Precision Medicine 1

10-742

Carl Kingsford, October 31, 2024

About Me

- Herbert A. Simon Professor of Computer Science, SCS
- Co-Director, Ph.D. program in Computational Biology
- Chair, NIH BDMA grant review study section (2023-)
- Co-founder, and current CEO, Ocean Genomics, Inc.
- All work including and since Ph.D. has been aimed at the question “How can computers advance a science like biology?”
    - Focused on molecular data.

Roadmap

- Precision Medicine (based on molecular data) overview
- Example PM Use Cases
- Big Challenges in AI for Molecular Biology
- Molecular Knowledge Graphs as one abstraction
- Ocean Genomics (Pittsburgh startup) as an example of PM in this space
- Q&A about commercialization, PM, etc. (aiming for >20 mins)
- Next lecture: more in-depth discussion of some genomic-based PM algorithms.

Drug Discovery and Development is Long, Expensive and High Risk

- Decisions are often made based on hypotheses
- The relevant biology under explored
- Targets and their environment are often not adequately understood
- Too many early trials fail – should be confirmatory
- Expensive late-stage failures are too common

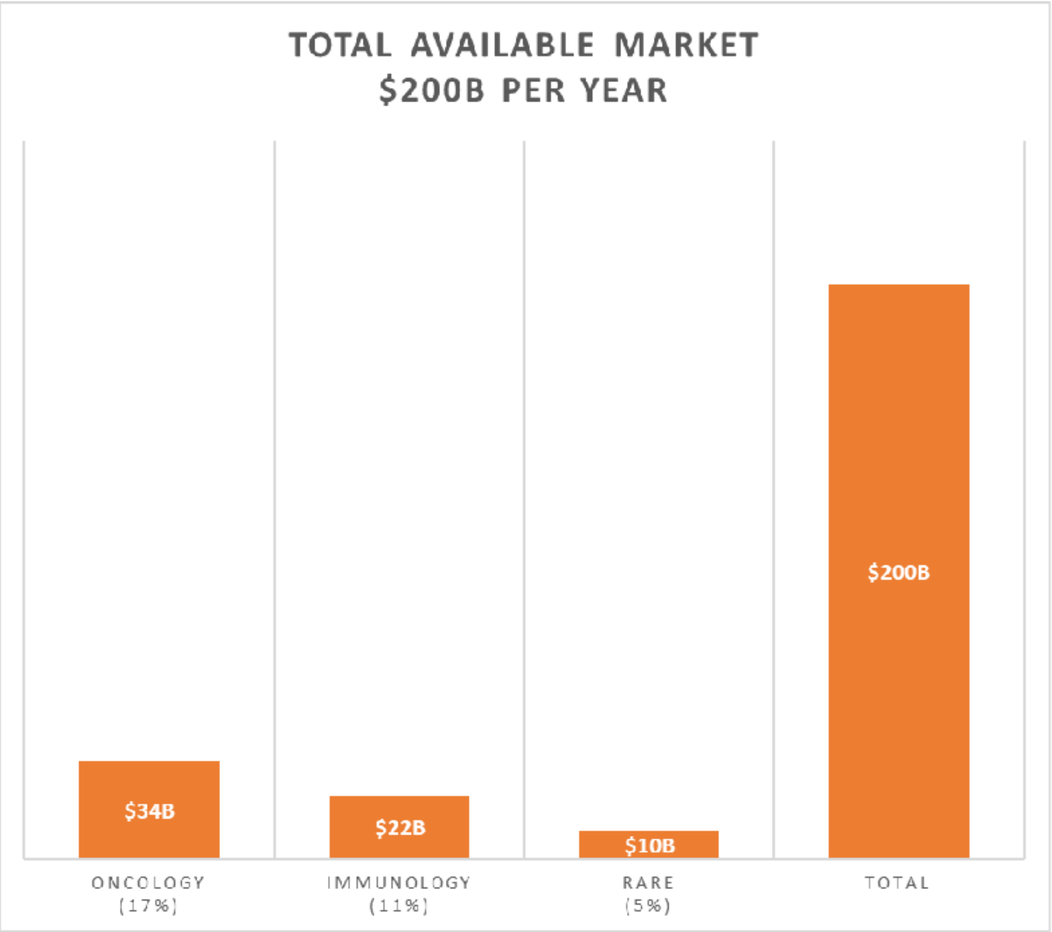
Evidence and Data-driven Insights and Decisioning are Needed at Every Step

---

To provide confidence in the underlying biology and improve probability of success

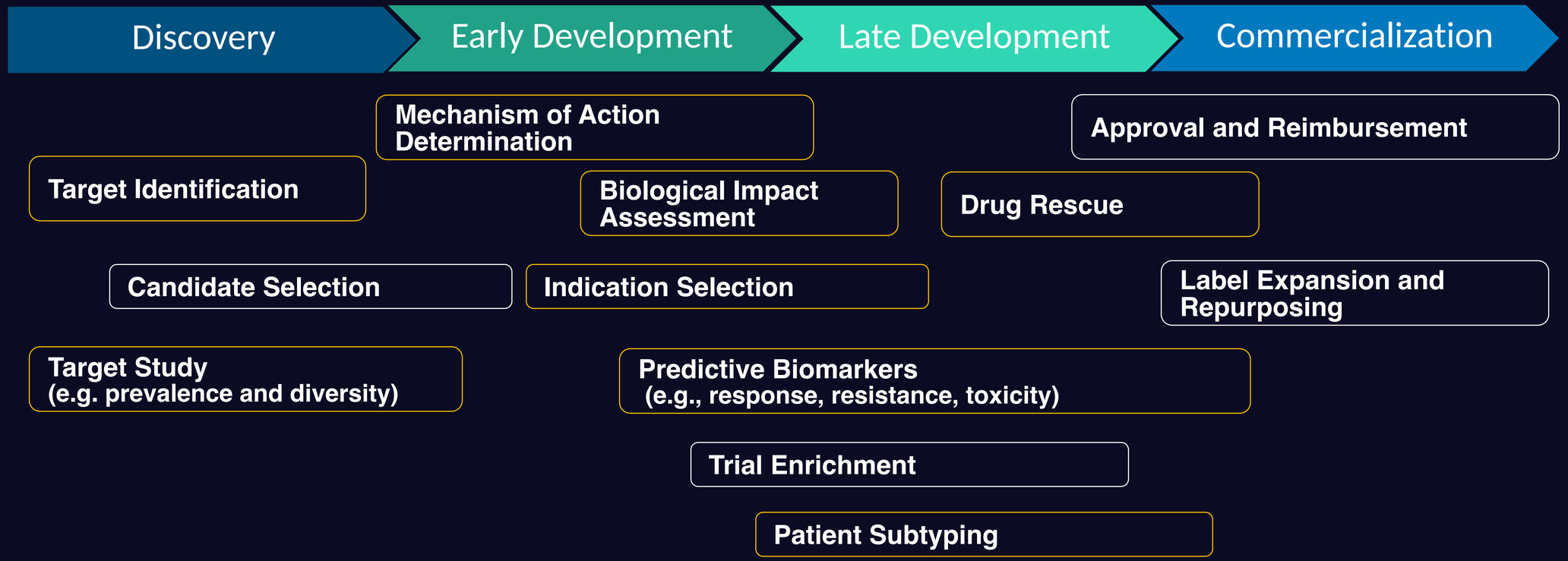
Artificial Intelligence (AI) is (slowly) Transforming Drug Development

