Performance Optimization of High-Throughput Virtual Screening Pipelines

Hyun-Myung Woo[a], Xiaoning Qian[a,b], Li Tan[b], Shantenu Jha[b,c], Francis J. Alexander[b], Edward R. Dougherty[a], and Byung-Jun Yoon[a,b]

[a]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843
[b]Computational Science Initiative, Brookhaven National laboratory, Upton, NY 11973
[c]Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08854

TEXAS A&M UNIVERSITY

Contents

Introduction
Screening problem

- Effective selection of the potential molecular candidates that meet certain conditions in an immense search space has been one of the major concerns in many real-world biochemistry applications.
  - Finding molecules that can proceed to later stages of the drug design protocol against the COVID-19[1].

1. Saadi, A.A., Alfe, D., Babuji, Y., Bhati, A., Blaiszik, B., Brace, A., Brettin, T., Chard, K., Chard, R., Clyde, A. and Coveney, P., 2021, August. Impeccable: Integrated modeling pipeline for covid cure by assessing better leads. In *50th International Conference on Parallel Processing* (pp. 1-12).

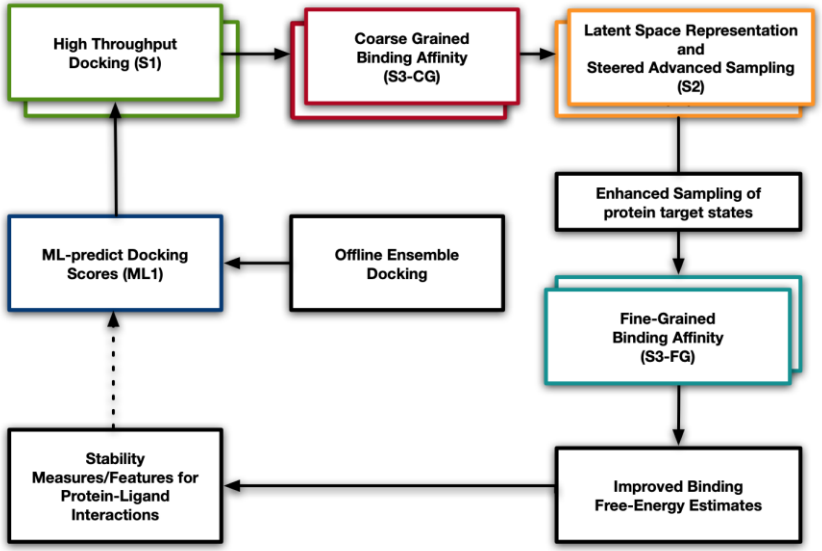
Introduction
Fundamental challenges in the screening problem

1. The number of drug molecules is enormous.
2. The screening cost based on the accurate evaluation platform is expensive.

- Accurate and efficient selection of the potential drug candidates from a huge set of drug molecules is the key factor determining the success of the screening problem.

Introduction
High-Throughput Virtual Screening (HTVS) Pipeline

- HTVS pipeline is one practical approach for the screening problem.



IMPECCABLE: HTVS for COVID cure[1]

1. Saadi, A.A., Alfe, D., Babuji, Y., Bhati, A., Blaiszik, B., Brace, A., Brettin, T., Chard, K., Chard, R., Clyde, A. and Coveney, P., 2021, August. Impeccable: Integrated modeling pipeline for covid cure by assessing better leads. In *50th International Conference on Parallel Processing* (pp. 1-12).

