with structured initialization, smaller context embeddings, dueling context heads, lineage-specific query scores, context fusion, and SELECT-only context.

Context from start with structured initialization reaches reward $1.025$, close to but below the tuned Structured DQN. A smaller context embedding reaches $1.023$, dueling context reaches $0.997$, and lineage-specific query scores reach $1.010$. The fusion variant performs worst, reaching $0.932$ while using $16.21$ queries on average and shifting toward CNA and damaging-mutation evidence.

| Policy | Reward | Dependency | Queries | Hit@3 | NDCG@3 |
|---|---|---|---|---|---|
| Random select | 0.164 | -0.164 | 0.00 | 0.483 | 0.221 |
| Query expression budget 12 | 0.921 | -1.161 | 12.00 | 0.976 | 0.761 |
| Query CNA budget 12 | 0.927 | -1.167 | 12.00 | 0.975 | 0.760 |
| Query expression full | 1.005 | -1.325 | 16.00 | 0.998 | 0.921 |
| Query CNA full | 1.011 | -1.331 | 16.00 | 0.998 | 0.919 |
| Structured narrow DQN | 1.023 | -1.261 | 11.90 | 1.000 | 0.829 |
| Structured DQN | 1.035 | -1.287 | 12.59 | 1.000 | 0.847 |
| Context DQN | 1.043 | -1.290 | 12.37 | 1.000 | 0.843 |
| Oracle select | 1.423 | -1.423 | 0.00 | 1.000 | 1.000 |

Table 2: Results across baselines and models.

The best model is SELECT-only Context DQN. It improves reward from $1.035$ to $1.043$ while using $12.37$ queries, slightly fewer than the Structured DQN's $12.59$. Its selected dependency $-1.290$ is also slightly stronger than the Structured DQN's $-1.287$, indicating that the gain comes from both better target quality and lower query cost. Both Structured and Context DQN maintain Hit@3 $= 1.0$, so both models consistently select high-ranked dependency candidates. Training converges quickly: reward rises from $0.65$ at episode $250$ to a peak of $1.041$ around episode $12750$, after which performance stabilizes near $1.0$ (Figure 6).

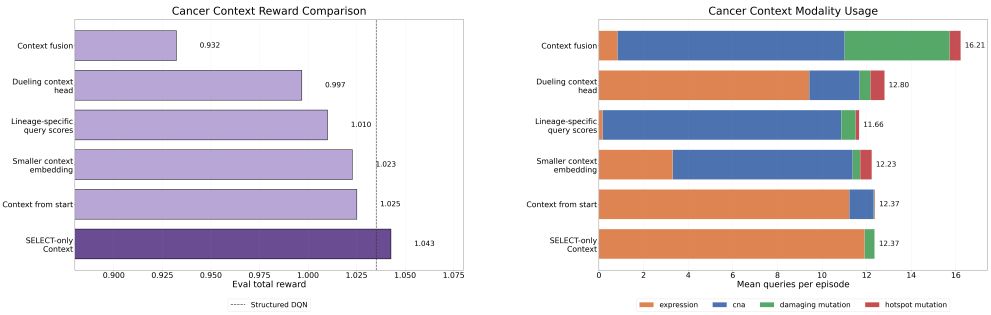


Figure 2: Cancer DQN architecture comparisons.

### 5.2 Qualitative Analysis

The two strongest Structured 1-step runs differ in modality preference: the ablation is CNA-heavy, while the tuned Structured DQN is expression-heavy, suggesting that CNA and expression may contain overlapping useful dependency signal.

Figure 3 shows repeated expression queries before selection, rare mutation queries, and almost no CNA. Query-efficiency summaries show that queries concentrate on genes with stronger true dependency ranks. The representative trajectories show the same pattern: acquisition is mostly expression-driven. Although Hit@3 is $1.0$, $27\%$ of episodes still have positive dependency regret, meaning the agent selected a dependent gene but not the most dependent one. The largest regrets occur when the agent selects a weaker dependent gene after a short expression-only protocol (e.g., regret $1.48$ on *KIF18A* and $1.43$ on *CCND1*).