- **AI is making biology "computable"** fueled by rapidly increasing data and computing power
- New "TechBio" companies are building AI platforms and disrupting drug discovery and development (e.g.)
  - Recursion Pharmaceuticals (RXRX) market cap $3.0B
  - Exscientia (EXAI) market cap $2.5B
  - Schrödinger (SDGR) market cap $2.7B
  - Insitro (private) valuation $2.5B post series-C - $400M round
  - Atomwise (private) valuation $423M post series - $123M round
- Each of these platforms leverage specific forms of data with specific forms of artificial intelligence
- The Total Available Market (TAM) is $200B per year

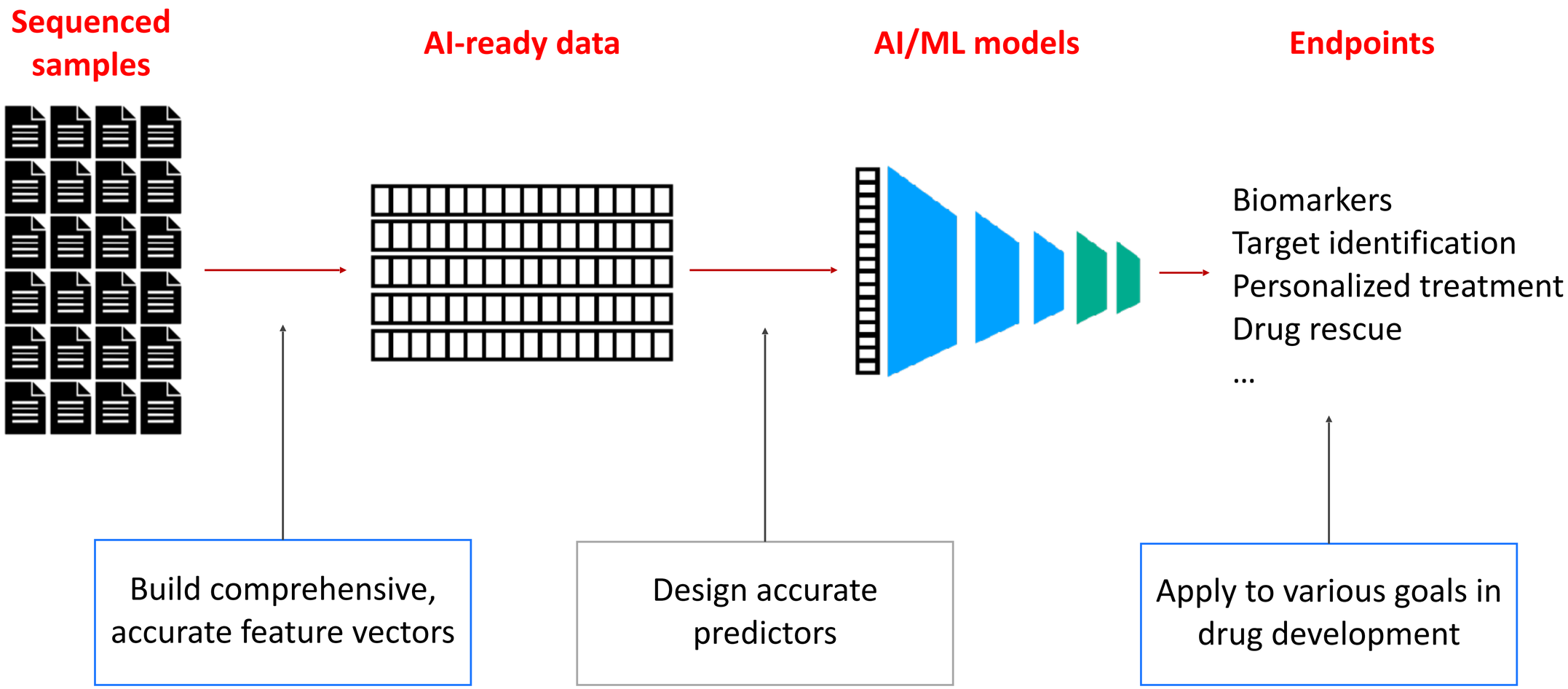


* Based on 20% available royalties assuming 100 new approvals per year with $500M average revenue per year, and 20-year patient life (source Evaluate Pharma, Bain)

Various Challenges Over the Drug Development Pipeline



Pipeline to build transcription-based bio-predictors raises many computational challenges



Example Precision Medicine Use Cases

Successful Programs Using Our Intelligent Transcriptome to Explore Complex Clinical Trial Data with Tier-1 Academic Partners

Novel Gene Targets in Pembrolizumab-Resistant Gastric Cancer

Dr. Jeeyun Lee,
Samsung Medical Center

"This is the first study to identify novel targets in pembrolizumab-resistant GC using RNA-seq algorithms beyond PDL-1."

SAMSUNG
SAMSUNG MEDICAL CENTER

ASCO 20 Virtual

Predictive Marker and Novel Gene Targets in BRAF V600E mCRC

Dr. Takayuki Yoshino, BRAVERY Trial, Japan

"This is the first finding for a potential biomarker in [pts with BRAF V600E mt mCRC] using RNA-seq analysis tools."

National Cancer Center Japan

2021 ASCO ANNUAL MEETING

Biomarker of Response to FOLFIRINOX in PDAC

COMPASS Trial Data

Identified 5 genes highly associated with response among thousands of molecular features.

2023 AACR ANNUAL MEETING Poster

Molecular Predictors and Immunomodulatory Role of Ipi/Nivo Blockade in SCLC

Yale / Illumina

Molecular predictor and biological insight into dual checkpoint inhibitor blockade in advanced small-cell lung cancer in novel clinical trial.

illumina

Yale CANCER CENTER
A Comprehensive Cancer Center Designated by the National Cancer Institute

2023 ASCO ANNUAL MEETING Poster

Case Study #1:
Identified Novel Gene Targets in ICI-Resistant Gastric Cancer

Capability: Find genes indicative of tumors that develop resistance as potential druggable targets.