Introduction
Motivation

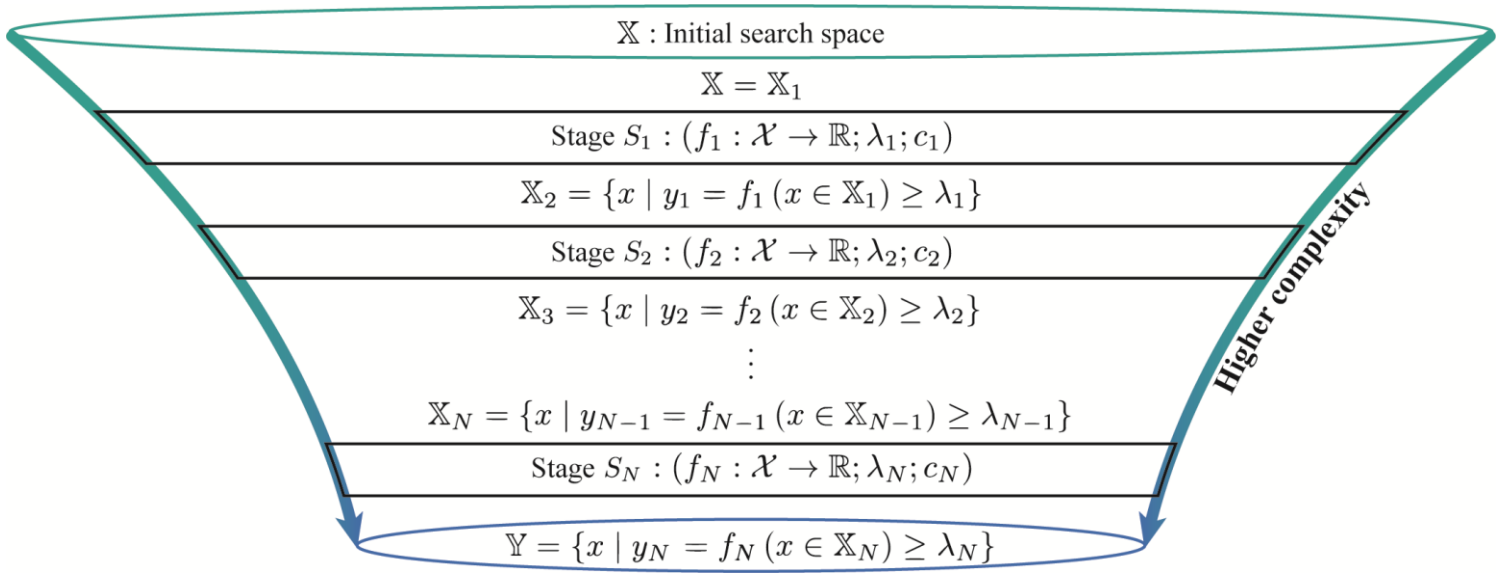
- To date, there has been no optimal rule to manage such HTVS pipelines.

- Can we optimize the performance of the HTVS pipeline?
  1. Can we minimize the (computational) cost?
  2. Can we maximize the throughput (the number of potential candidates)?

- We present two optimization frameworks for the HTVS pipeline.
  1. A framework that optimizes the throughput given a computational budget constraint.
  2. A framework that jointly optimizes the throughput and computational costs.

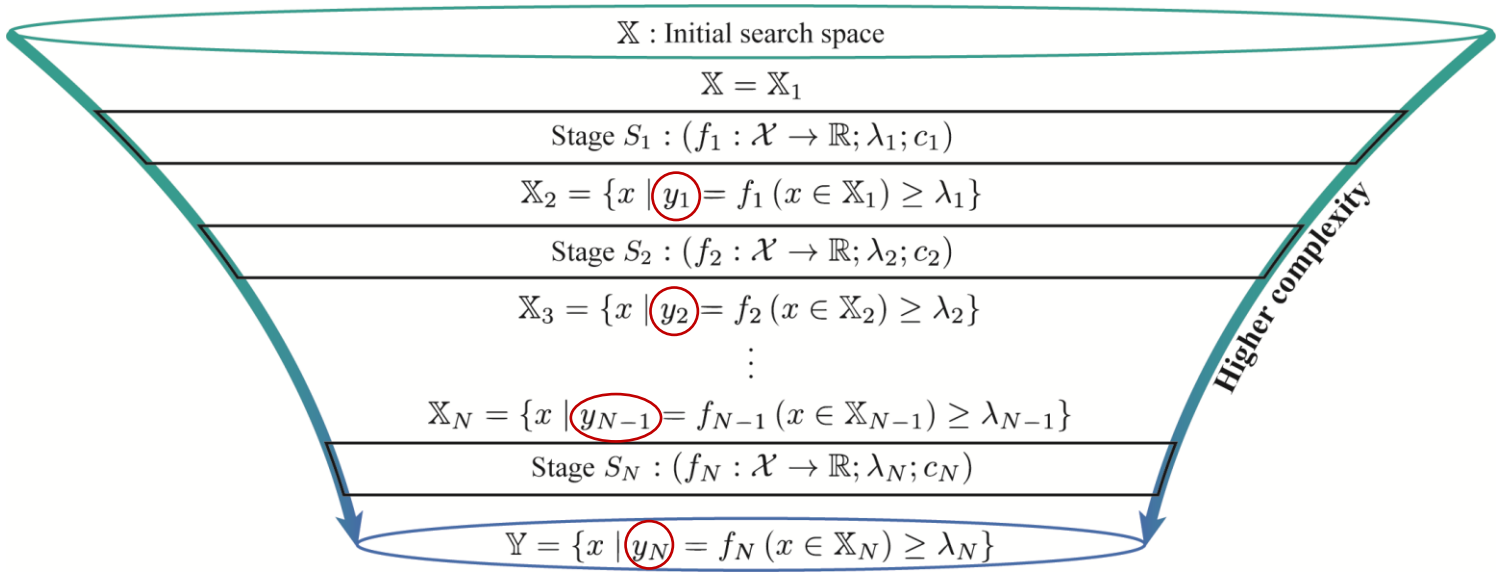
Methods
Illustration of the formal HTVS pipeline problem