We analyzed lineage-specific patterns in Figure 4 and found that the Context DQN remains expression-heavy across all cancers, so the gain comes from lineage-aware selection rather than a new query protocol. Some lineages lean more on mutation evidence: thyroid, lymphoid, liver, and skin show the highest damaging-mutation query rates. Hotspot mutation use is rare and only appears in liver, and CNA is unused.

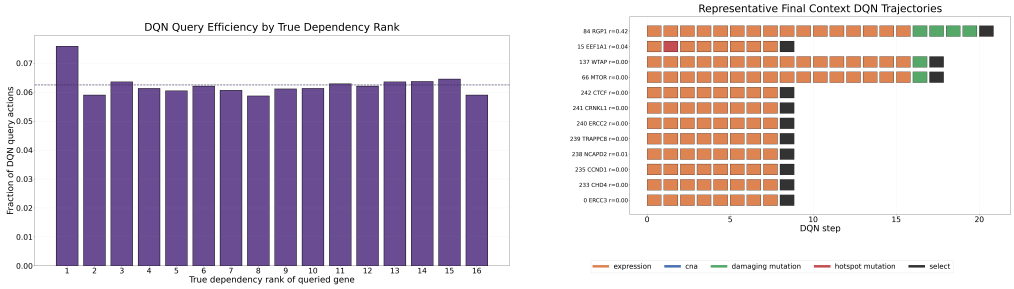


Figure 3: Context DQN behavior in query efficiency and example trajectories.

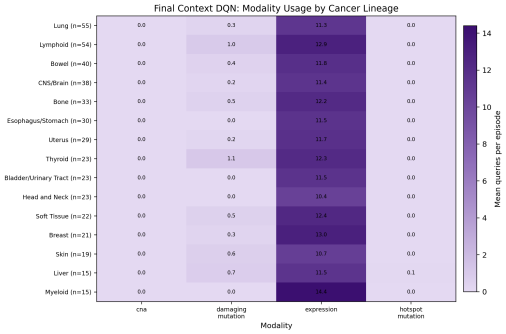


Figure 4: Lineage-level modality usage.

## 6 Discussion

The main empirical finding is that adaptive evidence acquisition improves the cost-adjusted target-selection objective. If the only goal were to maximize target quality after all evidence is available, full-query expression or CNA would be difficult to beat. SEACTS instead asks whether additional evidence justifies its cost, and under this objective the learned policies stop earlier while preserving high Hit@3.

The ablations show that the improvement is not simply due to applying DQN. The flat MLP performs substantially worse than the Structured DQN, indicating that action representation is central. By sharing parameters across repeated gene-modality decisions, the structured architecture provides a useful inductive bias for an 80-action problem with sparse terminal rewards.

The context experiments sharpen the biological interpretation. OncoTree lineage is useful, but not uniformly at every stage: acquisition-level context often shifts modality use without improving reward, while the strongest context model uses lineage only for final target scoring. This suggests that broad cancer lineage may be more helpful for interpreting accumulated evidence than for deciding which evidence to collect from scratch.

The ablations suggest several biological explanations for the learned modality use. The narrow Structured DQN leaned on CNA and the tuned model on expression while reaching similar reward. This provides evidence that CNA and expression may be partially redundant: copy-number changes often shift expression at the same genes (Shao et al., 2019), and both carry dependency signal in DepMap (Tsherniak et al., 2017; Pacini et al., 2024). Furthermore, damaging-mutation queries were highest in thyroid, lymphoid, liver, and skin (Figure 4), lineages where recurrent drivers define common subtypes–lymphoid signaling and epigenetic programs (Schmitz et al., 2018), melanoma in skin models (The Cancer Genome Atlas Network, 2015), MAPK-altered papillary thyroid carcinoma (The Cancer Genome Atlas Research Network, 2014), and TP53/CTNNB1-defined HCC (The Cancer Genome Atlas Research Network, 2017). In the broader literature, lymphoid, melanoma/skin, thyroid, and liver cancers are among types where somatic mutations are biologically important.