- Samsung Medical Center is one of the world's most advanced cancer centers and regularly uses RNA-seq clinically to help direct treatment
- Dr. Lee and her team have shown that single agent pembrolizumab provides durable response in MSI and EBV gastric cancer
- But most MSS patients progress after initial response
- Partnered to find novel gene targets to improve outcomes in ICI-resistant GC patients

"This is the first study to identify novel targets in pembrolizumab-resistant GC using RNA-seq algorithms beyond PDL-1. "

Identified initial gene candidates using our **txome.ai** platform within 18 hours

ASCO 20 Virtual

Novel target discovery in pembrolizumab-resistant gastric cancer using a comprehensive RNA-seq analysis pipeline

Jeeyun Lee and others, including the Ocean Genomics team

Source: https://meetinglibrary.asco.org/record/186530/abstract

Case Study #2:
Identified Predictive Marker and Novel Gene Targets in mCRC

Capability: Identify patients who will respond to specific treatments (biomarkers).

- BRAVERY is a multicenter PHASE II trial including the leading cancer centers in Japan to study eribulin in patients with BRAF V600E mutant metastatic colorectal cancer (mCRC)
- These patients have a poor prognosis when treated with encorafenib plus cetuximab and limited response with eribulin
- National Cancer Center Hospital East (NCCHE) and Ocean Genomics partnered to identify novel biomarkers to predict which patients would respond to eribulin

"This is the first finding for a potential biomarker in [pts with BRAF V600E mt mCRC] using RNA-seq analysis tools."

Identified potential predictive marker (10 genes) among thousands of molecular features within 24 hours

2021 ASCO ANNUAL MEETING

Discovery of a potential predictive marker for eribulin treatment and novel target genes in BRAF V600E mutant metastatic colorectal cancer using an AI-driven RNA-seq analysis platform: Translational research of the BRAVERY study (EPOC1701)

Toshiki Masuishi, Hiroya Taniguchi,Takayuki Yoshino and others including Ocean Genomics team

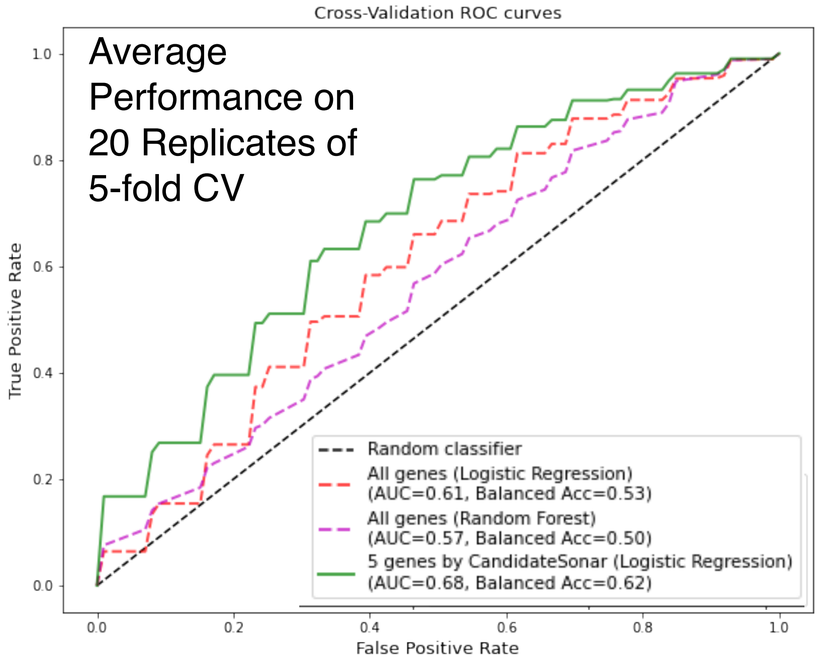
Source: https://meetinglibrary.asco.org/record/198553/abstract

Case Study #3:
Biomarker of Response to FOLFIRINOX in PDAC

Capability: Identify key genes related to drug response.

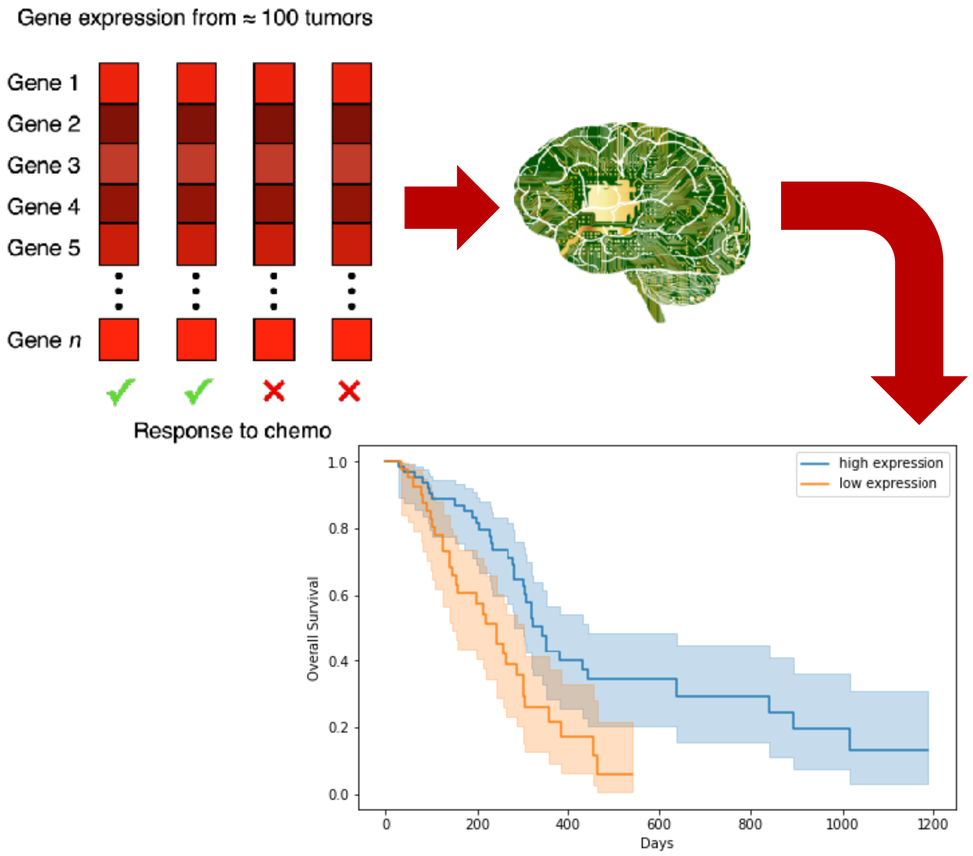
- A partner had 103 RNA-seq samples with pancreatic ductal adenocarcinoma patients who were treated with FOLFIRINOX.
- Engaged Ocean Genomics to predict positive response from gene expression and to identify genes relevant to response and survival.
- Ocean Genomics used its txome.ai pipeline, including computational feature selection, to identify a robust set of 5 genes highly associated with response.