- Objective: maximizing throughput $|\mathbb{Y}|$.
- Assumption: screening threshold of the last stage λ_N is given by experts.
- Optimization variables: screening thresholds $\lambda_1, \lambda_2, \ldots, \lambda_{N-1}$ of earlier stages $S_1, S_2, \ldots, S_{N-1}$.



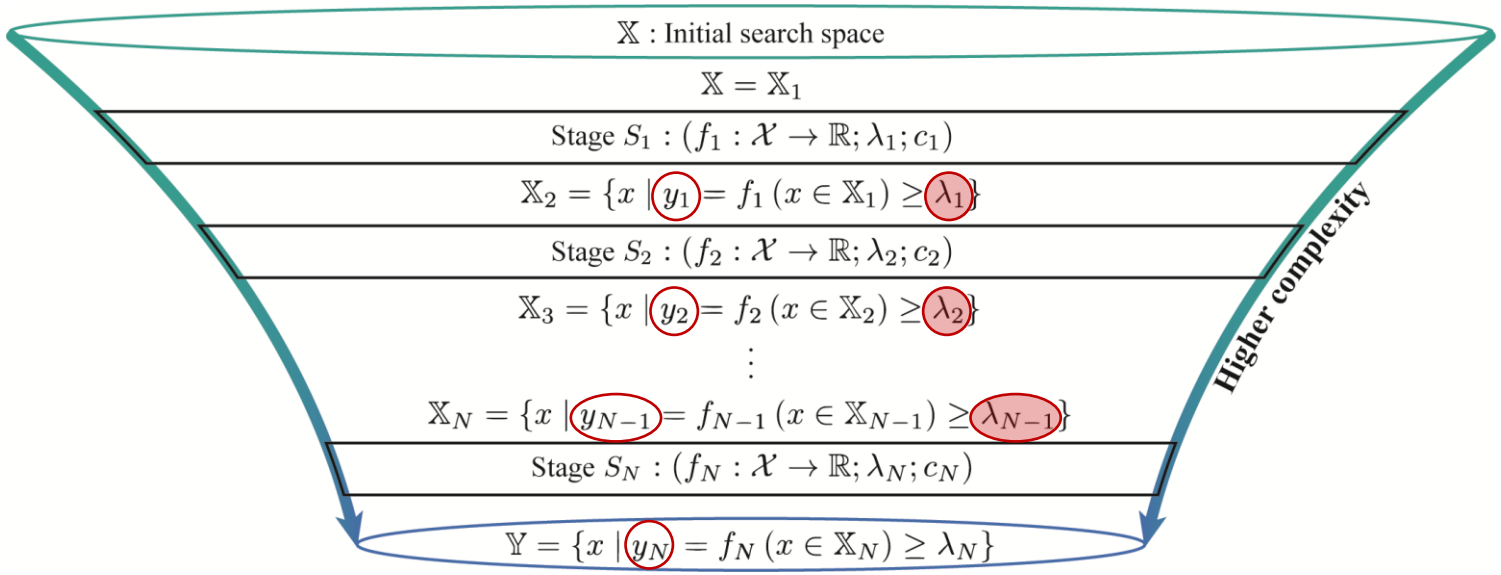
Methods
Key idea of the proposed approaches

1. **Estimating the joint score distribution $f_S(y_1, y_2, \ldots, y_N)$.**
2. Finding $\lambda_1, \lambda_2, \ldots, \lambda_{N-1}$ via the optimization framework.