In terms of limitations, query costs are normalized proxies rather than measured experimental or computational costs. Query returns are Ridge-predicted dependency scores, which makes modalities comparable but limits biological richness. Episodes also use small candidate sets with fixed positive/negative composition, so larger target pools may require additional action-space structure or retrieval. Finally, DepMap cell lines are an imperfect proxy for patient tumors, motivating evaluation on patient-derived or clinically annotated settings.

## 7 Conclusion

We introduced SEACTS, a sequential evidence-acquisition framework for cancer target selection. Instead of assuming all omics are available upfront, SEACTS models target discovery as a cost-aware decision process in which an agent chooses which gene-modality evidence to query before selecting a target. On DepMap-derived episodes, learned DQN policies improve total reward over fixed full-modality query baselines by selecting strong targets with fewer queries.

Three findings seem most worth carrying forward. First, representation dominates naive deep RL here: sharing parameters across repeated gene-modality decisions yields a much stronger policy than a flat MLP on the same data and reward. Second, cancer context is stage-specific: OncoTree lineage helps when it informs final target scoring, not when it is allowed to rewrite the entire acquisition policy. Third, the agent discovers biologically structured acquisition behavior without hand-coded rules—expression-led querying, partial interchangeability between CNA and expression, and slightly higher damaging-mutation use in lineages where somatic drivers are useful.

Three directions seem most promising. First, we will calibrate query costs to realistic assay burdens—for example, separating the expense of expression profiling, copy-number assays, and mutation panels rather than treating modalities as equally costly—and study how policies change when marginal evidence is priced like real experiments. Second, we will use richer mutation representations and lineage-conditioned acquisition routing, building on the emergent pattern we observe in thyroid, lymphoid, skin, and liver. Third, we will use stronger modality-specific predictors in place of Ridge evidence scores, and larger target pools to test whether high Hit@3 can be converted into lower dependency regret. Together, these steps would push SEACTS toward a target-discovery system that jointly reason about what to target and which evidence is worth paying for to decide.

## 8 Team Contributions

- **Ria Garg:** Contributed equally across problem formulation, data processing, environment design, training, evaluation, and writing. Took the lead on the cancer-context experiments, including OncoTree lineage conditioning, context DQN variants, lineage-specific evidence-score analysis, context-sweep interpretation, and biological analysis of modality usage across cancer lineages.
- **Nathan Zhou:** Contributed equally across problem formulation, data processing, environment design, training, evaluation, and writing. Took the lead on the structured reinforcement-learning implementation, including the query/select environment, Double DQN training pipeline, candidate-structured Q-network, architecture ablations, baseline comparisons, and analysis of learned query behavior.

**Changes from Proposal** The final project stayed close to the proposal's objective of modeling cancer target selection as sequential, cost-aware evidence acquisition. We narrowed evaluation toward target-selection reward, query cost, and ranking metrics, and we developed the proposed DQN into a candidate-structured Q-network. We also chose to expand our research on structured DQN ablations and a more detailed cancer-context sweep.

## References

Gabriel Bernardino, Anders Jonsson, Filip Loncaric, Pablo-Miki Martí Castellote, Marta Sitges, Patrick Clarysse, and Nicolas Duchateau. 2022. Reinforcement Learning for Active Modality Selection During Diagnosis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li (Eds.). Springer Nature Switzerland, Cham, 592–601.

Broad Institute. 2026. DepMap Portal. https://depmap.org/portal/.