Combining selected genes into a classifier yields improved predictive performance.



Improved biomarker for response to FOLFIRINOX treatment for pancreatic cancer

- Estimation of expression of every gene in 103 tumors.
- Rigorous, multi-round algorithm to identify genes predictive of response to FOLFIRINOX.
- Robust candidate set of 5 genes individually statistically significantly associated with response.
- Combined into a predictive model, more accurate than all-gene models.



Asghari et al. AACR (2023)

Goal

- Automatically identify potential biomarkers that are predictive for response to FOLFIRINOX (FFX) in PDAC patients
  - Enable appropriate patient selection for personalized treatment
  - Identify genes of relevance to FFX response

- Develop a binary classification model
  - Responders (CR/PR) vs. non-responders (SD/PD/NE)

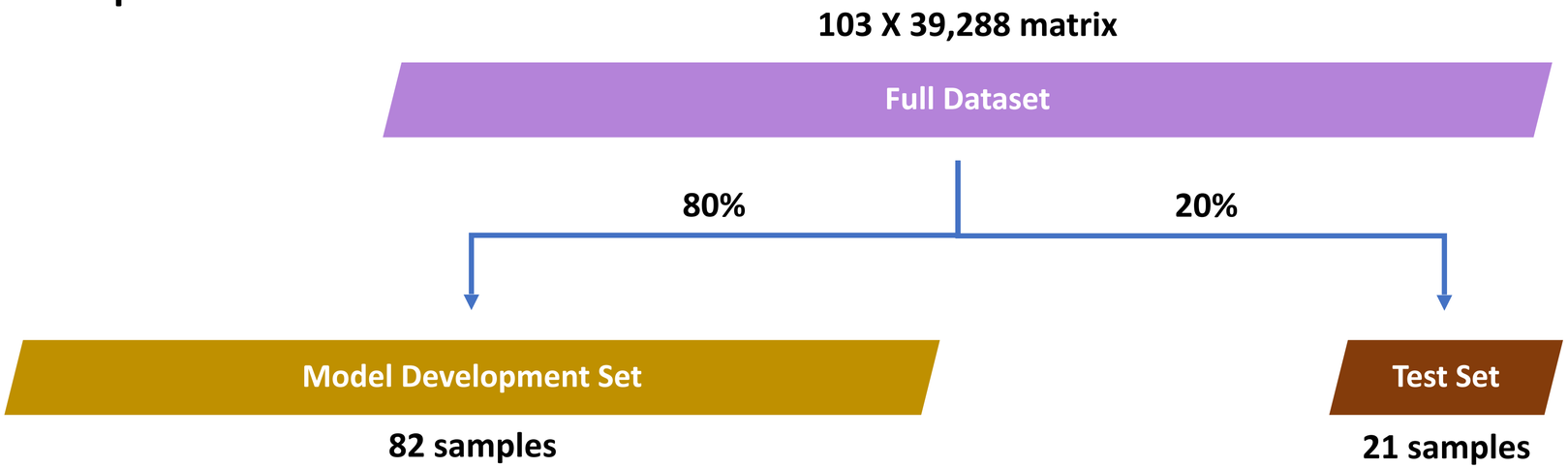
Challenge and Approach

- Standard machine learning models (e.g. Random Forest) tend to overfit on gene/transcript expression data due to high feature to sample ratio
  - > 60 k genes and > 200 k transcripts

- Need a robust computational feature selection method that is useful for building ML-based predictive models

Pre-processing

- Performed trimming on 103 RNA-Seq samples
  - Removes low quality bases
- Transcript/Gene quantification was done by **Salmon** using gencode v31 annotation and reference transcriptome
  - Gives normalized expression estimates in transcript per million (TPM)
- The input feature matrix was built from genes with non-zero expression in at least one of the samples



Model Development

- Set aside 21 samples as the held-out test set
- Model selection using 5-fold cross-validation repeated 20 times (100 folds in total)

- For each fold:
  1. Split model development set into 80% training and 20% validation set
  2. Select 5 most important genes using **CandidateSonar**
  3. Build a Logistic Regression model using the genes selected by CandidateSonar
  4. Evaluate the Logistic Regression model on the validation set and keep track of the performance for the final cross-validation metrics

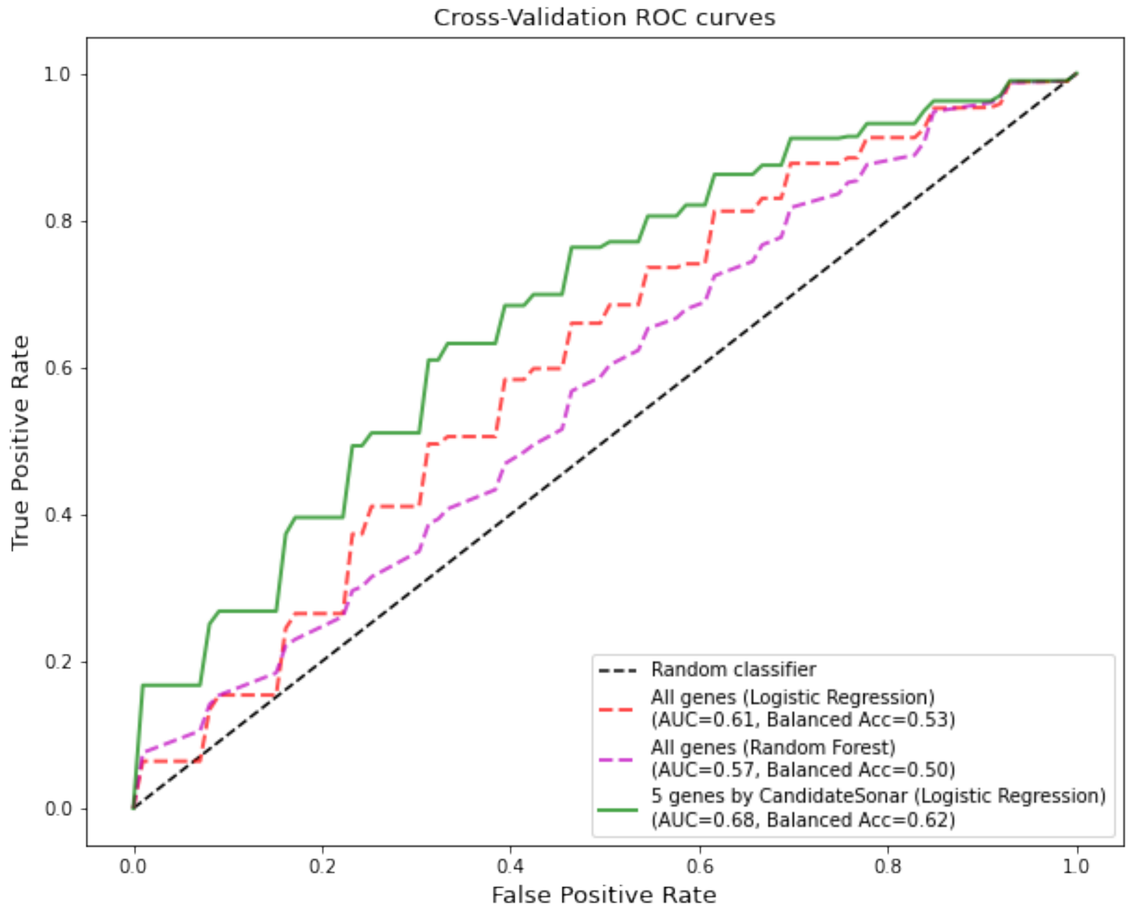
- Find the best model via hyper-parameter tuning using the above cross-validation
- Train the selected model on the whole model development set as the final model and evaluate on the test set