Methods
Key idea of the proposed approaches

1. Estimating the joint score distribution $f_S(y_1, y_2, \dots, y_N)$.
2. **Finding $\lambda_1, \lambda_2, \dots, \lambda_{N-1}$ via the optimization framework.**



Methods
Proposed optimization framework under a computation budget constraint

- Optimal screening thresholds $\boldsymbol{\psi}^* = [\lambda_1^*, \lambda_2^*, \dots, \lambda_{N-1}^*]$ under fixed computational budget C

$$\boldsymbol{\psi}^* = \underset{\boldsymbol{\psi} \in \mathbb{R}^{N-1}}{\arg\max}\, r([\boldsymbol{\psi}, \lambda_N])$$
$$\text{s. t.} \sum_{i=1}^{N} c_i \, |\mathbb{X}_i| \leq C.$$

Reward function: $r(\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \dots, \lambda_N]) = \int \cdots \int_{[\lambda_N, \lambda_{N-1}, \dots, \lambda_1]}^{\infty} f_{\boldsymbol{S}}(y_1, y_2, \dots, y_N) dy_1 dy_2 \cdots dy_N$

Cardinality of input set $\mathbb{X}_i$ of stage S_i: $|\mathbb{X}_i| = |\mathbb{X}| \int \cdots \int_{[\lambda_i, \lambda_{i-1}, \dots, \lambda_1]}^{\infty} f_{\boldsymbol{S}_{1:i-1}}(y_1, y_2, \dots, y_{i-1}) dy_1 dy_2 \cdots dy_{i-1}$

Methods
Proposed joint optimization framework

- Optimal screening thresholds $\boldsymbol{\psi}^* = [\lambda_1^*, \lambda_2^*, \ldots, \lambda_{N-1}^*]$ jointly optimizing efficiency and throughput

$$\boldsymbol{\psi}^* = \underset{\boldsymbol{\psi} \in \mathbb{R}^{N-1}}{\arg\min}\, \alpha g([\boldsymbol{\psi}, \lambda_N]) + (1-\alpha) h([\boldsymbol{\psi}, \lambda_N]).$$

Weight parameter: $\alpha \in [0,1]$

Relative reward function: $g([\boldsymbol{\psi}, \lambda_N]) = \frac{r([-\infty, \lambda_N]) - r([\boldsymbol{\psi}, \lambda_N])}{r([-\infty, \lambda_N])}$

Normalized total cost function: $h([\boldsymbol{\psi}, \lambda_N]) = \frac{1}{N|\mathbb{X}| \max_i c_i} \sum_{i=1}^{N} c_i \, |\mathbb{X}_i|$