Hung-Tien Huang, Dzung Dinh, and Junier B. Oliva. 2026. Information Templates: A New Paradigm for Intelligent Active Feature Acquisition. arXiv:2508.18380 [cs.AI] https://arxiv.org/abs/2508.18380

Jaromír Janisch, Tomáš Pevný, and Viliam Lisý. 2018. Classification with Costly Features using Deep Reinforcement Learning. arXiv:1711.07364 [cs.AI] https://arxiv.org/abs/1711.07364

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533. doi:10.1038/nature14236

Clare Pacini, Joshua M. Dempster, Ian Boyle, Emanuel Goncalves, Hanna Najgebauer, et al. 2024. A comprehensive clinically informed map of dependencies in cancer cells and framework for target prioritization. *Cancer Cell* (2024). doi:10.1016/j.ccell.2023.12.004

Roland Schmitz, George W. Wright, Da Wei Huang, Calvin A. Johnson, James D. Phelan, Jerry Q. Wang, Louis M. Staudt, et al. 2018. Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *New England Journal of Medicine* 378, 16 (2018), 1396–1407. doi:10.1056/NEJMoa1801445

Xin Shao, Ning Lv, Jie Liao, Jinbo Long, Rui Xue, Ni Ai, Donghang Xu, and Xiaohui Fan. 2019. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Medical Genomics* 20, 1 (2019), 175. doi:10.1186/s12881-019-0909-5

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1701.06538

Jingyi Shi, Andrew J. Aguirre, et al. 2024. Building a translational cancer dependency map for The Cancer Genome Atlas. *Nature Cancer* (2024). doi:10.1038/s43018-024-00789-y

Duxin Sun, Wei Gao, Hongxiang Hu, and Simon Zhou. 2022. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B* 12, 7 (2022), 3049–3062. doi:10.1016/j.apsb.2022.02.002

The Cancer Genome Atlas Network. 2015. Genomic Classification of Cutaneous Melanoma. *Cell* 161, 7 (2015), 1681–1696. doi:10.1016/j.cell.2015.05.044

The Cancer Genome Atlas Research Network. 2014. Integrated Genomic Characterization of Papillary Thyroid Carcinoma. *Cell* 159, 3 (2014), 676–690. doi:10.1016/j.cell.2014.09.038

The Cancer Genome Atlas Research Network. 2017. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell* 169, 7 (2017), 1327–1341. doi:10.1016/j.cell.2017.05.038

Aviad Tsherniak, Francis Vazquez, Philip G. Montgomery, Bryan A. Weir, Gregory Kryukov, Graham S. Cowley, Shantanu Gill, William F. Harrington, Sam Pantel, John M. Krill-Burger, et al. 2017. Defining a Cancer Dependency Map. *Cell* 170, 3 (2017), 564–576. doi:10.1016/j.cell.2017.06.010

Hado van Hasselt, Arthur Guez, and David Silver. 2016. Deep Reinforcement Learning with Double Q-Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30. 2094–2100. doi:10.1609/aaai.v30i1.10295

# A Additional Results

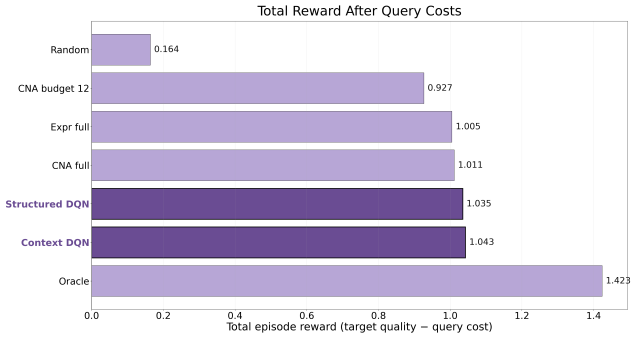


Figure 5: Reward comparison across baselines and models.

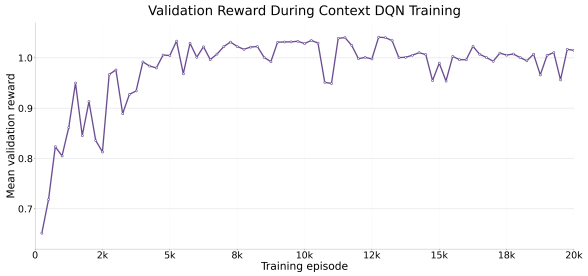


Figure 6: Validation reward of Context DQN.