CandidateSonar

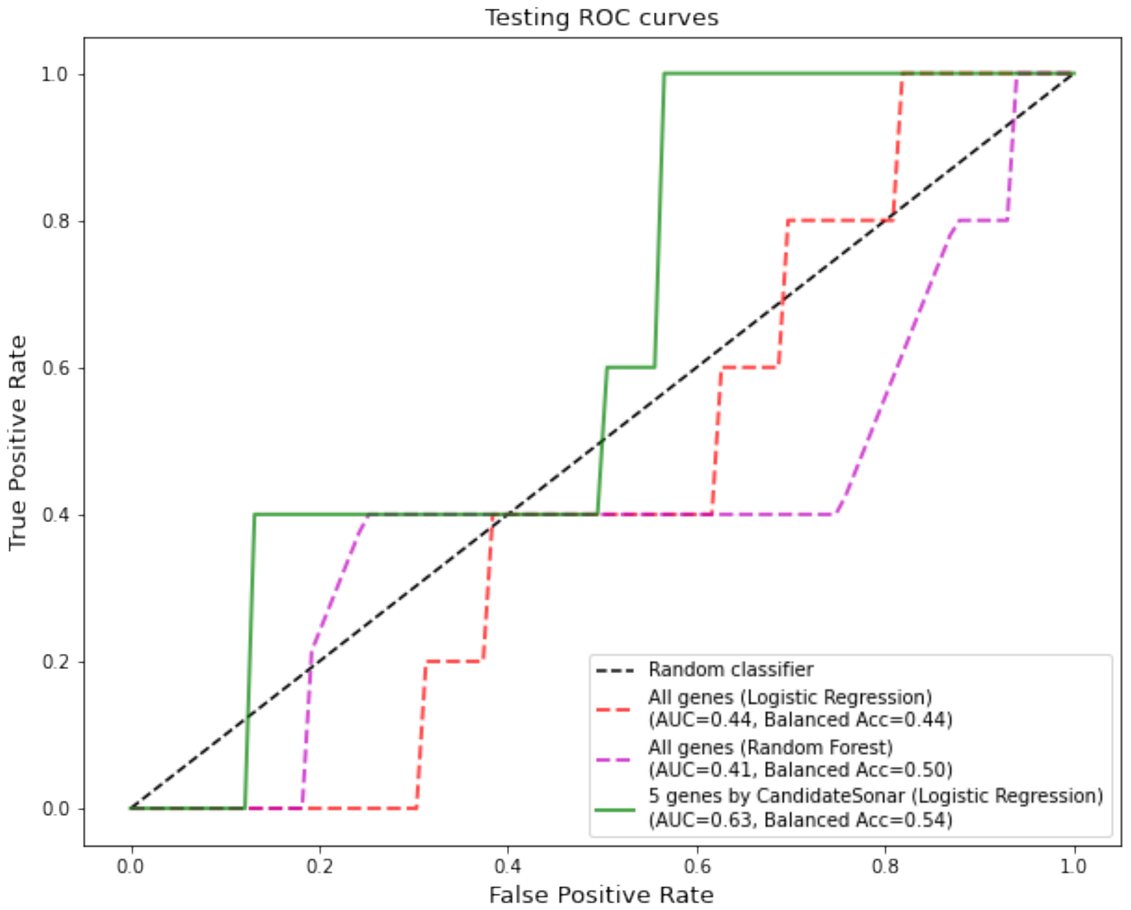
1. Perform differential gene expression (DGE) analysis between responders and non-responders
2. Genes that are differentially expressed with P-value < 0.01 and abs(Log2FC) > 0.5 were kept for further analysis
3. For each gene
   a) Find an expression cut-off using the maximal chi-squared statistic, and use the cutoff to define two groups with low and high expression of that gene
   b) Calculate the difference in response rate between low and high expression groups, as well as the statistical significance using Z-test for proportions
4. Select 30 genes with largest difference in response rate and Z-test P-value < 0.01
5. Select 5 most important genes using *permutation feature importance* based on a Logistic Regression model

Performance of Predictive Model

Average Performance on 20 Replicates of 5-fold CV



20% Held-out Test Set



Big Challenges in AI for Biology

Areas of Concern for AI in Biology

- **Federated learning** - How can multiple AIs be collaboratively trained and collaboratively make predictions when data sharing is constrained by privacy or data ownership policies?
- **Privacy & security in AI** - How can we ensure that predictive systems do not compromise the privacy and security of their users (or the data used to train them)?
- **Transfer learning & use of prior knowledge** - How can prevalent data or existing knowledge from one system improve modeling of a new system with less data?
- **Automated science** - How can experimentation be incorporated into the AI loop?
- **Generative AI** - How can large generative models and language models be exploited in biological investigations?
- **Fairness and social effects of AI** - How can we ensure that AI systems are fairly and reliably applicable across populations and situations?
- **Explainable AI and causality** - How can AI models be better used to increase human understanding via explaining their predictions and extracting causal relationships?
- **Scalability of AI in biology** - How can models be practically and efficiently trained on vast collections of data?

Federated Learning

Why it's Important

- Huge potential benefit in the clinical space, where data sharing is hard.
- Some key applications include:
  (a) protein structure prediction
  (b) drug discovery
  (c) single-cell RNA-seq modeling

Challenges & Directions

- Infrastructure is the biggest challenge: even industry efforts are stymied by the work required to set up and run such a system
- Need better incentives & cultural change for organizations to participate (e.g., participate-to-get).
- Data harmonization (biases, different distributions) and label uniformity is a big roadblock. This is especially challenging if institutions derive their data from different populations.
- Privacy guarantees in federated learning are still not sufficient.
- A technical challenge is minimizing communication while maximizing accuracy.
- A goal is to broaden what can be shared beyond models: can the underlying knowledge be shared?