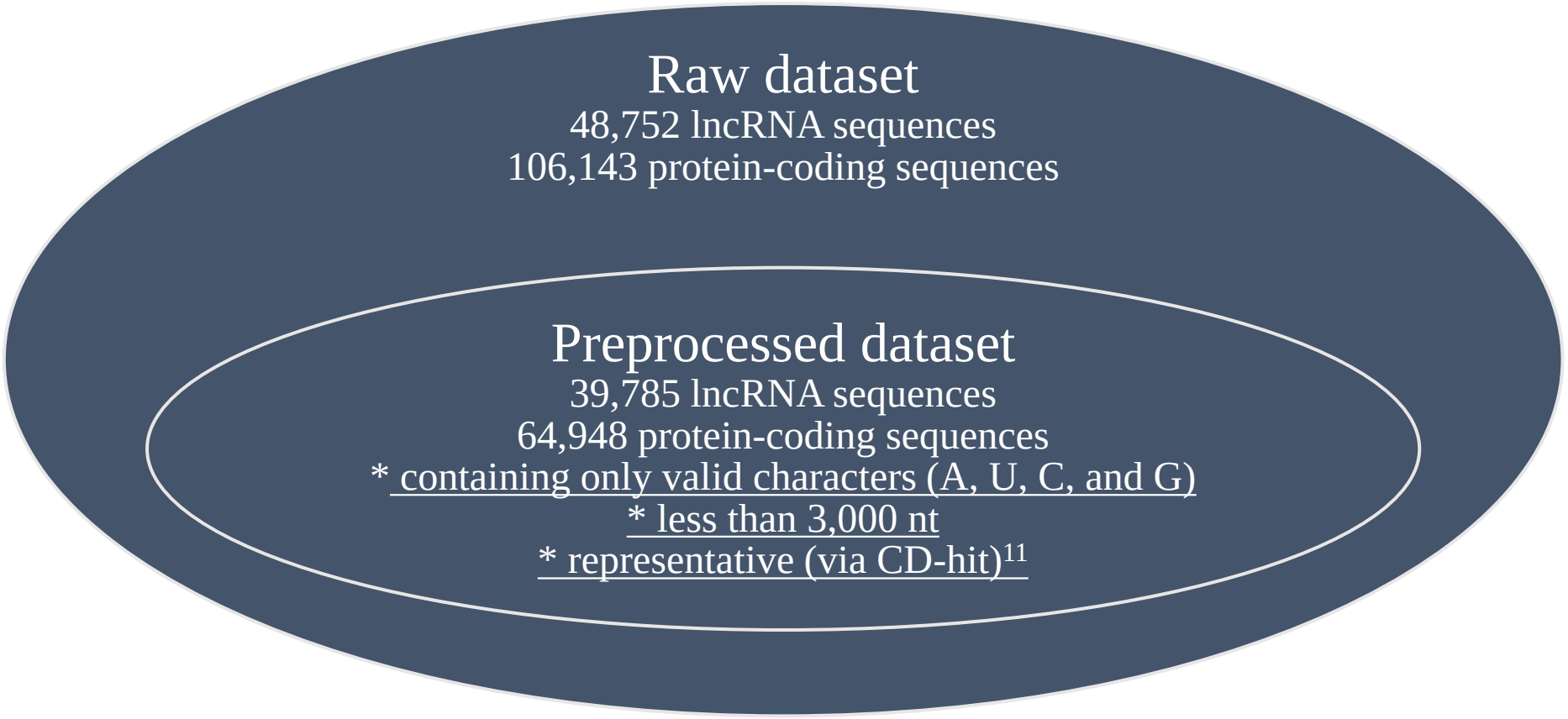
Results: Long non-coding RNA (lncRNA) screening
Motivation

- Long non-coding RNAs (lncRNAs) do not encode proteins.

- LncRNAs are closely related to hard-to-treat diseases including Alzheimer's disease[2,3,4], cardiovascular disease[5,6], and several types of cancers[7,8,9,10].

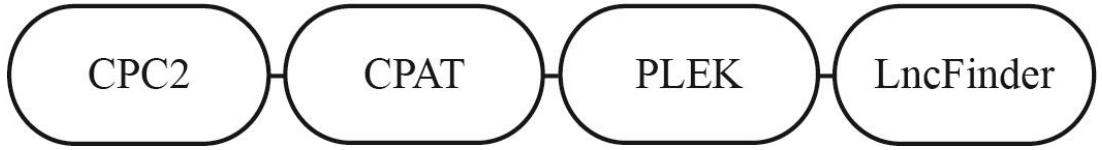
2. Ng, S.Y., Lin, L., Soh, B.S. and Stanton, L.W., 2013. Long noncoding RNAs in development and disease of the central nervous system. *Trends in Genetics*, *29*(8), pp.461-468.
3. Tan, L., Yu, J.T., Hu, N. and Tan, L., 2013. Non-coding RNAs in Alzheimer's disease. *Molecular neurobiology*, *47*(1), pp.382-393.
4. Luo, Q. and Chen, Y., 2016. Long noncoding RNAs and Alzheimer's disease. *Clinical interventions in aging*, *11*, p.867.
5. Congrains, A., Kamide, K., Oguro, R., Yasuda, O., Miyata, K., Yamamoto, E., Kawai, T., Kusunoki, H., Yamamoto, H., Takeya, Y. and Yamamoto, K., 2012. Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of ANRIL and CDKN2A/B. *Atherosclerosis*, *220*(2), pp.449-455.
6. Xue, Z., Hennelly, S., Doyle, B., Gulati, A.A., Novikova, I.V., Sanbonmatsu, K.Y. and Boyer, L.A., 2016. A G-rich motif in the lncRNA braveheart interacts with a zinc-finger transcription factor to specify the cardiovascular lineage. *Molecular cell*, *64*(1), pp.37-50.
7. Yang, G., Lu, X. and Yuan, L., 2014. LncRNA: a link between RNA and cancer. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, *1839*(11), pp.1097-1109.
8. Shi, X., Sun, M., Liu, H., Yao, Y., Kong, R., Chen, F. and Song, Y., 2015. A critical role for the long non-coding RNA GAS5 in proliferation and apoptosis in non-small-cell lung cancer. *Molecular carcinogenesis*, *54*(S1), pp.E1-E12.
9. Peng, W.X., Koirala, P. and Mo, Y.Y., 2017. LncRNA-mediated regulation of cell signaling in cancer. *Oncogene*, *36*(41), pp.5661-5667.
10. Carlevaro-Fita, J., Lanzós, A., Feuerbach, L., Hong, C., Mas-Ponte, D., Pedersen, J.S. and Johnson, R., 2020. Cancer LncRNA Census reveals evidence for deep functional conservation of long noncoding RNAs in tumorigenesis. *Communications biology*, *3*(1), pp.1-16.

Results: Long non-coding RNA (lncRNA) screening
Dataset (Human) - GENCODE (v38, May 2021)

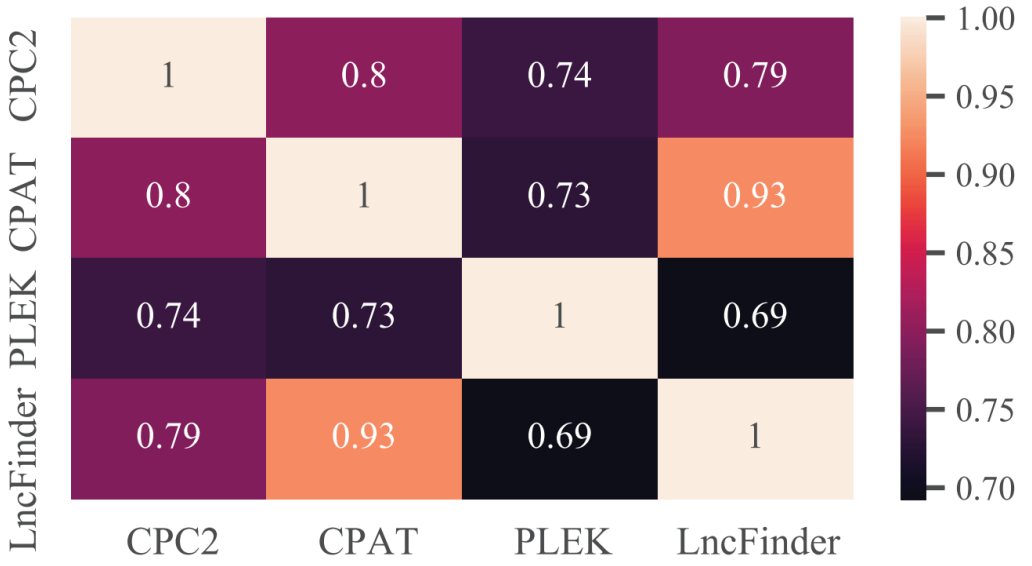


11. Li, W. and Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), pp.1658-1659.

Results: Long non-coding RNA (lncRNA) screening
Construction of the HTVS pipeline