Privacy & Security

Why it's Important

- Privacy is a particular challenge in the genomics space where sharing can compromise the privacy of related individuals.
- Still discovering ways that membership in training data can be revealed from models so it's hard to guarantee privacy even when just sharing trained models.

Challenges & Directions

- Improved synthetic data generation approaches can help with model development and training.
- Standards and requirements are lacking: it's not clear what the bar should be for privacy, especially for genomic data.
- Security / benefit tradeoffs are not well quantified: sharing nothing provides great security and no benefit.
- Use of blockchain or other encryption-based computation (e.g. homomorphic computation) has found some success, but not widely adopted or mature enough (and in some cases adds significant computational burden).
- Need more cross-talk between the pure security/privacy research community and the biological community.
- Reproducibility of results is a challenging problem with private data.

Transfer Learning & Use of Prior Knowledge

Why it's Important

- Transfer learning is essential. Perturbation experiments are necessary for causality and other tasks and can really only be done in model organisms.
  - Rare diseases are another key use case for this type of approach — usually do not have enough data.
- Lots of existing successes using prior knowledge from predictive models to data analysis (e.g., use of reference genomes or other features of model organisms).

Challenges & Directions

- Tremendous interest emerging in TL using deep networks, including fine-tuning foundation models, and "prompt engineering" (how do you engineer the right "question" for such a system?)
- Adapting large-language models (LLMs) to biomedical questions is an open and attractive direction.
- Adapting LLMs to molecular data (DNA, RNA, etc.) is also an emerging and active area of research.
- Transfer learning between data modalities is another important direction.
- Metrics for the "relatedness" of problems are also fairly open: how do we know (or measure) when a model trained in one setting can be applied to others?
- Determining the best way to integrate insights and theory from physics/chemistry/etc into ML models. e.g., learning dynamic equations in system biology.

Automated Science and Active Learning

Why it's Important

- Causality requires perturbation, which requires the right experiments to train models. This is the context of active learning applications.
- High throughput predictions are empowering – automation of experimentation increases the speed of discovery & increases reproducibility.
- Should perhaps be renamed "Augmented Science" to emphasize that the human needs to be in the loop.

Challenges & Directions

- Need better metrics to assess the success of automated science. Measuring things like novelty, transparency, cost, in addition to traditional accuracy measures.
- Need broader access to automated lab equipment.
- The question of interpretability is introduced at an even more fundamental level. Systems should be able to answer why an experiment was proposed by the system.
- Generally increasing Human-AI collaboration presents many new opportunities (reviewing manuscripts, literature search, replace repetitive tasks).
- Some debate over the proper role of AI in this setting: how much do we really want to automate?

Generative AI

Why it's Important

- A paradigm shift in the ML process: from search to generation.
- Huge successes in several domains (e.g. ChatGPT, AlphaFold).
- Molecule and protein design are examples of biomedical areas with a lot of success.

Challenges & Directions

- Uses of GenAI for new applications, e.g. LLM for biological sequences, prediction of gene expression, design of gene regulatory networks.
- A big challenge is trustworthiness and explainability: models are often confidently wrong and can't adequately give an explanation of their predictions.
- Scalability and computational resources are a big problem for the large models that are now required. Industry is really the primary source of these models since they are too costly for academics to build. Improving scalability and reducing model size are important directions.
- Hard to predict the success of GenAI efforts: approaches to estimate model power, data required, parameters / size required are needed.
- Biggest issues are not technical: politics, ethics, law — should increase the interaction between CS and those other fields.

Fairness & social effects of AI

Why it's Important

- AI has some promise to reduce biases and increase access to information and services, and to narrow the gap between low- and high-skilled people.
- Biases in data used could significantly reduce those benefits or cause harm.
- Biases in what questions are tackled with AI could also reduce the impact of AI.

Challenges & Directions

- Computational approaches for correcting biases in data is an active and important area of research
- Need additional work to understand how training data distribution affects biases in trained AI models.
- Approaches to select data and train models that optimize for diversity and fairness and not just accuracy are needed.
- Techniques for monitoring models over time are also needed, as data distributions can shift in the future.

Explainable AI and causality

Why it's Important

- Explainability important for collaboration —- experimentalists must trust model predictions.
- Critical from a regulatory perspective, essential to trust models enough to spend lots of effort based on their predictions.
- Some very general post-hoc explainable approaches have been developed with success: e.g. DeepLift, SHAP.
- Explainability remains hard.

Challenges & Directions

- Need better definitions of explainability and interpretability for different situations.
- Additional approaches to design neural networks that make the trained network interpretable are needed.
- Very hard to extract causality from observational data — need perturbation data to truly obtain causality and explanations.
- Need to identify problems where data + prior knowledge can be used to identify causality.

Scalability

Why it's Important

- As models get bigger and more data is used to train them, additional computational resources will be needed.
- Already expensive for non-commercial organizations to train some models, raising access, equity and bias issues.

Challenges & Directions

- Optimize hardware architectures to train models faster, especially tuned to models of use in biomedical sciences.
- For settings where existing foundation models do not exist, explore cross-model training: speed up training a model by using an already trained model on (very) different data.
- Measurement of and reduction to environmental impact (e.g. carbon emissions) from computational resource usage during training.
- Develop new techniques to reduce the size of the models.

Molecular Knowledge Graphs

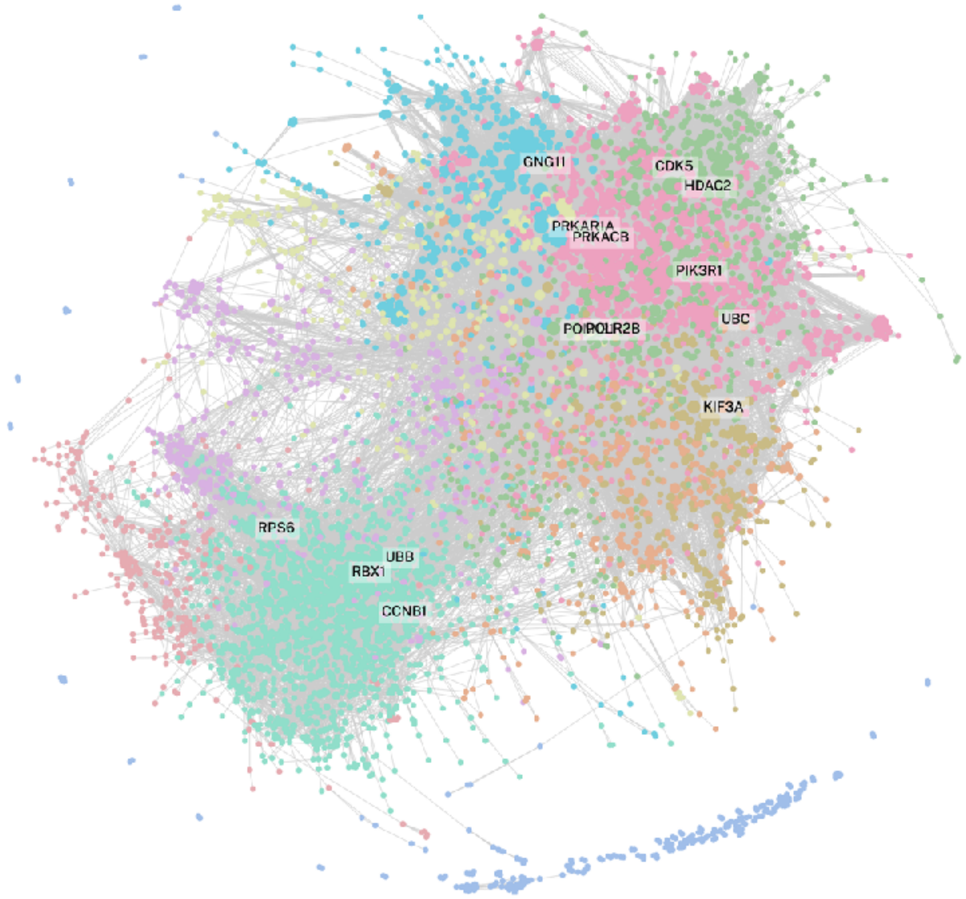
Molecular knowledge graphs will improve target identification

Knowledge graphs provide a way to encode and integrate diverse data types

Molecular knowledge graphs can connect “human” knowledge with inferred biology.

Integrates known biology (literature, databases) with inferred relationships from data.

Provides a framework for inferring new relationships between diseases, treatments, and transcriptomics.



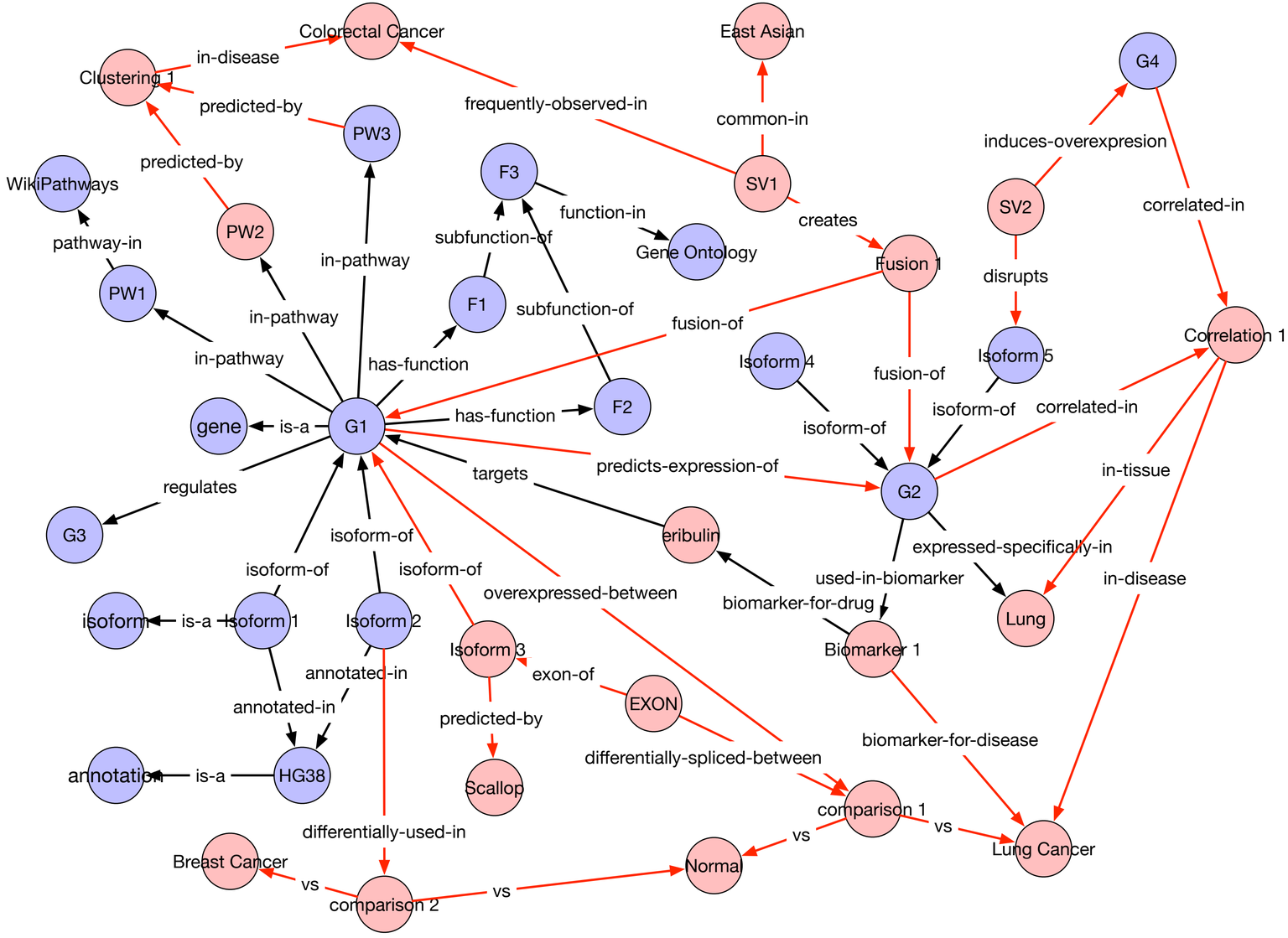
Transcriptomic knowledge graphs will improve target identification

Example schema of molecular knowledge graph.

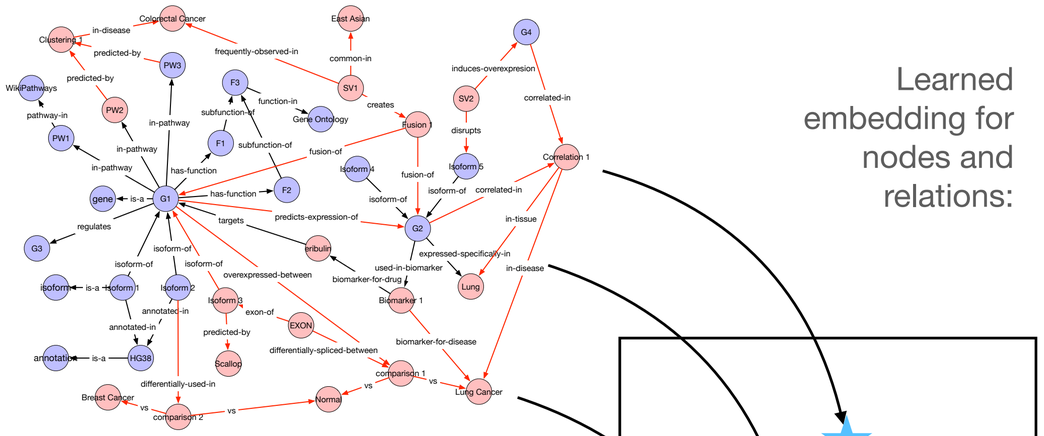
- Databases / literature
- AI / ML predicted

Distills and integrates insights from known biology and connects them to other databases and literature.