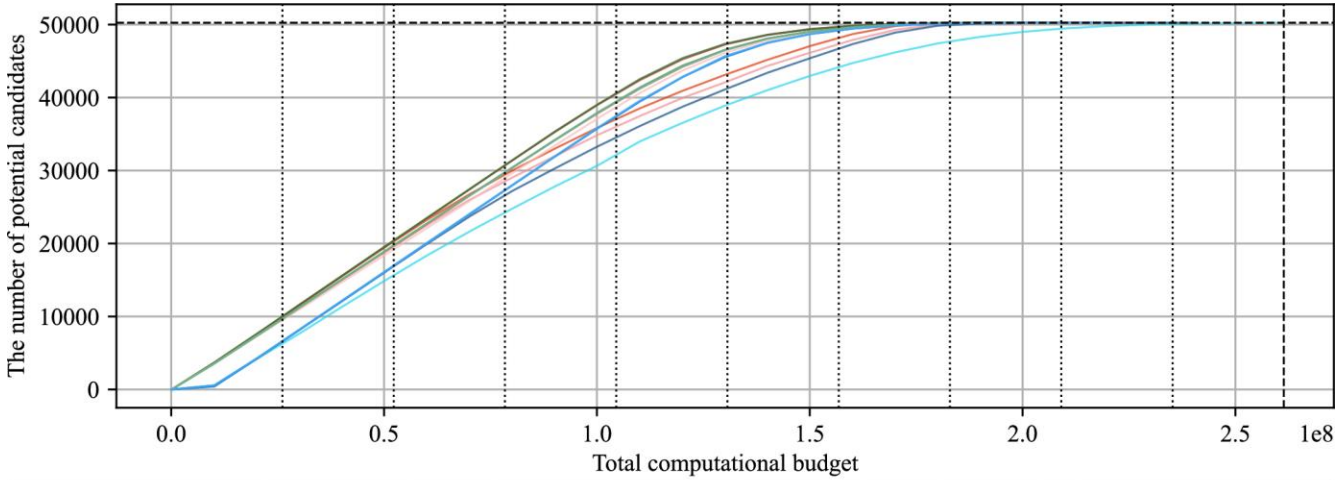
Algorithm	Accuracy	Sensitivity	Specificity	Time (ms)
CPC2	0.7154	0.5760	0.9493	2.5265
CPAT	0.8217	0.6861	0.9817	2.7336
PLEK	0.7050	0.5666	0.9478	83.1765
LncFinder	0.8329	0.7062	0.9678	2,495.623



- Learnt the joint score distribution with 4 % of samples via the EM algorithm

Results - Long non-coding RNA (lncRNA) screening

Performance of the optimized pipeline under the computational budget constraint



	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	100 %
$[S_1, S_4]$	9,973	20,280	29,452	37,071	43,250	48,210	50,220	50,265	50,266	50,266
$[S_2, S_4]$	10,020	20,418	30,749	40,615	47,400	49,719	50,244	50,266	50,266	50,266
$[S_3, S_4]$	6,218	15,643	24,246	32,032	39,017	44,202	47,696	49,447	50,091	50,254
$[S_1, S_2, S_4]$	10,033	20,364	30,692	40,537	47,365	49,713	50,244	50,266	50,266	50,266
$[S_1, S_3, S_4]$	9,563	19,570	28,678	36,164	42,218	47,364	50,125	50,252	50,266	50,266
$[S_2, S_1, S_4]$	10,025	20,397	30,765	40,695	47,466	49,717	50,244	50,266	50,266	50,266
$[S_2, S_3, S_4]$	9,688	19,753	29,760	39,431	46,680	49,467	50,198	50,264	50,266	50,266
$[S_3, S_1, S_4]$	6,530	16,866	26,590	34,663	41,332	46,705	49,974	50,254	50,265	50,266
$[S_3, S_2, S_4]$	6,607	16,997	27,335	37,464	45,742	49,239	50,162	50,264	50,266	50,266
$[S_1, S_2, S_3, S_4]$	9,647	19,711	29,728	39,335	46,647	49,456	50,194	50,264	50,266	50,266
$[S_1, S_3, S_2, S_4]$	9,518	19,184	28,983	38,729	46,271	49,347	50,160	50,264	50,266	50,266
$[S_2\ S_1, S_3\ S_4]$	9,692	19,741	29,768	39,393	46,758	49,464	50,197	50,264	50,266	50,266
$[S_2, S_3, S_1, S_4]$	9,677	19,734	29,730	39,393	46,745	49,460	50,197	50,264	50,266	50,266
$[S_3, S_1, S_2, S_4]$	6,448	16,937	27,276	37,429	45,697	49,222	50,158	50,264	50,266	50,266
$[S_3, S_2, S_1, S_4]$	6,601	16,964	27,329	37,429	45,709	49,230	50,159	50,264	50,266	50,266