Deep learning to embed graph & predict links.

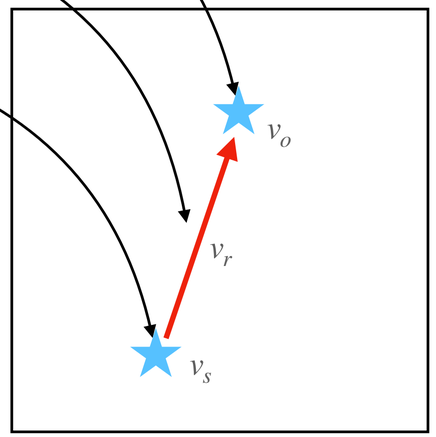


Expanded loss functions for embeddings can improve link prediction



Knowledge Graphs

Learned embedding for nodes and relations:



Low-dimensional Embedding

Relationships (edge labels) should move related entities (nodes) close together

$$\min_{\{v\}} \sum_{(s,r,o)\in\mathscr{C}} \left(w_{sro} \, || \, v_s + v_r - v_o \, || \right) + \sum_{(r_1,r_2)\in\mathscr{R}} \left(w_{r_1,r_2} \, || \, v_{r_1} - v_{r_2} \, || \right)$$

Weighting of facts based on confidence and relevance (< 0 for known falsehoods)

Similar relationships (edge labels) should have similar embeddings — can account for similarity between diseases, conditions, drugs, etc.)

Context-specific subset of facts specializes graph to particular data types, evidence, conditions, etc.

Example Startup: Ocean Genomics

Ocean Genomics

We are the Transcriptomics AI Company

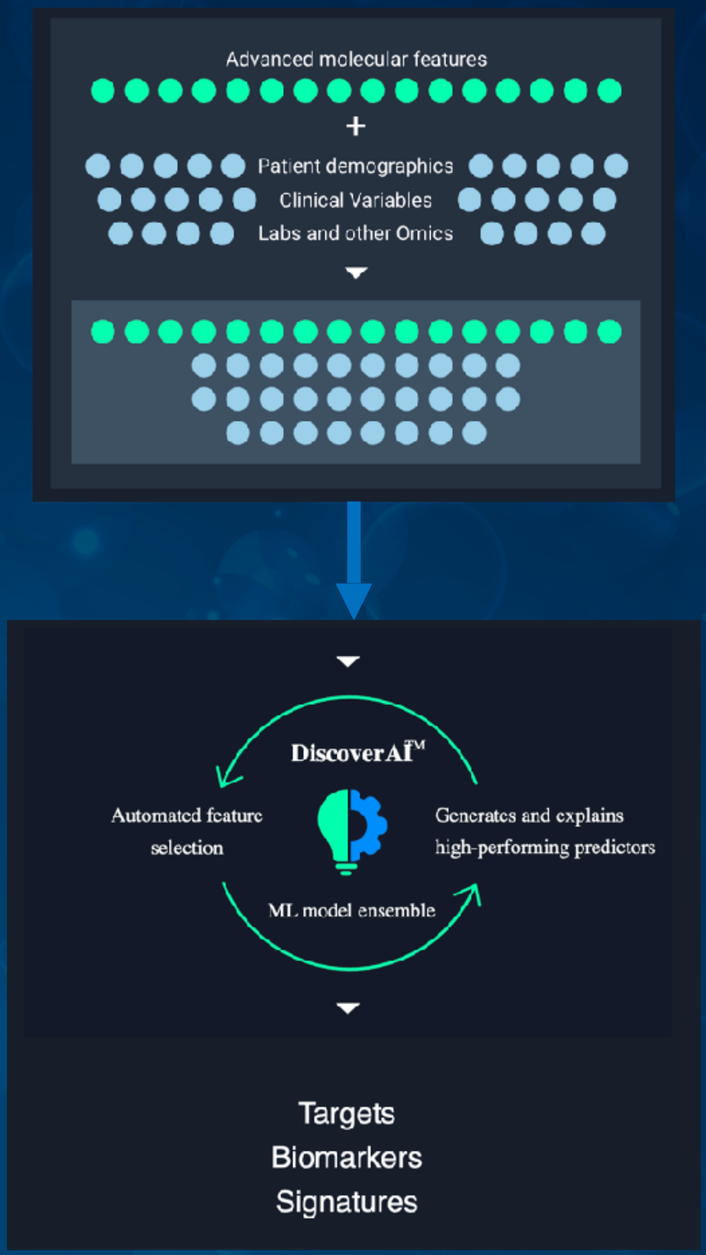
Our Intelligent Transcriptome AI platform brings evidence, and data-driven insights and decisioning to drug discovery and development.

We develop software, data, and AI models that enable advanced use of transcriptomics and related data.

Transcriptomics provides the best representation of cell function and is critical in discovery and development

Our Intelligent Transcriptome provides the essential elements required for success:

- sophisticated ML/AI for analysis to harness the size and complexity of this data
- large amounts of compute-ready data for discovery and to augment experiments – and the ability to rapidly create more
- powerful environments and tools integration for creating and maintaining analyses



An example AI application is talking with your transcriptomic data

How many structural variants are present in liver cancer samples? Same question for melanoma? List the disease and count.

index	disease	COUNT
0	hepatocellular carcinoma	1304
1	melanoma	475

What are the 3 most common fusions in the bladder cancer samples?

index	gene1	gene2	COUNT
0	UPK3B	SPDYE17	32
1	UPK3B	AC007000.4	31
2	AF254983.1	BAGE2	18

How many samples in each indication in the database have a fusion where one of the fusion partners is TP53?

index	disease	COUNT
0	gastric cancer	2
1	sarcoma	1
2	bladder cancer	1

How many UPK3B fusions have one gene expressed at > 10 tpm?

index	count(1)
0	149

Are any genes known to be associated with poor survival expressed highly (tpm > 200) in the melanoma samples?

index	gene_symbol	COUNT
0	NPIPP1	2
1	GLUL	8
2	SNRPA1	1
3	RPS9	8
4	COX6B1	8
5	RPS6	10