Long non-coding RNA (lncRNA) screening problem

Performance of the jointly optimized HTVS pipeline ($\alpha = 0.5$)

Configuration	Potential candidates	Total cost (ms)	Effective cost	Computational savings	Accuracy	Sensitivity	Specificity	F1
$[S_4]$	$50,266$	$261,374,090$	$5,200$	0%	0.8440	0.9264	0.7936	0.8186
$[S_1, S_4]$	$48,875$	$161,357,081$	$3,301$	36.52%	0.8429	0.9075	0.8034	0.8144
$[S_2, S_4]$	$47,950$	$134,366,143$	$2,802$	46.12%	0.8624	0.9215	0.8262	0.8357
$[S_3, S_4]$	$47,083$	$176,963,736$	$3,758$	27.73%	0.8450	0.8876	0.8188	0.8131
$[S_1, S_2, S_4]$	$48,210$	$134,748,992$	$2,795$	46.25%	0.8600	0.9216	0.8222	0.8333
$[S_1, S_3, S_4]$	$49,100$	$168,490,516$	$3,432$	34.00%	0.8442	0.9120	0.8026	0.8164
$[S_2, S_1, S_4]$	$48,214$	$134,812,024$	$2,796$	46.23%	0.8600	0.9216	0.8222	0.8334
$[S_2, S_3, S_4]$	$48,295$	$141,710,246$	$2,934$	43.58%	0.8602	0.9230	0.8218	0.8338
$[S_3, S_1, S_4]$	$49,119$	$171,803,403$	$3,498$	32.73%	0.8444	0.9124	0.8026	0.8166
$[S_3, S_2, S_4]$	$48,326$	$146,100,080$	$3,023$	41.86%	0.8600	0.9231	0.8214	0.8336
$[S_1, S_2, S_3, S_4]$	$48,402$	$140,954,256$	$2,912$	44.00%	0.8591	0.9228	0.8200	0.8326
$[S_1, S_3, S_2, S_4]$	$48,332$	$141,229,518$	$2,922$	43.81%	0.8587	0.9215	0.8203	0.8321
$[S_2, S_1, S_3, S_4]$	$48,409$	$141,022,859$	$2,913$	43.98%	0.8591	0.9229	0.8200	0.8326
$[S_2, S_3, S_1, S_4]$	$48,414$	$141,225,328$	$2,917$	43.90%	0.8591	0.9230	0.8200	0.8327
$[S_3, S_1, S_2, S_4]$	$48,424$	$145,321,388$	$3,001$	42.29%	0.8589	0.9228	0.8197	0.8324
$[S_3, S_2, S_1, S_4]$	$48,429$	$145,388,626$	$3,002$	42.27%	0.8589	0.9229	0.8197	0.8325

Concluding remarks

- We present two computational frameworks optimizing the performance of HTVS pipelines involving surrogate models with different complexity.

- The key idea is to estimate the joint distribution of scores computed at different stages of the pipeline, based on which the screening thresholds are optimized to maximize the throughput while minimizing the computational costs.

Concluding remarks

- We first consider the case where the computational budget is fixed, and the goal is to maximize the throughput within the given budget.

- Next, we consider the case where we aim to maximize the throughput of the HTVS pipeline while minimizing the overall computational costs at the same time.

Concluding remarks

- We demonstrated the performance of the proposed optimization schemes based on both synthetic and real-world pipeline data. We formed a high-throughput virtual screening (HTVS) pipeline for screening long non-coding RNAs (lncRNAs) by integrating various lncRNA prediction algorithms with different accuracy and computational costs. We showed that our proposed optimization frameworks can lead to significant computational savings at identical (or comparable) screening throughput/accuracy.

Thank you for watching!

Speaker (Hyun-Myung Woo) e-mail address: larcwind@tamu.edu

Manuscript: https://arxiv.org/abs/2109.11683

TEXAS A&M UNIVERSITY